School of Biomedical Sciences

An Analysis of the Class I Gene Region in the Sheep Major Histocompatibility Complex

(Sharon) Nitthiya Siva Subramaniam

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

February 2012

Declaration of Authenticity

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.

Signature: 

Date: 22nd Oct 2012

Acknowledgements

This thesis would not have been possible without the guidance and support from several individuals who had extended their valuable assistance throughout the course of this research.

I would like to convey utmost gratitude to my supervisors, A/Prof David Groth and Prof John Wetherall for their constant supervision and support. Both David and John have been of great help and generous in sharing their experience and knowledge in all avenues of my research such as laboratory work, writing of journal articles for publication and thesis preparation. Special thanks to David for always being positive and working hard to generate funds to cover my research cost and expenses. Thank you to John, for his help with various data analysis and being very understanding especially in the process of editing and writing of my thesis.

I am grateful to Mrs Eleanor Morgan for her help in bioinformatics analysis and working tirelessly to analyse and interpret the tremendous amount of data for publications and thesis, as well as for always being very enthusiastic about my research.

I am thankful to Dr Steve Bottomley for his help in structural bioinformatics analysis and for his assistance in form of advice, suggestions and concern regarding the progress of my research.

I would like to express my gratitude to Mr Adrian Paxman, for proof reading my thesis and for his constructive comments that helped improve the quality of my thesis. I also appreciate his support and willingness to help throughout my study.

My sincere thanks to Dr Brian Brestovac for his generous assistance and expert advice on immunology related queries as well as other aspects of my project. I would also like to thank Brian for his readiness to share his knowledge, and for all the interesting topics of conversation in the tearoom that was always entertaining.

It is my pleasure to thank my laboratory mates, in particular "Team David" and all other staff member at the School of Biomedical Science for being friendly and helpful in one way or another, and for making my time at Curtin University an enjoyable one.

I owe my deepest gratitude to my family for their unconditional love and support not only during this study but at all times. My heart-felt thanks to mum, dad, Leelan and Dinesh for always being there for me. I would also like to especially thank Ryan for his never-ending encouragement, support and help. Thank you also for your favour with Excel related work and editing of figures for my thesis.

Last but not least, I would like to thank Curtin University for providing financial assistance through Endeavour International Postgraduate Research Scholarship and Curtin University Postgraduate Scholarship, which enabled me to complete my project successfully, without having any financial stress.

Abstract

The major histocompatibility complex (MHC) is a chromosomal region associated with immune responsiveness in vertebrates. Over four decades many studies have demonstrated important associations between MHC loci and disease resistance or susceptibility in a variety of mammals, especially humans and mice. However, characterisation of the sheep MHC has not been widely studied compared to other domestic species. Since sheep provide food and fibre to many of the world's populations, and is a major industry in Australia a better understanding of the sheep MHC will deliver many benefits when used in conjunction with the new field of genomics, such as the marker assisted selective breeding. This will be of particular benefit for traits that are difficult to improve by conventional selection - especially those that have low heritability, or are expensive to breed for by phenotypic selection. The main aims of this study are to characterise better the genomic architecture of the sheep MHC class I region and to explore structure function relationships of some of the many loci therein. Essential to these aims is the discovery of polymorphic loci, especially SNPs, and the identification of haplotypic elements for association studies and patterns of recombination within this region.

Bacterial artificial chromosomes (BACs) containing sub-regions of the sheep MHC class I region were sub-cloned and sequenced. Contiguous sequences were then re-assembled to generate a physical map of MHC class I region. Single nucleotide polymorphisms (SNPs) within the sheep MHC class I region were identified by polymerase chain reaction (PCR) amplification and sequencing of specific loci in a small population of unrelated sheep. Subsequently, a panel of SNPs spanning the class I region were chosen to genotype a population of distantly related 108 animals to develop a linkage disequilibrium (LD) map, and identify possible recombination hotspots relative to the LD blocks within this region. In addition, the SNP genotypes were used to predict haplotypes within the sheep MHC class I region. In conjunction with these studies, a small population of homozygous sheep was produced by sire-daughter mating for identification of discrete immune related genes within the class I region. This strategy was considered

necessary to reduce allelic variation thereby facilitating the detection of discrete class I loci within a region noted for multiple copies of classical (Ia) and non-classical (Ib) class I genes, as well as probable pseudogenes. MHC class I gene sequences derived from homozygous sheep were supplemented with published reference sequences of different breeds. The many sheep sequences obtained were subjected to multiple sequence alignments and phylogenetic analysis in order to estimate the number of discrete loci present.

The SNP panel generated in this project was also used for association studies with wool production traits in sheep and class I haplotypes. In particular, intragenic genotypes and haplotypes from the skin and hair related *Corneodesmosin* (*CDSN*) gene were analysed in 107 sheep with known estimated breeding values (EBV) for clean fleece weight (CFW), fibre diameter (FD) and staple strength (SS).

The work described herein produced a comprehensive physical map of the sheep MHC class I region, which includes information relating to gene location and organisation. Immune related class I genes are clustered into 3 blocks; beta, kappa and a novel block not previously identified in other organisms. The organisation of other MHC class I genes is similar to that present in the cattle MHC except for a re-arrangement of a cluster of *TRIM* genes. Thirty two SNPs were identified from 14 distinct loci within the MHC class I region. Linkage analysis with a selected panel of 14 SNPs spanning approximately 505 kbp of the class I region revealed four blocks characterised by high linkage disequilibrium. Genotyping of 108 animals with this SNP panel permitted the prediction of thirty four unique haplotypes, which accounted for approximately 90% of haplotype frequency. Two of these haplotypes showed associations with wool production traits, suggesting that the SNPs analysed within MHC class I could be part of an extended haplotype influencing for wool traits. Classification of MHC classical (Ia) and non-classical genes into loci resulted in identification of 14 loci in the sequences analysed. A few of the loci identified show breed specific characteristics and explains possible evolutionary history of the loci in different breeds of sheep. Analysis of SNPs within *CDSN* showed that the gene, which is located in the MHC class I region has an association with

fineness of wool in sheep. Five SNPs within the coding region showed reduced EBV for fibre diameter when present in a homozygous state.

This project has resulted therefore in an improved physical map of the class I region in the sheep MHC, the identification and annotation of several new genes, together with genotypic and haplotypic associations with productivity traits in sheep that will be of immediate interest to the wool industry. As often is the case, the resulted obtained have generated more questions that relate to the structure, function and evolution of this fascinating genomic region that is a critical modulator of the adaptive immune response in vertebrates.

Table of Contents

List of Figures

List of Tables

Glossary of Non-standard Terms

BAC	Bacterial artificial chromosome
Bp / kbp / Mbp	Base pair / kilo base pair / mega base pair
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
E.coli	Escherichia *coli*
EDTA	Ethylenediaminetetraacetic acid
hpH_2O	High pure water
IPTG	Isopropyl β-D-1-thiogalactopyranoside
LB	Luria Bertani
$MgCl_2$	Magnesium chloride
MHC	Major histocompatibility complex
PCR	Polymerase chain reaction
pH	fluid acidity/ alkali indicator
Rpm	Revolutions per minute
SOC	Super optimal broth with catabolite repression
TE	Tris-EDTA buffer
X-GAL	5-bromo-4-chloro-indolyl-β-D-galactopyranoside
^oC	Degrees Celsius
g	Gram
mM / M	Millimolar / molar
ng / mg	Nanogram / milligram
µ	Micro
µL / mL / L	Microlitre / millilitre / litre
V / kV	volts / kilo volts

Chapter 1

Literature Review of MHC Class I

The major histocompatibility complex (MHC) is a gene-dense region associated with innate/adaptive immunity in vertebrates. The MHC was originally discovered by virtue of its critical role in determining the fate of skin allografts between inbred strains of mice (Gorer 1937). The first part of this chapter will describe the structure of the MHC using the human MHC as an archetypical model for comparison with the sheep MHC. Subsequent sections in this literature review will focus on the class I region of MHC and immune function related genes contained within (classical and non-classical class I genes). Functional aspects of the MHC class I region will then be reviewed. These include the role of classical class I molecules in antigen presentation pathways, structure of classical and non-classical class I genes and molecules, haplotypic variation within the class I region, and the identification of selection acting upon MHC class I genes to maintain sequence diversity.

1.1 Introduction

The most extensively characterised major histocompatibility complex (MHC) is the human MHC, and it will be used as a model MHC for this review. The human MHC regulates several aspects of the immune system and contributes significantly to the disease control mechanisms in the host (Beck & Trowsdale 2000). The human MHC is located on the short arm of chromosome 6 at p21.3 and is divided into three main regions, based primarily upon the main components of the respective regions (class I, III and II) (Trowsdale 2001). The extended class I and II regions flank these three major regions and gives rise to the organisation of MHC from telomere to centromere as extended class I, class I, class III, class II and extended class II regions respectively (Stephens *et al.* 1999). A diagrammatic representation of the organisation of the human MHC is shown in Figure 1.1.

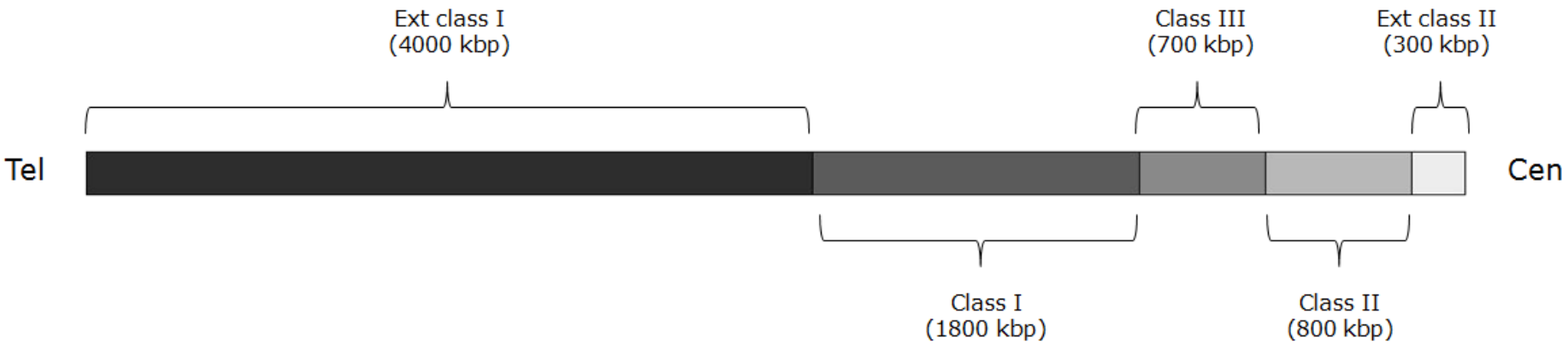


Figure 1.1: The general structure of MHC. Tel indicates telomeric end and Cen indicates centromeric end.

The human MHC spans a region of approximately 3.6 Mbp and contains an average of 1 gene every 16 kbp, making it one of the most gene dense regions identified in the human genome (Beck & Trowsdale 2000). A total of 224 loci have been annotated within the human MHC (excluding the extended class I region), of which 128 are expressed (Beck & Trowsdale 2000). The immune related loci comprise 40% of these 128 expressed loci (Beck & Trowsdale 2000). The overall organisation of MHC is relatively conserved in mammals, as there is evidence that the MHC evolved before mammalian evolutionary divergence from its ancestor species (Srivastava *et al.* 1985; Figueroa & Klein 1986; Klein & Figueroa 1986; Klein & O'Huigin 1993; Takahata 1995).

The extended class I region is telomeric to the classical class I region and is approximately 4000 kbp in length (Beck & Trowsdale 2000). This region of the human MHC exhibits significant linkage disequilibrium (Malfroy *et al.* 1997), which continues beyond *HLA-F* at the end of the classical class I region (Yoshino *et al.* 1997). It contains MHC-like genes such as *HFE*, which encode for membrane proteins that are similar to MHC class I-type proteins, and several copies of *butyrophilin* (Ruddy *et al.* 1997; Tazi-Ahnini *et al.* 1997). This region also contains a large repertoire of olfactory-receptor genes however, the actual number of loci in the extended class I region varies between species (Beck & Trowsdale 2000).

The class I region of the human MHC is approximately 1800 kbp in length and harbours more than 55 different genes (Beck & Trowsdale 2000). The region contains genes comprising the classical MHC genes, nonclassical MHC genes and other miscellaneous genes which have no apparent involvement in the immune response (Flajnik & Kasahara 2001). An evolutionary study has shown that the class I genes in mammals are paralogous as there is evidence of extensive duplication and structural re-organisation (Kelley *et al.* 2005). Almost half the genes within class I region are classified as non-functional pseudogenes (Beck & Trowsdale 2000).

The class III region spans the central portion of the human MHC and is approximately 700 kbp in length (Beck & Trowsdale 2000). It contains several genes with a role in regulating innate and adaptive immunity (Flajnik

& Kasahara 2001). This region contains genes encoding proteins such as the complement components C4, C2 and factor B (Bf), inflammatory cytokine tumour necrosis factor (TNF) and the heat shock protein HSP70, as well as other genes which encode non-immunologically related proteins (Flajnik & Kasahara 2001). The class III region in humans has a high gene density, with one expressed gene approximately every 15 kbp. It is the most gene-dense region of the MHC and possibly the entire human genome, with no pseudogenes observed (Beck & Trowsdale 2000).

The centromeric class II region is approximately 800 kbp in length and contains genes that among other things encode proteins involved in the presentation of peptides derived from extracellular pathogens (Trowsdale 1993; Beck & Trowsdale 2000). The class II region encodes for HLA-DR, -DQ, -DO, -DM and –DP, all of which are involved in this presentation of exogenous proteins. Expression of class II molecules occurs primarily on cells associated with the immune response such as B cells, macrophages and dendritic cells. Class II molecules present antigens to CD4 + cells and trigger the production of immunoglobins by B cells (Trowsdale 1993).

The extended class II region spans about 300 kbp between the *HSET* and *HLA-DP* genes (the centromeric end of the class II region). Aside from *TAPASIN*, most of the genes such as *BING1, BING3, BING4, BING5, RGL2, DAXX (BING2)* and *HKE2* (Herberg *et al.* 1998) residing in extended class II region, do not have an apparent immune related activity. The *TAPASIN* gene is located 180 kbp centromic of *HLA-DP* (Herberg *et al.* 1998) and its protein product has an integral role in the MHC class I-dependent pathway by directing the assembly of MHC-peptide accessory protein complexes (Stephens *et al.* 1999; Flajnik & Kasahara 2001).

1.2 Ovine Leukocyte Antigen Complex

The sheep/ovine major histocompatibility complex (MHC) was initially identified in 1978, approximately 40 years after the initial discovery of MHC in mice and termed OLA (ovine lymphocyte antigens) due to its serological history (Millot 1978). Later, the OLA was renamed as *Ovar* (***Ov****is* ***ar****ies*) MHC in an attempt to standardise the nomenclature of MHC in all vertebrates, using a four-letter abbreviation of the species' name (Klein *et al.* 1990). However, the proposed nomenclature system (Klein *et al.* 1990) was not universally used as most researchers in the field of sheep immunogenetics still used the original name; OLA (Ellis *et al.* 2006). In 2005, the Comparative MHC Nomenclature Committee established guidelines for MHC nomenclature in various species, a common framework and standard nomenclature for sequence submission and publication. This committee suggested that sheep MHC be designated as ovine leukocyte antigen (OLA) complex and that sheep sequences be prefixed according to their species (Ellis *et al.* 2006). Therefore, the MHC in sheep can now be described simply either as the sheep MHC or more specifically OLA.

The OLA complex, a gene dense region containing genes involved in both immunological and non-immunological functions, gained prominence in the late 1980s for the study of disease resistance (Dukkipati *et al.* 2006). Aside from its biomedical significance, the OLA complex is also a region with evolutionary interest for the understanding of its function and emergence of the current organisational structure amongst jawed vertebrates (Laird *et al.* 2000; Flajnik & Kasahara 2001; Danchin *et al.* 2003; Danchin *et al.* 2004; Kasahara *et al.* 2004). The OLA complex has been mapped to chromosome 20 between bands q15 and q23 (Mahdy *et al.* 1989; Hediger *et al.* 1991). Reviews on structure, function and gene polymorphisms within the OLA complex have previously been published (Schwaiger *et al.* 1996; Dukkipati *et al.* 2006). Since these reviews, there have been additional investigations published on the physical map of the OLA complex, the organisation of its genes and some preliminary data on haplotypic variation (Liu *et al.* 2006; Ballingall *et al.* 2008; Qin *et al.* 2008; Gao *et al.* 2010; Liu *et al.* 2010; Lee *et al.* 2011).

1.3 Organisation of OLA

The mammalian MHC shows a considerable level of conservation between species with respect to the three main regions (class I, class III or central region and class II) and the two sub-regions (extended class I and extended class II) (Herberg *et al.* 1998; Stephens *et al.* 1999; Kelley *et al.* 2005). However, the precise organisation of the MHC structure differs between different mammalian species, with some regions showing conservation whilst other areas displaying significant differences (Kelley *et al.* 2005). In sheep, the MHC class I region is poorly characterised (and understood) and the assumed gene organisation is often based upon comparisons with other mammalian MHC such as, human, or its evolutionarily closely related cattle MHC. The general structure of sheep MHC is known to be similar to cattle. A distinct feature of cattle MHC is an inversion involving part of the class II region to form distinct class IIa and class IIb sub-regions (Childers *et al.* 2006). The sheep class II is also split into 2 distinct sub-regions. In sheep, the class IIa and class IIb are separated by approximately 20cM and incorporate the human MHC equivalent extended class II and other non-MHC genes in between the class IIa and IIb regions (Liu *et al.* 2006; Gao *et al.* 2010; Lee *et al.* 2011).

Whilst, many studies in sheep have concentrated on specific MHC immune related genes and their relationship with diseases, there has been paucity of information relating to the characterisation of the structure of the OLA complex in the literature. The first physical map of OLA complex was published in 2006 and was based upon Bacterial Artificial Chromosome (BAC) clones derived from Chinese fine-wool Merino sheep (Liu *et al.* 2006). This map was incomplete and there was a gap between *Notch4* and *Btnl2* in the region between class III and IIa. This gap was later bridged by the identification of two overlapping BAC clones (Liu *et al.* 2010). Following this, a complete DNA sequence map of OLA complex was recently published (Gao *et al.* 2010). In this study, 26 overlapping BAC clones were sequenced using shotgun sequencing and the sequences assembled to form a sequence map of the OLA complex (Gao *et al.* 2010). Annotation was performed on approximately 2.4 Mbp of OLA sequence resulting in the identification of 177 protein-coding genes (Gao *et al.* 2010). The study by Gao *et al*. (2010)

suggested that 14 of the annotated genes had not been previously reported on sheep, and that 10 genes were ovine-specific as they were not present in other mammals. However, several genes that have been sequenced either partially or completely in other sheep studies were not present in the sheep MHC map proposed by Gao *et al*. (2010). For example, the following gene sequences have been published in National Center for Biotechnology Information (NCBI) GenBank database: within MHC class II region; *tapasin* (*TAPBP*) (EU814901), *prefoldin 6* (*PFDN6*) (GQ867665), *WDR repeat domain 46* (*WDR46*) (GU056180), and *ral guanine nucleotide dissociation stimulator-like 2* (*RGL2*) (GQ131514), within MHC class III region; the *tenascin XB* (*TNXB*) (EF197845) and *G6B* (EF197833). These genes were sequenced from Australian Merino sheep. It is unlikely the multiple genes that are present in Australian Merino and other breeds of sheep as well as orthologous cattle MHC have not been identified in the MHC map derived from Chinese Merino. Figure 1.2 shows the comparison between human and sheep MHC.

Although there is a current BAC map and DNA sequence based upon the Chinese Merino it is not known if this is representative for all breeds. It is unknown whether this can be used as a framework to understand the structure of class I region. Furthermore, sequences obtained from these BACs represent only one individual animal and may not be indicative of the diversity in the class I region. In addition, no linkage map for the sheep MHC is available and therefore relationships between actual MHC genes within and between populations are poorly understood. In order to answer these questions it is proposed that neutral markers such as single nucleotide polymorphisms (SNPs) be identified to allow a comprehensive study of structural variation in the sheep class I region.

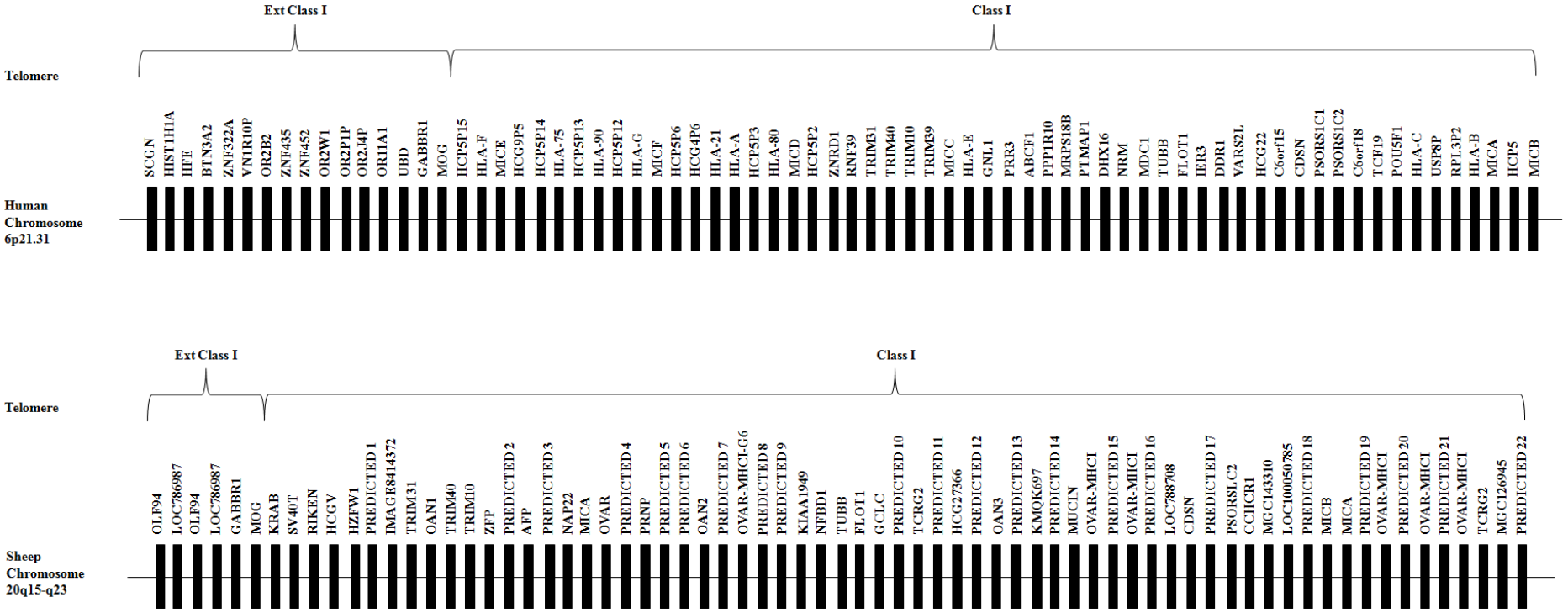


Figure 1.2(A): Comparison of gene content and organisation between human and sheep MHC class I region. Human MHC map in this figure is based on National Center for Biotechnology Information (NCBI) Genome Map; Build 36.3. Sheep MHC map is based on the map published by Gao *et al*. (2010).

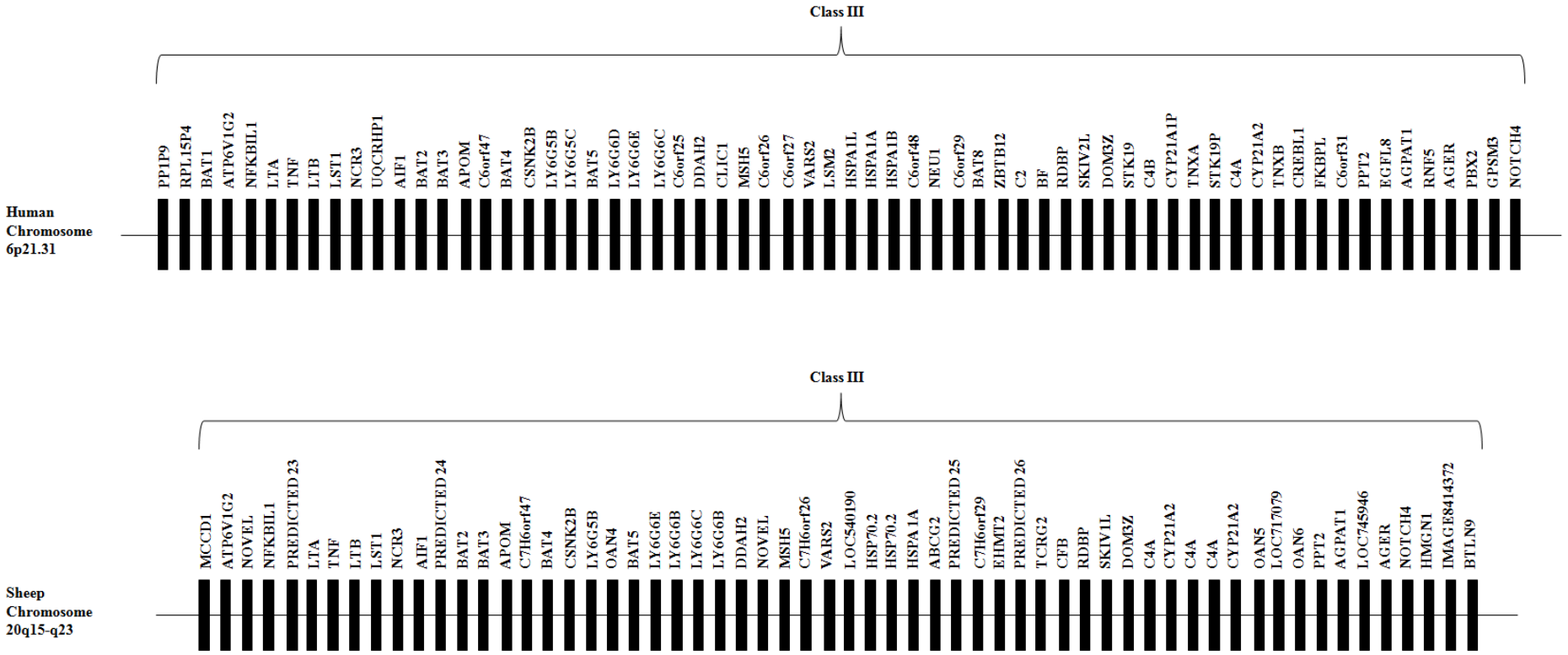


Figure 1.2(B): Comparison of gene content and organisation between human and sheep MHC class III region. Human MHC map in this figure is based on National Center for Biotechnology Information (NCBI) Genome Map; Build 36.3. Sheep MHC map is based on the map published by Gao *et al*. (2010).

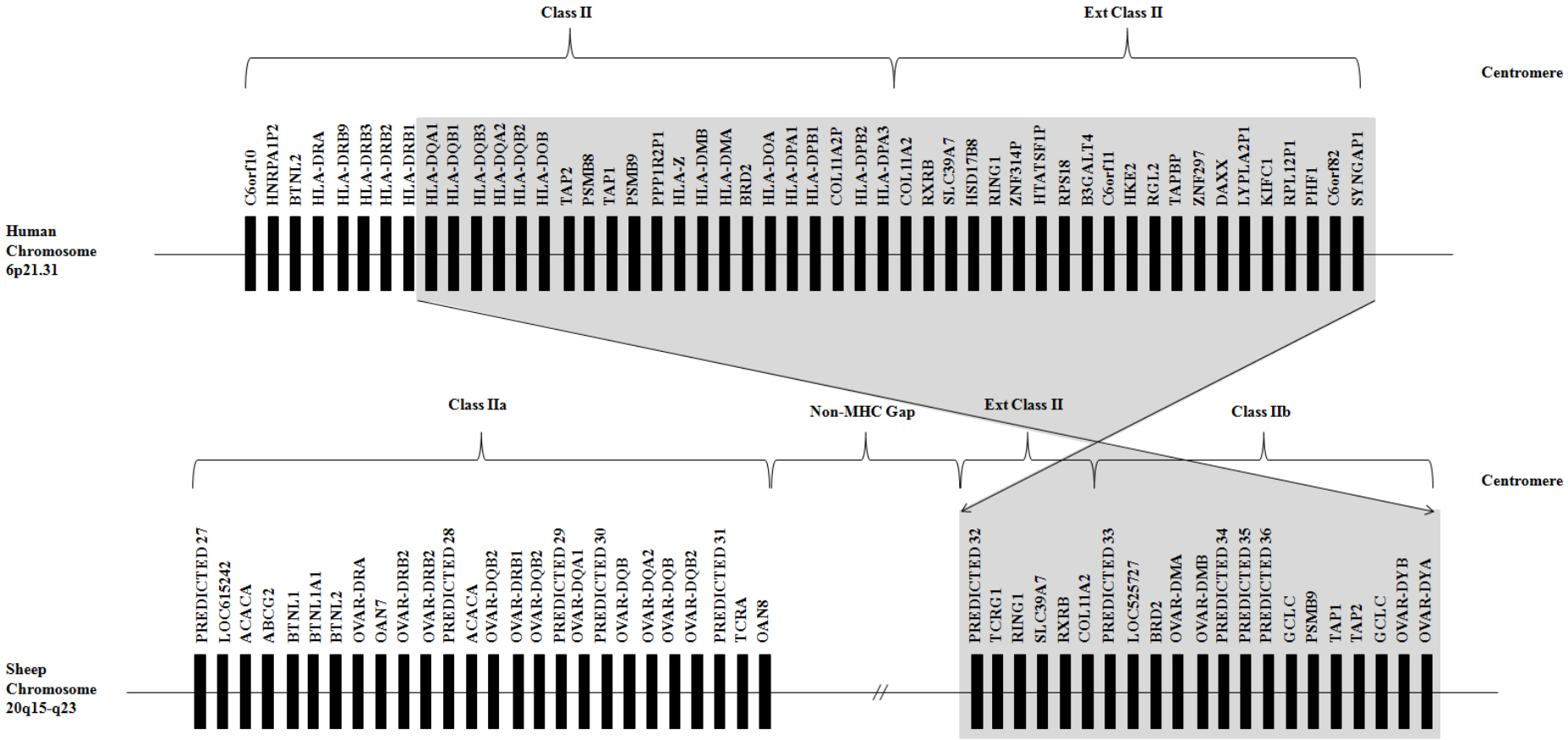


Figure 1.2(C): Comparison of gene content and organisation between human and sheep MHC class II region. Human MHC map in this figure is based on National Center for Biotechnology Information (NCBI) Genome Map; Build 36.3. Sheep MHC map is based on the map published by Gao *et al*. (2010).

1.4 Class I region of MHC

Major Histocompatibility Complex (MHC) is a region containing genes which are involved in both immune and non-immune related functions. Immune related genes in MHC class I region are classified into main 2 categories; classical and non-classical class I genes. Expression of the classical class I molecules occurs on the surface of nucleated cells in the body, except neurons. Classical class I gene products serve as a medium for the transportation and presentation of intracellular peptides to the CD8+ cytotoxic T cells (Trowsdale 1993). The next section of this chapter will describe in detail the MHC class I endogenous peptide presentation pathway.

1.4.1 Class I antigen presentation pathway

Somatic cells in higher vertebrates continuously display intracellular peptides at the cell surface for recognition by cytotoxic T cells. Intracellular peptides include fragments of the body's own 'house-keeping' proteins as well as peptides resulting from the degradation of non-self proteins, such as viral proteins (Carbone & Bevan 1990; Kovacsovics-Bankowski *et al.* 1993; Pfeifer *et al.* 1993; Parham & Ohta 1996). Antigenic peptide migration and presentation at the cell surface alerts the immune system to the existence of potential pathogenic organisms within the presenting cells. Infected cells are eliminated through the cell-mediated immunity (CMI) response that causes cellular destruction therefore, preventing cell-to-cell spread and systemic infection. The presentation of intracellular peptides is an intricate process involving a diverse array of cellular components.

A schematic antigen presentation pathway is shown in Figure 1.3. The process initiates with the proteolytic cleavage of endogenous proteins by the proteasome, a large protease complex in the cytosol (Rock *et al.* 1994). Large peptide fragments then undergo a series of degradation processes mediated by cytosolic peptidases such as puromycin-sensitive aminopeptidase (PSA), bleomycin hydrolase (BH) and tripeptidyl peptidase-II (TPP-II) (Stoltze *et al.* 2000; Reits *et al.* 2004). The heterodimeric transporter associated with antigen processing (TAP) translocates the resultant peptides into the endoplasmic reticulum (ER) (Cresswell & Howard

1999) where ER-luminal aminopeptidases 1 and 2 (ERAP 1/ 2) trim the peptides further to 8-11 amino acid residues (Saric *et al.* 2002; Serwold *et al.* 2002; Saveanu *et al.* 2005). TAP subunits (TAP1/TAP2), tapasin and calreticulin form a multimeric structure called the Peptide Loading Complex (PLC), where endogenous peptides are loaded onto the major histocompatibility complex (MHC) class I heavy chain (Howarth *et al.* 2004). Chaperone calnexin and ERp57, a thiol-dependent oxidoreductase that aids in the folding of the MHC class-I heavy chain prior to an interaction with PLC through the formation of disulfide bonds (Farmery *et al.* 2000). Correctly folded MHC class I heavy chain then associates with β_2-microglubolin to form a MHC class I-β_2-m dimer and this dimer subsequently dissociates from calnexin (Sadasivan *et al.* 1996; Ortmann *et al.* 1997; Morrice & Powis 1998; Antoniou *et al.* 2002). Following this, the MHC class I-β_2-m dimer associates with the PLC to form an assembled PLC, which includes class I heavy chain, β_2-microglubolin, ER chaperones calreticulin and ERp57, TAP1/TAP2 and tapasin (Sadasivan *et al.* 1996; Ortmann *et al.* 1997; Morrice & Powis 1998; Antoniou *et al.* 2002).

Loading of appropriately sized 8-11 residue endogenous peptides onto the class I molecules occurs in the ER and is optimised and controlled by the PLC (Tan *et al.* 2002). The peptide loaded MHC class I-β_2-m dimer also known as peptide-MHC class I complex then dissociates from the remaining PLC (Momburg *et al.* 1994; Ortmann *et al.* 1994). ERp57 which remains bound to MHC class I-β_2-m dimer during peptide loading and productive maturation is also released upon the exit of peptide-MHC class I complex from PLC into the ER (Frenkel *et al.* 2004). The resulting peptide-MHC class I complex is finally shuttled to the cell surface via the Golgi apparatus for recognition by CD8+ cytotoxic T lymphocytes (Momburg *et al.* 1994). Presentation of 'foreign' or nonself-peptides to CD8+ cytotoxic T lymohocytes activates CMI response, which causes destruction of abnormal or infected cells (Townsend & Bodmer 1989). However, the host immune system shows tolerance towards self-proteins, and as such, the presentation of self-peptides does not stimulate cellular destruction (Townsend & Bodmer 1989).

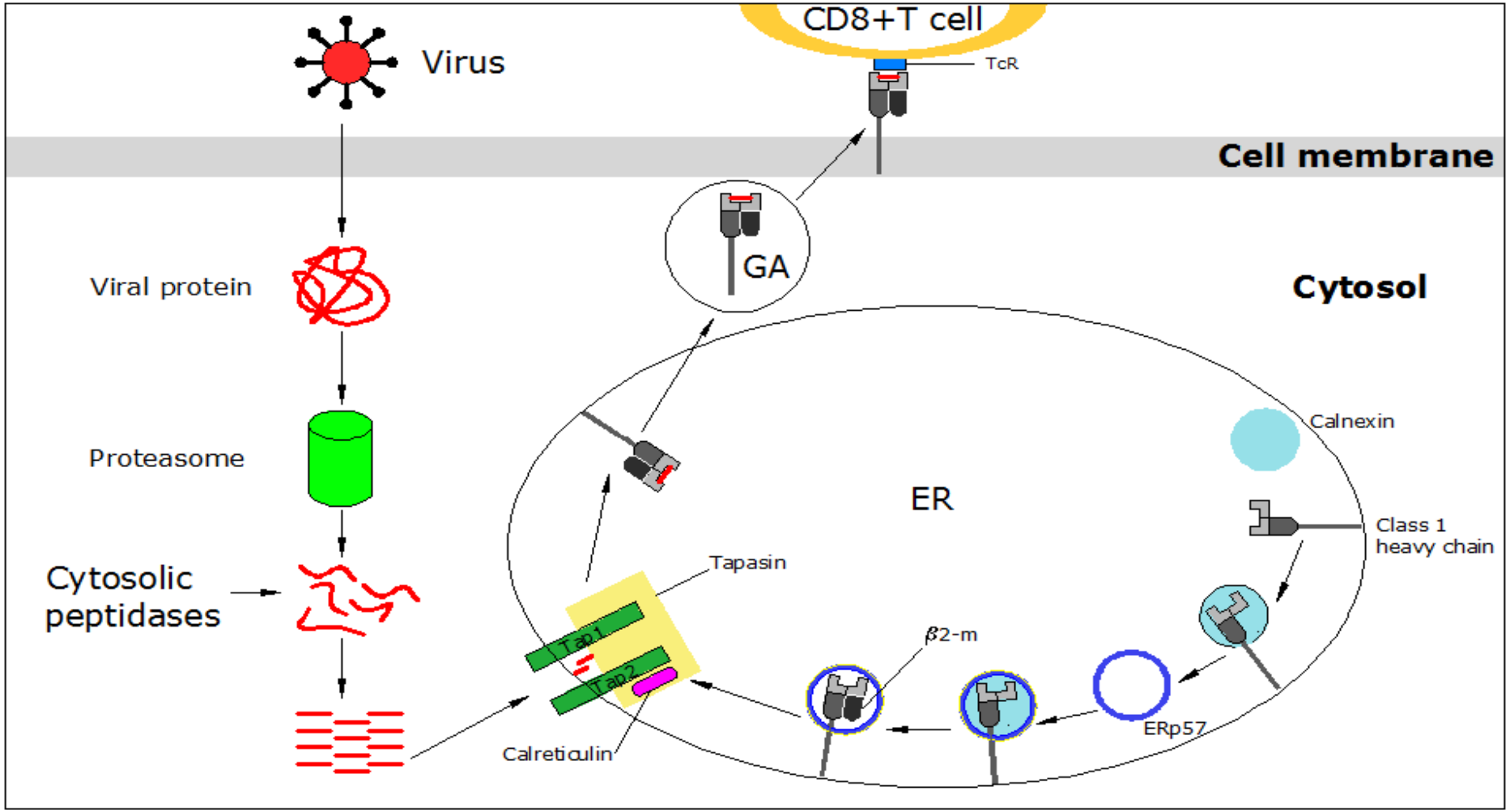


Figure 1.3: MHC class I presentation pathway.

1.4.2 Classical class I genes

Classical sheep class I genes, initially characterised as MHC cDNA clones derived from sheep thymus (Grossberger *et al.* 1990) and was found to be similar to sequences from other organisms such as humans and bovines (Grossberger *et al.* 1990). This region encodes the MHC class I heavy chains which heterodimerize with covalently linked β_2-microglobin to play a central role in peptide presentation during immune response (Bjorkman *et al.* 1987a, 1987b; Townsend & Bodmer 1989). The highly polymorphic nature of the classical class I genes allows a broad spectrum involvement in the activation and inhibition of natural killer cell responses (Trowsdale 2001).

Structurally, the classical class I MHC gene encodes for three α domains, a transmembrane domain and a cytoplasmic domain (Orr *et al.* 1979; Malissen *et al.* 1982; Bjorkman & Parham 1990) and is approximately 3000 base pairs in length. Of the eight exons composing the gene, exon 1 encodes the leader or signal peptide. Exon 2, 3 and 4 encode the α_1, α_2 and α_3 domains respectively, whilst the transmembrane domain of the heavy chain is encoded by exon 5. Exons 6, 7 and 8 encode the cytoplasmic tail region (Hughes & Nei 1989a, 1989b).

Figure 1.4 illustrates the basic structure of the MHC classical class I gene as exemplified by human HLA-A (allele HLA-A*020101XX, National Center for Biotechnology Information (NCBI) GenBank database; accession ID: AM943368 (Heinold *et al.* 2008).

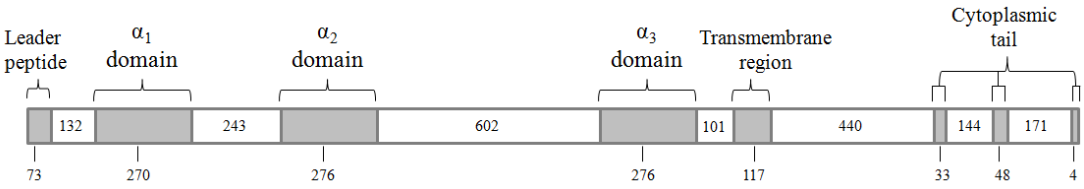


Figure 1.4: An example of MHC classical class I gene structure.

1.4.3 Classical class I molecule

Classical class I MHC (Ia) molecules are cell surface glycoproteins found on all nucleated somatic cells, with the highest concentrations on lymphocytes and macrophages (Bjorkman *et al.* 1987a, 1987b; Bjorkman & Parham 1990). Due to their involvement in peptide presentation, Ia proteins are classified as both peptide binding proteins and antigen presenting molecules (Teh *et al.* 1988; Parnes 1989). MHC class Ia protein binds antigenic peptides (of endogenous proteins derived from virus infected or cancerous cells) to form a peptide-MHC protein complex (Rammensee *et al.* 1995) that is recognised by T cell receptors (TCRs) and the consequential stimulation of CD8+ cytotoxic T lymphocytes (CTLs) or killer T cells (Teh *et al.* 1988; Parnes 1989; Bjorkman & Parham 1990).

Sheep MHC class I molecule is known to be structurally similar to its counterpart in other mammals (Gogolin-Ewens *et al.* 1985). The three dimensional molecular structure of a class I protein was first studied with HLA-A2, a human class I protein, using X-ray crystallography (Bjorkman *et al.* 1987a, 1987b). Classical class I MHC molecule is a heterodimer comprising a heavy or α chain, which is non-covalently bonded to a light chain. The heavy chain is a 44 kd protein encoded by class I MHC gene and consists 3 major domains based on their location; extracellular domain, transmembrane domain, and cytoplasmic domain (Grey *et al.* 1973; Peterson *et al.* 1974; Bjorkman *et al.* 1987a; Hughes & Yeager 1998). Whereas the light chain, also known as β_2-microglobin (β_2-m), is a non-MHC encoded protein of 12 kd (Grey *et al.* 1973; Peterson *et al.* 1974; Bjorkman *et al.* 1987a; Hughes & Yeager 1998).

The N-terminus of the class I MHC polypeptide heavy chain is composed of three separately encoded subunits comprising the extracellular domain (Orr *et al.* 1979; Malissen *et al.* 1982). Analysis of human HLA-A2 indicates these subunits to have lengths of approximately 90 amino acids each (α_1 - 90aa; α_2 - 92aa; α_3 - 92aa) (Malissen *et al.* 1982). The transmembrane region extending from the α_3 domain is approximately 25 amino acids in length, and the cytoplasmic tail contains around 30 amino acids (Bjorkman &

Parham 1990). The structure of MHC classical class I MHC molecule is shown in Figure 1.5.

The α_1 and α_2 domains form one of the two homologous regions in the classical class I MHC molecule, with each being composed of a four-stranded anti-parallel β-pleated sheet and two α-helices (Grey *et al.* 1973; Peterson *et al.* 1974). Despite minimal sequence similarity, the α_1 and α_2 subunits of the extracellular domain share a common tertiary structure and form the homologous peptide-binding region (PBR) that projects from the cell surface (distal to the cell membrane) during peptide presentation (Orr *et al.* 1979). Comparison of different classical class I allelic gene products has indicated that the α_1 and α_2 domains demonstrate the highest number of amino acid substitutions (Guillet *et al.* 1986; Parham *et al.* 1988).

The α_3 region, positioned between the cellular membrane and α_1/α_2 structure during presentation, is structurally similar to the light chain β_2-m, with which it constitutes the second homologous classical class I MHC molecule domain. Further, whilst the α_1 and α_2 domains are highly variable, the α_3 and β_2-m domains are relatively conserved (Hughes & Nei 1988; Parham *et al.* 1988), with the latter possessing the least number of amino acid substitutions between products of the various classical class I alleles (Hughes & Nei 1988; Parham *et al.* 1988). The α_3 and β_2-m domains each consist of two anti-parallel β-pleated sheets (one of four β-strands and the other of three strands) and both molecules fold to form β-sandwich structures similar to those seen in Ig regions. These proteins are therefore classified as members of the Ig gene superfamily (Peterson *et al.* 1972; Smithies & Poulik 1972; Orr *et al.* 1979; Tragardh *et al.* 1979; Williams 1987; Hunkapiller & Hood 1989).

Association of β_2-m with the class I MHC molecule is essential for the facilitation and stabilisation of the fold in the MHC heavy chain that in turn, is responsible for its structural conformation (Lancet *et al.* 1979; Krangel *et al.* 1983; Yokoyama *et al.* 1985; Allen *et al.* 1986; Hansen *et al.* 1988).

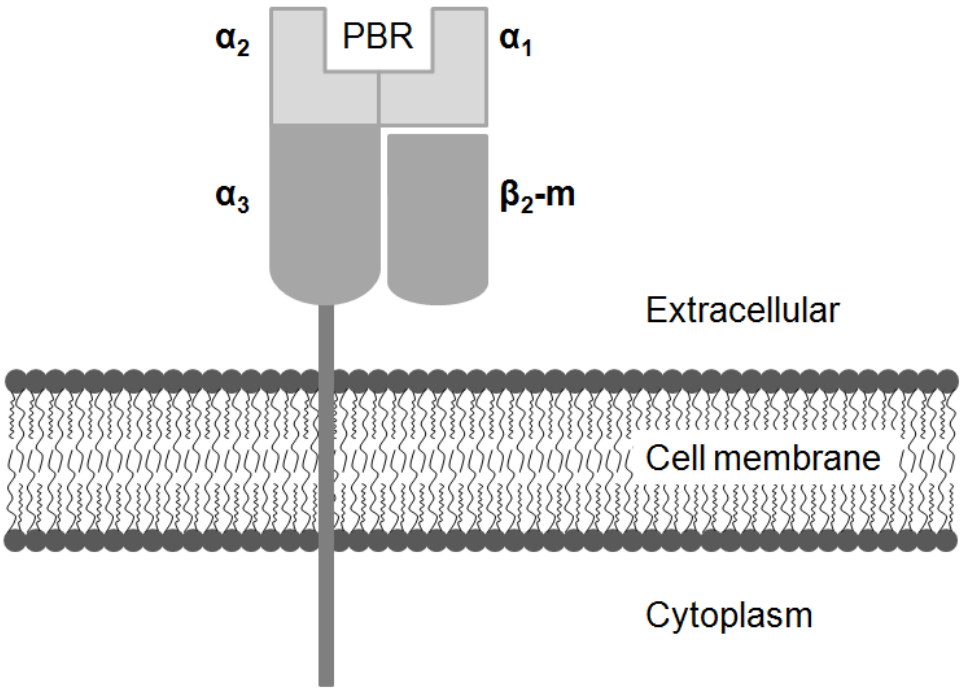


Figure 1.5: Schematic representation of classical class I MHC molecule.

1.4.4 Non-classical class I genes

Non-classical class I MHC genes also called class Ib genes in human share sequence homology with classical class I MHC genes (class Ia). The non-classical class I genes differ from classical class I genes based on a few distinct characteristic: non-ubiquitous expression or lower cell surface expression, truncated cytoplasmic domain and limited polymorphism (Geraghty *et al.* 1992; Birch *et al.* 2008a). There are three class Ib genes in human; HLA-E, HLA-F and HLA-G (Wei & Orr 1990; Shawar *et al.* 1994). Other class Ib genes include pseudogenes, MHC class I chain related (MIC) genes and non-MHC genes such as FcRn, CD1-a, CD1-b, CD1-c and CD1-d (Shawar *et al.* 1994; Braud *et al.* 1999), which will not be further discussed in this review. In cattle, four non-classical class I genes have been identified within the MHC; one located close to classical class I genes and the other three genes located about 500 kbp centromeric close to *MIC* genes (Birch *et al.* 2008b). The number of non-classical class I genes in sheep remains unconfirmed although the recently published sheep MHC map showed presence of at least two copies (Gao *et al.* 2010).

The overall arrangement and size of intron and exons of non-classical class I genes are similar with the classical class I genes, but with notable differences. Although the non-classical class I genes have limited polymorphism, exon 6-8 encoding the cytoplasmic domain of MHC heavy chain shows little homology between different paralogues (Stroynowski 1990; Srivastava & Lambert 1991; Cook *et al.* 1992). The HLA-E gene is distinguishable from other class I genes, in that, exon 7 has a deletion of 5 bases. This deletion leads to in an in-frame termination stop codon in the mRNA, thus encoding a shorter polypeptide compared to HLA-A and HLA-B chains (Srivastava *et al.* 1987; Koller *et al.* 1988; Mizuno *et al.* 1988). Another variation unique to the HLA-E gene is the three ALU elements present in the non-coding region, located in the 5' flanking region, intron 5 and 3' untranslated region (UTR) respectively (Srivastava *et al.* 1987; Koller *et al.* 1988; Mizuno *et al.* 1988). The most distinguishable variation in the HLA-F gene is the dinucleotide AA sequence instead of AG sequence at the 3' splice recognition site of intron 6 that disrupts completely splicing of exon 7 onwards, resulting in truncated mRNA sequence (Green 1986; Geraghty *et al.* 1990). The HLA-G is different from other class I genes due to the in-frame termination stop codon in exon 6, which causes translation to stop before most of exon 6 and all of exon 7 and 8 are translated (Geraghty *et al.* 1987). The cytoplasmic domain of HLA-G molecule consists of only 6 amino acids (Geraghty *et al.* 1987).

In cattle, four non-classical class I genes have been reported; NC1, NC2, NC3 and NC4. Study of these cattle non-classical class I genes showed that there is variation between loci and many alleles at any locus (Birch *et al.* 2008a). Non-classical class I amino acid sequence alignment in cattle revealed characteristic VPI, IPI or VLIK motif in the transmembrane domain that is not present in the classical class I genes (Birch *et al.* 2008a). The characteristic feature that describes non-classical class I gene NC1 is the truncated cytoplasmic domain due to early stop codon in the gene and/ or deletion in the region encoding transmembrane (TM) domain (Birch *et al.* 2008a). The NCI gene encoded amino acid sequence has a distinct VPI motif in the TM domain. Contrastingly, the NC2 gene encodes for full-length cytoplasmic domain, harbours IPI motif in its TM domain and has a single codon deletion in the α_2 domain (Birch *et al.* 2008a). Gene NC3 is

monomorphic with only one amino acid change in α1 domain between two alleles. However, the 3' end encoding cytoplasmic region has significant differences; one allele has an early stop codon whereas another allele encodes for full-length cytoplasmic domain (Birch *et al.* 2008a). The unique feature of amino acid sequence encoded by gene NC3 is the VLIK motif on the transmembrane domain (Birch *et al.* 2008a). Analysis of amino acid sequence of NC4 gene from 3 alleles have shown that two alleles encode for full-length TM and cytoplasmic domain whereas one allele has a single base insertion in exon 5, which results in early stop codon just before the end of TM domain (Birch *et al.* 2008a).

1.4.5 Non-classical class I molecules

Class Ib molecules are now acknowledged through various studies as a key mediator of immune recognition (Borrego *et al.* 1998; Braud *et al.* 1998b; Lee *et al.* 1998b). To date, the structure and function of class Ib molecules have been best characterised in human and mouse (Forman 1979; Kastner *et al.* 1979; Rich *et al.* 1979; Shawar *et al.* 1994). Class Ib and Ia molecules are similar in structure, although the cytoplasmic domain in class Ib molecules are generally shorter than class Ia (O'Callaghan & Bell 1998). All of the class Ib molecules bind with β_2-microglubolin, but each shows tissue-specific expression (Shimizu *et al.* 1988; Geraghty *et al.* 1990; Wei & Orr 1990; Ishitani *et al.* 2003; Lee & Geraghty 2003; Le Friec *et al.* 2004; Le Rond *et al.* 2004). Both HLA-E and HLA-G are involved in up and down regulation of natural killer (NK) activity (Borrego *et al.* 1998; Braud *et al.* 1998b; Lee *et al.* 1998b).

The HLA-E molecule binds peptide from the leader sequence of other class I molecules, specifically of HLA-A, HLA-B, HLA-C and HLA-G (Perez-Villar *et al.* 1997; O'Callaghan & Bell 1998; O'Callaghan 2000). Binding of the leader peptide to HLA-E molecule in the endoplasmic reticulum is a TAP-dependent process and as such interacts with TAP, calreticulin and tapasin, similar to the class Ia antigen presentation pathway (Aldrich *et al.* 1994; Braud *et al.* 1998a; Lee *et al.* 1998a). The peptide loaded HLA-E interacts with β_2-microglubolin to form a stable tetrameric complex and subsequently migrates to the cell surface for recognition of CD94/NKG2 receptors on NK

cells, and a subset of T cells (Aldrich *et al.* 1994; O'Callaghan & Bell 1998; Braud *et al.* 1998a; Lee *et al.* 1998a; Braud *et al.* 1998b; O'Callaghan 2000). Interaction of HLA-E tetramers with NK cells expressing inhibiting receptor CD94/NKG2A prevents lysis of cells with an intact class Ia antigen-processing pathway, whereas the interaction with activating receptor CD94/NKG2C triggers cytotoxic activity (Llano *et al.* 1998; Braud *et al.* 1998b; Lee *et al.* 1998b).

HLA-G, expressed mainly on the placental tissue invading the maternal uterus is known to play a role in maternal-foetal immune interaction (Loke & King 1991). It is suggested that HLA-G molecule inhibit NK lysis by the maternal immune response against the placenta in the first trimester, because HLA-G is the only HLA molecule expressed in high concentrations during this period (King *et al.* 1996; O'Callaghan & Bell 1998). Although HLA-F is primarily expressed in lymphoid and mature T-cell lines, its function is still unknown (O'Callaghan & Bell 1998; Braud *et al.* 1999).

1.5 Class I loci and haplotypes variation in OLA complex

The diversity in the number of OLA classical class I loci and haplotypic variation in sheep is a feature of sheep class I region, similar to that described in cattle (Holmes *et al.* 2003; Ellis 2004). Most of the initial studies investigating the OLA class I genes were based on serological methods. Some proposed two expressed class I loci whereas others suggested three loci (Stear & Spooner 1981; Cullen *et al.* 1982; Millot 1984; Garrido *et al.* 1995; Jugo & Vicario 2001; Jugo *et al.* 2002). A study using immunoprecipitation and one-dimensional isoelectric focusing (IEF) of OLA class I antigens to characterise the class I polymorphism at the product level, had revealed that serology has underestimated the OLA diversity level (Jugo *et al.* 2002). The result showed that IEF identified antigens previously not defined by serology and confirmed three serological specificities by biochemical typing (Jugo *et al.* 2002).

The first molecular based study performed to understand the OLA class I genes involved screening of sheep thymus cDNA library with human probe

under conditions of relaxed stringency (Grossberger *et al.* 1990). In this study 13 sequences were assigned to five different classes, suggesting expression of at least three loci (Grossberger *et al.* 1990). In 2005, molecular genetic analysis of haplotypes from two heterozygous Scottish Blackface rams led to identification of four polymorphic class I loci and contributed towards the breeding of an OLA defined resource flock (Miltiadou *et al.* 2005). Subsequently, sheep homozygous four each of the four MHC haplotypes was established to evaluate the functional significant of each of the class I genes (Ballingall *et al.* 2008). Complementary molecular genetic and proteomic approaches were used to determine the ability of each gene to express a class I MHC product and to give rise to a surface expressed protein. Based on the data generated, this study proposed that the OLA consists at least eight class I loci with considerable variation between haplotypes (Ballingall *et al.* 2008). In another study involving Chinese Merino, four classical class I genes and two non-classical class I genes have been reported (Gao *et al.* 2010).

1.6 Selection at the MHC class I region

First identified in a study involving tissue transplantation in mice (Gorer 1937), MHC genes were subsequently typed in humans using serology (Terasaki & McClelland 1964; Dyer & Martin 1991). However, serology has been insufficiently sensitive to distinguish between all alleles - particularly amongst those MHC subtypes that most commonly mark population differences (Imanishi *et al.* 1992; Bidwell 1994; Grubic *et al.* 2000; Grubic *et al.* 2008). Such allelic diversity within and between populations was only identified later through DNA typing procedures.

Answers to the questions pertaining to the generation and maintenance of MHC polymorphisms remain ambiguous even today, but generally fall into one of two camps:

Studies in the mid-seventies indicating that variant MHC alleles might present different antigens for recognition by cytotoxic T lymphocytes lead to the hypothesis of "heterozygote advantage" or "overdominant selection"

(Zinkernagel & Doherty 1974; Doherty & Zinkernagel 1975). This concept suggests that heterozygosity at MHC loci is evolutionarily advantageous, as it allows an individual to recognize a broader array of antigens, thereby conferring resistance to a greater diversity of pathogens.

The alternative view, put forward by population geneticists, took a more neutral approach and was encapsulated by "the theory of molecular evolution". Based upon the concept that synonymous amino acid substitutions outnumber non-synonymous substitutions (as non-synonymous substitutions would alter the amino acid composition and, potentially, result in deleterious consequences), this theory advocates that variation within a species is caused by random drift of effectively equivalent or neutral mutations (Kimura 1979).

In order to test the assumption that different alleles at a locus have equivalent effects upon fitness, and that, with increasing sample size, heterozygosity will decrease, Hendrick & Thomson (1983) analysed the HLA A and B loci from 22 human ethnic groups and found lower homozygosity than would be expected under the "neutrality model". This finding suggested that maintenance of genetic variation at the HLA loci A and B occurs as a result of balancing selection pressures and not random drift (Hedrick & Thomson 1983). The increasing availability of sequences for MHC alleles in the 1980s made it possible to further evaluate this conclusion.

Nucleotide sequence analysis revealed specific MHC allelic regions in which nonsynonymous substitutions clustered at greater frequency than synonymous substitutions (Jaulin *et al.* 1985). The significance of this finding was further elucidated via structural determinations of the MHC molecule through x-ray crystallography, which showed highly polymorphic α_1 and α_2 domains in the peptide binding region (PBR) of the MHC molecule (Bjorkman *et al.* 1987b). Such a high rate of nonsynonymous substitutions in a functional coding region does not support Kimura's (1979) neutrality theory but, rather, reflects the influence of natural or balancing selection (Hughes & Nei 1988). This suggestion has been further supported in a free living population of Soay sheep, wherein microsatellite markers used to

assay genetic variation demonstrated a high level of heterozygosity and relatively even allele frequency distribution across the MHC domains (Paterson & Pemberton 1997).

Whilst these studies have consistently indicated that the significant heterozygosity of the MHC loci is the result of balancing selection pressures, the mechanism through which the polymorphisms are maintained remains unclear, although several alternative hypotheses have been proposed:

The overdominant selection mechanism proposed in the 1970s (Zinkernagel & Doherty 1974; Doherty & Zinkernagel 1975) has since been investigated through a determination of the rate of codon substitution in the MHC PBR (Hughes & Nei 1988). Nonsynonymous substitutions were found to be significantly more prevalent than synonymous substitutions in the PBR region, whereas the reverse was true in regions outside of the PBR domain, a finding in agreement with the overdominance hypothesis (Hughes & Nei 1988).

Frequency-dependent selection is based upon the argument that overdominant selection does not sufficiently and quantitatively explain the persistence of polymorphic alleles (Takahata & Nei 1990). Computer simulations derived from mathematical modelling of frequency-dependent selection have demonstrated a similarly high degree of polymorphism as that observed at MHC loci, although strong evidence of overdominant selection was also indicated by the study (Takahata & Nei 1990).

Gene conversion theories assume that interlocus recombination diverging at the sequence level between members of a gene family could increase polymorphisms at each locus (Lopez de Castro *et al.* 1982; Ohta 1982). However, gene recombination more commonly occurs in a random pattern and cannot, therefore, explain the high rate of nonsynonymous substitutions clustered within the PBR-encoding region of the MHC (Hughes & Nei 1988).

Maternal-foetal interaction, where production of maternal antibodies to foetal class I molecules leads to selection favouring nonsynonymous substitution in the PBR (Clarke & Kirby 1966) has not been supported, due

to the absence of specific clustering of epitopes for maternal antibodies in the PBR – such epitopes being generally dispersed throughout the class I MHC molecules (James 1965; Clarke 1971; Wegmann 1984). Moreover, non-mammalian organisms (lacking maternal-foetal interaction at this level) have also been shown to possess a high level of MHC polymorphism (Hedrick 1998).

Disassortative mating to increase offspring fitness could also result in MHC diversity (Potts *et al.* 1991), as studies on the mating preferences in mice have demonstrated a preference amongst females for males with dissimilar MHC genotypes (distinguishable via classical class I MHC-derived phenotypical odour profiles) (Singh *et al.* 1987; Potts *et al.* 1991). This MHC-based mate-choice hypothesis has been further supported in fish (Olsen *et al.* 1998; Landry *et al.* 2001; Forsberg *et al.* 2007), birds (Freeman-Gallant *et al.* 2003; Bonneaud *et al.* 2006), reptiles (Olsson *et al.* 2003) and human (Wedekind *et al.* 1995; Chaix *et al.* 2008). However, no evidence to account for how this hypothesis would result in natural selection centred distinctively on the PBR has been published (Hedrick & Black 1997; Westerdahl 2004; Richardson *et al.* 2005). Further, a paternity-based likelihood approach employed to analyse the mating pattern amongst a population of free-living St Kildan Soay sheep (typed using five microsatellite markers positioned within or adjacent to the ovine MHC), showed no evidence of MHC-dependent mating preferences (Paterson & Pemberton 1997).

Overall, given the function of the MHC molecule in peptide presentation, and the broad spectrum advantage logically conferred by heterozygosity, maintenance of MHC polymorphisms through overdominant selection appears to remain the more reasonable model, as also supported by the available evidence.

The sheep MHC class I region exhibits high levels of genetic diversity and polymorphisms and the nature, co-inheritance and functional consequences of this diversity are still not fully understood. Specific aims of this research are to:

- Sub-clone and sequence CHORI Bacterial Artificial Chromosomes (BACs) known to contain sheep MHC class I sequences to identify genes present within each BAC and determine the organisation of genes within sheep MHC class I region.
- Compare the map generated from sub-cloning CHORI BACs with the existing cattle MHC reference map, and the recently published sheep MHC map derived from Chinese Merino sheep, (which was not available when this study commenced) to map the general architecture and gene organisation within class I region.

- Identify single nucleotide polymorphisms (SNPs) within the MHC class I region, including intergenic and intragenic loci and use the SNPs discovered to genotype a population of sheep and identify patterns of linkage disequilibrium (LD).

- Sequence various copies of MHC class I genes from a small population of homozygous animals and determine the number of loci present in each animal.

- Identify the number of MHC class I genes being expressed utilising mRNA analysis from homozygous animals.

- Analyse and define the haplotypic structure of the MHC class I region.

- Identify possible association between MHC class I region with production traits in sheep

Chapter 2

General Materials and Methods

This chapter describes various techniques and resources that were used throughout the described research projects. Methods that are specific to a particular research within a chapter are described in the respective chapter. In addition, a description of the general buffers and reagents used in this thesis are described in specific detail in Appendix A.

2.1 Sample collection

Blood samples used in this project were collected from animals maintained as part of the Rylington Merino Project by the Department of Agriculture and Food of Western Australia (DAFWA), Katanning, Western Australia. Blood samples were collected by venepuncture into 10mL K_3EDTA vacutubes (Vacuette) and stored at -20^{o}C. Blood samples were obtained from 3 cohorts of animals:

I. **Cohort 1:** This cohort comprised of 12 unrelated sheep and was used for SNP identification. Approximately 6 mL of blood was collected from each animal.

II. **Cohort 2:** This cohort comprised 38 lambs, obtained from father-daughter matings using artificial insemination (AI), to generate Major Histocompatibility Complex (MHC) homozygous animals. These lambs were generated and maintained by DAFWA. DNA samples from these animals were genotyped using four microsatellite markers SMHCC (Groth & Wetherall 1994), SKIV2LM (Groth & Wetherall 1995), OLADRB (Schwaiger *et al.* 1993) and OLADRBps (Blattman & Beh 1992) to identify animals that were potentially MHC homozygous. Homozygous animals were used for the analysis of MHC class I

loci. Four tubes of approximately 6 mL of blood were collected from each lamb. Two of the tubes were specifically reserved for mRNA isolation.

III. **Cohort 3:** This cohort comprised of 108 animals, among which, 107 had known estimated breeding values (EBV) for clean fleece weight (CFW), fibre diameter (FD) and staple strength (SS). The EBV for was calculated and provided by DAFWA. These animals were used for association of animals with known phenotype.

All animal experiments were performed according to the Australian Code of Practice for the care and use of the animals for scientific purposes. Blood samples were collected under approval of Curtin's Animal Ethics Committee.

2.2 DNA and RNA Extraction

2.2.1 Genomic DNA extraction

Genomic DNA was extracted using the AxyPrep^{TM} Blood Genomic DNA Miniprep Kit (Axygen). Whole sheep blood was used as a starting material without the need to remove the red blood cells. DNA was extracted by using the manufacturer's protocol except that each sample was eluted from the column twice to maximise DNA yield.

2.2.2 Plasmid DNA extraction

Plasmid vectors were used for cloning either digested fragments of Bacterial Artificial Chromosome (BAC) DNA or PCR. Purification of plasmid DNA was performed using Axyprep^{TM} Plasmid Miniprep Kit (Axygen). Bacterial cells containing recombinant plasmids were grown for 18 hours at 37^{o}C with gentle agitation in 5 mL of Luria Bertani (LB) media supplemented with 0.1 mg/mL ampicillin. The plasmids were isolated using the manufacturer's standard protocol.

2.2.3 Bacterial Artificial Chromosome (BAC) DNA extraction

Bacterial Artificial Chromosomes (BACs) used in this project were from the ovine CHORI-243 BAC library constructed by the Children's Hospital Oakland Research Institute (CHORI) from a ram of the Texel breed. CHORI 243-269M18, CHORI 243-454E19, and CHORI 243-360H16 were previously identified to contain sequence from the class I region of the MHC (J. Qin, D. Groth; unpublished 2006). Glycerol stock of a single BAC clone was used to inoculate 5 mL of LB media supplemented with 12.5 µg/mL chloramphenicol and this was incubated for 8 hours at 37^{o}C with vigorous agitation. Five 1 mL aliquots of the starter culture were added into five 100 mL LB supplemented with 12.5 µg/mL chloramphenicol. The culture was grown at 37^{o}C for 12 hours with constant shaking and BAC DNA extracted using QIAGEN® Large-Construct Kit. The standard protocol was used except an additional elution of BAC DNA was performed with 100 µL of pre-warmed TE buffer, pH 8. DNA yield was determined by UV spectrophotometry (at 260 nm) and quantitated on an agarose gel using size standards with a known concentration.

2.2.4 Total RNA extraction

Prior to extraction of total RNA, white blood cells were isolated from fresh blood samples. Whole EDTA blood (10 mL) was centrifuged at 1600 rpm for 15 minutes to separate the white blood cells. The buffy coat containing white blood cells was then transferred into a 3DT tube and 3 mL of TE buffer added and mixed gently. The sample was centrifuged again at 1600 rpm for 15 minutes and the supernatant discarded. The pellet of white blood cells was re-suspended in 500 µL of RNAlater (Sigma) and stored at -20^{o}C until required. Total RNA was extracted from the cells using the Isolate RNA Mini Kit (Bioline), following the standard protocol of RNA isolation from animal tissue.

2.3 Sub-cloning BAC into vector

2.3.1 pGEM® -3Z vector

Vector pGEM® -3Z which was used for sub-cloning BAC was purchased from Promega. PGEM® -3Z was used because it contained the restriction enzyme site compatible with the restriction digested BAC fragments, yields high copy numbers of cloned products, uses blue/ white screening system for recombinants, and harbours M13 forward and reverse primer binding site flanking multiple cloning site, suitable for PCR amplification and direct sequencing. Recombinant plasmid was used to transform into *E. coli* by electroporation.

2.3.2 Restriction and modifying enzyme treatment

Restriction enzyme (*Pst I*) was purchased from Promega and used to cut purified BAC DNA and cloning vector to provide sticky ends for sub-cloning. Total reaction volume and ratio of individual components added to each restriction digest reaction differed slightly depending on the type of DNA being digested. BAC DNA was digested in a 10 µL reaction consisting 8 µL of BAC DNA (500 ng – 1000 ng), 1 µL of restriction enzyme (10 units) and 1 µL of restriction enzyme buffer (10X). The pGEM® vector was digested in a 25 µL reaction as follows; 20 µL of vector (5 µg), 3 µL of restriction enzyme (30 units) and 2 µL of restriction enzyme buffer (10X). The mixture was mixed gently and incubated at 37°C for 2 – 3 hours and follow by heat inactivation of the restriction enzyme at 65°C for 15 minutes. Digested BAC and vector were analysed on agarose gel.

