

Department of Environment and Agriculture

**Developing and Applying Methodologies to Characterise
Biodiversity Using Ancient and Degraded DNA**

Dáithí Conall Murray

**This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University**

June 2016

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

The research presented and reported in Chapter Two and Chapter Five was conducted in compliance with the National Health and Medical Research Council Australian code for the care and use of animals for scientific purposes 8th edition (2013). The proposed research study received animal ethics approval from the Murdoch University Animal Ethics Committee (permit no. W2002/06).

I would like to acknowledge the support of the Wardandi People in the process of conducting the research outlined in Chapter Four and Chapter Six.

Signature:

Date:

Abstract

The rapid development of DNA sequencing technologies over the past decade has revolutionised the biological sciences. Accessible and affordable high-throughput DNA sequencing (HTS) platforms coupled with high-performance computing have transformed the field of genetics, with far-reaching applications across a wide variety of disciplines. This thesis explores the utility of HTS in characterising ancient and degraded DNA for environmental metabarcoding.

HTS used in tandem with universal primers represents a rapid way to ‘profile’ plant and animal DNA from complex, heterogeneous environmental samples including sediment, faecal material and herbivore middens. For the first time, this thesis applies environmental metabarcoding to study past and present ecosystems in Western Australia. Issues arising from a lack of DNA preservation, owing to the state’s hot climate, and the poor characterisation of Western Australian biodiversity are discussed. Despite these challenges, the environmental metabarcoding workflows applied herein resulted in a number of novel insights into diets, palaeoecology, archaeology and past biodiversity.

A primary goal in many ecological studies seeking to assess biodiversity is to establish quantitative estimates of species abundance. Prior to this thesis, no studies had assessed the ability of HTS to provide quantitative estimates of DNA in faecal samples. By comparing HTS results to those of quantitative PCR (qPCR), Chapter Two of this thesis demonstrates that such estimates are possible. However, careful attention must be paid to sample screening in terms of inhibition and DNA copy number prior to sequencing, regardless of any downstream analysis of HTS data, be it quantitative or qualitative. This screening strategy forms the fundamental basis of all workflows in this thesis and is explored further in Chapter Five.

Chapter three makes the transition from modern to ancient DNA (aDNA) and seeks to define the limits of preservation using Holocene and Pleistocene-aged herbivore middens. Midden material from hot, arid environments of Australia and South Africa

was found to be a valuable source of ancient plant and animal DNA. Chapter Three explores a number of considerations for the characterisation of ancient and degraded DNA from environments not typically thought of as conducive to long-term DNA survival.

The development of a new bulk-bone metabarcoding (BBM) methodology in Chapter Four of this thesis further illustrates the potential of environmental metabarcoding to profile former ecosystems. Fragmented bone is common at both archaeological and palaeontological excavations, however, due to a lack of diagnostic morphological features, it is rarely used in taxonomic analyses. The BBM method, first developed as part of this thesis research, makes use of otherwise overlooked fragmented fossil bone and provides a fast and cost-effective means to assess DNA preservation and taxonomic biodiversity at archaeological and palaeontological sites.

The extraction and characterisation of ancient plant DNA from cave sediments, alongside that of animal DNA using BBM, enabled a detailed molecular profile of several cave sites across southwest Australia. This exploratory study, conducted within one of the world's recognised biodiversity hotspots, made use of methods and considerations highlighted across all manuscripts in this thesis to assess the suitability of environmental metabarcoding studies in Australian archaeology. The work presented in Chapter Six describes a number of insights into the interactions of people, flora and fauna over the past 50,000 years in southwest Australia.

Together, the manuscripts within this thesis raise a series of universal considerations when embarking upon environmental metabarcoding studies, especially those using degraded DNA. They emphasise a need for careful attention to be paid during all stages of the environmental metabarcoding workflow from sample collection and screening through to data generation and analysis – these considerations are the focus of Chapter Five. The use of environmental DNA (eDNA), like that of aDNA, carries with it a set of unique challenges and, in this thesis, these limitations are critically addressed. Despite the methodical and analytical challenges of conducting environmental metabarcoding in warm environments, the research presented in this thesis demonstrates the future prospects of these methods across a wide variety of applications.

Table of Contents

List of Figures	xi
List of Tables	xiv
Acknowledgements	xv
Preamble	xvi

Chapter One – Introduction

1.1 Preface	1
1.2 The DNA sequencing revolution	2
1.2.1 High-throughput sequencing platforms	2
1.2.2 High-throughput sequencing strategies	5
1.3. High-throughput sequencing applications	9
1.3.1 Characterising ancient and environmental DNA	9
1.3.2 Challenges associated with aDNA and eDNA	13
1.4 Characterising a biodiversity hotspot	17
1.4.1 Southwest Australian biodiversity	17
1.4.2 Threats to southwest Australian biodiversity	18
1.4.3 Environmental metabarcoding southwest Australia	19
1.5 References	22
1.6 Synopsis: the aim and scope of this thesis	41

Chapter Two – A comparison of qPCR and HTS for diet assessment using modern faecal material.

2.1 Preface	43
2.1.1 Statement of contribution	44
2.2 DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches.	45
2.2.1 Abstract	45
2.2.2 Introduction	46
2.2.3 Materials and methods	49
2.2.3.1 Sample collection & storage	49

2.2.3.2 Sample preparation and DNA extraction	50
2.2.3.3 Sample screening and initial quantification	50
2.2.3.4 Cloning of amplified DNA	50
2.2.3.5 HTS library preparation.....	51
2.2.3.6 GS-Junior set-up and sequencing	51
2.2.3.7 Four fish qPCR assay	51
2.2.3.8. Data analysis.....	53
2.2.4 Results & Discussion	54
2.2.4.1 Overview and comparisons of Cloning and HTS approaches.....	54
2.2.4.2 Overview of qPCR approach.....	57
2.2.4.3 Comparison of HTS and qPCR approaches	58
2.2.4.4 Recommendation for future experimental design	62
2.2.5 Conclusion	64
2.2.6 Acknowledgements	65
2.2.7 References.....	65
2.2.8 Supplementary Information	73
2.3 Synopsis	75

Chapter Three – Herbivore middens as a source of palaeoecological and palaeogenetic data

3.1 Preface	77
3.1.1 Statement of Contribution	78
3.2 High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens.....	79
3.2.1 Abstract	79
3.2.2 Introduction	80
3.2.3 Collection sites.....	84
3.2.3.1. Truitjes Kraal, RSA (TK).....	85
3.2.3.2 Brockman Ridge, WA (BR)	85
3.2.3.3 Young Range, WA (YR)	86
3.2.3.4 Cavenagh Range, WA (CR)	86
3.2.4 Materials and Methods	86
3.2.4.1 Background to midden samples.....	87
3.2.4.2 DNA extraction and screening	87

3.2.4.3 DNA Sequencing.....	88
3.2.4.4 Data analysis.....	89
3.2.5 Results and Discussion	90
3.2.5.1 Overview of sequencing data	90
3.2.5.2 Site-specific analysis	93
Cavenagh Range.....	93
Young Range.....	94
Brockman Ridge.....	95
Truitjes Kraal.....	96
3.2.5.3 Limitations of study.....	96
3.2.5.4 Future considerations.....	100
3.2.6 Conclusion	102
3.2.7 Acknowledgements	103
3.2.8 References.....	104
3.2.9 Supplementary Information	115
3.3 Synopsis	120
Chapter Four – A novel method to analyse archaeological wastes material	
4.1 Preface	122
4.1.1 Statement of Contribution	123
4.2 Scrapheap Challenge: a novel bulk-bone metabarcoding method	
to investigate ancient DNA in faunal assemblages	124
4.2.1 Abstract	124
4.2.2 Introduction	125
4.2.3 Methods	127
4.2.3.1 Sample collection and processing	127
4.2.3.2 DNA extraction and screening	128
4.2.3.3 DNA sequencing	129
4.2.3.4 Sequence identification	130
4.2.3.5 Genetic biodiversity analysis.....	131
4.2.4 Results.....	132
4.2.4.1 Overview of data generated.....	132
4.2.4.2 Taxonomic identification.....	133
4.2.4.3 Genetic biodiversity analysis.....	135
4.2.5 Discussion	138

4.2.6 Acknowledgements	142
4.2.7 References.....	142
4.3 Synopsis	149
Chapter Five – The pitfalls of HTS and potential suggestions for how to address them	
5.1 Preface	150
5.1.1 Statement of Contribution	150
5.2 From benchtop to desktop: important considerations when designing amplicon sequencing workflows.....	151
5.2.1 Abstract	151
5.2.2 Introduction	152
5.2.3 Materials and Methods	154
5.2.3.1 General methods	154
DNA extraction and screening	154
Amplicon generation and sequencing	155
Data analysis.....	156
5.2.3.2 Specific methodologies	157
Experiment 1: Importance of sample screening	157
Experiment 2: Assessing the amplicon target region.	157
Experiment 3: Importance of experimental controls	159
Experiment 4: Library generation efficiency	159
Experiment 5: Analysis parameters and their impact.....	160
5.2.4 Results and Discussion	162
5.2.4.1 Experiment 1: Importance of sample screening	162
5.2.4.2 Experiment 2: Assessing the amplicon target region.	165
5.2.4.3 Experiment 3: Importance of experimental controls	168
5.2.4.4 Experiment 4: Library generation efficiency	172
5.2.4.5 Experiment 5: Analysis parameters and their impact.....	173
5.2.5 Conclusion	177
5.2.6 Acknowledgements	177
5.2.7 References.....	178
5.2.8 Supplementary Information	189
5.3 Synopsis	197

Chapter Six – Using HTS to explore past plant and animal assemblages in a biodiversity hotspot

6.1 Preface	198
6.2 Insights and challenges from combined palaeoecological reconstructions using fossils and sediment in southwest Australia.....	199
6.2.1 Abstract	199
6.2.2. Introduction	199
6.2.3 Background to sites	201
6.2.4 Materials and Methods	203
6.2.4.1 Sample collection, extraction and screening	204
Bone sampling and extraction	204
Sediment sampling and extraction.....	205
6.2.4.2 Sample screening, amplicon generation and DNA sequencing.....	205
6.2.4.3 Data analysis.....	207
6.2.5 Results and Discussion	209
6.2.5.1 Taxonomic insights from BBM and <i>sedaDNA</i>	210
Taxonomic identification of bulk-bone material.....	214
Taxonomic identification of <i>sedaDNA</i>	218
6.2.5.2 OTU analysis of bulk-bone and <i>sedaDNA</i>	220
Devil’s Lair and Tunnel Cave OTU diversity	223
Rainbow Cave and Wonijti Janga OTU diversity	226
6.2.6 Conclusion	228
6.2.7 References.....	229
6.2.8 Supplementary Information	240
6.3 Synopsis	271

Chapter Seven – General discussion and future directions of environmental metabarcoding

7.1 Preface	272
7.2. General discussion	273
7.2.1 Is quantitative HTS data possible?	273
7.2.2 Is plant DNA preserved in Australian middens?	274
7.2.3 Is bulk-bone metabarcoding feasible?	274
7.2.4 Metabarcoding workflows: is it time to change focus?.....	275
7.2.5 Is metabarcoding useful in Australian biodiversity assessment?	276

7.2.5 Is fine taxonomic resolution possible?	277
7.3 Future directions in environmental metabarcoding.....	279
7.3.1 Environmental sample handling and screening	279
7.3.2 Generation of HTS data.....	281
7.3.2 Analysis of HTS data	282
7.4 Concluding statement.....	284
7.5 References.....	284
Appendix I: Signed co-author permissions	292
Appendix II: Quaternary Science Reviews permission.....	310
Appendix III: Publications arising from PhD candidature.....	316

List of Figures

Figure 1.2.1	8
High-throughput DNA sequencing amplicon workflow.	
Figure 1.3.1	11
The potential sources of environmental DNA and the environments from which it has been reported.	
Figure 1.4.1	21
Map showing some of the regions studied and some abiotic and biotic information related to Western Australia.	
Figure 2.2.1	48
<i>Eudyptula minor</i> distribution and study site for faecal monitoring.	
Figure 2.2.2	56
Percentage contribution of identified prey items in the faecal DNA of <i>E. minor</i> .	
Figure 2.2.3	59
Comparison of HTS and qPCR methods determining the proportion of four major fish species.	
Figure 2.2.4	60
Correlation between four-fish data obtained via HTS and qPCR.	
Figure 3.2.1	84
Location of midden sites used in this study and associated information.	
Figure 4.2.1	126
Bulk-bone fragments ground to form a bulk-bone powder at two archaeological sites.	
Figure 4.2.2	134
Taxa identified in bulk-bone powder samples.	

Figure 4.2.3	135
DTUs shared across bulk-bone powder samples.	
Figure 4.2.4	136
Change in DTU number and composition over time at Tunnel Cave and Devil's Lair.	
Figure 4.2.5	137
Change in Macropodidae DTU number over time at Tunnel Cave and Devil's Lair.	
Figure 5.2.1	161
Definitions used in assessing the importance of analysis parameters.	
Figure 5.2.2	163
Quantitative PCR and sequencing results of the sample screening assay.	
Figure 5.2.3	166
Average sequencing error rates across a single amplicon region.	
Figure 5.2.4	175
Impact of analysis parameters on the numbers of taxonomic units obtained for a bulk-bone sample.	
Figure 6.2.1	203
Location of southwest Australian cave sites used in this study.	
Figure 6.2.2	212
Cladograms showing faunal diversity identified across Devil's Lair, Tunnel Cave, Rainbow Cave and Wonitji Janga.	
Figure 6.2.3	221
OTU number and diversity change over time at Devil's Lair and Tunnel Cave.	
Figure 6.2.4	222
OTU number and diversity change over time at Rainbow Cave, Wonitji Janga and Northcote Sinkhole.	

Figure 6.2.5	224
Clustering of Devil's Lair and Tunnel Cave bulk-bone and sediment samples according to LGM boundaries.	
Figure 6.2.6	226
Clustering of Tunnel Cave bulk-bone samples according to occupation and non-occupation layers.	
Figure 6.2.7	228
Clustering of Wonitji Janga and Northcote Sinkhole bulk-bone samples.	

List of Tables

Table 1.2.1	3
Comparison of Sanger and high-throughput sequencing platforms.	
Table 2.2.1	52
List of primer pairs used in this study.	
Table 3.2.1	91
Plant families identified in the midden samples using <i>trnL</i> plastid primers.	
Table 3.2.2	92
Mammalian taxa identified in midden samples using 16S and 12S rRNA primer sets.	
Table 5.2.1	153
Details for the experiments conducted.	
Table 6.2.1A-D	213
Presence and absence of select plant families detected at cave sites for <i>trnL</i> and <i>rbcl</i>	

Acknowledgements

Throughout my Ph.D. candidature, I have had the pleasure of meeting and working alongside some exceptional people from whom I have learnt a great deal. I would like to offer my deepest gratitude to all those who have guided me along the path to thesis submission. In particular, I would like to acknowledge the support of all members of the Trace and Environmental DNA (TrEnD) Lab at Curtin University, past and present, who have taught me a series of invaluable lessons along the way.

I would like to thank my primary supervisor, Professor Michael Bunce, who has involved me in research across a diverse range of projects within the TrEnD Lab. It has been a privilege to work under your guidance.

Finally, I would like to express the most sincerest thanks to Jesse Sounness who supported me throughout my entire studies and without whose encouragement I would probably never have decided to embark on a Ph.D. or have completed it.

It is a conscious decision not to mention anybody else by name but to all those I have met along the way please know that your advice and help have been very much appreciated. I hope some of you at least, like myself, managed to take something useful away from the experience – if even a little giggle.

Preamble

The thesis presented consists of an introductory literature review (Chapter One), four manuscripts published in scientific journals (Chapters Two–Five), one manuscript currently in the advanced stages of preparation for submission (Chapter Six) and a final general discussion with recommendations for future research (Chapter Seven). Manuscripts already in the scientific domain have been reproduced “as published” with minor exceptions to maintain consistency in formatting and allow cross-referencing throughout the thesis. For the purposes of continuity and flow, each manuscript is flanked by a preface introducing the work and a synopsis summarising the findings and how they inform the subsequent chapters. All manuscripts contain an abstract, introduction and methods section; however, due to specific journal requirements the results and discussion sections have been merged into a single section in some cases. As this is largely a “thesis by publication” every manuscript consists of a self-contained introduction and discussion. Therefore, to minimise repetition across the introduction (Chapter One) and discussion (Chapter Seven), cross-referencing of the manuscripts presented has been used throughout the chapters.

Due to the multi-disciplinary nature of this thesis, it was necessary to foster collaborations across a number of scientific disciplines, including, but not limited to, molecular biology, bioinformatics and archaeology. Signed declarations of author contributions have been included for Chapters Two-Five (Appendix I) and the roles of co-authors also stated at the conclusion of all manuscript chapter prefaces. Permission from *Quaternary Science Reviews* to reproduce the manuscript in Chapter Three is in Appendix II. Lastly, the title pages of published manuscripts on which I am a co-author are included in Appendix III. While these co-authored publications are not formally included in this thesis submission they speak to the entire scope of research that I undertook during the tenure of my Ph.D. candidature.

Chapter One – Introduction

1.1 Preface

The following introductory chapter (Chapter One) seeks to provide sufficient background to consider the manuscripts (Chapters Two-Six) presented and the scope of research contained within this thesis. Each manuscript presented in this thesis has its own introduction section and as such every effort has been made to avoid repetition and to direct the reader to other chapters where appropriate.

The introductory chapter has been divided into three primary sections (Sections 1.2-1.4) with a synopsis detailing the broad aim(s) and scope of the thesis and its chapters (Section 1.6). Each section of this introductory chapter revolves around themes that are relevant to the thesis manuscripts. Firstly, a brief overview of high-throughput DNA sequencing (HTS) technology and workflow considerations is provided (Section 1.2). Modern DNA sequencing technology has revolutionised the study of past and present environments — through ancient DNA (aDNA) and environmental DNA (eDNA) — but its application has also introduced fresh challenges, both of which are reflected upon in the introduction (Section 1.3). Finally, the application of HTS to study biodiversity in southwest Australia, one of only a handful of biodiversity hotspots worldwide, provides a promising tool to explore the rich regional biota. The benefits and difficulties of studying how such diversity may have changed over time and the insights this can provide for future ecological management are explored in the final section of the introduction (Section 1.4).

1.2 The DNA sequencing revolution

From the 1970's until the early 2000's Sanger sequencing involving chain-termination (Sanger & Coulson, 1975; Sanger *et al.*, 1977) remained the primary means by which the DNA code was biochemically derived (Shendure & Ji, 2008). During this period it underwent several refinements and iterations, became automated and was the chosen method for the generation of the first 3-billion base pair (bp) draft of the human genome (Lander *et al.*, 2001; Venter *et al.*, 2001; França *et al.*, 2002; Shendure & Ji, 2008). Nonetheless, DNA sequencing remained expensive and low-throughput (Table 1.2.1). The release of Roche's 454 pyrosequencing platform in 2005 (Margulies *et al.*, 2005) proved to be a disruptive force in the landscape of DNA sequencing technology and marks the beginning of an influx of high-throughput – or next-generation – DNA sequencing technologies (HTS and NGS respectively) onto the market. These new platforms made DNA sequencing affordable and, as such, accessible across a much wider range of laboratories and projects compared with previous technology (Metzker, 2010; Liu *et al.*, 2012; Tillmar *et al.*, 2013).

1.2.1 High-throughput sequencing platforms

High-throughput sequencing has been touted as a cost-effective means of genetic analysis across a range of disciplines that allows the genetic characterisation of many samples in parallel using short nucleotide barcodes assigned to each sample with a depth of coverage across samples simply not possible with previous technology (Binladen *et al.*, 2007; Shokralla *et al.*, 2012). Due to its utility across many disciplines HTS has undergone rapid development in the past decade; in this thesis alone there was a progression across three HTS platforms as sequencing technology remained in flux. Reduced cost, improved accuracy and adequate read length are a few of the major drivers of development and change within the HTS industry with both cost and accuracy the primary factors in changing HTS platforms during this thesis (Glenn, 2011; Liu *et al.*, 2012). A standardised industry metric to facilitate straightforward comparison of platforms, however, is still lacking (Fuller *et al.*, 2009).

Table 1.2.1 Comparison of Sanger and high-throughput sequencing platforms. A selection of high-throughput sequencing platforms and their sequencing method is given with output in read length and Gb, associated error and cost per million bases given for each (Glenn, 2011; Liu *et al.*, 2012; Quail *et al.*, 2012; Buermans & den Dunnen, 2014; Laver *et al.*, 2015).

Company/ Platform	Sequencing method	Max. Read length (base pairs)	Max. Output (Gb)	Error Rate (%)	Cost (USD)
Life Technologies/ Sanger 3730xl	Dideoxy chain termination	400-900	<0.001	~0.001	2400
Roche/ 454	Synthesis	700	~0.7	~1	10
Illumina/ MiSeq, HiSeq	Synthesis	2x300 (MiSeq) 2x150 (HiSeq)	~15 (MiSeq) ~600 (HiSeq)	~0.8	0.07
Life Technologies/ Ion Torrent PGM	Synthesis	500	~1-2	>1.5	1
Life Technologies/ SOLiD	Ligation	75	250	<0.5	0.13
Pacific Biosystems/ PacBio (RS II)	Single-molecule	>1000	~0.1-5 (~50k reads)	>10	0.13- 0.6
Oxford Nanopore / MinIon	Single-molecule (nanopores)	~1000	>21	>30	10

BOLD indicates platforms used in this thesis

Currently, a range of HTS systems exist (Table 1.2.1) from which researchers may choose (Mardis, 2008; Shendure & Ji, 2008; Ansorge, 2009; Tautz *et al.*, 2010; Buermans & den Dunnen, 2014) and the development of the future generation is already in advanced stages with novel protein nanopore (Stoddart *et al.*, 2009; Mikheyev & Tin, 2014) and silicon nanogap (Wang *et al.*, 2015) technologies offering glimpses of the future of HTS democratisation and consumerisation (Goodwin *et al.*, 2016). Different platforms have held the mantle of the “consumer’s preferred sequencer” during the past decade (Mikheyev & Tin, 2014). While many platforms exist, Roche’s 454 (Margulies *et al.*, 2005), Life Technologies’ Ion Torrent PGM (Rothberg *et al.*, 2011) and Illumina’s MiSeq (Bentley *et al.*, 2008) systems were used in the proceeding chapters. All three platforms employ sequencing-by-synthesis methods, in contrast to the chain termination strategy of Sanger sequencing

(França *et al.*, 2002) or the direct strand-sequencing employed in Oxford Nanopore's MinIon (Laver *et al.*, 2015).

In sequencing-by-synthesis, a complementary DNA strand is created and each nucleotide base incorporation is recorded. In both 454 and MiSeq, base incorporation is detected via fluorescence: pyrophosphate detection as dNTPs are flowed separately across the picotitre plate in 454 (Margulies *et al.*, 2005; van Dijk *et al.*, 2014) and through reversible terminator-bound dNTPs which are present simultaneously on the MiSeq flowcell (Bentley *et al.*, 2008; van Dijk *et al.*, 2014). An alternative strategy is employed on the Ion Torrent PGM whereby changes in pH are detected as protons are released upon nucleotide incorporation (Rothberg *et al.*, 2011; van Dijk *et al.*, 2014). All three platforms utilise PCR steps during library preparation: emulsion PCR in 454 and Ion Torrent PGM systems and bridge-amplification in the MiSeq (van Dijk *et al.*, 2014). There are a number of challenges associated with methods of sequencing-by-synthesis that largely fall into the categories of sample preparation (e.g. the use of clonal amplification which can inflate bias introduced in earlier stages of preparation), detection of nucleotide incorporation (e.g. stray dye signals due to adhesion to plate surfaces), sequencing accuracy (incorrect base calling in low complexity regions) and sequence phasing issues (e.g. incomplete extension of priming strand) to name a few (Fuller *et al.*, 2009). HTS platforms are reported to have an elevated error rate when compared to traditional Sanger sequencing (Liu *et al.*, 2012; Loman *et al.*, 2012; Quail *et al.*, 2012; Bragg *et al.*, 2013); although direct comparisons are problematic given that Sanger sequences are based on consensus sequences with error essentially muted through the averaging of signal intensities to produce a chromatograph. Nevertheless, a major goal in developing HTS platforms is achieving a balance between high-throughput and accuracy to ensure sufficient genetic coverage and fidelity. In the case of 454 and Ion Torrent PGM, relative to Illumina platforms, both platforms have high error rates particularly in regions of low complexity such as homopolymer stretches (Loman *et al.*, 2012; Quail *et al.*, 2012; Bragg *et al.*, 2013). However, in the case of 454, it offered some of the longest reads of all current HTS platforms (~800 bp), though as of 2016 it is no longer in production or supported. The error rates observed on Illumina platforms, compared to 454 and Ion Torrent platforms, are less influenced by homopolymer indels because nucleotide detection is performed one

base at a time (Loman *et al.*, 2012; Shokralla *et al.*, 2012). The sequencing error on Illumina is accumulative though and as such limits read length and quality in longer reads (Zhou *et al.*, 2010). The actual sequencing library preparation for MiSeq is relatively straight-forward but library concentration can impact on the distribution of sequence clusters on the flowcell and cause over- or under-clustering that can have a detrimental impact on data quality and the success of a run due to cluster overlap or a lack of cluster resolution (van Dijk *et al.*, 2014).

Despite the challenges associated with HTS, it is fast becoming an industry standard across many scientific disciplines (Buermans & den Dunnen, 2014) and methods to combat issues associated with sequencing error rate are being developed (Quince *et al.*, 2009; Coissac *et al.*, 2012). Ultimately, the choice of sequencing platform is largely dependent on the project in question and considerations include the level of sequencing coverage that might be needed in addition to whether short or long DNA fragments will be targeted. Concomitant with the decisions associated with the appropriate platform for a study is the choice of sequencing strategy that should be employed, e.g. amplicon or shotgun sequencing.

1.2.2 High-throughput sequencing strategies

There are two primary strategies to sequence DNA contained within samples: shotgun sequencing and amplicon sequencing – though targeted enrichment of DNA of interest can be overlaid on both strategies. Shotgun sequencing is essentially a non-targeted form of sequencing that was used extensively prior to the advent of HTS (Rizzi *et al.*, 2012). In shotgun sequencing the DNA within a sample is fragmented into smaller pieces post-extraction; however in some applications where the DNA is already quite fragmented, such as those using ancient and degraded DNA, this is not necessary. The shotgun sequencing approach has the capacity to sequence all the DNA within a sample and has been used to generate whole mitochondrial and nuclear genomes (Lander *et al.*, 2001; Venter *et al.*, 2001; Green *et al.*, 2006; Miller *et al.*, 2008) and to provide insights into species diversity in bacterial and viral metagenomics studies (Williamson *et al.*, 2008; Sharpton, 2014). The ability to sequence DNA in a non-targeted fashion while advantageous in some studies is a hindrance in others, especially when analysing complex, heterogeneous

samples such as faeces or sediment for taxa other than bacteria, as bacterial DNA is in much greater relative abundance than the DNA of interest.

Due to the non-targeted nature of shotgun sequencing, a second strategy is often employed – amplicon sequencing (Thomas *et al.*, 2006). This is the sole strategy used in the thesis chapters that follow as it allows the sequencing of plant and animal DNA from the substrates used in this study to the exclusion of ubiquitous microorganisms. In amplicon sequencing an appropriate gene region is chosen and primers designed to allow the specific PCR amplification of the selected region (explored further in Section 1.3). Once suitable primers have been selected a short, unique multiplex identifier (MID) DNA-based tag (alternatively called an index) is added to the 5-prime (5') end of the primer during synthesis to allow the multiplexing of samples facilitating parallel sequencing of multiple samples on a single HTS sequencing run (Binladen *et al.*, 2007; Roche, 2009). In addition to MID-tags, platform specific DNA-based adapter sequences from which the actual sequencing reaction is primed are added to amplicons. Once the target DNA region is amplified the PCR amplicons, or products, can be pooled and sequenced using the most appropriate sequencing platform and the resulting sequencing reads can be separated bioinformatically according to the sample from which they originated based on assigned MID-tags (Figure 1.2.1).

Multiple strategies may be employed to incorporate the above-mentioned MID-tags and adapter sequences but, briefly, there are two primary approaches: ligation-based approaches (e.g. Binladen *et al.*, 2007) and fusion-tagged primer based approaches (e.g. Sønstebo *et al.*, 2010; Clarke *et al.*, 2014). In the case of ligation approaches, generally, MID-tagged PCR amplicons are generated, then pooled and sequencing adapters subsequently ligated onto the products prior to sequencing. For fusion-tagged primer approaches there are two primary classes: those involving a single PCR reaction and those involving multiple rounds of PCR (Bronner *et al.*, 2001; Varley & Mitra, 2008; Bybee *et al.*, 2011; de Cárcer *et al.*, 2011; Archer *et al.*, 2012; Campo *et al.*, 2014). When using a single PCR approach the MID-tag and sequencing adapters are both present on the 5' end of the chosen primers and as such are incorporated into the amplicon during PCR amplification. Strategies involving multiple rounds of PCR tend to conduct amplification with MID-tagged primers

followed by subsequent PCR amplification to incorporate sequencing adaptors. Further exploration of these strategies and their merits are dealt with in Chapter Five.

Prior to the introduction of HTS technology, the sequencing of DNA within samples in parallel would not have been possible and instead PCR amplification of DNA within samples followed by subsequent cloning and Sanger sequencing on individual clones would have been necessary. Current HTS strategies, therefore, offer vast improvements in time, cost and efficiency through the elimination of cloning and low-throughput Sanger sequencing (Hudson, 2008; ten Bosch & Grody, 2008; van Dijk *et al.*, 2014). Alongside these advantages is the increase in the depth of sequencing and large quantities of data produced which carry a fresh set of challenges associated with the accurate screening, identification and analysis of sequences which is explored extensively throughout the following chapters in this thesis (Huson *et al.*, 2007; Quince *et al.*, 2009; Caporaso *et al.*, 2010; Hamady *et al.*, 2010; Quince *et al.*, 2011; Coissac *et al.*, 2012; Faircloth & Glenn, 2012; Gonzalez & Knight, 2012). However despite these challenges modern DNA sequencing has established itself as a necessary technology across a range of disciplines (Buermans & den Dunnen, 2014) and has revolutionised the field of genetics, none more so probably than the fields of ancient DNA (aDNA) and environmental metabarcoding (Knapp & Hofreiter, 2010).

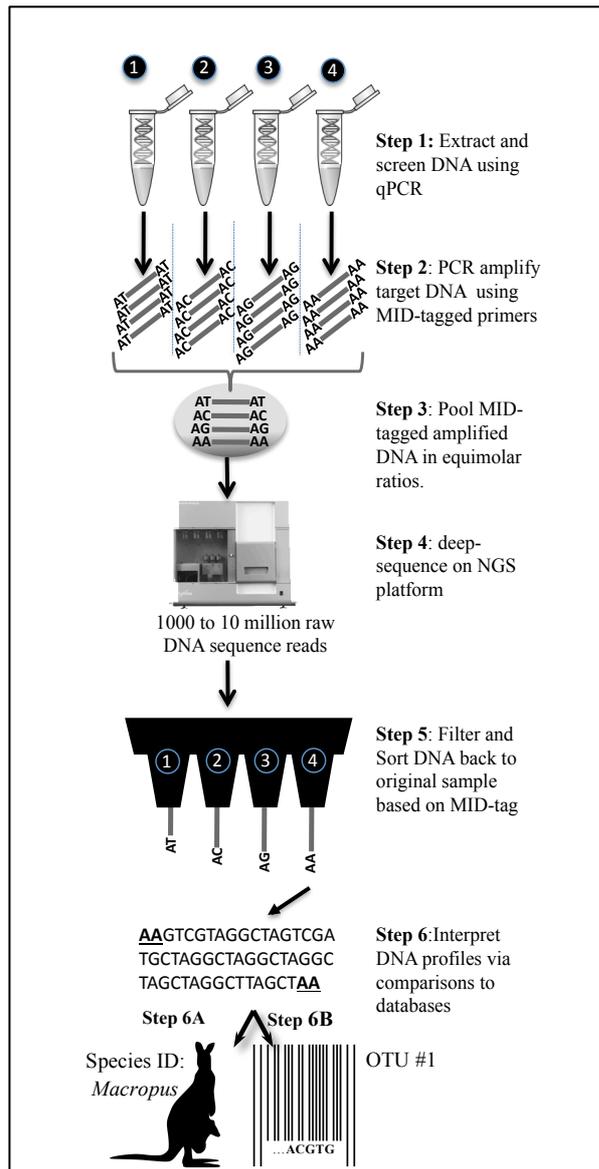


Figure 1.2.1 High-throughput DNA sequencing amplicon workflow. DNA is first extracted (**Step 1**) and the target DNA is amplified using multiplex identifier (MID) tagged primers specific to gene region of interest (**Step 2**). Tagged PCR products are pooled to form a sequencing library (**Step 3**) and sequenced (**Step 4**). Post-sequencing, reads are separated out into the samples from which they originated (**Step 5**) and the DNA profiles analysed (**Step 6**) by comparison to genetic reference databases (**Step 6A**) or by clustering sequences into operational taxonomic units (OTUs) to allow a degree of taxonomy-independent analysis when reference database are patchy (**Step 6B**).

1.3. High-throughput sequencing applications

High-throughput DNA sequencing has carved out a niche across many disciplines in the biological sciences from medicine (e.g. Roychowdhury *et al.*, 2011; Soon *et al.*, 2013) and forensics (e.g. Budowle *et al.*, 2014; Yang *et al.*, 2014) to ecology (e.g. Valentini *et al.*, 2009b; Pedersen *et al.*, 2014; Thomsen & Willerslev, 2015) and conservation (Angeloni *et al.*, 2012; Bohmann *et al.*, 2014). Moreover, the arrival of HTS had a profound impact on certain fields of research; particularly in the area of palaeogenetics where a sharp increase in the generation of ancient sequence data followed the introduction of HTS workflows (Knapp & Hofreiter, 2010; Rizzi *et al.*, 2012). Alongside the adoption of HTS in aDNA studies the technology became increasingly popular among ecologists as it offered a means by which to non-invasively, and non-lethally, study important species and environments while causing minimal impact (Soininen *et al.*, 2009; Pompanon *et al.*, 2012; Bohmann *et al.*, 2014). Most importantly, however, it has become increasingly apparent that the use of HTS in ecological and environmental studies complements, rather than supplants, traditional modes of ecological study and management (Andersen *et al.*, 2012; Jørgensen *et al.*, 2012; Yoccoz *et al.*, 2012; Parducci *et al.*, 2013; Pawlowska *et al.*, 2014).

1.3.1 Characterising ancient and environmental DNA

The characterisation of DNA extracted from both single-source (e.g. bone, hair or eggshell) and environmental samples (e.g. sediment, water or faeces) relies primarily on short sections of DNA that serve to differentiate taxa – a DNA barcode (Nanney, 1982; Hebert *et al.*, 2003). The use of DNA barcodes in the identification of taxa is well established and the barcoding community have designated the mitochondrial cytochrome oxidase 1 (*COXI*) gene and a combination of the plastid RuBisCO long chain/maturase K (*rbcl/matK*) genes for the identification of animals (Hebert *et al.*, 2003) and plants (Hollingsworth *et al.*, 2011) respectively. While DNA barcoding in the strictest sense involves the identification of single-source specimens using the approved DNA barcodes, alternative barcodes, such as 16S rRNA for animals (Deagle *et al.*, 2014) and plastid *trnL* for plants (Taberlet *et al.*, 1991; Taberlet *et al.*, 2007), are often used in cases where this is not feasible. Instances warranting the use

of alternative barcodes include those where the highly fragmented and degraded nature of endogenous DNA makes it difficult to design primers to successfully amplify short regions of *COXI* (Deagle *et al.*, 2014), for example, in the identification of samples using DNA extracted from “ancient” or historical specimens – ancient DNA (aDNA). An extension of the concept of DNA barcoding is that of analysing environmental samples to determine the species composition within a sample to explore matters pertaining to diet or biodiversity using DNA extracted from environmental samples without isolating any specific target – environmental DNA (eDNA) (Taberlet *et al.*, 2012a). The characterisation of eDNA from environmental samples necessitates the use of non-traditional DNA barcodes much like aDNA from single-source samples, again due to DNA degradation and fragmentation. Additionally, owing to the potential diversity of taxa within a sample, barcodes are often targeted using universal primers that are designed to maximise the number of taxa that can be amplified in PCR. The characterisation of eDNA from modern environmental samples using non-standard universal DNA barcodes – environmental metabarcoding (Taberlet *et al.*, 2012a) – is emerging as a promising tool to explore a range of pertinent ecological and biological questions (Valentini *et al.*, 2009a; Pompanon *et al.*, 2012; Taberlet *et al.*, 2012b; Bohmann *et al.*, 2014; Pedersen *et al.*, 2014; Thomsen & Willerslev, 2015). Environmental metabarcoding also represents a point at which the fields of aDNA and eDNA both converge and it has been applied with success to “ancient” samples and, in doing so, it has provided, at times novel, insights into past ecosystem composition and change (Hofreiter *et al.*, 2000; Kuch *et al.*, 2002; Matisoo-Smith *et al.*, 2008; Haile *et al.*, 2009; Hebsgaard *et al.*, 2009; Sønstebo *et al.*, 2010; Jørgensen *et al.*, 2011; Jørgensen *et al.*, 2012; Pedersen *et al.*, 2013; Giguet-Covex *et al.*, 2014; Willerslev *et al.*, 2014).

The variety of samples from which both ancient and modern eDNA has been successfully extracted is varied (Figure 1.3.1) and includes, amongst others: sediment (Hofreiter *et al.*, 2003; Willerslev *et al.*, 2003; Hebsgaard *et al.*, 2009), faeces and coprolites (Poinar *et al.*, 2001; Deagle *et al.*, 2010; Burgar *et al.*, 2014), urine (Valiere & Taberlet, 2000), and freshwater (Ficetola *et al.*, 2008; Jerde *et al.*, 2013; Santas *et al.*, 2013; Takahara *et al.*, 2013). The environmental metabarcoding of eDNA extracted from “ancient” sediment (*sedDNA*) has been a source of intense focus since early publications (Hofreiter *et al.*, 2003; Willerslev *et al.*, 2003) when it

was shown to successfully identify a range of plants and animals, some of which were extinct (Willerslev *et al.*, 2003). Since these initial studies, *sedaDNA* has provided a series of intriguing and sometimes challenging results, for instance, it revealed an unexpected stability in vegetation patterns in Late Pleistocene northern Siberia despite severe climate fluctuations, providing a possible explanation as to why the Taymyr Peninsula acted as a refugium for the last mainland woolly mammoth population (Jørgensen *et al.*, 2012). This study was also the first to demonstrate that *sedaDNA* analyses and traditional palynological and morphological analyses are complementary rather than mutually exclusive. Furthermore, the novel insights that *sedaDNA* can provide has been demonstrated through the detection of relict populations that are absent in the macrofossil record, namely Alaskan megafauna (Haile *et al.*, 2009).

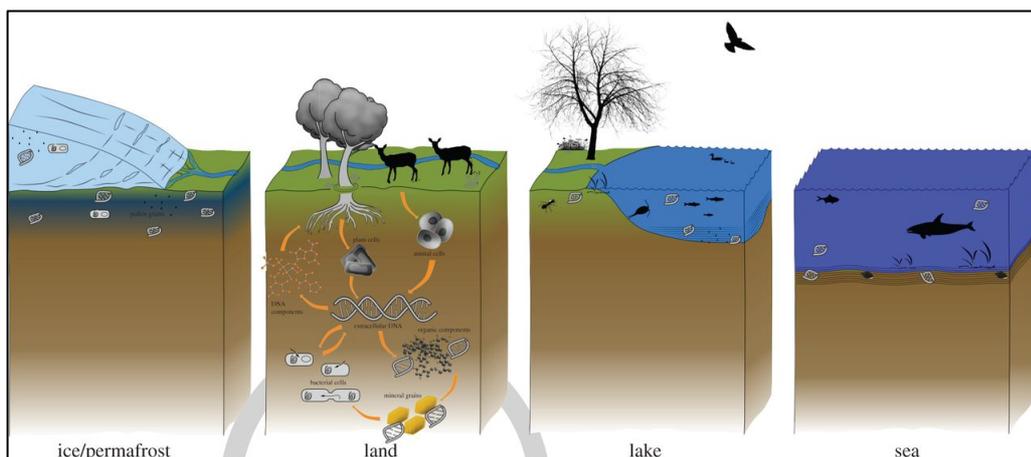


Figure 1.3.1 The potential sources of environmental DNA and the environments from which it has been reported. The successful extraction and characterisation of environmental DNA (eDNA) has been reported from permafrost, terrestrial sediment, lake sediment as well as both freshwater and seawater. The potential sources of eDNA include faeces, urine, pollen grains, epithelial cells, and other plant and animal micro- and macro-fossils and it may present extra-cellularly and intra-cellularly (reproduced from Pedersen *et al.*, 2014).

The ability of eDNA, in general, to detect rare species is a major advantage that is currently being exploited in ecosystem biomonitoring. One of the earliest studies using water samples highlighted the utility of eDNA in detecting invasive species for biomonitoring (Ficetola *et al.*, 2008). While this was not an environmental metabarcoding study – a species-specific PCR assay was used – the successful

detection of the invasive frog *Rana catesbeiana* (bullfrog) even when present at low densities in the environment demonstrated the sensitivity of eDNA approaches to ecosystem monitoring. Further studies have supported this finding and in some cases have demonstrated improved detection rates when compared to traditional methods (Jerde *et al.*, 2011; Dejean *et al.*, 2012). Such sensitivity in eDNA monitoring techniques, while variable across taxonomic groups (Thomsen *et al.*, 2012a), bodes well for the study of overall ecosystem biodiversity: another major focus of current environmental metabarcoding projects.

The assessment of biodiversity and the monitoring of vulnerable or threatened species is the *raison d'être* of conservation biology. Environmental metabarcoding is a tool that can assist in the assessment of biodiversity and aid in the detection of rare species while at the same time causing minimal impact to the surrounding environment (Jerde *et al.*, 2011; Schnell *et al.*, 2012; Thomsen *et al.*, 2012b; Bohmann *et al.*, 2014). Additionally, the use of eDNA has proven itself not just in the assessment of present-day diversity but it has also shown that it is capable of providing insights into past biodiversity. A recent study provided a record of vegetation change covering a period of 50, 000-years and in doing so revealed a stable system until the last glacial maximum (LGM) at which point the diversity dropped dramatically; a finding that contrasts with most pollen records (Willerslev *et al.*, 2014).

The ability to use environmental metabarcoding techniques to assess biodiversity in both the past and the present using eDNA and aDNA is a major strength of the method and lends itself well to areas of conservation and species management (Nichols *et al.*, 2012; Schnell *et al.*, 2012; Barnes & Turner, 2015). Moreover, the use of historical or ancient data are of great benefit when deciding and implementing species or ecosystem management strategies (Ficetola *et al.*, 2010; Jørgensen *et al.*, 2011; Barnes & Turner, 2015). However, despite the gains made in the last number of years, there remain many challenges when using environmental metabarcoding techniques to assess modern, historical and ancient samples. These challenges can prove to be particularly problematic when DNA preservation becomes compromised as is the case with degraded sources such as faecal material or with aDNA in samples

from temperate environments. Indeed, with a few exceptions, environmental metabarcoding using aDNA has been largely confined to cold environments.

1.3.2 Challenges associated with aDNA and eDNA

There are numerous challenges involved in working with degraded DNA extracted from environmental samples be they ancient or modern. While some of these challenges are historical and existed in the days of Sanger sequencing (e.g. DNA degradation and damage) others have been exacerbated more recently by the shift across into HTS workflows (e.g. filtering sequencing “noise” from genuine data). The field of aDNA has been dogged by issues of data fidelity since its inception and as such there are now strict guidelines in place when working with aDNA (Hofreiter *et al.*, 2001; Pääbo *et al.*, 2004; Gilbert *et al.*, 2005; Willerslev & Cooper, 2005); many of which are equally applicable to working with eDNA and have been loosely adopted by the environmental metabarcoding community.

The most pervasive issues across both aDNA and eDNA revolve around the use of samples where the level of endogenous DNA is, or can be, extremely low and severely degraded. Controlling and mitigating the risks of contamination is a key focal point in both aDNA and eDNA studies (Hofreiter *et al.*, 2001; Pääbo *et al.*, 2004; Gilbert *et al.*, 2005; Willerslev & Cooper, 2005; Champlot *et al.*, 2010). Contamination can arise at any point in the environmental metabarcoding workflow – from activity in the field through to sample extraction and sequencing (Thomsen & Willerslev, 2015). Plant and animal contaminants have also been detected in PCR and other laboratory reagents (Malmström *et al.*, 2005; Leonard *et al.*, 2007; Champlot *et al.*, 2010; Hofreiter *et al.*, 2010; Erlwein *et al.*, 2011; Tuke *et al.*, 2011; Boessenkool *et al.*, 2012), and as such the use of controls when conducting laboratory work is imperative and can greatly reduce the chance of false positives arising from reagent contamination. It has also been suggested that a contamination database of control sequences within a laboratory setting would be a useful addition allowing the detection of any contaminants within samples (Porter *et al.*, 2013; Pedersen *et al.*, 2014). An additional strategy may also be to sequence new reagents prior to use as sequencing technology becomes cheaper, easier to use and more routine which may determine the source of contamination as manufacturer derived or

in-house laboratory derived. The impact of contamination not only results in misleading data but can also cause the preferential amplification of contaminant sequences, seriously limiting the successful detection of endogenous DNA (Pääbo *et al.*, 2004; Gilbert *et al.*, 2005; Axelsson *et al.*, 2008). Contamination can be extremely problematic when dealing with certain taxa such as mammals when using universal primers whereby human DNA sequences (Malmström *et al.*, 2005) may swamp out any endogenous signal: this is a particularly pertinent issue when working with material from archaeological sites where the handling of bones without gloves is common.

A primary reason for the preferential amplification of contaminant DNA within ancient and modern environmental samples is due to the fact that, aside from endogenous DNA being in extremely low amounts, it can also be severely damaged and degraded (Pääbo *et al.*, 2004; Roberts & Ingham, 2008; Dabney *et al.*, 2013). The damage and degradation associated with aDNA, in particular, is problematic. Both increasing age and decreasing quality of preservation of specimens or samples tend to worsen issues associated with DNA degradation and damage, although different sources of aDNA, such as museum skins or archaeological bone, are degraded at different rates owing to differing levels of preservation and post-mortem environment (Leonard, 2008; Higgins *et al.*, 2015). The sources of DNA degradation and damage are many and include oxidative damage causing depurination (Lindahl, 1993; Hofreiter *et al.*, 2001) and DNA crosslinks preventing amplification and increasing risk of contamination (Poinar *et al.*, 1998). Additionally, miscoding lesions arising from hydrolysis, such as the deamination of cytosine to uracil, can result in incorrect bases being incorporated during PCR amplification (Hansen *et al.*, 2006). While a lot of research has been conducted into the sources, types and impacts of damage in and on aDNA very few studies have been conducted to assess similar aspects in modern environmental samples (Trevors, 1996; Deagle *et al.*, 2006). The lack of research into DNA damage specific to environmental samples is despite the obvious short, degraded nature of eDNA across the range of substrates to which it has been applied indicating the likelihood of DNA damage within sequences which can pose problems when analysing and interpreting HTS sequencing results.

The interplay between the triumvirate of damage, degradation and HTS sequencing error needs to be at the forefront of experimental plans when embarking upon environmental metabarcoding, be it on modern, historical or ancient samples. Likewise, HTS data requires careful management to limit the impact of all three drivers and prevent over- or under-estimation of taxonomic richness within samples. There is no one-size fits all approach to the removal of error arising from DNA damage, PCR artefacts or HTS sequencing but various strategies have been suggested to reduce the impacts associated with PCR and sequencing error that involve laboratory procedures such as PCR replicates (Pedersen *et al.*, 2014) and careful attention to bioinformatic filtering of sequences, such as abundance threshold cut-offs for low abundant unique sequence clusters (explored further in Chapter Five). Interestingly, despite attempts to overcome issues associated with error by increasing replicates this may prove to be a false reassurance of data fidelity as error profiles between replicates have been shown to correlate significantly (Coissac *et al.*, 2012). Nonetheless, replicates do serve to help identify less abundant taxa within samples that may only appear sporadically – although differentiating between ‘genuine’ rare taxa in samples and ‘obvious’ sequencing error is debatable. It seems that at present a balance must be struck between filtering data to remove error and relaxing filtering strategies to maintain sensitivity when attempting to detect rare taxa.

Isolating DNA damage and sequencing error from *bona fide* sequencing reads is yet further complicated by the lack of coverage of many plants, animals and other taxonomic groups on publically available genetic databases (Kvist, 2013). As mentioned previously, one of the primary methods of characterising the genetic diversity within a sample is through the use of universal primers which are designed to amplify a wide diversity of taxa during PCR. However, the conventional barcodes chosen by the barcoding community are wholly unsuitable for use in aDNA and eDNA studies. Firstly, the approved barcodes are much too long to be used to identify degraded and fragmented DNA. Secondly, designing primers to amplify short regions within the accepted barcodes to discriminate taxa is inherently difficult due to a lack of conserved regions flanking regions of variability suitable in the identification of taxa (Deagle *et al.*, 2014). As such, alternative primers must be used in aDNA and eDNA and often they target regions within either 16S rRNA or 12S

rRNA for animals (Deagle *et al.*, 2014) and *trnL* or *rbcl* genes for plants (Taberlet *et al.*, 1991; Taberlet *et al.*, 2007; Valentini *et al.*, 2009a). Such a discord between both the barcoding and the environmental metabarcoding community is at least partly to blame for the current patchy and biased genetic databases (Kvist, 2013). The problems surrounding a lack of database coverage for many Australian taxa are explored further throughout this thesis as it proved a major obstacle in the taxonomic characterisation of aDNA and eDNA extracted from Australian environmental samples.

As a result of the difficulties associated with taxonomically assigning genetic sequences in aDNA and eDNA studies, many researchers resort to the use of molecular operational taxonomic units (mOTUs, or simply OTUs) to analyse the diversity within a sample. This can be a fully or partially taxonomy-independent approach to classifying genetic sequences based on sequence nucleotide composition (Blaxter *et al.*, 2005; Ryberg, 2015). Briefly, it involves the clustering of sequences into OTUs based on a specified similarity threshold (generally 97 %) which is used to determine either the maximum within cluster or minimum between cluster similarity (Schloss & Handelsman, 2005; Schloss *et al.*, 2009; Hamady *et al.*, 2010; Edgar, 2013; Ryberg, 2015). Such a strategy, however, can prove to be particularly sensitive to error associated with HTS and DNA damage causing an inflation of true OTU numbers (Chapter Four and Five). In future, refinements of similarity threshold cut-offs will be likely needed that take into account known data on barcode mutation rates and intra/inter-species variation. Nonetheless, such a method can still be extremely useful when databases are known to be poorly populated and also when presented with samples sourced from regions of poorly characterised biodiversity (Blaxter *et al.*, 2005; Boyer *et al.*, 2015; Ryberg, 2015), such as southwest Australia, one of only a handful of recognised biodiversity hotspots worldwide (Myers *et al.*, 2000).

1.4 Characterising a biodiversity hotspot

The development of environmental metabarcoding techniques coupled with HTS presents a powerful tool for which to assess biodiversity and it has been applied extensively throughout many ecosystems worldwide with great success (Kuch *et al.*, 2002; Edwards *et al.*, 2006; Ficetola *et al.*, 2008; Haile *et al.*, 2009; Deagle *et al.*, 2010; Sønstebo *et al.*, 2010; Bohmann *et al.*, 2011; Jørgensen *et al.*, 2011; Nichols *et al.*, 2012; Thomsen *et al.*, 2012a; Pedersen *et al.*, 2014; Willerslev *et al.*, 2014; Thomsen & Willerslev, 2015). However, at the outset of this thesis, few studies had explored the potential of environmental metabarcoding in analysing Australian biodiversity, either past or present. This thesis has a particular focus on the southwest corner of Australia which is renowned for its rich flora and unique fauna (Hopper & Gioia, 2004). Despite its recognition as a biodiversity hotspot, the threat to its extensive native flora and fauna remains substantial. The development of tools such as environmental metabarcoding, however, can aid in the assessment of the region's past and present biodiversity which together can inform present and future conservation strategies in Australia's southwest biodiversity hotspot.

1.4.1 Southwest Australian biodiversity

The Australian continent is home to a large number of plants and animals not found elsewhere in the world (Hanson *et al.*, 2008). The Southwest Australia Ecoregion is particularly speciose and it includes the area designated the Southwest Australian Floristic Region (SAWFR) (Hopper & Gioia, 2004) and spans a number of Interim Biogeographical Regionalisation of Australia (IBRA) regions (Thackway & Cresswell, 1995) (Figure 1.4.1). The large plant biodiversity that is found in southwest Australia is possibly the result of nutrient-poor soils in the area requiring plants to be highly specialised to cope (Dortch, 2004b), in addition to the region's Mediterranean climate and the protracted periods of aridity and isolation it has experienced (Dortch, 2004b). As a result of these factors, there are over 5,500 species of vascular plant found in the region, possibly as high as 8,000, with approximately 80 % of those listed as endemic to Western Australia; however, these numbers do vary (Beard, 1995; Beard *et al.*, 2000; Hopper & Gioia, 2004). Much of this diversity is concentrated in woody plant families such as Myrtaceae, Proteaceae

and Fabaceae. Although there is a rich diversity in plants, and also invertebrates (Cooper *et al.*, 2011), the diversity of vertebrates is considerably less, particularly for mammals (Dortch, 2004b). However, there have been suggestions of a historic loss of diversity among reptiles, amphibians and small mammals (How *et al.*, 1987; Dortch, 2004b). Indeed, genetic studies using aDNA have recently shown declines in genetic diversity and connectivity in *Bettonia penicillata ogilbyi* (Western Australian woylie) (Pacioni *et al.*, 2011; Pacioni *et al.*, 2015). Despite current and past declines, southwest Australia still maintains a significant level of genetic, species and ecosystem diversity with a high degree of species endemism, particularly for vascular plants. However, Myers *et al.*, 2000 classify southwest Australia as a biodiversity hotspot of conservation priority recognising its extensive biodiversity but also the fact that at least 70 % of its primary vegetation has been lost (Myers *et al.*, 2000). Therefore, while southwest Australia is a region of great biological importance, it and its biodiversity are under serious threat due to habitat loss.

1.4.2 Threats to southwest Australian biodiversity

The loss of habitat in the Australian wheatbelt, to the east of the Southwest Australia Ecoregion, is estimated in region of 93 % of its original vegetation while the Swan Coastal Plain, an IBRA region within the Southwest Australia Ecoregion, has been shown to have lost approximately 80 % of its habitat (Beard, 1995). Currently, southwest Australia has over 50 ecological communities that are listed as threatened, most of which occur in the Swan Coastal Plain IBRA region (DPAW, 2015). In addition to high levels of habitat loss (Bradshaw, 2012) the ecoregion is also suffering as a result of a climate that is becoming increasingly arid (Klausmeyer & Shaw, 2009; Wardell-Johnson *et al.*, 2011).

Since European colonisation of Australia, 22 species of mammal have gone extinct (Johnson, 2006; Woinarski *et al.*, 2015); an unprecedented number that no other country in the world matches. In southwest Australia itself, nearly a third of mammals recorded as being present prior to European arrival have disappeared, while populations of many other mammal species, such as *Setonix brachyurus* (quokka) and *Bettongia penicillata ogilbyi* (Western Australian woylie), have been

drastically reduced to essentially relict populations and show high degrees of genetic loss (Alacs *et al.*, 2003; Hayward *et al.*, 2003; Pacioni *et al.*, 2011; Pacioni *et al.*, 2015). This contraction in populations and distributions is also seen in other non-mammalian taxa such as reptiles – e.g. *Pseudemydura umbrina* (Burbidge & Kuchling, 2004; Burbidge & Kuchling, 2007) and amphibians – *Geocrinia alba* (Wardell-Johnson *et al.*, 1995; Driscoll, 1997; Roberts *et al.*, 1999). Non-native invasive species such as foxes, cats and rabbits continue to pose serious threats to native wildlife. However, introduced species are not the only cause of concern, for example, the native Australian *Eolophus roseicapilla* (galah), which has colonised much of Australia due to large changes in the environmental landscape, is regarded as an invasive species throughout much of the continent as it competes with locally native birds such as *Calyptorhynchus latirostris* (Carnaby's black cockatoo) in southwest Australia (SoE, 1996). The threat to native flora is also serious with the introduction and spread of non-native weeds, e.g. *Zantedeschia aethiopica* (arum lily) (Parsons, 2001), and the susceptibility of native plants to *Phytophthora cinnamomi*: the causative agent of a form of dieback known as root rot. It is estimated that over 2000 of southwest Australian native plants are susceptible to dieback and over 50 % of flora listed as threatened appear susceptible (Shearer *et al.*, 2004).

1.4.3 Environmental metabarcoding southwest Australia

Southwest Australia represents a challenging ecosystem to study using ancient and degraded DNA obtained from environmental samples. However, the rapid loss of habitat and species diversity in the past and the continued threats posed by habitat modification, both natural and anthropogenic, necessitate the need for effective conservation management plans. Assessing both the current and historical status of a species or examining the past and present diversity of a region to facilitate informed conservation decisions is critical. Over the past decade, countless studies have shown the insights that can be gained from using aDNA and eDNA in ecological studies, conservation research and species management plans (Leonard, 2008; Kelly *et al.*, 2014) and aDNA, specifically, can help establish baseline targets for managing vulnerable species (Pacioni *et al.*, 2015).

It is, therefore, worthwhile to explore the current utility and limits of both aDNA and eDNA environmental metabarcoding in the context of southwest Australia using modern and ancient samples. In doing so, it may have immediate implications for the application of these techniques to similar regions such as the Western Cape of South Africa – also regarded as a biodiversity hotspot of conservation priority. The use of environmental metabarcoding in a region with high levels of undescribed biodiversity will undoubtedly prove challenging, particularly in the hot and temperate regions of Australia that are less than ideal for aDNA preservation (Lindahl, 1993; Willerslev & Cooper, 2005; Leonard, 2008). Nevertheless in exploring methods to improve sequence data fidelity using modern environmental samples (Chapter Two) and ancient samples that are known to preserve aDNA in other hot climates (Chapter Three) it may be possible to develop a strategy to effectively assess biodiversity through time at archaeologically significant sites in the region (Chapters Four and Six).

The sites chosen are found in the extensive limestone cave network of the Leeuwin-Naturaliste National Park situated in the Southwest Australia Ecoregion. Historically limestone caves preserve bones, and DNA, well and tend to buffer both pH and temperature fluctuation (Lindahl, 1993; Willerslev & Cooper, 2005; Leonard, 2008; Llamas *et al.*, 2015). For this thesis, five cave sites were selected that offer a unique insight into past faunal and floral turnover in southwest Australia over a combined 50, 000-year window against the backdrop of changing climate and episodic human occupation at the sites. Three of the sites, Devil’s Lair, Tunnel Cave and Rainbow Cave, have been studied previously using traditional archaeological methods (Dortch, 1979; Lilley, 1993; Turney & Bird, 2001; Dortch, 2004b; Dortch & Wright, 2010) while the final two, Wonitji Janga (meaning “spirits talking”) and Northcote Sinkhole, were not studied prior to the commencement of this research.

The Wonitji Janga deposit spans the period before and after European arrival in Australia while Northcote Sinkhole is devoid of any archaeological material and is located just 10 m from Wonitji Janga (Dortch *et al.*, 2014). The deposit at Rainbow Cave is essentially divided into a cultural upper section and a non-cultural bottom section, both of which containing multiple stratigraphical units (Lilley, 1993; Dortch, 2004b). At both Devil’s Lair and Tunnel Cave a substantial change in vegetation was

determined using bone material and charcoal (Dortch, 2004a, 2004b). Archaeological analyses indicated that approximately 13,000 BP the habitat became more closed with the replacement of the Jarrah forest with Karri forest. At Tunnel Cave, it was established that the time at which Karri forest had completely replaced the previous Jarrah forest coincided with the abandonment of the sites by people (Dortch, 2004b, 2004a).

Together these sites offer a unique opportunity with which to assess the suitability of environmental metabarcoding in the study of Australian archaeology and palaeoecology. Additionally, the exploration of these sites has the potential to inform future studies analysing past and present biodiversity elsewhere in Australia and also in other hot or temperate regions further afield.

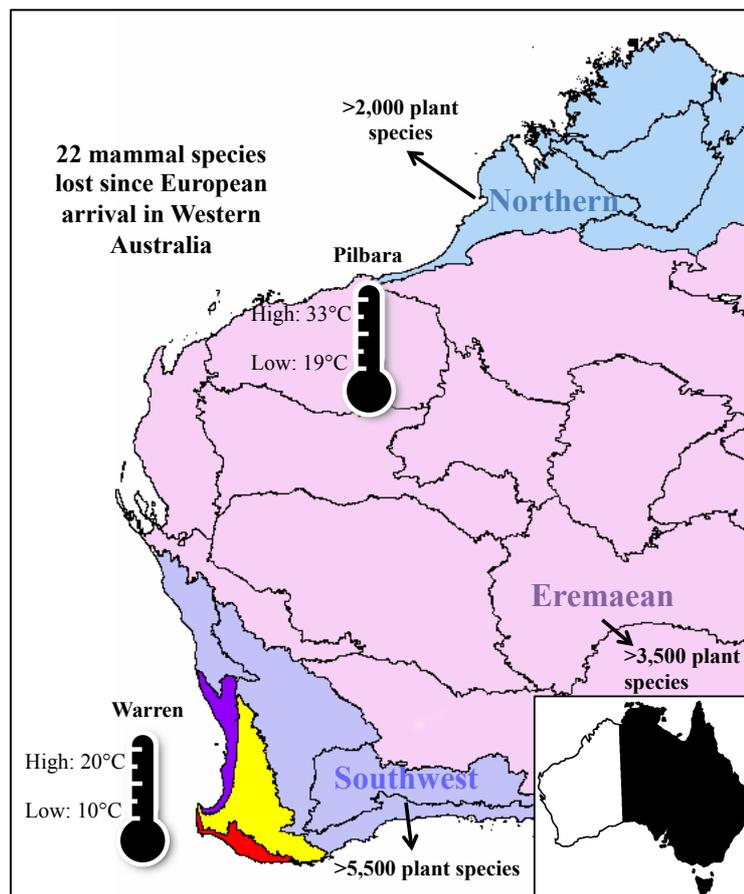


Figure 1.4.1 Map showing some of the regions studied and some abiotic and biotic information related to Western Australia. The three major botanical provinces are shown and the three regions that converge at the sites studied are highlighted. These are the Swan Coastal Plain (purple), Jarrah Forest (yellow) and Warren (red). (Data is taken from <https://florabase.dpaw.wa.gov.au>; <http://www.bom.gov.au>; Hopper & Gioia, 2004; Johnson, 2006)

1.5 References

- Alacs, E., Alpus, D., de Tores, P. J., Dillon, M., & Spencer, P. B. S. (2003). Identifying the presence of Quokkas (*Setonix brachyurus*) and other macropods using cytochrome b analysis from faeces. *Wildlife Research*, *30*, 41-47.
- Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K. H., . . . Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, *21*, 1966-1979.
- Angeloni, F., Wagemaker, N., Vergeer, P., & Ouborg, J. (2012). Genomic toolboxes for conservation biologists. *Evolutionary Applications*, *5*, 130-143.
- Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *Nature Biotechnology*, *25*, 195-203.
- Archer, J., Weber, J., Henry, K., Winner, D., Gibson, R., Lee, L., . . . Quiñones-Mateu, M. E. (2012). Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS One*, *7*, e49602.
- Axelsson, E., Willerslev, E., Gilbert, M. T. P., & Nielsen, R. (2008). The effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution*, *25*, 2181-2187.
- Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*, 1-17.
- Beard, J. S. (1995). South-west Botanical Province. In S. D. Davis, V. H. Heywood, & A. C. Hamilton (Eds.), *Centres of Plant Diversity. Volume 2. Asia, Australasia, and the Pacific*. (Vol. 2). Cambridge, UK: WWF/IUCN, IUCN Publications Unit.
- Beard, J. S., Chapman, A. R., & Gioia, P. (2000). Species richness and endemism in the Western Australian flora. *Journal of Biogeography*, *27*, 1257-1268.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., . . . Bignell, H. R. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*, 53-59.

- Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, *2*, e197.
- Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1935-1943.
- Boessenkool, S., Epp, L. S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., . . . Brochmann, C. (2012). Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, *21*, 1806-1815.
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., . . . de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, *29*, 358-367.
- Bohmann, K., Monadjem, A., Lehmkühl, N., Rasmussen, M., Zeale, M. R. K., Clare, E., . . . Gilbert, M. T. P. (2011). Molecular diet analysis of two African Free-tailed Bats (Molossidae) using High Throughput Sequencing. *PLoS One*, *6*, e21441.
- Boyer, S., Cruickshank, R. H., & Wratten, S. D. (2015). Faeces of generalist predators as 'biodiversity capsules': A new tool for biodiversity assessment in remote and inaccessible habitats. *Food Webs*, *3*, 1-6.
- Bradshaw, C. J. A. (2012). Little left to lose: deforestation and forest degradation in Australia since European colonization. *Journal of Plant Ecology*, *5*, 109-120.
- Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., & Tyson, G. W. (2013). Shining a light on dark sequencing: Characterising errors in Ion Torrent PGM data. *PLoS Computational Biology*, *9*, e1003031.
- Bronner, I. F., Quail, M. A., Turner, D. J., & Swerdlow, H. (2001). Improved protocols for Illumina sequencing *Current Protocols in Human Genetics*: John Wiley & Sons, Inc.

- Budowle, B., Connell, N. D., Bielecka-Oder, A., Colwell, R. R., Corbett, C. R., Fletcher, J., . . . Minot, S. (2014). Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics*, *5*, 1-18.
- Buermans, H. P., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta*, *1842*, 1932-1941.
- Burbidge, A., & Kuchling, G. (2004). Western Swamp Tortoise (*Pseudemydura umbrina*) recovery plan, Department of Conservation & Land Management, Western Australia.
- Burbidge, A. A., & Kuchling, G. (2007). The Western Swamp Tortoise - 50 years on. *Landscape*, *22*, 24-29.
- Burgar, J. M., Murray, D. C., Craig, M. D., Haile, J., Houston, J., Stokes, V., & Bunce, M. (2014). Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology*, *23*, 3605-3617.
- Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., . . . Crandall, K. A. (2011). Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution*, *3*, 1312-1323.
- Campo, D. S., Dimitrova, Z., Yamasaki, L., Skums, P., Lau, D. T., Vaughan, G., . . . Khudyakov, Y. (2014). Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC Genomics*, *15 Suppl 5*, S4.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335-336.
- Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One*, *5*, e13042.

- Clarke, L. J., Czechowski, P., Soubrier, J., Stevens, M. I., & Cooper, A. (2014). Modular tagging of amplicons using a single PCR for high-throughput sequencing. *Molecular Ecology Resources*, *14*, 117-121.
- Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*, 1834-1847.
- Cooper, S. J. B., Harvey, M. S., Saint, K. M., & Main, B. Y. (2011). Deep phylogeographic structuring of populations of the trapdoor spider *Moggridgea tingle* (Migidae) from southwestern Australia: evidence for long-term refugia within refugia. *Molecular Ecology*, *20*, 3219-3236.
- Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*, a012567.
- de Cárcer, D. A., Denman, S. E., McSweeney, C., & Morrison, M. (2011). Strategy for modular tagged high-throughput amplicon sequencing. *Applied Environmental Microbiology*, *77*, 6310-6312.
- Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.
- Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Frontiers in Zoology*, *3*, 10.
- Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*.
- Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology*, *49*, 953-959.

- Dortch, C. (1979). Devil's Lair, an example of prolonged cave use in southwestern Australia. *World Archaeology*, *10*, 258-279.
- Dortch, J. (2004a). Late Quaternary vegetation change and the extinction of Black-flanked Rockwallaby (*Petrogale lateralis*) at Tunnel Cave, southwestern Australia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *211*, 185-204.
- Dortch, J. (2004b). *Palaeo-environmental change and the persistence of human occupation in south-western Australian forests*. Oxford: Archaeopress.
- Dortch, J., Monks, C., Webb, W., & Balme, J. (2014). Intergenerational archaeology: Exploring niche construction in southwest Australian zooarchaeology. *Australian Archaeology*, *79*, 187-193.
- Dortch, J., & Wright, R. (2010). Identifying palaeo-environments and changes in Aboriginal subsistence from dual-patterned faunal assemblages, south-western Australia. *Journal of Archaeological Science*, *37*, 1053-1064.
- DPAW - Department of Parks and Wildlife . (2015). *Priority ecological communities for Western Australia*. Retrieved from Species and Communities Branch, Department of Parks and Wildlife
- Driscoll, D. A. (1997). Mobility and metapopulation structure of *Geocrinia alba* and *Geocrinia vitellina*, two endangered frog species from southwestern Australia. *Australian Journal of Ecology*, *22*, 185-195.
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*, 996-998.
- Edwards, R. A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D. M., . . . Rohwer, F. (2006). Using pyrosequencing to shed light on deep mine microbial ecology. *BMC Genomics*, *7*, 57-57.
- Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S., & McClure, M. O. (2011). DNA extraction columns contaminated with murine sequences. *PLoS One*, *6*, e23484.

- Faircloth, B. C., & Glenn, T. C. (2012). Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS One*, *7*, e42543.
- Ficetola, G. F., Maiorano, L., Falcucci, A., Dendoncker, N., Boitani, L., Padoa-Schioppa, E., . . . Thuiller, W. (2010). Knowing the past to predict the future: land-use change and the distribution of invasive bullfrogs. *Global Change Biology*, *16*, 528-537.
- Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423-425.
- França, L. T. C., Carrilho, E., & Kist, T. B. L. (2002). A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics*, *35*.
- Fuller, C. W., Middendorf, L. R., Benner, S. A., Church, G. M., Harris, T., Huang, X., . . . Vezenov, D. V. (2009). The challenges of sequencing by synthesis. *Nature Biotechnology*, *27*, 1013-1023.
- Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.-J., Griggo, C., Gielly, L., . . . Taberlet, P. (2014). Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, *5*, 3211.
- Gilbert, M. T. P., Bandelt, H. J., Hofreiter, M., & Barnes, I. (2005). Assessing ancient DNA studies. *Trends in Ecology and Evolution*, *20*, 541-544.
- Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, *5*, 759-769.
- Gonzalez, A., & Knight, R. (2012). Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology*, *23*, 64-71.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*, 333-351.

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., . . . Paabo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature*, *444*, 330-336.

Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., . . . Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *proceedings of the National Academy of Sciences*, *106*, 22352-22357.

Hamady, M., Lozupone, C., & Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal*, *4*, 17-27.

Hansen, A. J., Mitchell, D. L., Wiuf, C., Paniker, L., Brand, T. B., Binladen, J., . . . Willerslev, E. (2006). Crosslinks rather than strand breaks determine access to ancient DNA sequences from frozen sediments. *Genetics*, *173*, 1175-1179.

Hanson, T., Brooks, T. M., Fonseca, G. A. B. D., Hoffmann, M., Lamoreux, J. F., Machlis, G., . . . Pilgrim, J. D. (2008). Warfare in biodiversity hotspots. *Conservation Biology*, *23*, 578-587.

Hayward, M. W., de Tores, P. J., Dillon, M. J., & Fox, B. J. (2003). Local population structure of a naturally occurring metapopulation of the Quokka (*Setonix brachyurus* Macropodidae: Marsupialia). *Biological Conservation*, *110*, 343-355.

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*, 313-321.

Hebsgaard, Martin B., Gilbert, M. T. P., Arneborg, J., Heyn, P., Allentoft, Morten E., Bunce, M., . . . Willerslev, E. (2009). 'The Farm Beneath the Sand' – an archaeological case study on ancient 'dirt' DNA. *Antiquity*, *83*, 430-444.

Higgins, D., Rohrlach, A. B., Kaidonis, J., Townsend, G., & Austin, J. J. (2015). Differential nuclear and mitochondrial DNA preservation in post-mortem teeth with implications for forensic and ancient DNA studies. *PLoS One*, *10*, e0126935.

- Hofreiter, M., Kreuz, E., Eriksson, J., Schubert, G., & Hohmann, G. (2010). Vertebrate DNA in fecal samples from bonobos and gorillas: evidence for meat consumption or artefact? *PLoS One*, *5*, e9419.
- Hofreiter, M., Mead, J. I., Martin, P., & Poinar, H. N. (2003). Molecular caving. *Current Biology*, *13*, R693-R695.
- Hofreiter, M., Poinar, H. N., Spaulding, W. G., Bauer, K., Martin, P. S., Possnert, G., & Pääbo, S. (2000). A molecular analysis of ground sloth diet through the last glaciation. *Molecular Ecology*, *9*, 1975-1984.
- Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews Genetics*, *2*, 353-359.
- Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One*, *6*, e19254.
- Hopper, S. D., & Gioia, P. (2004). The southwest Australian floristic region: Evolution and conservation of a global hot spot of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 623-650.
- How, R. A., Dell, J., & Humphreys, W. F. (1987). The ground vertebrate fauna of coastal areas between Busselton and Albany, Western Australia. *Records of the Western Australian Museum*, *13*, 553-574.
- Hudson, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, *8*, 3-17.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*, 377-386.
- Jerde, C. L., Chadderton, W. L., Mahon, A. R., Renshaw, M. A., Corush, J., Budny, M. L., . . . Lodge, D. M. (2013). Detection of Asian carp DNA as part of a Great Lakes basin-wide surveillance program. *Canadian Journal of Fisheries and Aquatic Sciences*, *70*, 522-526.

- Jerde, C. L., Mahon, A. R., Chadderton, W. L., & Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, *4*, 150-157.
- Johnson, C. (2006). *Australia's mammal extinctions: a 50000 year history*: America; Cambridge University Press.
- Jørgensen, T., Haile, J., Möller, P. E. R., Andreev, A., Boessenkool, S., Rasmussen, M., . . . Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, *21*, 1989-2003.
- Jørgensen, T., Kjær, K. H., Haile, J., Rasmussen, M., Boessenkool, S., Andersen, K., . . . Willerslev, E. (2011). Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Molecular Ecology*, *21*, 1980-1988.
- Kelly, R. P., Port, J. A., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., . . . Crowder, L. B. (2014). Environmental monitoring. Harnessing DNA to improve environmental management. *Science*, *344*, 1455-1456.
- Klausmeyer, K. R., & Shaw, M. R. (2009). Climate change, habitat loss, protected areas and the climate adaptation potential of species in mediterranean ecosystems worldwide. *PLoS ONE*, *4*, e6392.
- Knapp, M., & Hofreiter, M. (2010). Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes*, *1*, 227-243.
- Kuch, M., Rohland, N., Betancourt, J. L., Latorre, C., Stepan, S., & Poinar, H. N. (2002). Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert, Chile. *Molecular Ecology*, *11*, 913-924.
- Kvist, S. (2013). Barcoding in the dark?: A critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. *Molecular Phylogenetics and Evolution*, *69*, 39-45.

- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860-921.
- Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, *3*, 1-8.
- Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. *Molecular Ecology*, *17*, 4186-4196.
- Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., . . . Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, *34*, 1361-1366.
- Lilley, I. (1993). Recent research in southwestern Australia: a summary of initial findings. *Australian Archaeology*, *36*, 34-41.
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*, 709-715.
- Liu, L., Li, Y., Li, S., Hu, Y., He, R., Pong, D., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of Biomedicine and Biotechnology*, *2012*, 251364.
- Llamas, B., Brotherton, P., Mitchell, K. J., Templeton, J. E., Thomson, V. A., Metcalf, J. L., . . . Camens, A. B. (2015). Late Pleistocene Australian marsupial DNA clarifies the affinities of extinct megafaunal kangaroos and wallabies. *Molecular Biology and Evolution*, *32*, 574-584.
- Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, *30*, 434-439.
- Malmström, H., Stora, J., Dalen, L., Holmlund, G., & Götherström, A. (2005). Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution*, *22*, 2040-2047.

- Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387-402.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bembien, L. A., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376-380.
- Matisoo-Smith, E., Roberts, K., Welikala, N., Tannock, G., Chester, P., Feek, D., & Flenley, J. (2008). Recovery of DNA and pollen from New Zealand lake sediments. *Quaternary International*, 184, 139-149.
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 31-46.
- Mikheyev, A. S., & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources*, 14, 1097-1102.
- Miller, W., Drautz, D. I., Ratan, A., Pusey, B., Qi, J., Lesk, A. M., . . . Schuster, S. C. (2008). Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456, 387-390.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853-858.
- Nanney, D. L. (1982). Genes and phenes in *Tetrahymena*. *Bioscience*, 32, 783-788.
- Nichols, R. V., Konigsson H Fau - Danell, K., Danell K Fau - Spong, G., & Spong, G. (2012). Browsed twig environmental DNA: diagnostic PCR to identify ungulate species. *Molecular Ecology Resources*, 12, 983-989.
- Pääbo, S., Poinar, H., Serre, D., Jaenicke-Després, V., Hebler, J., Rohland, N., . . . Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38, 645-679.
- Pacioni, C., Hunt, H., Allentoft, M. E., Vaughan, T. G., Wayne, A. F., Baynes, A., . . . Bunce, M. (2015). Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, 24, 5813-5828.

- Pacioni, C., Wayne, A. F., & Spencer, P. B. S. (2011). Effects of habitat fragmentation on population structure and long-distance gene flow in an endangered marsupial: the woylie. *Journal of Zoology*, 283, 98-107.
- Parducci, L., Matetovici, I., Fontana, S. L., Bennett, K. D., Suyama, Y., Haile, J., . . . Willerslev, E. (2013). Molecular- and pollen-based vegetation analysis in lake sediments from central Scandinavia. *Molecular Ecology*, 3511-3524.
- Parsons, W. T. (2001). *Noxious weeds of Australia / W.T. Parsons and E.G. Cuthbertson*. Collingwood, Vic: CSIRO Publishing.
- Pawlowska, J., Lejzerowicz, F., Esling, P., Szczucinski, W., Zajaczkowski, M., & Pawlowski, J. (2014). Ancient DNA sheds new light on the Svalbard foraminiferal fossil record of the last millennium. *Geobiology*, 12, 277-288.
- Pedersen, M. W., Ginolhac, A., Orlando, L., Olsen, J., Andersen, K., Holm, J., . . . Kjær, K. H. (2013). A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. *Quaternary Science Reviews*, 75, 161-168.
- Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., . . . Willerslev, E. (2014). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B*, 370, 20130383.
- Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., . . . Pääbo, P. (1998). Molecular coproscopy: dung and diet of the extinct Ground Sloth *Nothrotheriops shastensis*. *Science*, 281, 402-406.
- Poinar, H. N., Kuch, M., Sobolik, K. D., Barnes, I., Stankiewicz, A. B., Kuder, T., . . . Pääbo, S. (2001). A molecular analysis of dietary diversity for three archaic Native Americans. *Proceedings of the National Academy of Sciences of the United States of America*, 98, 4317-4322.
- Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, 21, 1931-1950.

- Porter, T. M., Golding, G. B., King, C., Froese, D., Zazula, G., & Poinar, H. N. (2013). Amplicon pyrosequencing late Pleistocene permafrost: the removal of putative contaminant sequences and small-scale reproducibility. *Molecular Ecology Resources*, *13*, 798-810.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341.
- Quince, C., Lanzen, A., Curtis, T. P., Davenport, R. J., Hall, N., Head, I. M., . . . Sloan, W. T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*, *6*, 639-641.
- Quince, C., Lanzen, A., Davenport, R., & Turnbaugh, P. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, *12*, 38.
- Rizzi, E., Lari, M., Gigli, E., De Bellis, G., & Caramelli, D. (2012). Ancient DNA studies: new perspectives on old samples. *Genetics Selection Evolution*, *44*, 21-39.
- Roberts, C., & Ingham, S. (2008). Using ancient DNA analysis in palaeopathology: a critical analysis of published papers, with recommendations for future work. *International Journal of Osteoarchaeology*, *18*, 600-613.
- Roberts, D., Conroy, S., & Williams, K. (1999). Conservation status of frogs in Western Australia. In A. Campbell (Ed.), *Declines and Disappearances of Australian Frogs* (pp. 177-184): Environment Australia, Canberra.
- Roche. (2009). Technical Bulletin: Amplicon fusion primer design guidelines for GS FLX Titanium series Lib-A chemistry. *TCB No. 013-2009*, 1-3.
- Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., . . . Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*, 348-352.

- Roychowdhury, S., Iyer, M. K., Robinson, D. R., Lonigro, R. J., Wu, Y. M., Cao, X., . . . Chinnaiyan, A. M. (2011). Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science Translational Medicine*, 3, 111ra121.
- Ryberg, M. (2015). Molecular operational taxonomic units as approximations of species in the light of evolutionary models and empirical data from Fungi. *Molecular Ecology*, 24, 5770-5777.
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 25, 441-448.
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74, 5463-5467.
- Santas, A. J., Persaud, T., Wolfe, B. A., & Bauman, J. M. (2013). Noninvasive method for a statewide survey of Eastern hellbenders *Cryptobranchus alleganiensis* using environmental DNA. *International Journal of Zoology*, 2013, 174056.
- Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied Environmental Microbiology*, 71, 1501-1506.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied Environmental Microbiology*, 75, 7537-7541.
- Schnell, I. B., Thomsen, P. F., Wilkinson, N., Rasmussen, M., Jensen, L. R. D., Willerslev, E., . . . Gilbert, M. T. P. (2012). Screening mammal biodiversity using DNA from leeches. *Current Biology*, 22, R262-R263.
- Sharpton, T. J. (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in Plant Science*, 5, 209.

- Shearer, B. L., Crane, C. E., & Cochrane, A. (2004). Quantification of the susceptibility of the native flora of the South-West Botanical Province, Western Australia, to *Phytophthora cinnamomi*. *Australian Journal of Botany*, *52*, 435-443.
- Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*, 1135-1145.
- Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*, 1794-1805.
- SoE - State of the Environment Advisory Council. (1996). *State of the Environment Report 1996*, CSIRO Publishing, Melbourne.
- Soininen, E. M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., . . . Taberlet, P. (2009). Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, *6*, 16.
- Sønstebø, J. H., Gielly, L., Brysting, A. K., Elven, R., Edwards, M., Haile, J., . . . Brochmann, C. (2010). Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, *10*, 1009-1018.
- Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, *9*, 640.
- Stoddart, D., Heron, A. J., Mikhailova, E., Maglia, G., & Bayley, H. (2009). Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 7702-7707.
- Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012a). Environmental DNA. *Molecular Ecology*, *21*, 1789-1793.

- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012b). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045-2050.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., . . . Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*, e14.
- Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three noncoding regions of chloroplast DNA. *Plant Molecular Biology*, *17*, 1105-1109.
- Takahara, T., Minamoto, T., & Doi, H. (2013). Using environmental DNA to estimate the distribution of an invasive fish species in ponds. *PLoS One*, *8*, e56584.
- Tautz, D., Ellegren, H., & Weigel, D. (2010). Next generation molecular ecology. *Molecular Ecology*, *19*, Supplement 1:1-3.
- ten Bosch, J. R., & Grody, W. W. (2008). Keeping up with the next generation: massively parallel sequencing in clinical diagnostics. *The Journal of Molecular Diagnostics* *10*, 484-492.
- Thackway, R., & Cresswell, I. D. (1995). *An interim biogeographic regionalisation for Australia: a framework for setting priorities in the National Reserves System Cooperative Program*.
- Thomas, R. K., Nickerson, E., Simons, J. F., Janne, P. A., Tengs, T., Yuza, Y., . . . Meyerson, M. (2006). Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Methods*, *12*, 852-855.
- Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., . . . Willerslev, E. (2012a). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565-2573.

- Thomsen, P. F., Kielgast, J., Iversen, L. L. n., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012b). Detection of a diverse marine fish fauna using Environmental DNA from seawater samples. *PLoS One*, *7*, e41732.
- Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4-18.
- Tillmar, A., Dell'Amico, Welander, J., & Holmlund. (2013). A universal method for species identification of mammals utilizing Next Generation Sequencing for the analysis of DNA mixtures. *PLoS One*, *8*, e83761.
- Trevors, J. T. (1996). Nucleic acids in the environment. *Current Opinion in Biotechnology*, *7*, 331-336.
- Tuke, P. W., Tettmar, K. I., Tamuri, A., Stoye, J. P., & Tedder, R. S. (2011). PCR Master Mixes Harbour Murine DNA Sequences. Caveat Emptor! *PLoS One*, *6*, e19953.
- Turney, C., & Bird, M. I. (2001). Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. *Quaternary Research*, *55*, 3-13.
- Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E. V. A., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009a). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, *9*, 51-60.
- Valentini, A., Pompanon, F., & Taberlet, P. (2009b). DNA barcoding for ecologists. *Trends in Ecology and Evolution*, *24*, 110-117.
- Valiere, N., & Taberlet, P. (2000). Urine collected in the field as a source of DNA for species and individual identification. *Molecular Ecology*, *9*, 2150-2152.
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, *30*, 418-426.
- Varley, K. E., & Mitra, R. D. (2008). Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Research*, *18*, 1844-1850.

- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science*, *291*, 1304-1351.
- Wang, Y., Yang, Q., & Wang, Z. (2015). The evolution of nanopore sequencing. *Frontiers in Genetics*, *5*, 449.
- Wardell-Johnson, G., Roberts, J. D., Driscoll, D., & Williams, K. (1995). Orange-bellied and white-bellied frogs recovery plan. *Western Australian Department of Conservation and Land Management*.
- Wardell-Johnson, G. W., Keppel, G., & Sander, J. (2011). Climate change impacts on the terrestrial biodiversity and carbon stocks of Oceania. *Pacific Conservation Biology*, *17*, 220-240.
- Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*, 3-16.
- Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., . . . Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, *506*, 47-51.
- Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., . . . Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, *300*, 791-795.
- Williamson, S. J., Rusch, D. B., Yooseph, S., Halpern, A. L., Heidelberg, K. B., Glass, J. I., . . . Venter, J. C. (2008). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*, *3*, e1456.
- Woinarski, J. C. Z., Burbidge, A. A., & Harrison, P. L. (2015). Ongoing unraveling of a continental fauna: Decline and extinction of Australian mammals since European settlement. *proceedings of the National Academy of Sciences*, *112*, 4531-4540.

Yang, Y., Xie, B., & Yan, J. (2014). Application of next-generation sequencing technology in forensic science. *Genomics, Proteomics & Bioinformatics*, *12*, 190-197.

Yoccoz, N. G., Brathen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., . . . Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*, 3647-3655.

Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., & Yu, J. (2010). The next-generation sequencing technology: a technology review and future perspective. *Science China Life Sciences*, *53*, 44-57.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

1.6 Synopsis: the aim and scope of this thesis

The landscape of DNA sequencing has changed dramatically from the early days of Sanger sequencing and continues to develop at a rapid pace. The cost of DNA sequencing has dropped dramatically over the past decade making the technology more readily accessible and affordable. The benefits offered by HTS have been seized upon across well-established disciplines (aDNA research) and given rise to new sub-disciplines (environmental metabarcoding). The ability to characterise difficult to study ecosystems and gain insight into ecological shifts over time has been, and to some extent still is, a novel use of HTS technologies. However, the application of such technology is not without issues with respect to experimental design and data analysis.

The manuscripts within this thesis use HTS to study a combination of both modern and ancient complex, heterogeneous substrates. Each manuscript addresses aims specific to the study therein but they can also be taken as a whole to give an insight into the applicability and use of HTS to address pertinent ecological questions: the central aim of this thesis.

Chapter Two presents a comprehensive comparison of quantitative PCR (qPCR) and HTS data to determine whether or not HTS is actually quantitative, i.e. whether sequence data obtained using HTS accurately reflects the proportion of prey in faecal samples determined via qPCR.

Chapter Three details the use of aDNA techniques and HTS to characterise the flora and fauna contained within ancient herbivore middens collected from sites in Western Australia and South Africa. Cognisant of the issues raised in Chapter Two this chapter also explores some of the difficulties in identifying taxa in regions of extensive biodiversity that are poorly characterised.

Chapter Four attempts to develop an efficient and cost-effective strategy to genetically identify hundreds of unidentifiable fragmentary bones. Such material is often found in large quantities at archaeological and paleontological sites but is seldom used in traditional analyses.

Chapter Five seeks to relay a number of important considerations when embarking on the use of HTS in ecological studies and others, using data generated from Chapters Two–Four and new data. It addresses what was at the time an imbalance in the literature whereby considerations for bioinformatic analyses took precedence over important considerations surrounding sample preparation and data generation.

Chapter Six uses the methods developed and considerations raised in the previous chapters to apply aDNA and HTS techniques to the study of ancient sediment and fragmentary bone across five archaeological cave sites in southwest Australia. Using modern sequencing technology and methods developed in this thesis, while remaining aware of the issues raised in previous chapters, the floral and faunal biodiversity across the sites is explored through time against the backdrop of episodic human occupation and environmental change.

The overarching theme of this thesis is to develop robust strategies to explore both present and past biodiversity using modern HTS technologies. It seeks to highlight the benefits of such strategies but also the shortcomings that accompany them in an attempt to critically evaluate and further develop the use of HTS in modern ecological studies.

Chapter Two – A comparison of qPCR and HTS for diet assessment using modern faecal material

2.1 Preface

*Chapter Two uses the locally endangered *Eudyptula minor* (Little Penguin) as a model organism to determine whether prey estimates in faecal samples obtained using a species-specific qPCR assay are statistically different to those obtained using universal primers and HTS. This study resulted in the published manuscript entitled 'DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches' (PLoS ONE 2011, 6, e25776). With the exception of formatting and in-thesis referencing this manuscript has been reproduced as published.*

The genetic analysis of complex, heterogeneous mixtures, including faeces, water and sediment, offers a relatively non-invasive means to study ecosystem health and trophic interactions. High-throughput sequencing overcame a major limitation to conducting genetic audits of such samples, namely efficient and cost-effective screening of samples for DNA signatures. Prior to HTS the methods of choice for the analysis of environmental samples had been either cloning followed by Sanger sequencing or qPCR using species-specific primers.

The following manuscript used a species-specific primer assay to estimate the abundance of four fish — primary prey items — within penguin faecal samples. The abundances determined were then compared to those obtained using universal fish primers and HTS. Despite efforts to determine if estimates of prey abundance from faecal samples accurately reflected what was consumed by the predator; a comparison between qPCR and HTS to actually determine if species abundance determined from HTS using universal primers was at all quantitative had not previously been conducted.

2.1.1 Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

2.2 DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches.

Dáithí C. Murray¹, Michael Bunce¹, Belinda L. Cannell², Rebecca Oliver¹, Jayne Houston¹, Nicole E. White¹, Roberto A. Barrero³, Matthew I. Bellgard³ and James Haile¹

¹*Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, South St, Murdoch, WA, 6150, Australia.*

²*School of Biological Sciences and Biotechnology, Murdoch University, South St, Murdoch, WA, 6150, Australia.*

³*Centre for Comparative Genomics, Murdoch University, South St, Murdoch, WA, 6150, Australia.*

2.2.1 Abstract

The genetic analysis of faecal material represents a relatively non-invasive way to study animal diet and has been widely adopted in ecological research. Due to the heterogeneous nature of faecal material the primary obstacle, common to all genetic approaches, is a means to dissect the constituent DNA sequences. Traditionally, bacterial cloning of PCR amplified products was employed; less common has been the use of species-specific quantitative PCR (qPCR) assays. Currently, with the advent of High-Throughput Sequencing (HTS) technologies and indexed primers it has become possible to conduct genetic audits of faecal material to a much greater depth than previously possible. To date, no studies have systematically compared the estimates obtained by HTS with that of qPCR. What are the relative strengths and weaknesses of each technique and how quantitative are deep-sequencing approaches that employ universal primers? Using the locally threatened Little Penguin (*Eudyptula minor*) as a model organism, it is shown here that both qPCR and HTS techniques are highly correlated and produce strikingly similar quantitative estimates of fish DNA in faecal material, with no statistical difference. By designing four species-specific fish qPCR assays and comparing the data to the same four fish in the

HTS data it was possible to directly compare the strengths and weaknesses of both techniques. To obtain reproducible quantitative data one of the key, and often overlooked, steps common to both approaches is ensuring that efficient DNA isolation methods are employed and that extracts are free of inhibitors. Taken together, the methodology chosen for long-term faecal monitoring programs is largely dependent on the complexity of the prey species present and the level of accuracy that is desired. Importantly, these methods should not be thought of as mutually exclusive, as the use of both HTS and qPCR in tandem will generate datasets with the highest fidelity.

2.2.2 Introduction

DNA-based dietary analysis of faecal material has emerged as a promising tool to study animal biology, ecology and archaeology (Poinar *et al.*, 1998; Kuch *et al.*, 2001; Symondson, 2002; Valentini *et al.*, 2009a). Dietary analysis is not limited to the discovery of what an animal consumes; it can also give an insight into ecosystem health (Deagle *et al.*, 2009; Clare *et al.*, 2011; Raye *et al.*, 2011), species' responses to environmental/anthropogenic stresses (Vila & Borrelli, 2011), and assist in the development of targeted strategies for conservation (Kowalczyk *et al.*, 2011). It is evident from the increase in the use of genetic techniques that there is a growing appreciation of the use of DNA-based faecal methods to investigate diet. The analysis of faecal material has proven to be a welcome move away from more invasive techniques used to study animal diet such as lethal sampling (Miller & McEwen, 1995) and stomach flushing (Montague & Cullen, 1985), both of which have undesirable effects on the sampled population (Chiaradia *et al.*, 2003). Moreover, a general move towards molecular based approaches, e.g. fatty acid, stable isotope or DNA analysis, has allowed a shift from more subjective morphological approaches (Casper *et al.*, 1997; Valentini *et al.*, 2009a). The extraction and sequencing of DNA from faecal samples is seen to be an effective and reliable indicator of species' diet, offering increased specificity and taxonomic resolution compared to other techniques (Soininen *et al.*, 2009; Williams & Buck, 2010; Bohmann *et al.*, 2011). The possibility of misidentification of species is greatly reduced (Huson *et al.*, 2007; Bohmann *et al.*, 2011) and the ability to account for a wider range of species within the actual diet is greatly increased when

compared to morphology which relies entirely on analysis of undigested remains, therefore neglecting prey that may leave little trace of its consumption (Sheppard & Harwood, 2005; King *et al.*, 2008; Tollit *et al.*, 2009).

DNA based quantitative estimates of diet, however, are not without problems. Issues have arisen as a result of primer biases and the problem of differential digestion still remains. Put simply, “is what goes in what comes out” (Deagle *et al.*, 2010)? Moreover, variability in the amount of DNA per unit biomass between species and different tissues is also difficult to quantify. Attempts to address such concerns have recently become an active area of research. Such efforts include; the use of blocking primers to circumvent the issue of predator DNA amplification (Vestheim & Jarman, 2008; Deagle *et al.*, 2009); the use of captive feeding trials to examine differential digestion; (Deagle *et al.*, 2010) and the introduction of correction factors to account for DNA amount variability within species and tissues (Bowles *et al.*, 2011). These confounding factors continue to be a contentious issue within analytical dietary research, however, DNA-based methods arguably still present the best way forward in the explication of species’ diet (King *et al.*, 2008; Valentini *et al.*, 2009a).

Little Penguins (*Eudyptula minor*) are ideal test subjects for molecular dietary analysis and have been the subject of previous research into diet (Klomp & Wooller, 1988; Wienecke, 1989; Wooller *et al.*, 1991; Bradley *et al.*, 1997; Deagle *et al.*, 2010). The use of seabirds as barometers of marine ecosystem health is widely acknowledged, and the use of facultative feeders such as Little Penguins, whose diet is limited by food availability, provides a good indication of changes in marine environments (Boersma *et al.*, 2009; Mallory *et al.*, 2010). Little Penguins are found across the coastal regions of Australia and New Zealand (Marchant & Higgins, 1990) (Figure 2.2.1) and their diet, which includes a variety of small (<20cm) schooling fish, varies throughout the year (Klomp & Wooller, 1988; Wienecke, 1989; Wooller *et al.*, 1991; Bradley *et al.*, 1997). The penguin population situated on Penguin and Garden Islands (32°S 115°E), located south of Perth, Western Australia, represent the northernmost and westernmost limits of the range of *E. minor* (Wienecke, 1993; Wienecke *et al.*, 1995) (Figure 2.2.1). As a fringe population, these penguins are more vulnerable to environmental changes such as rising sea temperatures and increased ocean acidification (Boersma, 2008; Dann & Chambers, 2009). Moreover,

Penguin Island's close proximity to human settlement also puts it under increased pressure due to anthropogenic stressors, such as commercial and recreational fishing, in addition to coastal development (Chape, 1984; Harrigan, 1992; Wienecke *et al.*, 1995; Cannell, 2001; Pichegru *et al.*, 2009). The development of a multi-year DNA-based study to investigate dietary preferences will prove an effective method to monitor *E. minor* and the marine environment.

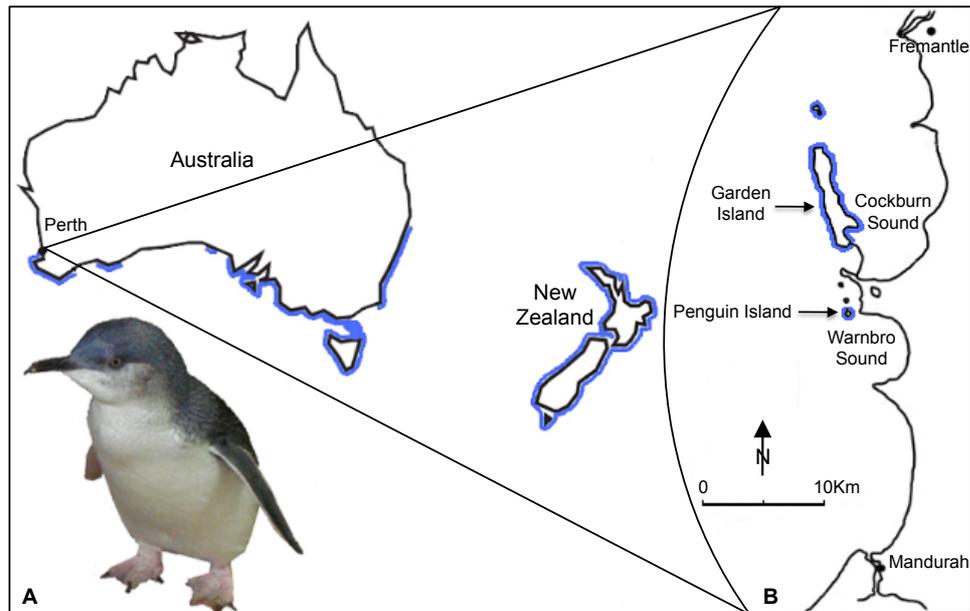


Figure 2.2.1 *Eudyptula minor* distribution and study site for faecal monitoring. (A) The coastal distribution (marked in blue) of *E. minor* across Australia and New Zealand. (B) Map of the study site in Western Australia; for this faecal monitoring study samples were collected from Penguin Island.

Three major DNA-based techniques have been used to varying degrees in the study of species' diet. Firstly, PCR amplification using universal primers with subsequent cloning and sequencing of amplicons, is a technique that has been used extensively in molecular dietary analyses, and to some extent still is (Casper *et al.*, 1997; Deagle *et al.*, 2005; Bohmann *et al.*, 2011). Secondly, quantitative PCR (qPCR), using species-specific primers has been purported to offer great promise in relation to dietary analysis, with the potential to determine estimates of diet composition (Deagle & Tollit, 2007; Matejusová *et al.*, 2008; Bowles *et al.*, 2011). Thirdly, a number of recent studies have highlighted the potential impact that High-Throughput

Sequencing (HTS) may have on dietary studies. HTS has been proposed as a cost-effective alternative in assessing and quantifying species' diet (Soininen *et al.*, 2009; Deagle *et al.*, 2010; Bohmann *et al.*, 2011), and using indexed primers enables a large number of samples to be processed in parallel (Binladen *et al.*, 2007; Valentini *et al.*, 2009b; Bohmann *et al.*, 2011). As yet, however, no study has validated the use of HTS in providing quantitative estimates similar to those obtained via qPCR.

This study sets out to determine the composition of Little Penguin faecal samples by comparing cloning, qPCR and HTS approaches. The primary purpose of this study was to develop an effective long-term strategy for the continual monitoring of diet in the penguin population. However, it is envisaged that the approach and recommendations advocated here will assist in experimental design for DNA-based faecal monitoring across a wide diversity of species.

2.2.3 Materials and methods

The handling of penguins and the collection of faecal samples was conducted by experienced handlers under a strict set of animal ethics guideline approved by the Murdoch University Animal Ethics Committee (permit no. W2002/06).

2.2.3.1 Sample collection & storage

A total of 47 penguin faecal samples were collected, for cloning analysis, over the period from August 2008 until September 2009 and a further 52 samples, for HTS and qPCR analyses over the period from October to December 2010. All samples were collected from free-living penguins inhabiting the study area (Figure 2.2.1). Samples were collected opportunistically from adults and chicks by checking artificial nest boxes or by intercepting penguins returning from the ocean to their nests. Adult penguins were placed in plastic-lined containers for a maximum of 15 minutes. Chicks were placed in a smaller container with a hot water bottle for a maximum of 15 minutes before being returned to their nest boxes. Upon collection the faecal samples were placed in a labelled vial and then stored at -20°C within 12 hours. All handling and sampling was carried out under Murdoch University Animal Ethics Committee permit W2002/06.

2.2.3.2 Sample preparation and DNA extraction

The penguin samples were extracted in batches with the appropriate extraction controls. Samples were weighed and collected into 2mL tubes, with between 26-330mg of sample being used in each extraction depending on the condition of the faecal material. Extractions were performed using QIAamp DNA Stool Mini Kit (QIAGEN) as per manufacturer's instructions. DNA was eluted in 100µL of AE buffer and dilutions of 1:10 and 1:50 were made using Milli-Q UV Pure H₂O for subsequent PCR reactions. DNA extracts were stored at -20°C until further analyses were performed.

2.2.3.3 Sample screening and initial quantification

Each faecal extract was screened using qPCR with 16S1F/2R primers in order to assess the DNA quality, quantity and to detect any possible PCR inhibition (Deagle *et al.*, 2007) (Table 2.2.1). Each extract was amplified at neat, 1:10 and 1:50 dilutions using the ABI Step One Real Time PCR machine. Each reaction was made up to 25µL, containing 12.5µL Power Sybr master mix (Applied Biosystems), 0.4µM of each primer, 8.5µL H₂O and 2µL DNA. Reaction conditions were as follows: initial heat denaturation at 95°C for 5mins, followed by 40 cycles of 95°C for 30s; 54°C for 30s; 72°C for 45s followed by final extension at 72°C for 10mins and a 1°C melt curve to assist in the identification of primer dimer and non-specific amplification.

2.2.3.4 Cloning of amplified DNA

PCR products were cloned into pGEM®-T vectors (Promega) following the manufacturer's protocol and a maximum of 10 positive clones were selected per sample and amplified using the M13F/M13R primer set. Each 25µL reaction contained 1X PCR buffer, 2mM MgCl₂, 0.4mg/mL BSA, 0.25mM each dNTP, 0.6µL SYBR Green (Invitrogen), 0.4µM of each primer, 0.25µL *Taq* polymerase and 2.0µL of template DNA. The cycling conditions were as follows: initial denaturation at 94°C for 5mins, followed by 35 cycles at 94°C for 15s; 55°C for 30s; 72°C for 30s. Amplicons were purified using an ACROPrep 10K 96 well plate (Pall) under a 25mmHg vacuum and screened via gel electrophoresis. Amplicons of the correct size were sequenced by Macrogen (Korea) using BigDye sequencing chemistry (Applied Biosystems) and analysed using Geneious v5.4.6 (Drummond *et al.*, 2011).

2.2.3.5 HTS library preparation

Prior to amplicon sequencing on the GS-Junior (454 Life Sciences), the 16S1F and 16S2R-degenerate primers were modified into fusion primers consisting of a GS FLX Titanium Primer A or B on the 5' end followed by one of 25 different 6bp Multiplex Identifier (MID) tags (allowing the simultaneous processing of 25 different PCR products) and then the template specific primer at the 3' end (Roche, 2009).

Extracts that successfully yielded DNA, as determined by the initial screening via qPCR, were assigned a unique tagged primer set. Fusion tagged PCR was carried out in 25µL reactions containing 1X PCR Gold Buffer, 2.5mM MgCl₂, 0.4mg/mL BSA, 0.25mM each dNTP, 0.4µM of each primer, 0.25µL AmpliTaq Gold (Applied Biosystems) and 2µL DNA. The cycling conditions were as follows: initial heat denaturation at 95°C for 5mins, followed by 40 cycles of 95°C for 30s; 54°C for 30s; 72°C for 45s followed by final extension at 72°C for 10mins. Amplicons were always generated in duplicate and pooled together to minimise the effects of PCR stochasticity. The resultant pooled amplicons were purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), and eluted in 40µL H₂O. Purified amplicons were electrophoresed on 2% agarose gel and amplicons were pooled in approximately equimolar ratios based on band intensity.

2.2.3.6 GS-Junior set-up and sequencing

To achieve the desired bead:template ratio, pooled amplicons were quantified using a synthetic 200bp oligonucleotide standard (of known molarity) with the Roche A and B primers engineered at either end. Quantitative PCR on a dilution series of both the standard and the pooled library, each run in duplicate, has enabled us to reproducibly normalise bead:template ratios. All procedures involved in the set up of the sequencing run (emulsion PCR and bead recovery), including the sequencing run itself, were carried out according to the Roche GS Junior protocols for amplicon sequencing (<http://www.454.com>).

2.2.3.7 Four fish qPCR assay

Based on previous diet studies (Klomp & Wooller, 1988; Cannell, 2001) and the DNA sequence data it was apparent that *Engraulis australis* (Australian Anchovy),

Spratelloides robustus (Blue Sprat), *Sardinops sagax* (Australian Pilchard) and *Hyperlophus vittatus* (Sandy Sprat) formed a major part of the Little Penguins' diet. Therefore, in order to quantitatively assess the abundance of each of these species within each faecal sample and also to compare the quantitative nature of HTS using degenerate primers to that of qPCR, species-specific primer pairs (Table 2.2.1) were designed for each of the four fish species using Geneious v5.4 (Drummond *et al.*, 2011). Primer sets for the four fish were designed using regions within the mitochondrial genes encoding for 16S rRNA based on sequence data obtained from local fish. Each primer pair was tested for efficiency and sensitivity on their target fish species. Importantly, the primer pairs were selected only if they did not cross-react with each other or other species detected in the area (Klomp & Wooller, 1988; Dept., 2008/2009). Once primer pairs were optimised, qPCR of faecal samples that successfully yielded DNA were performed in 25µL reactions containing 1X PCR Gold Buffer, 2.5mM MgCl₂, 0.4mg/mL BSA, 0.25mM each dNTP, 0.4µM of each primer, 0.25µL AmpliTaq Gold and 0.6µL SybrGreen (Invitrogen cat no S7563, 1:2000 dilution). Cycling conditions were as follows; initial denaturation at 95°C for 10min, followed by 40 cycles of 95°C for 15sec; 60°C for 45 sec.

Table 2.2.1 List of primer pairs used in this study. Primers listed include species specific pairs (*) used in the targeted four fish qPCR assays and the universal pairs (†) used in cloning and High Throughput Sequencing approaches.

Target species	Primer name	Sequence (5'-3')	Product Size (bp)	Annealing temp. (°C)	Reference
<i>Engraulis australis</i> (Australian Anchovy)	AN1F* AN2R*	CCTAAATACCCGCAGCCTTAT CAACTCTCGGCTTAAGGTTT	101	60	This study
<i>Spratelloides robustus</i> (Blue Sprat)	BS2F* BS2R*	GCGGCTACTGCCCTAACTATCGC CTGAGCTCCAGGCCGAAGGC	109	60	This study
<i>Sardinops sagax</i> (Australian Pilchard)	PIL1F* PIL1R*	CCTAACTGGAGCCCAAC GCTGTGGCTCTGGGTTTTAG	117	60	This study
<i>Hyperlophus vittatus</i> (Sandy Sprat)	SS2F* SS2R*	GGCCTCAAACAACATGACAGT TAGGGTGGCCCTAATCCACT	91	60	This study
All prey	16S1F-degenerate† 16S2R-degenerate†	GACGAKAAGACCCTA CGCTGTTATCCCTADRGTAACT	180-270	54	Deagle <i>et al.</i> , 2007

*Note the 16S1F/16S2R primers had 5' fusion and MID tags (Roche, 2009) if they were to be sequenced on the GS-Junior.

2.2.3.8. Data analysis

FASTA (.fna) and Quality (.qual) sequence files obtained from the GS FLX Junior sequencing runs were processed using the following programs; BARTAB (Frank, 2009) de-convoluted the reads into sample batches using a map file containing sample and primer-MID tag information, cross_match (de la Bastide & McCombie, 2007) masked the primer and MID-tag sequences contained in the map file, trimseq (Rice *et al.*, 2000) trimmed the masked primer and MID-tag sequences, and finally each sample of batched reads was then searched using BLASTN (Altschul *et al.*, 1990) without a low complexity sequence filter against the NCBI GenBank nucleotide database (Benson *et al.*, 2006). This was automated in the Internet-based bioinformatics workflow environment, YABI (<https://ccg.murdoch.edu.au/yabi/>). The BLAST results that were obtained using YABI were imported into MEtaGenome Analyzer (MEGAN) where they were taxonomically assigned using the LCA-assignment algorithm (parameters included: min. bit score = 65.0, top percentage = 10%, min. support= 1) (Huson *et al.*, 2007). Where MEGAN was unable to resolve the taxonomy of a sequence (due to multiple species' sequences matching the query sequence), taxonomies were assigned using a combination of FishBase (<http://fishbase.org>) and Atlas of Living Australia (<http://www.ala.org>) to determine the most likely species based on their geographic distribution. Where more than one species returned by GenBank occurred around the Perth coastal area the query sequence was assigned to a higher taxonomic level.

Upon successful classification of all sequences obtained via HTS the percentage contribution of each prey item identified within each faecal sample was calculated, in addition to the overall contribution of each prey item across all faecal samples. In the case of the cloning data, a presence/absence method was used to determine the abundance of prey items within faecal samples.

In order to calculate the percentage contribution of each of the four major fish species within each faecal sample during the Oct '10-Dec '10 sampling period, the C_T (Cycle threshold) values obtained for the four target species via qPCR (at the same dilution if deemed free of inhibition) were compared and converted into a percentage relative to each other. These individual percentages were then used to calculate the overall proportion of each of the four fish species across all faecal

samples. Due to the stochasticity associated with low copy number DNA and primer dimer accumulation above C_T values of 34, all C_T values recorded above this level were attributed a C_T value of 34. This approach enables the target amplicon's presence to be acknowledged, whilst still allowing for it to be expressed proportionally to the other fish species within that sample.

To enable comparison of the qPCR and HTS datasets, the proportions of each of the four major fish species within each faecal sample as determined via HTS were considered to the exclusion of all other prey species detected. Using these data in conjunction with that obtained via qPCR, the Pearson product-moment correlation coefficient (Pearson's r) was calculated to determine the degree of correlation between the datasets. Additionally, individual paired sample t -tests for each major fish species were used to determine if there was a significant difference between the data obtained via both methods for any of the four major fish species. Samples that recorded C_T values >34 were excluded from statistical analyses, due to the stochasticity of qPCR above this threshold. All statistical analyses were carried out using the program R.

