Received Date : 22-Nov-2015

Revised Date : 05-Jul-2016

Accepted Date : 19-Jul-2016

Article type : From the Cover

Molecular Ecology Resources

'From the Cover' submission

Draft genome of an iconic Red Sea reef fish, the blacktail butterflyfish (*Chaetodon austriacus*): current status and its characteristics

Running title: **Draft genome of the blacktail butterflyfish**

Joseph D. DiBattista[1,2*], Xin Wang[1], Pablo Saenz-Agudelo[1,3], Marek J. Piatek[4], Manuel Aranda[1], and Michael L Berumen[1]

[1]Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia, [2]Department of Environment and Agriculture, Curtin University, PO Box U1987, Perth, WA 6845, Australia, [3]Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Valdivia 5090000, Chile, [4]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

*Correspondance: Joseph DiBattista, Department of Environment and Agriculture, Curtin University, PO Box U1987, Perth, WA 6845, Australia

Phone: +61 478 514 097
E-mail: josephdibattista@gmail.com

Keywords: adaptation, aquarium trade, bony fish, endemism, genomics, Indo-Pacific

Abstract

Butterflyfish are among the most iconic of the coral reef fishes and represent a model system to study general questions of biogeography, evolution, and population genetics. We assembled and annotated the genome sequence of the blacktail butterflyfish (*Chaetodon austriacus*), an Arabian region endemic species that is reliant on coral reefs for food and shelter. Using available bony fish (superclass Osteichthyes) genomes as a reference, a total of 28,926 high-quality protein-coding genes were predicted from 13,967 assembled scaffolds. The quality and completeness of the draft genome of *C. austriacus* suggests that it has the potential to serve as a resource for studies on the co-evolution of reef fish adaptations to the unique Red Sea environment, as well as a comparison of gene sequences between closely related congeneric species of butterflyfish distributed more broadly across the tropical Indo-Pacific.

Introduction

Genetic research is rapidly moving from simple to increasingly complex systems, with great benefit for multiple fields of research. Genome sequences from the bony fishes (superclass Osteichthyes), which represent more than 50% of all known vertebrate species, were first obtained for well-studied model systems related to medical research (zebrafish, *Danio rerio*; Howe *et al.* 2013), aquaculture (common carp, *Cyprinus carpio*; Xu *et al.* 2014), adaptive radiations (family Cichlidae; Loh *et al.* 2008), or evolutionary novelties (coelacanth, genus *Latimeria*; Amemiya *et al.* 2014). Genome-scale resources for non-model organisms, particularly reef fishes, are still lacking, despite significant advances in next-generation sequencing (NGS) technologies (for review see Kosuri & Church 2014) and bioinformatic approaches (Chen 2015; Xiao *et al.* 2015).

There are at least 5,000 true reef fish species in tropical seas with many diverse traits and

adaptations, but thus far a cold water pufferfish from the northwest Pacific Ocean (*Takifugu rubripes*; van de Peer 2004), a brackish water pufferfish from the rivers of Sri Lanka to China (*Tetraodon nigroviridis*; van de Peer 2004), a pelagic predator in the northern Pacific Ocean (bluefin tuna, *Thunnus orientalis*; Nakamura *et al.* 2013), and the world's largest fish (whale shark, *Rhincodon typus*; Read *et al.* 2015) are the closest genomic resources (also see Spaink *et al.* 2014). This deficiency does not allow us to identify genes under selection or elucidate the mechanisms underlying the adaptations of some of these fish to coral reefs.

The conspicuous butterflyfish and bannerfish (family Chaetodontidae) in particular, a speciose group (> 130 species) of reef fish often targeted by the ornamental trade (Wabnitz 2003; Lawton *et al.* 2013), display a tremendous diversity of color pattern, body shape, feeding behavior, and ecological niche selection (Roberts & Ormond 1992; Zekeria *et al.* 2002; Littlewood *et al.* 2004; Pratchett 2005; Graham 2007; Cole *et al.* 2008; Gregson *et al.* 2008). Indeed, of the fish species that feed directly on scleractinian corals as their primary source of nutrition, 61% are butterflyfishes (Cole *et al.* 2008; Rotjan & Lewis 2008), with the next closest being the wrasses (family: Labridae) (20%; Bellwood *et al.* 2010). Molecular analyses indicate that the butterflyfish family diverged from its nearest relative, the angelfish family (Pomacanthidae), about 33 million years ago (Bellwood *et al.* 2010). Due to the diversity of specializations in the butterflyfishes, this family should serve as an ideal model for genome-based phylogenetic comparisons among the specialized reef fishes.

From a biogeographic perspective, butterflyfishes are prominent in fish assemblages of one of the most unique and geologically complex regions of the world's oceans, the Red Sea. This semi-enclosed basin at the northwestern boundary of the Indian Ocean is subject to minimal freshwater inflow, high rates of evaporation, and latitudinal gradients in environmental variables (Sofianos 2003; Ngugi *et al.* 2012; Raitsos *et al.* 2013). The Red Sea also harbors one of the highest levels of endemism for marine organisms, with 12.9% for

fishes, 12.6% for polychaetes, 8.1% for echinoderms, 16.5% for ascidians, and 5.8% for scleractinian corals (DiBattista *et al.* 2016a). Butterflyfishes are particularly interesting given that 6 of the 12 species recorded from the Red Sea were at one time described as endemic (Roberts *et al.* 1992), although some of these species are now known to extend their range into the adjacent Gulf of Aden or as far as the Gulf of Oman (DiBattista *et al.* 2016a).