Restriction enzyme digested vector was treated with modifying enzyme Shrimp Alkaline Phosphatase (SAP) (Promega) to catalyse dephosphorylation of 5' phosphates from the pGEM® vector. Dephosphorylation prevents recircularisation and religation of linearised vector and increases ligation efficiency of BAC DNA fragments into vector. Reaction for SAP treatment contained vectors digested with restriction enzyme (5 µg), 10x SAP buffer and SAP enzyme (5 units). The mixture was incubated at 37°C for 30 minutes and subsequently heat-denatured at 65°C

for 15 minutes. To validate the effectiveness of dephosphorylation, SAP treated vector was checked for self-religation using T4 DNA ligase. SAP treated vector (1 µL) was mixed with 7.5 µL sterile water, 1 µL 10x ligase buffer, and 0.5 µL T4 DNA ligase (1.5 units) and incubated for 1 hour at 14^{o}C. SAP and T4 DNA ligase treated vector was analysed on agarose gel.

2.3.3 Ligation

Ligation reaction of 10 µL volume was prepared as follow: 3 µL of restriction enzyme digested BAC DNA was mixed with 1 µL SAP treated pGEM® -3Z vector (50 ng), 4 µL sterile water, 1 µL 10x ligase buffer and 1 µL T4 DNA ligase (3 units) . The reaction was incubated for 15 hours at 14^{o}C to increase the number of transformants.

2.3.4 Ethanol precipitation and purification

Ligation reaction of 10 µL was precipitated with 25 µL 100% ethanol and left in freezer at -20^{o}C for approximately 18 hours, overnight. Sample was centrifuged at 1350 rpm for 30 minutes and discarded the supernatant without disturbing pellet. 125 µL chilled 70% ethanol was added to the pellet, spun again at 1350 rpm for 15 minutes and supernatant discarded. Pellet was air-dried and subsequently re-suspended in 5 µL of sterile water.

2.3.5 Transformation

Transformation of ligated recombinant vector into competent cells was performed by electroporation method using a Biorad Gene Pulser II. ELECTROMAX™ DH5α-ETM *E.coli* cells (20 µL) was pipetted into a sterile chilled 0.65 mL tube and placed on ice. Purified ligation mixture (1.5 µL) was added to the competent cells and gently mixed by pipetting. The entire content of the tube was then transferred into a pre-chilled electroporation cuvette (Cell Projects) and pulsed at setting of 1.8 kV, 25 µF. Immediately after electroporation, 700 µL of SOC medium was added the cuvette and the diluted cells transferred immediately to a sterile 15 mL falcon tube. The culture was incubated at 37^{o}C for 1 hour with gentle shaking. The transformation mixture (100 µL) was plated onto LB agar plates

supplemented with 0.1 mg/mL of ampicillin, 0.1 mM of IPTG and 40 µg/mL of X-GAL. The plates were incubated inverted overnight at 37°C.

2.3.5 Selection of recombinant clones

Bacterial colonies containing a recombinant plasmid were identified as white colonies on the plate. Recombinant clones were picked and streaked onto a gridded plate containing 0.1 mg/mL of ampicillin, 0.1 mM of IPTG and 40 µg/mL of X-GAL. The plates are incubated inverted at 37°C for 18 hours. Each of the recombinant clones on the gridded plate was inoculated in 500 µL of LB containing 0.1 mg/mL of ampicillin and grown at 37°C for 18 hours with shaking. Glycerol (500 µL of a sterile 60% glycerol solution) was added to each culture making a final concentration of 30% glycerol. These glycerol stocks were stored at -80 °C for future use.

2.4 Polymerase chain reaction (PCR)

Polymerase chain reaction (PCR) was used extensively within this project to specifically amplify various targets such as intron and exons within genes as well as intergenic sequences. Various templates were used for PCR and include genomic DNA, BAC DNA, plasmid DNA and cDNA. PCR was performed in 10 µl reactions using the Eppendorf Mastercycler. Each reaction contained 9 µl of master mix and 1 µl of DNA template (100 – 200 ng/µl). The content of the master mix varies according to the type of PCR being performed and specificity of the primers used. Standard master mix contained 200µM of each dNTP (Roche), 0.75mM – 2.95 mM $MgCl_2$ (Roche), 1 X PCR reaction buffer (Roche), 0.2 units of FastStart Taq DNA polymerase (Roche), 1 pM of each primer (Geneworks) and distilled water. The standard PCR protocol used is detailed in Table 2.1. Amplicons were analysed on agarose gel.

Table 2.1: Standard PCR cycling condition

Temperature (°C)	Time	Cycles
95	10 minutes	1
95 45 – 65 72	1 minutes 30 seconds 1 minutes	1
95 45 – 65 72	20 seconds 30 seconds 1 minutes	25 – 35
95 45 – 65 72	20 seconds 30 seconds 5 minutes	1

2.5 PCR clean-up

Prior to the PCR products being sent for sequencing, pre-sequencing clean-up was performed using the ExoSAP method (Bell 2008). Exonuclease I (0.5 µL or 10 units) (New England Biolabs) and 0.56 µL of Shrimp Alkaline Phosphotase (2.8 units) (New England Biolabs) were added directly to 20 µL of PCR product (pool of 2 x 10 µL PCR product). The mixture was incubated at 37°C for 30 minutes, followed by heat inactivation at 80°C for 20 minutes.

2.6 Cloning PCR product

PCR amplicons from multi-copy gene amplified using universal primers were not suitable for direct sequencing. Such PCR products were cloned into cloning vectors following the standard protocol of pGEM-T Easy Vector Systems (Promega). Recombinant clones identified as white colonies on plates were individually grown in 500 µL of LB media supplemented with 0.1 mg/mL ampicillin at 37°C for 18 hours. The overnight cultures were diluted into ultra-pure water at a ratio of 1:100 (culture: water) and subjected to PCR amplification. The PCR screening was performed using either of the following primers sets; universal primers used to PCR amplify the multi-copy gene, or primers designed to specifically amplify a small fragment the gene if the cloned insert was large. Positives clones were purified using Axyprep™ Plasmid Miniprep Kit (Axygen) and sequenced.

2.7 Sequencing

DNA sequencing was performed by Macrogen Inc. (Korea) using an ABI 3730XL sequencer. The following products were sequenced; purified PCR product, PCR amplicons cloned into plasmid vectors and recombinant plasmid inserts generated from sub-cloning of BACs containing sheep DNA. Generally, the sequencing of plasmid DNA fragments was performed using standard universal primers; M13 forward or M13 reverse. However, the direct sequencing of purified PCR products was performed with the same specific primers used for the PCR amplification. Internal primers were used to sequencing DNA fragments larger than 1000 bp to ensure contiguous and reliable sequence. Concentration of the templates and primers submitted to Macrogen Inc was 10 ng/µL and 5 pmol/µL respectively.

2.8 Agarose gel electrophoresis

The quality of DNA extractions and PCR products were analysed using agarose gel electrophoresis. Agarose gel was prepared with 1 X Tris-acetate-EDTA (TAE) buffer (Appendix A). The percentage of gel that was used depended on the fragment size of product analysed; 0.7% w/v agarose gel was used for large size fragments such as DNA and plasmid extractions, whereas PCR amplicons ranging from 200 bp to 3 kbp were separated on 1.5% w/v gel. Samples and loading buffer (Appendix A) were mixed at a ratio of 1:0.2 by volume prior to loading onto the gel. Samples were electrophoresed at 80 – 100 V for 45 minutes. Migrated samples were then stained with ethidium bromide and visualised with Bio-Vision Gel Documentation (GelDoc) System. Size standards were used on each gel for confirmation of the expected product size.

2.9 Sequence analysis

Vector NTI® (Invitrogen) software was used for contig assembly and viewing of the sequencing chromatograms and to identify single nucleotide polymorphisms (SNPs). SNPs were identified as double peaks

(heterozygous) in a sequence and confirmed by comparison with other sequences at the same locus (Figure 2.1). Multiple sequence alignment and phylogenetic analysis were predominantly performed using ClustalW (Thompson *et al.* 1994). TreeView (Page 1996) was used to view phylogenetic trees generated by ClustalW.

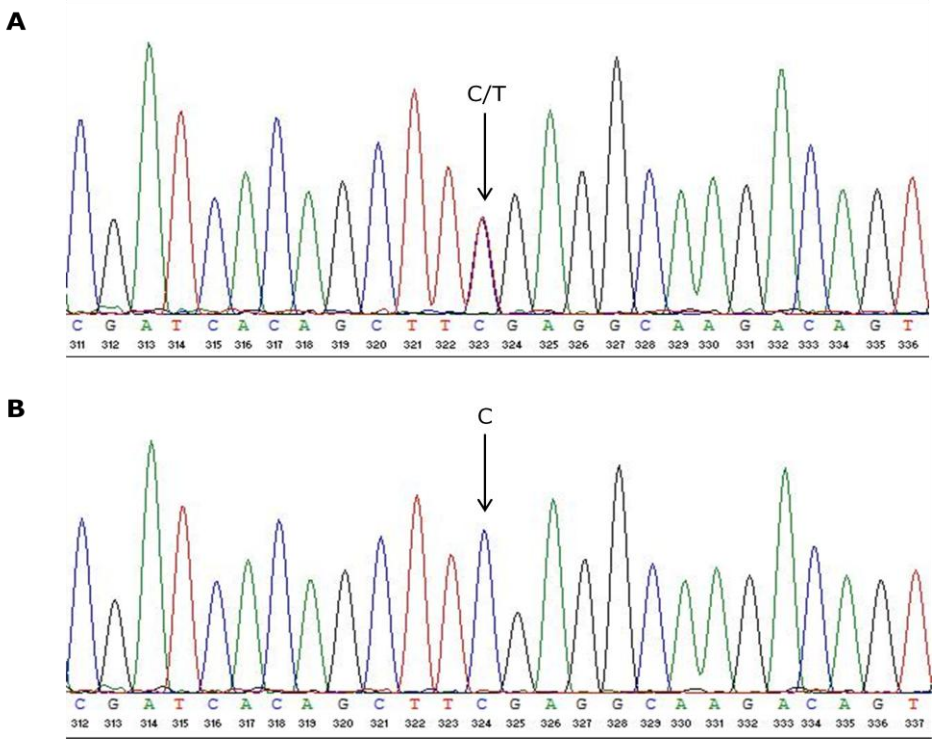


Figure 2.1: Comparison of sequences for SNP identification. (A) Chromatogram of a sequence with a double peak (heterozygous) at a specific locus. (B) The same locus in different individual has single peak (homozygous).

Chapter 3

Structure and Organisation of Sheep MHC Class I Region

This chapter describes the structure and gene organisation of sheep MHC class I region and identifies genetic markers within this region for association studies. When this study commenced in 2007, there was limited knowledge about the gene content within the sheep MHC class I region because the sheep genome had not been sequenced or annotated, and the cattle MHC Class I map was used as reference. CHORI bacterial artificial chromosomes (BACs) known to contain sequences from the sheep MHC class I region were sub-cloned, and the clones sequenced. The resulting sequences were analysed for sequence similarity with the cattle MHC and re-assembled to identify the gene content and organisation within each BAC. More recently a research group in China published the first sheep MHC map based on BAC clones derived from a Chinese Merino sheep and the sequences were made available on GenBank for public access (Gao et al. 2010). These sequences from GenBank were re-assembled to compare the sheep MHC map published by Gao et al. (2010) and the map constructed in this study.

3.1 Introduction

Characterisation of sheep MHC has been based predominantly on analysis of orthologous loci from the human and cattle MHCs. Early studies presumed that the basic structure of sheep MHC was similar to other mammals, consisting of the telomeric class I, central class III and centromeric class II. A later study of MHC structure in Chinese Merino sheep reported that the class II region is sub-divided into IIa and IIb due to a chromosomal inversion, a characteristic initially identified in cattle but not in other mammals (Childers *et al.* 2006; Liu *et al.* 2006). In recent years, low resolution physical maps of sheep MHC class II and III regions have been

constructed using a combination of sub-cloning and partial sequencing of Bacterial Artificial Chromosome (BAC) clones that were known to contain MHC sequences (Qin *et al.* 2008; Lee *et al.* 2011). In addition, a panel of single nucleotide polymorphisms (SNPs) spanning the sheep MHC class II and III regions have been developed to provide a framework for the identification and analysis of haplotypes for future use in association studies.

The human MHC map shows that the class I region is rich in pseudogenes, duplicated genes and genes showing copy number variations (Horton *et al.* 2004). There is a paucity of knowledge about the sheep MHC class I in terms of gene content, structural organisation and genetic variation. For instance there is a lack of understanding of its haplotypic structure. The dinucleotide microsatellite loci OHCCI (Groth & Wetherall 1994) has been the only non-classical polymorphic genetic marker frequently used in association studies in sheep (Gruszczynska *et al.* 2002; Worley *et al.* 2006; Bozkaya *et al.* 2007; Kaeuffer *et al.* 2007; Petroli *et al.* 2009).

A physical map of sheep MHC derived from a Chinese Merino sheep has been published by Gao and colleagues (Gao *et al.* 2010). However this map would seem to be ambiguous and poses new questions. Overlapping sequences between BAC sequences are not identified. Such overlaps would have enabled confirmation of the final contiguous architecture reported by this group (Gao *et al.* 2010). Furthermore, little evidence was provided for the accuracy of the specific gene identification procedures used and these are not reported for individual BACs. The research described in this chapter is intended to supplement the map proposed by Gao and colleagues and confirm the relative positions of the loci identified.

The aim of this study is to sub-clone CHORI BACs known to contain class I sequence and re-assemble the sequences to identify genes present within the sheep MHC class I region. The sequences from CHORI BACs will be used to generate a contig map for comparison with the cattle MHC class I map. The low resolution physical map constructed from CHORI BACs will also be used for comparison with the recently published sheep MHC map of Gao *et al*. (2010).

3.2 Materials and methods

3.2.1 Sub-cloning of BAC DNA

Three BAC clones (CHORI 243-269M18, CHORI 243-390H16 and CHORI 243-454E19) previously identified to contain MHC class I sequence by J. Qin/D.Groth (unpublished 2006) had their DNA extracted using the QIAGEN® Large-Construct Kit (Chapter 2.2.3). BAC DNA isolated from each of clones was digested with *Pst I* restriction enzyme (Promega) and sub-cloned into the *Pst I* site of the pGEM® vector (Chapter 2.3). Thirty to fifty random recombinant clones from each BAC were then sequenced using the standard universal M13 forward and reverse primers, to ensure good quality double pass sequences were obtained (Chapter 2.7).

3.2.2 Analysis of CHORI BAC sub-clones

The quality of CHORI BAC DNA sequences were analysed using Vector NTI® software (default setting). The sequences were aligned using BLAST to determine their location relative to the cattle reference genomic sequence database; NW_001494164.1. Cattle genome was used as reference for mapping the sequences and SNPs because a contiguous sheep map was not available when this work commenced and therefore, the location of the sequences relative to each other or with existing genes had not been assigned. The BAC sequences were assembled to form a contig map with cattle MHC as reference. The boundaries of each BAC were determined using the end sequences available in GenBank database. The map was then compared with the newly published sheep MHC map by Gao *et al*. (2010).

3.2.3 Re-analysis of Gao's sheep MHC map

Twenty BAC sequences representing the class I, II, and IIa regions in the sheep MHC map proposed by Gao *et al*. (2010) were downloaded from NCBI and re-assembled using Geneious Pro 5.5 (Drummond *et al*. 2011 - unpublished) in an attempt to form a contiguous assembly. The BAC sequences used for this analysis were: FJ985852, FJ985853, FJ985854, FJ985856, FJ985857, FJ985859, FJ985861, FJ985862, FJ985864, FJ985865,

FJ985866, FJ985867, FJ985868, FJ985869, FJ985870, FJ985872, FJ985873, FJ985874, FJ985875 and FJ985876. At the time of the analysis, version 1 of each BAC was available (uploaded 14 October, 2010) and identified as a working draft. The outcome of the assembly analysis was compared to the published sheep MHC map (Gao *et al.* 2010). Further analyses were performed to rectify discrepancies between the published map and the result obtained from assembly of BAC sequences as follows: Potential overlaps between BAC sequences were determined using the NCBI BLAST option to align two sequences (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Sequences that overlapped at the 5' or 3' ends were subsequently aligned with the CHAOS/DIALIGN software (Brudno *et al.* 2004) provided at http://dialign.gobics.de/chaos-dialign-submission. Alignments were examined and edited where required using Seaview 4.2.12 (Gouy *et al.* 2010) in order to provide an optimal alignment and precisely determine overlap boundaries. Thirteen of the BAC sequences examined were reverse complemented before pairwise sequence alignment in order to provide a contiguous assembly in the telomeric to centromeric direction as per the published map by Gao *et al.* (2010).

Ten of the BAC sequences published by Gao *et al*. (2010) proposed to cover the MHC Class I region were further analysed for gene content, these were: FJ985869, FJ985854, FJ985864, FJ985870, FJ985873, FJ985868, FJ985852, FJ985875, FJ985859 and FJ985874. Of the ten sequences, seven were reverse complemented before analysis in order to facilitate mapping in a contiguous 5' to 3' direction. Gene content analysis proceeded as follows: BAC sequences were masked for repeats with Repeatmasker, open version 3.2.9, then analysed with GENSCAN (http://genes.mit.edu/GENSCAN.html) and Softberry FGENESH (http://linux1.softberry.com/all.htm). Predicted transcripts were submitted to the NCBI BLAST to identity putative gene transcripts by homology to known genes previously reported in mammalian species, in particular *Ovis aries* or *Bos taurus*. To refine predictions for putative genes, BAC sequences were subsequently analysed with FGENESH+ using one or more of the best matching proteins as a homologue. *Bos taurus* was chosen as the model organism for both FGENESH and FGENESH+ and up to five variant transcripts were considered. In the case where there appeared to be multiple copies of the same or a similar gene in

a single BAC, FGENESH+ gene prediction was localised to each particular region of interest. The most suitable transcript for each gene was selected based on alignment with known genes from *Ovis aries* (when available), *Bos taurus*, *Sus scrofa* and *Homo sapiens*.

3.3 Results

3.3.1 Analysis of end sequences of CHORI BACs

Assembly of sub-cloned BAC sequences and BLAST analysis of end sequences revealed the gene content and coverage of each BAC. Table 3.1 shows the details of end sequence for each BAC and its position within the cattle reference sequence, NW_001494164.1. CHORI 243-390H16 extended from BOLA-NC1 (~691 kbp) to tubulin (~467 kbp). One end of CHORI 243-454E19 was located between loci *DDR1* and *IER3* (~434 kbp), whereas the other end located close to *RPP21* (~641 kbp). CHORI 243-390H16 spanned from LOC788634 (~110 kbp) to approximately close to LOC787188 (~228 kbp).

Table 3.1: Result of BLAST analysis of end sequences of BACs downloaded from GenBank. The sequences were BLAST with reference genomic sequence database; NW_001494164.1.

BAC ID	**End sequence GenBank ID**	**Location within NW_001494164.1**	**Description of locus**
CH243-390H16	DU202647.1	690703 - 691106	Close to non-classical MHC class I antigen (*BOLA-NC1*)
CH243-390H16	DU201205.1	466593 - 467149	tubulin, beta 2B (*TUBB*)
CH243-454E19	DU262410.1	433666 - 434658	Between discoidin domain receptor family, member 1 (*DDR1*) and immediate early response 3 (*IER3*)
CH243-454E19	DU252291.1	640354 - 641212	Close to ribonuclease P 21 (*RPP21*)
CH243-269M18	DU420388.1	109388 - 110185	Protein similar to non-classical MHC class I antigen (*LOC788634*)
CH243-269M18	DU418632.1	228169 - 228975	Close to hypothetical protein *LOC787188*

3.3.2 Re-assembly and analysis of CHORI BAC contigs map

CHORI 243-390H16 overlaps with CHORI 243-454E19 and were located in the middle of MHC class I region whereas, CHORI 243-269M18 were located further away from the telomeric end. Comparative analysis and sequence similarity with reference to the cattle MHC map indicated the estimated locations of the CHORI BACs in the sheep MHC map. These estimated positions were subsequently confirmed to be concordant with the Virtual Sheep Genome data (http://www.livestockgenomics.csiro.au/perl/gbrowse.cgi/vsheep2/).

Analysis of CHORI BAC sub-clone sequences identified 10 additional loci in the sheep MHC class I region that were not present in the recently published sheep MHC map (Gao *et al.* 2010) but present in the cattle MHC map. The additional loci include ribonuclease P 21-like isoform 2 (*RPP21*), guanine nucleotide-binding protein-like 1 (*GNL1*), ATP-binding cassette sub-family F

member 1 (*ABCF1*), chromosome 6 open reading frame 136 ortholog (*C23H6orf136*), DEAH (*Asp-Glu-Ala-His*) box polypeptide 16 (*DHX16*), nuclear envelope membrane protein or nurim (*NRM*), general transcription factor II H subunit 4 (*GTF2H4*), surfactant associated protein G (*SFTPG*), transcription factor 19 (*TCF19*) and POU class 5 homeobox 1 (*POU5F1*). Comparison of MHC class I region between cattle and sheep is illustrated in Figure 3.1. The relative locations of the additional loci were deduced based on the comparison of gene content and arrangement between BAC contigs map generated from assembly of BAC sub-cloned sequences and sheep MHC class I region. Figure 3.2 shows the possible positions of the additional loci identified through sequencing in the sheep MHC class I map.

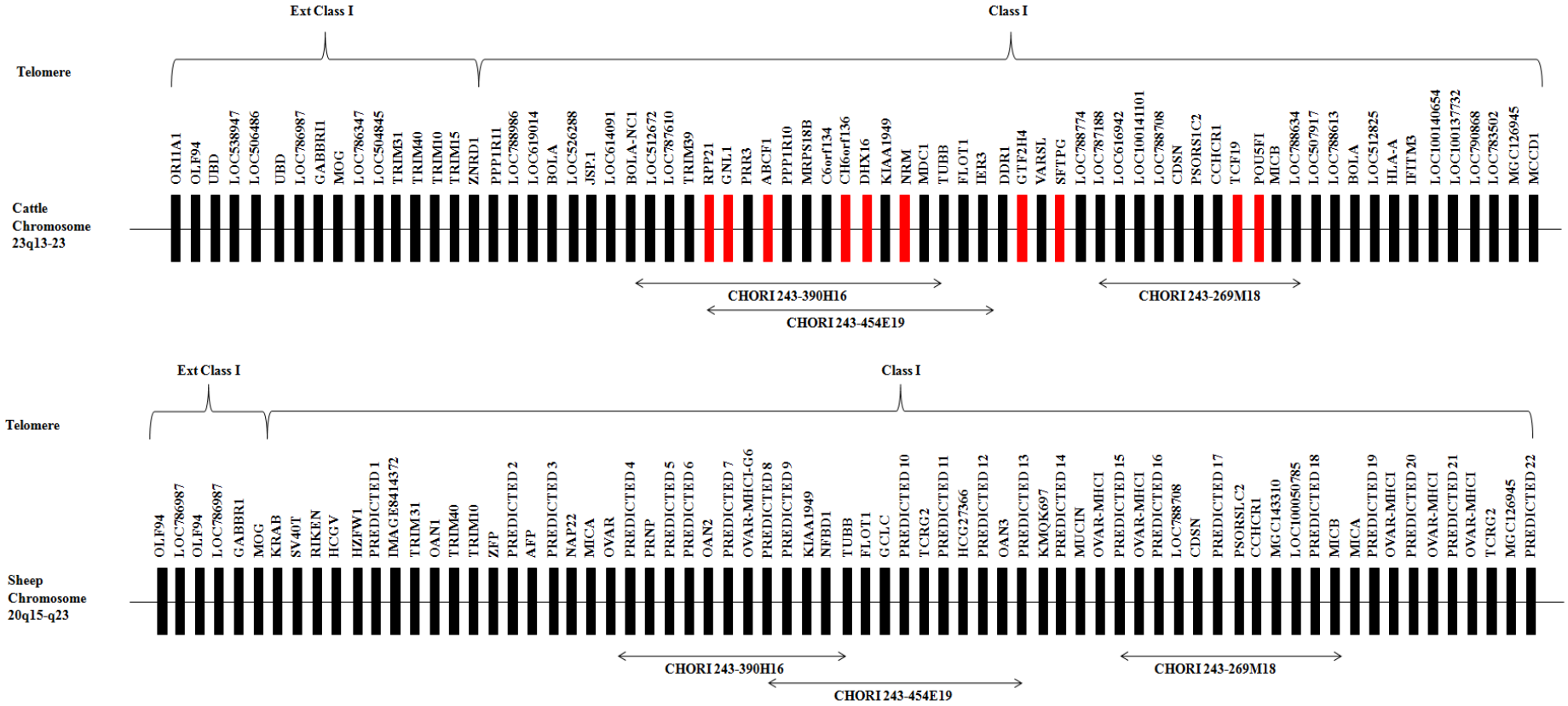


Figure 3.1: Comparison of sheep MHC class I map proposed by Gao *et al.* (2010) and cattle MHC class I map generated by NCBI Map Viewer. The loci highlighted in red in cattle map indicate loci identified within the CHORI 243-269M18, CHORI 243-390H16 and CHORI 243-454E19 but, not present in the sheep map.

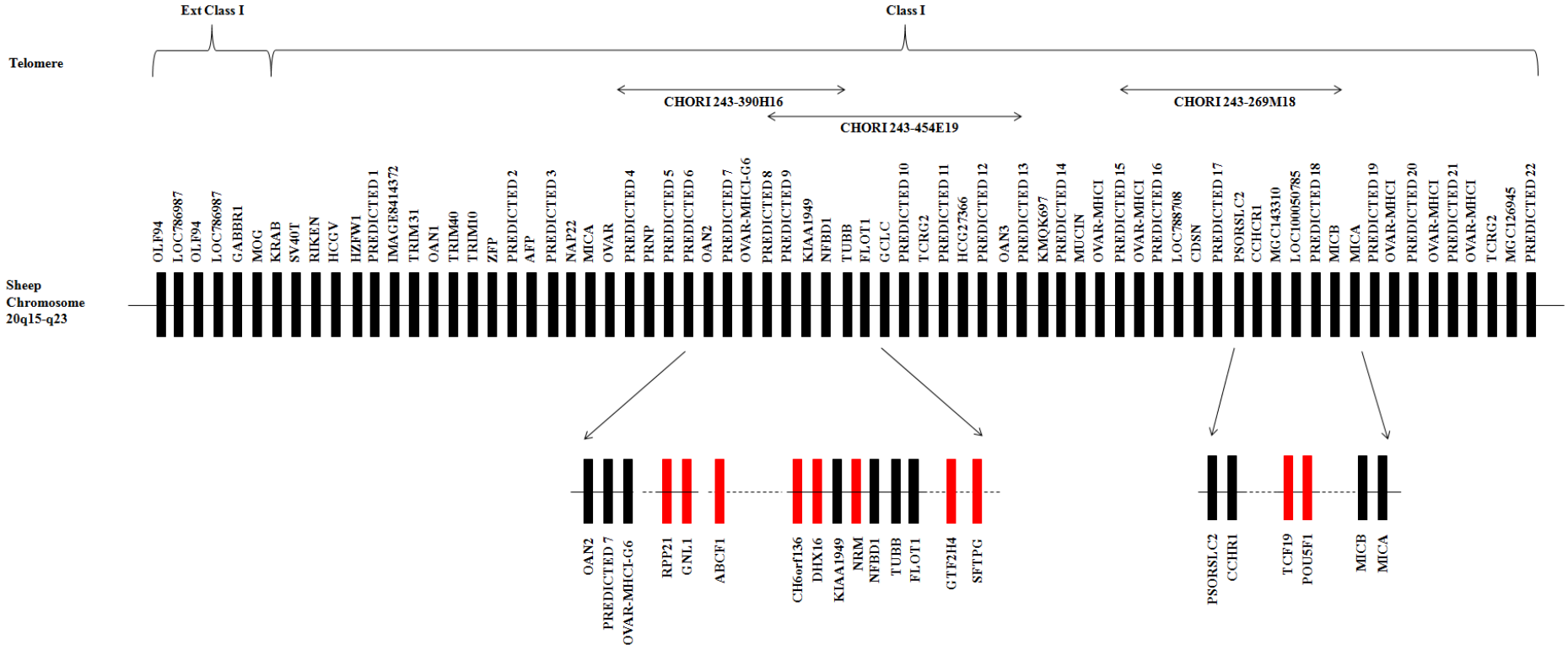


Figure 3.2: Identification of 10 loci in the MHC class I region through sub-cloning of CHORI 243-269M18, CHORI 243-390H16 and CHORI 243-454E19 and their relative position within the map proposed by Gao *et al*. (2010). Identity of these loci was confirmed by sequence homology using NCBI BLAST. Loci highlighted in red indicate loci previously not present on map proposed by Gao *et al*. (2010).

3.3.3 Re-analysis of Gao's sheep MHC map

3.3.3.1 Re-assembly of contig tiling path

Analysis with Geneious 5.5 software produced the 5 contigs listed in Table 3.2 instead of the single contig reported by Gao *et al*. (2010). Four reads were not incorporated into the assembly; FJ985852, FJ985862, FJ985865 and FJ985867. Comparing with Gao's map in the 5' to 3' direction: Contig 2 assembles four reads - FJ985869, FJ985854, FJ985864 and FJ985870. Geneious does not detect overlap between FJ985873 and either FJ985864 or FJ985870 as indicated in Gao's map (Figure 3.3). Contig 3 assembles 3 reads - FJ985873, FJ985868 and FJ985875. Geneious fails to include FJ985852 in an assembly with FJ985868 and FJ985875 as would be expected from Gao's map. Geneious fails to detect overlap between FJ985875 and FJ985859. Contig 4 assembles FJ985874 and FJ985859. Geneious fails to detect overlap between FJ985874 and FJ985856. Contig 1 assembles FJ985856, FJ985861, FJ985872, FJ985857 and FJ985853. Geneious fails to detect overlap between FJ985853 and FJ985867, FJ985867 and FJ985862, FJ985862 and FJ985866. Contig 5 assembles FJ985876 and FJ985866. Geneious fails to detect overlap between FJ985876 and FJ985865. Contigs 2, 3 and 4 are in the opposite orientation when compared to Gao's map.

Table 3.2: Geneious assembly of 20 BAC clones published by Gao *et al*. (2010).

Assembled	**Length**	**Reads Unused**	**Length**
Contig 1 – 5 Reads	**629701**	FJ985852	118738
FJ985856 (f)	127050	FJ985862	167309
FJ985861 (f)	162317	FJ985865	159959
FJ985872 (r)	169910	FJ985867	133881
FJ985857 (r)	165531		
FJ985853 (f)	134434		
Contig 2 - 4 Reads	**430168**		
FJ985870 (r)	138311		
FJ985864 (f)	142360		
FJ985854 (f)	145292		
FJ985869 (f)	134643		
Contig 3 – 3 Reads	**460095**		
FJ985875 (f)	173955		
FJ985868 (f)	140835		
FJ985873 (r)	196844		
Contig 4 – 2 Reads	**283944**		
FJ985874 (r)	141902		
FJ985859 (f)	160643		
Contig 5 – 2 Reads	214322		
FJ985876 (f)	88495		
FJ985866 (f)	155022		

(f) indicates assembly of sequence in the forward (5'->3') direction. (r) indicates assembly of the reverse complement sequence (3'->5').

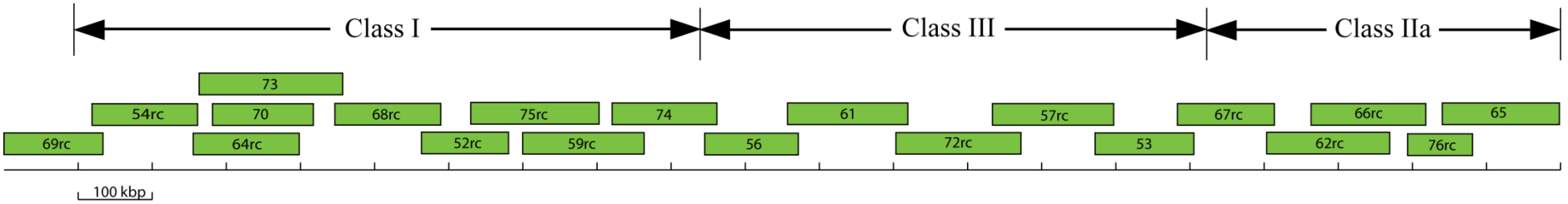
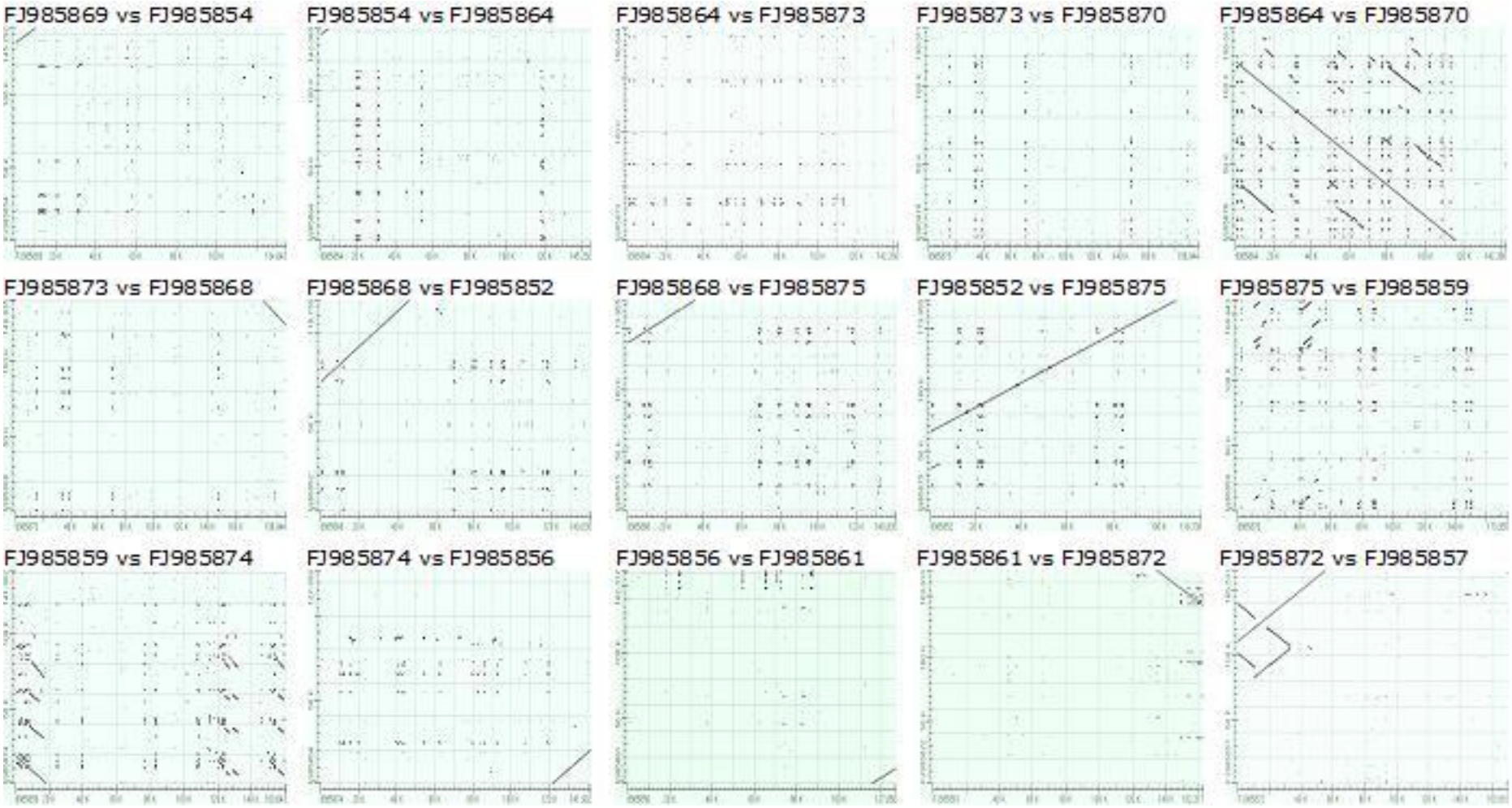


Figure 3.3 MHC sheep contig tiling map as published by Gao *et al*. (2010).

A series of dotplots illustrating results of pairwise alignment by the NCBI BLAST program is shown in Figure 3.4. The 5′ end of FJ985869 has a significant overlap with the 3′ end of FJ985854 in a plus/plus orientation. The 5′ end of FJ985854 has a short but significant overlap near the 3′ end of FJ985864 in a plus/plus orientation; however, the alignment ends at bp 141596 instead of at the 3′ end of the sequence (142360) as would be expected in a contiguous assembly as indicated in Gao's map. There is no signficant overlap between FJ985873 and either FJ985864 or FJ985870, contrary to what is shown in Gao's map. The 5′ end of FJ985864 (1-~116,000) overlaps with the 5′ end (1-~116,000) of FJ985870 in a plus/minus orientation. The 3′ end of FJ985873 overlaps with the 3′ end of FJ985868 in a plus/minus orientation. The 5′ end of FJ985868 overlaps with the 3′ end of FJ985852 in a plus/plus orientation. The 5′ end of FJ985868 overlaps with the 3′ end of FJ9858675 in a plus/plus orientation. The 5′ end of FJ985852 overlaps with the 3′ end of FJ9858675 in a plus/plus orientation. There is no significant overlap between FJ985875 and FJ985859, contrary to what is shown in Gao's map. The 5′ end of FJ985859 overlaps with the 5′ end of FJ985874 in a plus/minus orientation. The 3′ end of FJ985874 overlaps with the 5′ end of FJ985856 in a plus/plus orientation. The 3′ end of FJ985856 overlaps with the 5′ end of FJ985861 in a plus/plus orientation. The 3′ end of FJ985861 overlaps with the 3′ end of FJ985872. The orientation is plus/minus. The 5′ end of FJ985872 overlaps with the 3′ end of FJ985857 in a plus/plus orientation. The 5′ end of FJ985857 overlaps with the 5′ end of FJ985853 in a plus/minus orientation. The 3′ end of FJ985853 overlaps with the 3′ end of FJ985867 in a plus/minus orientation. There is no overlap between FJ985867 and FJ985862 contrary to what is shown in Gao's map. The 5′ end of FJ985862 overlaps with the 3′ end of FJ985866 in a plus/plus orientation. The 5′ end of FJ985866 overlaps with the 3′ end of FJ985876 in a plus/plus orientation. The 5′ end of FJ985876 overlaps with the 3′ end of FJ985865 in a plus/plus orientation. The overlap is dubious since there are several breaks in it. All alignments in the overlap region are on the order of 94% identity with numerous small indels and base substitutions. This is not likely to represent a true overlap of contigs, contrary to what is shown in Gao's map. There is no overlap between FJ985865 and FJ985855, as expected according to Gao's map. The 3′ end of FJ985855 overlaps with the 5′ end of FJ985860 (52604 bp). Finally, there is

a short overlap of 4401 bp between the 3’ end of FJ985860 and the 5’ end of FJ985871 in the plus/plus orientation. The overlaps between BAC sequences FJ985855, FJ985860 and FJ985871, which represent the class IIb region, are in agreement with Gao’s map.


FJ985869 vs FJ985854
FJ985854 vs FJ985864
FJ985864 vs FJ985873
FJ985873 vs FJ985870
FJ985864 vs FJ985870
FJ985873 vs FJ985868
FJ985868 vs FJ985852
FJ985868 vs FJ985875
FJ985852 vs FJ985875
FJ985875 vs FJ985859
FJ985859 vs FJ985874
FJ985874 vs FJ985856
FJ985856 vs FJ985861
FJ985861 vs FJ985872
FJ985872 vs FJ985857

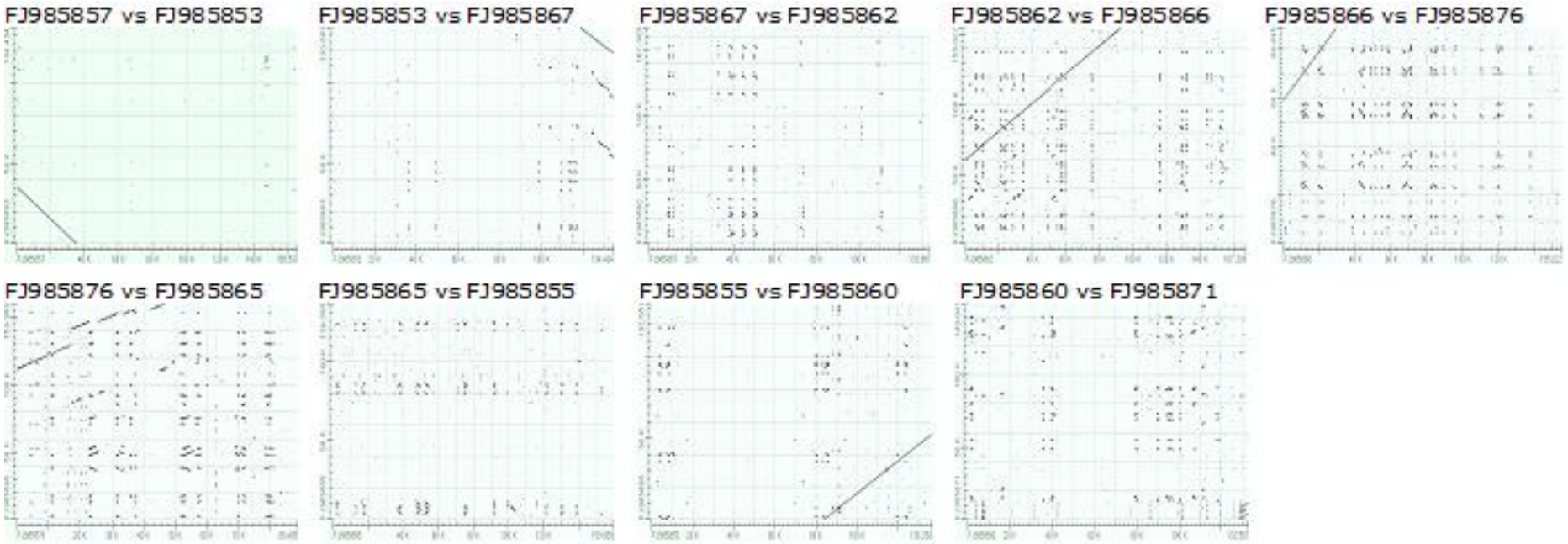
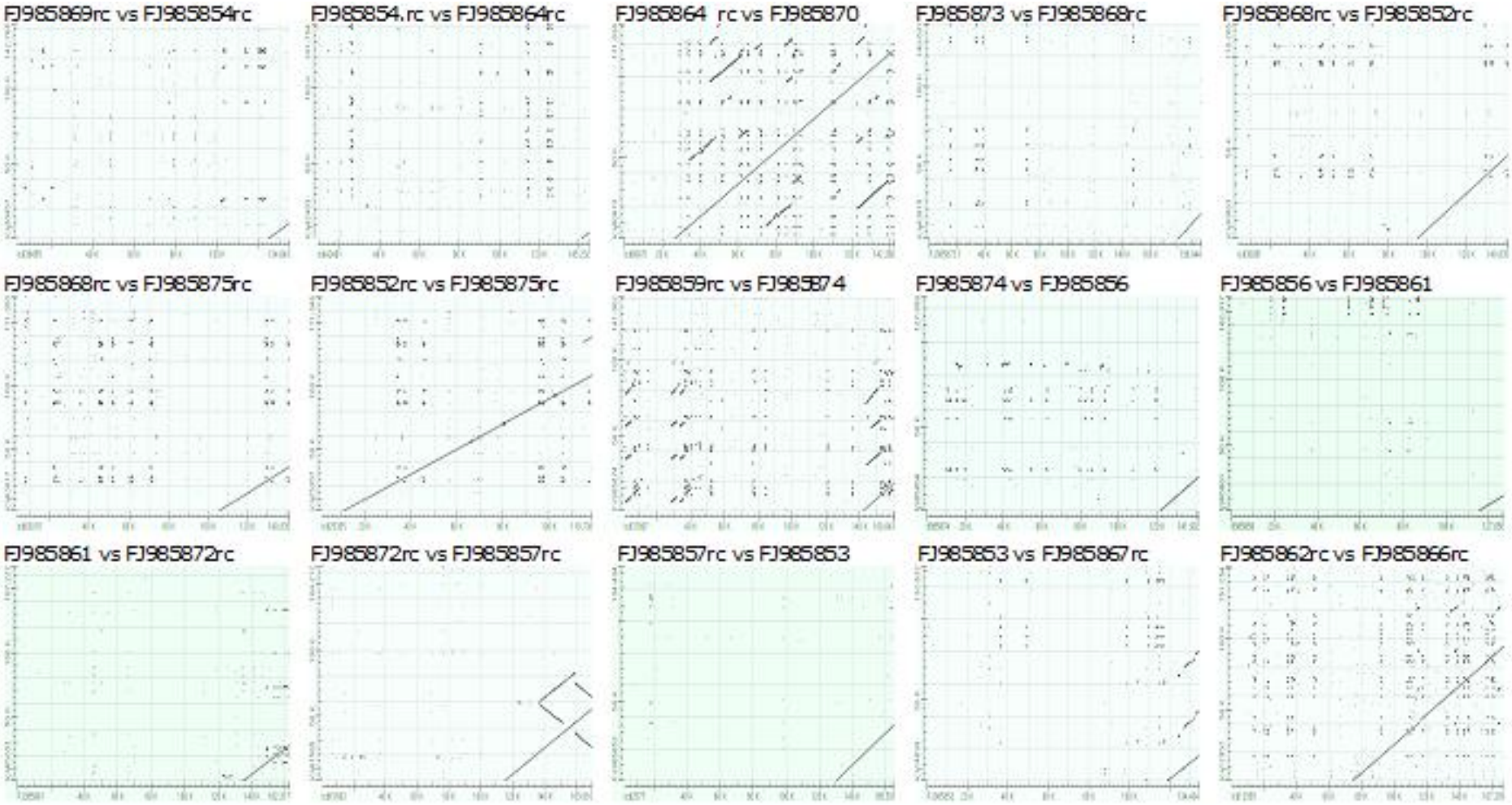


Figure 3.4: Dot plots of BAC sequence contigs published in the NCBI database by Gao *et al*. (2010). X axis represents the first sequence named in the 5'->3' direction left to right. The y axis represents the second sequence in the 5'->3' direction from bottom to top. Unbroken diagonal lines in the corners of the axes are indicative of overlap regions between BAC sequences. Diagonals drawn up and to the right indicate alignments in the plus/plus orientation. Diagonals drawn down and to the right indicate alignments in the plus/minus orientation.

BAC sequences that were in 3'-5' strand were reverse complemented and the dotplot analysis repeated for all of Gao's BAC sequences. Figure 3.5 shows the series of dotplot for the BACS in correct (telomeric to centromeric) orientation. The 3' end of FJ985869_rc aligns with the 5' end of FJ985854_rc in a plus/plus orientation with an overlap of 10399 bp. The 3' end of the reverse complement of FJ985854 has a significant pairwise alignment near the 5' end of the reverse complement of FJ985864 in a plus/plus orientation, but the alignment does not begin until bp 765 in FJ985864. This may represent a regional duplication. The 3' end of the reverse complement of FJ985864 aligns with the 5' end of FJ985870 in a plus/plus orientation with an overlap of 116126 bp. The 3' end of FJ985873 aligns with the 5' end of the reverse complement of FJ985868 in a plus/plus orientation with an overlap of 16758 bp. The 3' end of the reverse complement of FJ985868 aligns with the 5' end of the reverse complement of FJ985852 in a plus/plus orientation with an overlap of 46,198 bp. The 3' end of the reverse complement of FJ985868 aligns with the 5' end of the reverse complement of FJ9858675 in a plus/plus orientation with an overlap of 35324 bp. The 3' end of the reverse complement of FJ985852 aligns with the 5' end of the reverse complement of FJ9858675 in a plus/plus orientation with an overlap of 108,231 bp. The 3' end of the reverse complement of FJ985859 aligns with the 5'end of FJ985874 in a plus/plus orientation with an overlap of 18,619 bp. The 3' end of FJ985874 aligns with the 5' end of FJ985856 in a plus/plus orientation with an overlap of 20,418 bp. The 3' end of FJ985856 aligns with the 5' end of FJ985861 in a plus/plus orientation with an overlap of 11759 bp. The 3' end of FJ985861 aligns with the 5' end of the reverse complement of FJ985872 in a plus/plus orientation with an overlap of 27705 bp. The 3' end of the reverse complement of FJ985872 aligns with the 5' end of the reverse complement of FJ985857 in a plus/plus orientation with an overlap of 55107 bp. The 3' end of the reverse complement of FJ985857 aligns with the 5' end of FJ985853 in a plus/plus orientation with an overlap of 35410 bp. The 3' end of FJ985853 aligns with the 5' end of the reverse complement of FJ985867 in a plus/plus orientation with an overlap of 15709 bp. The 3' end of the reverse complement of FJ985862 aligns with the 5' end of the reverse complement of FJ985866 in a plus/plus orientation with an overlap of 94777 bp. The 3' end of the reverse complement of FJ985866 aligns with the 5' end of the reverse complement

of FJ985876 in a plus/plus orientation with an overlap of 29542 bp. The 3' end of the reverse complement of FJ985876 aligns with the 5' end of the reverse complement of FJ985865 in a plus/plus orientation. However, there are several large indels in the diagonal region of overlap. All pairwise alignments in the overlap diagonal share only approximately 94% identity and show a number of small indels and substitutions. In the Class IIb region, the 3' end of FJ985855 aligns with the 5' end of FJ985860 with an overlap of 52604 bp, and the 3' end of FJ985860 aligns with the 5' end of FJ985871 with an overlap of 4401 bp. The regions of overlaps between the BAC sequences are detailed in Table 3.3.


FJ985869rc vs FJ985854rc
FJ985854.rc vs FJ985864rc
FJ985864 rc vs FJ985870
FJ985873 vs FJ985868rc
FJ985868rc vs FJ985852rc
FJ985868rc vs FJ985875rc
FJ985852rc vs FJ985875rc
FJ985859rc vs FJ985874
FJ985874 vs FJ985856
FJ985856 vs FJ985861
FJ985861 vs FJ985872rc
FJ985872rc vs FJ985857rc
FJ985857rc vs FJ985853
FJ985853 vs FJ985867rc
FJ985862rc vs FJ985866rc

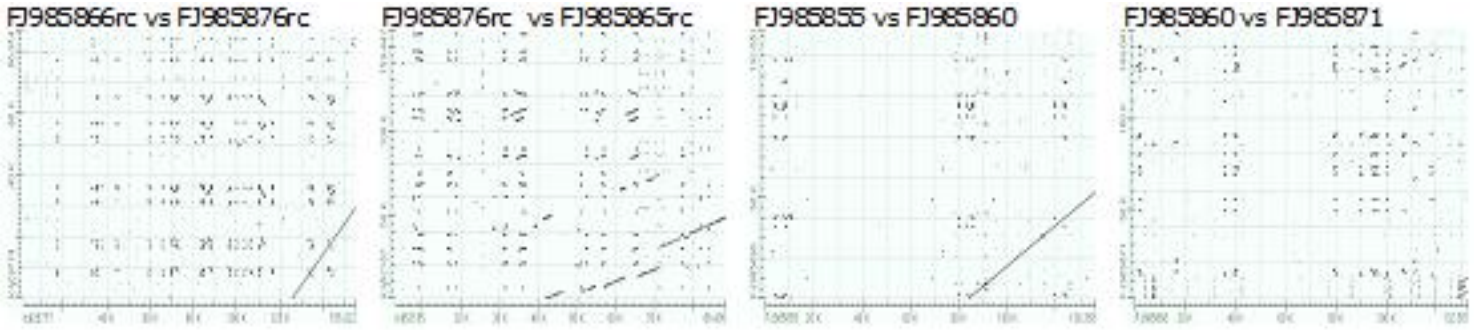


Figure 3.5: Dot plots of Gao's BAC sequence contigs in a telomeric to centromeric (5' to 3') orientation. Dotplots with no potential overlap are not shown in this figure. The text 'rc' appended to the accession number indicates the sequence was reverse complemented before alignment.

Table 3.3: Overlapping regions of Gao's BACs in a telomeric to centromeric (5' to 3') orientation.

BAC IDs	**3' Loc**	**5' Loc**		**BAC IDs**	**3' Loc**	**5' Loc**
69rc x 54rc	124243-134643	1-10399		56 x 61	115294-127050	1-11759
54rc x 64rc	140977-145292	765-5061*		61 x 72rc	134589-162317	1-27705
64rc x 70	-	-		72rc x 57rc	114820-169910	1-55107
64rc x 73	-	-		57rc x 53	130181-16553	1-35410
70 x 73	26368-142360	1-116126		53 x 67rc	118698-134434	1-15709
73 x 68rc	179850-196844	1-16758		67rc x 62rc	-	-
68rc x 52rc	94483-140835	1-46198		62rc x 66rc	74418-167309	1-94777
68rc x 75rc	105191-140835	1-35324		66rc x 76rc	125570-155022	1-29542
52rc x 75rc	10710-118738	1-108231		76rc x 65rc	-	-
75rc x 59rc	-	-		65rc x 55	NA	NA
59rc x 74	142004-160643	1-18619		55 x 60	83750-136358	1-52604
74 x 56	121429-141902	1-20418		60 x 71	128151-132551	1-4401

Precede BAC ID numerals with 'FJ9858' to determine NCBI accession. Appended 'rc' indicates the BAC was reverse complemented before alignment. '-' indicates there was no overlap identified. NA indicates no overlap was reported by Gao *et al.* (2010). *First 765 bp of FJ985864 do not overlap with FJ985854.

A pictorial view of the revised tiling map for Gao's BAC sequences is shown in Figure 3.6. Based on this analysis, there is a potential gap between BACs FJ985854 and FJ985864, and definitive gaps between BACs FJ985870 and FJ985873, FJ985875 and FJ985859, FJ985867 and FJ985862 and FJ985876 and FJ985865.

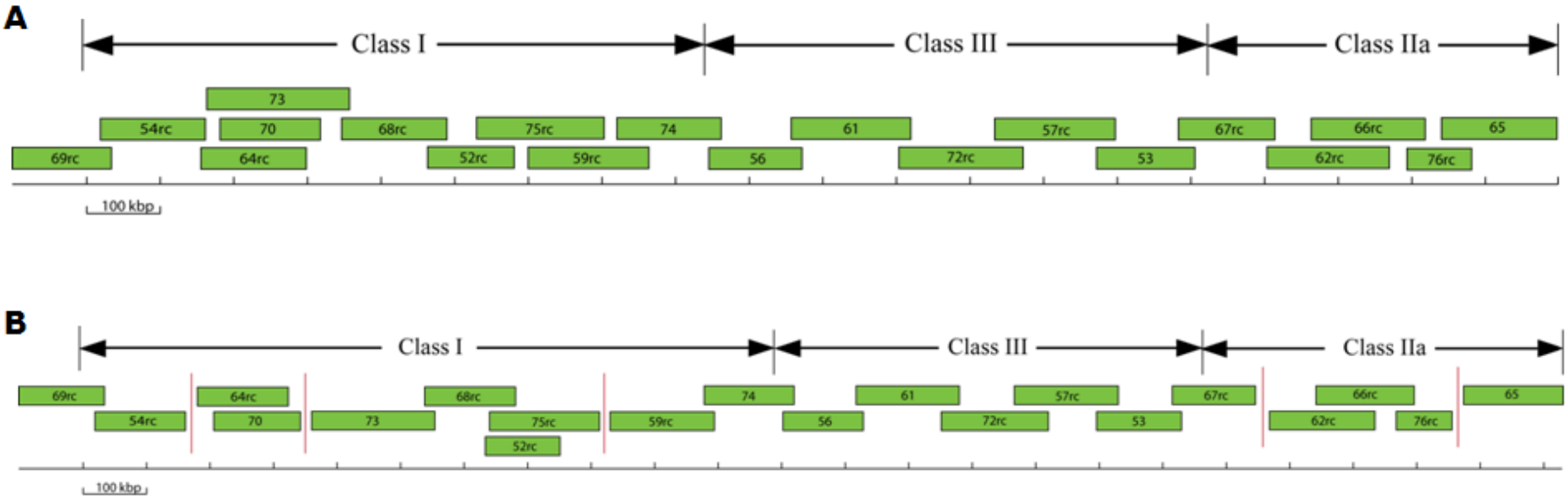


Figure 3.6: Comparison of the map published by Gao *et al*. (2010) and the new map proposed in this study. (A) Original tiling path published by Gao's map (Gao *et al*. 2010). (B) New tiling path of Gao's BAC sequences proposed in this study. A red vertical line between BAC contigs indicates a gap.

3.3.3.2 Identification of gene content

Table 3.4 lists the name and location of genes identified in Gao's BAC sequences. The majority of genes identified appear to have the same gene arrangement as that identified in cattle, along with a high level of similarity to their cattle orthologues. Identical gene predictions between adjacent BAC sequences corroborates with overlapping regions identified through dotplot and pairwise sequence alignment analysis. Likewise, non-overlapping regions show no similarity in gene content. Figure 3.7 illustrates a comparison between genes identified by Gao *et al*. (2010) and genes identified in this study. Forty of the sixty-six genes identified in this study were not annotated by Gao *et al*. (2010). The ten genes highlighted in red (Figure 3.7) have been identified both through sub-cloning and sequencing of the CHORI BAC sequences and through re-analysis of Gao's BAC sequences. These ten genes are included in the forty sequences not reported by Gao *et al*. (2010).

Table 3.4: Summary of gene content on each of Gao's BAC sequences.

Gene	Str	CDS Location	No. Exons	Len aa	Identity
FJ985869.1 reverse complement (134643 bp)					
1†	+	953 - 1423	1	156	Ubiquitin D (UBD)
2	-	3267 – 3521	1	84	olfactory receptor 94-like
3	+	12029 – 12544	1	171	ubiquitin D-like
4	-	14881 - 17031	2	170	olfactory receptor 94-like
5	+	23451 - 23840	1	129	Similar to LOC786987 (218 aa)
6	-	25868 - 33873	2	173	olfactory receptor 94-like
7*	-	36056 - 59082	18	862	gamma-aminobutyric acid type B receptor (GABBR1)
8*	+	82718 - 92594	8	284	myelin oligodendrocyte glycoprotein alpha-2 (MOG)
9*	-	93556 – 97185	4	563	Simlar to KRAB box
10*	-	109674 - 111825	3	361	Similar too C7H6orf12 (*Sus scrofa*) Hypothetical protein
11*	+	115700 - 118582	4	123	zinc ribbon domain containing protein 1 (DNA-directed RNA polymerase I subunit RPA12) (ZNRD1)
12*	+	124107 - 125982	3	126	protein phosphatase 1, , regulatory (inhibitor) subunit 11 (PPP1R11)
13*	-	127839 - 132244	4	353	RING finger protein 39 (HZFw1)
FJ985854.1 reverse complement (145292 bp)					
1	+	993 - 1739	2	103	protein phosphatase 1, regulatory (inhibitor) subunit 11 (PPP1R11)
2*	-	3596 - 8000	4	353	RING finger protein 39 (HZFw1)
3#	-	35569 - 35488	NA	81	tRNA Leucine (anticodon CAA)
4*	-	36264 – 46765	9	580	tripartite motif-containing 31 (Trim31)
5	+	74551 - 85943	5	310	tripartite motif-containing protein 40 – like (Trim40)
6*	-	89882 - 97620	7	489	tripartite motif-containing protein 10 (Trim10)
7*	+	101086 - 111003	7	461	tripartite motif-containing protein 15 (Trim15)
8	-	128422 - 136886	8	535	tripartite motif-containing 26 (Trim 26) - like
FJ985864.1 reverse complement (142360 bp)					
1	-	18376 - 18882	1	168	Similar to neuronal axonal membrane protein (NAP22) 227 aa
2 *	+	53040 - 56369	8	364	MHC Class 1 (classical)
3	+	97997 - 113333	5	229	MHC Class 1 like (pseudogene?)
4*	+	130285 - 133430	7	346	MHC class I antigen (non-classical)

FJ985870.1 (138311 bp)					
1*	+	26626 - 30134	8	364	MHC class I antigen (classical)
2	+	71762 - 87098	5	229	MHC class I antigen - like
3*	+	104051 - 107196	7	346	MHC class I antigen (non-classical)
FJ985873.1 (196844 bp)					
1 *	+	13714 - 22197	6	488	tripartite motif-containing 39 (488aa) (TRIM39)
2 *	+	24860 - 26427	5	146	ribonuclease P 21-like isoform 2 (RPP21)
3 *	-	51229 - 58671	12	607	guanine nucleotide-binding protein-like 1 (GNL1)
4*	+	59826 - 63665	4	187	proline rich 3 (187) (PRR3)
5 *	+	71954 - 84473	25	841	ATP-binding cassette sub-family F member 1 (ABCF1)
6 *	_	91053 - 99435	18	924	(serine/threonine) protein phosphatase 1, regulatory subunit 10 (PP1R10)
7 *	+	106890 - 112858	7	258	28S ribosomal protein S18b, mitochondrial-like (MRPS18b)
8 *	+	113644 - 26396	13	421	alpha tubulin acetyltransferase 1 (ATAT1) Similar to chromosome 6 open reading frame 134 (*Homo sapiens*)
9 *	+	126938 - 130352	6	315	Similar to hypothetical protein LOC538804 (Bos taurus) chromosome 6 ORF 136 ortholog
10*	-	130804 - 142735	20	1045	DEAH (Asp-Glu-Ala-His) box polypeptide 16 (DHX16)
11*	-	146685 - 154204	3	614	KIAA1949 (phostensin)
12*	-	156587 - 158890	4	262	Nurim (nuclear envelope membrane protein) (NRM)
13	-	162258 - 171968	14	1845	mediator of DNA-damage checkpoint 1-like (MDC1)
14*	+	177689 - 180933	4	444	Beta tubulin (TUBB)
15 *	-	182927 - 192800	12	427	Flotillin 1 (FLOT1)
FJ985868rc (140835 bp)					
1†	+	27 - 1084	1	351	Beta tubulin (TUBB)
2*	-	3081 - 12715	12	427	Flotillin 1 (FLOT1)
3*	-	14317 - 14887	2	156	Immediate early response 3 (IER3)
4*	+	115834 - 125877	17	915	Discoidin domain receptor tyrosine kinase 1 (DDR1)
5‡	+	134840 - 139463	12	406	general transcription factor IIH subunit 4 (GTF2H4)
FJ985852rc (140835 bp)					
1*	+	21197 - 31240	17	915	Discoidin domain receptor tyrosine kinase 1 (DDR1)
2*	+	40203 - 46789	13	463	general transcription factor IIH subunit 4 (GTF2H4)
3*	+	47699 - 59288	29	1065	valyl-tRNA synthetase (2), mitochondrial precursor VARS(2)
4*	-	62895 - 63483	3	78	surfactant-associated protein 2 (SFTA2)

5*	+	65744 - 69748	3	86	diffuse panbronchiolitis critical region 1-like ((DPCR1-like))
6	+	72864 - 82078	4	545	diffuse panbronchiolitis critical region 1-like (DPCR1-like)
7	+	99845 - 103977	3	463	Mucin 21-like (MUC21-like)
FJ985875rc (173955 bp)					
1	+	10324 - 2036	17	915	discoidin domain receptor tyrosine kinase 1 (DDR1)
2*	+	29329 - 35915	13	463	general transcription factor IIH subunit 4 (GTF2H4)
3*	+	36825 - 48413	29	1035	valyl-tRNA synthetase (2), mitochondrial precursor VARS(2)
4*	-	52020 - 52608	3	78	surfactant-associated protein 2 (SFTA2)
5*	+	54869 - 58872	3	86	diffuse panbronchiolitis critical region 1-like (DPCR1-like)
6	+	61988 – 71572	4	805	diffuse panbronchiolitis critical region 1-like (DPCR1-like)
7	+	89339 - 93471	3	463	Mucin 21-like (MUC21-like)
12*	-	127171 - 130033	7	354	MHC Class 1 (non-classical)
13	+	141258 - 141674	1	138	Hypothetical mucin-like
15	-	160117 - 162943	6	337	MHC Class 1-like
17	+	163405 - 163941	1	176	Hypothetical protein – mucin-like
18	+	165598 - 166975	2	441	eukaryotic translation elongation factor 1 alpha 1 (EEF1A1)
FJ985859rc (160643 bp)					
1*	-	1806 - 4642	7	356	MHC Class 1 (non-classical)
2	-	28878 - 34806	7	355	MHC Class1 – like
3	-	61177 - 62297	2	295	Similar to bovine hypothetical protein LOC788708
4*	-	66388 - 70063	2	545	Corneodesmosin (CDSN)
5*	-	85378 - 86424	2	135	psoriasis susceptibility 1 candidate gene 2 protein (PSORS1C2)
6*		88619 - 99042	15	724	coiled-coil alpha-helical rod protein 1 (CCHCR1)
7*	+	101464 - 103790	3	344	transcription factor 19 (TCF19)
8*	-	105620 - 110107	5	360	POU domain, class 5, transcription factor 1 (POU5F1)
9	+	112817 - 141996	6	307	MHC class I-related protein (MICB)
10	-	148998 - 151240	5	356	MHC class I - like
FJ985874 (173955 bp)					
1	-	6993- 9231	5	355	MHC Class I - like
2*	-	34390- 37396 res 30- 39000	8	360	MHC Class I
3	-	45384- 59320		290	MHC Class I - like
4	-	77708 - 81297	7	379	MHC class I - like

5	-	81735-82233	2	113	Similar to interferon-induced transmembrane protein 3 (IFITM3)
6	-	96704-100257	7	374	Uncharacterised protein MGC126945
7	+	125982 - 127105	2	126	mitochondrial coiled-coil domain 1 - like
8*#	-	127581-137035	10	428	HLA-B associated transcript 1 (BAT1)

†First exon not in BAC. ‡Last exon not in BAC. *Alignment to known homolog(s) shows high percent identity with no evidence of wrong or missing exons. #not included in Figure 3.7.

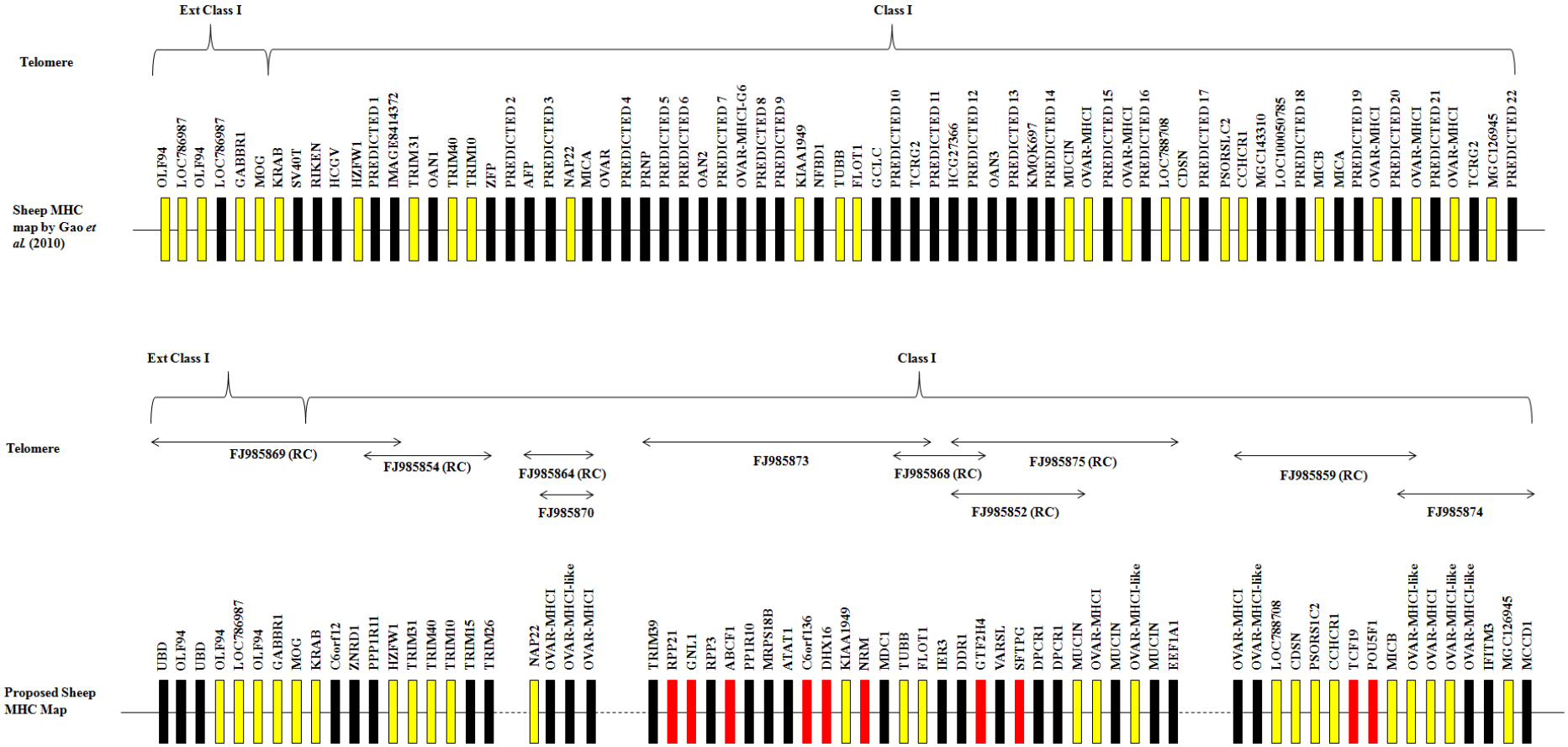


Figure 3.7: Comparison of sheep MHC class I map published by Gao *et al*. (2010) and the new map proposed in this study. Loci highlighted in red indicate the 10 loci identified through sub-cloning of CHORI 243-269M18, CHORI 243-390H16 and CHORI 243-

454E19 and confirmed by re-analysis of Gao's BACs. Loci found both in sheep map by Gao *et al*. (2010) and the newly proposed map constructed based on findings in this study were highlighted in yellow.