2.2.4 Results & Discussion

2.2.4.1 Overview and comparisons of Cloning and HTS approaches

Using the cloning approach, a total of nine fish species were identified from 129 sequences, in 22 of the 47 samples (47%) collected during the Aug '08-Sep '09 sampling period. Samples deemed to have failed either yielded no amplifiable DNA, were severely compromised by inhibitors, or had target copy numbers (as determined by qPCR C_T values >35.0) that were considered too low to be reliable. The dominant prey species detected within these samples was *H. vittatus*, present in 32% of samples, followed by *S. robustus*, found in 20% of samples, with *S. sagax*, *E. australis* and *Sardinella lemuru* (Scaly Mackerel) each found in 9.8% of samples (Figure 2.2.2A). A number of other minor prey items were also identified, however they were found to represent a small proportion of sequences (Figure 2.2.2A).

Of the 52 samples collected during the Oct '10-Dec '10 sampling period, only 27 samples (52%) were deemed to have yielded DNA of sufficient quality free of

inhibition (determined by qPCR) that they could advance to HTS analysis. The two independent GS-Junior runs generated a total of 7810 DNA sequences. Of these sequences ~93% were unambiguously attributed to eleven fish species and <0.1% were identified as belonging to the genus *Pelates* (Striped Grunters). There were low levels of human contamination and penguin DNA (~3%) and unassigned/uninformative sequences accounted for ~3.6% of sequences. There was notable variation in the number of sequences generated for each faecal sample (range= 35-1055), and this is likely due to inaccurate blending of amplicons (see Section 2.2.3). However, an average of ~300 reads per sample is more than sufficient coverage for dietary audits, especially when compared to the average number of sequences often generated per sample using bacterial cloning (Clare *et al.*, 2009; Kim *et al.*, 2011). HTS of the Oct '10-Dec '10 samples revealed that, of the prey items identified, *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* were the major species present within the faecal material, each contributing 49%, 32%, 11% and 5% respectively (Figure 2.2.2B). The remaining fish identified were minor contributors to the overall composition of the samples (ranging from 0.02% to 1.9%) (Figure 2.2.2B) and only in one sample did any of these fish constitute a significant proportion of the prey detected, that of PEN_42, where *Parequula melbournensis* (Silverbelly) contributed 48% to the sample composition for this individual (Table S2.2.1).

It is clear from the bacterial cloning and HTS data that there were four dominant fish species detected within the samples at this study site, those being *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* (Figure 2.2.2). The occurrence of other minor contributing prey items within the samples is consistent with previous findings and reflects the opportunistic feeding behaviour of the Little Penguins (Klomp & Wooller, 1988; Bradley *et al.*, 1997). A direct comparison of cloning and HTS is somewhat hampered by the fact that different faecal samples from different time periods were used for each method. However, it is clear that a number of important conclusions can be drawn from both datasets. Both methods provide a clear picture of the major prey species that are present within the collective faecal samples. Where they differ is in the relative contribution of each of these individual species (Figure 2.2.2), however this could be a result of temporal effects as it is well documented that the diet of Little Penguins varies throughout the year (Klomp & Wooller, 1988).

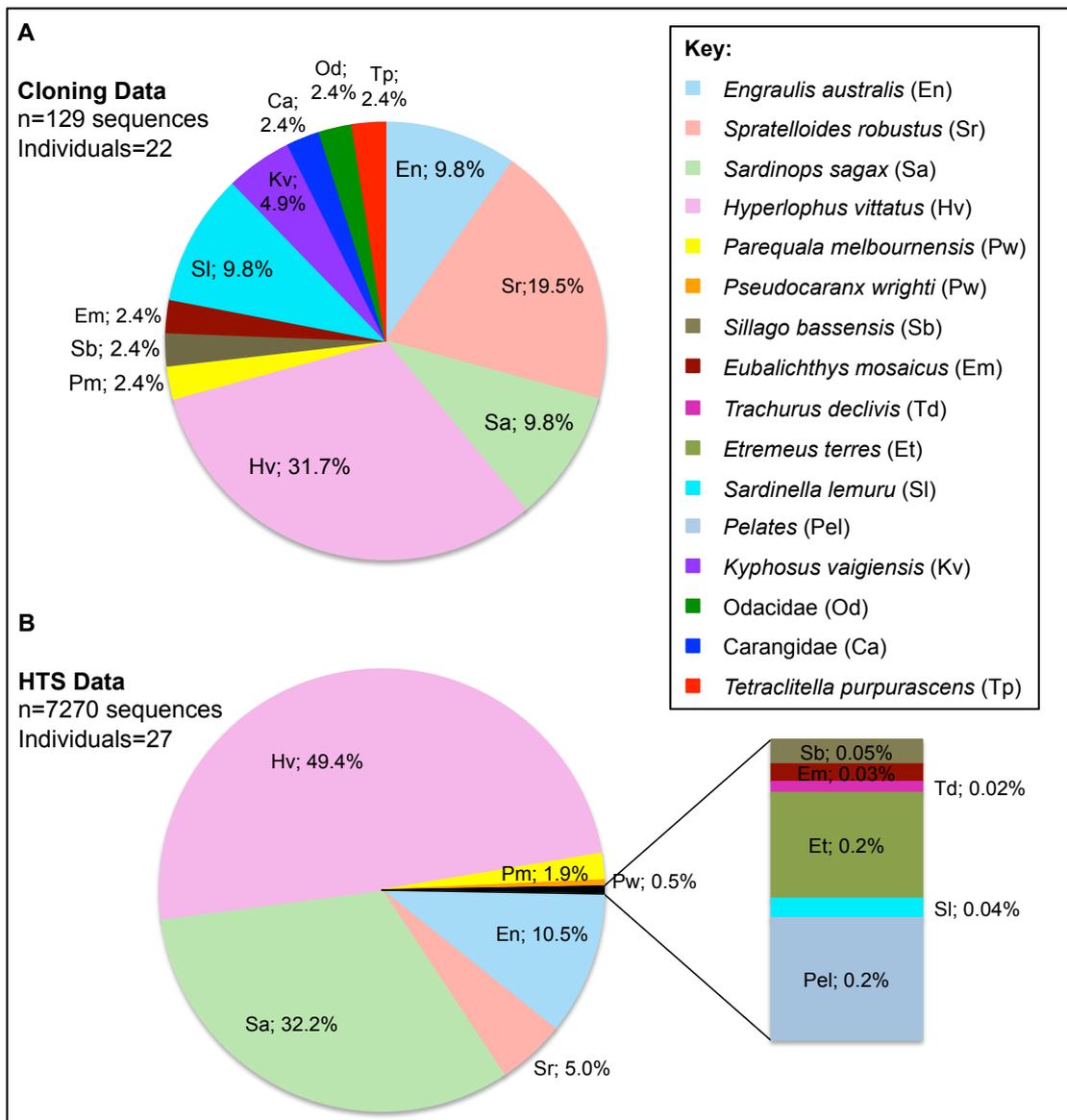


Figure 2.2.2 Percentage contribution of identified prey items in the faecal DNA of *E. minor*. (A) Graph showing fish identifications based on 16S rRNA sequence data obtained via cloning using universal primer set 16SF1/16S2R. Faecal samples (n=22) for this study were collected during the Sep '08/Aug '09 period. (B) Penguin faecal samples collected during Oct '10-Dec '10 period (n=27) that were audited using HTS methods. The 16SF1/16S2R set were MID-tagged and a total of 7270 sequences were assigned to prey items.

Cloning of universally amplified PCR products using bacteria, followed by DNA purification and Sanger sequencing is both expensive and time consuming. An additional issue, not entirely observed in this study, is that large numbers of clones are required in order to detect rare species (Clare *et al.*, 2009; Clare *et al.*, 2011), with the associated time and expense being inefficient for long-term monitoring of species' diet. For this reason, our Little Penguin monitoring program made the transition to HTS for the 2010 samples. Newly developed HTS platforms, especially small-scale systems such as the GS-Junior or IonTorrent, enable a quick, efficient and relatively inexpensive way to deep-sequence PCR amplicons generated from faecal DNA extracts (Soininen *et al.*, 2009; Deagle *et al.*, 2010; Bohmann *et al.*, 2011). Moreover, the use of MID-tagged primers makes it possible to run numerous samples in parallel, enabling not only an overview of the diet composition across a population, but also at the individual level (Valentini *et al.*, 2009b; Bohmann *et al.*, 2011). HTS can provide a wealth of information; greatly increasing the number of DNA sequences returned (129 sequences vs 7810 sequences) for a fraction of the labour and associated costs. Concomitant with the increases in sequencing depth is the prospect that HTS data might now provide better quantitative measures of the DNA targets within faecal material, much like estimates obtained using qPCR (Deagle *et al.*, 2007; Bowles *et al.*, 2011).

2.2.4.2 Overview of qPCR approach

In order to compare the quantitative nature of HTS to that of qPCR, a species-specific four fish qPCR assay was designed to estimate the relative abundance of each of the four major prey species determined within the collective samples (Figure 2.2.2, Table 2.2.1). Careful development of each of the four primer pairs was critical to data fidelity (Sipos *et al.*, 2007; King *et al.*, 2008), as was ensuring that the DNA extracts' C_T values behaved as desired when diluted (i.e. they were free from inhibition). From this four fish assay it was clear that *H. vittatus* and *S. sagax* were major constituents of the faecal samples; 49% and 32% respectively, with both *E. australis* and *S. robustus* each contributing 13% and 5% to the overall composition (Figure 2.2.3A). The ANF1/ANR2 assay encountered some primer dimer issues at low template copy numbers, however the melt curves enabled differentiation of product and dimer. Although not wholly representative of the *total* amount of prey

DNA within samples, the qPCR assays gave a good indication of the abundance of each of the four major fish species relative to each other.

2.2.4.3 Comparison of HTS and qPCR approaches

It is important to actively compare and contrast both HTS and qPCR approaches to enable an informed decision of the most suitable method to be used for genetic faecal screening. To allow a comparison between both approaches, the HTS data had to be transformed to focus on the same four fish species as the qPCR assay; *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus*. The proportion of these species to the exclusion of the other species present was determined to be 52%, 32%, 11% and 5% respectively (Figure 2.2.3B transformed from Figure 2.2.2B data). It is clear that there is a striking degree of similarity between the proportions identified for the four fish species determined by qPCR and HTS (Figure 2.2.3C). In order to investigate this further, the absolute differences between the results obtained individually by both methods were calculated. In the case of each fish species the overall difference in percentage abundance between the two techniques was negligible (*H. vittatus* - Median= 0.02, n= 19; *S. sagax* - Median= 0.31 n= 13; *E. australis* - Median= -0.18, n= 15; *S. robustus* - Median= -0.05, n= 7) (Figure 2.2.3C). These initial results demonstrate a high degree of similarity between individual measures obtained by both methods. Furthermore, Pearson's *r* calculations revealed strong correlations between both methods for all four fish species (*H. vittatus* – Pearson's *r* = 0.976, n= 19; *S. sagax* - Pearson's *r* = 0.996, n= 13; *E. australis* - Pearson's *r* = 0.973, n= 15; *S. robustus* - Pearson's *r* = 1.0, n= 7)* (Figure 2.2.4), whilst individual paired *t*-tests revealed no significant difference between the values obtained by either method for any of the major prey species (*H. vittatus* – *p* = 0.215, n= 19; *S. sagax* - *p* = 0.226, n= 13; *E. australis* - *p* = 0.100, n= 15; *S. robustus* - *p* = 0.266, n= 7).

**addendum*: all Pearson's correlation tests were statistically significant with $p < 0.001$ in each case.

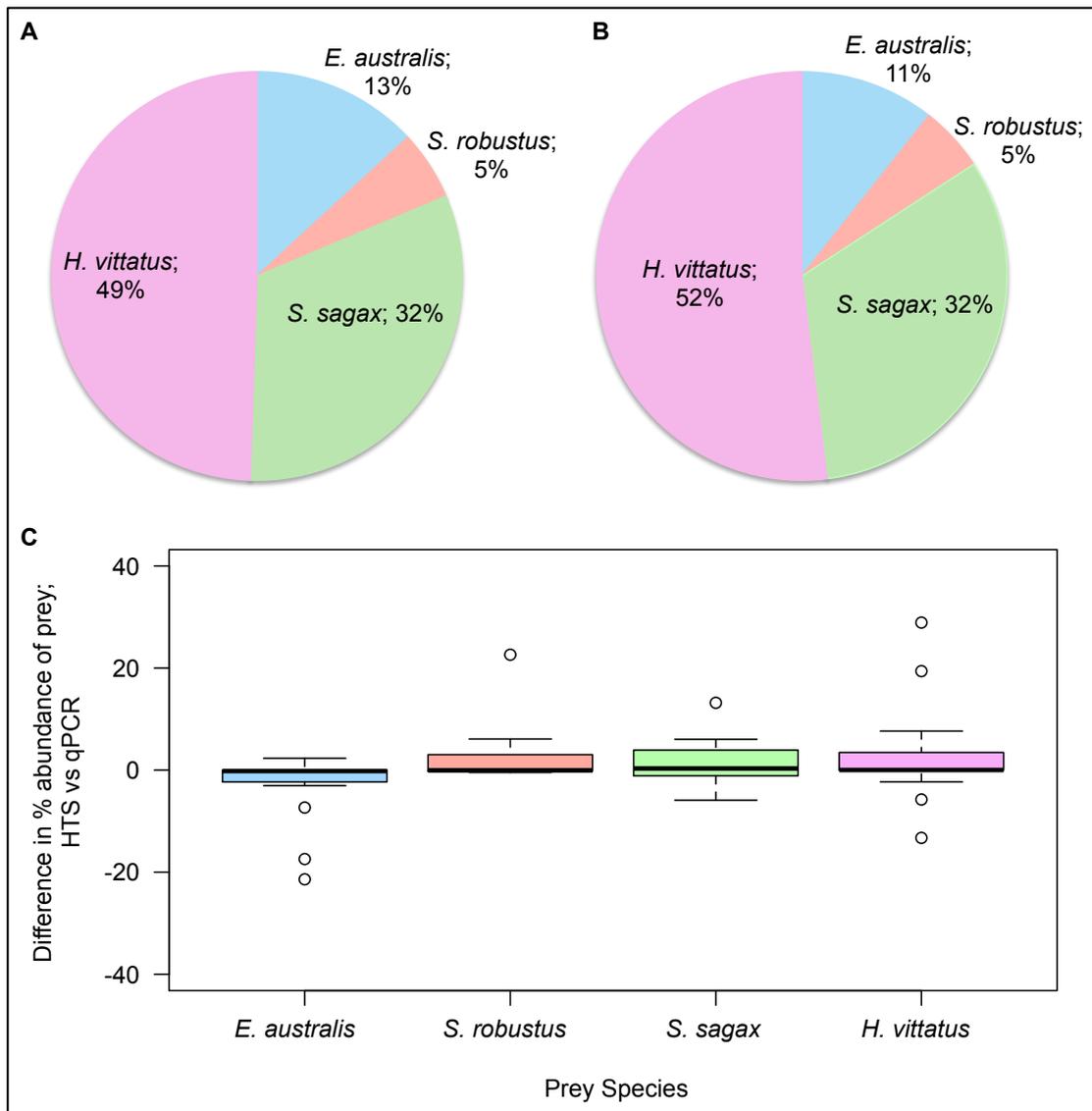


Figure 2.2.3. Comparison of HTS and qPCR methods determining the proportion of four major fish species. Graphs indicate the relative percentage composition of *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* within faecal samples of *E. minor* on Penguin Island, as determined by **(A)** qPCR and **(B)** HTS of samples collected during the period of Oct '10-Dec '10. **(C)** Box plot showing the difference between the results obtained by HTS and qPCR for each of the four major fish species found in the diet of *E. minor*. Samples whose C_T values were >34 have been excluded from the dataset (see Materials and Methods section 2.2.3.8).

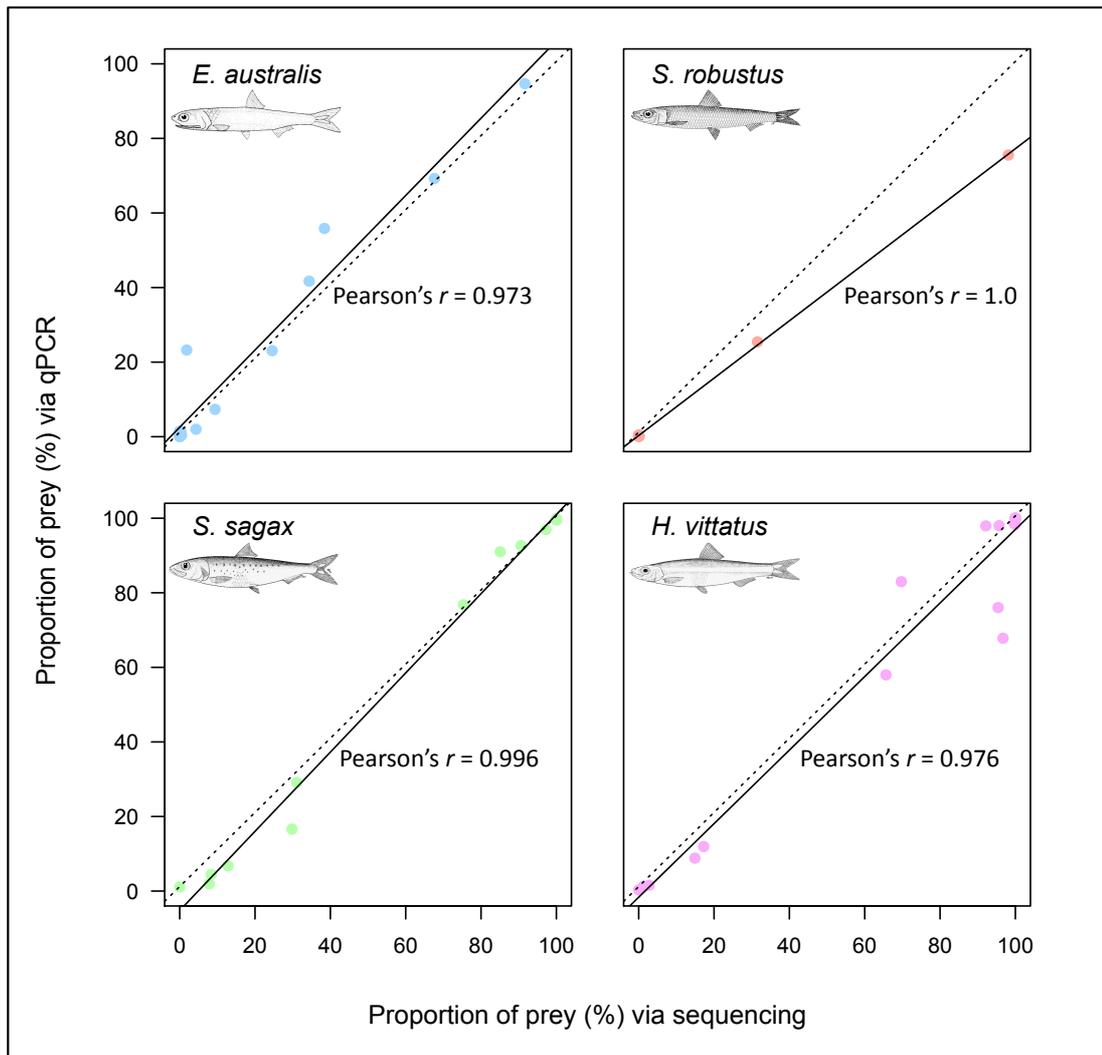


Figure 2.2.4. Correlation between four-fish data obtained via HTS and qPCR.*

Scatterplots include the percentage contributions obtained for each individual penguin via HTS and qPCR for each of the four major fish species detected within faecal samples. Solid line represents the line of best fit for individual species (Pearson's r values are shown), whilst the dotted line represents the overall correlation between both datasets with the data obtained for all fish species across all samples combined. Samples whose C_T values were >34 have been excluded from the dataset (see Materials and Methods). Fish images used in this figure can be reproduced freely for non-commercial purposes and are sourced from (Whitehead, 1985).

**addendum:* all Pearson's correlation tests were statistically significant with $p < 0.001$ in each case.

Although no statistical difference was detected in species composition in the combined analysis, it was apparent that there are slight differences between the datasets at the individual level (Table S2.2.2). There could be a number of reasons for such differences. Firstly, differential degradation of prey tissue DNA could account for some of the variance between datasets (Deagle *et al.*, 2005; Bowles *et al.*, 2011). In this study the amplicon sizes produced by the primer sets in qPCR were shorter than those for HTS (see Table 2.2.1), and so in some instances length biases may be present, especially in instances where there is differential degradation of prey tissue DNA in the gastrointestinal tract (Deagle & Tollit, 2007). Indeed, it would appear that in this study *E. australis* was slightly over-represented in qPCR relative to HTS, whilst *H. vittatus* was marginally under-represented in qPCR relative to HTS (Table S2.2.2). A second potential cause could be the fact that the targeted qPCR assay is more efficient than the universal 16S primers used in HTS, therefore enabling the detection of the four prey species' DNA at lower template amounts. This is best illustrated when considering the presence/absence data, where HTS vs qPCR detection rates are compared: 70.4% vs 88.9% (*H. vittatus*), 48.2% vs 81.5% (*S. sagax*), 40.7% vs 74.1% (*E. australis*) and 14.8% vs 40.7% (*S. robustus*). In all cases where a species was detected in qPCR but not in HTS the C_T values were either >34 or the relative abundance of that species was below 1.5% (Table S2.2.2). Taken together, these data do suggest that the shorter, targeted qPCR assays were, across all four fish species, more sensitive to low template amounts. However, the higher qPCR detection success did not drastically affect the overall estimates of both methods, due to the low abundance of prey species in these instances. This also highlights a very important advantage of species-specific qPCR over HTS, in that it can detect species at very low DNA abundances, whereas the nature of universal primers, such as those used in HTS, renders them less specific and less likely to efficiently amplify low copy number targets in the presence of abundant targets.

Whilst it is clear that there are slight differences between both methods, which are attributable to a variety of factors, it is also clear that in this case no single factor seemed to have a detrimental effect on the overall estimates of prey items within the collective faecal samples. It appears, however, that the difficulty arises when the penguins are considered on an individual basis. If, for instance, HTS were solely

used in this study then it is quite clear that a good idea of the overall breadth of species could be ascertained. However, in some cases the use of universal primers may result in the non-detection of certain dietary constituents, if present in low abundance. On the other hand, with the use of the targeted qPCR approach a possibly more accurate estimate of the relative contribution of the major fish species' DNA could be determined across the population and individually, provided an *a priori* knowledge of diet is known. However, the contribution of the other minor constituents is overlooked. It would appear that the effect of this is largely minimal, unless, as was the case with sample PEN_42, one of the 'minor contributors' accounts for a large proportion, or all, of any given sample.

2.2.4.4 Recommendation for future experimental design

The uptake of genetic techniques to analyse faecal material has provided important insights into animal diet. It is clear that the use of qPCR and the advent of affordable HTS technologies are proving to be a welcome addition to this field of research. Both of these techniques have the potential to eclipse the more traditional molecular methodology of bacterial cloning and/or direct sequencing, which is costly, laborious and time-consuming. In light of the results of this study, it is fair to assume that qPCR and HTS represent the best approaches currently available.

A key component of experimental design in this study was the methodical preparation and selection of samples for DNA extraction prior to qPCR or HTS. The extraction of DNA from faecal samples and the screening of samples for copy number and inhibition is a major bottleneck in the lab. However, the importance of this screening process cannot be under-stated, particularly when the samples being dealt with are complex, heterogeneous substrates containing severely degraded DNA in low copy numbers (Taberlet *et al.*, 1999; Deagle *et al.*, 2006). The initial qPCR screening strategy implemented in this study allowed the identification of suitable samples and DNA extract dilutions that contained the maximum concentration of amplifiable DNA and yet were inhibition free. There is no substitute for prior screening of samples; the congruence of qPCR and HTS in this study can be attributed largely to the fact that there is confidence in the amplifiability of the DNA extract dilution on which HTS and qPCR was conducted.

The ultimate choice of which method to opt for should be considered on a case-by-case basis, although the use of both methods in tandem would be the preferred option. If, for instance, an *a priori* knowledge of the species' diet in question were lacking then it would be more appropriate to use HTS with universal primer sets, thus giving an overview of the animal's diet. With this broad view of the animal's diet it can then be decided whether to pursue the use of targeted primers via the qPCR approach. If the number of prey species within the diet is of limited complexity qPCR may be preferable. Although not implemented here, in theory the quantitiveness of HTS using universal primers could be improved by using multiple universal primer sets in parallel (Deagle *et al.*, 2009; Deagle *et al.*, 2010).

If the goal of any dietary study is the long-term monitoring of diet, then it would be advisable to use HTS to determine the overall composition of the diet, and if possible a subsequent targeted qPCR approach to examine major prey items, to ensure that the diet remains consistent throughout the period of study. Ideally it would be beneficial to consider the use of both techniques in parallel to safeguard against erroneous results, as the removal of major contributors to the diet can have profound impacts on prey quantification. This is highlighted by the example of PEN_42 where *P. melbournensis* formed a major part of that individual penguin's faecal sample (Table S2.2.1). Therefore, in this case, the four fish qPCR assay is a poor representation of prey abundance.

Irrespective of the chosen method, primer design is crucial to the sensitivity of PCR, and careful consideration should be given to the design and testing of primers (King *et al.*, 2008). In the case of universal primers used in HTS, it is imperative that they are designed to allow taxonomic discrimination of amplicons, and yet also amplify a small enough region to circumvent issues of DNA degradation within faeces (King *et al.*, 2008). One additional issue is the fact that the coverage of certain animal groups in certain databases is not complete which will always make taxonomic assignments difficult (Bohmann *et al.*, 2011; Clare *et al.*, 2011). The study of bats is a case in point; in this instance the use of qPCR assays would not be able to account for the hundreds of insect species in bat guanos, however qPCR could still be used to validate the relative portion of a few target species (Bohmann *et al.*, 2011; Clare *et al.*, 2011).

The validation of the quantitative nature of HTS, as compared to qPCR, to detect the DNA in faecal material, bodes well for future dietary studies. However, it is acknowledged that the results obtained via DNA-based faecal analysis are not always directly correlated with the biomass of prey consumed (Sipos *et al.*, 2007) – a recent study referred to them as semi-quantitative at best (Bowles *et al.*, 2011). Much work is yet to be done to enable accurate reconstructions of the physical diet as estimates are currently confounded by a range of factors including; differential digestion rates of prey between species; DNA per unit biomass variability between tissues and the developmental stage of the prey species to name but a few issues (Pegard *et al.*, 2009; Valentini *et al.*, 2009b; Bowles *et al.*, 2011). It is also questionable whether digestion/faecal studies of captive birds will accurately recreate what is happening in the wild. Despite the many caveats regarding actual dietary intake, the accurate quantification of prey DNA actually contained in faecal matter represents an important developmental step.

2.2.5 Conclusion

Characterising the DNA preserved in faecal material is a powerful way to study both animal diet and also provide broader insights into ecosystem composition and health. In light of recent advances in DNA sequencing it was unclear which genetic auditing method(s) should be adopted for a multi-year monitoring program of Little Penguins. The results of qPCR and HTS approaches tested in this study demonstrate that the two methods are capable of generating high-fidelity datasets with no statistical difference between them. In the case of penguin diet, the use of both methods in parallel proved particularly useful with species-specific qPCR assays having better sensitivity, whilst HTS is able to detect species not targeted by qPCR. It is anticipated that the data and approaches presented here will be of benefit to other researchers intending to implement dietary monitoring programs and will assist in improving the accuracy of environmental audits based on faecal material.

2.2.6 Acknowledgements

We would like to thank staff at Perth Zoo and the Penguin Island Experience for their involvement in the initial control trials where molecular methodologies were developed. The authors acknowledge the support of Ms Frances Brigg at the State Agricultural Biotechnology Centre DNA sequencing facility and the iVEC Informatics Facility, for computational support.

2.2.7 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research*, 34, D16-D20.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, 2, e197.

Boersma, P. D. (2008). Penguins as Marine Sentinels. *BioScience*, 58, 597-605.

Boersma, P. D., Rebstock, G. A., Frere, E., & Moore, S. E. (2009). Following the fish: penguins and productivity in the South Atlantic. *Ecological Monographs*, 79, 59-76.

Bohmann, K., Monadjem, A., Lehmkuhl, N., Rasmussen, M., Zeale, M. R. K., Clare, E., . . . Gilbert, M. T. P. (2011). Molecular diet analysis of two African Free-tailed Bats (Molossidae) using High Throughput Sequencing. *PLoS One*, 6, e21441.

Bowles, E., Schulte, P. M., Tollit, D. J., Deagle, B. E., & Trites, A. W. (2011). Proportion of prey consumed can be determined from faecal DNA using real-time PCR. *Molecular Ecology Resources*, 11, 530-540.

Boyer, S., Wratten, S. D., Holyoake, A., Abdelkrim, J., & Cruickshank, R. H. (2013). Using next-generation sequencing to analyse the diet of a highly endangered

land snail (*Powelliphanta augusta*) feeding on endemic earthworms. *PLoS One*, 8, e75962.

Bradley, J. S., Cannell, B. L., & Wooller, R. D. (1997). A radio-tracking study of the movements at sea and diet of Little Penguins *Eudyptula minor* breeding on Penguin Island, Western Australia. *Final report for Bowman Bishaw Gorham*.

Cannell, B., Bunce, M., Murray, D., Pollock, K., & Valesini, F. (2015, 26-30 October). *Divorcing immediate and longer term impacts of a marine heatwave on Little Penguins from the effects of coastal development- is it possible?* Paper presented at the 2nd World Seabird Conference, Cape Town, South Africa, available from <http://www.worldseabirdconference.com>.

Cannell, B., Chambers, L., Bunce, M., & Murray, D. (2011, 1 - 5 July). *Little Penguins (Eudyptula minor) breeding and survival in Western Australia compromised by a "marine heat wave" in 2011*. Paper presented at the Australian Marine Sciences Association and New Zealand Marine Sciences Society Joint Conference, Hobart, Tasmania. available from <http://researchrepository.murdoch.edu.au/17919/>.

Cannell, B. L. (2001). Status of Little Penguins in Western Australia: A management review. *Department of Conservation and Land Management Report MMS/LNE/SIS-40/2001*.

Casper, R. M., Jarman, S. N., Deagle, B. E., Gales, N. J., & Hindell, M. A. (1997). Detecting prey from DNA in predator scats: A comparison with morphological analysis, using *Arctocephalus* seals fed a known diet. *Journal of Experimental Marine Biology and Ecology*, 347, 144-154.

Chape, S. (1984). *Penguin Island draft management plan*. Perth: National Parks Authority and Dept. of Conservation & Environment.

Chiaradia, A., Costalunga, A., & Kerry, K. (2003). The diet of Little Penguins (*Eudyptula minor*) at Phillip Island, Victoria, in the absence of a major prey - Pilchard (*Sardinops sagax*). *EMU*, 103, 43-48.

- Clare, E. L., Barber, B. R., Sweeney, B. W., Hebert, P. D. N., & Fenton, M. B. (2011). Eating local: influences of habitat on the diet of little brown bats (*Myotis lucifugus*). *Molecular Ecology*, *20*, 1772-1780.
- Clare, E. L., Fraser, E. E., Braid, H. E., Fenton, M. B., & Hebert, P. D. N. (2009). Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. *Molecular Ecology*, *18*, 2532-2542.
- Dann, P., & Chambers, L. (2009). Climate change and Little Penguins. *Report for Western Port Greenhouse Alliance*.
- de la Bastide, M., & McCombie, W. R. (2007). Assembling genomic DNA sequences with PHRAP. *Current Protocols in Bioinformatics, Chapter 11*, Unit 11.14.
- Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.
- Deagle, B., & Tollit, D. (2007). Quantitative analysis of prey DNA in pinniped faeces: potential to estimate diet composition? *Conservation Genetics*, *8*, 743-747.
- Deagle, B. E., Eveson, J. P., & Jarman, S. N. (2006). Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Frontiers in Zoology*, *3*, 10.
- Deagle, B. E., Gales, N. J., Evans, K., Jarman, S. N., Robinson, S., Trebilco, R., & Hindell, M. A. (2007). Studying seabird diet through genetic analysis of faeces: A case study on Macaroni Penguins (*Eudyptes chrysolophus*). *PloS ONE*, *2*, e831.
- Deagle, B. E., Kirkwood, R., & Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, *18*, 2022-2038.
- Deagle, B. E., Tollit, D. J., Jarman, S. N., Hindell, M. A., Trites, A. W., & Gales, N. J. (2005). Molecular scatology as a tool to study diet: analysis of prey DNA in scats from captive Steller sea lions. *Molecular Ecology*, *14*, 1831-1842.

Dept., W. A. F. (2008/2009). State of the fisheries report. Hillarys, W.A. : Dept. of Fisheries, Western Australia.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., . . . Wilson, A. (2011). Geneious v5.4, Available from <http://www.geneious.com/>. Retrieved from <http://www.geneious.com/>

Frank, D. (2009). BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics*, *10*, 362.

Harrigan, K. E. (1992). Causes of mortality of Little Penguins *Eudyptula minor* in Victoria. *EMU*, *91*, 273-277.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*, 377-386.

Kim, B. J., Lee, N. S., & Lee, S. D. (2011). Feeding diets of the Korean water deer (*Hydropotes inermis argyropus*) based on a 202 bp rbcL sequence analysis. *Conservation Genetics*, *12*, 851-856.

King, R. A., Read, D. S., Traugott, M., & Symondson, W. O. C. (2008). Molecular analysis of predation: a review of best practice for DNA-based approaches. *Molecular Ecology*, *17*, 947-963.

Klomp, N. I., & Wooller, R. D. (1988). Diet of Little Penguins, *Eudyptula minor*, from Penguin Island, Western Australia. *Australian Journal of Marine Freshwater Research*, *39*, 633-639.

Kowalczyk, R., Taberlet, P., Coissac, E., Valentini, A., Miquel, C., Kaminski, T., & Wojcik, J. M. (2011). Influence of management practices on large herbivore diet- Case of European bison in Bialowieza Primeval Forest (Poland). *Forest Ecology and Management*, *261*, 821-828.

Kuch, M., Sobolik, K., Barnes, I., Stankiewicz, B. A., Spaulding, G., Bryant, V., . . . Pääbo, S. (2001). A molecular analyses of the dietary diversity for three archaic

native americans. . *Proceedings of the National Academy of Sciences, USA*, 98, 4317-4322.

Mallory, M. L., Robinson, S. A., Hebert, C. E., & Forbes, M. R. (2010). Seabirds as indicators of aquatic ecosystem conditions: a case for gathering multiple proxies of seabird health. *Marine Pollution Bulletin*, 60, 7-12.

Marchant, S., & Higgins, P. J. (1990). Ratites to Ducks *In: Handbook of Australian, New Zealand and Antarctic Birds, Vol 1* (Vol. Vol. 1). Melbourne: Oxford University Press.

Matejusová, I., Doig, F., Middlemas, S. J., Mackay, S., Douglas, A., Armstrong, J. D., . . . Snow, M. (2008). Using quantitative real-time PCR to detect salmonid prey in scats of grey *Halichoerus grypus* and harbour *Phoca vitulina* seals in Scotland – an experimental and field study. *Journal of Applied Ecology*, 45, 632-640.

Miller, K. M., & McEwen, L. C. (1995). Diet of nesting Savannah Sparrows in interior Alaska. *Journal of Field Ornithology*, 66, 152-158.

Montague, T. L., & Cullen, J. M. (1985). Comparison of techniques to recover stomach contents from penguins. *Australian Wildlife Research*, 12, 327-330.

Pegard, A., Miquel, C., Valentini, A., Coissac, E., Bouvier, F., & *al., e.* (2009). Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from faeces. *Journal of Agricultural and Food Chemistry*, 57, 5700-5706.

Pichegru, L., Ryan, P. G., Le Bohec, C., van der Lingen, C. D., Navarro, R., Petersen, S., . . . Grémillet, D. (2009). Overlap between vulnerable top predators and fisheries in the Benguela upwelling system: implications for marine protected areas. *Marine Ecology-Progress Series*, 391, 199-208.

Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., . . . Pääbo, P. (1998). Molecular coproscopy: dung and diet of the extinct Ground Sloth *Nothrotheriops shastensis*. *Science*, 281, 402-406.

Raye, G., Miquel, C., Coissac, E., Redjadj, C., Loison, A., & Taberlet, P. (2011). New insights on diet variability revealed by DNA barcoding and high-throughput

pyrosequencing: chamois diet in autumn as a case study. *Ecological Research*, 26, 265-276.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16, 276-277.

Roche. (2009). Guidelines for amplicon fusion primer design for GS FLX Titanium Series Lib-A Chemistry TCB No. 013-2009. *Technical bulletin*, TCB No. 013-2009.

Sheppard, S. K., & Harwood, J. D. (2005). Advances in molecular ecology: tracking trophic links through predator-prey food-webs. *Functional Ecology*, 19, 751-762.

Sipos, R., Szekely, A. J., Palatinszky, M., Revesz, S., Marialigeti, K., & Nikolausz, M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *Fems Microbiology Ecology*, 60, 341-350.

Soininen, E. M., Valentini, A., Coissac, E., Miquel, C., Gielly, L., Brochmann, C., . . . Taberlet, P. (2009). Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Frontiers in Zoology*, 6, 16.

Symondson, W. O. C. (2002). Molecular identification of prey in predator diets. *Molecular Ecology*, 11, 627-641.

Taberlet, P., Waits, L. P., & Luikart, G. (1999). Noninvasive genetic sampling: look before you leap. *Trends in Ecology and Evolution*, 14, 323-327.

Tollit, D. J., Schulze, A. D., Trites, A. W., Olesiuk, P. F., Crockford, S. J., Gelatt, T. S., . . . Miller, K. M. (2009). Development and application of DNA techniques for validating and improving pinniped diet estimates. *Ecological Applications*, 19, 889-905.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E. V. A., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009a). New perspectives in diet analysis based on

DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, 9, 51-60.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E. V. A., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009b). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, 9, 51-60.

Vestheim, H., & Jarman, S. N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, 5, Article No.: 12.

Vila, A. R., & Borrelli, L. (2011). Cattle in the Patagonian forests: feeding ecology in Los Alerces National Reserve. *Forest Ecology and Management*, 261, 1306-1314.

Whitehead, P. J. P. (1985). FAO species catalogue. Vol 7: Clupeoid fishes of the world (suborder Clupeoidei). An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolf-herrings. Part 1 - Chirocentridae, Clupeidae and Pristigasteridae. *FAO Fish Synop*, 125, 1–303.

Wienecke, B. C. (1989). *The breeding patterns of little penguins in Penguin Island, Western Australia, in relation to dietary and oceanographic factors*. Murdoch University, Honours Thesis.

Wienecke, B. C. (1993). *The size and breeding patterns of the Little Penguin *Eudyptula minor* in Australia: a comparative study*. (PhD Thesis), Murdoch University, PhD Thesis.

Wienecke, B. C., Wooller, R. D., & Klomp, N. I. (1995). The ecology and management of Little Penguins on Penguin Island, Western Australia. In P. Dann, Norman, I., Reilly, R. (Ed.), *The penguins: ecology and management* (pp. 440–467). Sydney: Surrey Beatty.

Williams, C., & Buck, C. (2010). Using fatty acids as dietary tracers in seabird trophic ecology: theory, application and limitations. *Journal of Ornithology*, 151, 531- 543.

Wooller, R. D., Dunlop, J. N., Klomp, N. I., Meathrel, C. E., & Wienecke, B. C. (1991). Seabird abundance, distribution and breeding patterns in relation to the Leeuwin Current. . *Journal of the Royal Society of Western Australia*, 74, 129-132.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Table S2.2.1 Percentage contribution of prey items detected by HTS for each faecal sample. The percentage contribution of detected prey items within each individual faecal sample, as determined by HTS of samples collected during the period of Oct '10-Dec '10, using 16SF1/16S2R universal primers.

aDNA extraction no.	Total reads	En	Sr	Sa	Hv	Pm	Pw	Sb	Em	Td	Et	Sl	Pel
PEN_01	176			100.0									
PEN_04	64	9.4		90.6									
PEN_05	811	90.5		8.3		1.2							
PEN_08	114			100.0									
PEN_11	139	4.3			95.7								
PEN_12	96	34.4			65.6								
PEN_14	145			97.2	2.8								
PEN_15	35				100.0								
PEN_16	675		4.1		86.5		9.3						
PEN_17	101			100.0									
PEN_18	208	37.5	30.8	12.5	16.8		2.4						
PEN_19	187				100.0								
PEN_20	268			7.8	92.2								
PEN_21	62	3.2			95.2						1.6		
PEN_22	357				100.0								
PEN_23	119				100.0								
PEN_24	133			100.0									
PEN_28	230			84.3	14.8				0.9				
PEN_32	208	0.5		29.8	69.7								
PEN_35	136				100.0								
PEN_38	0												
PEN_39	113	1.8	93.8			0.9					3.5		
PEN_41	962	24.4	0.1	74.9						0.5			
PEN_42	84				45.2	47.6		1.2					6.0
PEN_50	630	67.0		30.8	1.3							1.0	
PEN_51	162				100.0								
PEN_40B	1055	0.3			99.7								
	Sum of %'s	273.2	128.8	836.4	1285.5	49.7	11.7	1.2	0.9	0.5	5.2	1.0	6.0
	Overall % contribution	10.5	5.0	32.2	49.4	1.9	0.5	0.0	0.0	0.0	0.2	0.0	0.2

En - *Engraulis australis* (Anchovy), Sr - *Spratelloides robustus* (Blue Sprat), Sa - *Sardinops sagax* (Sandy Sprat), Hv - *Hyperlophus vittatus* (Pilchard), Pm - *Parequula melbournensis* (Silverbelly), Pw - *Pseudocaranx wrighti* (Shipjack Trevally), Sb - *Sillago bassensis* (Southern School Whiting), Em - *Eubalichthys mosaicus* (Mosaic Leatherjacket), Td - *Trachurus declivis* (Common Jack Mackerel), Et - *Etrumeus teres* (Round Herring), Sl - *Sardinella lemuru* (Scaly Mackerel), Pel - *Pelates* (Striped Grunter)

Table S2.2.2 Percentage contribution of four major fish species determined by HTS and qPCR methods. The percentage composition of *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* within individual faecal samples of *E. minor* on Penguin Island, as determined by HTS and qPCR of samples collected during the period of Oct '10-Dec '10.

Extraction #	High-Throughput Sequencing				qPCR			
	<i>E. australis</i>	<i>S. robustus</i>	<i>S. sagax</i>	<i>H. vittatus</i>	<i>E. australis</i>	<i>S. robustus</i>	<i>S. sagax</i>	<i>H. vittatus</i>
PEN_01			100.0		0.1	0.5	99.5	
PEN_04	9.4		90.6		7.3		92.7	
PEN_05	91.6		8.4		94.7		4.4	0.9
PEN_08			100.0				100.0	
PEN_11	4.3			95.7	2.0		0.0	98.0
PEN_12	34.4			65.6	41.7	0.3	0.0	58.0
PEN_14			97.2	2.8	1.5	0.1	96.9	1.6
PEN_15				100.0	0.1		0.0	99.9
PEN_16		4.6		95.4		12.0	12.0	76.0
PEN_17			100.0				99.7	0.3
PEN_18	38.4	31.5	12.8	17.2	55.9	25.4	6.8	11.9
PEN_19				100.0	0.0		0.0	100.0
PEN_20			7.8	92.2	0.0	0.1	1.9	97.9
PEN_21	3.3			96.7	32.2			67.8
PEN_22				100.0				100.0
PEN_23				100.0	1.6			98.4
PEN_24			100.0		0.1	0.1	99.8	0.1
PEN_28			85.1	14.9	0.2		91.0	8.8
PEN_32	0.5		29.8	69.7	0.4		16.6	83.0
PEN_35				100.0	0.0	0.0	0.0	100.0
PEN_38						33.3	33.3	33.3
PEN_39	1.9	98.1			23.2	75.5	1.1	0.1
PEN_41	24.6	0.1	75.3		23.1	0.1	76.8	0.0
PEN_42				100.0				100.0
PEN_50	67.6		31.1	1.3	69.2	0.4	29.1	1.3
PEN_51				100.0				100.0
PEN_40B	0.3			99.7	1.5		0.0	98.5
Sums of %	276.2	134.4	838.2	1351.2	354.7	147.8	861.7	1335.8
Avg percentage	10.6	5.2	32.2	52.0	13.1	5.5	31.9	49.5

2.3 Synopsis

This manuscript demonstrates that both qPCR and HTS techniques are highly correlated and produce strikingly similar quantitative estimates of fish DNA in faecal material, with no statistical difference. This study led to the establishment of a multi-year monitoring programme to investigate the diet of *E. minor* on Penguin Island to assess the impacts of a newly installed boat ramp on the diet and health of the penguins and, by inference, the local fish population. Moreover, this study also underpinned findings during the following year 2011 off the coast of Western Australia that had a considerable impact on penguin diet (Cannell *et al.*, 2011; Cannell *et al.*, 2015)

Prior to this study, no publications had tested estimates obtained using HTS to validate its ability to accurately reflect the proportion of prey present within the actual faecal sample itself. Previous studies had instead focused on whether estimates of prey proportions obtained using HTS or qPCR reflected what was consumed by the predator (Deagle *et al.*, 2010; Bowles *et al.*, 2011). Such studies concluded that diet estimates obtained using HTS or qPCR were semi-quantitative at best often requiring the use of correction factors to account for variations in prey size, DNA per unit biomass and differential digestion and degradation rates.

Although quantitative estimates of diet are fraught with difficulties the detection of prey without the need of *a priori* knowledge of what a predator consumes is a major strength of HTS techniques, provided careful attention is paid to the quality of extracts used in analyses. Somewhat disappointingly at times this paper has been cited as an exemplar that HTS data are ‘quantitative’ — far from this sweeping conclusion the work shows that careful workflow (revisited again in Chapter Five) and validations are required to interpret data in a quantitative or semi-quantitative context.

Metabarcoding workflows employing HTS are a powerful tool to study regions of rich biodiversity or regions that may undergo significant fluctuations in diversity over time. Moreover, the detection of soft-bodied organisms (Clare *et al.*, 2009;

Boyer *et al.*, 2013) that leave few macroscopic traces in faecal material illustrates the utility of HTS in potentially identifying taxa that leave little to no trace within a range of ecological samples such as in herbivore middens, which are explored further in the following chapter (Chapter Three), or cave sediments (Chapter Six).

Chapter Three – Herbivore middens as a source of palaeoecological and palaeogenetic data

3.1 Preface

Chapter Three attempts to extract and sequence aDNA from herbivore middens located in hot, arid environments to provide a local palaeogenetic record of plant and animal taxa. This study resulted in the published manuscript entitled 'High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens' (Quaternary Science Reviews 2012, 58, 135). With the exception of formatting and in-thesis referencing this manuscript has been reproduced as published.

In Chapter Two the utility of HTS in detecting prey species overlooked using species-specific primers was highlighted through the use of HTS of faecal samples obtained from *Eudiptula minor* which revealed a range of minor prey items. Additionally, the importance of multi-year data to facilitate the best-practice monitoring and evaluation of a species population was briefly mentioned.

In Chapter Three aDNA and HTS techniques are used to characterise the DNA preserved in herbivore middens sourced from four southern hemisphere sites, one of which was dated to approximately 30,490 BP. In doing so, the genetic characterisation of substrates, middens in this case, from sites that possess a distinct lack of preserved fossil remains is explored. Herbivore middens are well known to preserve DNA although the exact means by which preservation is aided remains poorly understood. Additionally, studies to date using herbivore middens as genetic archives have used Sanger sequencing followed by cloning and are restricted to cool or temperate sites

In the following study, the preservational limits of hot, arid sites are explored while at the same time the novel use of HTS to screen herbivore middens is tested. The use of herbivore middens as an important source of material for reconstructing past

environments is evaluated in the context of the difficulties associated with analysing degraded plant and animal DNA from taxa using current patchy reference databases.

3.1.1 Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

3.2 High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens

Dáithí C. Murray¹, Stuart G. Pearson², Richard Fullagar³, Brian M. Chase^{4, 5}, Jayne Houston¹, Jennifer Atchison³, Nicole E. White¹, Matthew I. Bellgard⁶, Edward Clarke⁷, Mike Macphail⁸, M. Thomas P. Gilbert^{1,9}, James Haile¹ and Michael Bunce¹

¹ *Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, South Street, Murdoch, WA, 6150, Australia.*

² *PaleoLab, School of Physical, Environmental and Mathematical Science, University of New South Wales, Canberra, ACT, 2610, Australia*

³ *Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, NSW, 2522, Australia.*

⁴ *Institut des Sciences de l'Evolution de Montpellier, UMR 5554, Centre National de Recherche Scientifique/Université Montpellier 2, Bat.22, CC061, Place Eugène Bataillon, 34095 Montpellier, cedex5, France.*

⁵ *Department of Archaeology, History, Culture and Religion, University of Bergen, Postbox 7805, 5020, Bergen, Norway.*

⁶ *Centre for Comparative Genomics, Murdoch University, South Street, Murdoch, WA, 6150, Australia.*

⁷ *Rio Tinto, Dampier, WA, Australia.*

⁸ *Department of Archaeology and Natural History, College of Asia and the Pacific, Australian National University, Canberra, ACT, 0200, Australia.*

⁹ *Centre for GeoGenetics, Natural History Museum of Denmark, and Department of Biology, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark.*

3.2.1 Abstract

The study of arid palaeoenvironments is often frustrated by the poor or non-existent preservation of plant and animal material, yet these environments are of considerable environmental importance. The analysis of pollen and macrofossils isolated from herbivore middens has been an invaluable source of information regarding past

environments and the nature of ecological fluctuations within arid zones. The application of ancient DNA (aDNA) techniques to hot, arid zone middens remains unexplored. This paper attempts to retrieve and characterise aDNA from four Southern Hemisphere fossil middens; three located in hot, arid regions of Australia and one sample from South Africa's Western Cape province. The middens are dated to between 30,490 (± 380) and 710 (± 70) cal yr BP. The Brockman Ridge midden in this study is potentially the oldest sample from which aDNA has been successfully extracted in Australia. The application of high-throughput sequencing approaches to profile the biotic remains preserved in midden material has not been attempted to date and this study clearly demonstrates the potential of such a methodology. In addition to the taxa previously detected via macrofossil and palynological analyses, aDNA analysis identified unreported plant and animal taxa, some of which are locally extinct or endemic. The survival and preservation of DNA in hot, arid environments is a complex and poorly understood process that is both sporadic and rare, but the survival of DNA through desiccation may be important. Herbivore middens now present an important source of material for DNA metabarcoding studies of hot, arid palaeoenvironments and can potentially be used to analyse middens in these environments throughout Australia, Africa, the Americas and the Middle East.

3.2.2. Introduction

The field of ancient DNA (aDNA) has, since its infancy, been largely restricted to the study of substrates from cool and frozen environments, which are deemed most amenable to long-term DNA preservation (Lindahl, 1993a, 1993b). To date, a number of historical and ancient samples have been subject to genetic analyses, ranging from bone (Smith *et al.*, 2001) and hair (Bonnichsen *et al.*, 2001; Gilbert *et al.*, 2004) to more complex, heterogeneous substrates such as coprolites (Kuch *et al.*, 2001; Poinar *et al.*, 2001) and sediments (Hofreiter *et al.*, 2003c; Willerslev *et al.*, 2003; Haile *et al.*, 2009). A number of studies have also attempted the isolation of DNA from samples – including fossil rodent middens - collected in cool to cold, semi-arid or arid environments (Kuch *et al.*, 2002; Hofreiter *et al.*, 2003b) and at high altitudes (Poinar *et al.*, 1998; Hofreiter *et al.*, 2000; Poinar *et al.*, 2003). The application of molecular aDNA techniques to hot, semi-arid or arid environmental

samples has previously been considered unrealistic due to the extreme heat found in such areas and as such is somewhat rarer and controversial (Smith *et al.*, 2003; Gilbert *et al.*, 2005b; although see Gilbert, 2011; Hekkala *et al.*, 2011)

Hot, arid and semi-arid environments are often marked by periods of stasis fluctuating on the edge of environmental equilibrium (Moore, 1953; Van Devender, 1990), punctuated by potentially dramatic changes that are induced by various triggers (Friedel *et al.*, 1993; Tausch *et al.*, 1993). There exists a delicate ecological balance and complex interplay across various environmental and biological gradients in arid regions (Beadle, 1966; Hayward & Phillipson, 1979; Northcote & Wright, 1982; Ritchie, 1986), making them of considerable environmental and biological interest. Flora and fauna inhabiting such environments are often at the limits of their tolerance to various abiotic factors, including temperature and water conservation, and have evolved to cope with extreme environmental conditions (Tongway & Ludwig, 1990; Groves, 1994). The study of past and present arid zone environments, - and the distribution of species within them - allows for the exploration of how they have adapted and shifted in response to both natural and anthropogenic mechanisms (Van Devender & Spaulding, 1979; Fall *et al.*, 1990; Pearson & Betancourt, 2002). The study of arid environments, however, is extremely challenging owing to the costs of collection and analysis, paucity of research attention and the lower quantities of recovered macro- and microfossil material. Nevertheless, studies using herbivore middens show promise in examining temporal and spatial variation in arid zone climates and biota, and perhaps, in some cases, may be the only viable means of doing so (Scott, 1990; Pearson & Betancourt, 2002; Scott & Woodborne, 2007; Chase *et al.*, 2009; Chase *et al.*, 2011).

To date, the reconstruction of palaeoenvironments has involved the use of a variety of molecular and morphological techniques, usually applied to sediment cores. Such techniques have included macrofossil and pollen identification, stable isotope analysis and ^{14}C dating. The application of these techniques to middens, where pollen and macrofossils have been preserved for thousands of years (Pons & Quézel, 1958; Wells & Jorgensen, 1964; Van Devender & Spaulding, 1979; Fall *et al.*, 1990; Pearson & Betancourt, 2002; Scott *et al.*, 2004), has provided the bulk of palaeoecological information in arid environments, where macrofossils are sparse

and continuous fossil pollen records are largely unattainable. Midden material has therefore played a large part in our understanding of arid zone ecology and environment and act as archives of valuable information. Midden accumulations, usually as organic-rich nests in the case of American and Australian middens and latrines in the case of the African rock hyrax middens (Figure S3.2.1), consist of material from the surrounding environment for construction or dietary purposes by arid-zone adapted mammals, and for the most part, represent a localised picture of the flora and fauna (Dial & Czaplewski, 1990; Scott, 1990; Pearson & Dodson, 1993). In the case of American and Australian middens, the animals urinate and defecate on their nests during the course of habitation, and organic material such as plant and animal tissue, bone, hair and eggshell gathered from the local surroundings (Pearson *et al.*, 2001), become cemented together by means of crystallized urine or amberat, solidifying the mass into a hard, impermeable amalgam, referred to as a midden. Individually, these middens are generally recognised as reflecting sub-centennial-scale periods of construction and habitation. Conversely, African rock hyrax middens are latrines composed nearly exclusively of excrement. They are excellent traps for microfossils (pollen, phytoliths, etc.) from both regional and local environments as these are respectively brought in by the wind or adhere to the midden agent's fur. Hyrax middens, however, contain very little non-dietary macrofossil material (for a fuller comparison and description of hyrax latrines and rodent nest middens see Chase *et al.*, In press). Increasingly, the hyrax middens that are collected for analysis are composed predominantly of urine, and have been shown to accumulate continuously over many thousands of years (Chase *et al.*, 2009; Chase *et al.*, 2011).

Genetic profiling has previously been applied to midden contexts, with two aDNA profiling studies retrieving reliable, seemingly authentic aDNA sequences from cold, arid zone (BWk - Köppen climate classification, see Peel *et al.*, 2007) middens (Kuch *et al.*, 2002; Hofreiter *et al.*, 2003b). Since these studies, the fields of aDNA and environmental metabarcoding, whereby complex environmental samples are genetically audited (Valentini *et al.*, 2009a; Taberlet *et al.*, 2012), have rapidly evolved. With the advent of affordable and accessible Second Generation high-throughput sequencing (HTS) it is now possible to genetically screen a wide range of complex modern and ancient substrates, with an unprecedented depth of sequence

coverage (Shokralla *et al.*, 2012). Through the use of material as diverse as sediment (Haile *et al.*, 2009; Jørgensen *et al.*, 2012), water (Rusch *et al.*, 2007; Ficetola *et al.*, 2008; Thomsen *et al.*, 2012) and faeces (Deagle *et al.*, 2009; Valentini *et al.*, 2009b; Murray *et al.*, 2011) a wealth of data can be produced to aid in the understanding of pertinent ecological questions in relation to biodiversity (Andersen *et al.*, 2011; Griffiths *et al.*, 2011), dietary analysis (Pegard *et al.*, 2009; Deagle *et al.*, 2010) and anthropogenic impacts (Chariton *et al.*, 2010; Vila & Borrelli, 2011). It is now possible, therefore, to bypass traditional molecular cloning and Sanger sequencing techniques through the use of new DNA technologies (HTS) to supplement morphological (macrofossils and palynology) methods of midden analysis, to allow an even fuller investigation of arid zone ecology.

Using HTS and environmental DNA metabarcoding techniques, this study attempts to recover aDNA from herbivore midden material collected from three hot, arid Australian sites and one site in South Africa (Figure 3.2.1) that have been dated to between $30,490 \pm 380$ and 710 ± 70 cal yr BP. A comparison of the data obtained via HTS with complementary data on past and present species distributions, in addition to pollen and macrofossil analyses, allows for a critical examination and authentication of the genetic data. This study aims to demonstrate how genetic methods can be used to complement traditional methods of midden investigation for palaeoenvironmental reconstruction, to further our understanding of hot, arid environments.

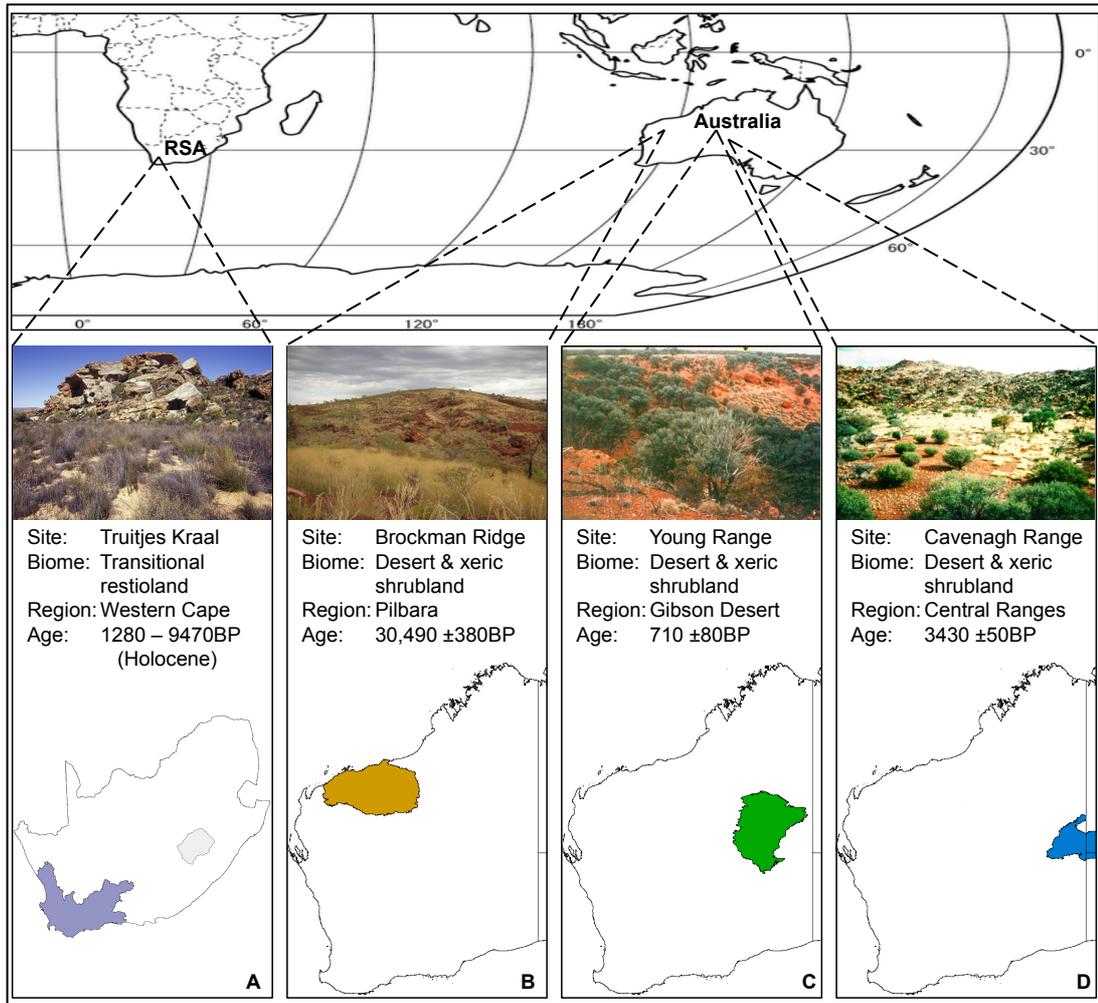


Figure 3.2.1 Location of midden sites used in this study and associated information. Location and image of Truitjes Kraal midden site, South Africa, with Western Cape highlighted (A). Locations and images of Western Australian midden sites, with IBRA regions highlighted (B-D).

3.2.3. Collection sites

Four Southern Hemisphere middens were sampled in this study; a single hyrax midden from South Africa's (RSA) Western Cape Province (Figure 3.2.1A) and three herbivore middens from separate Interim Biogeographic Regionalisation of Australia (IBRA) regions (Thackway & Cresswell, 1995) within Western Australia (WA) (Fig3.2.1B-D). The three midden samples collected in Western Australia were from hot, arid (BWh) zones (Köppen climate classification, see Peel *et al.*, 2007). The hot, arid zone collection sites are generally characterised by extreme hot summers and somewhat mild winters. Daytime summer temperatures average ~37-

38°C, but regularly exceed 40°C. In winter, average daytime highs are ~21-25°C, but can fall to ~6-7°C at night. Winter nighttime temperatures at or close to freezing are extremely rare in these zones (climate data from Giles and Tom Price weather stations, WA). This contrasts markedly with previous midden genetic studies (Kuch *et al.*, 2002; Hofreiter *et al.*, 2003a) where average daily highs in summer are ~24-28°C, although it can reach ~30°C, and winter daily highs average ~16-21°C, with nighttime temperatures at or below freezing more common (climate data from weather stations at Neuquén Airport, Argentina and Calama, Chile).

3.2.3.1. Truitjes Kraal, RSA (TK)

Truitjes Kraal (32.5123°S, 19.3112°E) is located in the Cape Floristic Region (CFR) in the Western Cape province of RSA (Figure 3.2.1A). The midden site lies in what is described as "restioland" (dominated by Restionaceae), within a few kilometres of the relatively sharp transition between the Fynbos and Succulent Karoo biomes, with a climate on the threshold between hot-summer Mediterranean (Csa) and cold, semi-arid (BSk). The site records a mean annual precipitation around 350 mm, a mean annual temperature of around 15°C (data from Hijmans *et al.*, 2005) and an Aridity Index value of 0.242 (Trabucco & Zomer, 2009). The vegetation at the site consists of a low shrub understorey with intermittent taller shrubs, in addition to dwarf succulent shrubs of Crassulaceae and Mesembryanthemaceae (Meadows *et al.*, 2010).

3.2.3.2 Brockman Ridge, WA (BR)

The Brockman Ridge, an ironstone-capped strike ridge, lies in the Pilbara IBRA region of northwestern WA (Figure 3.2.1B), approximately 60 km northwest of Mount Tom Price (22.68°S, 117.78°E). The Pilbara is a desert and xeric shrubland biome with a BWh climate that consists of scattered low trees of *Eucalyptus leucophloia* over *Acacia atkinsiana* open shrubland, over *Triodia wiseana* mid-dense hummock grassland. A number of other species are also associated with the site that includes *Acacia aneura*, *Hakea chordophylla*, *Paspalidium clementii*, *Ptilotus calostachyus* and *Solanum lasiophyllum* (Biota Environmental Sciences Pty Ltd, 2005). The nearest weather station is situated at Tom Price (-22.7°, 227.77°) recording a median annual precipitation around 313mm, a mean annual temperature of around 24°C (data from Hijmans *et al.*, 2005) and an Aridity Index value of 0.2 (Trabucco & Zomer, 2009).

3.2.3.3 Young Range, WA (YR)

The Young Range (25.05°S, 124.983°E), located in Western Australia, is a low breakway in the extremely isolated Gibson Desert IBRA region (Figure 3.2.1C). The Young Range is a desert and xeric shrubland biome that consists of shrubs, low shrubs and herbs. The dominant flora at the site is a mixture of Caesalpiniaceae, Myoporaceae, *Acacia*, *Grevillea* and species dominating hummock grassland (e.g. *Triodia* spp.) (Pearson, 1997). The Young Range also has a BWh climate and Giles is the nearest meteorological station (-25.03, 128.30) recording a median annual precipitation around 250mm, a mean annual temperature of around 23°C (data from Hijmans *et al.*, 2005) and an Aridity Index value of 0.101 (Trabucco & Zomer, 2009).

3.2.3.4 Cavenagh Range, WA (CR)

The Cavenagh Range (26.2°S, 127.9°E) is a rock pile situated in the Central Ranges IBRA region located in eastern WA (Figure 3.2.1D). The dominant vegetation at the site includes spinifex (*Triodia* spp.), with shrubs and *Eucalyptus* spp. along the drainage lines (Pearson, 1997). It is considered a desert and xeric shrubland biome with a BWh climate, with the nearest meteorological station (Giles: -25.03, 128.30) recording a median annual precipitation around 250mm, a mean annual temperature of around 22°C (data from Hijmans *et al.*, 2005) and an Aridity Index value of 0.1164 (Trabucco & Zomer, 2009).

3.2.4 Materials and Methods

In keeping with standard aDNA practice, pre-PCR work was conducted in a dedicated aDNA clean room, with all downstream post-PCR work conducted in a physically separate laboratory, thus minimising sample contamination (Cooper & Poinar, 2000). Each midden was sub-sampled at Murdoch University, Australia and subsequently sent to the Centre for GeoGenetics, Denmark for independent replication. For all samples, DNA extraction, amplification and sequencing were performed at both Murdoch University and the Centre for GeoGenetics. Whilst HTS was performed at Murdoch University, traditional cloning and direct Sanger sequencing were performed at the Centre for GeoGenetics.

3.2.4.1 Background to midden samples

The samples used in this study were collected, radiocarbon dated (Table S3.2.1) and analysed for pollen and macrofossils prior to this study (Pearson, 1997; Meadows *et al.*, 2010; Macphail, 2011). Large, intact samples were taken from the middens in this study to allow for sub-sampling, thus limiting the risk of environmental contamination. Middens that appeared to have been damaged as a result of weathering, digging or burrowing were avoided, although the BR midden was fractured along the base and had a honeycombed appearance (Atchison, 2010). The TK midden was collected in its entirety from a rock face overhang and spans the period from 1280 - 9470 cal yr BP (Meadows *et al.*, 2010). The sample used for aDNA analysis was not dated separately, but is certainly of Holocene age. The BR midden was collected from the rear of the Brock 12 rock shelter within an Aboriginal site complex in the Puutu Kunti Kurrama and Pinikura native title claimant area. Sections of the cave had been walled-in with the likely purpose of either the creation of an artificial habitat for the exploitation of, or the trapping of small animals (Figure S3.2.2). With the exception of the creation of these walls, no other evidence of cultural material or influence was identified at the site of sample collection (Clarke, 2010). The BR midden is the oldest in this study, radiocarbon dated to 30,490±380 cal yr BP (Macphail, 2011), although the age of the midden was not known before aDNA analysis took place. This midden consisted of three sub-samples obtained from one midden mound that were processed separately (Atchison, 2010). The YR midden was found in a rock shelter, protected from dissolution by moisture, and has been radiocarbon dated to 710±80 cal yr BP (Pearson, 1997). The CR midden, radiocarbon dated to 3430±50 cal yr BP, was collected from a small crevice and had few leaves and sticks, suggesting that an animal other than a stick-nest rat (*Leporillus* spp.) may have formed the midden (Pearson, 1997). The above radiocarbon dates were taken directly on organic materials and the age estimates they provide for midden accumulation carry the possibility of being on material both older and younger than the aDNA within the stratigraphic units sampled.

3.2.4.2 DNA extraction and screening

Between 0.16 - 0.31g of midden material was used for each sample DNA extraction using the Sergey Bulat extraction method optimised for small amounts of material, with extraction controls also included (Haile, 2011). Bulat buffer component

concentrations were as follows; 0.02 g/mL Sarcosyl, 50 mM Tris-HCL (pH 8.0), 20 mM NaCl, 3.5 % 2-mercaptoethanol, 50 mM DTT, 2 mM PTB, 0.8 g/mL Proteinase K. DNA was eluted in 100 µL and screened using quantitative PCR (qPCR) at multiple dilutions. DNA extracts were screened using multiple primer sets for both plants and mammals. The plant primer sets included both *trnL c/h* and *trnL g/h* plastid primers that amplify short sections of the *trnL* intron (Taberlet *et al.*, 1991; Taberlet *et al.*, 2007). In addition to these, both 12S A/O and 16Smam (Taylor, 1996) primer sets, designed to amplify a small region within mammalian mitochondrial 12S and 16S genes respectively, were also used. Each qPCR reaction was made up to a total volume of 25 µL, containing 12.5 µL ABI Power SYBR master mix (Applied Biosystems), 0.4µM of forward and reverse primer, 8.5 µL H₂O and 2 µL DNA extract. Reaction conditions for the plant primers were as follows: initial heat denaturation at 95 °C for 5mins, followed by 40 cycles of 95 °C for 30 s; 54 °C for 30 s (annealing step); 72 °C for 45 s followed by a 1 °C melt curve and final extension at 72 °C for 10 mins. Quantitative PCR cycling conditions for the 12S A/O and 16Smam primer sets were the same as those for both plant primers, except the annealing temperatures, which were 55 °C and 57 °C, respectively. For each qPCR assay, DNA extraction, negative PCR reagent and positive controls were included.

3.2.4.3 DNA Sequencing

DNA extracts that successfully yielded DNA of sufficient quality, free of inhibition, as determined by initial qPCR screening, were assigned a unique 6bp DNA tag (specifically a Multiplex Identifier-tag, MID-tag) (Roche, 2009) for each of the *trnL g/h*, 12S A/O and 16Smam primer sets. Independent MID-tagged qPCR for all midden samples were carried out using each primer set in 25 µL reactions containing 1X PCR Gold Buffer (Applied Biosystems), 2.5 mM MgCl₂ (Applied Biosystems), 0.4 mg/mL BSA (Fisher Biotech, Aus), 0.25 mM of each dNTP (Astral Scientific, Aus), 0.4 µM of forward and reverse primer, 0.25 µL AmpliTaq Gold (Applied Biosystems), 0.6 µL SYBR Green (1:2,000, Life Sciences gel stain solution) and 2µL of template. The cycling conditions for qPCR using the *trnL g/h* primer set were as follows: initial heat denaturation at 95 °C for 5mins, followed by 50 cycles of 95 °C for 30 s; 50 °C for 30s (annealing step); 72 °C for 45 s followed by final extension at 72°C for 10mins. The cycling conditions were the same for both 12S

A/O and 16Smam primer sets apart from the annealing temperatures, which were 50 °C and 57 °C respectively. Multiplex Identifier-tagged PCR amplicons were generated in duplicate and pooled together to minimise the effects of PCR stochasticity on low-temple samples. The resultant pooled amplicons were purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), according to the manufacturer's instructions and eluted in 40 µL H₂O. Purified amplicons were electrophoresed on 2 % agarose gel and pooled in approximately equimolar ratios based on ethidium-stained band intensity to form a sequencing library. For each MID-tagged qPCR assay, negative PCR controls were included and if found to contain amplifiable DNA these PCR amplicons were incorporated into the pooled sequencing library. Emulsion PCR and GS Junior 454 Sequencing were performed as per Roche GS Junior protocols for amplicon sequencing (<http://www.454.com>).

3.2.4.4 Data analysis

Processed emulsion PCR amplicon sequence reads (hereafter referred to as sequences) obtained from the GS Junior sequencing runs have been deposited in the Dryad Repository (doi:10.5061/dryad.7334t). Sequences were sorted into sample batches based on MID-tags using Geneious v5.6.4 (Drummond *et al.*, 2011). MID-tags and primers were trimmed from the sequences allowing for no mismatch in length or base composition, also performed using Geneious v5.6.4. Batched and trimmed sequences were then dereplicated using 454 Replicate Filter (Gomez-Alvarez *et al.*, 2009), clustering sequences of exact identity and length. Dereplicated sequence files were then searched for chimeras using the *de novo* method in UCHIME (Edgar *et al.*, 2011), and were removed. After the above post-sequencing screen, sequences occurring only once (i.e. singletons) were removed, to minimise false positives arising from sequencing error. Once complete, each batch of cleaned, de-noised sequences was searched using BLASTn version 2.2.23 (Altschul *et al.*, 1990), against the NCBI GenBank nucleotide database (Benson *et al.*, 2006) to enable the identification of reads. Sequences were searched without a low complexity filter, with a gap penalties existence of five and extension of two, expected alignment value less than 1e-10 and a word count of seven. This was automated in the internet-based bioinformatics workflow environment, YABI (Hunter *et al.*, 2012). The BLAST results obtained using YABI were imported into

MEtaGenome Analyzer v4 (MEGAN), where they were taxonomically assigned using the LCA-assignment algorithm (min. bit score = 65.0, top percentage = 5%, min. support = 1) (Huson *et al.*, 2007). Further analysis of Muridae sequences was conducted by determining Operational Taxonomic Units (OTUs) using OTUPIPE with default parameters (<http://drive5.com/otupipe/>), whilst a phylogenetic comparison of Muridae sequences between samples was conducted using MrBayes (Huelsenbeck & Ronquist, 2001) in Geneious v5.6.4 (Drummond *et al.*, 2011).

After sequences were processed, identified and parsed, the species identified were investigated to determine whether or not they currently occur in the region where they were detected, or have occurred in the past. To do this, the South African National Biodiversity Institute's (SANBI) Plants of Southern Africa online checklist [<http://posa.sanbi.org/searchspp.php>] was used for the RSA midden (Figure 3.2.1A), and a combination of FloraBase [<http://florabase.dec.wa.gov.au/>] and Atlas of Living Australia [<http://www.ala.org.au/>] were used for the Australian sites (Figure 3.2.1B-D).

3.2.5 Results and Discussion

3.2.5.1 Overview of sequencing data

Over 20,000 sequences were obtained via HTS that passed the post-sequencing screen and occurred at an abundance greater than one (see Section 3.4). DNA was amplified using *trnL* g/h (size variable product between ~90-120 bp – including MID-tags and primers), 12S A/O (~160 bp) and 16Smam (~150 bp) primer combinations, whilst amplicon generation using the longer *trnL* c/h (giving an expected product of variable length >200 bp) primer set failed at each of the four study sites. Appropriate control reactions (described in Section 3: Materials and Methods) throughout the process, with the exception of ubiquitous human DNA sequences, were found to be negative for contaminant DNA arising from laboratory processing procedures. It is acknowledged however that contamination can be cryptic and sporadic, and that low-level contamination can escape contamination controls (Champlot *et al.*, 2010). The strict adherence to aDNA protocols, the use of appropriate controls throughout, in addition to the critical analysis of the data (described in Section 3.4) (Cooper & Poinar, 2000; Gilbert *et al.*, 2005a), however,

greatly reduces the likelihood that contamination can account for the data presented here.

Previous studies involving the amplification of the hyper-variable p-loop region of the plastid *trnL* intron, using the *trnL* g/h primer set, have shown taxonomic assignment possible with sequences as short as 10 bp (Taberlet *et al.*, 2007). In this study, however, sequences less than 38 bp returned no taxonomic information and as such were discarded. Across the four midden samples, taxa representing 28 distinct families of plants were identified using *trnL* sequences that varied in length from 38-70 bp, minus MID-tags and primers (Table 3.2.1).

Table 3.2.1 Plant families identified in the midden samples using *trnL* plastid primers. For a more detailed comparison between plant taxa identified previously via morphological analysis and those identified via genetic means refer to Figure S3.2.3A-D.

Taxon	Midden Location			
	Cavenagh Range	Young Range	Pilbara	Truitjes Kraal
Acanthaceae		√		
Amaranthaceae	√	√ [#]		√
Amaryllidaceae			√ ^Ω	
Anacardiaceae		√ ^Ω	√ ^Ω	
Apocynaceae				√
Asteraceae	√		√ [#]	√
Brassicaceae	√			
Bromeliaceae			√ [¶]	
Campanulaceae				√ [#]
Casuarinaceae	√ [#]			
Ebenaceae				√
Fabaceae	√	√ [#]	√ [#]	√ [#]
Gesneriaceae		√ ^Ω		
Goodeniaceae		√		
Lamiaceae				√
Loranthaceae	√			
Malvaceae	√			
Melanthaceae				√
Oleaceae				√
Pinaceae	√ [¶]			
Poaceae	√		√ [#]	√ [#]
Podocarpaceae				√ [#]
Proteaceae		√ [#]		
Sapindaceae	√ [#]		√ [#]	
Scrophulariaceae				√ [#]
Solanaceae	√ [#]	√ [#]		√
Thymelaeaceae				√ [#]
Toricelliaceae				√ [¶]

Key: √ - Present in midden sample; # - Found previously in midden via morphological analysis;

Ω - Not found in region; ¶ - Not found natively in Aus/SA

Through the assignment of DNA sequences to GenBank a total of six mammalian families were identified using both mammalian mtDNA 12S and 16S rRNA PCR assays, which generated sequences ~95-105 bp and ~90-100 bp in length respectively, minus MID-tags and primers. Within these mammalian families, species could reasonably be assigned in three cases (Table 3.2.2).

Table 3.2.2 Mammalian taxa identified in midden samples using 16S and 12S rRNA primer sets.

Taxon	Midden Location			
	Cavenagh Rng	Young Rng	Pilbara	Truitjes Kraal
Dasyuridae - <i>Pseudantechinus</i>		√ ^B √		
Gliridae - <i>Graphiurus ocellaris</i>				√ √
Macropodidae	√ ^B			
Muridae		√ [#]	√ ^B	
Procaviidae - <i>Procavia capensis</i>				√ ^B √ ^{B%}
Phalangeridae - <i>Trichosurus vulpecula</i>	√ ^B √ ^{B%}			

√ - Present in midden sample; # - Found previously in midden via morphological analysis; B - Detected using both 16S and 12S rRNA primer sets; % - indicates top BLAST species match 100% similarity

To our knowledge this is the first study to focus on the retrieval and sequencing of aDNA from Southern Hemisphere fossil midden material located in hot, arid regions. Moreover, the application of HTS techniques to midden material has not been attempted to date, and the following findings clearly demonstrate the increase in resolution afforded by the use of such methodology. Of significance, the Brockman Ridge midden sample is the oldest environmental sample; quite possibly the oldest sample, from which aDNA has been successfully extracted in Australia (although see Adcock *et al.*, 2001; and subsequent critiques Cooper *et al.*, 2001; Smith *et al.*, 2003). For the Pilbara IBRA region in particular, aDNA work of this kind could be a critical addition to the assemblage of palaeoenvironmental data, as it is dated to a period for which almost no such regional data exists (Clarke, 2010; Macphail, 2011). This paper confirms that aDNA can be successfully recovered from midden deposits in hot, arid climates, suggesting that middens may be a valuable substrate for genetic analysis in such regions; it does not claim to be a comprehensive study of the sampled middens. Instead, an overview of the aDNA data is provided, focusing on

some of the more salient points related to taxa identified by HTS and comparing the results with previous pollen and macrofossil analyses.

3.2.5.2 Site-specific analysis

Cavenagh Range

At least eleven families of plants were identified in the CR midden (Table 3.2.1), all of which, with the exception of Pinaceae (Order: Pinales), occur in the Central Ranges IBRA. Of the plant families identified, three were previously detected via pollen analysis: Casuarinaceae, Sapindaceae and Solanaceae (Table 3.2.1) (Pearson, 1997). Pollen analysis was only able to identify the genus *Dodonaea* (Sapindaceae), whilst genetic analysis identified both *Casuarina* (Casuarinaceae) and *Solanum* (Solanaceae) (Fig S2.2.3A). However, although *Casuarina* is known to occur in the IBRA, it is recorded some distance from the site (ALA, FloraBase). The sequences assigned to *Casuarina* in this study are highly likely to be *Allocasuarina*, which does occur at the site and is known to occur alongside *Atriplex* (Mitchell & Wilcox, 1994), also detected via genetic analysis (Figure S3.2.3A). In addition to these taxa, Loranthaceae was identified via genetic analysis but not through previous pollen analysis of the fossil midden. A number of possible genera of Poaceae were also detected, including *Eriachne* and *Urochloa* (Figure S3.2.3A), both of which, although not formally recorded at the site, are recorded in the IBRA.

Previous analysis of the CR midden did not identify any macrofossil remains (Pearson, 1997). Through the use of mammal specific primers, however, it was possible to detect the presence of Phalangeridae, specifically *Trichosurus vulpecula* (the common brushtail possum) and Macropodidae (Table 3.2.2). *Trichosurus vulpecula* is no longer found at Cavenagh Range; last recorded in the area in the 1930's, and it is the only species of Phalangeridae known to have existed in the Central Ranges IBRA (ALA). The distribution of *T. vulpecula* has retracted considerably since European settlement, as a result of a range of issues including predation and overgrazing by introduced species (How & Hillcox, 2000). The identification of Macropodidae sequences to genus or species level proved difficult, with both 12S and 16S giving no clear indication past the family level. Currently there are only four species of Macropodidae known to exist in the Central Regions, with *Lagorchestes hirsutus* (the rufous hare-wallaby), *Macropus robustus* (the

common wallaroo) and *Petrogale lateralis* (the black-flanked rock-wallaby) all recorded specifically at Cavenagh Range (ALA). Whilst 16S indicated the presence of *Macropus* it was not possible to identify *M. robustus* using this primer set and *Macropus* sequence identities were quite low ($\leq 95\%$). Use of the 12S primer set again resulted in difficulties with assignment to a genus or species level, with both *Lagorchestes* and *Petrogale* identified with equal similarity (98%). However, currently no 16S or 12S sequences for *P. lateralis* exist on GenBank. It was initially suggested that the CR midden was constructed by an animal other than a stick-nest rat (*Leporillus* spp.), possibly a rock wallaby or possum (Pearson, 1997). The identification of Macropodidae, possibly *Petrogale*, and *T. vulpecula* DNA (Table 3.2.2) in the midden material therefore increase the likelihood of this being the case.

Young Range

All plant families detected in the YR midden (shown in Table 3.2.1), with the exception of Gesneriaceae, which has an eastern Australian distribution, are known to occur in the Gibson Desert IBRA (ALA, FloraBase). Previous pollen and macrofossil analysis had identified Amaranthaceae, Fabaceae, Proteaceae and Solanaceae (Figure S3.2.3B), all of which were detected via this genetic screening, and a number of other families not detected in this study (Pearson, 1997).

Previous macrofossil analysis of the YR midden found several species of mammal, that included the locally extinct *T. vulpecula* and *Isoodon auratus* (the golden bandicoot), in addition to both *Notomys* (hopping mice) and *Macropus robustus* (Pearson, 1997). Genetic screening of the midden did not detect any of the above specifically (Table 3.2.2). Muridae sequences were identified from the midden material, though it was not possible to assign such sequences to a genus level due to the absence of 12S, 16S and COI reference sequences for many of the Muridae species found in the area, however it seems that these sequences cluster to form a single OTU, although there is some variation in the collective sequences ($< 2\%$). Such variation, although minor, is unlikely to have arisen as a result of sequencing error or chimeras due to the post-sequencing screen removing such instances, and could indicate multiple individuals contributing to this midden. Additionally, sequence BLASTn matches group these Muridae sequences closest to other Australasian Muridae, e.g. *Melomys cervinipes* (Fawn-footed Mosaic-tailed Rat) and

Paramelomys rubex (Mountain Mosaic-tailed Rat), albeit with low percentage similarities (<93%). However, Dasyuridae, most likely *Pseudantechinus* (false antechinuses), currently found in the area, was detected in the midden material through DNA analysis (Table 3.2.2), and this was not previously identified via macrofossil analysis.

Brockman Ridge

The Brockman Ridge midden mound is the oldest midden deposit in this study, and for the purposes of this discussion the three sub-samples are treated as one.

Fossil pollen assemblages recovered from the samples were dominated by unidentified Poaceae and a number of taxa within Family Myrtaceae, leading Macphail (2011) to propose that if plant DNA were preserved in the amber that it would most likely be that of Myrtaceae (*Eucalyptus* and possibly *Melaleuca*) and Poaceae (possibly *Triodia*). Of these taxa, only Poaceae were detected using genetic techniques, although other less common taxa represented by pollen were identified via genetic screening, e.g. Asteraceae, Fabaceae and Sapindaceae (*Diplopeltis* and/or *Dodonaea*) (Table 3.2.1, Figure S3.2.3C).

The Brockman Ridge sample contained no identifiable macroscopic remains when analysed conventionally (Atchison, 2010), however the targeting of both 12S and 16S mammalian mitochondrial genes revealed the presence of Muridae sequences (Table 3.2.2). It was not possible to definitively say to which genera these sequences belong, owing to the lack of 12S and 16S sequences on GenBank for species that occur or are known to have occurred in the area, however BLASTn results group these Muridae signatures closest to other Australasian Muridae, e.g. *Uromys hadrourus* (Masked White-tailed Rat), albeit with low percentage similarities (<93 %). Additionally, for both primer sets, OTU analysis suggests that these sequences form a single OTU, although, as was the case with the YR midden, there is some minor variation between sequences within this clustering (<2 %). Based on phylogenetic analysis it is also possible to suppose that the Muridae sequences identified in this midden differ from those detected in the YR midden, and represent distinct species (Figure S3.2.4).

Truitjes Kraal

Initial pollen analysis of the TK midden revealed high levels of Asteraceae, Ericaceae (Order: Ericales) and Poaceae (Meadows *et al.*, 2010). Using genetic means a number of different possible genera of both Asteraceae and Poaceae were detected (Figure S3.2.3D), however, no Ericaceae was found. Alternatively, genetic analysis detected Ebenaceae of the same order Ericales. In addition to several species detected by both pollen and DNA analysis, a number of additional taxa were identified, solely through genetic analysis, such as Apocynaceae, Lamiaceae and Solanaceae (Table 3.2.1, Figure S3.2.3D). A few taxa were identified that do not occur specifically at the site, such as Melianthaceae and Oleaceae. However, both of these taxa are known to occur relatively close to the site (Melianthaceae occurrence id: NBG171075-0 and Oleaceae occurrence id: PRE320306-0) (SANBI), and considering the antiquity of the material it is possible that they grew at the site in the past

The TK midden contains no faunal macrofossils but targeting mammalian DNA revealed both the midden builders - *Procavia capensis*, the rock hyrax - and *Graphiurus ocularis* (the spectacled dormouse or namtap); a South African endemic species that inhabits a wide range of habitats including dry rocky outcrops and cliffs in South Africa (Table 3.2.2).

3.2.5.3 Limitations of study

Given the controversy surrounding previously purported aDNA retrieval from hot, arid zone specimens (see Cooper & Poinar, 2000; Gilbert *et al.*, 2005b; Schlumbaum *et al.*, 2008 but also; Gilbert, 2011; Hekkala *et al.*, 2011) a number of caveats need to be considered when interpreting the degraded and ancient DNA recovered in this study to allow for a proper evaluation of the authenticity of the presented results (Gilbert *et al.*, 2005a).

Ancient DNA, which by its nature is extremely degraded and often damaged, is typically quite short, fragmented and in low copy number. Various studies have shown that the average length of DNA recovered from ancient specimens is generally less than 100 bp (Poinar *et al.*, 2006), and this study is no exception. The DNA sequences retrieved from the middens in this study for all primer combinations

were less than 100bp. Moreover, the attempt to target and amplify a longer stretch of the *trnL* intron, using the *trnL* c/h primer set, universally failed. The degraded nature of aDNA sequences thus makes it difficult to use conventional barcoding primers, as the lengths of resultant amplicons far exceed that which is realistically possible in aDNA studies (Valentini *et al.*, 2009a). The use of short sections of mammalian genes is generally straightforward compared to that for plants, due to the coverage afforded them on GenBank and greater taxonomic certainty associated with this group. Nonetheless, the use of the hyper-variable p-loop region of the *trnL* intron for plants, although not without problems (Hollingsworth *et al.*, 2011), provides sufficient taxonomic resolution in the case of this study. In most samples taxonomic assignment was possible to the family level, as was the case with previous morphological studies on these middens (Pearson, 1997; Meadows *et al.*, 2010; Macphail, 2011). In several instances (Figure S3.2.3A-D), it was possible to provide greater taxonomic resolution, to the genus level, than is possible using pollen; as the taxonomic resolution provided by fossil pollen in most of the families common to the arid zone is low. This is of particular value for families such as Poaceae that are highly diverse, but which - based on their pollen - are morphologically indistinguishable. For the sake of remaining cautious and conservative, however, such assignments are only dealt with peripherally in this study and the establishment of much better databases of reference material than currently exists is required to allow for greater certainty in taxonomic assignment at this level. In other words, datasets, like that compiled here, will have greater resolution in the future as databases become more comprehensive and flaws in the underpinning taxonomic framework are resolved.

The middens in this study have previously been analysed for pollen and macrofossil remains (Pearson, 1997; Meadows *et al.*, 2010; Macphail, 2011) and thus provide a valuable point of comparison. The preservation of organic material is generally excellent in middens, with the presence and preservation of pollen and/or macrofossils varying from low and adequate in the BR midden to substantial and good in the TK midden. Whilst not guaranteeing the presence of aDNA, the survival of other biomolecular components in these samples suggests aDNA survival is at least plausible. Indeed, genetic analysis did detect the presence of a number of families previously identified in pollen and macrofossil analyses, as well as families

and possible genera not previously detected in the midden samples (Figure S3.2.A-D). The presence of additional taxa, and the absence of previously identified taxa, further highlights that discussed in Jørgensen *et al.* (2012), namely that pollen, macrofossil and aDNA analyses are complementary as opposed to mutually exclusive and each provide ecological overviews with varying levels of taxonomic information. Moreover, the detection of extirpated (e.g. *T. vulpecula* not recorded in the region since the 1930's) and endemic taxa (e.g. *G. ocularis*), in addition to results obtained independently at the Centre for GeoGenetics in Copenhagen, using cloning followed by Sanger sequencing, is strong evidence that argues for the authenticity of these aDNA sequences.

The lack of database coverage afforded certain taxa has proven problematic in this study. However, much of the difficulty associated with this issue is observed at a genus level and can be overcome through critical assessment of taxonomic assignments and the use of current, historical and modelled distribution data. Overall the database coverage problem, although cumbersome, has a limited impact upon the results of this particular study, and in general the results obtained in this study are plausible and in keeping with expected outcomes. In general, the taxa detected in the middens are known to occur in close proximity to the midden sites and reflect the climate at the sites, e.g. taxa detected in the Australian middens are generally all hot, arid or semi-arid adapted plants. In addition to this, there appears to be little overlap in taxa identified between samples, with the TK midden from South Africa, for instance, being noticeably distinct in terms of identified plant and mammalian taxa, when compared to the Australian middens. Finally, had there been significant modern environmental contamination of the samples arising from modern invasive taxa found in the area, urinating on the middens for example, such as *Mus musculus* (house mouse), *Rattus rattus* (black rat) or *Vulpes vulpes* (red fox), or indeed contamination arising from reagents (Erlwein *et al.*, 2011; Tuke *et al.*, 2011), DNA from these taxa should have been detected, but were not. It is noted that unidentifiable Muridae sequences were detected, however, it is clear from phylogenetic analysis that these sequences do not group with the common contaminant *Mus musculus*; they cluster, rather, with other native Australasian murids (Fig S3.2.4). Indeed, not only does the amberat help to create an impermeable mass but its properties enable it to seal breaks in the weathering rind and discourage

insect attack (Spaulding *et al.*, 1990), further reducing possible exogenous contamination. This does not completely remove the possibility of “old” contamination, arising from the movement of material up through the stratigraphy of the midden (Spaulding & Robinson, 1984; Pearson & Dodson, 1993; McCarthy *et al.*, 1996), although this is not an issue with the TK midden sample, as hyrax middens maintain stratigraphical integrity significantly better than rodent nest middens (Chase *et al.*, In press).

As noted previously there are a number of taxa that have been “detected” in the midden material that are somewhat problematic (Table 3.2.1). In some instances, such as the presence of Gesneriaceae in the YR midden or Amaryllidaceae in the BR midden, such taxa are not known to occur locally, at least in the present day flora. In other cases taxa have been “detected” that are not found natively in the country from which the midden was sampled, such as Torricelliaceae and Pinaceae in the TK and CR middens respectively. In the first instance, it is doubtful that there has been an extirpation or range contraction of the taxa identified. Gesneriaceae has a wholly east Australian distribution, whilst the closest record of Amaryllidaceae is over 350km from the BR site. As regards to non-local or exotic taxa, with the exception of Pinaceae, which is a common laboratory and environmental contaminant, it is highly improbable that this is the result of laboratory or environmental contamination. The most likely explanation for such irregularities is a lack of coverage afforded certain taxa in current DNA databases (Taylor & Harris, 2012). In all the cases where disputed taxa have been identified there are records of related taxa (i.e. families within the same order, or genera within the same families) occurring in the area. In these cases there is little or no representation of these taxa in current DNA databases for *trnL* or other commonly used loci. For instance, in the case of sequences identified as Torricelliaceae (Order: Apiales), there are only two genera of Apiales known to occur at the site, neither of which are represented on GenBank; *Centella* and the rare, Western Cape endemic *Nanobubon* (Magee *et al.*, 2008; Magee, 2012).

The genetic auditing of midden samples in this study also failed to identify families and genera, both plant and mammal, detected previously via morphological analyses. Previously identified plant taxa such as *Ptilotus* and Myoporaceae are not currently represented on GenBank, whilst mammalian taxa such as *Leporillus apicalis* and

Notomys have no 16S, 12S or *COXI* sequences on current databases either. However, insufficient database coverage of taxa fails to explain the absence of other important taxa such as *Acacia* (Family: Fabaceae) and *Eucalyptus* (Family: Myrtaceae). Both of these genera are useful indicators of habitat type and conditions and have been identified in previous analyses, at least to family level. In this study no *Eucalyptus* or *Acacia* sequences were identified, however Fabaceae sequences (possibly sub-family Mimosoideae) were detected. The *trnL* g/h primers used in this study have been tested successfully on *Acacia* and *Eucalyptus* reference samples and as such the absence of these taxa may be the result of primer biases, lack of genus-level resolution with the primers used, or simply a lack of DNA preservation and survival, which may vary between taxa or between preserved materials. Moreover, the presence of pollen from certain taxa does not guarantee the retrieval of DNA from such taxa. Previous studies have had difficulties in amplifying DNA from pollen due to the limited amount of DNA contained within pollen grains (Parducci *et al.*, 2005). Parducci *et al.* (2005) failed to retrieve plant DNA using *trnL* primers from horizons in which pollen from such plants was present. This may also serve to illustrate how it may be worthwhile to adopt a taxa specific approach in primer design to target important indicator species useful in the exploration in past environmental conditions and shifts, and highlights the value of multiple proxies in palaeoenvironmental reconstruction (discussed in detail in Jørgensen *et al.*, 2012). Additionally, the absence of previously detected taxa and the converse, may also suggest that the source of aDNA recovered from these middens may be macrofossil in origin or DNA bound to, or within, the urea matrix, as opposed to pollen.

3.2.5.4 Future considerations

The preservation of DNA is a complex process that is at the mercy of a number of biotic and abiotic factors, which act in unison causing DNA degradation and damage (Hofreiter *et al.*, 2001). Previous studies have shown that the survival of DNA is dependent not only on these factors but also the substrate in which DNA is found, which itself can mitigate the effects of DNA degradation and damage. Substrates such as hair (Gilbert *et al.*, 2004) and eggshell (Oskam *et al.*, 2010) are excellent at preserving DNA, with high levels of endogenous to microbial DNA, in comparison to bone for instance. In both of these cases, the substrate acts almost like a barrier to microbial attack, and in the case of hair in particular the substrate acts as a barrier to

water. Midden material from cold, arid environments has been shown to preserve DNA over time (Kuch *et al.*, 2002; Hofreiter *et al.*, 2003b), and this study now shows that this is also the case with midden material from hot, arid environments. Hot, arid zone middens have very little moisture and the urine cementing the midden into a hard impermeable mass is highly ureic (Spaulding *et al.*, 1990). These high levels of urea may serve as a means to further desiccate middens in environments that already lack a significant amount of moisture (Spaulding *et al.*, 1990), thus aiding in the preservation of DNA. Moreover, lack of moisture therein also limits microbial induced DNA damage and degradation as well reducing hydrolytic damage. It would appear that the desiccation of midden material plays an important role in the long-term survival of DNA in middens. Importantly, these sites were in caves, rock shelters or overhangs and as such would have limited exposure to direct UV and weathering. It has been stated that there is much controversy surrounding aDNA claims arising from the study of hot, arid zone specimens and that there is a contrast between success rates of aDNA retrieval from similar sites of different ages (Schlumbaum *et al.*, 2008; Gilbert, 2011). It is clear from these previous studies that the retrieval of aDNA from samples within hot, arid environments is much more sporadic than that involving samples obtained from frozen or cooler environments, possibly giving rise to these differing success rates.

Regardless of the issues surrounding the preservation of DNA in hot, arid environments, there are a number of practical recommendations that would aid in the exploration of present and past metabarcoding data. In order to benefit fully from the wealth of data produced by current sequencing technologies it is essential to have well-populated and informative DNA and environmental databases. Current DNA databases are not sufficient to allow fine resolution of sequencing data and this may prove to be a major obstacle in some studies (Taylor & Harris, 2012). However, as DNA sequencing methods become cheaper and more accessible, the issues associated with insufficient database coverage are likely to diminish. In order to partly overcome this issue it is strongly recommended that a multi-primer approach targeting multiple loci be employed in environmental metabarcoding studies. This would provide a more comprehensive audit of environmental samples by reducing the effects of database biases and primer skews arising from preferential amplification. Although not used for plant screening in this study, this multi-locus

approach was employed for mammal screening with clear benefits. The use of both 16S and 12S mammal specific primers allowed for the confirmation of the presence of certain taxa such as *P. capensis* in the TK midden, whilst also detecting taxa not identified through the use of one or the other, for example *G. ocularis* in the TK midden (Table 3.2.2). Moreover, the detection of *G. ocularis* would only have been possible using the 12S primer set, as neither 16S nor conventional COI sequences are on GenBank. This also holds true for many of the Muridae species known to occur in the areas where the Australian middens were found. For many, there exist no *COXI* sequences for the currently accepted and approved *COXI* barcode on GenBank or BOLD (Barcode of Life Database; <http://www.boldsystems.org/>), and the same can be said of 12S and 16S sequences. The patchy genetic database coverage for Muridae sequences further illustrates the importance of using multiple loci in metabarcoding studies at present, be they loci accepted by the barcoding community or otherwise.

In addition to genetic databases, environmental databases using current and historical records of taxa distribution are invaluable in environmental metabarcoding studies. Databases such as ALA and SANBI, coupled with historical records, are immensely useful to truth and validate data or to detect possible range shifts of identified taxa. In relation to historical and ancient samples, the macro- and microscopic examination of environmental samples is highly valuable in determining the likelihood of DNA preservation and in the corroboration of genetic results, thereby improving data fidelity. The use of a range of resources to validate sequence data highlights the need for co-operation and collaboration between multiple disciplines ranging from palaeontology and archaeology, molecular biology and biochemistry through to ecology and botany. Through this concerted cross-disciplinary effort it would be possible to gain a more robust insight into both past and present environments.

3.2.6 Conclusion

The survival and preservation of DNA in hot, arid environments is a complex and poorly understood process. Most of the few studies that have attempted to retrieve aDNA from samples in such environments have been a source of controversy and dispute. The results in this study have been dealt with critically and overall they are both plausible and consistent with predicted outcomes and previous analyses of the

same samples. Although further empirical research is needed to assess the survival of DNA in midden material, it appears that DNA survival through accumulation and desiccation may be important in relation to samples from hot environments, and middens in general. Furthermore, it is apparent that neither the age of the samples nor the temperature at which they have been preserved, albeit important, can be grounds for the rejection of results. The preservation of DNA from hot environments, it suffices to say, is sporadic and rare.

Nonetheless, herbivore middens with their excellent preservational qualities now present an important source of material for DNA metabarcoding studies of past hot, arid environments, especially when palaeoenvironmental data are lacking, as was the case with the Brockman Ridge sample. As such, sampling procedures should be revised to ensure samples are collected in such a way as to allow for aDNA techniques to be applied. The retrieval of aDNA from midden material is not unique to Australia, as is evidenced by the results from South Africa and previous South American genetic studies. This study has wider implications for the analysis of midden material throughout hot, arid and semi-arid environments across the globe. Multidisciplinary investigations of midden material using stable isotopes, aDNA, pollen, macrofossils and dating will build knowledge of palaeoenvironments and inform conservation and rehabilitation policies. Such data will ensure the maintenance and survival of ecologically important taxa and communities within fragile arid environments, which are increasingly under anthropogenic induced threats.

3.2.7 Acknowledgements

DNA studies of these midden samples were funded by Australian Research Council (ARC) grants (DP0771971 and FT0991741). Fieldwork (CR, YR) was possible with an ARC grant awarded to J. Dodson and the support of D. Pearson and the School of Geography, University of New South Wales. Fieldwork (BR) was undertaken by Scarp Archaeology with the approval of the Puutu Kunti Kurrama and Pinikura group who assisted in the archaeological work that was funded by Rio Tinto Iron Ore. Radiocarbon dating of CR & YR material was made possible by grants from Australian Institute of Nuclear Science (03/704) and permits from the Western

Australian Department of Conservation and Land Management. Support was also received from the European Research Council (ERC) Starting Grant project HYRAX (no. 258657), and the Leverhulme Trust grant F/08 773/C. The authors acknowledge the support and contribution of Eske Willerslev (Centre for GeoGenetics), Fred Ford (Department of Defence), Ms Frances Brigg (State Agricultural Biotechnology Centre) and computational support from the iVEC Informatics Facility.

3.2.8 References

Adcock, G. J., Dennis, E. S., Eastal, S., Huttley, G. A., Jermini, L. S., Peacock, W. J., & Thorne, A. (2001). Mitochondrial DNA sequences in ancient Australians: Implications for modern human origins. *proceedings of the National Academy of Sciences*, *98*, 537-542.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403-410.

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K. H., . . . Willerslev, E. (2011). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, *21*, 1966-1979.

Atchison, J. (2010). *Short report on Pilbara amberat samples from Brock 12, Pilbara, Western Australia*.

Beadle, N. C. W. (1966). Soil phosphate and its role in moulding segments of Australian flora and vegetation with special reference to xeromorphy and sclerophylly. *Ecology*, *47*, 992-1020.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research*, *34*, D16-D20.

Bonnichsen, R., Hodges, L., Ream, W., Field, K. G., Kirner, D. L., Selsor, K., & Taylor, R. E. (2001). Methods of the study of ancient hair: Radiocarbon dates and gene sequences from individual hairs. *Journal of Archaeological Science*, *28*, 775 - 785.

Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One*, *5*, e13042.

Chariton, A. A., Court, L. N., Hartley, D. M., Colloff, M. J., & Hardy, C. M. (2010). Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, *8*, 233-238.

Chase, B. M., Meadows, M. E., Scott, L., Thomas, D. S. G., Marais, E., Sealy, J., & Reimer, P. J. (2009). A record of rapid Holocene climate change preserved in hyrax middens from southwestern Africa. *Geology*, *37*, 703-706.

Chase, B. M., Quick, L. J., Meadows, M. E., Scott, L., Thomas, D. S. G., & Reimer, P. J. (2011). Late glacial interhemispheric climate dynamics revealed in South African hyrax middens. *Geology*, *39*, 19-22.

Chase, B. M., Scott, L., Meadows, M. E., Gil-Romera, G., Boom, A., Carr, A. S., . . . Quick, L. J. (In press). Rock hyrax middens: a palaeoenvironmental archive for southern African drylands. *Quaternary Science Reviews*.

Clarke, E. (2010). *A short report detailing the salvage of sites Brock-11 and Brock-12*.

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: Do it right or not at all. *Science*, *289*, 1139-1139.

Cooper, A., Rambaut, A., Macaulay, V., Willerslev, E., Hansen, A. J., & Stringer, C. (2001). Human origins and ancient human DNA. *Science*, *292*, 1655 - 1656.

Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.

Deagle, B. E., Kirkwood, R., & Jarman, S. N. (2009). Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, *18*, 2022-2038.

- Dial, K. P., & Czaplewski, N. J. (1990). Do woodrat middens accurately represent the animals' environment and diets? The Woodhouse Mesa study. In J. L. Betancourt, T. R. Van Devender, & P. S. Martin (Eds.), *Packrat Middens: The Last 40,000 Years of Biotic Change* (pp. 43-58). Tucson: University of Arizona.
- Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., . . . Wilson, A. (2011). Geneious v5.4, Available from <http://www.geneious.com/>. Retrieved from <http://www.geneious.com/>
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*, 2194-2200.
- Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S., & McClure, M. O. (2011). DNA extraction columns contaminated with murine sequences. *PLoS One*, *6*, e23484.
- Fall, P. L., Lindquist, C. A., & Falconer, S. E. (1990). Fossil Hyrax middens from the Middle East: A record of paleovegetation and human disturbance. In J. L. Betancourt, T. R. Van Devender, & P. S. Martin (Eds.), *Packrat Middens: The Last 40,000 Years of Biotic Change* (pp. 398-407). Tucson: University of Arizona.
- Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423-425.
- Friedel, M. H., Pickup, G., & Nelson, D. J. (1993). The interpretation of vegetation change in a spatially and temporally diverse arid Australian landscape. *Journal of Arid Environments*, *24*, 241-260.
- Gilbert, M. T. P. (2011). The mummy returns... and sheds new light on old questions. *Molecular Ecology*, *20*, 4195-4198.
- Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., & Barnes, I. (2005a). Assessing ancient DNA studies. *Trends in Ecology & Evolution*, *20*, 541-544.

- Gilbert, M. T. P., Barnes, I., Collins, M. J., Smith, C., Eklund, J., Goudsmit, J., . . . Cooper, A. (2005b). Long-term survival of ancient DNA in Egypt: Response to Zink and Nerlich (2003). *American Journal of Physical Anthropology*, *128*, 110-114.
- Gilbert, T. P., Wilson, A. S., Bunce, M., Hansen, A. J., Willerslev, E., Shapiro, B., . . . Cooper, A. (2004). Ancient mitochondrial DNA from hair. *Current Biology*, *14*, R463-R464.
- Gomez-Alvarez, V., Teal, T. K., & Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *The ISME Journal*, *3*, 1314-1317.
- Griffiths, R. I., Thomson, B. C., James, P., Bell, T., Bailey, M., & Whiteley, A. S. (2011). The bacterial biogeography of British soils. *Environmental Microbiology*, *13*, 1642-1654.
- Groves, R. H. (1994). *Australian Vegetation* (Second edition ed.). Cambridge: Cambridge University Press.
- Haile, J. (2011). Ancient DNA extraction from soils and sediments. In B. Shapiro, Hofreiter, M. (Ed.), *Methods in Molecular Biology - Ancient DNA* (pp. 57-63): Humana Press Series.
- Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., . . . Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *proceedings of the National Academy of Sciences*, *106*, 22352-22357.
- Hayward, G. F., & Phillipson, J. (1979). Community structure and the functional role of small mammals in ecosystems. In M. Stoddart (Ed.), *Ecology of Small Mammals* (pp. 135-211). London: Chapman & Hall.
- Hekkala, E., Shirley, M. H., Amato, G., Austin, J. D., Charter, S., Thorbjarnarson, J., . . . Blum, M. J. (2011). An ancient icon reveals new mysteries: mummy DNA resurrects a cryptic species within the Nile crocodile. *Molecular Ecology*, *20*, 4199-4215.

- Hijmans, R., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965-1978.
- Hofreiter, M., Betancourt, J. L., Pelliza Sbriller, A., Markgraf, V., & McDonald, H. G. (2003a). Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Curá, Neuquén Province, southwest Argentina. *Quaternary Research*, *59*, 364-378.
- Hofreiter, M., Betancourt, J. L., Pelliza Sbriller, A., Markgraf, V., & McDonalde, H. G. (2003b). Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Curá, Neuquén Province, southwest Argentina. *Quaternary Research*, *59*, 364-378.
- Hofreiter, M., Mead, J. I., Martin, P., & Poinar, H. N. (2003c). Molecular caving. *Current Biology*, *13*, R693-R695.
- Hofreiter, M., Poinar, H. N., Spaulding, W. G., Bauer, K., Martin, P. S., Possnert, G., & Pääbo, S. (2000). A molecular analysis of ground sloth diet through the last glaciation. *Molecular Ecology*, *9*, 1975-1984.
- Hofreiter, M., Serre, D., Poinar, H., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews Genetics*, *2*, 353-359.
- Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One*, *6*, e19254.
- How, R. A., & Hillcox, S. J. (2000). Brushtail possum, *Trichosurus vulpecula*, populations in south-western Australia: demography, diet and conservation status. *Wildlife Research*, *27*, 81-89.
- Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*, 754-755.
- Hunter, A. A., Macgregor, A. B., Szabo, T. O., Wellington, C. A., & Bellgard, M. I. (2012). Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine*, *7*, 1.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*, 377-386.

- Jørgensen, T., Haile, J., Möller, P. E. R., Andreev, A., Boessenkool, S., Rasmussen, M., . . . Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and microfossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, *21*, 1989-2003.
- Kuch, M., Rohland, N., Betancourt, J. L., Latorre, C., Steppan, S., & Poinar, H. N. (2002). Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert, Chile. *Molecular Ecology*, *11*, 913-924.
- Kuch, M., Sobolik, K., Barnes, I., Stankiewicz, B. A., Spaulding, G., Bryant, V., . . . Pääbo, S. (2001). A molecular analyses of the dietary diversity for three archaic native americans. . *Proceedings of the National Academy of Sciences, USA*, *98*, 4317-4322.
- Lindahl, T. (1993a). Instability and decay of the primary structure of DNA. *Nature*, *362*, 709-715.
- Lindahl, T. (1993b). Recovery of Antediluvian DNA. *Nature*, *365*, 700-700.
- Macphail, M. (2011). *Palynological analyses, 30 Ka 'amberat' deposit, Brockman Ridge, Pilbara region, Western Australia*.
- Magee, A. R. (2012). *Nanobubon hypogaeum* (Apiaceae), a new contractile-rooted species from the Western Cape Province of South Africa. *South African Journal of Botany*, *80*, 63-66.
- Magee, A. R., Van Wyk, B.-E., & Tilney, P. M. (2008). A taxonomic revision of the genus *Nanobubon* (Apiaceae: Apioideae). *South African Journal of Botany*, *74*, 713-719.
- McCarthy, L., Head, L., & Quade, J. (1996). Holocene palaeoecology of the northern Flinders Ranges, South Australia, based on stick-nest rat (*Leporillus* spp.) middens: a preliminary overview. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *123*, 1205-1218.
- Meadows, M. E., Chase, B. M., & Seliane, M. (2010). Holocene palaeoenvironments of the Cederberg and Swarttruggens mountains, Western Cape, South Africa: Pollen

and stable isotope evidence from hyrax dung middens. *Journal of Arid Environments*, 74, 789-793.