The origin of high levels of endemism in Red Sea reef fish is still subject to much debate (DiBattista *et al.* 2016b), but is thought to be, at least in part, the result of a narrow (18 km) and shallow (137 m) connection with the Indian Ocean that restricted water flow during Pleistocene glacial cycles (Rohling *et al.* 1998; Siddall *et al.* 2003; Bailey 2009). Even more uncertain is how the separation among extant species is maintained *within* the Red Sea, or what limits dispersal for these relatively sedentary organisms at the larval stage (see DiBattista *et al.* 2016a). Some possible barriers to larval dispersal include cold-water upwelling off the northeast African and southern Arabian coasts (Smeed 1997; Kemp 2000) and a turbid-water region south of 19° to 20° N in the Red Sea (Roberts *et al.* 1992; Nanninga *et al.* 2014; Giles *et al.* 2015). The importance of these barriers is supported by the disjunctive distribution of some reef fish species (Roberts *et al.* 1992; DiBattista *et al.* 2016a) and coral genera (DiBattista *et al.* 2016a; Sheppard & Sheppard 1991), as well as genetic differentiation between populations of coral reef organisms (Froukh & Kochzius 2008; Nanninga *et al.* 2014; Giles *et al.* 2015; Saenz-Agudelo *et al.* 2015). This area of research would benefit from the application of NGS approaches (e.g. restriction site associated DNA markers [RAD-Seq] for population genomics, ultra-conserved elements [UCEs] for phylogeny, or whole genome sequencing) to shed light on the exact timing and location of congruent isolation events, or to test finer-scale hypotheses related to reef fish connectivity.

Our objective in this study was to use NGS approaches and *de novo* assembly to provide the first draft genome of a Red Sea reef fish, the blacktail butterflyfish (*Chaetodon*

austriacus). *Chaetodon austriacus* Rüppell, 1836 is one of twelve shallow water members of the Chaetodontidae family found in the Red Sea. This iconic species is distributed almost exclusively in the northern and central Red Sea region (DiBattista *et al.* 2016a), and has three closely related sister taxa found together in the *Corallochaetodon* subgenus, *C. trifasciatus*, *C. lunulatus*, and *C. melapterus*, which live in the Indian Ocean, Pacific Ocean, and Arabian Sea, respectively (Waldrop *et al.* 2016), with rare areas of co-occurrence and hybridization (e.g. DiBattista *et al.* 2015). Future genomic comparisons among these closely related fish may provide a means to identify genes related to unique adaptations to the different environments.

Material and Methods

DNA sequencing

A single *C. austriacus* specimen was collected on March 5, 2013, near Ablo Island, Saudi Arabia (N 18.658617°, E 40.826967°). Portions of the gill filaments were removed for genomic library preparation (KAUST Tissue ID: RS3552), and the whole specimen was preserved and archived at the California Academy of Sciences (CAS 236290). Although it is preferred to use an inbred individual for genome sequencing to simplify assembly, an inbred animal was not available. Genomic DNA was extracted from approximately 100 mg of gill tissue and placed into four separate aliquots using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) following the manufacturer's protocol. Total DNA from each extracted aliquot was quantified using a Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA). A total of 10 μg of DNA was provided for library prep and downstream shotgun sequencing, 45 μg was provided for mate-pair library prep.

Genomic DNA libraries for shotgun sequencing were prepared at the King Abdullah University of Science and Technology (KAUST) Bioscience Core Laboratory (BCL) using

an Illumina TruSeq DNA Sample Preparation Kit (Illumina, San Diego, CA). The library was run PE on five lanes of an Illumina HiSeq2000 (v3 reagents) at KAUST. Mate-pair libraries subject to strand displacement of 3-5kb, 5-8kb, and 8-11kb were prepared separately at the KAUST BCL facility and Genotypic Technology genomics facility following a Nextera Mate Pair Gel Plus protocol outlined in "Illumina Nextera Mate Pair library preparation guide (Cat# FC-132-9001DOC, Part#15035209 Rev D.)". The 5-8kb library was run on one lane of PE Illumina HiSeq2000 at KAUST, and all three insert sizes (3-5kb, 5-8kb, and 8-11kb) were run on three lanes of PE Illumina NextSeq500 at Genotypic Technology.

De novo genome assembly

The shotgun sequences (101 bp read length) were quality filtered using SeqQC *vers.* 2.2, which included adapter trimming, B-trimming, and low-quality end trimming functions. A total of 808,641,943 paired-end reads were sequenced and subjected to quality filtering and trimming, resulting in 777,393,511 paired-reads containing 150.99 Gb of data (with a Phred quality score > 20). The mate-pair reads were processed using NextClip tool (Leggett *et al.* 2013) for quality filtering with the following parameters: -t 0, -y 32,17 (soft match), and the remaining parameters were kept as default. After trimming, 33,333,612, 19,362,215, and 11,207,168 read-pairs were retained for the 3-5 kb, 5-8 kb, and 8-11 kb insert libraries, respectively.

Paired-end libraries were screened for mitochondrial sequences after aligning to available fish mitogenomes prior to genome assembly with the short read assembler Abyss *vers.* 1.5.2.; the best assembly was obtained using a kmer size of 63. A further consequence of animal tissue extractions and normalizing the libraries was the increased potential to sequence non-fish transcripts (i.e. contaminants), such as bacteria. These putative bacterial contaminant sequences were identified by performing a BLAST search with the NCBI bacterial genome

sequences dataset. Assembled contigs were joined using SSPACE-BASIC-*vers.* 2.0 (Boetzer *et al.* 2011) with the following parameters: -x 1, -T 20, –g 1. The resulting scaffolds were gap filled using GapCloser *vers.* 1.12 (Luo *et al.* 2012). The NCBI BioProject accession for the genome sequence is PRJNA292048.