3.4 Discussion

The work described in this chapter contributes significantly to a physical map of the sheep MHC together with annotation. The study focuses on the gene arrangement within the class I region via sub-cloning and sequencing of CHORI BACs that contain class I sequences. The results obtained allow for a re-examination of the currently available data in the NCBI database, based on the recent report by Gao and colleagues (Gao *et al.* 2010) that was not available when this study commenced.

Assembly of Gao's BAC sequences with Geneious resulted in five contigs rather than the single long contig reported by Gao *et al*. (2010). These multiple contigs indicate the probable presence of gaps in the map inferred by Gao and colleagues. Analysis of overlapping regions between Gao's BAC sequences using manual methods (combination of BLAST, dotplot, genomic sequence alignment and various gene prediction programs) suggested that there are actually six contigs in the sheep MHC map published by Gao *et al*. (2010). Three gaps were present in the class I region while the remaining two gaps were identified in the class IIa region. The comparison of alignments generated by Geneious and the manual method, suggests that the latter produced a better contig tiling path. This may be explained by the low sensitivity of Geneious for a sequence containing a string of undefined nucleotides (N) and therefore omits the complete BAC sequence from the assembly. It seems that Genious is not capable of discriminating between real contigs and false positives. For instance, Geneious failed to include four BAC sequences (FJ985852, FJ985862, FJ985865 and FJ985867) in the contig tiling path and indicated that there is an overlap between FJ985854 and FJ985864. Closer examination of the overlapping region between FJ985854 and FJ985864 through the manual method revealed that there is a potential gap in this region, which encompasses a region between *TRIM26* and *NAP22*. The initial 765 bp in the 5' end of BAC sequence FJ985864 does not align with the 3' overlapping region of FJ985854, but aligns from bp 766 onwards. BLAST analysis indicates that the 765 bp sequence at the 5' end of BAC FJ985864 does not show a contiguous alignment with the downstream BAC sequence when examining the matching alignments with two BAC sequences from cattle. Instead this region shows a match with a region

elsewhere in the same cattle BAC sequence. The 765 bp sequences in 5' end of BAC FJ985864 could potentially be due to a mistake introduced in the initial re-assembly of the BAC sequence. Each of the 26 BAC sequences present in Gao's map was sequenced through a DNA shotgun sequencing method, which involves sub-cloning, sequencing and assembling randomised 0.5 – 2.0 kbp small fragments of DNA to form a full-length BAC sequence (Gao *et al.* 2010). If the 765 bp ambiguity is not present on BAC FJ985864, the rest of the BAC would align with no gap with BAC FJ985854. Despite the differences in the result produced by Geneious and the manual method, both showed that the MHC map published by Gao *et al*. (2010) is not contiguous and hence is possibly incomplete.

The BAC sequence FJ985873 does not overlap with either FJ985864 or FJ985870 and the gap size in this region relative to the cattle reference sequence map is expected to be approximately 150 kbp. It is likely that there are several genes missing in this region due to the gap. Comparison of the class I region in other species suggests that the gap region may account for at least one peptide-presenting MHC class I gene (Horton *et al.* 2004; Brinkmeyer-Langford *et al.* 2009). The gap between the *OVAR-MHCI* and *TRIM39* loci is notable. The other gap in the class I region is located between FJ985875 and FJ985859, which is between *EEF1A1* and an adjacent *OVAR-MHCI* locus. The size of this gap is estimated to be only a few thousand bp relative to the cattle map but the actual size is not known. The size of the gap between FJ985867 and FJ985862 and the genes at either end of the gap is not known because the class IIa region is yet to be annotated. The size of the gap in the class IIa region between FJ985876 and FJ985865 is also unknown.

Sequencing and re-assembly of CHORI BAC sub-clones provided a low resolution physical map of two separate areas spanning approximately 436 kbp within the class I region. Identification of ten new genes in this study, adds significantly to the incomplete annotation in Gao's map. These ten genes account for approximately 14% of gene content within the class I region relative to the cattle reference map. The genes identified in CHORI BAC sequences are also present in the class I region of other mammals such as cattle, horse, human and pig (Gustafson *et al.* 2003; Horton *et al.* 2004; Hurt *et al.* 2004; Demars *et al.* 2006; Brinkmeyer-Langford *et al.* 2009). The

sheep MHC map derived from Chinese Merino published by Gao *et al*. (2010), predicted 22 orthologous genes that have yet to be mapped to the cattle MHC.

In contrast, annotation of genes within the BAC sequences reported above for the class I region showed that there was a high level of sequence identity between genes within Gao's BAC sequences and known genes previously reported in sheep (*Ovis aries*) or cattle (*Bos Taurus*). Among the 66 genes predicted in this study from Gao's BAC sequences, only 26 (i.e. 39%) were reported by Gao *et al*. (2010). The ten genes identified in the CHORI BAC sequences were also identified through re-analysis of Gao's BAC sequences, indicating that the gene prediction methods (BLAST and Ensembl pipeline) used by Gao *et al*. (2010) were not up to the task. The Ensembl pipeline was initially designed to perform large-scale automated annotation of genomic sequence, in particular that of human and mouse (Potter *et al.* 2004). Although the Ensembl pipeline has been modified and used to produce gene predictions for rat, zebrafish, fly, mosquito, worm and fugu, the versatility of the pipeline is clearly limited when predicting genes in various other species (Curwen *et al.* 2004). Most integrated algorithms are limited to only a few organisms, and all gene prediction programs have their strengths and weaknesses; none are perfect (Fickett 1996) Therefore, a better result of gene prediction or annotation can be achieved by using several programs with a careful comparison of results (Fickett 1996).

Overall, the general structure and gene content in the sheep MHC class I region is similar to that of other mammals but with a slight difference in gene arrangement (Amadou 1999; Gustafson *et al.* 2003; Horton *et al.* 2004; Hurt *et al.* 2004; Demars *et al.* 2006; Brinkmeyer-Langford *et al.* 2009). The class I genes involved in peptide presentation functions in some mammalian species such as chimpanzee (Kulski *et al.* 2002), human (Leelayuwat *et al.* 1995), rhesus macaque (Kulski *et al.* 2004) and horse (Gustafson *et al.* 2003), are often clustered within three distinct locations designated as the alpha (between *MOG* and *PPP1R11*), beta (between *POU5F1* and *BAT1*, which borders the class III region) and kappa (between *TRIM26* and *GNL1*) blocks. However, the clustering of peptide-presenting MHC class I genes in sheep does not fit entirely into the alpha, beta and

kappa block theory. In a previous study of sheep MHC, the presence of an additional novel block located between *GTF2H4* and *CDSN* has been reported (Liu *et al.* 2006). Analysis of gene organisation within class I region in this chapter concurs with the presence of such a novel block. In this study, there are at least two definite MHC class I and two MHC class I-like genes between *GTF2H4* and *CDSN*. The exact number of peptide-presenting MHC class I genes is not known due to the presence of a gap in this block. In addition, this study reveals that there is no evidence for presence of peptide-presenting MHC class I genes between *MOG* and *PPP1R1* (alpha block) as reported in other organisms. This finding is also in agreement with the previously reported sheep MHC study by Liu *et al*. (2006). The blocks of peptide-presenting genes are separated by numerous other class I genes with immune and non-immune related functions. The organisation of other sheep genes in the class I region is similar to the evolutionarily closely related cattle MHC, except for the positioning of several *TRIM* genes. In cattle, a cluster of genes from the *TRIM* sub-family is located telomeric to *ZNRD1*. In contrast, a similar but not identical cluster of *TRIM* genes are positioned centromeric to the *ZNRD1* gene.

The organisation of MHC class I peptide-presenting genes in distinct blocks, which are interspersed between other genes located within the class I region, is most likely due to segment or tandem block duplication (Kulski *et al.* 1997; Gaudieri *et al.* 1999; Kulski *et al.* 1999). The framework hypothesis suggests that the MHC class I region is a "conserved ordered segment" that represents a dense region of genes with essential functions, whose alterations are deleterious (Amadou 1999).

The information obtained in this study updates the existing sheep MHC map and enhances annotation of the genes present in the MHC class I region. In particular, the telomeric to centromeric orientation of BACs used by Gao and colleagues has been resolved, overlapping sequence regions identified, gaps in the sheep MHC map mapped and the putative position of loci within each BAC detailed.

Chapter 4

An Analysis of Linkage Disequilibrium Across the MHC Class I Region

This chapter describes the discovery of single nucleotide polymorphisms (SNPs) within the sheep MHC class I region and genotyping of these SNPs in a population of sheep. Fragments of MHC class I sequence derived by sub-cloning of the BACs described in Chapter 3, and additional sequences from National Center for Biotechnology Information (NCBI), were used as target templates for identification of SNPs in a small population of animals. 32 SNPs were identified from this analysis. Fourteen SNPs spanning the class I region were used for genotyping 108 distantly related animals. Genotypic data for these animals were then analysed to estimate expected heterozygosity and linkage disequilibrium (LD) patterns across class I region of sheep MHC. Four blocks of high LD were observed within the sheep MHC class I region. Information from this chapter is being prepared for publication.

4.1 Introduction

Single nucleotide polymorphisms (SNPs) are a substitution of one nucleotide with another at a given position in a DNA sequence. SNPs are the most abundant type of polymorphism within the genome, and adjacent SNPs are usually inherited together in block-like structures called haplotypes (Sobrino *et al.* 2005). Often, individuals with a specific SNP allele would be very likely to share a haplotype with a correlated allele at a nearby variant site (HapMap 2005). This correlation between genetic markers is referred to as linkage disequilibrium (LD). Linkage disequilibrium decreases with increased physical and genetic distance between loci due to meiotic recombination between ancestral haplotypes leading to increased haplotype diversity (Dawson *et al.* 2002). Similarly, LD between neighbouring loci is inversely related to haplotype diversity (Dawson *et al.* 2002). The non-random

association of variants with neighbouring markers (linkage disequilibrium) can be statistically quantified through SNP genotyping, and hence characterise the concomitant haplotype diversity that reflects observed phenotypic manifestations. SNPs are preferred over other genetic markers in haplotype studies because they manifest very low mutation rates, are abundant in the genome and are suitable for analysis using automated high-throughput technologies (Daly *et al.* 2001; Sobrino *et al.* 2005). Recent advances in SNP genotyping technologies, together with the availability of economically viable genotyping methodologies have provided much impetus for SNP-based Genome Wide Association Studies (GWAS) and identification of important region specific haplotypes (Rafalski 2002; Botstein & Risch 2003).

Genetic information inherited as haplotype blocks are resistant to disruption by recombination over generations (Jeffreys *et al.* 2001; Stenzel *et al.* 2004; Gibson *et al.* 2006). In mammals, including humans, haplotypic blocks are often separated by meiotic recombination hotspots that occur in clusters every 60-90 kb apart, with approximately 1-7 kb region separating each hotspot within a cluster (Jeffreys *et al.* 2001). Haplotype blocks are therefore characterised as particular combinations of alleles in a population (Gabriel *et al.* 2002). Haplotypes diversity arises when a new mutation occurs on an ancestral haplotype and the haplotype is disrupted through recombination events, which are transmitted to subsequent generations (Gabriel *et al.* 2002).

The study of haplotype structure is important in identifying loci subject to evolution by natural selection. Haplotypic diversity also allows an understanding of the pathways of divergence from a common ancestor that is important in comparing intraspecies populations (HapMap 2005). Therefore, haplotypic relationships provide a powerful tool for mapping mutations associated with complex diseases and traits. For example, haplotype methods have not only contributed to the identification of genes associated with complex Mendelian diseases such as diastrophic dysplasia and age-related macular degeneration in human (Hastbacka *et al.* 1992; Cardon & Abecasis 2003; Edwards *et al.* 2005; Haines *et al.* 2005; Klein *et al.* 2005), but are highly useful in selective breeding for improved quality

and productivity in the agricultural sector (Schenkel *et al.* 2005; Schnabel *et al.* 2005; Konishi *et al.* 2006; Salvi *et al.* 2007).

Although there is considerable interest in mapping recombination hotspots and understanding patterns of LD in various mammalian genomes, the most intensive efforts (and funding) have been directed to human and mouse genomes. SNP genotyping has identified haplotypic structures within the human MHC region, and in the process identified at least eight independent recombinational hotspots (Stenzel *et al.* 2004). Three hotspots were identified in the class I region; between *GABBR1* and *MOG*, between *TRIM26* and *TRIM39*, and in the 3'UTR of the *DPCR1* (Stenzel *et al.* 2004). The hotspot in MHC class III region lay between two markers close to *NOTCH4* (Stenzel *et al.* 2004). The remaining four hotspots were identified in class II region; between *HLA-DRA* and *HLA-DQA2*, within *TAP2*, close to *HLA-DMB* and between *HLA-DMA* and *HLA-DOA* (Stenzel *et al.* 2004). These findings supported earlier work using human sperm typing of the MHC class II region, in which hotspots were identified close to *HLA-DOA* (also known as *HLA-DNA)*, *HLA-DMB* and *TAP2* (Jeffreys *et al.* 2001).

The aim of this study is to identify SNPs in non-immunological genes or gene areas within the sheep MHC class I region and generate a LD map describing recombination hotspots relative to the LD blocks in this region.

4.2 Materials and methods

4.2.1 Sample collection and DNA extraction

Blood samples were collected from 2 cohorts; 1 cohort consisted of 12 unrelated animals for identification of single nucleotide polymorphisms (SNPs) and the other cohort consisted of 108 animals, which was used for genotypic analysis. Samples were obtained from Department of Agriculture and Food of Western Australia (DAFWA), Katanning, Western Australia. Blood samples were collected using venepuncture method as described in Chapter 2.1. DNA was extracted using AxyPrep^{TM} Blood Genomic DNA Miniprep Kit (Chapter 2.2.1).

4.2.2 Primer design

Sequences of BAC sub-clones derived from CHORI 243-269M18, CHORI 243-390H16 and CHORI 243-454E19 identified in Chapter 3 were used as template for primer design. Specific primers for SNP identification were designed using Primer3, Primer-BLAST and NetPrimer program for sequences covering different locations across the MHC class I region. To expand the SNP identification beyond the coverage of BAC DNA sequences, additional primers based on sheep sequences in the National Center for Biotechnology Information (NCBI) GenBank database (http://www.ncbi.nlm.nih.gov/genbank/) were designed. The details of primers designed are given in Table 4.1.

Table 4.1: Primers designed for 22 loci located across the MHC class I region (from telomere to centromere). These primers were used for SNP identification.

Primer set	Primer sequence (5'-3')		Source	Product size (bp)	Location within NW_001494164.1
	Forward primer	Reverse primer			
SN21	TATGGTAGGTTGAGGAGGGA	GCAAGGAAACTGAAAAGATG	CHORI 243-390H16	515	649084 – 649599
SN6	GGATGTGACCCTGGACCCTG	AGAACCCCCTGAAAGGCTGT	GenBank FJ985873.1	885	635501 – 634616
SN20	ATCTGAGTTATTGTCCCCAT	GAATGGGGAGTAAAACAAGG	CHORI 243-454E19	660	626870 – 626214
SN19	CACGAAAAATCACCAGGAGC	CTTTTACACCCACATCTCTG	CHORI 243-454E19/ 390H16	635	624305 – 623670
SN18	CTGTGCATCTCAAGTAGGTC	ATTCAGTATTGTTGTGGAGG	CHORI 243-454E19	375	617701 – 618073
SN17	GCCATCTCCTCTACCTTCTC	TCAGCAAGGGCAAGACCACT	CHORI 243-454E19	490	565170 – 565659
SN29	GCCCACTCAGTTATCTCAAG	ATACAAGGGTGGTGAGTCAG	CHORI 243-390H16	445	560894 – 561338
SN16	GCCTTCTCTGCCCCATTGTA	TCCCACACACTCCCATCCTC	CHORI 243-454E19	920	529393 – 528475
SN15	AGGCAGGCTTTGTTTTAGTC	TGTAAGTTCGGTTCTGGTTC	CHORI 243-390H16	690	515468 – 514781
SN27	GAAATGGGTATCTGGCTCTA	TCTCCCAAAGTGAAGTGAAG	CHORI 243-390H16	625	514011 – 513387
NRM	AGCTCATGGGCCTGAAACAG	AGTGGAGGCTCAAGCAAGGA	CHORI 243-390H16	650	487636 – 488283
SN43A	GGGAGTGAAGCGGCAGTTAC	GGGACAGAACACCTGCCATC	GenBank FJ985868.1	820	462628 – 463445
SN42	GGGCTTTGTCTTTGACGGGA	TGCCTACTGGACCACTTGCT	GenBank FJ985868.1	690	408353 – 409042

SN41	GGGGCAGAGCGCTTCATAAA	GGGTTGCTGAGAATGCCACA	GenBank FJ985868.1	530	358169 – 358697
SN40	TCCCAGACCGCTTCTCACTT	CCGACCCTGGATACACACCT	GenBank FJ985875.1	490	313661 – 314149
SN25	AGGTGCTGAGAGTGAAGAGA	CTATCCGCAGAATCTTGAAC	CHORI 243-269M18	655	206913 – 206260
SN5	AGAGCCTCACTGACATCTCC	AAGCAAAGACCAGTCCAACA	CHORI 243-269M18	465	198122 – 198587
CDSN	CAGGAGCTGCTGTCAGTCAG	TCCACAGCCGTGTCACATGT	GenBank FJ985859.1	390	187966 – 187576
SN11	CTACCCCAAGAAGAGACAGT	GCTGCTCATTCTTTATTTGG	GenBank FJ985859.1	650	155465 – 154819
SN3A	GAGATGAACACCTTTGCCCA	ACACCAAGTCAGCACACTGT	CHORI 243-269M18	360	154070 – 154428
SN22	TTCGTTTCGGGGTCATTTAG	CTAGGTGGTCCGAGTGTGGT	CHORI 243-269M18	585	153831 – 153246
SN2	GACACTAGAGAAGGTTGCAA	CTGCAGCTGGCTAAGGTCAT	CHORI 243-269M18	624	144151 – 143587

4.2.3 PCR amplification and sequencing

PCR amplification of sequences for SNP identification was performed using the standard PCR protocol (Chapter 2.4). Six to twelve random Merino sheep DNA samples were amplified for each primer set. The details of PCR conditions specific to each primer set are shown in Table 4.2. PCR products were DNA sequenced by Macrogen Inc. (South Korea).

Table 4.2: Details of PCR conditions for each primer set used for SNP identification.

Primer	**Annealing temperature (^{o}C)**	**Number of cycles**	**$MgCl_2$ concentration (mM)**
SN21	50	35	1.85
SN6	52	27	1.5
SN20	55	32	1.85
SN19	50	35	1.85
SN18	55	35	1.85
SN17	55	30	1.85
SN29	54	28	1.85
SN16	55	30	1.85
SN15	50	35	1.85
SN27	55	30	1.85
NRM	65	35	1.5
SN43A	65	28	1.85
SN42	65	28	1.85
SN41	65	28	1.85
SN40	65	28	1.85
SN25	49	29	2.95
SN5	61	35	1.5
CDSN	60	33	1.85
SN11	46	30	1.95
SN3A	65	35	1.5
SN22	54	28	1.85
SN2	64	35	1.5

4.2.4 SNP discovery in class I region

DNA sequences were subjected to BLAST analysis against the cattle reference genomic sequence database; NW_001494164.1 to confirm the target region had been amplified and sequenced. Chromatograms of each sequence were analysed using Vector NTI® (Invitrogen) software for SNPs and used to determine if the sequence was heterozygous or homozygous at the respective base pair of interest. Base changes at a specific position were confirmed as SNPs when a minimum of two heterozygotes and one of each homozygote were identified.

4.2.5 SNP genotyping

Fourteen SNPs with good quality flanking regions on both sides of the locus representing different MHC class I regions were selected for genotyping of 108 unrelated Merino sheep. SNP genotyping was performed by KBioscience (http://www.kbioscience.co.uk/). SNPs were genotyped using the KASPar assay system, which is based on the discriminatory power of competitive allele specific PCR to determine the alleles at a specific locus within genomic DNA for SNP typing. The flanking sequences for SNPs submitted to KBioscience for genotyping are summarised in Table 4.3. Purified DNA samples with minimum concentration of 7ng/ul were sent in 96-well plates together with blind-duplicates and plate-identifying blank wells as within and between plate quality controls.

4.2.6 Analysis of SNP genotypic data

SNP genotypic data was analysed using the Genepop population genetics software package (Raymond & Rousset 1995; Rousset 2008) to test for Hardy-Weinberg equilibrium, and estimate allele frequencies, number of heterozygotes and number of homozygotes. The SNPstats program (Sole *et al.* 2006) was used to estimate measures of linkage disequilibrium between adjacent SNPs across the MHC class I region.

Table 4.3: Sequence flanking SNPs used for genotyping. The polymorphic nucleotide is highlighted in red.

SNP ID	Sequence
SN21_1	TGTTTAAATAAGACATTTAGAGCCTTTCCGGGTGTTTATTTTTTG**[R]**CCACAGGGCTTGCAGATCTTAGTTCCCCAACCGGAAATGAAACCC
SN6_1	TTTTACTGAGAAACTTTGGCCCCTCTTCTATCCAGGCATCCGGGC**[Y]**GGTCGGAAGAATGCTGCACCACTCACCATCAGGCCCCCGACAGAC
SN20_1	TATTCCTGGGAGCCACCCCAAACAGTGAGGGATGAGTATCGGTGG**[K]**TAAATGTTCTACATAGTCTCAGGGCCTCCAGCAGAGTGGAGCCCT
SN17_2	CCCTTCCAACTCACAGCCACAGGCTCTTCTGTTCTTGGCTGGCTC**[R]**CAGACCCCAGTGGTAGCCCCTCGCAAGTCCCTGAGTTTCCCCCCA
SN29_1	TTTCTGTGATGAGTCGGGCATCGTGGCTGACGACGATCACAGCTT**[Y]**GAGGCAAGACAGTGAGAATGGAGGACAGGCAGAAAGGAAGAGGGG
SN15_1	CGGCCCAGAGTAGAATCCCGAGAAATTCTTAGCCTAAGCCCACAC**[Y]**GGTAGTTCCTAAACTTTGCTGCACAGTGGATTCTTGGATTTTTTA
NRM_1	CCCAGGACGGGAGGAAAGAGGCAGTTAAGTTGTGACAAGGGCTCG**[R]**CTCCATTCCTAAGCCATTCTCCCTCTGCTAGGTGTACTACCATGT
SN43A_2	ACACTAAAGTGTCAGTTTAGRATGGGACTATGATATGACACCACC**[Y]**CTGCCCAGGTGACTTTTGCCTTAGGCCAAGAAAACCCACATACCC
SN42_2	ATTTCCTAACAAGGGTCATGCCTTGAAGACTCAGCAGTGATGCAA**[Y]**GACAGAAAAAGGAGAGGAATTTACATAGAAGGCAAAGATGCGAAC
SN41_4	AAATATCCTCACAGTCACATAGCAGAAATATTTTTATTTATAGAC**[R]**GAATCTTAGAATCTTTCTCTTTGCTATCATCCTTTGCCAAAGCCT
SN25_1	CTCCCCTGTGGGAGACAAGAGAAAATAAGAGGTGGGCAGTCTTAG**[Y]**TGCATTCAAGCCCTTGGTTYTAGTTTGTACCTGAGATTCACTTGC
SN5_1	TGACATAGCTAGCGTGTGACAGAGTCAGTGACTCAAATGAAGTAT**[Y]**TCTCATTATCTGTCCAGGGAGTGTTCAGGTATAATATGCAATTAA
CDSN_1	CTGGCTGGTCTCCTCCTGCAAGGTAGGAGGTTGGGGCCCTGGGAG**[Y]**GGGGAGAGTTGGGAGAGGAGGGAGATTCAGGCTCAGACAACTGGT
SN3A_1	AATCAAGCCTTTACTCATTTCTTCTATCAGTCCAACCAGAAACGC**[M]**ACTTTGTGCTTTCTTGATACTGGGGTCCATTTGGTTTTCACTTAA

4.3 Results

4.3.1 SNP discovery

Thirty-two SNPs were identified within the MHC class I region. These spanned a region of approximately 505 kbp. The region between SN41 and SN25 was approximately 152 kbp in length. Primer design in this region was difficult due to many repeating elements and multiple copy genes. The only primer pair (SN 40) amplified from this region was monomorphic, as confirmed by sequencing. The locations and sequence identity of fragments containing nucleotide variation were determined from BLAST alignments using the orthologous cattle sequence database as a reference sequence. Among twenty two sequences targeted for SNP discovery, seven were monomorphic and one proved to be a duplicated locus. From the resulting 32 SNPs, 14 that were not located within repeat regions, and with good quality flanking sequences of at least 50 bases long on either side of the SNP (as required by KBioscience) were selected for subsequent genotyping.
Many additional SNPs were also discovered within the *CDSN* coding and non-coding regions and are further described in Chapter 6. Table 4.4 shows the number of SNPs identified for each locus. Figure 4.1 shows a graphic representation of SNP containing fragments relative to each other with respect to the cattle reference genomic sequence database; NW_001494164.1.

Table 4.4: Summary of SNPs identified and their locus of origin each locus. SNPs with an asterisk (*) were used for genotyping the sheep cohort.

Primer set	SNP identity	Base change	Locus	Description of locus	GenBank accession ID
SN21	SN21_1 *	A/G	*RPP21 - LOC512672*	Between ribonuclease P 21 (*RPP21*) and major histocompatibility complex, class I	JQ433713
SN6	SN6_1 *	C/T	*RPP21 - LOC512672*	Between ribonuclease P 21 (*RPP21*) and major histocompatibility complex, class I	JQ433712
SN20	SN20_1 *	G/T	*GNL1 - RPP21*	Between guanine nucleotide binding protein-like 1 (*GNL1*) and ribonuclease P 21 isoform 1 (*RPP21*)	JQ433711
SN17	SN17_1	A/G	*PPP1R10 - PRR3*	Between protein phosphatase 1, regulatory (inhibitor) subunit 10 (*PPP1R10*) and proline-rich protein 3 (*PRR3*)	JQ433710
	SN17_2 *	A/G			
	SN17_3	A/G			
SN29	SN29_1 *	C/T	*PPP1R10 - PRR3*	Between protein phosphatase 1, regulatory (inhibitor) subunit 10 (*PPP1R10*) and proline-rich protein 3 (*PRR3*)	JQ433709
SN15	SN15_1 *	C/T	*C23H6orf136*	*C23H6orf136*	JQ433715
	SN15_2	C/G			
NRM	NRM_1 *	A/G	*NRM*	Nurim (*NRM*)	JQ433714
SN43A	SN43A_1	A/G	*FLOT1 - TUBB*	Between flotillin-1 (*FLOT1*) and tubulin, beta 2B (*TUBB*)	JQ433708
	SN43A_2 *	C/T			
	SN43A_3	C/T			

SN42	SN42_1	C/T	*DDR1 - IER3*	Between discoidin domain receptor family, member 1 (*DDR1*) and immediate early response 3 (*IER3*)	JQ433707
	SN42_2 *	C/T			
	SN42_3	C/T			
	SN42_4	C/T			
	SN42_5	C/T			
	SN42_6	A/G			
	SN42_7	C/T			
SN41	SN41_1	A/C	*DDR1 - IER3*	Between discoidin domain receptor family, member 1 (*DDR1*) and immediate early response 3 (*IER3*)	JQ394773
	SN41_2	C/T			
	SN41_3	C/G			
	SN41_4 *	A/G			
SN25	SN25_1 *	C/T	*LOC788708* - Hypothetical protein	Between *LOC788708* and hypothetical protein	JQ394772
SN5	SN5_1 *	C/T	*LOC788708* - Hypothetical protein	Between *LOC788708* and hypothetical protein	JQ394771
	SN5_2	C/G			
	SN5_3	C/T			
CDSN	CDSN_1 *	C/T	*CDSN*	Corneodesmosin (*CDSN*)	JQ394770
SN3A	SN3A_1 *	A/C	*POU5F1 - TCF1*	Between POU domain, class 5, transcription factor 1 (*POU5F1*) and transcription factor 19 (*TCF19*)	JQ394769
SN2	SN2_1	C/G	*MICB - POU5F1*	Between MHC class I polypeptide-related sequence B (*MICB*) and POU domain, class 5, transcription factor 1 (*POU5F1*)	JQ394768
	SN2_2	A/C			

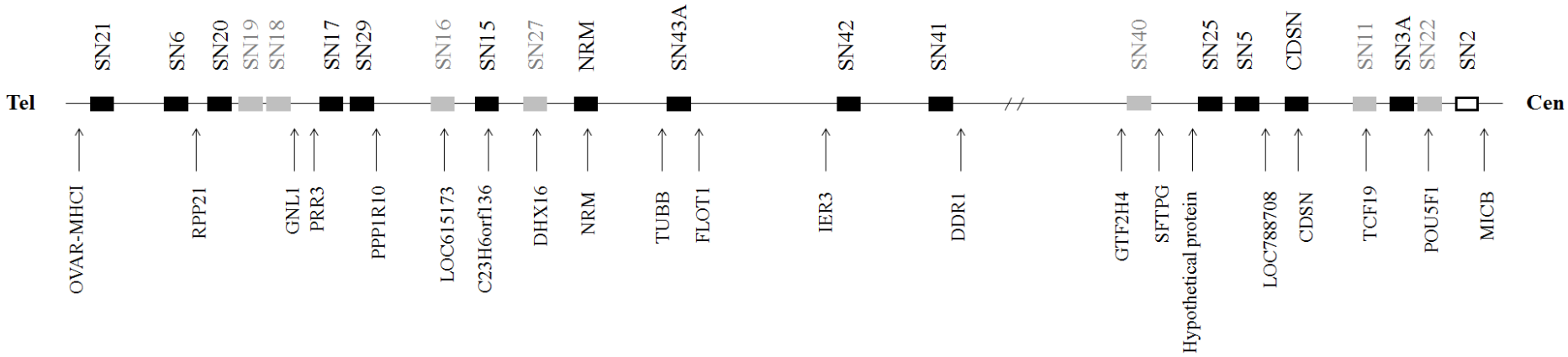


Figure 4.1: Graphical representation of loci within the MHC class I region used as templates for SNP discovery for which PCRs were developed (figure not drawn to scale). "Tel" indicates the telomeric end and "Cen" indicates the centromeric end. Grey shaded areas indicate PCR targets tested but which did not yield any SNPs. The empty non-shaded box indicates a duplicated locus. Numbers of SNPs discovered per locus are shown in Table 4.4.

4.3.2 Genotypic analysis

Since genotyping is error prone the accuracy of allele calling was estimated by the inclusion of four blind duplicates and seven DNA samples with known genotypes at each locus. The blind duplicates were all concordant and genotypes for the seven animals were accurately determined. The physical distances between adjacent sheep MHC class I region SNPs chosen for genotyping ranged from 4 kbp to 152 kbp, with an average frequency one SNP per 35 kbp. The cumulative distance between SNPs is shown in Figure 4.2. Figure 4.2 is based on approximate physical distances of orthologous loci in the better characterised cattle MHC class I region and shows that the SNPs chosen were relatively evenly distributed across MHC class I region. Table 4.5 shows the SNPs used for genotyping, location based on cattle reference genomic database (NW_001494164.1) and the distance between adjacent SNPs. Results from the Genepop analysis for each SNP are shown in Table 4.6. All SNPs exhibited Hardy-Weinberg proportions in the sample population.

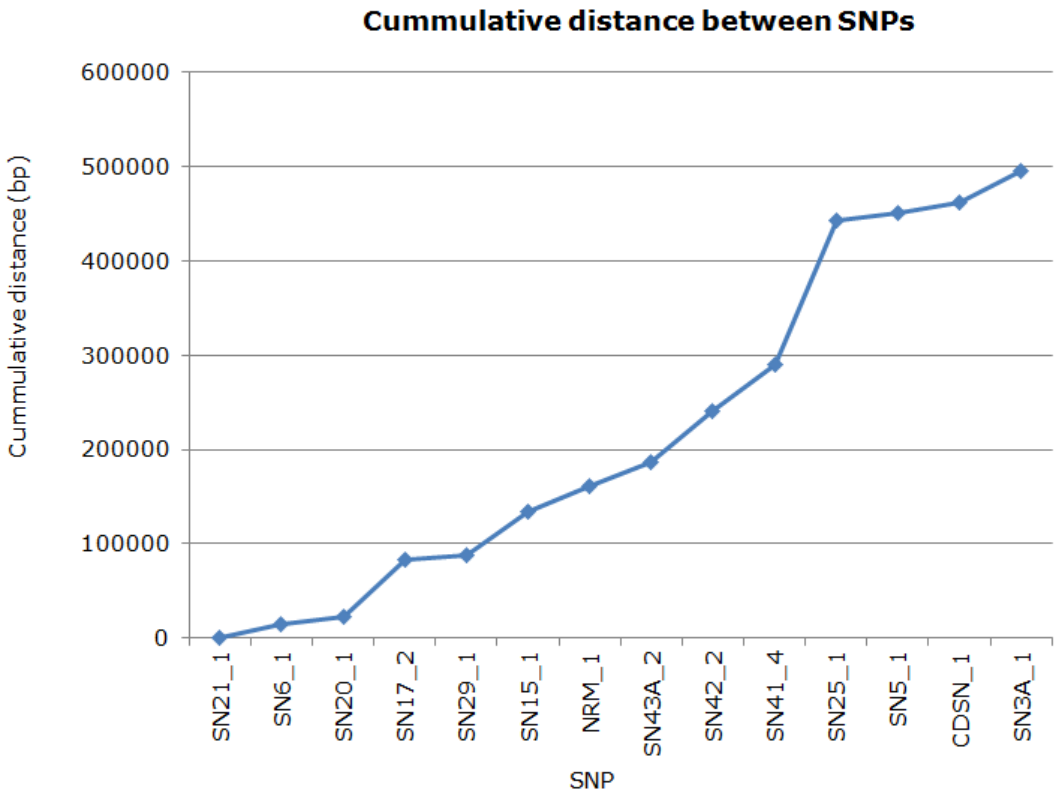


Figure 4.2: Cumulative distance between SNPs used for genotyping based on approximate physical distances of orthologous loci in the cattle MHC class I region.

Table 4.5: Details of SNPs used for genotyping across MHC class I region.

SNP identity	Base change	Locus	Location in NW_001494164	Gap between the next SNP (bp)
SN21_1	A/G	*RPP21 - OVAR-MHCI*	649260	14,208
SN6_1	C/T	*RPP21 - OVAR-MHCI*	635052	8,736
SN20_1	G/T	*GLN1 - RPP21*	626316	60,812
SN17_2	A/G	*PPP1R10 - PRR3*	565504	4,306
SN29_1	C/T	*PPP1R10 - PRR3*	561198	45,927
SN15_1	C/T	*C23H6orf136*	515271	27,428
NRM_1	A/G	*NRM*	487843	25,008
SN43A_2	C/T	*FLOT1 - TUBB*	462835	54,251
SN42_2	C/T	*DDR1 - IER3*	408584	49,950
SN41_4	A/G	*DDR1 - IER3*	358634	152,030
SN25_1	C/T	*LOC788708* - Hypothetical protein	206604	8,330
SN5_1	C/T	*LOC788708* - Hypothetical protein	198274	10,552
CDSN_1	C/T	*CDSN*	187722	33,555
SN3A_1	A/C	*POU5F1 - TCF19*	154167	-

Table 4.6: Summary of Genepop analysis on SNP genotypes across MHC class I region. P ≤ 0.0500 indicate rejection of the null hypothesis that mating was random (highlighted in red).

SNP	Samples typed	Frequency		Homozygosity		Heterozygosity		Hardy-Weinberg P-value
		Allele 1	Allele 2	Observed	Expected	Observed	Expected	
SN21_1	108	0.792	0.208	65	72.2	43	35.8	**0.0388**
SN6_1	108	0.903	0.097	87	89.0	21	19.1	0.5914
SN20_1	108	0.620	0.380	58	56.9	50	51.1	0.8392
SN17_2	108	0.449	0.551	49	54.3	59	53.7	0.3301
SN29_1	108	0.440	0.560	53	54.5	55	53.5	0.8454
SN15_1	108	0.727	0.273	65	64.9	43	43.1	1.0000
NRM_1	108	0.644	0.356	65	58.2	43	49.8	0.2028
SN43A_2	108	0.926	0.074	96	93.1	12	14.9	0.0999
SN42_2	108	0.079	0.921	91	92.3	17	15.7	1.0000
SN41_4	108	0.769	0.231	74	69.4	34	38.6	0.2760
SN25_1	108	0.532	0.468	45	54.0	63	54.0	0.1188
SN5_1	108	0.819	0.181	73	75.9	35	32.1	0.5238
CDSN_1	108	0.880	0.120	82	85.0	26	23.0	0.3565
SN3A_1	108	0.981	0.019	104	104.1	4	3.9	1.0000

4.3.3 Linkage disequilibrium

The SNPstats program was used to analyse linkage disequilibrium between adjacent SNPs across MHC class I region. In addition, the observed heterozygosity for each SNP locus was counted. The results of these analyses are summarised in Figure 4.3A, which shows the observed heterozygosity, pair-wise D' value, r^2 value and cumulative distance from the most telomeric SNPs. The observed heterozygosity is low at loci represented by SN6_1 (*RPP21-OVAR-MHCI*), SN43A_2 (*FLOT1-TUBB*), SN42_2 (*DDR1-IER3*) and SN3A_1 (*POU5F1-TCF19*). The observed heterozygosity is high at SN21_1 (*RPP21-OVAR-MHCI*), SN20_1 (*GLN1-RPP21*), SN17_2 (*PPP1R10-PRR3*), SN29_1 (*PPP1R10-PRR3*), SN15_1 (*C23H6orf136*), NRM_1 (*NRM*) and SN25_1 (*LOC788708*-hypothetical protein). Position SN41_4 (*DDR1-IER3*), SN5_1 (*LOC788708*-hypothetical protein) and CDSN_1 (*CDSN*) have intermediate heterozygosity.

Linkage disequilibrium, estimated by pair-wise D' values, (Figure 4.3B) are high between SN21_1 (*RPP21-OVAR-MHCI*) and SN6_1 (*RPP21-OVAR-MHCI*), between SN20_1 (*GLN1-RPP21*) and SN17_2 (*PPP1R10-PRR3*), between SN15_1 (*C23H6orf136*) and NRM_1 (*NRM*), between SN43A_2 (*FLOT1-TUBB*) and SN42_2 (*DDR1-IER*), and between SN25_1 (*LOC788708*-hypothetical protein) and SN5_1 (*LOC788708*-hypothetical protein). Low LD was observed between SN6_1 (*RPP21-OVAR-MHCI*) and SN20_1 (*GLN1-RPP21*), between SN17_2 (*PPP1R10-PRR3*) and SN29_1 (*PPP1R10-PRR3*), between NRM_1 (*NRM*) and SN43A_2 (*FLOT1-TUBB*), between SN42_2 (*DDR1-IER3*) and SN41_4 (*DDR1-IER3*), and between CDSN_1 and SN3A_1 (*POU5F1-TCF19*). Regions between SN29_1 (*PPP1R10-PRR3*) and SN15_1 (*C23H6orf136*), and between SN41_4 (*DDR1-IER3*) and SN25_1 (*LOC788708*-hypothetical protein) have intermediated pair-wise D' value. SN20_1 (*GLN1-PP21*) and SN43A_2 (*FLOT1-TUBB*) have sharp decreases in D' from 1.00 to 0.04 or less with the adjacent locus. In contrast, SN17_2 (*PPP1R10-PRR3*), NRM_1 (*NRM*) and SN5_1 (*LOC788708*-hypothetical protein) showed increases in D' from 0.1 to 1.00 with adjacent loci. Telomeric pair-wise D' values were generally higher than at the centromeric end.

Blocks with high LD levels and p-value are distributed randomly with no apparent trend of distribution within the class I region. Three short blocks of high LD with significant increase in D' and r^2 value are observed (grey box in Figure 4.3B); one at the telomeric end MHC class I region encompassing SN21_1 and SN6_1, which are located between *RPP21* and *OVAR-MHCI*, another encompassing SN17_2 (*PPP1R10-PRR3*) and SN29_1 (*PPP1R10-PRR3*), one other LD block at NRM_1 (*NRM*). A block of high LD with increased D' but intermediate r^2 was also observed at the centromeric end consisting SN5_1 (*LOC788708*-hypothetical protein) and CDSN_1 (*CDSN*) (represented as empty box in Figure 4.3B). LD for the data set analysed with the SNPstats program is also depicted as a heat map in Figure 4.4. Size of the high LD blocks range from approximately 30 kbp to 50 kbp, with reference to gap sizes between SNPs. These high LD blocks are interspersed with less significant blocks with varying p-values. The LD p-value between blocks is shown in Table 4.7. Regions with deceased LD are observed between SN6_1 (*RPP21-OVAR-MHCI*) and SN20_1 (*GLN1-PP21*), between SN29_1 (*PPP1R10-PRR3*) and SN15_1 (*C23H6orf136*), between NRM_1 (*NRM*) and SN43A_2 (*FLOT1-TUBB*), and between SN41_4 (*DDR1-IER3*) and SN25_1 (*LOC788708*-hypothetical protein). The relative size of the decreased LD blocks range from approximately 8.7 kbp to 50 kbp. The one exception is the LD block that extends from SN41_4 (*DDR1-IER3*) to SN25_1 (*LOC788708*-hypothetical protein), which is approximately 152 kbp.

A)

	SN21_1	SN6_1	SN20_1	SN17_2	SN29_1	SN15_1	NRM_1	SN43A_2	SN42_2	SN41_4	SN25_1	SN5_1	CDSN_1	SN3A_1
Allele 1	0.79	0.90	0.62	0.45	0.44	0.73	0.64	0.93	0.08	0.77	0.53	0.82	0.88	0.98
Allele 2	0.21	0.10	0.38	0.55	0.56	0.27	0.36	0.07	0.92	0.23	0.47	0.18	0.12	0.02
Obs het	0.39	0.20	0.47	0.55	0.51	0.40	0.40	0.11	0.16	0.32	0.59	0.33	0.24	0.04
D'	0.96	1.00	0.01	0.97	0.41	0.13	1.00	0.04	0.85	0.23	0.03	1.00	0.98	-
r^2	0.03	0.07	0.00	0.61	0.05	0.01	0.14	0.00	0.02	0.01	0.00	0.03	0.00	-
Distance (kbp)	0	14.20	8.70	60.80	4.30	45.90	27.40	25.00	54.20	50.00	152.00	8.30	10.60	33.60
Cum Dist kbp	0	14.20	22.90	83.70	88.00	133.90	161.30	186.30	240.50	290.50	442.50	450.80	461.40	495.00
Probability	0	0.01	0	0.91	0	0.001	0.11	0	0.57	0.04	0.08	0.82	0.01	0.47

B)

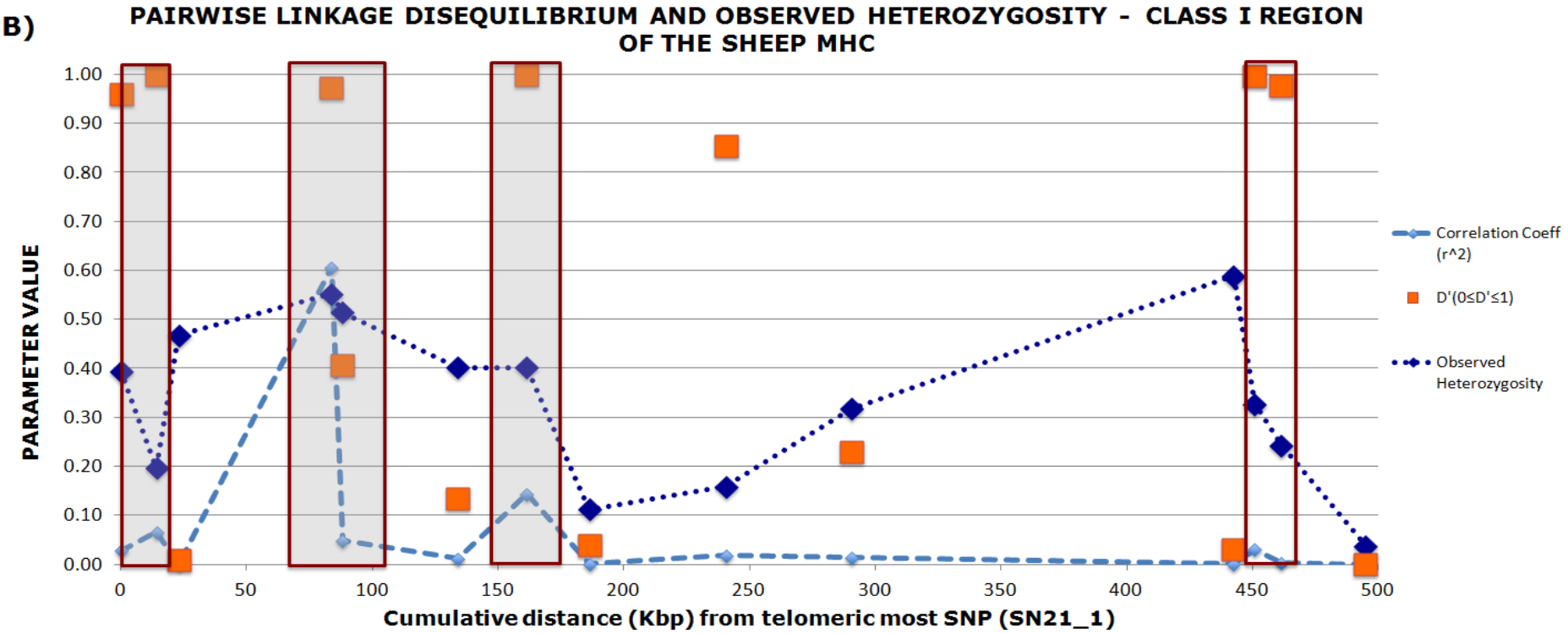


Figure 4.3: (A) Summary of SNP diversity across 14 polymorphic loci from the sheep MHC class I region showing allele frequencies, observed heterozygosity and LD estimates. (B) Graphical representation of observed heterozygosity and LD (D' and r^2) values for each adjacent pair of SNP loci across the MHC class I region. The grey boxes indicate regions where LD measures change significantly with increase in both D' and r^2 values. The empty box shows a region of high LD with significant increase in D' but intermediate value of r^2.

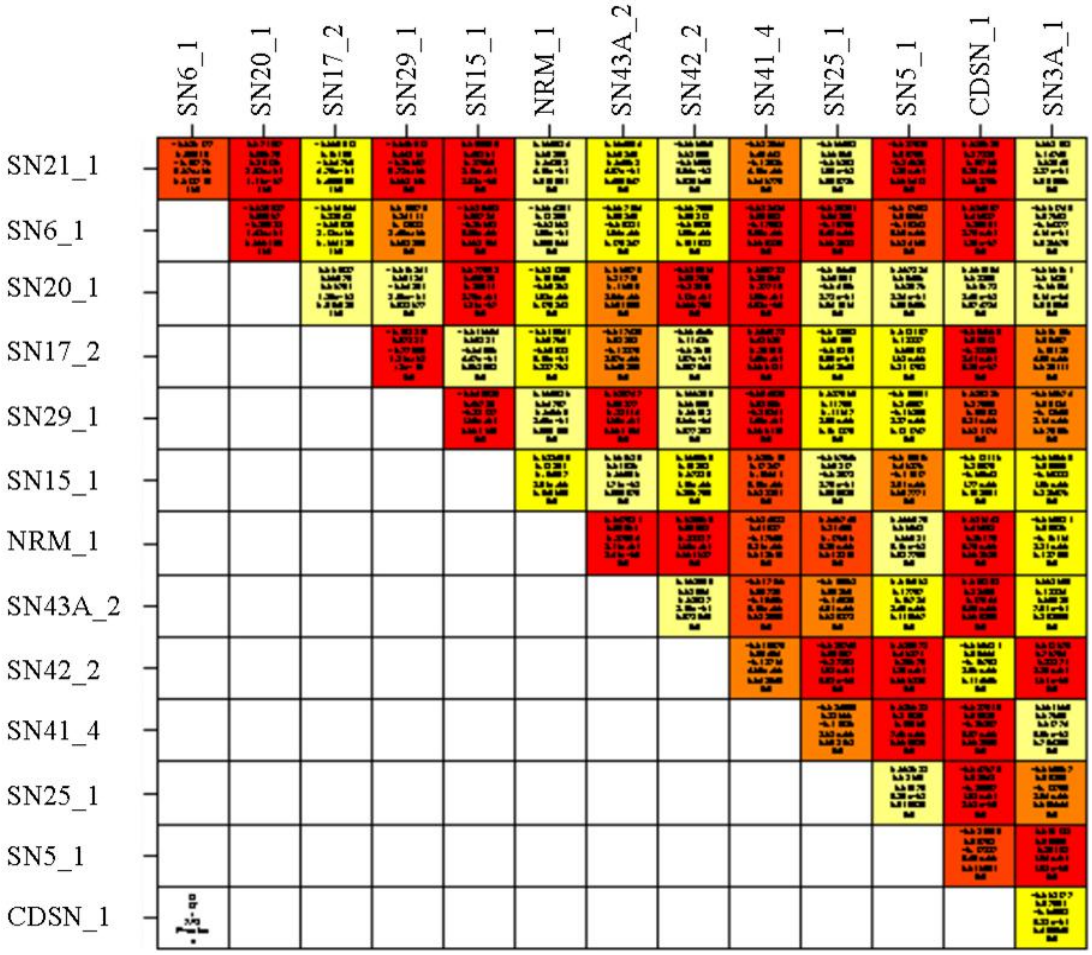


Figure 4.4: Heat map of LD pattern in sheep MHC class I region. Red represents block with highly significant p-value and very strong LD, orange represents block with significant p-value and moderate LD, dark yellow represents block with no significant p-value and weak LD. Light yellow represents block with no significant p-value and very weak LD.

Table 4.7: Pairwise LD probability values (LD≥0) extracted from the SNPstat analysis of 14 sheep SNPs used to generate Figure 4.4

	SN21_1	SN6_1	SN20_1	SN17_2	SN29_1	SN15_1	NRM_1	SN43A_2	SN42_2	SN41_4	SN25_1	SN5_1	CDSN_1	SN3A_1
SN21_1	.	0.0137	0	0.4889	0.0031	0	0.5196	0.4988	0.935	0.0408	0.9657	4e-04	0.0038	0.5617
SN6_1	.	.	2e-04	0.1441	0.0634	0.0032	0.6589	0.1753	0.1615	0.0083	0.0036	0.0241	0	0.5201
SN20_1	.	.	.	0.9109	0.5321	0	0.1762	0.0919	8e-04	0	0.5419	0.5695	0.8747	0.9819
SN17_2	.	.	.	.	0	0.5037	0.3377	0.0493	0.6575	1e-04	0.443	0.3118	0	0.0261
SN29_1	.	.	.	.	.	0.0011	0.5562	0.0012	0.9774	1e-04	0.1014	0.1317	0.0212	0.0762
SN15_1	.	.	.	.	.	.	0.1061	0.8959	0.2808	0.0233	0.5996	0.0878	0.1839	0.2207
NRM_1	.	.	.	.	.	.	.	0	0.001	0.012	0.0122	0.9378	0.003	0.1372
SN43A_2	.	.	.	.	.	.	.	.	0.5738	0.0229	0.0283	0.115	0.0084	0.383
SN42_2	.	.	.	.	.	.	.	.	.	0.0438	1e-04	2e-04	0.1141	0
SN41_4	.	.	.	.	.	.	.	.	.	.	0.0821	0.0065	0.0029	0.7943
SN25_1	.	.	.	.	.	.	.	.	.	.	.	0.8169	0	0.06
SN5_1	.	.	.	.	.	.	.	.	.	.	.	.	0.0109	0
CDSN_1	.	.	.	.	.	.	.	.	.	.	.	.	.	0.4658
SN3A_1	.	.	.	.	.	.	.	.	.	.	.	.	.	.

4.3.4 Class I haplotypes

Implementation of the expectation maximization algorithm in SNPstats predicted 34 haplotypes with a total frequency of approximately 90% from the 14 MHC associated loci. Rare haplotypes (<1%) accounted for the remaining 10%. The most common haplotype frequency was 8%. Three out of 108 animal cohort were homozygous at all loci.

4.4 Discussion

This chapter describes the discovery and annotation of 32 previously undescribed SNPs, from both intergenic and intragenic regions spanning approximately 505 kbp of the sheep MHC class I region. Fourteen SNPs were used to genotype the MHC class I region in distantly related sheep with an approximate average coverage of 1 SNP in every 35 kbp.

The observed heterozygosity of the 14 SNPs in this study varied from 0.04 to 0.59. These values are similar to those reported by Pariset *et al*. (2006) for 37 breed specific SNPs in 16 animals from 8 distinct breeds where the observed heterozygosity varied from 0.01 to 0.48. Miller *et al*. (2011) showed that in two North American wild sheep species (*Ovis canadensis* and *Ovis dalli*), SNP heterozygosity ranged from 0.02 to 0.83 with 14/308 markers deviating from HW proportions. In the present study one locus (SN21_1) deviated from HW proportions. However, after allowing for anomalous results due to chance and correction for multiple testing (after Bonferroni), the locus exhibited no significant deviation from Hardy-Weinberg proportions.

Data presented in this chapter extend considerably the SNP map for the sheep MHC (Lee *et al.* 2011) and provide preliminary evidence for regions of high and low LD that may define an internal block-like structure as has been reported by other studies (Qin *et al.* 2008; Lee *et al.* 2011). The number of SNPs used for genotyping the MHC class I region is similar to another recent study in sheep investigating the LD pattern across the MHC class II and III regions (Lee *et al.* 2011). The study by Lee *et al*. (2011) focused on the two

subsets of class II (class IIa and IIb) and class III region and a total of 30 SNPs were genotyped (10 SNPs for each of the three regions). Most SNP related studies analysing the human MHC have relied upon considerably larger numbers of SNPs (Stenzel *et al.* 2004; Miretti *et al.* 2005; Traherne *et al.* 2006) but very large panels of SNPs are not yet available for use in sheep.

A study of Australian cattle using a high density SNP panel reported that significant LD in cattle extends to 8.2 Mbp when estimated by D', which suggested therefore, that a minimum of 1 SNP is required in 8.2 Mbp for a low power LD scan (Khatkar *et al.* 2008). Mean marker density has been suggested to influence identification of structure or range of LD blocks (Phillips *et al.* 2003). Short blocks are known to be more visible with high marker density and long LD blocks emerge preferably when widely spaced markers are used (Phillips *et al.* 2003). However Stenzel *et al*. (2004) has also argued that estimates of LD block size are independent of local SNP density. Since estimates of LD between loci are very sensitive to low allele frequencies, identification of genomic blocks defined by high LD is more likely to depend on the choice of SNP loci and sequence stability rather than SNP density per se (Stenzel *et al.* 2004). SNPs with minor allele frequencies may well be recent mutations and exhibit longer-range LD (Stenzel *et al.* 2004).

This study has shown that pairwise LD varies across ≈500 kbp of the sheep MHC class I region and that regions of high LD are separated by regions of low LD. These results therefore support the hypothesis that the sheep MHC is composed of block like subregions as has been demonstrated for the human MHC (Stenzel *et al.* 2004). Higher levels of LD occurred closer to the telomeric end of the MHC, although this trend was not analysed statistically. It is consistent however with reports of higher levels of pairwise LD in telomeric subregions relative to centromeric subregions for the orthologous human MHC (Stenzel *et al.* 2004; Traherne *et al.* 2006).

Estimating LD parameters can be difficult. The two commonly used measures of LD are the D' and r^2. The D' estimates are very sensitive to low allele frequencies (and hence low heterozygosities) and can be easily

inflated when small sample sizes are used. The r^2, based on the common correlation coefficient between alleles at adjacent loci is less sensitive to allele frequency, but can be affected by low heterozygosities. Both measures have been used to interpret the results from this study. In his study of the human MHC, Stenzel *et al*. (2004) revealed that sample populations of greater than 45 unrelated individuals are sufficient to delineate haplotypic structures and that stable D' values are obtained for populations of approximately 45-550 individuals. It is possible however that within the breed selected sheep population, larger sample size is required to achieve the same degree of reproducibility reported by Stenzel *et al*. (2004) in humans. A recent study has also showed that r^2 has allele frequency dependence, and this is especially true when the minor allele frequency is less than 0.3, which occurs quite often when using SNPs (Van Liere & Rosenberg 2008). Van Liere and Rosenberg (2008) reported a direct relationship between r^2 and D', for some but not all domains in their study. The results reported herein are therefore consistent with the findings of the human MHC. The most likely explanation for the results is that the sheep MHC class I region contains subregions (i.e. blocks) characterised by high LD where recombination is rare, separated by usually smaller subregions where recombination is more frequent due to the presence of sequence motifs referred to as recombination hotspots. One such motif is the zinc finger protein PDRM9 which has been recently implicated in a family of "recombination hotspots" in a variety of mammals (Oliver *et al.* 2009; Thomas *et al.* 2009; Baudat *et al.* 2010; McVean & Myers 2010; Parvanov *et al.* 2010).

A heat map of LD between SNPs in the sheep MHC class I region showed some evidence for "block-like structures" with varying p-values and sizes. Three regions of very weak or decreased LD are present within the sheep MHC class I region indicating an abrupt breakdown of LD. The LD was reduced remarkably at *PPP1R10*-PRR3, *DDR1-IER*, and the junction of *LOC788708*-hypothetical and *CDSN*. The size of these blocks ranges from approximately 4 kbp to 50 kbp. The breakdown of LD could suggest that these loci are hotspots for recombination relative to the LD blocks. The actual size of recombination hotspots relative to the LD blocks identified within class I region of sheep MHC is uncertain because the exact

boundaries of each hotspot relative to the LD blocks could not be deduced. Therefore, the size of each hotspot relative to the LD block is an estimate based on the distance between SNPs that represents the adjacent locus. A study of meiotic recombination in human MHC class II region has reported that hotspots do not occur randomly, but in clusters of 60 – 90 kbp apart (Jeffreys *et al.* 2001).

A study of human MHC haplotypes comprising over 259 loci and using 20,000 SNPs has identified variation density within the MHC (Traherne *et al.* 2006). The result of this study showed that MHC displays varying levels of heterozygosity for different loci. The highest variation density was observed in the classical MHC class I and II loci, which is known to be maintained by balancing selection to provide heterozygous advantage for molecular function at the peptide-binding domains (Hughes & Nei 1988, 1989b). Locus telomeric of *HLA-C* from *CDSN* to *POU5F1* has been previously reported to show independent selection for variation (Stewart *et al.* 2004; Traherne *et al.* 2006). Analysis of LD in this study shows that levels of heterozygosity was high at SN21_1 (*RPP21-OVAR-MHCI*), SN20_1 (*GLN1-RPP21*), SN17_2 (*PPP1R10-PRR3*), SN29_1 (*PPP1R10-PRR3*), SN15_1 (*C23H6orf136*), NRM_1 (*NRM*) and SN25_1 (*LOC788708*-hypothetical protein). The high level of observed heterozygosity at SN21_1 could be due to the hitch-hiking effect from adjacent classical class I gene known to show high variability. Hitch-hiking alters the frequency of alleles at closely linked loci when a favourable mutation arises and becomes fixed in the population (Smith & Haigh 1974). Functional advantageous mutation at a locus increases the frequency of the allele and as such also lifts the frequency of neighbouring alleles (Smith & Haigh 1974). Direct comparison of other loci that have high level of heterozygosity in this study and previously performed MHC studies is not feasible because of availability of incoherent data, especially the difference in loci being genotyped. Other factors that prevent direct comparison are the difference in the number of molecular markers used for genotyping and the difference in sample size.

The general pattern of heterozygosity distribution observed in MHC class I region in this study is different compared to the previously analysed sheep class II region. The study investigated LD pattern by genotyping 10 loci

within each of the class II sub-regions (class IIa and IIb) of the sheep MHC in a population of half-sibling family groups (Lee, 2011). A consistently low level of observed heterozygosity was detected in the class IIa region, which extends from *DQB2-DQA2* to *BTNL2-C6orf10* (Lee, 2011). The stretch of low heterozygosity region in class IIa region has also been observed in human MHC class II region (Traherne *et al.* 2006). This region has been described as "SNP desert" due to its lack of polymorphisms (Traherne *et al.* 2006; Lee *et al.* 2011). The class IIb region in sheep showed variable levels of heterozygosity, similar to finding in the sheep MHC class I region in this study. Varying levels of heterozygosity in the sheep MHC class I region suggested that various loci in this region have been through independent balancing selection to maintain the vast variety of immune and non-immune related functions, similar to other mammals (Hedrick 1998; Aguilar *et al.* 2004). This characteristic of MHC makes it the most polymorphic region in the genome (Trowsdale 1993).

As will be reported in Chapter 6, the 14 SNPs used in this study were distributed across many haplotypes, of which none had a frequency greater than 8% in the 108 animals typed. This result suggests that LD, at least within the MHC, does not manifest across long physical distances (Mbps) as has been reported by others for other region (Austerlitz & Heyer 1999; McRae *et al.* 2002; Odani *et al.* 2006; Kijas *et al.* 2009; Miller *et al.* 2011; Kijas *et al.* 2012).

Further research is required to increase the saturation of the sheep SNP map to allow the elucidation of any hotspots for recombination. The identification and analysis of more markers using a larger cohort of defined pedigrees will enable a greater understanding of the fine structure of this region. In addition, analysis of different breeds of sheep will place the structure of the MHC into an evolutionary context.

Chapter 5

Analysis of MHC Class I Gene Diversity in Australian Merino Sheep

The work described in this chapter addresses the complex problem of estimating the number of class I loci in the sheep MHC. The approach used multiple sequence alignments (MSAs) of class I genomic sequences from sheep bred for MHC class I homozygosity by means of sire/daughter matings. Criteria for homozygosity of class I loci were homozygosity of four microsatellite loci within the class I region plus homozygosity of a further 14 single nucleotide polymorphisms (SNPs) within known class I loci. The MSAs used also contained class I sequences from the BAC sheep clones described by Gao et al. (2010). Interpretation of the MSAs generated was further facilitated by construction of phylogenetic trees derived from the MSAs using the maximum likelihood and neighbour joining algorithms. It was expected that sequence analysis of class I loci from homozygous sheep would minimize allelic variation thereby facilitating identification of inter locus variability. This strategy was partially successful, however the technical and analytical problems associated with MHC class I mutation, recombination, duplication and gene conversion were still present to a high degree thereby restricting the potential of this strategy. In this study, fourteen class I loci were observed in the sheep MHC. Evidence of transcribed pseudogenes have also been identified. At least 11 transcribed pseudogenes are present, 4 of which are transcribed into truncated alpha chains and are therefore unlikely to be functional. Classification of sequences based on the MHC Immuno Polymorphism Database defined groups described in this study indicates that the number of distinct loci is less (and possibly much less) than the total number of groups predicted. Information from this chapter is being prepared for publication.

5.1 Introduction

The Major Histocompatibility Complex (MHC) class I genes are highly polymorphic and encode class I molecules involved in presentation of a wide repertoire of self and pathogenic antigens for recognition by CD8+ T lymphocytes (Harty *et al.* 2000). The most variable regions within class I gene are exons 2 and 3, which code for the peptide binding domains $\alpha 2$ and $\alpha 3$ (Hughes & Nei 1988, 1989a; Madden *et al.* 1993; Ellis 2004). It is widely accepted that high levels of polymorphism in this region are maintained primarily through natural selection (Hughes & Nei 1988, 1989a). Generation of genetic diversity leads to functional advantage as a wider array of antigens derived from pathogens can be presented to T cells. This enhances immune responsiveness against a greater variety of infectious agents (Hughes & Nei 1988).

In addition to polymorphisms within the gene, variation in copy number and gene size of MHC class I genes occurs within and between species (Trowsdale 1995). In humans, all three class I genes (HLA-A, HLA-B and HLA-C) are consistently expressed (Parham *et al.* 1995). The MHC class I genes in other mammals have more complex expression patterns with haplotypes expressing different combinations of class I loci. In horse, there is evidence for at least four genes being expressed, although examination of a conserved region within the genes suggested that there may be five expressed genes (Ellis *et al.* 1995). Similarly, six or more class I loci have been reported in cattle (Ellis 2004). Studies on sheep based on cellular localisation, phylogenetic and expression analysis have indicated that eight or more expressed class I loci contribute to some sheep MHC haplotypes (Ballingall *et al.* 2008).

Assignment of MHC class I alleles to a specific locus is difficult because there is no evidence of a single gene that is consistently expressed (Ellis 2004). In addition, it has been shown that MHC class I alleles in ruminants can be 'hybrid' genes generated from inter-locus recombination, further complicating the assignment of a given allele to a designated locus (Holmes *et al.* 2003). A recently published physical map derived from Chinese Merino sheep reported discovery of five classical and two non-classical MHC class I

genes (Gao *et al.* 2010). However the actual number of class I loci in sheep is still uncertain and is very complex to elucidate. A greater understanding of MHC class I genes and their expression would be helpful in development of vaccines against intracellular pathogens (Miltiadou *et al.* 2005; Ballingall *et al.* 2008). One approach to resolving these issues would be to examine class I genes in sheep that are homozygous for the MHC, since allelic variation is eliminated (or at least much reduced). This may allow the number of expressed loci to be more easily identified.

In this project, blood samples from lambs produced through father-daughter mating were used to establish a flock of sheep homozygous at the MHC region. The objectives of this study are to amplify and sequence genomic DNA and complementary DNA (cDNA) from MHC class I genes present in these homozygous individuals, annotate the sequences and to compare the sequences with published MHC class I cDNA sequences (Miltiadou *et al.* 2005; Ballingall *et al.* 2008). This study will also analyse the MHC class I genes present in the BAC clones sequences submitted to GenBank by Gao *et al.* (2010) to assist in determining the number of loci present in sheep.

5.2 Materials and methods

5.2.1 Animals and DNA extraction

Homozygous animals were generated by artificial insemination (AI) through sire-daughter mating from a flock of Merino sheep maintained at the Department of Agriculture and Food, Katanning, Western Australia. Five different sires were mated with 20 daughter ewes (total 100 ewes). Blood samples were collected by venepuncture into 10mL K_3EDTA vacutubes (Vacuette) and the samples stored at -20^{o}C (Chapter 2.1). Genomic DNA was extracted according to manufacturer's protocols using the Fisher Biotech Easy DNA (EDNA) High-Speed Extraction Blood Kit.

5.2.2 Selection of MHC homozygous animals

All the samples were genotyped using previously identified microsatellites at four different MHC loci; primers SMHCC (Groth & Wetherall 1994), SKIV2LM (Groth & Wetherall 1995), and OLADRB (Schwaiger *et al.* 1993) and OLADRBps (Blattman & Beh 1992). PCR was performed in 10µl reactions using the following conditions; 50 ng of DNA, 200µM of each dNTP (Roche), 1.5mM $MgCl_2$, 1 X PCR reaction buffer (Roche), 0.2 units of FastStart Taq DNA polymerase (Roche) and 1 pM of each primer (Geneworks). PCRs were performed on an Eppendorf Mastercycler (Eppendorf). The annealing temperature (T_{ann}) used for the PCR reactions varied according to the melting temperature of the primer sets used. The typical cycling conditions were as follows: 95 ^{o}C for 10 minutes; 35 cycles of 94 ^{o}C for 30 s; T_{ann} for 30 s; and 72 ^{o}C for 30 s; and a final extension step at 72 ^{o}C for 5 minutes. Amplicons approximately 200 bp were electrophoresed on a 10% polyacrylamide gel (30% w/v acrylamide: bisacrylamide (29:1), 10% w/v ammonium persulfate, 1X Tris-boric-EDTA (TBE) and tetramethylethylenediamine (TEMED)). Homozygosity was confirmed for samples that were homozygous at all four microsatellite loci through genotyping of 14 MHC class I SNPs (described in Chapter 6).