Mitchell, A. A., & Wilcox, D. G. (1994). *Arid shrubland plants of Western Australia* (Second ed.). Perth: University of Western Australia Press.

Moore, C. W. E. (1953). The vegetation of the south-eastern Riverina, New South Wales. II. The disclimax communities. *Australian Journal Botany*, 1, 548-567.

Murray, D., Bunce, M., Cannell, B. L., Oliver, R., Houston, J., White, N. E., . . . Haile, J. (2011). DNA-based faecal dietary analysis: a comparison of qPCR and High Throughput Sequencing approaches. *PLoS One*, 6, e25776.

Northcote, K. H., & Wright, M. J. (1982). Soil landscapes of arid Australia. In W. R. Barker & P. J. M. Greenslade (Eds.), *Evolution of the Flora and Fauna of Arid Australia* (pp. 15-21). Adelaide: Peacock Publications.

Oskam, C. L., Haile, J., McLay, E., Rigby, P., Allentoft, M. E., Olsen, M. E., . . . Bunce, M. (2010). Fossil avian eggshell preserves ancient DNA. *Proceedings of the Royal Society B-Biological Sciences*, 277, 1991-2000.

Parducci, L., Suyama, Y., Lascoux, M., & Bennett, K. D. (2005). Ancient DNA from pollen: a genetic record of population history in Scots pine. *Molecular Ecology*, 14, 2873-2882.

Pearson, S. (1997). *Stick-nest rat middens as a source of palaeo-environmental data in central Australia*. (PhD), University of New South Wales, Sydney.

Pearson, S., & Betancourt, J. L. (2002). Understanding arid environments using fossil rodent middens. *Journal of Arid Environments*, 50, 499-511.

Pearson, S., & Dodson, J. (1993). Stick-Nest Rat middens as sources of paleoecological data in Australian deserts. *Quaternary Research*, 39.

Pearson, S. G., Triggs, B. E., & Baynes, A. (2001). The record of fauna, and accumulating agents of hair and bone, found in middens of stick-nest rats (Genus *Leporillus*) (Rodentia: Muridae). *Wildlife Research*, 28, 435-444.

- Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and Earth System Sciences*, *11*, 1633-1644.
- Pegard, A., Miquel, C., Valentini, A., Coissac, E., Bouvier, F., François, D., . . . F., P. (2009). Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from faeces. *Journal of Agricultural and Food Chemistry*, *57*, 5700-5706.
- Poinar, H., Kuch, M., McDonald, G., Martin, P., & Pääbo, S. (2003). Nuclear gene sequences from a late Pleistocene sloth coprolite. *Current Biology*, *12*, 1150-1152.
- Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., . . . Pääbo, P. (1998). Molecular coproscopy: dung and diet of the extinct Ground Sloth *Nothrotheriops shastensis*. *Science*, *281*, 402-406.
- Poinar, H. N., Kuch, M., Sobolik, K. D., Barnes, I., Stankiewicz, A. B., Kuder, T., . . . Pääbo, S. (2001). A molecular analysis of dietary diversity for three archaic Native Americans. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 4317-4322.
- Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., . . . Schuster, S. C. (2006). Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science*, *311*, 392-394.
- Pons, A., & Quézel, P. (1958). Premières remarques sur l'étude palynologique d'un guano fossile du Hoggar. *Comptes Rendus des Séances de l'Acad. Sci.*, *244*, 2290-2292.
- Ritchie, J. C. (1986). Climate change and vegetation response. *Vegetatio*, *67*, 65-74.
- Roche. (2009). Technical Bulletin: Amplicon fusion primer design guidelines for GS FLX Titanium series Lib-A chemistry. *TCB No. 013-2009*, 1-3.
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yoosheph, S., . . . Venter, J. C. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, *5*.

- Schlumbaum, A., Tensen, M., & Jaenicke-Després, V. (2008). Ancient plant DNA in archaeobotany. *Vegetation History and Archaeobotany*, *17*, 233-244.
- Scott, L. (1990). Hyrax (Procaviidae) and Dassie Rat (Petromuridae) middens in paleoenvironmental studies in Africa. In J. L. Betancourt, T. R. Van Devender, & P. S. Martin (Eds.), *Packrat Middens: The Last 40,000 Years of Biotic Change* (pp. 398-407). Tucson: University of Arizona.
- Scott, L., Marais, E., & Brook, G. A. (2004). Fossil hyrax dung and evidence of Late Pleistocene and Holocene vegetation types in the Namib Desert. *Journal of Quaternary Science*, *19*, 829-832.
- Scott, L., & Woodborne, S. (2007). Pollen analysis and dating of late Quaternary faecal deposits (hyraceum) in the Cederberg, Western Cape, South Africa. *Review of Palaeobotany and Palynology*, *144*, 123-134.
- Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*, 1794-1805.
- Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B., & Collins, M. J. (2001). Neanderthal DNA. Not just old but old and cold? *Nature*, *410*, 771-772.
- Smith, C. I., Chamberlain, A. T., Riley, M. S., Stringer, C., & Collins, M. J. (2003). The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution*, *45*, 203-217.
- Spaulding, W. G., Betancourt, J. L., Croft, L. K., & L., C. K. (1990). Packrat middens: Their composition and methods of analysis. In J. L. Betancourt, T. R. Van Devender, & P. S. Martin (Eds.), *Packrat Middens: The Last 40,000 Years of Biotic Change* (pp. 59-84). Tucson: University of Arizona.
- Spaulding, W. G., & Robinson, S. W. (1984). *Preliminary assessment of climatic change during the later Wisconsin Time, southern Great Basin and vicinity.*

- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045-2050.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., . . . Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*, e14.
- Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three noncoding regions of chloroplast DNA. *Plant Molecular Biology*, *17*, 1105-1109.
- Tausch, R. J., Wigand, P. E., & Burkhardt, J. W. (1993). Viewpoint - plant community thresholds, multiple steady states and multiple successional pathways: legacy of the Quaternary? *Journal of Rangeland Management*, *46*, 439-447.
- Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, *12*, 377-388.
- Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*, 283-285.
- Thackway, R., & Cresswell, I. D. (1995). *An interim biogeographic regionalisation for Australia: a framework for setting priorities in the National Reserves System Cooperative Program*.
- Thomsen, P. F., Kielgast, J. O. S., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., . . . Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565-2573.
- Tongway, D. J., & Ludwig, J. A. (1990). Vegetation and soil patterning in semi-arid mulga lands of Eastern Australia. *Australian Journal of Ecology*, *15*, 23-34.
- Trabucco, A., & Zomer, R. J. (2009). Global Aridity Index (Global-Aridity) and Global Potential Evapo-Transpiration (Global-PET) Geospatial Database. CGIAR

Consortium for Spatial Information. Published online, available from the CGIAR-CSI GeoPortal at: <http://www.csi.cgiar.org/>.

Tuke, P. W., Tettmar, K. I., Tamuri, A., Stoye, J. P., & Tedder, R. S. (2011). PCR master mixes harbour murine DNA sequences. Caveat emptor! *PLoS One*, *6*, e19953.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E. V. A., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009a). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, *9*, 51-60.

Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E. V. A., Coissac, E., Pompanon, F., . . . Taberlet, P. (2009b). New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Molecular Ecology Resources*, *9*, 51-60.

Van Devender, T. R. (1990). Late Quaternary vegetation and climate of the Sonoran Desert, United States and Mexico. In J. L. Betancourt, T. R. Van Devender, & P. S. Martin (Eds.), *Packrat Middens: The Last 40,000 Years of Biotic Change* (pp. 134-164). Tucson: University of Arizona.

Van Devender, T. R., & Spaulding, W. G. (1979). Development of vegetation and climate in the south western United States. *Science*, *204*, 701-710.

Vila, A. R., & Borrelli, L. (2011). Cattle in the Patagonian forests: feeding ecology in Los Alerces National Reserve. *Forest Ecology and Management*, *261*, 1306-1314.

Wells, P. V., & Jorgensen, C. D. (1964). Pleistocene wood rat middens and climatic change in Mohave Desert - a record of juniper woodlands. *Science*, *143*, 1171-1174.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., . . . Alan, C. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, *300*, 791-795.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

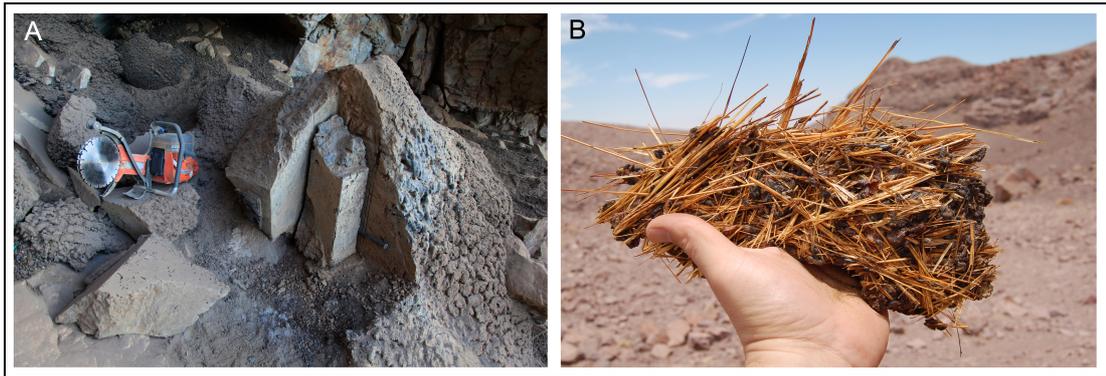


Figure S3.2.1 A comparison between African rock hyrax and American/Australian midden accumulation. A: African rock hyrax midden accumulation as a latrine. B: American/Australian midden accumulation as an organic-rich nest – *tut tut*...no gloves!



Figure S3.2.2 Walled middle entrance to Brockman Ridge rockshelter site. An example of the walled features at the cave complex where the Brockman Ridge midden was excavated.

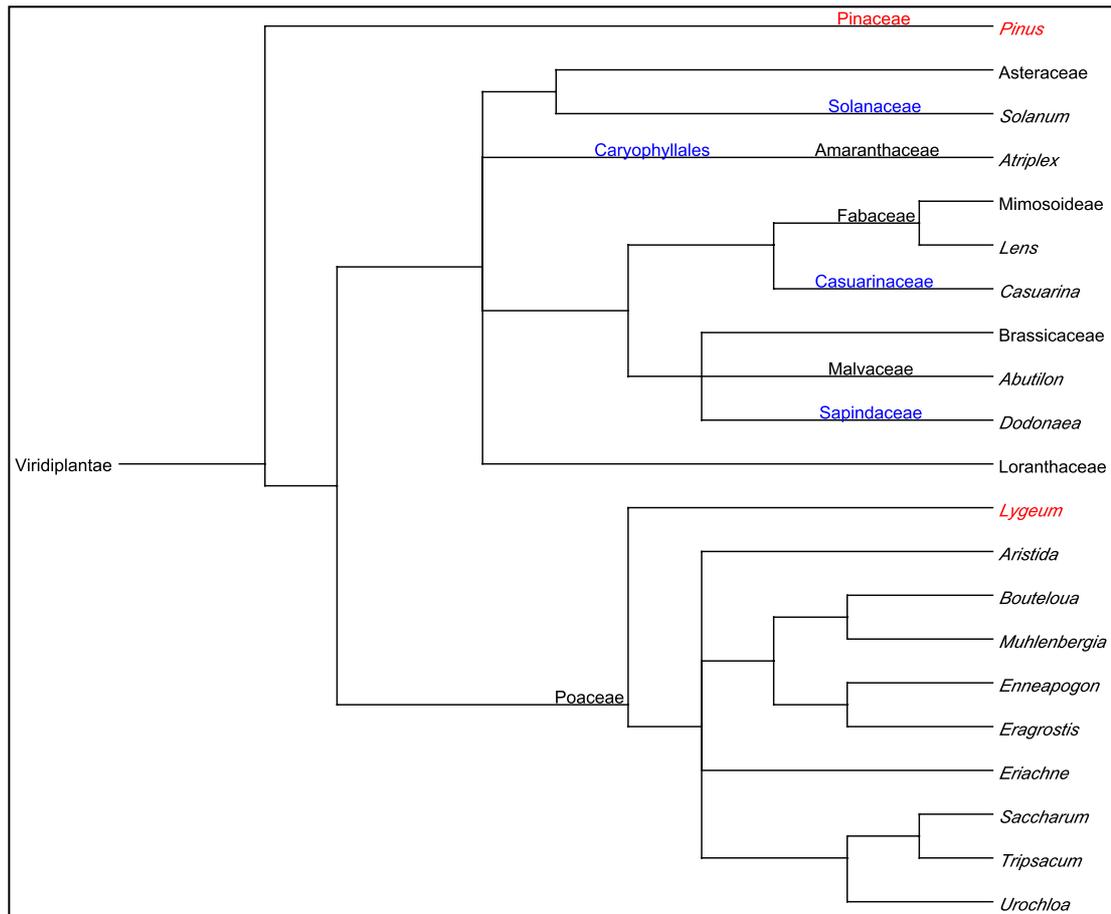


Figure S3.2.3A Cavenagh Range midden taxa identifications. Cladogram showing taxa identifications to genus level for the Cavenagh Range midden sample. Taxa in red are not found natively in Australia. Taxa in blue were identified via previous pollen analysis.

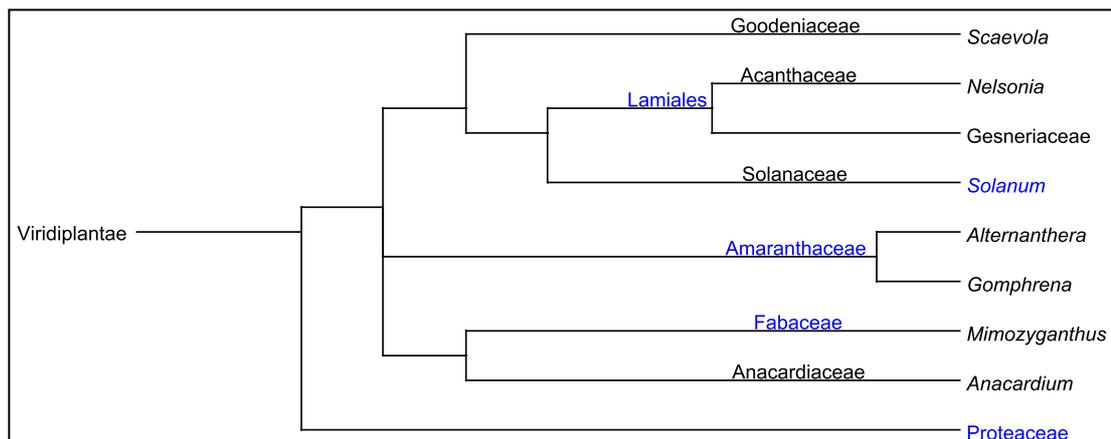


Figure S3.2.3B Young Range midden taxa identifications. Cladogram showing taxa identifications to genus level for the Young Range midden sample. Taxa in blue were identified via previous pollen and macrofossil analysis.

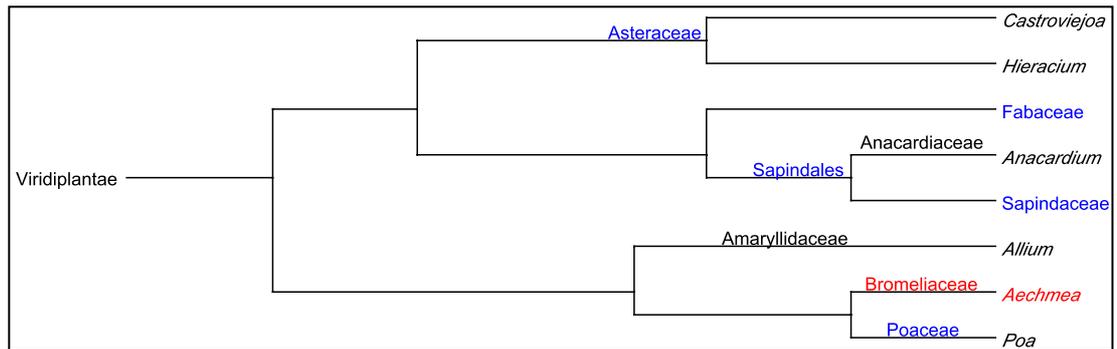


Figure S3.2.3C Pilbara midden taxa identifications. Cladogram showing taxa identifications to genus level for the Pilbara midden sample. Taxa in red are not found natively in Australia. Taxa in blue were identified via previous pollen analysis.

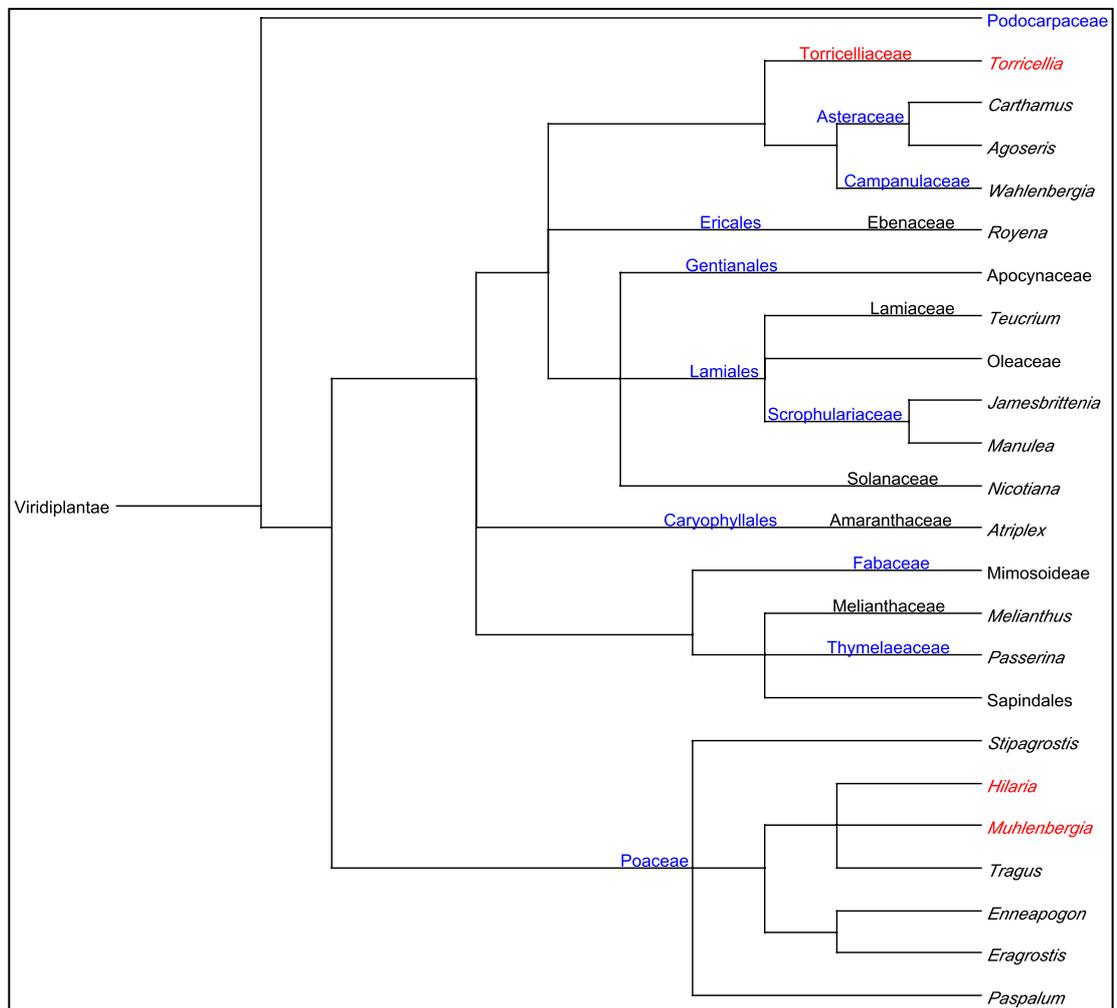


Figure S3.2.3D Truitjes Kraal taxa identifications. Cladogram showing taxa identifications to genus level for the Truitjes Kraal midden sample. Taxa in red are not found natively in South Africa. Taxa in blue were identified via previous pollen analysis.

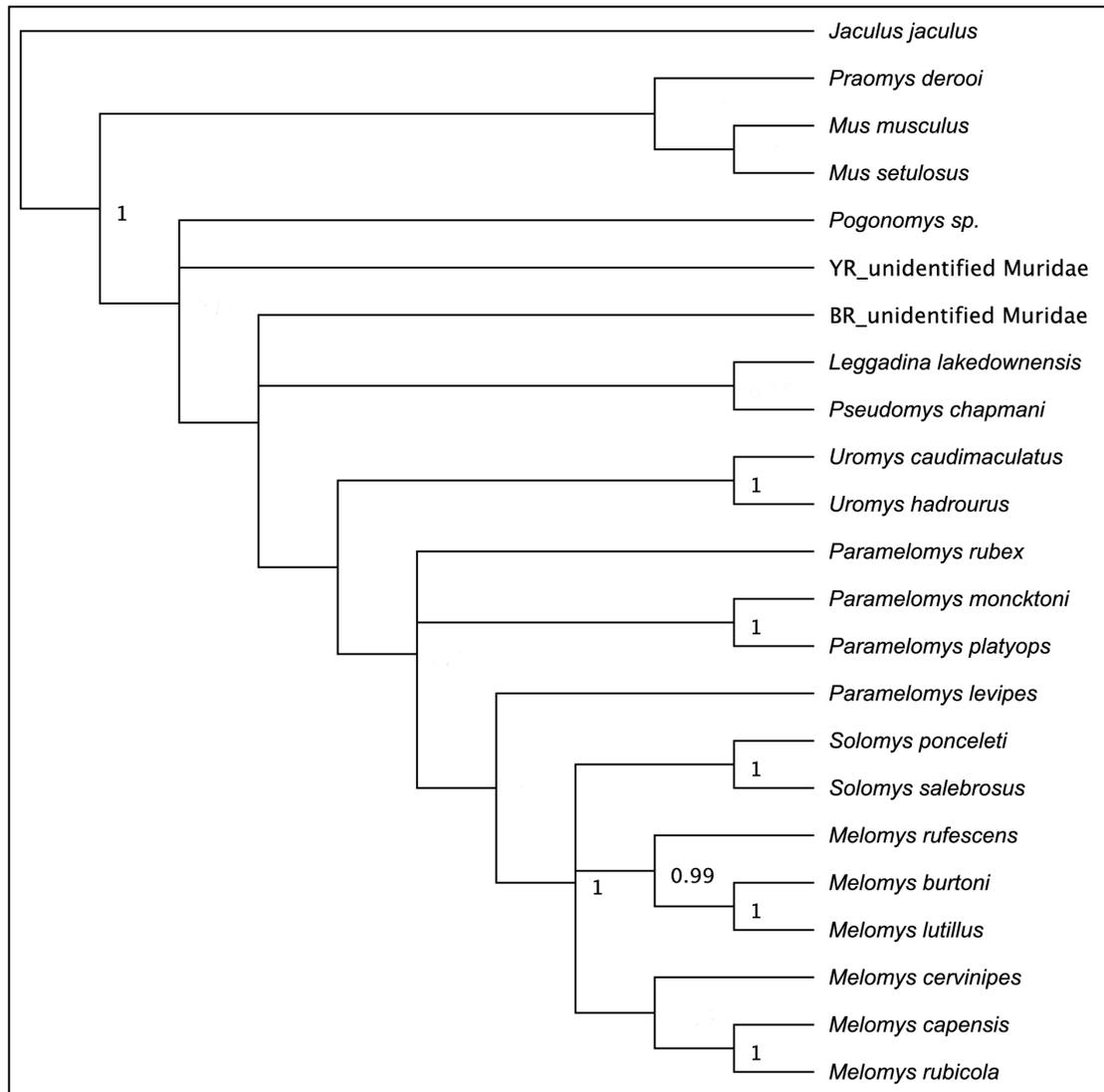


Figure S3.2.4 Phylogenetic tree comparing Muridae sequences obtained from Brockman Ridge and Young Range midden samples. Posterior output tree generated using MrBayes (Huelsenbeck and Ronquist, 2001) in Geneious v5.6.4 (Drummond et al., 2011), following 1,000,000 iterations using a HKY and gamma invariant model of evolution, with *Jaculus jaculus* selected as an outgroup. The difference between the Young Range and Brockman Ridge Muridae 16S OTUs from each other and to *Mus musculus* can be seen. Node labels indicate posterior probabilities, whilst YR and BR indicate Young Range and Brockman Ridge respectively.

Table S3.2.1 Radiocarbon age estimates of sampled middens.

Midden Sample	Lab code (submitter's code)	Submitter's code	Sample Type	Conventional Age (BP)
Cavenagh Range	OZB168U	CAA04	Amberat	3,430 ±50
Young Range	BETA30956	MO505	Plant fragment	710 ±80
Brockman Ridge	BETA268576	B12BASERAT	Amberat	30,490 ±380
Truitjes Kraal	See Meadows <i>et al.</i> , 2010			

3.3 Synopsis

Prior to this study, it was well-established that herbivore middens were valuable repositories of genetic information that could aid in the study of past environments. DNA preservation in herbivore midden material from cooler environments has on the whole been generally good, but that was no guarantee that it would be successful in hot environments such as Australia or South Africa.

This study was successful in extracting and characterising plant and animal DNA from herbivore middens sourced from locations often overlooked in aDNA studies due to the inhospitable environments in which they are found. Teasing ancient DNA from any hot environment, be it from midden material or otherwise is still surprising and is rare in the literature. In some instances, such as the Brockman Ridge midden from Australia's Pilbara region, it may be the only source of available aDNA in these regions. Moreover, the samples used are amongst some of the oldest environmental samples tested in Australia to date.

The successful application of aDNA and HTS techniques to midden material in Australia is promising in a continent where the study of palaeoenvironments is often frustrated by the poor or non-existent preservation of plant and animal material. It is not without its difficulties as discussed in the manuscript and throughout this thesis; however, the issues associated with it are not insurmountable. Indeed, means of analysing genetic signatures within samples such as middens or sediment in taxonomy-independent ways are possible thus allowing some degree of analysis to explore past changes in genetic diversity (explored further in Chapters Four-Seven). Such methods provide a useful tool to analyse changes in diversity in regions that are poorly characterised taxonomically, such as the Australian Southwest Floristic Province: a world-renowned biodiversity hotspot of ecological and archaeological importance (Chapter Six).

Chapter Two highlighted the importance of collecting data over a wide time period to gain a fuller insight into ecological and environmental shifts over time, while Chapter Three has shown that, despite the difficulties associated with the hot

Australian climate, it is possible to extend our knowledge of the past through the use of aDNA in less than desirable climates. Chapter Four continues with the theme of understanding past environments and builds upon both of these papers in order to develop a method to screen bulk-bone fragments excavated from archaeological sites to better understand how faunal assembles have changed over the past ~50,000 years.

Chapter Four – A novel method to analyse archaeological waste material

4.1 Preface

Chapter Four details the development of a bulk-bone metabarcoding strategy to sample and genetically identify fragmentary bone. This study resulted in the published manuscript 'Scrapheap challenge: a novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages' (Scientific Reports 2013, 3, 3371). With the exception of formatting and in-thesis referencing this manuscript has been reproduced as published.

Chapter Three demonstrated that it is possible to extract and successfully characterise aDNA from 'ancient' and historical environmental samples sourced from within Australia. In Chapter Four a novel bulk-bone metabarcoding (BBM) methodology is presented whereby hundreds of bone fragments spanning the past several thousand years may be sampled and characterised genetically in both a taxonomy-dependent and taxonomy-independent manner.

Fragmentary bone is found in abundance at palaeontological and archaeological sites worldwide. It is seldom used in site analyses as it lacks any morphologically identifiable features. This chapter sets out to determine whether it is possible to sample these bone fragments in a high-throughput manner by grouping bones from similar time periods into a single bulk sample; in essence creating a synthetic environmental sample akin to the faecal and midden samples in Chapters Two and Three. If possible, it would represent a novel, rapid and cost-effective means to characterise archaeological and palaeontological sites without the need to destructively sample valuable bone material. It would also make available a wealth of material to be used in aDNA and HTS analyses that would otherwise be discarded or left sitting unused in museum collections.

4.1.1 Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

4.2 Scrapheap Challenge: a novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages

Dáithí C. Murray^{1#}, James Haile^{1#}, Joe Dortch², Nicole E. White^{1#}, Dalal Haouchar¹, Matthew I. Bellgard³, Richard J. Allcock⁴, Gavin J. Prideaux⁵ and Michael Bunce^{1#}

¹ *Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, WA, 6150, Australia.*

² *Eureka Archaeological Research and Consulting, School of Social Sciences, The University of Western Australia, Crawley, Perth, WA, 6009, Australia.*

³ *Centre for Comparative Genomics, Murdoch University, South Street, Murdoch, WA, 6150, Australia.*

⁴ *LotteryWest State Biomedical Facility: Genomics, School of Pathology and Laboratory Medicine, The University of Western Australia, Nedlands, WA, 6009, Australia*

⁵ *School of Biological Sciences, Flinders University, Bedford Park, SA, 5042, Australia*

[#]*Current Address: Trace and Environmental DNA laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, 6845, Australia.*

4.2.1 Abstract

Highly fragmented and morphologically indistinct fossil bone is common in archaeological and paleontological deposits but unfortunately it is of little use in compiling faunal assemblages. The development of a cost-effective methodology to taxonomically identify bulk bone is therefore a key challenge. Here, an ancient DNA methodology using high-throughput sequencing is developed to survey and analyze thousands of archaeological bones from southwest Australia. Fossils were collectively ground together depending on which of fifteen stratigraphical layers they were excavated from. By generating fifteen synthetic blends of bulk bone powder, each corresponding to a chronologically distinct layer, samples could be collectively

analyzed in an efficient manner. A diverse range of taxa, including endemic, extirpated and hitherto unrecorded taxa, dating back to c.46,000 years BP were characterized. The method is a novel, cost-effective use for unidentifiable bone fragments and a powerful molecular tool for surveying fossils that otherwise end up on the taxonomic "scrapheap".

4.2.2 Introduction

Fossil assemblages offer insights into past biodiversity, paleoecology and human activities (Dortch & Wright, 2010; Archibald *et al.*, 2013; Colonesea *et al.*, 2013). However, the accuracy of fossil identifications relies on the preservation of taxonomically significant morphological features, which are often lacking in highly fragmented remains. Over the past decade, analyses of ancient DNA (aDNA) have developed in sophistication and the breadth of contexts in which they are applied. Ancient DNA has been used to address questions of speciation, extinction and disease (Raoult *et al.*, 2000; Worobey *et al.*, 2008; Haile *et al.*, 2009; Rohland *et al.*, 2010) using a variety of substrates, including bone (Smith *et al.*, 2001), hair (Bonnichsen *et al.*, 2001) and eggshell (Oskam *et al.*, 2010). However, to date, no study has attempted to use aDNA from taxonomically diverse fossils to map faunal assemblage data from a single site, largely due to the time and cost associated with generating aDNA sequences from each bone fragment.

The destructive nature of sampling also means researchers and collection managers may be reluctant to analyze valuable specimens. At the same time, most archaeological and paleontological excavations also collect large numbers of small, morphologically indistinct bone fragments (Figure 4.2.1A). Such material is of limited use in species identifications, although it may be important for some taphonomic analyses. Taxonomically, however, it is usually destined for the analytical "scrapheap".

It is now possible, largely due to second generation high-throughput DNA sequencing (HTS) methodologies, to genetically profile complex, heterogeneous samples (Figure 4.2.1B) in parallel, both cheaply and quickly (Binladen *et al.*, 2007; Shokralla *et al.*, 2012). This DNA metabarcoding (Taberlet *et al.*, 2012) approach to

genetically unravel complex substrates via HTS, as opposed to cloning, has transformed the analysis of substrates such as sediment (Jørgensen *et al.*, 2011; Jørgensen *et al.*, 2012) and fecal material (Deagle *et al.*, 2010; Murray *et al.*, 2012). To explore large HTS-generated genomic datasets from environmental samples researchers use tools that are either: 1) taxonomy-dependent, which involves searching DNA reference databases for query and reference sequence matches (Altschul *et al.*, 1990; Little, 2011), or 2) taxonomy-independent, which involves taxonomy-independent measures of sequence diversity and clustering such as Operational Taxonomic Unit (OTU) analysis or UniFrac-based methods (Schloss & Handelsman, 2005; Caporaso *et al.*, 2010; Hamady *et al.*, 2010).

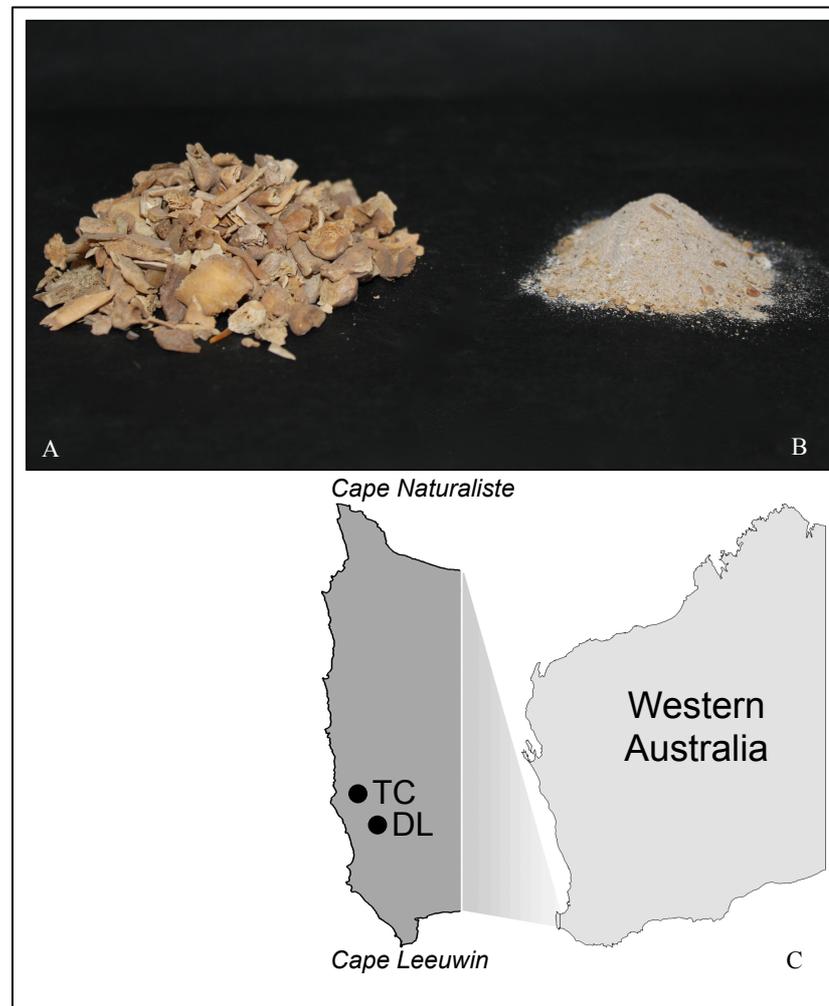


Figure 4.2.1 Bulk-bone fragments ground to form a bulk-bone powder at two archaeological sites. Morphologically indistinct bulk-bone fragments (A) were ground to form single bulk-bone powder samples (B). Bulk-bone fragments were excavated from Devil's Lair (DL) and Tunnel Cave (TC), two archaeologically significant sites in southwest Western Australia (C).

This study seeks to employ HTS technology to sequence and identify aDNA obtained from thousands of morphologically unidentifiable archaeological bone fragments freshly excavated from deposits at Tunnel Cave (115° 02' E, 34° 05' S) and Devil's Lair (115° 04' E, 30° 09' S), two archaeologically and culturally significant sites in southwestern Australia (Figure 4.2.1C). Taken together, these sites, used to explore this methodological approach, span the last c.50,000 years (Dortch, 2004) and provide an unparalleled opportunity to study past Australian biodiversity and Aboriginal occupation (Dortch, 2004) located within an internationally recognized biodiversity “hotspot” (Myers *et al.*, 2000). A new method for the bulk sampling of fragmented bone material that would otherwise remain an untapped taxonomic resource is presented. By grinding multiple bones (Figure 4.2.1A) into an artificial “bulk-bone powder” (Figure 4.2.1B), thus producing a single bulk-bone powder sample, a large amount of highly informative genetic data can be quickly extracted. Such an approach should become commonplace in archaeological and paleontological practice as it enables rapid assessment of DNA preservation and effectively maps zooarchaeological and paleontological assemblages without destructive sampling of more valuable fossils.

4.2.3 Methods

4.2.3.1 Sample collection and processing

Thousands of indistinct bone fragments were collected from both Tunnel Cave and Devil's Lair during excavations in February 2012. Approximately 150 L (0.15 m³) of sediment was analysed at both sites. Sediment was dry-sieved on site, using 2mm and 5mm sieves, and bagged according to well-defined and dated stratigraphical layers (Dortch, 2004). Each bagged sample was screened for bone fragments off-site, which were kept in groupings according to the layers in which they were found. Fifteen bulk-bone samples representing fifteen layers were processed: eight from Tunnel Cave, covering a period from 4,160 – 24,110 years BP (uncalibrated) (Dortch, 2004), and seven from Devil's Lair, covering a period from 6,200 – 46,890 years BP (uncalibrated) (Dortch, 2004). Small sections of the bones within each layer (typically 50-150 bones) were drilled (Dremel 114 drill bits) for a few seconds each and approximately equal amounts of drilled material from each bone fragment within a single layer was combined to form a “bulk-bone powder.” Owing to inherent

differences in the amount of DNA per unit of biomass between species and differential DNA preservation between individual bones, over-representation of certain bone material in terms of DNA amplicon sequences is unavoidable.

4.2.3.2 DNA extraction and screening

All laboratory work was conducted in keeping with standard aDNA protocols (Cooper & Poinar, 2000). Approximately 1g of bulk-bone powder from each sample, including a blank extraction control, was digested overnight on a lab rotator at 55°C in 5mL of digestion buffer containing: 2.5 mL EDTA (0.5M), 0.1 mL Tris-HCL (1 M), 5 mg Proteinase K powder, 50 µL DTT (1 M), 50 µL SDS (10 %) and made up to a final volume of 5 mL using EDTA. DNA digests were centrifuged at 6,000 rpm for 2 mins and the supernatant was concentrated to 50 µL using AMICON 30,000 MWCO columns (Millipore) as per the manufacturer's instructions. Each concentrate was transferred to a clean 2 mL eppendorf tube and PBi buffer (Qiagen) totalling 250 µL (i.e. 5X the volume of concentrate) was added. Each 300 µL PBi/concentrate mix was subsequently transferred to Qiagen silica spin columns and centrifuged at 13,000 rpm. Columns were washed with 700 µL of AW1 followed by AW2. A final dry spin at 13,000 rpm for 1 min followed. DNA was eluted from the columns in 60 µL EB with a 1 min incubation at room temperature prior to centrifugation at 13,000 rpm for 1 min.

Extracts were screened for amplifiable mtDNA using multiple primer sets via qPCR at three concentrations - undiluted, 1/10 and 1/50. Extracts were screened for mammalian mtDNA using 12S A/O and 16Smam primer sets, designed to amplify a small region within mammalian 12S and 16S mitochondrial genes respectively (Taylor, 1996; Cooper *et al.*, 2001). Extracts were also tested for avian mtDNA using 12S A/E and 12S A/H primer sets, designed to amplify a short and slightly longer overlapping region of the avian mitochondrial 12S gene respectively (Cooper *et al.*, 2001). Finally, extracts were tested for snake mtDNA using the following primers: 12s_tRNA_F1_S AAAGTATAGCACTGAAAATGCTAA and 12s_R1_Snake GTTAGCCTGATACCGGCTCCG, designed to amplify a short region within the mitochondrial 12S gene. Each qPCR reaction was made up to a total volume of 25 µL, containing 1X PCR Gold Buffer (Applied Biosystems), 2.5 mM MgCl₂ (Applied Biosystems), 0.4 mg/mL BSA (Fisher Biotech, Aus), 0.25 mM of each dNTP (Astral

Scientific, Aus), 0.4 μ M forward primer, 0.4 μ M reverse primer, 0.25 μ L AmpliTaq Gold (Applied Biosystems), 0.6 μ L SYBR Green (1:2,000, Life Sciences gel stain solution) and 2 μ L DNA extract. Quantitative PCR cycling conditions for the 12S A/O and snake 12S qPCR assays were as follows: initial heat denaturation at 95 °C for 5 mins, followed by 50 cycles of 95 °C for 30 s; 55 °C for 30 s (annealing step); 72 °C for 45 s followed by a 1 °C melt curve and final extension at 72 °C for 10 mins. Cycling conditions for 16Smam, 12S A/E and 12S A/H assays were the same as for the 12S A/O assay, except the annealing temperature, which was 57 °C in each case. For each qPCR assay, DNA extraction, negative PCR reagent and positive DNA template controls were included.

4.2.3.3 DNA sequencing

DNA extracts that successfully yielded DNA of sufficient quality, free of inhibition, as determined by initial qPCR screening (Bunce *et al.*, 2011), were prepared for amplicon sequencing. DNA extracts successful for all primer sets were sequenced on Roche's GS-Junior. Additional, separate, amplicon sequences were generated for extracts using mammalian 12S A/O and 16Smam primer sets for sequencing on Life Technologies' Ion Torrent Personal Genome Machine (PGM).

For each primer set, DNA extracts were assigned a unique DNA tag (Binladen *et al.*, 2007). Each sample was tagged at both the 5' and 3' end of the target sequence using separate tags at both ends, resulting in a unique forward and reverse tag combination for each sequence. Independent tagged qPCRs for all samples, across all primer sets, were carried out in 25 μ L reactions with reaction components and cycling conditions as described in 'Methods: DNA extraction and screening.' Tagged qPCR amplicons were generated in triplicate and combined, thus minimizing the effects of PCR stochasticity on low-template samples, purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), as per manufacturer's instructions and eluted in 40 μ L H₂O. Purified amplicons were pooled to form separate sequencing libraries according to primer set used and sequencing platform. GS-Junior libraries were quantified using qPCR to determine an appropriate volume of library for sequencing (described in Murray *et al.* 2011). Each 25 μ L reaction contained 12.5 μ L ABI Power SYBR master mix (Applied Biosystems), 0.4 μ M A-adapter primer, 0.4 μ M B-adapter primer, 8.5 μ L H₂O and 2 μ L pooled library, with

the following cycling conditions: 95 °C for 5mins; 40 cycles of 95 °C for 15 s, 56 °C for 1min followed by a 1 °C melt curve. The appropriate library volume for use on the Ion Torrent PGM was determined using a Bioanalyser 2100 (Agilent). For each tagged qPCR assay, negative qPCR controls were included and if found to contain amplifiable DNA these qPCR amplicons were incorporated into the appropriate pooled sequencing library. All sequencing was performed as per manufacturer's instructions, with the use of 200 bp reagents and a 314 chip on the PGM.

4.2.3.4 Sequence identification

Amplicon sequence reads (hereafter referred to as sequences) were sorted into sample batches based on unique DNA tags. Identification tags and primers were trimmed allowing for no mismatch in length or base composition using Geneious v6.0.5 (created by Biomatters, available from <http://www.geneious.com/>). Batched and trimmed sequences from both GS-Junior and Ion Torrent PGM sequencing runs were combined according to sample and primer used. Each combined file was dereplicated, thus grouping sequences of exact identity and length, using USEARCH (Edgar, 2010). Dereplicated sequence files were searched for artificial chimeric sequences using the UCHIME *de novo* method (Edgar *et al.*, 2011) in USEARCH and were removed, in addition to sequences occurring only once (i.e. singletons). The remaining sequences in each sample were subsequently clustered at an identity threshold of 97 % using USEARCH with the most abundant sequence within each cluster selected as the representative sequence. To reduce noise associated with sequencing error, low abundant clusters, classed as those that occur at less than 1 % of the total number of unique sequences when clustered at 100 % sequence identity, were removed from the dataset. While the selection of a 1 % cut-off is somewhat arbitrary, it should negate the possibility of clusters remaining that are the result of sequencing error. Additionally, the decision to class clusters as being in low abundance with respect to the total number of unique sequences (as opposed to total number of sequences or total number of sequences within the most abundant cluster) was made to minimize the effects of preferential DNA preservation and/or amplification. For each sample, every sequence assigned to the remaining clusters were queried against the NCBI GenBank nucleotide database using BLASTn (Benson *et al.*, 2006) in YABI (Hunter *et al.*, 2012), enabling taxonomic identification. Sequences were searched without a low complexity filter, with a gap

penalties existence of five and extension of two, expected alignment value less than $1e-10$ and a word count of seven. The BLASTn results obtained were imported into MEtaGenome Analyzer v4 (MEGAN), where they were mapped and visualised against the NCBI taxonomic framework (min. bit score = 35.0, top percentage = 5 %, min. support = 1) (Huson *et al.*, 2007). Sequences that were obviously the result of contamination (primarily human and cow) were eliminated from all subsequent downstream analysis steps.

Sequences that were truncated when queried against the NCBI GenBank nucleotide database were discarded from taxonomic analysis. Sequences with percentage similarity to a reference below 90 % were discarded. Where sequence similarities were between 90-95 % these were assigned to a family level, while those between 95-100 % were assigned to a genus. Owing to the difficulties in assigning taxa beyond the genus level for some families, in addition to issues associated with characterizing past biodiversity that has been lost, species identifications were avoided in this particular study. Sequences that provided high percentage similarity to query references at a species level may or may not be bona fide, however with current insufficient data it is prudent to categorise these sequences cautiously. Where multiple taxa had equal percentage similarity scores to a query sequence, such sequences were moved higher up the taxonomic rankings.

While the validity of filters and hard percentage cut-offs are always debatable, those chosen in the analysis of this dataset seemed to afford the best balance when accounting for low template amounts and post-mortem damage on short aDNA fragments.

4.2.3.5 Genetic biodiversity analysis

Cognisant of the difficulties associated with assigning sequences to lower taxonomic levels, a modified form of OTU analysis was applied to the 16S_{mam} sequences obtained in this study. This allowed changes in observed genetic diversity over time at both sites to be investigated independently of the above taxonomic classifications. Sequences within each sample were clustered at 97 % identity, filtered and representative sequences were selected as detailed in Methods: Sequence Identification. Representative sequences within each sample were aligned in

Geneious using MAFFT's G-INS-I algorithm and default parameters (Katoh *et al.*, 2002). MAFFT alignments were imported into MEGA5 (Tamura *et al.*, 2011) where a distance matrix between OTUs within a sample was calculated using a Kimura 2-parameter model (Kimura, 1980), with all positions containing gaps and missing data ignored. OTUs less than 3 % divergent from each other were collapsed into a single DTU. This serves the purpose of reducing the influence of HTS homopolymer sequencing error (Loman *et al.*, 2012; Quail *et al.*, 2012) by collapsing multiple homopolymer-derived OTUs into a single DTU, as errors in homopolymer stretches appear as gaps and are not included in the calculation of the distance matrix. Whilst this is first and foremost a largely taxonomic-independent analysis it is still nonetheless useful to identify coarsely to which family each DTU belongs, as this gives an idea of the diversity of DTUs within specific families. As such, all DTUs were searched against the NCBI GenBank nucleotide database using BLASTn (Benson *et al.*, 2006) to identify the family to which each DTU could be easily assigned. For the faunal specific Macropodidae DTU analysis the same method as above was followed except that only sequences assigned to Macropodidae were selected.

4.2.4 Results

4.2.4.1 Overview of data generated

In a 2012 excavation, thousands of small bone fragments were collected by dry-sieving sediment from 15 well-dated stratigraphic units or layers at Devil's Lair and Tunnel Cave (Figure 4.2.1C). Around 50–150 bone fragments from within each layer were each drilled for 10-15s to form 15 bulk-bone powder samples representing the 15 layers (Figures 4.2.1A & B). DNA was extracted from each bulk-bone powder sample using established extraction methods (described in Section 4.2.3) as if the bulk-bone sample were a single-source sample. The DNA extracts were screened for amplifiable mitochondrial DNA (mtDNA) using generic primers (tagged with HTS adaptors and unique barcodes) and subsequently sequenced using two HTS platforms: the GS-Junior (Roche) and the Ion Torrent PGM (Life Technologies).

Ancient DNA was successfully extracted from all bulk-bone powder samples, including a layer dated c.44,260 – 46,890 years BP (uncalibrated). The successful

amplification and sequencing of DNA from all 15 layers was a rapid, cheap and effective way to assess DNA preservation at the sites (Figure 4.2.1C).

Amplicon DNA sequences (hereafter referred to as sequences) obtained from collective GS-Junior and Ion Torrent PGM sequencing runs were analyzed for quality and possible chimeras. Except for ubiquitous human DNA sequences, control reactions throughout the process (described in Section 4.2.3) were negative for contaminating DNA arising from laboratory processing.

Short regions within the mammalian mitochondrial 12S and 16S rRNA genes were amplified generating products of 100-104 bp and 90-96 bp respectively (Taylor, 1996). Amplification and sequencing of avian mtDNA was successful for some samples, producing either a 106-121 bp or 227-239 bp region of the avian mtDNA 12S gene (Taylor, 1996). Some cross-species reactivity was observed when using both 12S and 16S mammalian primer sets, resulting in the amplification and sequencing of avian and reptilian DNA. A targeted quantitative PCR and HTS (qPCR) approach to identify snake species was successful for a single sample.

4.2.4.2 Taxonomic identification

Mammalian 12S and 16S assays identified eight mammalian families representing 16 genera, using assignment filters chosen for this study (see Section 4.2.3; Figure 4.2.2). The increase in sequencing depth afforded by the Ion Torrent PGM, as compared to the GS-Junior, did not increase the diversity of taxa identified. Mammalian taxa endemic to Australia were detected in multiple samples, in addition to taxa that have undergone significant range contraction and extirpation. The macropodid genus *Thylogale* (pademelon), provided the closest BLAST matches for many sequences across multiple samples, but to date no member of the genus has been recorded in this region. It was not possible to provide accurate taxonomic identifications for most of the Muridae sequences and for many *Macropus* sequences. While many sequences could be assigned with high confidence to a genus level, others could not be assigned beyond family or genus. A number of birds and reptiles were also identified and these have been collated at the family and genus level (Figure 4.2.2). While assignment to the species level is certainly possible in many instances a conservative approach is adopted here to showcase the approach.

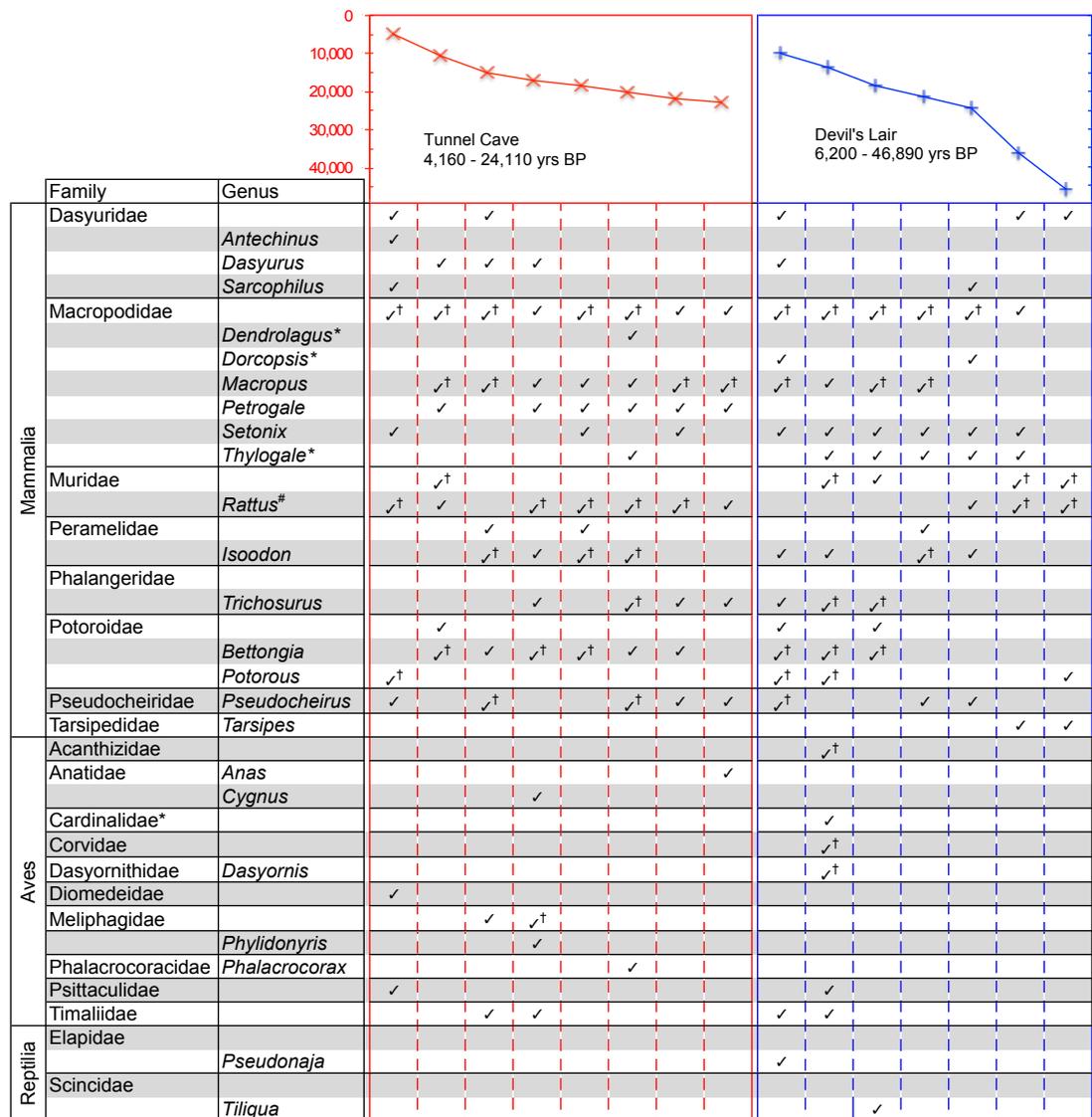


Figure 4.2.2 Taxa identified in bulk-bone powder samples. Mammals, birds and reptiles identified in each sample are listed. Samples are grouped according to site from youngest to oldest in years BP (uncalibrated), which is plotted on the same scale for both sites. The criteria used in taxonomic assignment are detailed in the Methods. Note that there is uncertainty surrounding taxonomy with regards to both Timaliidae and Cardinalidae (See Discussion).

Key: † Detected using multiple primer sets; * Taxa not historically known to occur in the study region; # Sequences assigned to *Rattus* aligned closest to native *Rattus fuscipes* (bush rat)

4.2.4.3 Genetic biodiversity analysis

A largely taxonomy-independent approach was adopted to examine fluctuations in observed genetic diversity over time at both sites. While the taxa identified using the GS-Junior and Ion Torrent PGM were mostly congruent, coverage dependent OTU inflation, arising from homopolymer sequencing error (see Section 4.2.3 and 4.2.5) was observed. A modified OTU analysis filter was designed to reduce the influence of HTS homopolymer sequencing error (Loman *et al.*, 2012; Quail *et al.*, 2012), by employing distance-based metrics obtained from sequence alignments, giving rise to a new method referred to here as Distance-based Taxonomic Units (DTUs).

A total of 72 DTUs were identified across all 15 samples, 23 of which were shared across multiple samples, and in some instances both archeological sites (Figure 4.2.2). The number of DTUs fluctuates noticeably with time (Figure 4.2.3). The number of DTUs shows a notable decrease that roughly coincides with the last glacial maximum (LGM), whilst also showing an increase post-LGM. The composition of DTUs also varies over time. For instance, Potoroidae (potoroids) DTUs appear around the LGM and show an increase in numbers, whilst numbers of Macropodidae (macropodids) DTUs show a decline post-LGM.

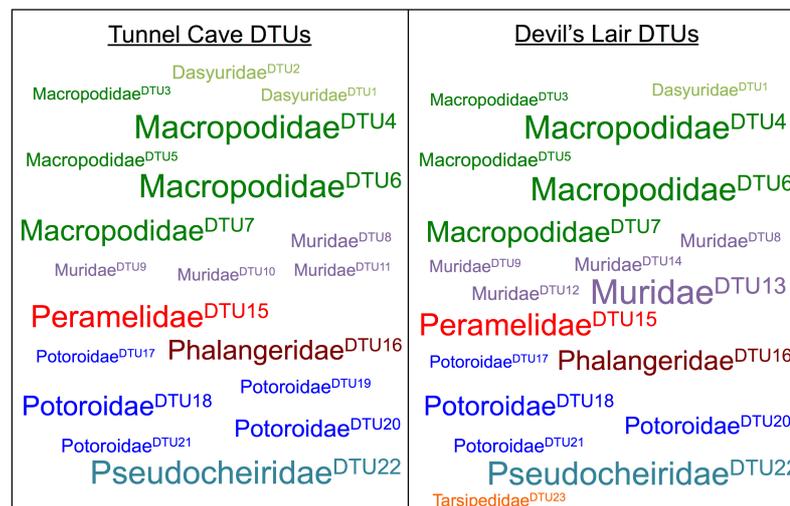


Figure 4.2.3 DTUs shared across bulk-bone powder samples. The DTUs shared between bulk-bone powder samples, and across both Tunnel Cave (left) and Devil's Lair (right), are shown. DTUs have been labeled with the closest BLAST family matches. Each DTU has been assigned a numeric identifier following the acronym 'DTU', shown in superscript. Font size is indicative of the total number of samples a DTU was detected in.

With obvious variation in DTU composition, macropodid sequences were selected to examine DTU number flux at a finer scale to examine whether or not this reflected the overall trends in biodiversity change. Macropodids exhibit a declining trend in DTU diversity post-LGM (Figure 4.2.5) that marginally increases near the Holocene/Pleistocene transition 11,700 years ago.

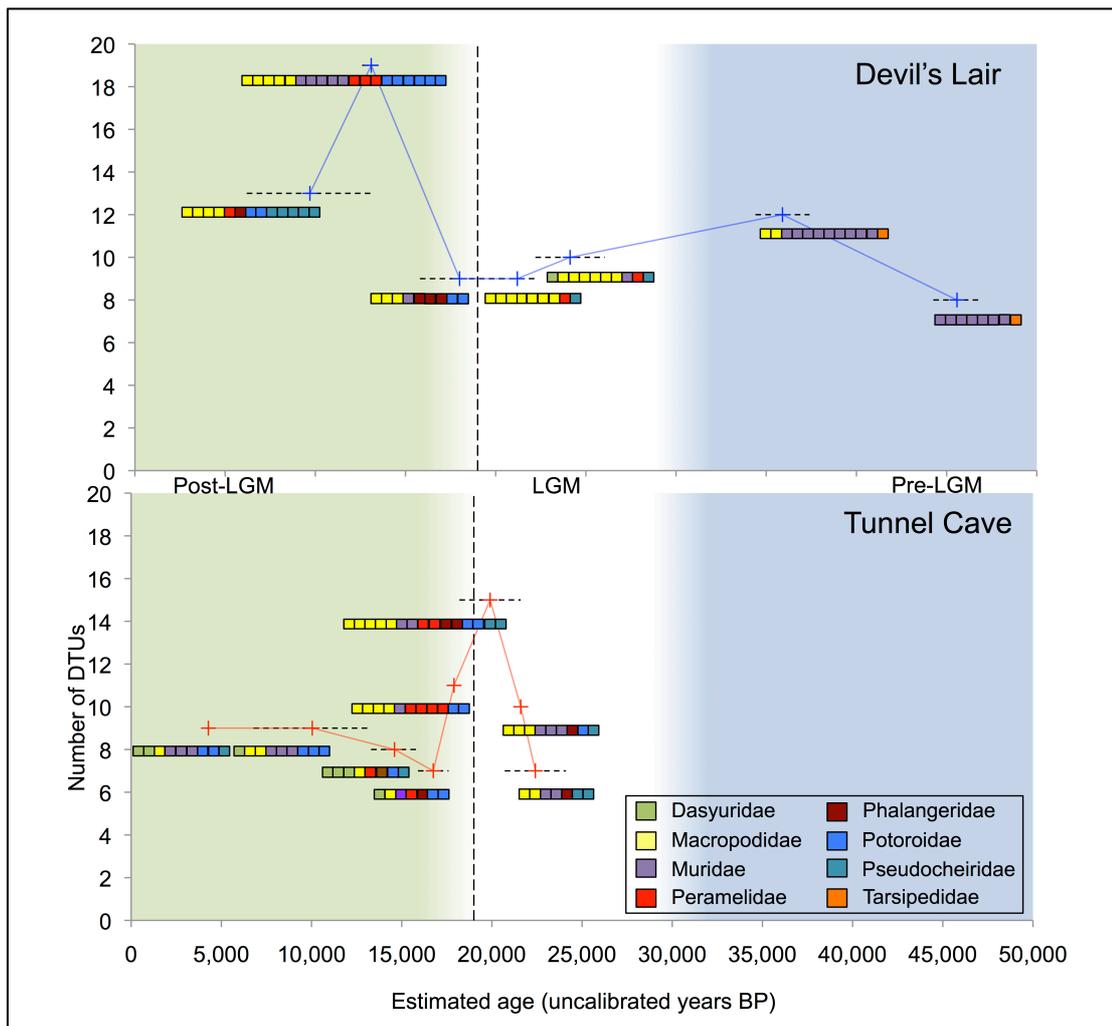


Figure 4.2.4 Change in DTU number and composition over time at Tunnel Cave and Devil's Lair. The fluctuation in DTU number and the change in DTU composition across samples and at both sites are plotted against the backdrop of the major climatic shift around the end of the Last Glacial Maximum (LGM). Dashed vertical line - approximate end of the LGM; Blue background – Pre-LGM; White background – LGM; Green background – Post-LGM. Median ages are plotted for each sample; dashed horizontal line indicates minimum and maximum accepted date range for each layer.

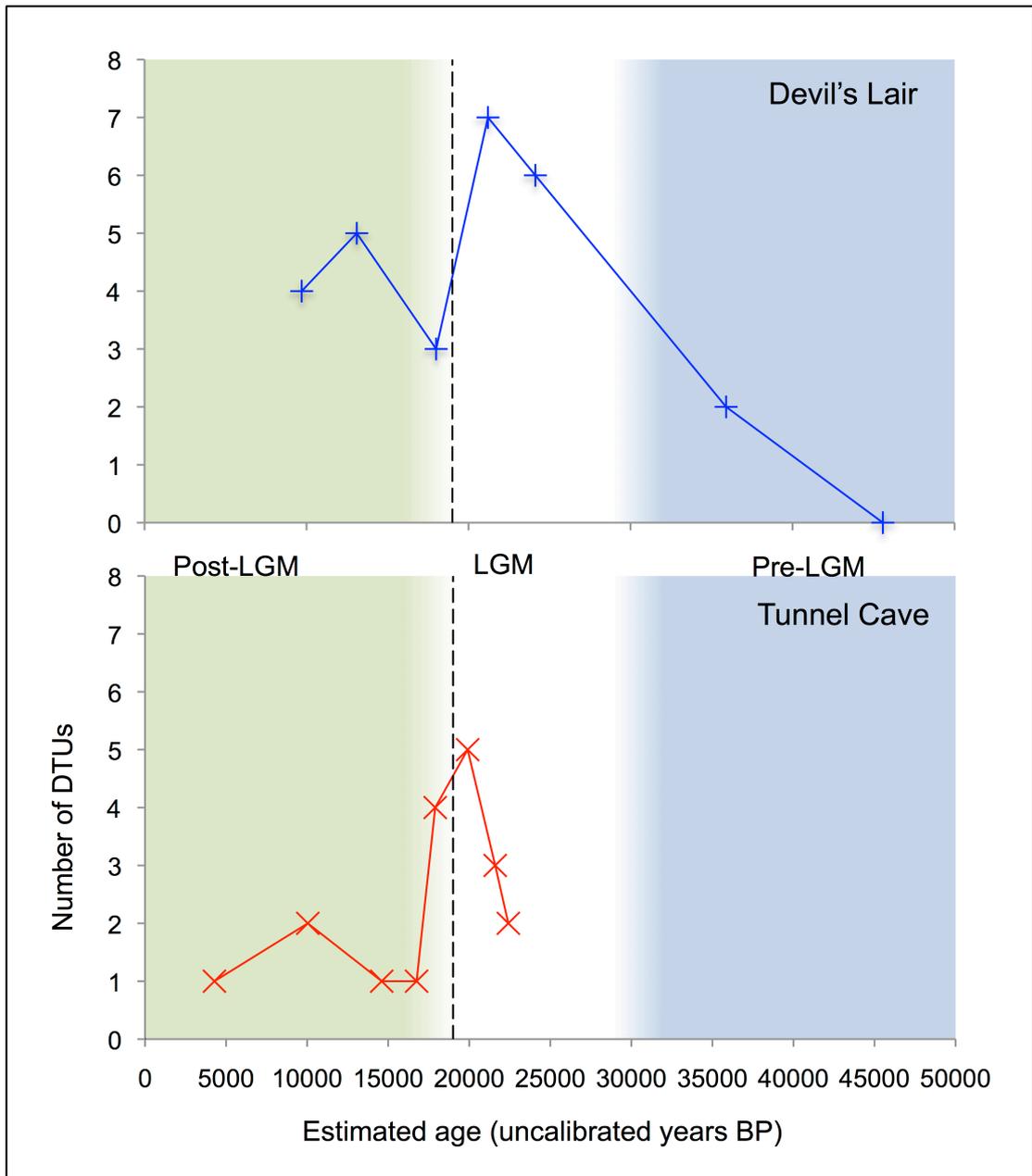


Figure 4.2.5: Change in Macropodidae DTU number over time at Tunnel Cave and Devil's Lair. The fluctuation in Macropodidae DTU number across samples and at both sites is illustrated. Dashed vertical line - approximate end of the LGM; Blue background – Pre LGM; White background – LGM; Green background – Post-LGM. Median ages are plotted for each sample.

4.2.5 Discussion

This study presents a novel HTS method using aDNA characterized from bulk-bone powder samples. It represents a powerful new approach to analyze unidentifiable fragments excavated from fossil deposits. Ancient DNA extracted from bones within a layer dated between 44,260 – 46,890 years BP (uncalibrated), is the oldest aDNA recovered from Australia to date. These HTS results and the initial exploration of this technique show promise for larger scale bulk-bone analyses of fossil deposits. Rapidly analyzing a bulk bone sample to determine if a site is conducive to DNA preservation will be valuable in excavations and test pits as DNA becomes increasingly incorporated into archeological and paleontological practices.

Even with the limited sampling, this first foray into bulk-bone analyses, has uncovered a significant amount of biological information that adds substantially to previous knowledge of the sites and surrounding biodiversity. Analyzing these data in the context of DNA damage, sequencing error, incomplete reference databases and the necessary use of short DNA sequences raises numerous challenges that must be systematically addressed (Cooper & Poinar, 2000; Coissac *et al.*, 2012; Murray *et al.*, 2012; Taylor & Harris, 2012). Nevertheless, when appropriate protocols and sequence filters are applied (see Methods) the method affords new insights into past biodiversity (Figure 4.2.2) and its temporal and spatial variation (Figures 4.2.3-4.2.5).

Raw DNA sequences obtained from HTS platforms can be sorted and screened using a combination of filters that collectively exclude low-quality reads (Q-scores), sequences with errors in known flanking regions (adaptors, primers, and barcodes), artificial chimeric sequences and low abundance reads (see Section 4.2.3). However, even sequences that pass these filters need to be interpreted with caution: the bird family Cardinalidae, which is not known to occur in Australia, is a case in point. The identification of birds also serves to illustrate the pitfalls associated with taxonomic revision. The taxonomy of the family Cardinalidae has been revised on a number of occasions, as has that of Timaliidae, which was also identified in some samples. Timaliidae has been regarded as a family consisting of Old World passerine birds, however the Australasian babblers (family: Pomatostomidae) were once within this

family and the typical white-eyes (*Zosterops*) are disputably within this family also (Jönsson & Fjeldså, 2006). The families and genera identified (Figure 4.2.1) within each of the 15 samples require further investigation to identify taxa to the species level. Nevertheless, most of the genera identified at both sites from fossil morphology were again successfully detected in the bulk bone (Dortch, 2004). The absence of some morphologically identified taxa from the genetically-determined faunal assemblage list is most likely due to sampling bias, as the present analysis derives from deposits representing less than one percent of the volume of the original excavations. Additionally, the possibility of primer binding bias contributing to the discontinuities between both aDNA and fossil assemblage datasets cannot be excluded. *In silico* analysis of variation in binding sites and the use of the multiple markers attempts to identify and minimize the impact of amplification bias. Finally, inherent differences between bones in terms of the preservation and quantum of mtDNA per unit biomass may also skew results between both methods of analysis causing artifactual over-representation of some taxa relative to others. However, taxa were also identified that were not detected in any previous morphology-based analyses, particularly small mammals, birds and reptiles, all of which require highly-specialized taxonomic skills to identify, are less likely to preserve diagnostic remains, and may be poorly represented in reference collections.

A high level of confidence surrounds the bulk of the taxonomic identifications; for instance, the majority of mammalian taxa identified are locally extant or known from the fossil record. The same generally holds true for avian and reptilian taxa identifications. The detection of sequences endemic to southwest Australia, such as a 100% match to *Tarsipes rostratus* (honey possum), further supports the *bona fide* nature of the sequences obtained. Moreover, the detection of extirpated taxa, such as *Setonix* (quokka) and *Sarcophilus* (Tasmanian devil), as far back as c.24,000 years BP (uncalibrated) illustrates the antiquity and authenticity of the sequences, as does the detection of species whose ranges have contracted and are no longer documented at the sites, e.g. *Bettongia* (bettongs). There appears to be little or no environmental contamination as evidenced by the absence of any sequences from highly abundant invasive taxa including *Mus musculus* (house mouse) or *Rattus rattus* (black rat). Whereas downward contamination may be an issue at some sites (Haile *et al.*, 2007), Devil's Lair contains several stratigraphical layers capped with calcite "flowstone"

(Turney & Bird, 2001) preventing the movement of fossils, and likely DNA (Dortch, 2004; Haile *et al.*, 2009). Whilst it is acknowledged that contamination can be cryptic and sporadic (Champlot *et al.*, 2010; Erlwein *et al.*, 2011; Tuke *et al.*, 2011), the strict adherence to aDNA protocols (Gilbert *et al.*, 2005), the use of sequence quality filters and the plausibility of the data (see Section 4.2.3), greatly reduces the likelihood that contamination contributed to the data presented here.

Although most taxonomic assignments from DNA sequences confirmed previous morphological identification (Dortch, 2004), some unexpected sequences resulted in distinct DTUs that were more difficult to assign. The issue is best exemplified by indeterminate Macropodidae sequences. It is unlikely that poor database coverage is the cause of this family-level assignment, as the Macropodidae database is nearly complete for both 16S and 12S rRNA mtDNA. In such cases sequencing error or DNA damage is also unlikely as the sequences are abundant and present across numerous samples at both sites, have passed all quality filters, form distinct DTUs and are unlikely to be nuclear copies (Figures 4.2.2-4.2.4). It is possible therefore that these sequences may arise from extinct lineages of present-day macropodids or indeed from extinct taxa. In some cases sequences mapped closest to species of the New Guinea forest wallaby (*Dorcopsis*) and the east Australian restricted pademelon (*Thylogale*). The presence of such 'indeterminate' DNA sequences in bulk-bone samples is intriguing. For example, two extinct tree-kangaroo species (genus *Bohra* (Prideaux & Warburton, 2008; Prideaux & Warburton, 2009)), have been described in caves along the Nullarbor Plain, yet tree-kangaroos of the genus *Dendrolagus* are only currently present in northeastern Queensland and New Guinea and were previously not thought to have occurred so far south (Prideaux & Warburton, 2008). It is a tantalizing prospect that 'indeterminate' DNA sequences could represent previously unknown species from southwest Western Australia, but it is also a problematic finding, as there is no easy way to uncover the fossils that contributed the DNA. It is likely that bulk-sampling methods such as this will generate genetically plausible taxa that lack morphological identifications. Arguably a similar result has already occurred with the single Denisovan finger bone from "X-woman" used to postulate a new lineage of archaic humans in Siberia (Krause *et al.*, 2010; Meyer *et al.*, 2012).

When dealing with past biodiversity and aDNA sequences from fossil assemblages, analyses that are largely independent of taxonomy will likely be crucial to mapping temporal and/or spatial variation in genetic signatures. Such an approach facilitates the use of sequences that would otherwise be labeled “indeterminate”, which will be commonly encountered when employing the bulk-bone HTS methodologies advocated here. While it is not possible to comprehensively analyze changes in biodiversity over time presented here from only a handful of samples such an analysis serves to illustrate how bulk-bone data could be approached. The data presented in Figures 4.2.2-4.2.5 should therefore be viewed tentatively, as further extensive replication and investigation is required to confirm any significant patterning over time.

Owing to the difficulties of definitively assigning sequences to a defined taxonomy, a modified OTU analysis (referred to as DTU), has been introduced to examine biodiversity change over time. It was clear from the initial analysis that OTU numbers were artificially inflated primarily by homopolymer error. When dealing with short sequences homopolymer errors can create a distinct OTU whereby the only difference between it and its closest OTU match is a base within a homopolymer stretch. It was observed that homopolymer-derived OTUs were more common in those samples with greater depth of sequencing coverage. To overcome this issue, an OTU alignment and Kimura 2-parameter distance matrix was adopted whereby errors in homopolymer stretches appear as gaps and homopolymer-derived OTUs collapse into a single DTU (See Section 4.2.3). Whilst at these particular sites, it is a challenge to disentangle the roles of climate, DNA decay and past anthropogenic influences; shifts in DTU composition appear at the LGM and at the Holocene-Pleistocene transition (Figures 4.2.4 and 4.2.5). Furthermore, specific Macropodidae DTU analysis showed a reduction in DTU diversity and abundance over time, with a drop in diversity around the LGM (Figure 4.2.5). With these tentative patterns of biodiversity being derived from only 14 DNA extractions it is easy to conceptualize how, with adequate sampling and appropriate genetic markers, a bulk-bone sampling method will facilitate detailed mapping of faunal changes over time. Moreover, the method is cheaper than single bone approaches (Shapiro *et al.*, 2004; Lorenzen *et al.*, 2011) while augmenting traditional morphological analysis.

The bulk-bone aDNA metabarcoding method used in this study presents a new, cost effective approach to identifying bulk quantities of morphologically indistinct bone fragments that otherwise end up in the taxonomic scrapheap. From modest amounts of sieved material across multiple layers at two study sites it was possible to detect equivalent diversity as described in previous morphological analyses (Dortch, 2004). While some taxa previously identified were not detected (most noticeably *Macropus* species), the converse was also true. This method is by no means an attempt to supplant traditional morphological approaches to taxonomic identification and analysis. Rather, it complements these approaches and by means of DTU analysis indicates changes in genetic diversity through time. Besides improving the identification of fossil assemblages the method allows researchers to rapidly assess the DNA preservation potential of freshly excavated material, which will vary from site to site. The approach will be equally applicable to archaeological and paleontological sites, providing snapshots of past faunal diversity and human subsistence in both taxonomic dependent and independent ways. As such, it is anticipated that a bulk-bone approach will become a valuable part of the archaeological and paleontological toolkit.

4.2.6 Acknowledgements

We thank the Wardandi people, and the Webb family in particular, as traditional owners and custodians of the region, for supporting our excavations and analyses. We thank student volunteers for their assistance in the field, Nina Kresoje (UWA) and Vanessa Atkinson (Pathwest Laboratory Medicine WA) and Frances Brigg for sequencing assistance. We also thank the WA Museum and the Curator of Anthropology Moya Smith for access to museum resources, and iVEC for computational support. Australian Research Council grants DP120103725 (to MB, JD and JH) and FT0991741 (to MB) funded the research.

4.2.7 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

- Archibald, S. B., Greenwood, D. R., & Mathewes, R. W. (2013). Seasonality, montane beta diversity, and Eocene insects: testing Janzen's dispersal hypothesis in an equable world. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *371*, 1-8.
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research*, *34*, D16-D20.
- Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, *2*, e197.
- Bonnichsen, R., Hodges, L., Ream, W., Field, K. G., Kirner, D. L., Selsor, K., & Taylor, R. E. (2001). Methods of the study of ancient hair: Radiocarbon dates and gene sequences from individual hairs. *Journal of Archaeological Science*, *28*, 775-785.
- Bunce, M., Oskam, C., & Allentoft, M. (2011). The use of quantitative real-time PCR in ancient DNA research. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA: Methods and Protocols* (pp. 121-132): Humana Press.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335-336.
- Champlot, S., Berthelot, C., Pruvost, M. I., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One*, *5*, e13042.
- Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*, 1834-1847.
- Coloneasa, A. C., Zanchettab, G., Fallicke, A., Manganellif, G., Sañag, M., Alcadeh, G., & Neboti, J. (2013). Holocene snail shell isotopic record of millennial-scale hydrological conditions in western Mediterranean: Data from Bauma del Serrat del Pont (NE Iberian Peninsula). *Quaternary International*, *303*, 43-53.

- Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, *409*, 704-707.
- Cooper, A., & Poinar, H. N. (2000). Ancient DNA: do it right or not at all. *Science*, *289*, 1139-1139.
- Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.
- Dortch, J. (2004). *Palaeo-environmental Change and the Persistence of Human Occupation in South-western Australian Forests*. Oxford: Archaeopress.
- Dortch, J., & Wright, R. (2010). Identifying palaeo-environments and changes in Aboriginal subsistence from dual-patterned faunal assemblages, south-western Australia. *Journal of Archaeological Science*, *37*, 1053-1064.
- Edgar, R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*, 2460-2461.
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*, 2194-2200.
- Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S., & McClure, M. O. (2011). DNA extraction columns contaminated with murine sequences. *PLoS ONE*, *6*, e23484.
- Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M., & Barnes, I. (2005). Assessing ancient DNA studies. *Trends in Ecology and Evolution*, *20*, 541-544.
- Grealy, A. C., McDowell, M. C., Scofield, P., Murray, D. i. C., Fusco, D. A., Haile, J., . . . Bunce, M. (2015). A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quaternary Science Reviews*, *128*, 37-47.

Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., . . . Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 22352-22357.

Haile, J., Holdaway, R., Oliver, K., Bunce, M., Gilbert, M. T. P., Nielsen, R., . . . Willerslev, E. (2007). Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? *Molecular Biology and Evolution*, *24*, 982-989.

Hamady, M., Lozupone, C., & Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal*, *4*, 17-27.

Haouchar, D., Haile, J., McDowell, M. C., Murray, D. C., White, N. E., Allcock, R. J. N., . . . Bunce, M. (2013). Thorough assessment of DNA preservation from fossil bone and sediments excavated from a late Pleistocene-Holocene cave deposit on Kangaroo Island, South Australia. *Quaternary Science Reviews*, *84*, 56-64.

Hunter, A. A., Macgregor, A. B., Szabo, T. O., Wellington, C. A., & Bellgard, M. I. (2012). Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine*, *7*, 1.

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*, 377-386.

Jönsson, K. A., & Fjeldså, J. (2006). A phylogenetic supertree of oscine passerine birds (Aves: Passeri). *Zoologica Scripta*, *35*, 149-186.

Jørgensen, T., Haile, J., Möller, P. E. R., Andreev, A., Boessenkool, S., Rasmussen, M., . . . Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, *21*, 1989-2003.

Jørgensen, T., Kjær, K. H., Haile, J., Rasmussen, M., Boessenkool, S., Andersen, K., . . . Willerslev, E. (2011). Islands in the ice: detecting past vegetation on Greenlandic

nunataks using historical records and sedimentary ancient DNA meta-barcoding. *Molecular Ecology*, 21, 1980-1988.

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30, 3059-3066.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111-120.

Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. V., Derevianko, A. P., & Pääbo, S. (2010). The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*, 464, 894-897.

Little, D. P. (2011). DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE*, 6, e20552.

Loman, N. J., Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30, 434-439.

Lorenzen, E. D., Nogues-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., . . . Willerslev, E. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, 479, 359-364.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., . . . Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338, 222-226.

Murray, D. C., Pearson, S. G., Fullagar, R., Chase, B. M., Houston, J., Atchison, J., . . . Bunce, M. (2012). High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews*, 58, 135-145.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., de Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853-858.

- Oskam, C. L., Haile, J., McLay, E., Rigby, P., Allentoft, M. E., Olsen, M. E., . . . Bunce, M. (2010). Fossil avian eggshell preserves ancient DNA. *Proceedings of the Royal Society Biological Sciences Series B*, 277, 1991-2000.
- Prideaux, G. J., & Warburton, N. (2009). *Bohra nullarbora* sp. nov., a second tree-kangaroo (Marsupialia:Macropodidae) from the Pleistocene of the Nullarbor Plain, Western Australia. *Records of the Western Australian Museum*, 25, 165-179.
- Prideaux, G. J., & Warburton, N. M. (2008). A new Pleistocene tree-kangaroo (Diprotodontia: Macropodidae) from the Nullarbor Plain of south-central Australia. *Journal of Vertebrate Paleontology*, 28, 463-478.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13, 341.
- Raoult, D., Aboudharam, G., Crubezy, E., Larrouy, G., Ludes, B., & Drancourt, M. (2000). Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of Medieval Black Death. *Proceedings of the National Academy of Sciences, USA*, 97, 12800-12803.
- Rohland, N., Reich, D., Mallick, S., Meyer, M., Green, R. E., Georgiadis, N. J., . . . Hofreiter, M. (2010). Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biology*, 8, e1000564.
- Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, 71, 1501-1506.
- Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P., Sher, A. V., . . . Cooper, A. (2004). Rise and fall of the Beringian steppe bison. *Science*, 306, 1561-1565.

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*, 1794-1805.

Smith, C. I., Chamberlain, A. T., Riley, M. S., Cooper, A., Stringer, C. B., & Collins, M. J. (2001). Neanderthal DNA. Not just old but old and cold? *Nature*, *410*, 771-772.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045-2050.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*, 2731-2739.

Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, *12*, 377-388.

Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*, 283-285.

Tuke, P. W., Tettmar, K. I., Tamuri, A., Stoye, J. P., & Tedder, R. S. (2011). PCR master mixes harbour murine DNA sequences. Caveat emptor! *PLoS ONE*, *6*, e19953.

Turney, C., & Bird, M. I. (2001). Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. *Quaternary Research*, *55*, 3-13.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., . . . Wolinsky, S. M. (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, *455*, 661-664.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

4.3 Synopsis

The proof-of-concept method developed in this chapter not only proved to be efficient and cost-effective but it was also shown that fragmentary bone can offer valuable insights into the faunal turnover at archaeological sites. The BBM strategy employed was able to identify much of the fauna previously identified at the sites while adding a new dimension through the identification of difficult to characterise taxa such as reptiles and murids, thus addressing a clear bias in the archaeological record at the sites.

The BBM methodology developed in this paper has since proved to be a useful adjunct to the morphological identification of fossil assemblages at a number of palaeontological sites and shown to be a ‘consistent, accurate and sensitive’ technique (Haouchar *et al.*, 2013; Grealy *et al.*, 2015, co-authored publications included in Appendix III).

Throughout Chapters Two–Four a number of methods, workflows and molecular ‘safeguards’ have been noted that should be considered when embarking on HTS projects, especially when using poor-quality substrates and working in areas of high diversity with poor reference databases. Chapter Five serves to underscore the considerations noted in the previous chapters while raising additional points that may be of value when embarking upon amplicon sequencing (i.e. metabarcoding) projects where PCR and HTS workflows are employed.

Chapter Five – The pitfalls of HTS and potential suggestions for how to address them

5.1 Preface

Chapter Five reviews the challenges associated with HTS methodologies and introduces some key considerations when embarking on HTS metabarcoding studies. This study resulted in the published manuscript 'From benchtop to desktop: important considerations when designing amplicon sequencing workflows' (PLoS One 2015, 10, e0124671). With the exception of formatting and in-thesis referencing this manuscript has been reproduced as published.

High-throughput sequencing as mentioned previously (Chapter 1) has truly revolutionised the field of molecular genetics; from aDNA and molecular ecology to bacterial metagenomics and medical genomics. While the technology has evolved at a rapid pace over the past decade many issues continue to remain (Chapter 1).

Due to the rapid development of HTS sequencing technology and the onslaught of numerous iterations of HTS platforms this thesis represents in itself a realisation of not just the potential of HTS but also the pitfalls associated with it. Chapter Five seeks to synthesise many of the ideas and considerations encountered throughout the work involved in Chapters Two–Four. These largely revolve around issues related to sample screening, the targeted gene region and library generation. As technology and analyses have improved a number of issues have come to light that were not fully realised at the beginning of this thesis and as such this chapter seeks to address this and offer suggestions for future projects.

5.1.1 Statement of Contribution

Conceived and designed the experiments: DCM, MB. Performed the experiments: DCM, MLC. Analysed the data: DCM. Contributed reagents/materials/analysis tools: MB. Wrote the paper: DCM, MB. Edited the manuscript: DCM, MB, MLC.

5.2 From benchtop to desktop: important considerations when designing amplicon sequencing workflows

Dáithí C. Murray¹, Megan L. Coghlan¹ and Michael Bunce¹

¹ *Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia.*

5.2.1 Abstract

Amplicon sequencing has been the method of choice in many high-throughput DNA sequencing (HTS) applications. To date there has been a heavy focus on the means by which to analyse the burgeoning amount of data afforded by HTS. In contrast, there has been a distinct lack of attention paid to considerations surrounding the importance of sample preparation and the fidelity of library generation. No amount of high-end bioinformatics can compensate for poorly prepared samples and it is therefore imperative that careful attention is given to sample preparation and library generation within workflows, especially those involving multiple PCR steps. This paper redresses this imbalance by focusing on aspects pertaining to the benchtop within typical amplicon workflows: sample screening, the target region, and library generation. Empirical data are provided to illustrate the scope of the problem. Lastly, the impact of various data analysis parameters is also investigated in the context of how the data was initially generated. It is hoped this paper may serve to highlight the importance of pre-analysis workflows in achieving meaningful, future-proof data that can be analysed appropriately. As amplicon sequencing gains traction in a variety of diagnostic applications from forensics to environmental DNA (eDNA) it is paramount workflows and analytics are both fit for purpose.

5.2.2 Introduction

The myriad of names and acronyms associated with high-throughput DNA sequencing (HTS) is undeniably impressive and the number of applications for which the technology itself has proven useful equally matches this. To date, amplicon sequencing (Thomas *et al.*, 2006), whereby PCR products are generated, converted to libraries, pooled and then sequenced, has been the method of choice in many HTS studies. Amplicon sequencing has been used in, or proposed for, a wide range of contexts that include, amongst others, biomonitoring (Ficetola *et al.*, 2008; Andersen *et al.*, 2012; Baird & Hajibabaei, 2012; Shokralla *et al.*, 2012; Taberlet *et al.*, 2012; Thomsen *et al.*, 2012), diet analysis (Deagle *et al.*, 2010; Bohmann *et al.*, 2011; Razgour *et al.*, 2011; Pompanon *et al.*, 2012; Quéméré *et al.*, 2013; Burgar *et al.*, 2014) and bacterial metagenomics (Fierer *et al.*, 2010; Consortium, 2012; Meadow *et al.*, 2013; Ding & Schloss, 2014; Liu *et al.*, 2014; Mason *et al.*, 2014; Sun *et al.*, 2014). As a result of the ease with which the technology can be applied across an array of disciplines, it can at times prove to be a minefield for scientists seeking to avail of it. This is especially true for those with limited experience in either wet-lab molecular biology skills or computational bioinformatics. The latter of these areas has received much attention; the importance of the former is often under-appreciated.

Currently, most primary literature, reviews and opinion articles surrounding HTS tend to focus on the applications of the technology (Kircher & Kelso, 2010; Ekblom & Galindo, 2011; Baird & Hajibabaei, 2012; Pompanon *et al.*, 2012; Shokralla *et al.*, 2012; Taberlet *et al.*, 2012; Murray *et al.*, 2013; Clarke *et al.*, 2014), platform evaluations (Loman *et al.*, 2012; Quail *et al.*, 2012b) and bioinformatic approaches to data analysis (Binladen *et al.*, 2007; Huson *et al.*, 2007; Caporaso *et al.*, 2010; Hamady *et al.*, 2010; Quince *et al.*, 2011; Faircloth & Glenn, 2012; Gonzalez & Knight, 2012). While all three are extremely important in the generation of high fidelity data, a heavy focus on these aspects fails to address the need to pay close attention to the implementation of protocols and procedures at the bench. The data one has to work with is, and will only ever be, as good as the quality of experimental procedures implemented and no amount of high-end bioinformatics can compensate for poorly prepared samples, artefacts or contamination. It is therefore imperative

that careful consideration is given to the ways in which samples are screened for sequencing, in addition to the method used to generate the amplicon sequencing library. These aspects are independent of the equally important need to carefully choose extraction methods that are optimised for the chosen substrates. While DNA isolation methods are a key consideration, this is dealt with extensively elsewhere. Instead, this paper focuses on how best to approach amplicon workflows following DNA extraction to generate robust and representative datasets for a given DNA isolation.

Through a series of simple experiments (Table 5.2.1), various aspects that should be considered when preparing to embark on the use of amplicon sequencing are highlighted, some aspects of which are equally as applicable to shotgun sequencing. These experiments focus primarily on three areas of experimental design or benchwork within the typical amplicon sequencing workflow: sample screening, the target region, and library generation. Finally, although not a focus of the paper, certain pertinent considerations in relation to data analysis that are seldom acknowledged in other literature will also be addressed. It is hoped that the following may address the distinct lack of literature in relation to sample preparation and library generation. It is advocated that closer attention is required at the bench when conducting amplicon sequencing. Ultimately, it may be appropriate to define a set of flexible guidelines, such as the MIQE guidelines used for qPCR data (Bustin *et al.*, 2009), for the reporting of amplicon data generation and analysis.

Table 5.2.1 Details for the experiments conducted. The purpose of each numbered experiment is shown in addition to the title used for each one in the methods and results section. The appropriate methods sections, results sections and figures to consult for each experiment are also given.

Experiment	Purpose	Methods	Results
Experiment 1: Importance of sample screening	Illustrate the importance of quantifying samples using a dilution series to select an appropriate working dilution free of inhibition containing a sufficient quantity of input template DNA	Main: 2.2.1 (see also: Section 2.1.1, S1A Fig & S1 Table)	Section 3.1. Fig 2
Experiment 2: Assessing the amplicon target region	Explore the potential benefits to the downstream processing of high-throughput sequencing data arising from the inclusion of amplicon-specific single-source samples embedded into sequencing runs	Main: 2.2.2 (see also: S1B Fig & S1 Table)	Section 3.2. Fig 3
Experiment 3: Importance of experimental controls	Demonstrate the importance of control reactions in bacterial metagenomics and other fields using samples with a high propensity for environmental contamination	Main: 2.2.3 (see also: S1C Fig & S1 Table)	Section 3.3. S2 Table
Experiment 4: Library generation efficiency	Assess the efficiency drop-off associated with the use of fusion tagged primers of different 'architecture' when compared to standard non-fusion tagged template specific primers	Main: 2.2.4 (see also: Section 2.1.1, S1D and S1E Fig & S1 Table)	Section 3.4. S3 Table
Experiment 5: Analysis parameters and their impact	Highlight the difficulties in choosing appropriate quality and abundance filtering parameters when analysing complex, heterogeneous samples; the composition of which are unknown.	Main: 2.2.5 (see also: Fig 1, S1F Fig & S1 Table)	Section 3.5. Fig 4, S4 Table

*Please note that due to thesis preparation requirements, the links indicated in this table have not been maintained. If using the table please refer to the original publication included in Appendix III to which the section referencing applies.

5.2.3 Materials and Methods

Some of the following methodologies were specifically designed for this study; others have utilised samples and/or data drawn from previous studies (Murray *et al.*, 2011; Coghlan *et al.*, 2012; Murray *et al.*, 2013; Tridico *et al.*, 2014; White *et al.*, 2014). The materials and methods below provide an overview of the methodologies and the reader is referred to the original publications and also the supplementary online information where schematics of all experiments conducted are presented (Figure S5.2.1A-F Fig.). Each of four important steps in amplicon workflows: sample screening (Figure S5.2.1A), the target region (Figure S5.2.1B), library generation (Figure S5.2.1C-E) and data analysis (Figure S5.2.1F), is addressed separately in the materials and methods that follow (Section 5.2.3.2). General methods employed during sample screening, amplicon generation, DNA sequencing and data analysis that were common to all areas are detailed first (Section 5.2.3.1) before more focused information on each of the four aforementioned steps (Section 5.2.3.2). Any further detailed information on the samples or experimental workflows used is available in previous publications (Murray *et al.*, 2011; Coghlan *et al.*, 2012; Murray *et al.*, 2013; Tridico *et al.*, 2014; White *et al.*, 2014) or from the authors upon request. Where applicable amplicon sequence reads have been uploaded to Data Dryad (doi:10.5061/dryad.2qf0t).

5.2.3.1 General methods

DNA extraction and screening

A variety of samples and extraction methods are used throughout these experiments. Extraction protocols followed can be found in the original publications where indicated (Murray *et al.*, 2011; Coghlan *et al.*, 2012; Murray *et al.*, 2013; Tridico *et al.*, 2014; White *et al.*, 2014), but typically involved silica-based purification methods to isolate DNA. Where sample extraction has not been reported previously, the details of the extraction procedure are found below in Section 5.2.3.2.

All samples used were screened to determine the appropriate working dilution containing sufficient DNA free of inhibition using quantitative PCR (qPCR) on a SYBR-based STEP-ONE Applied Biosystems Real-Time PCR instrument (Bunce *et al.*, 2011; Murray *et al.*, 2011). Samples were assessed based on Cycle Threshold

(C_T) values, curve form and melt-curves. Extraction controls were conducted for each batch of extractions and screened using qPCR to test for contamination arising from laboratory practice, reagents, or the environment. If positive for the presence of DNA, extraction controls were included in tagged qPCR assays. All qPCR reaction conditions and reagent components can be found in previous publications where indicated below, and primer details can be found in Table S5.2.1. Details are provided below for any qPCR reactions not previously reported.

Amplicon generation and sequencing

For samples deemed to have sufficient DNA copy number and determined to be free of inhibition, amplicon sequences were always generated in triplicate via qPCR using a unique combination of forward and reverse Multiplex Identifier (MID-) tagged (i.e. indexed) primers (Binladen *et al.*, 2007; Roche, 2009) (for the only exceptions to this see Section 5.2.3.2 and Figure S5.2.1A). For each tagged qPCR assay, negative reaction controls were included and, if found to contain amplifiable DNA, were incorporated into the appropriate sequencing library. Resultant amplicon products were purified following the Agencourt AMPure XP PCR Purification Kit protocol (Beckman Coulter Genomics, NSW, Aus.) and were eluted in 40 µL of Ultrapure H₂O. Purified amplicon products for each sequencing library for each platform were electrophoresed on ethidium bromide stained 2% agarose gel and pooled in equimolar ratios based on band intensity to form sequencing libraries.

In order to determine an appropriate volume of library for sequencing, each amplicon library was serially diluted and quantified using qPCR against a serial dilution of a custom synthetic oligonucleotide of known molarity. Reaction components and conditions were the same for each sequencing platform with the exception of platform specific primers appropriate to the sequencing adaptors. Each 25 µL reaction contained 2X ABI Power SYBR master mix (Applied Biosystems, CA, USA), 0.4 µM each of platform specific forward and reverse primer (IDT), and 2 µL of pooled library. Each reaction underwent the following cycling conditions: 95 °C for 5 mins; 40 cycles of 95 °C for 15 s, 56 °C for 1 min followed by a 1 °C melt curve. All sequencing was conducted according to manufacturer's protocols using one of three sequencing platforms: GS Junior (Roche), Ion Torrent PGM (Life Technologies) and MiSeq (Illumina). Sequencing on Roche was conducted using

LibA chemistry. Ion Torrent PGM emulsion PCR (emPCR) was conducted on a OneTouch2 using 400bp chemistry and sequencing was performed on 314 chips. Finally, Illumina MiSeq sequencing used V2 300 cycle chemistry on nano flow cells. To enable direct comparisons both PGM and MiSeq used single direction sequencing only, despite the fact that paired-end sequencing is available in the latter.

Data analysis

Regardless of the platform, amplicon sequence reads were deconvoluted in Geneious v7.1.3 (this version of Geneious is used throughout this paper) (Drummond *et al.*, 2011) based on unique primer indexes. As a first step in deconvolution any sequences found to contain ambiguous base calls (e.g. N) were discarded. Identification tags and primer sequences were trimmed from all reads in Geneious, allowing for no mismatch in either length or base composition as a means of quality filtering, using the inbuilt “Separate Reads by Barcode” and “Trim Ends” functions respectively. The only exception to this can be found in Section 5.2.3.2 where in some instances two base mismatches in the primer sequences were allowed (see also Figure 5.2.1 and Section 5.2.3.2). Unless otherwise stated in Section 5.2.3.2, Quality Score (Q-Score) filtering was not performed. Sequences were subsequently dereplicated at 100 % identity across their full length using USEARCH v7 (this version of USEARCH is used throughout this paper) (Edgar, 2010; Edgar, 2013), and low abundant sequence clusters, defined as those below 1 % of the total number of unique sequences, were removed using USEARCH also. Dereplicated sequences were clustered at a 97 % threshold using the UPARSE (Edgar, 2013) algorithm implemented in USEARCH. Chimeric sequences were also identified and removed using USEARCH (Edgar, 2010; Edgar *et al.*, 2011). At all stages of dereplication and OTU clustering abundance information was retained and used when calculating taxa/sequence abundance or error rates. Where appropriate, sequences were queried against the NCBI GenBank nucleotide database (Benson *et al.*, 2006) using BLASTn (Altschul *et al.*, 1990) in YABI (Hunter *et al.*, 2012), enabling taxonomic identification. Sequences were searched without a low complexity filter, with a gap penalties existence of five and extension of two, expected alignment value less than $1e-10$ and a word count of seven. The BLASTn results obtained were imported into MEtaGenome ANalyzer v4 (MEGAN) (Huson *et al.*, 2007), where they were mapped and visualised against the NCBI taxonomic framework (min. bit score =

35.0, top percentage = 5%, min. support = 1). In cases where taxonomic identification was necessary, a genus or family level assignment of a query sequence was required to have a BLASTn percentage similarity to a reference sequence of 97 % or 95 % respectively. Instances where data analysis deviated from the above steps are detailed where necessary below.

5.2.3.2 Specific methodologies

Experiment 1: Importance of sample screening

To evaluate the importance of screening samples for inhibition and low target template amount, an environmental faecal sample was obtained from a *Eudyptula minor* (Little Penguin) individual. DNA was extracted from the faecal sample, serially diluted, and screened via qPCR as described in Murray *et al.* (2011) using 16S1F/16S2R degenerate fish primers (Deagle *et al.*, 2007) (see also Figure S5.2.1A and Table S5.2.1). An appropriate working dilution of the sample deemed to have sufficient DNA copy number and free of inhibition (see Section 5.2.3.1) was used for sequencing on both the Ion Torrent PGM and GS Junior. In addition to this, both an aliquot of the working dilution spiked with an extremely inhibited soil DNA extract, to mimic inhibition, and a dilution classed as “Low Template” were selected for sequencing. For each sample, the detection and percentage abundance of two baitfish genera, *Sardinops* (specifically *S. sagax* – Australian pilchard) and *Engraulis* (specifically *E. australis* – Australian anchovy) were examined. The former being in the highest abundance: the latter in lowest abundance, as determined by a taxon-specific qPCR assay (see Table S5.2.1 and Murray *et al.*, 2011)).

The handling of the penguin, and the collection and use of the faecal sample was conducted by experienced handlers under a strict set of animal ethics guidelines approved by the Murdoch University Animal Ethics Committee (permit no. W2002/06) as part of a long-term study into *Eudyptula minor* (Little Penguin) diet. Faecal sampling and DNA extraction were performed as part of a previously published study (Murray *et al.*, 2011) and not as a part of this study, however ethics approval covers the use of the faecal sample DNA extract in this study.

Experiment 2: Assessing the amplicon target region.

Five single-source bird tissue samples were used to assess error profiles associated

with a specific amplicon target region (see Figure S5.2.1B). *Calyptorhynchus latirostris* (Carnaby's Black Cockatoo) and *C. lathami* (Glossy Black Cockatoo) samples were collected, and DNA extracted, as detailed in White *et al.*, 2014 (White *et al.*, 2014). Tissue samples of *Gallus gallus* (Chicken), *Dromaius novaehollandiae* (Emu) and *Struthio camelus* (Ostrich) were bought commercially and DNA was extracted using a Qiagen DNeasy Blood and Tissue Kit following the manufacturer's protocol. For each sample an approximately 250 bp region of the mitochondrial 12S rRNA gene was amplified and MID-tagged using 12SA/H avian primers (see Table S5.2.1 and Cooper, 1994; Cooper *et al.*, 2001) via qPCR (reaction components and conditions as detailed in Murray *et al.*, 2013, and then sequenced on both Ion Torrent PGM and Illumina MiSeq platforms.

Amplicon sequence reads for each bird were randomly sub-sampled a total of 25 times to a depth of 1,000 sequences using seqtk (available from <https://github.com/lh3/seqtk>) following deconvolution into sample batches (see 5.2.3.1). Each sub-sample was dereplicated at 100 % identity to determine the most abundant sequence, with the abundance of each unique sequence appended to sequence names for use in calculating error rates. The most abundant sequence was taken as the reference sequence. For both platforms the most abundant sequence was identical thus meaning it is likely 'correct.' Each set of sub-sampled sequences was individually aligned using MUSCLE with default parameters (Edgar, 2004). Alignments were imported into excel and for each sample the error associated with each base was calculated as a percentage of the total number of non-dereplicated sequences that differed from the reference sequence at that specific base. This was performed using an in-house macro; the output of which can be seen in File S5.2.1. The error associated with each sub-sample was subsequently calculated as the mean error across all bases. The overall percentage error rate for each bird species on both the Ion Torrent and MiSeq was taken as the mean error rate across all 25 sub-samples of each species.

The collection and use of DNA material from Cockatoos was approved by, and conducted under, Department of Parks and Wildlife (Western Australia) scientific purposes licences SC000357, SC000920, SC001230, Australian Bird and Bat Banding Authority 1862 and Animal Ethics Committee approvals DEC AEC

11/2005 and 32/2008 held by P. R. Mawson. Samples of Chicken, Emu and Ostrich (all non-endangered) were purchased from Franks Gourmet Meats, Perth, WA, Australia, and are exempt from a collection permit.

Experiment 3: Importance of experimental controls

To illustrate the importance of control reactions in bacterial metagenomics and other fields dealing with samples with a high likelihood of environmental contamination, bacterial 16S data from hair samples were generated and analysed as detailed in Tridico *et al.*, 2014 (see also Figure S5.2.1C and Table S5.2.1). Briefly, pubic and scalp hair were self-sampled by male and female volunteers. Hair samples were prepared and extracted as detailed in Tridico *et al.* (2014). Samples were screened using Bact_16S_F515 and Bact_16S_R806 primers (Turner *et al.*, 1999; Caporaso *et al.*, 2011) and amplicon libraries were generated, sequenced and analysed as per Tridico *et al.*, 2014.

The collection of human hairs for bacterial profiling was approved by, and conducted in accordance with, Murdoch University Human Research Ethics Committee Policies and Guidelines (Project Number 2011/139). Each volunteer was made aware of the nature of the study and gave written, informed consent. Hairs were self-collected from two somatic origins and placed in sample bags bearing no information that would allow the identification of any individual participant in the study (Tridico *et al.*, 2014).

Experiment 4: Library generation efficiency

Quantitative PCR using the plant plastid trnLg/h primer set (Taberlet *et al.*, 2007) was carried out to investigate the issues surrounding efficiency drop-off associated with the use of “full” fusion tagged primers (see Figure S5.2.1D and Table S5.2.1), i.e. those with MID tags and sequencing adapters upstream of the template specific primer (TSP) (see Figure S5.2.1E and Roche, 2009). A single-source plant extract in addition to two complex, heterogeneous Traditional Chinese Medicines (TCM) were used; a MoBio Plant DNA Isolation kit was used following the manufacturer’s protocol for the single-source plant sample DNA extraction, while sampling and extraction of TCMs are detailed in Coghlan *et al.* (2012). Each sample was amplified in triplicate using either (1) standard non-fusion TSP; (2) MID encoded TSP (3)

“full” fusion tagged TSP or (4) “full” fusion tagged TSP with standard non-fusion TSP spiked in (see Figure S5.2.1D-E). For (1-3) each qPCR reaction was carried out in a total volume of 25 μ L containing 2X ABI Power SYBR master mix (Applied Biosystems, CA, USA), 0.4 μ M each of the appropriate forward and reverse TSP (IDT) and 2 μ L DNA extract. For (4) the previous components were also used but an additional 0.04 μ M spike-in of each the forward and reverse standard non-fusion TSP (IDT) was also used. For each reaction C_T threshold was set at 0.1.

TCM samples were obtained from, and approved for use by, the Wildlife trade section of the Department of Sustainability, Environment, Water, Population and Communities (Australia) after being seized by Australian Customs and Border Protection Service at airports and seaports across Australia. The samples were seized because they contravened Australia's international wildlife trade laws as outlined under Part 13A of the Environment Protection and Biodiversity Conservation Act 1999 (EPBC Act). The samples were stored in a quarantine-approved facility within the laboratory after being catalogued. The samples were patent medicines available over the counter and were donated by Australian Customs and Border Protection Service under no ethics or quarantine requirements and were deemed suitable to be used for specific and general research purposes by the Customs service (Coghlan *et al.*, 2012).

Experiment 5: Analysis parameters and their impact

To demonstrate the variability in calculated OTU (operational taxonomic unit) diversity within a sample, a single bulk-bone sample, comprising ~50 individual bones and containing an unknown number of taxa, was extracted and screened using the 16Smam1 and 16SMam2 mammalian specific primer set (Taylor, 1996). Amplicon sequences were generated for short sections within the mammalian mitochondrial 16S gene using the 16Smam1 and 16SMam2 primer set and sequenced using the Ion Torrent PGM as described in Murray *et al.* (2013) (see also Figure S5.2.1F and Table S5.2.1). After deconvolution following the method detailed in Section 5.2.3.1 the data were analysed using various quality filtering methods (QFM), abundance filtering methods (AFM), and taxonomy-independent methods (TIM) of diversity analysis as shown in Figure 5.2.1. Quality Score filtering was conducted in Galaxy (Blankenberg *et al.*, 2001; Giardine *et al.*, 2005; Goecks *et al.*,

2010) using the FASTQ Quality Filter tool. Maximum expected error (maxee) quality filtering, set at 0.5, was conducted using the fastq_filter command in USEARCH. Summary quality statistics were calculated in excel using fastq files post quality filtering for QFM1 and QFM4, prior to any further abundance filtering. Dereplication and OTU clustering at 97 % was conducted using USEARCH also. DTU's were determined post OTU clustering as described in Murray *et al.*, 2013. Briefly, for DTU analyses, OTU's were aligned using MAFFT (Kato *et al.*, 2002) and alignments imported into MEGA v6.06 (Tamura *et al.*, 2011) where a distance matrix was created and exported. To determine OTU's that differed from each other by less than 3 % distance matrices were analysed in excel using an in-house macro, an example output of which is shown in File S5.2.2 (Murray *et al.*, 2013). The impacts of DNA preservation, DNA degradation, mode of bone accumulation and deposit setting will have negligible impact on the results of this experiment as the exact same set of amplicon sequences, from the exact same DNA extract, are used for each combination of QFM, AFM and TIM used. The dataset in this experiment is therefore static throughout and any biases introduced by any of the aforementioned factors will be consistent across all methods.

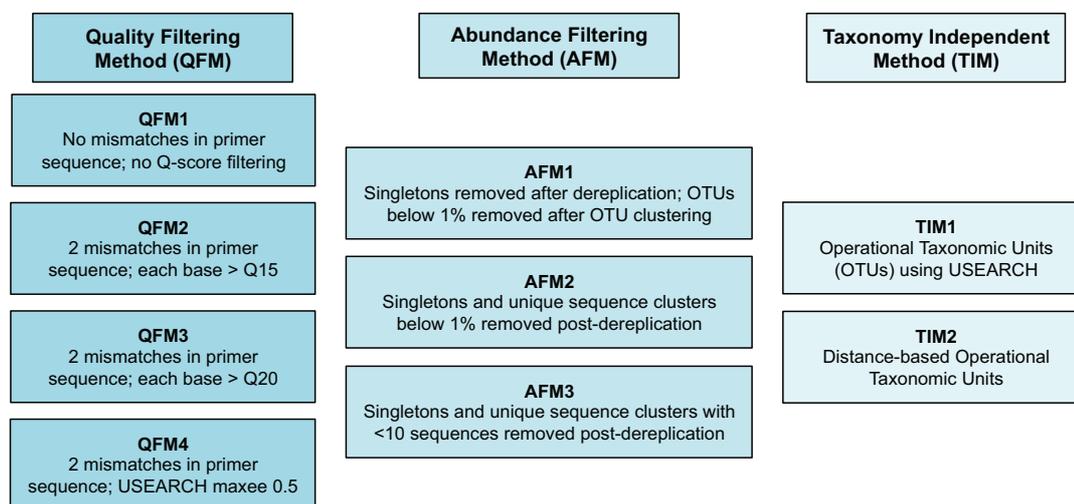


Figure 5.2.1 Definitions used in assessing the importance of analysis parameters. Shown are the definitions for quality and abundance filtering methods used in assessing their impact on both the number of operational taxonomic units (OTUs) and distance-based operational taxonomic units (DTUs) (Murray *et al.*, 2013) obtained for a given sample. maxee – Maximum Expected Error

5.2.4 Results and Discussion

Much attention has been devoted to the bioinformatic challenges associated with the analysis of amplicon sequencing data. There are a suite of programs, tools and pipelines available to assist in the deconvolution, filtering and parsing of data. As a relatively new field there is no obvious consensus on how data should, or should not, be handled bioinformatically, with the exception that sequence clusters in very low abundance should be filtered. Likewise there is no consensus on what is best-practice for data generation. Arguably the importance of data generation has taken a backseat to the computational workflows that surround bioinformatics. Bioinformaticians, rightly so, ask key questions of researchers with regard to replicates, coverage and filtering. They are less likely to ask questions about input copy number, PCR inhibition, contamination and the appropriateness of benchtop protocols. This study, through the presentation of new and existing empirical data, seeks to demonstrate the importance of both benchwork and bioinformatics. The purpose of this study is to raise awareness of potential pitfalls associated with amplicon-based workflows. The workflows dealt with in this paper do not include the process of actual DNA extraction, itself undeniably important, as this has been dealt with extensively elsewhere. The workflows presented here take as their starting point a working, amplifiable DNA extract, which can only be achieved through the careful consideration of both the scope of the project and type of substrate.