Genome annotation - repeat identification and phylogenetic comparison

We identified interspersed repeats and low complexity DNA sequences in the genome using Recon *vers.* 1.0.8 (Bao & Eddy 2002), RepeatScout *vers.* 1.0.5 (Price *et al.* 2005), and PILER *vers.* 1.0 (Edgar & Myers 2005). The output was subsequently used to refine and classify consensus models using Repeatmodeler *vers.* 1.0.8 (http://www.repeatmasker.org/RepeatModeler.html). Identified repeat elements were annotated based on a combination of RepBase *vers.* 20.02 (Bao *et al.* 2015) and Dfam *vers.* 1.4 (Wheeler *et al.* 2013). Repeatmasker *vers.* 4.0.5 (Chen 2004) was also used in conjunction with BLAST and Crossmatch to annotate the identified repeats in the genome assembly. We used a hierarchical system to classify *de novo* repeat elements into the different categories: DNA transposon, LTR, SINEs, LINEs, and simple repeats. Data for the comparative analysis with selected fish species was retrieved from Amemiya *et al.* (2013), Brawand *et al.* (2014), Howe *et al.* (2013), Kasahara *et al.* (2007), and Wang *et al.* (2015). Evolutionary distances for the different species to *C. austriacus* were retrieved from http://www.timetree.org/.

Gene annotation - function annotation

Gene models were predicted using a combined approach of homology-based (SPALN *vers.* 2.1; Iwata & Gotoh 2012) and *ab initio* gene prediction (Augustus *vers.* 3.0.3, Stanke *et al.* 2006; GlimmerHMM *vers.* 3.0.3, Kelley *et al.* 2012). For homology-based gene prediction

we used SPALN and HSP-searcher with default parameters in order to align the UniProt Reference Cluster (UniRef100) and four reference fish proteomes (*Electrophorus electricus* [Gallant *et al.* 2014], *Esox lucius* [Rondeau *et al.* 2014], *Latimeria chalumnae* [Amemiya *et al.* 2013], and *Danio rerio* [Howe *et al.* 2013]) to our assembled genome; this provided clusters of homologous sequences and sub-fragments. Augustus and GlimmerHMM were run using the available (zebrafish) species-specific parameters for *ab initio* gene prediction. The results of the homology-based and *ab initio* gene prediction were integrated using EVM (EvidenceModeler; Haas *et al.* 2008) to form a comprehensive set of consensus gene models.

After removal of gene models with more than two stop codons and < 20 bp in length, we retrieved a total number of 28,926 high quality gene sets (Table S1 and Table S2 in Supporting Information) with an average size of 10,232 bp, and a gene density of approximately 40.6/Mb. We identified 270,127 exons with an average coding sequence length of 143.5 bp; the GC content was 53.7% and 39.81% for exons and introns, respectively. These gene models were then annotated using a three-step approach: 1) all genes were blasted (BlastP) against the Swissprot database; the best hit for each gene model with a minimum cut-off of $< e^{-5}$ was retained whereas gene models with no hit or with hits $> e^{-5}$ were blasted against the TreMBL database, 2) all gene models with no suitable hits against the Swissprot and TreMBL database were blasted against the nr protein database and the best hit with $< e^{-5}$ was retained, and 3) the remaining gene models without annotation were blasted against gene sets of several fish species with fully sequenced genomes in order to identify fish-specific genes that might not be represented in the public databases that we used; these included the species *E. electricus* (Gallant *et al.* 2014), *E. lucius* (Rondeau *et al.* 2014), *L. chalumnae* (Amemiya *et al.* 2013), and *D. rerio* (Howe *et al.* 2013). To provide further functional annotations we used InterProScan (Jones *et al.* 2014) to predict domains based on the Pfam database (Table S1 in Supporting Information). A final validation of the models was

performed by analyzing the distribution of hit lengths of 20,794 gene models with significant hits against highly curated proteins from the SWISS-PROT database and calculating their median coverage (~ 82%; see Fig. S1 in Supporting Information).

Genome validation

In order to evaluate the quality of the assembled genome, we assessed the mapping rates of all libraries using BWA *vers.* 0.7.5 (Li & Durbin 2010) with default parameters. The mapping rate of the short-insert paired-end library, which ranged from 200 to 400 bp, was 96.5% in total (Fig. S2 in Supporting Information). Owing to the different qualities of the mate-pair libraries, the mapping rates had a broad range, but were all substantially higher than 65%. The distribution of mapped read-pair insert sizes on the assembled genome indicated a satisfying accuracy and continuance of our assembly (Fig. S2 in Supporting Information).

We estimated the approximate genome size using a kmer-based approach. Briefly, we used the small insert paired-end library to determine the kmer frequency distribution using different kmer sizes (15, 17, and 19). For each kmer size, we counted the kmer frequency and plotted with a frequency histogram. The distribution should be unimodal, where the peak value is the mean kmer coverage for the data. Based on this value and the actual genome coverage, we applied the formula $M = N * (L – K + 1) / L$, where M is the mean kmer coverage, N is the actual coverage of the genome, L is the mean read length, and K is the kmer size. This approach generated comparable genome size estimates of 650.8 Mb (kmer 15), 656.4 Mb (kmer 17), and 654.9 Mb (kmer 19) (Fig. S3 in Supporting Information).