5.2.3 Polymerase Chain Reaction to amplify MHC class I genes

Three PCR primer sets (MHC 1AF and MHC 1AR, MHC 1BF and MHC 1BR, and MHC 1CF and MHC 1CR shown in Table 5.1) were designed to amplify full-length MHC class I genes based on alignment of different copies of the gene previously identified by Gao *et al*. (2010). Primer set MHC 1A and MHC 1B had the same forward primer sequence. Primers published in previous sheep studies were also attempted (Miltiadou *et al*. 2005; Ballingall *et al*. 2008). PCR amplification was performed using FastStart High Fidelity PCR System (Roche). Final volume of 10 µL reaction contains 9 µL of master mix (1.8 mM $MgCl_{2,}$ 10% DMSO, 200 µM of each dNTP, 0.4 µM of each primer and 2.5 U FastStart High Fidelity Enzyme Blend) and 1 µL of DNA template (100 – 200 ng). 5 X 10 µL reactions were prepared for each animal and the products pooled before cloning. The cycling conditions for the PCR were as follow: initial denaturation for 2 minutes at 94 ^{o}C, 10 cycles of 30 s at 94 ^{o}C,

30 s 59 ^{o}C and 5 minutes at 68 ^{o}C, 30 cycles of gradual increase in extension time of 30 s at 94 ^{o}C, 30 s 59 ^{o}C and 5 minutes + 20 s cycle extension for each successive cycle at 68 ^{o}C, and final extension for 7 minutes at 68 ^{o}C. PCRs were performed on an Eppendorf Mastercycler (Eppendorf). Agarose gel electrophoresis was used to confirm the correct size product is amplified.

Table 5.1: Primers used amplify complete MHC class I gene, exon 2 of MHC class I gene and sequence clones containing MHC class I gene.

Locus	**Forward Primer (5'-3')**	**Reverse Primer (5'-3')**
MHC 1A	CTCCTCGAGTTTCACTTTCT	GTAAGGGCTGACATTCTCCA
MHC 1B	CTCCTCGAGTTTCACTTTCT	GTAAGCACTCATATTCTCCA
MHC 1C	CTCCCCGACTTTCACTTTCT	GTAAGGGCTSACATTCTCCA
MHC Ex2	GGCTCCCACTCCCAGAGG	CGGCCTCGCTCTGGTTGTAGTAGCC
MHC I3	CTGTACTAGACGGTGACTTG	TACAGAGCAATGGTCCTGAC
MHC Ex4	AGATCTCACTGACCTGGCAG	CTGCCAGGTCAGTGAGATCT

5.2.4 Amplification of exon 2 of MHC class I genes

Pooled product from PCR amplification of full-length MHC class I genes were diluted 1:100 and used as template for re-PCR with MHC exon 2 primers (MHC Ex2F and MHC Ex2R in Table 5.1) to confirm the presence of MHC sequence prior to cloning. Exon 2 amplification was performed using the standard PCR protocol described in Chapter 2.4.

5.2.5 Reverse transcriptase PCR

Synthesis of complementary DNA (cDNA) from total RNA and subsequent PCR of the cDNA was performed using MyTaq^{TM} One-Step RT-PCR Kit (Bioline). Primer MHC Ex4 (Table 5.1) was used to amplify the cDNA template. In a 10 µL PCR reaction, there were 5 µL of 2x My Taq One-Step mix, 1 µL of forward primer (10 µM), 1 µL of reserve primer (10 µM), 0.1 µL of reverse transcriptase, 0.2 µL of RiboSafe RNAse inhibitor, 1.7 µL of DEPC-water and 1 µL of total RNA template. The cycling condition was as follows: 1 cycle of 45^{o}C for 45 minutes, 1 cycle of 95^{o}C for 5 minutes and 40 cycles

of 95°C for 20 sec, 60°C for 30 sec, 72°C for 1 minutes. PCR product was analysed on 0.7% agarose gel.

5.2.6 Cloning MHC class I genes

MHC class I gene PCR products were ligated into pGEM –T Easy Vector System (Promega). Reaction components consist of 5 µL of 2X Rapid Ligation Buffer, 1 µL pGEM –T Easy Vector, 3 µL PCR product, and 1 µL T4 DNA Ligase. Ligation reaction was incubated overnight at 4 °C. The ligation reaction (1 µL) was transformed into 20 µL of ELECTROMAX DH5a-E *Escherichia coli* (Invitrogen) cells using electroporation. SOC media (800 µL) was added to the transformation mix and the mixture incubated at 37 °C for 1 hour with gentle agitation. The transformation mixture (100 µL) was plated onto 1.5% w/v Luria Bertani (LB) agar plate supplimented with 0.1 mM isopropyl beta-D-thiogalactopyranoside or IPTG (Sigma), 40 µg/mL 5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside or Xgal (Sigma), and 75 µg/mL ampicillin and incubated at 37 °C for 18 hours.

5.2.7 Selection of recombinant clones

White colonies indicating successful recombination were identified on plates. To confirm the correct insert had been cloned, individual white colonies were inoculated into 500 µL LB broth containing 75 µg/mL ampicillin and grown at 37 °C for 18 hours with shaking. The overnight cultures were then diluted 1:100 in high pure water and used as template for PCR with MHC exon 2 primers (MHC Ex2F and MHC Ex2R) using the standard PCR protocol described in Chapter 2.4. Positive clones were extracted using Axyprep™ Plasmid Miniprep Kit (Axygen) as described in Chapter 2.2.2.

5.2.8 Sequencing MHC class I genes

Purified plasmid DNA was sent to Macrogen Inc., Korea for direct sequencing with MHC class I gene specific primers; MHC 1F, MHC 1R, MHC Ex2F, MHC Ex2R, MHC I3F, MHC I3R, MHC Ex4F and MHC Ex4R (shown in Table 5.1). Sequencing of each clone was done multiple times. Sequences were

assembled and screened for vector contamination using either Vector NTI (Invitrogen) or Geneious (Drummond *et al*. 2011 – unpublished).

5.2.9 Identification and annotation of the class I genes

Gao *et al*. (2010) published a DNA sequence map of the ovine MHC. BAC clone sequences from the class I region of the map were downloaded from NCBI and screened for the location of MHC class I genes. BACs FJ985869, FJ985854, FJ985864, FJ985868, FJ985852, FJ985875 and FJ985859 were reverse complemented before analysis in order to provide for a contiguous gene map in the 5' to 3' direction. BAC sequences were masked for repeats with Repeatmasker, open version 3.2.9, then analysed with GENSCAN (http://genes.mit.edu/GENSCAN.html) and Softberry FGENESH (http://linux1.softberry.com/all.htm). Predicted transcripts were submitted to the basic local alignment search tool (BLAST) at the National Center for Biotechnology Information (NCBI) to identity putative MHC class I transcripts by homology to known class I genes previously reported in *Ovis aries* or *Bos taurus* (http://blast.ncbi.nlm.nih.gov/Blast.cgi). To refine predictions for putative class I genes, BAC sequences were subsequently analysed with FGENESH+ using one or more of the best matching MHC class I proteins as a homologue. *Bos taurus* was chosen as the model organism for both FGENESH and FGENESH+ and up to five variant transcripts were considered. In the case where multiple potential class I genes were identified in a single BAC sequence, FGENESH+ gene prediction was localised to each particular region of interest. Suitable transcripts were selected based on alignment with known class I genes.

Genomic contigs isolated in this study were also analysed with Softberry FGENESH+ (http://linux1.softberry.com/all.htm) using a panel of homologues consisting of translated MHC class 1 mRNAs from sheep – accessions NM_001130934, AJ874675, AJ874678, AJ874678, AJ874683 and U03092, along with MHC class I predicted proteins from GAO's BACs.

5.2.10 Identification of protein domains

Predicted MHC class I proteins were checked for the presence of signal peptide (leader sequence) using the SignalP 4.0 server (http://www.cbs.dtu.dk/services/SignalP/). Predicted proteins were screened for the presence of MHC class I domains using the NCBI Conserved Domain Database (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and Pfam (http://pfam.sanger.ac.uk/). The Pfam database proved more sensitive for detection of the cytoplasmic (CP) domain. Transmembrane (TM) domain was predicted using the TmPred program provided by EMBnet (http://www.ch.embnet.org/software/TMPRED_form.html).

5.2.11 Assignment of groups to MHC class I sequences

The six homozygous animals used in this study generated multiple MHC class I sequences. For each animal, several types of DNA sequences were generated, namely full-length genomic, cDNA exons 1-3 and genomic exons 1-3. Unique amino acid sequences derived from each of these above sequence types were selected for multiple sequence alignments and construction of phylogenetic trees. In preliminary analysis, the sequences used to assign groups using the MHC - Immuno Polymorphism Database (IPD) classification (http://www.ebi.ac.uk/ipd/mhc/) were limited to sequences generated from homozygous animals and reference sequences from BACs to simplify the interpretation of results. Subsequent analysis included all MHC class I sequences. MHC-IPD nomenclature system specifies that alleles belonging to a group must have no more than four differences in amino acid sequence in the α1 (exon 2) and α2 (exon 3) domains and no more than four sequence differences elsewhere. In addition, putative class Ib (non-classical) sequences are distinguished from class Ia (classical) based on the presence of a VPI or IPI motif in the TM domain and/or the presence of a truncation at the same location in the CP domain as in previous studies in sheep (Miltiadou *et al.* 2005) and cattle (Holmes *et al.* 2003; Davies *et al.* 2006; Birch *et al.* 2008a). The number of class I groups present in individual animals was determined from an analysis of the multiple sequence alignments of predicted amino acid sequences.

5.2.12 Sequence alignment and phylogenetic analysis

Predicted amino acid sequences were aligned using MUSCLE (Edgar 2004) with default parameter settings. Alignments were edited manually (as required) to provide optimal alignments in Seaview version 4.2.12 (Gouy *et al.* 2010). Maximum likelihood phylogenetic trees in the Seaview package were inferred using PhyML (Guindon *et al.* 2010) based on the LG matrix model (Le & Gascuel 2008). MEGA5-1.05 (Tamura *et al.* 2011) was used to infer maximum likelihood (ML) and neighbor-joining (NJ) phylogenetic trees based on the JTT matrix model (Jones *et al.* 1992). Sequence AAB69337 from *Sus scrofa* was selected as the out-group for all trees.

5.3 Results

5.3.1 Generation of homozygous animals

The sire-daughter matings resulted in 38 live births. DNA typing showed that 8 animals were homozygous for all 4 MHC microsatellite loci. Subsequent genotyping using an additional 14 MHC class I region SNP loci also showed that these 8 animals were homozygous for all loci (microsatellite and 14 SNP loci). The 8 MHC homozygous animals observed is a significant departure from the 19 expected (X^2 = 11.1, 1 df, $p < 0.001$).

5.3.2 Identification of predicted class I genes from BAC sequences

Analysis of the BAC sequences (Gao *et al.* 2010) from GenBank indicated the presence of OVAR MHC class I gene sequences in five BACs. Table 5.2 shows the predicted MHC class I peptide and the estimated positions of the class I domains within each predicted peptide sequence. Three putative MHC class I-like genes were found in BACs FJ985864 and FJ985870; in each case the corresponding gene from these two BAC sequences was identical. This was not unexpected as these two BAC sequences overlap. BAC sequences FJ985875, FJ985859, FJ985874 contained two, three and three putative MHC class I genes respectively. The last gene in BAC FJ985859 sequence is

identical to the first gene identified in BAC FJ985874 except for a 3 nucleotide indel, which most likely represents an allelic variation. Pairwise alignment of the two BACs indicates an overlap in the region containing the putative MHC class I gene. Table 5.3 summarises the predicted MHC class I gene structures identified from the BAC sequences.

Table 5.2: MHC class 1-like genes predicted from genomic BAC sequences published by Gao *et al*. (2010).

Name	**S**	**Location**	**#Ex**	**AA len**	**SP**	**α1/α2**	**α3**	**TM**	**C-Term**	**Type**	**Homologue**
FJ985864rc_C1a	+	53040 - 56369	8	364	Y, 25	26-203	207-299	310-331	336-363	Ia	NP_001124406
FJ985864rc_C1b	+	97997 - 113333	5*	229	N	27-93	95-186	194 215		Ib	NP_001124406
FJ985864rc_C1c	+	130285 -133430	7	346	Y/22	23-200	205-296	304- 325	329-346**	Ib	NP_001124406
FJ985870_C1a	+	26626 - 30134	8	364	Y, 25	26-203	207-299	310-331	336-363	Ia	NP_001124406
FJ985870_C1b	+	71762 - 87098	5*	229	N	27-93	95-186	194 215		Ib	NP_001124406
FJ985870_C1c	+	104051 - 107196	7	346	Y/22	23-200	205-296	304- 325	329-346**	Ib	NP_001124406
FJ985875rc_C1a	-	127171 - 130033	7	354	Y/25	26-203	207-299	307 -329	332-354	Ib	NP_001035644
FJ985875rc_C1b	-	160117 - 162943	6	337	Y/25	26-187	189-233	242-262	267-279**	Ia/Ib	NP_001035644
FJ985859rc_C1a	-	1806 - 4642	7	356	Y/25	26-203	207-299	308 327	332-356**	Ia/Ib	CAI43976
FJ985859rc_C1b	-	28878 - 34806	7	355	Y/18	28-204	208-300	309 329	333-355	Ia/Ib	NP_001124406
FJ985859rc_C1c	-	148998- 151240	5	356	Y/25	52-229	233-325	333 350		Ib	AAZ74696
FJ985874C1a	-	6993 - 9231	5	355	Y/24	51-228	232-324	332 349		Ib	AAZ74696
FJ985874C1b1***	-	34484-37396	8	383	Y/22	23-200	204-296	305 328	331-359	Ia/Ib	CAI43976
FJ985874C1b2***	-	34390 - 37396	8	360	Y/22	23-200	204-296	305 328	331-359	Ia/Ib	FJ985864rc_C1a
FJ985874C1c	-	77708 - 81297	7	379	Y/25	26-203	207-299	305 328	333-355**	Ia/Ib	NP_998933

S: strand. #Ex: number of exons identified. AA len: Length of predicted amino acid sequence. SP: signal peptide. α1: alpha 1 domain. α2: alpha 2 domain. α3: alpha 3 domain. TM: transmembrane domain. C-Term: C-terminus end. N: no signeal peptide identified. Y: Signal peptide identified and location of cleavage site indicated. *Terminal exon not predicted. **Weak match. ***Alternative transcript predictions.

Table 5.3: Comparison of exon structure of MHC class 1 genes isolated from BAC sequences (exon length predicted by FGENESH+).

Name	**S**	**Ex 1**	**Ex 2**	**Ex 3**	**Ex 4**	**Ex5**	**Ex 6**	**Ex 7**	**Ex 8**
FJ985864rc_C1a	+	72	267	273	273	111	30	45	3
FJ985864rc_C1b	+	72	210		270	99	27		
FJ985864rc_C1c	+	63	267	273	273	102	30	15	
FJ985870_C1a	+	72	267	273	273	111	30	45	3
FJ985870_C1b	+	72	210		270	99	27		
FJ985870_C1c	+	63	267	273	273	102	30	15	
FJ985875rc_C1a	-	72	267	273	273	102	30	30	
FJ985875rc_C1b	-	72	264	219	135	102	210		
FJ985859rc_C1a	-	72	267	273	273	102	30	36	
FJ985859rc_C1b	-	78	267	270	273	102	30	30	
FJ985859rc_C1c	-	72	345	273	273	96			
FJ985874_C1a	-	69	345	273	273	96			
FJ985874_C1b1*	-	63	267	273	273	108	30	120	
FJ985874_C1b2*	-	63	267	273	273	108	30	45	3
FJ985874_C1c	-	72	267	270	273	102	21	114	

S: strand. Ex: exon. *Alternative transcript predictions

The gene identified as C1a in BAC sequences FJ985864 and FJ985870 has eight predicted exons (Table 5.3). The first exon encodes for a signal peptide with a predicted cleavage site located at the 24th amino acid. Search of the NCBI Conserved Domain Database (CDD) and Sanger Institute Pfam confirmed the presence of MHC class I antigen binding domains ($\alpha 1$ and $\alpha 2$) located in exons two and three, alpha chain immunoglobulin ($\alpha 3$) domain located in exon 4, a transmembrane (TM) domain located in exon 5, and a cytoplasmic (CP) domain located in exons 6, 7 and 8. The predicted protein has the classical FLT amino acid motif in TM domain.

The C1b gene from BACs FJ985864 and FJ985870 has five predicted exons (Table 5.3). However, the first exon does not show significant sequence similarity to the reference exon 1 sequences. SignalP predicts no cleavage site, indicating that it is unlikely that this represents a signal sequence. Exon two is shorter than expected and more variable in comparison with other class I sequences. However, a match to MHC class I histocompatibility antigen domain $\alpha 1$ was present but the $\alpha 2$ domain is missing and replaced

with the α3 domain. Exon 4 contains the TM domain with a non-classical IPI amino acid motif. The fifth (and last) exon corresponds to a truncated CP domain, however this has not been identified as a terminal exon by FGENESH+. An alternative FGENESH prediction includes a short terminal exon which does not share amino acid similarity with any of the available class I reference sequences in this region. This sequence is either a highly divergent pseudogene or a product from a distinct but related MHC class I-like locus.

The C1c gene from BACs FJ985864 and FJ985870 has seven predicted exons (Table 5.3). The predicted protein aligns well with the reference sequences, sharing the most significant similarity with N*50301. The gene has an identifiable signal peptide, α1, α2 and α3 domains, a TM domain and a weak match to the CP domain. The TM domain contains a non-classical IPI amino acid motif. The CP domain is truncated in the seventh exon (length of 15 bp compared to 45 bp seen in FJ985864rc_C1a).

The FJ985875rc_C1a gene has seven exons (Table 5.3). The predicted protein aligns well with the reference sequences, sharing the most sequence similarity with N*50101. The gene has an identifiable signal peptide, α1, α2 and α3 domains, a TM domain and a CP domain. The TM domain contains a non-classical VPI amino acid motif. The CP domain is truncated in the seventh exon (length of 30 bp compared to 45 bp seen in FJ985864rc_C1a).

FJ985875rc_C1b gene has six exons, with shorter than expected exons 3 and 4 (Table 5.3). An atypical amino acid motif is seen at the end of exon 2 and beginning of exon 3. There are matches in exons 2 and 3 with the α1 and α2 domains, and the shortened exon 4 shows a match to the α3 domain. The TM domain (exon 5) contains a classical FLT amino acid motif. Exon 6 is elongated to extend through the expected exon 7 in a continuous open reading frame, with the last seven amino acids aligning as expected with exon 7 from other predictions, bar a few atypical amino acids also observed in the FJ985859rc_C1a gene.

The FJ985859rc_C1a gene has seven exons (Table 5.3). The predicted protein aligns well with the reference sequences, sharing the most sequence

similarity with N*006001. The gene has an identifiable signal peptide, $\alpha 1$, $\alpha 2$ and $\alpha 3$ domains, a TM domain and a weak match to the CP domain. The TM domain contains a classical FLT motif. The CP domain is truncated in the seventh exon (length of 36 bp compared to 45 bp seen in FJ985864rc_C1a) and has several atypical amino acids in this region.

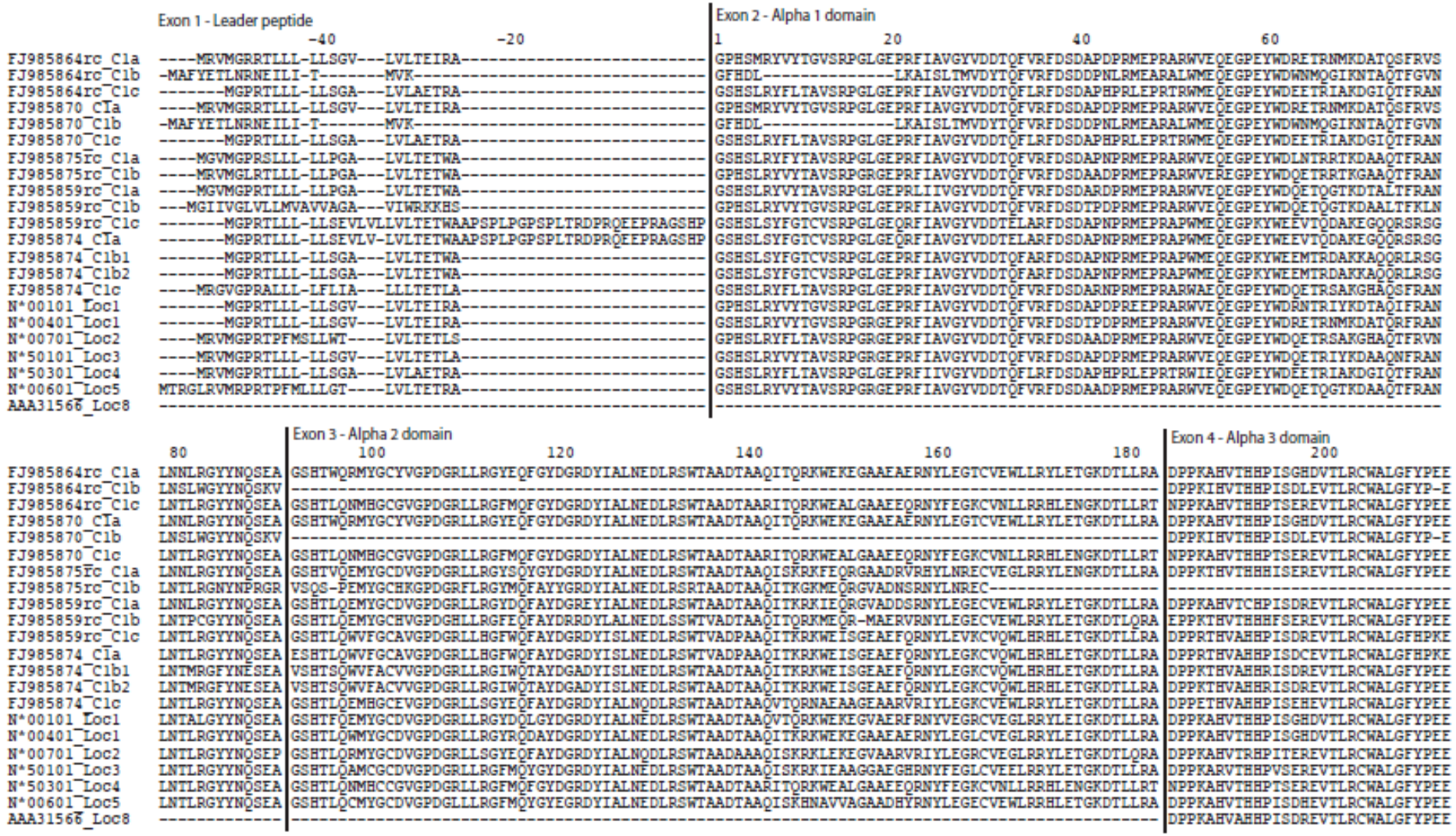
The FJ985859rc_C1b gene has seven exons (Table 5.3). The predicted protein has an atypical exon 1 which is identified as a signal peptide, however the predicted cleavage site is at position 18, 8 amino acids before the start of exon 2. Exons 2 to 7 of the predicted protein align well with the reference sequences, sharing the most sequence similarity with N*006001. The gene has identifiable $\alpha 1$, $\alpha 2$ and $\alpha 3$ domains, a TM domain and a CP domain. The TM domain contains a classical FLT amino acid motif. The CP domain is truncated in the seventh exon (length of 30 bp compared to 45 bp seen in FJ985864rc_C1a) and is the same length as the CP domain found in FJ985875rc_C1a.

The FJ985859rc_C1c gene has five exons (Table 5.3). The gene has an identifiable signal peptide in exon 1. Exon 2 contains an insertion of 27 amino acids at the N-terminal end. Following this, the sequence aligns well through the remainder of exon 2 and exons 3, 4 and 5, with identifiable matches to $\alpha 1$, $\alpha 2$ and $\alpha 3$ domains, as well as a predicted TM domain in exon 5. There is a non-classical IPI amino acid motif in exon 5 and the gene is truncated toward the end of this exon. Genes FJ985859rc_C1c and FJ985874_C1a are identical except for a single codon deletion in the first exon.

Two alternative transcripts are predicted for the FJ985874_C1b gene, differing in the CP region (Table 5.3). FJ985874_C1b1 and FJ985874_C1b2 have seven and eight exons respectively. The FJ985874_C1b2 predicted protein has sequence identity with the reference sequences, sharing the most sequence similarity with N*50101. The gene has an identifiable signal peptide, $\alpha 1$, $\alpha 2$ and $\alpha 3$ domains, a TM domain and a CP domain. There is a non-classical IPI amino acid motif in exon 5. The gene has a full-length CP domain. The FJ985874_C1b1 predicted protein differs from the C1b2 transcript in that it has an extended seventh exon resulting in a long CP

domain and no eighth exon, representing a possible alternative transcript. The CP domain for this transcript is highly similar to that of a partial transcript isolated by Grossberger *et al*. (1990).

The FJ985874_C1c gene contains eight exons showing sequence identity with the reference sequences (Table 5.3). The TM domain contains an atypical FLI amino acid motif. The first few amino acids of exon 6 are missing due to a deletion in the genomic sequence at the 5' end of the exon. Similarly, there is a deletion at the splice site for exon 7; the predicted exon begins 3' to this after the next 'AG' splice signal. The seventh exon is longer than expected, although there is a 'GT' splice signal at the location of the expected termination point determined through genomic sequence alignment with FJ985874_C1b and FJ985864rc_C1a. There is also a SNP altering the AG splice signal before the eighth exon. These results suggest that FJ985874_C1c is a pseudogene. Alignment of the predicted MHC class I-like protein sequences derived from BAC sequences together with the seven MHC-IPD reference sequences is shown in Figure 5.1.


Exon 1 - Leader peptide
Exon 2 - Alpha 1 domain
Exon 3 - Alpha 2 domain
Exon 4 - Alpha 3 domain
FJ985864rc_C1a
FJ985864rc_C1b
FJ985864rc_C1c
FJ985870_C1a
FJ985870_C1b
FJ985870_C1c
FJ985875rc_C1a
FJ985875rc_C1b
FJ985859rc_C1a
FJ985859rc_C1b
FJ985859rc_C1c
FJ985874_C1a
FJ985874_C1b1
FJ985874_C1b2
FJ985874_C1c
N*00101_Loc1
N*00401_Loc1
N*00701_Loc2
N*50101_Loc3
N*50301_Loc4
N*00601_Loc5
AAA31566_Loc8

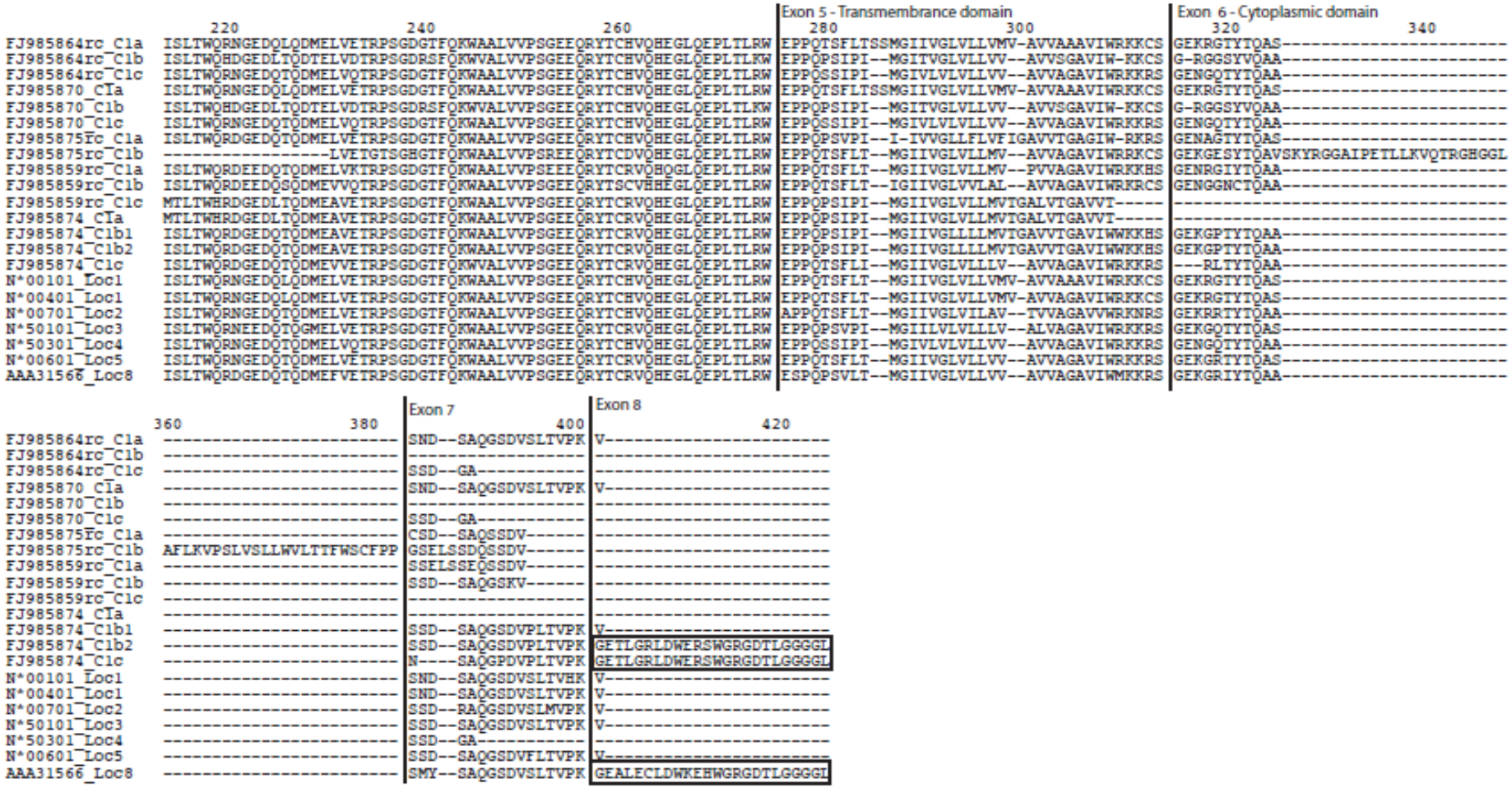


Figure 5.1: Alignment of predicted amino acid sequence of MHC CI genes identified in BAC sequences and selected MHC-IPD reference sequences.

5.3.3 Characteristics of selected reference sequences

Of the ten unique MHC class I sequences identified in the BAC sequences, five were chosen for subsequent multiple and phylogenetic sequence analysis. The genes selected were FJ985864rc_C1a (FJ985870_C1a), FJ985864rc_C1c (FJ985870_C1c), FJ985875rc_C1a, FJ985859rc_C1a and FJ985874_C1b2. Criteria used for selection were that the predicted sequence represented a full length gene (containing an identified initial and terminal exon), the expected MHC class I domains were present in the predicted amino acid sequence, and that the predicted amino acid sequence manifested a high degree of sequence similarity with known translated MHC class I reference mRNA sequences (shown in Table 5.4). The five sequences selected have strong amino acid similarity to previously reported class Ia (classical) and class Ib (non-classical) loci for all domains - signal (leader) peptide, $\alpha 1$ (exon 2), $\alpha 2$ (exon 3), $\alpha 3$ (exon 4), transmembrane (TM) domain (exon 5) and cytoplasmic (CP) domain (exons 6 and 7 for class Ia and Ib genes, and include an eighth exon for class Ia genes).

Table 5.4: Accession numbers and locus information of MHC class 1 sequences from previously published sheep studies.

mRNA Acc	Bp	Protein Acc	AA	IPD Loc	Locus/ID	Source	Breed
AJ874673	1104	CAI43966	361	N*00301	1 / D5	Miltiadou *et al*. 2005	Scottish Blackface
AJ874674	1098	CAI43967	359	N*00101	1 / D3	Miltiadou *et al*. 2005	Scottish Blackface
AJ874675	1123	CAI43968	361	N*00701	2 / C7	Miltiadou *et al*. 2005	Scottish Blackface
AJ874676	1113	CAI43969	357	N*00801	5 / E2	Miltiadou *et al*. 2005	Scottish Blackface
AJ874677	1123	CAI43970	361	N*50101	3 / C6	Miltiadou *et al*. 2005	Scottish Blackface
AJ874678	1123	CAI43971	361	N*50201	3 / C11	Miltiadou *et al*. 2005	Scottish Blackface
AJ874679	1098	CAI43972	359	N*00401	1 / F8	Miltiadou *et al*. 2005	Scottish Blackface
AJ874680	1098	CAI43973	359	N*00501	1 / F12	Miltiadou *et al*. 2005	Scottish Blackface
AJ874681	1135	CAI43974	365	N*00601	5 / H10	Miltiadou *et al*. 2005	Scottish Blackface
AJ874682	1123	CAI43975	361	N*50001	3 / F9	Miltiadou *et al*. 2005	Scottish Blackface
AJ874683	1123	CAI43976	349	N*50301	4 / G6	Miltiadou *et al*. 2005	Scottish Blackface
AJ874684**	283	CAI43977	94		6 / E9	Miltiadou *et al*. 2005	Scottish Blackface
M34676	1396	AAA31568	368		1	Grossberger *et al*. 1990	Unknown
U03092	1042	AAA03455	346		8	Garber *et al*. unpublished	Unknown
U03093	1042	AAA03456	334		4	Garber *et al*. unpublished	Unknown
U03094	1051	AAA03457	349		5	Garber *et al*. unpublished	Unknown
EF489537	1104	ABP37900	368	N*00901		Wu *et al*. unpublished	Unknown
EF489539	1107	ABP37902	368			Wu *et al*. unpublished	Unknown
EF489538	1101	ABP37901	366	N*01001		Wu *et al*. unpublished	Unknown
AM181175	1119	CAJ57269	366	N*00201		Ballingall. unpublished	Scottish Blackface
GQ150751	1107	ACS66687+	368	N*01101		Wu *et al*. unpublished	Finnish Landrace Dorset
NM_001130934	1107	NP_001124406+	368			Wu *et al*. 2008	Soay
M34673*	841	AAA31565	180		7	Grossberger *et al*. 1990	Unknown
AY188824*	468	AAO92004	155		7	Holmes *et al*. 2003	Unknown
M34675*	1050	AAA31567	284		2	Grossberger *et al*. 1990	Unknown
AY188825*	717	AAO92005	238			Holmes *et al*. 2003	Unknown
AY188826*	717	AAO92006	238			Holmes *et al*. 2003	Unknown
M34674	995	AAA31566	178*		8	Grossberger *et al*. 1990	Unknown

*C-terminal region exons 4-8 only. **Partial sequence part of exons 2 and 3. +Identical protein sequences, NP_001124406 was not included in the phylogenetic analysis.

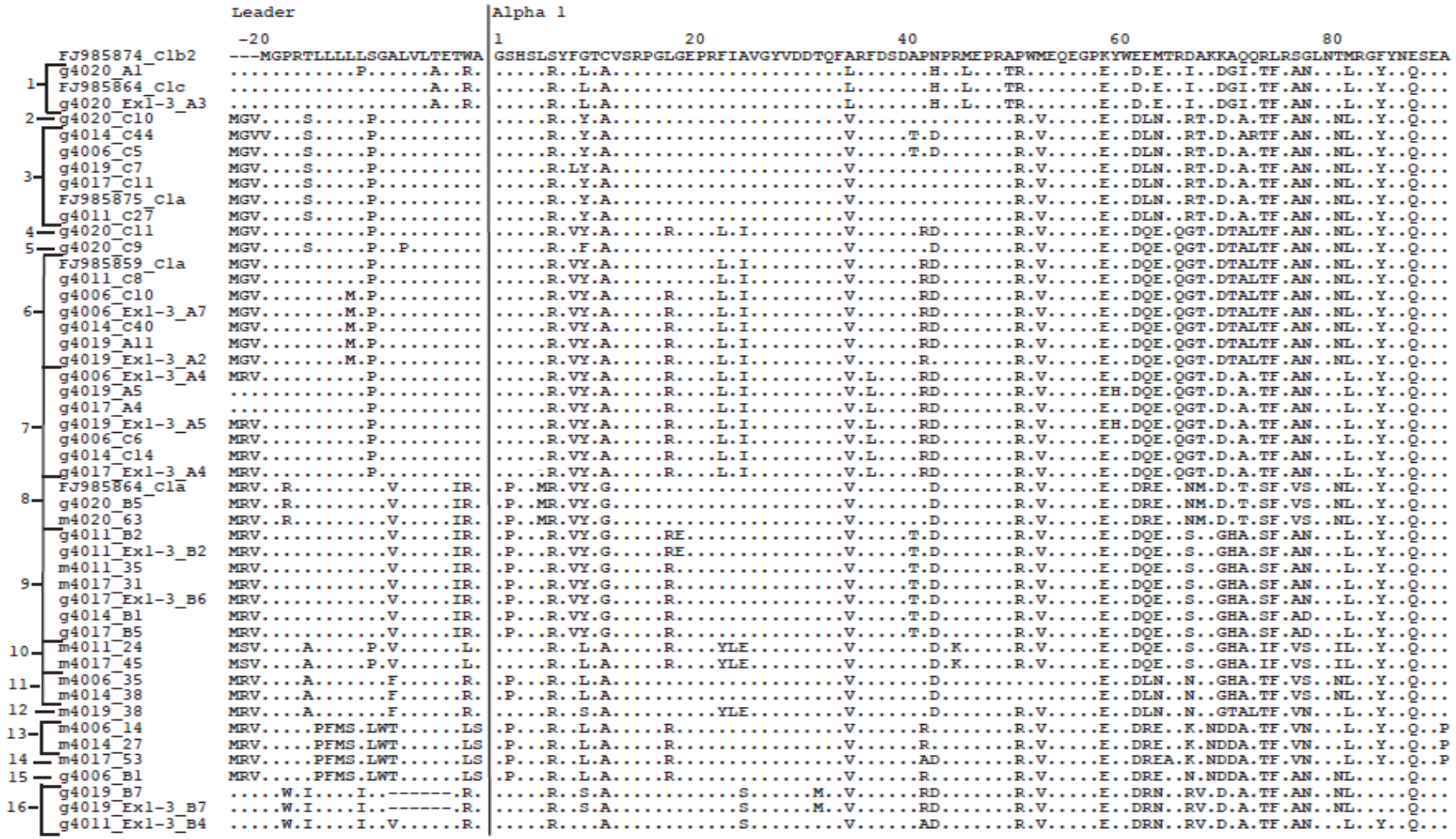
5.3.4 Grouping of MHC class I sequences based upon MHC - Immuno Polymorphism Database (IPD) classification

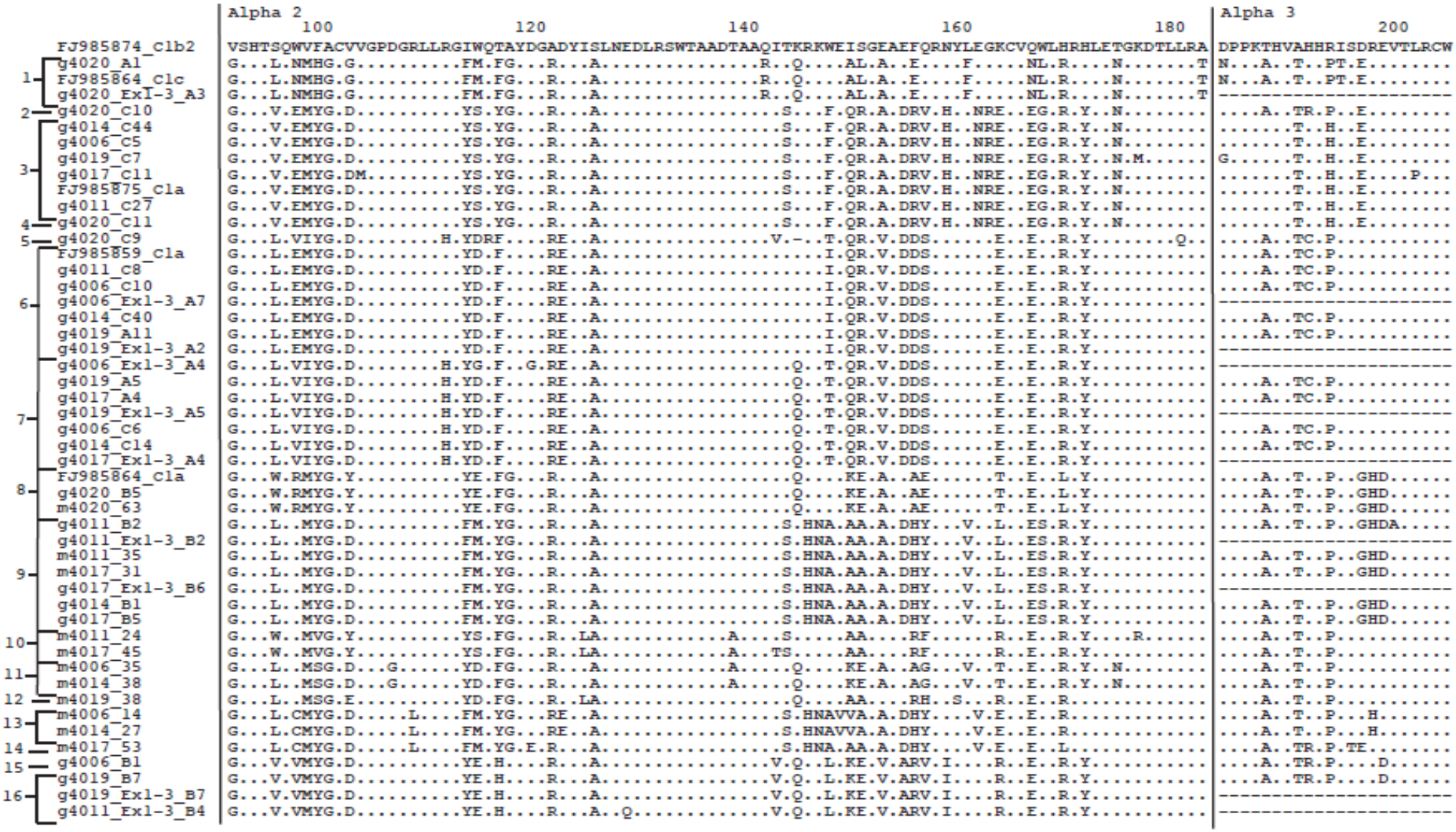
5.3.4.1 Sequences from BAC and homozygous animals

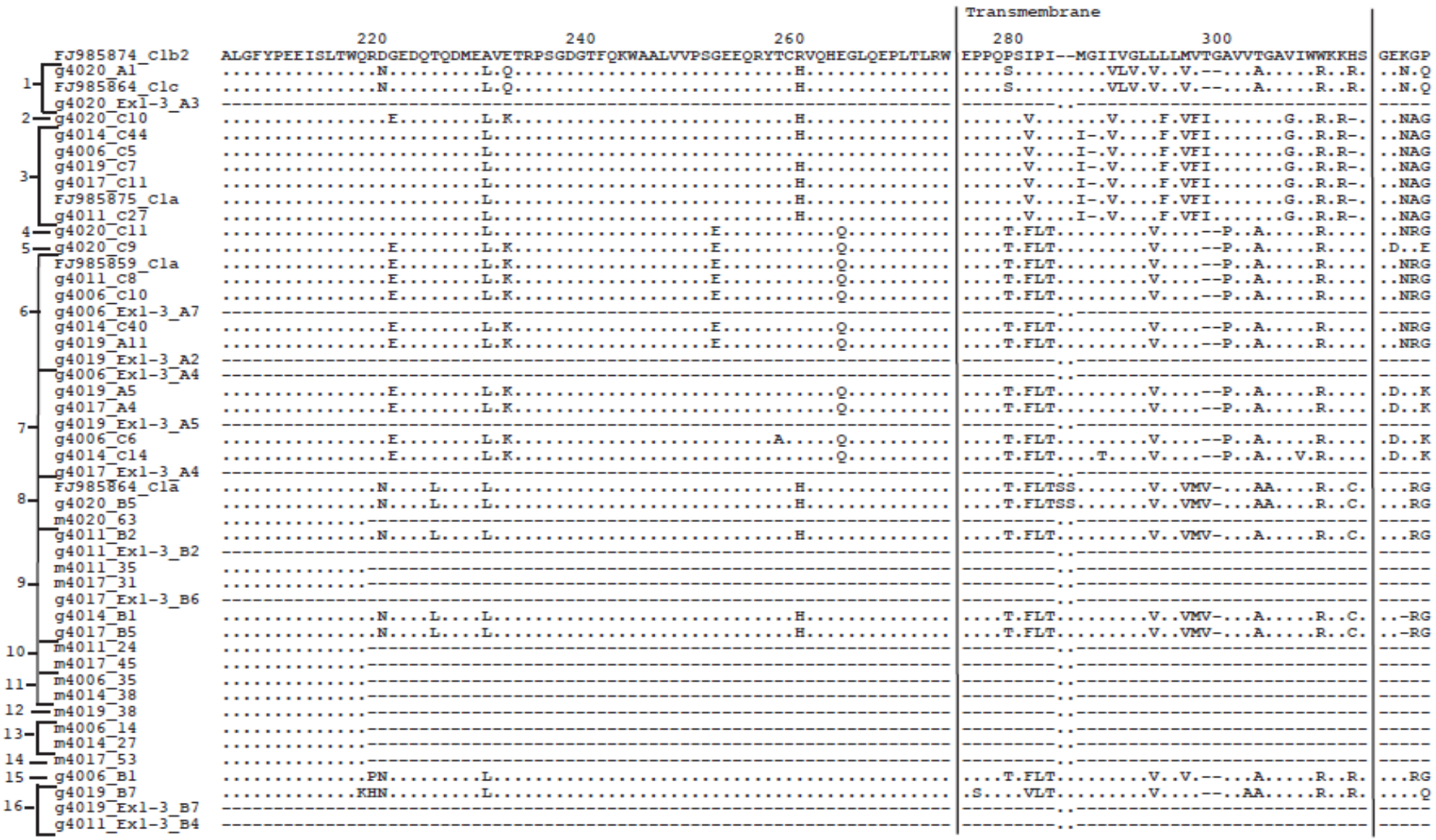
In order to derive information from the large number of sequences available for comparison, a preliminary multiple sequence alignment using predicted amino acid sequences from the homozygous animals and the reference sequences identified from BACs was performed. A total of sixteen groups were identified among the 6 homozygous sheep. Figure 5.2 shows the multiple amino acid sequence alignment of MHC class I. The phylogenetic tree representing the alignment is shown in Figure 5.3. Eight of the sixteen groups contained sequences from more than one animal and the remaining groups represented in only one animal. For sequences extending to the end of exon 3, the number of differences found in the signal peptide, $\alpha 1$ and $\alpha 2$ domains was considered. The number of unique groups identified in each animal varies from five to six (Table 5.5).

Classification of MHC class I sequences based on MHC-IPD criteria showed that groups 1 and 8 are represented by genomic and mRNA sequences from one animal (4020) and include the class I genes FJ985864_C1c and FJ985864_C1a, respectively. Group 3 contains representatives from five animals. Groups 6 and 7 have representatives from four animals. Three animals (4006, 4014 and 4019) have sequences belonging to both groups, and since these animals are homozygous, the groups must represent different but similar loci rather than allelic variants at the same locus. A genomic sequence from animal 20 (g4020_C9), appears to be an intermediate and represents Group 5. Group 9 has representatives from three animals, with mRNA sequences from two of the three animals. It is likely that Group 9 represents an independent locus as it is well separated from other clades. Group 16 has representatives from two animals (4011 and 4019). Groups 2, 4, 5 and 15 are each represented by a single predicted protein derived from genomic sequence. Groups 12 and 14 are represented by a single translated mRNA sequence. All predicted proteins from Groups 1, 3, 6, 7 and 16 are derived from genomic sequences, with no

mRNA representatives. All predicted amino acid sequences from Groups 10, 11, 12, 13 and 14 are derived from mRNA.


Leader
Alpha 1
-20
1
20
40
60
80
FJ985874_C1b2 ---MGPRTLLLLLSGALVLTETWA GSHSLSYFGTCVSRPGLGEPRFIAVGYVDDTQFARFDSDAPNPRMEPRAPWMEQEGPKYWEEMTRDAKKAQQRLRSGLNTMRGFYNESEA
g4020_A1 .....P.....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..Q...
1 FJ985864_C1c .....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..Q...
g4020_Ex1-3_A3 .....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..Q...
2 g4020_C10 MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
g4014_C44 MGVV...S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.ARTF.AN..NL..Y..Q...
g4006_C5 MGV....S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
3 g4019_C7 MGV....S.....P..... .....R.LY.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
g4017_C11 MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
FJ985875_C1a MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
g4011_C27 MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q...
4 g4020_C11 MGV.........P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
5 g4020_C9 MGV....S.....P..P..... .....R..F.A.....V.....D.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
FJ985859_C1a MGV.........P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4011_C8 MGV.........P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4006_C10 MGV........M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
6 g4006_Ex1-3_A7 MGV........M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4014_C40 MGV........M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4019_A11 MGV........M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4019_Ex1-3_A2 MGV........M.P..... .....R.VY.A.....R....L.I.....V.....R.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q...
g4006_Ex1-3_A4 MRV.........P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q...
g4019_A5 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....EH.DQE.QGT.D.A.TF.AN...L..Y..Q...
g4017_A4 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q...
7 g4019_Ex1-3_A5 MRV.........P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....EH.DQE.QGT.D.A.TF.AN...L..Y..Q...
g4006_C6 MRV.........P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q...
g4014_C14 MRV.........P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q...
g4017_Ex1-3_A4 MRV.........P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q...
FJ985864_C1a MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..Q...
8 g4020_B5 MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..Q...
m4020_63 MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..Q...
g4011_B2 MRV.........V.....IR. .P...R.VY.G.....RE.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q...
g4011_Ex1-3_B2 MRV.........V.....IR. .P...R.VY.G.....RE.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q...
m4011_35 MRV.........V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q...
9 m4017_31 MRV.........V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q...
g4017_Ex1-3_B6 MRV.........V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q...
g4014_B1 MRV.........V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..Q...
g4017_B5 MRV.........V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..Q...
10 m4011_24 MSV....A.....P.V.....L. .....R..L.A.....R....YLE.....V.....D.K.....R.V.....E..DQE..S..GHA.IF.VS..IL..Y..Q...
m4017_45 MSV....A.....P.V.....L. .....R..L.A.....R....YLE.....V.....D.K.....R.V.....E..DQE..S..GHA.IF.VS..IL..Y..Q...
11 m4006_35 MRV....A.....F.....R. .P...R..L.A.....V.....D.....E..DLN..N..GHA.TF.VS..NL..Y..Q...
m4014_38 MRV....A.....F.....R. .P...R..L.A.....V.....D.....E..DLN..N..GHA.TF.VS..NL..Y..Q...
12 m4019_38 MRV....A.....F.....R. .....R..S.A.....YLE.....V.....D.....R.V.....E..DLN..N..GTALTF.VN...L..Y..Q...
13 m4006_14 MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..Q..P
m4014_27 MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..Q..P
14 m4017_53 MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DREA.K.NDDA.TF.VN...L..Y..Q..P
15 g4006_B1 MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..N.NDDA.TF.AN..NL.....Q...
g4019_B7 .....W.I....I.-------.R. .....R..S.A.....S.....M..V.....RD.....R.V.....E..DRN..RV.D.A.TF.AN..NL.....Q...
16 g4019_Ex1-3_B7 .....W.I....I..------.R. .....R..S.A.....S.....M..V.....RD.....R.V.....E..DRN..RV.D.A.TF.AN..NL.....Q...
g4011_Ex1-3_B4 .....W.I....I..V.....R. .....R....A.....S.....V.....AD.....R.V.....E..DRN..RV.D.A.TF.AN...L..Y..Q...


Alpha 2
100
120
140
160
180
Alpha 3
200
FJ985874_C1b2
g4020_A1
FJ985864_C1c
g4020_Ex1-3_A3
g4020_C10
g4014_C44
g4006_C5
g4019_C7
g4017_C11
FJ985875_C1a
g4011_C27
g4020_C11
g4020_C9
FJ985859_C1a
g4011_C8
g4006_C10
g4006_Ex1-3_A7
g4014_C40
g4019_A11
g4019_Ex1-3_A2
g4006_Ex1-3_A4
g4019_A5
g4017_A4
g4019_Ex1-3_A5
g4006_C6
g4014_C14
g4017_Ex1-3_A4
FJ985864_C1a
g4020_B5
m4020_63
g4011_B2
g4011_Ex1-3_B2
m4011_35
m4017_31
g4017_Ex1-3_B6
g4014_B1
g4017_B5
m4011_24
m4017_45
m4006_35
m4014_38
m4019_38
m4006_14
m4014_27
m4017_53
g4006_B1
g4019_B7
g4019_Ex1-3_B7
g4011_Ex1-3_B4


Transmembrane
220
240
260
280
300
FJ985874_C1b2 ALGFYPEEISLTWQRDGEDQTQDMEAVETRPSGDGTFQKWAALVVPSGEEQRYTCRVQHEGLQEPLTLRW EPPQPSIPI--MGIIVGLLLLMVTGAVVTGAVIWWKKHS GEKGP
g4020_A1 ..............N.........L.Q.....................H.............. ....S.........VLV.V..V.--...A.....R..R. ..N.Q
FJ985864_C1c ..............N.........L.Q.....................H.............. ....S.........VLV.V..V.--...A.....R..R. ..N.Q
g4020_Ex1-3_A3 ---------- ---------..---------- -----
g4020_C10 ..............E........L.K.....................H.............. ......V.......V....F.VFI.......G..R.R-. ..NAG
g4014_C44 ..............................L.....................H.............. ......V....I-.V....F.VFI.......G..R.R-. ..NAG
g4006_C5 ....L..................................... ......V....I-.V....F.VFI.......G..R.R-. ..NAG
g4019_C7 ..............................L.....................H.............. ......V....I-.V....F.VFI.......G..R.R-. ..NAG
g4017_C11 ..............................L.....................H.............. ......V....I-.V....F.VFI.......G..R.R-. ..NAG
FJ985875_C1a ..............................L.....................H.............. ......V....I-.V....F.VFI.......G..R.R-. ..NAG
g4011_C27 ..............................L.....................H.............. ......V....I-.V....F.VFI.......G..R.R-. ..NAG
g4020_C11 ..............................L.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4020_C9 ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... .D..E
FJ985859_C1a ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4011_C8 ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4006_C10 ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4006_Ex1-3_A7 ---------- ---------..---------- -----
g4014_C40 ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4019_A11 ..............E........L.K.....................E..........Q..... ....T.FLT.........V.....--P..A.....R..... ..NRG
g4019_Ex1-3_A2 ---------- ---------..---------- -----
g4006_Ex1-3_A4 ---------- ---------..---------- -----
g4019_A5 ..............E........L.K.....................................Q..... ....T.FLT.........V.....--P..A.....R..... .D..K
g4017_A4 ..............E........L.K.....................................Q..... ....T.FLT.........V.....--P..A.....R..... .D..K
g4019_Ex1-3_A5 ---------- ---------..---------- -----
g4006_C6 ..............E........L.K.....................A.....Q..... ....T.FLT.........V.....--P..A.....R..... .D..K
g4014_C14 ..............E........L.K.....................................Q..... ....T.FLT....T....V.....--P..A...V.R..... .D..K
g4017_Ex1-3_A4 ---------- ---------..---------- -----
FJ985864_C1a ..............N....L...L.....................H.............. ....T.FLTSS.......V..VMV-...AA....R..C. ...RG
g4020_B5 ..............N....L...L.....................H.............. ....T.FLTSS.......V..VMV-...AA....R..C. ...RG
m4020_63 ....---------- ---------..---------- -----
g4011_B2 ..............N....L...L.....................H.............. ....T.FLT.........V..VMV-...A.....R..C. ...RG
g4011_Ex1-3_B2 ---------- ---------..---------- -----
m4011_35 ....---------- ---------..---------- -----
m4017_31 ....---------- ---------..---------- -----
g4017_Ex1-3_B6 ---------- ---------..---------- -----
g4014_B1 ..............N....L...L.....................H.............. ....T.FLT.........V..VMV-...A.....R..C. ..-RG
g4017_B5 ..............N....L...L.....................H.............. ....T.FLT.........V..VMV-...A.....R..C. ..-RG
m4011_24 ....---------- ---------..---------- -----
m4017_45 ....---------- ---------..---------- -----
m4006_35 ....---------- ---------..---------- -----
m4014_38 ....---------- ---------..---------- -----
m4019_38 ....---------- ---------..---------- -----
m4006_14 ....---------- ---------..---------- -----
m4014_27 ....---------- ---------..---------- -----
m4017_53 ....---------- ---------..---------- -----
g4006_B1 ..............PN.........L..................................... ....T.FLT.........V..V.--...A.....R..R. ...RG
g4019_B7 ............KHN.........L..................................... .S....VLT.........V.....--..AA.....R..R. ....Q
g4019_Ex1-3_B7 ---------- ---------..---------- -----
g4011_Ex1-3_B4 ---------- ---------..---------- -----

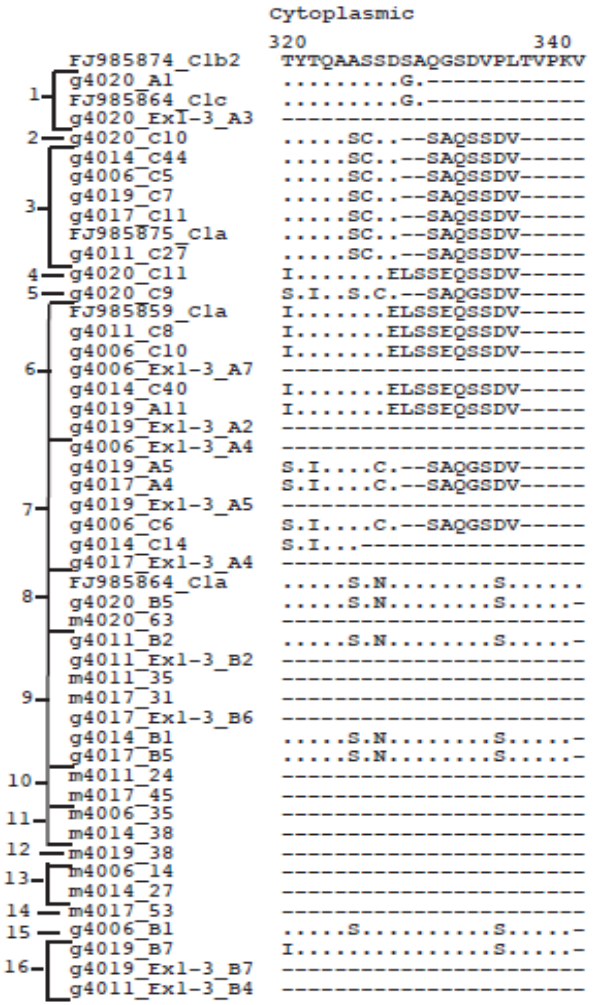


Figure 5.2: Alignment of translated genomic and mRNA sequences from each of the homozygous animals.

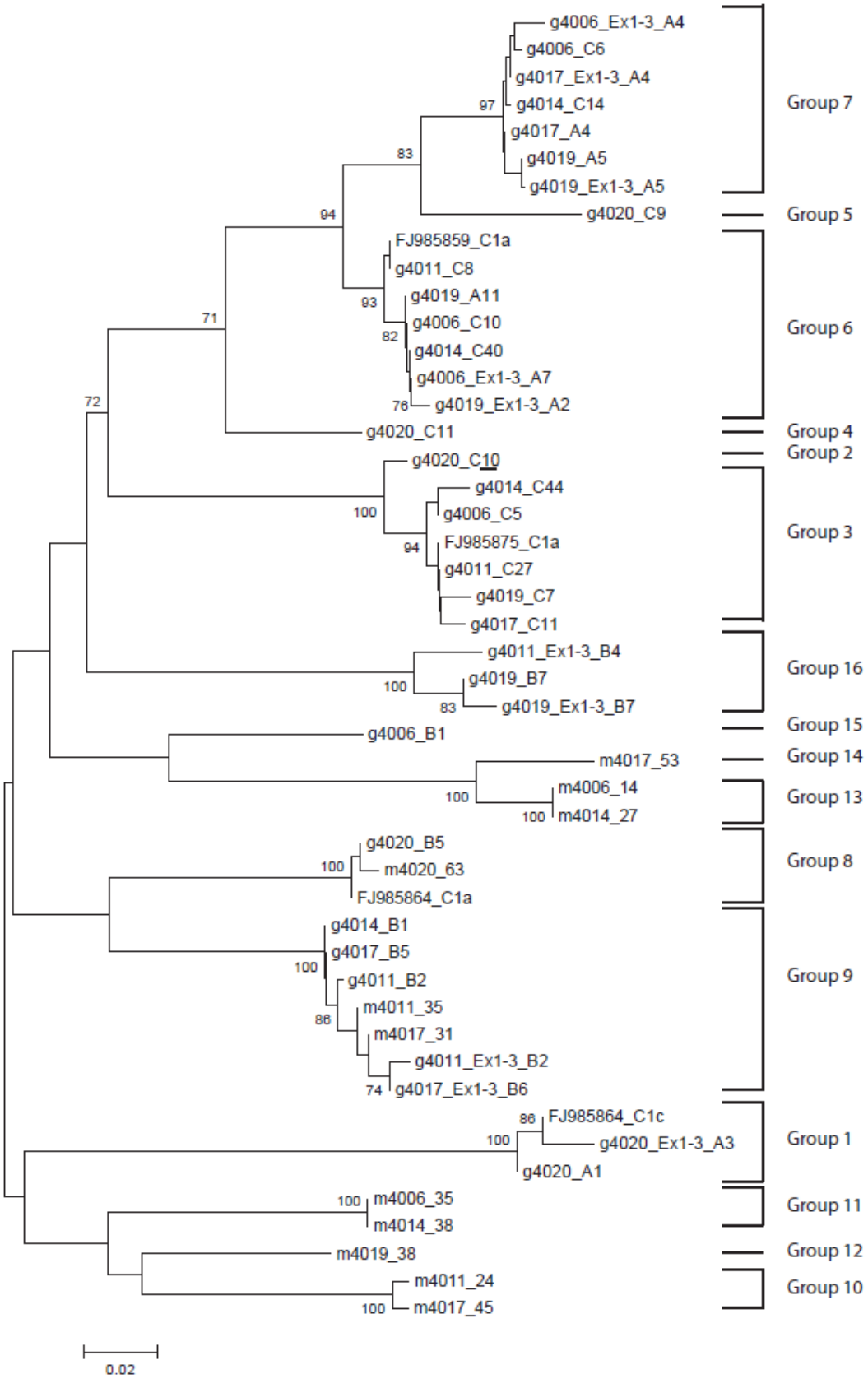


Figure 5.3: Phylogenetic tree of sequences from each of the homozygous sheep showing correlations with group assignment.

Table 5.5: Number of genomic and mRNA sequences isolated from each of the homozygous animals, as well as the number of MHC-IPD groups classified for both genomic and mRNA sequences.

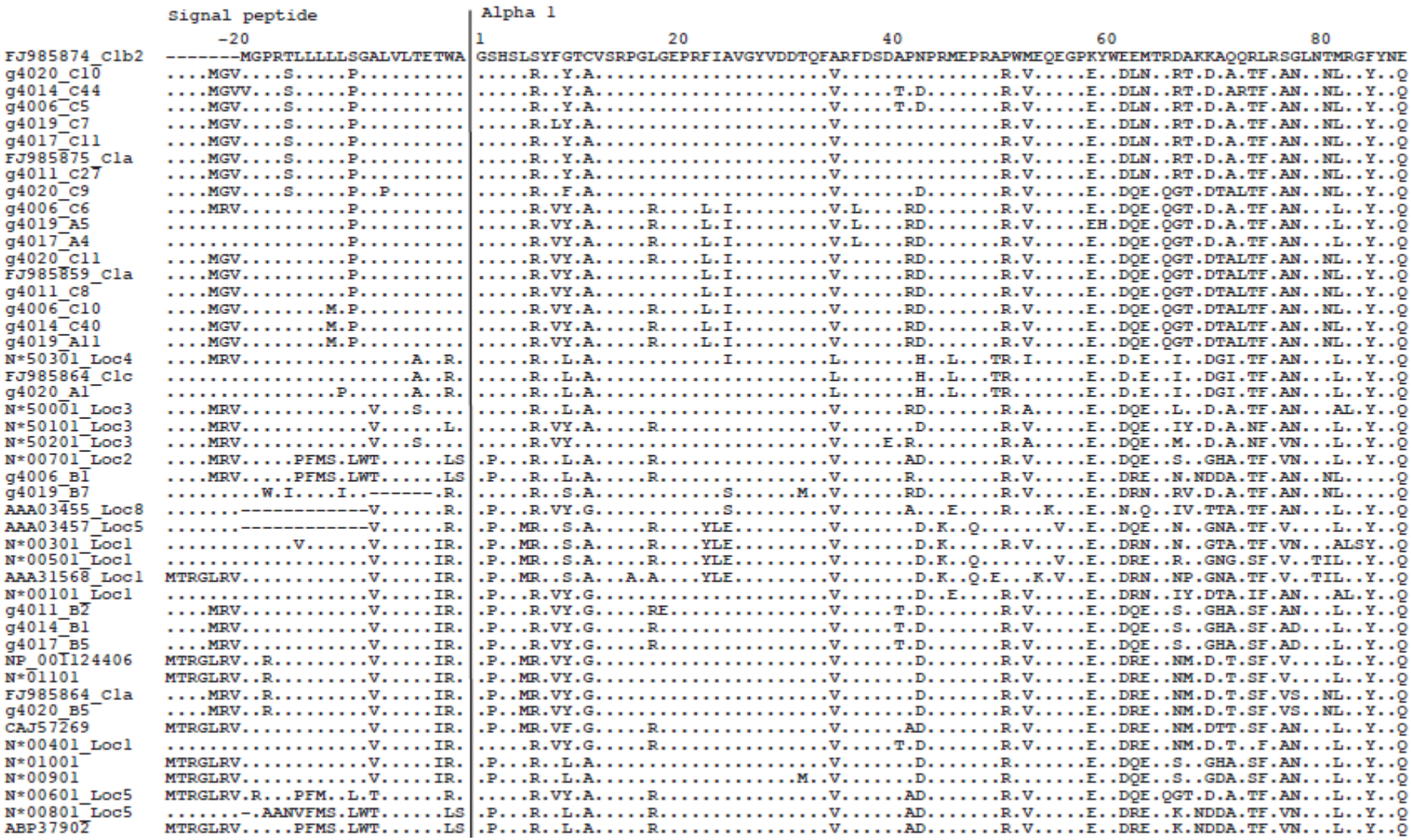
Animal	Gen fl	Gen ex 1-3	# gen groups	mRNA ex1-3	# mRNA groups	Total groups
4006	5	8	4	14	2	6
4011	6	8	4	13	2	5
4014	5	0	4	12	2	6
4017	4	10	3	13	3	5
4019	4	9	4	13	1	5
4020	6	10	5	12	1	5

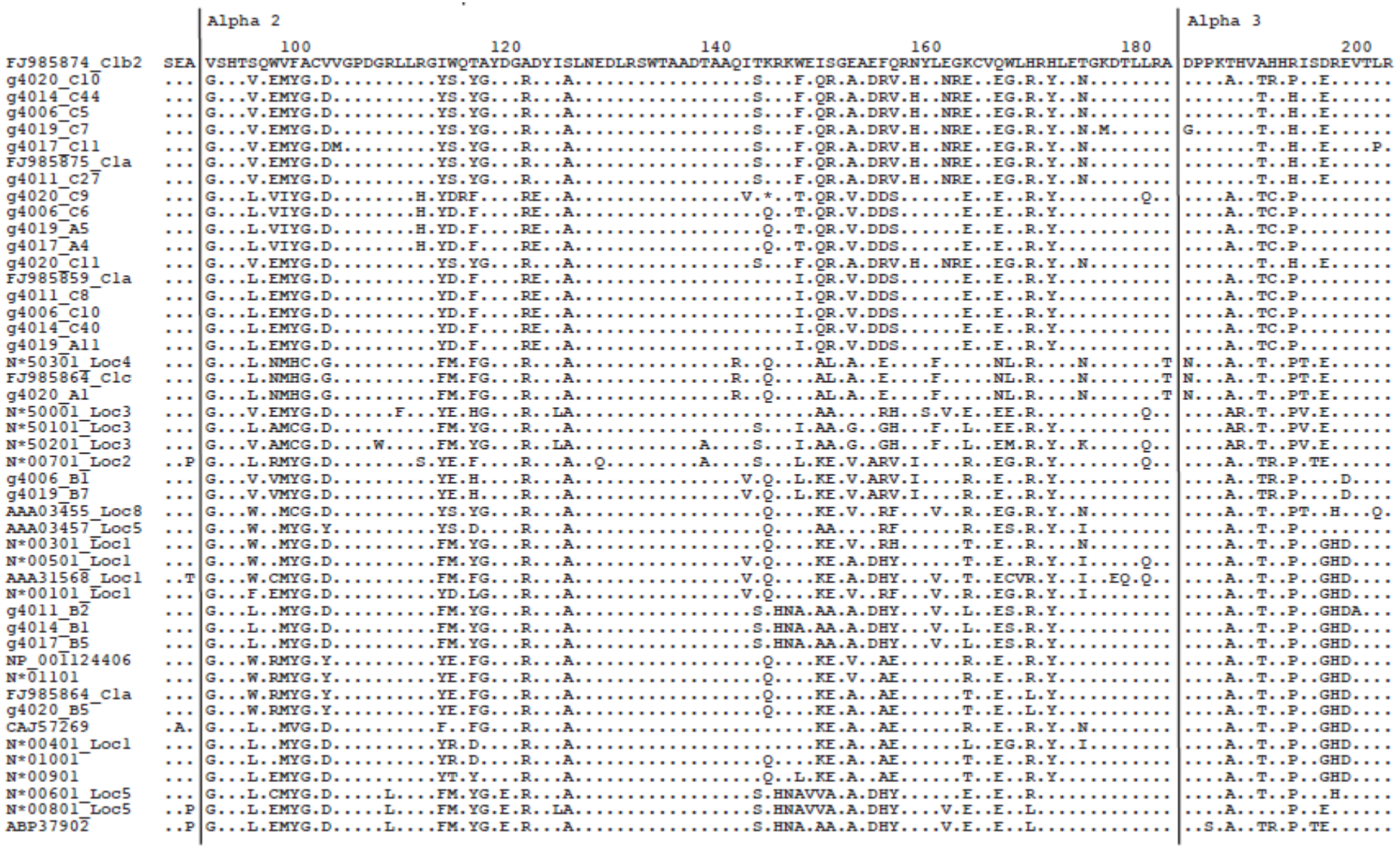
Gen fl: genomic full-length sequences. Gen ex 1-3: genomic sequences covering exons 1-3. # gen groups: number of genomic groups identified according to MHC-IPD classification. mRNA ex 1-3: messenger RNA covering exons 1-3. # mRNA groups: number of mRNA MHC-IPD groups identified. Total groups: total number of MHC-IPD groups classified combining genomic and mRNA groups (refer Figure 5.3).