5.2.4.1 Experiment 1: Importance of sample screening

Adequate screening of samples prior to sequencing is an important task, yet fails to be routinely implemented in amplicon workflows. It is particularly prudent to assess the quality of samples when dealing with complex, heterogeneous substrates that may contain a variety of taxa or when examining samples that may contain highly degraded or low copy number DNA. There are arguably two primary factors that should be considered when evaluating samples for sequencing: the extent of inhibition, and the number of target input DNA template molecules used in generating an amplicon sequencing library. Both inhibition and low template number can have a negative impact upon the results obtained from amplicon sequencing workflows and failure to account for both can exacerbate other biases associated with amplicon sequencing. Common methods of screening samples include quantitative

PCR (qPCR) and PCR end-point assays such as gel electrophoresis or capillary electrophoresis (e.g. Agilent Bioanalyzer). The advantage of using qPCR over end-point electrophoresis lies in the fact that it is easy to determine whether or not a sample is inhibited through the analysis of the Cycle Threshold (C_T) values in a dilution series and the resultant curves. Traditional end-point assays such as electrophoresis are a blunt binary-state tool to assess inhibition and low-template samples; both will still produce bands on a gel (see gel image in Figure 5.2.2) or peaks on a Bioanalyzer trace. A case is not being made that samples should not be subjected to electrophoretic analysis, as this is a useful means for determining the presence of PCR artefacts. Rather, it would be practical to consider the additional use of qPCR or other similar methods of quantification (e.g. digital PCR), to assess the levels of inhibition and the absolute, or relative, number of target template molecules that are the input for amplicon sequencing workflows.

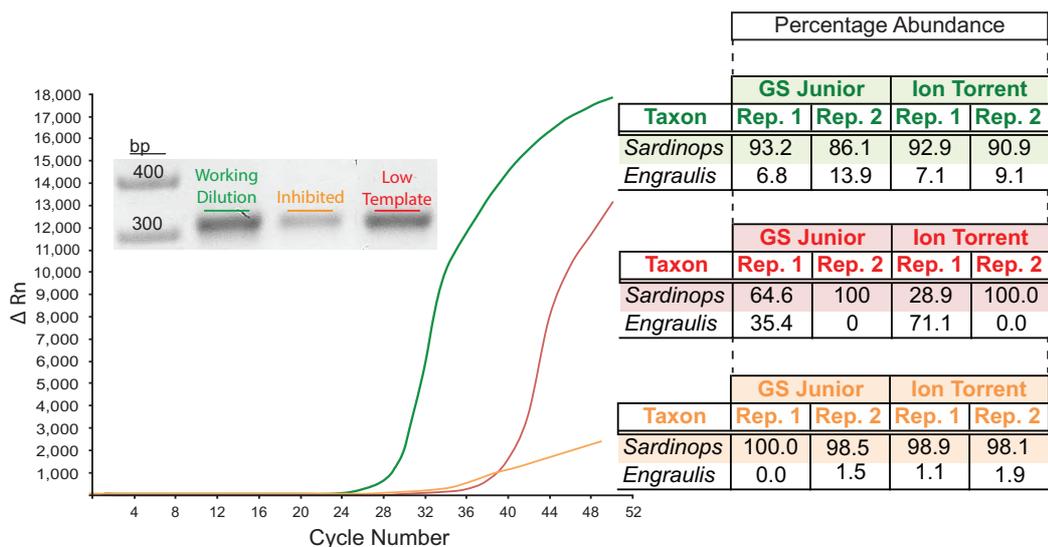


Figure 5.2.2 Quantitative PCR and sequencing results of the sample screening assay. Quantitative PCR curves indicating the presence of DNA and the degree of inhibition (**LEFT**) with agarose gel electrophoresis clearly indicating the presence of DNA post amplification via means of strong bands (**INSET ON GRAPH**). Samples were subsequently sequenced and the percentage abundance of two fish genera is indicated, where, based on taxa-specific quantitative PCR results, *Sardinops* (specifically *S. sagax* – Australian pilchard) should be in the highest abundance, with *Engraulis* (specifically *E. australis* – Australian anchovy) being in the lowest abundance. (**RIGHT**).

In a simple experiment involving the detection of two genera of fish, *Sardinops* (specifically *S. sagax* – Australian pilchard), in high abundance, and *Engraulis* (specifically *E. australis* – Australian anchovy), in low abundance, the effects of not being cognisant of inhibition or low DNA copy number are clearly demonstrated. When an appropriate working dilution exhibiting a sufficient number of input template copies and deemed free of inhibition (as determined by qPCR), was sequenced both fish species were detected in all PCR replicates, across two platforms (Figure 5.2.2, green line and shaded table). Furthermore, *Sardinops* was consistently detected as the fish species in the greater sequence abundance. In the case of the inhibited aliquot (Figure 5.2.2, orange line and shaded table) *Sardinops* was detected in all replicates and across both platforms, however *Engraulis* was not, and in those instances where it was detected it was typically at abundances <1 %. When the low-template sample dilutions (Figure 5.2.2, red line and shaded table) were sequenced a similar pattern was observed, with again *Sardinops* being detected in all replicates and across both platforms and *Engraulis* being detected in only a few (see Ficetola *et al.*, 2008) for a further example of the non-detection across multiple replicates of a target species known to be in a sample). In this instance, the abundances were vastly different between the replicates and in one instance *Engraulis* appeared to be the fish species in the highest abundance.

The inclusion of PCR and/or sequencing replicates is without doubt an important aspect of any amplicon workflow serving to improve confidence and reliability in data interpretation (Ficetola *et al.*, 2014; Robasky *et al.*, 2014) although see (Smith & Peay, 2014). Efforts have been made to determine the optimum level of PCR replicates, but it is acknowledged that the degree of replication required is dependent on the complexity of the sample in question and the objective of the study (Ficetola *et al.*, 2014). Additionally, it is also clear that simply increasing the depth of sequencing does not necessarily translate into an increased ability to detect low abundant taxa. In this study the increase in sequence depth afforded by the Ion Torrent did not improve *Engraulis* detection success. Arguably an extremely important, yet somewhat overlooked, aspect in generating an accurate species profile contained within any given sample is paying close attention to template input amount and quality, i.e. the level of amplifiable DNA and the degree of inhibition. This is becoming increasingly important as research efforts are moving towards quantitative

interpretations of sequence abundance. Simply replicating PCRs using poor quality extracts is a blunt means of increasing the fidelity of amplicon sequence data.

It is acknowledged that PCR bias can greatly skew amplicon sequencing workflows (Aird *et al.*, 2011; Schloss *et al.*, 2011; Lee *et al.*, 2012), this is especially true when little or no attention is paid to input template amount or a sample's amplifiable limits. Although only a small-scale experiment, the above serves to illustrate the importance of screening samples prior to sequencing (Figure 5.2.2). Amplicon sequencing results can clearly be obtained with low-template and inhibited samples but the reproducibility of these results is questionable: even more so if they are subsequently used in weighted analyses. Even when not interested in the relative abundance of taxa, OTUs or sequence variants, it is still nonetheless useful to screen samples for inhibition and low template amounts, as both of which can increase the possibility of false negatives. Whilst the absence of something in a sample can never truly be proven, being aware of the level of inhibition inherent within a sample or an estimate (however crude) of relative input can greatly improve the confidence surrounding presence, possible absence and/or abundance conclusions based off amplicon data. A common theme in the literature, including work by the authors, is to report the number of amplicon sequence reads obtained, but in reality a much more useful metric is to state the relative or absolute number of target templates provided to the reaction per replicate. In other words sequencing coverage is often a meaningless statistic — a PCR reaction that starts off a single molecule being the case in point. An increase in the use and reporting of quantitative data in amplicon workflows using qPCR or digital PCR can only assist in data fidelity and meaningful downstream analyses.

5.2.4.2 Experiment 2: Assessing the amplicon target region.

Irrespective of the gene region chosen for investigation it is advisable to be aware of the composition of that region. This holds true especially for methods that rely on a small amount of data from the target region to infer conclusions, such as SNP data or taxonomic assignments between closely related taxa based off a few nucleotides. The primary reason for such attention is due to the fact that not all gene regions are “created” equal. Some gene regions can be more prone to error due to the occurrence of homopolymer stretches or secondary structures within the target area, particularly

when dealing with 454 or Ion Torrent data. There are also well-recognised issues with quality and fidelity when dealing with target regions that are GC rich (Benjamini & Speed, 2012; Dabney & Meyer, 2012; Chen *et al.*, 2013; Ross *et al.*, 2013). Both of these issues are in addition to the typical drop off in sequence quality and increase in potential error observed towards the sequencing length limitations of any given platform. The error rate, in addition to the quality of an amplicon sequence, is not uniform across the length of itself (Figure 5.2.3) nor is there necessarily a common error rate across different amplicon targets. Also worth noting is the potential for error rates to fluctuate between runs on the same platform on the same control DNA.

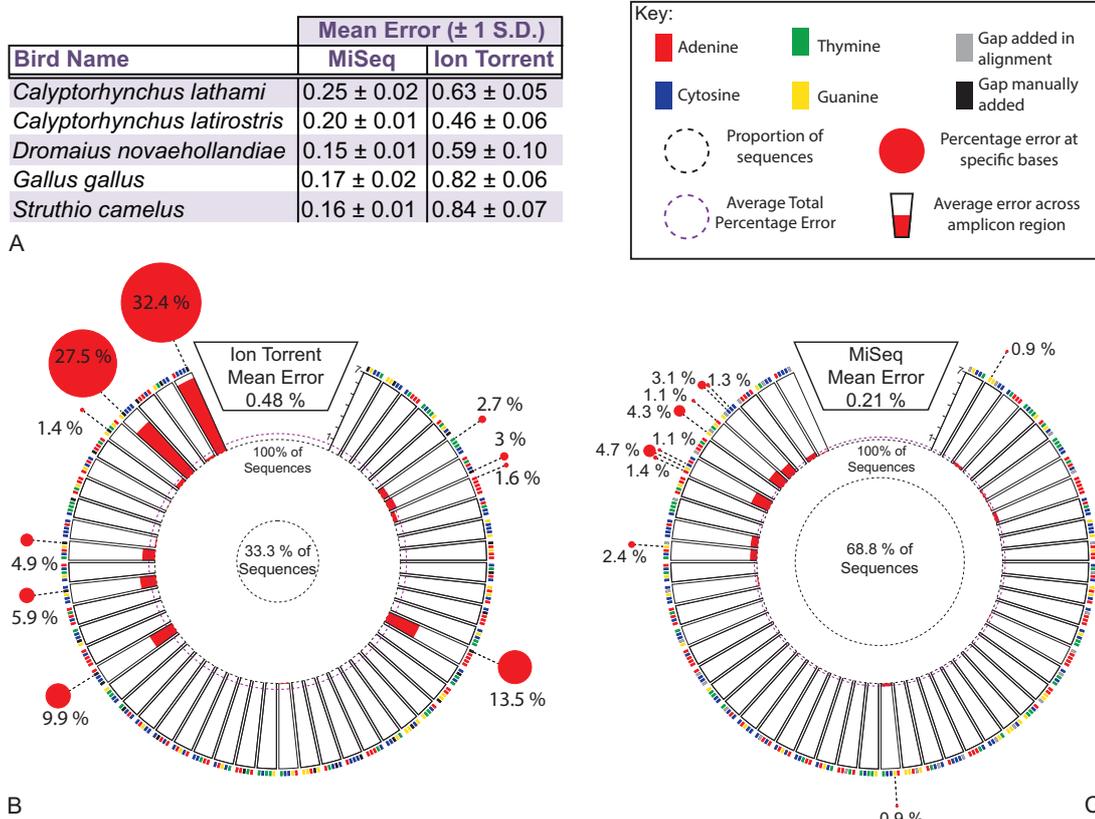


Figure 5.2.3 Average sequencing error rates across a single amplicon region. Average sequencing error rates are shown for multiple bird species across the whole of a short 12S rRNA gene region (A). Additionally, the error profile across the gene region is shown for *Calyptorhynchus lathami* for both the Ion Torrent PGM (B) and MiSeq (C) with key. The error patterns observed were similar across all species sequenced. Error rates are shown across 5 bp segments and where error rates were above 1% for a single base this is indicated through the red circles.

Some amplicon regions will undoubtedly sequence better than others due to the presence or absence of homopolymer regions and the complexity of the base composition. Rather than relying on generic error rates reported by the manufacturers or in the literature in the case of amplicons it is preferable to determine the error rate for the target region. In a small-scale experiment where single source samples for multiple bird species were sequenced, the error profile of the chosen ~250 bp target region of the 12S gene can be seen (Figure 5.2.3). It is evident that on both platforms the overall error varies slightly from species to species, yet a much greater range of mean error rates is observed in the case of the Ion Torrent PGM relative to MiSeq sequencing (Figure 5.2.3A). The variation in error rates observed across species is likely due to overall error rates associated with each platform. In addition to this it is observed that the percentage error for certain regions and specific bases far exceed the reported error rates cited for the platforms and in some cases, most notably with the Ion Torrent, certain regions recorded error rates as high as 7% (Figure 5.2.3B & C). Moreover, the increased error beyond that reported for the platform, and in some instances greater than 1%, often cited as a level used to eliminate erroneous sequences, is not solely confined to the 3' end of the amplicon read. In the case of the Ion Torrent an error rate of 13.5% was calculated just 80 bases into the amplicon read (Figure 5.2.3B). Although significantly lower error rates at specific bases and in specific regions was observed in the MiSeq, bases and regions recording error rates approaching the 1% mark were found mid-way through the amplicon. In both cases this is despite average error rates for those sub-sampled sequences being calculated as 0.48% for Ion Torrent and 0.21% for the MiSeq (Figure 5.2.3B & C). The propensity for error is again highlighted in the case of the Ion Torrent whereby only 33.3% of sequences obtained for that sub-sample were contained within the highest unique cluster, which is alarming given that it is a single source sample, with theoretically only one possible sequence composition, yet two thirds of the sequences differed from the most common. Although the error profile for only one sub-sample for a single species (*C. lathami*) is shown for both the Ion Torrent and MiSeq in Fig. 3B and 3C a similar error profile was found across all species on both platforms.

When dealing with amplicon sequencing, determining not only the overall error rate for the target region but also calculating an error spectrum can have many benefits. In doing this, certain error “hot-spots” can be detected, and being aware of the

presence of such areas can enable more informed decisions in relation to determining OTUs, calling SNPs and verifying taxonomic identifications. Having a good understanding of the composition of the chosen target region can also be of benefit. If the area of the amplicon that proves to be most informative is at the 3' end of the amplicon sequence for instance, it is possible to optimally position the direction of sequencing. The profile may also dictate if a paired end strategy is more appropriate. Single-source samples specific to the targeted gene region can also facilitate the monitoring of run-to-run variation in error rates specifically for the amplicon of choice.

Awareness of the error profile and composition of an amplicon gene region is an important consideration that can impact upon one's ability to taxonomically discriminate taxa. If an amplicon sequencing approach is adopted some of the biases associated with PCR and primer skews may also be minimised, or can at least be highlighted, by ensuring that the primer binds on all taxa of interest through the use of *in silico* bioinformatics (Ficetola *et al.*, 2010). It is also worth being aware of the fact that no primer is truly universal. It is therefore worthwhile to consider the use of a multi-locus approach especially given the current patchy state of reference databases where some taxa may be present for one gene region but not another (Murray *et al.*, 2012; Taylor & Harris, 2012). Lastly, it is worth noting that just because a primer set is said to "work well" in one study (or because it is a currently accepted DNA barcode) it does not necessarily follow that it will also be fit for purpose in another study. This issue is clearly highlighted in the case of Australian mammals where the conventional barcode COI is wholly insufficient due to the poor representation of Australian marsupials and rodents for this gene in current databases such as GenBank or BOLD (Murray *et al.*, 2012; Murray *et al.*, 2013; Deagle *et al.*, 2014).

5.2.4.3 Experiment 3: Importance of experimental controls

Once an appropriate target region(s) is selected and DNA extracts are screened for copy number and inhibition, decisions then turn to how best to build a library free of artefacts and contamination. The issue of contamination and artefact formation should always be considered when PCR is involved. Amplicon sequencing on 454, Illumina or Ion Torrent, always involves the manipulation of PCR products, thus

workflows are susceptible to contamination. Amplicon sequencing workflows on current second generation platforms involve multiple rounds of PCR (Bybee *et al.*, 2011; de Cárcer *et al.*, 2011), many published workflows utilise three rounds of PCR (Bronner *et al.*, 2001; Varley & Mitra, 2008; Archer *et al.*, 2012; Campo *et al.*, 2014): a primary PCR, an MID (Multiplex Identifier) tagging PCR (i.e. indexing) and then amplification within emulsions (454, Ion Torrent) or on a flow cell (Illumina). Unlike Sanger sequencing when low-level contaminants presented as a ‘bumpy’ baseline, HTS will show these as unambiguous sequences. In many respects high-throughput amplicon sequencing should be viewed as the “white-glove” test of laboratory cleanliness.

A major potential source of contamination is due to the handling of amplicon products post-PCR. Thus it is strongly recommended (where possible) to conduct pre-PCR and post-PCR work in independent, dedicated spaces or labs, preferably physically separated from each other. It is advisable to minimise the handling of untagged amplicon products as much as possible to prevent cross-contamination of samples. It is for this reason that methods such as nested- or hemi-nested PCR, reamplification, and ligation of ‘sequencing adapter-MID tag’ sequences to untagged amplicons can be problematic. Employing nested-PCR approaches to enrich for low abundant taxa may be more prone to contamination and/or artefactual sequences when compared to PCR-free targeted enrichment of amplicons.

It goes without saying that minimising contamination is essential in all studies where amplicon sequencing is used, especially those that seek to explore diversity in instances where it arises as a result of low-abundant taxa or variants (Budowle *et al.*, 2014; Sajantila, 2015). The increased sequencing depth afforded by HTS should not be viewed as a means by which to “cut-through” potential contamination be it environmentally derived or otherwise. This is particularly true in scenarios where endogenous DNA is highly degraded or in low copy number, as is the case for ancient or environmental DNA, where modern or well-preserved DNA sequences will amplify more readily. The degree to which a sample has been contaminated cannot be known *a priori* and such contamination, especially environmentally derived, may not always be low-level. Increased sequencing depth, therefore, will do nothing to dilute the level of contaminant sequences, and neither will arbitrary cut-

offs designed to remove low-abundant unique sequence clusters or OTUs. There is no substitute for environmental, extraction and PCR blank reaction controls. The failure to use controls can never be justified and nor can the failure to report the use of controls, even when they turn up negative results. Controls are the only true means by which it can be determined whether or not the fidelity of samples have been maintained throughout processing. Controls are seldom reported in papers using HTS (De Barba *et al.*, 2014), especially in the fields of environmental DNA and microbial metagenomics. The lack of reporting of controls in bacterial metagenomics studies is alarming given the ubiquitous nature of bacteria. In the absence of such controls it is impossible to say what bacteria are endogenous to the samples collected or even the extent to which bacteria common to the environment contribute to the microbiome from which the sample was collected. This is particularly true when dealing with coarse taxonomic assignments at an ordinal or family level, not to mention when making claims about the presence, absence and/or abundance of OTUs. The importance of controls in bacterial metagenomics is clearly shown when considering that after OTU sequences present in control reactions conducted during bacterial profiling of hairs (Tridico *et al.*, 2014) were removed the number of OTU sequences present in scalp hair samples dropped by ~60-70% (Table S5.2.2). Moreover, it is clear that this is not a simple case of PCR contamination arising from poor lab practice as the drop off for pubic hair, conducted within the same PCR plate was much lower at ~30% (see Table S5.2.2 and Tridico *et al.*, 2014) for further details and also (Porter *et al.*, 2013) for another example of using controls to filter sequences for contamination). High-throughput sequencing serves to hold up a magnifying glass to the laboratory practices of any lab that makes use of it. The depth at which a sample can be sequenced can result in even the lowest levels of contamination being revealed. This can be problematic where analyses and conclusions rely on low abundant sequences and the only assured means of retaining confidence in results and conclusions in these cases is through careful library preparation and considered data analysis. While it is easy to pick out common laboratory contaminants or aberrant sequences when such amplicons assign taxonomically to taxa not found in the study area, it is more difficult to account for cross-sample, environmental or laboratory contamination that closely resembles the taxa or sequence variants of interest.

The use of indexed (or MID tagged) primer sequences is not only useful in allowing the processing of multiple samples in parallel but it is also a convenient means by which to filter. This can be achieved by only allowing amplicon sequences with the exact MID tag to be used in further analyses. However, the use of the word “unique,” and other related terms, with respect to these MID tags is slightly misleading as in reality MID tags are often recycled across many samples. This may prove problematic due to sample carry-over that is observed with some platforms or potential library contamination by means of aerosolised particles during library generation. The issues surrounding the possibility of sample carry-over is best illustrated when considering the first Ion Torrent PGM run that the authors of this paper outsourced to a sequencing facility where, when the data was analysed, 25 tags not used in the preparation of the amplicon library were detected, amounting to 0.02 % of the total number of reads returned. Out of these 25 tags, if the tag that was present in the greatest abundance had been used in the experiment, approximately 1.2 % of the reads belonging to the sample to which it was assigned could have been indistinguishable contamination. In this instance it was clear that the contamination might have arisen at the sequencing facility itself as none of the tags detected were ever used in the laboratory where the amplicon library was generated. This highlights an important issue when considering the outsourcing of DNA sequencing to other labs, commercial or otherwise. It may be necessary in future to provide statistics of run-to-run carry over and the timeframe between the re-use of tags when such a sequencing facility also generates the amplicon for sequencing. Numerous studies are now beginning to highlight the issue of contamination arising from the laboratory, reagents and commercial kits (Salter *et al.*, 2014; Sajantila, 2015). Anecdotally, researchers also talk about contaminating data from sequencing facilities but it is rarely, if ever, reported in the literature.

A simple strategy to limit issues associated with this is to increase the timeframe between the first use and subsequent re-use of an MID tag. While it is tempting when dealing with a small number of core loci to re-use a limited number of tags, such as those officially released by the platform manufacturers, it nonetheless increases the likelihood of contamination creeping in from run to run and building up over time. Expanding the number of MID tags used in a lab greatly reduces the potential of MID tag contamination with little extra cost. A further means of ensuring tag

contamination is kept to a minimum is the use of differing MID tags at the 5' and 3' end of the amplicon sequences (see Section 5.2.3.1), which can also benefit in terms of data filtering to increase the likelihood of only high quality sequences being retained. Additionally, the use of different 5' and 3' MID tags on an amplicon greatly increases the number of possible combinations at a laboratory's disposal. Finally, the use of different 5' and 3' MID tagged amplicons may also help in the detection of chimeric sequences. The downside of a method such as this however is the cost associated with ordering primers; although this can be kept to a minimum by not ordering HPLC purified primers as synthesis errors are easily managed by post-run filtering. Moreover quality control validation by mass spectrometry is now commonplace and serves to minimise the likelihood of primers with high proportions of incorrect bases.

While some might argue that the purchase of MID tagged primers is expensive the counter argument is that so too is repeating runs where the researcher believes the data is compromised. In our lab six reads were detected of a Chinese herbal plant from one study (Coghlan *et al.*, 2012) that turned up in a palaeosediment sample from Australia. In this instance both samples shared the same MID tags despite being many runs apart. In sensitive applications the re-use of MID tags may be a false economy. Low-template samples necessitate sensitivity and single-use of tag combinations. This has the added benefit that each amplicon product generated is unique to the originating sample and contamination can be removed bioinformatically.

5.2.4.4 Experiment 4: Library generation efficiency

The opening and closing of PCR-tubes or plates post-PCR and the handling of untagged amplicon products serve to increase the chances of untraceable contamination as a result of poor laboratory technique or the release of aerosolised amplicons. It is for this reason that a single "full" fusion tagged TSP (see Figure S5.2.1E) PCR approach (Sonstebo *et al.*, 2010; Clarke *et al.*, 2014) or sequencing adapter ligation post-MID tagging (Binladen *et al.*, 2007) via PCR method is preferable from the perspective of contamination control. The drawbacks associated with a "full" fusion tagged TSP PCR approach centre around a loss of PCR efficiency due to the long fusion primers required and also the problems surrounding

primer-dimer. However, careful size selection can assist with dimer removal (DeAngelis *et al.*, 1995; Lundin *et al.*, 2010; Borgström *et al.*, 2011; Quail *et al.*, 2012a). The ligation of sequencing adapters post-MID tagging via PCR itself can be inefficient and may be biased towards the preferential ligation of certain amplicons or terminal bases. In some cases the efficiency drop-off associated with a “full” fusion tagged TSP approach can be mitigated through the use of the modular tagging of amplicons using a single PCR (MoTASP) method (Clarke *et al.*, 2014) or by simply spiking in some standard non-fusion TSP into the PCR reaction containing “full” fusion tagged TSP (see Figure S5.2.1E). The latter showed generally modest efficiency improvements when compared to qPCR in the absence of spiking in standard non-fusion TSP, however the C_T value shifts in qPCR varied considerably for each platform (Table S5.2.3). Additionally, the spiking in of standard non-fusion TSP when using “full” fusion tagged TSP still showed a general increase in C_T values when compared to qPCR containing only standard non-fusion TSP, particularly in the case of the MiSeq (Table S5.2.3). Although the MoTASP method has been reported to improve PCR efficiency, it is unclear as to the extent this may be the case as qPCR was not carried out and neither was a direct comparison of sequencing results (Clarke *et al.*, 2014).

The use of a “full” fusion tagged TSP approach where a library is generated in a single step is theoretically the cleanest way to generate amplicon libraries. The downside to this is the drop in PCR efficiency discussed above. A common alternative pathway is a series of primary PCRs which are pooled and followed by a secondary PCR to amplify sequencing adapters and/or MID tags onto the target sequences. Notwithstanding the contamination risk inherent to this two-step approach it is also the source of inter-sample chimeras, presumably through incomplete extension and/or ‘jumping’ PCR (Pääbo *et al.*, 1990). Practitioners need to carefully weigh the benefits and drawbacks of each library building method and be cognisant of how the method impacts on the conclusions they hope to draw from the resultant data.

5.2.4.5 Experiment 5: Analysis parameters and their impact

It is beyond the scope of this study to delve into the complexities of data analysis. It is however relevant to note that amplicon data can be analysed in many different

ways, sometimes subtly so, that can result in quite dissimilar outcomes. It is also worth noting that analysis parameters are contingent on the benchwork component of amplicon sequencing workflows. To date there is no currently accepted best practice pipeline or approach to the analysis of amplicon sequencing output, although many do exist (Schloss *et al.*, 2009; Caporaso *et al.*, 2010; Edgar, 2010; Piry *et al.*, 2012). Nevertheless one of the few agreements on the way in which both shotgun and amplicon sequencing data are handled is the necessity to filter sequences for error and potential contamination in a manner that strikes a balance between overly relaxed and unnecessarily stringent filtering. The manner in which such filtering is done and the definitions associated with various processes along the filtering pipeline can have a marked impact on the final result. Naturally, the stringency and type of filtering method employed is both platform dependent and sensitive to the library building methodology.

The difficulty of analysing the diversity of samples whilst accounting for sequence quality, abundance and attempting a taxonomy-independent measure of analysis is illustrated in Figure 5.2.4. Depending on the quality filtering method (QFM), abundance filtering method (AFM) and taxonomy-independent method (TIM) used (Figure 5.2.1 & 5.2.4) the number of taxonomic units detected varied between 3 and 22 operational taxonomic units (OTUs) or between 3 and 14 distance-based operational taxonomic units (DTUs) (Murray *et al.*, 2013) (Figure 5.2.4). In each case the minimum average Quality Scores (Q-Scores) for all sequences post-filtering were well above the standard cut-off of Q15. Tellingly however, when considering QFM1 and QFM4 (see Figure 5.2.1 for definitions and also Table S5.2.4) where individual bases below Q15 were permissible, a sizeable proportion of sequences contained bases below Q15 (57.0% and 42.3% respectively) and there was a noticeable percentage of bases below Q15 overall (2.6% and 0.9% respectively) (Table S5.2.4).

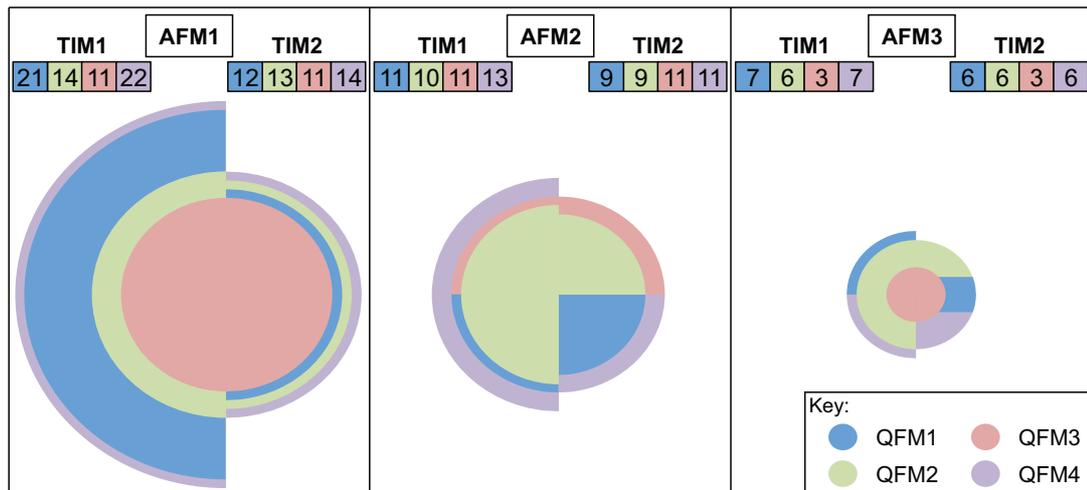


Figure 5.2.4 Impact of analysis parameters on the numbers of taxonomic units obtained for a bulk-bone sample. A number of analysis parameters were used to analyse a complex mixture containing numerous taxa. Different quality and abundance filtering methods were used in addition to two taxonomy-independent measures of analysis, full definitions and explanations of which are in Fig. 1. The spread in the numbers of taxonomic units obtained across the combinations of parameters chosen is seen. The radius of each semicircle represents the number of taxonomic units obtained given a set combination of the parameters used. The number of taxonomic units is also indicated above each semicircle. Each semicircle is proportional to all others. AFM – abundance filtering method; QFM – quality filtering method; TIM – taxonomy-independent method.

The use of Phred Q-Scores, as noted above, is one means by which to filter sequence data for error. Many papers, including those by the authors, make mention of how the data contained within has been filtered for quality, however, few make mention of how this is done thus making it difficult to reproduce data from the pipeline used. It is an open question as to what truly constitutes a high quality sequence. For instance, is it one where the average Q-Score across its length is $>Q20$ or should it be a requirement that all bases within the sequence be at least $Q15$? Q-scores are also complicated by the fact that different platforms use different methods when generating Q-scores. An issue surrounding the use of a stringent Q-Score cut-off that all bases must meet is the fact that the Q-Score of a base is impacted by the Q-Scores of the bases immediately surrounding it. Homopolymers are generally areas of quite low quality and this low quality can extend for a number of bases beyond the homopolymer stretch itself. In an extreme example, a Q-score based filtering method

might actively discard amplicon variants that contain homopolymer stretches in favour of those that do not, thereby warping the composition of the resultant data.

In addition to Q-score cut-offs, filtering of sequence reads below a certain abundance is often employed. This is often cited as an attempt to reduce the possibility of erroneous and artefactual sequences as well as to remove instances of low-level contamination. At times such an approach could be seen as the molecular biology equivalent of “sweeping the dirt under the carpet” — simply moving a baseline until one is happy with the data will ultimately reduce sensitivity and reduce transparency of data fidelity. As with Q-score quality filtering, abundance filtering can be performed in a variety of ways with no accepted definition of what should be classed as a low abundant grouping of sequences. Methods of abundance filtering vary from the removal of singletons only, to the use of, at times, arbitrary cut-offs or inferred cut-offs defining a low abundance cluster (see Figure 5.2.1 for examples and Figure 5.2.4 for impacts). The choice of an appropriate abundance filter is no easy task especially in cases where there is unequal sequencing depth that may necessitate the need for sample specific abundance filters.

The fluidity of the definition of a high quality sequence and what constitutes a low abundance cluster as well as the order in which filtering steps are performed (see Figure 5.2.1 for examples and Figure 5.2.4 for impact) can all combine to create a rather difficult analysis of the diversity of a sample when dealing with heterogeneous samples of unknown composition. This holds true not only when dealing with the abundance of sample constituents but also when dealing with presence and/or absence. These factors are exacerbated further when weighted analyses are employed. In reality there is no means by which to determine the “correct” number of OTUs within a sample. For instance, with regards to a pool of single-source bird samples containing a single sample of only one representative of the family Dromaiidae, *Dromaius novaehollandiae* (emu), a total of four distinct OTUs were obtained post-filtering (data available from authors upon request). Also worth noting is the importance of ensuring samples are free of inhibition and have sufficient copy number of DNA when conducting OTU analyses that involves a requirement for a particular OTU to occur in a certain proportion of uniquely tagged replicates before it is accepted (Willerslev *et al.*, 2014). If such a criterion were used in the two-fish

screening assay (Figure 5.2.2), the genus *Engraulis* would have been excluded at times as it only occurred in a single replicate in certain cases, even though its presence was confirmed using *Engraulis* specific primers. Notwithstanding the above, when used appropriately, OTUs can be a useful index for species diversity provided parameters are both transparent and consistent across samples and studies.

5.2.5 Conclusion

It is proving to be the case in amplicon sequencing that a one-size-fits-all approach is ill-advised and unwise, due to differing budgets, scopes and end-goals. It is therefore not the aim of this article to call for definitive guidelines with regard to best practice when generating amplicon libraries or sequencing them, although a set of flexible reporting guidelines may be appropriate. It is hoped that this paper may instead prove to be a catalyst ultimately aiding in the development of robust amplicon sequencing workflows. The generation of amplicon data is easy, however the generation of high-fidelity data free of contamination, artefacts and appropriately analysed, is far more complex. It is important to be aware of the limitations of amplicon data and know that with the advances afforded by it there are many hurdles. It is imperative that more attention be paid to the processes involved in preparing amplicon libraries to limit some of the pitfalls highlighted in this paper. While published data can be analysed and re-analysed time and again, such as when reference databases improve, the library generation step is not as easily, quickly or cheaply repeated. It is widely acknowledged that amplicon sequencing will continue to play an important role across a wide range of applications. Taken together these data suggest that, in order to get the most out of amplicon datasets, careful attention should be paid to workflows at both benchtop and desktop.

5.2.6 Acknowledgements

The authors acknowledge sequencing assistance from the State Agricultural Biotechnology Centre (SABC, Murdoch University, Perth, Australia), technical expertise and assistance during PGM experiments from staff at the LotteryWest State Biomedical Facility Genomics (Perth, Australia) in addition to the Centre for Comparative Genetics (Murdoch University, Perth, Australia) and iVEC for

computational support. We would also like to thank the various members of the TrEnD Lab (Curtin University, Perth, Australia) and Centre for GeoGenetics (University of Copenhagen, Copenhagen, Denmark) for their varied contributions.

5.2.7 References

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnrke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, *12*, R18.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403-410.

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K. H., . . . Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, *21*, 1966-1979.

Archer, J., Weber, J., Henry, K., Winner, D., Gibson, R., Lee, L., . . . Quiñones-Mateu, M. E. (2012). Use of four next-generation sequencing platforms to determine HIV-1 coreceptor tropism. *PLoS One*, *7*, e49602.

Baird, D. J., & Hajibabaei, M. (2012). Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, *21*, 2039-2044.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research*, *34*, D16-D20.

Binladen, J., Gilbert, M. T. P., Bollback, J. P., Panitz, F., Bendixen, C., Nielsen, R., & Willerslev, E. (2007). The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, *2*, e197.

Blankenberg, D., Kuster, G. V., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., . . . Taylor, J. (2001). Galaxy: A web-based genome analysis tool for experimentalists *Current Protocols in Molecular Biology*: John Wiley & Sons, Inc.

Bohmann, K., Monadjem, A., Lehmkühl Noer, C., Rasmussen, M., Zeale, M. R. K., Clare, E., . . . Gilbert, M. T. P. (2011). Molecular diet Analysis of two African free-tailed bats (Molossidae) using high throughput sequencing. *PLoS One*, 6, e21441.

Borgström, E., Lundin, S., & Lundeberg, J. (2011). Large scale library generation for high throughput sequencing. *PLoS One*, 6, e19119.

Bronner, I. F., Quail, M. A., Turner, D. J., & Swerdlow, H. (2001). Improved protocols for Illumina sequencing *Current Protocols in Human Genetics*: John Wiley & Sons, Inc.

Budowle, B., Connell, N., Bielecka-Oder, A., Colwell, R., Corbett, C., Fletcher, J., . . . Minot, S. (2014). Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics*, 5, 1-18.

Bunce, M., Oskam, C., & Allentoft, M. (2011). The use of quantitative real-time PCR in ancient DNA research. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA: Methods and Protocols* (pp. 121-132): Humana Press.

Burgar, J. M., Murray, D. C., Craig, M. D., Haile, J., Houston, J., Stokes, V., & Bunce, M. (2014). Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology*, 23, 3605-3617.

Bustin, S. A., Benes, V., Garson, J. A., Hellems, J., Huggett, J., Kubista, M., . . . Wittwer, C. T. (2009). The MIQE guidelines: minimum information for publication of quantitative Real-Time PCR experiments. *Clinical Chemistry*, 55, 611-622.

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., . . . Crandall, K. A. (2011). Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol Evol*, 3, 1312-1323.

- Campo, D. S., Dimitrova, Z., Yamasaki, L., Skums, P., Lau, D. T., Vaughan, G., . . . Khudyakov, Y. (2014). Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC Genomics*, *15 Suppl 5*, S4.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., . . . Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*, 335-336.
- Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., . . . Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *proceedings of the National Academy of Sciences*, *108*, 4516-4522.
- Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y., & Hwang, C.-C. (2013). Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS One*, *8*, e62856.
- Clarke, L. J., Czechowski, P., Soubrier, J., Stevens, M. I., & Cooper, A. (2014). Modular tagging of amplicons using a single PCR for high-throughput sequencing. *Molecular Ecology Resources*, *14*, 117-121.
- Coghlán, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P., . . . Bunce, M. (2012). Deep sequencing of plant and animal DNA contained within traditional chinese medicines reveals legality issues and health safety concerns. *PLoS Genetics*, *8*, e1002657.
- Consortium, T. H. M. P. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*, 207-214.
- Cooper, A. (1994). DNA from museum specimens. In B. Herrmann & S. Hummel (Eds.), *Ancient DNA* (pp. 149-165): Springer New York.
- Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, *409*, 704-707.

Dabney, J., & Meyer, M. (2012). Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques*, *52*, 87-94.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, *14*, 306-323.

de Cárcer, D. A., Denman, S. E., McSweeney, C., & Morrison, M. (2011). Strategy for modular tagged high-throughput amplicon sequencing. *Applied Environmental Microbiology*, *77*, 6310-6312.

Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.

Deagle, B. E., Gales, N. J., Evans, K., Jarman, S. N., Robinson, S., Trebilco, R., & Hindell, M. A. (2007). Studying seabird diet through genetic analysis of faeces: a case study on macaroni penguins (*Eudyptes chrysolophus*). *PLoS One*, *2*, e831.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*.

DeAngelis, M. M., Wang, D. G., & Hawkins, T. L. (1995). Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research*, *23*, 4742-4743.

Ding, T., & Schloss, P. D. (2014). Dynamics and associations of microbial community types across the human body. *Nature*, *509*, 357-360.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., . . . Wilson, A. (2011). Geneious v7.1, created by Biomatters. Available from <http://www.geneious.com/>. Retrieved from <http://www.geneious.com/>

- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792-1797.
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*, 2460-2461.
- Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*, 996-998.
- Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*, 2194-2200.
- Ekblom, R., & Galindo, G. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, *107*, 1-15.
- Faircloth, B. C., & Glenn, T. C. (2012). Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS One*, *7*, e42543.
- Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., . . . Pompanon, F. (2010). An In silico approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 1-10.
- Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423-425.
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguët-Covex, C., De Barba, M., . . . Taberlet, P. (2014). Replication levels, false presences, and the estimation of presence / absence from eDNA metabarcoding data. *Molecular Ecology Resources*, n/a-n/a.
- Fierer, N., Lauber, C. L., Zhou, N., McDonald, D., Costello, E. K., & Knight, R. (2010). Forensic identification using skin bacterial communities. *proceedings of the National Academy of Sciences*, *107*, 6477-6481.

- Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., . . . Nekrutenko, A. (2005). Galaxy: A platform for interactive large-scale genome analysis. *Genome Research, 15*, 1451-1455.
- Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology, 11*, R86-R86.
- Gonzalez, A., & Knight, R. (2012). Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology, 23*, 64-71.
- Hamady, M., Lozupone, C., & Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal, 4*, 17-27.
- Hunter, A. A., Macgregor, A. B., Szabo, T. O., Wellington, C. A., & Bellgard, M. I. (2012). Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine, 7*, 1.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research, 17*, 377-386.
- Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30*, 3059-3066.
- Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing – concepts and limitations. *BioEssays, 32*, 524-536.
- Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R., & Cary, S. C. (2012). Groundtruthing next-gen sequencing for microbial ecology—biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One, 7*, e44224.
- Liu, J., Sui, Y., Yu, Z., Shi, Y., Chu, H., Jin, J., . . . Wang, G. (2014). High throughput sequencing analysis of biogeographical distribution of bacterial

communities in the black soils of northeast China. *Soil Biology and Biochemistry*, 70, 113-122.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, 30, 434-439.

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D., & Lundeberg, J. (2010). Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One*, 5, e10029.

Mason, O. U., Scott, N. M., Gonzalez, A., Robbins-Pianka, A., Balum, J., Kimbrel, J., . . . Jansson, J. K. (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *The ISME Journal*, 8, 1464-1475.

Meadow, J. F., Bateman, A. C., Herkert, K. M., O'Connor, T. K., & Green, J. L. (2013). Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ*, 1, e53.

Murray, D. C., Bunce, M., Cannell, B. L., Oliver, R., Houston, J., White, N. E., . . . Haile, J. (2011). DNA-based faecal dietary analysis: a comparison of qPCR and High Throughput Sequencing approaches. *PLoS One*, 6, e25776.

Murray, D. C., Haile, J., Dortch, J., White, N. E., Haouchar, D., Bellgard, M. I., . . . Bunce, M. (2013). Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Sci. Rep.*, 3.

Murray, D. C., Pearson, S. G., Fullagar, R., Chase, B. M., Houston, J., Atchison, J., . . . Bunce, M. (2012). High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews*, 58, 135-145.

Pääbo, S., Irwin, D. M., & Wilson, A. C. (1990). DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry*, 265, 4718-4721.

- Piry, S., Guivier, E., Realini, A., & Martin, J. F. (2012). |SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization – application to species and environmental barcoding. *Molecular Ecology Resources*, *12*, 1151-1157.
- Pompanon, F., Deagle, B. E., Symondson, W. O. C., Brown, D. S., Jarman, S. N., & Taberlet, P. (2012). Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, *21*, 1931-1950.
- Porter, T. M., Golding, G. B., King, C., Froese, D., Zazula, G., & Poinar, H. N. (2013). Amplicon pyrosequencing late Pleistocene permafrost: the removal of putative contaminant sequences and small-scale reproducibility. *Molecular Ecology Resources*, *13*, 798-810.
- Quail, M. A., Gu, Y., Swerdlow, H., & Mayho, M. (2012a). Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. *Electrophoresis*, *33*, 3521-3528.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012b). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341.
- Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., . . . Chikhi, L. (2013). A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS One*, *8*, e58971.
- Quince, C., Lanzen, A., Davenport, R., & Turnbaugh, P. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, *12*, 38.
- Razgour, O., Clare, E. L., Zeale, M. R. K., Hanmer, J., Schnell, I. B., Rasmussen, M., . . . Jones, G. (2011). High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecology and Evolution*, *1*, 556-570.
- Robasky, K., Lewis, N. E., & Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, *15*, 56-62.

- Roche. (2009). Technical bulletin: amplicon fusion primer design guidelines for GS FLX Titanium series Lib-A chemistry. *TCB No. 013-2009*, 1-3.
- Ross, M., Russ, C., Costello, M., Hollinger, A., Lennon, N., Hegarty, R., . . . Jaffe, D. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*, R51.
- Sajantila, A. (2015). Editors' Pick: Contamination has always been the issue! *Investigative Genetics*, *5*, 17.
- Salter, S., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., . . . Walker, A. W. (2014). Reagent contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, 87.
- Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, *6*, e27310.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied Environmental Microbiology*, *75*, 7537-7541.
- Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*, 1794-1805.
- Smith, D. P., & Peay, K. G. (2014). Sequence Depth, Not PCR Replication, improves ecological inference from next generation DNA sequencing. *PLoS One*, *9*, e90234.
- Sonstebo, J. H., Gielly, L., Brysting, A. K., Elven, R., Edwards, M., Haile, J., . . . Brochmann, C. (2010). Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol Ecol Resour*, *10*, 1009-1018.

- Sun, B., Wang, F., Jiang, Y., Li, Y., Dong, Z., Li, Z., & Zhang, X.-X. (2014). A long-term field experiment of soil transplantation demonstrating the role of contemporary geographic separation in shaping soil microbial community structure. *Ecology and Evolution*, *4*, 1073-1087.
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045-2050.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., . . . Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*, e14.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, *28*, 2731-2739.
- Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, *12*, 377-388.
- Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*, 283-285.
- Thomas, R. K., Nickerson, E., Simons, J. F., Janne, P. A., Tengs, T., Yuza, Y., . . . Meyerson, M. (2006). Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Methods*, *12*, 852-855.
- Thomsen, P. F., Kielgast, J. O. S., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., . . . Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565-2573.
- Tridico, S. R., Murray, D. C., Addison, J., Kirkbride, K. P., & Bunce, M. (2014). The application of metagenomic analyses of human hair shafts in forensic

investigations using next generation sequencing: a qualitative assessment. *Investigative Genetics*, 5, 16.

Turner, S., Pryer, K. M., Miao, V. P., & Palmer, J. D. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol*, 46, 327-338.

Varley, K. E., & Mitra, R. D. (2008). Nested patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Research*, 18, 1844-1850.

White, N. E., Bunce, M., Mawson, P. R., Dawson, R., Saunders, D. A., & Allentoft, M. E. (2014). Identifying conservation units after large-scale land clearing: a spatio-temporal molecular survey of endangered white-tailed black cockatoos (*Calyptorhynchus* spp.). *Diversity and Distributions*, 20, 1208-1220.

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., . . . Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, 506, 47-51.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

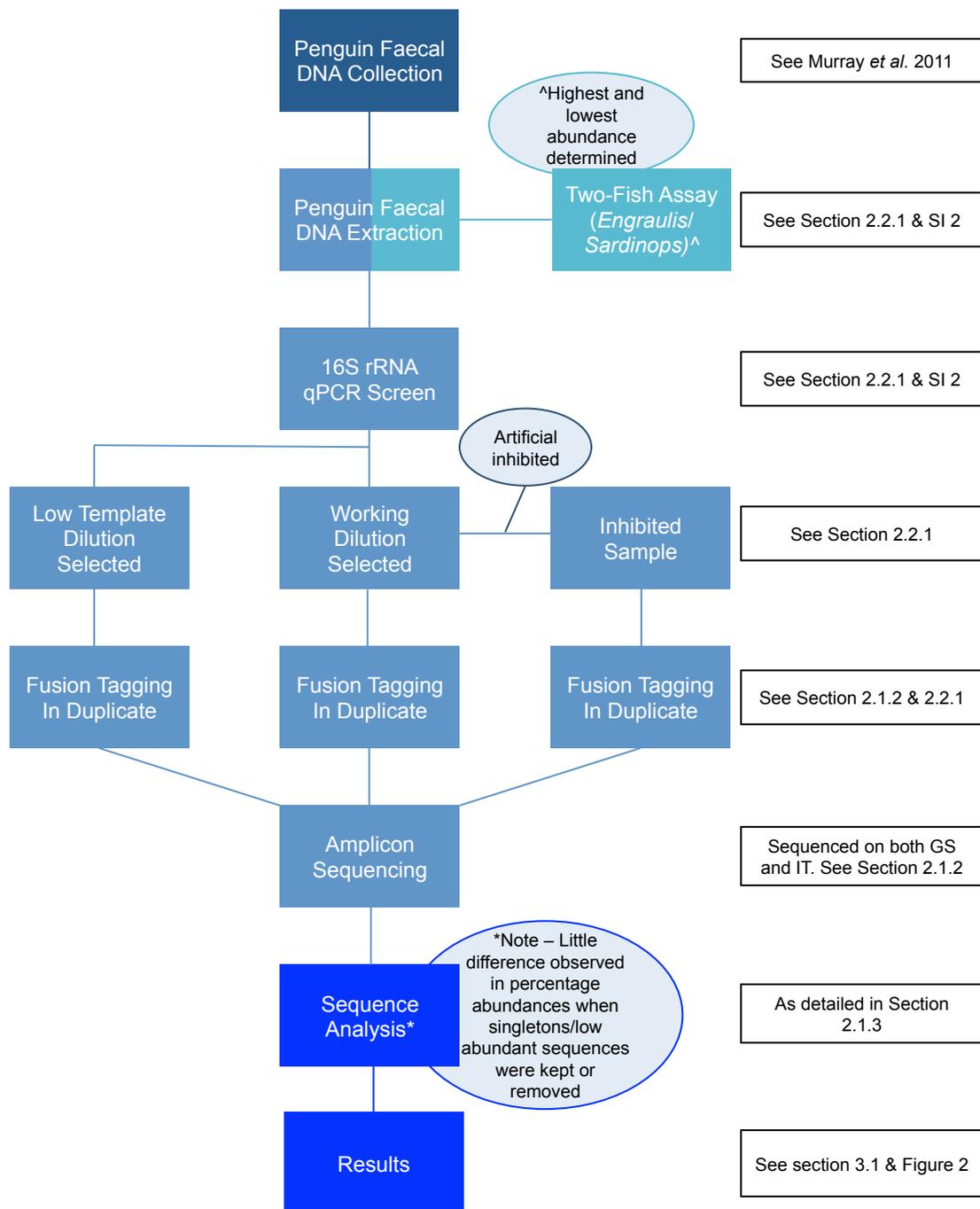


Figure S5.2.1A Experiment 1: Importance of sample screening. Schematic showing steps involved in the experiment determining the impact of inhibition and low template amount on the successful detection of two fish genera.

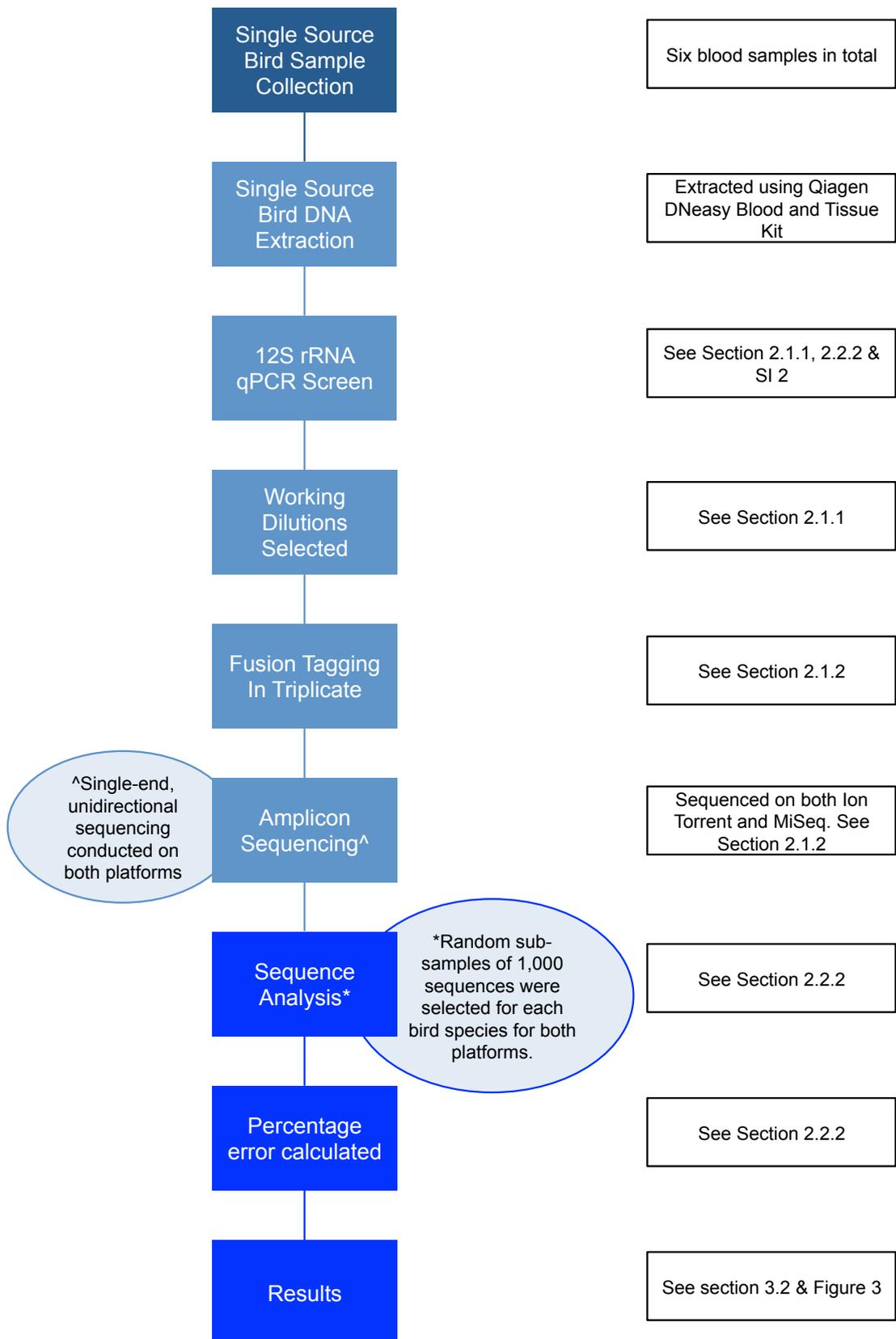


Figure S5.2.1 Experiment 2: Assessing the amplicon target region. Schematic showing steps involved in the experiment illustrating the benefits of characterising and understanding the target region in amplicon sequencing

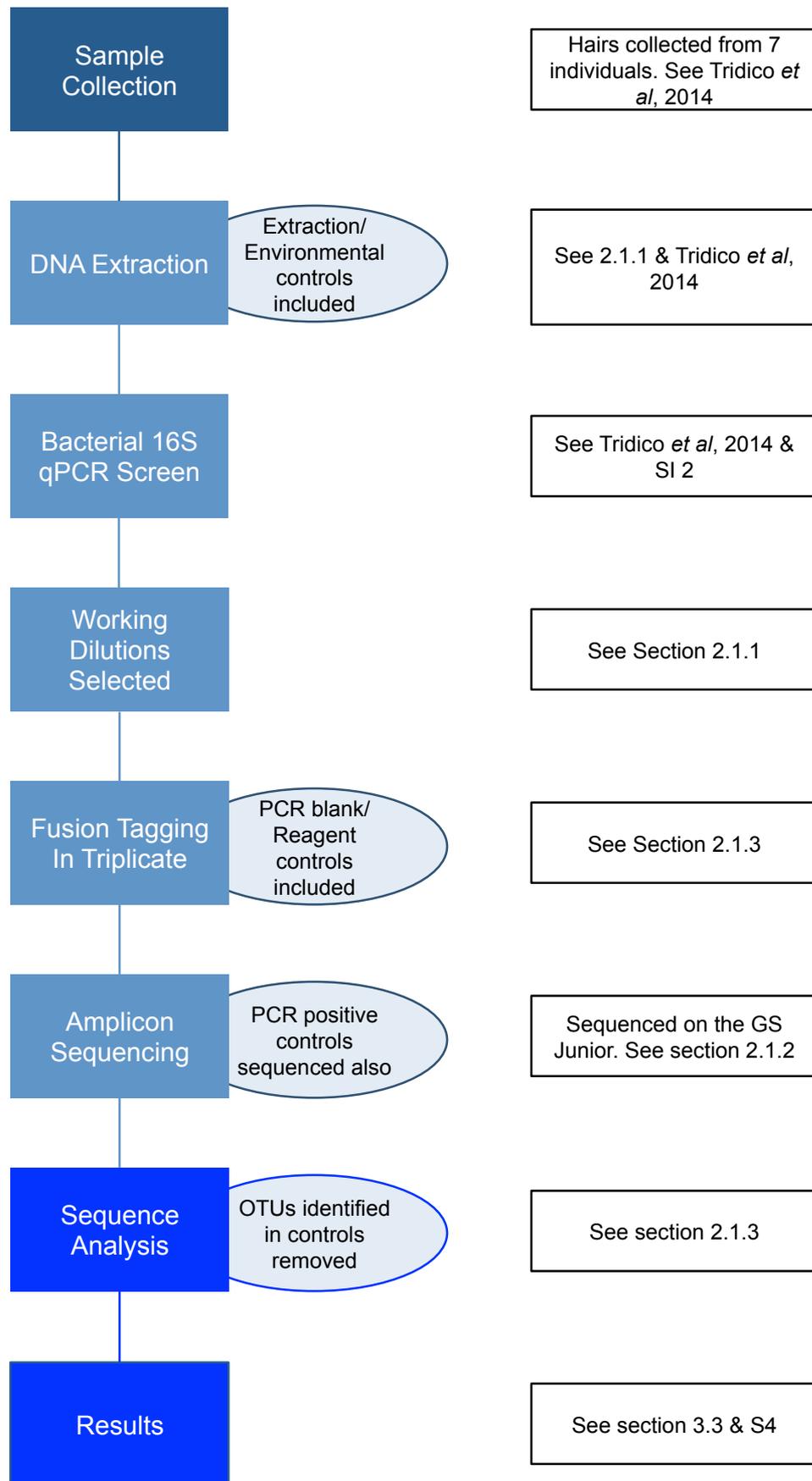


Figure S5.2.1C Experiment 3: Importance of experimental controls. Schematic showing steps involved in the experiment illustrating the importance of controls along each step during the preparation of amplicon libraries

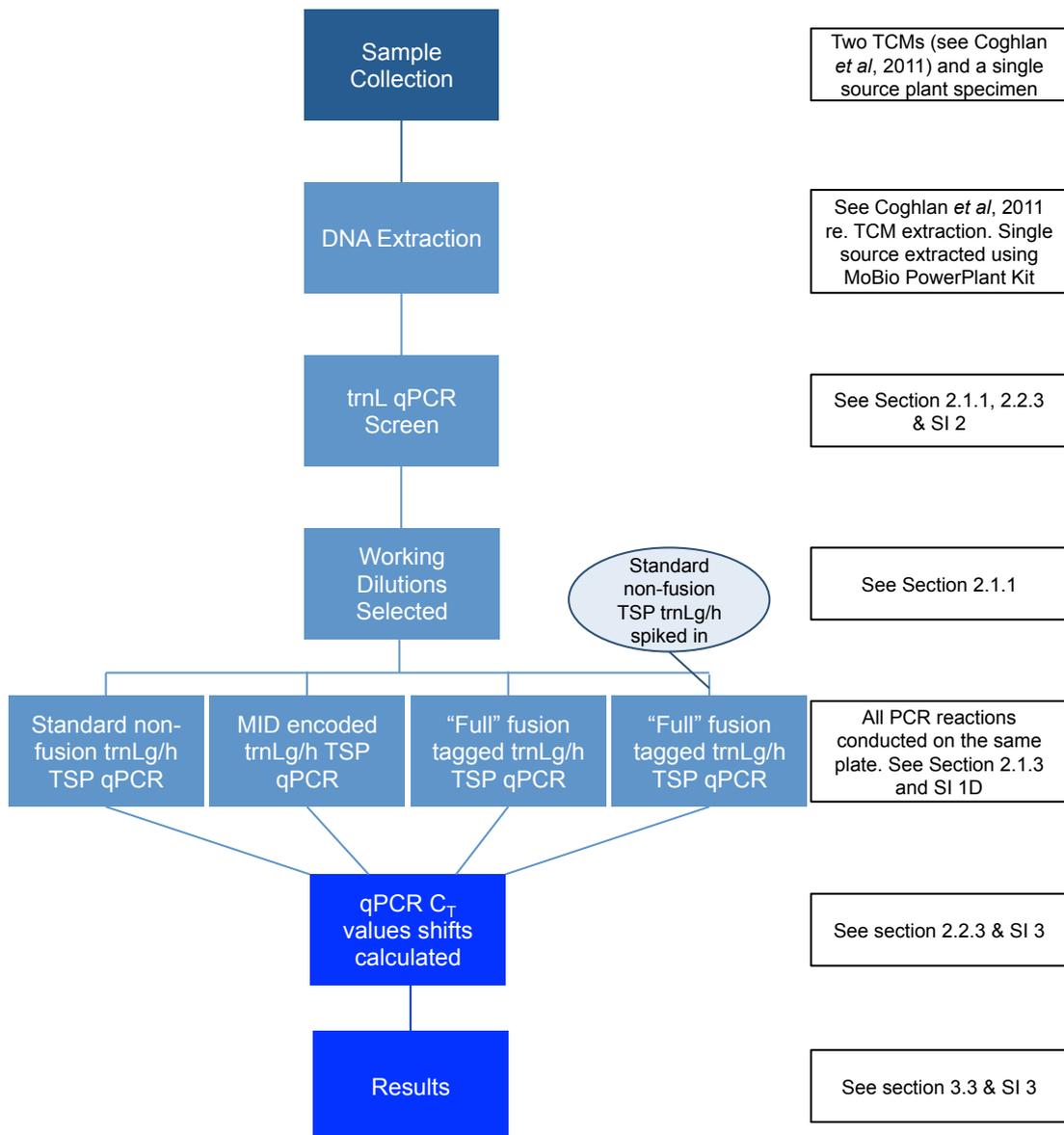


Figure S5.2.1D Experiment 4: Library generation efficiency. Schematic showing steps involved in the experiment assessing the reduced efficiency of PCR amplicon generation due to long fusion-tagged primers and the amelioration of.



Figure S5.2.1E Primer Architecture. Diagram showing the architecture of the primers used in experiments. TSP – Template specific sequence (e.g. trnLg primer); MID – Multiplex Identifier Tag (i.e. unique DNA index); Sequencing Adapters – Platform specific adapters required for clustering (MiSeq) and/or sequencing (all platforms).

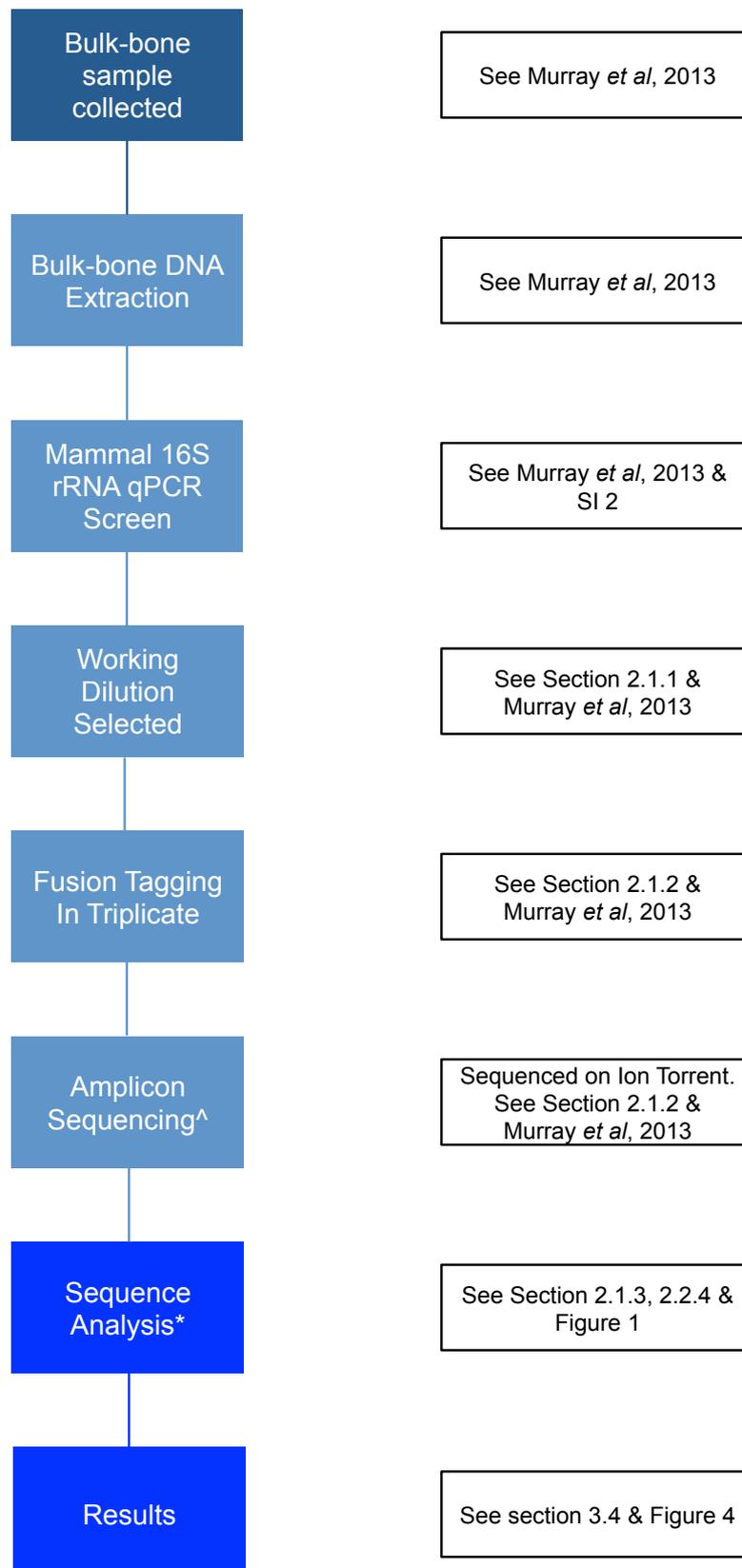


Figure S5.2.1F Experiment 5: Analysis parameters and their impact. Schematic showing steps involved in the experiment illustrating how choosing different analysis parameters can impact greatly on the number of taxonomic units determined to be in a sample.

Table S5.2.1 Table Primer Information. Details are provided for each primer set used in this paper including the sequence, annealing temperature and taxa targeted. Additionally, the experiment in which each primer was used is given.

Primer name	Target Taxa	Sequence (5'-3')	Annealing temp. (°C)	Experiment	Reference
16S1F-degenerate 16S2R-degenerate	Fish	GACGAKAAGACCCTA CGCTGTTATCCCTADRGTAAC	54	1	1
PIL1F PIL1R	<i>Sardinops sagax</i> (Australian Pilchard)	CCTAACTGGAGCCCCAAAC GCTGTGGCTCTGGGTTTAG	60	1	2
AN1F AN2R	<i>Engraulis australis</i> (Australian Anchovy)	CCTAAATACCCGCAGCCTTAT CAACTCTCGGCTTAAGGGTTT	60	1	2
16Smam1 16Smam2	Mammals	CGGTTGGGGTGACCTCGGA GCTGTTATCCCTAGGGTAACT	55	5	3
12SA 12SH	Aves	CTGGGATTAGATACCCACTAT CCTTGACCTGCTTGTTAGC	57	2	4 5
Bact_16S_F515 Bact_16S_R806	Bacteria	GTGCCAGCMGCCGCGGTAA GGACTACHVGGGTWCTAAT	54	4	6 7
trnLg trnLh	Plants	GGGCAATCCTGAGCCAA CCATTGAGTCTCTGCACCTATC	52	3	8

1. Deagle BE, Gales NJ, Evans K, Jarman SN, Robinson S, et al. (2007) Studying Seabird Diet through Genetic Analysis of Faeces: A Case Study on Macaroni Penguins (*Eudyptes chrysolophus*). PLoS ONE 2: e831.
2. Murray D, Bunce M, Cannell BL, Oliver R, Houston J, et al. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and High Throughput Sequencing approaches. PLoS One 6: e25776.
3. Taylor PG (1996) Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. Molecular Biology and Evolution 13: 283-285.
4. Cooper A (1994) DNA from Museum Specimens. In: Herrmann B, Hummel S, editors. Ancient DNA: Springer New York. pp. 149-165.
5. Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, et al. (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature 409: 704-707.
6. Turner S, Pryer KM, Miao VP, Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol 46: 327-338.
7. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proceedings of the National Academy of Sciences 108: 4516-4522.
8. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, et al. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Research 35: e14.

Table S5.2.2 Table showing the number of sequences removed due to control filtering of bacterial data. The proportion of bacterial amplicon sequences lost are shown after filtering operational taxonomic units from samples that were also found to be in DNA extraction and PCR negative controls.

Sample	Number of Sequences Without Control Filtering	Number of Sequences With Control Filtering	% of Sequences Remaining After Control Filtering
Female Scalp Hair	19762	6420	32.5
Female Pubic Hair	26714	18802	70.4
Male Scalp Hair	15776	6881	43.6
Male Pubic Hair	17514	12618	72.0

Table S5.2.3 Cycle threshold value shifts when performing fusion-tagged PCR. Cycle threshold values are shown for quantitative PCR reactions using one of the following: (1) - Standard non-fusion TSP; (2) - MID encoded TSP; (3) - “Full” fusion tagged TSP; (4) - “Full” fusion tagged TSP with standard non-fusion TSP spiked in (for further clarification see Section 2.2.4 of main article and S1 E Fig.) For (3) and (4) TSP sequences specific for each of the GS-Junior (GS), IonTorrent (IT) and MiSeq (MS) were used. Any efficiency drop off associated with using "full" fusion tagged primers (3) when compared to standard non-fusion TSP (1) is shown as is whether any efficiency drop-off can be ameliorated using a spike in of standard non-fusion TSP when using "full" fusion tagged TSP (4).

			(1)	(2)	GS (3)	GS (4)	IT (3)	IT (4)	MS (3)	MS (4)
Sample	Single Source Plant	Avg. C_T	25.63	25.75	28.21	27.61	26.65	26.40	30.61	27.96
		Diff. in C_T 1		0.12	2.58	1.98	1.02	0.78	4.98	2.33
		Diff. in C_T 2			-0.60		-0.24		-2.65	
	CAM Plant Screen 1	Avg. C_T	22.13	22.57	23.94	23.63	23.77	22.88	25.72	25.65
		Diff. in C_T 1		0.44	1.80	1.50	1.63	0.75	3.59	3.52
		Diff. in C_T 2			-0.31		-0.88		-0.07	
	CAM Plant Screen 2	Avg. C_T	21.37	20.76	22.46	22.50	22.79	21.77	26.00	23.76
		Diff. in C_T 1		-1.37	0.33	0.37	0.66	-0.37	3.87	1.63
		Diff. in C_T 2			0.04		-1.03		-2.24	
Diff. in C_T 1	Difference in C_T value compared to 1									
Diff. in C_T 2	Difference in C_T value between 3 and 4									

Table S5.2.4 Summary Quality Statistics. Table showing quality statistics of sequences resulting from quality filtering methods QFM1 and QFM4. QFM1 – No mismatches in primer sequence and no Q score filtering; QFM4 – 2 mismatches in primer sequence; UPPARSE maxee 0.5.

	QFM1	QFM4
Minimum Average Q-Score of a Sequence	27	27
Minimum Q-Score for a Base	4	7
No. of Sequences with a Base of Q-Score \leq 15	44.1%	16.2%
No. of Bases with Q-Score \leq 15	3.8%	1.9%

File S5.2.1 Single-source bird error output example.

Available from:

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124671>

doi:10.1371/journal.pone.0124671.s002

File S5.2.2 DTU calculation example.

Available from:

<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124671>

doi:10.1371/journal.pone.0124671.s003

5.3 Synopsis

High-throughput sequencing has been met with enthusiasm in many disciplines, most noticeably in molecular ecology, metagenomics and aDNA. However, such enthusiasm should not mask the difficulties associated with developing robust studies in which HTS is used because without careful consideration in the early stages of experimental design the analysis of the data generated could be seriously impacted.

However, it is by no means clear-cut as to what constitutes ‘best-practice’ across all projects. Indeed, a one-size fits all approach to experimental design is strongly discouraged in the same sense that a unified approach to data analysis is wholly inappropriate across the wide range of disciplines making use of HTS, each of which have their own set of issues compounding those associated with the use of HTS technology. Since the publication of this manuscript other reviews of HTS have also acknowledged that the considerations necessary when embarking on HTS studies and the mechanisms put in place to reduce untoward results are project aim dependent.

The use of HTS in ecological studies both past and present, despite the current difficulties associated with it, is proving to be a key tool by which to assess ecosystem change and explore ecosystem dynamics spatially and temporally. Provided the caveats presented in Chapters One–Five are acknowledged and where possible addressed, it is a powerful tool capable of generating fast and cost-effective genetic profiles across practically all environments. Regions rich in biodiversity, such as southwest Australia (Chapter Six), stand to gain from studies making use of HTS as it provides a means by which to easily obtain non-invasive environmental samples thus causing minimal impact to perhaps fragile ecosystems

Chapter Six – Using HTS to explore past plant and animal assemblages in a biodiversity hotspot

6.1 Preface

Chapter Six applies the methodologies and considerations developed in this thesis to embark upon an ambitious project to study past diversity in southwest Australia. The study involves isolating DNA from ~150 sediment samples and >6,000 bones with the aim of providing a detailed picture of past flora and fauna in the southwest including extinct and extirpated species. This study is presented as a “manuscript in preparation” and has been formatted in the style for Quaternary Science Reviews.

The previous chapters in this thesis have developed strategies to deal with, and highlighted considerations when using, highly degraded ancient and environmental DNA in ecological studies. Chapter Four also introduced the BBM methodology which has been further developed and applied in other publications (Haouchar *et al.*, 2013; Grealy *et al.*, 2015; Grealy *et al.*, 2016). Chapter Six applies the strategies and methodologies developed during this thesis while remaining cognisant of the issues associated with using ancient and degraded DNA to profile past environments.

In Chapter Six fragmentary bone and cave sediment is used to identify both animal and plant DNA, respectively, at five cave sites in southwest Australia. Four of the cave sites are of archaeological significance and collectively all five represent a combined history of the past 50,000 years in southwest Australia; a history set against the backdrop of episodic human occupation and environmental change.

6.2 Insights and challenges from combined palaeoecological reconstructions using fossils and sediment in southwest Australia.

6.2.1 Abstract

Environmental metabarcoding of ancient DNA in sediment and fossil bone is a promising approach but it has been largely confined to cool environments due to the poor preservation of DNA in warmer climates. Regions such as southwest Australia have therefore been largely overlooked in environmental metabarcoding studies to date. This is despite its recognition as a world biodiversity hotspot of conservation priority and the continued threat posed to its unique biota as a result of habitat loss, the spread of feral species and aridification. Both the conservation and restoration of biodiversity is best achieved with knowledge of the former composition and connectivity of flora and fauna. By metabarcoding ~150 sediment samples (*sedaDNA*) and ~6,000 bones (using bulk-bone) this study set out to explore the prospects and limitations of paleoenvironmental data derived from five cave sites in the Leeuwin-Naturaliste National Park (southwest Australia). Poor DNA preservation necessitated the use of short amplicon barcodes and this, coupled with a poor reference database, resulted in a number of challenges when generating and analysing data. Nonetheless, this exploratory study was able to detect changes in biodiversity that potentially relate to a change in forest habitat around the end of the Last Glacial Maximum. Together with other insights into the past 50,000 years of human habitation in the region, the data presented demonstrate that in tandem, *sedaDNA* and bulk-bone metabarcoding can act to complement previous and future archaeological and palaeontological studies into past biodiversity in warm, temperate environments.

6.2.2. Introduction

Faithful palaeoecological reconstructions play a key role in understanding both past and present biodiversity. By this means, changes in plant and animal diversity can be studied in light of ecological and environmental shifts, giving an insight into how species within a current ecosystem have responded in the past, and how they may

respond in the future to natural or anthropogenic-induced stresses (Leonard, 2008; Ramakrishnan & Hadly, 2009; Paplinska *et al.*, 2011; Terry *et al.*, 2011; Gavin *et al.*, 2014; Dietl *et al.*, 2015; Mann *et al.*, 2015). The morphological and molecular characterisation of fossil material has provided key insights into the evolutionary history of species (Donoghue *et al.*, 1989; Prideaux & Warburton, 2010; Shapiro & Hofreiter, 2014; Pacioni *et al.*, 2015), past human interactions with their surroundings (Campos *et al.*, 2010; Dortch & Wright, 2010; Lorenzen *et al.*, 2011; Golyeva & Andrič, 2014) and the impact of climate change on floral and faunal diversity, range extent and survival (Van Devender & Spaulding, 1979; Moody, 2005; Willerslev *et al.*, 2014; Mann *et al.*, 2015).

The molecular study of fossil material through the analysis of ancient DNA (aDNA) has proven to be a useful adjunct to traditional morphological studies; albeit provided contamination issues and authentication are carefully considered (Hofreiter *et al.*, 2001; Gilbert *et al.*, 2005; Willerslev & Cooper, 2005; Pedersen *et al.*, 2014). It is now also generally accepted (although see Birks & Birks, 2016) that palaeogenetics can add a new dimension to the study of fossils via the analysis of genetic diversity through time at a species, population and ecological (environmental metabarcoding) level (Shapiro *et al.*, 2004; Jørgensen *et al.*, 2011; Jørgensen *et al.*, 2012; Willerslev *et al.*, 2014; Pacioni *et al.*, 2015). However, there is still a host of challenges associated with aDNA metabarcoding studies using high-throughput DNA sequencing (HTS) technology. The burgeoning amount of data presents a number of challenges including those associated with DNA damage and degradation (Hofreiter *et al.*, 2001; Willerslev & Cooper, 2005; Dabney *et al.*, 2013), contamination at sample preparation and library generation (Champlot *et al.*, 2010; Hofreiter *et al.*, 2010; Boessenkool *et al.*, 2012; Pedersen *et al.*, 2014; Thomsen & Willerslev, 2015), low-template samples (Ficetola *et al.*, 2014; Pedersen *et al.*, 2014; Thomsen & Willerslev, 2015) and appropriate data filtering (Coissac *et al.*, 2012; Faircloth & Glenn, 2012; De Barba *et al.*, 2014; Philippe *et al.*, 2015). Nonetheless, certain environments have proven especially conducive to aDNA survival; namely high-latitude Arctic and Antarctic regions where a large number of aDNA projects have been undertaken (Willerslev *et al.*, 2007; Hebsgaard *et al.*, 2009; Sønstebo *et al.*, 2010; Jørgensen *et al.*, 2011; Willerslev *et al.*, 2014). However, aDNA studies in more temperate and hot regions, while challenging, have also been successful. This

is particularly true of studies using eggshell (Oskam *et al.*, 2010), hair (Rasmussen *et al.*, 2011) and midden material (Murray *et al.*, 2012). While such substrates have proven to be excellent sources of well-preserved aDNA, samples such as bone (Thomas *et al.*; Heupink *et al.*, 2011; Llamas *et al.*, 2015) and sediment (Haouchar *et al.*, 2013), although successful, have been more variable in their success. Generally, for both bone and sediment samples, those taken from cave systems tend to be more conducive to DNA preservation due to favourable conditions (Lindahl, 1993; Willerslev & Cooper, 2005; Leonard, 2008).

This study seeks to exploit such conditions to characterise aDNA extracted from fossil bone material and cave sediment obtained from archaeological sites in southwest Australia; a biodiversity hotspot of conservation priority (Myers *et al.*, 2000) where reference genetic databases are limited. First, a bulk-bone metabarcoding (BBM) approach (Murray *et al.*, 2013; Grealy *et al.*, 2015) is applied to determine the faunal diversity across sites and whether changes in this diversity can be detected using aDNA. Second, complementing the BBM approach, aDNA is extracted from cave sediment (*sedaDNA*) collected from the sites across multiple stratigraphical layers to explore floristic changes. This study sets out to explore if this two-pronged approach can add to the existing morphological record and enrich our understanding of plant and animal diversity against the backdrop of episodic human occupation over the past 50,000 years. The aim of the research is to explore the limits of DNA preservation in a temperate environment where there is a typically poor representation of species within the region on genetic reference databases.

6.2.3 Background to sites

Southwest Australia is a species-rich region with a moderate Mediterranean climate (Myers *et al.*, 2000; Hopper & Gioia, 2004). There is an extensive limestone cave network along the Leeuwin-Naturaliste Ridge of the region in which there are a number of archaeologically and culturally significant sites (Dortch, 2004b). The five cave sites within this network chosen for this study are all located at the intersection of three Interim Biogeographical Regionalisation of Australia (IBRA) regions in southwest Australia: Warren, Swan Coastal Plain and Jarrah Forest (Figure 6.2.1) (Thackway & Cresswell, 1995).

These caves provide an opportunity to study a combined ~50,000-year record of past biodiversity and Aboriginal occupation against the backdrop of major climatic shifts around the last glacial maximum (LGM) c.22-18, 000 BP (Dortch, 1979; Lilley, 1993; Dortch, 1996; Turney & Bird, 2001; Dortch, 2004b, 2004a; Dortch & Wright, 2010). Three of the chosen sites have been studied in the past using traditional morphological methods: Tunnel Cave (TC) (Dortch, 1996, 2004b, 2004a), Devil's Lair (DL) (Dortch, 1979; Turney & Bird, 2001; Dortch, 2004b) and Rainbow Cave (RC) (Lilley, 1993; Dortch, 2004b). Both DL and TC are located in the Warren biogeographical region while RC is a coastal site in the Swan Coastal Plain. The final two sites, Wonitji Janga (WJ) and Northcote Sinkhole (NS), are both located in the Warren and have only recently been described in scientific literature (Dortch *et al.*, 2014).

Devil's Lair (115° 04' E, 30° 09' S) covers the most extensive time period going back ~50,000 years and spans the LGM (Turney & Bird, 2001). In addition to this, DL shows signs of episodic human occupation between ~45, 000 BP to 12, 000 BP and is among the earliest sites of human occupation in Australia (Turney & Bird, 2001). The deposit at TC (115° 02' E, 34° 05' S) extends back to ~25, 000 BP and covers the LGM and the Holocene/Pleistocene transition in addition to a number of periods of human occupation interspersed with distinct periods of non-occupation (Dortch, 1996, 2004b). Rainbow Cave (114° 59' E, 33° 58' S), a Holocene site, covers a much shorter and rapid period of human occupation during the last 300-800 BP (Lilley, 1993; Dortch, 2004b). Wonitji janga (115° 02' E, 33° 39' S) has only recently been described (Dortch *et al.*, 2014) and covers ~100-1300 BP with human occupation evidenced throughout while NS covers a similar age range (380-2120 BP) but is devoid of any archaeological material to indicate past human occupation and is ~10 m south of WJ.

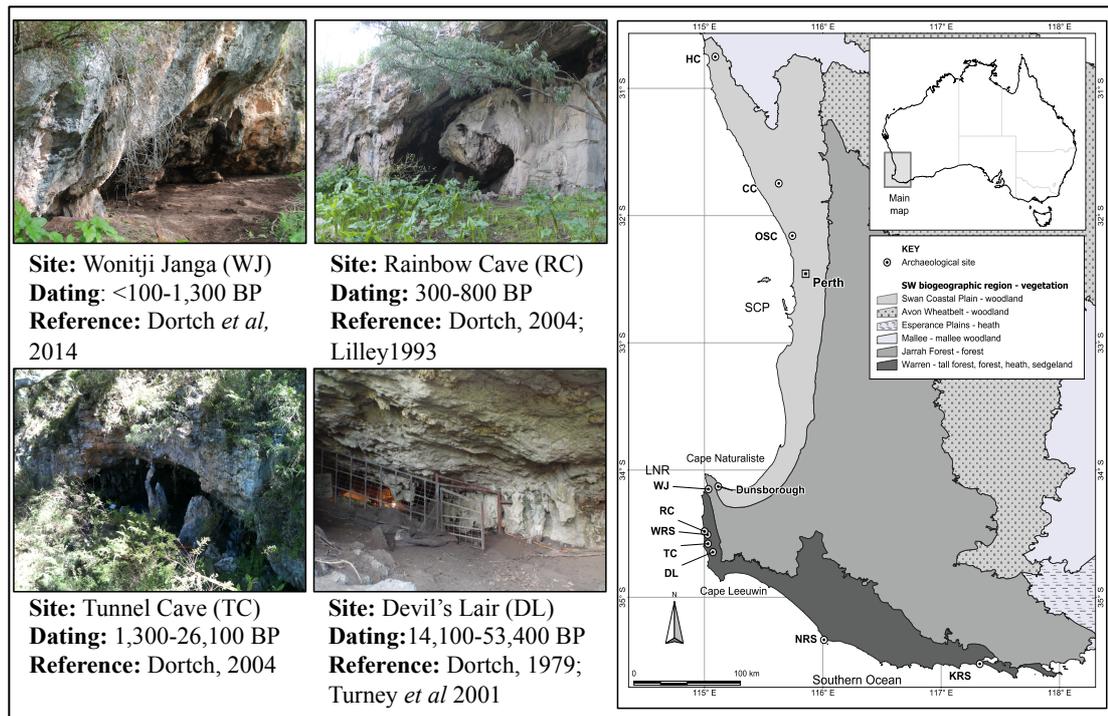


Figure 6.2.1 Location of southwest Australian cave sites used in this study. Four of the sites are show with additional meta-information associated with each. Northcote Sinkhole is not included but it is ~10 m south of Wonitji Janga.

6.2.4 Materials and Methods

Excavation and sampling of DL and TC deposits took place in February 2012 while that of RC, WJ and NS took place in March the following year in the presence of archaeologists and traditional representatives of the local Wardandi Noongar people. Additionally, the laboratory workflow described below took place across three labs with a one-way flow in place at Curtin University, Australia. All pre-PCR work was conducted in modules within a specialised clean-air facility. Sample preparation (pre-extraction) took place in a dedicated ‘sample preparation module’ that was separate to an ‘extraction module’ where DNA extraction took place. Post-extraction PCR set-up was conducted in a separate room within the extraction module. All post-PCR workflows were conducted in a laboratory in a separate building to that of both sample preparation and extraction. This workflow was designed to minimise sample contamination along various stages of the workflow as per routine aDNA guidelines (Gilbert *et al.*, 2005).

6.2.4.1 Sample collection, extraction and screening

Prior to all sampling and between all sampling and sub-sampling all surfaces and equipment were decontaminated with 10 % bleach followed by 70 % ethanol.

Bone sampling and extraction

At all sites fragmentary fossil bone (bulk-bone) was collected post sieving prior to sediment sampling with some minor differences in the bone collection between sites. The sampling of bone material at DL and TC has been described previously in Chapter Four (Murray *et al.*, 2013) and this was the method followed at NS. Briefly, all fragmentary bone at DL, TC and NS was collected from sediment sieved on-site using 2mm and 5mm sieves and bagged according to stratigraphical layers. At both RC and WJ fragmentary bone was sampled after trench excavations. At both sites, two of the four trench walls were pared back a further 5-10 cm after excavation and the sediment bagged according to stratigraphical units. The sediment was then sieved in the dedicated 'sample preparation module' at Curtin University and the bone sorted according to the method described in Chapter Four (Murray *et al.*, 2013).

For all sites, the fragmentary bone material was grouped according to stratigraphical units. A total of 92 bulk-bone samples (~6,000 bones total) were sampled across stratigraphical units (TC – 21, DL – 19, RC – 21, WJ – 17, NS – 14). All bones within each bulk-bone sample were sampled if the total was less than 50 bones. If more than 50 bones were present, samples were divided into batches of no more than three with each containing a maximum of 50 bones. Each bone was weighed and if found to be less than 20 mg the whole bone was included in subsequent grinding. Bones weighing over 20 mg were sub-sampled. For each bulk-bone sample or each batch where appropriate, the bone was ground into a fine powder in a stainless steel pot using a Retsch PM200 planetary ball mill at 2,000 rpm for a maximum of 5 minutes. A total of 100 mg of bulk-bone powder was sub-sampled in duplicate for each sample, or sample batch where appropriate. Each sub-sample was then extracted, with the inclusion of negative extraction controls, following the method described in Chapter Four (Murray *et al.*, 2013). Where possible, all sub-samples and sites were extracted in separate extraction batches.

Sediment sampling and extraction

At all sites sediment sampling was conducted after bone collection following strict aDNA protocols. At each site, after trench excavation was complete, prior to the day of sediment sampling, each trench wall was cleaned by means of removing a further >2.5 cm layer of sediment. At DL and TC the trenches were vacuumed prior to sampling in an attempt to remove as much loose sediment as possible around the trench shoring. On the day before sampling, at each site, a tarpaulin was erected and the sites were left dormant overnight to allow for the settling of dust and sediment. The following day only those involved in sediment sampling were permitted on site and a single person conducted all sediment coring under the guidance of an archaeologist: the only two people permitted under the tarpaulin in the trench during sampling. Once in the trench, all sampling was conducted without any breaks. Those involved in sampling were also required to wear a forensic body suit, gloves and a face mask. Prior to sampling, approximately 1 cm of sediment was removed using a sterile scalpel blade from the exact area to be sampled and a falcon tube used to obtain a sediment core. Samples were stored in a freezer at -25 °C within 6 hours of collection where they remained until transferred to laboratory freezer storage -25 °C.

Of the 248 samples collected across the five sites 149 were selected for DNA extraction (DL – 50, TC – 44, RC – 28, WJ – 23, NS – 5). A total of 5 g of sediment was used for each extraction and negative extraction controls were included. Samples were digested overnight (~15 hours) at 55 °C with rotation and extracts were cleaned the following day using a phenol/chloroform and silica extraction method detailed in File S6.2.1. Where possible, samples were extracted according to the site and in batches of no more than 12 samples.

6.2.4.2 Sample screening, amplicon generation and DNA sequencing

Bulk-bone and sediment extracts were screened using qPCR to enable the selection of samples with sufficient amplifiable DNA free of inhibition. For the purposes of this study, a maximum C_T value of 32 was set across all samples and any samples not meeting this requirement were not included in further analyses owing to the difficulties of achieving reliable taxonomic profiles from samples with low amounts of DNA (Murray *et al.*, 2015, reproduced as Chapter Five). All bulk-bone extracts were screened for faunal DNA using the 16Smam primer set (Taylor, 1996) designed

to amplify a small region within the 16S mitochondrial gene, using the protocol in Murray *et al.*, 2013 (Chapter Four). All sediment samples were screened using the *trnL* g/h plastid primer set that amplifies a short section of the *trnL* intron (Taberlet *et al.*, 1991; Taberlet *et al.*, 2007) using the protocol in Murray *et al.*, 2012 (Chapter Three).

Extracts deemed suitable for further processing based on the criteria set out above and cognisant of the issues raised in Murray *et al.*, 2015 (Chapter Five) were prepared for amplicon sequencing. For bulk-bone samples extracts, fusion-tagged amplicons were generated for 16Smam and 12SA/O primer sets (Taylor, 1996; Cooper *et al.*, 2001) at all sites using a unique forward and reverse DNA-based tag combination for each extract, with DNA extraction negative and negative PCR reagent controls included, as described in Murray *et al.*, 2012 (Chapter Three). An additional 12S primer set – 12SV5 (Riaz *et al.*, 2011) – was used for samples from RC due to the high proportion of fish bone at RC. The qPCR protocol for 12SV5 was as that for both 16Smam and 12SA/O with the exception of the annealing temperature which was set at 60 °C. All primers used in amplicon generation for bulk-bone samples while detecting mammalian DNA preferentially also detect a range of other groups to varying degrees. All bulk-bone extracts that amplified successfully were grouped together based on qPCR to allow extracts with similar C_T values, end-point and melt curves to be pooled together into a mini-pool prior to final pooling of amplicon mini-pools into a single sequencing library using a LabChip GX (PerkinElmer) following the manufacturer's instructions to create a sequencing library. The correct volume of sequencing library to use in sequencing was determined using the LabChip GX (PerkinElmer). Sequencing was carried out in-house on the Illumina MiSeq using 150 cycle V3 kits and single direction sequencing following the manufacturer's protocols.

In the case of sediment extracts a ligation method was adopted to create tagged amplicons following the protocol in Kozarewa & Turner, 2011. The method was adopted for the sediments because of a noticeable loss of sensitivity with the fusion-tagged method. All sediment extracts that passed initial screening were amplified using tagged *trnL* g/h primers and tagged *rbcl* h1aF/h2aR primers (IDT, Australia) using qPCR reaction components and conditions for both, with DNA extraction negative

and negative PCR reagent controls included, as described in Murray *et al.*, 2012 (Chapter Three). All extracts for both primer sets were assigned a unique forward and reverse DNA-based tag combination as for BBM described above. After amplicon generation, qPCR products were purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), according to the manufacturer's instructions and eluted in 40 μ L H₂O. Purified amplicon products were pooled into a single sequencing library following the protocol in Murray *et al.*, 2012 (Chapter Three). After pooling of amplicons in a sequencing library, Illumina MiSeq sequencing adaptors (IDT, Australia) were ligated onto the amplicon products following the protocol in Kozarewa & Turner, 2011. The post-ligation sequencing library was quantified as per the protocol outlined in Murray *et al.*, 2015 (Chapter Five) to determine the appropriate amount for sequencing on the Illumina MiSeq using 150 cycle V3 kits and single direction sequencing following the manufacturer's protocols.

6.2.4.3 Data analysis

Amplicon sequences for both bulk-bone and sediment samples were downloaded in FastQ format from the MiSeq. Amplicons were initially separated based on unique forward tags in Geneious v8.1.4 (Drummond *et al.*, 2011). After initial separation based on forward tag all tags were exported in FastQ format for further deconvolution based on the reverse tag using the ShortRead and Biostrings packages in the statistical program R (Pages *et al.*; R Development Core Team, 2008; Morgan *et al.*, 2009). Without exception, only perfect matches in base composition and length were accepted for both forward and reverse tags and primers. Any sequences that were found to have any tag other than the assigned tag were removed from analyses. Sequences found to have more than two primers, the incorrect combination of primers or both primers in the same orientation were also discarded. After deconvolution FastQ files were filtered to remove all sequences that had any bases below Q15 and any sequences whose average quality was below Q20.

Once deconvolution and quality filtering were complete, sequences were filtered to remove potential contamination, PCR artefacts and sequencing error. To do this sequences were dereplicated using USEARCH v8 (Edgar, 2004) to create unique sequence clusters with abundances appended to each sequence cluster name. This was done for each sample individually and any unique sequence cluster within a

sample that fell below 1% of the total number of unique sequences within that sample was discarded. Once this filtering was complete the remaining sequences were searched against the NCBI GenBank nucleotide database (Benson *et al.*, 2006) using BLASTn v2.2.3 (Altschul *et al.*, 1990) to enable the identification of sequences. Sequences were searched without a low complexity filter, with a gap penalties existence of five and extension of two, expected alignment value less than 1e-10 and a word count of seven. Additionally, BLASTn hits below 90 % similarity and those below 90 % query coverage were not considered. BLAST results were then parsed in METaGenome Aalyzer v5.11 (Huson *et al.*, 2007) as described previously in Murray *et al.*, 2013 (Chapter Four). Sequence matches were then assessed based on percentage similarity requiring a >90 % match for family, >97 % match for genus and 100 % match for species. In addition to this, both Atlas of Living Australia (<http://www.ala.org.au>) and Florabase (<http://www.florabase.dpaw.wa.gov.au>) were used to determine whether or not identified taxa occurred locally (sites accessed in June 2016). These percentage cut-offs, while somewhat arbitrary have empirically worked well at distinguishing various taxonomic levels and, at present, offer the best solution to a number of issues associated with taxonomic assignment based on percentage similarity and bit scores (see Chapter Seven for further discussion).

In addition to the above, an OTU analysis was conducted on the *trnL* g/h sediment dataset and the 16Smam bulk-bone dataset – other primer datasets were not considered due to a large drop-off in successful samples. Due to the unequal representation of some taxa within samples which can lead to the skewing of OTU analyses in the direction of either low or high abundant taxa, OTUs were not calculated with all taxa grouped into one. Rather, after deconvolution and quality filtering sequences were searched against the NCBI Genbank database, as described above. BLASTn results were then parsed in MEGAN, as described above, and collapsed at specific taxonomic levels. In the case of the bulk-bone samples sequences matching to eight Metatherian families, Muridae and higher ranking taxa whose sequence abundances were not very high (Aves, Amphibia, Squamata and Actinopterygii) were extracted and OTU analysis was conducted on the sequences within each family independently of the other. In the case of the sediment samples, a selection of 18 abundant families were chosen that included archaeologically important plant families and the most abundant plant families in southwest Australia

(Dortch, 2004b). The remaining steps involved in the OTU analysis pipeline were kept consistent across all plant and animal families selected. Sequences were dereplicated to produce clusters of identical unique sequences with the abundance appended to sequence names. The unique clusters within each family were subjected to an initial filter based on abundance on a per sample basis to remove any error associated with PCR and potential low-level contamination. Both a strict and relaxed abundance cut-off was selected across the board. For the strict analysis, any unique sequence clusters below 1 % of the total number of sequences were deleted which has the potential to skew OTUs towards more abundant taxa. For the relaxed analysis, unique sequence clusters below 1 % of the total number of unique sequences were deleted. This has the potential to capture OTUs that may otherwise be discarded using the stringent filter that may be genuine while at the same time allowing some error to creep in thus inflating OTUs. This approach was adopted to strike a balance between discarding error and damage while still acknowledging the fact that some *bona fide* OTUs are indeed in low abundance. Lastly, after the two-step filtering process unique sequences were clustered into OTUs using the UPARSE algorithm (Edgar, 2013) in USEARCH v8 (Edgar, 2010). For clarity, a summary of both stringent and relaxed data processing workflows are presented in Figure S6.2.1.

6.2.5 Results and Discussion

Approximately 6,000 bones and approximately 150 sediments constituted the samples in this aDNA survey of cave sites in southwest Australia. This dataset represents a sizable temporal and spatial survey for the region and, to date, the largest aDNA study conducted in Australia, to the author's knowledge.

The overall success rates of bulk-bone metabarcoding (BBM) across sites and faunal primer sets were variable. The more recent sites proved most successful with RC and NS having 100 % sample and replicate success rates for 16Smam and WJ having a sample success rate of 83 % and replicate success rate of 89 % - hereafter replicate success rate is given in brackets after sample success rate. At both DL and TC, success rates remained high at 79 % (84 %) and 94 % (75 %), respectively. However, this may be a false estimate of DNA preservation at the sites as, empirically, the 16Smam primer sets amplify human contaminants more readily than that of 12S/O.

Indeed, 12S/O success rates dropped dramatically at some sites, e.g. TC – 33 % (25 %) and WJ – 61 % (57 %). Furthermore, the sole contaminants observed in extraction and PCR negative controls for BBM were that of human, while for sediment samples it was that for pine, both commonly encountered contaminants (Boessenkool *et al.*, 2012; Pedersen *et al.*, 2014). Such a scenario proves problematic when screening samples for DNA amount using qPCR, or indeed any method of DNA quantitation. Future studies may be required to test existing, or to develop new, blocking primers for human sequences (Vestheim & Jarman, 2008; Boessenkool *et al.*, 2012), or others, as ensuring data fidelity through accurate quantitation is essential in ancient and environmental studies (Murray *et al.*, 2015, reproduced as Chapter Five). Future developments in sample quantitation, such as digital PCR (Hindson *et al.*, 2011), and HTS though may overcome the necessity for such a blocking strategy (discussed further in thesis Chapter Seven).

Similarly to BBM, when sediments were screened with plant primer sets success rates were quite variable and a marked decrease in success was observed when using the longer *rbcl h1aF/h2aR* primer set when compared to the shorter *trnL g/h* primer set. Interestingly, at NS only a single sample worked for *trnL g/h* assays while none worked for *rbcl h1aF/h2aR* despite the promising results from primer sets used in BBM. Perhaps this may be due to the fairly closed nature of the sinkhole and lack of accessibility. Alternatively, it could be due to soil characteristics and chemistry. The sediment sampled at NS was sand-like in nature which has been shown to bind DNA less effectively than clay-like sediment (Lorenz & Wackernagel, 1987). It has been speculated that a clay-rich soil may result in better *sedaDNA* preservation as the binding of DNA to clay particles offers some protection from degradation (Blum *et al.*, 1997; Crecchio & Stotzky, 1998; Pietramellara *et al.*, 2007; Huang *et al.*, 2014). At present, despite studies positing a local origin (Jørgensen *et al.*, 2012), little is known regarding the provenance of plant *sedaDNA* or the means by which it survives. With such a large gap in our understanding, it is difficult to reason why NS would have yielded such uniformly poor results.

6.2.5.1 Taxonomic insights from BBM and *sedaDNA*

There are numerous challenges associated with identifying ancient DNA sequences from material stretching back almost 50, 000 years (explored in Chapter One). The

retrieval of ancient DNA from specimens of this age in Australia must be treated with caution, however, every effort was made to adhere to aDNA protocols (Gilbert *et al.*, 2005). A cautious interpretation of the results is presented where an overview of the family level assignment for faunal taxa is shown (Figure 6.2.2) and, where possible, genus and species level identification are provided across layers and discussed further (Table S6.2.1A-E & S6.2.2). For *seadaDNA* analysis a brief overview of plant family assignments is given (Table 6.2.1A-D) and taxonomic assignments are explored further in the supplementary information (Table S6.2.3A-D & S6.2.4), however owing to difficulties associated with the genetic assignment of plant taxa the major focus of the analysis for *seadaDNA* will be an OTU-based analysis. In light of this, the oldest successful bulk-bone sample was from layer 39 within Period I which is dated to between 43, 000-51, 600 years BP at DL (Turney & Bird, 2001; Dortch, 2004b). Additionally, it appears that samples such as this might be a rarity in Australia as from four samples taken within this period only a single sample was successful for both replicates while another two were successful for a single replicate. The oldest sediment sample that proved successful was from layer 23 within Period II at DL, however, this layer is near the top of the period and despite a date between 28, 400-44, 800 years BP for the period, the age is likely closer to that of the upper limit of c.30, 000 years BP (Turney & Bird, 2001; Dortch, 2004b).

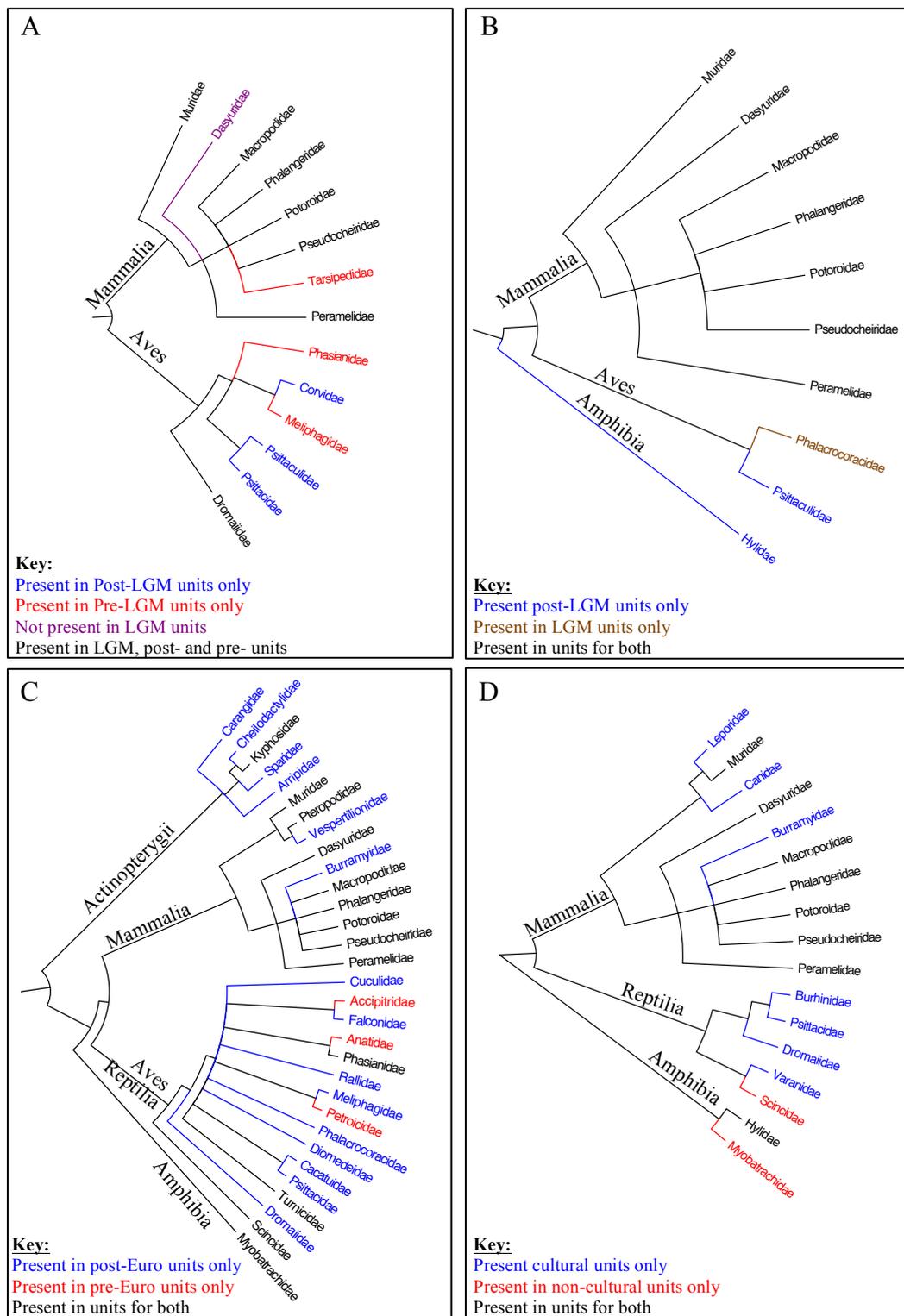


Figure 6.2.2 Cladograms showing faunal diversity identified across Devil's Lair, Tunnel Cave, Rainbow Cave and Wonitji Janga. Family level assignments at each site are presented. Devil's Lair (A) is colour coded according to pre-LGM, LGM and post-LGM layers. Tunnel Cave (B) is coloured according to LGM and post-LGM layers. Rainbow Cave (C) is coloured according to suspected cultural and pre-cultural layers. Wonitji Janga (D) is coloured according to layers pre-European arrival and post-European arrival. For possible assignments to genera and species refer to S6.2.1A-E.

Table 6.2.1A Presence and absence of plant families detected at DL pre-LGM, during the LGM and post LGM (Dortch 2001).

Order	Family	Post-LGM	LGM	Pre-LGM
Asparagales	Asparagaceae			
Asterales	Asteraceae			
Caryophyllales	Amaranthaceae			
Fabales	Fabaceae			
Laurales	Lauraceae			
Malpighiales	Euphorbiaceae			
Malvales	Malvaceae			
Myrtales	Myrtaceae			
Poales	Poaceae			
Proteales	Proteaceae			
Solanales	Solanaceae			

Families in **BOLD** indicate non-resource families in high abundance in Western Australia.

Table 6.2.1C Presence and absence of plant families detected at RC in layers signs of cultural material and no cultural material (Dortch 2001).

Order	Family	Cultural	Not Cultural
Apiales	Apiaceae		
Araucariales	Podocarpaceae		
Asparagales	Xanthorrhoeaceae		
Asterales	Asteraceae		
Asterales	Goodeniaceae		
Caryophyllales	Amaranthaceae		
Cycadales	Zamiaceae		
Ericales	Ericaceae		
Fabales	Fabaceae		
Malvales	Thymelaeaceae		
Myrtales	Myrtaceae		
Poales	Cyperaceae		
Poales	Poaceae		
Poales	Restionaceae		
Polypodiales	Dennstaedtiaceae		
Proteales	Proteaceae		
Ranunculales	Ranunculaceae		
Rosales	Rhamnaceae		
Santalales	Santalaceae		
Sapindales	Rutaceae		
Solanales	Solanaceae		

Families in **BOLD** indicate non-resource families in high abundance in Western Australia.

Table 6.2.1B Presence and absence of plant families detected at TC during the LGM, post LGM and in hearth features (Dortch 2001).

Order	Family	Post-LGM	LGM	Hearth Features
Apiales	Apiaceae			
Asterales	Asteraceae			
Caryophyllales	Amaranthaceae			
Fabales	Fabaceae			
Myrtales	Myrtaceae			
Poales	Poaceae			
Proteales	Proteaceae			
Rosales	Rhamnaceae			
Solanales	Solanaceae			

Families in **BOLD** indicate non-resource families in high abundance in Western Australia.

Table 6.2.1D Presence and absence of plant families detected at WJ pre- and post-European arrival (Dortch *et al.* 2014).

Order	Family	Post-European	Pre-European
Apiales	Apiaceae		
Araucariales	Podocarpaceae		
Asparagales	Xanthorrhoeaceae		
Asterales	Asteraceae		
Asterales	Goodeniaceae		
Caryophyllales	Amaranthaceae		
Cycadales	Zamiaceae		
Ericales	Ericaceae		
Fabales	Fabaceae		
Fagales	Casuarinaceae		
Myrtales	Myrtaceae		
Poales	Cyperaceae		
Poales	Poaceae		
Poales	Restionaceae		
Polypodiales	Dennstaedtiaceae		
Proteales	Proteaceae		
Ranunculales	Ranunculaceae		
Rosales	Rhamnaceae		
Sapindales	Rutaceae		
Solanales	Solanaceae		

Families in **BOLD** indicate non-resource families in high abundance in Western Australia.

Taxonomic identification of bulk-bone material

The bulk-bone metabarcoding method employed was able to identify a range of taxa across numerous families at the five study sites including amphibians, avifauna, mammals and fish (Figure 6.2.2, Table S6.2.1 & S6.2.2). The morphological identification of taxa such as amphibians and fish can be a difficult task requiring expert knowledge and as such is an aspect of site analysis to which BBM can greatly add. Moreover, taxa such as amphibians can be very sensitive to environmental change.

At the non-archaeological NS site, amphibians were detected throughout the deposit, and these included the endemic *Pseudophryne guentheri*, *Limnodynastes dorsalis* and unidentified *Litoria*. Amphibians were not detected at any other site except in a single layer WJ. It is unclear as to why they should be absent from the other deposits considering that amphibians, alongside reptiles, represent easy food resources that people of all abilities could hunt. This absence may be due, in part, to the abundance of mammalian bones recovered from the other sites compared to those of other taxonomic groups and a possible preferential amplification of mammalian taxa. The use of a primer assay targeted at, for example, amphibians may prove to be a solution to this issue. Nonetheless, the ability to identify at least some amphibians and reptiles from NS bones well for future research into habitat loss, range contraction and extinction of such taxa. The ability to identify amphibians and reptiles is especially relevant in southwest Australia where land clearances since European arrival (Beard, 1995; Bradshaw, 2012), causing the potential loss of many species of amphibians and reptiles, has made it difficult to assess the truth of reports suggesting extensive historical diversity in these groups (How et al., 1987; Dortch, 2004b).