To further assess the quality of our assembly we performed microsynteny analyses against the recently published Nile Tilapia genome (Brawand *et al.* 2014) using the synteny analysis tool SynChro (Drillon *et al.* 2014), allowing for 5, 10, and 15 intervening genes between gene pairs (Table S3 in Supporting Information). We also assessed the completeness

of the genome using CEGMA *vers.* 2.5 (Parra *et al.* 2007) in order to determine the presence of eukaryotic core genes. The results showed that 98% of the core genes assessed by CEGMA were either completely or partially represented in our genome assembly (Table S4 in Supporting Information).

Results

Genome assembly

Quality-filtered Illumina reads (> 97% of bases had Phred quality scores > 20) of the diploid *C. austriacus* genome generated 150.99 Gb of data, which represented approximately 212-fold coverage of the genome after processing (Table 1). Raw Illumina reads are available through the National Center for Biotechnology Information (NCBI) Short Read Archive under Bioproject PRJNA292048.

After trimming and filtering, the remaining reads were assembled into 118,995 contigs (Table 1). The N50 and N90 lengths of the contigs were 21,086 and 3,026, respectively; these values increased 8-fold in the scaffold genome (N50 = 170,231 bp, N90 = 24,561 bp) based on scaffolds > 200 bp in length. These values are similar to other published bony fish genomes. For example, an N50 of 136,950 bp was achieved in bluefin tuna based on scaffolds greater than 2,000 bp (Nakamura *et al.* 2013). Total assembly size of the scaffolded butterflyfish genome was approximately 0.71 Gb, which was slightly larger than the kmer-based estimated genome size of 0.65 Gb (Fig. S2 in Supporting Information), but similar to estimates from other *Chaetodon* species such as *Chaetodon aureofasciatus* (685 Mb) and *Chaetodon rainfordi* (733 Mb) (Hardie & Hebert 2004). Kmer-based genome size estimates can often vary by +/- 10% from short read sequencing assemblies (Guo et al. 2015) due to lower amounts of identified repeat elements.

Annotation

Using a combined approach of homology-based and *de novo* gene prediction we identified 28,926 high confidence gene models. BLAST-based annotation against multiple databases provided annotations for 24,433 genes (Table S1 in Supporting Information). Contigs that did not have a BLAST match in any of the databases were designated unannotated. There were few singleton contigs in our genome, meaning that all scaffolds we present here were assembled with high coverage and none are indicative of contamination. Indeed, there was no observable bacterial contamination in the final gap-closed assembly. The number of genes per assembled scaffold ranged from 1 to 126, with 7% of the scaffolds containing only one gene. BLAST results showed high confidence database matches ($< e^{-5}$) for ~82 % of the gene models, of which more than 97.5% were known sequences from Actinopterygii. The majority of the annotated genes (> 70%) matched sequences from the fish genera *Larimichthys* (~51.3%), *Stegastes* (~16.2%), and *Oreochromis* (~5.1%), which, like *Chaetodon*, belong to the order Perciformes (see Fig. 1).

Analysis of the repetitive element content showed that the *Chaetodon austriacus* genome contained 17.86% (127 Mb) repeat elements with an average GC content of 42.48%. The most abundant elements were DNA transposons, which accounted for > 54% of the classified elements in our analysis (Table S4 and Table S5 in Supporting Information). The overall content of repeat elements, as well as the relative proportions of the most abundant transposable elements (TEs), were similar to the closer related Perciformes species such as *Oreochromis niloticus*, *Neolamprologus brichardi*, *Astatotilapia burtoni*, *Metriaclima zebra*, and *Pundamilia nyererei* (~114 Mya divergence**)**. However, clear differences were observed in comparison to *Oryzias latipes* (Japanese rice fish), which showed reduced proportions of DNA transposons and an increase in Short (SINEs) and Long (LINEs) Interspersed Nuclear Elements, as well as Long Terminal Repeat (LTRs) elements, despite the marginally deeper

phylogenetic distance (~139 Mya divergence). The overall content and relative proportions of TEs was substantially different in the more distantly related species such as *D. rerio* (~232 Mya divergence), *Ctenopharyngodon idellus* (~237 Mya divergence), and *L. chalumnae* (~429 Mya divergence) (Fig. 2; Table S5 and S6 in Supporting Information). Comparisons of repetitive elements among species must be viewed with caution given that the other genomes referred to here were derived from short read sequencing assemblies only, which means that the observed differences could simply be due to the inherent biases associated with each assembly approach.

Discussion

Using an NGS approach we were able to sequence, assemble, and annotate a butterflyfish genome; the first draft genome available for a tropical reef fish. This publically available resource (http://caus.reefgenomics.org) has the potential to elevate the Chaetodontidae family to 'model group' status for future genomic studies with increased short-read sequencing coverage, longer reads to close more of the gaps, transcriptome sequencing to identify expressed gene sets, a linkage map to linearize the scaffolds, and new assembly algorithms as they are released. That said, the high genome coverage (~212-fold), large N50 average scaffold size (170,231 bp), and a relatively high proportion of annotated protein-coding genes (82% of the predicted genes) will facilitate the immediate use of these sequences for comparison with closely related butterflyfish species (DiBattista *et al.* submitted). In the short-term, this resource will enable comparative genomics research with the congeners of *C. austriacus* in the greater Indo-Pacific (Waldrop *et al.* 2016), which may reveal a genetic basis for dispersal capacity, habitat selection, and adaptation to the unique oceanic environments that they inhabit. In the long-term, this genome may help shape new hypotheses related to the

evolution of a broader set of taxa within the specialized but complex butterflyfish family (Bellwood *et al.* 2010), as well as reveal a genetic basis for the relative high rates of hybridization within this group (Hobbs *et al.* 2013; DiBattista *et al.* 2015).