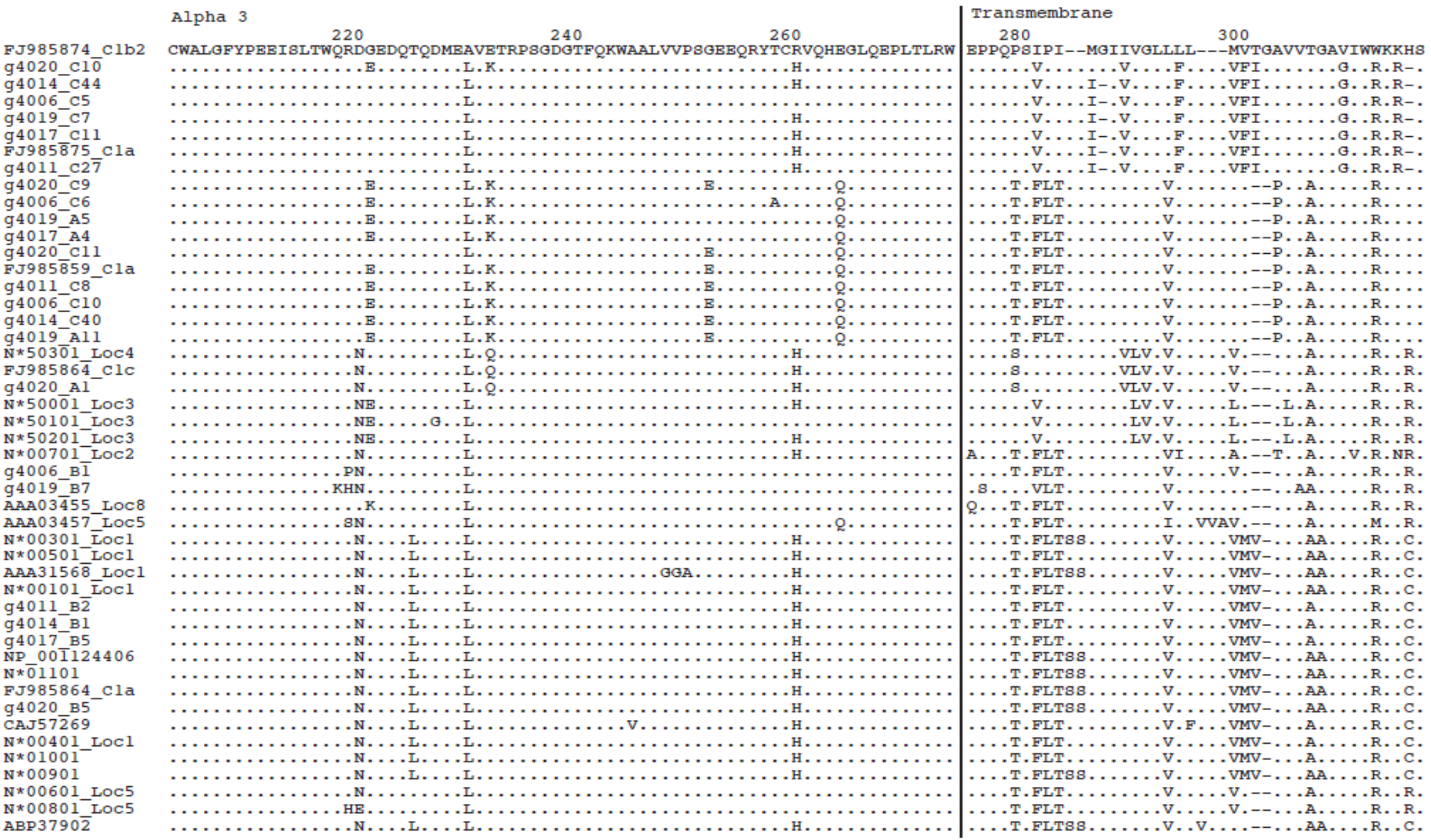
5.3.4.2 Grouping of all full length sequences

This analysis shown above was further extended to include 14 MHC-IPD reference sequences and an additional five sequences downloaded from the NCBI database representing a separate group based on the MHC-IPD criteria. Of these, only reference sequences N*50301 and N*01101 could be classified in the same MHC-IPD-defined group with sequences isolated in this study. The predicted amino acid sequences from the full-length genomic sequences isolated in this study comprised a signal peptide and α1, α2, α3, TM and CP domains (Figure 5.4). All the predicted genes contained 7 exons except for sequence g4014_C14, which has missing exon 7 and hence has truncated predicted amino acid sequence (this sequence is not included in Figure 5.4). Genome sequences amplified using primer sets A and C have truncated CP domains, terminating after exon 7. Genome sequences amplified using primer set B have full-length CP domains, except for a single amino acid (valine) encoded by the eighth exon. The reverse primers were designed immediately outside of exon 8 but failed to amplify the short exon. The signal peptides varied in length due to variation in the N-terminal region, but for all sequences, cleavage sites were identified immediately

before exon 2 as expected. Considerable variation was observed in the alpha 1, 2 and 3 regions. TM domains also showed variation, ranging in length from 35-38 amino acids, with variation seen in the pattern of indels and substitutions that were consistent among groups. Groups assigned in Figure 5.3 have also been shown in Figure 5.5, which illustrates the TM and CP domains and the classification of the sequences into class Ia or Ib. Sequence g4014_C14 was not included in Figure 5.5 because it has a truncated CP domain. Comparison of group assignment in Figure 5.5 and previous analysis (Figure 5.4 and Figure 5.3) showed that there was similar arrangement of groups, except for Group 2. Group 2 is represented by a single genomic sequence (g4020_C10), which shows the same amino acid sequence pattern as Group 3 in the first three exons, a unique pattern through most of exon 4, and returning to the same pattern as Group 3 from exon five onwards. Therefore, the sequence (g4020_C10) was grouped with Group 3 in Figure 5.5 (which is an alignment of MHC class I sequences representing the TM and CP domains) instead of presenting Group 2 on its own as seen in the Figure 5.4. Groups 10, 11, 12, 13 and 14 in Figure 5.4 and Figure 5.3 consisted of predicted amino acid sequences derived from mRNA, which do not extend until TM and CP domains, and as such are not included in Figure 5.5.


Signal peptide
Alpha 1
-20
1
20
40
60
80
FJ985874_C1b2 -------MGPRTLLLLLSGALVLTETWA GSHSLSYFGTCVSRPGLGEPRFIAVGYVDDTQFARFDSDAPNPRMEPRAPWMEQEGPKYWEEMTRDAKKAQQRLRSGLNTMRGFYNE
g4020_C10 ....MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
g4014_C44 ....MGVV...S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.ARTF.AN..NL..Y..Q
g4006_C5 ....MGV....S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
g4019_C7 ....MGV....S.....P..... .....R.LY.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
g4017_C11 ....MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
FJ985875_C1a ....MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
g4011_C27 ....MGV....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..Q
g4020_C9 ....MGV....S.....P..P..... .....R..F.A.....V.....D.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
g4006_C6 ....MRV.....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q
g4019_A5 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....EH.DQE.QGT.D.A.TF.AN...L..Y..Q
g4017_A4 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q
g4020_C11 ....MGV.....P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
FJ985859_C1a ....MGV.....P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
g4011_C8 ....MGV.....P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
g4006_C10 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
g4014_C40 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
g4019_A11 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..Q
N*50301_Loc4 ....MRV.....A..R. .....R..L.A.....I.....L.....H..L...TR.I.....E..D.E..I..DGI.TF.AN...L..Y..Q
FJ985864_C1c .....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..Q
g4020_A1 .....P.....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..Q
N*50001_Loc3 ....MRV.....V...S.... .....R..L.A.....V.....RD.....R.A.....E..DQE..L..D.A.TF.AN...AL.Y..Q
N*50101_Loc3 ....MRV.....V.....L. .....R.VY.A.....R.....V.....D.....R.V.....E..DQE..IY.D.A.NF.AN...L..Y..Q
N*50201_Loc3 ....MRV.....V...S.... .....R.VY.....V....E.R.....R.A.....E..DQE..M..D.A.NF.VN...L..Y..Q
N*00701_Loc2 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DQE..S..GHA.TF.VN...L..Y..Q
g4006_B1 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..N.NDDA.TF.AN...NL.....Q
g4019_B7 .....W.I....I..------.R. .....R..S.A.....S.....M..V.....RD.....R.V.....E..DRN..RV.D.A.TF.AN...NL.....Q
AAA03455_Loc8 .....----------V.....R. .P...R.VY.G.....S.....V.....A...E....R...K...E..N.Q..IV.TTA.TF.AN...L..Y..Q
AAA03457_Loc5 .....----------V.....R. .P...MR..S.A.....R....YLE.....V.....D.K..Q.....V..E..DQE..N..GNA.TF.V....L..Y..Q
N*00301_Loc1 .....V.....V.....IR. .P...MR..S.A.....R....YLE.....V.....D.K.....R.V.....E..DRN..N..GTA.TF.VN...ALSY..Q
N*00501_Loc1 .....V.....IR. .P...MR..S.A.....R....YLE.....V.....D.K..Q.....V..E..DRE..R..GNG.SF.V..TIL..Y..Q
AAA31568_Loc1 MTRGLRV.....V.....IR. .P...MR..S.A...A.A....YLE.....V.....D.K..Q.E...K.V..E..DRN..NP.GNA.TF.V..TIL..Y..Q
N*00101_Loc1 .....V.....IR. .P...R.VY.G.....V.....D..E....R.V.....E..DRN..IY.DTA.IF.AN...AL.Y..Q
g4011_B2 ....MRV.....V.....IR. .P...R.VY.G.....RE.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..Q
g4014_B1 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..Q
g4017_B5 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..Q
NP_001124406 MTRGLRV..R.....V.....IR. .P...MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.V....L..Y..Q
N*01101 MTRGLRV..R.....V.....IR. .P...MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.V....L..Y..Q
FJ985864_C1a ....MRV..R.....V.....IR. .P...MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..Q
g4020_B5 ....MRV..R.....V.....IR. .P...MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..Q
CAJ57269 MTRGLRV.....V.....IR. .P...MR.VF.G.....R.....V.....AD.....R.V.....E..DRE..NM.DTT.SF.AN...L..Y..Q
N*00401_Loc1 .....V.....IR. .....R.VY.G.....R.....V.....T.D.....R.V.....E..DRE..NM.D.T..F.AN...L..Y..Q
N*01001 MTRGLRV.....V.....IR. .P...R..L.A.....V.....D.....R.....E..DQE..S..GHA.SF.AN...L..Y..Q
N*00901 MTRGLRV.....V.....IR. .P...R..L.A.....M..V.....D.....R.....E..DQE..S..GDA.SF.AN...L..Y..Q
N*00601_Loc5 MTRGLRV.R...PFM..L.T.....R. .....R.VY.A.....R.....V.....AD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..Q
N*00801_Loc5 .....-.AANVFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..Q
ABP37902 MTRGLRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..Q


Alpha 2
Alpha 3
100
120
140
160
180
200
FJ985874_C1b2 SEA VSHTSQWVFACVVGPDGRLLRGIWQTAYDGADYISLNEDLRSWTAADTAAQITKRKWEISGEAEFQRNYLEGKCVQWLHRHLETGKDTLLRA DPPKTHVAHHRISDREVTLR
g4020_C10 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
g4014_C44 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
g4006_C5 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
g4019_C7 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.M.....
g4017_C11 ... G...V.EMYG.DM.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
FJ985875_C1a ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
g4011_C27 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
g4020_C9 ... G...L.VIYG.D.....H.YDRF....RE..A.....V.*..T.QR.V.DDS.....E..E..R.Y.....Q..
g4006_C6 ... G...L.VIYG.D.....H.YD.F....RE..A.....Q..T.QR.V.DDS.....E..E..R.Y.....
g4019_A5 ... G...L.VIYG.D.....H.YD.F....RE..A.....Q..T.QR.V.DDS.....E..E..R.Y.....
g4017_A4 ... G...L.VIYG.D.....H.YD.F....RE..A.....Q..T.QR.V.DDS.....E..E..R.Y.....
g4020_C11 ... G...V.EMYG.D.....YS.YG...R...A.....S...F.QR.A.DRV.H..NRE..EG.R.Y..N.....
FJ985859_C1a ... G...L.EMYG.D.....YD.F....RE..A.....I.QR.V.DDS.....E..E..R.Y.....
g4011_C8 ... G...L.EMYG.D.....YD.F....RE..A.....I.QR.V.DDS.....E..E..R.Y.....
g4006_C10 ... G...L.EMYG.D.....YD.F....RE..A.....I.QR.V.DDS.....E..E..R.Y.....
g4014_C40 ... G...L.EMYG.D.....YD.F....RE..A.....I.QR.V.DDS.....E..E..R.Y.....
g4019_A11 ... G...L.EMYG.D.....YD.F....RE..A.....I.QR.V.DDS.....E..E..R.Y.....
N*50301_Loc4 ... G...L.NMHC.G.....FM.FG...R...A.....R..Q....AL.A..E....F.....NL.R....N.....T N...A..T..PT.E.....
FJ985864_C1c ... G...L.NMHG.G.....FM.FG...R...A.....R..Q....AL.A..E....F.....NL.R....N.....T N...A..T..PT.E.....
g4020_A1 ... G...L.NMHG.G.....FM.FG...R...A.....R..Q....AL.A..E....F.....NL.R....N.....T N...A..T..PT.E.....
N*50001_Loc3 ... G...V.EMYG.D.....F...YE.HG...R..LA.....AA....RH..S.V.E..EE.R.....Q..
N*50101_Loc3 ... G...L.AMCG.D.....FM.YG...R...A.....S...I.AA.G..GH...F..L..EE.R.Y.....
N*50201_Loc3 ... G...V.AMCG.D....W.....FM.YG...R..LA.....A....S...I.AA.G..GH...F..L..EM.R.Y..K.....Q..
N*00701_Loc2 ..P G...L.RMYG.D.....S.YE.F....R...A..Q.....A....S...L.KE.V.ARV.I....R..EG.R.Y.....Q..
g4006_B1 ... G...V.VMYG.D.....YE.H....R...A.....V.Q..L.KE.V.ARV.I....R..E..R.Y.....
g4019_B7 ... G...V.VMYG.D.....YE.H....R...A.....V.Q..L.KE.V.ARV.I....R..E..R.Y.....
AAA03455_Loc8 ... G...W..MCG.D.....YS.YG...R...A.....Q....KE.V..RF...V..R..EG.R.Y..N.....
AAA03457_Loc5 ... G...W..MYG.Y.....YS.D....R...A.....Q....AA....RF.....R..ES.R.Y..I.....
N*00301_Loc1 ... G...W..MYG.D.....FM.YG...R...A.....Q....KE.V..RH.....T..E..R....N.....
N*00501_Loc1 ... G...W..MYG.D.....FM.YG...R...A.....V.Q....KE.A.DHY.....T..E..R.Y..I.....Q..
AAA31568_Loc1 ..T G...W.CMYG.D.....FM.FG...R...A.....V.Q....KE.A.DHY...V..T..ECVR.Y..I..EQ.Q..
N*00101_Loc1 ... G...F.EMYG.D.....YD.LG...R...A.....V.Q....KE.V..RF....V..R..EG.R.Y..I.....
g4011_B2 ... G...L..MYG.D.....FM.YG...R...A.....S.HNA.AA.A.DHY...V..L..ES.R.Y.....
g4014_B1 ... G...L..MYG.D.....FM.YG...R...A.....S.HNA.AA.A.DHY...V..L..ES.R.Y.....
g4017_B5 ... G...L..MYG.D.....FM.YG...R...A.....S.HNA.AA.A.DHY...V..L..ES.R.Y.....
NP_001124406 ... G...W.RMYG.Y.....YE.FG...R...A.....Q....KE.V..AE.....R..E..R.Y.....
N*01101 ... G...W.RMYG.Y.....YE.FG...R...A.....Q....KE.V..AE.....R..E..R.Y.....
FJ985864_C1a ... G...W.RMYG.Y.....YE.FG...R...A.....Q....KE.A..AE.....T..E..L.Y.....
g4020_B5 ... G...W.RMYG.Y.....YE.FG...R...A.....Q....KE.A..AE.....T..E..L.Y.....
CAJ57269 .A. G...L..MVG.D.....F..FG...R...A.....KE.A..AE.....R..E..R.Y..N.....
N*00401_Loc1 ... G...L..MYG.D.....YR.D....R...A.....KE.A..AE.....L..EG.R.Y..I.....
N*01001 ... G...L..MYG.D.....YR.D....R...A.....Q....KE.A..AE.....T..E..R.Y.....
N*00901 ... G...L.EMYG.D.....YT.Y....R...A.....Q..L.KE.A..AE.....T..E..R.Y.....
N*00601_Loc5 ... G...L.CMYG.D.....L....FM.YG.E.R...A.....S.HNAVVA.A.DHY.....E..E..R.....
N*00801_Loc5 ..P G...L.EMYG.D.....L....FM.YG.E.R..LA.....S.HNAVVA.A.DHY....V.E..E..L.....
ABP37902 ..P G...L.EMYG.D.....L....FM.YG.E.R...A.....S.HNA.AA.A.DHY....V.E..E..L.....


Alpha 3
Transmembrane
220
240
260
280
300
FJ985874_C1b2 CWALGFYPEEISLTWQRDGEDQTQDMEAVETRPSGDGTFQKWAALVVPSGEEQRYTCRVQHEGLQEPLTLRW EPPQPSIPI--MGIIVGLLLL---MVTGAVVTGAVIWWKKHS
g4020_C10
g4014_C44
g4006_C5
g4019_C7
g4017_C11
FJ985875_C1a
g4011_C27
g4020_C9
g4006_C6
g4019_A5
g4017_A4
g4020_C11
FJ985859_C1a
g4011_C8
g4006_C10
g4014_C40
g4019_A11
N*50301_Loc4
FJ985864_C1c
g4020_A1
N*50001_Loc3
N*50101_Loc3
N*50201_Loc3
N*00701_Loc2
g4006_B1
g4019_B7
AAA03455_Loc8
AAA03457_Loc5
N*00301_Loc1
N*00501_Loc1
AAA31568_Loc1
N*00101_Loc1
g4011_B2
g4014_B1
g4017_B5
NP_001124406
N*01101
FJ985864_C1a
g4020_B5
CAJ57269
N*00401_Loc1
N*01001
N*00901
N*00601_Loc5
N*00801_Loc5
ABP37902

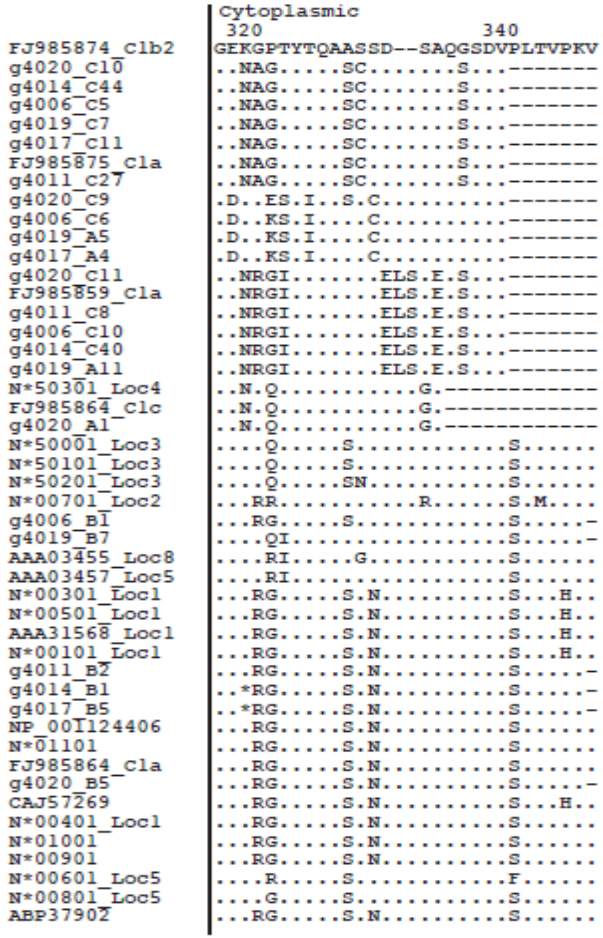


Figure 5.4: The predicted amino acid sequences from the full-length genomic sequences isolated in this study.

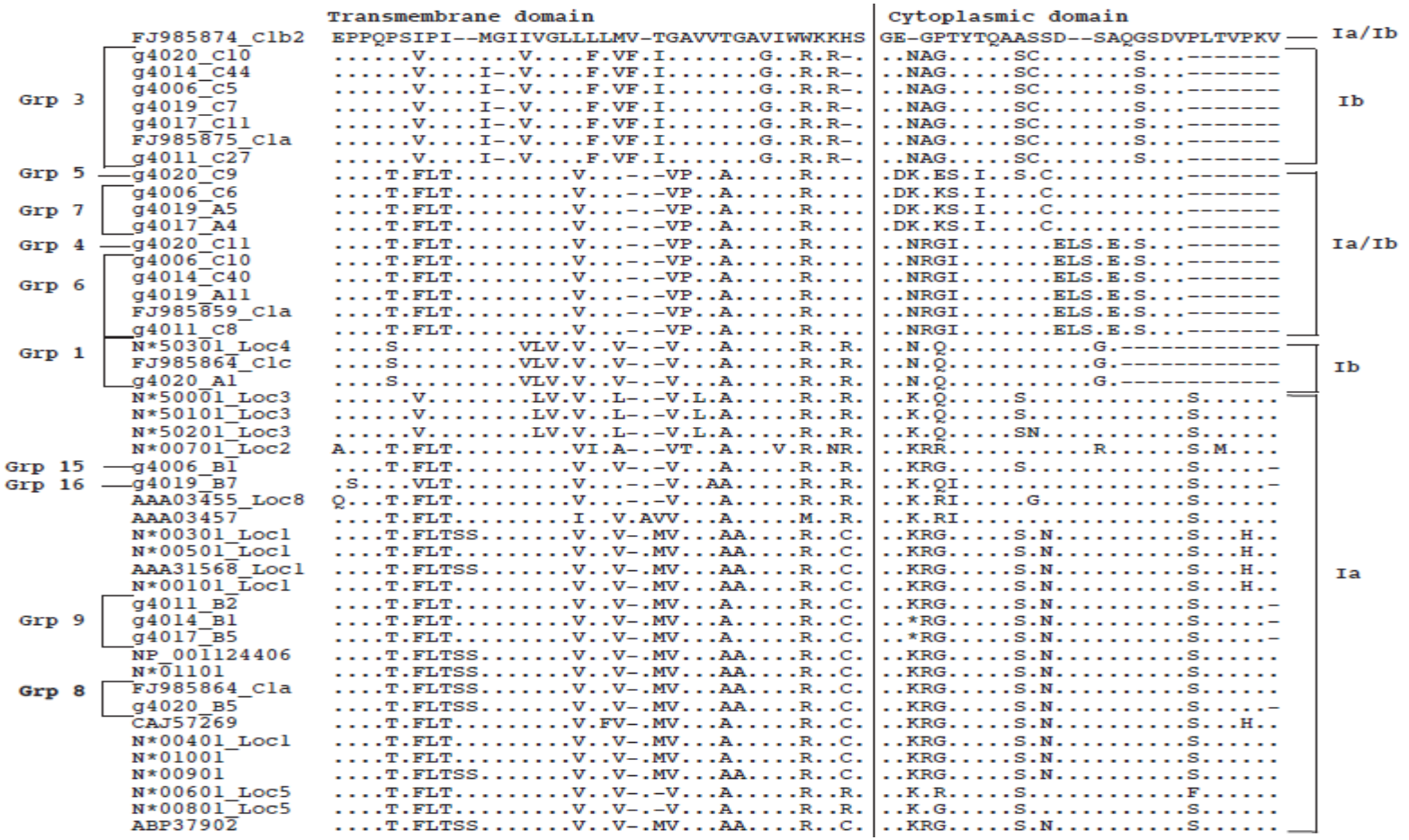


Figure 5.5: Alignment of the TM and CP domains and the classification of the sequences into classical (Ia) or non-classical (Ib) genes.

Group 1 is represented in one animal (4020) and also includes gene FJ985864rc_C1c and the MHC-IPD reference sequence N*50301. The TM domain of Group 1 has the non-classical VPI motif and is 34 amino acids in length. The exons six and seven of the CP domain are 12 and 6 amino acids in length respectively, hence the gene is truncated earlier in the exon seven than has been observed in other groups. Group 1 is a likely candidate for a non-classical class I locus.

Group 3 includes FJ985875rc_C1a and these sequences have a VPI motif in the TM domain (exon 5) of 36 amino acids and a CP domain consisting of an exon 6 of 12 amino acids together with a truncated exon 7 of 9 amino acids.

Group 7 is represented in three animals (4011, 4017 and 4019) and has a classical FLT motif in the TM domain, and has exons six and seven the same size as those seen in Group 3, but with a different pattern of amino acid substitutions.

Group 6 is represented in four animals and includes BAC gene FJ985859rc_C1a. Groups 6 and 7 are identical in the TM domain having the classical FLT motif and are 34 amino acids in length; however the groups differ in the CP domain. Groups 6 and 7 have a unique characteristic in that they have the classical FLT motif in the TM domain, but a truncated CP domain. Group 6 also has exon 6 of the same length (12 amino acids) but a different pattern of substitution, and a unique motif (ELS) in exon 7 which appears in the alignment as a substitution E/D followed by an insertion of 2 amino acids (LS), giving a total of 11 amino acids. Exon seven is truncated at the same position seen in Groups 3 and 7.

Group 9 is represented in three animals (4011, 4014 and 4017) and Group 8 is represented in one animal and also includes BAC sequence FJ985864rc_C1a and the reference sequence N*01101. Groups 8 and 9 both have FLT motifs in the TM domain and full-length exons 6 and 7 of 17 and 12 amino acids respectively. The substitution pattern in the TM and CP domain for these two groups is shared with a number of reference sequences identified as belonging to Locus 1 by Ballingall *et al.* (2008). Group 8 has an insertion of two amino acids (SS) immediately following the

FLT motif that is absent in Group 9, hence Group 8 has a TM domain of 36 amino acids in length compared to 34 amino acids in Group 9. Related substitution patterns are evident for alleles N*50001, N*50101 and N*50201 (Locus 3), alleles AAA03455 (Locus 5) and AAA03457 (Locus 8) and alleles N*00601 and N*00801 (Locus 5). It is noteworthy that accession AAA03455, proposed as an allele of Locus 5 by Ballingall, *et al.* (2008) is more similar to AAA03457 (Locus 8) in this region. The class I gene FJ985874_C1b identified from BAC sequence has an IPI motif in the TM domain, but a full-length CP domain.

Groups 15 and 16 are each represented by one sequence; they share similar but not identical TM domain of 34 amino acids and CP domain of 17 and 12 amino acids in length respectively. Both sequences are similar to reference sequence N*00701 (Locus 2) and are likely to be classical class I genes, although sequence g4019_B7 has an unusual VLT motif instead of the FLT motif seen in sequence g4006_B1. Also, g4006_B1 appears to be more similar to reference sequences N*00601 and N*00801 (Locus 5) in this region.

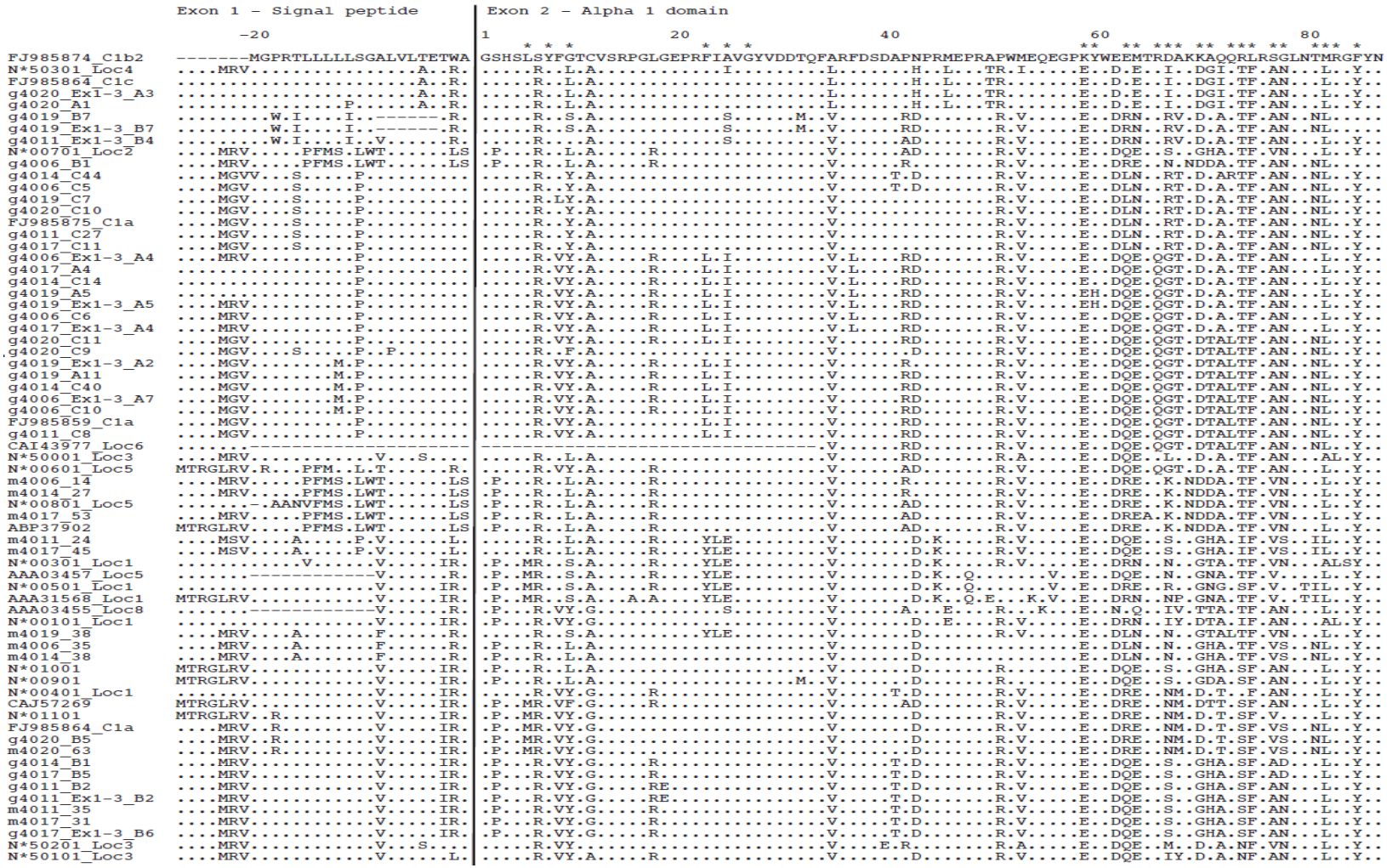
5.3.5 Phylogenetic analysis of sequences containing exons 1 to 3

Neighbor-joining (NJ) and Maximum Likelihood (ML) phylogenetic trees were inferred from a multiple sequence alignment of the predicted amino acid from the first three exons of MHC class I genes derived from genomic and cDNA sequences in this study, along with reference sequences shown in Table 5.4 (alignment shown Figure 5.6). The NJ and ML trees had minor rearrangement of some sequences within clades, but the overall topologies were the same. The bootstrapped NJ tree is shown in Figure 5.7. Assignment of clades to independent loci, as opposed to allelic variants at the same locus, was based on the following criteria.

a. The presence of one of the class I genes identified in the BAC sequences clearly indicates a separate locus since these genes have been identified in disparate locations in the genomic contigs.

b. Presence of a locus assigned by Ballingall *et al*. (2008) for which one or more full-length sequences is available in this study to corroborate previous evidence.

c. Observance in separate clades of a sequence from the same homozygous animal indicates that these clades represent separate loci. Eleven potential loci were identified in this tree based on clade arrangement and branch lengths. Loci 7 and 8, as assigned by Ballingall *et al*. (2008), are not included in this tree as they are partial sequences containing only exons 4-8.


Exon 1 – Signal peptide
Exon 2 – Alpha 1 domain
FJ985874_C1b2 -------MGPRTLLLLLSGALVLTETWA GSHSLSYFGTCVSRPGLGEPRFIAVGYVDDTQFARFDSDAPNPRMEPRAPWMEQEGPKYWEEMTRDAKKAQQRLRSGLNTMRGFYN
N*50301_Loc4 ....MRV.....A..R. .....R..L.A.....I.....L.....H..L...TR.I.....E..D.E..I..DGI.TF.AN...L..Y..
FJ985864_C1c .....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..
g4020_Ex1-3_A3 .....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..
g4020_A1 .....P.....A..R. .....R..L.A.....L.....H..L...TR.....E..D.E..I..DGI.TF.AN...L..Y..
g4019_B7 .....W.I....I.-------R. .....R..S.A.....S.....M..V.....RD.....R.V.....E..DRN..RV.D.A.TF.AN..NL.....
g4019_Ex1-3_B7 .....W.I....I.-------R. .....R..S.A.....S.....M..V.....RD.....R.V.....E..DRN..RV.D.A.TF.AN..NL.....
g4011_Ex1-3_B4 .....W.I....I.V.....R. .....R...A.....S.....V.....AD.....R.V.....E..DRN..RV.D.A.TF.AN...L..Y..
N*00701_Loc2 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DQE..S..GHA.TF.VN...L..Y..
g4006_B1 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..N.NDDA.TF.AN..NL.....
g4014_C44 ....MGVV....S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.ARTF.AN..NL..Y..
g4006_C5 ....MGV.....S.....P..... .....R..Y.A.....V.....T.D.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
g4019_C7 ....MGV.....S.....P..... .....R.LY.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
g4020_C10 ....MGV.....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
FJ985875_C1a ....MGV.....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
g4011_C27 ....MGV.....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
g4017_C11 ....MGV.....S.....P..... .....R..Y.A.....V.....R.V.....E..DLN..RT.D.A.TF.AN..NL..Y..
g4006_Ex1-3_A4 ....MRV.....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
g4017_A4 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
g4014_C14 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
g4019_A5 .....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....EH.DQE.QGT.D.A.TF.AN...L..Y..
g4019_Ex1-3_A5 ....MRV.....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....EH.DQE.QGT.D.A.TF.AN...L..Y..
g4006_C6 ....MRV.....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
g4017_Ex1-3_A4 ....MRV.....P..... .....R.VY.A.....R....L.I.....V.L....RD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
g4020_C11 ....MGV.....P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4020_C9 ....MGV....S.....P..P..... .....R..F.A.....V.....D.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4019_Ex1-3_A2 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....R.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4019_A11 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4014_C40 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4006_Ex1-3_A7 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4006_C10 ....MGV.....M.P..... .....R.VY.A.....R....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
FJ985859_C1a ....MGV.....P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
g4011_C8 ....MGV.....P..... .....R.VY.A.....L.I.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
CAI43977_Loc6 .....------------------- ----------.....V.....RD.....R.V.....E..DQE.QGT.DTALTF.AN..NL..Y..
N*50001_Loc3 ....MRV.....V...S..... .....R..L.A.....V.....RD.....R.A.....E..DQE..L..D.A.TF.AN...AL.Y..
N*00601_Loc5 MTRGLRV.R...PFM..L.T.....R. .....R.VY.A.....R.....V.....AD.....R.V.....E..DQE.QGT.D.A.TF.AN...L..Y..
m4006_14 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..
m4014_27 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....R.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..
N*00801_Loc5 .....-.AANVFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..
m4017_53 ....MRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DREA.K.NDDA.TF.VN...L..Y..
ABP37902 MTRGLRV.....PFMS.LWT.....LS .P...R..L.A.....R.....V.....AD.....R.V.....E..DRE..K.NDDA.TF.VN...L..Y..
m4011_24 ....MSV....A.....P.V.....L. .....R..L.A.....R....YLE.....V.....D.K.....R.V.....E..DQE..S..GHA.IF.VS..IL..Y..
m4017_45 ....MSV....A.....P.V.....L. .....R..L.A.....R....YLE.....V.....D.K.....R.V.....E..DQE..S..GHA.IF.VS..IL..Y..
N*00301_Loc1 .....V.....V.....IR. .P..MR...S.A.....R....YLE.....V.....D.K.....R.V.....E..DRN...N..GTA.TF.VN...ALSY..
AAA03457_Loc5 .....---------V.....R. .P..MR...S.A.....R....YLE.....V.....D.K..Q.....V..E..DQE..N..GNA.TF.V....L..Y..
N*00501_Loc1 .....V.....IR. .P..MR...S.A.....R....YLE.....V.....D.K..Q.....V..E..DRE..R..GNG.SF.V..TIL..Y..
AAA31568_Loc1 MTRGLRV.....V.....IR. .P..MR...S.A...A.A....YLE.....V.....D.K..Q.E...K.V..E..DRN..NP.GNA.TF.V..TIL..Y..
AAA03455_Loc8 .....---------V.....R. .P...R.VY.G.....S.....V.....A...E....R...K...E..N.Q..IV.TTA.TF.AN...L..Y..
N*00101_Loc1 .....V.....IR. .P...R.VY.G.....V.....D..E....R.V.....E..DRN..IY.DTA.IF.AN...AL.Y..
m4019_38 ....MRV....A.....F.....R. .....R...S.A.....YLE.....V.....D.....R.V.....E..DLN...N..GTALTF.VN...L..Y..
m4006_35 ....MRV....A.....F.....R. .P...R..L.A.....V.....D.....E..DLN...N..GHA.TF.VS..NL..Y..
m4014_38 ....MRV....A.....F.....R. .P...R..L.A.....V.....D.....E..DLN...N..GHA.TF.VS..NL..Y..
N*01001 MTRGLRV.....V.....IR. .P...R..L.A.....V.....D.....R.....E..DQE..S..GHA.SF.AN...L..Y..
N*00901 MTRGLRV.....V.....IR. .P...R..L.A.....M..V.....D.....R.....E..DQE..S..GDA.SF.AN...L..Y..
N*00401_Loc1 .....V.....IR. .....R.VY.G.....R.....V.....T.D.....R.V.....E..DRE..NM.D.T..F.AN...L..Y..
CAJ57269 MTRGLRV.....V.....IR. .P..MR.VF.G.....R.....V.....AD.....R.V.....E..DRE..NM.DTT.SF.AN...L..Y..
N*01101 MTRGLRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.V....L..Y..
FJ985864_C1a ....MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..
g4020_B5 ....MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..
m4020_63 ....MRV..R.....V.....IR. .P..MR.VY.G.....V.....D.....R.V.....E..DRE..NM.D.T.SF.VS..NL..Y..
g4014_B1 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..
g4017_B5 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AD...L..Y..
g4011_B2 ....MRV.....V.....IR. .P...R.VY.G.....RE.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..
g4011_Ex1-3_B2 ....MRV.....V.....IR. .P...R.VY.G.....RE.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..
m4011_35 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..
m4017_31 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..
g4017_Ex1-3_B6 ....MRV.....V.....IR. .P...R.VY.G.....R.....V.....T.D.....R.V.....E..DQE..S..GHA.SF.AN...L..Y..
N*50201_Loc3 ....MRV.....V...S..... .....R.VY.....V....E.R.....R.A.....E..DQE..M..D.A.NF.VN...L..Y..
N*50101_Loc3 ....MRV.....V.....L. .....R.VY.A.....R.....V.....D.....R.V.....E..DQE..IY.D.A.NF.AN...L..Y..

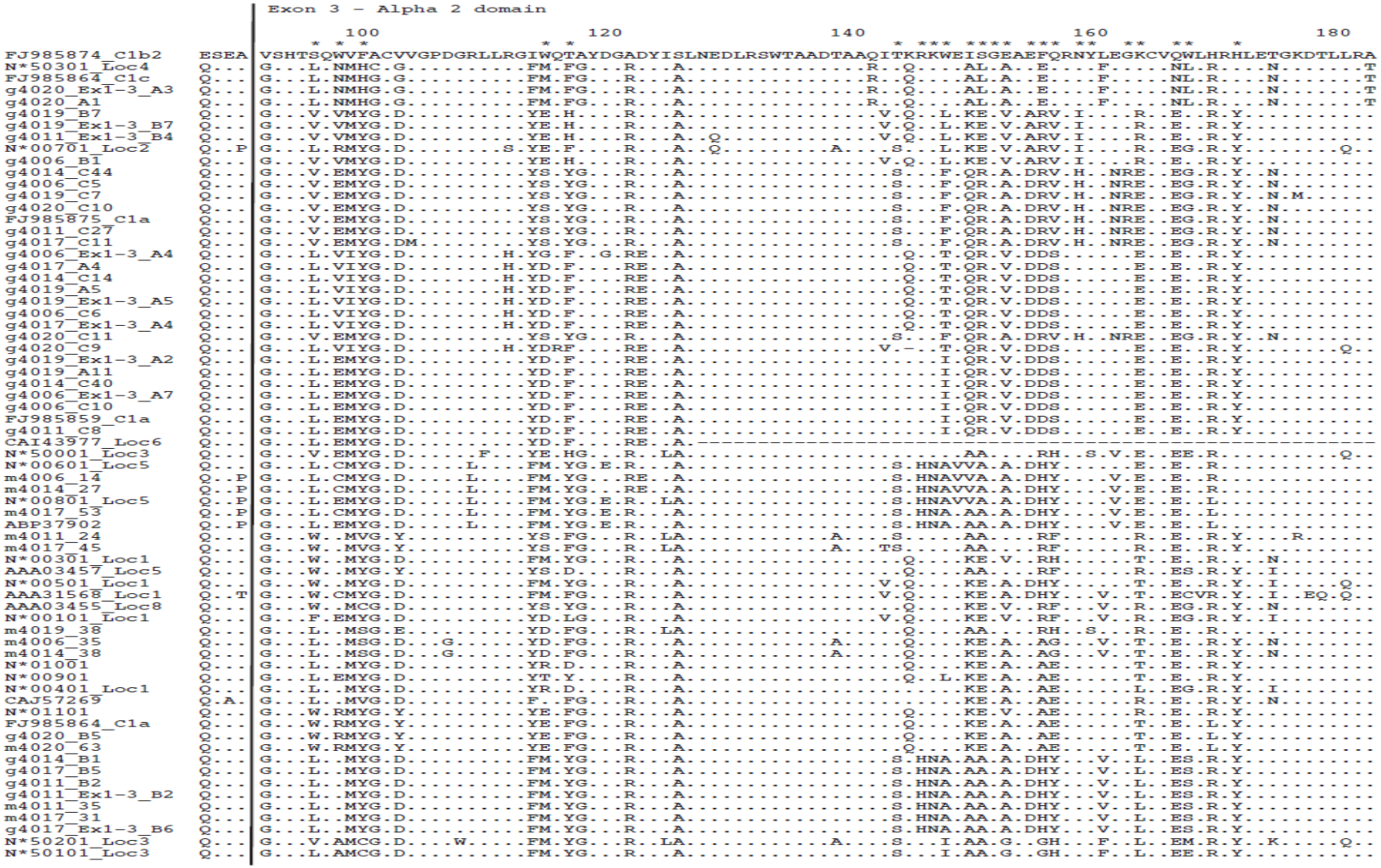


Figure 5.6: Multiple alignment of MHC class I amino acid sequences (exon 1 to 3).

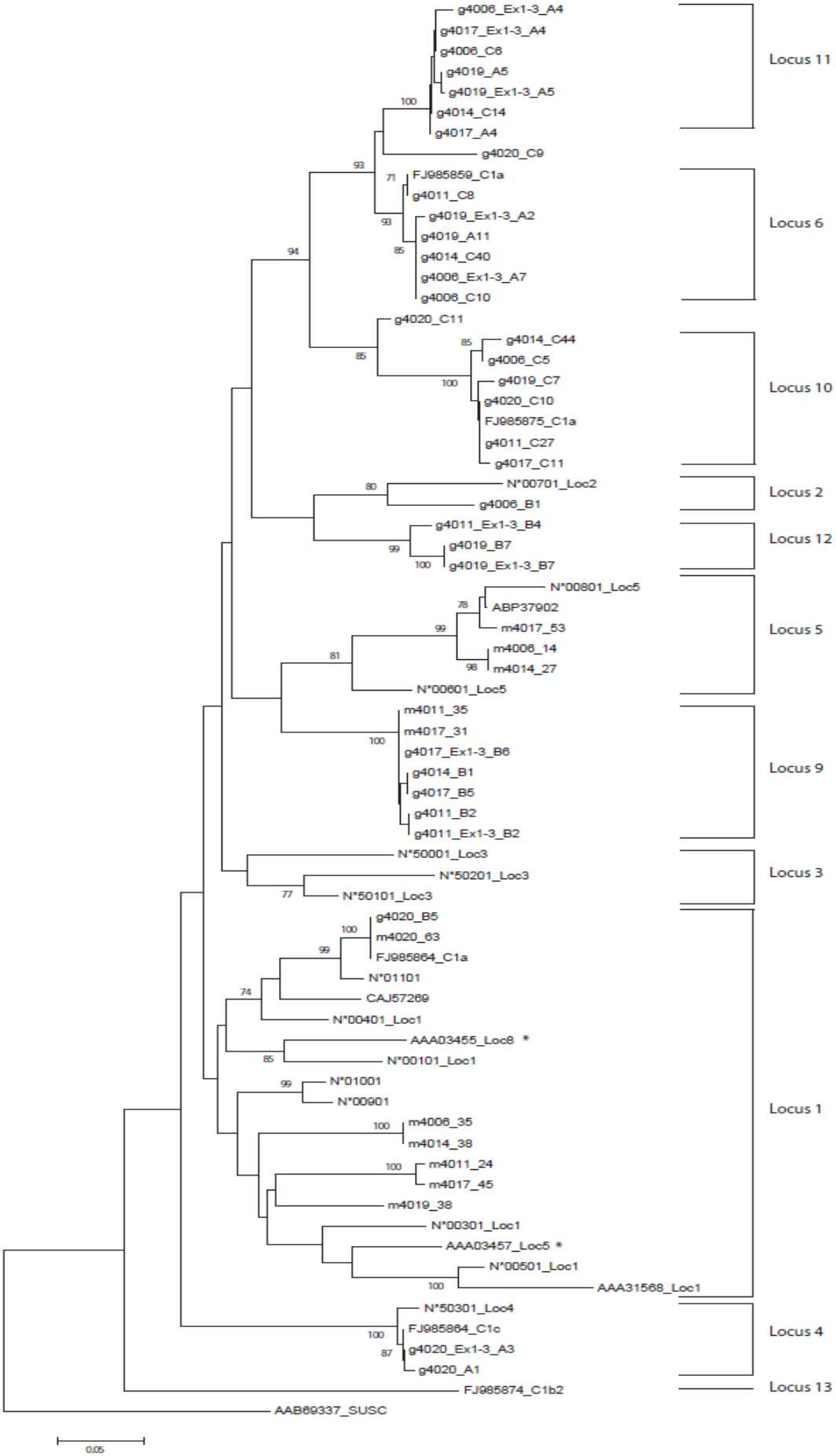

g4006_Ex1-3_A4
g4017_Ex1-3_A4
g4006_C6
g4019_A5
g4019_Ex1-3_A5
g4014_C14
g4017_A4
g4020_C9
FJ985859_C1a
g4011_C8
g4019_Ex1-3_A2
g4019_A11
g4014_C40
g4006_Ex1-3_A7
g4006_C10
g4020_C11
g4014_C44
g4006_C5
g4019_C7
g4020_C10
FJ985875_C1a
g4011_C27
g4017_C11
N*00701_Loc2
g4006_B1
g4011_Ex1-3_B4
g4019_B7
g4019_Ex1-3_B7
N*00801_Loc5
ABP37902
m4017_53
m4006_14
m4014_27
N*00601_Loc5
m4011_35
m4017_31
g4017_Ex1-3_B6
g4014_B1
g4017_B5
g4011_B2
g4011_Ex1-3_B2
N*50001_Loc3
N*50201_Loc3
N*50101_Loc3
g4020_B5
m4020_63
FJ985864_C1a
N*01101
CAJ57269
N*00401_Loc1
AAA03455_Loc8 *
N*00101_Loc1
N*01001
N*00901
m4006_35
m4014_38
m4011_24
m4017_45
m4019_38
N*00301_Loc1
AAA03457_Loc5 *
N*00501_Loc1
AAA31568_Loc1
N*50301_Loc4
FJ985864_C1c
g4020_Ex1-3_A3
g4020_A1
FJ985874_C1b2
AAB69337_SUSC
0.05
Locus 11
Locus 6
Locus 10
Locus 2
Locus 12
Locus 5
Locus 9
Locus 3
Locus 1
Locus 4
Locus 13

Figure 5.7: Phylogenetic tree generated based on alignment of unique MHC class I sequence identified in homozygous animals and MHC class I reference sequences for exon 1 to 3. Locus assignments reported by Ballingall, *et al*. (2008) are indicated in the sequence identifier, for example, Loc1 corresponds to a sequence believed to be a Locus 1 allele. Loci 7 and 8 as assigned by Ballingall *et al*. (2008) are not included in this tree.

5.3.6 Phylogenetic analysis of full-length sequences

The NJ and ML phylogenetic trees were inferred from an alignment of amino acid predictions derived from the genomic sequences from this study (MHC class I exons 1-7), together with full-length reference sequences previously shown in Table 5.4. The alignment is shown in Figure 5.4. The amino acid sequences represent full-length protein sequences except for the truncated CP domains in the class Ib genes, and a single amino acid from the short eighth exon typically present in class Ia genes. The NJ and ML trees had minor rearrangement of some sequences within clades, but the overall topologies were the same. The bootstrapped NJ tree is shown in Figure 5.8a. Thirteen putative independent loci have been identified in this tree, eleven of which correspond to loci identified in the tree inferred from an alignment of exons 1-3. AAA03455 representing Locus 8 as assigned by Ballingall *et al*. (2008) appears as a separate long branch on this tree, and is thought to represent an independent locus. The sequence from accession AAA03457 also appears on a long branch in this tree and may also represent an independent locus (identified as Locus 14). This sequence has been proposed as an allelic variant of Locus 5 by Ballingall *et al*. (2008) based on previous phylogenetic analysis (Miltiadou *et al.* 2005). However in this full-length tree it is clearly separated from the other Locus 5 sequences, appearing in a clade with AAA03455 but with long separating branches; thus it represents a possible independent locus. Locus 7 as assigned by Ballingall *et al*. (2008) is not represented in this tree as it represents partial sequences from exons 4-8.

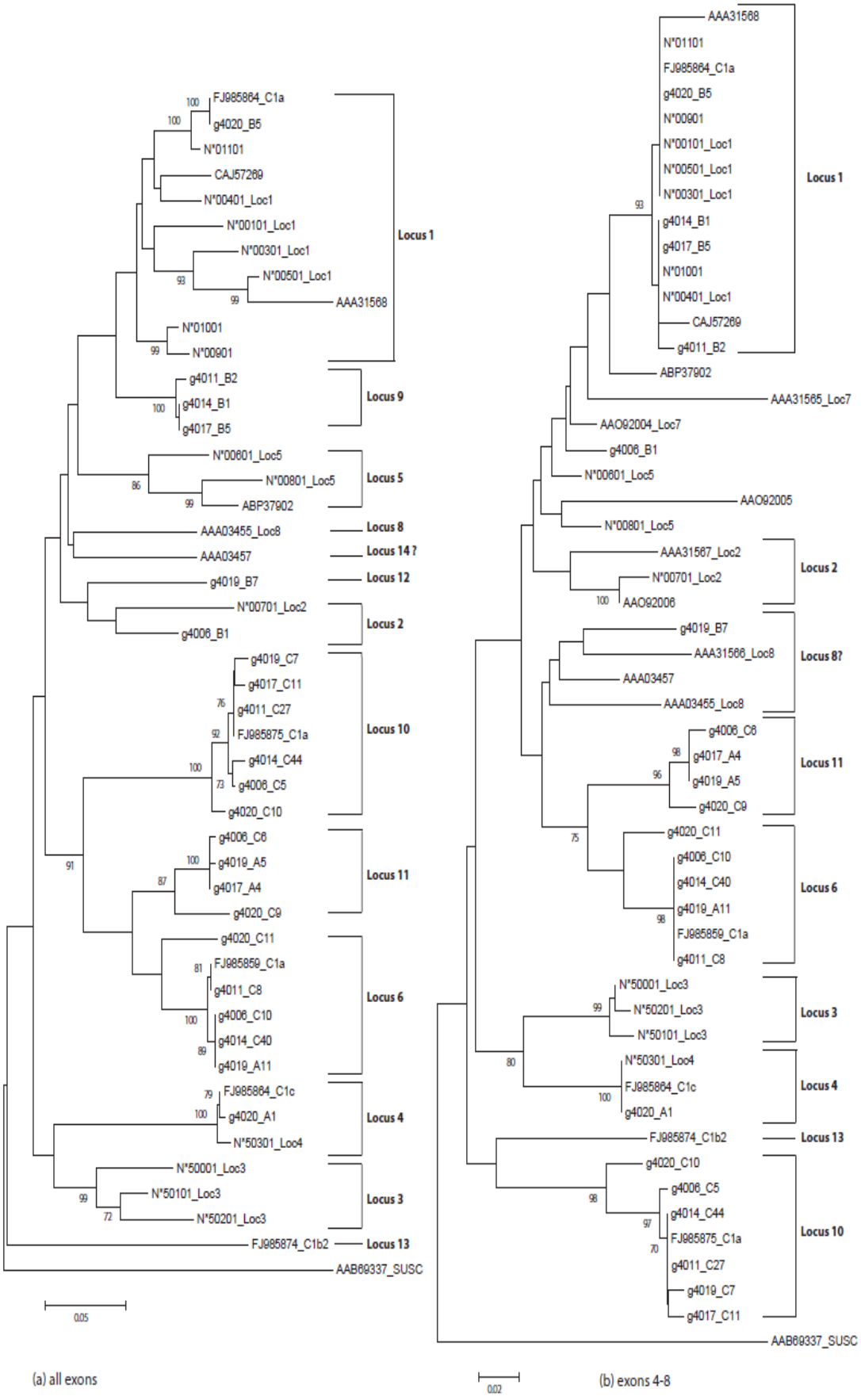


Figure 5.8: (a) NJ tree representing full-length sequences. (b) NJ tree representing sequence consisting of exon 4 to 8.

5.3.7 Phylogenetic analysis of exons 4-8 from full-length sequences

The NJ and ML trees based on the JTT matrix model were inferred from an alignment of exons 4-7 from representative class I sequences isolated in this study, along with the reference sequences, including sequences from Locus 7 assigned by Ballingall *et al*. (2008) and an additional partial sequence from Ballingall's assigned Locus 8. The general topology was similar in both trees. However, the overall topology of the exon 4-7 NJ tree (Figure 5.8b) was not consistent in some areas when compared to the trees inferred from alignments of the full-length sequences (Figure 5.8a). Locus 1 in Figure 5.8b appears to have merged with Locus 9 identified in Figure 5.8a, with the clade flattening into multi-furcated or very short branches. Sequence g4006_B1 (previously in clade with Locus 2) appears on a separate branch to N*00701 (Locus 2) instead of being in the same clade. Sequences N*00601 (Locus 5) appears on a separate branch to sequence N*00801 (Locus 5) and ABP37902 (previously in a clade with Locus 5). ABP37902 is well separated from N*00601 and N*00801 and is present adjacent to the branch separating the Locus 1 sequences. Sequence AAO92005 appears in a clade with N*00801, and may represent another Locus 5 sequence. ABP37902 was placed on an adjacent branch to the clade identified as Locus 1 in Figure 5.8b, whereas it appeared in a clade with Locus 5 sequences in tree inferred from full-length sequences (Figure 5.8a). Sequences of the proposed Locus 7 (shown only in Figure 5.8b), were not clustered in a definitive clade, although they did appear on adjacent branches. Notably, Locus 8 in Figure 5.8b consists of 4 sequences (g4019_B7, AAA31566_Loc8, AAA03457 and AAA03455_Loc8). Sequence g4019_B7, proposed as a possible Locus 12 in Figure 5.8a was placed in a clade with the proposed Locus 8 sequence AAA31566, although with long branches. AAA03457 appeared on a separate branch adjacent to the AAA03455 (Locus 8) branch in the tree. Clades containing the remaining proposed loci (3, 4, 6, 10, 11 and 13) appear in a similar arrangement to that seen in Figure 5.8a.

5.3.8 Number of loci in homozygous animals

The number of loci identified for each homozygous animal and for the BAC sequences derived from the Chinese Merino sheep are shown in Table 5.6. Loci are based on clades identified in Figures 5.7 and 5.8. Animal 4006 has six MHC class I loci, of which, three appear to represent MHC class Ia genes. Two of the four loci in animal 4011 have evidence of expression of MHC class Ia genes. In animal 4014, mRNA evidence for MHC class Ia gene is available for two out of six loci. Animal 4017 has five MHC class I loci expressing only three MHC class Ia genes. Animals 4019 and 4020 may only have one classical locus expressed, although genomic sequence evidence exists for another class Ia gene in animal 4019, for which mRNA evidence have not been obtained.

No two animals have exactly the same loci according to the data generated in this study. Based on isolation of mRNA, individual sheep may contain between one and three expressed classical class I loci. There is genomic evidence to indicate the presence of between two and four intermediate or non-classical class I loci but no mRNA evidence were identified for these loci. Similar patterns of class I gene expression have been reported in sheep (Miltiadou *et al.* 2005; Ballingall *et al.* 2008) and in cattle (Birch *et al.* 2008a; Birch *et al.* 2008b).

Table 5.6: Genomic DNA and mRNA evidence for loci in homozygous animals and BAC sequences.

Animal	Locus 1 (Ia)	Locus 2 (Ia)	Locus 4 (Ib)	Locus 5 (Ia)	Locus 6 (Ia/1b)	Locus 9 (Ia)	Locus 10 (Ib)	Locus 11 (Ia/Ib)	Locus 12 (Ia)	Locus 13 (Ia/Ib)
4006	m	g		m	g		g	g		
4011	m				g	g/m	g			
4014	m			m	g	g	g	g		
4017	m			m		g/m	g	g		
4019	m				g		g	g	g	
4020	g/m		g		g		g	g		
BACs	64rc_C1a		64rc_C1c		59rc_C1a		75rc_C1a			74_C1b

Loci are identified based on phylogenetic analysis (Figures 5.7 and 5.8.). Ia: classical class I gene based on sequence analysis. Ib: non-classical class I gene. Ia/Ib: intermediate features. g: genomic sequence. m: mRNA sequence. BACs: Class I sequences identified from BAC sequences. Accession prefix for BAC sequences: 'FJ9858'. rc: sequence was reverse complemented before gene prediction analysis.

5.3.9 Frequency of mRNA transcripts

The distribution of mRNA sequences isolated in this study among identified groups and loci is presented in Table 5.7. Clones were randomly chosen for sequencing, so it is quite possible that less frequent transcripts are under-represented or completely missed. However, it suggests that each animal appears to have a unique mRNA expression pattern. The mRNA sequences obtained from animals 4017 and 4011 have transcripts that include intronic sequence between exons 1 and 2. Animals 4006, 4011 and 4114 have transcripts that represent two functional proteins. Animals 4019 and 4020 both have transcripts that represent only one functional protein. Animal 4017 has transcripts that represent three functional proteins. Animals 4006 and 4014 have functional transcripts from the same two groups/loci. Animals 4011 and 4017 have transcripts from the same groups/loci bar an additional group/locus found only in 4017 at low frequency. Locus 1 (identified as such according to phylogenetic analysis) is the only locus represented by mRNA from all animals. Locus 5 is represented by expressed transcripts in three of the six animals. Locus 9 (a new locus identified in this study) is represented in two animals.

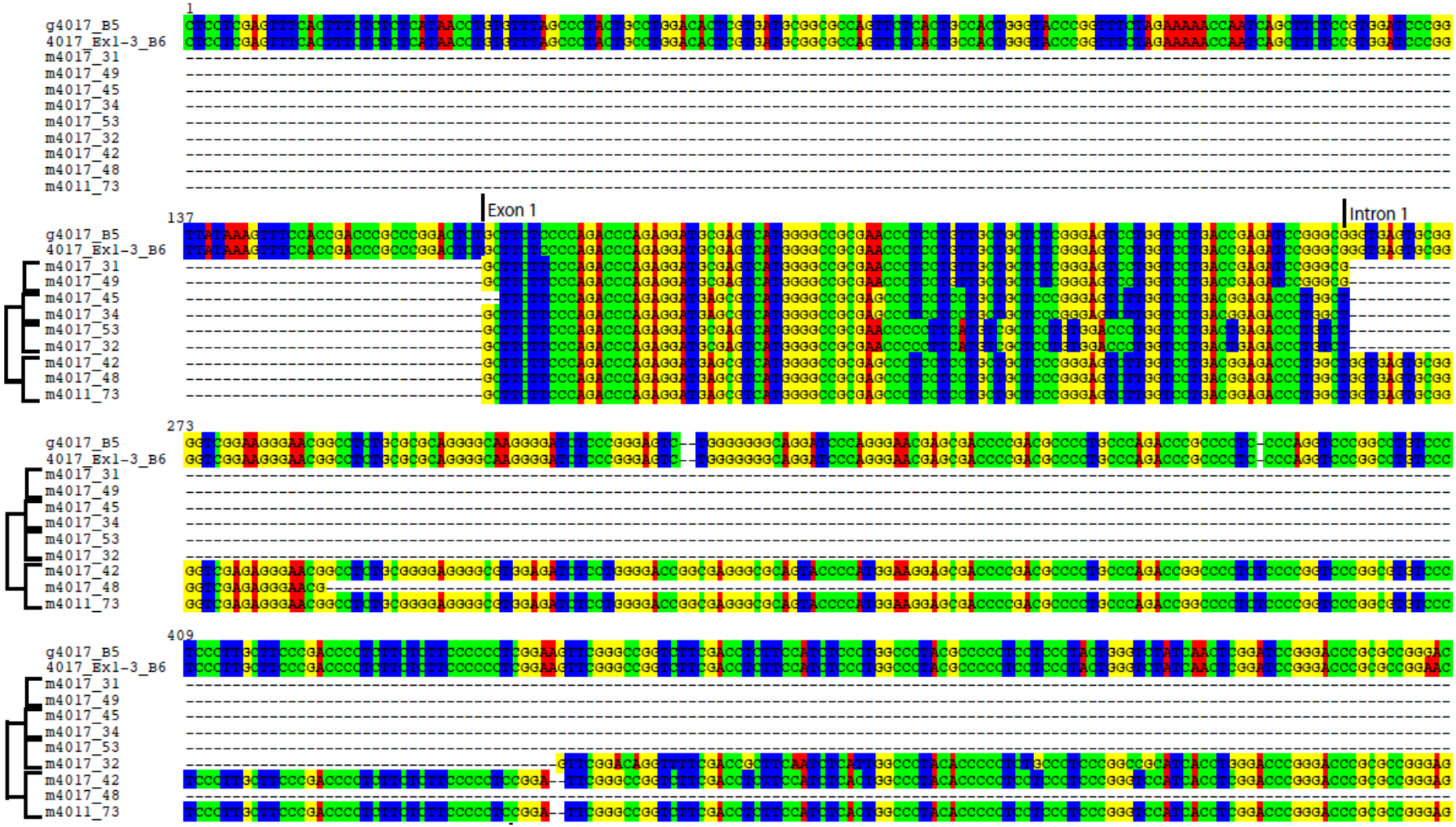
Table 5.7: Frequency of mRNA at expressed loci in homozygous animals.

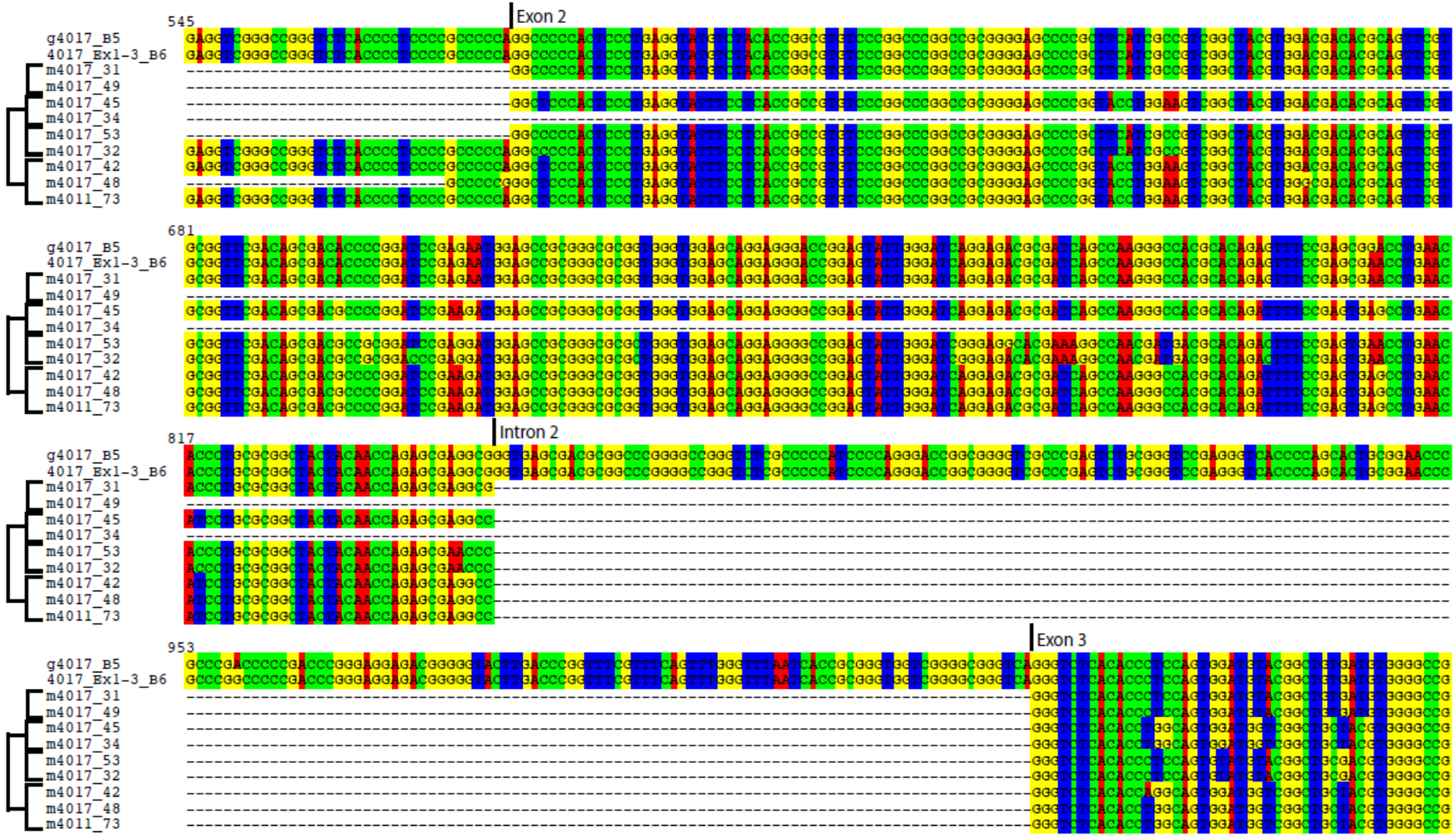
Animal	**Total seqs**	**Grp/Loc/ #**	**Grp/Loc/ #**	**Grp/Loc/ #**	**Pseudo1 / #**	**Pseudo2 / #**
4006	14	11 / 1 / 6	13 / 5 / 7		Stop / 1	
4011	13	9 / 9 / 5 and a splice variant	10 / 1 / 4		Stop / 2	Transcribed intron / 1
4014	11	11 /1 / 5	13 / 5 / 6			
4017	13	9 / 9 / 3	10 / 1 / 4	14 / 5 / 1	Splice variant / 2	Transcribed intron / 3
4019	13	12 / 1 / 12			Stop / 1	
4020	12	8 / 1 / 11			Bad splicing / 1	

Total seqs: total number of mRNA sequences. Grp: MHC-IPD group. Loc: locus. #: number of sequence. Pseudo: pseudogene. Stop: premature stop codon.