Identification of fish bones is important archaeologically as it has the potential to give insight into the seasonal use of past sites, historic fishing practices and is a clear indication of human occupation at sites such as RC. It was possible to identify several, primarily estuarine or shallow reef, fish genera at RC. This finding is supported by the fact that fish not close to shore were inaccessible to the Nyoongar peoples (Dortch, 2004b). Genera *Kyposus* and *Girella* (both sea chubs) were detected using both 12SA/O and 12SV5, while *Kyposus* was also detected using 16Smam;

though none of these primers could distinguish them further. *Kyposus sydneyanus* (silver drummer) and *K. cornelii* (Western buffalo bream) could both be possibilities, although *K. sydneyanus* may be most likely as it has previously been identified as a food resource at sites in the region. In southwest Australia there are two species of *Girella*: *G. zebra* (zebrafish) and *G. tephraeops* (western rock blackfish). The former can be found in estuaries, bays and coastal reefs while the latter usually inhabits shallow coastal rock reefs or headlands. Two fish were also identified that spawn in the southern hemisphere autumnal or winter seasons: *Rhabdosargus sarba* (silver bream) and *Arripis* (Australian salmon). Although it was not possible to identify *Arripis* to species only two are recorded around the southwest coast close to RC: *A. georgianus* (Australian herring) and *A. truttaceus* (Western Australian salmon). Finally, the genus *Cheilodactylus* (morwongs) was detected, of which *C. rubrolabiatus* (redlip morwong), a WA endemic, and *C. gibbosus* (Western crested morwong) could be possibilities although the former is a rare occurrence in the region. A more likely candidate in this instance may be *Dactylophora nigricans* which is within the same family, has been recognised as a food source in the region in the past and for which no sequences for either 16S or 12S genes exist on Genbank at present. It has been determined from previous research (Lilley, 1993; Dortch, 2004b) that the extent of cultural stratigraphical units (i.e. those in which cultural material is found) is from stratigraphical units 001-004, with some limited evidence of cultural material in unit 005. However, in this BBM analysis DNA from fish was detected in units 015 and 016 but in no others between these units and 005. While the possibility of vertical movement of bone material cannot be ruled out there appeared to be no disturbance of the stratigraphical units at the site. DNA leaching may also be unlikely as no fish DNA was detected in the units between these and the recognised cultural units. Previous research assessing DNA movement detected the presence of the “leached” DNA in all samples collected down until the last sample in which the “leached” DNA was detected (Andersen *et al.*, 2012). Further work in analysing the individual bones from these units is currently on-going.

Avian taxa, as with fish and amphibians, are quite difficult to morphologically assign at archaeological deposits and are often fragile (Dortch, 2004b). Although we did not specifically target birds with 12SA/O it, along with 12SV5, do detect some avian taxa. Using the BBM method it was possible to identify a range of birds, the majority

of which were detected at RC in stratigraphical units 004-006. It is unclear why a large number of the birds identified at the site would occur in these layers alone, however, a similar grouping of bird bones was found near the middle of the excavation in previous morphological analyses (Dortch, 1996, 2004b). Many of the avian taxa identified across these layers were detected using both 12SV5 and 12SA/O. Taxa identified for birds include those known to have been targeted for hunting such as Anatidae and Accipitridae whose eggs were eaten as well as Cacatuidae (genus: *Calyptorhynchus* – black cockatoo) and Phasianidae (Dortch, 2004b). A noticeable absence from the genetic identifications was *Dromaius novaehollandiae* (emu) for which numerous eggshell was found at several of the archaeological sites (Dortch, 2004b; Dortch *et al.*, 2014). It was only found in a single layer at WJ. A simple explanation for this may lie in the fact that, when sampling, eggshell was actively not sampled due to a lack of optimisation of the BBM extraction protocol for eggshell samples. Unfortunately, for many of the bird families detected that have been identified as being historical resources, it was not possible to identify them past a family or genus level.

The issues associated with an inability to assign taxa past a family or genus level is a major drawback at present in environmental metabarcoding and it has not been possible to conduct finer scale studies of patterns across sites and taxa (discussed further in Chapter Seven). The lack of resolution is particularly true for members of the Muridae family and some Macropodidae species and this is due to a combination of poor database coverage for some taxa such as native mice and rodents which are severely underrepresented for 16S and 12S on Genbank (discussed in Chapter Three and Chapter Seven) and an uncertain taxonomic framework. Assignment problems are further exacerbated by the fact that with extinction and expiration many of the current day reference barcodes will not match what was there in the past. For these reasons, it was not possible for robust rodent assignment past a family level with the exception of *Rattus fuscipes* (bush rat) which was identified in many layers across sites. For most mammalian taxa, however, a genus level identification was possible due to the presence of only a single species known to occur or to have occurred in southwest Australia. Such taxa include *Tarsipes rostratus* (honey possum) which was exclusively found in pre-Last Glacial Maximum (LGM) Periods at Devil's Lair in keeping with previous findings at the site suggesting a decline in “non-forest”

animals post-LGM (Dortch, 2004b). Conversely, Dasyuridae genera (*Antechinus*, *Dasyurus* – quoll – and *Phascogale* – wambengers) appear to occur more frequently in post-LGM Periods. A somewhat surprising discovery was DNA identified to the Dasyurid *Sarcophilus harrisii* with 100% similarity in layers that are after the time it is meant to have been extirpated from the area. While uncertainty surrounds the exact timing of the extirpation of *S. harrisii* from mainland Australia it is put at somewhere around 3, 000-years ago (Brown, 2006; Hunter *et al.*, 2015). However, a single *S. harrisii* tooth was discovered in a cave site near Augusta and dated to 430 ±160 years BP (Archer & Baynes, 1972). Potential *S. harrisii* DNA was detected at both WJ and RC. In the WJ deposit, it was found in stratigraphical unit 003 and stratigraphical unit 010 for 16Smamm while for 12SA/O it was found in stratigraphical unit 003. Stratigraphical unit 003 and 010 have been dated to between 450-540 calibrated years BP and 1180-1280 calibrated years BP, respectively. At RC, potential *S. harrisii* DNA was detected in stratigraphical units 005 for 12SA/O and units 005 and 006 for 16Smam, units which have been dated to between 300-800 calibrated years BP. In all but one replicate, *S. harrisii* was the sole Dasyuridae detected within the sample thus reducing the likelihood that such DNA sequences could be the result of error arising from genetically similar taxa within the same family. While the discovery of the single tooth has proven controversial (Brown, 2006) the regular occurrence of *S. harrisii* in late Holocene layers adds weight to the notion of late-surviving taxa in the southwest forests. This finding demonstrates the ability of BBM to generate hypotheses that require further testing.

The difficulty in assigning some mammalian taxa to a species level proved somewhat problematic when addressing some questions regarding the changing ecology of the cave sites such as attempting to detect the decline of non-forest mammals (e.g. *Bettongia lesueur*) and an increase in the occurrence of forest mammals (e.g. *B. penicillata*) post-LGM (Dortch, 2004b). While both were detected at some sites it is difficult to definitively state the detection of one or the other due to the high similarity in sequences between both species for 16Smam. Indeed, while both were detected for 16S across layers only *B. penicillata* was identified using 12SA/O despite full mitochondrial genomes being available for both on Genbank. In the case of 16Smam, while *B. penicillata* occurred in samples without *B. lesueur* the opposite was not true, suggesting a greater potential for one to be erroneous in instances where

B. lesueur is less abundant or *vice versa*. Similar difficulties arose when attempting to identify *Macropus* species. Of the three *Macropus* species identified at the sites, *M. fuliginosus* is of most interest as it is believed to have been preferentially targeted by people. For 16Smam *M. fuliginosus* is identical to *M. rufogriseus*, however, the latter is not found in the area and has never been recorded at any archaeological sites in the area to the best of the author's knowledge. At TC *M. fuliginosus* was identified without fail in all hearth samples that were successful. It was also detected in all occupation stratigraphical units with a large number of sequences compared to other taxa detected but it was decidedly absent in stratigraphical units 003 and 008, both of which show no evidence of occupation, and in quite low sequence numbers for 5-lower and 006 (limited occupation). While this is not being stated as a definitive result it is supported by previous studies (Dortch, 1996, 2004b; Dortch & Wright, 2010) and may warrant further work to increase sample numbers around occupation and non-occupation units to allow some future statistical tests of ubiquity or frequency of occurrence. Future studies using primers that are specific to *Bettongia* species or *Macropus* species may help to elucidate these trends further and perhaps even have the potential to reveal genetic changes at a population level in the case of *Bettongia* (Pacioni *et al.*, 2015).

Taxonomic identification of *seda*DNA

This study represents the first foray into plant aDNA identifications at an Australian archaeological site. Despite the associated challenges this is significant because, currently, little information can be gained on flora subsistence practice using morphological methods at the sites and others in Australia. Together with low DNA yields, one of the biggest challenges in this particular study was the current genetic databases (and underpinning taxonomic frameworks) which made it inherently difficult to assign a range of taxa to species level. Database issues are a recurring problem with many studies dealing with plant identifications but are even more pronounced when using short DNA barcodes (Chapters Three, Four and Seven). A major drawback in the use of environmental barcoding and *seda*DNA in southwest Australia is the fact that it is a region of high biodiversity in relation to plants, most of which is uncharacterised genetically. There are over 700 genera (13 % endemic) and more than 5,710 species (52.5 % endemic) (Paczkowska & Chapman, 2000); in

the Warren biogeographic region alone there are over 2,500 species. This study did not set out to identify plants beyond a family level necessarily. In the case of *rbcl h1af/h2ar* it was not possible to identify any plants beyond a family level while genus level identification using *trnL g/h* with aDNA can be problematic due to DNA damage and sequencing error in such a short amplicon. To remain conservative, plants were considered at a family level and full explanations are provided for each family assignment and the likelihood of it being an artefactual sequence (i.e. contamination, damage or error) is provided in the supplementary information (Table S6.2.3A-D & S6.2.14). It was possible to detect a number of plant families that are known to contain specific resources used in the past by Noongar communities (Dortch, 2004b) and these include Podocarpaceae (of which only one species occurs in the area *Podocarpus drouynianus*) and Zamiaceae (of which only *Macrozamia riedlei* is found in the region). This bodes well for future studies using *sedaDNA* in southwest Australia as it indicates that it is possible to extract DNA from cave sediment, however, the ~150 samples examined here demonstrate that preservation is highly variable and high fidelity taxonomic identifications are difficult. It may be advisable in future studies to use universal primers such as *trnL g/h* and *rbcl h1aF/h2aR* to assess the likelihood of retrieving DNA and to identify what plant families are present and then armed with this *a priori* knowledge proceed to design specific primers to target genera or species of interest in samples for which the chosen families were detected. This *a priori* approach may enable more effective use of existing barcoding regions such as *matK* and the internal transcribed spacer region (ITS) – typically the length of amplicons required make them unsuitable for most *sedaDNA* studies employing universal primers. Such an approach may, in future, allow a closer inspection of known changes around the LGM boundaries such as the transition between different *Eucalyptus* species (Dortch, 2004b) and the associated changes in under-canopy genera such as *Acacia*.

For the reasons outlined it was decided from the outset of this study to attempt a measure of within family diversity using an OTU-based approach for both plants and animals; however this method too is limited to some degree as key species and genera for both *rbcl h1aF/h2aR* and *trnL g/h* are identical for the gene regions targeted (e.g. *Eucalyptus* species).

6.2.5.2 OTU analysis of bulk-bone and *sedaDNA*

Due to the limitations of species-level assignments across a range of important taxa an OTU approach was adopted in an attempt to determine if there were any changes in biodiversity at the sites through time (Figure 6.2.3, Figure 6.2.4). The stringent and relaxed OTU approaches adopted showed strong, statistically significant correlations in the overall number of OTUs detected in bulk-bone samples across layers at each site (Spearman's rho for WJ $r_s = .91$, $p < 0.01$; DL $r_s = .83$, $p < 0.05$; TC $r_s = .91$, $p < 0.01$; NS $r_s = .95$, $p < 0.05$; RC $r_s = .88$, $p < 0.01$). As expected the total number of OTUs obtained using the relaxed OTU approach was considerably higher than that for the stringent approach – the strict approach was favoured as a conservative first foray into BBM and *sedaDNA* datasets from Australia. In addition to this, despite adopting the stringent approach to present the data it was not possible to assess any correlation between the amount of sediment sieved and the number of bones sampled from stratigraphical units at any site other than WJ where sampled sediment weights per unit were recorded. With that said, at WJ bucket weight did not have a statistically significant correlation with the number of bones ($r_s = .75$, $p > 0.05$). In culmination caution is urged when assessing any patterns in OTU diversity through time for the bulk-bone data.

With regard to the *sedaDNA* analysis of sediment samples, a consistent weight of samples was used in each case for each sample, however, the number of successful samples across the different stratigraphical units at each site differed. To test the impact of unequal numbers of successful samples for each layer, the total amount of sediment was calculated for each layer by summing the amount extracted for each sample within a particular layer. No correspondence was detected between the weight of sediment sampled per layer and the number of OTUs at WJ ($r_s = .5$, $p > 0.05$) or RC ($r_s = .2$, $p > 0.05$). At TC there was no statistically significant correlation between OTUs and weight of sediment sampled for each non-hearth layers ($r_s = .56$, $p > 0.05$), although there was for hearth features ($r_s = .83$, $p < 0.05$). At DL there was a statistically significant correlation between OTU number and the weight of sediment sampled per period ($r_s = .64$, $p < 0.05$) and this is likely due to the large number of samples that were successful for Period VII (nine samples) compared to the other periods which were either one or two samples each – significance disappears when Period VII is removed from the dataset ($r_s = .52$, $p > 0.05$). Therefore, caution is urged when interpreting the plant data presented for DL.

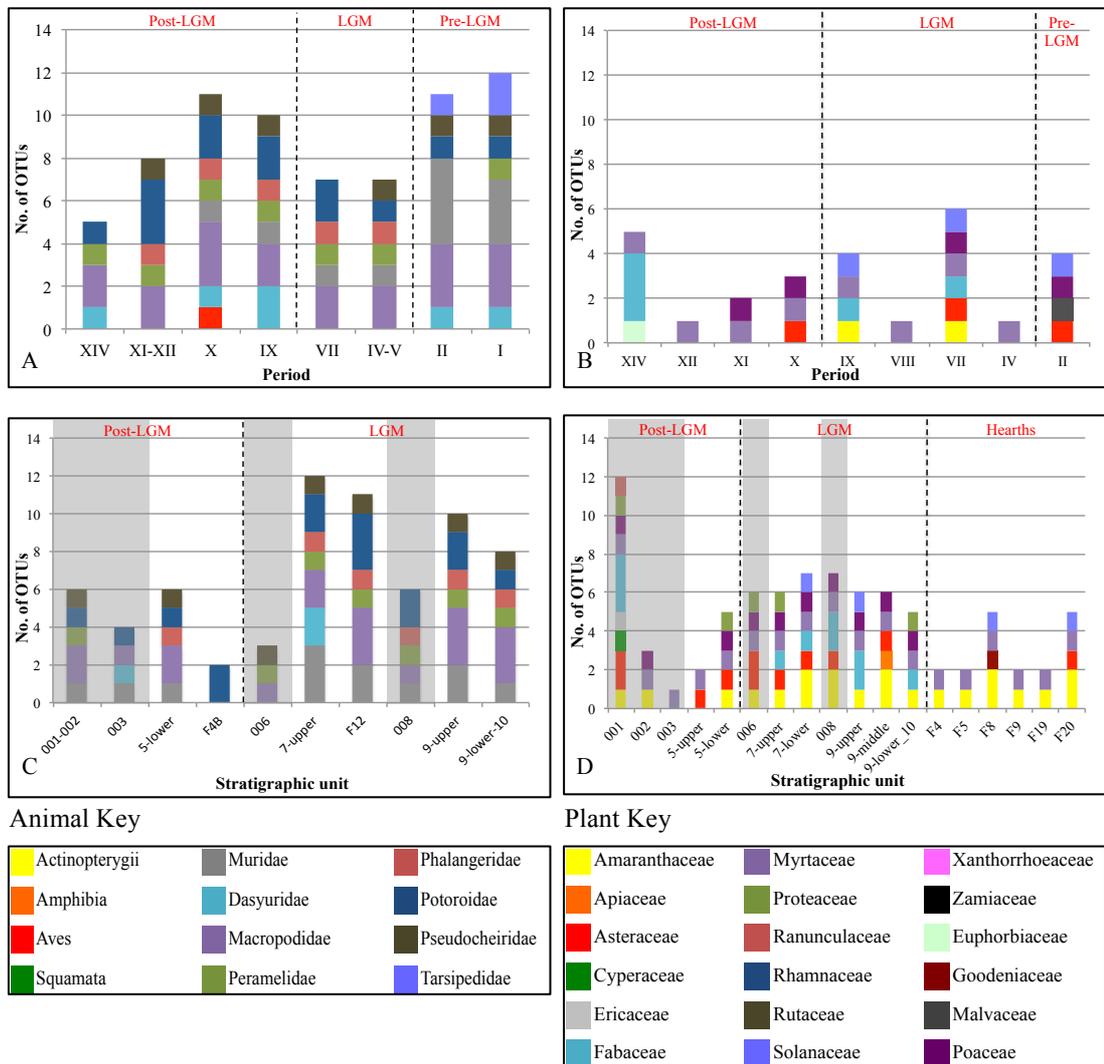


Figure 6.2.3 OTU number and diversity change over time at Devil's Lair and Tunnel Cave. The change in the number of animal (A) and plant (B) OTUs is shown for Devil's Lair with dashed lines showing the boundaries between pre-LGM, LGM and post-LGM. The change in the number of animal (C) and plant (D) OTUs at Tunnel Cave is shown with dashed lines indicating the boundaries between pre-LGM and post-LGM. Shaded rectangles in C and D represent periods of non-occupation at Tunnel Cave (Dortch 2001).

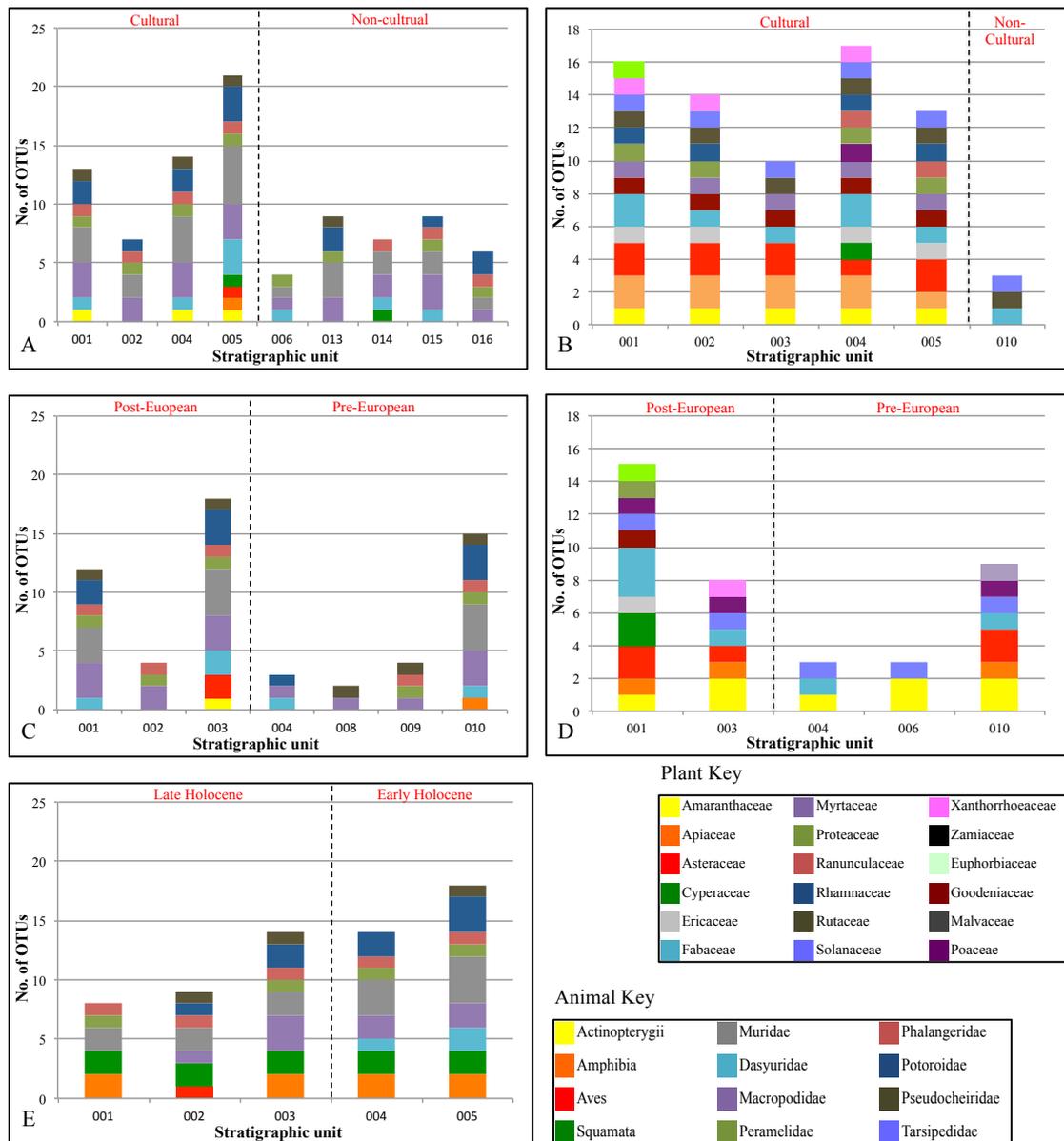


Figure 6.2.4 OTU number and diversity change over time at Rainbow Cave, Wonitji Janga and Northcote Sinkhole. The change in the number of animal (A) and plant (B) OTUs is shown for RC with the dashed line indicating the boundary between suspected cultural and non-cultural layers. The change in the number of animal (C) and plant (D) OTUs is shown for Wonitji Janga with the dashed line indicating the boundary between pre-European and post-European arrival. The change in the number of animal (E) OTUs only is shown from Northcote Sinkhole with the dashed line indicating the boundary between early and late Holocene layers.

Devil's Lair and Tunnel Cave OTU diversity

At both DL and TC there were some notable shifts in the diversity and number of OTUs through time observed (Figure 6.2.3A & B and Figure 6.2.3C & D, respectively). At DL there is a clear distinction between pre-LGM and LGM/post-LGM when looking at OTU composition (Figure 6.2.3A & B) and which is also identified through OTU clustering (Figure 6.2.5A & B). In addition to the previously identified and supported disappearance of Tarsipedidae with the onset of the LGM, three of four Muridae OTUs identified at DL are only detected in Periods I, II and IV-V. The loss of Muridae OTUs may be as a result of owl predation that has been documented throughout Period I and II at the site but that became quite reduced above layer 18 within Period III (Balme *et al.*, 1978; Dortch, 1979; Dortch, 2004b). Detecting a possible change in Muridae OTU diversity highlights the potential of BBM in identifying small fragmentary bone possibly resulting from owl deposits which can be difficult and time-consuming. Indeed, distinguishing taxa found in owl deposits from other taxa was not initially attempted at DL (Balme *et al.*, 1978; Dortch, 2004b). However, it is noted that no avian DNA was detected in these periods but a future avian-targeted PCR assay may remedy this.

While a difference was observed between pre-LGM and the onset of the LGM, it was not possible to differentiate between post-LGM and LGM at DL as both the composition of OTUs and the diversity within family level OTUs seems to remain fairly constant. However, the inability to detect a shift from post-LGM to LGM at DL may not be unusual as the occupation at DL is unclear and not well-defined unlike at TC. However, there are signs of occupation throughout DL from circa Period II through to Period XI-XII (Dortch, 1979; Dortch, 2004b). It has been stated previously that DL may provide little information on human reactions to changing vegetation around the Pleistocene/Holocene boundary (Dortch, 2004b). This study also suggests that it is difficult to tease out changes in OTU diversity and composition as a result of changing climate due to continued human occupation at DL. Both issues in tandem seem to mask each other due to natural 'noise.'

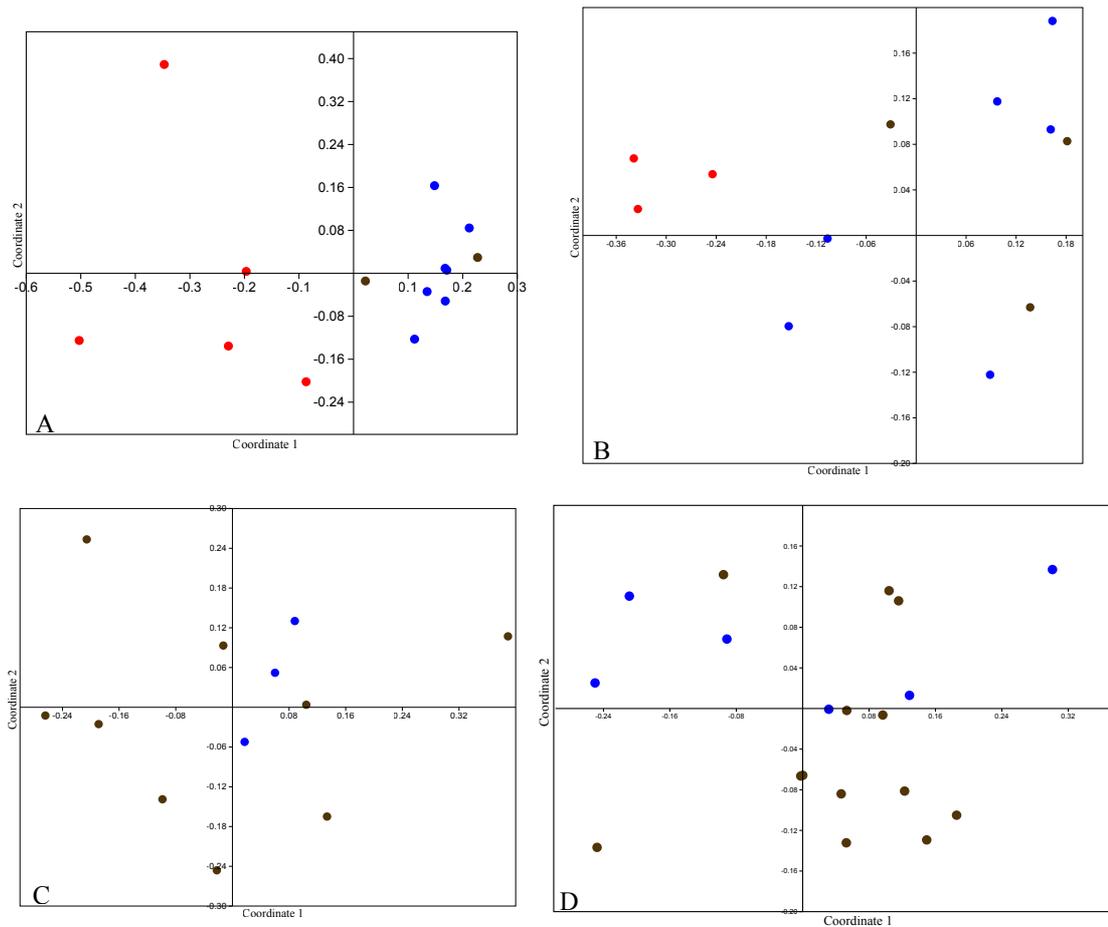


Figure 6.2.5 Clustering of Devil's Lair and Tunnel Cave bulk-bone and sediment samples according to LGM boundaries. Devil's Lair and Tunnel Cave bulk-bone samples (A,C respectively) in addition to Devil's Lair and Tunnel Cave sediment samples (B,D respectively) are clustered via nMDS using jaccard similarity index. Red circles (●) indicate pre-LGM samples, brown circles (●) indicate LGM samples and blue circles indicate (●) post-LGM samples.

As with DL, for the bulk-bone samples, it was difficult to differentiate LGM and post-LGM layers at TC despite a noticeable drop in OTU numbers post-LGM, however, there is some loose clustering of LGM samples (Figure 6.2.3C & D, and Figure 6.2.5C & D). It has been noted previously that attempting to detect differences between occupation and non-occupation or other factors at both DL and TC have been difficult when considering total taxa diversity (Dortch, 2004b). It seems likely that this is proving the case in this study also when using OTUs and this previous observation (Dortch, 2004b) influenced the decision to conduct OTU

analysis at family level assignments or higher independently of each other. Nevertheless, there is a drop in OTU numbers in non-occupation units when compared to surrounding occupation units during the LGM. With the post-LGM this pattern is difficult to determine as there may have been some limited occupation in stratigraphical unit 001 and there are too few layers that were sampled that worked (Dortch, 1996, 2004b). Furthermore, when samples are clustered using nMDS (Figure 6.2.6) there appears to be some clustering around confirmed occupation layers. Interestingly, of the three Macropodidae OTUs detected a single OTU is present in half of the samples with documented occupation but is absent from all non-occupation layers across both LGM and post-LGM. Further to this, it appears that at TC the number of 'burnt white bones' – a strong indication of human occupation (Dortch, 2004b) – and number of OTUs detected are significantly correlated ($r_s = 0.80$, $p < 0.05$); however no correlation was found between overall total burnt bone or artefacts.

The analysis of plant OTUs at TC is somewhat more promising than that of DL owing to difficulties in getting working sediment extracts at DL and variation in the number of extracts per period. At TC there seems to be a partially clearer separation of LGM and post-LGM samples for the sediment plant data when compared to that for bulk-bone. This is further supported by examining OTU diversity where a decrease in OTU number and reduction in diversity is observed post-LGM and the high levels of OTUs during the LGM seem independent of occupation. This loss in diversity correlates well with previous findings at both DL and TC where a shift in vegetation from more diverse Jarrah-dominated to less diverse Karri-dominated took place around the LGM/post-LGM boundary centred between stratigraphical unit 5-lower and stratigraphical unit 5-upper which coincides with the late Pleistocene and early Holocene, respectively (Dortch, 2004b, 2004a; Dortch & Wright, 2010). It is possible that plant OTU data may provide a clearer snapshot of past environmental change that is not as biased as the accumulation of zooarcheological bone which may be more influenced by human resource use. Connecting the changes in the environment to whether or not human subsistence changes accordingly is an important aspect of archaeology and can be difficult to determine without clear records. In the case of both DL and TC, sediment samples that were analysed for pollen returned no data (Dortch, 2004b). While it was possible to identify some

pollen grains from coprolites this may be biased towards the predators diet. As such the identification of plant DNA within sediment at TC, and although less successfully at DL, provide a potentially less biased indication of the local environment which can further substantiate findings based on charcoal analysis (Dortch, 2004b). Such an approach as that adopted in this study may potentially add a new dimension to archaeological plant analyses in the form of non-woody plants, provided future population of genetic databases with missing plant data takes place.

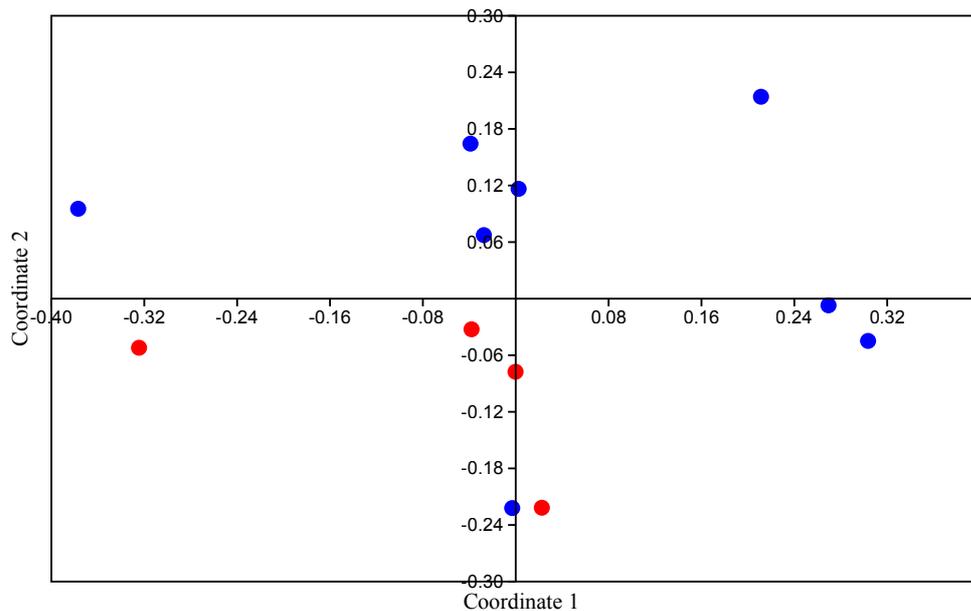


Figure 6.2.6 Clustering of Tunnel Cave bulk-bone samples according to occupation and non-occupation layers. Tunnel Cave bulk-bone samples are clustered based on the OTU composition identified in each sample, with the exclusion of hearth samples. Samples are coloured according to layers of occupation (•) and non-occupation (•). All nMDS clustering was performed using jaccard similarity index

Rainbow Cave and Wonijti Janga OTU diversity

Rainbow Cave showed a distinct difference in OTU number and to some degree diversity between cultural and non-cultural stratigraphical units for bulk-bone samples (Figure 6.2.4A), but it shows a shallow degree of clustering using nMDS with a lot of noise arising from layer 005 which is a transitional unit that shows limited cultural evidence (Lilley, 1993; Dortch, 2004b). The plant OTU diversity showed little variation over time in cultural stratigraphical units, with the exception

perhaps of unit 003 (Figure 6.2.4B), however for OTU analysis, only one non-cultural unit successfully yielded *sedaDNA*. Although six individual samples were successfully sequenced for this stratigraphical, which is the greatest number of samples used in *sedaDNA* analysis for any one layer at RC, a single stratigraphical unit will not give a nuanced view of change.

While there are some differences between stratigraphical units at WJ in terms of OTU composition and diversity for both plants and animals (Figure 6.2.4C & D) the samples show limited clustering according to stratigraphical unit when plotted using nMDS. The lack of clustering may indicate little change over time and a continuity in subsistence even post-European arrival within local communities. Such a continuity in subsistence is supported by research into the role of *habitus* and the preservation of traditional knowledge involving Noongar elders in southwest Australia (Rusack *et al.*, 2011). With that, however, when the plant OTU analysis was plotted there was some differentiation between samples classed as post-European and pre-European with the exception of two of six samples from stratigraphical unit 010 which grouped with the post-European layers (Figure S6.2.2). Alternatively, perhaps a different approach such as looking at the ubiquity of OTUs may yield better clustering in future studies; the number of samples and replicates for WJ at present do not permit this.

The use of nearby non-archaeological sites is of great benefit to archaeological analyses as they can be used to distinguish human influence on the fossil assemblage at a site from natural bone accumulation, in essence, serving as a control. In the case of WJ, layer 002 and 003 of the nearby NS are contemporaneous to the whole of the WJ deposit and both cluster fairly distinctly from each other when all OTUs are considered (Figure S6.2.3). Due to the possibility of the presence of amphibians at the NS site largely contributing to this separation all non-mammalian taxa were removed. When re-clustered unfortunately only two contemporaneous samples of six clustered separately from those of WJ (Figure 6.2.7). All remaining samples at NS clustered distinctly apart from the WJ deposit, however, there appears to be some slight variation in OTU number and perhaps composition at the site (Figure 6.2.4E).

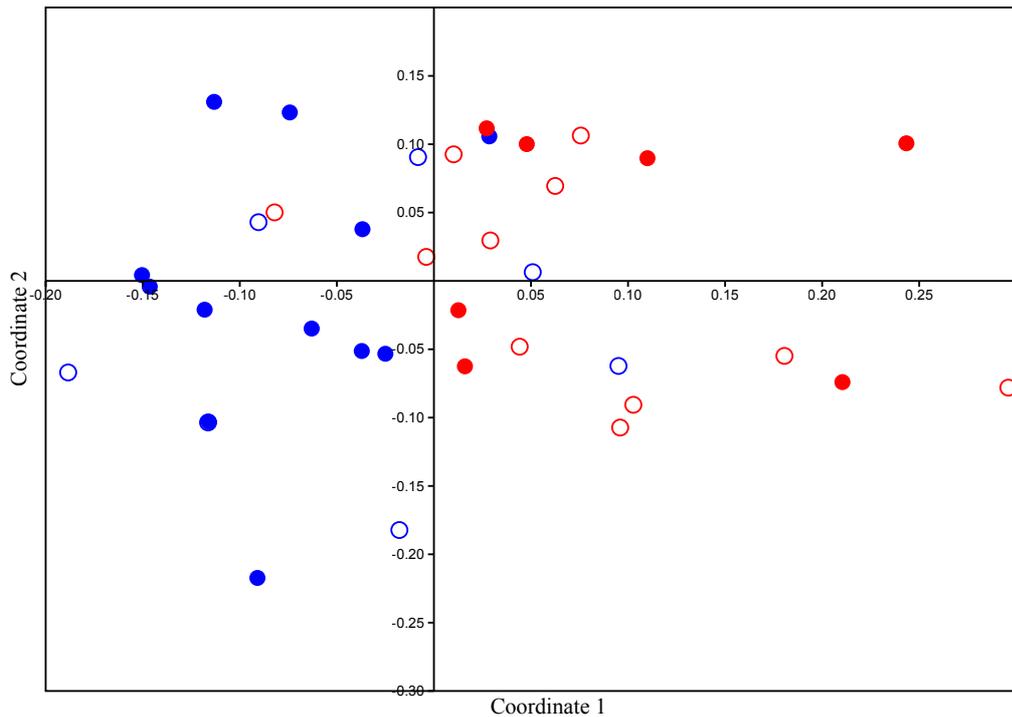


Figure 6.2.7 Clustering of Wonitji Janga and Northcote Sinkhole bulk-bone samples. Wonitji Janga (•) and Northcote Sinkhole (•) samples clustered based on the OTU composition (with non-mammalian OTUs excluded) identified in each sample. Open red circles indicate post-European arrival samples while filled red circles indicate pre-European arrival samples for Wonitji Janga. Open blue circles indicate Northcote Sinkhole samples that are contemporaneous to WJ samples, while filled circles indicate non-contemporaneous samples. All nMDS clustering was performed using jaccard similarity index.

6.2.6 Conclusion

This study set out to combine bulk-bone and *sed*aDNA metabarcoding approaches to analyse five sites in southwest Australia spanning a combined 50, 000-year record. The size and scope of the study coupled with unknown DNA preservation presented a number of challenges. Future studies may benefit from a more targeted approach to analysing samples such as sediment and bulk-bone material in southwest Australia. Through the use of *a priori* knowledge, it may be possible to examine loss of diversity in species such as *Bettongia* (woylie) or study changes resulting from the shifting canopy around LGM boundaries using *Eucalyptus*- or *Acacia*-specific primers. However, while this may solve some issues for taxa that are difficult to identify using frequently adopted gene regions it will not solve all issues in analysing

past biodiversity that has been lost or that is not currently characterised. The challenges presented by poor DNA preservation, DNA damage, sequencing error, contamination and lost biodiversity are all factors that will limit the resolving power of these aDNA approaches.

Despite these inherent difficulties, the data from this study have revealed some intriguing results that warrant further investigation and confirmation. In addition to this, many of the findings in this paper are further supported by previous morphology-based research into these sites and as such this paper serves to illustrate that BBM and *sedaDNA* strategies are capable of adding to current and past archaeological, palaeontological and ecological studies; albeit with the caution inherent to studies of ancient DNA.

6.2.7 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403-410.

Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K. H., . . . Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, 21, 1966-1979.

Archer, M., & Baynes, A. (1972). Prehistoric mammal faunas from two small caves in the extreme southwest of Western Australia. *Journal of the Royal Society of Western Australia*, 55, 80-89.

Balme, J., Merrilees, D., & Porter, J. (1978). Late Quaternary mammal remains spanning about 30,000 years from excavations in Devil's Lair, Western Australia. *Journal of the Royal Society of Western Australia*, 6, 33-65.

Beard, J. S. (1995). South-west Botanical Province. In S. D. Davis, V. H. Heywood, & A. C. Hamilton (Eds.), *Centres of Plant Diversity. Volume 2. Asia, Australasia, and the Pacific*. (Vol. 2). Cambridge, UK: WWF/IUCN, IUCN Publications Unit.

- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research*, *34*, D16-D20.
- Birks, H. J., & Birks, H. H. (2016). How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? *New Phytologist*, *209*, 499-506.
- Blum, S. A. E., Lorenz, M. G., & Wackernagel, W. (1997). Mechanism of retarded DNA degradation and Prokaryotic origin of DNases in nonsterile soils. *Systematic and Applied Microbiology*, *20*, 513-521.
- Boessenkool, S., Epp, L. S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., . . . Brochmann, C. (2012). Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, *21*, 1806-1815.
- Bradshaw, C. J. A. (2012). Little left to lose: deforestation and forest degradation in Australia since European colonization. *Journal of Plant Ecology*, *5*, 109-120.
- Brown, O. J. F. (2006). Tasmanian devil (*Sarcophilus harrisii*) extinction on the Australian mainland in the mid-Holocene: multicausality and ENSO intensification. *Alcheringa: An Australasian Journal of Palaeontology*, *30*, 49-57.
- Campos, P. F., Willerslev, E., Sher, A., Orlando, L., Axelsson, E., Tikhonov, A., . . . Gilbert, M. T. P. (2010). Ancient DNA analyses exclude humans as the driving force behind late Pleistocene musk ox (*Ovibos moschatus*) population dynamics. *proceedings of the National Academy of Sciences*, *107*, 5675-5680.
- Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An Efficient Multistrategy DNA Decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One*, *5*, e13042.
- Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*, 1834-1847.

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., & Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, *409*, 704-707.

Crecchio, C., & Stotzky, G. (1998). Binding of DNA on humic acids: Effect on transformation of *Bacillus subtilis* and resistance to DNase. *Soil Biology and Biochemistry*, *30*, 1061-1067.

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*, a012567.

De Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E., & Taberlet, P. (2014). DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Molecular Ecology Resources*, *14*, 306-323.

Dietl, G. P., Kidwell, S. M., Brenner, M., Burney, D. A., Flessa, K. W., Jackson, S. T., & Koch, P. L. (2015). Conservation paleobiology: leveraging knowledge of the past to inform conservation and restoration. *Annual Review of Earth and Planetary Sciences*, *43*, 79-103.

Donoghue, M. J., Doyle, J. A., Gauthier, J., Kluge, A. G., & Rowe, T. (1989). The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics*, *20*, 431-460.

Dortch, C. (1979). Devil's Lair, an example of prolonged cave use in southwestern Australia. *World Archaeology*, *10*, 258-279.

Dortch, J. (1996). Late Pleistocene and recent Aboriginal occupation of Tunnel Cave and Witchcliffe Rockshelter, southwestern Australia. *Australian Aboriginal Studies*, *2*, 51-60.

Dortch, J. (2004a). Late Quaternary vegetation change and the extinction of Black-flanked Rockwallaby (*Petrogale lateralis*) at Tunnel Cave, southwestern Australia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *211*, 185-204.

Dortch, J. (2004b). *Palaeo-environmental Change and the Persistence of Human Occupation in South-western Australian Forests*. Oxford: Archaeopress.

Dortch, J., Monks, C., Webb, W., & Balme, J. (2014). Intergenerational archaeology: Exploring niche construction in southwest Australian zooarchaeology. *Australian Archaeology*, 79, 187-193.

Dortch, J., & Wright, R. (2010). Identifying palaeo-environments and changes in Aboriginal subsistence from dual-patterned faunal assemblages, south-western Australia. *Journal of Archaeological Science*, 37, 1053-1064.

Drummond, A. J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., . . . Wilson, A. (2011). Geneious v8.1, created by Biomatters. Available from <http://www.geneious.com/>. Retrieved from <http://www.geneious.com/>

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792-1797.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460-2461.

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10, 996-998.

Faircloth, B. C., & Glenn, T. C. (2012). Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS One*, 7, e42543.

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguët-Covex, C., De Barba, M., . . . Taberlet, P. (2014). Replication levels, false presences, and the estimation of presence / absence from eDNA metabarcoding data. *Molecular Ecology Resources*, n/a-n/a.

Gavin, D. G., Fitzpatrick, M. C., Gugger, P. F., Heath, K. D., Rodríguez-Sánchez, F., Dobrowski, S. Z., . . . Williams, J. W. (2014). Climate refugia: joint inference from fossil records, species distribution models and phylogeography. *New Phytologist*, 204, 37-54.

- Gilbert, M. T. P., Bandelt, H. J., Hofreiter, M., & Barnes, I. (2005). Assessing ancient DNA studies. *Trends in Ecology and Evolution*, *20*, 541-544.
- Golyeva, A., & Andrič, M. (2014). Palaeoecological reconstruction of wetlands and Eneolithic land use in Ljubljansko barje (Slovenia) based on biomorphic and pollen analysis. *CATENA*, *112*, 38-47.
- Grealy, A., Macken, A., Allentoft, M. E., Rawlence, N. J., Reed, E., & Bunce, M. (2016). An assessment of ancient DNA preservation in Holocene–Pleistocene fossil bone excavated from the world heritage Naracoorte Caves, South Australia. *Journal of Quaternary Science*, 1-13.
- Grealy, A. C., McDowell, M. C., Scofield, P., Murray, D. i. C., Fusco, D. A., Haile, J., . . . Bunce, M. (2015). A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quaternary Science Reviews*, *128*, 37-47.
- Haouchar, D., Haile, J., McDowell, M. C., Murray, D. C., White, N. E., Allcock, R. J. N., . . . Bunce, M. (2013). Thorough assessment of DNA preservation from fossil bone and sediments excavated from a late Pleistocene-Holocene cave deposit on Kangaroo Island, South Australia. *Quaternary Science Reviews*, *84*, 56-64.
- Hebsgaard, Martin B., Gilbert, M. T. P., Arneborg, J., Heyn, P., Allentoft, Morten E., Bunce, M., . . . Willerslev, E. (2009). ‘The Farm Beneath the Sand’ – an archaeological case study on ancient ‘dirt’ DNA. *Antiquity*, *83*, 430-444.
- Heupink, T. H., Huynen, L., & Lambert, D. M. (2011). Ancient DNA suggests dwarf and 'giant' emu are conspecific. *PLoS One*, *6*, e18728.
- Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., . . . Colston, B. W. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, *83*, 8604-8610.
- Hofreiter, M., Kreuz, E., Eriksson, J., Schubert, G., & Hohmann, G. (2010). Vertebrate DNA in fecal samples from bonobos and gorillas: evidence for meat consumption or artefact? *PLoS One*, *5*, e9419.

- Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews Genetics*, 2, 353-359.
- Hopper, S. D., & Gioia, P. (2004). The Southwest Australian Floristic Region: evolution and conservation of a global hot spot of biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 35, 623-650.
- How, R. A., Dell, J., & Humphreys, W. F. (1987). The ground vertebrate fauna of coastal areas between Busselton and Albany, Western Australia. *Records of the Western Australian Museum*, 13, 553-574.
- Huang, Y.-T., Lowe, D. J., Churchman, G. J., Schipper, L. A., Rawlence, N. J., & Cooper, A. (2014). Carbon storage and DNA adsorption in allophanic soils and paleosols. In A. E. Hartemink & K. McSweeney (Eds.), *Soil Carbon* (pp. 163-172). Switzerland: Springer International Publishing.
- Hunter, D. O., Britz, T., Jones, M., & Letnic, M. (2015). Reintroduction of Tasmanian devils to mainland Australia can restore top-down control in ecosystems where dingoes have been extirpated. *Biological Conservation*, 191, 428-435.
- Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17, 377-386.
- Jørgensen, T., Haile, J., Möller, P. E. R., Andreev, A., Boessenkool, S., Rasmussen, M., . . . Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, 21, 1989-2003.
- Jørgensen, T., Kjær, K. H., Haile, J., Rasmussen, M., Boessenkool, S., Andersen, K., . . . Willerslev, E. (2011). Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Molecular Ecology*, 21, 1980-1988.
- Kozarewa, I., & Turner, D. J. (2011) Amplification-free library preparation for paired-end Illumina sequencing.

- Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. *Molecular Ecology*, *17*, 4186-4196.
- Lilley, I. (1993). Recent research in southwestern Australia: a summary of initial findings. *Australian Archaeology*, *36*, 34-41.
- Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*, 709-715.
- Llamas, B., Brotherton, P., Mitchell, K. J., Templeton, J. E., Thomson, V. A., Metcalf, J. L., . . . Camens, A. B. (2015). Late Pleistocene Australian marsupial DNA clarifies the affinities of extinct megafaunal kangaroos and wallabies. *Molecular Biology and Evolution*, *32*, 574-584.
- Lorenz, M. G., & Wackernagel, W. (1987). Adsorption of DNA to sand and variable degradation rates of adsorbed DNA. *Applied Environmental Microbiology*, *53*, 2948-2952.
- Lorenzen, E. D., Nogues-Bravo, D., Orlando, L., Weinstock, J., Binladen, J., Marske, K. A., . . . Willerslev, E. (2011). Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, *479*, 359-364.
- Mann, D. H., Groves, P., Reanier, R. E., Gaglioti, B. V., Kunz, M. L., & Shapiro, B. (2015). Life and extinction of megafauna in the ice-age Arctic. *proceedings of the National Academy of Sciences*, *112*, 14301-14306.
- Moody, J. (2005). Unravelling the threads: climate changes in the Late Bronze III Aegean. In A. L. D'Agata & J. Moody (Eds.), *Ariadne's threads: connections between Crete and the Greek mainland in Late Minoan III (LM IIIA2 to LM IIIC)*. (pp. pp 443–474). Athens: Scuola Archeologica Italiana di Atene.
- Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pages, H., & Gentleman, R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, *25*, 2607-2608.

- Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PLoS One*, *10*, e0124671.
- Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., . . . Bunce, M. (2013). Scrapheap challenge: a novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports.*, *3*, 3371.
- Murray, D. C., Pearson, S. G., Fullagar, R., Chase, B. M., Houston, J., Atchison, J., . . . Bunce, M. (2012). High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews*, *58*, 135-145.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*, 853-858.
- Oskam, C. L., Haile, J., McLay, E., Rigby, P., Allentoft, M. E., Olsen, M. E., . . . Jacomb, C. (2010). Fossil avian eggshell preserves ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, *277*.
- Pacioni, C., Hunt, H., Allentoft, M. E., Vaughan, T. G., Wayne, A. F., Baynes, A., . . . Bunce, M. (2015). Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, *24*, 5813-5828.
- Paczkowska, G., & Chapman, A. R. (2000). *The Western Australian flora : a descriptive catalogue*. Perth, W.A.: Wildflower Society of Western Australia : Western Australian Herbarium : CALM : Western Australian Botanic Gardens & Parks Authority.
- Pages, H., Aboyoun, P., Gentleman, R., & DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms R package version 2.38.4.
- Paplinska, J. Z., Taggart, D. A., Corrigan, T., Eldridge, M. D. B., & Austin, J. J. (2011). Using DNA from museum specimens to preserve the integrity of evolutionarily significant unit boundaries in threatened species. *Biological Conservation*, *144*, 290-297.

- Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., . . . Willerslev, E. (2014). Ancient and modern environmental DNA. *Philosophical Transactions Royal Society B*, *370*, 20130383.
- Philippe, E., Franck, L., & Jan, P. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*.
- Pietramellara, G., Ascher, J., Ceccherini, M. T., Nannipieri, P., & Wenderoth, D. (2007). Adsorption of pure and dirty bacterial DNA on clay minerals and their transformation frequency. *Biology and Fertility of Soils*, *43*, 731-739.
- Prideaux, G. J., & Warburton, N. M. (2010). An osteology-based appraisal of the phylogeny and evolution of kangaroos and wallabies (Macropodidae: Marsupialia). *Zoological Journal of the Linnean Society*, *159*, 954-987.
- Ramakrishnan, U. M. A., & Hadly, E. A. (2009). Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies. *Molecular Ecology*, *18*, 1310-1330.
- Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., . . . Willerslev, E. (2011). An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science*, *334*, 94-98.
- Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Research*, *39*, e145-e145.
- Rusack, E. M., Dortch, J., Hayward, K., Renton, M., Boer, M., & Grierson, P. (2011). The role of *habitus* in the maintenance of traditional Noongar plant knowledge in southwest Western Australia. *Human Ecology*, *39*, 673-682.
- Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P., Sher, A. V., . . . Cooper, A. (2004). Rise and fall of the Beringian steppe bison. *Science*, *306*, 1561-1565.
- Shapiro, B., & Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, *343*.

- Sønstebø, J. H., Gielly, L., Brysting, A. K., Elven, R., Edwards, M., Haile, J., . . . Brochmann, C. (2010). Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Molecular Ecology Resources*, *10*, 1009-1018.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., . . . Willerslev, E. (2007). Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*, e14.
- Taberlet, P., Gielly, L., Pautou, G., & Bouvet, J. (1991). Universal primers for amplification of three noncoding regions of chloroplast DNA. *Plant Molecular Biology*, *17*, 1105-1109.
- Taylor, P. G. (1996). Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution*, *13*, 283-285.
- R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- Terry, R. C., Li, C., & Hadly, E. A. (2011). Predicting small-mammal responses to climatic warming: autecology, geographic range, and the Holocene fossil record. *Global Change Biology*, *17*, 3019-3034.
- Thackway, R., & Cresswell, I. D. (1995). *An interim biogeographic regionalisation for Australia: a framework for setting priorities in the National Reserves System Cooperative Program*. Retrieved from Canberra:
- Thomas, R. H., Schaffner W Fau - Wilson, A. C., Wilson Ac Fau - Paabo, S., & Paabo, S. DNA phylogeny of the extinct marsupial wolf. *Nature*, *340*, 465-467.
- Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4-18.
- Turney, C., & Bird, M. I. (2001). Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. *Quaternary Research*, *55*, 3-13.

Van Devender, T. R., & Spaulding, W. G. (1979). Development of vegetation and climate in the south western United States. *Science*, 204, 701-710.

Vestheim, H., & Jarman, S. N. (2008). Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Frontiers in Zoology*, 5, Article No.: 12.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M. B., Brand, T. B., . . . Collins, M. J. (2007). Ancient biomolecules from deep ice cores reveal a forested southern greenland. *Science*, 317, 111-114.

Willerslev, E., & Cooper, A. (2005). Ancient DNA. *Proceedings of the Royal Society of London B: Biological Sciences*, 272, 3-16.

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., . . . Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, 506, 47-51.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

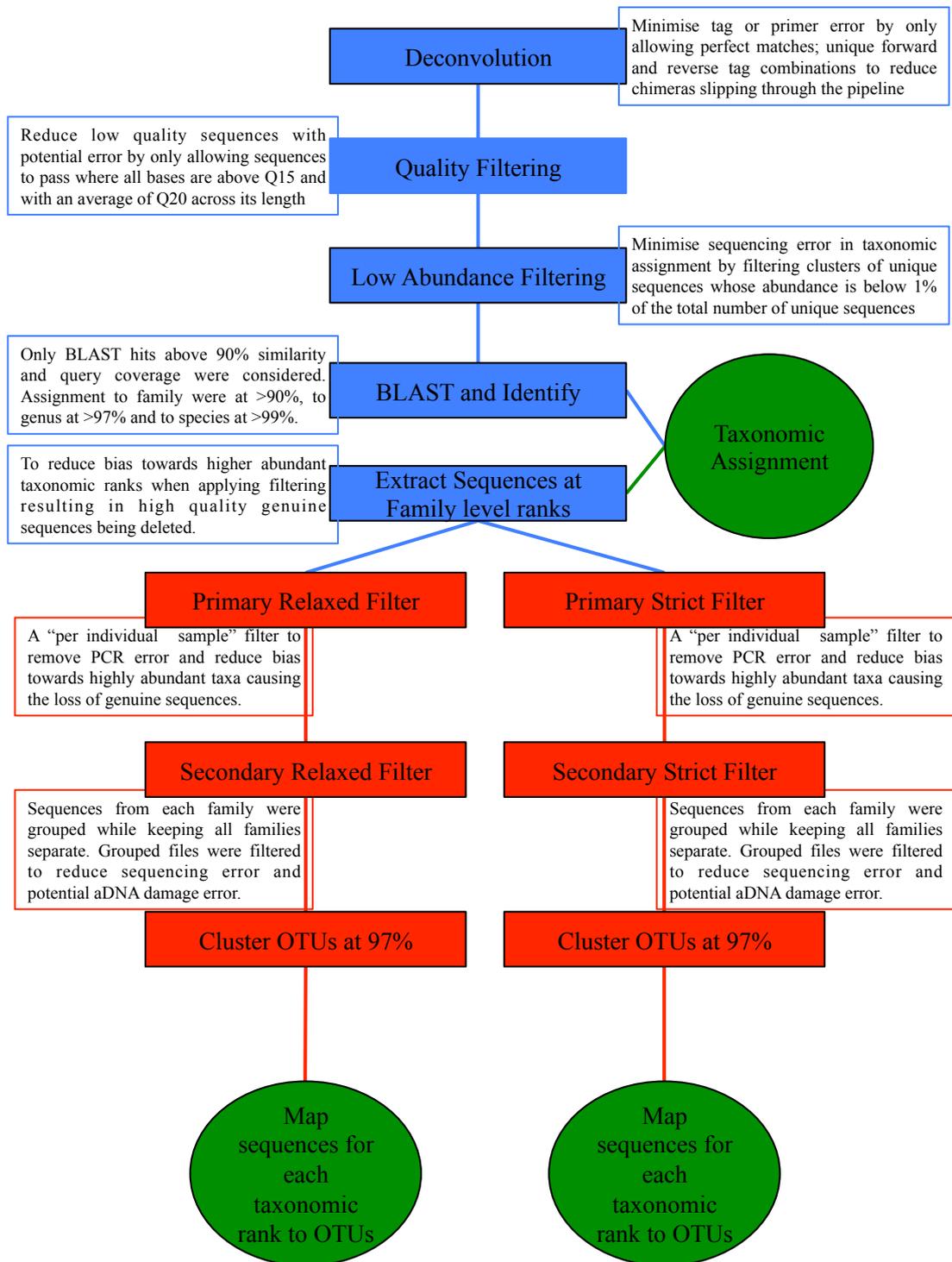


Figure S6.2.1 Schematic of the stringent and relaxed approach to OTU analysis adopted in this study.

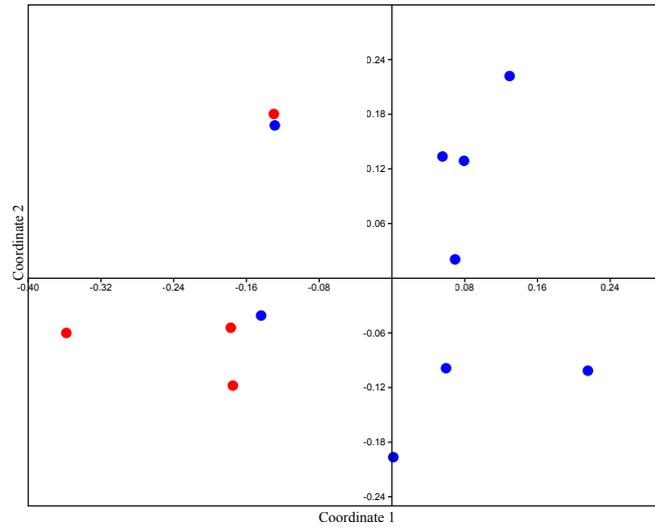


Figure S6.2.2 Clustering of Wonitji Janga samples according to European arrival. Wonitji Janga samples are clustered based on the OTU composition identified in each sample. Red circles (•) indicate post-European arrival samples while blue circles (•) indicate pre-European arrival samples. All nMDS clustering was performed using jaccard similarity index.

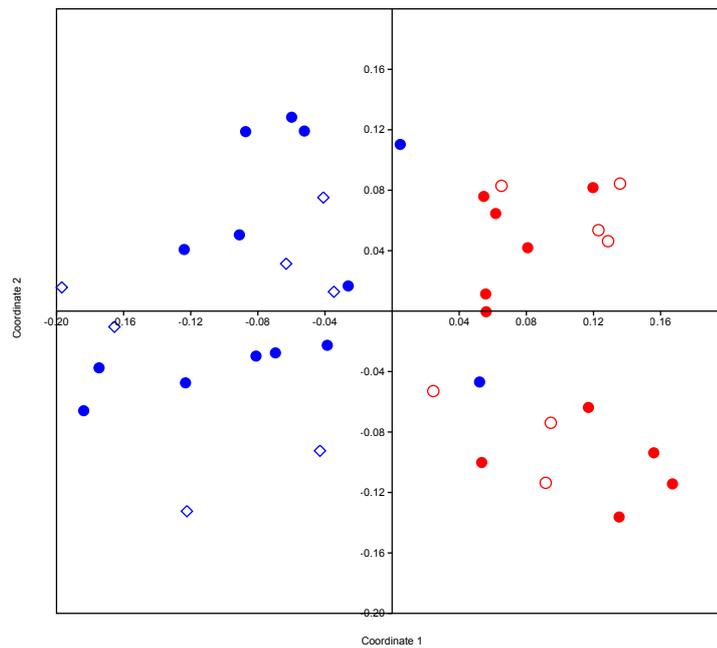


Figure S6.2.3 Clustering of Wonitji Janga and Northcote Sinkhole bulk-bone samples. Wonitji Janga (•) and Northcote Sinkhole (•) samples clustered based on the OTU composition (all OTUs included) identified in each sample. Open red circles indicate post-European arrival samples while filled red circles indicate pre-European arrival samples for Wonitji Janga. Open blue diamonds indicate Northcote Sinkhole samples that are contemporaneous to WJ samples, while filled circles indicate non- contemporaneous samples. All nMDS clustering was performed using jaccard similarity index.

Table S6.2.1A Presence and absence of all faunal taxa identified at Devil's Lair.

Class	Family	Genus	Period							
			Post-LGM				LGM		Pre-LGM	
			XIV	XI-XII	X	IX	VII	IV-V	II	I
Aves	Corvidae				√1					
	Dromaiidae				√1					√1
		<i>Dromaius</i>			√1					√1
	Meliphagidae									√1
	Phasianidae									√1
		<i>Coturnix</i>								√1
	Psittacidae					√1				
	Psittaculidae				√1					
Mammalia	Dasyuridae		√1	√1	√1	√2			√2	√2
		<i>Dasyurus</i>	√1	√1		√1				
		<i>Phascogale</i>			√1					√2
		<i>Sminthopsis</i>								√1
	Macropodidae		√2	√2	√2	√2	√2	√2	√2	√2
		<i>Macropus</i>	√2	√2	√2	√2	√2		√2	√2
		<i>Setonix</i>	√1	√1	√1	√1	√1	√1	√1	√1
	Peramelidae		√1	√2	√2	√2	√1	√1		√2
		<i>Isoodon</i>	√1	√2	√2	√2	√1	√1		√2
	Phalangeridae			√2	√2	√2	√2	√2		
		<i>Trichosurus</i>		√2	√2	√2	√2	√2		
	Potoroidae		√2	√2	√2	√2	√2	√2	√1	√2
		<i>Bettongia</i>	√2	√2	√2	√2	√2	√2		√1
		<i>Potorous</i>		√1	√1				√1	
	Pseudocheiridae			√2	√2	√2		√2	√2	√1
		<i>Pseudocheirus</i>		√2	√2	√1		√2	√1	√1
	Tarsipedidae								√1	√2
	<i>Tarsipes</i>							√1	√2	
Muridae		√1	√1	√1	√1	√1	√2	√2	√2	
	<i>Rattus</i>	√1	√1		√1	√1	√2	√2	√2	

√1 Detected with one primer , √2 Detected with two primers

Table S6.2.1B Presence and absence of all faunal taxa identified at Tunnel Cave.

Order	Family	Genus	Stratigraphical unit									
			Post-LGM				LGM					
			001-002	003	5-lower	F4B	006	7-upper	F12	008	9-upper	9-lower-10
Amphibia	Hylidae				√1							
		<i>Litoria</i>			√1							
Aves	Phalacrocoracidae									√1		
	<i>Phalacrocorax</i>									√1		
	Psittaculidae		√1									
Mammalia	Dasyuridae			√1				√2	√1			
	<i>Antechinus</i>			√1								
	<i>Dasyurus</i>							√2				
	<i>Sarcophilus</i>							√1				
	Macropodidae		√2	√1	√1			√1	√2	√2	√1	√1
	<i>Macropus</i>		√2		√1			√1	√2	√2		√1
	<i>Setonix</i>		√1	√1	√1					√1	√1	√1
	Peramelidae		√2	√1				√1	√1	√1	√1	√1
	<i>Isoodon</i>		√2					√1	√1	√1	√1	√1
	Phalangeridae		√1		√1				√2	√2	√1	√1
	<i>Trichosurus</i>		√1		√1				√2	√2	√1	√1
	Potoroidae		√2	√1	√1	√1			√2	√2	√1	√1
	<i>Bettongia</i>		√2		√1	√1			√2	√2	√1	√1
	<i>Potorous</i>		√1							√1		
	Pseudocheiridae		√2		√1			√1	√1	√2		√1
	<i>Pseudocheirus</i>		√1		√1			√1	√1	√2		√1
	Muridae		√2	√1	√1				√2	√2	√1	√1
	<i>Mus</i>			√1								
	<i>Rattus</i>		√2	√1					√2	√2	√1	√1

√1 Detected with one primer , √2 Detected with two primers

Table S6.2.1C Presence and absence of all faunal taxa identified at Rainbow Cave.

Class	Family	Genus	Stratigraphical unit											
			Cultural					Not Cultural						
			001	002	004	005	006	013	014	015	016			
Actinopterygii	Arripidae					√1								
	Carangidae				√1	√1								
	Cheilodactylidae		√1		√1									
		<i>Cheilodactylus</i>	√1		√1									
		Kyphosidae	√1	√2	√1	√2					√2	√2		
		<i>Girella</i>	√1	√2	√1	√1					√2	√1		
		<i>Kyphosus</i>	√1	√1	√1	√2					√1	√2		
	<i>Scorpis</i>	√1												
	Sparidae				√1									
Amphibia	Myobatrachidae		√1	√1		√1			√1					
Aves	Accipitridae								√1					
	Anatidae								√1					
	Cacatuidae					√2								
		<i>Calyptorhynchus</i>					√2							
	Cuculidae					√1								
	Diomedeidae		√1											
	Dromaiidae			√1		√1								
		<i>Dromaius</i>		√1		√1								
	Falconidae				√2									
	Meliphagidae					√2								
	Petroicidae								√2					
		<i>Petroica</i>							√2					
	Phalacrocoracidae						√1							
		<i>Phalacrocorax</i>					√1							
	Phasianidae						√2			√1				
		<i>Coturnix</i>					√2							
Psittacidae				√1										
	<i>Platycercus</i>			√1										
Rallidae				√2										
	<i>Lewinia</i>			√2										
Turnicidae						√2	√2	√2						
	<i>Turnix</i>					√2	√2	√2						
Reptilia	Scincidae		√1			√1			√1	√1				
	<i>Ctenotus</i>		√1						√1					
	<i>Hemiergis</i>		√1			√1								
Mammalia	Burramyidae				√1	√1								
	Dasyuridae				√1	√2	√2	√1	√2	√1				
		<i>Antechinus</i>					√2			√1				
		<i>Dasyurus</i>					√1		√1	√1				
		<i>Sarcophilus</i>					√2	√1						
		<i>Sminthopsis</i>					√1							
	Macropodidae		√3	√3	√3	√3	√3	√3	√3	√2	√3	√3		
		<i>Macropus</i>	√3	√3	√3	√3	√1	√3	√2					
		<i>Setonix</i>	√1	√1	√1	√1	√1	√1	√1	√1	√1	√1		
	Phalangeridae		√1		√1	√1			√1	√1	√3	√1		
		<i>Trichosurus</i>	√1		√2	√1				√1	√3	√2		
	Potoroidae		√3	√3	√3	√3	√1	√3				√3		
		<i>Bettongia</i>	√3	√3	√3	√3	√1	√3				√3		
		<i>Potorous</i>	√2		√1	√2						√1		
	Pseudocheiridae		√3		√3	√3			√2					
		<i>Pseudocheirus</i>	√3		√3	√3			√2					
	Peramelidae		√2	√2	√2	√2	√2	√2			√1	√2		
	<i>Isoodon</i>	√2	√1	√2	√2	√2	√2			√1	√2			
Pteropodidae					√1					√1				
Vespertilionidae		√1												
Muridae		√3	√3	√3	√3	√3	√3	√1	√3	√3	√3			
	<i>Rattus</i>	√3	√3	√3	√3	√3	√3		√3	√3	√3			

√ 1 Detected with one primer, √ 2 Detected with two primers, √ 3 Detected with three primers

Table S6.2.1D Presence and absence of all faunal taxa identified at Wonitji Janga.

Order	Family	Genus	Stratigraphical unit						
			Post-Euro			Pre-Euro			
			001	002	003	004	008	009	010
Amphibia	Hylidae				√1				√1
		<i>Litoria</i>			√1				√1
	Myobatrachidae								√2
		<i>Pseudophryne</i>							√1
Reptilia	Scincidae								√1
	Varanidae		√1						
		<i>Varanus</i>	√1						
Aves	Burhinidae				√1				
		<i>Burhinus</i>			√1				
	Dromaiidae				√1				
		<i>Dromaius</i>			√1				
	Psittacidae		√1						
Mammalia	Dasyuridae		√2		√2	√2			√1
		<i>Dasyurus</i>	√2		√1				
		<i>Phascogale</i>				√1			
		<i>Sarcophilus</i>			√2				√1
	Burramyidae		√1		√1				
	Macropodidae		√2	√2	√2	√1	√1	√1	√2
		<i>Macropus</i>	√2	√2	√2	√2	√1		√2
		<i>Setonix</i>	√1	√1	√1			√1	√1
	Peramelidae		√2	√1	√2			√1	√2
		<i>Isoodon</i>	√2	√1	√2			√1	√2
	Phalangeridae		√2	√2	√2			√1	√2
		<i>Trichosurus</i>	√1	√1	√1			√1	√1
	Potoroidae		√2	√1	√2	√1			√2
		<i>Bettongia</i>	√2		√2	√1			√2
		<i>Potorous</i>		√1					√1
	Pseudocheiridae		√2		√1		√1	√1	√1
		<i>Pseudocheirus</i>	√2		√1		√1	√1	√2
	Canidae		√1						
		<i>Vulpes</i>	√1						
	Leporidae		√2		√2				
	<i>Oryctolagus</i>	√2		√2					
Muridae		√2	√1	√2				√2	
	<i>Mus</i>			√1				√1	
	<i>Rattus</i>	√2	√1	√2				√2	

√1 Detected with one primer , √2 Detected with two primers

Table S6.2.1D Presence and absence of all faunal taxa identified at Northcote Sinkhole.

Order	Family	Genus	Stratigraphical unit				
			Late-Hol			Early-Hol	
			001	002	003	004	005
Anura	Hylidae		√1	√1	√1	√1	√1
		<i>Litoria</i>	√1	√1	√1	√1	√1
	Leptodactylidae		√1		√1		
	Myobatrachidae		√2	√1	√2	√2	√2
		<i>Limnodynastes</i>	√1		√1	√1	√1
		<i>Pseudophryne</i>	√1		√1	√1	√1
Squamata	Elapidae		√1				
	Scincidae		√2	√1	√1	√1	√1
		<i>Ctenotus</i>		√1		√1	
	Varanidae	<i>Varanus</i>	√1				
Passeriformes	Pardalotidae			√1			
		<i>Sericornis</i>		√1			
	Phasianidae		√1				
		<i>Coturnix</i>	√1				
	Podargidae		√1		√1		
		<i>Podargus</i>	√1		√1		
	Turnicidae					√1	
		<i>Turnix</i>				√1	
Diprotodontia	Burramyidae		√1				
	Dasyuridae					√2	√2
		<i>Antechinus</i>					√1
		<i>Dasyurus</i>					√1
		<i>Phascogale</i>				√1	√1
		<i>Sminthopsis</i>				√1	√1
	Macropodidae		√1	√2	√2	√1	√2
		<i>Macropus</i>		√1		√1	
		<i>Setonix</i>					√1
	Peramelidae		√2		√1	√2	√2
		<i>Isoodon</i>	√2		√1	√2	√2
	Phalangeridae		√2	√2	√1	√1	√1
		<i>Trichosurus</i>	√2	√2	√2	√2	√2
	Potoroidae		√1	√2	√2	√2	√2
		<i>Bettongia</i>	√1	√2	√2	√2	√2
		<i>Potorous</i>					√1
	Pseudocheiridae			√2	√1		√2
		<i>Pseudocheirus</i>		√1	√1		√2
	Tachyglossidae						√2
	Vespertilionidae	<i>Chalinolobus</i>			√2		
Leporidae		√2	√1				
	<i>Oryctolagus</i>	√2					
Muridae		√2	√2	√2	√2	√2	
	<i>Rattus</i>	√2	√2	√2	√2	√2	

√1 Detected with one primer, √2 Detected with two primers

Table S6.2.2 Information regarding the taxonomic assignment of faunal taxa.

Class	Family	Genus	Comments
Actinopterygii	Kyphosidae	<i>Scorpis</i>	Both <i>S. aequipinnis</i> and <i>S. georgiana</i> are found off the southwest coast of Australia. Neither species has 16S or 12S sequences on Genbank. <i>S. aequipinnis</i> may be the most likely candidate for assignment as it has previously been identified as a potential food resource (Dortch, 2004). It is a marine species and it was speared from rocky shores and as such supports previous findings on past fishing practices (Dortch, 2004).
Amphibia	Hylidae	<i>Litoria</i>	Many <i>Litoria</i> species are found in Western Australia. However, at the sites where the genus was detected the most likely species, as determined from occurrence records, is either <i>L. adelaidensis</i> (present on Genbank for both 12S and 16S) or <i>L. moorei</i> (not present on Genbank for 12S or 16S).
Aves	Cacatuidae	<i>Calyptorhynchus</i>	The taxonomy of <i>Calyptorhynchus</i> is uncertain and it could be any one of possibly three species.
	Cuculidae		<i>Cuculus</i> is not currently found in the region, however <i>Cuculus</i> fossil material has been identified from the region in another study (Baird, 1991).
	Diomedeidae		This is an arguably unusual finding due to the fact that species within this family spend most of their time out at sea. However, sightings of, for example, <i>Thalassarche causta</i> , which occurs in the region, have been recorded in bays and harbours. This was found at the coastal Rainbow Cave site (Garnett & Crowley, 2000).
	Dromaiidae	<i>Dromaius</i>	Only <i>D. novaehollandiae</i> is known to occur in southwest Australia.
	Petroicidae	<i>Petroica</i>	Many species of <i>Petroica</i> have been recorded in the region and they have varying degrees of representation on Genbank across both 16S and 12S. As such, it is not possible to make a confident assignment.

Phalacrocoracidae	<i>Phalacrocorax</i>	An arguably unusual find due to the fact that these birds are found primarily around aquatic habitats. However, it was detected at the coastal Rainbow Cave site. Members of this genus can be found quite far inland, generally around lakes, swamps or other bodies of water. There are a number of species in the region with varying degrees of representation on Genbank across both 16S and 12S. As such, it is not possible to make a confident assignment.
Phasianidae		In four samples predating the introduction of the species into Australia, <i>Gallus gallus</i> was identified with 100 % similarity. This is most likely laboratory contamination as it has been found sporadically in controls in other studies (Leonard <i>et al.</i> , 2007). It should be noted that in instances where <i>G.gallus</i> occurred NO <i>Coturnix</i> sequences (below) were present and vice versa. Therefore the presence of <i>G. gallus</i> does not necessarily cast doubt on the presence of <i>Coturnix</i> . However, further confirmation of <i>Coturnix</i> in samples where it occurred should take place to ensure its validity.
Phasianidae	<i>Coturnix</i>	Unidentified Phasianidae have been recorded as potential past food resources in the southwest. There are two possible species in the immediate area, <i>C. pectoralis</i> and <i>C. ypsilophor</i> , neither of which have sequences for 16S nor 12S on Genbank. Therefore, a confident assignment cannot be made other than at the genus level.
Psittacidae	<i>Platycercus</i>	The only records of <i>Platycercus</i> in the area is that of <i>P. icterotis</i> for which there are no 16S or 12S sequences on Genbank.
Rallidae	<i>Lewinia</i>	The only species of <i>Lewinia</i> recorded in Australia is <i>L. pectoralis</i> . Due to sequence homology, however, it was not dissimilar enough to the other non-Australian species to allow it to be identified to species level in MEGAN.

	Turnicidae	<i>Turnix</i>	The only records of <i>Turnix</i> in the area is that of <i>T. varius</i> for which there are no 16S or 12S sequences on Genbank.
Reptilia	Scincidae	<i>Ctenotus</i>	A number of <i>Ctenotus</i> species occur in southwest Australia. Those represented on Genbank for 12S and 16S have a high degree of similarity thus complicating assignment. Furthermore, many other species are not represented on Genabank for 16S and/or 12S.
	Scincidae	<i>Hemiergus</i>	There are two species of <i>Hemiergus</i> known to occur near the cave sites: <i>H. peronii</i> and <i>H. gracilipes</i> . The latter has no 16S or 12S sequences on Genbank while the former is present on Genbank for both. Therefore, it is not possible to make a confident assignment
	Varanidae	<i>Varanus</i>	<i>Varanus</i> was only detected using 12S. There are two possible species: <i>V. gouldii</i> and <i>V. rosenbergi</i> . The former is represented on Genbank for 12S while the latter is not. <i>Varanus gouldii</i> has been recorded as a potential food resource in the southwest, however, <i>V. rosenbergi</i> also occurs in the immediate vicinity of the cave sites.
Mammalia	Pteropodidae		This family while occurring in Western Australia is not known to have ever occurred in southwest Australia. The percentage similarity of the DNA sequences to members of this family was quite poor in comparison to most other faunal taxa - between 91-94 %. This could very well represent error or damage or a lack of database coverage for some native southwest bats.

Vespertilionidae		While Vespertilionidae does occur in southwest Australia, for Rainbow Cave, the percentage similarity of the DNA sequences to members of this family was quite poor in comparison to most other faunal taxa - around 90 %. It could therefore indicate damage, error or a lack of database coverage for members of the family e.g. <i>Falsistrellus</i> for which only a single 12S sequence exists on Genbank for a species not found in the area and no 16S for any species within this genus is present on Genbank. However, at Northcote Sinkhole the percentage similarity of Vespertilionidae sequences to <i>Chalinolobus</i> , which does occur in the area, was 100 %, though it also mapped to a species in another genus with 100 % similarity that is not known to occur in the area.
Burramyidae		The only member of this family found in southwest Australia is <i>Cercartetus concinnus</i> . However, the sequence similarity was low at 95 % and as such this is being interpreted cautiously despite the fact that it has been recorded at the sites previously. This may be a case of DNA damage or error.
Dasyuridae	<i>Antechinus</i>	<i>Antechinus flavipes</i> is the only member of this genus known to occur or have occurred in southwest Australia. The sequences mapped with 100 % similarity to this species but this was also the case for other members of the genus.
Dasyuridae	<i>Dasyurus</i>	<i>Dasyurus geoffroii</i> is the only member of this genus known to occur or have occurred in southwest Australia. While the sequences mapped to this species they also mapped to other members of the genus.
Dasyuridae	<i>Sminthopsis</i>	<i>Sminthopsis</i> was only detected with 16S matching to a species not found in the region. However, there is no 16S sequence on Genbank for <i>S. griseoventer</i> which is found in the area

Dasyuridae	<i>Phascogale</i>	Two species of <i>Phascogale</i> occur in southwest Australia: <i>P. calura</i> and <i>P. tapoatafa</i> . The latter occurs in the immediate vicinity of the site. While both are present on Genbank, the sequences mapped closest in similarity to those of <i>P. calura</i> . It is not possible at this time to provide a confident taxonomic assignment given that <i>P. calura</i> is not found in the immediate vicinity of the caves.
Macropodidae	<i>Macropus</i>	Three species of <i>Macropus</i> have been found in the cave deposits: <i>M. irma</i> , <i>M. eugenii</i> and <i>M. fuliginosus</i> . <i>Macropus fuliginosus</i> is discussed in Chapter Six. Neither of the other two species of <i>Macropus</i> were detected, however this could be due to a large degree of genetic homology between different species of <i>Macropus</i> (and Macropodidae taxa in general) across 16S and 12S in some cases.
Macropodidae	<i>Setonix</i>	<i>Setonix brachyurus</i> is the only member of this genus known to occur or have occurred in southwest Australia. It is only present on Genbank for 16S. Therefore it may be the contributing factor to a large number of unidentified Macropodidae sequences for 12S. It was identified with 100 % similarity to reference sequences for 16S.
Phalangeridae	<i>Trichosurus</i>	<i>Trichosurus vulpecula</i> is the only member of this genus known to occur or have occurred in southwest Australia.
Potoroidae	<i>Potorous</i>	The sequences identified as <i>Potorous</i> matched <i>P. gilbertii</i> with 100% similarity for 12S, however, no 16S sequences for this species exists on Genbank. For 16S the sequences identified as <i>Potorous</i> matched <i>P. tridactylus</i> with only 92 % similarity and despite the presence of 12S for <i>P. tridactylus</i> on Genbank this, as explained above, was not the closest match for the sequences.

Pseudocheiridae	<i>Pseudocheirus</i>	<i>Pseudocheirus occidentalis</i> (also classified as <i>Pseudocheirus peregrinus occidentalis</i>) is the only member of this genus found in the southwest Australia.
Peramelidae	<i>Isoodon</i>	<i>Isoodon obesulus</i> is only member of this genus found in the southwest Australia but its DNA sequences have a high degree of homology with other <i>Isoodon</i> species.
Muridae		The 12S and 16S sequence representation on Genbank for southwest Australian Muridae species is practically non-existent. The representation for the mitochondrial cytochrome b gene is somewhat better. A rodent primer targeting cytochrome b was attempted, however, it performed poorly on qPCR and was not quantitative (data not shown). As such the use of this primer has been temporarily abandoned.
Muridae	<i>Rattus</i>	The only native <i>Rattus</i> to occur in the region is <i>Rattus fuscipes</i> as noted in Chapter Six. No sequences mapped to either <i>R. rattus</i> or <i>R. norvegicus</i> .
Canidae	<i>Vulpes</i>	<i>Vulpes vulpes</i> is a non-native invasive species that was only detected in stratigraphical units after European arrival. At this moment in time it is difficult to determine whether this detection is as a result of leaching or re-working of the sediment. However, the fact that it was not detected at any sites or in any layers pre-dating European arrival suggests that leaching or re-working may be an unlikely source of this DNA.
Leporidae	<i>Oryctolagus</i>	<i>Oryctolagus cuniculus</i> was detected with 100 % similarity to Genbank reference sequences. As with <i>Vulpes vulpes</i> , it was only detected in stratigraphical units after European arrival and there is little suggestion of DNA leaching or vertical movement of bone through the deposit being the source of this DNA.

Muridae	<i>Mus</i>	<i>Mus musculus</i> was detected in three stratigraphical units that pre-date European arrival with 100 % similarity. This is very likely to be contamination as it has been shown to be a "common" sporadic contaminant in PCR reagents and other laboratory consumables (Leonard <i>et al.</i> , 2007; Erlwein <i>et al.</i> , 2011; Tuke <i>et al.</i> , 2011).
Tachyglossidae		Tachyglossidae was detected in a single layer at Northcote sinkhole. This was the bottom-most layer of the deposit and the site has been dated to ~2,000 years BP. Results for both 12S and 16S for Tachyglossidae were conflicting. For 12S the sequences showed greatest similarity to <i>Tachyglossus</i> whereas for 16S they showed greatest similarity to <i>Zaglossus</i> . However, it appears that there is a great degree of homology across both species for these genes, particularly for 16S. Currently, only the genus <i>Tachyglossus</i> is found in southwest Australia. However, <i>Zaglossus</i> fossils (<i>Z. hacketti</i>) have been recovered from nearby cave sites (Augee <i>et al.</i> , 2006). These fossils have been dated to the upper Pleistocene however. Moreover the placement of <i>Z. hacketti</i> in <i>Zaglossus</i> is uncertain. While the DNA sequences detected are most likely <i>Tachyglossus</i> , a confident taxonomic assignment cannot be made at present.

References

- Augee, M. L., Gooden, B., & Musser, A. (2006). *Echidna: Extraordinary Egg-laying Mammal*. Victoria, Australia: CSIRO Publishing.
- Baird, R. F. (1991). Avian fossils from the Quaternary of Australia. In P. Vickers-Rich, J. M. Monaghan, R. F. Baird, & T. H. Rich (Eds.), *Vertebrate palaeontology of Australasia* (pp. 809-855). Melbourne, Australia: Pioneer Design Studio and Monash University Publications Committee.
- Dortch, J. (2004). *Palaeo-environmental Change and the Persistence of Human Occupation in South-western Australian Forests*. Oxford: Archaeopress.
- Erlwein, O., Robinson, M. J., Dustan, S., Weber, J., Kaye, S., & McClure, M. O. (2011). DNA extraction columns contaminated with murine sequences. *PLoS One*, 6, e23484.
- Garnett, S. T., & Crowley, G. M. (2000). The action plan for Australian birds 2000. Retrieved from <http://www.environment.gov.au/biodiversity/threatened/publications/action/birds2000/index.html>
- Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., . . . Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, 34, 1361-1366.
- Tuke, P. W., Tettmar, K. I., Tamuri, A., Stoye, J. P., & Tedder, R. S. (2011). PCR master mixes harbour murine DNA sequences. Caveat emptor! *PLoS One*, 6, e19953.

Table S6.2.3A Presence and absence of faunal taxa identified at Devil's Lair.

Order	Family	Stratigraphical unit										
		Post-LGM					LGM			Pre-LGM		
		XIV	XII	XI	X	IX	VIII	VII	IV	II		
Asparagales	Amaryllidaceae		√1									
Asparagales	Asparagaceae							√2				
Asterales	Asteraceae				√1			√1			√1	
Brassicales	Brassicaceae							√1				
Caryophyllales	Amaranthaceae					√2	√1	√2	√1			
Cupressales	Cupressaceae					√1						√1
Fabales	Fabaceae	√1						√1				
Fagales	Betulaceae							√1				√1
Fagales	Fagaceae	√1						√2				
Laurales	Lauraceae	√1										
Malpighiales	Euphorbiaceae	√1										
Malpighiales	Salicaceae							√2				
Malvales	Malvaceae											√1
Myrtales	Lythraceae							√1				
Myrtales	Myrtaceae	√1	√2	√1	√1	√2	√2	√2	√1			
Pinales	Pinaceae					√1		√1	√2			√1
Poales	Poaceae			√1	√1	√1		√1				√1
Proteales	Platanaceae						√1					
Proteales	Proteaceae							√1				
Rosales	Moraceae							√1				
Rosales	Rosaceae	√1										
Rosales	Urticaceae							√1	√1			
Sapindales	Anacardiaceae					√1		√2				
Saxifragales	Crassulaceae							√1				
Solanales	Convolvulaceae					√1						
Solanales	Solanaceae					√1		√1				√1

√1 Detected with one primer, √2 Detected with two primers

Table S6.2.3B Presence and absence of faunal taxa identified at Tunnel Cave.

Order	Family	Stratigraphical unit																	
		Post-LGM					LGM					Hearths							
		001	002	003	5-upper	5-lower	006	7-upper	7-lower	008	9-upper	9-middle	9-lower_10	F4	F5	F8	F9	F19	F20
Apiales	Apiaceae											√1							
Apiales	Pittosporaceae	√1																	
Asparagales	Amoryllidaceae							√1											
Asterales	Asteraceae	√1			√1	√1	√1	√1	√1		√1								√1
Asterales	Goodeniaceae	√2																	
Boraginales	Boraginaceae	√1						√1								√1			
Brassicales	Brassicaceae			√1				√1		√1									
Caryophyllales	Amaranthaceae	√2	√1					√1		√1									
Caryophyllales	Polygonaceae				√2	√2	√2	√2	√2	√1	√2	√1	√1	√1	√1	√2	√1	√1	√1
Cucurbitales	Cucurbitaceae				√1														
Cupressales	Cupressaceae				√1					√1									
Cupressales	Taxaceae							√1											
Dilleniales	Dilleniaceae	√1																	
Ericales	Ericaceae	√2																	
Ericales	Theaceae									√1	√1								
Fabales	Fabaceae	√2				√1	√1	√1	√1	√1	√1		√1				√1		
Fagales	Betulaceae				√1	√1	√1	√1											
Fagales	Fagaceae								√1										
Fagales	Juglandaceae						√1												
Gentianales	Loganiaceae	√2										√1							
Gentianales	Rubiaceae	√2				√1													
Hypnodendrales	Racopilaceae	√1																	
Malpighiales	Salicaceae							√1					√1						
Malvales	Thymelaeaceae	√1																	
Myrtales	Lythraceae							√1											
Myrtales	Melastomataceae						√1	√1											
Myrtales	Myrtaceae	√2	√1	√1	√1	√2	√2	√2	√2	√1	√1	√2	√1	√1	√1	√2	√1	√1	√1
Pinales	Pinaceae				√1	√2		√1	√1		√1								
Poales	Cyperaceae	√1																	
Poales	Poaceae	√1	√1		√1	√2	√2	√2	√2	√1	√1	√1							√1
Poales	Restionaceae	√1						√1											
Polypodiales	Aspleniaceae	√1			√1	√1	√1												√1
Polypodiales	Dennstaedtiaceae	√1			√1			√1											
Proteales	Proteaceae	√1			√1	√1	√1	√2				√1	√1						
Ranunculales	Ranunculaceae	√1																	
Rosales	Moraceae					√1													
Rosales	Rhamnaceae												√1						
Rosales	Rosaceae				√1														
Rosales	Urticaceae						√1											√1	√1
Sapindales	Rutaceae							√1	√1										
Solanales	Solanaceae					√1		√1	√1	√1						√1			√1
Zingiberales	Musaceae			√1															

√ 1 Detected with one primer , √ 2 Detected with two primers

Table S6.2.3B Presence and absence of faunal taxa identified at Rainbow Cave.

Order	Family	Stratigraphical unit					
		Cultural					Not-Cultural
		001	002	003	004	005	010
Alismatales	Araceae					√1	
Apiales	Apiaceae	√2	√2	√1	√2	√2	
Apiales	Araliaceae	√1					
Apiales	Pittosporaceae	√1					
Araucariales	Podocarpaceae					√1	
Asparagales	Amaryllidaceae				√1		
Asparagales	Hypoxidaceae				√1		
Asparagales	Xanthorrhoeaceae	√1	√1		√1		
Asterales	Asteraceae	√2	√2	√2	√1	√2	
Asterales	Campanulaceae		√1				
Asterales	Goodeniaceae	√2	√2	√1	√1	√1	
Brassicales	Brassicaceae	√1		√1	√2	√2	
Caryophyllales	Amaranthaceae	√2	√1	√2	√2	√2	
Caryophyllales	Caryophyllaceae			√2			
Caryophyllales	Polygonaceae	√2	√2	√2	√2	√2	√1
Cucurbitales	Cucurbitaceae			√1	√1		
Cupressales	Cupressaceae					√1	
Cycadales	Zamiaceae	√1					
Dilleniales	Dilleniaceae		√1				
Ericales	Ericaceae	√1	√1		√1	√1	
Ericales	Theaceae				√1		
Fabales	Fabaceae	√2	√1	√1	√2	√2	√1
Fagales	Fagaceae				√1		
Gentianales	Rubiaceae	√2	√2		√1	√1	
Lamiales	Lamiaceae					√1	
Lamiales	Plantaginaceae	√1	√2				
Malpighiales	Phyllanthaceae			√1		√1	
Malpighiales	Picrodendraceae		√1		√1		
Malpighiales	Salicaceae				√1		
Malvales	Thymelaeaceae		√1				
Myrtales	Melastomataceae			√1	√1	√1	
Myrtales	Myrtaceae	√2	√2	√2	√2	√2	
Myrtales	Onagraceae			√1			
Oxalidales	Connaraceae			√1		√1	
Pinales	Pinaceae		√1	√2			
Poales	Cyperaceae				√1		
Poales	Poaceae			√2	√2	√2	
Poales	Restionaceae		√1				
Polypodiales	Dennstaedtiaceae					√1	
Polypodiales	Pteridaceae				√1		
Proteales	Proteaceae	√2	√2	√1	√1	√2	
Ranunculales	Ranunculaceae	√1			√1	√2	
Rosales	Rhamnaceae	√1	√1		√2	√1	
Rosales	Ulmaceae					√1	
Santalales	Santalaceae	√1					
Sapindales	Anacardiaceae				√1		
Sapindales	Rutaceae	√2	√2	√2	√2	√2	√1
Sapindales	Sapindaceae	√1	√1		√1	√1	
Saxifragales	Crassulaceae		√1				
Solanales	Convolvulaceae		√1			√1	
Solanales	Solanaceae	√2	√2	√2	√2	√2	√1

√1 Detected with one primer, √2 Detected with two primers

Table S6.2.3C Presence and absence of faunal taxa identified at Wonitji Janga.

Order	Family	Stratigraphical unit				
		Post-Euro		Pre-Euro		
		001	003	004	006	010
Apiales	Apiaceae	√1	√1		√1	√2
Araucariales	Podocarpaceae	√1				
Asparagales	Xanthorrhoeaceae		√1			
Asterales	Asteraceae	√2	√1			√2
Asterales	Goodeniaceae	√1				
Asterales	Menyanthaceae	√1	√1			
Boraginales	Boraginaceae	√1				
Brassicales	Brassicaceae				√1	
Caryophyllales	Amaranthaceae	√1	√2	√2	√2	√2
Caryophyllales	Caryophyllaceae				√1	
Caryophyllales	Polygonaceae	√2	√2	√1		√2
Cucurbitales	Cucurbitaceae				√1	
Cupressales	Cupressaceae			√1	√1	
Cycadales	Zamiaceae	√1				
Dilleniales	Dilleniaceae	√1				
Ericales	Ericaceae	√1				
Fabales	Fabaceae	√2	√2	√1		√1
Fagales	Betulaceae			√1		
Fagales	Casuarinaceae	√2	√1			
Gentianales	Rubiaceae	√2	√2			√2
Geraniales	Geraniaceae	√1				
Lamiales	Plantaginaceae	√1				
Malpighiales	Phyllanthaceae	√1				
Malpighiales	Salicaceae					√1
Myrtales	Myrtaceae	√2	√2	√1	√1	√1
Poales	Cyperaceae	√2	√1			
Poales	Poaceae	√1	√1		√1	√1
Poales	Restionaceae	√1				
Polypodiales	Dennstaedtiaceae	√1	√1	√1	√1	√1
Pottiales	Pottiaceae	√1				
Proteales	Proteaceae	√1				
Ranunculales	Ranunculaceae	√1				
Rosales	Rhamnaceae	√1				
Rosales	Urticaceae	√1				
Sapindales	Rutaceae					√1
Saxifragales	Haloragaceae	√1				
Solanales	Convolvulaceae	√1				
Solanales	Solanaceae					√1
Vitales	Vitaceae				√1	

√1 Detected with one primer, √2 Detected with two primers

Table S6.2.4 Information regarding the taxonomic assignment of plant taxa.

Family	Comments
Amaranthaceae	Shrubs and herbs. Includes the sometimes accepted Chenopodiaceae family. Includes some plant species recognised as resources in southwest Australia including <i>Rhagodia baccata</i> which is currently common at several of the sites.
Amaryllidaceae	Perennial herbs. Includes <i>Allium</i> (onion, garlic, chives). Very likely contamination.
Anacardiaceae	Trees, or shrubs. Family not found in the southwest. However the family Nitrariaceae is in the same order and there are no <i>rbcl</i> or <i>trnL</i> sequences on Genbank for this family.
Apiaceae	Includes some plant species recognised as resources in southwest Australia.
Araceae	Likely environmental contamination as the invasive weed <i>Zantedeschia aethiopica</i> was present at some cave sites. However, there is also a single native species that occurs in the region - <i>Lemna disperma</i> . It was not possible to identify either of the genera or species.
Asparagaceae	Shrubs, or lianas, or herbs and includes a number of native genera found in the southwest
Aspleniaceae	Only the genus <i>Asplenium</i> is recorded in the area
Asteraceae	Includes some plant species recognised as resources in southwest Australia.
Betulaceae	Member of the order Fagales. No members of this family known to exist in the area. Only family known to exist in the area is Casuarinaceae
Boraginaceae	A number of native genera found in WA but most with ranges lying just outside the area of the cave sites
Brassicaceae	Trees, or shrubs, or herbs, or lianas. Includes some native species such as those in the genus <i>Stenopetalum</i>
Campanulaceae	Mostly herbs, with some trees and shrubs. Includes native genera such as <i>Lobelia</i> and <i>Isotoma</i>
Caryophyllaceae	Mostly herbs with some small trees or shrubs. No native species in the immediate area of the sites although there are some just outside the region. It is in the order Caryophyllales and there are many families within this order with varying degrees of representation on Genbank for <i>trnL</i> and <i>rbcl</i> .
Casuarinaceae	Trees and shrubs. Includes some plant species recognised as resources in southwest Australia.

Connaraceae	Trees, shrubs or lianas. No genera within this family are found in Australia and there is no indication of any invasive species from the family. Additionally, they do not seem to be highly cultivated genera within the family. The family is in the order Oxalidales which includes a number of native Australian plants found in the southwest for which database representation is patchy. This family was detected using <i>rbcl</i> and there are no <i>rbcl</i> sequences for either of the native <i>Tetralthea</i> or <i>Tremandra</i> species.
Convolvulaceae	Herbs (mostly, climbing or trailing), or shrubs, or lianas, or trees (a few) native <i>Calystegia</i> no <i>rbcl</i> or <i>trnl</i> and native <i>Dichondra</i> but this is represented on genbank
Crassulaceae	Mostly herbs and some shrubs. A number of native <i>Crassula</i> species are found in the area.
Cucurbitaceae	Mostly herbs and a few shrubs. No members of this family are found in the area, native or introduced. However, native <i>Pilostyles</i> in family Apodanthaceae does occur in the area and there is no reference <i>rbcl</i> or <i>trnL</i> sequences on Genbank.
Cupressaceae	Includes the genus <i>Callitris</i> . No apparent recording of any invasive species from the family occurring in the area. <i>Callitris acuminata</i> is native to the area but there is no <i>rbcl</i> or <i>trnl</i> reference on Genbank.
Cyperaceae	Herbs. A large number of native species occur in the area with varying degrees of representation on Genbank.
Dennstaedtiaceae	Most likely the native <i>Pteridium</i> which is also recognised as a resource in southwest Australia.
Dilleniaceae	Trees, shrubs, and lianas, or a few herbs. Most likely the genus <i>Hibbertia</i> as it is the only one occurring in the area. A large genus with many species that have varying degrees of representation on Genbank.
Ericaceae	Small trees or shrubs. Includes some plant species recognised as resources in southwest Australia.
Euphorbiaceae	Trees, shrubs, herbs, lianas. Many species across a number of genera found natively in the area including those in <i>Euphorbia</i> and <i>Amperea</i> . Species have varying degrees of representation on Genbank.
Fabaceae	Includes some plant species recognised as resources in southwest Australia include <i>Acacia</i> species which may prove also to be good indicators of environmental change in future studies. Species have varying degrees of representation on Genbank.
Fagaceae	Trees and shrubs. Member of the order Fagales. No members of this family known to exist in the area. Only family known to exist in the area is Casuarinaceae

Geraniaceae	Mostly herbs or shrubs. A few native species such as those in the genera <i>Pelargonium</i> and also a few natives in the genus <i>Geranium</i> .
Goodeniaceae	Herbs, shrubs or some trees. Several native species in the area including those in the genus <i>Dampiera</i> . Species have varying degrees of representation on Genbank.
Haloragaceae	Several native species in the genera <i>Gonocarpus</i> and <i>Haloragis</i> , amongst other, some of which have no <i>trnL</i> or <i>rbcl</i> sequences on Genbank
Hypoxidaceae	Perennial herbs. <i>Pauridia</i> is the mostly likely genus as it is the only one recorded in the area.
Juglandaceae	Trees and some shrubs. Member of the order Fagales. No members of this family known to exist in the area. Only family known to exist in the area is Casuarinaceae
Lamiaceae	Large family with a number of species found natively in the region. Species have varying degrees of representation on Genbank.
Lauraceae	Trees and shrubs. Only one genus known in the area - <i>Cassytha</i> .
Loganiaceae	Herbs or shrubs. Most likely <i>Logania</i> or <i>Phyllangium</i> species as these are the only genera in the area.
Lythraceae	Not recorded in the region. Within the order Myrtales and only one family in the order occurs in the area - Myrtaceae. It could represent error or a lack of Genbank references for members within the family
Malvaceae	Herbs, shrubs and some trees. A few native species in genera such as <i>Lasiopetalum</i> and <i>Alyogyne</i>
Melastomataceae	Not recorded in the region. Within the order Myrtales and only one family in the order occurs in the area - Myrtaceae. It could represent error or a lack of Genbank references for members within the family
Menyanthaceae	Herbs. The only recorded genera in the area are <i>Liparophyllum</i> and <i>Ornduffia</i>
Moraceae	No native species recorded in the area. Likely contamination possibly from <i>Ficus carica</i> .
Musaceae	Very large herbs. No recorded sightings of family in southwest Australia. No recorded sightings of any of the other families within the same order, Zingiberales, either.
Myrtaceae	Trees and shrubs. Includes some plant species recognised as resources in southwest Australia. Includes <i>Eucalyptus</i> species, discussed in Chapter Six
Onagraceae	Shrubs and herbs, or trees. A number of native <i>Epilobium</i> species are known to occur in the area.
Phyllanthaceae	A number of native species found in the area including those from the genera <i>Poranthera</i> and <i>Phyllanthus</i>

Picrodendraceae	Trees. No species from this family recorded in the area, however it is part of the order Malpighiales which includes, amongst others, the families Phyllanthaceae and Euphorbiaceae
Pinaceae	This family occurred in controls and as such was excluded from any analyses. It likely represents laboratory contamination as it has been encountered in a number of other projects.
Pittosporaceae	Trees, shrubs, and lianas. Includes some plant species recognised as resources in southwest Australia.
Plantaginaceae	Mostly herbs and some shrubs. Includes native <i>Gratiola</i> but also non-native <i>Plantago</i> .
Platanaceae	Not recorded in the area. Proteaceae is within the same order and as such it could represent error or degradation.
Poaceae	Many native species that occur in the area and further resolution within this family may be improved with a family-specific primer approach.
Podocarpaceae	Includes some plant species recognised as resources in southwest Australia.
Polygonaceae	Mostly herbs and trees, shrubs or lianas. Includes native genera such as <i>Persicaria</i> and <i>Muehlenbeckia</i> that are found in the area.
Pottiaceae	Family found in the area and includes native species in genera such as <i>Barbula</i> .
Proteaceae	Trees, shrubs or herbs. Recognised as a potential resource plant in southwest Australia.
Pteridaceae	Family found in the region and includes native genera such as <i>Cheilanthes</i> and <i>Adiantum</i>
Racopilaceae	Recorded in the area, little information on the Family and genus in the area (<i>Racopilum</i>).
Ranunculaceae	Mostly herbs with some shrubs or lianas. Includes some plant species recognised as resources in southwest Australia.
Restionaceae	Herbs. Includes many native genera such as <i>Leptocarpus</i> and <i>Chaetanthus</i>
Rhamnaceae	Trees, shrubs, lianas, or herbs. Many native species recorded in the area including those in the genus <i>Cryptandra</i> .
Rosaceae	Trees, or shrubs, or herbs. Only one native species (<i>Acaena echinata</i>) recorded in the southwest and a few invasive species.
Rubiaceae	Trees and shrubs, lianas or herbs. Family includes the southwest native <i>Opercularia</i> genus
Rutaceae	Mostly trees and shrubs and some herbs. Includes some plant species recognised as resources in southwest Australia.
Salicaceae	Not recorded at the cave sites. However, it has been detected as contamination in previous studies.

Santalaceae	Trees, shrubs, and herbs. Includes some plant species recognised as resources in southwest Australia.
Sapindaceae	Trees, shrubs, lianas or herbs. Includes the native <i>Dodonaea</i> genus
Solanaceae	Herbs, shrubs, trees, and lianas. Includes some plant species recognised as resources in southwest Australia however it is also a commonly encountered laboratory contaminant.
Taxaceae	No recordings of this family in Western Australia. It is in the order Pinales and as such could represent contamination or error associated with Pinaceae
Theaceae	No recordings of this family in Western Australia. It could represent laboratory or environmental contamination as it contains the genus <i>Camellia</i> , which includes <i>C. sinensis</i> which is used to make tea
Thymelaeaceae	Mostly shrubs and occasional lianas and herbs. Includes some plant species recognised as resources in southwest Australia.
Ulmaceae	Trees and shrubs. Only one invasive species recorded in Western Australia and not near the cave sites.
Urticaceae	Shrubs, lianas, and herbs or trees. Could be the native <i>Parietaria debilis</i> as it is the only species of this genus recorded in the area.
Vitaceae	This is very likely to be contamination as it has been picked up in controls previously.
Xanthorrhoeaceae	Shrubs. Includes plant species recognised as resources in southwest Australia (<i>X. gracilis</i> and <i>X. preissii</i>).
Zamiaceae	Only one species found in southwest Australia (<i>Macrozamia riedlei</i>) and it is recognised as a plant resource.

File S6.2.1 Ancient DNA extraction from sediment protocol

DNA EXTRACTION FROM ANCIENT SEDIMENT

INTRODUCTION

This document details the method to be employed for extracting DNA from ancient sediment using a phenol/chloroform method. It is highly recommended for ancient sediments with little DNA or that are extremely inhibited.

PRECAUTIONS

Lab users must wear a laboratory coat, safety glasses, covered shoes, facemask (if required) and clean disposable gloves while doing DNA extractions within the Green Lab (311.212). Lab users must read SOP 0.3 TrACE Facility Entry & PPE Requirements on the correct PPE and suiting up protocol for DNA extractions within the TrACE Facility (206.203). For quarantine samples, please refer to SOP 2.2 and 3.06 for handling and DNA extraction. All work with phenol and/or chloroform must be conducted in a fume hood with waste clearly marked and disposed of appropriately in purple cytotoxic waste bin at Building 311. Care must be taken when disposing of Dabney Binding Buffer (SOP 1.26) as it contains guanidine hydrochloride, which adversely reacts with bleach. Lab users must read the MSDS for all reagents to ensure that any additional precautions are taken.

CONSUMABLES, REAGENTS & EQUIPMENT

Sediment Digest Buffer	50mL Falcon tubes (lysis matrix optional)
Phenol (buffered, pH 7.2)	50mL Light Phase Lock Tubes
Chloroform	Vivaspin columns (Sartorius)
10mg/mL Proteinase K	MinElute columns (Qiagen)
Dabney Binding Buffer	1.5mL and 0.5mL Eppendorf tubes
Buffer AW1 (Qiagen)	Parafilm
Buffer AW2 (Qiagen)	Tris/Tween Solution

METHOD

1. Add 15mL of Sediment Digest Buffer to 5g of sediment in a 50mL Falcon tube (with or without lysis matrix). Seal the tube with parafilm.
2. Vortex for 30 seconds at maximum speed.

3. Incubate with continuous gentle mixing (setting 6) at 55°C overnight in the hybridization oven (15 hours total).
4. Remove the sample from the oven and add 200µL of 10mg/mL Proteinase K.
5. Vortex tube for 30 seconds at maximum speed.
6. Place tubes in hybridization oven for a final 1 hour digestion at 55°C with continuous gentle mixing (setting 6).
7. Prior to using the Phase Lock Tubes, they must be spun in a centrifuge for 10 minutes at maximum speed.
8. Remove the sample from the oven and centrifuge for 10 minutes at maximum speed and transfer the supernatant to a clean 50mL Phase Lock tube. Standard Falcon tubes can be used instead of Phase Lock tubes if necessary.
9. Add an equal volume of Phenol to the tube and incubate for 5 minutes under continuous gentle mixing (setting 4) at room temperature in the hybridization oven. Seal the tube with parafilm to prevent any leaking.
10. Remove the tube from the hybridization oven and centrifuge for 5 minutes at 4,000g (or maximum rpm) to allow the phases to separate.
11. Repeat steps 9-10 once more with an equal volume of Phenol. The same Phase Lock tube may be used again provided the maximum allowable volume has not been reached. However, if using a standard Falcon tube a new standard Falcon tube must be used for this step.
12. Repeat Steps 9-10 once more with an equal volume of Chloroform using a new Phase Lock (or standard Falcon) tube.
13. Transfer the DNA aqueous layer (top layer), to a 30kDa Vivaspin 20 Centrifugal Filter (maximal volume 14mL) and centrifuge at 8,000g until the desired volume of extract is left in the filter (e.g. ~1 hour for 50µL). Please consult user guide if using a swing bucket rotor centrifuge for an equivalent speed.
14. Prepare the Dabney Binding Buffer according to SOP 1.26 and mix well.
15. Once the desired volume of concentrate has been achieved, transfer the concentrate to a clean 1.5mL Eppendorf tube.
16. Aliquot an appropriate volume of Dabney Binding Buffer to the tube for a 13:1 ratio of binding buffer to DNA extract (e.g. to 50µL of DNA extract add 650µL of binding buffer).
17. Mix thoroughly, but gently, and allow the mixture to incubate for 5 minutes at room temperature.

18. Pipette the mixture into a Qiagen MinElute column and centrifuge for 1 minute at 13,000 rpm and then discard the flow-through. Keep the flow-through tube for use in steps 19-21.
19. Place the column back into the flow-through tube. Wash the MinElute column with 500 μ L of AW1, incubate for 2 minutes, centrifuge for 1 minute at 13,000 rpm and discard the flow-through.
20. Place the column back into the flow-through tube. Wash the MinElute column with 500 μ L of AW2, incubate for 2 minutes, centrifuge for 3 minutes at 13,000 rpm and discard the flow-through.
21. Place the column back into the flow-through tube. Centrifuge the MinElute column for 1 minute at 13,000 rpm to dry the membrane.
22. Discard the flow-through tube and place the column in a new 1.5mL Eppendorf tube with the lid cut off.
23. Add 20 μ L of Buffer EB directly to centre of the membrane and incubate at room temperature for 5 minutes. To enhance DNA recovery, the EB buffer should be heated to 37°C prior to use.
24. Centrifuge the MinElute column for 1 minute at 10,000 rpm to elute the DNA.
25. Transfer the eluted DNA to a new 0.5 μ L low-bind tube and place the column back on the tube used previously.
26. Repeat steps 23-24 once more, transferring the eluted DNA to the same 0.5 μ L low-bind tube.
27. Store at -20°C.

PREPARATION OF SEDIMENT DIGEST BUFFER

INTRODUCTION

This document details the method to be employed for preparation of 100mL of Sediment Digest Buffer Stock and 40mL of Sediment Final Digest Buffer. Sediment Digest Buffer Stock may be stored for up to 1 week and Sediment Final Digest Buffer should be used immediately.

PRECAUTIONS

Lab users must wear a laboratory coat, safety glasses, covered shoes, facemask (if required) and clean disposable gloves while preparing solutions and buffers. Lab users must read the MSDS for all reagents to ensure that any additional precautions are taken.

PROPERTIES OF SEDIMENT DIGEST BUFFER STOCK

2% SDS	0.02M EDTA
0.05M Tris/HCL	1.5M Sodium Chloride

ADDITIONAL PROPERTIES OF SEDIMENT FINAL DIGEST BUFFER

0.05 M DTT	1mg/mL Proteinase K
------------	---------------------

CONSUMABLES, REAGENTS & EQUIPMENT

10% SDS	1M DTT (SOP 1.08)
1M Tris/HCl	Proteinase-K
0.5M EDTA	100mL Schott Bottle
5M Sodium Chloride	50mL Falcon tube

METHOD

1. To make Sediment Digest Buffer Stock, combine the following in a 100mL Schott bottle:

- 20mL SDS
- 5mL Tris/HCL
- 4mL EDTA
- 3mL Sodium Chloride

Fill to 100mL with ultrapure water.

2. Mix thoroughly and place in the UV oven for 1 hour to sterilize.
3. To make Sediment Final Digest Buffer, combine the following in a 50mL Falcon tube:
2.01mL DTT
40mg Proteinase K
Fill to 50mL with Sediment Digest Buffer Stock.
5. Incubate with rotation for 5 minutes at 55°C to activate enzymes prior to addition to samples.

PREPARATION OF DABNEY BINDING BUFFER (DBB)

INTRODUCTION

This document details the method to be employed for preparation of 13mL of Dabney Binding Buffer (DBB) (Dabney *et al.*, 2013). If the volume of DNA extract is 50 μ L and a 13:1 ratio of DBB to DNA extract is used this will provide enough for 20 reactions (i.e. 650 μ L of DBB for each 50 μ L DNA extract). This buffer may be prepared a day prior to its use.