We predicted 28,926 protein coding genes in *C. austriacus*, which is slightly higher than the 24,559 genes identified in *O. niloticus*, the next closest related species with a fully sequenced genome (Brawand *et al.* 2014), but still lower than the 31,059 genes predicted in *T. rubripes* (Aparicio *et al.* 2002). The BLAST-based annotation of these genes provided further support for most, but not all of the gene models. This might in part be explained by the lack of closely related genomes in public repositories, since most of the recently sequenced fish genomes are not yet available on NCBI. The phylogenetic classification of the *C. austriacus* gene set using the NCBI nr database, however, matched the vast majority of the gene models from Perciformes species, which lends further support to the fidelity of the genome assembly and gene models. Moreover, our comparison of the repertoire of repetitive elements across selected fish species showed high similarity to closer related species, but also highlighted substantial differences in the overall content and relative proportion of the most abundant TEs with increasing phylogenetic distance. Specifically we find that all Perciformes species contained substantially less transposable elements than the zebrafish genome (Table S6 in Supplementary Material), although this may be due, in part, to a bias in sequencing and assembly approaches among studies.

This assembly here provides one of the first draft "genomic blueprints" for breeders in the ornamental fish trade (Delbeek 2013; Lawton et al. 2013). This information, along with further sequencing refinements, would then have broad applications to captive breeding programs, experimental study under controlled conditions, and, in some cases, the production of evolutionary curios (http://reefbuilders.com/tag/hybrid-butterflyfish/). Genome scans to identify the genetic underpinnings of traits involved in the ability to spawn and survive in

aquaria, particularly for the obligate corallivores, may mitigate the ongoing harvest of these fish (Lawton *et al.* 2013), in addition to the nefarious fishing methods that sometimes come with it (i.e. cyanide fishing). More generally, ongoing perturbations to shallow-water coastal environments that result in a loss of habitat and coral bleaching are a major threat to the specialized butterflyfish. Recent studies have also shown that the degree of diet specialization in this family may depend on the role of CYP3A proteins in the detoxification of allelochemicals found in some their food sources (i.e. soft corals; Maldonado et al. 2016), which represents an annotated fraction of our draft genome. This butterflyfish draft genome can therefore provide a critical first step in determining whether these species can adapt to natural and anthropogenic pressures.

Acknowledgements. This research was supported by the KAUST Office of Competitive Research Funds (OCRF) under Award No. CRG-1-2012-BER-002 and baseline research funds to M.L.B., as well as a National Geographic Society Grant 9024-11 to J.D.D. For logistic support, we thank Eric Mason at Dream Divers in Saudi Arabia and the KAUST Coastal and Marine Resources Core Lab and Amr Gusti. We also acknowledge important contributions from Luiz Rocha and David Catania for assistance with specimen archiving at the California Academy of Sciences; Yi Jin Liew for providing essential scripts for annotation and species classification, as well as provisioning the genome browser; the KAUST Bioscience Core Laboratory with Sivakumar Neelamegam and Hicham Mansour for their assistance with library prep and Illumina sequencing; and Genotypic Technology Pvt. Ltd. with Dinesh Velayutham, Vinay Pantulwar, and Subashini for their assistance with library prep, Illumina sequencing, genome assembly, and genome annotation.

References

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

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, MacCallum I, et al (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature*, **496,** 311-316.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, **297,** 1301-1310.

Bailey R (2009) *Ecosystem Geography: From Ecoregions to Sites*. 2nd edition, New York, Springer Publishing.

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Research*, **12,** 1269-1276.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, **6,** 11.

Bellwood DR, Klanten S, Cowman PF, Pratchett MS, Konow N, van Herwerden L (2010) Evolutionary history of the butterflyfishes (f: Chaetodontidae) and the rise of coral feeding fishes. *Journal of Evolutionary Biology*, **23,** 335-349.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27,** 578-579.

Brawand D, Wagner CE, Li YI, Malinsky M, Keller I, Fan S, *et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, **513,** 375-381.

Chen N (2004) Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics*, **Chapter 4,** Unit 4 10.

Chen R (2015) On Bioinformatic Resources. *Genomics Proteomics Bioinformatics*.

Cole AJ, Pratchett MS, Jones GP (2008) Diversity and functional importance of coral-feeding fishes on tropical coral reefs. *Fish and Fisheries*, **9,** 286-307.

Delbeek JC (2013) Captive care and breeding of coral reef butterflyfishes. *Biology of Butterflyfishes*, 292.

DiBattista JD, Rocha LA, Hobbs JP, He S, Priest MA, Sinclair-Taylor TH, Bowen BW, Berumen ML (2015) When biogeographical provinces collide: hybridization of reef fishes at the crossroads of marine biogeographical provinces in the Arabian Sea. *Journal of Biogeography*, **42,** 1601-1614.

DiBattista JD, Roberts M, Bouwmeester J, Bowen BW, Coker DF, Lozano-Cortés DF, Choat JH, Gaither MR, Hobbs JP, Kahil M, Kochzius M, Myers R, Paulay G, Robitzch V, Saenz-Agudelo P, Salas E, Sinclair-Taylor TH, Toonen RJ, Westneat M, Williams S, Berumen ML (2016a) A review of contemporary patterns of endemism for shallow water reef fauna in the

Red Sea. *Journal of Biogeography*, **43,** 423-439.