5.3.10 Evidence for transcribed pseudogenes

Most of the homozygous animals generated at least one mRNA sequence with a premature stop codon indicating a probable transcribed pseudogene (Table 5.7). Many of these had stop codons that would result in a truncated protein if translated. Animals 4017 and 4011 had expressed transcripts (m4011_73, m4017_32, m4017_42 and m4017_48) that incorporated sequence from the first intron (Figure 5.9). Translated proteins from these transcripts are out of frame and are therefore unlikely to be functional proteins. Two transcripts from animal 4017 (m4014_34 and m4014_49) appeared to be splice variants missing exon 2 (Figure 5.9).


1
g4017_B5
4017_Ex1-3_B6
m4017_31
m4017_49
m4017_45
m4017_34
m4017_53
m4017_32
m4017_42
m4017_48
m4011_73
Exon 1
Intron 1
137
273
409


Exon 2
Intron 2
Exon 3
545
681
817
953
g4017_B5
4017_Ex1-3_B6
m4017_31
m4017_49
m4017_45
m4017_34
m4017_53
m4017_32
m4017_42
m4017_48
m4011_73

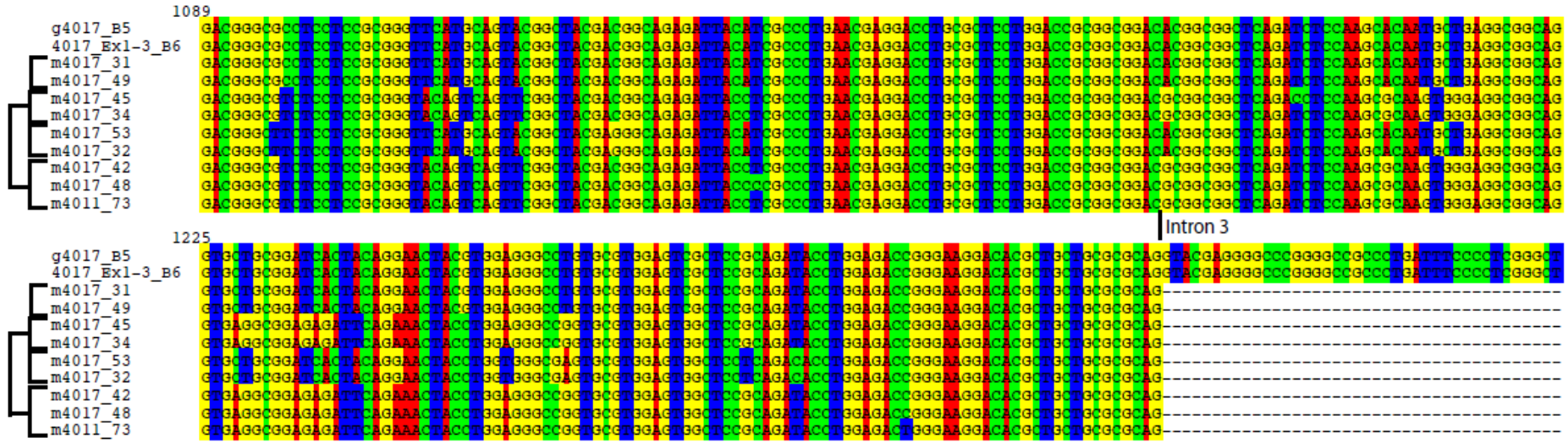


Figure 5.9: Alignment of transcribed pseudogenes from animal 4011 and 4014. [Sequences grouped according to similarity.

5.4 Discussion

This study intended to define the genomic architecture of the complex class I region in Australian Merino sheep. In order to establish the number of genes, both classical and non-classical, and describe the diversity of MHC class I sequences, homozygous animals were generated. The use of homozygous animals eliminates allelic variation and greatly simplifies the assignment of discrete loci. Of the eight animals initially identified to be homozygous at MHC, only six were subsequently used. These animals were selected because they had different microsatellite profiles suggesting different haplotypes. Further analysis using SNPs from the MHC class I region showed that all six animals had a unique SNP haplotype (see Chapter 6).

It was found that the high level of sequence variation between different copies of the MHC class I genes from within and between sheep breeds (Miltiadou *et al.* 2005; Ballingall *et al.* 2008; Gao *et al.* 2010) was a drawback in terms of the design of universal primers. In this study, three sets of primers were designed based upon the BAC sequences published by Gao *et al*. (2010) and used to amplify full-length genomic and cDNA sequences. The use of multiple sets of primers to amplify MHC class I sequences has also been reported in other sheep studies (Miltiadou *et al.* 2005; Ballingall *et al.* 2008). It appears to be very difficult to design a single universal MHC class I gene primer set in sheep. Previously published primers from other breeds of sheep (Miltiadou *et al.* 2005; Ballingall *et al.* 2008) were not able to amplify class I genes in Australian Merino sheep. Primers designed based upon potential class I loci identified from Chinese Merino sheep derived BAC sequences (Gao *et al.* 2010) was successful in amplifying full-length genomic DNA (except for a valine in exon 8), but was found not to amplify cDNA. This is because the primer sets used in this study were located 3' to exon 8 but were probably not located in an expressed region of the DNA. The use of primers from other breeds (Miltiadou *et al.* 2005; Ballingall *et al.* 2008) to amplify full-length cDNA also yielded no result. Furthermore, it was found that it was difficult to design primers using the genomic sequences generated in this study due to the high sequence diversity. This indicates that it is difficult to design sheep specific primers and future efforts should be focussed on breed specific primers. Additional

work is required to establish the extent of class I interbreed variation in sheep. Hence, full-length cDNA sequences were not used in this study and truncated cDNA sequences encompassing exons 1-4 (amplified from full length cDNA) were used instead.

The primer design based upon the BAC sequences had limitations; only one classical class I gene was identified in any of the MHC homozygous animals. Using mRNA analysis, it was clear that additional classical MHC class I loci were present in these animals but were not identified from the genomic DNA amplified using the three sets of primers. Previous dotplot and pairwise sequence alignment analysis of these BAC sequences showed some discrepancies from the published map (Gao *et al.* 2010). A comparison with the bovine MHC map suggests there may be a cryptic gap in Gao's contigs where one or more class I genes usually reside (see also Chapter 3). Therefore, it is likely that the primers designed in this study are not specific for all of the possible MHC class I sequences present. The complexity of the MHC class I region (a likely result of not only a high degree of point mutations, but of significant gene recombination, duplication and conversion events) creates many difficulties both technically and analytically.

In situations where an mRNA transcript has not been identified for a particular locus, assignment of MHC class I genes into classical class I (Ia) or non-classical class I (Ib) loci is difficult. MHC class I genes are often classified as non-classical class I genes when the gene sequence fulfil one or more of certain criteria: a non-classical motif in the TM domain (VPI, IPI, VLIK); a truncated CP domain; monomorphic or with limited polymorphism (Birch *et al.* 2008a).

Group 1 (assigned Locus 4) is a non-classical class I gene candidate in that it has the IPI motif and is truncated in the CP domain. It is difficult to assess polymorphism as there are only three sequences in this group, one representative from homozygous animal (g4020_A1), an IPD reference sequence N*50301 and FJ985864_C1c from BAC sequence. The differences observed are in the signal peptide region (N*50301 has an MRV at the N-terminus not observed in the other two sequences and g4020A has a proline/leucine substitution) and in exon 2 (N*50301 has an

isoleucine/alanine substitution). No representative mRNAs for this gene were isolated in this study, however previous mRNA evidence (IPD reference N*50301) corroborates evidence from homozygous animals that this is a functional gene.

Group 3 from this study (assigned Locus 10), which includes FJ985875_C1a from BAC sequence, is a non-classical class I gene candidate. The group has a VPI motif in the TM domain and a truncated CP domain (seventh exon). Genomic sequences for this gene were found in all six animals studied. Amino acid substitutions in this locus were observed primarily in exons 2 and 3. The genomic sequence g4011_27, which was also present in Group 3 was identical to FJ985875_C1a. No mRNAs were isolated for this group from the homozygous animals. This could be due to low expression levels, non-expression in lymphocytes, or, alternatively, it is a pseudogene.

Groups 6 and 7 (assigned Locus 6 and 11) have classical FLT motifs in the TM domain, but have truncated CP domains. These two groups represent different loci since they appear to have been co-isolated from the same homozygous animals. The patterns of substitution observed in the signal peptide, α1, α2, and in the CP domains are markedly different between the two groups, and they consistently form distinct but adjacent clades in all phylogenetic trees. Group 6 contains sequence FJ985859_C1a isolated from BAC. No expression evidence is currently available to confirm that these are functional genes, as no known reference proteins appear to be closely related and no mRNAs belonging to either group were successfully isolated in this study. Again this may be due to low expression levels, lack of expression in lymphocytes or problems with primer specificity. Alternatively, this may be a pseudogene; however there is no supporting sequence evidence to indicate this (i.e. no introduced stop codons, no missing exon splice sites, etc).

Assignment of MHC class I sequences to a particular locus is also difficult. Classification of MHC class I sequence into MHC-IPD groups does not equate to classification into loci, as there could be allelic diversity within a locus and differences in the number of loci present and/or expressed in different animals (Miltiadou *et al.* 2005; Ballingall *et al.* 2008). Complicating matters

further, there appears to be a number of transcribed class I pseudogenes. This makes definitive identification of loci difficult in the absence of genome mapping data from a wide variety of animals. The total number of groups identified from the alignment of sequences isolated in this study along with the reference sequences is indicative of the high diversity of MHC class I sequences in sheep. However, the phylogenetic analysis suggests that the number of probable loci is less. The classification of groups is arguably too narrow for separation into loci.

In this study, three types of phylogenetic trees were analysed for assignment of loci; the tree of exons 1 to 3 (Figure 5.7), the tree of full-length sequences (Figure 5.8a) and the tree of exons 4 to 8 (Figure 5.8b). This approach was followed because only exons 1 to 3 of cDNA sequences were identified instead of the full-length cDNA sequences derived from reverse transcribed mRNA. Similarly, not all genomic sequences generated in this study were full-length sequence. Some of the MHC class I clones were only sequenced from exon 1 to 3. The unique sequences among the short genomic sequences (exons 1 to 3) were then selected and sequenced completely. So initially, a tree based on exons 1 to 3 was generated to align cDNA sequence and exons 1 to 3 of amino acid sequence predicted from genomic DNA to determine the expressed loci. A full-length genomic DNA sequence tree was created to validate the assignment of locus based on exons 1 to 3. The trees derived from exons 1 to 3 and full-length sequences complimented each other; no re-arrangement of clades was observed. A tree of exons 4 to 8 was also created to assign locus as according to the method used in previous studies (Holmes *et al.* 2003; Miltiadou *et al.* 2005). These previous studies (Holmes *et al.* 2003; Miltiadou *et al.* 2005) have used exons 4 to 8 (TM and CP domains) to assign locus because those regions have less sequences variation compared to the exons 1-3 (α1, α2 and α3 domains), which manifest extensive sequence diversity related to their antigen presenting function (Hughes & Nei 1988). When the tree based on exons 4-8 was compared with the tree based on full length peptides, a number of clades were re-arranged. The assignment of a given sequence to a possible distinct locus was based finally on a conservative interpretation of all three tree types constructed and not solely on a specific phylogenetic tree.

This means of identifying distinct loci resulted in the detection of 14 putative class I loci in the sheep MHC. The phylogenetic analysis has shown that loci 6, 10, 11 are independent loci as they are carried on the same haplotype (in animals 4006 and 4019). Locus 13 is found only in the Chinese Merino but it is physically separated from the loci 6, 10 and 11 in the BAC contig map. All animals from which these loci are derived are either Australian Merino or Chinese Merino. This suggests that these loci may reflect a breeding history or in other terms an evolutionary history. Loci 3 and 4 are shown to cluster close together and Locus 4 has representation from all three breeds (Australian Merino, Chinese Merino and Scottish Blackface). These two loci are physically separated as demonstrated by their co-localisation on haplotype 501A in the Scottish Blackface (Miltiadou *et al.* 2005). However, no Merino sheep sequences in this study clustered with Locus 3 suggesting that this locus may be present in Scottish Blackface but not Merino breeds. Locus 3 has an IPI motif in the TM domain, but has a full-length CP domain and is classified by Ballingall *et al*. (2008) as intermediate class I (Ia/Ib). Locus 4 has been classified as a non-classical class I (Ib) gene (Ballingall *et al.* 2008). No cDNA was indentified in this study for Locus 4. Locus 5 contains two sequences from the Scottish Blackface (N*00601 and N*00801) and an unknown breed (ABP37902). The two Scottish Blackface derived sequences (N*00601 and N*00801) have been characterised as transcribed pseudogenes, which fail to be translated into protein (Ballingall *et al.* 2008). This present study failed to amplify Locus 5 in genomic DNA although amplification from cDNA was achieved in three animals (4006, 4014 and 4017). This illustrates the lack of universality of the primers used. Locus 1 has representative sequences derived from many different breeds and clusters close to locus 9 in this study (shown in Figure 5.8a). Locus 1 has sequences come from Scottish Blackface (N*00401, N*00101, N*00301, N*00501, Finnish Landrace (N*01101), Soay (same as N*01101), unknown breed (AAA31568), Australian Merino (g4020_B5) and Chinese Merino (FJ985864_C1a). Locus 9 has sequences from only Australian Merino. Loci 1 and 9 are also both classical loci as expression evidence is available in both breeds. Although loci 1 and 9 cluster close together phylogenetically, they appear to be independent loci, as expression of these loci in the same homozygous animals has been observed (4011 and 4017, see Figure 5.7). Thus, it would be tempting to speculate that Locus 1 and Locus 9 are either

recent duplications or simply reflect the mixed heritage of the Australian Merino breed. Locus 2 contains one representative sequence from Scottish Blackface sheep as well as the Australian Merino and appears to be a classical class I (Ia) gene based on expression evidence from Ballingall *et al*. (2008). Sequences (AAA03455, AAA03457 and g4019_B7) cluster close together although the long-branch lengths indicate they could represent independent loci.

Intra-locus homologous recombination is caused by the exchange of very similar sequences in close physical chromosomal proximity (Martinsohn *et al.* 1999; Chen *et al.* 2007). In the present study there were two animals in which evidence of possible recombination was observed. For instance, clone g4020_C11, showed similarity with sequences from IPD Group 6 from the leader sequence but at the α2 domain it manifests similarity with IPD Group 3 and then reverts back to IPD Group 6 similarity midway through the α3 domain. Likewise, clone g4006_B1 has sequence similarity with Group 13/14 from the leader sequence until midway in the α1 domain then changes to be similar to Group 16 and appears to change near the TM junction to become more similar to Group 9 or possibly 13/14. In both cases the patterns observed support an intra-molecular recombination event, such as gene conversion (Martinsohn *et al.* 1999; Chen *et al.* 2007), rather than a simple PCR induced artefact. This is supported by previous description of intra molecular gene conversion for several other genes within the MHC (Roos *et al.* 1984; Campbell *et al.* 1986; Urabe *et al.* 1990; Yu 1991; White *et al.* 1994; Hickford *et al.* 2004; Qin *et al.* 2011).

Chapter 6

MHC Class I Located Sheep *Corneodesmosin*: Gene Structure, Polymorphisms and Association with Phenotype

Sheep Corneodesmosin (CDSN) gene sequence was initially obtained through sub-cloning of BAC CHORI 243-269M18, which also contains other MHC class I genes as described in Chapter 3. This chapter describes the CDSN sequence, variation within the gene and its association with known phenotype. Protein encoded by CDSN is known to be involved in skin and hair (wool) physiology. Primers were designed to completely sequence the CDSN and identify single nucleotide polymorphisms (SNPs) within this gene. Fifty eight SNPs were identified in a fragment of 4579 bp, which includes the entire CDSN sequence and both 3' and 5' untranslated regions. Structural bioinformatics tools were used to analyse the effects of these SNPs on predicted protein structure. Several SNPs, mainly those in coding sequences were used to genotype a population of 107 animals with known univariate and multivariate estimated breeding value (EBV) wool traits. Haplotypes within CDSN associated with both univariate and multivariate EBV for wool traits were identified. Some of the work described herein has been published and is shown in Appendix B (Siva Subramaniam et al. 2011; doi: 10.1111/j.1740-0929.2011.00975.x). An additional manuscript describing information from this chapter is currently in preparation for publication.

6.1 Introduction

In addition to its involvement in immune related functions, the Major Histocompatibility Complex (MHC) also has been shown to contain loci associated with production traits in domestic animals (Schook & Lamony 1996; Vaiman *et al.* 1998). A previous study has shown that clean fleece weight in sheep is associated with microsatellite alleles from two loci located within the MHC (Bot 2000). In this study, microsatellite markers were used

to investigate the influence of the MHC on various traits including clean fleece weight (CFW) and greasy fleece weight (GFW) (Bot 2000). Two alleles at the SKIV2LM and OLADRB microsatellite loci showed significant correlation with higher wool yield, resulting in an approximate increase of 20% in mean CFW (Bot 2000). This study suggested that *S* gene, later designated as *Corneodesmosin* (*CDSN)* gene, because of its skin and wool related functions was a plausible candidate gene to explain this observed association with the wool production trait (Bot 2000). A study using New Zealand Wiltshire sheep, aimed at assessing genes not previously shown to have a role in wool physiology, demonstrated that the desmosome component was involved (Rufaut *et al.* 1999). This study in sheep suggested that the desmosome had a significant influence in the wool follicle growth cycle.

The human *CDSN* gene located 160 kb telomeric of HLA-C (6p21.3), encodes for corneodesmosin, a 52- to 56-kDa basic glycoprotein specific to the cornified epithelia and inner hair follicle root sheath (Zhou & Chaplin 1993; Jonca *et al.* 2002). Corneodesmosin is a component of desmosome, which helps to form specialised intercellular junction in the epidermal layer of the skin and thus maintain the skin integrity (Guerrin *et al.* 1998). Human CDSN is a 529 amino acid long protein with a very high serine and glycine content, 27.5% and 16% respectively, a feature shared with several other epidermal proteins (Guerrin *et al.* 1998; Simon *et al.* 2001; Jonca *et al.* 2002).

Studies have shown that *CDSN* gene polymorphisms are associated with psoriasis (Allen *et al.* 1999; Jenisch *et al.* 1999; Tazi Ahnini *et al.* 1999a; Tazi Ahnini *et al.* 1999b; Schmitt-Egenolf *et al.* 2001; Capon *et al.* 2003; McGrath & Wessagowit 2005; Orru *et al.* 2005; Matsumoto *et al.* 2008) and with an autosomal dominant disorder, hypotrichosis simplex of the scalp (Jonca *et al.* 2002; Levy-Nissenbaum *et al.* 2003b, a). Affected individuals experience gradual loss of the scalp hair starting in the middle of the first decade, resulting in almost complete baldness by the third decade (McGrath & Wessagowit 2005). In three families suffering from hypotrichosis of the scalp, nonsense mutations have been identified in the *CDSN* gene resulting in the accumulation of truncated corneodesmosin aggregates in the

superficial dermis and at the periphery of hair follicles (Levy-Nissenbaum *et al.* 2003a). It has been suggested that the accumulation of abnormal corneodesmosin aggregates associated with protein misfolding is toxic to the hair follicle cells (Kalinin *et al.* 2001; McGrath & Wessagowit 2005).

The aims of this study were to PCR amplify, sequence and annotate the entire sheep *CDSN* gene, and compare the sheep sequence with other species. A small population of animals were analysed to identify SNPs within *CDSN*. These SNPs will be used to genotype larger population of animals with known estimated breeding value for clean fleece weight to identify any association between wool yield and haplotypes within *CDSN*.

6.2 Materials and methods

6.2.1 Comparative Analysis and primer design

The BLAST programme (Altschul *et al.* 1990) was used for comparative analysis of *CDSN* DNA sequences from cattle (NW_001494146), dog (NW_876254), gray mouse lemur (AB480748), gray short-tailed opossum (NW_001581878), horse (NW_001867389), human (NW_001838980) and pig (NW_001886435).

Based on the alignment of cattle *CDSN* gene sequence with the other sequences given above, primers were designed to amplify a series of 500 bp fragments, with an overlap of 100 bp (Table 6.1) spanning *CDSN*. These primers were used for PCR amplification and sequencing. Primers were synthesized by either Geneworks (Australia) or Invitrogen (Australia).

Table 6.1: Details of primers used to amplify sheep CDSN sequence.

Primer set	Primer sequence (5' to 3')	
	Forward primer	**Reverse primer**
CDSN1	CAGGAGCTGCTGTCAGTCAG	TCCACAGCCGTGTCACATGT
CDSN2	CTGATGCTTGAGTCTAGAGG	CTCACACAGTTCTTCTACTG
CDSN3	ACATGTGACACGGCTGTGGA	TCTCCACAATTCGTGCAGTC
CDSN4	TGCTGATTGGTGCAGCGGAG	GCACGCACTGTTGCTTCAGA
CDSN5	GTACTGAGGAAGGTCAGAGG	TGAGATGGCAGCTTATTCTG
CDSN6	GAGATCAGAGTGTCAGATGC	AGTGGGAGCTGGAGATGTAG
CDSN7	CTCTACTCTACAAGGTGCAC	AAGTACTTGCCCTCAGAGAC
CDSN8	GTCTCTGAGGGCAAGTACTT	GACCTCCACTCAATGTCGAG
CDSN9	TTCCTCCTCGACACTGAGTG	AGCATATCTCGCTGGTTGAC
CDSN10	CACTCCAGTGTTCTCGCATG	TCTCACTTGTTGACCAGATG

6.2.2 BAC DNA extraction

The BAC clone (CHORI 243-269M18), used in this study was originally identified to contain *CDSN* by J. Qin/D.Groth (unpublished 2006). The BAC clone was extracted using QIAGEN Large-Construct Kit, following manufacturer's standard protocol (Chapter 2.2.3).

6.2.3 Polymerase chain reaction and sequencing

Primers were used for PCR amplification of BAC DNA and sequencing of the PCR product. PCR was performed using standard protocol (Chapter 2.4). The PCR products from amplified BAC DNA were cleaned with ExoSAP (Biolab) protocol (Chapter 2.5) and electrophoresed (Chapter 2.8). Sequencing reactions were performed at Macrogen Inc. (South Korea).

6.2.4 Sequence analysis and SNP identification

The PCR derived BAC sequences were assembled using VectorNTI (Invitrogen). The resulting sequence was used to design additional primers specific for sheep *CDSN* gene. For amplification of sheep genomic DNA, second pass sequencing primers of 20 bp were designed based upon the assembled first pass BAC contigs. The second pass primers were optimised

to amplify regions of 500 bp to 1000 bp in length. Genomic DNA from twelve Merino sheep samples were independently amplified and sequenced. For each animal, 5 X 10 µL PCR reactions were prepared and the products pooled before sequencing. The resulting sequences were interrogated using VectorNTI software to identify SNPs. Sheep *CDSN* and each of the PCR fragments with potential SNPs were aligned to identify the region and the exact location of the SNPs with respect to the sheep *CDSN* start codon. The sequence was also aligned with data generated by the International Sheep Genome Consortium (ISGC) (https://isgcdata.agresearch.co.nz/). The SNPs identified within *CDSN* coding sequence (CDs) were used to determine putative amino acid (aa) changes. Synonymous and Non-synonymous Analysis Program (SNAP) analysis (www.hiv.lanl.gov and http://hcv.lanl.gov/content/sequence/SNAP/SNAP.html) (Nei & Gojobori 1986; Korber 2000) was performed to determine the rate of synonymous and non-synonymous change in multiple sequences.

6.2.5 Gene prediction and phylogenetic analysis

Gene prediction tools including GENSCAN (Burge & Karlin 1997), FGENESH (http://www.softberry.com), HMM gene (http://www.cbs.dtu.dk/services/HMMgene/), Augustas (Stanke *et al.* 2006) and GeneMark (Lomsadze *et al.* 2005) were used to determine the number of exons, mRNA and amino acid sequence from the DNA sequence generated from the assembly of the contigs by VectorNTI. Basic pairwise alignment, ExPASy (Swiss Institute of Bioinformatics) and GAP (Pairwise Sequence Alignment Server) were also used to analyse the sequence. The resulting structure of the sheep *CDSN* gene was compared with the structure of cattle (NW_001494164), chimpanzee (NW_001236523), dog (NW_876254), gray mouse lemur (AB480748), gray short-tailed opossum (NW_001581878), horse (NW_001867389), human (NW_001838980), mouse (NT_039663), pig (NW_001886435), rat (NW_001084776) and rhesus monkey (NW_001116482). ClustalW (Thompson *et al.* 1994) was used independently for multiple sequence alignment of the sheep *CDSN* with other mammals. Analyses were performed for both mRNA and amino acid sequences. The resulting alignments were used for phylogenetic analyses of the *CDSN*.

6.2.6 Structural bioinformatics analysis

Non-synonymous SNPs were analyzed using automated methods that derive a consensus view on the potential effect of the observed amino acid substitutions on CDSN structure or function. In particular, the web server versions of the iPTree-STAB algorithm (Huang *et al.* 2007) and iMutant (Capriotti *et al.* 2004, 2005; Capriotti *et al.* 2008) were used. Both of these methods use a thermodynamic approach to predict the effect of amino acid substitutions on protein stability. iMutant also uses a binary classification of stability called SVM2 (Support Vector Machine 2) or ternary classification of stability called SVM3 (support Vector Machine 3).

6.2.7 Analysis of internal structure and nucleotide divergence

DNAStrider software (Marck 1988) was used also to determine internal structure or features of *CDSN*. The intron/exon architecture was predicted by dotplot mapping. The polymorphism and divergence of *CDSN* genes was predicted from pairwise or multiple alignments of the *CDSN* coding regions using DNASP software (Librado & Rozas 2009).

6.2.8 SNP-typing within *CDSN*

Primers previously used to amplify exon 1 of *CDSN* (labelled as 1F2 and 1R) and additional primers designed specifically to amplify exon 2 of *CDSN* were used to sequence the two exonic regions in 107 sheep. The univariate estimated breeding value (EBV) for clean fleece weight (CFW) and multivariate EBV for CFW, fibre diameter (FD) and staple strength (SS) are known for this population of 107 animals from the Katanning Baseflocks. Univariate EBV is a single measurement of a trait, whereas multivariate EBV is adjusted for other variables influencing wool production. The EBVs were estimated using a multi-trait mixed model on a population of more than 20,000 individuals born between 1982 up to 2005. Full pedigrees and environmental factors were available on all the animals (107) used in this study. Both univariate and multivariate EBV for wool traits were obtained Dr Johan Greeff from the Department of Agriculture and Food Western Australia (DAFWA), from where the sheep samples were also collected. All 16 SNPs

identified in the coding sequence of *CDSN* and 1 SNP in intron 1 (at 109 bp within *CDSN*) were PCR amplified using standard protocol (Chapter 2.4) and the products subsequently sequenced at Macrogen Inc. (South Korea). The SNP within intron 1 was analysed because the primer set for exon 1 also covered part of the intron in which this SNP was located. Sequences were interrogated using VectorNTI software (Invitrogen) and details of SNPs in each animal were noted. Genepop, a population genetics software package (Raymond & Rousset 1995; Rousset 2008) was used to obtain the basic information for each locus typed in the population. Haplotype frequency and association with phenotype was analysed using the SNPstat program (Sole *et al.* 2006). Associations between *CDSN* haplotypes and univariate EBV for CFW, as well as multivariate EBV for CFW, FD and SS were analysed. The SNPs were also fitted as random factors to a multivariate regression model using ASREML. Factors were dropped from the model using log likelihood ratio tests to determine which factors contribution significantly to the total variation.

Table 6.2: Primers used for SNP-typing coding sequence of *CDSN* gene.

Exon	**Forward (5' to 3')**	**Reverse (5' to 3')**
1	CAGGAGCTGCTGTCAGTCAG	TCCACAGCCGTGTCACATGT
2	ACCTTTTCCACCCCAGACTC	AGATTTTGCCCCCACTGTAG
	TCTCCAGCAGTTCCAGCATT	CATGCGAGAACACTGGAGTG

6.2.9 Genotyping across MHC class I region

Analysis of associations between MHC class I region haplotypes and wool traits was also performed using both univariate and multivariate EBV. The 107 animals were genotyped using the panel of 14 SNPs used for linkage disequilibrium analysis as described in Chapter 4. Although a total of 108 animals were genotyped for SNPs across MHC class I region, only 107 were included in the association study for wool yield because of a missing EBV for CFW. Genotyping was performed by KBioscience using KASPar assay system (http://www.kbioscience.co.uk/). SNPstat program (Sole *et al.* 2006) was

used to analyse haplotypic correlation between genotypic data and EBV for CFW.

6.3 Results

6.3.1 Comparative analysis and gene annotation

An overview of the *CDSN* gene was obtained by comparative analyses of several other mammalian organisms such as dog (NW_876254), gray short-tailed opossum (NW_001581878), gray mouse lemur (AB480748), horse (NW_001867389), human (NW_001838980) and pig (NW_001886435). The basic structure of the *CDSN* gene in all the organisms was similar to the cattle gene with 2 exons and a large intron. Variation in the size of the gene ranged from 3302 bp in pig to 5350 bp in human.

Assembly of triple pass DNA sequencing of BAC and sheep genomic DNA with VectorNTI resulted in a contig fragment of 4579 bp in length. The sheep *CDSN* gene was located between positions 159 bp and 3841 bp within the fragment. The results obtained using various gene prediction tools were compared and a consensus sheep *CDSN* gene sequence generated. Comparative analysis of the sheep *CDSN* gene showed that the gene has two exons of 85 bp and 1553 bp respectively and an intron of 2045 bp. The sheep *CDSN* gene is 3683 bp in length and encodes a protein of 545 amino acids. Nucleotide and predicted peptide sequence of sheep CDSN are available as NCBI GenBank accession number GU591411 and ADD84518.1, respectively.

6.3.2 SNP identification

Sequence analysis resulted in identification of 51 SNPs within the genomic DNA from twelve Merino sheep. Alignment of the 4579 bp sequence with other breed of sheep from 454 data revealed an additional 7 SNPs. Sixteen SNPs identified with sequence analysis were also identified in 454 data. In total, 58 SNPs were identified within the entire fragment. Sixteen SNPs are located within the coding sequence of the *CDSN* gene. The other 30 and 12

SNPs are located within the intron 1 of *CDSN* gene and after the stop codon respectively. The frequency of SNP in sheep *CDSN* is approximately 1 in 80bp. The details of all the SNPs identified are shown in Table 6.3.

Table 6.3: SNPs indentified in within and outside sheep CDSN gene.

Location in the fragment (bp)	**Location within *CDSN* (bp)**	**Base change**	**Description**
241	83	A/G	Exon 1
267	109	C/T	Intron 1
748	590	A/G	Intron 1
823	665	C/T	Intron 1
929 ‡	771	A/G	Intron 1
1029 ‡	871	A/G	Intron 1
1246	1088	G/T	Intron 1
1355	1097	C/G	Intron 1
1389	1231	A/G	Intron 1
1416	1258	C/G	Intron 1
1418	1260	A/C	Intron 1
1440 †	1282	G/C	Intron 1
1526 ‡	1368	C/T	Intron 1
1562 ‡	1404	C/T	Intron 1
1566 ‡	1408	A/G	Intron 1
1600	1442	C/T	Intron 1
1643 ‡	1485	A/G	Intron 1
1681 ‡	1523	C/T	Intron 1
1711	1553	C/T	Intron 1
1753	1595	C/T	Intron 1
1782	1624	C/T	Intron 1
1789	1631	C/T	Intron 1
1799	1641	C/T	Intron 1
1825	1667	G/T	Intron 1
1848	1690	G/T	Intron 1
1978 ‡	1820	G/T	Intron 1
1993 ‡	1835	C/T	Intron 1
2087 ‡	1929	C/T	Intron 1
2097 †	1939	G/T	Intron 1
2149 ‡	1991	C/T	Intron 1
2160 ‡	2002	G/T	Intron 1
2405	2247	A/G	Exon 2

2433 ‡	2275	C/G	Exon 2
2536	2378	C/T	Exon 2
2538	2380	A/G	Exon 2
2539	2381	C/T	Exon 2
2611	2453	C/T	Exon 2
2630 ‡	2472	A/G	Exon 2
2736	2578	A/G	Exon 2
2756 ‡	2598	C/T	Exon 2
2827	2669	A/C	Exon 2
2845 ‡	2687	C/T	Exon 2
2968 †	2810	C/T	Exon 2
3262 †	3104	C/T	Exon 2
3502 †	3344	A/G	Exon 2
3608	3450	A/G	Exon 2
3873	3715	C/T	After stop codon
3957	3799	G/T	After stop codon
4160	4002	C/T	After stop codon
4161	4003	A/G	After stop codon
4174	4016	C/T	After stop codon
4213	4055	C/T	After stop codon
4222	4064	A/G	After stop codon
4303	4145	C/T	After stop codon
4384 †	4226	A/C	After stop codon
4427 †	4269	C/T	After stop codon
4537	4379	A/G	After stop codon
4557	4399	C/T	After stop codon

† SNP identified only in 454 data.
‡ SNP identified in both genomic sheep sequences and ISGC data.

The 16 SNPs identified within the coding sequence of *CDNS* gene would result in eight synonymous and eight nonsynonymous changes. Table 6.4 shows the details of amino acid changes. The statistics of SNAP analysis of pairwise comparisons of *CDSN* for all organisms, which takes into account Jukes-Cantor correction, showed that the average ds/dn value is 1.4426 with ds and dn of 1.9007 and 1.4149 respectively (Table 6.5).

Table 6.4: Location of SNP in coding sequence and the corresponding amino acid changes within sheep CDSN protein sequence.

Location within *CDSN* (bp)	**Location in CDs**	**Change of base**	**Location in AAs**	**Details of AA substitution**		**Type of mutation**
				Change of AA	**AA classification**	
83	83	A/G	28	Glutamine (Q)	Polar, neutral	Non-synonymous
				Arginine (R)	Polar, positively charged	
2247	202	A/G	68	Serine (S)	Polar, neutral	Non-synonymous
				Glycine (G)	Nonpolar	
2275	230	C/G	77	Serine (S)	Polar, neutral	Non-synonymous
				Threonine (T)	Polar, neutral	
2378	333	C/T	111	Glycine (G)	Nonpolar	Synonymous
				Glycine (G)	Nonpolar	
2380	335	A/G	112	Histidine (H)	Polar, positively charged	Non-synonymous
				Arginine (R)	Polar, positively charged	
2381	336	C/T	112	Histidine (H)	Polar, positively charged	Synonymous
				Histidine (H)	Polar, positively charged	
				Arginine (R)	Polar, positively charged	Synonymous
				Arginine (R)	Polar, positively charged	
2453	408	C/T	136	Glycine (G)	Nonpolar	Synonymous
				Glycine (G)	Nonpolar	
2472	427	A/G	143	Glycine (G)	Nonpolar	Non-synonymous
				Serine (S)	Polar, neutral	
2578	533	A/G	178	Aspartic acid (D)	Polar, negatively charged	Non-synonymous
				Glycine (G)	Nonpolar	

2598	553	C/T	185	Proline (P)	Nonpolar	Non-synonymous
				Serine (S)	Polar, neutral	
2669	624	A/C	208	Threonine (T)	Polar, neutral	Synonymous
				Threonine (T)	Polar, neutral	
2687	642	C/T	214	Serine (S)	Polar, neutral	Synonymous
				Serine (S)	Polar, neutral	
2810	765 †	C/T	255	Serine (S)	Polar, neutral	Synonymous
				Serine (S)	Polar, neutral	
3104	1059 †	C/T	353	Serine (S)	Polar, neutral	Synonymous
				Serine (S)	Polar, neutral	
3344	1299 †	A/G	433	Glycine (G)	Nonpolar	Synonymous
				Glycine (G)	Nonpolar	
3450	1405	A/G	469	Glycine (G)	Nonpolar	Non-synonymous
				Serine (S)	Polar, neutral	

† SNP identified only in ISGC data.

Table 6.5: Result of Synonymous Non-synonymous Analysis Program (SNAP). Averages of all pairwise comparisons: ds = 1.9007, dn = 1.4149, ds/dn = 1.4426, ps/pn = 1.1720.

Sequences Names		ds	dn	ds/dn
OVAR	BOTA	1.1076	1.1841	0.9354
OVAR	EQCA	1.4732	1.5144	0.9728
OVAR	MAMU	2.0765	1.5056	1.3792
OVAR	SUSC	1.984	1.391	1.4263
OVAR	HOSA	1.8536	1.6309	1.1365
OVAR	PATR	1.7946	1.6321	1.0996
OVAR	CAFA	2.2992	1.7004	1.3521
OVAR	MODO	2.2017	1.9844	1.1095
OVAR	MUMU	1.5885	1.4703	1.0805
OVAR	MIMU	1.6718	1.3711	1.2193
OVAR	RANO	1.671	1.3729	1.2172
BOTA	EQCA	1.53	1.2663	1.2082
BOTA	MAMU	2.0873	1.4799	1.4105
BOTA	SUSC	1.5163	1.285	1.18
BOTA	HOSA	1.8427	1.3146	1.4017
BOTA	PATR	1.8109	1.3203	1.3717
BOTA	CAFA	3.6985	1.6305	2.2682
BOTA	MODO	2.3111	1.5932	1.4506
BOTA	MUMU	1.3355	0.9675	1.3805
BOTA	MIMU	1.7758	1.3487	1.3167
BOTA	RANO	1.7	1.4353	1.1844
EQCA	MAMU	1.6173	1.3262	1.2195
EQCA	SUSC	1.6136	1.3228	1.2199
EQCA	HOSA	1.5182	1.3921	1.0906
EQCA	PATR	1.515	1.3878	1.0916
EQCA	CAFA	3.5077	1.5617	2.246
EQCA	MODO	2.3508	2.0706	1.1353
EQCA	MUMU	1.3176	1.4005	0.9408
EQCA	MIMU	2.6972	1.5811	1.7059
EQCA	RANO	1.8381	1.5278	1.2031
MAMU	SUSC	1.5028	1.2045	1.2476
MAMU	HOSA	1.1702	1.2806	0.9138
MAMU	PATR	1.1543	1.2867	0.8971
MAMU	CAFA	NA	0	NA
MAMU	MODO	4.8187	1.9108	2.5218
MAMU	MUMU	1.5814	1.4449	1.0945
MAMU	MIMU	1.3354	1.2007	1.1122

MAMU	RANO	1.372	1.4907	0.9204
SUSC	HOSA	1.9989	1.1007	1.816
SUSC	PATR	1.976	1.1032	1.7912
SUSC	CAFA	1.8143	1.3052	1.3901
SUSC	MODO	NA	0	NA
SUSC	MUMU	NA	0	NA
SUSC	MIMU	1.1025	0.7883	1.3986
SUSC	RANO	2.0181	1.4062	1.4352
HOSA	PATR	0.0176	0.0024	7.2349
HOSA	CAFA	2.6581	1.4463	1.8379
HOSA	MODO	NA	0	NA
HOSA	MUMU	1.788	1.3917	1.2848
HOSA	MIMU	1.594	1.3927	1.1446
HOSA	RANO	1.6418	1.3874	1.1833
PATR	CAFA	2.505	1.4463	1.732
PATR	MODO	NA	0	NA
PATR	MUMU	1.7972	1.4005	1.2832
PATR	MIMU	1.5872	1.3997	1.1339
PATR	RANO	1.6057	1.3909	1.1544
CAFA	MODO	NA	0	NA
CAFA	MUMU	NA	0	NA
CAFA	MIMU	1.4204	1.3306	1.0675
CAFA	RANO	3.8809	1.3089	2.9649
MODO	MUMU	2.8023	2.3301	1.2027
MODO	MIMU	NA	0	NA
MODO	RANO	2.1204	1.9964	1.0621
MUMU	MIMU	2.0328	1.455	1.3971
MUMU	RANO	1.389	1.3649	1.0176
MUMU	RANO	2.2525	1.5262	1.4759

ps: The proportion of observed synonymous substitutions.
pn: The proportion of observed non-synonymous substitutions.
ds: The Jukes-Cantor correction for multiple hits of ps.
dn: The Jukes-Cantor correction for multiple hits of pn.
ds/dn: The ratio of synonymous to non-synonymous substitutions.

6.3.3 Phylogenetic analysis

Multiple alignments of amino acid sequences showed that the sheep *CDSN* gene has 92% DNA sequence identity with the cattle gene in comparison to 69% observed with that of human. The percent identity of amino acid sequence in the sheep gene compared to other species ranges from 58% in gray short-tailed opossum to 84% in pig. Table 6.6 shows the details of percent identity of sheep CDSN amino acid sequence compared with other species. The phylogenetic tree (Figure 6.1) generated from multiple alignment of CDSN amino acid sequence from various species showed significant identity and confidence level between sheep and cattle. The overall topology of the tree showed 3 major clades where the CDSN sequence from various organisms has been grouped; primates, rodents and ruminant with other higher order mammals.

Table 6.6: Percent identity of sheep CDSN amino acid sequence compared with other species, created with ClustalX. OVAR: sheep CDSN, BOTA: cattle (NW_001494164), PATR: chimpanzee (NW_001236523), CAFA: dog (NW_876254), MIMU: gray mouse lemur (AB480748), MODO: gray short-tailed opossum (NW_001581878), EQCA: horse (NW_001867389), HOSA: human (NW_001838980), MUMU: mouse (NT_039663), SUSC: pig (NW_001886435), RANO: rat (NW_001084776) and MAMU: rhesus monkey (NW_001116482).

	HOSA	**PATR**	**MAMU**	**EQCA**	**CAFA**	**OVAR**	**BOTA**	**SUSC**	**MIMU**	**MUMU**	**RANO**	**MODO**
HOSA	100	99	93	79	80	69	69	75	78	66	66	59
PATR	99	100	94	80	80	69	69	75	78	66	66	59
MAMU	93	94	100	80	82	69	69	76	77	66	67	59
EQCA	79	80	80	100	87	71	72	81	77	67	67	58
CAFA	80	80	82	87	100	75	76	82	81	69	66	60
OVAR	69	69	69	71	75	100	92	84	69	61	59	58
BOTA	69	69	69	72	76	92	100	85	70	62	60	59
SUSC	75	75	76	81	82	84	85	100	75	64	64	60
MIMU	78	78	77	77	81	69	70	75	100	66	66	59
MUMU	66	66	66	67	69	61	62	64	66	100	89	55
RANO	66	66	67	67	66	59	60	64	66	89	100	54
MODO	59	59	59	58	60	58	59	60	59	55	54	100

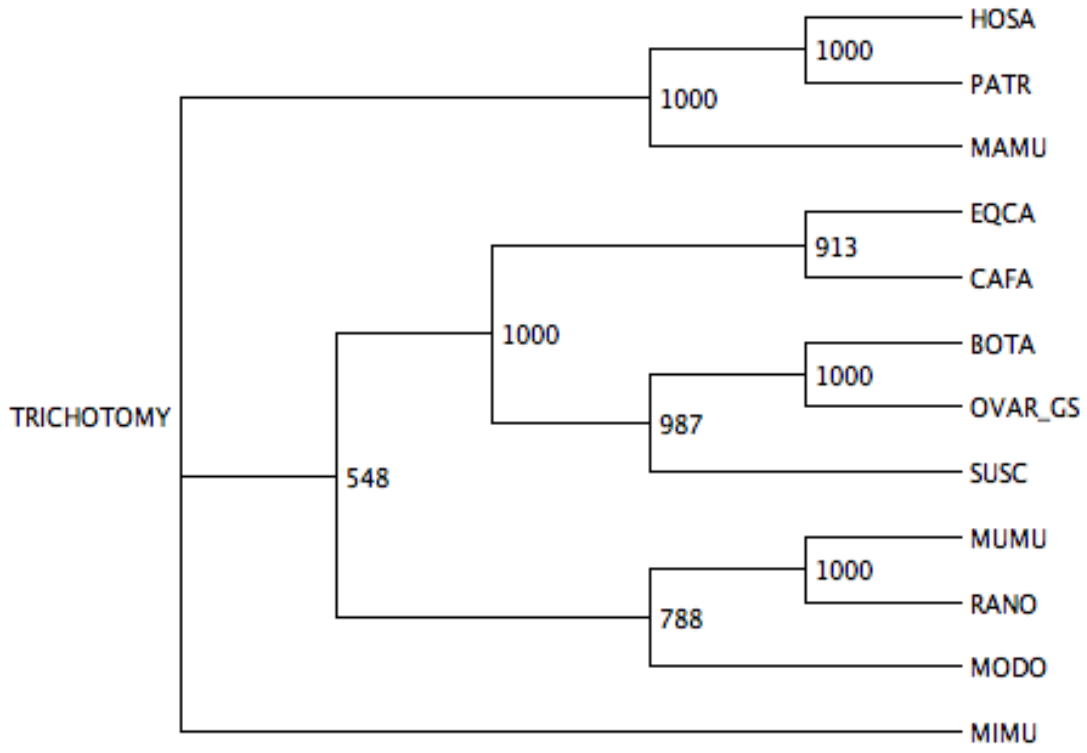


Figure 6.1: Neighbor-Joining Tree constructed using ClustalX after aligning amino acid sequences (default parameter settings). 1000 bootstraps. OVAR_GS=Genscan prediction for sheep CDSN. BOTA: cattle (NW_001494164), PATR: chimpanzee (NW_001236523), CAFA: dog (NW_876254), MIMU: gray mouse lemur (AB480748), MODO: gray short-tailed opossum (NW_001581878), EQCA: horse (NW_001867389), HOSA: human (NW_001838980), MUMU: mouse (NT_039663), SUSC: pig (NW_001886435), RANO: rat (NW_001084776) and MAMU: rhesus monkey (NW_001116482).

6.3.4 Structural bioinformatics analysis

All except three (S68G, D178G, and P185S) of the non-synonymous SNPs seen in this protein are generally conservative amino acid substitutions according to their physicochemical properties and the Blosum62 evolutionary matrix (Henikoff & Henikoff 1992). The thermodynamic approach of the iPTree-STAB algorithm shows that all, except the substitutions G143S and G469S were destabilising to protein structure (Table 6.7). Table 6.7 also shows that according to SVM3, most substitutions, except D178G and G469S, were neutral to the stability of protein structure. However, according to the SVM2 classification all substitutions appeared to be destabilizing to the protein structure. The G469S substitution was classified by the iMutant algorithm as a potential 'disease-related mutation' based on comparisons with other genes associated with disease. A diagrammatic comparison of the SNP in the coding sequence between human and sheep is shown in Figure 6.2.

Table 6.7: Thermodynamic Stability of Protein with indicated substitution.

Method	**iPTree-STAB**		**iMutant data**		
Substitution	**PredictedΔΔG kcal/mol**	**Predicted Stabilising/ Destabilising**	**Predicted ΔΔG kcal/mol**	**Predicted (SVM3) Stabilising/ Destabilising/ Neutral**	**Predicted (SVM2) Stabilising/ Destabilising**
Q28R	-0.0458	Destabilising	-0.04	Neutral	Destabilising
S68G	-1.96	Destabilising	-0.64	Neutral	Destabilising
S77T	-1.1936	Destabilising	-0.59	Neutral	Destabilising
H112R	-1.1814	Destabilising	-0.29	Neutral	Destabilising
G143S	-1.0663	Stabilising	-0.72	Neutral	Destabilising
D178G	-0.0163	Destabilising	-1.21	Destabilising	Destabilising
P185S	-1.0663	Destabilising	-1.33	Neutral	Destabilising
G469S	-0.4685	Stabilising	-0.93	Destabilising	Destabilising

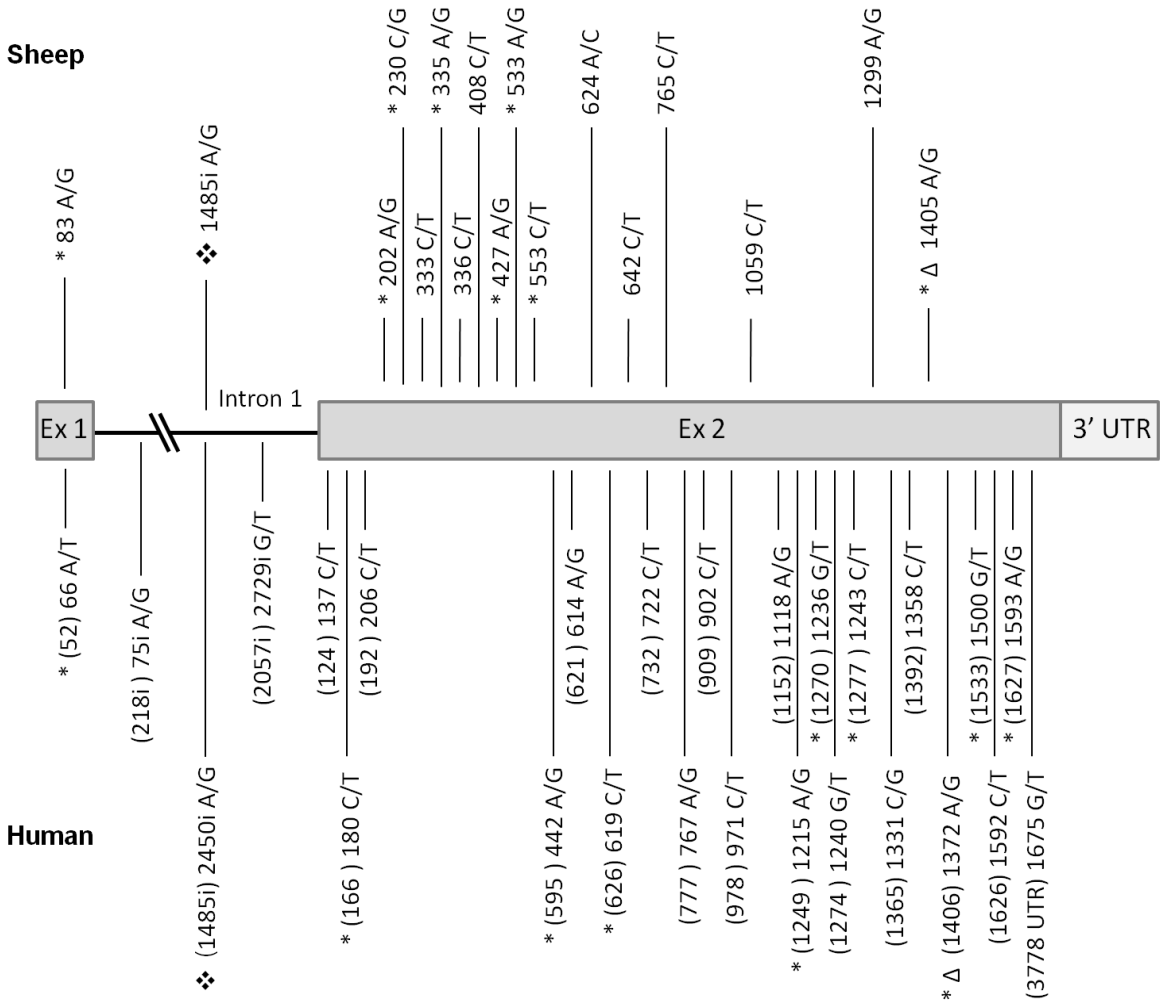


Figure 6.2: SNPs identified within sheep (this study) and human *CDSN* (Guerrin *et al.* 2001). * Non-synonymous substitution. ❖ Identical SNP identified in both human and sheep sequence. ( ) Relative location of human SNP in the sheep coding sequence. Δ SNP located next to each other in coding sequence that effect the same amino acid. Location of SNPs in the intron region is based on the genomic sequence.

6.3.5 Internal structure and nucleotide divergence

The internal structure of the Ovar CDSN gene was predicted using dot plots generated with DNAStrider software (Marck 1988). A plot of the CDSN coding sequence (1638 bp) versus the genomic CDSN sequence is shown in Figure 6.3. It can be seen that the gene contains two exons. Exon 1 is a short sequence of ≈100 bp that encodes a signal peptide. A long Exon 2 (≈1600 bp) is separated from exon 1 by an intron of ≈1900 bp. Within exon 2 a subregion of ≈240 bp (80 aa) is present that manifests a tandemly repetitive structure exhibiting an excess of glycine and serine residues.

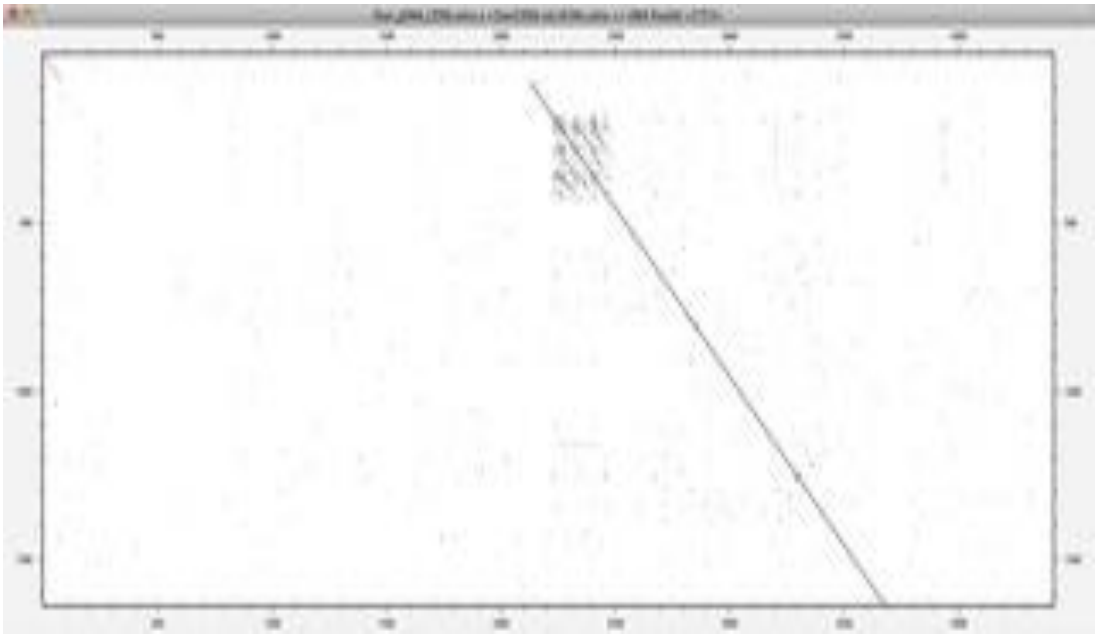


Figure 6.3: Dotplot of *CDSN* coding sequence versus genomic sequence showing the short exon 1 and the long exon 2. Within exon 2 a region manifesting tandemly repeating internal units is present close to the amino terminal end of the protein.

Multiple sequence alignments of CDSN peptides from Ovar, Hosa and Bota showed several regions of interest (Figure 6.4 and Figure 6.5). These included three regions with a predominance of glycine and serine residues. It is believed that these subregions are targets for the progressive protease sensitivity that characterises the CDSN protein (Jonca *et al.* 2010).

Pairwise alignment of the coding regions for Ovar and Hosa *CDSN* were compared to investigate nucleotide sequence divergence and to determine if these regions were targets for selection. DNASP (Librado & Rozas 2009) estimated the nucleotide diversity (Pi) for synonymous sites/synonymous substitution and nonsynonymous sites/nonsynonymous substitution within CDSN. From these the ratios of the number of nonsynonymous substitutions per nonsynonymous site (Ka), and the number of synonymous substitutions per synonymous site (Ks), corrected for saturation using the Jukes Cantor model, were estimated using the pairwise alignment and the method of Nei and Gojobori (1986) as implemented in DNASP. The results are presented as a moving average along the length of the sequence as shown in Figure 6.4. The first exon encoding a conserved signal peptide manifested low Ka/Ks ratios. Elevated Ka/Ks ratios (>1.0) were present in four regions at nucleotide positions ≈145, ≈529, ≈1153, and ≈ 1393. These elevated Ka/Ks ratios suggest that natural selection was responsible for the divergence observed within these regions. In view of these results, it was decided to compare nucleotide diversity between the exon 2 sequences of Ovar and the more closely related Bota (Figure 6.5). This analysis showed a similar pattern of nucleotide diversity. However, as expected, the Ka/Ks ratios observed were smaller than for the comparison between the more distantly related Ovar/Hosa exon 2s. It seems clear that the *CDSN* gene in mammals is subject to strong selection perhaps reflecting the diversity of tissues or skin types protecting these species from environmental challenges.

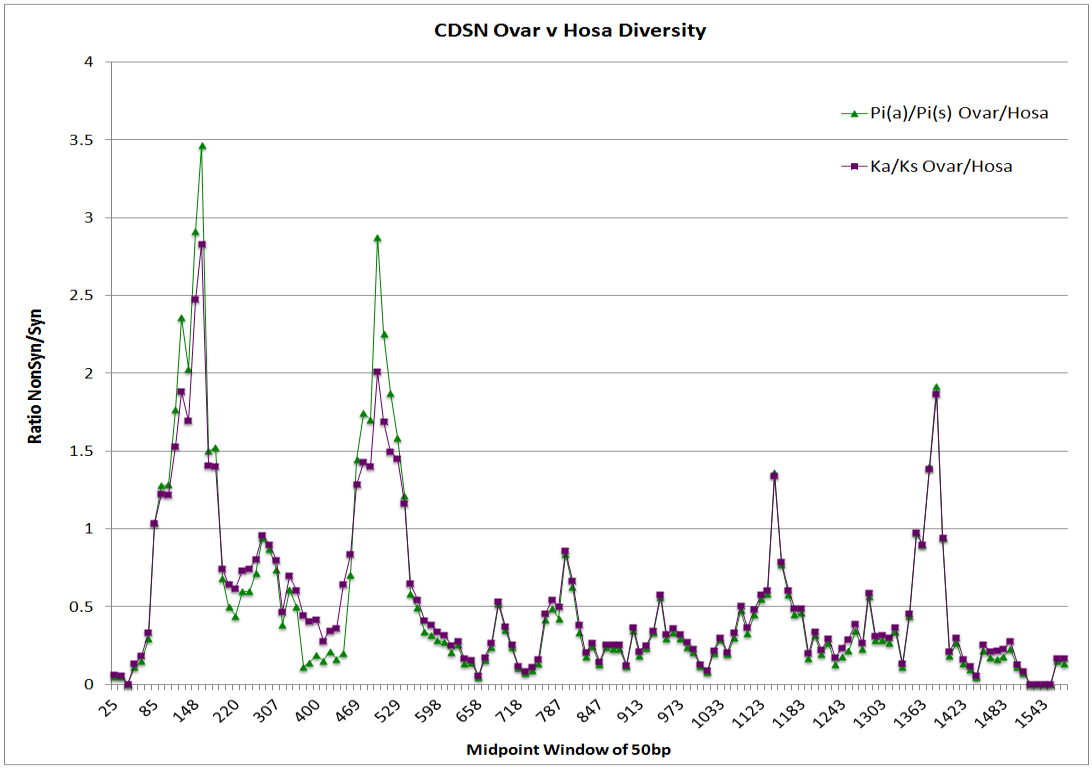


Figure 6.4: DNASP result for multiple sequence alignment of *CDSN* coding region for Ovar versus Hosa.

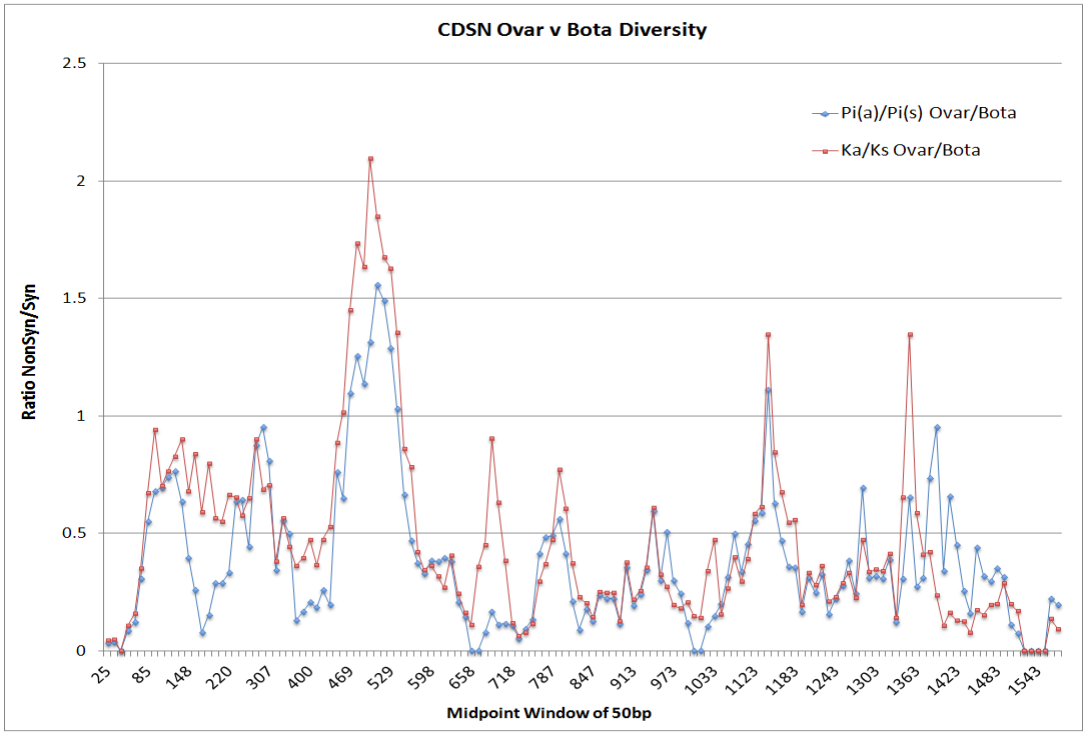


Figure 6.5: DNASP result for multiple sequence alignment of *CDSN* coding region for Ovar versus Bota.

6.3.6 Associations between *CDSN* haplotypes and univariate EBV

Analysis of the 17 SNPs in 107 animals with known univariate EBV for CFW showed that 16 SNPs were polymorphic apart from the first SNP in the coding sequence (CD1), which appeared to be monomorphic in these animals. The genotypes for each SNP locus typed in this population are shown in Appendix C. Allele frequency, observed and expected homozygosity, Fis estimates and p-values for Hardy-Weinberg equilibrium (HWE) estimated using Genepop are summarised in Table 6.8. Six SNPs (CD2, CD3, CD7, CD8, CD10 and CD16) showed significant deviation from HWE. SNPstats analysis was used to identify any associations between SNPs and phenotype of animals typed and the results are shown in Table 6.9. A graphical representation of each haplotype with its 95% confidence level is shown in Figure 6.6. Excluding undefined rare haplotypes, which accounted for approximately 9% of the haplotypes observed within this population, 10 major haplotypes were identified with varying frequencies and accounted for 91% of all haplotypes observed. Haplotype 1, 2, 3 and 4 have combined frequency of 73%. Haplotype 3, which has frequency of 17%, showed a significant increase in wool yield relative to the most common estimated haplotype (haplotype 1). The univariate EBV of CFW for 95% of animals with haplotype 3 ranged between 0.11 and 0.58. No other haplotype had significant association with wool yield.

Table 6.8: Basic data for each locus in the population. SNPs not in Hardy-Weinberg equilibrium are highlighted in red.

SNP	Location within CDs (bp)	Samples typed	Frequency		Homozygosity		Heterozygosity		Fis estimates		Hardy-Weinberg P-value
			Allele 1	Allele 2	Observed	Expected	Observed	Expected	W & C	R & H	
CD1	83	107	1.000	0.000	107	107.0	0	0.0	0.0000	0.0000	Monomorphic
INT1	109 *	107	0.879	0.121	81	84.1	26	22.9	-0.1337	-0.1342	0.36
CD2	202	106	0.693	0.307	75	60.7	31	45.2	0.3165	0.3184	**0.0024**
CD3	230	104	0.255	0.745	73	64.3	31	40.0	0.2197	0.2210	**0.037**
CD4	333	107	0.005	0.995	106	106.0	1	1.0	0.0000	0.0000	1
CD5	335	107	0.575	0.425	50	54.5	57	52.5	-0.0851	-0.0855	0.55
CD6	336	107	0.855	0.145	78	80.4	29	26.6	-0.0893	-0.0897	0.21
CD7	408	107	0.850	0.150	89	79.7	18	27.3	0.3427	0.3449	**0.0019**
CD8	427	107	0.308	0.692	75	61.1	32	45.9	0.3032	0.3051	**0.0028**
CD9	533	107	0.986	0.014	104	104.0	3	3.0	-0.0095	-0.0096	1
CD10	553	107	0.687	0.313	74	60.8	33	46.2	0.2873	0.2890	**0.0061**
CD11	624	107	0.262	0.738	65	65.5	42	41.5	-0.0111	-0.0112	1
CD12	642	107	0.981	0.019	103	103.1	4	3.9	-0.0144	-0.0144	1
CD13	765	107	0.411	0.589	53	55.0	54	52.1	-0.0375	-0.0377	0.84
CD14	1059	107	0.136	0.864	78	81.8	29	25.2	-0.1522	-0.1528	0.21
CD15	1299	107	0.150	0.850	75	79.7	32	27.3	-0.1713	-0.1719	0.12
CD16	1405	107	0.248	0.752	84	66.9	23	40.1	0.4270	0.4299	**<0.0001**

* Location of SNP within *CDSN* instead of coding sequence as the SNP is located in intronic region.

Table 6.9: Haplotype association between SNPs within *CDSN* and phenotype. Haplotype that shows an association is highlighted in red.

	CD 1	INT 1	CD 2	CD 3	CD 4	CD 5	CD 6	CD 7	CD 8	CD 9	CD 10	CD 11	CD 12	CD 13	CD 14	CD 15	CD 16	Frequency	Difference (95% CI)	P-value
1	A	C	G	G	T	G	C	C	A	A	T	A	C	C	T	G	G	0.2199	0.00	---
2	A	C	A	C	T	A	C	C	G	A	C	C	C	T	T	G	G	0.2197	-0.02 (-0.25 - 0.21)	0.87
3	A	C	A	G	T	A	C	C	G	A	C	C	C	T	T	G	G	0.1772	0.35 (0.11 - 0.58)	**0.0043**
4	A	T	A	G	T	A	C	T	G	A	C	C	C	T	T	G	A	0.1074	0.09 (-0.21 - 0.39)	0.57
5	A	C	A	G	T	A	C	C	G	A	C	C	C	T	T	G	A	0.0563	-0.03 (-0.37 - 0.31)	0.86
6	A	C	G	G	T	G	C	C	A	A	T	A	C	C	T	G	A	0.0371	0.21 (-0.26 - 0.68)	0.38
7	A	C	G	G	T	G	T	C	A	A	T	C	C	C	C	A	G	0.028	0.11 (-0.38 - 0.61)	0.65
8	A	C	A	G	T	G	T	C	G	A	C	C	C	C	C	A	G	0.0234	-0.24 (-0.73 - 0.25)	0.34
9	A	C	A	C	T	G	T	C	G	A	C	C	C	C	C	A	G	0.0187	0.41 (-0.12 - 0.95)	0.13
10	A	C	A	G	T	G	T	T	G	A	C	C	C	C	C	A	A	0.0187	0.1 (-0.49 - 0.69)	0.74
Rare	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0.0936	0.19 (-0.09 - 0.46)	0.19

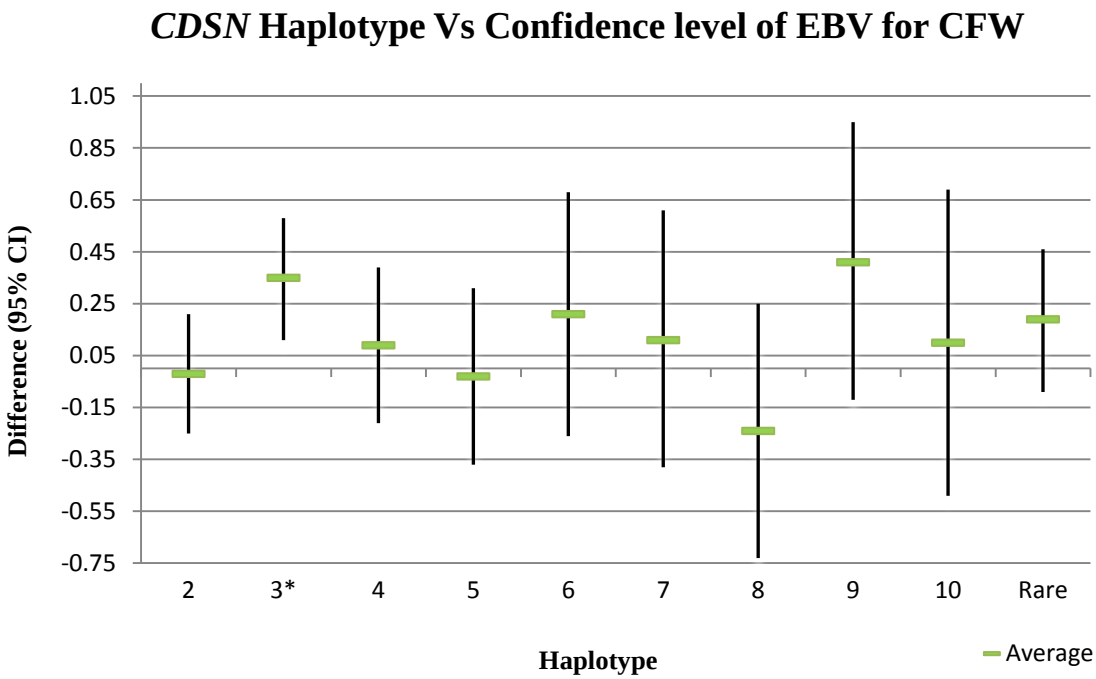


Figure 6.6: Haplotypes within *CDSN* gene and confidence level. All haplotypes were plotted relative to the most common haplotype (haplotype 1) which forms the baseline 0. * Haplotype with significant association.