PRECAUTIONS

Lab users must wear a laboratory coat, safety glasses, covered shoes, facemask (if required) and clean disposable gloves while preparing solutions and buffers. Lab users must read the MSDS for all reagents to ensure that any additional precautions are taken. DBB contains Guanidine Hydrochloride (GuHCl) and as such **MUST NOT** be mixed with bleach. Please dispose of waste buffer, tips and any other consumables carefully and appropriately. Thoroughly remove DBB from utensils and/or waste buckets prior to bleaching.

PROPERTIES OF THE FINAL BUFFER

5M Guanidine Hydrochloride

40% Isopropanol

0.05% Tween-20

90mM Sodium Acetate

CONSUMABLES, REAGENTS & EQUIPMENT

Guanidine Hydrochloride powder (MW 95.53 g/mol)	15mL Falcon tube
100% Isopropanol	Ultrapure water
100% Tween-20	P100 Positive
Displacement Pipette	
Sodium Acetate Powder (MW 82.03 g/mol, pH 5.2) or 3M Sodium Acetate Solution (pH 5.2)	

METHOD

1. To make 13mL of DBB add the following, in order, to a 15mL Schott Bottle:
6.209g Guanidine Hydrochloride Powder

5.2mL Isopropanol

6.5µL Tween-20

0.095g Sodium Acetate Powder **or** 390µL Sodium Acetate Solution

Fill to a total volume of 13mL with ultrapure water.

2. Mix thoroughly by inverting.

Dabney, J., Knapp, M., Glocke, I., Gansauge, M. T., Weihmann, A., Nickel, B., . . . Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 15758-15763.

6.3 Synopsis

Chapter Six represents one of the few ancient sedimentary DNA studies of archaeological deposits in Australia. The ability to extract and characterise plant DNA from cave sediments in Australia represents a significant challenge; not least because it is in an area of high diversity and species endemism. However, it provides an insight into past plant assemblages that would otherwise not be possible as the sites studied possess scant pollen records and the preservation of non-woody plants is virtually nil. While challenges still remain in the genetic identification of plants it is possible in some cases to partially circumvent this. The ability to identify plant DNA will no doubt improve with increased effort to populate the patchy genetic reference databases that exist for the region as well as attempts to solidify the underpinning taxonomic framework.

The identification of animal DNA from fragmentary bone, while not perfect, proved much more straightforward than that for plants. As in previous studies utilising the BBM method, it was capable of providing greater insight into lesser studied taxa at the archaeological sites selected.

The successful genetic profiling of five cave sites across the Australian southwest bodes well for future studies into ancient sedimentary DNA and BBM studies and it has raised several avenues for future research which are discussed in more detail in the future directions section (Chapter Seven).

Chapter Seven – General discussion and future directions of environmental metabarcoding

7.1. Preface

The manuscripts presented in this thesis showcase the applicability of ancient and degraded DNA to study past and present Australian ecology. The rapid evolution of sequencing technology and the gradual appreciation of the pitfalls associated with sample preparation, DNA sequence generation and HTS data analysis are evident throughout this thesis with each manuscript building upon the previous in an iterative process in an attempt to develop wet-lab and bioinformatic workflows suitable across a variety of applications.

The following general discussion seeks to connect the major discussion points raised across the manuscripts in this thesis in light of recent literature since their initial publication. Additionally, future prospects in the application of HTS to address spatial and temporal ecological shifts are offered.

7.2. General discussion

At the outset of this thesis few environmental metabarcoding studies, ancient or modern, had been conducted in Australia and, despite its fragile ecosystems and unique biota, to some extent this is still the case. The research papers presented, in addition to those co-authored during candidature (Appendix III), illustrate the practicality of environmental metabarcoding studies in Australia and other similar regions whose climatic conditions pose problems for the long-term survival of DNA. The findings from this thesis addressed some key questions in relation to the genetic analysis of ancient and degraded DNA in environmental samples. It was the overarching goal of this thesis research to provide empirical environmental metabarcoding data across a number of different biological substrates. Together with some commentary on future directions, this discussion summarises a number of the main findings of this thesis research.

7.2.1 Is quantitative HTS data possible?

Some of the earliest studies using environmental metabarcoding were those applying it to the analysis of faecal material. An early concern raised was whether or not reliable estimates of prey abundance could be determined from predator scats using HTS (Deagle *et al.*, 2010). However, it hadn't been determined if the proportion of DNA sequences obtained using HTS actually reflected the proportion of DNA sequences contained within the extract itself.

Through the use of qPCR, it was determined that the proportion of sequences obtained via HTS correlated significantly with those of qPCR thus proving that quantitative estimates of DNA sequences within a DNA extract using HTS are possible. This analysis did not set out to investigate whether prey abundance as determined by HTS reflected that consumed – other studies had established that molecular quantitative estimates of diet were likely semi-quantitative (Bowles *et al.*, 2011). Since the publication of these papers, further studies have shown correlations between HTS data and the proportion of taxa in water and in sediment for both animals and plants (Andersen *et al.*, 2012; Thomsen *et al.*, 2012; Yoccoz *et al.*,

2012). However, limitations to quantitative estimates of taxonomic diversity and abundance do still exist (Bohmann *et al.*, 2014).

7.2.2 Is plant DNA preserved in Australian middens?

A major strength of environmental metabarcoding is its ability to detect DNA from taxa that are difficult to identify via morphological means: be that due to degradation as it transits through the gastrointestinal tract or taphonomic processes affecting fossilisation. For this reason, environmental metabarcoding has particular relevance in Australia where pollen records going back in time are virtually non-existent for some areas, e.g. the Pilbara region in north-west Australia (Dortch, 2004b; Denham *et al.*, 2009).

Using a multi-locus universal primer approach, ancient plant and animal DNA was extracted from four arid-zone middens: three Australian and one South African. The successful extraction of aDNA from midden material from Australia's Pilbara region represents some of the oldest DNA extracted from Australian samples to date and was the first to successfully extract and characterise ancient plant DNA in Australia. While the successful extraction of aDNA from midden material from elsewhere in the world had been reported (Kuch *et al.*, 2002; Hofreiter *et al.*, 2003) it was uncertain whether this would be the successful for midden material from Australia where aDNA preservation had not been explored. However, this study also highlights the difficulties in identifying the extracted plant DNA past a family level; a non-trivial issue (discussed further in Section 7.1.6). Nonetheless, this study demonstrates that successfully extracting aDNA from old environmental samples in Australia is possible and may provide insights into past environments if issues associated with the taxonomic assignment of DNA sequences can be overcome.

7.2.3 Is bulk-bone metabarcoding feasible?

One of the challenges associated with aDNA analyses in archaeology and palaeontology is the destructive nature of sampling and this, rightly so, leads to apprehension in allowing bones to be sampled when there is a high probability of non-success. However, archaeological and palaeontological excavations generate a

large volume of fragmentary bone that is largely overlooked and sometimes discarded when taxonomically profiling fossil assemblages.

With the advent of HTS and the rapid evolution of metabarcoding workflows, it was possible to create, what were in essence, synthetic environmental samples composed of bone powder from ground fragmentary fossil bone collected at cave sites in southwest Australia. This proved to be a remarkably quick and efficient means by which to analyse hundreds of bone fragments as opposed to analysing each one individually. It provided an overview of taxonomic diversity at the study sites reflecting that determined by morphological analyses. Importantly, however, it was also possible to identify taxa for which a decision was made not to identify them traditionally due to morphological similarities. Further refinement of the method revealed the strengths and limitations of it (Haouchar *et al.*, 2013; Grealy *et al.*, 2015, both co-authored publications included in Appendix III) and have detailed its potential as a means to examine the extent of DNA preservation and damage at palaeontological sites (Grealy *et al.*, 2016). However, as with the analysis of plant DNA in Chapter Three, difficulties arose in identifying certain taxonomic groups thus revealing yet another large gap in the ability to provide thorough taxonomic profiles of environmental samples in Australia – this time for certain animal groups.

7.2.4 Metabarcoding workflows: is it time to change focus?

The first decade of environmental metabarcoding studies has progressed at a blistering pace and while some exciting results have been published, a number of challenges have emerged. The early focus in the field was in finding efficient and effective ways to cope with the ever-increasing amount of data produced with each new HTS platform. This focus in some ways overshadowed discussions on the challenges associated with the generation of robust datasets at the lab bench.

Based on empirical data presented in Chapters Two-Four, Chapter Five attempted to openly discuss some important considerations for experimental design in terms of sample screening and amplicon generation while at the same time resisting an urge to call for a rigid set of guidelines or “authentication criteria.” An important aspect of all HTS workflows is in ensuring that extracts are screened appropriately for

inhibition and DNA copy number. This is particularly pertinent as the field moves increasingly towards efforts to quantify occurrence or abundance using presence/absence or actual DNA concentrations (Bohmann *et al.*, 2014; Doi *et al.*, 2015). Such analyses will not be possible without adequate screening of samples to limit PCR stochasticity associated with inhibited or low copy number extracts. This holds true regardless of the number of replicates performed in a study, as the chance of detection will still be reduced and there is no satisfactory way of determining the “correct” number of replicates a taxon or OTU must occur in to be deemed as “present,” despite some attempts to accomplish this (Ficetola *et al.*, 2008; Ficetola *et al.*, 2014; Robasky *et al.*, 2014; Smith & Peay, 2014). For instance, if the correct number of replicates is determined by means of using a synthetic blend of single-source samples, unless this blend is as inhibited, or as low in DNA amount, as the actual sample being screened, any attempt to determine the minimum number of replicates will not be immediately comparable between the two.

7.2.5 Is metabarcoding useful in Australian biodiversity assessment?

Australia is a diverse continent with a wealth of unique animals and plants, yet to date it has been largely overlooked in aDNA and eDNA studies. It represents an extremely challenging environment to characterise genetically; not least because of its hot climate. In terms of biodiversity, the southwest corner of Australia is a rich but poorly characterised region that is under increasing threat from land clearance, invasive species and climate change.

The environmental metabarcoding of bulk-bone and sediment samples collected from five caves in Australia’s southwest proved extremely challenging. Nonetheless, it was still possible to analyse biodiversity through time using a largely taxonomy-independent approach that supported previous findings of a change in forest habitat at the site around the end of the LGM (Dortch, 2004b, 2004a; Dortch & Wright, 2010). Additionally, it revealed some intriguing results that warrant further investigation. Chapter Six of this thesis highlights the ability of environmental metabarcoding to generate hypotheses for future study and represents one of the most comprehensive environmental genetic audits of sediment and bone in Australia. However, as with previous chapters, in this exploratory study, difficulties arose in

taxonomically assigning DNA sequences beyond a family level for plants, as well as many faunal taxonomic groups.

7.2.5 Is fine taxonomic resolution possible?

The ability to assign sequences taxonomically has been a major limitation in the papers presented and it is an issue that is not easily remedied. To assign taxonomy to sequences effectively and confidently a well-populated reference database must exist and the underpinning taxonomic framework must be robust. Unfortunately, for Australian plants and animals, available genetic reference databases are far from adequate, particularly for vascular plants. Likewise, there still remains a lot of uncertainty in the systematics of Australian flora and, to some extent, fauna. Despite some efforts to populate databases, the representation of the generally accepted animal barcode – *COXI* – falls well short with, for example, approximately 20 % of Australian metatherians represented. Due to the rarity of some Australian flora and fauna, the granting of permits for sampling with the goal of building a genetic reference database of local taxa for environmental metabarcoding studies is not straightforward or common, although this may improve as the discipline becomes more established. Missing taxonomic groups, e.g. species, genera and families, on the database can not only lead to misidentifications but may also result in the loss of data as genuine sequences may map to the closest existing taxa on the database below the imposed similarity cut-off (Quéméré *et al.*, 2013). A further issue associated with some publically accessible databases such as Genbank is the fact that they are not curated and as such may suffer from incorrect taxonomic assignments of uploaded sequences (Nilsson *et al.*, 2006). While a simple solution to some of these issues is to create a custom, curated database, possibly of vouchered local specimens, such a strategy in southwest Australia may not be entirely feasible due to its large amount of uncharacterised diversity. As such, for the time being at least, limiting assignments to higher taxonomic levels and the use of OTU approaches, while not ideal, may be the most appropriate strategy (e.g. Burgar *et al.*, 2014, co-authored paper included in Appendix III).

Concomitant with the above issues is that the commonly accepted or “approved” DNA barcodes for both plants and animals are wholly inadequate for most

environmental metabarcoding studies that necessitate short hypervariable barcodes flanked by regions of conservation capable of identifying a breadth of taxa (Hollingsworth *et al.*, 2011; Deagle *et al.*, 2014; Staats *et al.*, 2016). Overcoming this attachment to a *COXI* barcode for animals and a combination of *matK/rbcL* for plants is essential to enable the full potential of environmental metabarcoding to be realised. The difficulty of selecting an appropriate gene region for which to design universal primers, however, is also acknowledged and the skewing of taxonomic profiles towards certain taxa is possible (Aird *et al.*, 2011; Linacre & Tobe, 2011; Schloss *et al.*, 2011; Lee *et al.*, 2012). A strategy to overcome this may be to use a multi-locus approach thus improving the breadth of taxa covered and limit the impact of primer skews. This may also serve to further support taxonomic identifications in cases where taxa are identified using both primers. However, an understanding of the taxa that are present on and absent from the databases being used is critical for both single- and multi-locus approaches. It is not possible, in most instances, to confidently assign a sequence to a given species unless all of the species within that genus occurring in the study area are all represented on the database being used. An important example of this is evident within the genus *Eucalyptus*. All *Eucalyptus* species in the study area in Chapter Six are identical in sequence length and base composition for both *trnL g/h* and *rbcL h1aF/h2aR*, rendering it impossible to differentiate them at a species level. If for instance, only one of these species were represented on Genbank, the *Eucalyptus* sequences would assign (using BLAST) to this species. However, clearly this assignment would be incorrect as, with the benefit of knowing the sequences for the other species of *Eucalyptus*, it is not possible to differentiate them using either primer set. Therefore, while universal primer approaches are useful they do have limitations and in some cases perhaps a universal primer approach followed by a targeted assay using family or genus specific primers may be a preferable strategy. This would enable the identification of samples for which the target genus or family occurs and enable the appropriate selection of samples for further analysis. However, *in silico* analyses to ensure primer bias is not impacting on the detection of the target family or genus is essential (Ficetola *et al.*, 2010).

7.3. Future directions in environmental metabarcoding

The use of HTS coupled with environmental metabarcoding is a powerful tool with which to study both past and present ecosystems. Over the relatively short tenure of this thesis research, the field has already undergone a rapid development and is gradually becoming an accepted facet of present-day ecosystem management and conservation research (Leonard, 2008; Bohmann *et al.*, 2014; Kelly *et al.*, 2014). Also, through the use of historical and ancient environmental samples it has led to a revision of our understandings of past ecosystems (de Vernal & Hillaire-Marcel, 2008; Haile *et al.*, 2009) and complemented or supported palynological and macrofossil analyses (Jørgensen *et al.*, 2011; Jørgensen *et al.*, 2012; Parducci *et al.*, 2013; Willerslev *et al.*, 2014), but it has also proven controversial (Birks *et al.*, 2012; Parducci *et al.*, 2012; Birks & Birks, 2016). Despite the recent advancements within the discipline, there remain many fundamental questions to address with regards to its accuracy, robustness and utility. Many of these issues are to be expected in an emerging field that is yet to define what constitutes best-practice across the many aspects of environmental metabarcoding workflows. In this discussion of future directions some outstanding issues and possible future directions are addressed.

7.3.1 Environmental sample handling and screening

As discussed in Chapter One, the origins of eDNA are many and questions remain with regard to DNA leaching, especially in relation to sediment samples. The extent of DNA leaching can have major implications for future environmental metabarcoding studies. In the context of the papers presented, there appeared to be little to no reworking of midden material (Chapter Three) or stratigraphical units (Chapters Four and Six) and there were clear distinctions between certain important periods in time. However, this does not remove the possibility of modern taxa contributing to DNA leaching and impacting results on the lower layers of deposits. Leaching to depths as low as 70 cm have been reported (Andersen *et al.*, 2012), however, this was conducted in zoo enclosures where natural animal ranges are reduced and there may be a biased accumulation of modern eDNA from taxa. Moreover, leaching was only observed in the elephant enclosure and below tiger latrines; both situations that are incredibly different to those encountered at the cave

sites in Chapter Four and Six. Nevertheless, similar studies into leaching of eDNA associated with Australian fauna may prove beneficial to future environmental metabarcoding studies in Australia as the extent of leaching of modern DNA may be dependent on a range of factors including soil characteristics, volume of species' excreta, intensity of grazing or roaming range and possibly even environmental conditions such as temperature.

A major consideration in environmental metabarcoding studies is the adequate screening of samples prior to amplicon generation and sequencing (Chapter Five). Screening methods such as qPCR are advocated with the intention of determining levels of inhibition and DNA copy number which can exert significant impacts on both qualitative and quantitative estimates of taxonomic diversity. The emergence of digital droplet PCR (ddPCR), whereby PCR reactions are partitioned into thousands of nanolitre sized droplets (Hindson *et al.*, 2011), offers a potential solution to some of the issues surrounding inhibition, primer bias and accurate quantitation. It has been used successfully to estimate the abundance and biomass of carp and is reportedly more accurate than qPCR, particularly when DNA concentration is low (Doi *et al.*, 2015). Moreover, it may be possible to generate amplicon libraries using ddPCR, breaking the emulsions post ddPCR and extracting the amplicon products for sequencing.

A final consideration in sample handling and preparation centres around the issue of contamination. Many studies have reported extensive contamination of laboratory reagents with animal and plant DNA (Malmström *et al.*, 2005; Leonard *et al.*, 2007; Champlot *et al.*, 2010; Hofreiter *et al.*, 2010; Jørgensen *et al.*, 2011; Boessenkool *et al.*, 2012; Pedersen *et al.*, 2013; Pedersen *et al.*, 2014; Willerslev *et al.*, 2014; Thomsen & Willerslev, 2015). This is of great concern in environmental metabarcoding and one that may become increasingly apparent with increased depths of sequencing. Low-level contamination may have to be accepted as a routine aspect of environmental metabarcoding studies and strategies developed to account for and eliminate it – particularly when dealing with apparently frequent contaminants such as *Homo*, *Sus* and *Bos* for animals and *Pinus*, *Salix* and *Solanum* for plants (Pedersen *et al.*, 2014). Such strategies may involve the creation of laboratory specific contamination databases consisting of both OTUs and those sequences for which

taxonomic assignment is possible (Porter *et al.*, 2013). The idea of creating a universal common contamination database, in the style of a publicly accessible database, may be inadvisable. The source of contamination may be difficult to determine, i.e. whether it arose from unique laboratory practices that are in place within a particular lab or a more general lack of safeguards to limit contamination at reagent processing plants. Controls along the data generation pipeline are non-negotiable and the sequencing of fresh stock reagents on arrival may be necessary, in addition to the routine sequencing of those with long shelf lives that are used frequently, e.g. primer stocks. Adding to this, it may be worthwhile, where possible, to segregate laboratory reagents among projects or even researchers to limit cross-contamination upon reagent arrival. Contamination control is vital for the validity of results and to foster trust in environmental metabarcoding data and the reporting of contamination should not be sidestepped as this will act as an impediment to gaining an appreciation of the extent of the issue.

7.3.2 Generation of HTS data

Most environmental metabarcoding studies to date have focused on biodiversity estimates or profiles, i.e. community level exploration. In Chapter Four and Six, both *Bettongia* (woylie) and *Setonix* (quokka) were routinely detected and both hold potential in exploring environmental metabarcoding applications beyond the community level. Recent studies in southwest Australia using the mitochondrial control region have shown a massive genetic decline and loss of connectivity in *Bettongia* using both historical and modern samples (Pacioni *et al.*, 2015). It is a natural extension to explore the potential of analysing bulk-bone to determine whether it is possible to detect the same decline and loss of connectivity. Alternatively, the signal-to-noise ratio arising from PCR and sequencing error may be too great in a synthetic heterogeneous sample of fossil bone to detect such changes. If it proves to be the case that it is possible, then it may result in the possibility of exploring similar effects in other taxa such as *Setonix* whose current distribution is highly fragmented with little inter-population genetic flow (Alacs *et al.*, 2003; Hayward *et al.*, 2003).

The identification of key species that serve as useful environmental proxies and indicators of human occupation proved difficult in this thesis. As such, an important area of future research must involve the discovery of suitable genetic regions to distinguish between, for example, species of *Eucalyptus*, *Acacia* and *Macropus*. Unsuccessful attempts (data not shown) were made to address this issue through the use of the ITS region for both *Eucalyptus* and *Acacia* and the use of a previously published marsupial mini-barcode (Grealy *et al.*, 2016). To effectively address these issues, vouchered specimens for some plants must be sourced to complete genetic reference databases for other regions of the plant genome – chloroplast and possibly multi-copy nuclear genes.

A final consideration worth mentioning is the application of shotgun sequencing which has been used successfully in previous environmental metabarcoding studies to effectively analyse diet (Srivathsan *et al.*, 2014) and to construct the full mitochondrial genome of *Pagophilus grienlandicus* (harp seal), albeit at low coverage (Seersholm, 2015). Such an approach, particularly for plants, may allow the detection of certain regions of the chloroplast genome that are useful for discriminating species but for which it is difficult to design primers to target. However, the high level of non-endogenous bacteria would pose a large hurdle in terms of sequence coverage. This issue is not insurmountable though and methods to enrich samples for specific targets (e.g. ‘bait capture’), or to exclude viral or bacterial DNA even, could enhance the detection of target taxa in an environmental sample (Thurber *et al.*, 2009; Feehery *et al.*, 2013).

7.3.2 Analysis of HTS data

The processes involved in generating HTS data introduce obvious biases into the data that can complicate data analysis (Coissac *et al.*, 2012; Philippe *et al.*, 2015). At present most studies tend to employ arbitrary cut-offs in relation to the abundance of sequences or percentage similarity to a given taxon to eliminate sources of PCR error and artefacts as well as sequencing error (Coissac *et al.*, 2012; Philippe *et al.*, 2015), however, exceptions do exist (Andersen, 2014; Callahan *et al.*, 2016). The sources of sequencing error are reasonably well established. In the case of eDNA sourced from “ancient” samples the damage pattern and method of degradation are reasonably well

covered also. However, the causes of eDNA damage and error for modern samples are understudied as is the damage profile associated with eDNA and whether or not this damage profile is similar to that observed in ancient samples. With the current knowledge of the patterns associated with sequencing error and aDNA damage, and hopefully in the future that associated with eDNA, better methods of modelling error within a dataset may be feasible. With a better understanding of error patterns, it will be possible to perform simulations on DNA sequences *in silico* or by using synthetic DNA mixtures which can, in turn, be used to inform the bioinformatic analysis and error filtering of actual samples.

Improved error modelling of potential HTS datasets may also aid in the taxonomic assignment of sequences and could potentially be incorporated into a probabilistic framework aimed at estimating the veracity of assignments. As it stands, the use of similarity scores or bit scores on which to base taxonomic assignments is fairly naïve and such a strategy neglects the associated metadata that exists around the taxonomic assignment of a sequence. A means by which to systematically and consistently include metadata across studies is sorely needed. Such metadata includes the presence or absence of a taxon in the area of study, the taxonomic coverage afforded by the database for a particular taxonomic ranking, the likelihood of a sequence being contamination, the degree of error inherent in the sequence and its similarity to other sequences in a dataset. An additional aspect that may be somewhat more difficult to account for at present, particularly in regions of high biodiversity, is the degree of inter- and intra-species variation associated with a given locus which can potentially lead to over- or under-estimation of true diversity (Coissac *et al.*, 2012). In some studies, including in this thesis, efforts have been made to employ a multi-locus approach aimed at improving the validity of taxonomic assignments and mitigate some of these issues. While this strategy is no doubt beneficial, a better means by which to incorporate probabilistic assignments of sequences across multiple loci is also needed to account for the differences in the propensity for error that may be locus specific, the differences in database coverage between each locus and the discriminatory power of each locus.

7.4 Concluding statement

This thesis research has, for the first time, explored the utility of environmental metabarcoding within the context of southwest Australia and has shown the potential prospects for future studies utilising such techniques in similar regions with a rich biota that is largely uncharacterised. While the initial research presented led to the successful characterisation of a biologically and culturally rich region of Australia it has also highlighted a set of challenges that remain. As databases and workflows improve, however, it is highly likely that metabarcoding the past and present biota of an ecosystem can contribute substantially to a number of fields including archaeology, palaeontology, ecology and conservation.

7.5 References

- Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., . . . Gnrke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*, R18.
- Alacs, E., Alpus, D., de Tores, P. J., Dillon, M., & Spencer, P. B. S. (2003). Identifying the presence of quokkas (*Setonix brachyurus*) and other macropods using cytochrome b analysis from faeces. *Wildlife Research*, *30*, 41-47.
- Andersen, K. (2014). *Methods for high-throughput characterisation of environmental DNA*. (Ph.D), Copenhagen University.
- Andersen, K., Bird, K. L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K. H., . . . Willerslev, E. (2012). Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, *21*, 1966-1979.
- Birks, H. H., Giesecke, T., Hewitt, G. M., Tzedakis, P. C., Bakke, J., & Birks, H. J. B. (2012). Comment on "Glacial survival of boreal trees in northern Scandinavia". *Science*, *338*, 742.
- Birks, H. J., & Birks, H. H. (2016). How have studies of ancient DNA from sediments contributed to the reconstruction of Quaternary floras? *New Phytologist*, *209*, 499-506.

- Boessenkool, S., Epp, L. S., Haile, J., Bellemain, E., Edwards, M., Coissac, E., . . . Brochmann, C. (2012). Blocking human contaminant DNA during PCR allows amplification of rare mammal species from sedimentary ancient DNA. *Molecular Ecology*, *21*, 1806-1815.
- Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., . . . de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, *29*, 358-367.
- Bowles, E., Schulte, P. M., Tollit, D. J., Deagle, B. E., & Trites, A. W. (2011). Proportion of prey consumed can be determined from faecal DNA using real-time PCR. *Molecular Ecology Resources*, *11*, 530-540.
- Burgar, J. M., Murray, D. C., Craig, M. D., Haile, J., Houston, J., Stokes, V., & Bunce, M. (2014). Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology*, *23*, 3605-3617.
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, advance online publication.
- Champlot, S., Berthelot, C., Pruvost, M., Bennett, E. A., Grange, T., & Geigl, E.-M. (2010). An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One*, *5*, e13042.
- Coissac, E., Riaz, T., & Puillandre, N. (2012). Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, *21*, 1834-1847.
- de Vernal, A., & Hillaire-Marcel, C. (2008). Natural variability of Greenland climate, vegetation, and ice volume during the past million years. *Science*, *320*, 1622-1625.
- Deagle, B., Chiaradia, A., McInnes, J., & Jarman, S. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics*, *11*, 2039-2048.

- Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, *10*.
- Denham, T., Atchison, J., Austin, J., Bestel, S., Bowdery, D., Crowther, A., . . . Matthews, P. (2009). Archaeobotany in Australia and New Guinea: practice, potential and prospects. *Australian Archaeology*, *68*, 1-10.
- Doi, H., Uchii, K., Takahara, T., Matsubishi, S., Yamanaka, H., & Minamoto, T. (2015). Use of droplet digital PCR for estimation of fish abundance and biomass in environmental DNA surveys. *PLoS One*, *10*, e0122763.
- Dortch, J. (2004a). Late Quaternary vegetation change and the extinction of Black-flanked Rockwallaby (*Petrogale lateralis*) at Tunnel Cave, southwestern Australia. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *211*, 185-204.
- Dortch, J. (2004b). *Palaeo-environmental change and the persistence of human occupation in south-western Australian forests*. Oxford: Archaeopress.
- Dortch, J., & Wright, R. (2010). Identifying palaeo-environments and changes in Aboriginal subsistence from dual-patterned faunal assemblages, south-western Australia. *Journal of Archaeological Science*, *37*, 1053-1064.
- Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., . . . Pradhan, S. (2013). A method for selectively enriching microbial DNA from contaminating vertebrate host DNA. *PLoS One*, *8*, e76096.
- Ficetola, G. F., Coissac, E., Zundel, S., Riaz, T., Shehzad, W., Bessière, J., . . . Pompanon, F. (2010). An *In silico* approach for the evaluation of DNA barcodes. *BMC Genomics*, *11*, 1-10.
- Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*, 423-425.
- Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., . . . Taberlet, P. (2014). Replication levels, false presences, and the estimation of

presence / absence from eDNA metabarcoding data. *Molecular Ecology Resources*, 15, 543-556.

Grealy, A., Macken, A., Allentoft, M. E., Rawlence, N. J., Reed, E., & Bunce, M. (2016). An assessment of ancient DNA preservation in Holocene–Pleistocene fossil bone excavated from the world heritage Naracoorte Caves, South Australia. *Journal of Quaternary Science*, 1-13.

Grealy, A. C., McDowell, M. C., Scofield, P., Murray, D. i. C., Fusco, D. A., Haile, J., . . . Bunce, M. (2015). A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quaternary Science Reviews*, 128, 37-47.

Haile, J., Froese, D. G., MacPhee, R. D. E., Roberts, R. G., Arnold, L. J., Reyes, A. V., . . . Willerslev, E. (2009). Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proceedings of the National Academy of Sciences*, 106, 22352-22357.

Haouchar, D., Haile, J., McDowell, M. C., Murray, D. C., White, N. E., Allcock, R. J. N., . . . Bunce, M. (2013). Thorough assessment of DNA preservation from fossil bone and sediments excavated from a late Pleistocene-Holocene cave deposit on Kangaroo Island, South Australia. *Quaternary Science Reviews*, 84, 56-64.

Hayward, M. W., de Tores, P. J., Dillon, M. J., & Fox, B. J. (2003). Local population structure of a naturally occurring metapopulation of the quokka (*Setonix brachyurus* Macropodidae: Marsupialia). *Biological Conservation*, 110, 343-355.

Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., . . . Colston, B. W. (2011). High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83, 8604-8610.

Hofreiter, M., Betancourt, J. L., Pelliza Sbriller, A., Markgraf, V., & McDonald, H. G. (2003). Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Curá, Neuquén Province, southwest Argentina. *Quaternary Research*, 59, 364-378.

Hofreiter, M., Kreuz, E., Eriksson, J., Schubert, G., & Hohmann, G. (2010). Vertebrate DNA in fecal samples from bonobos and gorillas: evidence for meat consumption or artefact? *PLoS One*, *5*, e9419.

Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *PLoS One*, *6*, e19254.

Jørgensen, T., Haile, J., Möller, P. E. R., Andreev, A., Boessenkool, S., Rasmussen, M., . . . Willerslev, E. (2012). A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Molecular Ecology*, *21*, 1989-2003.

Jørgensen, T., Kjær, K. H., Haile, J., Rasmussen, M., Boessenkool, S., Andersen, K., . . . Willerslev, E. (2011). Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA meta-barcoding. *Molecular Ecology*, *21*, 1980-1988.

Kelly, R. P., Port, J. A., Yamahara, K. M., Martone, R. G., Lowell, N., Thomsen, P. F., . . . Crowder, L. B. (2014). Environmental monitoring. Harnessing DNA to improve environmental management. *Science*, *344*, 1455-1456.

Kuch, M., Rohland, N., Betancourt, J. L., Latorre, C., Stepan, S., & Poinar, H. N. (2002). Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert, Chile. *Molecular Ecology*, *11*, 913-924.

Lee, C. K., Herbold, C. W., Polson, S. W., Wommack, K. E., Williamson, S. J., McDonald, I. R., & Cary, S. C. (2012). Groundtruthing next-gen sequencing for microbial ecology – biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One*, *7*, e44224.

Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. *Molecular Ecology*, *17*, 4186-4196.

Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., . . . Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science*, *34*, 1361-1366.

- Linacre, A., & Tobe, S. S. (2011). An overview to the investigative approach to species testing in wildlife forensic science. *Investigative Genetics*, 2, 2-2.
- Malmström, H., Stora, J., Dalen, L., Holmlund, G., & Götherström, A. (2005). Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution*, 22, 2040-2047.
- Nilsson, R. H., Ryberg, M., Kristiansson, E., Abarenkov, K., Larsson, K. H., & Koljalg, U. (2006). Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One*, 1, e59.
- Pacioni, C., Hunt, H., Allentoft, M. E., Vaughan, T. G., Wayne, A. F., Baynes, A., . . . Bunce, M. (2015). Genetic diversity loss in a biodiversity hotspot: ancient DNA quantifies genetic decline and former connectivity in a critically endangered marsupial. *Molecular Ecology*, 24, 5813-5828.
- Parducci, L., Jørgensen, T., Tollefsrud, M. M., Elverland, E., Alm, T., Fontana, S. L., . . . Willerslev, E. (2012). Glacial survival of boreal trees in northern Scandinavia. *Science*, 335, 1083-1086.
- Parducci, L., Matetovici, I., Fontana, S. L., Bennett, K. D., Suyama, Y., Haile, J., . . . Willerslev, E. (2013). Molecular- and pollen-based vegetation analysis in lake sediments from central Scandinavia. *Molecular Ecology*, 3511-3524.
- Pedersen, M. W., Ginolhac, A., Orlando, L., Olsen, J., Andersen, K., Holm, J., . . . Kjær, K. H. (2013). A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. *Quaternary Science Reviews*, 75, 161-168.
- Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., . . . Willerslev, E. (2014). Ancient and modern environmental DNA. *Philosophical Transactions Royal Society B*, 370, 20130383.
- Philippe, E., Franck, L., & Jan, P. (2015). Accurate multiplexing and filtering for high-throughput amplicon-sequencing. *Nucleic Acids Research*.

- Porter, T. M., Golding, G. B., King, C., Froese, D., Zazula, G., & Poinar, H. N. (2013). Amplicon pyrosequencing late Pleistocene permafrost: the removal of putative contaminant sequences and small-scale reproducibility. *Molecular Ecology Resources*, *13*, 798-810.
- Quéméré, E., Hibert, F., Miquel, C., Lhuillier, E., Rasolondraibe, E., Champeau, J., . . . Chikhi, L. (2013). A DNA metabarcoding study of a primate dietary diversity and plasticity across its entire fragmented range. *PLoS One*, *8*, e58971.
- Robasky, K., Lewis, N. E., & Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, *15*, 56-62.
- Schloss, P. D., Gevers, D., & Westcott, S. L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One*, *6*, e27310.
- Seersholm, F. V. (2015). *Three thousand years of resource economy in Greenland: an ancient DNA approach to study archaeological kitchen middens*. (Masters), University of Copenhagen.
- Smith, D. P., & Peay, K. G. (2014). Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS One*, *9*, e90234.
- Srivathsan, A., Sha J. C. M., Vogler A. P. & Meier, R. (2014). Comparing the effectiveness of metagenomics and metabarcoding for diet analysis of a leaf-feeding monkey (*Pygathrix nemaeus*). *Molecular Ecology Resources*, *15*, 250-261.
- Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., . . . Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry*, 1-16.
- Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., . . . Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*, 2565-2573.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4-18.

Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L., & Rohwer, F. (2009). Laboratory procedures to generate viral metagenomes. *Nature protocols*, *4*, 470-483.

Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M. E., . . . Taberlet, P. (2014). Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature*, *506*, 47-51.

Yoccoz, N. G., Brathen, K. A., Gielly, L., Haile, J., Edwards, M. E., Goslar, T., . . . Taberlet, P. (2012). DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, *21*, 3647-3655.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

Appendix I: Signed co-author permissions

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I RICHARD ALLCOCK confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:



Date: 19.5.2016

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Jennifer Atchison confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:



Date: 18.05.16

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Roberto Barrero confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 

Date: 3 June 2016

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Matthew Bellgard confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 24/05/2016

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Matthew Bellgard confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 24/05/16

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I Mattew Bellgard confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 24/05/16

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Prof. Michael Bunce confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19/5/16

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Prof. Michael Bunce confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19/5/16

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I Prof. Michael Bunce confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 

Date: 19/5/16

Murray, D. C., Coghlan, M.L. & Bunce, M. (2015) From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE* 10(4): e0124671.

Statement of Contribution

Conceived and designed the experiments: DCM, MB. Performed the experiments: DCM, MLC. Analysed the data: DCM. Contributed reagents/materials/analysis tools: MB. Wrote the paper: DCM, MB. Edited the manuscript: DCM, MB, MLC.

I Prof. Michael Bunce confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 

Date: 19/5/16

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Belinda Cannell confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:

Date:


23/5/16

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

Brian Chase 18 May 2016 6:51 pm
To: Dáithí Conall Gerard Murray [Hide Details](#)
RE: Curtin University (Mike Bunce's Lab) inclusion of co-authored publication in thesis permission 1

Hi Dáithí,

You certainly have my permission. Congrats! All my details are below in my signature.

Cheers!
b

[Dr. Brian M. Chase](#)
Director of Research, Centre National de la Recherche Scientifique (CNRS)
Secretary General, International Union for Quaternary Research (INQUA)
Institut des Sciences de l'Evolution-Montpellier, UMR 5554
Université de Montpellier, Bat.22, CC061, Place Eugène Bataillon, 34095 Montpellier cedex 5, France
Email: Brian.Chase@univ-montp2.fr
Telephone: [REDACTED]

Home page: <http://www.isem.univ-montp2.fr/recherche/equipes/environnement/personnel/chase-brian/>
European Research Project HYRAX website: <http://www.hyrax.univ-montp2.fr/>

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Edward Clarke confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19/05/2016

Murray, D. C., Coghlan, M.L. & Bunce, M. (2015) From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS ONE* 10(4): e0124671.

Statement of Contribution

Conceived and designed the experiments: DCM, MB. Performed the experiments: DCM, MLC. Analysed the data: DCM. Contributed reagents/materials/analysis tools: MB. Wrote the paper: DCM, MB. Edited the manuscript: DCM, MB, MLC.

I Megan Coghlan confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 23/05/2016

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I Joe Dortch confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19 May 2016

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Richard L.K. Fullagar confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 18th May 2016

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Tom Gilbert confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:

Date: 19/05/2016

A handwritten signature in black ink, appearing to read 'Tom Gilbert', with a large, sweeping flourish extending to the right.

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I James Haile confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:

Date:

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I James Haile confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:

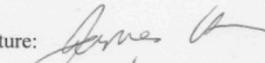
Date:

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I James Haile confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19/5/16

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I _____ Dalal Haouchar _____ confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 

Date: May 18, 2016

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Jayne Addison (née Houston) confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 23/05/16

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Jayne Addison (née Houston) confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 23/05/16

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Mike Macphail confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature:  (Hon.) Assoc. Prof.
Date: 6 June 2016 Dept. of Archaeology & Natural History
College of Asia & the Pacific
Australian National University

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Stuart Pearson confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19 May 2016

Murray, D. C., Bunce, M., Cannell, B., Oliver, R., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Rebecca Oliver confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: *Rebecca Oliver*

Date: 23/05/16

Murray, D. C., Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

Statement of contribution

Conceived and designed the experiments: DCM, MB, BLC, J. Haile. Performed the experiments: DCM, BLC, RO, J. Houston, NEW, MB, J. Haile. Analysed the data: DCM, MB, BLC, RAB, MIB, J. Haile. Contributed reagents/materials/analysis tools: BC, MIB, RB, Wrote the paper: DCM, MB, J. Haile.

I Nicole White confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19 May 2016

Murray, D. C., Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145

Statement of Contribution

Conceived and designed the experiments: DCM, MB, J. Haile. Performed the experiments: DCM, J. Houston, J. Haile. Analysed the data: DCM, SGP, RF, BMC, JA. Contributed reagents/materials/analysis tools: MB, SGP, RF, BMC, EC, MM. Wrote the paper: DCM, MB. with edits from co-authors.

I Nicole White confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19 May 2016

Murray, D. C., Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371

Statement of Contribution

DCM, MB and JH designed the experiments. DCM, JH, NW, DH and JD excavated and prepared samples. DCM, JH, MIB, DH and RA contributed to HTS data generation and bioinformatics. JD provided stratigraphic interpretations and GP and JD provided fossil and taxon interpretations. DCM and MB wrote the paper.

I Nicole White confirm my contribution to the cited publication and give permission for its inclusion in this thesis.

Signature: 
Date: 19 May 2016

Appendix II: Quaternary Science Reviews permission to reproduce manuscript

RightsLink Printable License

4/04/2016 4:35 pm

ELSEVIER LICENSE TERMS AND CONDITIONS

Apr 04, 2016

This is a License Agreement between Daithi C Murray ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard, Langford Lane Kidlington, Oxford, OX5 1GB, UK
Registered Company Number	1982084
Customer name	Daithi C Murray
Customer address	TrEnD laboratory Perth, WA 6102
License number	3841820088878
License date	Apr 04, 2016
Licensed content publisher	Elsevier
Licensed content publication	Quaternary Science Reviews
Licensed content title	High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens
Licensed content author	Dáithí C. Murray, Stuart G. Pearson, Richard Fullagar, Brian M. Chase, Jayne Houston, Jennifer Atchison, Nicole E. White, Matthew I. Bellgard, Edward Clarke, Mike Macphail, M. Thomas P. Gilbert, James Haile, Michael Bunce
Licensed content date	14 December 2012
Licensed content volume number	58
Licensed content issue number	n/a
Number of pages	11
Start Page	135
End Page	145
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Title of your thesis/dissertation	Developing and applying methodologies to characterise biodiversity using ancient and degraded DNA
Expected completion date	May 2016

<https://s100.copyright.com/App/PrintableLicenseFrame.jsp?publisherl...fecf6-0ff5-44d8-b051-05e812947474%20&targetPage=printablelicense>

Page 1 of 6

Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 GBP
VAT/Local Sales Tax	0.00 GBP / 0.00 GBP
Total	0.00 GBP
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement

and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. **Warranties:** Publisher makes no representations or warranties with respect to the licensed material.

10. **Indemnity:** You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. **No Transfer of License:** This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. **No Amendment Except in Writing:** This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. **Objection to Contrary Terms:** Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. **Revocation:** Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation:** This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website:** The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxx> or the Elsevier homepage for books at <http://www.elsevier.com>; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at <http://www.elsevier.com>. All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
 - o via their non-commercial person homepage or blog
 - o by updating a preprint in arXiv or RePEc with the accepted manuscript
 - o via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
 - o directly by providing copies to their students or to research collaborators for their personal use
 - o for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- after the embargo period
 - o via non-commercial hosting platforms such as their institutional repository
 - o via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional

private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a [CrossMark logo](#), the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's [posting policy](#) for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our [open access license policy](#) for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at <http://creativecommons.org/licenses/by/4.0>.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at <http://creativecommons.org/licenses/by-nc-sa/4.0>.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of

the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at <http://creativecommons.org/licenses/by-nc-nd/4.0>. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee. Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. Other Conditions:

v1.8

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix III: Publications arising from PhD candidature

1. Albornoz, F., Teste, F., Lambers, H., Bunce, M., **Murray, D. C.**, White, N., Laliberté, E. Changes in ectomycorrhizal fungal community composition and declining diversity along a 2-million year soil chronosequence. *Molecular Ecology*, DOI: 10.1111/mec.13778
2. Grealy A. C., McDowell, M. C., Scofield, P., **Murray D. C.**, Fusco D. A., Haile, J., Prideaux G. J., Bunce, M. (2015) A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages. *Quaternary Science Reviews*, 128:27
3. Coghlan, M. L., Maker, G., Crighton, E., Haile, J., **Murray, D. C.**, White, N. E., ... Bunce, M. (2015). Combined DNA, toxicological and heavy metal analyses provides an auditing toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM). *Scientific Reports*, 5:17475.
4. Berry, O., Bulman C., Bunce, M., Coghlan, M. L., **Murray, D. C.**, Ward, R. D. (2015) Comparison of morphological and DNA metabarcoding analyses of diets in exploited marine fishes. *Marine Ecology Progress Series*, 540:167
5. Gofton, A. W., Oskam, C. L., Lo, N., Beninati, T., Wei, H., McCarl, V., **Murray, D. C.**, ... Irwin, P. (2015). Inhibition of the endosymbiont “*Candidatus* Midichloria mitochondrii” during 16S rRNA gene profiling reveals potential pathogens in *Ixodes* ticks from Australia. *Parasites & Vectors*, 8:345.
6. **Murray, D. C.**, Coghlan, M.L. & Bunce, M. (2015) From benchtop to desktop: important considerations when designing amplicon sequencing workflows. *PLoS ONE* 10(4): e0124671.
7. Tridico, S. R., **Murray, D. C.**, Addison, J., Kirkbride, K. P. & Bunce, M. (2014) Metagenomic analyses of bacteria on human hairs: a qualitative assessment for applications in forensic science. *Investigative Genetics* 5:16
8. **Murray, D. C.**, Haile, J., Dortch, J., White, N., Haouchar, D., Bellgard, M. I., Allcock, R. J., Prideaux, G. J., & Bunce, M. (2013) Scrapheap challenge: a novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Scientific Reports* 3:3371
9. Coghlan, M. L., White, N. E., **Murray, D. C.**, Houston, J., Rutherford, W., Bellgard, M. I., Haile, J. & Bunce, M. (2013) Metabarcoding avian diets at airports: implications for birdstrike hazard management plans. *Investigative Genetics* 4:27

10. Haouchar, D., Haile, J., McDowell, M., **Murray, D. C.**, White, N., Allcock, R., Phillips, M., Prideaux, G., & Bunce, M. (2013) Thorough assessment of DNA preservation from fossil bone and sediments excavated from a Quaternary cave deposit on Kangaroo Island, South Australia. *Quaternary Science Reviews* 84:56-64
11. Burgar, J. M., **Murray, D. C.**, Craig M. D., Haile, J., Houston, J., Stokes, V., & Bunce, M. (2013) Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology* 23:3605-3617
12. **Murray, D. C.**, Pearson, S. G., Fullagar, R., Chase B. M., Houston, J., Atchison, J., White, N. E., Bellgard, M. I., Clarke, E., Macphail, M., Gilbert, M. T. P., Haile, J., & Bunce, M. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58:135-145
13. Coghlan, M., Haile, J., Houston, J., **Murray, D.**, White, N., Moolhuijzen, P., Bellgard, M., and & Bunce, M., (2011) Deep sequencing of plant and animal DNA contained within Traditional Chinese Medicines reveals legality issues and health safety concerns. *PLoS Genetics* 8:e1002657
14. **Murray, D. C.**, Bunce, M., Cannell, B., Oliver, B., Houston, J., White, N., Barrero, R., Bellgard, M. & Haile, J. (2011) DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776

DNA-Based Faecal Dietary Analysis: A Comparison of qPCR and High Throughput Sequencing Approaches

Dáithí C. Murray¹, Michael Bunce^{1*}, Belinda L. Cannell², Rebecca Oliver¹, Jayne Houston¹, Nicole E. White¹, Roberto A. Barrero³, Matthew I. Bellgard³, James Haile¹

1 Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, Murdoch, Western Australia, Australia, **2** School of Biological Sciences and Biotechnology, Murdoch University, Murdoch, Western Australia, Australia, **3** Centre for Comparative Genomics, Murdoch University, Murdoch, Western Australia, Australia

Abstract

The genetic analysis of faecal material represents a relatively non-invasive way to study animal diet and has been widely adopted in ecological research. Due to the heterogeneous nature of faecal material the primary obstacle, common to all genetic approaches, is a means to dissect the constituent DNA sequences. Traditionally, bacterial cloning of PCR amplified products was employed; less common has been the use of species-specific quantitative PCR (qPCR) assays. Currently, with the advent of High-Throughput Sequencing (HTS) technologies and indexed primers it has become possible to conduct genetic audits of faecal material to a much greater depth than previously possible. To date, no studies have systematically compared the estimates obtained by HTS with that of qPCR. What are the relative strengths and weaknesses of each technique and how quantitative are deep-sequencing approaches that employ universal primers? Using the locally threatened Little Penguin (*Eudyptula minor*) as a model organism, it is shown here that both qPCR and HTS techniques are highly correlated and produce strikingly similar quantitative estimates of fish DNA in faecal material, with no statistical difference. By designing four species-specific fish qPCR assays and comparing the data to the same four fish in the HTS data it was possible to directly compare the strengths and weaknesses of both techniques. To obtain reproducible quantitative data one of the key, and often overlooked, steps common to both approaches is ensuring that efficient DNA isolation methods are employed and that extracts are free of inhibitors. Taken together, the methodology chosen for long-term faecal monitoring programs is largely dependent on the complexity of the prey species present and the level of accuracy that is desired. Importantly, these methods should not be thought of as mutually exclusive, as the use of both HTS and qPCR in tandem will generate datasets with the highest fidelity.

Citation: Murray DC, Bunce M, Cannell BL, Oliver R, Houston J, et al. (2011) DNA-Based Faecal Dietary Analysis: A Comparison of qPCR and High Throughput Sequencing Approaches. PLoS ONE 6(10): e25776. doi:10.1371/journal.pone.0025776

Editor: Carles Lalueza-Fox, Institut de Biologia Evolutiva - Universitat Pompeu Fabra, Spain

Received: July 29, 2011; **Accepted:** September 9, 2011; **Published:** October 6, 2011

Copyright: © 2011 Murray et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding was obtained from the Foundation for National Parks and Wildlife, the Department of Environment and Conservation, and the Centre for Fish and Fisheries Research at Murdoch University. NEW received support from the Robert Hammond Research Studentship and MB from the Australian Research Council grant FT0991741. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.bunce@murdoch.edu.au

Introduction

DNA-based dietary analysis of faecal material has emerged as a promising tool to study animal biology, ecology and archaeology [1–4]. Dietary analysis is not limited to the discovery of what an animal consumes; it can also give an insight into ecosystem health [5–7], species' responses to environmental/anthropogenic stresses [8], and assist in the development of targeted strategies for conservation [9]. It is evident from the increase in the use of genetic techniques that there is a growing appreciation of the use of DNA-based faecal methods to investigate diet. The analysis of faecal material has proven to be a welcome move away from more invasive techniques used to study animal diet such as lethal sampling [10] and stomach flushing [11], both of which have undesirable effects on the sampled population [12]. Moreover, a general move towards molecular based approaches, e.g. fatty acid, stable isotope or DNA analysis, has allowed a shift from more subjective morphological approaches [1,13]. The extraction and sequencing of DNA from faecal samples is seen to be an effective and reliable indicator of species' diet, offering increased specificity

and taxonomic resolution compared to other techniques [14–16]. The possibility of misidentification of species is greatly reduced [14,17] and the ability to account for a wider range of species within the actual diet is greatly increased when compared to morphology which relies entirely on analysis of undigested remains, therefore neglecting prey that may leave little trace of its consumption [18–20].

DNA based quantitative estimates of diet, however, are not without problems. Issues have arisen as a result of primer biases and the problem of differential digestion still remains. Put simply, “is what goes in what comes out” [21]? Moreover, variability in the amount of DNA per unit biomass between species and different tissues is also difficult to quantify. Attempts to address such concerns have recently become an active area of research. Such efforts include; the use of blocking primers to circumvent the issue of predator DNA amplification [7,22]; the use of captive feeding trials to examine differential digestion; [21] and the introduction of correction factors to account for DNA amount variability within species and tissues [23]. These confounding factors continue to be a contentious issue within analytical dietary

research, however, DNA-based methods arguably still present the best way forward in the explication of species' diet [1,19].

Little Penguins (*Eudyptula minor*) are ideal test subjects for molecular dietary analysis and have been the subject of previous research into diet [21,24–27]. The use of seabirds as barometers of marine ecosystem health is widely acknowledged, and the use of facultative feeders such as Little Penguins, whose diet is limited by food availability, provides a good indication of changes in marine environments [28,29]. Little Penguins are found across the coastal regions of Australia and New Zealand [30] (Fig. 1) and their diet, which includes a variety of small (<20cm) schooling fish, varies throughout the year [24–27]. The penguin population situated on Penguin and Garden Islands (32°S 115°E), located south of Perth, Western Australia, represent the northernmost and westernmost limits of the range of *E. minor* [31,32] (Fig. 1). As a fringe population, these penguins are more vulnerable to environmental changes such as rising sea temperatures and increased ocean acidification [33,34]. Moreover, Penguin Island's close proximity to human settlement also puts it under increased pressure due to anthropogenic stressors, such as commercial and recreational fishing, in addition to coastal development [31,35–38]. The development of a multi-year DNA-based study to investigate dietary preferences will prove an effective method to monitor *E. minor* and the marine environment.

Three major DNA-based techniques have been used to varying degrees in the study of species' diet. Firstly, PCR amplification using universal primers with subsequent cloning and sequencing of amplicons, is a technique that has been used extensively in molecular dietary analyses, and to some extent still is [13,14,39]. Secondly, quantitative PCR (qPCR), using species-specific primers has been purported to offer great promise in relation to dietary analysis, with the potential to determine estimates of diet

composition [23,40,41]. Thirdly, a number of recent studies have highlighted the potential impact that High-Throughput Sequencing (HTS) may have on dietary studies. HTS has been proposed as a cost-effective alternative in assessing and quantifying species' diet [14,16,21], and using indexed primers enables a large number of samples to be processed in parallel [14,42,43]. As yet, however, no study has validated the use of HTS in providing quantitative estimates similar to those obtained via qPCR.

This study sets out to determine the composition of Little Penguin faecal samples by comparing cloning, qPCR and HTS approaches. The primary purpose of this study was to develop an effective long-term strategy for the continual monitoring of diet in the penguin population. However, it is envisaged that the approach and recommendations advocated here will assist in experimental design for DNA-based faecal monitoring across a wide diversity of species.

Materials and Methods

The handling of penguins and the collection of faecal samples was conducted by experienced handlers under a strict set of animal ethics guideline approved by the Murdoch University Animal Ethics Committee (permit no. W2002/06).

Sample collection & storage

A total of 47 penguin faecal samples were collected, for cloning analysis, over the period from August 2008 until September 2009 and a further 52 samples, for HTS and qPCR analyses over the period from October to December 2010. All samples were collected from free-living penguins inhabiting the study area (Fig. 1). Samples were collected opportunistically from adults and chicks by checking artificial nest boxes or by intercepting penguins

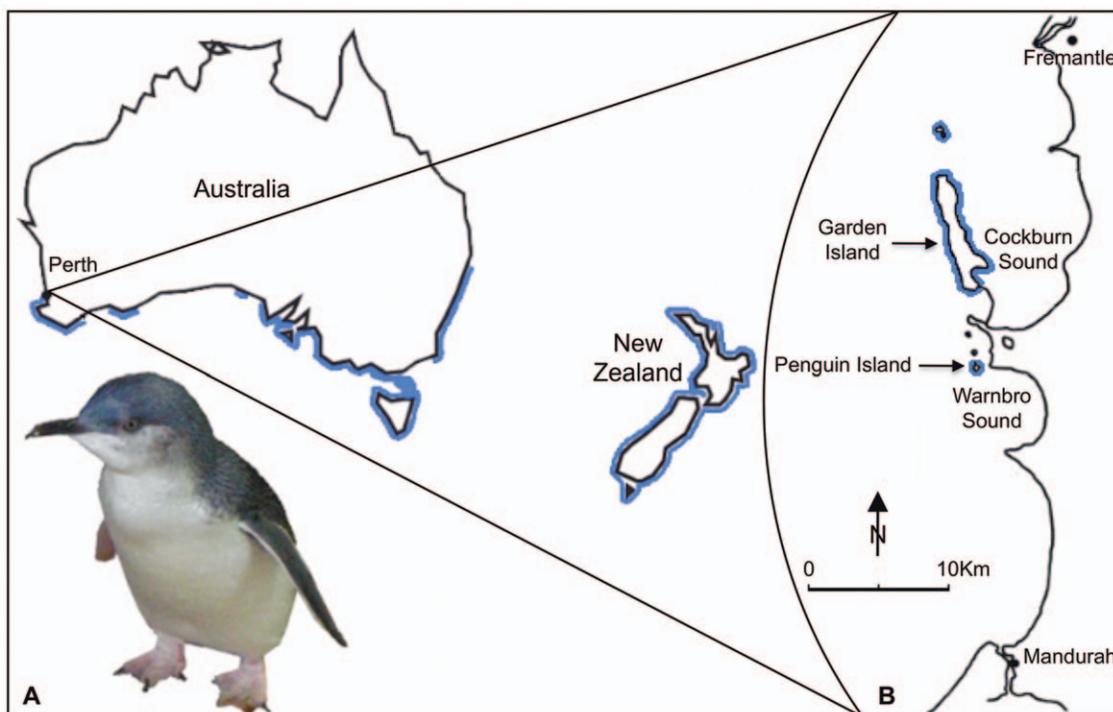


Figure 1. *Eudyptula minor* distribution and study site for faecal monitoring. (A) The coastal distribution (marked in blue) of *E. minor* across Australia and New Zealand. (B) Map of the study site in Western Australia; for this faecal monitoring study samples were collected from Penguin Island.

doi:10.1371/journal.pone.0025776.g001

returning from the ocean to their nests. Adult penguins were placed in plastic-lined containers for a maximum of 15 minutes. Chicks were placed in a smaller container with a hot water bottle for a maximum of 15 minutes before being returned to their nest boxes. Upon collection the faecal samples were placed in a labelled vial and then stored at -20°C within 12 hours. All handling and sampling was carried out under Murdoch University Animal Ethics Committee permit W2002/06.

Sample preparation & DNA extraction

The penguin samples were extracted in batches with the appropriate extraction controls. Samples were weighed and collected into 2mL tubes, with between 26–330mg of sample being used in each extraction depending on the condition of the faecal material. Extractions were performed using QIAamp DNA Stool Mini Kit (QIAGEN) as per manufacturer's instructions. DNA was eluted in 100 μL of AE buffer and dilutions of 1:10 and 1:50 were made using Milli-Q UV Pure H_2O for subsequent PCR reactions. DNA extracts were stored at -20°C until further analyses were performed.

Sample screening & initial quantification

Each faecal extract was screened using qPCR with 16S1F/2R primers in order to assess the DNA quality, quantity and to detect any possible PCR inhibition [44] (Table 1). Each extract was amplified at neat, 1:10 and 1:50 dilutions using the ABI Step One Real Time PCR machine. Each reaction was made up to 25 μL , containing 12.5 μL Power Sybr master mix (Applied Biosystems), 0.4 μM of each primer, 8.5 μL H_2O and 2 μL DNA. Reaction conditions were as follows: initial heat denaturation at 95°C for 5mins, followed by 40 cycles of 95°C for 30s; 54°C for 30s; 72°C for 45s followed by final extension at 72°C for 10mins and a 1°C melt curve to assist in the identification of primer dimer and non-specific amplification.

Cloning of amplified DNA

PCR products were cloned into pGEM[®]-T vectors (Promega) following the manufacturer's protocol and a maximum of 10 positive clones were selected per sample and amplified using the M13F/M13R primer set. Each 25 μL reaction contained 1X PCR buffer, 2mM MgCl_2 , 0.4mg/mL BSA, 0.25mM each dNTP, 0.6 μL SYBR Green (Invitrogen), 0.4 μM of each primer, 0.25 μL

Taq polymerase and 2.0 μL of template DNA. The cycling conditions were as follows: initial denaturation at 94°C for 5mins, followed by 35 cycles at 94°C for 15s; 55°C for 30s; 72°C for 30s. Amplicons were purified using an ACROPrep 10K 96 well plate (Pall) under a 25mmHg vacuum and screened via gel electrophoresis. Amplicons of the correct size were sequenced by Macrogen (Korea) using BigDye sequencing chemistry (Applied Biosystems) and analysed using Geneious v5.4.6 [45].

HTS library preparation

Prior to amplicon sequencing on the GS-Junior (454 Life Sciences), the 16S1F and 16S2R-degenerate primers were modified into fusion primers consisting of a GS FLX Titanium Primer A or B on the 5' end followed by one of 25 different 6bp Multiplex Identifier (MID) tags (allowing the simultaneous processing of 25 different PCR products) and then the template specific primer at the 3' end [46].

Extracts that successfully yielded DNA, as determined by the initial screening via qPCR, were assigned a unique tagged primer set. Fusion tagged PCR was carried out in 25 μL reactions containing 1X PCR Gold Buffer, 2.5mM MgCl_2 , 0.4mg/mL BSA, 0.25mM each dNTP, 0.4 μM of each primer, 0.25 μL AmpliTaq Gold (Applied Biosystems) and 2 μL DNA. The cycling conditions were as follows: initial heat denaturation at 95°C for 5mins, followed by 40 cycles of 95°C for 30s; 54°C for 30s; 72°C for 45s followed by final extension at 72°C for 10mins. Amplicons were always generated in duplicate and pooled together to minimise the effects of PCR stochasticity. The resultant pooled amplicons were purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), and eluted in 40 μL H_2O . Purified amplicons were electrophoresed on 2% agarose gel and amplicons were pooled in approximately equimolar ratios based on band intensity.

GS-Junior set-up and sequencing

To achieve the desired bead:template ratio, pooled amplicons were quantified using a synthetic 200bp oligonucleotide standard (of known molarity) with the Roche A and B primers engineered at either end. Quantitative PCR on a dilution series of both the standard and the pooled library, each run in duplicate, has enabled us to reproducibly normalise bead:template ratios. All procedures involved in the set up of the sequencing run (emulsion

Table 1. List of primer pairs used in this study.

Target species	Primer name	Sequence (5'-3')	Product Size (bp)	Annealing temp. ($^{\circ}\text{C}$)	Ref.
<i>Engraulis australis</i> (Australian Anchovy)	AN1F*	CCTAAATACCCGAGCCTTAT	101	60	This study
	AN2R*	CAACTCTCGGCTTAAGGGTTT			
<i>Spratelloides robustus</i> (Blue Sprat)	B52F*	GCGGCTACTGCCCTAATATCGC	109	60	This study
	B52R*	CTGAGCTCCAGGCCGAAGGC			
<i>Sardinops sagax</i> (Australian Pilchard)	PIL1F*	CCTAACTGGAGCCCCAAAC	117	60	This study
	PIL1R*	GCTGTGGCTCTGGTTTTAG			
<i>Hyperlophus vittatus</i> (Sandy Sprat)	SS2F*	GGCCTCAACAACATGACAGT	91	60	This study
	SS2R*	TAGGGTGCCCTAATCCACT			
All prey	16S1F-degenerate [†]	GACGAKAAGACCCTA	180–270	54	[44]
	16S2R-degenerate [†]	CGCTGTTATCCCTADRGTAACT			

Primers listed include species specific pairs (*) used in the targeted four fish qPCR assays and the universal pairs ([†]) used in cloning and High Throughput Sequencing approaches. Note the 16S1F/16S2R primers had 5' fusion and MID tags [46] if they were to be sequenced on the GS-Junior.

doi:10.1371/journal.pone.0025776.t001

PCR and bead recovery), including the sequencing run itself, were carried out according to the Roche GS Junior protocols for amplicon sequencing (<http://www.454.com>).

2.7 Four fish qPCR assay

Based on previous diet studies [24–27,31] and the DNA sequence data it was apparent that *Engraulis australis* (Australian Anchovy), *Spratelloides robustus* (Blue Sprat), *Sardinops sagax* (Australian Pilchard) and *Hyperlophus vittatus* (Sandy Sprat) formed a major part of the Little Penguins' diet. Therefore, in order to quantitatively assess the abundance of each of these species within each faecal sample and also to compare the quantitative nature of HTS using degenerate primers to that of qPCR, species-specific primer pairs (Table 1) were designed for each of the four fish species using Geneious v5.4 [45]. Primer sets for the four fish were designed using regions within the mitochondrial genes encoding for 16S rRNA based on sequence data obtained from local fish. Each primer pair was tested for efficiency and sensitivity on their target fish species. Importantly, the primer pairs were selected only if they did not cross-react with each other or other species detected in the area [27,47]. Once primer pairs were optimised, qPCR of faecal samples that successfully yielded DNA were performed in 25 μ L reactions containing 1X PCR Gold Buffer, 2.5mM MgCl₂, 0.4mg/mL BSA, 0.25mM each dNTP, 0.4 μ M of each primer, 0.25 μ L AmpliTaq Gold and 0.6 μ L SybrGreen (Invitrogen cat no S7563, 1:2000 dilution). Cycling conditions were as follows; initial denaturation at 95°C for 10min, followed by 40 cycles of 95°C for 15sec; 60°C for 45 sec.

Data analysis

FASTA (.fna) and Quality (.qual) sequence files obtained from the GS FLX Junior sequencing runs were processed using the following programs; BARTAB [48] de-convoluted the reads into sample batches using a map file containing sample and primer-MID tag information, cross_match [49] masked the primer and MID-tag sequences contained in the map file, trimseq [50] trimmed the masked primer and MID-tag sequences, and finally each sample of batched reads was then searched using BLASTN [51] without a low complexity sequence filter against the NCBI GenBank nucleotide database [52]. This was automated in the Internet-based bioinformatics workflow environment, YABI [<https://cgc.murdoch.edu.au/yabi/>]. The BLAST results that were obtained using YABI were imported into MEtaGenome Analyzer (MEGAN) where they were taxonomically assigned using the LCA-assignment algorithm (parameters included: min. bit score = 65.0, top percentage = 10%, min. support = 1) [17]. Where MEGAN was unable to resolve the taxonomy of a sequence (due to multiple species' sequences matching the query sequence), taxonomies were assigned using a combination of FishBase [<http://fishbase.org>] and Atlas of Living Australia [<http://www.ala.org>] to determine the most likely species based on their geographic distribution. Where more than one species returned by GenBank occurred around the Perth coastal area the query sequence was assigned to a higher taxonomic level.

Upon successful classification of all sequences obtained via HTS the percentage contribution of each prey item identified within each faecal sample was calculated, in addition to the overall contribution of each prey item across all faecal samples. In the case of the cloning data, a presence/absence method was used to determine the abundance of prey items within faecal samples.

In order to calculate the percentage contribution of each of the four major fish species within each faecal sample during the Oct '10-Dec '10 sampling period, the C_T (Cycle threshold) values obtained for the four target species via qPCR (at the same dilution

if deemed free of inhibition) were compared and converted into a percentage relative to each other. These individual percentages were then used to calculate the overall proportion of each of the four fish species across all faecal samples. Due to the stochasticity associated with low copy number DNA and primer dimer accumulation above C_T values of 34, all C_T values recorded above this level were attributed a C_T value of 34. This approach enables the target amplicon's presence to be acknowledged, whilst still allowing for it to be expressed proportionally to the other fish species within that sample.

To enable comparison of the qPCR and HTS datasets, the proportions of each of the four major fish species within each faecal sample as determined via HTS were considered to the exclusion of all other prey species detected. Using these data in conjunction with that obtained via qPCR, the Pearson product-moment correlation coefficient (Pearson's r) was calculated to determine the degree of correlation between the datasets. Additionally, individual paired sample t -tests for each major fish species were used to determine if there was a significant difference between the data obtained via both methods for any of the four major fish species. Samples that recorded C_T values >34 were excluded from statistical analyses, due to the stochasticity of qPCR above this threshold. All statistical analyses were carried out using the program R.

Results and Discussion

Overview and comparisons of Cloning and HTS approaches

Using the cloning approach, a total of nine fish species were identified from 129 sequences, in 22 of the 47 samples (47%) collected during the Aug '08-Sep '09 sampling period. Samples deemed to have failed either yielded no amplifiable DNA, were severely compromised by inhibitors, or had target copy numbers (as determined by qPCR C_T values >35.0) that were considered too low to be reliable. The dominant prey species detected within these samples was *H. vittatus*, present in 32% of samples, followed by *S. robustus*, found in 20% of samples, with *S. sagax*, *E. australis* and *Sardinella lemuru* (Scaly Mackerel) each found in 9.8% of samples (Fig. 2A). A number of other minor prey items were also identified, however they were found to represent a small proportion of sequences (Fig. 2A).

Of the 52 samples collected during the Oct '10-Dec '10 sampling period, only 27 samples (52%) were deemed to have yielded DNA of sufficient quality free of inhibition (determined by qPCR) that they could advance to HTS analysis. The two independent GS-Junior runs generated a total of 7810 DNA sequences. Of these sequences ~93% were unambiguously attributed to eleven fish species and <0.1% were identified as belonging to the genus *Pelates* (Striped Grunters). There were low levels of human contamination and penguin DNA (~3%) and unassigned/uninformative sequences accounted for ~3.6% of sequences. There was notable variation in the number of sequences generated for each faecal sample (range = 35–1055), and this is likely due to inaccurate blending of amplicons (see Materials & Methods). However, an average of ~300 reads per sample is more than sufficient coverage for dietary audits, especially when compared to the average number of sequences often generated per sample using bacterial cloning [53,54]. HTS of the Oct '10-Dec '10 samples revealed that, of the prey items identified, *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* were the major species present within the faecal material, each contributing 49%, 32%, 11% and 5% respectively (Fig. 2B). The remaining fish identified were minor contributors to the overall composition of

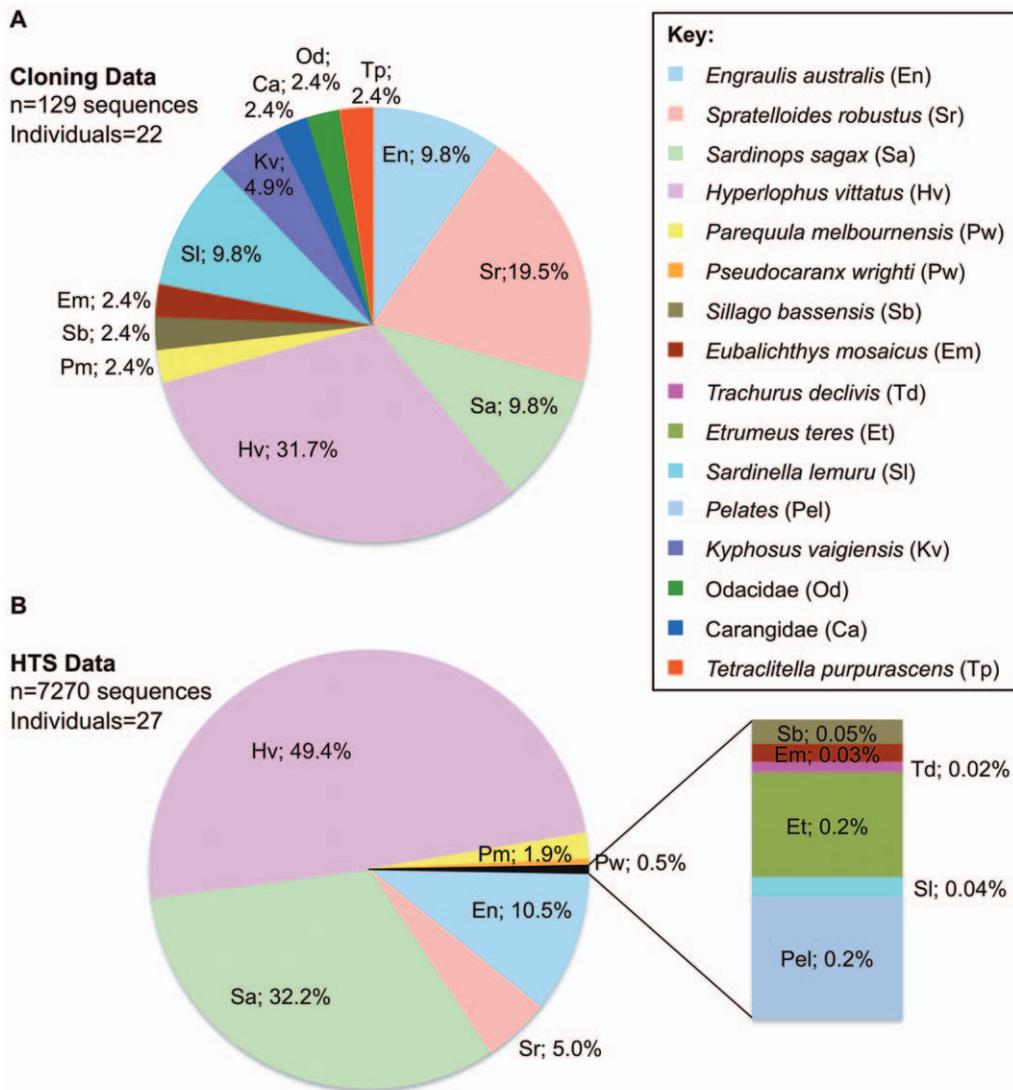


Figure 2. Percentage contribution of identified prey items in the faecal DNA of *E. minor*. (A) Graph showing fish identifications based on 16S rRNA sequence data obtained via cloning using universal primer set 16SF1/16S2R. Faecal samples (n = 22) for this study were collected during the Sep '08/Aug '09 period. (B) Penguin faecal samples collected during Oct '10-Dec '10 period (n = 27) that were audited using HTS methods. The 16SF1/16S2R set were MID-tagged and a total of 7270 sequences were assigned to prey items. doi:10.1371/journal.pone.0025776.g002

the samples (ranging from 0.02% to 1.9%) (Fig. 2B) and only in one sample did any of these fish constitute a significant proportion of the prey detected, that of PEN_42, where *Parequula melbournensis* (Silverbelly) contributed 48% to the sample composition for this individual (Table S1).

It is clear from the bacterial cloning and HTS data that there were four dominant fish species detected within the samples at this study site, those being *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* (Fig. 2). The occurrence of other minor contributing prey items within the samples is consistent with previous findings and reflects the opportunistic feeding behaviour of the Little Penguins [24,27]. A direct comparison of cloning and HTS is somewhat hampered by the fact that different faecal samples from different time periods were used for each method. However, it is clear that a number of important conclusions can be drawn from both datasets. Both methods provide a clear picture of the major prey species that are present within the collective faecal samples. Where they differ is in the relative contribution of each of these individual species (Fig. 2),

however this could be a result of temporal effects as it is well documented that the diet of Little Penguins varies throughout the year [27].

Cloning of universally amplified PCR products using bacteria, followed by DNA purification and Sanger sequencing is both expensive and time consuming. An additional issue, not entirely observed in this study, is that large numbers of clones are required in order to detect rare species [5,53], with the associated time and expense being inefficient for long-term monitoring of species' diet. For this reason, our Little Penguin monitoring program made the transition to HTS for the 2010 samples. Newly developed HTS platforms, especially small-scale systems such as the GS-Junior or IonTorrent, enable a quick, efficient and relatively inexpensive way to deep-sequence PCR amplicons generated from faecal DNA extracts [14,16,21]. Moreover, the use of MID-tagged primers makes it possible to run numerous samples in parallel, enabling not only an overview of the diet composition across a population, but also at the individual level [14,42]. HTS can provide a wealth of

information; greatly increasing the number of DNA sequences returned (129 sequences vs 7810 sequences) for a fraction of the labour and associated costs. Concomitant with the increases in sequencing depth is the prospect that HTS data might now provide better quantitative measures of the DNA targets within faecal material, much like estimates obtained using qPCR [23,44].

Overview of qPCR approach

In order to compare the quantitative nature of HTS to that of qPCR, a species-specific four fish qPCR assay was designed to estimate the relative abundance of each of the four major prey species determined within the collective samples (Fig. 2, Table 1). Careful development of each of the four primer pairs was critical to data fidelity [19,55], as was ensuring that the DNA extracts' C_T values behaved as desired when diluted (i.e. they were free from inhibition). From this four fish assay it was clear that *H. vittatus* and *S. sagax* were major constituents of the faecal samples; 49% and 32% respectively, with both *E. australis* and *S. robustus* each contributing 13% and 5% to the overall composition (Fig. 3A). The ANF1/ANR2 assay encountered some primer dimer issues at low template copy numbers, however the melt curves enabled

differentiation of product and dimer. Although not wholly representative of the *total* amount of prey DNA within samples, the qPCR assays gave a good indication of the abundance of each of the four major fish species relative to each other.

Comparison of HTS & qPCR approaches

It is important to actively compare and contrast both HTS and qPCR approaches to enable an informed decision of the most suitable method to be used for genetic faecal screening. To allow a comparison between both approaches, the HTS data had to be transformed to focus on the same four fish species as the qPCR assay; *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus*. The proportion of these species to the exclusion of the other species present was determined to be 52%, 32%, 11% and 5% respectively (Fig. 3B transformed from fig 2B data). It is clear that there is a striking degree of similarity between the proportions identified for the four fish species determined by qPCR and HTS (Fig. 3C). In order to investigate this further, the absolute differences between the results obtained individually by both methods were calculated. In the case of each fish species the overall difference in percentage abundance between the two techniques was negligible (*H. vittatus* -

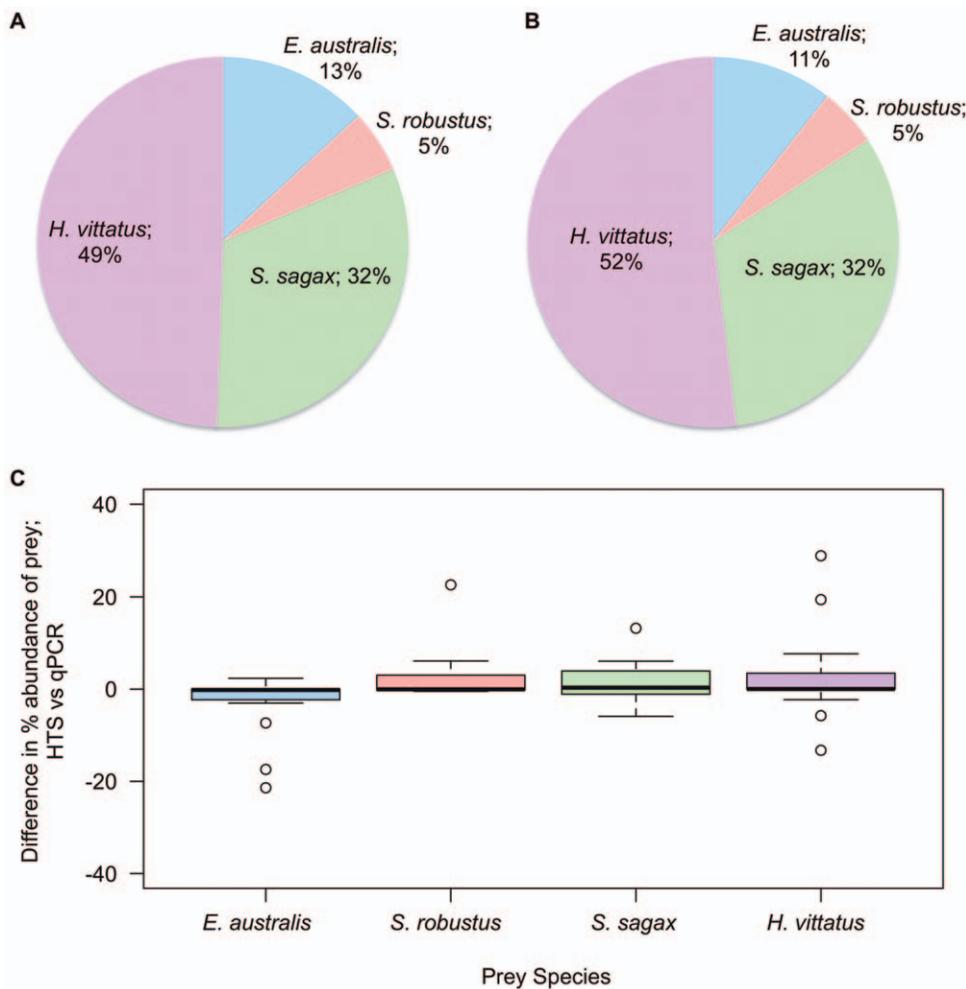


Figure 3. Comparison of HTS and qPCR methods determining the proportion of four major fish species. Graphs indicate the relative percentage composition of *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* within faecal samples of *E. minor* on Penguin Island, as determined by (A) qPCR and (B) HTS of samples collected during the period of Oct '10-Dec '10. (C) Box plot showing the difference between the results obtained by HTS and qPCR for each of the four major fish species found in the diet of *E. minor*. Samples whose C_T values were >34 have been excluded from the dataset (see Materials and Methods). doi:10.1371/journal.pone.0025776.g003

Median = 0.02, $n = 19$; *S. sagax* - Median = 0.31 $n = 13$; *E. australis* - Median = -0.18, $n = 15$; *S. robustus* - Median = -0.05, $n = 7$) (Fig. 3C). These initial results demonstrate a high degree of similarity between individual measures obtained by both methods. Furthermore, Pearson's r calculations revealed strong correlations between both methods for all four fish species (*H. vittatus* - Pearson's $r = 0.976$, $n = 19$; *S. sagax* - Pearson's $r = 0.996$, $n = 13$; *E. australis* - Pearson's $r = 0.973$, $n = 15$; *S. robustus* - Pearson's $r = 1.0$, $n = 7$) (Fig. 4), whilst individual paired t -tests revealed no significant difference between the values obtained by either method for any of the major prey species (*H. vittatus* - $p = 0.215$, $n = 19$; *S. sagax* - $p = 0.226$, $n = 13$; *E. australis* - $p = 0.100$, $n = 15$; *S. robustus* - $p = 0.266$, $n = 7$).

Although no statistical difference was detected in species composition in the combined analysis, it was apparent that there are slight differences between the datasets at the individual level (Table S2). There could be a number of reasons for such differences. Firstly, differential degradation of prey tissue DNA could account for some of the variance between datasets [23,39]. In this study the amplicon sizes produced by the primer sets in qPCR were shorter than those for HTS (see Table 1), and so in

some instances length biases may be present, especially in instances where there is differential degradation of prey tissue DNA in the gastrointestinal tract [41]. Indeed, it would appear that in this study *E. australis* was slightly over-represented in qPCR relative to HTS, whilst *H. vittatus* was marginally under-represented in qPCR relative to HTS (Table S2). A second potential cause could be the fact that the targeted qPCR assay is more efficient than the universal 16S primers used in HTS, therefore enabling the detection of the four prey species' DNA at lower template amounts. This is best illustrated when considering the presence/absence data, where HTS vs qPCR detection rates are compared: 70.4% vs 88.9% (*H. vittatus*), 48.2% vs 81.5% (*S. sagax*), 40.7% vs 74.1% (*E. australis*) and 14.8% vs 40.7% (*S. robustus*). In all cases where a species was detected in qPCR but not in HTS the C_T values were either >34 or the relative abundance of that species was below 1.5% (Table S2). Taken together, these data do suggest that the shorter, targeted qPCR assays were, across all four fish species, more sensitive to low template amounts. However, the higher qPCR detection success did not drastically affect the overall estimates of both methods, due to the low abundance of prey species in these instances. This also highlights a

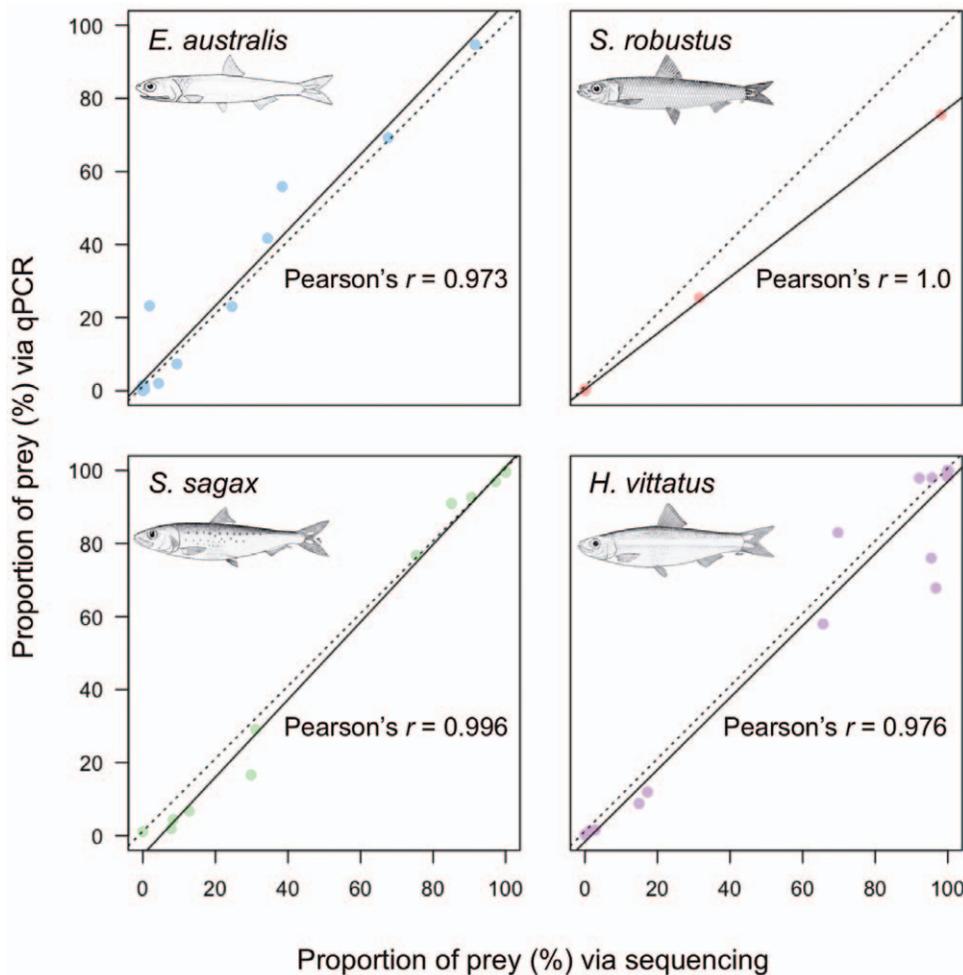


Figure 4. Correlation between four-fish data obtained via HTS and qPCR. Scatterplots include the percentage contributions obtained for each individual penguin via HTS and qPCR for each of the four major fish species detected within faecal samples. Solid line represents the line of best fit for individual species (Pearson's r values are shown), whilst the dotted line represents the overall correlation between both datasets with the data obtained for all fish species across all samples combined. Samples whose C_T values were >34 have been excluded from the dataset (see Materials and Methods). Fish images used in this figure can be reproduced freely for non-commercial purposes and are sourced from [59]. doi:10.1371/journal.pone.0025776.g004

very important advantage of species-specific qPCR over HTS, in that it can detect species at very low DNA abundances, whereas the nature of universal primers, such as those used in HTS, renders them less specific and less likely to efficiently amplify low copy number targets in the presence of abundant targets.

Whilst it is clear that there are slight differences between both methods, which are attributable to a variety of factors, it is also clear that in this case no single factor seemed to have a detrimental effect on the overall estimates of prey items within the collective faecal samples. It appears, however, that the difficulty arises when the penguins are considered on an individual basis. If, for instance, HTS were solely used in this study then it is quite clear that a good idea of the overall breadth of species could be ascertained. However, in some cases the use of universal primers may result in the non-detection of certain dietary constituents, if present in low abundance. On the other hand, with the use of the targeted qPCR approach a possibly more accurate estimate of the relative contribution of the major fish species' DNA could be determined across the population and individually, provided an *a priori* knowledge of diet is known. However, the contribution of the other minor constituents is overlooked. It would appear that the effect of this is largely minimal, unless, as was the case with sample PEN_42, one of the 'minor contributors' accounts for a large proportion, or all, of any given sample.

Recommendation for future experimental design

The uptake of genetic techniques to analyse faecal material has provided important insights into animal diet. It is clear that the use of qPCR and the advent of affordable HTS technologies are proving to be a welcome addition to this field of research. Both of these techniques have the potential to eclipse the more traditional molecular methodology of bacterial cloning and/or direct sequencing, which is costly, laborious and time-consuming. In light of the results of this study, it is fair to assume that qPCR and HTS represent the best approach currently available.

A key component of experimental design in this study was the methodical preparation and selection of samples for DNA extraction prior to qPCR or HTS. The extraction of DNA from faecal samples and the screening of samples for copy number and inhibition is a major bottleneck in the lab. However, the importance of this screening process cannot be under-stated, particularly when the samples being dealt with are complex, heterogeneous substrates containing severely degraded DNA in low copy numbers [56,57]. The initial qPCR screening strategy implemented in this study allowed the identification of suitable samples and DNA extract dilutions that contained the maximum concentration of amplifiable DNA and yet were inhibition free. There is no substitute for prior screening of samples; the congruence of qPCR and HTS in this study can be attributed largely to the fact that there is confidence in the amplifiability of the DNA extract dilution on which HTS and qPCR was conducted.

The ultimate choice of which method to opt for should be considered on a case-by-case basis, although the use of both methods in tandem would be the preferred option. If, for instance, an *a priori* knowledge of the species' diet in question were lacking then it would be more appropriate to use HTS with universal primer sets, thus giving an overview of the animal's diet. With this broad view of the animal's diet it can then be decided whether to pursue the use of targeted primers via the qPCR approach. If the number of prey species within the diet is of limited complexity qPCR may be preferable. Although not implemented here, in theory the quantitiveness of HTS using universal primers could be improved by using multiple universal primer sets in parallel [7,21].

If the goal of any dietary study is the long-term monitoring of diet, then it would be advisable to use HTS to determine the overall composition of the diet, and if possible a subsequent targeted qPCR approach to examine major prey items, to ensure that the diet remains consistent throughout the period of study. Ideally it would be beneficial to consider the use of both techniques in parallel to safeguard against erroneous results, as the removal of major contributors to the diet can have profound impacts on prey quantification. This is highlighted by the example of PEN_42 where *P. melbournensis* formed a major part of that individual penguin's faecal sample (Table S1). Therefore, in this case, the four fish qPCR assay is a poor representation of prey abundance.

Irrespective of the chosen method, primer design is crucial to the sensitivity of PCR, and careful consideration should be given to the design and testing of primers [19]. In the case of universal primers used in HTS, it is imperative that they are designed to allow taxonomic discrimination of amplicons, and yet also amplify a small enough region to circumvent issues of DNA degradation within faeces [19]. One additional issue is the fact that the coverage of certain animal groups in certain databases is not complete which will always make taxonomic assignments difficult [5,14]. The study of bats is a case in point; in this instance the use of qPCR assays would not be able to account for the hundreds of insects species in bat guanos, however qPCR could still be used to validate the relative portion of a few target species [5,14].

The validation of the quantitative nature of HTS, as compared to qPCR, to detect the DNA in faecal material, bodes well for future dietary studies. However, it is acknowledged that the results obtained via DNA-based faecal analysis are not always directly correlated with the biomass of prey consumed [55] – a recent study referred to them as semi-quantitative at best [23]. Much work is yet to be done to enable accurate reconstructions of the physical diet as estimates are currently confounded by a range of factors including; differential digestion rates of prey between species; DNA per unit biomass variability between tissues and the developmental stage of the prey species to name but a few issues [23,42,58]. It is also questionable whether digestion/faecal studies of captive birds will accurately recreate what is happening in the wild. Despite the many caveats regarding actual dietary intake, the accurate quantification of prey DNA actually contained in faecal matter represents an important developmental step.

Conclusion

Characterising the DNA preserved in faecal material is a powerful way to study both animal diet and also provide broader insights into ecosystem composition and health. In light of recent advances in DNA sequencing it was unclear which genetic auditing method(s) should be adopted for a multi-year monitoring program of Little Penguins. The results of qPCR and HTS approaches tested in this study demonstrate that the two methods are capable of generating high-fidelity datasets with no statistical difference between them. In the case of penguin diet, the use of both methods in parallel proved particularly useful with species-specific qPCR assays having better sensitivity, whilst HTS is able to detect species not targeted by qPCR. It is anticipated that the data and approaches presented here will be of benefit to other researchers intending to implement dietary monitoring programs and will assist in improving the accuracy of environmental audits based on faecal material.

Supporting Information

Table S1 Percentage contribution of prey items detected by HTS for each faecal sample. The percentage

contribution of detected prey items within each individual faecal sample, as determined by HTS of samples collected during the period of Oct '10-Dec '10, using 16SF1/16S2R universal primers. (XLS)

Table S2 Percentage contribution of four major fish species determined by HTS and qPCR methods. The percentage composition of *H. vittatus*, *S. sagax*, *E. australis* and *S. robustus* within individual faecal samples of *E. minor* on Penguin Island, as determined by HTS and qPCR of samples collected during the period of Oct '10-Dec '10. (XLS)

References

- Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. *Trends Ecol Evol* 24: 110–117.
- Symondson WOC (2002) Molecular identification of prey in predator diets. *Mol Ecol* 11: 627–641.
- Kuch M, Sobolik K, Barnes I, Stankiewicz BA, Spaulding G, et al. (2001) A molecular analyses of the dietary diversity for three archaic native americans. *Proc Natl Acad Sci USA* 98: 4317–4322.
- Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, et al. (1998) Molecular coproscopy: dung and diet of the extinct Ground Sloth *Nathrotheriops shastensis*. *Science* 281: 402–406.
- Clare EL, Barber BR, Sweeney BW, Hebert PDN, Fenton MB (2011) Eating local: influences of habitat on the diet of little brown bats (*Myotis lucifugus*). *Mol Ecol* 20: 1772–1780.
- Raye G, Miquel C, Coissac E, Redjadj C, Loison A, et al. (2011) New insights on diet variability revealed by DNA barcoding and high-throughput pyrosequencing: chamois diet in autumn as a case study. *Ecol Res* 26: 265–276.
- Deagle BE, Kirkwood R, Jarman SN (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Mol Ecol* 18: 2022–2038.
- Vila AR, Borrelli L (2011) Cattle in the Patagonian forests: Feeding ecology in Los Alerces National Reserve. *For Ecol Manage* 261: 1306–1314.
- Kowalczyk R, Taberlet P, Coissac E, Valentini A, Miquel C, et al. (2011) Influence of management practices on large herbivore diet-Case of European bison in Bialowieza Primeval Forest (Poland). *For Ecol Manage* 261: 821–828.
- Miller KM, McEwen LC (1995) Diet of nesting Savannah Sparrows in interior Alaska. *J Field Ornithol* 66: 152–158.
- Montague TL, Cullen JM (1985) Comparison of techniques to recover stomach contents from penguins. *Aust Wildl Res* 12: 327–330.
- Chiaradia A, Costalunga A, Kerry K (2003) The diet of Little Penguins (*Eudyptula minor*) at Phillip Island, Victoria, in the absence of a major prey - Pilchard (*Sardinops sagax*). *Emu* 103: 43–48.
- Casper RM, Jarman SN, Deagle BE, Gales NJ, Hindell MA (1997) Detecting prey from DNA in predator scats: A comparison with morphological analysis, using *Arctocephalus* seals fed a known diet. *J Exp Mar Biol Ecol* 347: 144–154.
- Bohmann K, Monadjem A, Lehmkühl N, Rasmussen M, Zeale MRK, et al. (2011) Molecular diet analysis of two African Free-tailed Bats (Molossidae) using High Throughput Sequencing. *PLoS ONE* 6: e21441.
- Williams C, Buck C (2010) Using fatty acids as dietary tracers in seabird trophic ecology: theory, application and limitations. *J Ornithol* 151: 531–543.
- Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, et al. (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. *Front Zool* 6: 16.
- Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17: 377–386.
- Tollit DJ, Schulze AD, Trites AW, Olesiuk PF, Crockford SJ, et al. (2009) Development and application of DNA techniques for validating and improving pinniped diet estimates. *Ecol Appl* 19: 889–905.
- King RA, Read DS, Traugott M, Symondson WOC (2008) Molecular analysis of predation: a review of best practice for DNA-based approaches. *Mol Ecol* 17: 947–963.
- Sheppard SK, Harwood JD (2005) Advances in molecular ecology: tracking trophic links through predator-prey food-webs. *Funct Ecol* 19: 751–762.
- Deagle B, Chiaradia A, McInnes J, Jarman S (2010) Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conserv Genet* 11: 2039–2048.
- Vestheim H, Jarman SN (2008) Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. *Front Zool* 5: Article No.: 12.
- Bowles E, Schulte PM, Tollit DJ, Deagle BE, Trites AW (2011) Proportion of prey consumed can be determined from faecal DNA using real-time PCR. *Mol Ecol Resour* 11: 530–540.
- Bradley JS, Cannell BL, Wooller RD (1997) A radio-tracking study of the movements at sea and diet of Little Penguins *Eudyptula minor* breeding on Penguin Island, Western Australia. Final report for Bowman Bishaw Gorham.
- Wooller RD, Dunlop JN, Klomp NI, Meathrel CE, Wienecke BC (1991) Seabird abundance, distribution and breeding patterns in relation to the Leeuwin Current. *J R Soc West Aust* 74: 129–132.
- Wienecke BC (1989) The breeding patterns of little penguins in Penguin Island, Western Australia, in relation to dietary and oceanographic factors: Murdoch University, Honours Thesis.
- Klomp NI, Wooller RD (1988) Diet of Little Penguins, *Eudyptula minor*, from Penguin Island, Western Australia. *Aust J Mar Fresh Res* 39: 633–639.
- Mallory ML, Robinson SA, Hebert CE, Forbes MR (2010) Seabirds as indicators of aquatic ecosystem conditions: A case for gathering multiple proxies of seabird health. *Mar Pollut Bull* 60: 7–12.
- Boersma PD, Rebstock GA, Frere E, Moore SE (2009) Following the fish: penguins and productivity in the South Atlantic. *Ecol Monogr* 79: 59–76.
- Marchant S, Higgins PJ (1990) *Ratites to Ducks*. In: *Handbook of Australian, New Zealand and Antarctic Birds*, Vol 1. Melbourne: Oxford University Press.
- Wienecke BC, Wooller RD, Klomp NI (1995) The ecology and management of Little Penguins on Penguin Island, Western Australia. In: Dann P, Norman I, Reilly R, eds. *The penguins: ecology and management*. Sydney: Surrey Beatty. pp 440–467.
- Wienecke BC (1993) The size and breeding patterns of the Little Penguin *Eudyptula minor* in Australia: a comparative study: Murdoch University, PhD Thesis.
- Dann P, Chambers L (2009) Climate change and Little Penguins. Report for Western Port Greenhouse Alliance.
- Boersma PD (2008) Penguins as Marine Sentinels. *Bioscience* 58: 597–605.
- Pichegru L, Ryan PG, Le Bohec C, van der Lingen CD, Navarro R, et al. (2009) Overlap between vulnerable top predators and fisheries in the Benguela upwelling system: implications for marine protected areas. *Mar Ecol Prog Ser* 391: 199–208.
- Cannell BL (2001) Status of Little Penguins in Western Australia: A management review. Department of Conservation and Land Management Report MMS/LNE/SIS-40/2001.
- Harrigan KE (1992) Causes of mortality of Little Penguins *Eudyptula minor* in Victoria. *Emu* 91: 273–277.
- Chape S (1984) Penguin Island draft management plan. Perth: National Parks Authority and Dept. of Conservation & Environment.
- Deagle BE, Tollit DJ, Jarman SN, Hindell MA, Trites AW, et al. (2005) Molecular scatology as a tool to study diet: analysis of prey DNA in scats from captive Steller sea lions. *Mol Ecol* 14: 1831–1842.
- Matejusková I, Doig F, Middlemas SJ, Mackay S, Douglas A, et al. (2008) Using quantitative real-time PCR to detect salmonid prey in scats of grey *Halichoerus grypus* and harbour *Phoca vitulina* seals in Scotland – an experimental and field study. *J Appl Ecol* 45: 632–640.
- Deagle B, Tollit D (2007) Quantitative analysis of prey DNA in pinniped faeces: potential to estimate diet composition? *Conserv Genet* 8: 743–747.
- Valentini A, Miquel C, Nawaz MA, Bellemain EVA, Coissac E, et al. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Mol Ecol Resour* 9: 51–60.
- Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, et al. (2007) The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. *PLoS ONE* 2: e197.
- Deagle BE, Gales NJ, Evans K, Jarman SN, Robinson S, et al. (2007) Studying seabird diet through genetic analysis of faeces: A case study on Macaroni Penguins (*Eudyptes chrysolophus*). *PLoS ONE* 2: e831.
- Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, et al. (2011) Geneious v5.4. Available: <http://www.geneious.com/>. Accessed 2011 Sep 20.

Acknowledgments

We would like to thank staff at Perth Zoo and the Penguin Island Experience for their involvement in the initial control trials where molecular methodologies were developed. The authors acknowledge the support of Ms Frances Brigg at the State Agricultural Biotechnology Centre DNA sequencing facility and the iVEC Informatics Facility, for computational support.

Author Contributions

Conceived and designed the experiments: DCM MB BLC J. Haile. Performed the experiments: DCM BLC RO J. Houston NEW MB J. Haile. Analyzed the data: DCM MB BLC RAB MIB J. Haile. Contributed reagents/materials/analysis tools: BC MIB RB. Wrote the paper: DCM MB J. Haile.

46. Roche (2009) Guidelines for Amplicon Fusion Primer Design for GS FLX Titanium Series Lib-A Chemistry TCB No. 013-2009. Technical bulletin: TCB No. 013-2009.
47. Dept. WAF (2008) State of the fisheries report. Hillarys WA, ed. Dept. of Fisheries, Western Australia.
48. Frank D (2009) BARCRAWL and BARTAB: software tools for the design and implementation of barcoded primers for highly multiplexed DNA sequencing. *BMC Bioinformatics* 10: 362.
49. de la Bastide M, McCombie WR (2007) Assembling Genomic DNA Sequences with PHRAP. *Curr Protoc Bioinformatics* Chapter 11: Unit 11.14.
50. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16: 276–277.
51. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
52. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. *Nucleic Acids Res* 34: D16–D20.
53. Clare EL, Fraser EE, Braid HE, Fenton MB, Hebert PDN (2009) Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. *Mol Ecol* 18: 2532–2542.
54. Kim BJ, Lee NS, Lee SD (2011) Feeding diets of the Korean water deer (*Hydropotes inermis argyropus*) based on a 202 bp rbcL sequence analysis. *Conserv Genet* 12: 851–856.
55. Sipos R, Szekely AJ, Palatinszky M, Revesz S, Marialigeti K, et al. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targeting bacterial community analysis. *FEMS Microbiol Ecol* 60: 341–350.
56. Deagle BE, Eveson JP, Jarman SN (2006) Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Front Zool* 3: 10.
57. Taberlet P, Waits LP, Luikart G (1999) Noninvasive genetic sampling: look before you leap. *Trends Ecol Evol* 14: 323–327.
58. Pegard A, Miquel C, Valentini A, Coissac E, Bouvier F, et al. (2009) Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from faeces. *J Agric Food Chem* 57: 5700–5706.
59. Whitehead PJP (1985) FAO species catalogue. Vol 7: Clupeoid fishes of the world (suborder Clupeioidi). An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, anchovies and wolf-herrings. Part 1 - Chirocentridae, Clupeidae and Pristigasteridae. *FAO Fish Synop* 125: 1–303.

Deep Sequencing of Plant and Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health Safety Concerns

Megan L. Coghlan¹, James Haile¹, Jayne Houston¹, Dáithí C. Murray¹, Nicole E. White¹, Paula Moolhuijzen², Matthew I. Bellgard², Michael Bunce^{1*}

1 Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, Murdoch, Australia, **2** Centre for Comparative Genomics, Murdoch University, Murdoch, Australia

Abstract

Traditional Chinese medicine (TCM) has been practiced for thousands of years, but only within the last few decades has its use become more widespread outside of Asia. Concerns continue to be raised about the efficacy, legality, and safety of many popular complementary alternative medicines, including TCMs. Ingredients of some TCMs are known to include derivatives of endangered, trade-restricted species of plants and animals, and therefore contravene the Convention on International Trade in Endangered Species (CITES) legislation. Chromatographic studies have detected the presence of heavy metals and plant toxins within some TCMs, and there are numerous cases of adverse reactions. It is in the interests of both biodiversity conservation and public safety that techniques are developed to screen medicinals like TCMs. Targeting both the p-loop region of the plastid *trnL* gene and the mitochondrial 16S ribosomal RNA gene, over 49,000 amplicon sequence reads were generated from 15 TCM samples presented in the form of powders, tablets, capsules, bile flakes, and herbal teas. Here we show that second-generation, high-throughput sequencing (HTS) of DNA represents an effective means to genetically audit organic ingredients within complex TCMs. Comparison of DNA sequence data to reference databases revealed the presence of 68 different plant families and included genera, such as *Ephedra* and *Asarum*, that are potentially toxic. Similarly, animal families were identified that include genera that are classified as vulnerable, endangered, or critically endangered, including Asiatic black bear (*Ursus thibetanus*) and Saiga antelope (*Saiga tatarica*). Bovidae, Cervidae, and Bufonidae DNA were also detected in many of the TCM samples and were rarely declared on the product packaging. This study demonstrates that deep sequencing via HTS is an efficient and cost-effective way to audit highly processed TCM products and will assist in monitoring their legality and safety especially when plant reference databases become better established.

Citation: Coghlan ML, Haile J, Houston J, Murray DC, White NE, et al. (2012) Deep Sequencing of Plant and Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health Safety Concerns. *PLoS Genet* 8(4): e1002657. doi:10.1371/journal.pgen.1002657

Editor: Robert DeSalle, American Museum of Natural History, United States of America

Received: August 12, 2011; **Accepted:** March 2, 2012; **Published:** April 12, 2012

Copyright: © 2012 Coghlan et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding for this research was provided by the Australian Research Council (FT0991741) and Murdoch University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: m.bunce@murdoch.edu.au

Introduction

Traditional Chinese medicines (TCMs) have been an integral part of Chinese culture and the primary medicinal treatment for a large portion of the population for more than 3000 years [1,2]. Outside of Asia there has been, in recent decades, a growing use of TCMs where they are being taken in conjunction with, or as an alternative to, conventional Western medicine [3,4]. The increasing popularity of TCM products has seen the monetary value of the industry increase to hundreds of millions of dollars *per annum* [5], its growth paralleled by the global increase in the use of complementary and alternative medicines. Despite its increased uptake, the therapeutic benefits of only a small number of TCM products have been scientifically validated [6], with their perceived efficacy being based largely on long-standing beliefs [7].

Chinese herbal medicines often contain numerous different plant and animal-derived products that combine to act synergistically to affect a desired outcome [8,9]. However, due to the proprietary nature of TCM manufacture, coupled with a lack of

industry regulation, the biological origin of contents can be difficult to determine with confidence, leading to questions regarding TCM quality, efficacy and safety [10,11]. Undeclared or misidentified TCM ingredients and adulterants can pose serious health risks to consumers [10,12,13]. These include: allergenic substances [14], plant toxins [7], heavy metals such as mercury, lead, copper and arsenic [15], and pharmaceutically active compounds of undetermined concentration [5]. In the early 1990s the misidentification of the toxic herb *Aristolochia fangchi* for the anti-inflammatory agent *Stephania tetrandra* led more than a hundred women to suffer kidney failure, with many later developing cancer of the urinary system [13].

In addition to safety concerns, issues of legality also surround TCMs. These concerns fall into three main categories: matters relating to the trade of endangered species; issues pertaining to honesty of food labelling; and adulteration of samples with drugs. Some TCMs contain plant and animal species [16–18] that fall under the jurisdiction of the Convention on International Trade in Endangered Species (CITES). CITES-listed species (see appendi-



Contents lists available at [SciVerse ScienceDirect](http://www.sciencedirect.com)

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev



High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens

Dáithí C. Murray^a, Stuart G. Pearson^b, Richard Fullagar^c, Brian M. Chase^{d,e}, Jayne Houston^a, Jennifer Atchison^c, Nicole E. White^a, Matthew I. Bellgard^f, Edward Clarke^g, Mike Macphail^h, M. Thomas P. Gilbert^{a,i}, James Haile^a, Michael Bunce^{a,*}

^a Ancient DNA Laboratory, School of Biological Sciences and Biotechnology, Murdoch University, South Street, Murdoch, WA 6150, Australia

^b PaleoLab, School of Physical, Environmental and Mathematical Science, University of New South Wales, Canberra, ACT 2610, Australia

^c Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, NSW 2522, Australia

^d Institut des Sciences de l'Evolution de Montpellier, UMR 5554, Centre National de Recherche Scientifique/Université Montpellier 2, Bat.22, CC061, Place Eugène Bataillon, 34095 Montpellier, cedex 5, France

^e Department of Archaeology, History, Culture and Religion, University of Bergen, Postbox 7805, 5020 Bergen, Norway

^f Centre for Comparative Genomics, Murdoch University, South Street, Murdoch, WA 6150, Australia

^g Rio Tinto, Dampier, WA, Australia

^h Department of Archaeology and Natural History, College of Asia and the Pacific, Australian National University, Canberra, ACT 0200, Australia

ⁱ Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark

ARTICLE INFO

Article history:

Received 22 June 2012

Received in revised form

17 October 2012

Accepted 18 October 2012

Available online 14 November 2012

Keywords:

Ancient DNA

Herbivore midden

Palaeoenvironment

Arid zone

Metabarcoding

High throughput sequencing

ABSTRACT

The study of arid palaeoenvironments is often frustrated by the poor or non-existent preservation of plant and animal material, yet these environments are of considerable environmental importance. The analysis of pollen and macrofossils isolated from herbivore middens has been an invaluable source of information regarding past environments and the nature of ecological fluctuations within arid zones. The application of ancient DNA (aDNA) techniques to hot, arid zone middens remains unexplored. This paper attempts to retrieve and characterise aDNA from four Southern Hemisphere fossil middens; three located in hot, arid regions of Australia and one sample from South Africa's Western Cape province. The middens are dated to between 30,490 (± 380) and 710 (± 70) cal yr BP. The Brockman Ridge midden in this study is potentially the oldest sample from which aDNA has been successfully extracted in Australia. The application of high-throughput sequencing approaches to profile the biotic remains preserved in midden material has not been attempted to date and this study clearly demonstrates the potential of such a methodology. In addition to the taxa previously detected via macrofossil and palynological analyses, aDNA analysis identified unreported plant and animal taxa, some of which are locally extinct or endemic. The survival and preservation of DNA in hot, arid environments is a complex and poorly understood process that is both sporadic and rare, but the survival of DNA through desiccation may be important. Herbivore middens now present an important source of material for DNA metabarcoding studies of hot, arid palaeoenvironments and can potentially be used to analyse middens in these environments throughout Australia, Africa, the Americas and the Middle East.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

The field of ancient DNA (aDNA) has, since its infancy, been largely restricted to the study of substrates from cool and frozen environments, which are deemed most amenable to long-term DNA preservation (Lindahl, 1993a,b). To date, a number of

historical and ancient samples have been subject to genetic analyses, ranging from bone (Smith et al., 2001) and hair (Bonnichsen et al., 2001; Gilbert et al., 2004) to more complex, heterogeneous substrates such as coprolites (Kuch et al., 2001; Poinar et al., 2001) and sediments (Hofreiter et al., 2003b; Willerslev et al., 2003; Haile et al., 2009). A number of studies have also attempted the isolation of DNA from samples – including fossil rodent middens – collected in cool to cold, semi-arid or arid environments (Kuch et al., 2002; Hofreiter et al., 2003a) and at high altitudes (Poinar et al., 1998; Hofreiter et al., 2000; Poinar et al., 2003). The application of

* Corresponding author. Tel.: +61 406998025.

E-mail address: m.bunce@murdoch.edu.au (M. Bunce).

molecular aDNA techniques to hot, semi-arid or arid environmental samples has previously been considered unrealistic due to the extreme heat found in such areas and as such is somewhat rarer and controversial (Smith et al., 2003; Gilbert et al., 2005b; although see; Gilbert, 2011; Hekkala et al., 2011).