DiBattista JD, Choat JH, Gaither MR, Hobbs JP, Lozano-Cortés DF, Myers R, Paulay G, Rocha LA, Toonen RJ, Westneat M, Berumen ML (2016b) On the origin of endemic species in the Red Sea. *Journal of Biogeography*, **43,** 13-30.

DiBattista JD, Saenz-Agudelo P, Piatek M, Wang X, Aranda M, Berumen ML (submitted) Using a butterflyfish genome as a general tool for RAD-Seq studies in specialized reef fish. *Molecular Ecology Resources*.

Drillon G, Carbone A, Fischer G (2014). SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PloS one*, **9,** e92621.

Edgar RC, Myers EW (2005) PILER: identification and classification of genomic repeats. *Bioinformatics*, **21,** i152-158.

Froukh T, Kochzius M: Species boundaries and evolutionary lineages in the blue green damselfishes *Chromis viridis* and *Chromis atripectoralis* (Pomacentridae). *Journal of Fish Biology* 2008, **72,** 451-457.

Gallant JR, Traeger LL, Volkening JD, Moffett H, Chen PH, Novina CD, Phillips GN, Jr., Anand R, Wells GB, Pinch M *et al.* (2014) Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science*, **344,** 1522-1525.

Giles EC, Saenz-Agudelo P, Ravasi T, Hussey N, Berumen ML (2015) Exploring seascape genetics and kinship in the reef sponge *Stylissa carteri* in the Red Sea. *Ecology and Evolution*, **5,** 2487-2502.

Graham NAJ (2007) Ecological versatility and the decline of coral feeding fishes following climate driven coral mortality. *Marine Biology*, **153,** 119-127.

Gregson MA, Pratchett MS, Berumen ML, Goodman BA (2008) Relationships between butterflyfish (Chaetodontidae) feeding rates and coral consumption on the Great Barrier Reef. *Coral Reefs*, **27,** 583-591.

Guo LT, Wang SL, Wu QJ, Zhou XG, Xie W, Zhang YJ (2015) Flow cytometry and K-mer analysis estimates of the genome sizes of *Bemisia tabaci* B and Q (Hemiptera: Aleyrodidae). *Frontiers in Physiology*, **6,** 144.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, **9,** R7.

Hardie DC, Hebert PD (2004) Genome-size evolution in fishes. *Canadian Journal of Fisheries and Aquatic Sciences*, **61,** 1636-1646.

Hobbs J-PA, van Herwerden L, Pratchett MS, Allen GR (2013) Hybridization among butterflyfishes. pp. 48-69 In: Pratchett MS, Berumen ML, Kapoor B (eds.) *Biology of* Butterflyfishes, CRC Press, Boca Raton, FL.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al (2013) The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, **496,** 498-503.

Iwata H, Gotoh O (2012) Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research*, **40,** e161.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30,** 236-1240.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, *et al.* (2007) The medaka draft genome and insights into vertebrate genome evolution. *Nature*, **447,** 714-719.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Research* , **40,** e9.

Kemp J (2000) Zoogeography of the coral reef fishes of the northeastern Gulf of Aden, with eight new records of coral reef fishes from Arabia. *Fauna of Arabia*, **18,** 293-321.

Kosuri S, Church, GM (2014) Large-scale *de novo* DNA synthesis: technologies and applications. *Nature Methods*, **11,** 499-507.

Lawton RJ, Pratchett MS, Delbeek JC (2013) Harvesting of butterflyfishes for aquarium and artisanal fisheries. In: *Biology of Butterflyfishes* (eds: Pratchett MS, Berumen ML, Kapoor BG), 269-291.

Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M (2013) NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. *Bioinformatics*, btt702.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26,** 589-595.

Littlewood, D.T.J., McDonald SM, Gill AC, Cribb TH (2004) Molecular phylogenetics of *Chaetodon* and the Chaetodontidae (Teleostei : Perciformes) with reference to morphology. *Zootaxa,* **779,** 1-20.

Loh YHE, Katz LS, Mims MC, Kocher TD, Yi SV, Streelman JT (2008) Comparative analysis reveals signatures of differentiation amid genomic polymorphism in Lake Malawi cichlids. *Genome Biology*, **9,** R113.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience*, **1,** 18.

Maldonado A, Lavado R, Knuston S, Slattery M, Ankisetty S, Goldstone JV, Watanabe K, Hoh E, Gadepalli RS, Rimoldi JM, Ostrander GK (2016) Biochemical mechanisms for geographical adaptations to novel toxin exposures in butterflyfish. *PloS one*, **11,** e0154208.

Nakamura Y, Mori K, Saitoh K, Oshima K, Mekuchi M, Sugaya T, et al (2013) Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proceedings of the National Academy of Sciences USA*, **110,** 11061-11066.

Nanninga GB, Saenz-Agudelo P, Manica A, Berumen ML (2014) Environmental gradients predict the genetic population structure of a coral reef fish in the Red Sea. *Molecular Ecology*, **23,** 591-602.

Ngugi DK, Antunes A, Brune A, Stingl U (2012) Biogeography of pelagic bacterioplankton across an antagonistic temperature–salinity gradient in the Red Sea. *Molecular Ecology*, **21,** 388-405.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23,** 1061-1067.

Pratchett MS (2005) Dietary overlap among coral-feeding butterflyfishes (Chaetodontidae) at Lizard Island, northern Great Barrier Reef. *Marine Biology*, **148,** 373-382.

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21,** i351-358.