6.3.7 Analysis of MHC class I SNPs and univariate EBV for CFW

Single-SNP SNPstats analysis for each SNP that span the MHC class I region showed no highly significant associations with wool yield (Table 6.10). However, SNP SN25_1 showed a weak but significant association with univariate EBV for CFW, with a p-value of 0.043. SNP labelled as CDSN_1 for genotyping across MHC class I region is the same SNP identified as INT1 for genotyping SNPs within *CDSN*. This SNP has C to T change and is located at intron 1 at 109 bp within *CDSN*. The location of each SNP used for genotyping, relative to telomeric and centromeric ends and with each other is shown in Figure 6.7.

Table 6.10: Analysis of individual SNP with univariate EBV for CFW. SNP that shows an association is highlighted in red.

SNP	Difference (95% CI)	P-value
SN21_1	0.04 (-0.15 - 0.23)	0.68
SN6_1	0.06 (-0.19 - 0.31)	0.66
SN20_1	0.02 (-0.12 - 0.16)	0.78
SN17_2	0.00 (-0.15 - 0.15)	0.98
SN29_1	0.00 (-0.14 - 0.14)	0.99
SN15_1	-0.14 (-0.30 - 0.01)	0.075
NRM_1	-0.04 (-0.17 - 0.10)	0.61
SN43A_2	-0.08 (-0.33 - 0.16)	0.5
SN42_2	0.17 (-0.10 - 0.44)	0.21
SN41_4	0.03 (-0.12 - 0.19)	0.67
SN25_1	**-0.16 (-0.31 - -0.01)**	**0.043**
SN5_1	0.03 (-0.16 - 0.23)	0.72
CDSN_1	0.05 (-0.18 - 0.28)	0.67
SN3A_1	0.09 (-0.43 - 0.62)	0.73

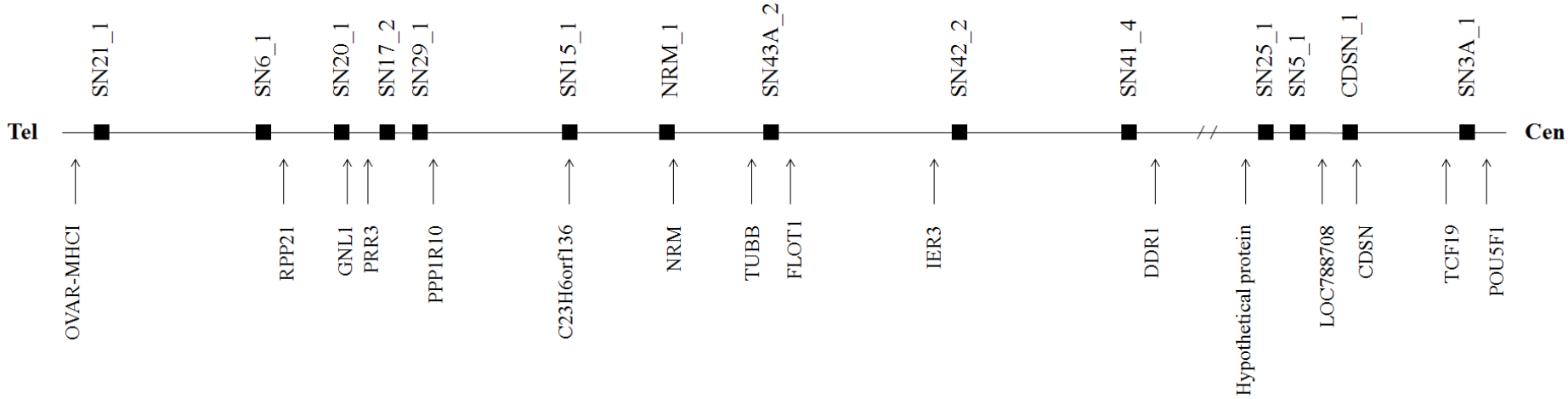


Figure 6.7: Position of each SNP within MHC class I region. The position of each SNP was deduced based on comparative analysis with cattle reference genome sequence NW_001494164. Telomeric end is indicated as Tel and centromeric as Cen.

6.3.8 Associations between MHC class I haplotypes and univariate EBV

Haplotypic analysis of SNPs that span across MHC class I region showed that there were 34 haplotypes with total frequency of approximately 90% (Table 6.11). Frequency of rare haplotypes within the population was approximately 10%. The most common haplotype (haplotype 1) accounts for approximately 8% within the population. Haplotype 5 and 6, which account for a combined frequency of approximately 7% of the observed population haplotypes showed an increase in wool yield relative to haplotype 1. The p-value of haplotype 5 and 6 was 0.031 and 0.029 respectively. Figure 6.8 shows the graphical representation of 95% confidence level for each haplotype. Short sub-sets of haplotypes were also analysed to identify the respective regions for associations with univariate EBV for CFW (Table 6.12). Analysis of the 5 SNPs (SN21_1, SN6_1, SN20_1, SN17_2 and SN29_1) located toward the telomeric end of MHC class I region (close to extended class I) showed that haplotype 7 had significant association whereas, haplotype 9 had weak association with wool yield relative to the most common haplotype (Table 6.12). Haplotype 7 and 9 had a frequency of approximately 5% and 3% respectively. Haplotype 7 had 95% confidence value of -0.56 and p-value of 0.003. Haplotype 9 had 95% confidence value of -0.48 and p-value of 0.033. Haplotype analysis of sub-set in the middle of MHC class I region (SN5_1, NRM_1, SN43A_2, SN42_2 and SN41_4) showed none of the haplotypes had significant association with wool yield except for haplotype 3, which had showed weak association with p-value of 0.033 (Table 6.12). Analysis of the haplotype sub-set derived from the centromeric end of MHC class I region (SN41_4, SN25_1, SN5_1, CDSN_1 and SN3A_1), which is closer to the MHC class III region, showed that 4 haplotypes had weak albeit significant associations with the univariate EBV for CFW (Table 6.12). Haplotypes 3, 7, 8 and 9 had p-value of 0.022, 0.02. 0.018 and 0.025 respectively.

Table 6.11: Haplotypic analysis of SNPs spread across MHC class I region and univariate EBV for CFW. Haplotypes that show an association are highlighted in red.

Haplotype association with univariate EBV for CFW (n=107)																	
	SN21_1	**SN6_1**	**SN20_1**	**SN17_2**	**SN29_1**	**SN15_1**	**NRM_1**	**SN43A_2**	**SN42_2**	**SN41_4**	**SN25_1**	**SN5_1**	**CDSN_1**	**SN3A_1**	**Frequency**	**Difference (95% CI)**	**P-value**
1	G	T	T	A	T	T	G	C	C	G	C	T	C	C	0.0783	0.00	-
2	G	T	G	A	T	T	A	C	C	G	C	T	C	C	0.0656	0.24 (-0.15 - 0.63)	0.23
3	G	T	G	G	C	T	G	C	C	A	T	T	C	C	0.0509	0.18 (-0.26 - 0.62)	0.42
4	A	T	T	G	C	C	G	C	C	A	C	T	C	C	0.0462	-0.2 (-0.65 - 0.25)	0.39
5	G	T	T	G	C	T	G	C	C	A	T	T	C	C	0.0417	**0.43 (0.04 - 0.83)**	**0.031**
6	G	T	G	A	T	T	G	C	C	G	T	T	C	C	0.0362	**0.53 (0.06 - 1)**	**0.029**
7	G	C	G	A	T	T	G	C	C	G	T	T	T	C	0.0355	0.39 (-0.11 - 0.89)	0.13
8	G	T	G	G	C	T	A	C	C	G	C	T	C	C	0.0327	-0.19 (-0.71 - 0.32)	0.47
9	G	T	T	A	T	T	A	C	C	A	C	C	C	C	0.0324	0.29 (-0.26 - 0.84)	0.31
10	G	T	G	G	C	T	G	C	C	G	C	T	C	C	0.0322	-0.02 (-0.56 - 0.52)	0.94
11	G	T	G	G	C	T	G	C	C	G	T	T	C	C	0.0311	0.03 (-0.51 - 0.57)	0.91
12	A	T	T	A	C	C	G	C	C	G	C	T	C	C	0.0278	0.03 (-0.51 - 0.56)	0.93
13	A	T	G	A	T	T	G	C	C	G	T	T	C	C	0.0262	0.27 (-0.35 - 0.89)	0.4
14	G	C	G	G	C	T	G	C	C	G	T	T	C	C	0.0237	-0.26 (-0.86 - 0.35)	0.41
15	A	T	T	A	C	C	A	C	C	G	C	T	C	C	0.0231	0.12 (-0.45 - 0.69)	0.67
16	G	T	G	G	C	C	A	C	T	G	T	T	C	C	0.0231	0 (-0.59 - 0.58)	0.99
17	G	T	G	A	T	T	A	C	C	G	C	T	T	C	0.0199	-0.44 (-1.05 - 0.16)	0.15
18	G	T	G	A	T	T	G	C	T	G	T	T	C	C	0.0196	0.01 (-0.63 - 0.65)	0.97
19	G	T	G	A	T	T	G	C	C	G	C	C	C	C	0.0195	-0.03 (-0.66 - 0.61)	0.94
20	G	T	G	G	C	T	A	C	C	G	C	C	C	C	0.0194	-0.31 (-0.94 - 0.33)	0.35
21	A	T	G	G	C	T	G	C	C	G	C	T	C	C	0.0189	0.16 (-0.5 - 0.81)	0.64
22	A	T	T	A	T	C	A	T	C	G	T	T	T	C	0.0185	0.06 (-0.55 - 0.67)	0.84
23	G	T	G	A	T	T	A	T	C	G	T	T	C	C	0.0185	0.1 (-0.57 - 0.76)	0.77

Table 6.11 (Continued)

Haplotype association with univariate EBV for CFW (n=107)																	
	SN21_1	**SN6_1**	**SN20_1**	**SN17_2**	**SN29_1**	**SN15_1**	**NRM_1**	**SN43A_2**	**SN42_2**	**SN41_4**	**SN25_1**	**SN5_1**	**CDSN_1**	**SN3A_1**	**Frequency**	**Difference (95% CI)**	**P-value**
24	G	T	T	G	C	C	G	C	C	G	T	T	C	C	0.0185	-0.21 (-0.89 - 0.48)	0.56
25	G	T	T	G	C	T	A	T	C	G	T	C	C	C	0.0185	0.02 (-0.62 - 0.67)	0.94
26	G	T	T	A	T	C	G	C	C	A	T	T	C	C	0.014	-0.16 (-0.9 - 0.58)	0.68
27	A	T	T	A	C	C	A	C	C	G	T	T	T	C	0.0139	0.13 (-0.56 - 0.83)	0.7
28	A	T	T	A	C	T	A	C	C	G	T	T	T	C	0.0139	0.3 (-0.41 –1.01)	0.41
29	G	T	G	G	C	T	G	C	T	G	T	C	C	C	0.0128	0.42 (-0.35 - 1.19)	0.28
30	G	T	G	A	C	C	G	C	C	A	T	C	C	C	0.0118	-0.17 (-0.9 - 0.57)	0.66
31	G	T	G	A	C	C	G	C	C	A	C	C	C	C	0.0102	-0.77 (-1.57 - 0.02)	0.06
32	G	T	G	G	C	T	G	C	T	G	T	C	C	A	0.0096	0 (-0.9 - 0.89)	0.99
33	G	T	G	A	T	T	A	T	C	G	C	T	C	C	0.0096	-0.11 (-0.98 - 0.75)	0.8
34	G	T	G	G	C	C	A	C	C	G	T	T	C	C	0.007	0.64 (-0.29 - 1.58)	0.18
Rare	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0.1192	0.02 (-0.39 - 0.42)	0.94

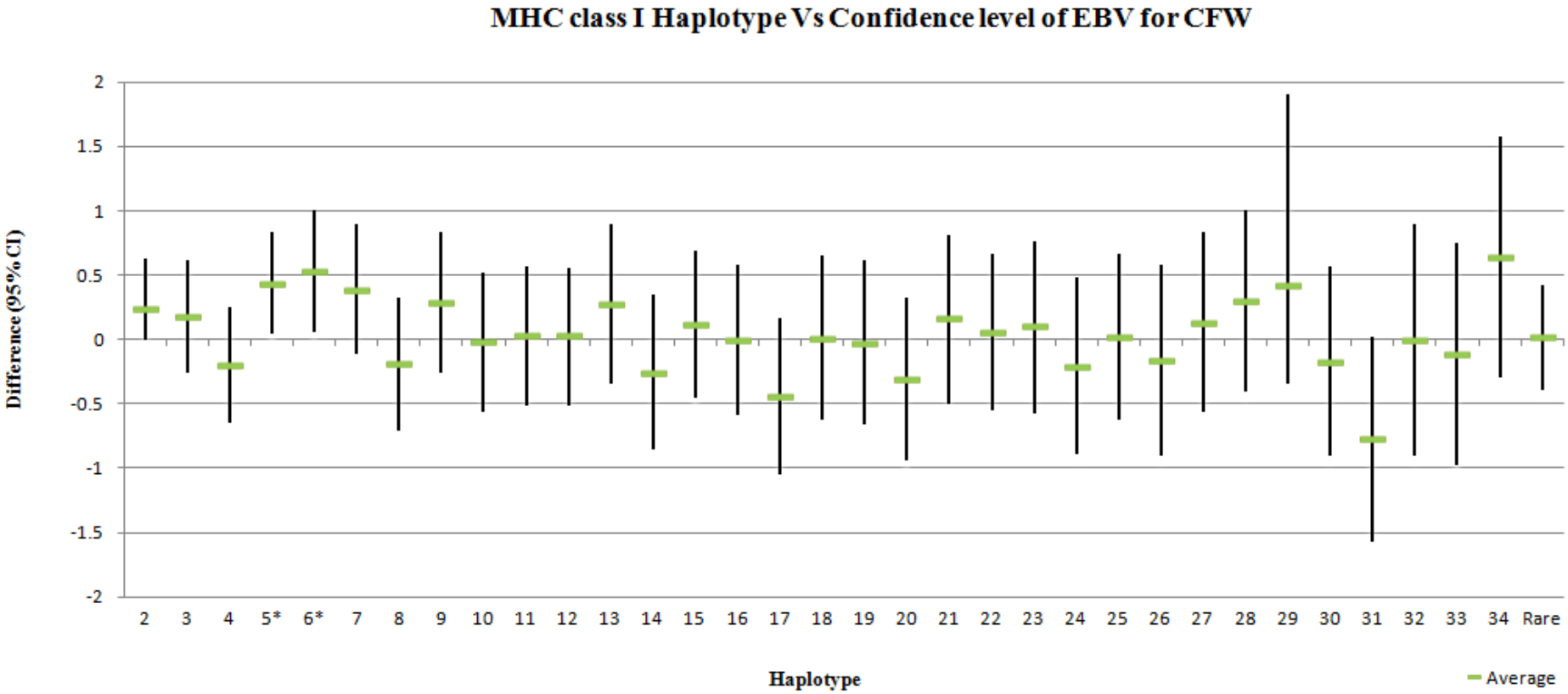


Figure 6.8: Haplotypes across MHC class I region. Confidence level for each haplotype was obtained from SNPstat analysis. All haplotypes were plotted relative to the most common haplotype (haplotype 1) which forms the baseline 0. * Haplotype with significant association.

Table 6.12: Analysis of haplotype across MHC class I region that were divided into 3 sub-sets; telomeric end, middle and centromeric end. Haplotypes that show an association are highlighted in red.

	Telomeric end								Middle								Centromeric end							
	SN21_1	**SN6_1**	**SN20_1**	**SN17_2**	**SN29_1**	**Frequency**	**Difference (95% CI)**	**P-value**	**SN15_1**	**NRM_1**	**SN43A_2**	**SN42_2**	**SN41_4**	**Frequency**	**Difference (95% CI)**	**P-value**	**SN41_4**	**SN25_1**	**SN5_1**	**CDSN_1**	**SN3A_1**	**Frequency**	**Difference (95% CI)**	**P-value**
1	G	T	G	G	C	0.2523	0.00	-	T	G	C	C	G	0.3798	0.00	-	G	C	T	C	C	0.3444	0.00	-
2	G	T	G	A	T	0.1855	-0.21 (-0.42 - 0.01)	0.061	T	A	C	C	G	0.1238	-0.07 (-0.32 - 0.18)	0.57	G	T	T	C	C	0.2202	0.01 (-0.18 - 0.2)	0.89
3	G	T	T	A	T	0.1363	-0.05 (-0.29 - 0.2)	0.72	C	G	C	C	A	0.0916	**-0.3 (-0.57 - -0.03)**	0.033	A	T	T	C	C	0.1243	**0.26 (0.04 - 0.48)**	**0.022**
4	G	T	T	G	C	0.0918	0.08 (-0.21 - 0.37)	0.57	T	G	C	C	A	0.0849	0.27 (-0.01 - 0.54)	0.059	G	T	T	T	C	0.0846	0.29 (-0.01 - 0.58)	0.06
5	A	T	T	A	C	0.0873	-0.01 (-0.31 - 0.28)	0.94	C	A	C	C	G	0.071	-0.08 (-0.39 - 0.23)	0.63	G	C	C	C	C	0.048	0.01 (-0.39 - 0.41)	0.96
6	G	C	G	A	T	0.0725	0 (-0.34 - 0.35)	0.98	C	G	C	C	G	0.0602	-0.13 (-0.5 - 0.24)	0.49	A	T	C	C	C	0.0354	-0.22 (-0.65 - 0.2)	0.3
7	A	T	T	G	C	0.0458	**-0.56 (-0.92 - -0.2)**	0.003	T	A	C	C	A	0.0467	0 (-0.4 - 0.4)	0.99	A	C	T	C	C	0.034	**-0.48 (-0.88 - -0.08)**	**0.02**
8	A	T	G	G	C	0.0298	-0.21 (-0.68 - 0.26)	0.38	T	A	T	C	G	0.0437	-0.09 (-0.51 - 0.34)	0.7	A	C	C	C	C	0.0293	**0.58 (0.1 - 1.05)**	**0.018**

Table 6.12 (Continued)

	Telomeric end								Middle								Centromeric end							
	SN21_1	SN6_1	SN20_1	SN17_2	SN29_1	Frequency	Difference (95% CI)	P-value	SN15_1	NRM_1	SN43A_2	SN42_2	SN41_4	Frequency	Difference (95% CI)	P-value	SN41_4	SN25_1	SN5_1	CDSN_1	SN3A_1	Frequency	Difference (95% CI)	P-value
9	G	T	G	A	C	0.0286	**-0.48 (-0.92 - -0.04)**	0.033	T	G	C	T	G	0.027	0.19 (-0.33 - 0.7)	0.48	G	T	C	C	C	0.0254	**0.62 (0.08 - 1.16)**	**0.025**
10	A	T	G	A	T	0.027	0.36 (-0.12 - 0.84)	0.14	C	A	C	T	G	0.0256	0.4 (-0.12 - 0.92)	0.13	G	T	C	T	C	0.0239	-0.04 (-0.48 - 0.4)	0.85
11	G	C	G	G	C	0.0247	-0.47 (-0.99 - 0.05)	0.078	C	A	T	C	G	0.0196	-0.17 (-0.82 - 0.48)	0.6	G	T	C	C	A	0.0101	0.49 (-0.25 - 1.24)	0.19
12	A	T	T	A	T	0.0137	0.31 (-0.48 - 1.09)	0.44	T	A	T	T	G	0.0108	0.18 (-0.65 - 1.01)	0.67	-	-	-	-	-	-	-	-
Rare	*	*	*	*	*	0.0047	-0.11 (-1.08 - 0.86)	0.82	*	*	*	*	*	0.0153	-0.2 (-0.87 - 0.48)	0.57	*	*	*	*	*	0.0203	-0.27 (-0.79 - 0.24)	0.3

6.3.9 Association of *CDSN* and MHC CI SNPs with multivariate EBV for wool traits

Significant associations between genotypes of individual SNPs (within and close to *CDSN*) and respective multivariate EBV for traits are given in Table 6.13. The loci CD2 (G/G), CD5 (G/G), CD8 (A/A), CD10 (T/T), CD13 (C/C) and SN25_1 (C/C) showed significant association with FD. Closer analysis of the five loci within *CDSN* showed that they were all in the same 17 animals, which were homozygous at these 5 loci. This finding most likely explains the identical mean response seen for these traits. CD7 and CD11 also showed significant associations with increased CFW and reduced fibre diameter (FD) respectively. In addition, the seven animals homozygous 'A' for CD11 were also homozygous for CD2 (G/G), CD5 (G/G), CD8 (A/A), CD10 (T/T) and CD13 (C/C). Furthermore, when the individual SNPs CD2 and CD11 were fitted as random factors to a multivariate regression model using ASREML they respectively accounted for 16.8% and 30.4% of the variation in FD observed in this population. The residual effect remaining was 52.8%.

Table 6.13: Association of individual SNP located within and close to *CDSN* with multivariate EBV for wool traits.

Locus name	SNP	No. of animals	Mean response (SE)	Type of response	P-value
CD2	G/G	17	-0.04 (0.04)	CFW	ns
	G/G	17	-1.09 (0.2)	FD	0.0017
CD5	G/G	17	-0.04 (0.04)	CFW	ns
	G/G	17	-1.09 (0.2)	FD	0.0015
CD7	T/T	7	0.22 (0.08)	CFW	0.049
	T/T	7	-0.68 (0.18)	FD	ns
CD8	A/A	17	-0.04 (0.04)	CFW	ns
	A/A	17	-1.09 (0.2)	FD	0.0015
CD10	T/T	17	-0.04 (0.04)	CFW	ns
	T/T	17	-1.09 (0.2)	FD	0.0015
CD11	A/A	7	-0.03 (0.05)	CFW	ns
	A/A	7	-1.47 (0.32)	FD	0.0017
CD13	C/C	17	-0.04 (0.04)	CFW	ns
	C/C	17	-1.09 (0.2)	FD	0.0015
SN25_1	C/C	19	-0.07 (0.04)	CFW	0.0052
	C/C	19	-0.84 (0.2)	FD	0.045

ns: No significance

Haplotype analysis was also performed on the SNPs within *CDSN* and the result is shown in Table 6.14. The estimated frequency of the most common haplotype was 22%. Ten common haplotypes accounted for 91% of haplotypes observed in this population. Rare haplotypes accounted for approximately 9% of the observed haplotypes within this population. The ten major haplotypes had varying frequencies in the population. Haplotypes 1, 2, 3 and 4 have combined frequency of 73%. Haplotype 3, with a frequency of 17% in the population, showed a significant association with increase in wool yield relative to the most common haplotype ($p = 0.048$). The mean multivariate EBV of CFW of animals with haplotype 3 was 0.1 and the 95% confidence ranged between 0.0 and 0.21. Five animals (4.6%) were homozygous for haplotype 3 with a mean CFW of 0.16 and 3/5 of these animals having a positive CFW. Interestingly, of the six animals with the highest multivariate EBV in the population, two were heterozygous for haplotype 3 and two were homozygous. No other haplotype appeared to show any significant association with CFW. A similar analysis using FD showed that there were four haplotypes having a significant increase in fibre diameter relative to the most common haplotype. No associations were identified for staple strength (SS).

Table 6.14: Haplotype associations between SNPs within *CDSN* and phenotype. Haplotypes that show an association are highlighted in red.

Haplotype	CD1	INT1	CD2	CD3	CD4	CD5	CD6	CD7	CD8	CD9	CD10	CD11	CD12	CD13	CD14	CD15	CD16	Frequency	Trait	Difference (95% CI)	P-value
1	A	C	G	G	T	G	C	C	A	A	T	A	C	C	T	G	G	0.2205	CFW	0.00	---
																			FD	0.00	---
2	A	C	A	C	T	A	C	C	G	A	C	C	C	T	T	G	G	0.2199	CFW	0.01 (-0.09 - 0.1)	0.92
																			FD	**0.44 (0.07 - 0.8)**	**0.021**
3	A	C	A	G	T	A	C	C	G	A	C	C	C	T	T	G	G	0.1716	CFW	**0.1 (0 - 0.21)**	**0.048**
																			FD	**0.63 (0.26 - 1.01)**	**0.0014**
4	A	T	A	G	T	A	C	T	G	A	C	C	C	T	T	G	A	0.1074	CFW	0.04 (-0.09 - 0.17)	0.56
																			FD	**0.62 (0.11 - 1.12)**	**0.018**
5	A	C	A	G	T	A	C	C	G	A	C	C	C	T	T	G	A	0.057	CFW	-0.02 (-0.17 - 0.12)	0.74
																			FD	0.32 (-0.21 - 0.86)	0.24
6	A	C	G	G	T	G	C	C	A	A	T	A	C	C	T	G	A	0.0365	CFW	0.1 (-0.11 - 0.31)	0.34
																			FD	**0.92 (0.17 - 1.68)**	**0.018**
7	A	C	G	G	T	G	T	C	A	A	T	C	C	C	C	A	G	0.028	CFW	-0.04 (-0.26 - 0.17)	0.71
																			FD	0.5 (-0.31 - 1.31)	0.23
8	A	C	A	G	T	G	T	C	G	A	C	C	C	C	C	A	G	0.0234	CFW	-0.18 (-0.4 - 0.03)	0.092
																			FD	0.42 (-0.38 - 1.22)	0.3
9	A	C	A	C	T	G	T	C	G	A	C	C	C	C	C	A	G	0.0187	CFW	0.15 (-0.08 - 0.38)	0.21
																			FD	0.39 (-0.49 - 1.26)	0.39
10	A	C	A	G	T	G	T	T	G	A	C	C	C	C	C	A	A	0.0187	CFW	0.09 (-0.17 - 0.35)	0.49
																			FD	-0.34 (-1.31 - 0.63)	0.49
Rare	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	0.0982	CFW	0.06 (-0.05 - 0.18)	0.28
																			FD	0.4 (-0.02 - 0.83)	0.066

Genotyping of SNPs localised within the centromeric region of MHC Class I close to *CDSN* was also performed. Eleven major haplotypes with a combined frequency of approximately 98% were observed in this population. SNPstats analysis of the genotypes is shown in Table 6.15. Four haplotypes had a significant association with CFW relative to the most common haplotype; three of which were associated with a positive multivariate EBV CFW and one with a negative multivariate EBV CFW. All four haplotypes contained the SNP allele (CDSN_INT1) also found in haplotype 3 described above.

Table 6.15: Haplotype data for four SNPs located at the centromeric of MHC class I region close to *CDSN* and one SNP within intron one of *CDSN*. Haplotypes that show an association are highlighted in red.

Haplotype	SN41_4	SN25_1	SN5_1	CDSN_1	SN3A_1	Frequency	Trait	Difference (95% CI)	P-value
1	G	C	T	C	C	0.3359	CFW	0.00	-
2	G	T	T	C	C	0.2238	CFW	0 (-0.08 - 0.07)	0.93
3	A	T	T	C	C	0.125	CFW	0.05 (-0.04 - 0.14)	0.0
							FD	**0.5 (0.08 - 0.91)**	**0.02**
4	G	T	T	T	C	0.0904	CFW	**0.12 (0.01 - 0.23)**	**0.041**
5	A	C	C	C	C	0.0488	CFW	-0.12 (-0.27 - 0.03)	0.13
6	G	C	C	C	C	0.0385	CFW	-0.13 (-0.3 - 0.05)	0.17
7	G	T	C	C	C	0.036	CFW	0.1 (-0.09 - 0.28)	0.31
8	A	C	T	C	C	0.0297	CFW	**-0.24 (-0.42 - -0.05)**	**0.012**
9	A	T	C	C	C	0.0236	CFW	**0.29 (0.1 - 0.49)**	**0.0039**
10	G	T	C	T	C	0.0152	CFW	0.01 (-0.24 - 0.25)	0.95
11	G	T	C	C	A	0.0141	CFW	**0.27 (0.04 - 0.49)**	**0.023**
Rare	*	*	*	*	*	0.0192	CFW	**-0.35 (-0.56 - -0.15)**	**0.00097**

6.4 Discussion

This chapter has analysed the *CDSN* sequence located within the sheep class I region, identified polymorphisms, and predicted the effects of mutations with respect to the protein structure. Association of *CDSN* and MHC class I haplotypes with wool yield breeding values was investigated. Significant findings from the work described in this chapter are summarised below.

- Sheep *CDSN* was annotated and is known to contain 2 exons. The sheep *CDSN* is 3683 bp in length and encodes a protein of 545 amino acids.

- Fifty-eight SNPs were identified in total within *CDSN*; 16 of these were located in the coding sequence.

- 8 of the 16 SNPs in the coding sequence cause changes in amino acid sequence. One of the non-synonymous changes was identified as 'potential disease-related mutation' through structural bioinformatics analysis.

- Sheep *CDSN* gene is conserved between various species and is maintained via purifying selection.

- One of the haplotype within *CDSN* showed a significant increase in wool yield relative to the most common estimated haplotype.

- Five SNPs within *CDSN* appeared to have a consistent effect on wool traits when found in a homozygous state.

The 1 in 80 bp SNP frequency of sheep *CDSN* indicates that the gene is highly polymorphic (Guerrin *et al.* 2001; Capon *et al.* 2003). High content of serine and glycine observed in the sheep gene is similar to that previously reported for human *CDSN* (Guerrin et al. 1998; Simon et al. 2001; Jonca et al. 2002). Sixteen SNPs were identified within the coding sequence of *CDSN* in this study, some of which may have an impact upon the function of the

protein. An additional 42 SNPs identified in this study will contribute to a comprehensive analysis of sheep MHC haplotypes. These SNPs located within the gene-rich MHC region will serve as a mapping tool for traits and diseases in sheep and especially with respect to skin and fleece. *CDSN* gene polymorphisms in humans are associated with psoriasis, a chronic inflammatory and hyper-proliferative skin disease (Allen *et al.* 1999; Jenisch *et al.* 1999; Tazi Ahnini *et al.* 1999a; Tazi Ahnini *et al.* 1999b; Schmitt-Egenolf *et al.* 2001; Capon *et al.* 2003; McGrath & Wessagowit 2005; Orru *et al.* 2005; Matsumoto *et al.* 2008). The initial genetic study linking the pathogenesis of human psoriasis to *CDSN* had identified 9 SNPs within the gene (Ishihara *et al.* 1996). More recent studies have identified a total of 23 SNPs and 3 trinucleotide indels in the exons of human *CDSN* (Jenisch *et al.* 1999; Guerrin *et al.* 2001). Like sheep *CDSN*, the human gene is regarded as a highly polymorphic, with an average SNP density of 1 SNP per 100bp (Guerrin *et al.* 2001; Capon *et al.* 2003).

The S68G, D178G, and P185S substitutions change the physicochemical properties of the residue at those positions in the peptide sequence and may therefore influence function. However, no obvious phenotypic alterations were observed in the sheep that carry these mutations. Nevertheless, the absence of obvious phenotypic changes does not mean absence of a functional or structural alteration in the protein. Unfortunately, the thermodynamic approach using iPTree-STAB and iMutant appeared to produce contradictory results with only the D178G substitution being classed as 'destabilizing' by both methods.

Selective pressure on protein encoding genes is often identified by estimating the ratio of substitution rates at non-synonymous and synonymous sites (Ka/Ks) corrected for saturation. This ratio is widely used and compares the rate of substitution at silent (i.e. synonymous) sites which are presumed neutral, to the rate of substitution at sites that result in replacement of the encoded amino acid. Values of Ka/Ks less than 1.0 are indicative of purifying selection, whereas ratios greater than 1.0 are a signature for positive selection resulting in sequence divergence and modified function. Kryazhimskiy and Plotkin (2008) have demonstrated that the ratio is relatively insensitive to the selection coefficient and that for

within species sequences (i.e. polymorphisms) the ratio is not monotonic. Furthermore, although this ratio is reasonably robust, it is now known that selection can act on synonymous sites by virtue of codon bias and the relative proportions of codon specific transfer RNA within a cell (Kimchi-Sarfaty *et al.* 2007). In this study only interspecies alignments were used to estimate Ka/Ks ratios. The results described above show an interesting picture of the internal composition of the *CDSN* gene and the selective pressures acting upon it. Abundant glycine and serine residues are present in four regions manifesting elevated Ka/Ks ratios. It is believed these are responsible for the progressive proteolysis of the *CDSN* protein within the epidermis. Since the skin of sheep and humans manifest significant differences with respect to their outer dermal covering (wool versus hair) it is not surprising that positive selection occurs in these subregions of the *CDSN*.

The Synonymous Non-synonymous Analysis Program (SNAP) was used to compare *CDSN* in various organisms, and this analysis showed that the average ratio of synonymous to non-synonymous changes (ds/dn) value (inverse of Ka/Ks) is 1.4426. This indicates that there are more synonymous than non-synonymous mutations within *CDSN*, and the gene is most likely undergoing purifying selection. The selection against mutations that lead to amino acid change could suggest that is it important to conserve the gene sequence and thus maintain the structure of CDSN for its skin and hair related functions. A previous study has shown that deletion of *CDSN* in adult knockout mice resulted in a chronic defect in the epidermal layer, suggesting that *CDSN* is essential for the preservation of structural integrity of the skin through the maintenance of the desmosome (Leclerc *et al.* 2009). In neonates, a deficiency in *CDSN* results in reduced mechanical resistance and a lethal barrier defect of the skin due to mechanical stresses is encountered after birth (Leclerc *et al.* 2009). Fleece rot and dermatophilosis are examples of bacterial infection in sheep associated with the degradation of a protective skin barrier (Norris *et al.* 2008). In cattle, there is a strong association between BoLA-DRB3-DQB class II haplotype and susceptibility to dermatophilosis (Maillard *et al.* 2003). It is also known that in cattle, *CDSN* is located within the relatively distant class I region of the MHC. Alternatively, the BoLA-DRB3-DQB association observed, rather

than conferring susceptibility, may form an extended haplotype in which variations in *CDSN* increase sensitivity to skin infections.

Analysis of sheep *CDSN* haplotypes in a population of animals with known univariate EBV for CFW showed that one haplotype was significantly associated with increased wool yields. Furthermore, an analysis of MHC class I haplotypes with the univariate EBV also showed that two haplotypes were significantly associated with wool yield in the same population of animals. Stronger associations observed in analysis of *CDSN* haplotypes could be due to the higher frequency of SNPs used for genotyping this area of interest. Analysis of haplotypes for univariate EBV for CFW within *CDSN* was performed using 17 SNPs in a region of 3683 bp in length (1 SNP in approximately 200 bp). In comparison, the haplotype analysis of SNPs within MHC class I region for univariate EBV had SNP frequency of 1: 29 kbp. Therefore, the apparently weaker association for the MHC class I haplotypes, which had only one SNP representing *CDSN* is more dispersed or diluted possibly representing more than one ancestral haplotype. Interestingly the same SNP within *CSDN* (CDSN_1) appears to be in the longer haplotype. However, the stronger haplotypic association observed between SNPs genotyped within *CDSN* compared to SNPs within MHC class I region suggests that *CDSN* gene may have an association with wool yield in sheep rather than the MHC class I region *per se*.

Analysis of haplotype sub-sets showed that there was stronger association between univariate EBV for CFW with SNPs at either ends of MHC class I region compared to the middle block. Although the analysis of haplotype sub-set from the telomeric end of MHC class I region showed one haplotype with the strongest association with wool production in terms of p-value, the overall number of haplotypes showing an association were fewer compared to the centromeric region closer to MHC class III. Multiple haplotypes from the centromeric end of MHC class I region closer to class III were associated with the trait. A previous study investigated the association between the MHC with wool production traits in sheep. Their study genotyped the SKIV2LM microsatellite marker (Groth & Wetherall 1995), adjacent to complement factor B gene in the MHC class III region and showed that several alleles were associated with a significant increase in greasy and

clean fleece weights (Bot 2000). Therefore, it could be speculated that *CDSN* gene which encodes for a protein involved in skin and wool related functions, or alternatively other genes with similar function such as *tenascin*, which are located either at the centromeric region of MHC class I or within the class III region could be involved with wool yield in sheep.

Analysis of individual SNPs with multivariate EBV for wool traits (CFW, FD and SS) has shown that there are several SNPs with significant associations with reduced FD. Five SNPs appeared to have a consistent effect when found in a homozygous state (CD2 G/G, CD5 G/G, CD8 A/A, CD10 T/T, CD13 C/C). The analysis also suggested that the effect was most likely recessive. These findings are suggestive of an ancestral haplotype having a negative effect on FD. Further analysis with multivariate EBV showed that SN25_1 was also homozygous C in animals bearing the above SNPs. Interestingly the seven animals that were homozygous for CD11 'A allele' formed a subset of these animals splitting the haplotype (CD2 G/G, CD5 G/G, CD8 A/A, CD10 T/T, CD13 C/C) into two subsets. CD11 A/A was also associated with decreased FD. Previously, some of these coding region SNPs (CD2, CD5, CD8 and CD10) have been shown to result in an amino acid change in the protein (S68G, H112R, G143S and P185S respectively).

Haplotype analysis using multivariate EBV for wool traits showed that these SNPs (CD2 'G', CD5 'G', CD8 'A', CD10 'T', CD13 'C') were also carried on the most frequent haplotype (haplotype 1) in addition to haplotypes 6 and 7 in this population. As all these SNPs had a negative effect on FD (when in a homozygous state), this suggests that there may be an extended *CDSN* haplotype associated with a decreased FD. SNPstats analysis for all *CDSN* SNPs showed that haplotype 3 was significantly associated with increased CFW and FD relative to most common haplotype (haplotype 1). There were three other haplotypes also showing significant positive associations with FD (haplotypes 2, 4, 6). A possible cause of increased FD associations seen for other haplotypes is the relative negative effect that SNPs (CD2 'G', CD5 'G', CD8 'A', CD10 'T', CD13 'C') in haplotype 1 have on FD.

CDSN forms an integral part of the desmosome (Matsumoto *et al.* 2008) so it is interesting that the data presented herein suggests that *CDSN*

polymorphism(s) affects fineness of sheep wool. As the variation has a relatively minor effect on clean fleece weight, it suggests that the desmosome may play a role in wool fineness. *CDSN* is expressed in the human and mouse hair follicles (HF) and the hair inner root sheaths (IRS) (Mils *et al.* 1992; Levy-Nissenbaum *et al.* 2003a). In hair, *CDSN* is expressed in the later keratinisation stages of the IRS and it has been suggested that *CDSN* has a pivotal role in the terminal differentiation of the IRS (Shimomura & Christiano 2010). The expression of *CDSN* and its role in wool development has not yet been studied. However, it is known that *CDSN* has strong adhesive properties, mainly through its N-terminal glycine rich domains and it is likely to be a major component in maintenance of wool follicle morphology and integrity.

The strong haplotypic association observed between SNPs within *CDSN* suggests that *CDSN* gene may have a direct effect on wool traits in Merino sheep. Further studies into the role of desmosomes in wool structure and in particular wool traits are required. The characterisation and identification of SNPs within and close to sheep *CDSN* gene will also provide an essential resource for the future identification of extended sheep MHC haplotypes. These haplotypes will be useful for linkage and association studies and will confirm a direct or indirect role of *CDSN* polymorphisms in skin related infections and wool production.

Chapter 7

General Discussion and Conclusions

The sheep industry is an important source of income to the Australian economy providing both food and fibre to the nation and the world at large. Over many decades the quality of the national flock has steadily improved by selective breeding. In the present era, it is possible to identify genetic markers associated with productivity traits in a variety of farm animals. Such markers hold promise for the more rapid development of sheep with desirable traits and also provide fundamental knowledge of the biological mechanisms underlying these traits. One genetic region that has attracted much attention in many species is the region known to control immune responsiveness due to its role in disease prevention and immunity. This region, referred to as the major histocompatibility complex or MHC is the focus of the study reported in this thesis. Characterisation of the sheep MHC will contribute to understanding the underlying genetic mechanism that affect the well-being of the animals and influence productivity traits. The sheep MHC is poorly characterised relative to the MHCs of other commercial species. This project seeks to address that shortcoming.

The broad objectives of this study and the main outcomes achieved are summarised as follows:

i. *Sub-clone and sequence CHORI Bacterial Artificial Chromosomes (BACs) known to contain sheep MHC class I sequence to identify genes present within each BAC and to determine the organisation of genes within the class I region.*

   Sub-cloning, sequencing and re-assembling of CHORI BACs resulted in construction of a low resolution physical map of sheep MHC class I region. Genes present within the physical map were annotated and the gene organisation determined. The general structure of sheep MHC class I gene content is similar to the class I region of cattle, with a few re-

arrangements in gene organisation. The sequences generated from sub-cloning CHORI BACs were also used to design specific primer to amplify intergenic and intragenic regions within a small population of sheep for single nucleotide polymorphism (SNP) identification.

ii. *Compare the map generated from sub-cloning CHORI BACs with the existing cattle MHC reference map and the recently published sheep MHC map derived from Chinese Merino sheep, to understand better the genomic organisation within the class I region.*

During the course of this project Gao and colleagues published a map of the extended sheep MHC based on a BAC library from a Chinese Merino sheep (Gao *et al.* 2010). Gao's map of the class I region, together with the existing cattle class I reference map were used as resources for comparison with the sheep class I region map determined in this study. Ten new genes that were not included in the sheep MHC map published by Gao *et al*. (2010) were identified in the class I region from CHORI BACs. A further 22 predicted genes were annotated, which were not described previously in Gao's map. Re-examination of the Gao's BAC sequences permitted resolution of several ambiguous gene annotations and a new refined sheep MHC class I map was proposed as described in Chapter 3. The peptide-presenting class I genes are clustered within 3 sub-regions or "blocks" in the sheep MHC class I region; beta (between *POU5F1* and *BAT1*, which is located in the class III region), kappa (between *TRIM26* and *GNL1*) and a novel block between *GTF2H4* and *CDSN*. No peptide-presenting MHC class I genes were identified in the alpha block (between *MOG* and *PPP1R11*), as reported in other mammals. Organisation of other genes located within MHC class I region is similar to cattle map, except for the re-arrangement of a cluster of *TRIM* genes.

iii. *Identify single nucleotide polymorphisms (SNPs) within sheep class I region, including intergenic and intragenic regions, and use these SNPs*

to genotype a population of sheep to identify patterns of linkage disequilibrium (LD).

Thirty-two SNPs were indentified within the MHC class I region, of which 14 were used to genotype a population of 108 animals. Three short blocks of high LD with significant increase in D' and r^2 value were observed. These occurred between *RPP21* and *OVAR-MHCI*, between *PPP1R10* and *PRR3* and at *NRM*. Another block of high LD with increased D' but intermediate r^2 was also observed at the centromeric end of MHC class I region between *LOC788708* and *CDSN*. These results show that for the sheep population studied, LD within the MHC does not persist across long physical distances (Mbps) as has been reported by others for non-MHC regions (Austerlitz & Heyer 1999; McRae *et al.* 2002; Odani *et al.* 2006; Miller *et al.* 2011; Kijas *et al.* 2012). The LD map of the sheep MHC class I region reported herein is the first work of its kind and will provide useful information for the identification of recombination hotspots within the sheep MHC class I region.

iv. Sequence MHC class I genes present in a small population of homozygous sheep to estimate the number of distinct loci present in each animal.

Three different sets of primers were designed that amplified many copies of full-length MHC class I sequences, except a few bases at the 3' end that encodes a sole valine comprising exon 8. The sequences generated included classical (Ia) and non-classical (Ib) class I genes. These sequences were annotated, translated and aligned with other MHC class I amino acid sequences obtained from the Immuno Polymorphism Database (IPD) and from previously published sheep work that were used as reference sequences (Grossberger *et al.* 1990; Holmes *et al.* 2003; Miltiadou *et al.* 2005; Ballingall *et al.* 2008; Wu *et al.* 2008). In addition, MHC class I sheep genomic sequences from NCBI submitted by Gao *et al*. (2010) were annotated and also used as reference sequences for alignment purposes. Assignment of MHC class I sequences into groups based on the MHC - Immuno Polymorphism

Database (IPD) criteria identified 16 independent groups. Further analysis of these sequences using phylogenetic analysis identified 14 possible loci. Four of the loci identified were unique to the homozygous animals used in this study. Three other loci were only present in the homozygous animals and genomic sequences published by Gao *et al*. (2010). Three loci characterised in previous studies (Miltiadou *et al.* 2005; Ballingall *et al.* 2008) were also present in the sequences generated from homozygous animals. The remaining four loci were not identified in the homozygous animals and are restricted to the reference sequences obtained from MHC-IPD. Hence, three to five distinct copies (possibly loci) of genomic MHC class I sequences were identified in each of the homozygous animals used.

v. *Estimate the number of expressed MHC class I genes in the cohort of homozygous sheep based on the detection of mRNA transcripts.*

Partial cDNA sequences (exons 1 to 4) derived from mRNA transcripts were amplified and sequenced from all 6 animals. The cDNA sequences were annotated and used for alignment with amino acid sequences translated from MHC class I genomic DNA sequences isolated from the same animals, as well as other reference sequences from NCBI and MHC-IPD. One to three copies (or loci) of expressed MHC class I sequences were identified in each of the six homozygous animals. Analysis of the cDNA sequences also revealed the presence of expressed pseudogenes.

vi. *Analyse and define haplotypes within the MHC class I region.*

Thirty-four definitive MHC class I haplotypes with a total frequency of ≈90% were identified within the population of 108 animal studied. Two haplotypes within the class I region had significant association with estimated breeding value for clean fleece weight in sheep.

vii. *Identify possible associations between MHC class I region SNPs and haplotypes with production traits in sheep.*

Corneodesmosin (CDSN) is a non-immunological gene located with the MHC class I region. This gene has been suggested to influence the wool production traits in sheep due to its skin and hair related functions, but has not been characterised in sheep (Bot 2000). In this study, the complete *CDSN* gene has been sequenced and annotated. Sheep *CDSN* is 3683 bp in length and encodes a protein of 545 amino acids. A total of fifty-eight SNPs were identified within *CDSN* and in the 3' untranslated region of this gene. Sixteen SNPs were identified in the coding sequence, of which eight caused non-synonymous substitutions. One of the SNPs in the coding region was characterised as a potential 'disease-related mutation' by structural bioinformatics analysis. Analysis of haplotypes within *CDSN* showed that one haplotype (haplotype 3), which has a frequency of 17%, correlated with significant increase in wool yield relative to the most common estimated haplotype (haplotype 1). Association of SNPs within *CDSN* and additional SNPs within the MHC class I region with multivariate estimated breeding value for wool traits showed that 5 loci had significant associations with fibre diameter when in a homozygous state. The strong haplotypic association observed suggests the *CDSN* gene may have a direct effect on wool traits in Merino sheep.

In completing the main objectives of this project, a detailed map of the sheep MHC class I has been obtained. Until recently most studies have focused primarily upon the characterisation of individual classical and non-classical class I loci (Miltiadou *et al.* 2005; Ballingall *et al.* 2008). It is expected the sequence data resulting from this project can be used for more efficient mapping especially when used in conjunction with newer techniques such as "capture technologies" combined with next generation sequencing. Furthermore, additional sequencing of the CHORI BACs using new sequencing techniques such as single-strand sequencing (Lieberman *et al.* 2010) will generate a comprehensive map of the sheep MHC without gaps.

The MHC class I linkage map generated in this study is probably the first of its kind. Other LD maps in sheep have targeted extensive genomic-distances not specifically within the class I region (McRae *et al.* 2002; Kijas *et al.* 2009; Miller *et al.* 2011; Kijas *et al.* 2012). The results presented in this thesis have revealed that LD within class I does not manifest across long physical distances. In the future, this should be extended to include a more extensive SNP panel and more individuals incorporating family groups. This could be achieved in conjunction with sequencing, capture and re-sequencing using next generation sequencing technology. Such studies would also permit the accurate identification of haplotypes.

Analysis of SNPs in the coding sequence of *CDSN* in sheep has indicated that there is an association between this gene and fibre diameter. This is the first study to investigate *CDSN* in sheep and to provide data suggesting that five homozygous SNP genotypes may affect wool production. Future work will be needed to determine the expression pattern of *CDSN* in the wool follicles and fibre. This study should be extended to include other breeds of sheep and to identify additional SNPs to add to the panel of existing SNPs for a more detailed description of haplotypes within *CDSN* across breeds. These SNPs should also be used to genotype other populations of sheep exhibiting variation in wool quality phenotypes. It would be interesting also to investigate polymorphisms within the promoter region of *CDSN* and their effects on expression of *CDSN*.

Identification, annotation and classification of loci for genomic DNA and cDNA of MHC class I genes is very complex. The 14 MHC class I loci identified in this study clearly reflect a high level diversity in sheep. It is possible that not all 14 loci are present in all breeds of sheep. Some breed specific loci were identified in this work that is consistent with gene duplication events occurring independently in different breeds of sheep. Given the long history of sheep breeding in the world, it is clear that MHC comparisons between distinct breeds will uncover a fascinating evolutionary history of this economically important species.

Such studies will also identify ancestral loci and haplotypes that will further facilitate the identification of loci critical for the application of marker

assisted selection in the industry. Experience to date suggests that multiple markers will be required since each marker has a much smaller effect than that obtained by a conventional dominant or recessive locus influencing a Mendelian phenotype. Together with the comprehensive map of the sheep MHC that will eventuate, the genetic resources available will then allow researchers to assess the more fundamental question of whether marker assisted selection is capable of achieving gains in productivity commensurate with the costs of developing and/or applying the technology. It may then be envisaged that other important questions will arise. For example, how is the appropriate balance achieved between the requirement for genetic diversity within important regions like the MHC (to maintain adequate immunity) and the reduced diversity implicit from introgression of multiple specific loci? Another interesting question is: "Why are high levels of diversity maintained within the MHC in domestic animals that have been selectively bred over long periods?" Future work of MHC classical (Ia) and non-classical (Ib) class I genes should also include investigation of gene expression pattern in different tissue types instead of just lymphocytes.

References

Aguilar A., Roemer G., Debenham S., Binns M., Garcelon D. & Wayne R.K. (2004) High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 3490-4.

Aldrich C.J., DeCloux A., Woods A.S., Cotter R.J., Soloski M.J. & Forman J. (1994) Identification of a Tap-dependent leader peptide recognized by alloreactive T cells specific for a class Ib antigen. *Cell* **79**, 649-58.

Allen H., Fraser J., Flyer D., Calvin S. & Flavell R. (1986) Beta 2-microglobulin is not required for cell surface expression of the murine class I histocompatibility antigen H-2Db or of a truncated H-2Db. *Proc Natl Acad Sci U S A* **83**, 7447-51.

Allen M.H., Veal C., Faassen A., Powis S.H., Vaughan R.W., Trembath R.C. & Barker J.N.W.N. (1999) A non-HLA gene within the MHC in psoriasis. *The Lancet* **353**, 1589-90.

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403-10.

Amadou C. (1999) Evolution of the <i>Mhc</i> class I region: the framework hypothesis. *Immunogenetics* **49**, 362-7.

Antoniou A.N., Ford S., Pilley E.S., Blake N. & Powis S.J. (2002) Interactions formed by individually expressed TAP1 and TAP2 polypeptide subunits. *Immunology* **106**, 182-9.

Austerlitz F. & Heyer E. (1999) Impact of demographic distribution and population growth rate on haplotypic diversity linked to a disease gene and their consequences for the estimation of recombination rate: Example of a French Canadian population. *Genetic Epidemiology* **16**, 2-14.

Ballingall K., Miltiadou D., Chai Z.-W., McLean K., Rocchi M., Yaga R. & McKeever D. (2008) Genetic and proteomic analysis of the MHC class I repertoire from four ovine haplotypes. *Immunogenetics* **60**, 177-84.

Baudat F., Buard J., Grey C., Fledel-Alon A., Ober C., Przeworski M., Coop G. & de Massy B. (2010) PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**, 836-40.

Beck S. & Trowsdale J. (2000) The Human Major Histocompatibility Complex: Lessons from the DNA Sequence. *Annual Review of Genomics and Human Genetics* **1**, 117-37.

Bell J. (2008) A simple way to treat PCR products prior to sequencing using ExoSAP-IT. *Biotechniques* **44**, 834.

Bidwell J. (1994) Advances in DNA-based HLA-typing methods. *Immunology Today* **15**, 303-7.

Birch J., Codner G., Guzman E. & Ellis S.A. (2008a) Genomic location and characterisation of nonclassical MHC class I genes in cattle. *Immunogenetics* **60**, 267-73.

Birch J., De Juan Sanjuan C., Guzman E. & Ellis S.A. (2008b) Genomic location and characterisation of MIC genes in cattle. *Immunogenetics* **60**, 477-83.

Bjorkman P.J. & Parham P. (1990) Structure, function, and diversity of class I major histocompatibility complex molecules. *Annu Rev Biochem* **59**, 253-88.

Bjorkman P.J., Saper M.A., Samraoui B., Bennett W.S., Strominger J.L. & Wiley D.C. (1987a) Structure of the human class I histocompatibility antigen, HLA-A2. *Nature* **329**, 506-12.

Bjorkman P.J., Saper M.A., Samraoui B., Bennett W.S., Strominger J.L. & Wiley D.C. (1987b) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature* **329**, 512-8.

Blattman A.N. & Beh K.J. (1992) Dinucleotide repeat polymorphism within the ovine major histocompatibility complex. *Animal Genetics* **23**, 392.

Bonneaud C., Chastel O., Federici P., Westerdahl H. & Sorci G. (2006) Complex Mhc-based mate choice in a wild passerine. *Proceedings. Biological sciences / The Royal Society* **273**, 1111-6.

Borrego F., Ulbrecht M., Weiss E.H., Coligan J.E. & Brooks A.G. (1998) Recognition of human histocompatibility leukocyte antigen (HLA)-E complexed with HLA class I signal sequence-derived peptides by CD94/NKG2 confers protection from natural killer cell-mediated lysis. *The Journal of Experimental Medicine* **187**, 813-8.

Bot J. (2000) MHC Studies Relating To Parasite Resistance In Merino Sheep. In: *School of Biomedical Sciences*, p. 177. Curtin University of Technology.

Botstein D. & Risch N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics* **33 Suppl**, 228-37.

Bozkaya F., Kuss A.W. & Geldermann H. (2007) DNA variants of the MHC show location-specific convergence between sheep, goat and cattle. *Small Ruminant Research* **70**, 174-82.

Braud V.M., Allan D.S. & McMichael A.J. (1999) Functions of nonclassical MHC and non-MHC-encoded class I molecules. *Current Opinion in Immunology* **11**, 100-8.

Braud V.M., Allan D.S., O'Callaghan C.A., Soderstrom K., D'Andrea A., Ogg G.S., Lazetic S., Young N.T., Bell J.I., Phillips J.H., Lanier L.L. & McMichael A.J. (1998b) HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. *Nature* **391**, 795-9.

Braud V.M., Allan D.S., Wilson D. & McMichael A.J. (1998a) TAP- and tapasin-dependent HLA-E surface expression correlates with the binding of an MHC class I leader peptide. *Curr Biol* **8**, 1-10.

Brinkmeyer-Langford C.L., Childers C.P., Fritz K.L., Gustafson-Seabury A.L., Cothran M., Raudsepp T., Womack J.E. & Skow L.C. (2009) A high resolution RH map of the bovine major histocompatibility complex. *BMC Genomics* **10**, 182.

Brudno M., Steinkamp R. & Morgenstern B. (2004) The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. *Nucleic Acids Research* **32**, W41-W4.

Burge C. & Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol* **268**, 78-94.

Campbell R.D., Carroll M.C. & Porter R.R. (1986) The molecular genetics of components of complement. *Advances in Immunology* **38**, 203-44.

Capon F., Toal I.K., Evans J.C., Allen M.H., Patel S., Tillman D., Burden D., Barker J.N.W.N. & Trembath R.C. (2003) Haplotype analysis of distantly related populations implicates corneodesmosin in psoriasis susceptibility. *Journal of Medical Genetics* **40**, 447-52.

Capriotti E., Fariselli P. & Casadio R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20 Suppl 1**, i63-8.

Capriotti E., Fariselli P. & Casadio R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* **33**, W306-10.
Capriotti E., Fariselli P., Rossi I. & Casadio R. (2008) A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9 Suppl 2**, S6.
Carbone F.R. & Bevan M.J. (1990) Class I-restricted processing and presentation of exogenous cell- associated antigen in vivo. *The Journal of Experimental Medicine* **171**, 377-87.
Cardon L.R. & Abecasis G.R. (2003) Using haplotype blocks to map human complex trait loci. *Trends in Genetics* **19**, 135-40.
Chaix R., Cao C. & Donnelly P. (2008) Is mate choice in humans MHC-dependent? *PLoS Genetics* **4**, e1000184.
Chen J.-M., Cooper D.N., Chuzhanova N., Ferec C. & Patrinos G.P. (2007) Gene conversion: mechanisms, evolution and human disease. *Nature Reviews. Genetics* **8**, 762-75.
Childers C.P., Newkirk H.L., Honeycutt D.A., Ramlachan N., Muzney D.M., Sodergren E., Gibbs R.A., Weinstock G.M., Womack J.E. & Skow L.C. (2006) Comparative analysis of the bovine MHC class IIb sequence1 identifies inversion breakpoints and three unexpected genes. *Animal Genetics* **37**, 121-9.
Clarke A.G. (1971) The effects of maternal pre-immunization on pregnancy in the mouse. *Journal of Reproduction and Fertility* **24**, 369-75.
Clarke B. & Kirby D.R. (1966) Maintenance of histocompatibility polymorphisms. *Nature* **211**, 999-1000.
Cook R.G., Leone B., Leone J.W., Widacki S.M. & Zavell P.J. (1992) Characterization of T cell proliferative responses induced by anti-Qa-2 monoclonal antibodies. *Cellular Immunology* **144**, 367-81.
Cresswell P. & Howard J. (1999) Antigen recognition: Editorial overview. *Current Opinion in Immunology* **11**, 61-3.
Cullen P.R., Bunch C., Brownlie J. & Morris P.J. (1982) Sheep lymphocyte antigens: a preliminary study. *Animal Blood Groups and Biochemical Genetics* **13**, 149-59.
Curwen V., Eyras E., Andrews T.D., Clarke L., Mongin E., Searle S.M.J. & Clamp M. (2004) The Ensembl Automatic Gene Annotation System. *Genome Research* **14**, 942-50.
Daly M.J., Rioux J.D., Schaffner S.F., Hudson T.J. & Lander E.S. (2001) High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229-32.
Danchin E., Vitiello V., Vienne A., Richard O., Gouret P., McDermott M.F. & Pontarotti P. (2004) The major histocompatibility complex origin. *Immunological Reviews* **198**, 216-32.
Danchin E.G., Abi-Rached L., Gilles A. & Pontarotti P. (2003) Conservation of the MHC-like region throughout evolution. *Immunogenetics* **55**, 141-8.
Davies C.J., Eldridge J.A., Fisher P.J. & Schlafer D.H. (2006) Evidence for Expression of Both Classical and Non-Classical Major Histocompatibility Complex Class I Genes in Bovine Trophoblast Cells. *American Journal of Reproductive Immunology* **55**, 188-200.
Dawson E., Abecasis G.R., Bumpstead S., Chen Y., Hunt S., Beare D.M., Pabial J., Dibling T., Tinsley E., Kirby S., Carter D., Papaspyridonos M., Livingstone S., Ganske R., Lohmussaar E., Zernant J., Tonisson N., Remm M., Magi R., Puurand T., Vilo J., Kurg A., Rice K., Deloukas P., Mott R., Metspalu A., Bentley D.R., Cardon L.R. & Dunham I.

(2002) A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544-8.
Demars J., Riquet J., Feve K., Gautier M., Morisson M., Demeure O., Renard C., Chardon P. & Milan D. (2006) High resolution physical map of porcine chromosome 7 QTL region and comparative mapping of this region among vertebrate genomes. *BMC Genomics* **7**, 13.
Doherty P.C. & Zinkernagel R.M. (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature* **256**, 50-2.
Dukkipati V.S., Blair H.T., Garrick D.J. & Murray A. (2006) Ovar-Mhc--ovine major histocompatibility complex: role in genetic resistance to diseases. *New Zealand Veterinary Journal* **54**, 153-60.
Dyer P.A. & Martin S. (1991) Techniques used to define human MHC antigens: serology. *Immunology Letters* **29**, 15-21.
Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-7.
Edwards A.O., Ritter R., Abel K.J., Manning A., Panhuysen C. & Farrer L.A. (2005) Complement Factor H Polymorphism and Age-Related Macular Degeneration. *Science* **308**, 421-4.
Ellis S. (2004) The cattle major histocompatibility complex: is it unique? *Veterinary Immunology and Immunopathology* **102**, 1-8.
Ellis S., Bontrop R., Antczak D., Ballingall K., Davies C., Kaufman J., Kennedy L., Robinson J., Smith D., Stear M., Stet R., Waller M., Walter L. & Marsh S. (2006) ISAG/IUIS-VIC Comparative MHC Nomenclature Committee report, 2005. *Immunogenetics* **57**, 953-8.
Ellis S.A., Martin A.J., Holmes E.C. & Morrison W.I. (1995) At least four MHC class I genes are transcribed in the horse: phylogenetic analysis suggests an unusual evolutionary history for the MHC in this species. *European Journal of Immunogenetics* **22**, 249-60.
Farmery M.R., Allen S., Allen A.J. & Bulleid N.J. (2000) The Role of ERp57 in Disulfide Bond Formation during the Assembly of Major Histocompatibility Complex Class I in a Synchronized Semipermeabilized Cell Translation System. *Journal of Biological Chemistry* **275**, 14933-8.
Fickett J.W. (1996) The gene identification problem: An overview for developers. *Computers and Chemistry* **20**, 103-18.
Figueroa F. & Klein J. (1986) The evolution of MHC class II genes. *Immunology Today* **7**, 78-81.
Flajnik M.F. & Kasahara M. (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. *Immunity* **15**, 351-62.
Forman J. (1979) H-2 unrestricted cytotoxic T cell activity against antigens controlled by genes in the QA/TLA region. *Journal of Immunology* **123**, 2451-5.
Forsberg L.A., Dannewitz J., Petersson E. & Grahn M. (2007) Influence of genetic dissimilarity in the reproductive success and mate choice of brown trout - females fishing for optimal MHC dissimilarity. *Journal of Evolutionary Biology* **20**, 1859-69.
Freeman-Gallant C.R., Meguerdichian M., Wheelwright N.T. & Sollecito S.V. (2003) Social pairing and female mating fidelity predicted by restriction fragment length polymorphism similarity at the major histocompatibility complex in a songbird. *Molecular Ecology* **12**, 3077-83.

Frenkel Z., Shenkman M., Kondratyev M. & Lederkremer G.Z. (2004) Separate roles and different routing of calnexin and ERp57 in endoplasmic reticulum quality control revealed by interactions with asialoglycoprotein receptor chains. *Molecular Biology of the Cell* **15**, 2133-42.

Gabriel S.B., Schaffner S.F., Nguyen H., Moore J.M., Roy J., Blumenstiel B., Higgins J., DeFelice M., Lochner A., Faggart M., Liu-Cordero S.N., Rotimi C., Adeyemo A., Cooper R., Ward R., Lander E.S., Daly M.J. & Altshuler D. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225-9.

Gao J., Liu K., Liu H., Blair H.T., Li G., Chen C., Tan P. & Ma R.Z. (2010) A complete DNA sequence map of the ovine major histocompatibility complex. *BMC Genomics* **11**, 466.

Garrido J.J., de Andres D.F., Pintado C.O., Llanes D. & Stear M.J. (1995) Serologically defined lymphocyte alloantigens in Spanish sheep. *Experimental and Clinical Immunogenetics* **12**, 268-71.

Gaudieri S., Kulski J.K., Dawkins R.L. & Gojobori T. (1999) Different Evolutionary Histories in Two Subgenomic Regions of the Major Histocompatibility Complex. *Genome Research* **9**, 541-9.

Geraghty D.E., Koller B.H., Hansen J.A. & Orr H.T. (1992) The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. *The Journal of Immunology* **149**, 1934-46.

Geraghty D.E., Koller B.H. & Orr H.T. (1987) A human major histocompatibility complex class I gene that encodes a protein with a shortened cytoplasmic segment. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 9145-9.

Geraghty D.E., Wei X.H., Orr H.T. & Koller B.H. (1990) Human leukocyte antigen F (HLA-F). An expressed HLA gene composed of a class I coding sequence linked to a novel transcribed repetitive element. *The Journal of Experimental Medicine* **171**, 1-18.

Gibson J., Morton N.E. & Collins A. (2006) Extended tracts of homozygosity in outbred human populations. *Hum Mol Genet* **15**, 789-95.

Gogolin-Ewens K.J., Mackay C.R., Mercer W.R. & Brandon M.R. (1985) Sheep lymphocyte antigens (OLA). I. Major histocompatibility complex class I molecules. *Immunology* **56**, 717-23.

Gorer P.A. (1937) The genetic and antigenic basis of tumour transplantation. *Journal of Pathology and Bacteriology* **44**, 691-7.

Gouy M., Guindon S.p. & Gascuel O. (2010) SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution* **27**, 221-4.

Green M.R. (1986) Pre-mRNA splicing. *Annual Review of Genetics* **20**, 671-708.

Grey H.M., Kubo R.T., Colon S.M., Poulik M.D., Cresswell P., Springer T., Turner M. & Strominger J.L. (1973) The small subunit of HL-A antigens is beta 2-microglobulin. *The Journal of Experimental Medicine* **138**, 1608-12.

Grossberger D., Hein W. & Marcuz A. (1990) Class I major histocompatibility complex cDNA clones from sheep thymus: alternative splicing could make a long cytoplasmic tail. *Immunogenetics* **32**, 77-87.

Groth D.M. & Wetherall J.D. (1994) Dinucleotide repeat polymorphism within the ovine major histocompatibility complex class I region. *Animal Genetics* **25**, 61.

Groth D.M. & Wetherall J.D. (1995) Dinucleotide repeat polymorphism adjacent to sheep complement factor B. *Animal Genetics* **26**, 282-3.
Grubic Z., Kerhin-Brkljacic V., Cecuk-Jelicic E., Kuci S. & Kastelan A. (2000) HLA class I polymorphism in the Albanian population. *Collegium Antropologicum* **24**, 303-7.
Grubic Z., Zunec R., Stingl K., Svilicic D. & Kerhin-Brkljacic V. (2008) [Haplotipic associations of the two most common HLA-B*27 alleles in the Croatian population]. *Reumatizam* **55**, 5-9.
Gruszczynska J., Charon K.M., Swiderek W. & Sawera M. (2002) Microsatellite polymorphism in locus OMHC1 (MHC Class I) in Polish Heath Sheep and Polish Lowland Sheep (Zelazna variety). *Journal of Applied Genetics* **43**, 217-22.
Guerrin M., Simon M., Montezin M., Haftek M., Vincent C. & Serre G. (1998) Expression Cloning of Human Corneodesmosin Proves Its Identity with the Product of the S Gene and Allows Improved Characterization of Its Processing during Keratinocyte Differentiation. *Journal of Biological Chemistry* **273**, 22640-7.
Guerrin M., Vincent C., Simon M., Tazi Ahnini R., Fort M. & Serre G. (2001) Identification of six novel polymorphisms in the human corneodesmosin gene. *Tissue Antigens* **57**, 32-8.
Guillet J.G., Lai M.Z., Briner T.J., Smith J.A. & Gefter M.L. (1986) Interaction of peptide antigens and class II major histocompatibility complex antigens. *Nature* **324**, 260-2.
Guindon S.p., Dufayard J.-F.o., Lefort V., Anisimova M., Hordijk W. & Gascuel O. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307-21.
Gustafson A.L., Tallmadge R.L., Ramlachan N., Miller D., Bird H., Antczak D.F., Raudsepp T., Chowdhary B.P. & Skow L.C. (2003) An ordered BAC contig map of the equine major histocompatibility complex. *Cytogenetic and Genome Research* **102**, 189-95.
Haines J.L., Hauser M.A., Schmidt S., Scott W.K., Olson L.M., Gallins P., Spencer K.L., Kwan S.Y., Noureddine M., Gilbert J.R., Schnetz-Boutaud N., Agarwal A., Postel E.A. & Pericak-Vance M.A. (2005) Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science* **308**, 419-21.
Hansen T.H., Myers N.B. & Lee D.R. (1988) Studies of two antigenic forms of Ld with disparate beta 2-microglobulin (beta 2m) associations suggest that beta 2m facilitate the folding of the alpha 1 and alpha 2 domains during de novo synthesis. *Journal of Immunology* **140**, 3522-7.
HapMap C. (2005) A haplotype map of the human genome. *Nature* **437**, 1299-320.
Harty J.T., Tvinnereim A.R. & White D.W. (2000) CD8+ T cell effector mechanisms in resistance to infection. *Annual Review of Immunology* **18**, 275.
Hastbacka J., de la Chapelle A., Kaitila I., Sistonen P., Weaver A. & Lander E. (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics* **2**, 204-11.
Hediger R., Ansari H.A. & Stranzinger G.F. (1991) Chromosome banding and gene localizations support extensive conservation of chromosome structure between cattle and sheep. *Cytogenetics and Cell Genetics* **57**, 127-34.