Hot, arid and semi-arid environments are often marked by periods of stasis fluctuating on the edge of environmental equilibrium (Moore, 1953; Van Devender, 1990), punctuated by potentially dramatic changes that are induced by various triggers (Friedel et al., 1993; Tausch et al., 1993). There exists a delicate ecological balance and complex interplay across various environmental and biological gradients in arid regions (Beadle, 1966; Hayward and Phillipson, 1979; Northcote and Wright, 1982; Ritchie, 1986), making them of considerable environmental and biological interest. Flora and fauna inhabiting such environments are often at the limits of their tolerance to various abiotic factors, including temperature and water conservation, and have evolved to cope with extreme environmental conditions (Tongway and Ludwig, 1990; Groves, 1994). The study of past and present arid zone environments – and the distribution of species within them – allows for the exploration of how they have adapted and shifted in response to both natural and anthropogenic mechanisms (Van Devender and Spaulding, 1979; Fall et al., 1990; Pearson and Betancourt, 2002). The study of arid environments, however, is extremely challenging owing to the costs of collection and analysis, paucity of research attention and the lower quantities of recovered macro- and microfossil material. Nevertheless, studies using herbivore middens show promise in examining temporal and spatial variation in arid zone climates and biota, and perhaps, in some cases, may be the only viable means of doing so (Scott, 1990; Pearson and Betancourt, 2002; Scott and Woodborne, 2007; Chase et al., 2009, 2011).

To date, the reconstruction of palaeoenvironments has involved the use of a variety of molecular and morphological techniques, usually applied to sediment cores. Such techniques have included macrofossil and pollen identification, stable isotope analysis and ^{14}C dating. The application of these techniques to middens, where pollen and macrofossils have been preserved for thousands of years (Pons and Quézel, 1958; Wells and Jørgensen, 1964; Van Devender and Spaulding, 1979; Fall et al., 1990; Pearson and Betancourt, 2002; Scott et al., 2004), has provided the bulk of palaeoecological information in arid environments, where macrofossils are sparse and continuous fossil pollen records are largely unattainable. Midden material has therefore played a large part in our understanding of arid zone ecology and environment and act as archives of valuable information. Midden accumulations, usually as organic-rich nests in the case of American and Australian middens and latrines in the case of the African rock hyrax middens (Fig. S1), consist of material from the surrounding environment for construction or dietary purposes by arid-zone adapted mammals, and for the most part, represent a localised picture of the flora and fauna (Dial and Czaplewski, 1990; Scott, 1990; Pearson and Dodson, 1993). In the case of American and Australian middens, the animals urinate and defecate on their nests during the course of habitation, and organic material such as plant and animal tissue, bone, hair and eggshell gathered from the local surroundings (Pearson et al., 2001), become cemented together by means of crystallised urine or amberat, solidifying the mass into a hard, impermeable amalgam, referred to as a midden. Individually, these middens are generally recognised as reflecting sub-centennial-scale periods of construction and habitation. Conversely, African rock hyrax middens are latrines composed nearly exclusively of excrement. They are excellent traps for microfossils (pollen, phytoliths, etc.) from both regional and local environments as these are respectively brought in by the wind or adhere to the midden agent's fur. Hyrax

middens, however, contain very little non-dietary macrofossil material (for a fuller comparison and description of hyrax latrines and rodent nest middens see Chase et al., 2012). Increasingly, the hyrax middens that are collected for analysis are composed predominantly of urine, and have been shown to accumulate continuously over many thousands of years (Chase et al., 2009, 2011).

Genetic profiling has previously been applied to midden contexts, with two aDNA profiling studies retrieving reliable, seemingly authentic aDNA sequences from cold, arid zone (BWk – Köppen climate classification, see Peel et al., 2007) middens (Kuch et al., 2002; Hofreiter et al., 2003a). Since these studies, the fields of aDNA and environmental metabarcoding, whereby complex environmental samples are genetically audited (Valentini et al., 2009; Taberlet et al., 2012), have rapidly evolved. With the advent of affordable and accessible second generation high-throughput sequencing (HTS) it is now possible to genetically screen a wide range of complex modern and ancient substrates, with an unprecedented depth of sequence coverage (Shokralla et al., 2012). Through the use of material as diverse as sediment (Haile et al., 2009; Jørgensen et al., 2012), water (Rusch et al., 2007; Ficetola et al., 2008; Thomsen et al., 2012) and faeces (Deagle et al., 2009; Valentini et al., 2009; Murray et al., 2011) a wealth of data can be produced to aid in the understanding of pertinent ecological questions in relation to biodiversity (Andersen et al., 2011; Griffiths et al., 2011), dietary analysis (Pegard et al., 2009; Deagle et al., 2010) and anthropogenic impacts (Chariton et al., 2010; Vila and Borrelli, 2011). It is now possible, therefore, to bypass traditional molecular cloning and Sanger sequencing techniques through the use of new DNA technologies (HTS) to supplement morphological (macrofossils and palynology) methods of midden analysis, to allow an even fuller investigation of arid zone ecology.

Using HTS and environmental DNA metabarcoding techniques, this study attempts to recover aDNA from herbivore midden material collected from three hot, arid Australian sites and one site in South Africa (Fig. 1) that have been dated to between $30,490 \pm 380$ and 710 ± 70 cal yr BP. A comparison of the data obtained via HTS with complementary data on past and present species distributions, in addition to pollen and macrofossil analyses, allows for a critical examination and authentication of the genetic data. This study aims to demonstrate how genetic methods can be used to complement traditional methods of midden investigation for palaeoenvironmental reconstruction, to further our understanding of hot, arid environments.

2. Collection sites

Four Southern Hemisphere middens were sampled in this study; a single hyrax midden from South Africa's (RSA) Western Cape Province (Fig. 1A) and three herbivore middens from separate Interim Biogeographic Regionalisation of Australia (IBRA) regions (Thackway and Cresswell, 1995) within Western Australia (WA) (Fig. 1B–D). The three midden samples collected in Western Australia were from hot, arid zones (BWh Köppen climate classification, see Peel et al., 2007). The hot, arid zone collection sites are generally characterised by extreme hot summers and somewhat mild winters. Daytime summer temperatures average $\sim 37\text{--}38$ °C, but regularly exceed 40 °C. In winter, average daytime highs are $\sim 21\text{--}25$ °C, but can fall to $\sim 6\text{--}7$ °C at night. Winter nighttime temperatures at or close to freezing are extremely rare in these zones (climate data from Giles and Tom Price weather stations, WA). This contrasts markedly with previous midden genetic studies (Kuch et al., 2002; Hofreiter et al., 2003a) where average daily highs in summer are $\sim 24\text{--}28$ °C, although it can reach ~ 30 °C, and winter daily highs average $\sim 16\text{--}21$ °C, with nighttime

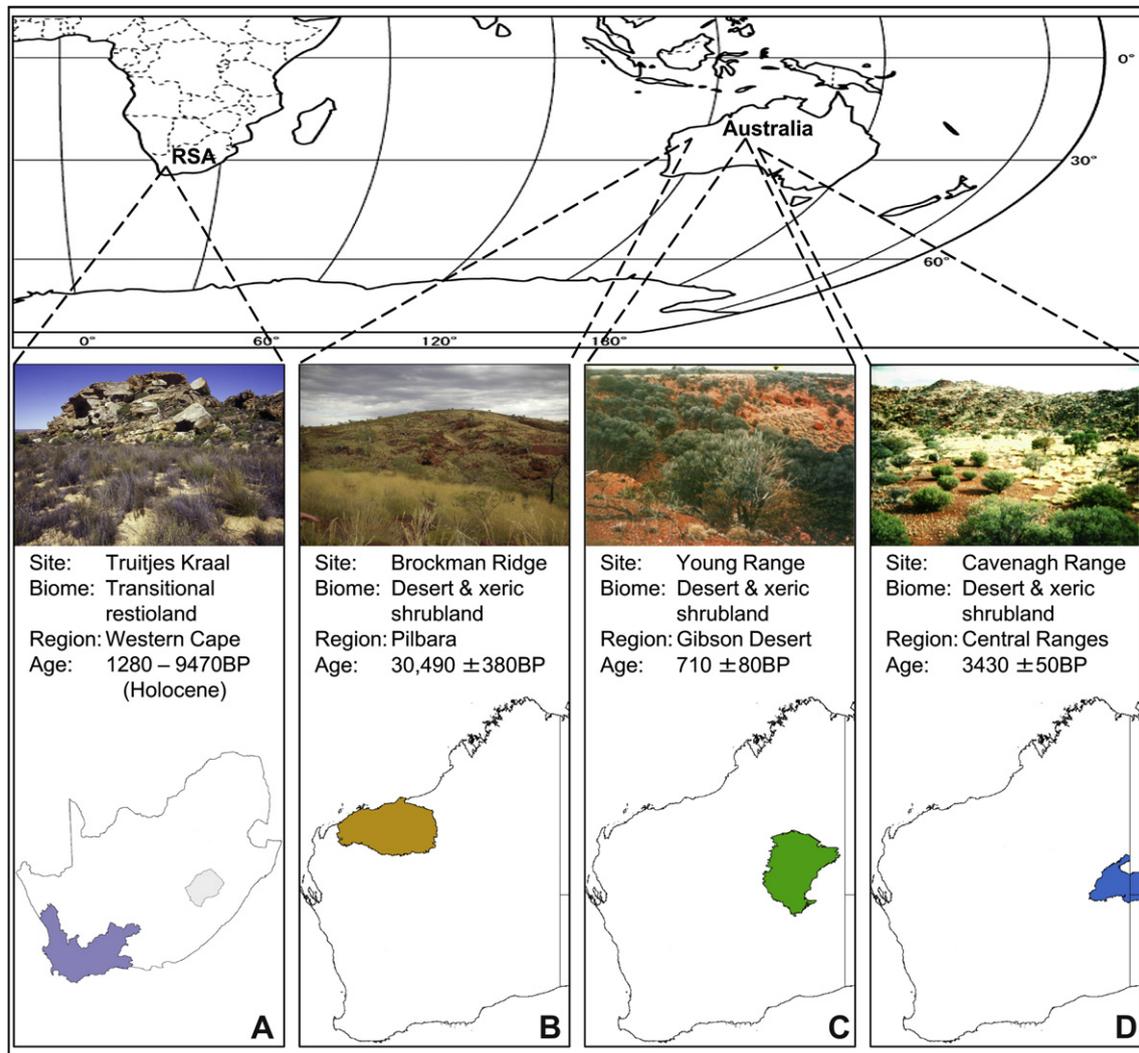


Fig. 1. Location of midden sites used in this study and associated information. A: Location and image of Truitjes Kraal midden site, South Africa, with Western Cape highlighted. B–D: Locations and images of Western Australian midden sites, with IBRA regions highlighted.

temperatures at or below freezing more common (climate data from weather stations at Neuquén Airport, Argentina and Calama, Chile).

2.1. Truitjes Kraal, RSA (TK)

Truitjes Kraal (32.5123°S, 19.3112°E) is located in the Cape Floristic Region (CFR) in the Western Cape province of RSA (Fig. 1A). The midden site lies in what is described as “restioliand” (dominated by Restionaceae), within a few kilometres of the relatively sharp transition between the Fynbos and Succulent Karoo biomes, with a climate on the threshold between hot-summer Mediterranean (Csa) and cold, semi-arid (BSk). The site records a mean annual precipitation around 350 mm, a mean annual temperature of around 15 °C (data from Hijmans et al., 2005) and an Aridity Index value of 0.242 (Trabucco and Zomer, 2009). The vegetation at the site consists of a low shrub understorey with intermittent taller shrubs, in addition to dwarf succulent shrubs of Crassulaceae and Mesembryanthemaceae (Meadows et al., 2010).

2.2. Brockman Ridge, WA (BR)

The Brockman Ridge, an ironstone-capped strike ridge, lies in the Pilbara IBRA region of northwestern WA (Fig. 1B),

approximately 60 km northwest of Mount Tom Price (22.68°S, 117.78°E). The Pilbara is a desert and xeric shrubland biome with a BWh climate that consists of scattered low trees of *Eucalyptus leucophloia* over *Acacia atkinsiana* open shrubland, over *Triodia wiseana* mid-dense hummock grassland. A number of other species are also associated with the site that includes *Acacia aneura*, *Hakea chordophylla*, *Paspalidium clementii*, *Ptilotus calostachyus* and *Solanum lasiophyllum* (Biota Environmental Sciences Pty Ltd, 2005). The nearest weather station is situated at Tom Price (−22.7°, 127.77°) recording a median annual precipitation around 313 mm, a mean annual temperature of around 24 °C (data from Hijmans et al., 2005) and an Aridity Index value of 0.2 (Trabucco and Zomer, 2009).

2.3. Young Range, WA (YR)

The Young Range (25.05°S, 124.983°E), located in Western Australia, is a low breakway in the extremely isolated Gibson Desert IBRA region (Fig. 1C). The Young Range is a desert and xeric shrubland biome that consists of shrubs, low shrubs and herbs. The dominant flora at the site is a mixture of Caesalpiniaceae, Myoporaceae, *Acacia*, *Grevillea* and species dominating hummock grassland (e.g. *Triodia* spp.) (Pearson, 1997). The Young Range also has a BWh climate and Giles is the nearest meteorological station

(−25.03°, 128.30°) recording a median annual precipitation around 250 mm, a mean annual temperature of around 23 °C (data from Hijmans et al., 2005) and an Aridity Index value of 0.101 (Trabucco and Zomer, 2009).

2.4. Cavenagh Range, WA (CR)

The Cavenagh Range (26.2°S, 127.9°E) is a rock pile situated in the Central Ranges IBRA region located in eastern WA (Fig. 1D). The dominant vegetation at the site includes spinifex (*Triodia* spp.), with shrubs and *Eucalyptus* spp. along the drainage lines (Pearson, 1997). It is considered a desert and xeric shrubland biome with a BWh climate, with the nearest meteorological station (Giles: −25.03°, 128.30°) recording a median annual precipitation around 250 mm, a mean annual temperature of around 22 °C (data from Hijmans et al., 2005) and an Aridity Index value of 0.1164 (Trabucco and Zomer, 2009).

3. Materials and methods

In keeping with standard aDNA practice, pre-PCR work was conducted in a dedicated aDNA clean room, with all downstream post-PCR work conducted in a physically separate laboratory, thus minimising sample contamination (Cooper and Poinar, 2000). Each midden was sub-sampled at Murdoch University, Australia and subsequently sent to the Centre for GeoGenetics, Denmark for independent replication. For all samples, DNA extraction, amplification and sequencing were performed at both Murdoch University and the Centre for GeoGenetics. Whilst HTS was performed at Murdoch University, traditional cloning and direct Sanger sequencing were performed at the Centre for GeoGenetics.

3.1. Background to midden samples

The samples used in this study were collected, radiocarbon dated (Table S1) and analysed for pollen and macrofossils prior to this study (Pearson, 1997; Meadows et al., 2010; Macphail, 2011). Large, intact samples were taken from the middens in this study to allow for sub-sampling, thus limiting the risk of environmental contamination. Middens that appeared to have been damaged as a result of weathering, digging or burrowing were avoided, although the BR midden was fractured along the base and had a honeycombed appearance (Atchison, 2010). The TK midden was collected in its entirety from a rock face overhang and spans the period from 1280 to 9470 cal yr BP (Meadows et al., 2010). The sample used for aDNA analysis was not dated separately, but is certainly of Holocene age. The BR midden was collected from the rear of the Brock 12 rock shelter within an Aboriginal site complex in the Puutu Kuntj Kurrama and Pinikura native title claimant area. Sections of the cave had been walled-in with the likely purpose of either the creation of an artificial habitat for the exploitation of, or the trapping of small animals (Fig. S2). With the exception of the creation of these walls, no other evidence of cultural material or influence was identified at the site of sample collection (Clarke, 2010). This is the oldest midden in this study, radiocarbon dated to 30,490 ± 380 cal yr BP (Macphail, 2011), although the age of the midden was not known before aDNA analysis took place. This midden consisted of three sub-samples obtained from one midden mound that were processed separately (Atchison, 2010). The YR midden was found in a rock shelter, protected from dissolution by moisture, and has been radiocarbon dated to 710 ± 80 cal yr BP (Pearson, 1997). The CR midden, radiocarbon dated to 3430 ± 50 cal yr BP, was collected from a small crevice and had few leaves and sticks, suggesting that an animal other than a stick-nest rat (*Leporillus* spp.) may have formed the midden (Pearson, 1997). The above radiocarbon dates were taken

directly on organic materials and the age estimates they provide for midden accumulation carry the possibility of being on material both older and younger than the aDNA within the stratigraphic units sampled.

3.2. DNA extraction and screening

Between 0.16 and 0.31 g of midden material was used for each sample DNA extraction using the Sergey Bulat extraction method optimised for small amounts of material, with extraction controls also included (Haile, 2011). Bulat buffer component concentrations were as follows; 0.02 g/mL Sarcosyl, 50 mM Tris–HCl (pH 8.0), 20 mM NaCl, 3.5% 2-mercaptoethanol, 50 mM DTT, 2 mM PTB, 0.8 g/mL Proteinase K. DNA was eluted in 100 µL and screened using quantitative PCR (qPCR) at multiple dilutions. DNA extracts were screened using multiple primer sets for both plants and mammals. The plant primer sets included both *trnLc/h* and *trnLg/h* plastid primers that amplify short sections of the *trnL* intron (Taberlet et al., 1991, 2007). In addition to these, both 12SA/O and 16Smam (Taylor, 1996) primer sets, designed to amplify a small region within mammalian mitochondrial 12S and 16S genes respectively, were also used. Each qPCR reaction was made up to a total volume of 25 µL, containing 12.5 µL ABI Power SYBR master mix (Applied Biosystems), 0.4 µM of forward and reverse primer, 8.5 µL H₂O and 2 µL DNA extract. Reaction conditions for the plant primers were as follows: initial heat denaturation at 95 °C for 5 min, followed by 40 cycles of 95 °C for 30 s; 54 °C for 30 s (annealing step); 72 °C for 45 s followed by a 1 °C melt curve and final extension at 72 °C for 10 min. Quantitative PCR cycling conditions for the 12SA/O and 16Smam primer sets were the same as those for both plant primers, except the annealing temperatures, which were 55 °C and 57 °C, respectively. For each qPCR assay, DNA extraction, negative PCR reagent and positive controls were included.

3.3. DNA sequencing

DNA extracts that successfully yielded DNA of sufficient quality, free of inhibition, as determined by initial qPCR screening, were assigned a unique 6 bp DNA tag (specifically a Multiplex Identifier-tag, MID-tag) (Roche, 2009) for each of the *trnLg/h*, 12SA/O and 16Smam primer sets. Independent MID-tagged qPCR for all midden samples were carried out using each primer set in 25 µL reactions containing 1 × PCR Gold Buffer (Applied Biosystems), 2.5 mM MgCl₂ (Applied Biosystems), 0.4 mg/mL BSA (Fisher Biotech, Aus), 0.25 mM of each dNTP (Astral Scientific, Aus), 0.4 µM of forward and reverse primer, 0.25 µL AmpliTaq Gold (Applied Biosystems), 0.6 µL SYBR Green (1:2,000, Life Sciences gel stain solution) and 2 µL of template. The cycling conditions for qPCR using the *trnLg/h* primer set were as follows: initial heat denaturation at 95 °C for 5 min, followed by 50 cycles of 95 °C for 30 s; 50 °C for 30 s (annealing step); 72 °C for 45 s followed by final extension at 72 °C for 10 min. The cycling conditions were the same for both 12SA/O and 16Smam primer sets apart from the annealing temperatures, which were 50 °C and 57 °C respectively. Multiplex Identifier-tagged PCR amplicons were generated in duplicate and pooled together to minimise the effects of PCR stochasticity on low-temple samples. The resultant pooled amplicons were purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), according to the manufacturer's instructions and eluted in 40 µL H₂O. Purified amplicons were electrophoresed on 2% agarose gel and pooled in approximately equimolar ratios based on ethidium-stained band intensity to form a sequencing library. For each MID-tagged qPCR assay, negative PCR controls were included and if found to contain amplifiable DNA these PCR amplicons were incorporated into the pooled sequencing library.

Emulsion PCR and GS Junior 454 Sequencing were performed as per Roche GS Junior protocols for amplicon sequencing (<http://www.454.com>).

3.4. Data analysis

Processed emulsion PCR amplicon sequence reads (hereafter referred to as sequences) obtained from the GS Junior sequencing runs have been deposited in the Dryad Repository (<http://dx.doi.org/10.5061/dryad.7334t>). Sequences were sorted into sample batches based on MID-tags using Geneious v5.6.4 (Drummond et al., 2011). MID-tags and primers were trimmed from the sequences allowing for no mismatch in length or base composition, also performed using Geneious v5.6.4. Batched and trimmed sequences were then dereplicated using 454 Replicate Filter (Gomez-Alvarez et al., 2009), clustering sequences of exact identity and length. Dereplicated sequence files were then searched for chimeras using the *de novo* method in UCHIME (Edgar et al., 2011), and were removed. After the above post-sequencing screen, sequences occurring only once (i.e. singletons) were removed, to minimise false positives arising from sequencing error. Once complete, each batch of cleaned, de-noised sequences was searched using BLASTn version 2.2.23 (Altschul et al., 1990), against the NCBI GenBank nucleotide database (Benson et al., 2006) to enable the identification of reads. Sequences were searched without a low complexity filter, with a gap penalties existence of five and extension of two, expected alignment value less than $1e-10$ and a word count of seven. This was automated in the internet-based bioinformatics workflow environment, YABI (Hunter et al., 2012). The BLAST results obtained using YABI were imported into MEtaGenome Analyzer v4 (MEGAN), where they were taxonomically assigned using the LCA-assignment algorithm (min. bit score = 65.0, top percentage = 5%, min. support = 1) (Huson et al., 2007). Further analysis of Muridae sequences was conducted by determining Operational Taxonomic Units (OTUs) using OTUPIPE with default parameters (<http://drive5.com/otupipe/>), whilst a phylogenetic comparison of Muridae sequences between samples was conducted using MrBayes (Huelsenbeck and Ronquist, 2001) in Geneious v5.6.4 (Drummond et al., 2011).

After sequences were processed, identified and parsed, the species identified were investigated to determine whether or not they currently occur in the region where they were detected, or have occurred in the past. To do this, the South African National Biodiversity Institute's (SANBI) Plants of Southern Africa online checklist [<http://posa.sanbi.org/searchspp.php>] was used for the RSA midden (Fig. 1A), and a combination of FloraBase [<http://florabase.dec.wa.gov.au/>] and Atlas of Living Australia [<http://www.ala.org.au/>] were used for the Australian sites (Fig. 1B–D).

4. Results and discussion

4.1. Overview of sequencing data

Over 20,000 sequences were obtained via HTS that passed the post-sequencing screen and occurred at an abundance greater than one (see Section 3.4). DNA was amplified using *trmLg/h* (size variable product between ~90 and 120 bp – including MID-tags and primers), 12SA/O (~160 bp) and 16Smam (~150 bp) primer combinations, whilst amplicon generation using the longer *trmLc/h* (giving an expected product of variable length >200 bp) primer set failed at each of the four study sites. Appropriate control reactions (described in Section 3: Materials and Methods) throughout the process, with the exception of ubiquitous human DNA sequences, were found to be negative for contaminant DNA arising from laboratory processing procedures. It is acknowledged however that

contamination can be cryptic and sporadic, and that low-level contamination can escape contamination controls (Champlot et al., 2010). The strict adherence to aDNA protocols, the use of appropriate controls throughout, in addition to the critical analysis of the data (described in Section 3.4) (Cooper and Poinar, 2000; Gilbert et al., 2005a), however, greatly reduces the likelihood that contamination can account for the data presented here.

Previous studies involving the amplification of the hyper-variable p-loop region of the plastid *trmL* intron, using the *trmLg/h* primer set, have shown taxonomic assignment possible with sequences as short as 10 bp (Taberlet et al., 2007). In this study, however, sequences less than 38 bp returned no taxonomic information and as such were discarded. Across the four midden samples, taxa representing 28 distinct families of plants were identified using *trmL* sequences that varied in length from 38 to 70 bp, minus MID-tags and primers (Table 1).

Through the assignment of DNA sequences to GenBank a total of six mammalian families were identified using both mammalian mtDNA 12S and 16S rRNA PCR assays, which generated sequences ~95–105 bp and ~90–100 bp in length respectively, minus MID-tags and primers. Within these mammalian families, species could reasonably be assigned in three cases (Table 2).

To our knowledge this is the first study to focus on the retrieval and sequencing of aDNA from Southern Hemisphere fossil midden material located in hot, arid regions. Moreover, the application of HTS techniques to midden material has not been attempted to date, and the following findings clearly demonstrate the increase in resolution afforded by the use of such methodology. Of significance, the Brockman Ridge midden sample is the oldest environmental sample; quite possibly the oldest sample, from which aDNA has

Table 1

Plant families identified in the midden samples using *trmL* plastid primers. For a more detailed comparison between plant taxa identified previously via morphological analysis and those identified via genetic means refer to Fig. S3a–d.

Taxon	Midden Location			
	Cavenagh Range	Young Range	Brockman Ridge	Truitjes Kraal
Acanthaceae		√		
Amaranthaceae	√	√#		√
Amaryllidaceae			√Ω	
Anacardiaceae		√Ω	√Ω	
Apocynaceae				√
Asteraceae	√		√#	√
Brassicaceae	√			
Bromeliaceae			√¶	
Campanulaceae				√#
Casuarinaceae	√#			
Ebenaceae				√
Fabaceae	√	√#	√#	√#
Gesneriaceae		√Ω		
Goodeniaceae		√		
Lamiaceae				√
Loranthaceae	√			
Malvaceae	√			
Meliastaceae				√
Oleaceae				√
Pinaceae	√¶			
Poaceae	√		√#	√#
Podocarpaceae				√#
Proteaceae		√#		
Sapindaceae	√#		√#	
Scrophulariaceae				√#
Solanaceae	√#	√#		√
Thymelaeaceae				√#
Toricelliaceae				√¶

Key: √ – Present in midden sample; # – Found previously in midden via morphological analysis; Ω – Not found in region; ¶ – Not found natively in Australia/RSA.

Table 2
Mammalian taxa identified in midden samples using 16S and 12S rRNA primer sets.

Taxon	Midden location			
	Cavenagh Rng	Young Rng	Brockman Ridge	Truitjes Kraal
Dasyuridae		√ ^β		
<i>Pseudantechinus</i>		√		
Gliridae				√
<i>Graphiurus ocellaris</i>				√
Macropodidae	√ ^β			
Muridae		√ [#]	√ ^β	
Procyonidae				√ ^β
<i>Procyon capensis</i>				√ ^β %
Phalangeridae	√ ^β			
<i>Trichosurus vulpecula</i>	√ ^β %			

Key: √ – Present in midden sample; # – Found previously in midden via morphological analysis; β – Detected using both 16S and 12S rRNA primer sets; % – indicates top BLAST species match 100% similarity.

been successfully extracted in Australia (although see Adcock et al., 2001; and subsequent critiques Cooper et al., 2001; Smith et al., 2003). For the Pilbara IBRA region in particular, aDNA work of this kind could be a critical addition to the assemblage of palaeoenvironmental data, as it is dated to a period for which almost no such regional data exists (Clarke, 2010; Macphail, 2011). This paper confirms that aDNA can be successfully recovered from midden deposits in hot, arid climates, suggesting that middens may be a valuable substrate for genetic analysis in such regions; it does not claim to be a comprehensive study of the sampled middens. Instead, an overview of the aDNA data is provided, focussing on some of the more salient points related to taxa identified by HTS and comparing the results with previous pollen and macrofossil analyses.

4.2. Site-specific analysis

4.2.1. Cavenagh Range

At least eleven families of plants were identified in the CR midden (Table 1), all of which, with the exception of Pinaceae (Order: Pinales), occur in the Central Ranges IBRA. Of the plant families identified, three were previously detected via pollen analysis: Casuarinaceae, Sapindaceae and Solanaceae (Table 1) (Pearson, 1997). Pollen analysis was only able to identify the genus *Dodonaea* (Sapindaceae), whilst genetic analysis identified both *Casuarina* (Casuarinaceae) and *Solanum* (Solanaceae) (Fig. S3a). However, although *Casuarina* is known to occur in the IBRA, it is recorded some distance from the site (ALA, FloraBase). The sequences assigned to *Casuarina* in this study are highly likely to be *Allocasuarina*, which does occur at the site and is known to occur alongside *Atriplex* (Mitchell and Wilcox, 1994), also detected via genetic analysis (Fig. S3a). In addition to these taxa, Loranaceae was identified via genetic analysis but not through previous pollen analysis of the fossil midden. A number of possible genera of Poaceae were also detected, including *Eriachne* and *Urochloa* (Fig. S3a), both of which, although not formally recorded at the site, are recorded in the IBRA.

Previous analysis of the CR midden did not identify any macrofossil remains (Pearson, 1997). Through the use of mammal specific primers, however, it was possible to detect the presence of Phalangeridae, specifically *Trichosurus vulpecula* (the common brushtail possum) and Macropodidae (Table 2). *Trichosurus vulpecula* is no longer found at Cavenagh Range; last recorded in the area in the 1930's, and it is the only species of Phalangeridae known to have existed in the Central Ranges IBRA (ALA). The distribution of *T. vulpecula* has retracted considerably since

European settlement, as a result of a range of issues including predation and overgrazing by introduced species (How and Hillcox, 2000). The identification of Macropodidae sequences to genus or species level proved difficult, with both 12S and 16S giving no clear indication past the family level. Currently there are only four species of Macropodidae known to exist in the Central Regions, with *Lagorchestes hirsutus* (the rufous hare-wallaby), *Macropus robustus* (the common wallaroo) and *Petrogale lateralis* (the black-flanked rock-wallaby) all recorded specifically at Cavenagh Range (ALA). Whilst 16S indicated the presence of *Macropus* it was not possible to identify *M. robustus* using this primer set and *Macropus* sequence identities were quite low ($\leq 95\%$). Use of the 12S primer set again resulted in difficulties with assignment to a genus or species level, with both *Lagorchestes* and *Petrogale* identified with equal similarity (98%). However, currently no 16S or 12S sequences for *P. lateralis* exist on GenBank. It was initially suggested that the CR midden was constructed by an animal other than a stick-nest rat (*Leporillus* spp.), possibly a rock wallaby or possum (Pearson, 1997). The identification of Macropodidae, possibly *Petrogale*, and *T. vulpecula* DNA (Table 2) in the midden material therefore increase the likelihood of this being the case.

4.2.2. Young Range

All plant families detected in the YR midden (shown in Table 1), with the exception of Gesneriaceae, which has an eastern Australian distribution, are known to occur in the Gibson Desert IBRA (ALA, FloraBase). Previous pollen and macrofossil analysis had identified Amaranthaceae, Fabaceae, Proteaceae and Solanaceae (Fig. S3b), all of which were detected via this genetic screening, and a number of other families not detected in this study (Pearson, 1997).

Previous macrofossil analysis of the YR midden found several species of mammal, that included the locally extinct *T. vulpecula* and *Isodon auratus* (the golden bandicoot), in addition to both *Notomys* (hopping mice) and *M. robustus* (Pearson, 1997). Genetic screening of the midden did not detect any of the above specifically (Table 2). Muridae sequences were identified from the midden material, though it was not possible to assign such sequences to a genus level due to the absence of 12S, 16S and COI (Cytochrome Oxidase I) reference sequences for many of the Muridae species found in the area, however it seems that these sequences cluster to form a single OTU, although there is some variation in the collective sequences ($< 2\%$). Such variation, although minor, is unlikely to have arisen as a result of sequencing error or chimeras due to the post-sequencing screen removing such instances, and could indicate multiple individuals contributing to this midden. Additionally, sequence BLASTn matches group these Muridae sequences closest to other Australasian Muridae, e.g. *Melomys cervinipes* (Fawn-footed Mosaic-tailed Rat) and *Paramelomys ruber* (Mountain Mosaic-tailed Rat), albeit with low percentage similarities ($< 93\%$). However, Dasyuridae, most likely *Pseudantechinus* (false antechinus), currently found in the area, was detected in the midden material through DNA analysis (Table 2), and this was not previously identified via macrofossil analysis.

4.2.3. Brockman Ridge

The Brockman Ridge midden mound is the oldest midden deposit in this study, and for the purposes of this discussion the three sub-samples are treated as one.

Fossil pollen assemblages recovered from the samples were dominated by unidentified Poaceae and a number of taxa within Family Myrtaceae, leading Macphail (2011) to propose that if plant DNA were preserved in the amberat that it would most likely be that of Myrtaceae (*Eucalyptus* and possibly *Melaleuca*) and Poaceae (possibly *Triodia*). Of these taxa, only Poaceae were detected using genetic techniques, although other less common taxa represented

by pollen were identified via genetic screening, e.g. Asteraceae, Fabaceae and Sapindaceae (*Diplopeltis* and/or *Dodonaea*) (Table 1, Fig. S3c).

The Brockman Ridge sample contained no identifiable macroscopic remains when analysed conventionally (Atchison, 2010), however the targeting of both 12S and 16S mammalian mitochondrial genes revealed the presence of Muridae sequences (Table 2). It was not possible to definitively say to which genera these sequences belong, owing to the lack of 12S and 16S sequences on GenBank for species that occur or are known to have occurred in the area, however BLASTn results group these Muridae signatures closest to other Australasian Muridae, e.g. *Uromys hadrourus* (Masked White-tailed Rat), albeit with low percentage similarities (<93%). Additionally, for both primer sets, OTU analysis suggests that these sequences form a single OTU, although, as was the case with the YR midden, there is some minor variation between sequences within this clustering (<2%). Based on phylogenetic analysis it is also possible to suppose that the Muridae sequences identified in this midden differ from those detected in the YR midden, and represent distinct species (Fig. S4).

4.2.4. *Truitjes Kraal*

Initial pollen analysis of the TK midden revealed high levels of Asteraceae, Ericaceae (Order: Ericales) and Poaceae (Meadows et al., 2010). Using genetic means a number of different possible genera of both Asteraceae and Poaceae were detected (Fig. S3d), however, no Ericaceae was found. Alternatively, genetic analysis detected Ebenaceae of the same order Ericales. In addition to several species detected by both pollen and DNA analysis, a number of additional taxa were identified, solely through genetic analysis, such as Apocynaceae, Lamiaceae and Solanaceae (Table 1, Fig. S3d). A few taxa were identified that do not occur specifically at the site, such as Melianthaceae and Oleaceae. However, both of these taxa are known to occur relatively close to the site (Melianthaceae occurrence id: NBG171075-0 and Oleaceae occurrence id: PRE320306-0) (SANBI), and considering the antiquity of the material it is possible that they grew at the site in the past.

The TK midden contains no faunal macrofossils but targeting mammalian DNA revealed both the midden builders – *Procapra capensis*, the rock hyrax – and *Graphiurus ocellaris* (the spectacled dormouse or namtap); a South African endemic species that inhabits a wide range of habitats including dry rocky outcrops and cliffs in South Africa (Table 2).

4.3. Limitations of study

Given the controversy surrounding previously purported aDNA retrieval from hot, arid zone specimens (see Cooper and Poinar, 2000; Gilbert et al., 2005b; Schlumbaum et al., 2008 but also; Gilbert, 2011; Hekkala et al., 2011) a number of caveats need to be considered when interpreting the degraded and ancient DNA recovered in this study to allow for a proper evaluation of the authenticity of the presented results (Gilbert et al., 2005a).

Ancient DNA, which by its nature is extremely degraded and often damaged, is typically quite short, fragmented and in low copy number. Various studies have shown that the average length of DNA recovered from ancient specimens is generally less than 100 bp (Poinar et al., 2006), and this study is no exception. The DNA sequences retrieved from the middens in this study for all primer combinations were less than 100 bp. Moreover, the attempt to target and amplify a longer stretch of the *trnL* intron, using the *trnLc/h* primer set, universally failed. The degraded nature of aDNA sequences thus makes it difficult to use conventional barcoding primers, as the lengths of resultant amplicons far exceed that which is realistically possible in aDNA studies (Valentini et al., 2009). The

use of short sections of mammalian genes is generally straightforward compared to that for plants, due to the coverage afforded them on GenBank and greater taxonomic certainty associated with this group. Nonetheless, the use of the hyper-variable p-loop region of the *trnL* intron for plants, although not without problems (Hollingsworth et al., 2011), provides sufficient taxonomic resolution in the case of this study. In most samples taxonomic assignment was possible to the family level, as was the case with previous morphological studies on these middens (Pearson, 1997; Meadows et al., 2010; Macphail, 2011). In several instances (Fig. S3a–d), it was possible to provide greater taxonomic resolution, to the genus level, than is possible using pollen; as the taxonomic resolution provided by fossil pollen in most of the families common to the arid zone is low. This is of particular value for families such as Poaceae that are highly diverse, but which – based on their pollen – are morphologically indistinguishable. For the sake of remaining cautious and conservative, however, such assignments are only dealt with peripherally in this study and the establishment of much better databases of reference material than currently exists is required to allow for greater certainty in taxonomic assignment at this level. In other words, datasets, like that compiled here, will have greater resolution in the future as databases become more comprehensive and flaws in the underpinning taxonomic framework are resolved.

The middens in this study have previously been analysed for pollen and macrofossil remains (Pearson, 1997; Meadows et al., 2010; Macphail, 2011) and thus provide a valuable point of comparison. The preservation of organic material is generally excellent in middens, with the presence and preservation of pollen and/or macrofossils varying from low and adequate in the BR midden to substantial and good in the TK midden. Whilst not guaranteeing the presence of aDNA, the survival of other biomolecular components in these samples suggests aDNA survival is at least plausible. Indeed, genetic analysis did detect the presence of a number of families previously identified in pollen and macrofossil analyses, as well as families and possible genera not previously detected in the midden samples (Fig. S3a–d). The presence of additional taxa, and the absence of previously identified taxa, further highlights that discussed in Jørgensen et al. (2012), namely that pollen, macrofossil and aDNA analyses are complementary as opposed to mutually exclusive and each provide ecological overviews with varying levels of taxonomic information. Moreover, the detection of extirpated (e.g. *T. vulpecula* not recorded in the region since the 1930's) and endemic taxa (e.g. *G. ocellaris*), in addition to results obtained independently at the Centre for GeoGenetics in Copenhagen, using cloning followed by Sanger sequencing, is strong evidence that argues for the authenticity of these aDNA sequences.

The lack of database coverage afforded certain taxa has proven problematic in this study. However, much of the difficulty associated with this issue is observed at a genus level and can be overcome through critical assessment of taxonomic assignments and the use of current, historical and modelled distribution data. Overall the database coverage problem, although cumbersome, has a limited impact upon the results of this particular study, and in general the results obtained in this study are plausible and in keeping with expected outcomes. In general, the taxa detected in the middens are known to occur in close proximity to the midden sites and reflect the climate at the sites, e.g. taxa detected in the Australian middens are generally all hot, arid or semi-arid adapted plants. In addition to this, there appears to be little overlap in taxa identified between samples, with the TK midden from South Africa, for instance, being noticeably distinct in terms of identified plant and mammalian taxa, when compared to the Australian middens. Finally, had there been significant modern environmental contamination of the samples arising from modern invasive taxa

found in the area, urinating on the middens for example, such as *Mus musculus* (house mouse), *Rattus rattus* (black rat) or *Vulpes vulpes* (red fox), or indeed contamination arising from reagents (Erlwein et al., 2011; Tuke et al., 2011), DNA from these taxa should have been detected, but were not. It is noted that unidentifiable Muridae sequences were detected, however, it is clear from phylogenetic analysis that these sequences do not group with the common contaminant *M. musculus*; they cluster, rather, with other native Australasian murids (Fig. S4). Indeed, not only does the amberat help to create an impermeable mass but its properties enable it to seal breaks in the weathering rind and discourage insect attack (Spaulding et al., 1990), further reducing possible exogenous contamination. This does not completely remove the possibility of “old” contamination, arising from the movement of material up through the stratigraphy of the midden (Spaulding and Robinson, 1984; Pearson and Dodson, 1993; McCarthy et al., 1996), although this is not an issue with the TK midden sample, as hyrax middens maintain stratigraphical integrity significantly better than rodent nest middens (Chase et al., 2012).

As noted previously there are a number of taxa that have been “detected” in the midden material that are somewhat problematic (Table 1). In some instances, such as the presence of Gesneriaceae in the YR midden or Amaryllidaceae in the BR midden, such taxa are not known to occur locally, at least in the present day flora. In other cases taxa have been “detected” that are not found natively in the country from which the midden was sampled, such as Torricelliaceae and Pinaceae in the TK and CR middens respectively. In the first instance, it is doubtful that there has been an extirpation or range contraction of the taxa identified. Gesneriaceae has a wholly east Australian distribution, whilst the closest record of Amaryllidaceae is over 350 km from the BR site. As regards to non-local or exotic taxa, with the exception of Pinaceae, which is a common laboratory and environmental contaminant, it is highly improbable that this is the result of laboratory or environmental contamination. The most likely explanation for such irregularities is a lack of coverage afforded certain taxa in current DNA databases (Taylor and Harris, 2012). In all the cases where disputed taxa have been identified there are records of related taxa (i.e. families within the same order, or genera within the same families) occurring in the area. In these cases there is little or no representation of these taxa in current DNA databases for *trnL* or other commonly used loci. For instance, in the case of sequences identified as Torricelliaceae (Order: Apiales), there are only two genera of Apiales known to occur at the site, neither of which are represented on GenBank; *Centella* and the rare, Western Cape endemic *Nanobubon* (Magee et al., 2008; Magee, 2012).

The genetic auditing of midden samples in this study also failed to identify families and genera, both plant and mammal, detected previously via morphological analyses. Previously identified plant taxa such as *Ptilotus* and Myoporaceae are not currently represented on GenBank, whilst mammalian taxa such as *Leporillus apicalis* and *Notomys* have no 16S, 12S or COI sequences on current databases either. However, insufficient database coverage of taxa fails to explain the absence of other important taxa such as *Acacia* (Family: Fabaceae) and *Eucalyptus* (Family: Myrtaceae). Both of these genera are useful indicators of habitat type and conditions and have been identified in previous analyses, at least to family level. In this study no *Eucalyptus* or *Acacia* sequences were identified, however Fabaceae sequences (possibly sub-family Mimosoideae) were detected. The *trnLg/h* primers used in this study have been tested successfully on *Acacia* and *Eucalyptus* reference samples and as such the absence of these taxa may be the result of primer biases, lack of genus-level resolution with the primers used, or simply a lack of DNA preservation and survival, which may vary between taxa or between preserved materials. Moreover,

the presence of pollen from certain taxa does not guarantee the retrieval of DNA from such taxa. Previous studies have had difficulties in amplifying DNA from pollen due to the limited amount of DNA contained within pollen grains (Parducci et al., 2005). Parducci et al. (2005) failed to retrieve plant DNA using *trnL* primers from horizons in which pollen from such plants was present. This may also serve to illustrate how it may be worthwhile to adopt a taxa specific approach in primer design to target important indicator species useful in the exploration in past environmental conditions and shifts, and highlights the value of multiple proxies in palaeoenvironmental reconstruction (discussed in detail in Jørgensen et al., 2012). Additionally, the absence of previously detected taxa and the converse, may also suggest that the source of aDNA recovered from these middens may be macrofossil in origin or DNA bound to, or within, the urea matrix, as opposed to pollen.

4.4. Future considerations

The preservation of DNA is a complex process that is at the mercy of a number of biotic and abiotic factors, which act in unison causing DNA degradation and damage (Hofreiter et al., 2001). Previous studies have shown that the survival of DNA is dependent not only on these factors but also the substrate in which DNA is found, which itself can mitigate the effects of DNA degradation and damage. Substrates such as hair (Gilbert et al., 2004) and eggshell (Oskam et al., 2010) are excellent at preserving DNA, with high levels of endogenous to microbial DNA, in comparison to bone for instance. In both of these cases, the substrate acts almost like a barrier to microbial attack, and in the case of hair in particular the substrate acts as a barrier to water. Midden material from cold, arid environments has been shown to preserve DNA over time (Kuch et al., 2002; Hofreiter et al., 2003a), and this study now shows that this is also the case with midden material from hot, arid environments. Hot, arid zone middens have very little moisture and the urine cementing the midden into a hard impermeable mass is highly ureic (Spaulding et al., 1990). These high levels of urea may serve as a means to further desiccate middens in environments that already lack a significant amount of moisture (Spaulding et al., 1990), thus aiding in the preservation of DNA. Moreover, lack of moisture therein also limits microbial induced DNA damage and degradation as well reducing hydrolytic damage. It would appear that the desiccation of midden material plays an important role in the long-term survival of DNA in middens. Importantly, these sites were in caves, rock shelters or overhangs and as such would have limited exposure to direct UV and weathering. It has been stated that there is much controversy surrounding aDNA claims arising from the study of hot, arid zone specimens and that there is a contrast between success rates of aDNA retrieval from similar sites of different ages (Schlumbaum et al., 2008; Gilbert, 2011). It is clear from these previous studies that the retrieval of aDNA from samples within hot, arid environments is much more sporadic than that involving samples obtained from frozen or cooler environments, possibly giving rise to these differing success rates.

Regardless of the issues surrounding the preservation of DNA in hot, arid environments, there are a number of practical recommendations that would aid in the exploration of present and past metabarcoding data. In order to benefit fully from the wealth of data produced by current sequencing technologies it is essential to have well-populated and informative DNA and environmental databases. Current DNA databases are not sufficient to allow fine resolution of sequencing data and this may prove to be a major obstacle in some studies (Taylor and Harris, 2012). However, as DNA sequencing methods become cheaper and more accessible, the

issues associated with insufficient database coverage are likely to diminish. In order to partly overcome this issue it is strongly recommended that a multi-primer approach targeting multiple loci be employed in environmental metabarcoding studies. This would provide a more comprehensive audit of environmental samples by reducing the effects of database biases and primer skews arising from preferential amplification. Although not used for plant screening in this study, this multi-locus approach was employed for mammal screening with clear benefits. The use of both 16S and 12S mammal specific primers allowed for the confirmation of the presence of certain taxa such as *P. capensis* in the TK midden, whilst also detecting taxa not identified through the use of one or the other, for example *G. ocellaris* in the TK midden (Table 2). Moreover, the detection of *G. ocellaris* would only have been possible using the 12S primer set, as neither 16S nor conventional COI sequences are on GenBank. This also holds true for many of the Muridae species known to occur in the areas where the Australian middens were found. For many, there exist no COI sequences for the currently accepted and approved COI barcode on GenBank or BOLD (Barcode of Life Database; <http://www.boldsystems.org/>), and the same can be said of 12S and 16S sequences. This further illustrates the importance of using multiple loci in metabarcoding studies at present, be they loci accepted by the barcoding community or otherwise.

In addition to genetic databases, environmental databases using current and historical records of taxa distribution are invaluable in environmental metabarcoding studies. Databases such as ALA and SANBI, coupled with historical records, are immensely useful to truth and validate data or to detect possible range shifts of identified taxa. In relation to historical and ancient samples, the macro- and microscopic examination of environmental samples is highly valuable in determining the likelihood of DNA preservation and in the corroboration of genetic results, thereby improving data fidelity. This highlights the need for co-operation and collaboration between multiple disciplines ranging from palaeontology and archaeology, molecular biology and biochemistry through to ecology and botany. Through this concerted cross-disciplinary effort it would be possible to gain a more robust insight into both past and present environments.

4.5. Conclusion

The survival and preservation of DNA in hot, arid environments is a complex and poorly understood process. Most of the few studies that have attempted to retrieve aDNA from samples in such environments have been a source of controversy and dispute. The results in this study have been dealt with critically and overall they are both plausible and consistent with predicted outcomes and previous analyses of the same samples. Although further empirical research is needed to assess the survival of DNA in midden material, it appears that DNA survival through accumulation and desiccation may be important in relation to samples from hot environments, and middens in general. Furthermore, it is apparent that neither the age of the samples nor the temperature at which they have been preserved, albeit important, can be grounds for the rejection of results. The preservation of DNA from hot environments, it suffices to say, is sporadic and rare.

Nonetheless, herbivore middens with their excellent preservational qualities now present an important source of material for DNA metabarcoding studies of past hot, arid environments, especially when palaeoenvironmental data is lacking, as was the case with the Brockman Ridge sample. As such, sampling procedures should be revised to ensure samples are collected in such a way as to allow for aDNA techniques to be applied. The retrieval

of aDNA from midden material is not unique to Australia, as is evidenced by the results from South Africa and previous South American genetic studies. This study has wider implications for the analysis of midden material throughout hot, arid and semi-arid environments across the globe. Multidisciplinary investigations of midden material using stable isotopes, aDNA, pollen, macrofossils and dating will build knowledge of palaeoenvironments and inform conservation and rehabilitation policies. Such data will ensure the maintenance and survival of ecologically important taxa and communities within fragile arid environments, which are increasingly under anthropogenic induced threats.

Acknowledgements

DNA studies of these midden samples were funded by Australian Research Council (ARC) grants (DP0771971 and FT0991741). Fieldwork (CR, YR) was possible with an ARC grant awarded to J. Dodson and the support of D. Pearson and the School of Geography, University of New South Wales. Fieldwork (BR) was undertaken by Scarp Archaeology with the approval of the Puutu Kunti Kurrama and Pinikura group who assisted in the archaeological work that was funded by Rio Tinto Iron Ore. Radiocarbon dating of CR & YR material was made possible by grants from the Australian Institute of Nuclear Science (03/704) and permits from the Western Australian Department of Conservation and Land Management. Support was also received from the European Research Council (ERC) Starting Grant project HYRAX (no. 258657), and the Leverhulme Trust grant F/08 773/C. The authors acknowledge the support and contribution of Eske Willerslev (Centre for GeoGenetics), Fred Ford (Department of Defence), Ms Frances Brigg (State Agricultural Biotechnology Centre) and computational support from the iVEC Informatics Facility.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.quascirev.2012.10.021>.

References

- Adcock, G.J., Dennis, E.S., Easteal, S., Huttley, G.A., Jermin, L.S., Peacock, W.J., Thorne, A., 2001. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc. Natl. Acad. Sci.* 98, 537–542.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kær, K.H., Orlando, L., Gilbert, M.T.P., Willerslev, E., 2011. Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Mol. Ecol.* 21, 1966–1979.
- Atchison, J., 2010. Short Report on Pilbara Amberat Samples from Brock 12, Pilbara, Western Australia. Unpublished, Prepared for Scarp Archaeology.
- Beadle, N.C.W., 1966. Soil phosphate and its role in moulding segments of Australian flora and vegetation with special reference to xeromorphy and sclerophylly. *Ecology* 47, 992–1020.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2006. GenBank. *Nucleic Acids Res.* 34, D16–D20.
- Bonnichsen, R., Hodges, L., Ream, W., Field, K.G., Kirner, D.L., Selsor, K., Taylor, R.E., 2001. Methods of the study of ancient hair: radiocarbon dates and gene sequences from individual hairs. *J. Archaeol. Sci.* 28, 775–785.
- Champlot, S., Berthelot, C., Pruvost, M.L., Bennett, E.A., Grange, T., Geigl, E.-M., 2010. An efficient multistrategy DNA decontamination procedure of PCR reagents for hypersensitive PCR applications. *PLoS One* 5, e13042.
- Chariton, A.A., Court, L.N., Hartley, D.M., Colloff, M.J., Hardy, C.M., 2010. Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Front. Ecol. Environ.* 8, 233–238.
- Chase, B.M., Meadows, M.E., Scott, L., Thomas, D.S.G., Marais, E., Sealy, J., Reimer, P.J., 2009. A record of rapid Holocene climate change preserved in hyrax middens from southwestern Africa. *Geology* 37, 703–706.
- Chase, B.M., Quick, L.J., Meadows, M.E., Scott, L., Thomas, D.S.G., Reimer, P.J., 2011. Late glacial interhemispheric climate dynamics revealed in South African hyrax middens. *Geology* 39, 19–22.

- Chase, B.M., Scott, L., Meadows, M.E., Gil-Romera, G., Boom, A., Carr, A.S., Reimer, P.J., Truc, L., Valsecchi, V., Quick, L.J., 2012. Rock hyrax middens: a palaeoenvironmental archive for southern African drylands. *Quat. Sci. Rev.* 56, 107–125.
- Clarke, E., 2010. A Short Report Detailing the Salvage of Sites Brock-11 and Brock-12. Unpublished, Prepared for Rio Tinto.
- Cooper, A., Poinar, H.N., 2000. Ancient DNA: do it right or not at all. *Science* 289, 1139.
- Cooper, A., Rambaut, A., Macaulay, V., Willerslev, E., Hansen, A.J., Stringer, C., 2001. Human origins and ancient human DNA. *Science* 292, 1655–1656.
- Deagle, B., Chiaradia, A., McInnes, J., Jarman, S., 2010. Pyrosequencing faecal DNA to determine diet of little penguins: Is what goes in what comes out? *Conserv. Genet.* 11, 2039–2048.
- Deagle, B.E., Kirkwood, R., Jarman, S.N., 2009. Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Mol. Ecol.* 18, 2022–2038.
- Dial, K.P., Czaplewski, N.J., 1990. Do woodrat middens accurately represent the animals' environment and diets? The Woodhouse Mesa study. In: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), *Packrat Middens: the Last 40,000 Years of Biotic Change*. University of Arizona, Tucson, pp. 43–58.
- Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moir, R., Stones-Havas, S., Sturrock, S., Thierer, T., Wilson, A., 2011. Geneious V5.4. Available from: <http://www.geneious.com/>.
- Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200.
- Erlwein, O., Robinson, M.J., Dustan, S., Weber, J., Kaye, S., McClure, M.O., 2011. DNA extraction columns contaminated with murine sequences. *PLoS One* 6, e23484.
- Fall, P.L., Lindquist, C.A., Falconer, S.E., 1990. Fossil Hyrax middens from the Middle East: a record of paleovegetation and human disturbance. In: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), *Packrat Middens: the Last 40,000 Years of Biotic Change*. University of Arizona, Tucson, pp. 398–407.
- Ficetola, G.F., Miaud, C., Pompanon, F., Taberlet, P., 2008. Species detection using environmental DNA from water samples. *Biol. Lett.* 4, 423–425.
- Friedel, M.H., Pickup, G., Nelson, D.J., 1993. The interpretation of vegetation change in a spatially and temporally diverse arid Australian landscape. *J. Arid Environ.* 24, 241–260.
- Gilbert, M.T.P., 2011. The mummy returns... and sheds new light on old questions. *Mol. Ecol.* 20, 4195–4198.
- Gilbert, M.T.P., Bandelt, H.-J., Hofreiter, M., Barnes, I., 2005a. Assessing ancient DNA studies. *Trends Ecol. Evol.* 20, 541–544.
- Gilbert, M.T.P., Barnes, I., Collins, M.J., Smith, C., Eklund, J., Goudsmit, J., Poinar, H., Cooper, A., 2005b. Long-term survival of ancient DNA in Egypt: response to Zink and Nerlich (2003). *Am. J. Phys. Anthropol.* 128, 110–114.
- Gilbert, T.P., Wilson, A.S., Bunce, M., Hansen, A.J., Willerslev, E., Shapiro, B., Higham, T.F., Richards, M.P., O'Connell, T.C., Tobin, D.J., Janaway, R.C., Cooper, A., 2004. Ancient mitochondrial DNA from hair. *Curr. Biol.* 14, R463–R464.
- Gomez-Alvarez, V., Teal, T.K., Schmidt, T.M., 2009. Systematic artifacts in metagenomes from complex microbial communities. *ISME J.* 3, 1314–1317.
- Griffiths, R.I., Thomson, B.C., James, P., Bell, T., Bailey, M., Whiteley, A.S., 2011. The bacterial biogeography of British soils. *Environ. Microbiol.* 13, 1642–1654.
- Groves, R.H., 1994. *Australian Vegetation*, second ed. Cambridge University Press, Cambridge.
- Haile, J., 2011. Ancient DNA extraction from soils and sediments. In: Shapiro, B., Hofreiter, M. (Eds.), *Methods in Molecular Biology – Ancient DNA*. Humana Press Series, pp. 57–63.
- Haile, J., Froese, D.G., MacPhee, R.D.E., Roberts, R.G., Arnold, L.J., Reyes, A.V., Rasmussen, M., Nielsen, R., Brook, B.W., Robinson, S., Demuro, M., Gilbert, M.T.P., Munch, K., Austin, J.J., Cooper, A., Barnes, I., Moller, P., Willerslev, E., 2009. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci.* 106, 22352–22357.
- Hayward, G.F., Phillipson, J., 1979. Community structure and the functional role of small mammals in ecosystems. In: Stoddart, M. (Ed.), *Ecology of Small Mammals*. Chapman & Hall, London, pp. 135–211.
- Hekkala, E., Shirley, M.H., Amato, G., Austin, J.D., Charter, S., Thorbjarnarson, J., Vliet, K.A., Houck, M.L., Desalle, R.O.B., Blum, M.J., 2011. An ancient icon reveals new mysteries: mummy DNA resurrects a cryptic species within the Nile crocodile. *Mol. Ecol.* 20, 4199–4215.
- Hijmans, R., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25, 1965–1978.
- Hofreiter, M., Betancourt, J.L., Pelliza Sbriller, A., Markgraf, V., McDonalde, H.G., 2003a. Phylogeny, diet, and habitat of an extinct ground sloth from Cuchillo Curá, Neuquén Province, southwest Argentina. *Quatern. Res.* 59, 364–378.
- Hofreiter, M., Mead, J.I., Martin, P., Poinar, H.N., 2003b. Molecular caving. *Curr. Biol.* 13, R693–R695.
- Hofreiter, M., Poinar, H.N., Spaulding, W.G., Bauer, K., Martin, P.S., Possnert, G., Pääbo, S., 2000. A molecular analysis of ground sloth diet through the last glaciation. *Mol. Ecol.* 9, 1975–1984.
- Hofreiter, M., Serre, D., Poinar, H., Kuch, M., Pääbo, S., 2001. Ancient DNA. *Nat. Rev. Genet.* 2, 353–359.
- Hollingsworth, P.M., Graham, S.W., Little, D.P., 2011. Choosing and using a plant DNA barcode. *PLoS One* 6, e19254.
- How, R.A., Hillcox, S.J., 2000. Brushtail possum, *Trichosurus vulpecula*, populations in south-western Australia: demography, diet and conservation status. *Wildl. Res.* 27, 81–89.
- Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755.
- Hunter, A.A., Macgregor, A.B., Szabo, T.O., Wellington, C.A., Bellgard, M.I., 2012. Yabi: an online research environment for grid, high performance and cloud computing. *Source Code Biol. Med.* 7, 1.
- Huson, D.H., Auch, A.F., Qi, J., Schuster, S.C., 2007. MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386.
- Jørgensen, T., Haile, J., Möller, P.E.R., Andreev, A., Boessenkool, S., Rasmussen, M., Kienast, F., Coissac, E., Taberlet, P., Brochmann, C., Bigelow, N.H., Andersen, K., Orlando, L., Gilbert, M.T.P., Willerslev, E., 2012. A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Mol. Ecol.* 21, 1989–2003.
- Kuch, M., Rohland, N., Betancourt, J.L., Latorre, C., Steppan, S., Poinar, H.N., 2002. Molecular analysis of a 11 700-year-old rodent midden from the Atacama Desert, Chile. *Mol. Ecol.* 11, 913–924.
- Kuch, M., Sobolik, K., Barnes, I., Stankiewicz, B.A., Spaulding, G., Bryant, V., Cooper, A., Pääbo, S., 2001. A molecular analyses of the dietary diversity for three archaic native Americans. *Proc. Natl. Acad. Sci. U S A* 98, 4317–4322.
- Lindahl, T., 1993a. Instability and decay of the primary structure of DNA. *Nature* 362, 709–715.
- Lindahl, T., 1993b. Recovery of antediluvian DNA. *Nature* 365, 700.
- Macphail, M., 2011. Palynological Analyses, 30 ka 'Amberat' Deposit, Brockman Ridge, Pilbara Region, Western Australia. Unpublished, Prepared for Scarp Archaeology, Canberra.
- Agee, A.R., 2012. *Nanobubon hypogaeum* (Apiaceae), a new contractile-rooted species from the Western Cape Province of South Africa. *S. Afr. J. Bot.* 80, 63–66.
- Agee, A.R., Van Wyk, B.-E., Tilney, P.M., 2008. A taxonomic revision of the genus *Nanobubon* (Apiaceae: Apioidae). *S. Afr. J. Bot.* 74, 713–719.
- McCarthy, L., Head, L., Quade, J., 1996. Holocene palaeoecology of the northern Flinders Ranges, South Australia, based on stick-nest rat (*Leporillus* spp.) middens: a preliminary overview. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 123, 1205–1218.
- Meadows, M.E., Chase, B.M., Seliane, M., 2010. Holocene palaeoenvironments of the Cederberg and Swarttruggens mountains, Western Cape, South Africa: pollen and stable isotope evidence from hyrax dung middens. *J. Arid Environ.* 74, 789–793.
- Mitchell, A.A., Wilcox, D.G., 1994. *Arid Shrubland Plants of Western Australia*, second ed. University of Western Australia Press, Perth.
- Moore, C.W.E., 1953. The vegetation of the south-eastern Riverina, New South Wales. II. The disclimax communities. *Aust. J. Bot.* 1, 548–567.
- Murray, D., Bunce, M., Cannell, B.L., Oliver, R., Houston, J., White, N.E., Barrero, R.A., Bellgard, M.I., Haile, J., 2011. DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6, e25776.
- Northcote, K.H., Wright, M.J., 1982. Soil landscapes of arid Australia. In: Barker, W.R., Greenslade, P.J.M. (Eds.), *Evolution of the Flora and Fauna of Arid Australia*. Peacock Publications, Adelaide, pp. 15–21.
- Oskam, C.L., Haile, J., McLay, E., Rigby, P., Allentoft, M.E., Olsen, M.E., Bengtsson, C., Miller, G.H., Schwenninger, J.L., Jacomb, C., Walter, R., Baynes, A., Dortch, J., Parker-Pearson, M., Gilbert, M.T.P., Holdaway, R.N., Willerslev, E., Bunce, M., 2010. Fossil avian eggshell preserves ancient DNA. *Proc. R. Soc. B-biol. Sci.* 277, 1991–2000.
- Parducci, L., Suyama, Y., Lascoux, M., Bennett, K.D., 2005. Ancient DNA from pollen: a genetic record of population history in Scots pine. *Mol. Ecol.* 14, 2873–2882.
- Pearson, S., 1997. Stick-nest Rat Middens as a Source of Palaeo-environmental Data in Central Australia. School of Geography, Faculty of Science. University of New South Wales, Sydney.
- Pearson, S., Betancourt, J.L., 2002. Understanding arid environments using fossil rodent middens. *J. Arid Environ.* 50, 499–511.
- Pearson, S., Dodson, J., 1993. Stick-nest rat middens as sources of paleoecological data in Australian deserts. *Quatern. Res.* 39.
- Pearson, S.G., Triggs, B.E., Baynes, A., 2001. The record of fauna, and accumulating agents of hair and bone, found in middens of stick-nest rats (Genus *Leporillus*) (Rodentia: Muridae). *Wildl. Res.* 28, 435–444.
- Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* 11, 1633–1644.
- Pegard, A., Miquel, C., Valentini, A., Coissac, E., Bouvier, F., François, D., Taberlet, P., Engel, E.F.P., 2009. Universal DNA-based methods for assessing the diet of grazing livestock and wildlife from faeces. *J. Agric. Food Chem.* 57, 5700–5706.
- Poinar, H., Kuch, M., McDonald, G., Martin, P., Pääbo, S., 2003. Nuclear gene sequences from a late Pleistocene sloth coprolite. *Curr. Biol.* 12, 1150–1152.
- Poinar, H.N., Hofreiter, M., Spaulding, W.G., Martin, P.S., Stankiewicz, B.A., Bland, H., Evershed, R.P., Possnert, G., Pääbo, P., 1998. Molecular coproscopy: dung and diet of the extinct Ground Sloth *Nothrotheriops shastensis*. *Science* 281, 402–406.
- Poinar, H.N., Kuch, M., Sobolik, K.D., Barnes, I., Stankiewicz, B.A., Kuder, T., Spaulding, W.G., Bryant, V.M., Cooper, A., Pääbo, S., 2001. A molecular analysis of dietary diversity for three archaic Native Americans. *Proc. Natl. Acad. Sci. U S A* 98, 4317–4322.
- Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R.D.E., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., Rampp, M., Miller, W., Schuster, S.C., 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 311, 392–394.
- Pons, A., Quézel, P., 1958. Premières remarques sur l'étude palynologique d'un guano fossile du Hoggar. *C. R. Acad. Sci.* 244, 2290–2292.

- Ritchie, J.C., 1986. Climate change and vegetation response. *Vegetation* 67, 65–74.
- Roche, 2009. Technical Bulletin: Amplicon Fusion Primer Design Guidelines for GS FLX Titanium Series Lib-a Chemistry, TCB No. 013-2009, pp. 1–3.
- Rusch, D.B., Halpern, A.L., Sutton, G., Heidelberg, K.B., Williamson, S., Yooseph, S., Wu, D., Eisen, J.A., Hoffman, J.M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J.E., Li, K., Kravitz, S., Heidelberg, J.F., Utterback, T., Rogers, Y.H., Falcón, L.L., Souza, V., Bonilla-Rosso, G., Eguiarte, L.E., Karl, D.M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M.R., Strausberg, R.L., Nealson, K., Friedman, R., Frazier, M., Venter, J.C., 2007. The sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5
- Schlumbaum, A., Tensen, M., Jaenicke-Després, V., 2008. Ancient plant DNA in archaeobotany. *Veg. Hist. Archaeobot.* 17, 233–244.
- Scott, L., 1990. Hyrax (Procaviidae) and Dassie Rat (Petromuridae) middens in paleoenvironmental studies in Africa. In: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), *Packrat Middens: the Last 40,000 Years of Biotic Change*. University of Arizona, Tucson, pp. 398–407.
- Scott, L., Marais, E., Brook, G.A., 2004. Fossil hyrax dung and evidence of Late Pleistocene and Holocene vegetation types in the Namib Desert. *J. Quat. Sci.* 19, 829–832.
- Scott, L., Woodborne, S., 2007. Pollen analysis and dating of late Quaternary faecal deposits (hyraceum) in the Cederberg, Western Cape, South Africa. *Rev. Palaeobot. Palynol.* 144, 123–134.
- Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* 21, 1794–1805.
- Smith, C.I., Chamberlain, A.T., Riley, M.S., Cooper, A., Stringer, C.B., Collins, M.J., 2001. Neanderthal DNA. Not just old but old and cold? *Nature* 410, 771–772.
- Smith, C.I., Chamberlain, A.T., Riley, M.S., Stringer, C., Collins, M.J., 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. *J. Hum. Evol.* 45, 203–217.
- Spaulding, W.G., Betancourt, J.L., Croft, L.K., Cole, K.L., 1990. Packrat middens: their composition and methods of analysis. In: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), *Packrat Middens: the Last 40,000 Years of Biotic Change*. University of Arizona, Tucson, pp. 59–84.
- Spaulding, W.G., Robinson, S.W., 1984. Preliminary Assessment of Climatic Change During the Later Wisconsin Time, Southern Great Basin and Vicinity. U.S. Geological Survey, Denver.
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., Willerslev, E., 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* 21, 2045–2050.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermet, T., Corthier, G., Brochmann, C., Willerslev, E., 2007. Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* 35, e14.
- Taberlet, P., Gielly, L., Pautou, G., Bouvet, J., 1991. Universal primers for amplification of three noncoding regions of chloroplast DNA. *Plant Mol. Biol.* 17, 1105–1109.
- Tausch, R.J., Wigand, P.E., Burkhardt, J.W., 1993. Viewpoint – plant community thresholds, multiple steady states and multiple successional pathways: legacy of the Quaternary? *J. Rangeland Manag.* 46, 439–447.
- Taylor, H.R., Harris, W.E., 2012. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol. Ecol. Resour.* 12, 377–388.
- Taylor, P.G., 1996. Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Mol. Biol. Evol.* 13, 283–285.
- Thackway, R., Cresswell, I.D., 1995. An Interim Biogeographic Regionalisation for Australia: a Framework for Setting Priorities in the National Reserves System Cooperative Program. Reserve System Unit, Australian Nature Conservation Agency, Canberra.
- Thomsen, P.F., Kielgast, J.O.S., Iversen, L.L., Wiuf, C., Rasmussen, M., Gilbert, M.T.P., Orlando, L., Willerslev, E., 2012. Monitoring endangered freshwater biodiversity using environmental DNA. *Mol. Ecol.* 21, 2565–2573.
- Tongway, D.J., Ludwig, J.A., 1990. Vegetation and soil patterning in semi-arid mulga lands of Eastern Australia. *Aust. J. Ecol.* 15, 23–34.
- Trabucco, A., Zomer, R.J., 2009. Global Aridity Index (Global-aridity) and Global Potential Evapo-transpiration (Global-PET) Geospatial Database. CGIAR Consortium for Spatial Information. Published online, Available from: the CGIAR-CSI GeoPortal at: <http://www.csi.cgiar.org/>.
- Tuke, P.W., Tettmar, K.I., Tamuri, A., Stoye, J.P., Tedder, R.S., 2011. PCR master mixes harbour murine DNA sequences. Caveat emptor! *PLoS One* 6, e19953.
- Valentini, A., Miquel, C., Nawaz, M.A., Bellemain, E.V.A., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J.E., Taberlet, P., 2009. New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the *trnL* approach. *Mol. Ecol. Resour.* 9, 51–60.
- Van Devender, T.R., 1990. Late Quaternary vegetation and climate of the Sonoran Desert, United States and Mexico. In: Betancourt, J.L., Van Devender, T.R., Martin, P.S. (Eds.), *Packrat Middens: the Last 40,000 Years of Biotic Change*. University of Arizona, Tucson, pp. 134–164.
- Van Devender, T.R., Spaulding, W.G., 1979. Development of vegetation and climate in the south western United States. *Science* 204, 701–710.
- Vila, A.R., Borrelli, L., 2011. Cattle in the Patagonian forests: feeding ecology in Los Alerces National Reserve. *For. Ecol. Manag.* 261, 1306–1314.
- Wells, P.V., Jorgensen, C.D., 1964. Pleistocene wood rat middens and climatic change in Mohave Desert – a record of juniper woodlands. *Science* 143, 1171–1174.
- Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T.P., Shapiro, B., Bunce, M., Wiuf, C., Gilichinsky, D.A., Alan, C., 2003. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science* 300, 791–795.

SPECIAL ISSUE: MOLECULAR DETECTION OF TROPHIC INTERACTIONS

Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed

JOANNA M. BURGAR,* DAITHI C. MURRAY,† MICHAEL D. CRAIG,*‡ JAMES HAILE,† JAYNE HOUSTON,† VICKI STOKES§ and MICHAEL BUNCE†

*School of Veterinary and Life Sciences, Murdoch University, 90 South Street, Murdoch, WA 6150, Australia, †Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, 90 South Street, Murdoch, WA 6150, Australia, ‡School of Plant Biology, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia, §Alcoa of Australia Ltd., PO Box 252, Applecross, WA 6953, Australia

Abstract

Effective management and conservation of biodiversity requires understanding of predator–prey relationships to ensure the continued existence of both predator and prey populations. Gathering dietary data from predatory species, such as insectivorous bats, often presents logistical challenges, further exacerbated in biodiversity hot spots because prey items are highly speciose, yet their taxonomy is largely undescribed. We used high-throughput sequencing (HTS) and bioinformatic analyses to phylogenetically group DNA sequences into molecular operational taxonomic units (MOTUs) to examine predator–prey dynamics of three sympatric insectivorous bat species in the biodiversity hotspot of south-western Australia. We could only assign between 4% and 20% of MOTUs to known genera or species, depending on the method used, underscoring the importance of examining dietary diversity irrespective of taxonomic knowledge in areas lacking a comprehensive genetic reference database. MOTU analysis confirmed that resource partitioning occurred, with dietary divergence positively related to the ecomorphological divergence of the three bat species. We predicted that bat species' diets would converge during times of high energetic requirements, that is, the maternity season for females and the mating season for males. There was an interactive effect of season on female, but not male, bat species' diets, although small sample sizes may have limited our findings. Contrary to our predictions, females of two ecomorphologically similar species showed dietary convergence during the mating season rather than the maternity season. HTS-based approaches can help elucidate complex predator–prey relationships in highly speciose regions, which should facilitate the conservation of biodiversity in genetically uncharacterized areas, such as biodiversity hotspots.

Keywords: *Chalinolobus gouldii*, dietary differentiation, molecular scatology, next-generation sequencing, *Nyctophilus gouldi*, *Vespadelus regulus*

Received 18 February 2013; revision received 5 September 2013; accepted 13 September 2013

Introduction

To effectively manage and conserve biodiversity, it is critical to understand predator–prey relationships so that both predator and prey populations can be

conserved. This is becoming increasingly important as continuing habitat loss and degradation may lead to trophic collapse (Dobson *et al.* 2006). Accurate dietary studies can contribute greatly to understanding predator–prey relationships and can also provide integral knowledge concerning food webs and trophic interactions, which in turn influence ecological processes such as niche partitioning and interspecific competition

Correspondence: Joanna Burgar, Fax: +61 (0)8 9360 6306; E-mail: joburgar@gmail.com

Contents lists available at [ScienceDirect](http://www.sciencedirect.com)

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

Thorough assessment of DNA preservation from fossil bone and sediments excavated from a late Pleistocene–Holocene cave deposit on Kangaroo Island, South Australia



Dalal Haouchar^a, James Haile^a, Matthew C. McDowell^b, Dáithí C. Murray^a,
Nicole E. White^a, Richard J.N. Allcock^c, Matthew J. Phillips^d, Gavin J. Prideaux^b,
Michael Bunce^{a,e,*}

^a Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, Murdoch, WA 6150, Australia

^b School of Biological Sciences, Flinders University, Bedford Park, SA 5042, Australia

^c LotteryWest State Biomedical Facility: Genomics, School of Pathology and Laboratory Medicine, The University of Western Australia, Nedlands, WA 6009, Australia

^d School of Earth, Environmental and Biological Sciences, Queensland University of Technology, Brisbane, QLD 4001, Australia

^e Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, WA 6845, Australia

ARTICLE INFO

Article history:

Received 9 August 2013

Received in revised form

6 November 2013

Accepted 10 November 2013

Available online

Keywords:

Pleistocene–Holocene

Quaternary

Fossils

Ancient DNA

Biodiversity

ABSTRACT

Fossils and sediments preserved in caves are an excellent source of information for investigating impacts of past environmental changes on biodiversity. Until recently studies have relied on morphology-based palaeontological approaches, but recent advances in molecular analytical methods offer excellent potential for extracting a greater array of biological information from these sites. This study presents a thorough assessment of DNA preservation from late Pleistocene–Holocene vertebrate fossils and sediments from Kelly Hill Cave Kangaroo Island, South Australia. Using a combination of extraction techniques and sequencing technologies, ancient DNA was characterised from over 70 bones and 20 sediment samples from 15 stratigraphic layers ranging in age from >20 ka to ~6.8 ka. A combination of primers targeting marsupial and placental mammals, reptiles and two universal plant primers were used to reveal genetic biodiversity for comparison with the mainland and with the morphological fossil record for Kelly Hill Cave. We demonstrate that Kelly Hill Cave has excellent long-term DNA preservation, back to at least 20 ka. This contrasts with the majority of Australian cave sites thus far explored for ancient DNA preservation, and highlights the great promise Kangaroo Island caves hold for yielding the hitherto-elusive DNA of extinct Australian Pleistocene species.

© 2013 Elsevier Ltd. All rights reserved.

1. Introduction

Islands have long provided a natural laboratory for the study of evolutionary processes because evolutionary changes on them are often magnified, simplified and therefore more readily interpretable (e.g., Darwin and Wallace, 1858; MacArthur and Wilson, 1967; Losos and Ricklefs, 2010). The study of genetic variation on islands also has a long history (Lomolino et al., 1989; Van der Geer et al., 2010). However, ancient DNA (aDNA) analyses applied to stratified, dated faunal successions can add a temporal context, allowing

the ebb and flow of genes, species and communities to be assessed, particularly in combination with more traditional analyses of vertebrate and plant macrofossils and pollen. A necessary prerequisite for aDNA research is adequate biomolecular preservation. Cave systems represent an ideal environment for palaeontological investigations as they often contain relatively complete and undisturbed stratigraphic deposits that harbour several environmental proxies (White, 2007; Butzer, 2008) that have been subjected to minimal temperature and humidity fluctuations; conditions that favour DNA persistence (Stone, 2000). Such caves represent archives of well-preserved Quaternary vertebrate assemblages (Prideaux et al., 2007, 2010), with the ability to preserve invaluable repositories of past biodiversity. All samples (bones and sediments) analysed in this study were obtained directly from Kelly Hill Cave (KHC), Kangaroo Island (KI) with the aim of conducting a

* Corresponding author. Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, WA 6845, Australia.

E-mail address: michael.bunce@curtin.edu.au (M. Bunce).

RESEARCH

Open Access

Metabarcoding avian diets at airports: implications for birdstrike hazard management planning

Megan L Coghlan¹, Nicole E White¹, Dáithí C Murray¹, Jayne Houston¹, William Rutherford², Matthew I Bellgard³, James Haile¹ and Michael Bunce^{1*}

Abstract

Background: Wildlife collisions with aircraft cost the airline industry billions of dollars per annum and represent a public safety risk. Clearly, adapting aerodrome habitats to become less attractive to hazardous wildlife will reduce the incidence of collisions. Formulating effective habitat management strategies relies on accurate species identification of high-risk species. This can be successfully achieved for all strikes either through morphology and/or DNA-based identifications. Beyond species identification, dietary analysis of birdstrike gut contents can provide valuable intelligence for airport hazard management practices in regards to what food is attracting which species to aerodromes. Here, we present birdstrike identification and dietary data from Perth Airport, Western Australia, an aerodrome that saw approximately 140,000 aircraft movements in 2012. Next-generation high throughput DNA sequencing was employed to investigate 77 carcasses from 16 bird species collected over a 12-month period. Five DNA markers, which broadly characterize vertebrates, invertebrates and plants, were used to target three animal mitochondrial genes (12S rRNA, 16S rRNA, and COI) and a plastid gene (*trnL*) from DNA extracted from birdstrike carcass gastrointestinal tracts.

Results: Over 151,000 DNA sequences were generated, filtered and analyzed by a fusion-tag amplicon sequencing approach. Across the 77 carcasses, the most commonly identified vertebrate was *Mus musculus* (house mouse). Acrididae (grasshoppers) was the most common invertebrate family identified, and Poaceae (grasses) the most commonly identified plant family. The DNA-based dietary data has the potential to provide some key insights into feeding ecologies within and around the aerodrome.

Conclusions: The data generated here, together with the methodological approach, will greatly assist in the development of hazard management plans and, in combination with existing observational studies, provide an improved way to monitor the effectiveness of mitigation strategies (for example, netting of water, grass type, insecticides and so on) at aerodromes. It is hoped that with the insights provided by dietary data, airports will be able to allocate financial resources to the areas that will achieve the best outcomes for birdstrike reduction.

Keywords: Birdstrike, Diet analysis, Species identification, Birdstrike management, Airport, Food chain

* Correspondence: m.bunce@icloud.com

¹Australian Wildlife Forensic Services and Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, Murdoch, Western Australia 6150, Australia

Full list of author information is available at the end of the article



OPEN

SUBJECT AREAS:

PALAEOLOGY

ARCHAEOLOGY

BIODIVERSITY

Received
3 May 2013Accepted
5 November 2013Published
28 November 2013

Correspondence and requests for materials should be addressed to M.B. (michael.bunce@curtin.edu.au)

* Current address: Trace and Environmental DNA laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, 6845, Australia.

Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages

Dáithí C. Murray^{1*}, James Haile^{1*}, Joe Dortch², Nicole E. White^{1*}, Dalal Haouchar¹, Matthew I. Bellgard³, Richard J. Allcock⁴, Gavin J. Prideaux⁵ & Michael Bunce^{1*}

¹Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, South Street, Murdoch, WA, 6150, Australia, ²Eureka Archaeological Research and Consulting, School of Social Sciences, The University of Western Australia, Crawley, Perth, WA, 6009, Australia, ³Centre for Comparative Genomics, Murdoch University, South Street, Murdoch, WA, 6150, Australia, ⁴LotteryWest State Biomedical Facility: Genomics, School of Pathology and Laboratory Medicine, The University of Western Australia, Nedlands, WA, 6009, Australia, ⁵School of Biological Sciences, Flinders University, Bedford Park, SA, 5042, Australia.

Highly fragmented and morphologically indistinct fossil bone is common in archaeological and palaeontological deposits but unfortunately it is of little use in compiling faunal assemblages. The development of a cost-effective methodology to taxonomically identify bulk bone is therefore a key challenge. Here, an ancient DNA methodology using high-throughput sequencing is developed to survey and analyse thousands of archaeological bones from southwest Australia. Fossils were collectively ground together depending on which of fifteen stratigraphical layers they were excavated from. By generating fifteen synthetic blends of bulk bone powder, each corresponding to a chronologically distinct layer, samples could be collectively analysed in an efficient manner. A diverse range of taxa, including endemic, extirpated and hitherto unrecorded taxa, dating back to c.46,000 years BP was characterized. The method is a novel, cost-effective use for unidentifiable bone fragments and a powerful molecular tool for surveying fossils that otherwise end up on the taxonomic “scrapheap”.

Fossil assemblages offer insights into past biodiversity, palaeoecology and human activities^{1–3}. However, the accuracy of fossil identifications relies on the preservation of taxonomically significant morphological features, which are often lacking in highly fragmented remains. Over the past decade, analyses of ancient DNA (aDNA) have developed in sophistication and the breadth of contexts in which they are applied. Ancient DNA has been used to address questions of speciation, extinction and disease^{4–7} using a variety of substrates, including bone⁸, hair⁹ and eggshell¹⁰. However, to date, no study has attempted to use aDNA from taxonomically diverse fossils to map faunal assemblage data from a single site, largely due to the time and cost associated with generating aDNA sequences from each bone fragment.

The destructive nature of sampling also means researchers and collection managers may be reluctant to analyse valuable specimens. At the same time, most archaeological and palaeontological excavations also collect large numbers of small, morphologically indistinct bone fragments (Figure 1a). Such material is of limited use in species identifications, although it may be important for some taphonomic analyses. Taxonomically, however, it is usually destined for the analytical “scrapheap”.

It is now possible, largely due to second generation high-throughput DNA sequencing (HTS) methodologies, to genetically profile complex, heterogeneous samples (Figure 1b) in parallel, both cheaply and quickly^{11,12}. This DNA metabarcoding¹³ approach to genetically unravel complex substrates via HTS, as opposed to cloning, has transformed the analysis of substrates such as sediment^{14,15} and faecal material^{16,17}. To explore large HTS-generated genomic datasets from environmental samples researchers use tools that are either: 1) taxonomy-dependent, which involves searching DNA reference databases for query and reference sequence matches^{18,19}, or 2) taxonomy-independent, which involves taxonomy-independent measures of sequence diversity and clustering such as Operational Taxonomic Unit (OTU) analysis or UniFrac-based methods^{20–22}.

This study seeks to employ HTS technology to sequence and identify aDNA obtained from thousands of morphologically unidentifiable archaeological bone fragments freshly excavated from deposits at Tunnel Cave (115° 02' E, 34° 05' S) and Devil's Lair (115° 04' E, 30° 09' S), two archaeologically and culturally significant sites

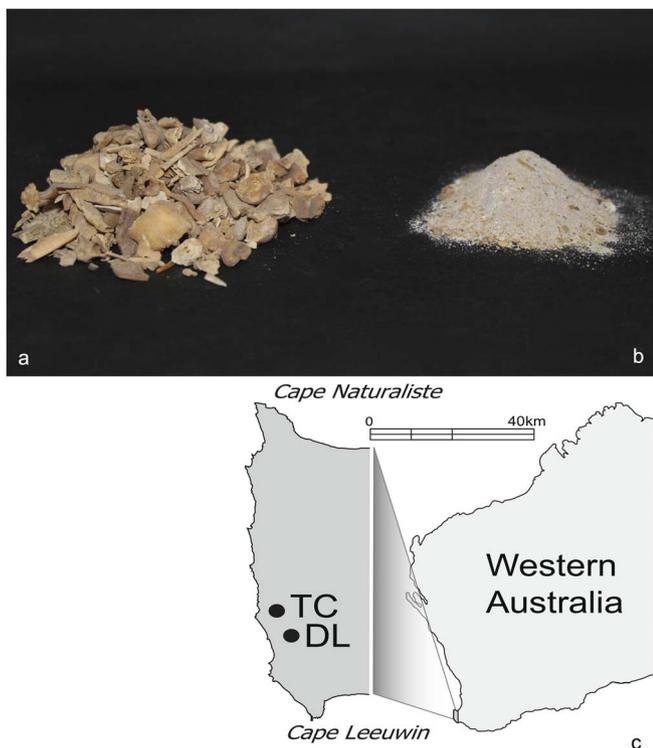


Figure 1 | Bulk-bone fragments ground to form a bulk-bone powder at two archaeological sites. Morphologically indistinct bulk-bone fragments (a) were ground to form single bulk-bone powder samples (b). Bulk-bone fragments were excavated from Devil's Lair (DL) and Tunnel Cave (TC), two archaeologically significant sites in southwest Western Australia (c). The map used in (c) was sourced from www.openclipart.org and was modified by J.H in Adobe Illustrator.

in southwestern Australia (Figure 1c). Taken together, these sites, used to explore this methodological approach, span the last c.50,000 years²³ and provide an unparalleled opportunity to study past Australian biodiversity and Aboriginal occupation²³ located within an internationally recognised biodiversity “hotspot”²⁴. A new method for the bulk sampling of fragmented bone material that would otherwise remain an untapped taxonomic resource is presented. By grinding multiple bones (Figure 1a) into an artificial “bulk-bone powder” (Figure 1b), thus producing a single bulk-bone powder sample, a large amount of highly informative genetic data can be quickly extracted. Such an approach should become commonplace in archaeological and palaeontological practice as it enables rapid assessment of DNA preservation and effectively maps zooarchaeological and palaeontological assemblages without destructive sampling of more valuable fossils.

Results

Overview of data generated. In a 2012 excavation, thousands of small bone fragments were collected by dry-sieving sediment from 15 well-dated stratigraphic units or layers at Devil's Lair and Tunnel Cave (Figure 1c). Around 50–150 bone fragments from within each layer were each drilled for 10–15s to form 15 bulk-bone powder samples representing the 15 layers (Figures 1a, 1b). DNA was extracted from each bulk-bone powder sample using established extraction methods (described in Methods) as if the bulk-bone sample were a single-source sample. The DNA extracts were screened for amplifiable mitochondrial DNA (mtDNA) using generic primers (tagged with HTS adaptors and unique barcodes) and subsequently sequenced using two HTS platforms: the GS-Junior (Roche) and the Ion Torrent PGM (Life Technologies).

Ancient DNA was successfully extracted from all bulk-bone powder samples, including a layer dated c.44,260–46,890 years BP (uncalibrated). The successful amplification and sequencing of DNA from all 15 layers was a rapid, cheap and effective way to assess DNA preservation at the sites (Figure 1c).

Amplicon DNA sequences (hereafter referred to as sequences) obtained from collective GS-Junior and Ion Torrent PGM sequencing runs were analysed for quality and possible chimeras. Except for ubiquitous human DNA sequences, control reactions throughout the process (described in Methods) were negative for contaminating DNA arising from laboratory processing.

Short regions within the mammalian mitochondrial 12S and 16S rRNA genes were amplified generating products of 100–104 bp and 90–96 bp respectively²⁵. Amplification and sequencing of avian mtDNA was successful for some samples, producing either a 106–121 bp or 227–239 bp region of the avian mtDNA 12S gene²⁵. Some cross-species reactivity was observed when using both 12S and 16S mammalian primer sets, resulting in the amplification and sequencing of avian and reptilian DNA. A targeted quantitative PCR (qPCR) and HTS approach to identify snake species was successful for a single sample.

Taxonomic identification. Mammalian 12S and 16S assays identified eight mammalian families representing 16 genera, using assignment filters chosen for this study (see Methods; Figure 2). The increase in sequencing depth afforded by the Ion Torrent PGM, as compared to the GS-Junior, did not increase the diversity of taxa identified. Mammalian taxa endemic to Australia were detected in multiple samples, in addition to taxa that have undergone significant range contraction and extirpation. The macropodid genus *Thylogale* (pademelon), provided the closest BLAST matches for many sequences across multiple samples, but to date no member of the genus has been recorded in this region. It was not possible to provide accurate taxonomic identifications for most of the Muridae sequences and for many *Macropus* sequences. While many sequences could be assigned with high confidence to a genus level, others could not be assigned beyond family or genus. A number of birds and reptiles were also identified and these have been collated at the family and genus level (Figure 2). While assignment to the species level is certainly possible in many instances a conservative approach is adopted here to showcase the approach.

Genetic biodiversity analysis. A largely taxonomy-independent approach was adopted to examine fluctuations in observed genetic diversity over time at both sites. While the taxa identified using the GS-Junior and Ion Torrent PGM were mostly congruent, coverage dependent OTU inflation, arising from homopolymer sequencing error (see Methods; Discussion) was observed. A modified OTU analysis filter was designed to reduce the influence of HTS homopolymer sequencing error^{26,27}, by employing distance-based metrics obtained from sequence alignments, giving rise to a new method referred to here as Distance-based Taxonomic Units (DTUs).

A total of 72 DTUs were identified across all 15 samples, 23 of which were shared across multiple samples, and in some instances both archaeological sites (Figure 3). The number of DTUs fluctuates noticeably with time (Figure 4). The number of DTUs shows a notable decrease that roughly coincides with the last glacial maximum (LGM), whilst also showing an increase post-LGM. The composition of DTUs also varies over time. For instance, Potoroidae (potoroids) DTUs appear around the LGM and show an increase in numbers, whilst numbers of Macropodidae (macropodids) DTUs show a decline post-LGM.

With obvious variation in DTU composition, macropodid sequences were selected to examine DTU number flux at a finer scale to examine whether or not this reflected the overall trends in biodiversity change. Macropodids exhibit a declining trend in DTU diversity post-LGM (Figure 5) that marginally increases near the Holocene/Pleistocene transition 11,700 years ago.

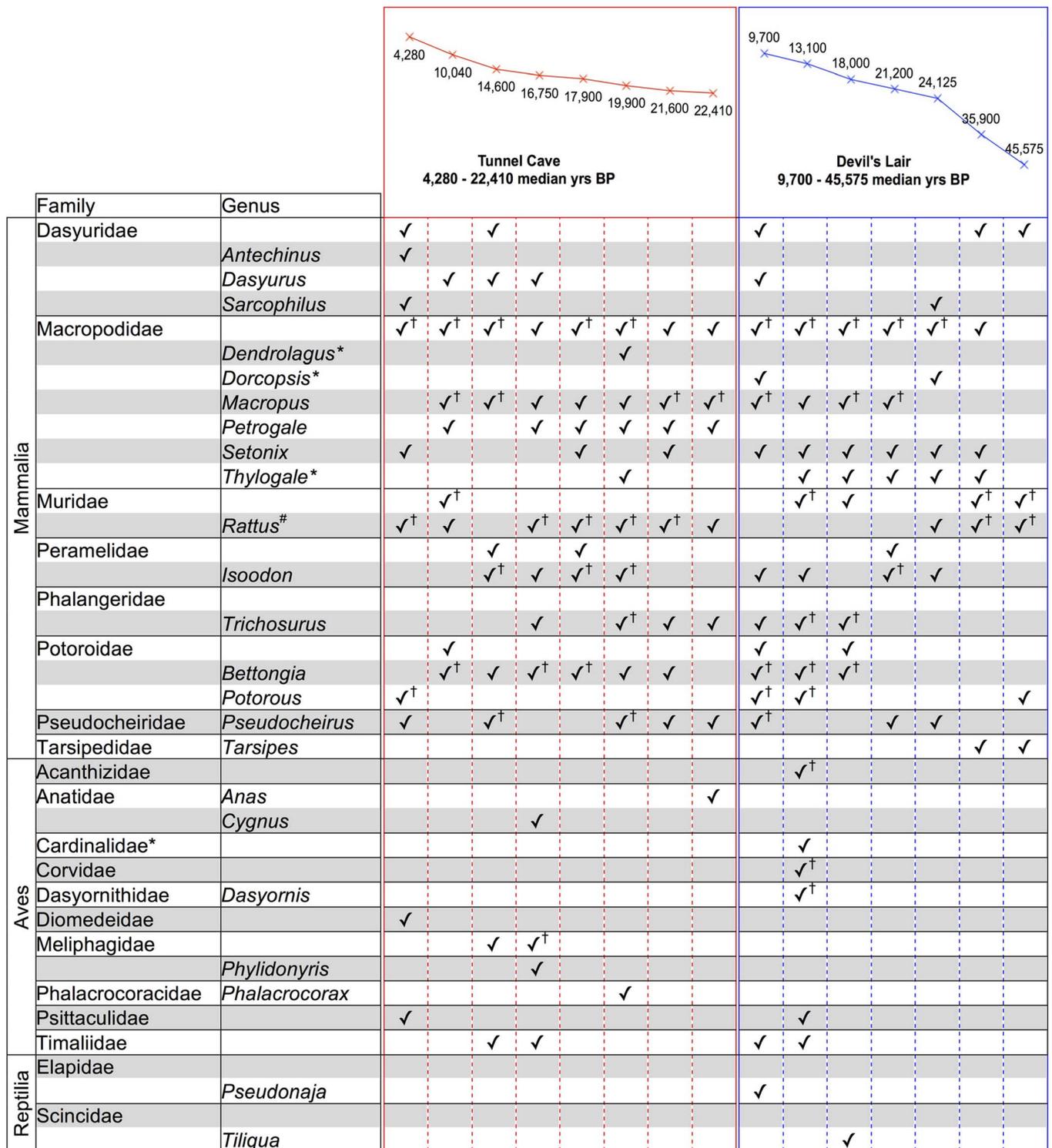


Figure 2 | Taxa identified in bulk-bone powder samples. Mammals, birds and reptiles identified in each sample are listed. Samples are grouped according to site from youngest to oldest in years BP (uncalibrated), which is plotted on the same scale for both sites. The criteria used in taxonomic assignment are detailed in the Methods. Note that there is uncertainty surrounding taxonomy with regards to both Timaliidae and Cardinalidae (See Discussion). Key: [†]Detected using multiple primer sets; *Taxa not historically known to occur in the study region; #Sequences assigned to *Rattus* aligned closest to native *Rattus Fuscipes* (bush rat).

Discussion

This study presents a novel HTS method using aDNA characterised from bulk-bone powder samples. It represents a powerful new approach to analyse unidentifiable fragments excavated from fossil deposits. Ancient DNA extracted from bones within a layer dated

between 44,260–46,890 years BP (uncalibrated), is the oldest aDNA recovered from Australia to date. These HTS results and the initial exploration of this technique show promise for larger scale bulk-bone analyses of fossil deposits. Rapidly analysing a bulk bone sample to determine if a site is conducive to DNA preservation will be

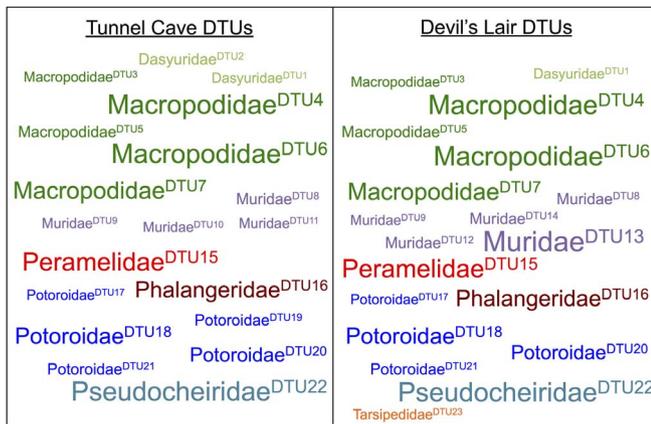


Figure 3 | DTUs shared across bulk-bone powder samples. The DTUs shared between bulk-bone powder samples, and across both Tunnel Cave (left) and Devil's Lair (right), are shown. DTUs have been labeled with the closest BLAST family matches. Each DTU has been assigned a numeric identifier following the acronym 'DTU', shown in superscript. Font size is indicative of the total number of samples a DTU was detected in.

valuable in excavations and test pits as DNA becomes increasingly incorporated into archaeological and palaeontological practices.

Even with the limited sampling, this first foray into bulk-bone analyses, has uncovered a significant amount of biological

information that adds substantially to previous knowledge of the sites and surrounding biodiversity. Analysing these data in the context of DNA damage, sequencing error, incomplete reference databases and the necessary use of short DNA sequences raises numerous challenges that must be systematically addressed^{17,28–30}. Nevertheless, when appropriate protocols and sequence filters are applied (see Methods) the method affords new insights into past biodiversity (Figure 2) and its temporal and spatial variation (Figures 3, 4 and 5).

Raw DNA sequences obtained from HTS platforms can be sorted and screened using a combination of filters that collectively exclude low-quality reads (Q-scores), sequences with errors in known flanking regions (adaptors, primers, and barcodes), artificial chimeric sequences and low abundance reads (see Methods). However, even sequences that pass these filters need to be interpreted with caution: the bird family Cardinalidae, which is not known to occur in Australia, is a case in point. The identification of birds also serves to illustrate the pitfalls associated with taxonomic revision. The taxonomy of the family Cardinalidae has been revised on a number of occasions, as has that of Timaliidae, which was also identified in some samples. Timaliidae has been regarded as a family consisting of Old World passerine birds, however the Australasian babbler (family: Pomatostomidae) were once within this family and the typical white-eyes (*Zosterops*) are disputably within this family also³¹. The families and genera identified (Figure 2) within each of the 15 samples require further investigation to identify taxa to the species level. Nevertheless, most of the genera identified at both sites from fossil morphology were again successfully detected in

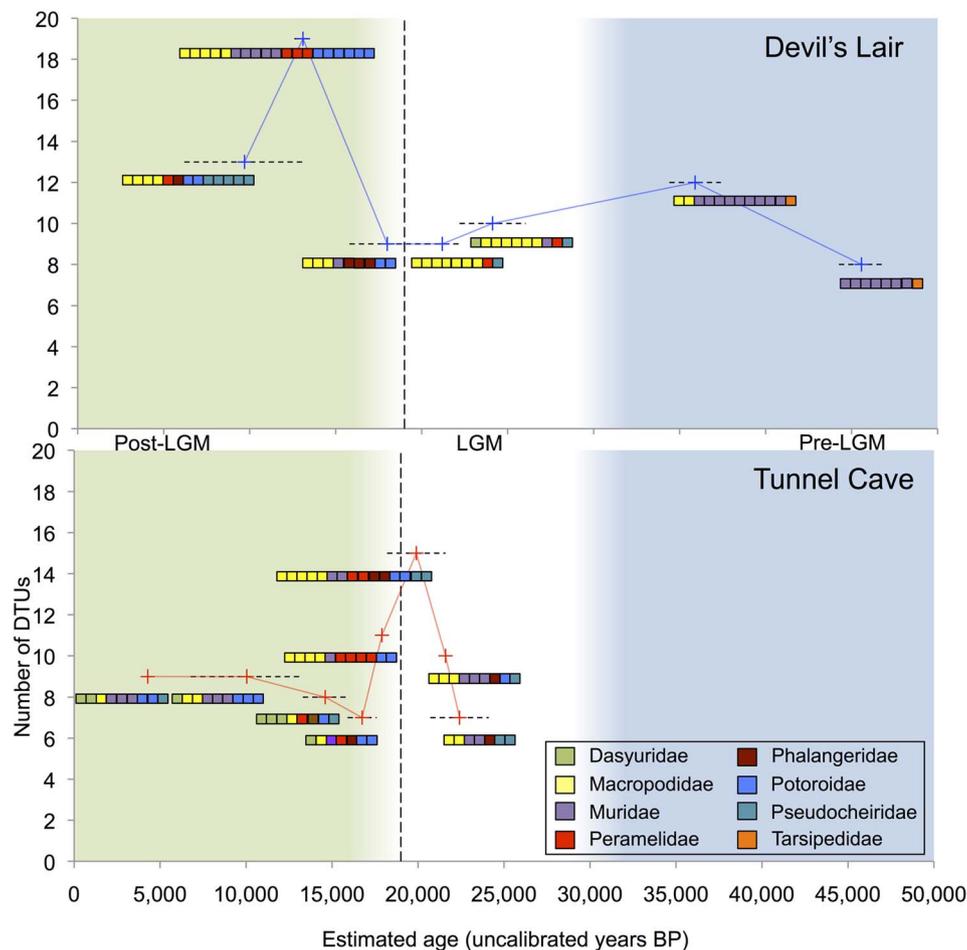


Figure 4 | Change in DTU number and composition over time at Tunnel Cave and Devil's Lair. The fluctuation in DTU number and the change in DTU composition across samples and at both sites are plotted against the backdrop of the major climatic shift around the end of the Last Glacial Maximum (LGM). Dashed vertical line - approximate end of the LGM; Blue background - Pre-LGM; White background - LGM; Green background - Post-LGM. Median ages are plotted for each sample; dashed horizontal line indicates minimum and maximum accepted date range for each layer.

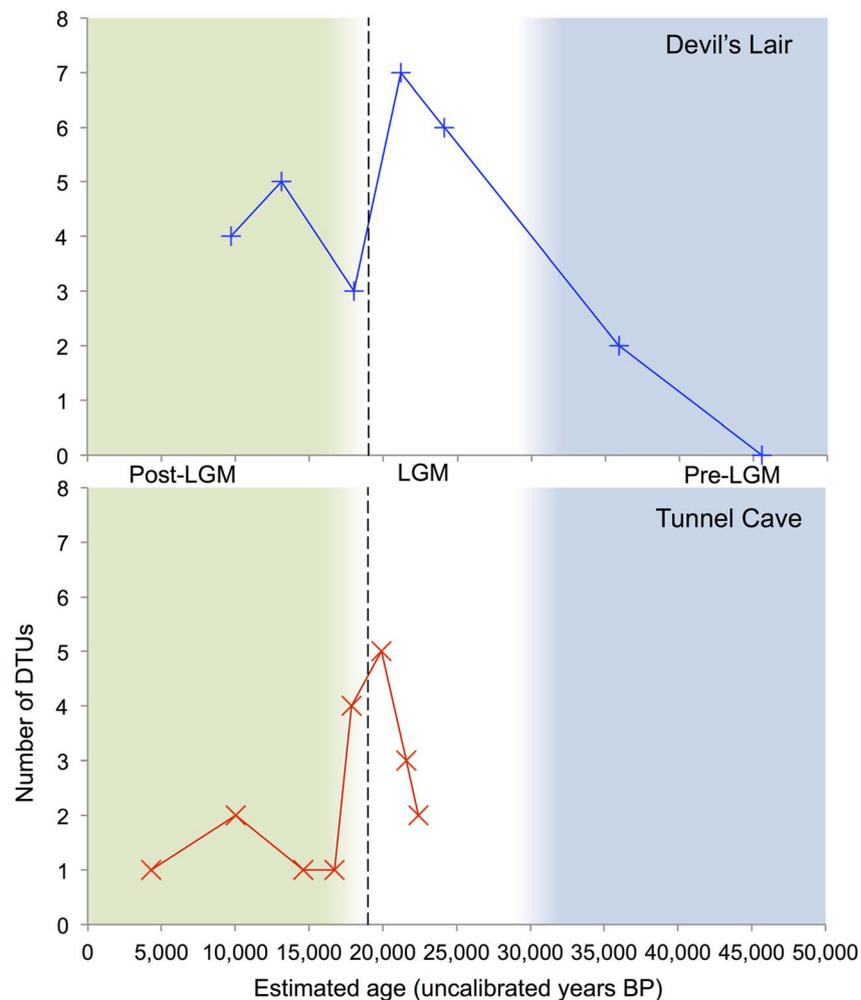


Figure 5 | Change in Macropodidae DTU number over time at Tunnel Cave and Devil's Lair. The fluctuation in Macropodidae DTU number across samples and at both sites is illustrated. Dashed vertical line - approximate end of the LGM; Blue background - Pre-LGM; White background - LGM; Green background - Post-LGM. Median ages are plotted for each sample.

the bulk-bone²³. The absence of some morphologically identified taxa from the genetically-determined faunal assemblage list is most likely due to sampling bias, as the present analysis derives from deposits representing less than one percent of the volume of the original excavations. Additionally, the possibility of primer binding bias contributing to the discontinuities between both aDNA and fossil assemblage datasets cannot be excluded. *In silico* analysis of variation in binding sites and the use of the multiple markers attempts to identify and minimize the impact of amplification bias. Finally, inherent differences between bones in terms of the preservation and quantum of mtDNA per unit biomass may also skew results between both methods of analysis causing artificial over-representation of some taxa relative to others. However, taxa were also identified that were not detected in any previous morphology-based analyses, particularly small mammals, birds and reptiles, all of which require highly-specialised taxonomic skills to identify, are less likely to preserve diagnostic remains, and may be poorly represented in reference collections.

A high level of confidence surrounds the bulk of the taxonomic identifications; for instance, the majority of mammalian taxa identified are locally extant or known from the fossil record. The same generally holds true for avian and reptilian taxa identifications. The detection of sequences endemic to southwest Australia, such as a 100% match to *Tarsipes rostratus* (honey possum), further supports the *bona fide* nature of the sequences obtained. Moreover, the detec-

tion of extirpated taxa, such as *Setonix* (quokka) and *Sarcophilus* (Tasmanian devil), as far back as c.24,000 years BP (uncalibrated) illustrates the antiquity and authenticity of the sequences, as does the detection of species whose ranges have contracted and are no longer documented at the sites, e.g. *Bettongia* (bettongs). There appears to be little or no environmental contamination as evidenced by the absence of any sequences from highly abundant invasive taxa including *Mus musculus* (house mouse) or *Rattus rattus* (black rat). Whereas downward contamination may be an issue at some sites³², Devil's Lair contains several stratigraphical layers capped with calcite "flowstone"^{23,33} preventing the movement of fossils, and likely DNA^{5,23}. Whilst it is acknowledged that contamination can be cryptic and sporadic³⁴⁻³⁶, the strict adherence to aDNA protocols³⁷, the use of sequence quality filters and the plausibility of the data (see Methods), greatly reduces the likelihood that contamination contributed to the data presented here.

Although most taxonomic assignments from DNA sequences confirmed previous morphological identification²³, some unexpected sequences resulted in distinct DTUs that were more difficult to assign. The issue is best exemplified by indeterminate Macropodidae sequences. It is unlikely that poor database coverage is the cause of this family-level assignment, as the Macropodidae database is nearly complete for both 16S and 12S rRNA mtDNA. In such cases sequencing error or DNA damage is also unlikely as the sequences are abundant and present across numerous samples at both sites, have passed all

quality filters, form distinct DTUs and are unlikely to be nuclear copies (Figures 2, 3 and 4). It is possible therefore that these sequences may arise from extinct lineages of present-day macropodids or indeed from extinct taxa. In some cases sequences mapped closest to species of the New Guinea forest wallaby (*Dorcopsis*) and the east Australian restricted pademelon (*Thylogale*). The presence of such ‘indeterminate’ DNA sequences in bulk-bone samples is intriguing. For example, two extinct tree-kangaroo species (genus *Bohra*^{38,39}), have been described in caves along the Nullarbor Plain, yet tree-kangaroos of the genus *Dendrolagus* are only currently present in northeastern Queensland and New Guinea and were previously not thought to have occurred so far south³⁸. It is a tantalizing prospect that ‘indeterminate’ DNA sequences could represent previously unknown species from southwest Western Australia, but it is also a problematic finding, as there is no easy way to uncover the fossils that contributed the DNA. It is likely that bulk-sampling methods such as this will generate genetically plausible taxa that lack morphological identifications. Arguably a similar result has already occurred with the single Denisovan finger bone from “X-woman” used to postulate a new lineage of archaic humans in Siberia^{40,41}.

When dealing with past biodiversity and aDNA sequences from fossil assemblages, analyses that are largely independent of taxonomy will likely be crucial to mapping temporal and/or spatial variation in genetic signatures. Such an approach facilitates the use of sequences that would otherwise be labeled “indeterminate”, which will be commonly encountered when employing the bulk-bone HTS methodologies advocated here. While it is not possible to comprehensively analyse changes in biodiversity over time presented here from only a handful of samples such an analysis serves to illustrate how bulk-bone data could be approached. The data presented in Figures 3–5 should therefore be viewed tentatively, as further extensive replication and investigation is required to confirm any significant patterning over time.

Owing to the difficulties of definitively assigning sequences to a defined taxonomy, a modified OTU analysis (referred to as DTU), has been introduced to examine biodiversity change over time. It was clear from the initial analysis that OTU numbers were artificially inflated primarily by homopolymer error. When dealing with short sequences homopolymer errors can create a distinct OTU whereby the only difference between it and its closest OTU match is a base within a homopolymer stretch. It was observed that homopolymer-derived OTUs were more common in those samples with greater depth of sequencing coverage. To overcome this issue, an OTU alignment and Kimura 2-parameter distance matrix was adopted whereby errors in homopolymer stretches appear as gaps and homopolymer-derived OTUs collapse into a single DTU (See Methods). Whilst at these particular sites, it is a challenge to disentangle the roles of climate, DNA decay and past anthropogenic influences; shifts in DTU composition appear at the LGM and at the Holocene-Pleistocene transition (Figures 4 and 5). Furthermore, specific Macropodidae DTU analysis showed a reduction in DTU diversity and abundance over time, with a drop in diversity around the LGM (Figure 5). With these tentative patterns of biodiversity being derived from only 15 DNA extractions it is easy to conceptualize how, with adequate sampling and appropriate genetic markers, a bulk-bone sampling method will facilitate detailed mapping of faunal changes over time. Moreover, the method is cheaper than single bone approaches^{42,43} while augmenting traditional morphological analysis.

The bulk-bone aDNA metabarcoding method used in this study presents a new, cost effective approach to identifying bulk quantities of morphologically indistinct bone fragments that otherwise end up in the taxonomic scrapheap. From modest amounts of sieved material across multiple layers at two study sites it was possible to detect equivalent diversity as described in previous morphological analyses²³. While some taxa previously identified were not detected (most noticeably *Macropus* species), the converse was also true.

This method is by no means an attempt to supplant traditional morphological approaches to taxonomic identification and analysis. Rather, it complements these approaches and by means of DTU analysis indicates changes in genetic diversity through time. Besides improving the identification of fossil assemblages the method allows researchers to rapidly assess the DNA preservation potential of freshly excavated material, which will vary from site to site. The approach will be equally applicable to archaeological and palaeontological sites, providing snapshots of past faunal diversity and human subsistence in both taxonomic dependent and independent ways. As such, it is anticipated that a bulk-bone approach will become a valuable part of the archaeological and palaeontological toolkit.

Methods

Sample collection and processing. Thousands of indistinct bone fragments were collected from both Tunnel Cave and Devil’s Lair during excavations in February 2012. Approximately 150 L (0.15 m³) of sediment was analysed at both sites. Sediment was dry-sieved on site, using 2 mm and 5 mm sieves, and bagged according to well-defined and dated stratigraphical layers²³. Each bagged sample was screened for bone fragments off-site, which were kept in groupings according to the layers in which they were found. Fifteen bulk-bone samples representing fifteen layers were processed: eight from Tunnel Cave, covering a period from 4,160–24,110 years BP (uncalibrated)²³, and seven from Devil’s Lair, covering a period from 6,200–46,890 years BP (uncalibrated)²³. Small sections of the bones within each layer (typically 50–150 bones) were drilled (Dremel 114 drill bits) for a few seconds each and approximately equal amounts of drilled material from each bone fragment within a single layer was combined to form a “bulk-bone powder”. Owing to inherent differences in the amount of DNA per unit of biomass between species and differential DNA preservation between individual bones, over-representation of certain bone material in terms of DNA amplicon sequences is unavoidable.