Raitsos DE, Pradhan Y, Brewin RJW, Stenchikov G, Hoteit I (2013) Remote sensing the phytoplankton seasonal succession of the Red Sea. *PloS one*, **8,** e64909.

Read TD, Petit III RA, Joseph SJ, Alam MT, Weil R, Ahmad M, Bhimani R, Vuong JS, Haase CP, Webb DH, Dove AD (2015) Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828. *PeerJ PrePrints*, e1036.

Roberts CM, Ormond RFG (1992) Butterflyfish social-behavior, with special reference to the incidence of territoriality – a review. *Environmental Biology of Fishes*, **34,** 79-93.

Roberts CM, Shepherd ARD, Ormond RFG (1992) Large-scale variation in assemblage structure of Red Sea butterflyfishes and angelfishes. *Journal of Biogeography*, **19,** 239-250.

Rohling EJ, Fenton M, Jorissen FJ, Bertrand P, Gnassen G, Caulet JP (1998) Magnitudes of sea-level lowstands of the past 500,000 years. *Nature*, **394,** 162-165.

Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF (2014) The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. *PLoS One*, **9,** e102089.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19,** 1572-1574.

Rotjan RD, Lewis SM (2008) Impact of coral predators on tropical reefs. *Marine Ecology Progress Series*, **367,** 73-91.

Saenz-Agudelo P, DiBattista JD, Piatek M, Gaither M, Harrison H, Nanninga G, Berumen ML (2015) Seascape genetics along environmental gradients in the Arabian Peninsula:

insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, **24,** 6241-6255.

Sheppard CRC, Sheppard ALS (1991) Corals and coral communities of Arabia. *Fauna of Arabia*, **12,** 3-170.

Siddall M, Rohling EJ, Almogi-Labin A, Hemleben C, Meischner D, Schmeltzer I, Smeed DA (2003) Sea-level fluctuations during the last glacial cycle. *Nature*, **423,** 853-858.

Smeed DA (1997) Seasonal variation of the flow in the strait of Bab al Mandab. *Oceanologica Acta*, **20,** 773-781.

Sofianos SS (2003) An Oceanic General Circulation Model (OGCM) investigation of the Red Sea circulation: 2. Three-dimensional circulation in the Red Sea. *Journal of Geophysical Research*, **108,** 3066.

Spaink HP, Jansen HJ, Dirks RP (2014) Advances in genomics of bony fish. *Briefings in Functional Genomics*, **13,** 144-156.

Stanke M, Steinkamp R, Waack S, Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Research*, **32,** W309-W312.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research*, **34,** W435-439.

van de Peer Y (2004) Tetraodon genome confirms *Takifugu* findings: most fish are ancient polyploids. *Genome Biology*, **5,** 250.

Wabnitz C (2003) *From Ocean to Aquarium: The Global Trade in Marine Ornamental Species* (No. 17) Cambridge, UNEP/Earthprint.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22,** 787-798.

Waldrop E, Hobbs JP, Randall JE, DiBattista JD, Rocha LA, Kosaki RK, Berumen ML, Bowen BW (2016) Phylogeography, population structure and evolution of coral-feeding butterflyfishes (Subgenus *Corallochaetodon*). *Journal of Biogeography*, **43,** 1116-1129.

Wang Y, Lu Y, Zhang Y, Ning Z, Li Y, Zhao Q., *et al.* (2015) The draft genome of the grass carp (*Ctenopharyngodon idellus*) provides insights into its evolution and vegetarian adaptation. *Nature genetics*, **47,** 625-631.

Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, Smit AF, Finn RD (2013) Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Research*, **41,** D70-82.

Xiao J, Zhang Z, Wu J, Yu J (2015) A brief review of software tools for pangenomics. *Genomics Proteomics Bioinformatics*.

Xu P, Zhang X, Wang X, Li J, Liu G, Kuang Y, et al (2014) Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genetics*, **46,** 1212-1219.

Zekeria ZA, Dawit Y, Ghebremedhin S, Naser M, Videler JJ (2002) Resource partitioning among four butterflyfish species in the Red Sea. *Marine and Freshwater Research,* **53,** 163-168.

Data Accessibility

The data sets supporting the results of this article, including the *Chaetodon austriacus* genome sequence (in FASTA format), are available in the NCBI repository, BioProject: PRJNA292048.

Genome browser URL: http://caus.reefgenomics.org

Author contributions

J.D.D. and M.L.B. conceived of and designed the genome study. J.D.D., P.S.-S., and M.J.P. analyzed the data. X.W. and M.A. validated, annotated, and analyzed the genome. All authors developed the manuscript and approve of the final paper.

Table 1. *Chaetodon austriacus* genome sequencing and assembly statistics.

Sequencing statistics	**Raw data**		**Processed data**	
Library	**Size**	**Coverage**	**Size**	**Coverage**
Shotgun sequences	163 Gb	230.06 X	151 Gb	212.66 X
Mate pair 3-5 Kb inserts	3.80 Gb	4.85 X	2.70 Gb	3.40 X
Mate pair 5-8 Kb inserts	2.30 Gb	2.88 X	1.50 Gb	1.95 X
Mate pair 8-11 Kb inserts	1.47 Gb	1.47 X	0.70 Gb	0.95 X

Assembly statistics	**Contigs**	**Scaffolds**
Total assembly size	663,525,437 bp	712,378,618 bp
Sequences generated	118,995	13,967
Maximum length	282,798 bp	1,950,664 bp
Average length	5,641 bp	51,004 bp
Percentage of non-ATGC characters	0 %	6.85 %
Sequences >= 500 bp	84,217	13,441
Sequences >= 1 Kb	66,176	12,937
Sequences >= 10 Kb	16,071	8,012
Sequences >= 1 Mb	0	18
N50 index[a]	21,086 bp	170,231 bp
N90 index[a]	3,026 bp	24,561 bp

[a]The N50 or N90 index indicate the shortest sequence length (contig and scaffold of the final genome) above which 50 or 90% of the genome, respectively, are assembled.

Fig. 1 Genomic composition of *Chaetodon austriacus*. Genes were classified by best hits against nr and grouped by genus to depict phylogenomic relationships between the predicted gene models and known proteins. The vast majority of *C. austriacus* genes have their best match to known sequences from species of closely related fish genera within the order Perciformes. The letters and associated colors depict the following genera: A) *Chaetodon*, B) *Larimichthys*, C) *Stegastes*, D) *Oreochromis*, E) *Notothenia*, F) *Haplochromis*, G) *Dicentrarchus*, H) *Neolamprologus*, I) *Poecilia*, J) *Takifugu*, K) *Maylandia*, L) *Pundamilia*, M) *Cynoglossus*, N) *Tetraodon*, O) *Fundulus*, P) *Oryzias*, Q) *Danio*, and R) Others. The inner numbers are the number of genes and the outer numbers are their percentage with respect to all the genes. The different colors in the outer ring correspond to the same colors in the Circos plot, which provides information regarding the percentage of genes matching to a specific genus.

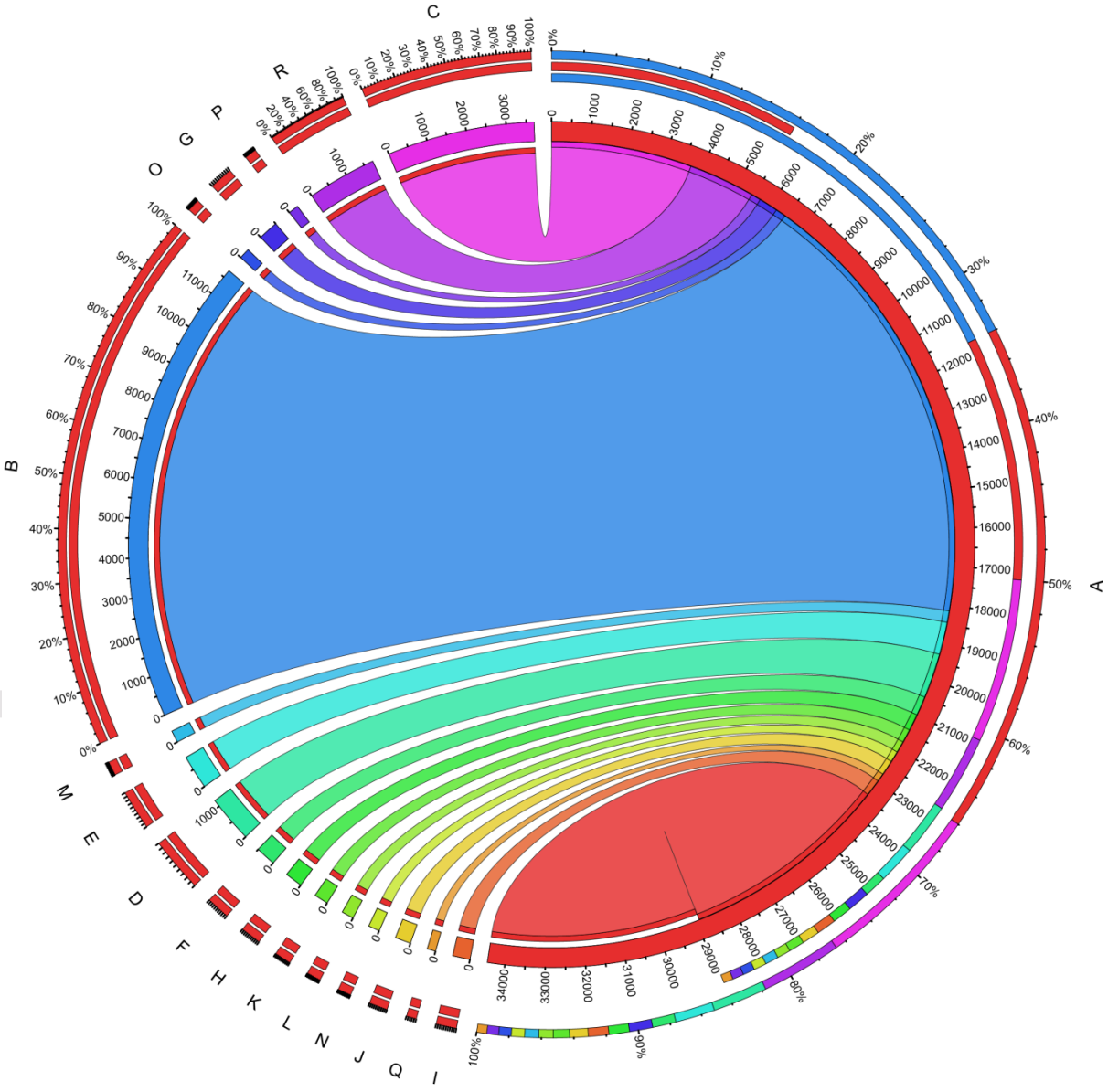

Fig. 2 Relative proportion of the most prevalent transposable elements in selected fish species. Note the generally higher amount of DNA transposons in species from the order Perciformes (below the dashed line), as well as the change in proportions with increasing phylogenetic distance from our focal species (*Chaetodon austriacus*).