DNA extraction and screening. All laboratory work was conducted in keeping with standard aDNA protocols²⁸. Approximately 1 g of bulk-bone powder from each sample, including a blank extraction control, was digested overnight on a lab rotator at 55 °C in 5 mL of digestion buffer containing: 2.5 mL EDTA (0.5 M), 0.1 mL Tris-HCL (1 M), 5 mg Proteinase K powder, 50 µL DTT (1 M), 50 µL SDS and made up to a final volume of 5 mL using EDTA. DNA digests were centrifuged at 6,000 rpm for 2 mins and the supernatant was concentrated to 50 µL using AMICON 30,000 MWCO columns (Millipore) as per the manufacturer’s instructions. Each concentrate was transferred to a clean 2 mL eppendorf tube and PBI buffer (Qiagen) totalling 250 µL (i.e. 5× the volume of concentrate) was added. Each 300 µL PBI/concentrate mix was subsequently transferred to Qiagen silica spin columns and centrifuged at 13,000 rpm. Columns were washed with 700 µL of AW1 followed by AW2. A final dry spin at 13,000 rpm for 1 min followed. DNA was eluted from the columns in 60 µL EB with a 1 min incubation at room temperature prior to centrifugation at 13,000 rpm for 1 min.

Extracts were screened for amplifiable mtDNA using multiple primer sets via qPCR at three concentrations - undiluted, 1/10 and 1/50. Extracts were screened for mammalian mtDNA using 12SA/O and 16Smam primer sets, designed to amplify a small region within mammalian 12S and 16S mitochondrial genes respectively^{25,44}. Extracts were also tested for avian mtDNA using 12SA/E and 12SA/H primer sets, designed to amplify a short and slightly longer overlapping region of the avian mitochondrial 12S gene respectively⁴⁴. Finally, extracts were tested for snake mtDNA using the following primers: 12s_tRNA_F1_S AAAGTATAGCACTGAAAATGC TAA and 12s_R1_Snake GTTAGCCTGATACCGGCTCCG, designed to amplify a short region within the mitochondrial 12S gene. Each qPCR reaction was made up to a total volume of 25 µL, containing 1× PCR Gold Buffer (Applied Biosystems), 2.5 mM MgCl₂ (Applied Biosystems), 0.4 mg/mL BSA (Fisher Biotech, Aus), 0.25 mM of each dNTP (Astral Scientific, Aus), 0.4 µM forward primer, 0.4 µM reverse primer, 0.25 µL AmpliTaq Gold (Applied Biosystems), 0.6 µL SYBR Green (1:2,000, Life Sciences gel stain solution) and 2 µL DNA extract. Quantitative PCR cycling conditions for the 12SA/O and snake 12S qPCR assays were as follows: initial heat denaturation at 95 °C for 5 mins, followed by 50 cycles of 95 °C for 30 s; 55 °C for 30 s (annealing step); 72 °C for 45 s followed by a 1 °C melt curve and final extension at 72 °C for 10 mins. Cycling conditions for 16Smam, 12SA/E and 12SA/H assays were the same as for the 12SA/O assay, except the annealing temperature, which was 57 °C in each case. For each qPCR assay, DNA extraction, negative PCR reagent and positive DNA template controls were included.

DNA sequencing. DNA extracts that successfully yielded DNA of sufficient quality, free of inhibition, as determined by initial qPCR screening⁴⁵, were prepared for amplicon sequencing. DNA extracts successful for all primer sets were sequenced on Roche’s GS-Junior. Additional, separate, amplicon sequences were generated for extracts using mammalian 12SA/O and 16Smam primer sets for sequencing on Life Technologies’ Ion Torrent Personal Genome Machine (PGM).

For each primer set, DNA extracts were assigned a unique DNA tag⁴¹. Each sample was tagged at both the 5’ and 3’ end of the target sequence using separate tags at both

ends, resulting in a unique forward and reverse tag combination for each sequence. Independent tagged qPCRs for all samples, across all primer sets, were carried out in 25 μ L reactions with reaction components and cycling conditions as described in 'Methods: DNA extraction and screening'. Tagged qPCR amplicons were generated in triplicate and combined, thus minimizing the effects of PCR stochasticity on low-template samples, purified using Agencourt AMPure XP PCR Purification Kit (Beckman Coulter Genomics, NSW, Aus), as per manufacturer's instructions and eluted in 40 μ L H₂O. Purified amplicons were pooled to form separate sequencing libraries according to primer set used and sequencing platform. GS-Junior libraries were quantified using qPCR to determine an appropriate volume of library for sequencing (described in Murray *et al.* 2011). Each 25 μ L reaction contained 12.5 μ L ABI Power SYBR master mix (Applied Biosystems), 0.4 μ M A-adapter primer, 0.4 μ M B-adapter primer, 8.5 μ L H₂O and 2 μ L pooled library, with the following cycling conditions: 95°C for 5 mins; 40 cycles of 95°C for 15 s, 56°C for 1 min followed by a 1°C melt curve. The appropriate library volume for use on the Ion Torrent PGM was determined using a Bioanalyser 2100 (Agilent). For each tagged qPCR assay, negative qPCR controls were included and if found to contain amplifiable DNA these qPCR amplicons were incorporated into the appropriate pooled sequencing library. All sequencing was performed as per manufacturer's instructions, with the use of 200 bp reagents and a 314 chip on the PGM.

Sequence identification. Amplicon sequence reads (hereafter referred to as sequences) were sorted into sample batches based on unique DNA tags. Identification tags and primers were trimmed allowing for no mismatch in length or base composition using Geneious v6.0.5 (created by Biomatters, available from <http://www.geneious.com/>). Batched and trimmed sequences from both GS-Junior and Ion Torrent PGM sequencing runs were combined according to sample and primer used. Each combined file was dereplicated, thus grouping sequences of exact identity and length, using USEARCH⁴⁶. Dereplicated sequence files were searched for artificial chimeric sequences using the UCHIME *de novo* method⁴⁷ in USEARCH and were removed, in addition to sequences occurring only once (i.e. singletons). The remaining sequences in each sample were subsequently clustered at an identity threshold of 97% using USEARCH with the most abundant sequence within each cluster selected as the representative sequence. To reduce noise associated with sequencing error, low abundant clusters, classed as those that occur at less than 1% of the total number of unique sequences when clustered at 100% sequence identity, were removed from the dataset. While the selection of a 1% cut-off is somewhat arbitrary, it should negate the possibility of clusters remaining that are the result of sequencing error. Additionally, the decision to class clusters as being in low abundance with respect to the total number of unique sequences (as opposed to total number of sequences or total number of sequences within the most abundant cluster) was made to minimize the effects of preferential DNA preservation and/or amplification. For each sample, every sequence assigned to the remaining clusters were queried against the NCBI GenBank nucleotide database using BLASTn⁴⁸ in YABI⁴⁹, enabling taxonomic identification. Sequences were searched without a low complexity filter, with a gap penalties existence of five and extension of two, expected alignment value less than 1e-10 and a word count of seven. The BLASTn results obtained were imported into MEtaGenome Analyzer v4 (MEGAN), where they were mapped and visualised against the NCBI taxonomic framework (min. bit score = 35.0, top percentage = 5%, min. support = 1)⁵⁰. Sequences that were obviously the result of contamination (primarily human and cow) were eliminated from all subsequent downstream analysis steps.

Sequences that were truncated when queried against the NCBI GenBank nucleotide database were discarded from taxonomic analysis. Sequences with percentage similarity to a reference below 90% were discarded. Where sequence similarities were between 90–95% these were assigned to a family level, while those between 95–100% were assigned to a genus. Owing to the difficulties in assigning taxa beyond the genus level for some families, in addition to issues associated with characterizing past biodiversity that has been lost, species identifications were avoided in this particular study. Sequences that provided high percentage similarity to query references at a species level may or may not be bona fide, however with current insufficient data it is prudent to categorise these sequences cautiously. Where multiple taxa had equal percentage similarity scores to a query sequence, such sequences were moved higher up the taxonomic rankings.

While the validity of filters and hard percentage cut-offs are always debatable, those chosen in the analysis of this dataset seemed to afford the best balance when accounting for low template amounts and post-mortem damage on short aDNA fragments.

Genetic biodiversity analysis. Cognisant of the difficulties associated with assigning sequences to lower taxonomic levels, a modified form of OTU analysis was applied to the 16S_{mm} sequences obtained in this study. This allowed changes in observed genetic diversity over time at both sites to be investigated independently of the above taxonomic classifications. Sequences within each sample were clustered at 97% identity, filtered and representative sequences were selected as detailed in Methods: Sequence Identification. Representative sequences within each sample were aligned in Geneious using MAFFT's G-INS-I algorithm and default parameters⁵¹. MAFFT alignments were imported into MEGA5⁵² where a distance matrix between OTUs within a sample was calculated using a Kimura 2-parameter model⁵³, with all positions containing gaps and missing data ignored. OTUs less than 3% divergent from each other were collapsed into a single DTU. This serves the purpose of reducing the influence of HTS homopolymer sequencing error^{26,27} by collapsing multiple

homopolymer-derived OTUs into a single DTU, as errors in homopolymer stretches appear as gaps and are not included in the calculation of the distance matrix. Whilst this is first and foremost a largely taxonomic-independent analysis it is still nonetheless useful to identify coarsely to which family each DTU belongs, as this gives an idea of the diversity of DTUs within specific families. As such, all DTUs were searched against the NCBI GenBank nucleotide database using BLASTn⁴⁸ to identify the family to which each DTU could be easily assigned. For the faunal specific Macropodidae DTU analysis the same method as above was followed except that only sequences assigned to Macropodidae were selected.

1. Archibald, S. B., Greenwood, D. R. & Mathewes, R. W. Seasonality, montane beta diversity, and Eocene insects: Testing Janzen's dispersal hypothesis in an equable world. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **371**, 1–8 (2013).
2. Coloneasa, A. C. *et al.* Holocene snail shell isotopic record of millennial-scale hydrological conditions in western Mediterranean: Data from Bauma del Serrat del Pont (NE Iberian Peninsula). *Quat. Int.* **303**, 43–53 (2013).
3. Dortch, J. & Wright, R. Identifying palaeo-environments and changes in Aboriginal subsistence from dual-patterned faunal assemblages, south-western Australia. *J. Archaeol. Sci.* **37**, 1053–1064 (2010).
4. Rohland, N. *et al.* Genomic DNA Sequences from Mastodon and Woolly Mammoth Reveal Deep Speciation of Forest and Savanna Elephants. *PLoS Biol.* **8**, e1000564 (2010).
5. Haile, J. *et al.* Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22352–22357 (2009).
6. Worobey, M. *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature.* **455**, 661–664 (2008).
7. Raoult, D. *et al.* Molecular identification by "suicide PCR" of *Yersinia pestis* as the agent of Medieval Black Death. *Proc. Natl. Acad. Sci. USA.* **97**, 12800–12803 (2000).
8. Smith, C. I. *et al.* Neanderthal DNA. Not just old but old and cold? *Nature.* **410**, 771–772 (2001).
9. Bonnicksen, R. *et al.* Methods of the study of ancient hair: Radiocarbon dates and gene sequences from individual hairs. *J. Archaeol. Sci.* **28**, 775–785 (2001).
10. Oskam, C. L. *et al.* Fossil avian eggshell preserves ancient DNA. *Proc. R. Soc. Biol. Sci. Ser. B.* **277**, 1991–2000 (2010).
11. Binladen, J. *et al.* The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE.* **2**, e197 (2007).
12. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21**, 1794–1805 (2012).
13. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).
14. Jørgensen, T. *et al.* A comparative study of ancient sedimentary DNA, pollen and macrofossils from permafrost sediments of northern Siberia reveals long-term vegetational stability. *Mol. Ecol.* **21**, 1989–2003 (2012).
15. Jørgensen, T. *et al.* Islands in the ice: detecting past vegetation on Greenlandic nunataks using historical records and sedimentary ancient DNA Meta-barcoding. *Mol. Ecol.* **21**, 1980–1988 (2011).
16. Deagle, B., Chiaradia, A., McInnes, J. & Jarman, S. Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conserv. Genet.* **11**, 2039–2048 (2010).
17. Murray, D. C. *et al.* High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quat. Sci. Rev.* **58**, 135–145 (2012).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
19. Little, D. P. DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS ONE.* **6**, e20552 (2011).
20. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods.* **7**, 335–336 (2010).
21. Schloss, P. D. & Handelsman, J. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.* **71**, 1501–1506 (2005).
22. Hamady, M., Lozupone, C. & Knight, R. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal.* **4**, 17–27 (2010).
23. Dortch, J. *Palaeo-environmental Change and the Persistence of Human Occupation in South-western Australian Forests.* (Archaeopress, Oxford, 2004).
24. Myers, N., Mittermeier, R. A., Mittermeier, C. G., de Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature.* **403**, 853–858 (2000).
25. Taylor, P. G. Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Mol. Biol. Evol.* **13**, 283–285 (1996).
26. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
27. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* **13**, 341 (2012).
28. Cooper, A. & Poinar, H. N. Ancient DNA: Do it right or not at all. *Science.* **289**, 1139–1139 (2000).



29. Coissac, E., Riaz, T. & Puillandre, N. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Mol. Ecol.* **21**, 1834–1847 (2012).
30. Taylor, H. R. & Harris, W. E. An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol. Ecol. Resour.* **12**, 377–388 (2012).
31. Jönsson, K. A. & Fjeldså, J. A phylogenetic supertree of oscine passerine birds (Aves: Passeri). *Zoologica Scripta.* **35**, 149–186 (2006).
32. Haile, J. *et al.* Ancient DNA chronology within sediment deposits: Are paleobiological reconstructions possible and is DNA leaching a factor? *Mol. Biol. Evol.* **24**, 982–989 (2007).
33. Turney, C. & Bird, M. I. Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. *Quatern. Res.* **55**, 3–13 (2001).
34. Champlot, S. *et al.* An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. *PLoS ONE.* **5**, e13042 (2010).
35. Erlwein, O. *et al.* DNA Extraction Columns Contaminated with Murine Sequences. *PLoS ONE.* **6**, e23484 (2011).
36. Tuke, P. W., Tettmar, K. I., Tamuri, A., Stoye, J. P. & Tedder, R. S. PCR Master Mixes Harbour Murine DNA Sequences. Caveat Emptor! *PLoS ONE.* **6**, e19953 (2011).
37. Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends. Ecol. Evol.* **20**, 541–544 (2005).
38. Prideaux, G. J. & Warburton, N. M. A new Pleistocene tree-kangaroo (Diprotodontia: Macropodidae) from the Nullarbor Plain of south-central Australia. *J. Vert. Paleontol.* **28**, 463–478 (2008).
39. Prideaux, G. J. & Warburton, N. *Bohra nullarbor* sp. nov., a second tree-kangaroo (Marsupialia: Macropodidae) from the Pleistocene of the Nullarbor Plain, Western Australia. *Rec. West. Aust. Mus.* **25**, 165–179 (2009).
40. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature.* **464**, 894–897 (2010).
41. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science.* **338**, 222–226 (2012).
42. Shapiro, B. *et al.* Rise and fall of the Beringian steppe bison. *Science.* **306**, 1561–1565 (2004).
43. Lorenzen, E. D. *et al.* Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature.* **479**, 359–364 (2011).
44. Cooper, A. *et al.* Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature.* **409**, 704–707 (2001).
45. Bunce, M., Oskam, C. & Allentoft, M. in *Ancient DNA: Methods and Protocols Methods in Molecular Biology* (eds Shapiro, B. & Hofreiter, M.) Ch. **16**, 121–132 (Humana Press, New York City, 2011).
46. Edgar, R. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* **26**, 2460–2461 (2010).
47. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics.* **27**, 2194–2200 (2011).
48. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **34**, D16–D20 (2006).
49. Hunter, A. A., Macgregor, A. B., Szabo, T. O., Wellington, C. A. & Bellgard, M. I. Yabi: An online research environment for grid, high performance and cloud computing. *Source Code Biol. Med.* **7**, 1 (2012).
50. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377–386 (2007).
51. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
52. Tamura, K. *et al.* MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011).
53. Kimura, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111–120 (1980).

Acknowledgments

We thank the Wardandi people, and the Webb family in particular, as traditional owners and custodians of the region, for supporting our excavations and analyses. We thank student volunteers for their assistance in the field, Nina Kresoje (UWA) and Vanessa Atkinson (Pathwest Laboratory Medicine WA) and Frances Brigg for sequencing assistance. We also thank the WA Museum and the Curator of Anthropology Moya Smith for access to museum resources, and iVEC for computational support. Australian Research Council grants DP120103725 (to M.B., J.D. and J.H.) and FT0991741 (to M.B.) funded the research.

Author contributions

D.C.M., M.B. and J.H. designed the experiments. D.C.M., J.H., N.W., D.H. and J.D. excavated and prepared samples. D.C.M., J.H., M.I.B., D.H. and R.A. contributed to HTS data generation and bioinformatics. J.D. provided stratigraphic interpretations and G.P. and J.D. provided fossil and taxon interpretations. D.C.M. and M.B. wrote the paper with assistance from all co-authors.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Murray, D.C. *et al.* Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Sci. Rep.* **3**, 3371; DOI:10.1038/srep03371 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0>

RESEARCH

Open Access

Metagenomic analyses of bacteria on human hairs: a qualitative assessment for applications in forensic science

Silvana R Tridico^{1,2*}, Dáithí C Murray^{1,2}, Jayne Addison¹, Kenneth P Kirkbride³ and Michael Bunce^{1,2}

Abstract

Background: Mammalian hairs are one of the most ubiquitous types of trace evidence collected in the course of forensic investigations. However, hairs that are naturally shed or that lack roots are problematic substrates for DNA profiling; these hair types often contain insufficient nuclear DNA to yield short tandem repeat (STR) profiles. Whilst there have been a number of initial investigations evaluating the value of metagenomics analyses for forensic applications (e.g. examination of computer keyboards), there have been no metagenomic evaluations of human hairs—a substrate commonly encountered during forensic practice. This present study attempts to address this forensic capability gap, by conducting a qualitative assessment into the applicability of metagenomic analyses of human scalp and pubic hair.

Results: Forty-two DNA extracts obtained from human scalp and pubic hairs generated a total of 79,766 reads, yielding 39,814 reads post control and abundance filtering. The results revealed the presence of unique combinations of microbial taxa that can enable discrimination between individuals and signature taxa indigenous to female pubic hairs. Microbial data from a single co-habiting couple added an extra dimension to the study by suggesting that metagenomic analyses might be of evidentiary value in sexual assault cases when other associative evidence is not present.

Conclusions: Of all the data generated in this study, the next-generation sequencing (NGS) data generated from pubic hair held the most potential for forensic applications. Metagenomic analyses of human hairs may provide independent data to augment other forensic results and possibly provide association between victims of sexual assault and offender when other associative evidence is absent. Based on results garnered in the present study, we believe that with further development, bacterial profiling of hair will become a valuable addition to the forensic toolkit.

Keywords: Forensic, Metagenomics, Bacteria, Scalp hairs, Pubic hairs, Sexual assaults, Next-generation sequencing, 16S DNA

Background

Over the last decade, the development of bacterial culture-independent approaches (metagenomics), based on 16S rRNA genes (hereafter referred to as 16S), sequences has become the cornerstone of microbial ecology [1]. The advent of next-generation sequencing (NGS) technologies and platforms capable of generating millions

of sequences per sample facilitated assessments of microbial communities between body sites and individuals [2,3]. The increased sequencing power stimulated the development of robust computational programmes capable of processing large, complex sequencing data sets [4] and enabled phylogenetic analyses of human and environmental genomes [5,6].

Studies on the human microbiome (the collective genomes present in the human body) suggest that there are significant differences in bacterial composition not only between different body sites but also between individuals

* Correspondence: silvanatrindico@yahoo.com

¹Veterinary and Life Sciences, Murdoch University, Perth, WA 6150, Australia

²Trace and Environmental DNA laboratory, Department of Environment and Agriculture, Curtin University, Perth, WA 6845, Australia

Full list of author information is available at the end of the article

RESEARCH ARTICLE

From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows

Dáithí C. Murray, Megan L. Coghlan, Michael Bunce*

Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia

* michael.bunce@curtin.edu.au**OPEN ACCESS**

Citation: Murray DC, Coghlan ML, Bunce M (2015) From Benchtop to Desktop: Important Considerations when Designing Amplicon Sequencing Workflows. PLoS ONE 10(4): e0124671. doi:10.1371/journal.pone.0124671

Academic Editor: Tom Gilbert, Natural History Museum of Denmark, DENMARK

Received: December 1, 2014

Accepted: March 16, 2015

Published: April 22, 2015

Copyright: © 2015 Murray et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper, Supporting Information files or are available from Data Dryad (doi:10.5061/dryad.2qf0t).

Funding: MB was funded from the Australian Research Council grant FT0991741 and DP120103725. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Amplicon sequencing has been the method of choice in many high-throughput DNA sequencing (HTS) applications. To date there has been a heavy focus on the means by which to analyse the burgeoning amount of data afforded by HTS. In contrast, there has been a distinct lack of attention paid to considerations surrounding the importance of sample preparation and the fidelity of library generation. No amount of high-end bioinformatics can compensate for poorly prepared samples and it is therefore imperative that careful attention is given to sample preparation and library generation within workflows, especially those involving multiple PCR steps. This paper redresses this imbalance by focusing on aspects pertaining to the benchtop within typical amplicon workflows: sample screening, the target region, and library generation. Empirical data is provided to illustrate the scope of the problem. Lastly, the impact of various data analysis parameters is also investigated in the context of how the data was initially generated. It is hoped this paper may serve to highlight the importance of pre-analysis workflows in achieving meaningful, future-proof data that can be analysed appropriately. As amplicon sequencing gains traction in a variety of diagnostic applications from forensics to environmental DNA (eDNA) it is paramount workflows and analytics are both fit for purpose.

Introduction

The myriad of names and acronyms associated with high-throughput DNA sequencing (HTS) is undeniably impressive and the number of applications for which the technology itself has proven useful equally matches this. To date, amplicon sequencing [1], whereby PCR products are generated, converted to libraries, pooled and then sequenced, has been the method of choice in many HTS studies. Amplicon sequencing has been used in, or proposed for, a wide range of contexts that include, amongst others, biomonitoring [2–7], diet analysis [8–13] and bacterial metagenomics [14–20]. As a result of the ease with which the technology can be applied across an array of disciplines, it can at times prove to be a minefield for scientists seeking to avail of it. This is especially true for those with limited experience in either wet-lab molecular

biology skills or computational bioinformatics. The latter of these areas has received much attention; the importance of the former is often under-appreciated.

Currently, most primary literature, reviews and opinion articles surrounding HTS tend to focus on the applications of the technology [3,5,6,9,21–24], platform evaluations [25,26] and bioinformatic approaches to data analysis [27–33]. While all three are extremely important in the generation of high fidelity data, a heavy focus on these aspects fails to address the need to pay close attention to the implementation of protocols and procedures at the bench. The data one has to work with is, and will only ever be, as good as the quality of experimental procedures implemented and no amount of high-end bioinformatics can compensate for poorly prepared samples, artefacts or contamination. It is therefore imperative that careful consideration is given to the ways in which samples are screened for sequencing, in addition to the method used to generate the amplicon sequencing library. These aspects are independent of the equally important need to carefully choose extraction methods that are optimised for the chosen substrates. While DNA isolation methods are a key consideration, this is dealt with extensively elsewhere. Instead, this paper focuses on how best to approach amplicon workflows following DNA extraction to generate robust and representative datasets for a given DNA isolation.

Through a series of simple experiments (Table 1), various aspects that should be considered when preparing to embark on the use of amplicon sequencing are highlighted, some aspects of which are equally as applicable to shotgun sequencing. These experiments focus primarily on three areas of experimental design or benchwork within the typical amplicon sequencing workflow: sample screening, the target region, and library generation. Finally, although not a focus of the paper, certain pertinent considerations in relation to data analysis that are seldom acknowledged in other literature will also be addressed. It is hoped that the following may address the distinct lack of literature in relation to sample preparation and library generation. It is advocated that closer attention is required at the bench when conducting amplicon sequencing. Ultimately, it may be appropriate to define a set of flexible guidelines, such as the MIQE guidelines used for qPCR data [34], for the reporting of amplicon data generation and analysis.

Table 1. Details for the experiments conducted.

Experiment	Purpose	Methods	Results
Experiment 1: Importance of sample screening	Illustrate the importance of quantifying samples using a dilution series to select an appropriate working dilution free of inhibition containing a sufficient quantity of input template DNA	Main: 2.2.1 (see also: Section 2.1.1. S1A Fig & S1 Table)	Section 3.1. Fig 2
Experiment 2: Assessing the amplicon target region	Explore the potential benefits to the downstream processing of high-throughput sequencing data arising from the inclusion of amplicon-specific single-source samples embedded into sequencing runs	Main: 2.2.2 (see also: S1B Fig & S1 Table)	Section 3.2. Fig 3
Experiment 3: Importance of experimental controls	Demonstrate the importance of control reactions in bacterial metagenomics and other fields using samples with a high propensity for environmental contamination	Main: 2.2.3 (see also: S1C Fig & S1 Table)	Section 3.3. S2 Table
Experiment 4: Library generation efficiency	Assess the efficiency drop-off associated with the use of fusion tagged primers of different ‘architecture’ when compared to standard non-fusion tagged template specific primers	Main: 2.2.4 (see also: Section 2.1.1. S1D and S1E Fig & S1 Table)	Section 3.4. S3 Table
Experiment 5: Analysis parameters and their impact	Highlight the difficulties in choosing appropriate quality and abundance filtering parameters when analysing complex, heterogeneous samples; the composition of which are unknown.	Main: 2.2.5 (see also: Fig 1, S1F Fig & S1 Table)	Section 3.5. Fig 4, S4 Table

The purpose of each numbered experiment is shown in addition to the title used for each one in the methods and results section. The appropriate methods sections, results sections and figures to consult for each experiment are also given.

doi:10.1371/journal.pone.0124671.t001

Materials and Methods

Some of the following methodologies were specifically designed for this study; others have utilised samples and/or data drawn from previous studies [24,35–38]. The materials and methods below provide an overview of the methodologies and the reader is referred to the original publications and also the supplementary online information where schematics of all experiments conducted are presented (S1A–S1F Fig). Each of four important steps in amplicon workflows: sample screening (Section 2.2.1 and S1A), the target region (Section 2.2.2 and S1B Fig), library generation (Sections 2.2.3, 2.2.4 and S1C–S1E Fig) and data analysis (Section 2.2.5 and S1F Fig), is addressed separately in the materials and methods that follow. General methods employed during sample screening, amplicon generation, DNA sequencing and data analysis that were common to all areas are detailed first (Section 2.1) before more focused information on each of the four aforementioned steps (Section 2.2). Any further detailed information on the samples or experimental workflows used is available in previous publications [24,35–38] or from the authors upon request. Where applicable amplicon sequence reads have been uploaded to Data Dryad (doi:10.5061/dryad.2qf0t).

2.1. General methods

2.1.1. DNA extraction and screening. A variety of samples and extraction methods are used throughout these experiments. Extraction protocols followed can be found in the original publications where indicated [24,35–38], but typically involved silica-based purification methods to isolate DNA. Where sample extraction has not been reported previously, the details of the extraction procedure are found below in Section 2.2.

All samples used were screened to determine the appropriate working dilution containing sufficient DNA free of inhibition using quantitative PCR (qPCR) on a SYBR-based STEP-ONE Applied Biosystems Real-Time PCR instrument [35,39]. Samples were assessed based on Cycle Threshold (C_T) values, curve form and melt-curves. Extraction controls were conducted for each batch of extractions and screened using qPCR to test for contamination arising from laboratory practice, reagents, or the environment. If positive for the presence of DNA, extraction controls were included in tagged qPCR assays (see Section 2.1.2). All qPCR reaction conditions and reagent components can be found in previous publications where indicated below, and primer details can be found in S1 Table. Details are provided below for any qPCR reactions not previously reported.

2.1.2. Amplicon generation and sequencing. For samples deemed to have sufficient DNA copy number and determined to be free of inhibition, amplicon sequences were always generated in triplicate via qPCR using a unique combination of forward and reverse Multiplex Identifier (MID-) tagged (i.e. indexed) primers [27,40] (for the only exceptions to this see Section 2.2.1 and S1A Fig). For each tagged qPCR assay, negative reaction controls were included and, if found to contain amplifiable DNA, were incorporated into the appropriate sequencing library. Resultant amplicon products were purified following the Agencourt AMPure XP PCR Purification Kit protocol (Beckman Coulter Genomics, NSW, Aus.) and were eluted in 40 μ L of Ultrapure H₂O. Purified amplicon products for each sequencing library for each platform were electrophoresed on ethidium bromide stained 2% agarose gel and pooled in equimolar ratios based on band intensity to form sequencing libraries.

In order to determine an appropriate volume of library for sequencing, each amplicon library was serially diluted and quantified using qPCR against a serial dilution of a custom synthetic oligonucleotide of known molarity. Reaction components and conditions were the same for each sequencing platform with the exception of platform specific primers appropriate to the sequencing adaptors. Each 25 μ L reaction contained 2X ABI Power SYBR master mix

(Applied Biosystems, CA, USA), 0.4 μM each of platform specific forward and reverse primer (IDT), and 2 μL of pooled library. Each reaction underwent the following cycling conditions: 95°C for 5 mins; 40 cycles of 95°C for 15 s, 56°C for 1 min followed by a 1°C melt curve. All sequencing was conducted according to manufacturer’s protocols using one of three sequencing platforms: GS Junior (Roche), Ion Torrent PGM (Life Technologies) and MiSeq (Illumina). Sequencing on Roche was conducted using LibA chemistry. Ion Torrent PGM emulsion PCR (emPCR) was conducted on a OneTouch2 using 400bp chemistry and sequencing was performed on 314 chips. Finally, Illumina MiSeq sequencing used V2 300 cycle chemistry on nano flow cells. To enable direct comparisons both PGM and MiSeq used single direction sequencing only, despite the fact that paired-end sequencing is available in the latter.

2.1.3. Data analysis. Regardless of the platform, amplicon sequence reads were deconvoluted in Geneious v7.1.3 (this version of Geneious is used throughout this paper) [41] based on unique primer indexes. As a first step in deconvolution any sequences found to contain ambiguous base calls (e.g. N) were discarded. Identification tags and primer sequences were trimmed from all reads in Geneious, allowing for no mismatch in either length or base composition as a means of quality filtering, using the inbuilt “Separate Reads by Barcode” and “Trim Ends” functions respectively. The only exception to this can be found in Section 2.2.5 where in some instances two base mismatches in the primer sequences were allowed (see also Fig 1 and Section 2.2.5). Unless otherwise stated in Section 2.2, Quality Score (Q-Score) filtering was not performed. Sequences were subsequently dereplicated at 100% identity across their full length using USEARCH v7 (this version of USEARCH is used throughout this paper) [42,43], and low abundant sequence clusters, defined as those below 1% of the total number of unique sequences, were removed using USEARCH also. Dereplicated sequences were clustered at a 97% threshold using the UPARSE [43] algorithm implemented in USEARCH. Chimeric sequences were also identified and removed using USEARCH [42,44]. At all stages of dereplication and OTU clustering abundance information was retained and used when calculating taxa/sequence abundance or error rates. Where appropriate, sequences were queried against the NCBI GenBank nucleotide database [45] using BLASTn [46] in YABI [47], enabling taxonomic identification. Sequences were searched without a low complexity filter, with a gap penalties existence of five and

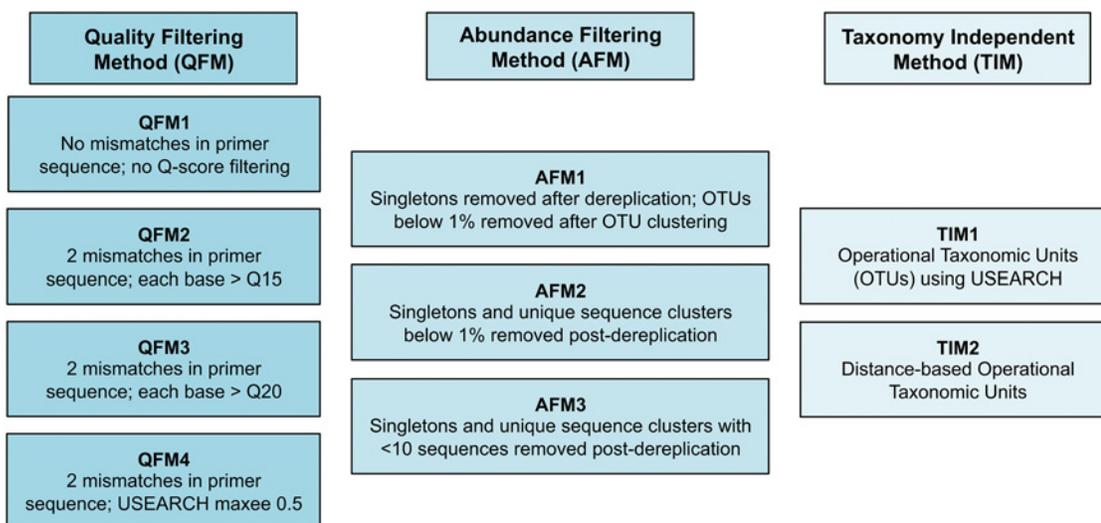


Fig 1. Definitions used in assessing the importance of analysis parameters. Shown are the definitions for quality and abundance filtering methods used in assessing their impact on both the number of operational taxonomic units (OTUs) and distance-based operational taxonomic units (DTUs) [24] obtained for a given sample. maxee—Maximum Expected Error

doi:10.1371/journal.pone.0124671.g001

extension of two, expected alignment value less than $1e-10$ and a word count of seven. The BLASTn results obtained were imported into MEtaGenome ANalyzer v4 (MEGAN) [32], where they were mapped and visualised against the NCBI taxonomic framework (min. bit score = 35.0, top percentage = 5%, min. support = 1). In cases where taxonomic identification was necessary, a genus or family level assignment of a query sequence was required to have a BLASTn percentage similarity to a reference sequence of 97% or 95% respectively. Instances where data analysis deviated from the above steps are detailed where necessary below.

2.2. Specific methodologies

2.2.1. Experiment 1: Importance of sample screening. To evaluate the importance of screening samples for inhibition and low target template amount, an environmental faecal sample was obtained from a *Eudyptula minor* (Little Penguin) individual. DNA was extracted from the faecal sample, serially diluted, and screened via qPCR as described in Murray et al. [35] using 16S1F/16S2R degenerate fish primers [48] (see also S1A Fig and S1 Table). An appropriate working dilution of the sample deemed to have sufficient DNA copy number and free of inhibition (see Section 2.1.1) was used for sequencing on both the Ion Torrent PGM and GS Junior. In addition to this, both an aliquot of the working dilution spiked with an extremely inhibited soil DNA extract, to mimic inhibition, and a dilution classed as “Low Template” were selected for sequencing. For each sample, the detection and percentage abundance of two baitfish genera, *Sardinops* (specifically *S. sagax*—Australian pilchard) and *Engraulis* (specifically *E. australis*—Australian anchovy) were examined. The former being in the highest abundance: the latter in lowest abundance, as determined by a taxon-specific qPCR assay (see S1 Table and [35]).

The handling of the penguin, and the collection and use of the faecal sample was conducted by experienced handlers under a strict set of animal ethics guidelines approved by the Murdoch University Animal Ethics Committee (permit no. W2002/06) as part of a long-term study into *Eudyptula minor* (Little Penguin) diet. Faecal sampling and DNA extraction were performed as part of a previously published study [35] and not as a part of this study, however ethics approval covers the use of the faecal sample DNA extract in this study.

2.2.2. Experiment 2: Assessing the amplicon target region. Five single-source bird tissue samples were used to assess error profiles associated with a specific amplicon target region (see S1B Fig). *Calyptorhynchus latirostris* (Carnaby's Black Cockatoo) and *C. lathami* (Glossy Black Cockatoo) samples were collected, and DNA extracted, as detailed in White et al., 2014 [38]. Tissue samples of *Gallus gallus* (Chicken), *Dromaius novaehollandiae* (Emu) and *Struthio camelus* (Ostrich) were bought commercially and DNA was extracted using a Qiagen DNeasy Blood and Tissue Kit following the manufacturer's protocol. For each sample an approximately 250 bp region of the mitochondrial 12S rRNA gene was amplified and MID-tagged using 12SA/H avian primers (see S1 Table and [49,50]) via qPCR (reaction components and conditions as detailed in [24]), and then sequenced on both Ion Torrent PGM and Illumina MiSeq platforms.

Amplicon sequence reads for each bird were randomly sub-sampled a total of 25 times to a depth of 1,000 sequences using seqtk (available from <https://github.com/lh3/seqtk>) following deconvolution into sample batches (see Section 2.1.3). Each sub-sample was dereplicated at 100% identity to determine the most abundant sequence, with the abundance of each unique sequence appended to sequence names for use in calculating error rates. The most abundant sequence was taken as the reference sequence. For both platforms the most abundant sequence was identical thus meaning it is likely ‘correct.’ Each set of sub-sampled sequences was individually aligned using MUSCLE with default parameters [51]. Alignments were imported into excel and for each sample the error associated with each base was calculated as a percentage of the total number of non-dereplicated sequences that differed from the reference sequence at that

specific base. This was performed using an in-house macro; the output of which can be seen in [S1 File](#). The error associated with each sub-sample was subsequently calculated as the mean error across all bases. The overall percentage error rate for each bird species on both the Ion Torrent and MiSeq was taken as the mean error rate across all 25 sub-samples of each species.

The collection and use of DNA material from Cockatoos was approved by, and conducted under, Department of Parks and Wildlife (Western Australia) scientific purposes licences SC000357, SC000920, SC001230, Australian Bird and Bat Banding Authority 1862 and Animal Ethics Committee approvals DEC AEC 11/2005 and 32/2008 held by P. R. Mawson. Samples of Chicken, Emu and Ostrich (all non-endangered) were purchased from Franks Gourmet Meats, Perth, WA, Australia, and are exempt from a collection permit.

2.2.3. Experiment 3: Importance of experimental controls. To illustrate the importance of control reactions in bacterial metagenomics and other fields dealing with samples with a high likelihood of environmental contamination, bacterial 16S data from hair samples were generated and analysed as detailed in Tridico *et al.*, 2014 ([37], see also [S1C Fig](#) and [S1 Table](#)). Briefly, pubic and scalp hair were self-sampled by male and female volunteers. Hair samples were prepared and extracted as detailed in Tridico *et al.*, 2014. Samples were screened using Bact_16S_F515 and Bact_16S_R806 primers [52,53] and amplicon libraries were generated, sequenced and analysed as per Tridico *et al.*, 2014.

The collection of human hairs for bacterial profiling was approved by, and conducted in accordance with, Murdoch University Human Research Ethics Committee Policies and Guidelines (Project Number 2011/139). Each volunteer was made aware of the nature of the study and gave written, informed consent. Hairs were self-collected from two somatic origins and placed in sample bags bearing no information that would allow the identification of any individual participant in the study [37].

2.2.4. Experiment 4: Library generation efficiency. Quantitative PCR using the plant plastid trnLg/h primer set [54] was carried out to investigate the issues surrounding efficiency drop-off associated with the use of “full” fusion tagged primers (see [S1D Fig](#) and [S1 Table](#)), i.e. those with MID tags and sequencing adapters upstream of the template specific primer (TSP) (see [S1E Fig](#) and [40]). A single-source plant extract in addition to two complex, heterogeneous Traditional Chinese Medicines (TCM) were used; a MoBio Plant DNA Isolation kit was used following the manufacturer’s protocol for the single-source plant sample DNA extraction, while sampling and extraction of TCMs are detailed in Coghlan *et al.* [36]. Each sample was amplified in triplicate using either (1) standard non-fusion TSP; (2) MID encoded TSP (3) “full” fusion tagged TSP or (4) “full” fusion tagged TSP with standard non-fusion TSP spiked in (see [S1D](#) and [S1E Fig](#)). For (1–3) each qPCR reaction was carried out in a total volume of 25 μ L containing 2X ABI Power SYBR master mix (Applied Biosystems, CA, USA), 0.4 μ M each of the appropriate forward and reverse TSP (IDT) and 2 μ L DNA extract. For (4) the previous components were also used but an additional 0.04 μ M spike-in of each the forward and reverse standard non-fusion TSP (IDT) was also used. For each reaction C_T threshold was set at 0.1.

TCM samples were obtained from, and approved for use by, the Wildlife trade section of the Department of Sustainability, Environment, Water, Population and Communities (Australia) after being seized by Australian Customs and Border Protection Service at airports and seaports across Australia. The samples were seized because they contravened Australia’s international wildlife trade laws as outlined under Part 13A of the Environment Protection and Biodiversity Conservation Act 1999 (EPBC Act). The samples were stored in a quarantine-approved facility within the laboratory after being catalogued. The samples were patent medicines available over the counter and were donated by Australian Customs and Border Protection Service under no ethics or quarantine requirements and were deemed suitable to be used for specific and general research purposes by the Customs service [36].

2.2.5. Experiment 5: Analysis parameters and their impact. To demonstrate the variability in calculated OTU (operational taxonomic unit) diversity within a sample, a single bulk-bone sample, comprising ~50 individual bones and containing an unknown number of taxa, was extracted and screened using the 16Smam1 and 16SMam2 mammalian specific primer set [55]. Amplicon sequences were generated for short sections within the mammalian mitochondrial 16S gene using the 16Smam1 and 16SMam2 primer set and sequenced using the Ion Torrent PGM as described in Murray *et al.*, 2013 [24] (see also [S1F Fig](#) and [S1 Table](#)). After deconvolution following the method detailed in 2.1.3 the data were analysed using various quality filtering methods (QFM), abundance filtering methods (AFM), and taxonomy-independent methods (TIM) of diversity analysis as shown in [Fig 1](#). Quality Score filtering was conducted in Galaxy [56–58] using the FASTQ Quality Filter tool. Maximum expected error (maxee) quality filtering, set at 0.5, was conducted using the `fastq_filter` command in USEARCH. Summary quality statistics were calculated in excel using `fastq` files post quality filtering for QFM1 and QFM4, prior to any further abundance filtering. Dereplication and OTU clustering at 97% was conducted using USEARCH also. DTU's were determined post OTU clustering as described in Murray *et al.*, 2013. Briefly, for DTU analyses, OTU's were aligned using MAFFT [59] and alignments imported into MEGA v6.06 [60] where a distance matrix was created and exported. To determine OTU's that differed from each other by less than 3% distance matrices were analysed in excel using an in-house macro, an example output of which is shown in [S2 File](#). [24]. The impacts of DNA preservation, DNA degradation, mode of bone accumulation and deposit setting will have negligible impact on the results of this experiment as the exact same set of amplicon sequences, from the exact same DNA extract, are used for each combination of QFM, AFM and TIM used. The dataset in this experiment is therefore static throughout and any biases introduced by any of the aforementioned factors will be the consistent across all methods.

Results and Discussion

Much attention has been devoted to the bioinformatic challenges associated with the analysis of amplicon sequencing data. There are a suite of programs, tools and pipelines available to assist in the deconvolution, filtering and parsing of data. As a relatively new field there is no obvious consensus on how data should, or should not, be handled bioinformatically, with the exception that sequence clusters in very low abundance should be filtered. Likewise there is no consensus on what is best-practice for data generation. Arguably the importance of data generation has taken a backseat to the computational workflows that surround bioinformatics. Bioinformaticians, rightly so, ask key questions of researchers with regard to replicates, coverage and filtering. They are less likely to ask questions about input copy number, PCR inhibition, contamination and the appropriateness of benchtop protocols. This study, through the presentation of new and existing empirical data, seeks to demonstrate the importance of both benchwork and bioinformatics. The purpose of this study is to raise awareness of potential pitfalls associated with amplicon-based workflows. The workflows dealt with in this paper do not include the process of actual DNA extraction, itself undeniably important, as this has been dealt with extensively elsewhere. The workflows presented here take as their starting point a working, amplifiable DNA extract, which can only be achieved through the careful consideration of both the scope of the project and type of substrate.

3.1. Experiment 1: Importance of sample screening

Adequate screening of samples prior to sequencing is an important task, yet fails to be routinely implemented in amplicon workflows. It is particularly prudent to assess the quality of samples when dealing with complex, heterogeneous substrates that may contain a variety of taxa or

when examining samples that may contain highly degraded or low copy number DNA. There are arguably two primary factors that should be considered when evaluating samples for sequencing: the extent of inhibition, and the number of target input DNA template molecules used in generating an amplicon sequencing library. Both inhibition and low template number can have a negative impact upon the results obtained from amplicon sequencing workflows and failure to account for both can exacerbate other biases associated with amplicon sequencing. Common methods of screening samples include quantitative PCR (qPCR) and PCR end-point assays such as gel electrophoresis or capillary electrophoresis (e.g. Agilent Bioanalyzer). The advantage of using qPCR over end-point electrophoresis lies in the fact that it is easy to determine whether or not a sample is inhibited through the analysis of the Cycle Threshold (C_T) values in a dilution series and the resultant curves. Traditional end-point assays such as electrophoresis are a blunt binary-state tool to assess inhibition and low-template samples; both will still produce bands on a gel (see gel image in Fig 2) or peaks on a Bioanalyzer trace. A case is not being made that samples should not be subjected to electrophoretic analysis, as this is a useful means for determining the presence of PCR artefacts. Rather, it would be practical to consider the additional use of qPCR or other similar methods of quantification (e.g. digital PCR), to assess the levels of inhibition and the absolute, or relative, number of target template molecules that are the input for amplicon sequencing workflows.

In a simple experiment involving the detection of two genera of fish, *Sardinops* (specifically *S. sagax*—Australian pilchard), in high abundance, and *Engraulis* (specifically *E. australis*—Australian anchovy), in low abundance, the effects of not being cognisant of inhibition or low DNA copy number are clearly demonstrated. When an appropriate working dilution exhibiting a sufficient number of input template copies and deemed free of inhibition (as determined by qPCR), was sequenced both fish species were detected in all PCR replicates, across two platforms (Fig 2, green line and shaded table). Furthermore, *Sardinops* was consistently detected as the fish species in the greater sequence abundance. In the case of the inhibited aliquot (Fig 2,

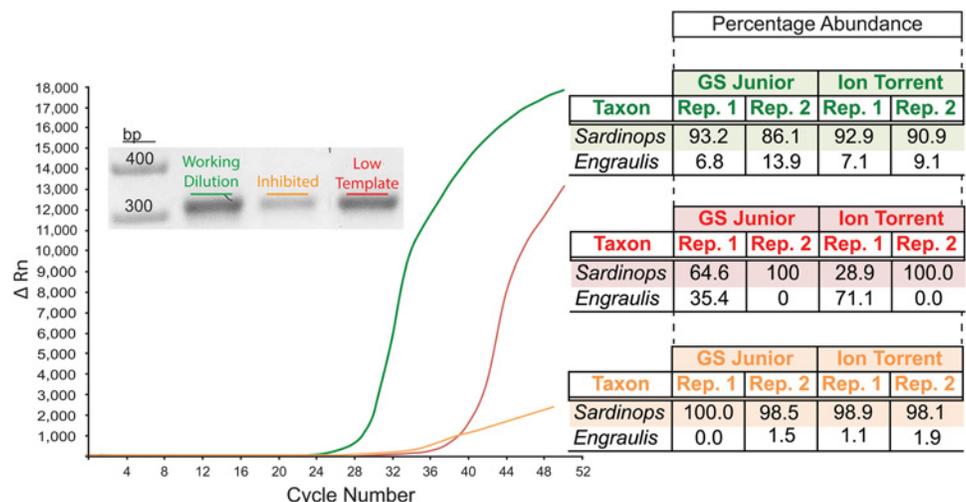


Fig 2. Quantitative PCR and sequencing results of the sample screening assay. Quantitative PCR curves indicating the presence of DNA and the degree of inhibition (LEFT) with agarose gel electrophoresis clearly indicating the presence of DNA post amplification via means of strong bands (INSET ON GRAPH). Samples were subsequently sequenced and the percentage abundance of two fish genera is indicated, where, based on taxa-specific quantitative PCR results, *Sardinops* (specifically *S. sagax*—Australian pilchard) should be in the highest abundance, with *Engraulis* (specifically *E. australis*—Australian anchovy) being in the lowest abundance. (RIGHT).

doi:10.1371/journal.pone.0124671.g002

orange line and shaded table) *Sardinops* was detected in all replicates and across both platforms, however *Engraulis* was not, and in those instances where it was detected it was typically at abundances <1%. When the low-template sample dilutions (Fig 2, red line and shaded table) were sequenced a similar pattern was observed, with again *Sardinops* being detected in all replicates and across both platforms and *Engraulis* being detected in only a few (see [4] for a further example of the non-detection across multiple replicates of a target species known to be in a sample). In this instance, the abundances were vastly different between the replicates and in one instance *Engraulis* appeared to be the fish species in the highest abundance.

The inclusion of PCR and/or sequencing replicates is without doubt an important aspect of any amplicon workflow serving to improve confidence and reliability in data interpretation ([61,62] although see [63]). Efforts have been made to determine the optimum level of PCR replicates, but it is acknowledged that the degree of replication required is dependent on the complexity of the sample in question and the objective of the study [61]. Additionally, it is also clear that simply increasing the depth of sequencing does not necessarily translate into an increased ability to detect low abundant taxa. In this study the increase in sequence depth afforded by the Ion Torrent did not improve *Engraulis* detection success. Arguably an extremely important, yet somewhat overlooked, aspect in generating an accurate species profile contained within any given sample is paying close attention to template input amount and quality, i.e. the level of amplifiable DNA and the degree of inhibition. This is becoming increasingly important as research efforts are moving towards quantitative interpretations of sequence abundance. Simply replicating PCRs using poor quality extracts is a blunt means of increasing the fidelity of amplicon sequence data.

It is acknowledged that PCR bias can greatly skew amplicon sequencing workflows [64–66], this is especially true when little or no attention is paid to input template amount or a sample's amplifiable limits. Although only a small-scale experiment, the above serves to illustrate the importance of screening samples prior to sequencing (Fig 2). Amplicon sequencing results can clearly be obtained with low-template and inhibited samples but the reproducibility of these results is questionable: even more so if they are subsequently used in weighted analyses. Even when not interested in the relative abundance of taxa, OTUs or sequence variants, it is still nonetheless useful to screen samples for inhibition and low template amounts, as both of which can increase the possibility of false negatives. Whilst the absence of something in a sample can never truly be proven, being aware of the level of inhibition inherent within a sample or an estimate (however crude) of relative input can greatly improve the confidence surrounding presence, possible absence and/or abundance conclusions based off amplicon data. A common theme in the literature, including work by the authors, is to report the number of amplicon sequence reads obtained, but in reality a much more useful metric is to state the relative or absolute number of target templates provided to the reaction per replicate. In other words sequencing coverage is often a meaningless statistic—a PCR reaction that starts off a single molecule being the case in point. An increase in the use and reporting of quantitative data in amplicon workflows using qPCR or digital PCR can only assist in data fidelity and meaningful downstream analyses.

3.2. Experiment 2: Assessing the amplicon target region

Irrespective of the gene region chosen for investigation it is advisable to be aware of the composition of that region. This holds true especially for methods that rely on a small amount of data from the target region to infer conclusions, such as SNP data or taxonomic assignments between closely related taxa based off a few nucleotides. The primary reason for such attention is due to the fact that not all gene regions are “created” equal. Some gene regions can be more prone to error due to the occurrence of homopolymer stretches or secondary structures within the target area, particularly when dealing with 454 or Ion Torrent data. There are also well-

recognised issues with quality and fidelity when dealing with target regions that are GC rich [67–70]. Both of these issues are in addition to the typical drop off in sequence quality and increase in potential error observed towards the sequencing length limitations of any given platform. The error rate, in addition to the quality of an amplicon sequence, is not uniform across the length of itself (Fig 3) nor is there necessarily a common error rate across different amplicon targets. Also worth noting is the potential for error rates to fluctuate between runs on the same platform on the same control DNA.

Some amplicon regions will undoubtedly sequence better than others due to the presence or absence of homopolymer regions and the complexity of the base composition. Rather than relying on generic error rates reported by the manufacturers or in the literature in the case of amplicons it is preferable to determine the error rate for the target region. In a small-scale experiment where single source samples for multiple bird species were sequenced, the error profile of the chosen ~250 bp target region of the 12S gene can be seen (Fig 3). It is evident that on both platforms the overall error varies slightly from species to species, yet a much greater range of mean error rates is observed in the case of the Ion Torrent PGM relative to MiSeq sequencing (Fig 3A). The variation in error rates observed across species is likely due to overall error rates associated with each platform. In addition to this it is observed that the percentage error for certain regions and specific bases far exceed the reported error rates cited for the platforms and in some cases, most notably with the Ion Torrent, certain regions recorded error rates as high as 7% (Fig 3B and 3C). Moreover, the increased error beyond that reported for the platform, and in some instances greater than 1%, often cited as a level used to eliminate erroneous sequences, is not solely confined to the 3' end of the amplicon read. In the case of the Ion Torrent an error rate of 13.5% was calculated just 80 bases into the amplicon read (Fig 3B). Although significantly lower error rates at specific bases and in specific regions was observed in the MiSeq, bases and regions recording error rates approaching the 1% mark were found mid-way through the amplicon. In both cases this is despite average error rates for those sub-sampled sequences being calculated as 0.48% for Ion Torrent and 0.21% for the MiSeq (Fig 3B and 3C). The propensity for error is again highlighted in the case of the Ion Torrent whereby only 33.3% of sequences obtained for that sub-sample were contained within the highest unique cluster, which is alarming given that it is a single source sample, with theoretically only one possible sequence composition, yet two thirds of the sequences differed from the most common. Although the error profile for only one sub-sample for a single species (*C. lathami*) is shown for both the Ion Torrent and MiSeq in Fig 3B and 3C a similar error profile was found across all species on both platforms.

When dealing with amplicon sequencing, determining not only the overall error rate for the target region but also calculating an error spectrum can have many benefits. In doing this, certain error “hot-spots” can be detected, and being aware of the presence of such areas can enable more informed decisions in relation to determining OTUs, calling SNPs and verifying taxonomic identifications. Having a good understanding of the composition of the chosen target region can also be of benefit. If the area of the amplicon that proves to be most informative is at the 3' end of the amplicon sequence for instance, it is possible to optimally position the direction of sequencing. The profile may also dictate if a paired end strategy is more appropriate. Single-source samples specific to the targeted gene region can also facilitate the monitoring of run-to-run variation in error rates specifically for the amplicon of choice.

Awareness of the error profile and composition of an amplicon gene region is an important consideration that can impact upon one's ability to taxonomically discriminate taxa. If an amplicon sequencing approach is adopted some of the biases associated with PCR and primer skews may also be minimised, or can at least be highlighted, by ensuring that the primer binds on all taxa of interest through the use of *in silico* bioinformatics [71]. It is also worth being aware of the fact that no primer is truly universal. It is therefore worthwhile to consider the use of a

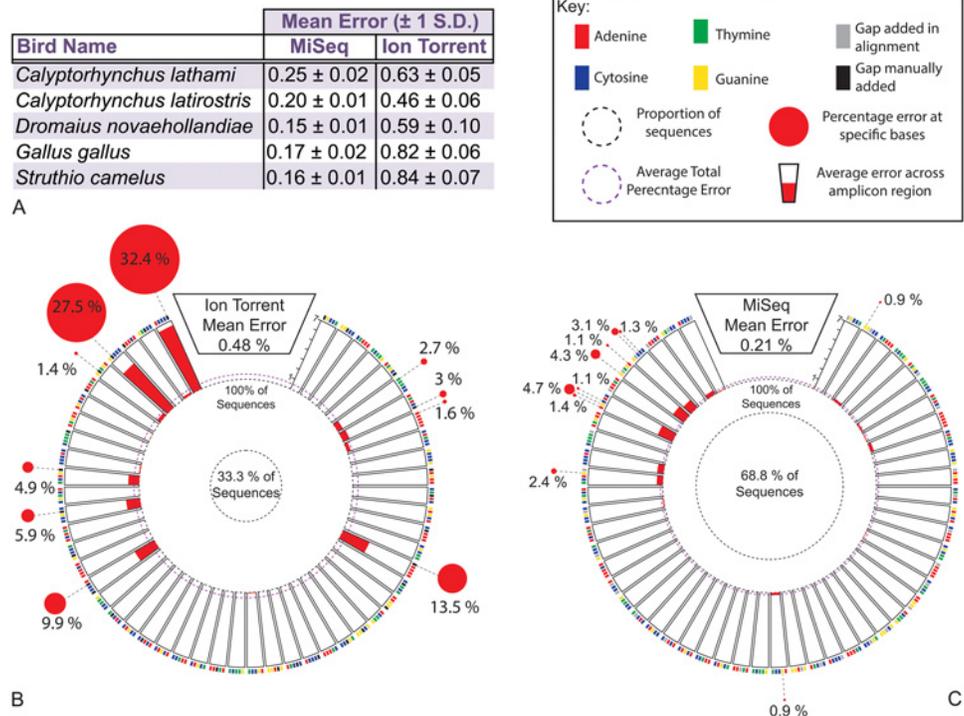


Fig 3. Average sequencing error rates across a single amplicon region. Average sequencing error rates are shown for multiple bird species across the whole of a short 12S rRNA gene region (A). Additionally, the error profile across the gene region is shown for *Calyptorhynchus lathami* for both the Ion Torrent PGM (B) and MiSeq (C) with key. The error patterns observed were similar across all species sequenced. Error rates are shown across 5 bp segments and where error rates were above 1% for a single base this is indicated through the red circles.

doi:10.1371/journal.pone.0124671.g003

multi-locus approach especially given the current patchy state of reference databases where some taxa may be present for one gene region but not another [72,73]. Lastly, it is worth noting that just because a primer set is said to “work well” in one study (or because it is a currently accepted DNA barcode) it does not necessarily follow that it will also be fit for purpose in another study. This issue is clearly highlighted in the case of Australian mammals where the conventional barcode COI is wholly insufficient due to the poor representation of Australian marsupials and rodents for this gene in current databases such as GenBank or BOLD [24,72,74].

3.3. Experiment 3: Importance of experimental controls

Once an appropriate target region(s) is selected and DNA extracts are screened for copy number and inhibition, decisions then turn to how best to build a library free of artefacts and contamination. The issue of contamination and artefact formation should always be considered when PCR is involved. Amplicon sequencing on 454, Illumina or Ion Torrent, always involves the manipulation of PCR products, thus workflows are susceptible to contamination. Amplicon sequencing workflows on current second generation platforms involve multiple rounds of PCR [75,76], many published workflows utilise three rounds of PCR [77–80]: a primary PCR, an MID (Multiplex Identifier) tagging PCR (i.e. indexing) and then amplification within emulsions (454, Ion Torrent) or on a flow cell (Illumina). Unlike Sanger sequencing when low-level contaminants presented as a ‘bumpy’ baseline, HTS will show these as unambiguous sequences. In many respects high-throughput amplicon sequencing should be viewed as the “white-glove” test of laboratory cleanliness.

A major potential source of contamination is due to the handling of amplicon products post-PCR. Thus it is strongly recommended (where possible) to conduct pre-PCR and post-PCR work in independent, dedicated spaces or labs, preferably physically separated from each other. It is advisable to minimise the handling of untagged amplicon products as much as possible to prevent cross-contamination of samples. It is for this reason that methods such as nested- or hemi-nested PCR, reamplification, and ligation of 'sequencing adapter-MID tag' sequences to untagged amplicons can be problematic. Employing nested-PCR approaches to enrich for low abundant taxa may be more prone to contamination and/or artefactual sequences when compared to PCR-free targeted enrichment of amplicons.

It goes without saying that minimising contamination is essential in all studies where amplicon sequencing is used, especially those that seek to explore diversity in instances where it arises as a result of low-abundant taxa or variants [81,82]. The increased sequencing depth afforded by HTS should not be viewed as a means by which to "cut-through" potential contamination be it environmentally derived or otherwise. This is particularly true in scenarios where endogenous DNA is highly degraded or in low copy number, as is the case for ancient or environmental DNA, where modern or well-preserved DNA sequences will amplify more readily. The degree to which a sample has been contaminated cannot be known *a priori* and such contamination, especially environmentally derived, may not always be low-level. Increased sequencing depth, therefore, will do nothing to dilute the level of contaminant sequences, and neither will arbitrary cut-offs designed to remove low-abundant unique sequence clusters or OTUs. There is no substitute for environmental, extraction and PCR blank reaction controls. The failure to use controls can never be justified and nor can the failure to report the use of controls, even when they turn up negative results. Controls are the only true means by which it can be determined whether or not the fidelity of samples have been maintained throughout processing. Controls are seldom reported in papers using HTS [83], especially in the fields of environmental DNA and microbial metagenomics. The lack of reporting of controls in bacterial metagenomics studies is alarming given the ubiquitous nature of bacteria. In the absence of such controls it is impossible to say what bacteria are endogenous to the samples collected or even the extent to which bacteria common to the environment contribute to the microbiome from which the sample was collected. This is particularly true when dealing with coarse taxonomic assignments at an ordinal or family level, not to mention when making claims about the presence, absence and/or abundance of OTUs. The importance of controls in bacterial metagenomics is clearly shown when considering that after OTU sequences present in control reactions conducted during bacterial profiling of hairs [37] were removed the number of OTU sequences present in scalp hair samples dropped by ~60–70% (S2 Table). Moreover, it is clear that this is not a simple case of PCR contamination arising from poor lab practice as the drop off for pubic hair, conducted within the same PCR plate was much lower at ~30% (see S2 Table and [37] for further details and also [84] for another example of using controls to filter sequences for contamination). High-throughput sequencing serves to hold up a magnifying glass to the laboratory practices of any lab that makes use of it. The depth at which a sample can be sequenced can result in even the lowest levels of contamination being revealed. This can be problematic where analyses and conclusions rely on low abundant sequences and the only assured means of retaining confidence in results and conclusions in these cases is through careful library preparation and considered data analysis. While it is easy to pick out common laboratory contaminants or aberrant sequences when such amplicons assign taxonomically to taxa not found in the study area, it is more difficult to account for cross-sample, environmental or laboratory contamination that closely resembles the taxa or sequence variants of interest.

The use of indexed (or MID tagged) primer sequences is not only useful in allowing the processing of multiple samples in parallel but it is also a convenient means by which to filter. This

can be achieved by only allowing amplicon sequences with the exact MID tag to be used in further analyses. However, the use of the word “unique,” and other related terms, with respect to these MID tags is slightly misleading as in reality MID tags are often recycled across many samples. This may prove problematic due to sample carry-over that is observed with some platforms or potential library contamination by means of aerosolised particles during library generation. The issues surrounding the possibility of sample carry-over is best illustrated when considering the first Ion Torrent PGM run that the authors of this paper outsourced to a sequencing facility where, when the data was analysed, 25 tags not used in the preparation of the amplicon library were detected, amounting to 0.02% of the total number of reads returned. Out of these 25 tags, if the tag that was present in the greatest abundance had been used in the experiment, approximately 1.2% of the reads belonging to the sample to which it was assigned could have been indistinguishable contamination. In this instance it was clear that the contamination might have arisen at the sequencing facility itself as none of the tags detected were ever used in the laboratory where the amplicon library was generated. This highlights an important issue when considering the outsourcing of DNA sequencing to other labs, commercial or otherwise. It may be necessary in future to provide statistics of run-to-run carry over and the timeframe between the re-use of tags when such a sequencing facility also generates the amplicon for sequencing. Numerous studies are now beginning to highlight the issue of contamination arising from the laboratory, reagents and commercial kits [82,85]. Anecdotally, researchers also talk about contaminating data from sequencing facilities but it is rarely, if ever, reported in the literature.

A simple strategy to limit issues associated with this is to increase the timeframe between the first use and subsequent re-use of an MID tag. While it is tempting when dealing with a small number of core loci to re-use a limited number of tags, such as those officially released by the platform manufacturers, it nonetheless increases the likelihood of contamination creeping in from run to run and building up over time. Expanding the number of MID tags used in a lab greatly reduces the potential of MID tag contamination with little extra cost. A further means of ensuring tag contamination is kept to a minimum is the use of differing MID tags at the 5' and 3' end of the amplicon sequences (see Section 2.1.2), which can also benefit in terms of data filtering to increase the likelihood of only high quality sequences being retained. Additionally, the use of different 5' and 3' MID tags on an amplicon greatly increases the number of possible combinations at a laboratory's disposal. Finally, the use of different 5' and 3' MID tagged amplicons may also help in the detection of chimeric sequences. The downside of a method such as this however is the cost associated with ordering primers; although this can be kept to a minimum by not ordering HPLC purified primers as synthesis errors are easily managed by post-run filtering. Moreover quality control validation by mass spectrometry is now commonplace and serves to minimise the likelihood of primers with high proportions of incorrect bases.

While some might argue that the purchase of MID tagged primers is expensive the counter argument is that so too is repeating runs where the researcher believes the data is compromised. In our lab six reads were detected of a Chinese herbal plant from one study [36] that turned up in a palaeosediment sample from Australia. In this instance both samples shared the same MID tags despite being many runs apart. In sensitive applications the re-use of MID tags may be a false economy. Low-template samples necessitate sensitivity and single-use of tag combinations. This has the added benefit that each amplicon product generated is unique to the originating sample and contamination can be removed bioinformatically.

3.4. Experiment 4: Library generation efficiency

The opening and closing of PCR-tubes or plates post-PCR and the handling of untagged amplicon products serve to increase the chances of untraceable contamination as a result of poor

laboratory technique or the release of aerosolised amplicons. It is for this reason that a single “full” fusion tagged TSP (see [S1E Fig](#)) PCR approach [21,86] or sequencing adapter ligation post-MID tagging [27] via PCR method is preferable from the perspective of contamination control. The drawbacks associated with a “full” fusion tagged TSP PCR approach centre around a loss of PCR efficiency due to the long fusion primers required and also the problems surrounding primer-dimer. However, careful size selection can assist with dimer removal [87–90]. The ligation of sequencing adapters post-MID tagging via PCR itself can be inefficient and may be biased towards the preferential ligation of certain amplicons or terminal bases. In some cases the efficiency drop-off associated with a “full” fusion tagged TSP approach can be mitigated through the use of the modular tagging of amplicons using a single PCR (MoTASP) method [21] or by simply spiking in some standard non-fusion TSP into the PCR reaction containing “full” fusion tagged TSP (see [S1E](#)). The latter showed generally modest efficiency improvements when compared to qPCR in the absence of spiking in standard non-fusion TSP, however the C_T value shifts in qPCR varied considerably for each platform ([S3 Table](#)). Additionally, the spiking in of standard non-fusion TSP when using “full” fusion tagged TSP still showed a general increase in C_T values when compared to qPCR containing only standard non-fusion TSP, particularly in the case of the MiSeq ([S3 Table](#)). Although the MoTASP method has been reported to improve PCR efficiency, it is unclear as to the extent this may be the case as qPCR was not carried out and neither was a direct comparison of sequencing results [21].

The use of a “full” fusion tagged TSP approach where a library is generated in a single step is theoretically the cleanest way to generate amplicon libraries. The downside to this is the drop in PCR efficiency discussed above. A common alternative pathway is a series of primary PCRs which are pooled and followed by a secondary PCR to amplify sequencing adapters and/or MID tags onto the target sequences. Notwithstanding the contamination risk inherent to this two-step approach it is also the source of inter-sample chimeras, presumably through incomplete extension and/or ‘jumping’ PCR [91]. Practitioners need to carefully weigh the benefits and drawbacks of each library building method and be cognisant of how the method impacts on the conclusions they hope to draw from the resultant data.

3.5. Experiment 5: Analysis parameters and their impact

It is beyond the scope of this study to delve into the complexities of data analysis. It is however relevant to note that amplicon data can be analysed in many different ways, sometimes subtly so, that can result in quite dissimilar outcomes. It is also worth noting that analysis parameters are contingent on the benchwork component of amplicon sequencing workflows. To date there is no currently accepted best practice pipeline or approach to the analysis of amplicon sequencing output, although many do exist [28,42,92,93]. Nevertheless one of the few agreements on the way in which both shotgun and amplicon sequencing data is handled is the necessity to filter sequences for error and potential contamination in a manner that strikes a balance between overly relaxed and unnecessarily stringent filtering. The manner in which such filtering is done and the definitions associated with various processes along the filtering pipeline can have a marked impact on the final result. Naturally, the stringency and type of filtering method employed is both platform dependent and sensitive to the library building methodology.

The difficulty of analysing the diversity of samples whilst accounting for sequence quality, abundance and attempting a taxonomy-independent measure of analysis is illustrated in [Fig 4](#). Depending on the quality filtering method (QFM), abundance filtering method (AFM) and taxonomy-independent method (TIM) used ([Figs 1 and 4](#)) the number of taxonomic units detected varied between 3 and 22 operational taxonomic units (OTUs) or between 3 and 14 distance-based operational taxonomic units (DTUs) [24] ([Fig 4](#)). In each case the minimum

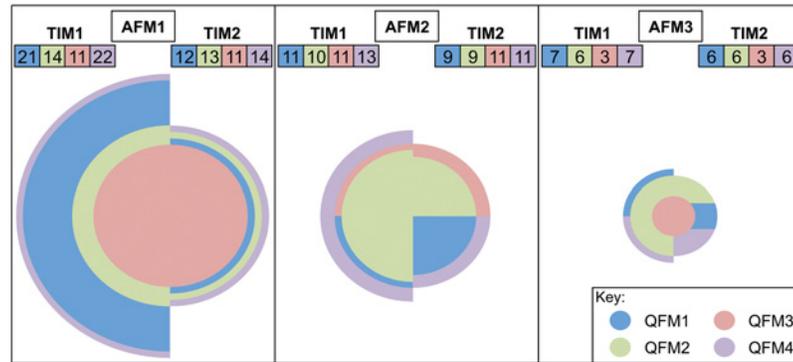


Fig 4. Impact of analysis parameters on the numbers of taxonomic units obtained for a bulk-bone sample. A number of analysis parameters were used to analyse a complex mixture containing numerous taxa. Different quality and abundance filtering methods were used in addition to two taxonomy-independent measures of analysis, full definitions and explanations of which are in Fig 1. The spread in the numbers of taxonomic units obtained across the combinations of parameters chosen is seen. The radius of each semicircle represents the number of taxonomic units obtained given a set combination of the parameters used. The number of taxonomic units is also indicated above each semicircle. Each semicircle is proportional to all others. AFM—abundance filtering method; QFM—quality filtering method; TIM—taxonomy-independent method.

doi:10.1371/journal.pone.0124671.g004

average Quality Scores (Q-Scores) for all sequences post-filtering were well above the standard cut-off of Q15. Tellingly however, when considering QFM1 and QFM4 (see Fig 1 for definitions and also S4 Table) where individual bases below Q15 were permissible, a sizeable proportion of sequences contained bases below Q15 (57.0% and 42.3% respectively) and there was a noticeable percentage of bases below Q15 overall (2.6% and 0.9% respectively) (S4 Table).

The use of Phred Q-Scores, as noted above, is one means by which to filter sequence data for error. Many papers, including those by the authors, make mention of how the data contained within has been filtered for quality, however, few make mention of how this is done thus making it difficult to reproduce data from the pipeline used. It is an open question as to what truly constitutes a high quality sequence. For instance, is it one where the average Q-Score across its length is >Q20 or should it be a requirement that all bases within the sequence be at least Q15? Q-scores are also complicated by the fact that different platforms use different methods when generating Q-scores. An issue surrounding the use of a stringent Q-Score cut-off that all bases must meet is the fact that the Q-Score of a base is impacted by the Q-Scores of the bases immediately surrounding it. Homopolymers are generally areas of quite low quality and this low quality can extend for a number of bases beyond the homopolymer stretch itself. In an extreme example, a Q-score based filtering method might actively discard amplicon variants that contain homopolymer stretches in favour of those that do not, thereby warping the composition of the resultant data.

In addition to Q-score cut-offs, filtering of sequence reads below a certain abundance is often employed. This is often cited as an attempt to reduce the possibility of erroneous and artefactual sequences as well as to remove instances of low-level contamination. At times such an approach could be seen as the molecular biology equivalent of “sweeping the dirt under the carpet”—simply moving a baseline until one is happy with the data will ultimately reduce sensitivity and reduce transparency of data fidelity. As with Q-score quality filtering, abundance filtering can be performed in a variety of ways with no accepted definition of what should be classed as a low abundant grouping of sequences. Methods of abundance filtering vary from the removal of singletons only, to the use of, at times, arbitrary cut-offs or inferred cut-offs defining a low abundance cluster (see Fig 1 for examples and Fig 4 for impacts). The choice of an

appropriate abundance filter is no easy task especially in cases where there is unequal sequencing depth that may necessitate the need for sample specific abundance filters.

The fluidity of the definition of a high quality sequence and what constitutes a low abundance cluster as well as the order in which filtering steps are performed (see [Fig 1](#) for examples and [Fig 4](#) for impact) can all combine to create a rather difficult analysis of the diversity of a sample when dealing with heterogeneous samples of unknown composition. This holds true not only when dealing with the abundance of sample constituents but also when dealing with presence and/or absence. These factors are exacerbated further when weighted analyses are employed. In reality there is no means by which to determine the “correct” number of OTUs within a sample. For instance, with regards to a pool of single-source bird samples containing a single sample of only one representative of the family Dromaiidae, *Dromaius novaehollandiae* (Emu), a total of four distinct OTUs were obtained post-filtering (data available from authors upon request). Also worth noting is the importance of ensuring samples are free of inhibition and have sufficient copy number of DNA when conducting OTU analyses that involves a requirement for a particular OTU to occur in a certain proportion of uniquely tagged replicates before it is accepted [94]. If such a criterion were used in the two-fish screening assay ([Fig 2](#)), the genus *Engraulis* would have been excluded at times as it only occurred in a single replicate in certain cases, even though its presence was confirmed using *Engraulis* specific primers. Notwithstanding the above, when used appropriately, OTUs can be a useful index for species diversity provided parameters are both transparent and consistent across samples and studies.

Conclusion

It is proving to be the case in amplicon sequencing that a one-size-fits-all approach is ill-advised and unwise, due to differing budgets, scopes and end-goals. It is therefore not the aim of this article to call for definitive guidelines with regard to best practice when generating amplicon libraries or sequencing them, although a set of flexible reporting guidelines may be appropriate. It is hoped that this paper may instead prove to be a catalyst ultimately aiding in the development of robust amplicon sequencing workflows. The generation of amplicon data is easy, however the generation of high-fidelity data free of contamination, artefacts and appropriately analysed, is far more complex. It is important to be aware of the limitations of amplicon data and know that with the advances afforded by it there are many hurdles. It is imperative that more attention be paid to the processes involved in preparing amplicon libraries to limit some of the pitfalls highlighted in this paper. While published data can be analysed and re-analysed time and again, such as when reference databases improve, the library generation step is not as easily, quickly or cheaply repeated. It is widely acknowledged that amplicon sequencing will continue to play an important role across a wide range of applications. Taken together these data suggest that, in order to get the most out of amplicon datasets, careful attention should be paid to workflows at both benchtop and desktop.

Supporting Information

S1 Fig. Schematics of the steps involved in each of the experiments performed.
(PDF)

S1 File. Single-source bird error output example.
(XLS)

S2 File. DTU calculation example.
(XLS)

S1 Table. Primer information.

(PDF)

S2 Table. Proportion of sequences removed post control filtering.

(PDF)

S3 Table. Cycle threshold value shifts when performing fusion tagged PCR.

(PDF)

S4 Table. Summary quality statistics.

(PDF)

Acknowledgments

The authors acknowledge sequencing assistance from the State Agricultural Biotechnology Centre (SABC, Murdoch University, Perth, Australia), technical expertise and assistance during PGM experiments from staff at the LotteryWest State Biomedical Facility Genomics (Perth, Australia) in addition to the Centre for Comparative Genetics (Murdoch University, Perth, Australia) and iVEC for computational support. We would also like to thank the various members of the TrEnD Lab (Curtin University, Perth, Australia) and Centre for GeoGenetics (University of Copenhagen, Copenhagen, Denmark) for their varied contributions.

Author Contributions

Conceived and designed the experiments: DCM MB. Performed the experiments: DCM MLC. Analyzed the data: DCM. Contributed reagents/materials/analysis tools: MB. Wrote the paper: DCM MB. Edited the manuscript: DCM MB MLC.

References

1. Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, Yuza Y, et al. (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat Med* 12: 852–855. PMID: [16799556](#)
2. Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kær KH, et al. (2011) Meta-barcoding of 'dirt' DNA from soil reflects vertebrate biodiversity. *Molecular Ecology* 21: 1966–1979. doi: [10.1111/j.1365-294X.2011.05261.x](#) PMID: [21917035](#)
3. Baird DJ, Hajibabaei M (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology* 21: 2039–2044. PMID: [22590728](#)
4. Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. *Biology Letters* 4: 423–425. doi: [10.1098/rsbl.2008.0118](#) PMID: [18400683](#)
5. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology* 21: 1794–1805. doi: [10.1111/j.1365-294X.2012.05538.x](#) PMID: [22486820](#)
6. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* 21: 2045–2050. doi: [10.1111/j.1365-294X.2012.05470.x](#) PMID: [22486824](#)
7. Thomsen PF, Kielgast JOS, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP, et al. (2012) Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology* 21: 2565–2573. doi: [10.1111/j.1365-294X.2011.05418.x](#) PMID: [22151771](#)
8. Deagle B, Chiaradia A, McInnes J, Jarman S (2010) Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? *Conservation Genetics* 11: 2039–2048.
9. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology* 21: 1931–1950. doi: [10.1111/j.1365-294X.2011.05403.x](#) PMID: [22171763](#)
10. Burgar JM, Murray DC, Craig MD, Haile J, Houston J, Stokes V, et al. (2014) Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. *Molecular Ecology* 23: 3605–3617. doi: [10.1111/mec.12531](#) PMID: [24118181](#)

11. Bohmann K, Monadjem A, Lehmkuhl Noer C, Rasmussen M, Zeale MRK, Clare E, et al. (2011) Molecular Diet Analysis of Two African Free-Tailed Bats (Molossidae) Using High Throughput Sequencing. *PLoS ONE* 6: e21441. doi: [10.1371/journal.pone.0021441](https://doi.org/10.1371/journal.pone.0021441) PMID: [21731749](https://pubmed.ncbi.nlm.nih.gov/21731749/)
12. Razgour O, Clare EL, Zeale MRK, Hanmer J, Schnell IB, Rasmussen M, et al. (2011) High-throughput sequencing offers insight into mechanisms of resource partitioning in cryptic bat species. *Ecology and Evolution* 1: 556–570. doi: [10.1002/ece3.49](https://doi.org/10.1002/ece3.49) PMID: [22393522](https://pubmed.ncbi.nlm.nih.gov/22393522/)
13. Quéméré E, Hibert F, Miquel C, Lhuillier E, Rasolondraibe E, Champeau J, et al. (2013) A DNA Metabarcoding Study of a Primate Dietary Diversity and Plasticity across Its Entire Fragmented Range. *PLoS ONE* 8: e58971. doi: [10.1371/journal.pone.0058971](https://doi.org/10.1371/journal.pone.0058971) PMID: [23527060](https://pubmed.ncbi.nlm.nih.gov/23527060/)
14. Ding T, Schloss PD (2014) Dynamics and associations of microbial community types across the human body. *Nature* 509: 357–360. doi: [10.1038/nature13178](https://doi.org/10.1038/nature13178) PMID: [24739969](https://pubmed.ncbi.nlm.nih.gov/24739969/)
15. Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Balum J, Kimbrel J, et al. (2014) Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISME J* 8: 1464–1475. doi: [10.1038/ismej.2013.254](https://doi.org/10.1038/ismej.2013.254) PMID: [24451203](https://pubmed.ncbi.nlm.nih.gov/24451203/)
16. Consortium THMP (2012) Structure, function and diversity of the healthy human microbiome. *Nature* 486: 207–214. doi: [10.1038/nature11234](https://doi.org/10.1038/nature11234) PMID: [22699609](https://pubmed.ncbi.nlm.nih.gov/22699609/)
17. Meadow JF, Bateman AC, Herkert KM, O'Connor TK, Green JL (2013) Significant changes in the skin microbiome mediated by the sport of roller derby. *PeerJ* 1: e53. doi: [10.7717/peerj.53](https://doi.org/10.7717/peerj.53) PMID: [23638391](https://pubmed.ncbi.nlm.nih.gov/23638391/)
18. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R (2010) Forensic identification using skin bacterial communities. *Proceedings of the National Academy of Sciences* 107: 6477–6481. doi: [10.1073/pnas.1000162107](https://doi.org/10.1073/pnas.1000162107) PMID: [20231444](https://pubmed.ncbi.nlm.nih.gov/20231444/)
19. Sun B, Wang F, Jiang Y, Li Y, Dong Z, Li Z, et al. (2014) A long-term field experiment of soil transplantation demonstrating the role of contemporary geographic separation in shaping soil microbial community structure. *Ecology and Evolution* 4: 1073–1087. doi: [10.1002/ece3.1006](https://doi.org/10.1002/ece3.1006) PMID: [24772284](https://pubmed.ncbi.nlm.nih.gov/24772284/)
20. Liu J, Sui Y, Yu Z, Shi Y, Chu H, Jin J, et al. (2014) High throughput sequencing analysis of biogeographical distribution of bacterial communities in the black soils of northeast China. *Soil Biology and Biochemistry* 70: 113–122.
21. Clarke LJ, Czechowski P, Soubrier J, Stevens MI, Cooper A (2014) Modular tagging of amplicons using a single PCR for high-throughput sequencing. *Molecular Ecology Resources* 14: 117–121. doi: [10.1111/1755-0998.12162](https://doi.org/10.1111/1755-0998.12162) PMID: [24028345](https://pubmed.ncbi.nlm.nih.gov/24028345/)
22. Ekblom R, Galindo G (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity* 107: 1–15. doi: [10.1038/hdy.2010.152](https://doi.org/10.1038/hdy.2010.152) PMID: [21139633](https://pubmed.ncbi.nlm.nih.gov/21139633/)
23. Kircher M, Kelso J (2010) High-throughput DNA sequencing—concepts and limitations. *BioEssays* 32: 524–536. doi: [10.1002/bies.200900181](https://doi.org/10.1002/bies.200900181) PMID: [20486139](https://pubmed.ncbi.nlm.nih.gov/20486139/)
24. Murray DC, Haile J, Dortch J, White NE, Haouchar D, Bellgard MI, et al. (2013) Scrapheap Challenge: A novel bulk-bone metabarcoding method to investigate ancient DNA in faunal assemblages. *Sci Rep* 3.
25. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology* 30: 434–439. doi: [10.1038/nbt.2198](https://doi.org/10.1038/nbt.2198) PMID: [22522955](https://pubmed.ncbi.nlm.nih.gov/22522955/)
26. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 13: 341. doi: [10.1186/1471-2164-13-341](https://doi.org/10.1186/1471-2164-13-341) PMID: [22827831](https://pubmed.ncbi.nlm.nih.gov/22827831/)
27. Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C, Nielsen R, et al. (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS ONE* 2: e197. PMID: [17299583](https://pubmed.ncbi.nlm.nih.gov/17299583/)
28. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335–336. doi: [10.1038/nmeth.f.303](https://doi.org/10.1038/nmeth.f.303) PMID: [20383131](https://pubmed.ncbi.nlm.nih.gov/20383131/)
29. Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: Designing and validating sequence identification tags robust to indels. *PLoS ONE* 7: e42543. doi: [10.1371/journal.pone.0042543](https://doi.org/10.1371/journal.pone.0042543) PMID: [22900027](https://pubmed.ncbi.nlm.nih.gov/22900027/)
30. Gonzalez A, Knight R (2012) Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology* 23: 64–71. doi: [10.1016/j.copbio.2011.11.028](https://doi.org/10.1016/j.copbio.2011.11.028) PMID: [22172529](https://pubmed.ncbi.nlm.nih.gov/22172529/)
31. Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *The ISME Journal* 4: 17–27. doi: [10.1038/ismej.2009.97](https://doi.org/10.1038/ismej.2009.97) PMID: [19710709](https://pubmed.ncbi.nlm.nih.gov/19710709/)
32. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Research* 17: 377–386. PMID: [17255551](https://pubmed.ncbi.nlm.nih.gov/17255551/)

33. Quince C, Lanzen A, Davenport R, Turnbaugh P (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* 12: 38. doi: [10.1186/1471-2105-12-38](https://doi.org/10.1186/1471-2105-12-38) PMID: [21276213](https://pubmed.ncbi.nlm.nih.gov/21276213/)
34. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. (2009) The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry* 55: 611–622. doi: [10.1373/clinchem.2008.112797](https://doi.org/10.1373/clinchem.2008.112797) PMID: [19246619](https://pubmed.ncbi.nlm.nih.gov/19246619/)
35. Murray DC, Bunce M, Cannell BL, Oliver R, Houston J, White NE, et al. (2011) DNA-based faecal dietary analysis: A comparison of qPCR and High Throughput Sequencing approaches. *PLoS One* 6: e25776. doi: [10.1371/journal.pone.0025776](https://doi.org/10.1371/journal.pone.0025776) PMID: [21998697](https://pubmed.ncbi.nlm.nih.gov/21998697/)
36. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, et al. (2012) Deep Sequencing of Plant and Animal DNA Contained within Traditional Chinese Medicines Reveals Legality Issues and Health Safety Concerns. *PLoS Genet* 8: e1002657. doi: [10.1371/journal.pgen.1002657](https://doi.org/10.1371/journal.pgen.1002657) PMID: [22511890](https://pubmed.ncbi.nlm.nih.gov/22511890/)
37. Tridico SR, Murray DC, Addison J, Kirkbride KP, Bunce M (2014) The Application of Metagenomic Analyses of Human Hair Shafts in Forensic Investigations using Next Generation Sequencing: A qualitative assessment. *Investigative Genetics* 5: 16. doi: [10.1186/s13323-014-0016-5](https://doi.org/10.1186/s13323-014-0016-5) PMID: [25516795](https://pubmed.ncbi.nlm.nih.gov/25516795/)
38. White NE, Bunce M, Mawson PR, Dawson R, Saunders DA, Allentoft ME (2014) Identifying conservation units after large-scale land clearing: a spatio-temporal molecular survey of endangered white-tailed black cockatoos (*Calyptorhynchus* spp.). *Diversity and Distributions* 20: 1208–1220.
39. Bunce M, Oskam C, Allentoft M (2011) The use of quantitative real-time PCR in ancient DNA research. In: Shapiro B, Hofreiter M, editors. *Ancient DNA: Methods and Protocols*: Humana Press. pp. 121–132.
40. Roche (2009) Technical Bulletin: Amplicon fusion primer design guidelines for GS FLX Titanium series Lib-A chemistry. TCB No. 013–2009: 1–3. doi: [10.1038/ajg.2008.97](https://doi.org/10.1038/ajg.2008.97) PMID: [19262518](https://pubmed.ncbi.nlm.nih.gov/19262518/)
41. Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, et al. (2011) Geneious v7.1, created by Biomatters. Available: <http://www.geneious.com/>.
42. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461. doi: [10.1093/bioinformatics/btq461](https://doi.org/10.1093/bioinformatics/btq461) PMID: [20709691](https://pubmed.ncbi.nlm.nih.gov/20709691/)
43. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* 10: 996–998.
44. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194–2200. doi: [10.1093/bioinformatics/btr381](https://doi.org/10.1093/bioinformatics/btr381) PMID: [21700674](https://pubmed.ncbi.nlm.nih.gov/21700674/)
45. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. *Nucleic Acids Research* 34: D16–D20. PMID: [16381837](https://pubmed.ncbi.nlm.nih.gov/16381837/)
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410. PMID: [2231712](https://pubmed.ncbi.nlm.nih.gov/2231712/)
47. Hunter AA, Macgregor AB, Szabo TO, Wellington CA, Bellgard MI (2012) Yabi: An online research environment for grid, high performance and cloud computing. *Source Code for Biology and Medicine* 7: 1. doi: [10.1186/1751-0473-7-1](https://doi.org/10.1186/1751-0473-7-1) PMID: [22333270](https://pubmed.ncbi.nlm.nih.gov/22333270/)
48. Deagle BE, Gales NJ, Evans K, Jarman SN, Robinson S, Trebilco R, et al. (2007) Studying Seabird Diet through Genetic Analysis of Faeces: A Case Study on Macaroni Penguins (*Eudyptes chrysolophus*). *PLoS ONE* 2: e831. PMID: [17786203](https://pubmed.ncbi.nlm.nih.gov/17786203/)
49. Cooper A, Lalueza-Fox C, Anderson S, Rambaut A, Austin J, Ward R (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409: 704–707. PMID: [11217857](https://pubmed.ncbi.nlm.nih.gov/11217857/)
50. Cooper A (1994) DNA from Museum Specimens. In: Herrmann B, Hummel S, editors. *Ancient DNA*: Springer New York. pp. 149–165.
51. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797. PMID: [15034147](https://pubmed.ncbi.nlm.nih.gov/15034147/)
52. Turner S, Pryer KM, Miao VP, Palmer JD (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *J Eukaryot Microbiol* 46: 327–338. PMID: [10461381](https://pubmed.ncbi.nlm.nih.gov/10461381/)
53. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. (2011) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* 108: 4516–4522. doi: [10.1073/pnas.1000080107](https://doi.org/10.1073/pnas.1000080107) PMID: [20534432](https://pubmed.ncbi.nlm.nih.gov/20534432/)
54. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, et al. (2007) Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* 35: e14. PMID: [17169982](https://pubmed.ncbi.nlm.nih.gov/17169982/)
55. Taylor PG (1996) Reproducibility of ancient DNA sequences from extinct Pleistocene fauna. *Molecular Biology and Evolution* 13: 283–285. PMID: [8583902](https://pubmed.ncbi.nlm.nih.gov/8583902/)

56. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. (2005) Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* 15: 1451–1455. PMID: [16169926](#)
57. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11: R86–R86. doi: [10.1186/gb-2010-11-8-r86](#) PMID: [20738864](#)
58. Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, et al. (2001) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*: John Wiley & Sons, Inc.
59. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30: 3059–3066. PMID: [12136088](#)
60. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731–2739. doi: [10.1093/molbev/msr121](#) PMID: [21546353](#)
61. Ficetola GF, Pansu J, Bonin A, Coissac E, Giguët-Covex C, De Barba M, et al. (2014) Replication levels, false presences, and the estimation of presence / absence from eDNA metabarcoding data. *Molecular Ecology Resources*: n/a-n/a.
62. Robasky K, Lewis NE, Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat Rev Genet* 15: 56–62. doi: [10.1038/nrg3655](#) PMID: [24322726](#)
63. Smith DP, Peay KG (2014) Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing. *PLoS ONE* 9: e90234. doi: [10.1371/journal.pone.0090234](#) PMID: [24587293](#)
64. Schloss PD, Gevers D, Westcott SL (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS ONE* 6: e27310. doi: [10.1371/journal.pone.0027310](#) PMID: [22194782](#)
65. Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR, et al. (2012) Groundtruthing Next-Gen Sequencing for Microbial Ecology—Biases and Errors in Community Structure Estimates from PCR Amplicon Pyrosequencing. *PLoS ONE* 7: e44224. doi: [10.1371/journal.pone.0044224](#) PMID: [22970184](#)
66. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12: R18. doi: [10.1186/gb-2011-12-2-r18](#) PMID: [21338519](#)
67. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*.
68. Chen Y-C, Liu T, Yu C-H, Chiang T-Y, Hwang C-C (2013) Effects of GC Bias in Next-Generation-Sequencing Data on *De Novo* Genome Assembly. *PLoS ONE* 8: e62856. doi: [10.1371/journal.pone.0062856](#) PMID: [23638157](#)
69. Ross M, Russ C, Costello M, Hollinger A, Lennon N, Hegarty R, et al. (2013) Characterizing and measuring bias in sequence data. *Genome Biology* 14: R51. doi: [10.1186/gb-2013-14-5-r51](#) PMID: [23718773](#)
70. Dabney J, Meyer M (2012) Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* 52: 87–94. doi: [10.2144/000113809](#) PMID: [22313406](#)
71. Ficetola G, Coissac E, Zundel S, Riaz T, Shehzad W, Bessière J, et al. (2010) An In silico approach for the evaluation of DNA barcodes. *BMC Genomics* 11: 1–10. doi: [10.1186/1471-2164-11-1](#) PMID: [20044946](#)
72. Murray DC, Pearson SG, Fullagar R, Chase BM, Houston J, Atchison J, et al. (2012) High-throughput sequencing of ancient plant and mammal DNA preserved in herbivore middens. *Quaternary Science Reviews* 58: 135–145.
73. Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources* 12: 377–388. doi: [10.1111/j.1755-0998.2012.03119.x](#) PMID: [22356472](#)
74. Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters* 10.
75. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, et al. (2011) Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biol Evol* 3: 1312–1323. doi: [10.1093/gbe/evr106](#) PMID: [22002916](#)
76. de Cárcer DA, Denman SE, McSweeney C, Morrison M (2011) Strategy for Modular Tagged High-Throughput Amplicon Sequencing. *Applied and Environmental Microbiology* 77: 6310–6312. doi: [10.1128/AEM.05146-11](#) PMID: [21764953](#)

77. Bronner IF, Quail MA, Turner DJ, Swerdlow H (2001) Improved Protocols for Illumina Sequencing. *Current Protocols in Human Genetics*: John Wiley & Sons, Inc.
78. Archer J, Weber J, Henry K, Winner D, Gibson R, Lee L, et al. (2012) Use of Four Next-Generation Sequencing Platforms to Determine HIV-1 Coreceptor Tropism. *PLOS ONE* 7: e49602. doi: [10.1371/journal.pone.0049602](https://doi.org/10.1371/journal.pone.0049602) PMID: [23166726](https://pubmed.ncbi.nlm.nih.gov/23166726/)
79. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, et al. (2014) Next-generation sequencing reveals large connected networks of intra-host HCV variants. *BMC Genomics* 15 Suppl 5: S4. doi: [10.1186/1471-2164-15-S5-S4](https://doi.org/10.1186/1471-2164-15-S5-S4) PMID: [25081811](https://pubmed.ncbi.nlm.nih.gov/25081811/)
80. Varley KE, Mitra RD (2008) Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res* 18: 1844–1850. doi: [10.1101/gr.078204.108](https://doi.org/10.1101/gr.078204.108) PMID: [18849522](https://pubmed.ncbi.nlm.nih.gov/18849522/)
81. Budowle B, Connell N, Bielecka-Oder A, Colwell R, Corbett C, Fletcher J, et al. (2014) Validation of high throughput sequencing and microbial forensics applications. *Investigative Genetics* 5: 1–18. doi: [10.1186/2041-2223-5-1](https://doi.org/10.1186/2041-2223-5-1) PMID: [24386986](https://pubmed.ncbi.nlm.nih.gov/24386986/)
82. Sajantila A (2015) Editors' Pick: Contamination has always been the issue! *Investigative Genetics* 5: 17.
83. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, et al. (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. *Mol Ecol Resour* 14: 306–323. doi: [10.1111/1755-0998.12188](https://doi.org/10.1111/1755-0998.12188) PMID: [24128180](https://pubmed.ncbi.nlm.nih.gov/24128180/)
84. Porter TM, Golding GB, King C, Froese D, Zazula G, Poinar HN (2013) Amplicon pyrosequencing late Pleistocene permafrost: the removal of putative contaminant sequences and small-scale reproducibility. *Mol Ecol Resour* 13: 798–810. doi: [10.1111/1755-0998.12124](https://doi.org/10.1111/1755-0998.12124) PMID: [23694692](https://pubmed.ncbi.nlm.nih.gov/23694692/)
85. Salter S, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. (2014) Reagent contamination can critically impact sequence-based microbiome analyses. *bioRxiv*.
86. Sonstebo JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, et al. (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. *Mol Ecol Resour* 10: 1009–1018. doi: [10.1111/j.1755-0998.2010.02855.x](https://doi.org/10.1111/j.1755-0998.2010.02855.x) PMID: [21565110](https://pubmed.ncbi.nlm.nih.gov/21565110/)
87. Lundin S, Stranneheim H, Pettersson E, Klevebring D, Lundeberg J (2010) Increased Throughput by Parallelization of Library Preparation for Massive Sequencing. *PLoS ONE* 5: e10029. doi: [10.1371/journal.pone.0010029](https://doi.org/10.1371/journal.pone.0010029) PMID: [20386591](https://pubmed.ncbi.nlm.nih.gov/20386591/)
88. DeAngelis MM, Wang DG, Hawkins TL (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research* 23: 4742–4743. PMID: [8524672](https://pubmed.ncbi.nlm.nih.gov/8524672/)
89. Borgström E, Lundin S, Lundeberg J (2011) Large Scale Library Generation for High Throughput Sequencing. *PLoS ONE* 6: e19119. doi: [10.1371/journal.pone.0019119](https://doi.org/10.1371/journal.pone.0019119) PMID: [21589638](https://pubmed.ncbi.nlm.nih.gov/21589638/)
90. Quail MA, Gu Y, Swerdlow H, Mayho M (2012) Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. *Electrophoresis* 33: 3521–3528. doi: [10.1002/elps.201200128](https://doi.org/10.1002/elps.201200128) PMID: [23147856](https://pubmed.ncbi.nlm.nih.gov/23147856/)
91. Pääbo S, Irwin DM, Wilson AC (1990) DNA damage promotes jumping between templates during enzymatic amplification. *Journal of Biological Chemistry* 265: 4718–4721. PMID: [2307682](https://pubmed.ncbi.nlm.nih.gov/2307682/)
92. Piry S, Guivier E, Realini A, Martin JF (2012) |SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization—application to species and environmental barcoding. *Molecular Ecology Resources* 12: 1151–1157. doi: [10.1111/j.1755-0998.2012.03171.x](https://doi.org/10.1111/j.1755-0998.2012.03171.x) PMID: [22823139](https://pubmed.ncbi.nlm.nih.gov/22823139/)
93. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537–7541. doi: [10.1128/AEM.01541-09](https://doi.org/10.1128/AEM.01541-09) PMID: [19801464](https://pubmed.ncbi.nlm.nih.gov/19801464/)
94. Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, et al. (2014) Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* 506: 47–51. doi: [10.1038/nature12921](https://doi.org/10.1038/nature12921) PMID: [24499916](https://pubmed.ncbi.nlm.nih.gov/24499916/)

RESEARCH

Open Access



Inhibition of the endosymbiont “*Candidatus* *Midichloria mitochondrii*” during 16S rRNA gene profiling reveals potential pathogens in *Ixodes* ticks from Australia

Alexander W. Gofton¹, Charlotte L. Oskam¹, Nathan Lo², Tiziana Beninati³, Heng Wei², Victoria McCarl², Dáithí C. Murray⁴, Andrea Paparini¹, Telleasha L. Greay¹, Andrew J. Holmes⁵, Michael Bunce⁴, Una Ryan¹ and Peter Irwin^{1*}

Abstract

Background: The Australian paralysis tick (*Ixodes holocyclus*) is of significant medical and veterinary importance as a cause of dermatological and neurological disease, yet there is currently limited information about the bacterial communities harboured by these ticks and the risk of infectious disease transmission to humans and domestic animals. Ongoing controversy about the presence of *Borrelia burgdorferi* sensu lato (the aetiological agent of Lyme disease) in Australia increases the need to accurately identify and characterise bacteria harboured by *I. holocyclus* ticks.

Methods: Universal PCR primers were used to amplify the V1-2 hyper-variable region of bacterial 16S rRNA genes present in DNA samples from *I. holocyclus* and *I. ricinus* ticks, collected in Australia and Germany respectively. The 16S amplicons were purified, sequenced on the Ion Torrent platform, and analysed in USEARCH, QIIME, and BLAST to assign genus and species-level taxonomy. Initial analysis of *I. holocyclus* and *I. ricinus* identified that > 95 % of the 16S sequences recovered belonged to the tick intracellular endosymbiont “*Candidatus* *Midichloria mitochondrii*” (CMM). A CMM-specific blocking primer was designed that decreased CMM sequences by approximately 96 % in both tick species and significantly increased the total detectable bacterial diversity, allowing identification of medically important bacterial pathogens that were previously masked by CMM.

Results: *Borrelia burgdorferi* sensu lato was identified in German *I. ricinus*, but not in Australian *I. holocyclus* ticks. However, bacteria of medical significance were detected in *I. holocyclus* ticks, including a *Borrelia* relapsing fever group sp., *Bartonella henselae*, novel “*Candidatus* *Neoehrlichia*” spp., *Clostridium histolyticum*, *Rickettsia* spp., and *Leptospira inadai*.

Conclusions: Abundant bacterial endosymbionts, such as CMM, limit the effectiveness of next-generation 16S bacterial community profiling in arthropods by masking less abundant bacteria, including pathogens. Specific blocking primers that inhibit endosymbiont 16S amplification during PCR are an effective way of reducing this limitation. Here, this strategy provided the first evidence of a relapsing fever *Borrelia* sp. and of novel “*Candidatus* *Neoehrlichia*” spp. in Australia. Our results raise new questions about tick-borne pathogens in *I. holocyclus* ticks.

Keywords: Tick, Vector-borne disease, Zoonoses, Metagenomics, 16S community profiling, *Ixodes holocyclus*, *Ixodes ricinus*, *Candidatus* *Midichloria*, *Borrelia*, *Candidatus* *Neoehrlichia*

* Correspondence: p.irwin@murdoch.edu.au

¹Vector and Water-Borne Pathogen Research Laboratory, School of Veterinary and Life Sciences, Murdoch University, Perth, Western Australia, Australia
Full list of author information is available at the end of the article

SCIENTIFIC REPORTS



OPEN

Combined DNA, toxicological and heavy metal analyses provides an auditing toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM)

Received: 11 August 2015
Accepted: 30 October 2015
Published: 10 December 2015

Megan L. Coghlan¹, Garth Maker^{2,3}, Elly Crighton^{2,3}, James Haile¹, Dáithí C. Murray¹, Nicole E. White¹, Roger W. Byard^{4,5}, Matthew I. Bellgard⁶, Ian Mullaney^{2,3}, Robert Trengove², Richard J.N. Allcock^{7,8}, Christine Nash⁵, Claire Hoban⁴, Kevin Jarrett⁹, Ross Edwards⁹, Ian F. Musgrave⁴ & Michael Bunce¹

Globally, there has been an increase in the use of herbal remedies including traditional Chinese medicine (TCM). There is a perception that products are natural, safe and effectively regulated, however, regulatory agencies are hampered by a lack of a toolkit to audit ingredient lists, adulterants and constituent active compounds. Here, for the first time, a multidisciplinary approach to assessing the molecular content of 26 TCMs is described. Next generation DNA sequencing is combined with toxicological and heavy metal screening by separation techniques and mass spectrometry (MS) to provide a comprehensive audit. Genetic analysis revealed that 50% of samples contained DNA of undeclared plant or animal taxa, including an endangered species of *Panthera* (snow leopard). In 50% of the TCMs, an undeclared pharmaceutical agent was detected including warfarin, dexamethasone, diclofenac, cyproheptadine and paracetamol. Mass spectrometry revealed heavy metals including arsenic, lead and cadmium, one with a level of arsenic >10 times the acceptable limit. The study showed 92% of the TCMs examined were found to have some form of contamination and/or substitution. This study demonstrates that a combination of molecular methodologies can provide an effective means by which to audit complementary and alternative medicines.

The use of complementary and alternative medicine is becoming increasingly popular worldwide^{1,2}. It is generally believed that since herbal remedies are of natural origin, they are therefore safe, that there

¹Trace and Environmental DNA laboratory, Department of Environment and Agriculture, Curtin University, Kent St, Bentley, WA, 6102, Australia. ²Separation Science and Metabolomics Laboratory and the Advanced Mass Spectrometry Facility, Murdoch University, South St, Murdoch, WA, 6150, Australia. ³School of Veterinary and Life Sciences, Murdoch University, South St, Murdoch, WA, 6150, Australia. ⁴School of Medical Sciences, The University of Adelaide, Frome Rd, Adelaide, SA, 5005, Australia. ⁵Forensic Science SA, Adelaide, SA, 5000, Australia. ⁶Centre for Comparative Genomics, Murdoch University, South St, Murdoch, WA, 6150, Australia. ⁷LotteryWest State Biomedical Facility Genomics, School of Pathology and Laboratory Medicine, University of Western Australia, 35 Stirling Hwy, Crawley, WA, 6009, Australia. ⁸Department of Diagnostic Genomics, Pathwest Laboratory Medicine WA, QEII Medical Centre, Hospital Ave, Nedlands, WA, 6009, Australia. ⁹Trace Research Advanced Clean Environment (TRACE) Facility, Department of Physics, Astronomy and Medical Radiation Sciences, Curtin University, Kent St, Bentley, WA, 6102, Australia. Correspondence and requests for materials should be addressed to M.B. (email: michael.bunce@curtin.edu.au)



Contents lists available at ScienceDirect

Quaternary Science Reviews

journal homepage: www.elsevier.com/locate/quascirev

A critical evaluation of how ancient DNA bulk bone metabarcoding complements traditional morphological analysis of fossil assemblages



Alicia C. Grealy^{a, b}, Matthew C. McDowell^c, Paul Scofield^d, Dáithí C. Murray^{a, b},
Diana A. Fusco^c, James Haile^{a, b}, Gavin J. Prideaux^c, Michael Bunce^{a, b, *}

^a Trace and Environmental DNA (TrEnd) Laboratory, Department of Environment and Agriculture, Curtin University, Kent St, Bentley, WA 6102, Australia

^b Ancient DNA Laboratory, School of Veterinary and Life Sciences, Murdoch University, South St, Murdoch, WA 6150, Australia

^c School of Biological Sciences, Flinders University, Bedford Park, SA 5042, Australia

^d Canterbury Museum, Christchurch 8013, New Zealand

ARTICLE INFO

Article history:

Received 22 April 2015

Received in revised form

6 September 2015

Accepted 11 September 2015

Available online 29 September 2015

Keywords:

aDNA

ancient DNA

Archaeology

Biodiversity

Bulk bone

Experimental error

Fossil

Metabarcoding

Next-generation sequencing

Palaeontology

ABSTRACT

When pooled for extraction as a bulk sample, the DNA within morphologically unidentifiable fossil bones can, using next-generation sequencing, yield valuable taxonomic data. This method has been proposed as a means to rapidly and cost-effectively assess general ancient DNA preservation at a site, and to investigate temporal and spatial changes in biodiversity; however, several caveats have yet to be considered. We critically evaluated the bulk bone metabarcoding (BBM) method in terms of its: (i) repeatability, by quantifying sampling and technical variance through a nested experimental design containing subsamples and replicates at several stages; (ii) accuracy, by comparing morphological and molecular family-level identifications; and (iii) overall utility, by applying the approach to two independent Holocene fossil deposits, Bat Cave (Kangaroo Island, Australia) and Finsch's Folly (Canterbury, New Zealand). For both sites, bone and bone powder sub-sampling were found to contribute significantly to variance in molecularly identified family assemblage, while the contribution of library preparation and sequencing was almost negligible. Nevertheless, total variance was small. Sampling over 80% fewer bones than was required to morphologically identify the taxonomic assemblages, we found that the families identified molecularly are a subset of the families identified morphologically and, for the most part, represent the most abundant families in the fossil record. In addition, we detected a range of extinct, extant and endangered taxa, including some that are rare in the fossil record. Given the relatively low sampling effort of the BBM approach compared with morphological approaches, these results suggest that BBM is largely consistent, accurate, sensitive, and therefore widely applicable. Furthermore, we assessed the overall benefits and caveats of the method, and suggest a workflow for palaeontologists, archaeologists, and geneticists that will help mitigate these caveats. Our results show that DNA analysis of bulk bone samples can be a universally useful tool for studying past biodiversity, when integrated with existing morphology-based approaches. Despite several limitations that remain, the BBM method offers a cost-effective and efficient way of studying fossil assemblages, offering complementary insights into evolution, extinction, and conservation.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

For over a century, the study of fossils has played a major role in understanding prehistoric life and evolutionary processes. In

particular, morphological analyses of fossils can reveal species that existed in the past, help elucidate the evolutionary relationships of extinct and extant species (e.g., Donoghue et al., 1989; Deméré et al., 2005; Manos et al., 2007), and assist the development of palaeoenvironment reconstructions that provide insights into the evolutionary and ecological impacts of environmental changes (e.g., Rodríguez-Aranda and Calvo, 1998; Zhang et al., 2008). However, such traditional methods have limitations. For instance, taxonomic assignments of fossils have been necessarily reliant on morphological distinctions, making the identification of

* Corresponding author. Trace and Environmental DNA (TrEnd) Laboratory, Department of Environment and Agriculture, Curtin University, Kent St, Bentley, WA 6102, Australia.

E-mail address: michael.bunce@curtin.edu.au (M. Bunce).

Comparison of morphological and DNA metabarcoding analyses of diets in exploited marine fishes

Oliver Berry^{1,*}, Cathy Bulman², Michael Bunce³, Megan Coghlan³,
Dáithí C. Murray³, Robert D. Ward²

¹CSIRO Oceans and Atmosphere Flagship, PMB 5, Wembley, Western Australia 6913, Australia

²CSIRO Oceans and Atmosphere Flagship, GPO Box 1538, Hobart, Tasmania 7001, Australia

³Trace and Environmental DNA (TrEnD) Laboratory, Department of Environment and Agriculture, Curtin University, Bentley, Western Australia 6102, Australia

ABSTRACT: Ecosystem-based management (EBM) is a framework for managing marine resources. EBM strategies can be evaluated with ecosystem models that represent functional components of ecosystems, including anthropogenic factors. Foodwebs are at the core of ecosystem models, but because dietary data can be difficult to obtain, they are often coarsely characterised. High-throughput DNA sequencing (HTS) of diets is a rapid way to parameterise foodwebs at enhanced taxonomic resolution, and potentially, to optimise the functioning of ecosystem models. We evaluated the relative merits of microscopic and HTS analyses of the diets of 8 fish species harvested in Australia's most intensive fishery, viz. the southeast trawl fishery. We compare the taxonomic resolution and phylogenetic breadth of diets yielded by these methods and include a comparison of 3 DNA barcoding markers (mtDNA COX1 Minibar, mtDNA 16S Chord-cephA, nDNA 18S Bilateria). Using paired individual gut samples (n = 151), we demonstrate that HTS typically identified similar taxon richness but at significantly higher taxonomic resolution than microscopy. However, DNA barcode selection significantly affected both the resolution and phylogenetic breadth of estimated diets. Both COX1 Minibar and 16S Chord-cephA barcodes provided higher taxonomic resolution than morphological analysis, but the resolution varied between taxonomic groups primarily due to availabilities of reference data. However, neither barcode recovered the full dietary spectrum revealed by the 18S Bilateria barcode. HTS also revealed the presence of dietary items not previously recorded for target species, and diverse parasite assemblages. We conclude that HTS has the potential to improve structure and function of ecosystem models and to facilitate best-practice EBM.

KEY WORDS: Foodweb · Ecosystem-based management · DNA sequencing · Fisheries management · Ecosystem modelling

— Resale or republication not permitted without written consent of the publisher —

INTRODUCTION

Increasingly, fisheries managers are adopting the principles of ecosystem-based management (EBM). This provides a holistic approach to managing marine resources that contrasts with the more conventional focal-species approach to management (Pikitch

et al. 2004). As a result, tools to monitor the complex and diverse interactions between fisheries and the environment are more in demand (Levin et al. 2009). Ecosystem modelling is among the most important emerging tools for understanding ecosystem dynamics and highlighting major knowledge gaps, and for evaluating EBM strategies (Fulton et al. 2011). Eco-

*Corresponding author: oliver.berry@csiro.au