Hedrick P.W. (1998) Balancing selection and MHC. *Genetica* **104**, 207-14.
Hedrick P.W. & Black F.L. (1997) HLA and mate selection: no evidence in South Amerindians. *American Journal of Human Genetics* **61**, 505-11.
Hedrick P.W. & Thomson G. (1983) Evidence for balancing selection at HLA. *Genetics* **104**, 449-56.
Heinold A., Schaller-Suefling E., Opelz G., Scherer S. & Tran T.H. (2008) Identification of two novel HLA alleles, HLA-A*02010103 and HLA-B*4455, and characterization of the complete genomic sequence of HLA-A*290201. *Tissue Antigens* **72**, 397-400.
Henikoff S. & Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10915-9.
Herberg J.A., Beck S. & Trowsdale J. (1998) TAPASIN, DAXX, RGL2, HKE2 and four new genes (BING 1, 3 to 5) form a dense cluster at the centromeric end of the MHC. *Journal of Molecular Biology* **277**, 839-57.
Hickford J.G.H., Zhou H., Slow S. & Fang Q. (2004) Diversity of the ovine DQA2 gene. *Journal of Animal Science* **82**, 1553-63.
Holmes E.C., Roberts A.F.C., Staines K.A. & Ellis S.A. (2003) Evolution of major histocompatibility complex class I genes in Cetartiodactyls. *Immunogenetics* **55**, 193-202.
Horton R., Wilming L., Rand V., Lovering R.C., Bruford E.A., Khodiyar V.K., Lush M.J., Povey S., Talbot C.C., Jr., Wright M.W., Wain H.M., Trowsdale J., Ziegler A. & Beck S. (2004) Gene map of the extended human MHC. *Nature Reviews. Genetics* **5**, 889-99.
Howarth M., Williams A., Tolstrup A.B. & Elliott T. (2004) Tapasin enhances MHC clss I peptide presentation according to peptide half-life. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11737-42.
Huang L.T., Gromiha M.M. & Ho S.Y. (2007) iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics* **23**, 1292-3.
Hughes A.L. & Nei M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167-70.
Hughes A.L. & Nei M. (1989a) Evolution of the major histocompatibility complex: independent origin of nonclassical class I genes in different groups of mammals. *Molecular Biology and Evolution* **6**, 559-79.
Hughes A.L. & Nei M. (1989b) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 958-62.
Hughes A.L. & Yeager M. (1998) Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics* **32**, 415-35.
Hunkapiller T. & Hood L. (1989) Diversity of the immunoglobulin gene superfamily. *Advances in Immunology* **44**, 1-63.
Hurt P., Walter L., Sudbrak R., Klages S., MÃ¼ller I., Shiina T., Inoko H., Lehrach H., GÃ¼nther E., Reinhardt R. & Himmelbauer H. (2004) The Genomic Sequence and Comparative Analysis of the Rat Major Histocompatibility Complex. *Genome Research* **14**, 631-9.
Imanishi T., Akaza T., Kimura A., Tokunaga K. & Gojobori T. (1992) Allele and haplotype frequencies for HLA and complement loci in various ethnic groups. P 1065 in K. Tsun, M. Aizawa and T. Sasazuki, eds.

HLA 1991: proceedings of the eleventh international histocompatibility workshop and conference. Oxford University Press, Oxford.

Ishihara M., Yamagata N., Ohno S., Naruse T., Ando A., Kawata H., Ozawa A., Ohkido M., Mizuki N., Shiina T., Ando H. & Inoko H. (1996) Genetic polymorphisms in the keratin-like S gene within the human major histocompatibility complex and association analysis on the susceptibility to psoriasis vulgaris. *Tissue Antigens* **48**, 182-6.

Ishitani A., Sageshima N., Lee N., Dorofeeva N., Hatake K., Marquardt H. & Geraghty D.E. (2003) Protein expression and peptide binding suggest unique and interacting functional roles for HLA-E, F, and G in maternal-placental immune recognition. *Journal of Immunology* **171**, 1376-84.

James D.A. (1965) Effects of Antigenic Dissimilarity between Mother and Foetus on Placental Size in Mice. *Nature* **205**, 613-4.

Jaulin C., Perrin A., Abastado J.P., Dumas B., Papamatheakis J. & Kourilsky P. (1985) Polymorphism in mouse and human class I H-2 and HLA genes is not the result of random independent point mutations. *Immunogenetics* **22**, 453-70.

Jeffreys A.J., Kauppi L. & Neumann R. (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics* **29**, 217-22.

Jenisch S., Koch S., Henseler T., Nair R.P., Elder J.T., Watts C.E., Westphal E., Voorhees J.J., Christophers E. & Krönke M. (1999) Corneodesmosin gene polymorphism demonstrates strong linkage disequilibrium with HLA and association with psoriasis vulgaris. *Tissue Antigens* **54**, 439-49.

Jonca N., Caubet C., Guerrin M., Simon M. & Serre G. (2010) Corneodesmosin: Structure, Function and Involvement in Pathophysiology. *The Open Dermatology Journal* **4**, 36-45.

Jonca N., Guerrin M., Hadjiolova K., Caubet C.c., Gallinaro H.l.n., Simon M. & Serre G. (2002) Corneodesmosin, a Component of Epidermal Corneocyte Desmosomes, Displays Homophilic Adhesive Properties. *Journal of Biological Chemistry* **277**, 5024-9.

Jones D.T., Taylor W.R. & Thornton J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275-82.

Jugo B.M., Joosten I., Grosfeld-Stulemeyer M., Amorena B. & Hensen E.J. (2002) Immunoprecipitation and isoelectric focusing of sheep MHC class I antigens reveal higher complexity than serology. *European Journal of Immunogenetics* **29**, 391-9.

Jugo B.M. & Vicario A. (2001) Lymphocyte antigens in sheep: linkage to the MHC class II DRB1 gene. *European Journal of Immunogenetics* **28**, 451-8.

Kaeuffer R., Coltman D.W., Chapuis J.L., Pontier D. & Reale D. (2007) Unexpected heterozygosity in an island mouflon population founded by a single pair of individuals. *Proceedings. Biological sciences / The Royal Society* **274**, 527-33.

Kalinin A., Marekov L.N. & Steinert P.M. (2001) Assembly of the epidermal cornified cell envelope. *Journal of Cell Science* **114**, 3069-70.

Kasahara M., Suzuki T. & Pasquier L.D. (2004) On the origins of the adaptive immune system: novel insights from invertebrates and cold-blooded vertebrates. *Trends in Immunology* **25**, 105-11.

Kastner D.L., Rich R.R. & Shen F.W. (1979) Qa-1-associated antigens. I. Generation of H-2-nonrestricted cytotoxic T lymphocytes specific for determinants of the Qa-1 region. *Journal of Immunology* **123**, 1232-8.

Kelley J., Walter L. & Trowsdale J. (2005) Comparative genomics of major histocompatibility complexes. *Immunogenetics* **56**, 683-95.

Khatkar M.S., Nicholas F.W., Collins A.R., Zenger K.R., Cavanagh J.A., Barris W., Schnabel R.D., Taylor J.F. & Raadsma H.W. (2008) Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**, 187.

Kijas J.W., Lenstra J.A., Hayes B., Boitard S., Porto Neto L.R., San Cristobal M., Servin B., McCulloch R., Whan V., Gietzen K., Paiva S., Barendse W., Ciani E., Raadsma H., McEwan J., Dalrymple B. & other members of the International Sheep Genomics C. (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals High Levels of Historic Mixture and Strong Recent Selection. *PLoS Biol* **10**, e1001258.

Kijas J.W., Townley D., Dalrymple B.P., Heaton M.P., Maddox J.F., McGrath A., Wilson P., Ingersoll R.G., McCulloch R., McWilliam S., Tang D., McEwan J., Cockett N., Oddy V.H., Nicholas F.W., Raadsma H. & for the International Sheep Genomics C. (2009) A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. *PLoS ONE* **4**, e4668.

Kimchi-Sarfaty C., Oh J.M., Kim I.-W., Sauna Z.E., Calcagno A.M., Ambudkar S.V. & Gottesman M.M. (2007) A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. *Science* **315**, 525-8.

Kimura M. (1979) The neutral theory of molecular evolution. *Scientific American* **241**, 98-100.

King A., Boocock C., Sharkey A.M., Gardner L., Beretta A., Siccardi A.G. & Loke Y.W. (1996) Evidence for the expression of HLAA-C class I mRNA and protein by human first trimester trophoblast. *Journal of Immunology* **156**, 2068-76.

Klein J., Bontrop R.E., Dawkins R.L., Erlich H.A., Gyllensten U.B., Heise E.R., Jones P.P., Parham P., Wakeland E.K. & Watkins D.I. (1990) Nomenclature for the major histocompatibility complexes of different species: a proposal. *Immunogenetics* **31**, 217-9.

Klein J. & Figueroa F. (1986) The evolution of class I MHC genes. *Immunology Today* **7**, 41-4.

Klein J. & O'Huigin C. (1993) Composite origin of major histocompatibility complex genes. *Current Opinion in Genetics & Development* **3**, 923-30.

Klein R.J., Zeiss C., Chew E.Y., Tsai J.-Y., Sackler R.S., Haynes C., Henning A.K., SanGiovanni J.P., Mane S.M., Mayne S.T., Bracken M.B., Ferris F.L., Ott J., Barnstable C. & Hoh J. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385-9.

Koller B.H., Geraghty D.E., Shimizu Y., DeMars R. & Orr H.T. (1988) HLA-E. A novel HLA class I gene expressed in resting T lymphocytes. *Journal of Immunology* **141**, 897-904.

Konishi S., Izawa T., Lin S.Y., Ebana K., Fukuta Y., Sasaki T. & Yano M. (2006) An SNP Caused Loss of Seed Shattering During Rice Domestication. *Science* **312**, 1392-6.

Korber B. (2000) HIV Signature and Sequence Variation Analysis. In: *Computational Analysis of HIV Molecular Sequences* (eds. by Rodrigo

AG & Learn GH), pp. 55-72. Kluwer Academic Publishers, Dordrecht, Netherlands.

Kovacsovics-Bankowski M., Clark K., Benacerraf B. & Rock K.L. (1993) Efficient major histocompatibility complex class I presentation of exogenous antigen upon phagocytosis by macrophages. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 4942-6.

Krangel M.S., Biddison W.E. & Strominger J.L. (1983) Comparative structural analysis of HLA-A2 antigens distinguishable by cytotoxic T lymphocytes. II. Variant DK1: evidence for a discrete CTL recognition region. *Journal of Immunology* **130**, 1856-62.

Kulski J.K., Anzai T., Shiina T. & Inoko H. (2004) Rhesus Macaque Class I Duplicon Structures, Organization, and Evolution Within the Alpha Block of the Major Histocompatibility Complex. *Molecular Biology and Evolution* **21**, 2079-91.

Kulski J.K., Gaudieri S., Bellgard M., Balmer L., Giles K., Inoko H. & Dawkins R.L. (1997) The evolution of MHC diversity by segmental duplication and transposition of retroelements. *Journal of Molecular Evolution* **45**, 599-609.

Kulski J.K., Gaudieri S., Martin A. & Dawkins R.L. (1999) Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. *Journal of Molecular Evolution* **49**, 84-97.

Kulski J.K., Shiina T., Anzai T., Kohara S. & Inoko H. (2002) Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunological Reviews* **190**, 95-122.

Laird D.J., De Tomaso A.W., Cooper M.D. & Weissman I.L. (2000) 50 million years of chordate evolution: seeking the origins of adaptive immunity. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6924-6.

Lancet D., Parham P. & Strominger J.L. (1979) Heavy chain of HLA-A and HLA-B antigens is conformationally labile: a possible role for beta 2-microglobulin. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 3844-8.

Landry C., Garant D., Duchesne P. & Bernatchez L. (2001) 'Good genes as heterozygosity': the major histocompatibility complex and mate choice in Atlantic salmon (Salmo salar). *Proceedings. Biological sciences / The Royal Society* **268**, 1279-85.

Le Friec G., Gros F., Sebti Y., Guilloux V., Pangault C., Fauchet R. & Amiot L. (2004) Capacity of myeloid and plasmacytoid dendritic cells especially at mature stage to express and secrete HLA-G molecules. *Journal of Leukocyte Biology* **76**, 1125-33.

Le Rond S., Le Maoult J., Creput C., Menier C., Deschamps M., Le Friec G., Amiot L., Durrbach A., Dausset J., Carosella E.D. & Rouas-Freiss N. (2004) Alloreactive CD4+ and CD8+ T cells express the immunotolerant HLA-G molecule in mixed lymphocyte reactions: in vivo implications in transplanted patients. *European Journal of Immunology* **34**, 649-60.

Le S.Q. & Gascuel O. (2008) An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**, 1307-20.

Leclerc E.A., Huchenq A., Mattiuzzo N.R., Metzger D., Chambon P., Ghyselinck N.B., Serre G., Jonca N. & Guerrin M. (2009) Corneodesmosin gene ablation induces lethal skin-barrier disruption

and hair-follicle degeneration related to desmosome dysfunction. *Journal of Cell Science* **122**, 2699-709.

Lee C.Y., Qin J., Munyard K.A., Siva Subramaniam N., Wetherall J.D., Stear M.J. & Groth D.M. (2011) Conserved haplotype blocks within the sheep MHC and low SNP heterozygosity in the Class IIa subregion. *Animal Genetics*, no-no; 10.1111/j.365-2052.11.02268.x.

Lee N. & Geraghty D.E. (2003) HLA-F surface expression on B cell and monocyte cell lines is partially independent from tapasin and completely independent from TAP. *Journal of Immunology* **171**, 5264-71.

Lee N., Goodlett D.R., Ishitani A., Marquardt H. & Geraghty D.E. (1998a) HLA-E surface expression depends on binding of TAP-dependent peptides derived from certain HLA class I signal sequences. *Journal of Immunology* **160**, 4951-60.

Lee N., Llano M., Carretero M., Ishitani A., Navarro F., Lopez-Botet M. & Geraghty D.E. (1998b) HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 5199-204.

Leelayuwat C., Pinelli M. & Dawkins R.L. (1995) Clustering of diverse replicated sequences in the MHC. Evidence for en bloc duplication. *The Journal of Immunology* **155**, 692-8.

Levy-Nissenbaum E., Betz R.C., Frydman M., Simon M., Lahat H., Bakhan T., Goldman B., Bygum A., Pierick M., Hillmer A.M., Jonca N., Toribio J., Kruse R., Dewald G., Cichon S., Kubisch C., Guerrin M., Serre G., Nothen M.M. & Pras E. (2003a) Hypotrichosis simplex of the scalp is associated with nonsense mutations in CDSN encoding corneodesmosin. *Nature Genetics* **34**, 151-3.

Levy-Nissenbaum E., Betz R.C., Frydman M., Simon M., Lahat H., Bakhan T., Goldman B., Bygum A., Pierick M., Hillmer A.M., Jonca N., Toribio J., Kruse R., Dewald G., Cichon S., Kubisch C., Guerrin M., Serre G., Nothen M.M. & Pras E. (2003b) Hypotrichosis simplex of the scalp is associated with nonsense mutations in CDSN encoding corneodesmosin. *Nat Genet* **34**, 151-3.

Librado P. & Rozas J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-2.

Lieberman K.R., Cherf G.M., Doody M.J., Olasagasti F., Kolodji Y. & Akeson M. (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society* **132**, 17961-72.

Liu H., Liu K., Wang J. & Ma R.Z. (2006) A BAC clone-based physical map of ovine major histocompatibility complex. *Genomics* **88**, 88-95.

Liu K., Zhang P., Gao J., Liu H., Li G., Qiu Z., Zhang Y., Ren J., Tan P. & Ma R.Z. (2010) Closing a gap in the physical map of the ovine major histocompatibility complex. *Animal Genetics*. In press.

Llano M., Lee N., Navarro F., Garcia P., Albar J.P., Geraghty D.E. & Lopez-Botet M. (1998) HLA-E-bound peptides influence recognition by inhibitory and triggering CD94/NKG2 receptors: preferential response to an HLA-G-derived nonamer. *European Journal of Immunology* **28**, 2854-63.

Loke Y.W. & King A. (1991) Recent developments in the human maternal-fetal immune interaction. *Current Opinion in Immunology* **3**, 762-6.

Lomsadze A., Ter-Hovhannisyan V., Chernoff Y.O. & Borodovsky M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* **33**, 6494-506.

Lopez de Castro J.A., Strominger J.L., Strong D.M. & Orr H.T. (1982) Structure of crossreactive human histocompatibility antigens HLA-A28 and HLA-A2: possible implications for the generation of HLA polymorphism. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 3813-7.

Madden D.R., Garboczi D.N. & Wiley D.C. (1993) The antigenic identity of peptide-MHC complexes: A comparison of the conformations of five viral peptides presented by HLA-A2. *Cell* **75**, 693-708.

Mahdy E.A., Makinen A., Chowdhary B.P., Andersson L. & Gustavsson I. (1989) Chromosomal localization of the ovine major histocompatibility complex (OLA) by in situ hybridization. *Hereditas* **111**, 87-90.

Maillard J.C., Berthier D., Chantal I., Thevenon S., Sidibe I., Stachurski F., Belemsaga D., Razafindraibe H. & Elsen J.M. (2003) Selection assisted by a BoLA-DR/DQ haplotype against susceptibility to bovine dermatophilosis. *Genetics, Selection, Evolution* **35 Suppl 1**, S193-200.

Malfroy L., Roth M.P., Carrington M., Borot N., Volz A., Ziegler A. & Coppin H. (1997) Heterogeneity in Rates of Recombination in the 6-Mb Region Telomeric to the Human Major Histocompatibility Complex. *Genomics* **43**, 226-31.

Malissen M., Malissen B. & Bertrand R.J. (1982) Exon/Intron Organization and Complete Nucleotide Sequence of an HLA Gene. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 893-7.

Marck C. (1988) 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Research* **16**, 1829-36.

Martinsohn J.T., Sousa A.B., Guethlein L.A. & Howard J.C. (1999) The gene conversion hypothesis of MHC evolution: a review. *Immunogenetics* **50**, 168-200.

Matsumoto M., Zhou Y., Matsuo S., Nakanishi H., Hirose K., Oura H., Arase S., Ishida-Yamamoto A., Bando Y., Izumi K., Kiyonari H., Oshima N., Nakayama R., Matsushima A., Hirota F., Mouri Y., Kuroda N., Sano S. & Chaplin D.D. (2008) Targeted deletion of the murine corneodesmosin gene delineates its essential role in skin and hair physiology. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6720-4.

McGrath J.A. & Wessagowit V. (2005) Human hair abnormalities resulting from inherited desmosome gene mutations. *The Keio journal of Medicine* **54**, 72-9.

McRae A.F., McEwan J.C., Dodds K.G., Wilson T., Crawford A.M. & Slate J. (2002) Linkage Disequilibrium in Domestic Sheep. *Genetics* **160**, 1113-22.

McVean G. & Myers S. (2010) PRDM9 marks the spot. *Nature Genetics* **42**, 821-2.

Miller J.M., Poissant J., Kijas J.W., Coltman D.W. & the International Sheep Genomics C. (2011) A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. *Molecular Ecology Resources* **11**, 314-22.

Millot P. (1978) The major histocompatibility complex of sheep (OLA) and two minor loci. *Animal Blood Groups and Biochemical Genetics* **9**, 115-21.

Millot P. (1984) The OLA major histocompatibility complex of sheep. Study of six new factors and evidence of a third locus of the complex: OLA-C. *Experimental and Clinical Immunogenetics* **1**, 31-42.
Mils V., Vincent C., Croute F. & Serre G. (1992) The expression of desmosomal and corneodesmosomal antigens shows specific variations during the terminal differentiation of epidermis and hair follicle epithelia. *Journal of Histochemistry & Cytochemistry* **40**, 1329-37.
Miltiadou D., Ballingall K., Ellis S., Russell G. & McKeever D. (2005) Haplotype characterization of transcribed ovine major histocompatibility complex (MHC) class I genes. *Immunogenetics* **57**, 499-509.
Miretti M.M., Walsh E.C., Ke X., Delgado M., Griffiths M., Hunt S., Morrison J., Whittaker P., Lander E.S., Cardon L.R., Bentley D.R., Rioux J.D., Beck S. & Deloukas P. (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am J Hum Genet* **76**, 634-46.
Mizuno S., Trapani J.A., Koller B.H., Dupont B. & Yang S.Y. (1988) Isolation and nucleotide sequence of a cDNA clone encoding a novel HLA class I gene. *Journal of Immunology* **140**, 4024-30.
Momburg F., Roelse J., Hammerling G.J. & Neefjes J.J. (1994) Peptide size selection by major histocompatibility complex-encoded peptide transporter. *The Journal of Experimental Medicine* **179**, 1613-23.
Morrice N.A. & Powis S.J. (1998) A role for the thiol-dependent reductase ERp57 in the assembly of MHC class I molecules. *Current Biology* **8**, 713-6.
Nei M. & Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418-26.
Norris B.J., Colditz I.G. & Dixon T.J. (2008) Fleece rot and dermatophilosis in sheep. *Veterinary Microbiology* **128**, 217-30.
O'Callaghan C.A. (2000) Natural killer cell surveillance of intracellular antigen processing pathways mediated by recognition of HLA-E and Qa-1b by CD94/NKG2 receptors. *Microbes and Infection* **2**, 371-80.
O'Callaghan C.A. & Bell J.I. (1998) Structure and function of the human MHC class Ib molecules HLA-E, HLA-F and HLA-G. *Immunological Reviews* **163**, 129-38.
Odani M., Narita A., Watanabe T., Yokouchi K., Sugimoto Y., Fujita T., Oguni T., Matsumoto M. & Sasaki Y. (2006) Genome-wide linkage disequilibrium in two Japanese beef cattle breeds. *Animal Genetics* **37**, 139-44.
Ohta T. (1982) Allelic and nonallelic homology of a supergene family. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 3251-4.
Oliver P.L., Goodstadt L., Bayes J.J., Birtle Z., Roach K.C., Phadnis N., Beatson S.A., Lunter G., Malik H.S. & Ponting C.P. (2009) Accelerated Evolution of the <italic>Prdm9</italic> Speciation Gene across Diverse Metazoan Taxa. *PLoS Genetics* **5**, e1000753.
Olsen K.H., Grahn M., Lohm J. & Langefors A. (1998) MHC and kin discrimination in juvenile Arctic charr, Salvelinus alpinus (L.). *Anim Behaviour* **56**, 319-27.
Olsson M., Madsen T., Nordby J., Wapstra E., Ujvari B. & Wittsell H. (2003) Major histocompatibility complex and mate choice in sand lizards.

Proceedings. Biological sciences / The Royal Society **270 Suppl 2**, S254-6.

Orr H.T., Lopez de Castro J.A., Lancet D. & Strominger J.L. (1979) Complete amino acid sequence of a papain-solubilized human histocompatibility antigen, HLA-B7. 2. Sequence determination and search for homologies. *Biochemistry* **18**, 5711-20.

Orru S., Giuressi E., Carcassi C., Casula M. & Contu L. (2005) Mapping of the Major Psoriasis-Susceptibility Locus (PSORS1) in a 70-Kb Interval around the Corneodesmosin Gene (CDSN). *The American Journal of Human Genetics* **76**, 164-71.

Ortmann B., Androlewicz M.J. & Cresswell P. (1994) MHC class I/beta 2-microglobulin complexes associate with TAP transporters before peptide binding. *Nature* **368**, 864-7.

Ortmann B., Copeman J., Lehner P.J., Sadasivan B., Herberg J.A., Grandea A.G. & et al. (1997) A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes. *Science* **277**, 1306-9.

Page R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357-8.

Parham P., Adams E.J. & Arnett K.L. (1995) The origins of HLA-A,B,C polymorphism. *Immunological Reviews* **143**, 141-80.

Parham P., Lomen C.E., Lawlor D.A., Ways J.P., Holmes N., Coppin H.L., Salter R.D., Wan A.M. & Ennis P.D. (1988) Nature of polymorphism in HLA-A, -B, and -C molecules. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 4005-9.

Parham P. & Ohta T. (1996) Population Biology of Antigen Presentation by MHC Class I Molecules. *Science* **272**, 67-74.

Parnes J.R. (1989) Molecular biology and function of CD4 and CD8. *Advances in Immunology* **44**, 265-311.

Parvanov E.D., Petkov P.M. & Paigen K. (2010) Prdm9 Controls Activation of Mammalian Recombination Hotspots. *Science* **327**, 835.

Paterson S. & Pemberton J.M. (1997) No evidence for major histocompatibility complex-dependent mating patterns in a free-living ruminant population. *Proceedings. Biological Sciences / The Royal Society* **264**, 1813-9.

Perez-Villar J.J., Melero I., Navarro F., Carretero M., Bellon T., Llano M., Colonna M., Geraghty D.E. & Lopez-Botet M. (1997) The CD94/NKG2-A inhibitory receptor complex is involved in natural killer cell-mediated recognition of cells expressing HLA-G1. *The Journal of Immunology* **158**, 5736-43.

Peterson P.A., Cunningham B.A., Berggard I. & Edelman G.M. (1972) 2 -Microglobulin--a free immunoglobulin domain. *Proceedings of the National Academy of Sciences of the United States of America* **69**, 1697-701.

Peterson P.A., Rask L. & Lindblom J.B. (1974) Highly purified papain-solubilized HL-A antigens contain beta2-microglobulin. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 35-9.

Petroli C.s.D., Paiva S.R., CorrÃªa M.P.c.C. & McManus C. (2009) Genetic monitoring of a Santa Ines herd using microsatellite markers near or linked to the sheep MHC. *Revista Brasileira de Zootecnia* **38**, 670-5.

Pfeifer J.D., Wick M.J., Roberts R.L., Findlay K., Normark S.J. & Harding C.V. (1993) Phagocytic processing of bacterial antigens for class I MHC presentation to T cells. *Nature* **361**, 359-62.

Phillips M.S., Lawrence R., Sachidanandam R., Morris A.P., Balding D.J., Donaldson M.A., Studebaker J.F., Ankener W.M., Alfisi S.V., Kuo F.S., Camisa A.L., Pazorov V., Scott K.E., Carey B.J., Faith J., Katari G., Bhatti H.A., Cyr J.M., Derohannessian V., Elosua C., Forman A.M., Grecco N.M., Hock C.R., Kuebler J.M., Lathrop J.A., Mockler M.A., Nachtman E.P., Restine S.L., Varde S.A., Hozza M.J., Gelfand C.A., Broxholme J., Abecasis G.R., Boyce-Jacino M.T. & Cardon L.R. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genetics* **33**, 382-7.

Potter S.C., Clarke L., Curwen V., Keenan S., Mongin E., Searle S.M.J., Stabenau A., Storey R. & Clamp M. (2004) The Ensembl Analysis Pipeline. *Genome Research* **14**, 934-41.

Potts W.K., Manning C.J. & Wakeland E.K. (1991) Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature* **352**, 619-21.

Qin J., Mamotte C., Cockett N.E., Wetherall J.D. & Groth D.M. (2008) A map of the class III region of the sheep major histocompatibilty complex. *BMC Genomics* **9**, 409.

Qin J., Munyard K., Lee C.Y., Wetherall J.D. & Groth D.M. (2011) Characterization of the sheep Complement Factor B gene (CFB). *Veterinary Immunology and Immunopathology* **140**, 170-4.

Rafalski A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* **5**, 94-100.

Rammensee H.G., Friede T. & Stevanoviic S. (1995) MHC ligands and peptide motifs: first listing. *Immunogenetics* **41**, 178-228.

Raymond M. & Rousset F. (1995) GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *Journal of Heredity* **86**, 248-9.

Reits E., Neijssen J., Herberts C., Benckhuijsen W., Janssen L., Drijfhout J.W. & Neefjes J. (2004) A Major Role for TPPII in Trimming Proteasomal Degradation Products for MHC Class I Antigen Presentation. *Immunity* **20**, 495-506.

Rich R.R., Sedberry D.A., Kastner D.L. & Chu L. (1979) Primary in vitro cytotoxic response of NZB spleen cells to Qa-1b-associated antigenic determinants. *The Journal of Experimental Medicine* **150**, 1555-60.

Richardson D.S., Komdeur J., Burke T. & von Schantz T. (2005) MHC-based patterns of social and extra-pair mate choice in the Seychelles warbler. *Proceedings. Biological sciences / The Royal Society* **272**, 759-67.

Rock K.L., Gramm C., Rothstein L., Clark K., Stein R., Dick L., Hwang D. & Goldberg A.L. (1994) Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. *Cell* **78**, 761-71.

Roos M.H., Giles C.M., Demant P., Mollenhauer E. & Rittner C. (1984) Rodgers (Rg) and Chido (Ch) determinants on human C4: characterization of two C4 B5 subtypes, one of which contains Rg and Ch determinants. *Journal of Immunology* **133**, 2634-40.

Rousset F. (2008) genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-6.

Ruddy D.A., Kronmal G.S., Lee V.K., Mintier G.A., Quintana L., Domingo R., Jr., Meyer N.C., Irrinki A., McClelland E.E., Fullan A., Mapa F.A., Moore T., Thomas W., Loeb D.B., Harmon C., Tsuchihashi Z., Wolff R.K., Schatzman R.C. & Feder J.N. (1997) A 1.1-Mb Transcript Map of the Hereditary Hemochromatosis Locus. *Genome Research* **7**, 441-56.

Rufaut N.W., Pearson A.J., Nixon A.J., Wheeler T.T. & Wilkins R.J. (1999) Identification of Differentially Expressed Genes During a Wool Follicle Growth Cycle Induced by Prolactin. *Journal of Investigative Dermatology* **113**, 865-72.

Sadasivan B., Lehner P.J., Ortmann B., Spies T. & Cresswell P. (1996) Roles for Calreticulin and a Novel Glycoprotein, Tapasin, in the Interaction of MHC Class I Molecules with TAP. *Immunity* **5**, 103-14.

Salvi S., Sponza G., Morgante M., Tomes D., Niu X., Fengler K.A., Meeley R., Ananiev E.V., Svitashev S., Bruggemann E., Li B., Hainey C.F., Radovic S., Zaina G., Rafalski J.A., Tingey S.V., Miao G.H., Phillips R.L. & Tuberosa R. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 11376-81.

Saric T., Chang S.-C., Hattori A., York I.A., Markant S., Rock K.L., Tsujimoto M. & Goldberg A.L. (2002) An IFN-[gamma]-induced aminopeptidase in the ER, ERAP1, trims precursors to MHC class I-presented peptides. *Nature Immunology* **3**, 1169.

Saveanu L., Carroll O., Lindo V., Del V.M., D. L., Lepelletier Y., Greer F., Schomburg L., Fruci D., Niedermann G. & van Endert P.M. (2005) Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nature Immunology* **6**, 689.

Schenkel F.S., Miller S.P., Ye X., Moore S.S., Nkrumah J.D., Li C., Yu J., Mandell I.B., Wilton J.W. & Williams J.L. (2005) Association of single nucleotide polymorphisms in the leptin gene with carcass and meat quality traits of beef cattle. *Journal of Animal Science* **83**, 2009-20.

Schmitt-Egenolf M., Windemuth C., Hennies H.C., Albis-Camps M., Engelhardt B.v., Wienker T., Reis A., Traupe H. & Blasczyk R. (2001) Comparative association analysis reveals that corneodesmosin is more closely associated with psoriasis than HLA-Cw*0602-B*5701 in German families. *Tissue Antigens* **57**, 440-6.

Schnabel R.D., Kim J.J., Ashwell M.S., Sonstegard T.S., Van Tassell C.P., Connor E.E. & Taylor J.F. (2005) Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 6896-901.

Schook L.B. & Lamony S.J. (1996) *The Major Histocompatibility Complex Region of Domestic Animal Species*. CRC Press, Boca Raton.

Schwaiger F.W., Buitkamp J., Weyers E. & Epplen J.T. (1993) Typing of artiodactyl MHC-DRB genes with the help of intronic simple repeated DNA sequences. *Molecular Ecology* **2**, 55-9.

Schwaiger F.W., Maddox J., Ballingall K., Buitkamp J., Crawford A.M., Dutia B.M., Epplen J.T., Ferguson E.D., Groth D., Hopkins J., Rhind S.M., Sargan D., Wetherall J. & Wright H. (1996) *The ovine major histocompatibility complex. In: The major histocompatibility complex region of domestic animal species*. CRC Press, Inc., Boca Raton, FL, USA.

Serwold T., Gonzalez F., Kim J., Jacob R. & Shastri N. (2002) ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. *Nature* **419**, 480-3.

Shawar S.M., Vyas J.M., Rodgers J.R. & Rich R.R. (1994) Antigen Presentation by Major Histocompatibility Complex Class I-B Molecules. *Annual Review of Immunology* **12**, 839-80.

Shimizu Y., Geraghty D.E., Koller B.H., Orr H.T. & DeMars R. (1988) Transfer and expression of three cloned human non-HLA-A,B,C class I major histocompatibility complex genes in mutant lymphoblastoid cells. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 227-31.

Shimomura Y. & Christiano A.M. (2010) Biology and Genetics of Hair. *Annual Review of Genomics and Human Genetics* **11**, 109-32.

Simon M., Jonca N., Guerrin M., Haftek M., Bernard D., Caubet C.c., Egelrud T.r., Schmidt R. & Serre G. (2001) Refined Characterization of Corneodesmosin Proteolysis during Terminal Differentiation of Human Epidermis and Its Relationship to Desquamation. *Journal of Biological Chemistry* **276**, 20292-9.

Singh P.B., Brown R.E. & Roser B. (1987) MHC antigens in urine as olfactory recognition cues. *Nature* **327**, 161-4.

Smith J.M. & Haigh J. (1974) The hitch-hiking effect of a favourable gene. *Genetics Research* **23**, 23-35.

Smithies O. & Poulik M.D. (1972) Initiation of protein synthesis at an unusual position in an immunoglobulin gene? *Science* **175**, 187-9.

Sobrino B., Brion M. & Carracedo A. (2005) SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Science International* **154**, 181-94.

Sole X., Guino E., Valls J., Iniesta R. & Moreno V. (2006) SNPStats: a web tool for the analysis of association studies. *Bioinformatics* **22**, 1928-9.

Srivastava R., Chorney M.J., Lawrance S.K., Pan J., Smith Z., Smith C.L. & Weissman S.M. (1987) Structure, expression, and molecular mapping of a divergent member of the class I HLA gene family. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 4224-8.

Srivastava R., Duceman B.W., Biro P.A., Sood A.K. & Weissman S.M. (1985) Molecular Organization of the Class I Genes of Human Major Histocompatibility Complex. *Immunological Reviews* **84**, 93-122.

Srivastava R. & Lambert M.E. (1991) Molecular organization and expression of class I HLA gene family. In *Immunogenetics of the Major Histocompatibility Complex*, ed. Srivastava R, Ram BP, Tyle P, pp. 100–54. New York: VCH Publ.

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-W9.

Stear M.J. & Spooner R.L. (1981) Lymphocyte antigens in sheep. *Animal Blood Groups and Biochemical Genetics* **12**, 265-76.

Stenzel A., Lu T., Koch W.A., Hampe J., Guenther S.M., De La Vega F.M., Krawczak M. & Schreiber S. (2004) Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. *Human Genetics* **114**, 377-85.

Stephens R., Horton R., Humphray S., Rowen L., Trowsdale J. & Beck S. (1999) Gene organisation, sequence variation and isochore structure at the centromeric boundary of the human MHC. *Journal of Molecular Biology* **291**, 789-99.

Stewart C.A., Horton R., Allcock R.J.N., Ashurst J.L., Atrazhev A.M., Coggill P., Dunham I., Forbes S., Halls K., Howson J.M.M., Humphray S.J., Hunt S., Mungall A.J., Osoegawa K., Palmer S., Roberts A.N., Rogers J., Sims S., Wang Y., Wilming L.G., Elliott J.F., de Jong P.J., Sawcer S., Todd J.A., Trowsdale J. & Beck S. (2004) Complete MHC Haplotype Sequencing for Common Disease Gene Mapping. *Genome Research* **14**, 1176-87.

Stoltze L., Schirle M., Schwarz G., Schröter C., Thompson M.W., Hersh L.B., Kalbacher H., Stevanovic S., Rammensee H.-G. & Schild H. (2000) Two new proteases in the MHC class I processing pathway. *Nature Immunology* **1**, 413.

Stroynowski I. (1990) Molecules related to class-I major histocompatibility complex antigens. *Annual Review of Immunology* **8**, 501-30.

Takahata N. (1995) A Genetic Perspective on the Origin and History of Humans. *Annual Review of Ecology and Systematics* **26**, 343-72.

Takahata N. & Nei M. (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* **124**, 967-78.

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**, 2731-9.

Tan P., Kropshofer H., Mandelboim O., Bulbuc N., Hammerling G.J. & Momburg F. (2002) Recruitment of MHC Class I Molecules by Tapasin into the Transporter Associated with antigen Processing-Associated Complex Is Essential for Optimal Peptide Loading. *Immunology* **168**, 1950-60.

Tazi-Ahnini R., Henry J., Offer C., Bouissou-Bouchouata C., Mather I.H. & Pontarotti P. (1997) Cloning, localization, and structure of new members of the butyrophilin gene family in the juxta-telomeric region of the major histocompatibility complex. *Immunogenetics* **47**, 55-63.

Tazi Ahnini R., Camp N.J., Cork M.J., Mee J.B., Keohane S.G., Duff G.W. & di Giovine F.S. (1999a) Novel genetic association between the corneodesmosin (MHC S) gene and susceptibility to psoriasis. *Human Molecular Genetics* **8**, 1135-40.

Tazi Ahnini R., di Giovine F.S., Cox A., Keohane S.G. & Cork M.J. (1999b) Corneodesmosin (MHC S) gene in guttate psoriasis. *The Lancet* **354**, 597-.

Teh H.S., Kisielow P., Scott B., Kishi H., Uematsu Y., Bluthmann H. & von Boehmer H. (1988) Thymic major histocompatibility complex antigens and the alpha beta T-cell receptor determine the CD4/CD8 phenotype of T cells. *Nature* **335**, 229-33.

Terasaki P.I. & McClelland J.D. (1964) Microdroplet Assay Of Human Serum Cytotoxins. *Nature* **204**, 998-1000.

Thomas J.H., Emerson R.O. & Shendure J. (2009) Extraordinary Molecular Evolution in the PRDM9 Fertility Gene. *PLoS ONE* **4**, e8505.

Thompson J.D., Higgins D.G. & Gibson T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673-80.

Townsend A. & Bodmer H. (1989) Antigen Recognition by Class I-Restricted T Lymphocytes. *Annual Review of Immunology* **7**, 601-24.

Tragardh L., Rask L., Wiman K., Fohlman J. & Peterson P.A. (1979) Amino Acid Sequence of an Immunoglobulin-Like HLA Antigen Heavy Chain

Domain. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 5839-42.
Traherne J.A., Horton R., Roberts A.N., Miretti M.M., Hurles M.E., Stewart C.A., Ashurst J.L., Atrazhev A.M., Coggill P., Palmer S., Almeida J., Sims S., Wilming L.G., Rogers J., de Jong P.J., Carrington M., Elliott J.F., Sawcer S., Todd J.A., Trowsdale J. & Beck S. (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genetics* **2**, e9.
Trowsdale J. (1993) Genomic structure and function in the MHC. *Trends in Genetics* **9**, 117-22.
Trowsdale J. (1995) "Both man & bird & beast": comparative organization of MHC genes. *Immunogenetics* **41**, 1-17.
Trowsdale J. (2001) Genetic and functional relationships between MHC and NK receptor genes. *Immunity* **15**, 363-74.
Urabe K., Kimura A., Harada F., Iwanaga T. & Sasazuki T. (1990) Gene conversion in steroid 21-hydroxylase genes. *American Journal of Human Genetics* **46**, 1178-86.
Vaiman M., Chardon P. & Rothschild M.F. (1998) Porcine major histocompatibility complex. *Revue Scientifique et Technique* **17**, 95-107.
Van Liere J.M. & Rosenberg N.A. (2008) Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical Population Biology* **74**, 130-7.
Wedekind C., Seebeck T., Bettens F. & Paepke A.J. (1995) MHC-dependent mate preferences in humans. *Proceedings. Biological sciences / The Royal Society* **260**, 245-9.
Wegmann T.G. (1984) Foetal protection against abortion: is it immunosuppression or immunostimulation? *Ann Immunol (Paris)* **135D**, 309-12.
Wei X.H. & Orr H.T. (1990) Differential expression of HLA-E, HLA-F, and HLA-G transcripts in human tissue. *Human Immunology* **29**, 131-42.
Westerdahl H. (2004) No evidence of an MHC-based female mating preference in great reed warblers. *Molecular Ecology* **13**, 2465-70.
White P.C., Tusie-Luna M.T., New M.I. & Speiser P.W. (1994) Mutations in steroid 21-hydroxylase (CYP21). *Human Mutation* **3**, 373-8.
Williams A.F. (1987) A year in the life of the immunoglobulin superfamily. *Immunology Today* **8**, 298-303.
Worley K., Carey J., Veitch A. & Coltman D.W. (2006) Detecting the signature of selection on immune genes in highly structured populations of wild sheep (Ovis dalli). *Molecular Ecology* **15**, 623-37.
Wu C., McConnell I. & Blacklaws B. (2008) Cloning and characterization of ovine beta2-microglobulin cDNAs. *Veterinary Immunology and Immunopathology* **123**, 360-5.
Yokoyama K., Geier S.S., Uehara H. & Nathenson S.G. (1985) Secondary structure of the murine histocompatibility alloantigen H-2Kb: relationship between heavy chain, beta 2-microglobulin, and antigenic reactivity. *Biochemistry* **24**, 3002-6.
Yoshino M., Xiao H., Jones E.P., Kumanovics A., Amadou C. & Fischer L.K. (1997) Genomic evolution of the distal Mhc class I region on mouse Chr 17. *Hereditas* **127**, 141 - 8.
Yu C.Y. (1991) The complete exon-intron structure of a human complement component C4A gene. DNA sequences, polymorphism, and linkage to the 21-hydroxylase gene. *Journal of Immunology* **146**, 1057-66.

Zhou Y. & Chaplin D.D. (1993) Identification in the HLA class I region of a gene expressed late in keratinocyte differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 9470-4.

Zinkernagel R.M. & Doherty P.C. (1974) Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytic choriomeningitis. *Nature* **251**, 547-8.

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 acknowledge.

Appendix A

General buffers and reagents

1x TAE or TBE buffer

Mixed 20 mL of 50x TAE buffer with 980 mL of hpH_2O to make up to a final volume of 1L.

1x TBE buffer

Mixed 100 mL of 10x TBE buffer with 900 mL of hpH_2O to make up to a final volume of 1L.

10x TBE buffer

162 g Tris
9.3 g $EDTA.2H_2O$
27.5 g Boric acid

Dissolved in hpH_2O and adjusted the volume to 1 L.

50x TAE buffer

242 g Tris
37.2 g $Na_2EDTA.2H_2O$
57.1 mL Glacial acetic acid

Dissolved in hpH_2O and adjusted the volume to 1 L.

500 mM EDTA

208.1 g EDTA

Dissolve in 800 mL of hpH_2O and adjusted the pH to 8.0. Made up to 1 L with hpH_2O. Sterilised by autoclaving.

Loading buffer

12.5 mg Bromophenol blue
12.5 mg Xylene cyanol
1.5 mL Glycerol

Dissolved in hpH_2O, and adjusted to 5 mL.

TE buffer

1.21 g Tris-HCl
0.372 g EDTA
Dissolved in hpH_2O, and adjusted pH to 8.0. Adjusted volume to 1 L.

Luria Bertani agar

10 g NaCl
5 g Yeast extract
10 g Bacto-Tryptone
15 g Agar (1.5 g into 100 mL)

Dissolved in 800 mL of hpH_2O. Adjusted pH to 7.5. Made up to 1 L with hpH_2O. The broth was divided into 100 mL aliquots and 1.5 g of agar was added to each aliquot. The broth was sterilised by autoclaving and stored 4 ^{o}C.

Luria Bertani broth

10 g NaCl
5 g Yeast extract
10 g Bacto-Tryptone

Dissolved in 800 mL of hpH_2O. Adjusted pH to 7.5. Made up to final volume of 1 L with hpH_2O. The broth was divided into 100 mL aliquots, sterilised by autoclaving and stored 4 ^{o}C.

Appendix B

Published Paper

Predictive mutational bioinformatic analysis of variation in the skin and wool associated corneodesmosin (CDSN) gene in sheep.

Siva Subramaniam N., Morgan E., Bottomley S., Tay S., Gregg K., Lee C.Y., Wetherall J. & Groth D.
Animal Science Journal (2011); doi: 10.1111/j.740-0929.2011.00975.x.

Animal Science Journal

Animal Science Journal (2011) doi: 10.1111/j.1740-0929.2011.00975.x

ORIGINAL ARTICLE

Predictive mutational bioinformatic analysis of variation in the skin and wool associated corneodesmosin (CDSN) gene in sheep

Nitthiya SIVA SUBRAMANIAM,[1] Eleanor MORGAN,[1] Steven BOTTOMLEY,[1] Sharon TAY,[2]* Keith GREGG,[1] Chee Yang LEE,[1] John WETHERALL[1] and David GROTH[1]

[1]*Western Australian Biomedical Research Institute (WABRI) & Centre for Health Innovation Research Institute (CHIRI), School of Biomedical Sciences, Curtin University, Perth, and* [2]*School of Biological Sciences, Murdoch University, Murdoch, Australia*

ABSTRACT

Corneodesmosin (CDSN) is an important component of the desmosome in the epidermal cornified stratum and inner root sheath of hair follicles. DNA from a sheep BAC clone previously identified by us to contain *CDSN* was PCR amplified using cattle-derived primers and the product sequenced. A region of 4579 bp containing *CDSN* was shown to contain two exons separated by one intron and spanning 3683 bp. The DNA encodes a predicted protein of 546 amino acids. Phylogenetic analysis shows that sheep CDSN falls within a clade containing cattle and other ruminant-like species. Comparison of sequences generated from 12 unrelated merino sheep and the International Sheep Genome Consortium (ISGC) data identified 58 single nucleotide polymorphisms (SNPs) within the 4579 bp region of which 16 are contained within coding sequences (1 in 80 bp). The SNPs identified in this study will add to the Major Histocompatibility Complex (MHC) SNP panel, which will allow extensive haplotyping of the sheep MHC in future studies.

Key words: *corneodesmosin, genetics, MHC, polymorphism, sheep.*

INTRODUCTION

Cornification is a late stage of epidermal differentiation, transforming keratinocytes into anucleated, flattened corneocytes (Guerrin *et al.* 1998). Corneodesmosomes, the altered desmosomes of the uppermost epidermal layer, form specialized keratinocyte intercellular junctions and play an essential role in cornoecyte cohesion (Guerrin *et al.* 1998). Accumulation of corneocyte forms the cornified layer of the epidermis, or stratum corneum, and provides a physical barrier from the environment (Montezin *et al.* 1997; Guerrin *et al.* 1998). The transmembrane corneodesmosomal proteins consist of desmoglein 1 (Dsg 1), desmocollin and corneodesmosin (CDSN) (Matsumoto *et al.* 2008).

Human *CDSN*, initially designated as the *S gene*, is located 160 kb telomeric of HLA-C (6p21.3) and encodes a 52- to 56-kDa basic glycoprotein specific to the cornified epithelia and inner hair follicle root sheath (Zhou & Chaplin 1993; Jonca *et al.* 2002). Human *CDSN* is a 529 amino acid protein with very high serine and glycine content of 27.5% and 16%, respectively, a feature shared with several other epidermal proteins (Guerrin *et al.* 1998; Simon *et al.* 2001; Jonca *et al.* 2002).

A study in New Zealand Wiltshire sheep focused on identifying genes previously not suspected of having a role in wool physiology, resulted in identification of a desmosome component (Rufaut *et al.* 1999). This study in sheep showed the involvement of desmosome in the wool follicle growth cycle. Another study in Merino sheep has shown association between MHC and wool production traits (Bot 2000). Therefore, it can be hypothesized that the major histocompatibility complex (MHC) located *CDSN* gene, which is involved in skin and hair physiology could be associated with wool production traits in sheep.

Correspondence: David Groth, School of Biomedical Sciences, Curtin University, Perth 6845, Australia. (Email: d.groth@curtin.edu.au)
*Present address: School of Animal Biology, University of Western Australia, Perth, Australia.
Received 29 October 2010; accepted for publication 15 June 2011.

In this study we intended to amplify and sequence the entire sheep *CDSN* gene and compare the sheep sequence with other species. Furthermore, we have characterized and identified single nucleotide polymorphisms (SNPs) within the gene and these polymorphisms can be used for defining MHC haplotypes.

MATERIALS AND METHODS

All animal experiments were performed according to the Australian Code of Practice for the care and use of animals for scientific purposes. Blood samples were collected under approval of Curtin University's Animal Ethics Committee.

The Basic Local Alignment Search Tool (BLAST) program (Altschul *et al.* 1990) was used for comparative analysis of *CDSN* DNA sequences from cattle (NW_001494146), dog (NW_876254), gray mouse lemur (AB480748), gray short-tailed opossum (NW_001581878), horse (NW_001867389), human (NW_001838980) and pig (NW_001886435).

Based on the alignment of cattle *CDSN* gene sequence with the other sequences given above, primers were designed to amplify a series of 500 bp fragments, with an overlap of 100 bp (Supporting Table S1) spanning *CDSN*. These primers were used for PCR amplification and sequencing. Primers were synthesized by either Geneworks (Adelaide, SA, Australia) or Invitrogen (Melbourne, Vic., Australia).

The BAC clone (CHORI 243–269 M18), used in this study was originally identified to contain *CDSN* by J. Qin/D. Groth (20 October, 2005). The BAC library (CHORI-243) was constructed by the Children's Hospital Oakland Research Institute (CHORI). The BAC clone was extracted using the manufacturer's standard protocol (Qiagen Large-Construct Kit; Melbourne, Vic., Australia). Plasmid vector (pGEM3Zf-) used for subcloning PCR product was purchased from Promega Corporation (Madison, WI, USA).Plasmid vectors pGEM-3Zf(-) were purified from bacterial cultures using the QIAprep Spin Miniprep Kit (Qiagen).

PCR was performed in 10 µL reactions, comprising 10 ng of DNA, 200 µmol of each deoxynucleotide triphosphate (dNTP) (Invitrogen), 1.5 mmol $MgCl_2$, 1 × reaction buffer (Taq Platinum, Invitrogen), 0.5 units of Taq DNA polymerase (Taq Platinum, Invitrogen) and 10 pmol of each primer. PCR was performed on an Eppendorf Mastercycler (Eppendorf, Sydney, NSW, Australia). The annealing temperature (T_{ann}) used for the PCR reactions varied according to the calculated melting temperature of the primer sets used. The typical cycling conditions were as follows: 95°C for 10 min; 30 cycles of 94°C for 30 s; T_{ann} for 30 s; and 72°C for 30 s; and a final extension step at 72°C for 5 min.

The PCR products from amplified BAC DNA were cleaned with ExoSAP protocol (New England Biolabs, Brisbane, Qld., Australia), electrophoresed in a 1 × Tris-acetate-ethylenediaminetetraacetic acid agarose gel and the DNA visualized by ethidium bromide (15 µg/mL) staining. Purified PCR products were sequenced with primers designed from the cattle *CDSN* gene sequence. Sequencing reactions were performed at Macrogen Inc. (Seoul, South Korea). PCR products that were difficult to sequence were subcloned into pGEM3Zf-. Insert-containing clones were extracted using the QIAprep Spin Miniprep Kit and sequenced.

The PCR-derived BAC sequences were assembled using VectorNTI (Invitrogen). The resulting sequence was used to design additional primers specific for sheep *CDSN* gene. For amplification of sheep genomic DNA, second-pass sequencing primers of 20 bp were designed based upon the assembled first-pass BAC contigs. The second-pass primers were optimised to amplify regions of 500 bp to 1000 bp in length. Genomic DNA from 12 merino sheep samples was independently amplified. The PCR products were purified with ExoSAP protocol (Biolab), followed by sequencing at Macrogen Inc (Seoul, South Korea). The resulting sequences were interrogated using VectorNTI software to identify the presence of any prominent SNP variation. Sheep *CDSN* and each of the PCR fragments with potential SNPs were aligned to identify the region and the exact location of the SNPs with respect to the sheep *CDSN* start codon. The sequence was also aligned with data generated by the International Sheep Genome Consortium (ISGC) (https://isgcdata.agresearch.co.nz/). The SNPs identified within *CDSN* coding sequence (CDs) were used to determine possible amino acid (A.A.) changes. Synonymous and Non-synonymous Analysis Program (SNAP) analysis (http://www.hiv.lanl.gov and http://hcv.lanl.gov/content/sequence/SNAP/SNAP.html) (Nei & Gojobori 1986; Korber 2000) was performed to determine the rate of synonymous and non-synonymous change in multiple sequences.

Gene prediction tools such as GENSCAN (Burge & Karlin 1997), FGENESH (http://www.softberry.com), HMM gene (http://www.cbs.dtu.dk/services/HMMgene/), Augustas (Stanke *et al.* 2006) and GeneMark (Lomsadze *et al.* 2005) were used to determine the number of exons, messenger RNA (mRNA) and amino acid sequence from the DNA sequence generated from the assembly of the contigs by VectorNTI. Basic pairwise alignment, ExPASy (Swiss Institute of Bioinformatics) and GAP (Pairwise Sequence Alignment Server) were also used to analyze the sequence. The resulting structure of sheep *CDSN* gene was compared with the structure of cattle (NW_001494164), chimpanzee (NW_001236523), dog (NW_876254), gray mouse lemur (AB480748), gray short-tailed opossum (NW_001581878), horse (NW_001867389), human (NW_001838980), mouse (NT_039663), pig (NW_001886435), rat (NW_001084776) and rhesus monkey (NW_001116482). ClustalW (Thompson *et al.* 1994) and GBlocks (Talavera & Castresana 2007) were used in combination for multiple sequence alignment of the sheep *CDSN* with other mammals. Analyses were performed for both mRNA and amino acid sequences. The resulting alignments were used for phylogenetic analyses of the *CDSN*.

Non-synonymous SNPs were analyzed using automated methods that derive a consensus view on the potential effect of the observed amino acid substitutions on CDSN structure or function. In particular, we used the web server versions of the iPTree-STAB algorithm (Huang *et al.* 2007) and 'iMutant' (Capriotti *et al.* 2004, 2005, 2008). Both of these methods use a thermodynamic approach to predict the effect of amino acid substitutions on protein stability (Gromiha 2007). iMutant also uses a binary classification of stability called SVM2 (Support Vector Machine 2) or ternary classification of stability called SVM3 (support Vector Machine 3).

RESULTS

The overview of the *CDSN* gene was obtained by comparative analyses of several other mammalian organisms such as dog (NW_876254), gray short-tailed opossum (NW_001581878), gray mouse lemur (AB480748), horse (NW_001867389), human (NW_001838980) and pig (NW_001886435). The basic structure of the *CDSN* gene in all the organisms

was similar to the cattle gene with two exons and a large intron. Variation in the size of the gene ranged from 3302 bp in pig to 5350 bp in human.

Assembly of triple pass DNA sequencing of BAC and sheep genomic DNA with VectorNTI resulted in a contig fragment of 4579 bp in length. The sheep *CDSN* gene was located between positions 159 bp and 3841 bp within the fragment. The results obtained using various gene prediction tools were compared and a consensus sheep *CDSN* gene sequence generated. Comparative analysis of the sheep *CDSN* gene showed that the gene has two exons of 85 bp and 1553 bp, respectively, and an intron of 2045 bp. The sheep *CDSN* gene is 3683 bp in length and encodes a protein of 545 amino acids. Supporting Figure S1 shows the nucleotide sequence of sheep *CDSN* gene, and Supporting Figure S2 shows the predicted peptide sequence.

Sequence analysis resulted in identification of 51 SNPs within the genomic DNA from 112 Merino sheep. Alignment of the 4579 bp gene sequence with other breeds of sheep from 454 data revealed an additional seven SNPs. Sixteen SNPs identified with sequence analysis were also identified in 454 data. In total, 58 SNPs were identified within the entire fragment. Sixteen SNPs are located within the coding sequence of the *CDSN* gene. The other 30 and 12 SNPs are located within the intron 1 of *CDSN* gene and after the stop codon, respectively. The frequency of SNP in sheep *CDSN* is approximately 1/80 bp. The details of all the SNPs identified are shown in Table 1.

The 16 SNPs identified within the coding sequence of *CDNS* gene resulted in eight synonymous and eight non-synonymous changes. Table 2 shows the details of amino acid changes. The statistics of SNAP analysis of pairwise comparisons of *CDSN* for all organisms, which takes into account Jukes-Cantor correction, showed that the average ds/dn value is 1.4426 with ds and dn of 1.9007 and 1.4149, respectively (Supplementary Table 2).

Multiple alignments of amino acid sequences showed that the sheep *CDSN* gene has 92% DNA sequence identity with the cattle gene in comparison to 69% observed with that of human. The percent identity of amino acid sequence in the sheep protein compared to other species ranges from 58% in gray short-tailed opossum to 84% in pig. Table 3 shows the details of percent identity of sheep CDSN amino acid sequence compared with other species. The phylogenetic tree (Fig. 1) generated from multiple alignment of CDSN amino acid sequence from various species showed significant identity and confidence levels between sheep and cattle. The overall topology of the tree showed three major clades in which the CDSN sequence from various organisms were grouped: primates, rodents and ruminant with other higher order mammals.

Table 1 Single nucleotide polymorphisms (SNPs) indentified in within and outside sheep *CDSN* gene

Location in the fragment (bp)	Location within *CDSN* (bp)	Base change	Description
241	83	A/G	Exon 1
267	109	C/T	Intron 1
748	590	A/G	Intron 1
823	665	C/T	Intron 1
929†	771	A/G	Intron 1
1029†	871	A/G	Intron 1
1246	1088	G/T	Intron 1
1355	1097	C/G	Intron 1
1389	1231	A/G	Intron 1
1416	1258	C/G	Intron 1
1418	1260	A/C	Intron 1
1440‡	1282	G/C	Intron 1
1526†	1368	C/T	Intron 1
1562†	1404	C/T	Intron 1
1566†	1408	A/G	Intron 1
1600	1442	C/T	Intron 1
1643†	1485	A/G	Intron 1
1681†	1523	C/T	Intron 1
1711	1553	C/T	Intron 1
1753	1595	C/T	Intron 1
1782	1624	C/T	Intron 1
1789	1631	C/T	Intron 1
1799	1641	C/T	Intron 1
1825	1667	G/T	Intron 1
1848	1690	G/T	Intron 1
1978†	1820	G/T	Intron 1
1993†	1835	C/T	Intron 1
2087†	1929	C/T	Intron 1
2097‡	1939	G/T	Intron 1
2149†	1991	C/T	Intron 1
2160†	2002	G/T	Intron 1
2405	2247	A/G	Exon 2
2433†	2275	C/G	Exon 2
2536	2378	C/T	Exon 2
2538	2380	A/G	Exon 2
2539	2381	C/T	Exon 2
2611	2453	C/T	Exon 2
2630†	2472	A/G	Exon 2
2736	2578	A/G	Exon 2
2756†	2598	C/T	Exon 2
2827	2669	A/C	Exon 2
2845†	2687	C/T	Exon 2
2968‡	2810	C/T	Exon 2
3262‡	3104	C/T	Exon 2
3502‡	3344	A/G	Exon 2
3608	3450	A/G	Exon 2
3873	3715	C/T	After stop codon
3957	3799	G/T	After stop codon
4160	4002	C/T	After stop codon
4161	4003	A/G	After stop codon
4174	4016	C/T	After stop codon
4213	4055	C/T	After stop codon
4222	4064	A/G	After stop codon
4303	4145	C/T	After stop codon
4384‡	4226	A/C	After stop codon
4427‡	4269	C/T	After stop codon
4537	4379	A/G	After stop codon
4557	4399	C/T	After stop codon

†SNP identified in both genomic sheep sequences and International Sheep Genomics Consortium data. ‡SNP identified only in 454 data.

Table 2 Location of Single nucleotide polymorphisms (SNPs) in coding sequence and the corresponding amino acid changes within sheep CDSN protein sequence

Location in CDs	Change of base	Location in AAs	Details of AA substitution		Type of mutation
			Change of AA	AA classification	
83	A/G	28	Glutamine (Q) Arginine (R)	Polar, neutral Polar, positively charged	Non-synonymous
202	A/G	68	Serine (S) Glycine (G)	Polar, neutral Nonpolar	Non-synonymous
230	C/G	77	Serine (S) Threonine (T)	Polar, neutral Polar, neutral	Non-synonymous
333	C/T	111	Glycine (G) Glycine (G)	Nonpolar Nonpolar	Synonymous
335	A/G	112	Histidine (H) Arginine (R)	Polar, positively charged Polar, positively charged	Non-synonymous
336	C/T	112	Histidine (H) Histidine (H)	Polar, positively charged Polar, positively charged	Synonymous
			Arginine (R) Arginine (R)	Polar, positively charged Polar, positively charged	Synonymous
408	C/T	136	Glycine (G) Glycine (G)	Nonpolar Nonpolar	Synonymous
427	A/G	143	Glycine (G) Serine (S)	Nonpolar Polar, neutral	Non-synonymous
533	A/G	178	Aspartic acid (D) Glycine (G)	Polar, negatively charged Nonpolar	Non-synonymous
553	C/T	185	Proline (P) Serine (S)	Nonpolar Polar, neutral	Non-synonymous
624	A/C	208	Threonine (T) Threonine (T)	Polar, neutral Polar, neutral	Synonymous
642	C/T	214	Serine (S) Serine (S)	Polar, neutral Polar, neutral	Synonymous
765†	C/T	255	Serine (S) Serine (S)	Polar, neutral Polar, neutral	Synonymous
1059†	C/T	353	Serine (S) Serine (S)	Polar, neutral Polar, neutral	Synonymous
1299†	A/G	433	Glycine (G) Glycine (G)	Nonpolar Nonpolar	Synonymous
1405	A/G	469	Serine (S) Glycine (G)	Polar, neutral Nonpolar	Non-synonymous

†SNP identified only in International Sheep Genomics Consortium data.

Table 3 Percent identity of sheep CDSN amino acid sequence compared with other species, created with ClustalX

	HOSA	PATR	MAMU	EQCA	CAFA	OVAR	BOTA	SUSC	MIMU	MUMU	RANO	MODO
HOSA	100	99	93	79	80	69	69	75	78	66	66	59
PATR	99	100	94	80	80	69	69	75	78	66	66	59
MAMU	93	94	100	80	82	69	69	76	77	66	67	59
EQCA	79	80	80	100	87	71	72	81	77	67	67	58
CAFA	80	80	82	87	100	75	76	82	81	69	66	60
OVAR	69	69	69	71	75	100	92	84	69	61	59	58
BOTA	69	69	69	72	76	92	100	85	70	62	60	59
SUSC	75	75	76	81	82	84	85	100	75	64	64	60
MIMU	78	78	77	77	81	69	70	75	100	66	66	59
MUMU	66	66	66	67	69	61	62	64	66	100	89	55
RANO	66	66	67	67	66	59	60	64	66	89	100	54
MODO	59	59	59	58	60	58	59	60	59	55	54	100

OVAR: sheep CDSN, BOTA: cattle (NW_001494164), PATR: chimpanzee (NW_001236523), CAFA: dog (NW_876254), MIMU: gray mouse lemur (AB480748), MODO: gray short-tailed opossum (NW_001581878), EQCA: horse (NW_001867389), HOSA: human (NW_001838980), MUMU: mouse (NT_039663), SUSC: pig (NW_001886435), RANO: rat (NW_001084776) and MAMU: rhesus monkey (NW_001116482).

All except three (S68G, D178G, P185S) of the non-synonymous SNPs seen in this protein are generally conservative amino acid substitutions according to their physicochemical properties and the Blosum62 evolutionary matrix (Henikoff & Henikoff 1992). The thermodynamic approach of the iPTree-STAB algorithm shows that all, except the substitution G143S, were destablizing to the protein structure (Table 4). However, according to SVM3, most substitutions, except D178G and S469G, were neutral to the stability of protein structure (Table 4). Furthermore, according to the SVM2 classification, all of the substitutions appeared to be destabilizing to the protein structure. The S469G substitution was classified by the iMutant algorithm as a potential 'disease-related mutation' based on comparisons with other genes associated with disease. A diagrammatic comparison of the SNP in the coding sequence between human and sheep is shown in Figure 2.

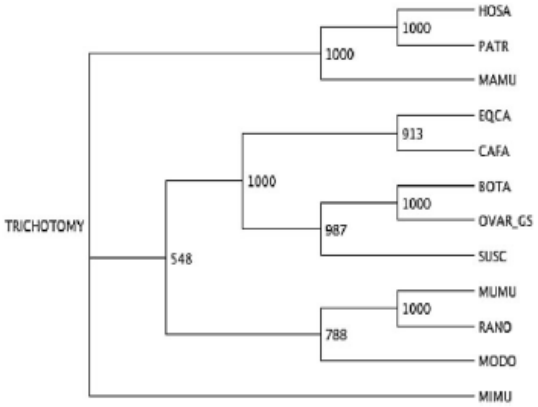


Figure 1 Neighbor-joining tree constructed using ClustalX after aligning amino acid sequences (default parameter settings). 1000 bootstraps. OVAR_GS = Genscan prediction for sheep CDSN. BOTA: cattle (NW_001494164), PATR: chimpanzee (NW_001236523), CAFA: dog (NW_876254), MIMU: gray mouse lemur (AB480748), MODO: gray short-tailed opossum (NW_001581878), EQCA: horse (NW_001867389), HOSA: human (NW_001838980), MUMU: mouse (NT_039663), SUSC: pig (NW_001886435), RANO: rat (NW_001084776) and MAMU: rhesus monkey (NW_001116482).

DISCUSSION

The 1/80 bp SNP frequency of sheep *CDSN* indicates that the gene is highly polymorphic (Guerrin *et al.* 2001; Capon *et al.* 2003). The high content of serine and glycine predicted in the protein is similar to that reported previously for human CDSN (Guerrin *et al.* 1998; Simon *et al.* 2001; Jonca *et al.* 2002). There were 16 SNPs which were identified to be within the coding sequence of *CDSN* in this study, some of which may have an impact upon the protein structurally. There was an additional 42 SNPs identified that may contribute in future to a comprehensive analysis of sheep MHC haplotypes. These SNPs located in the gene-rich MHC region can serve as a mapping tool for sheep traits and diseases involving skin. Studies in humans have shown that *CDSN* gene polymorphisms are associated with psoriasis, a chronic inflammatory and hyperproliferative skin disease (Allen *et al.* 1999; Jenisch *et al.* 1999; Tazi Ahnini *et al.* 1999; Schmitt-Egenolf *et al.* 2001; Capon *et al.* 2003; McGrath & Wessagowit 2005; Orru *et al.* 2005; Matsumoto *et al.* 2008). The initial genetic study linking the pathogenesis of human psoriasis to *CDSN*, identified nine SNPs within the gene (Ishihara *et al.* 1996). More recent studies have identified a total of 23 SNPs and three trinucleotide indels in the exons of human *CDSN* (Jenisch *et al.* 1999; Guerrin *et al.* 2001). Like sheep *CDSN*, the human gene is regarded as highly polymorphic, with an average SNP density of one SNP per 100 bp (Guerrin *et al.* 2001; Capon *et al.* 2003).

Corneodesmosin has also been associated with the autosomal-dominant disorder hypotrichosis simplex of the scalp (Jonca *et al.* 2002; Levy-Nissenbaum *et al.* 2003). Affected individuals experience gradual loss of the scalp hair starting in the middle of the first decade, resulting in almost complete baldness by the third decade (McGrath & Wessagowit 2005). Nonsense

Table 4 Predicted changes to thermodynamic stability of protein with the indicated substitution

Method	iPTree-STAB		iMutant data		
Substitution	PredictedΔΔG kcal/mol	Predicted stabilising/ destabilising	Predicted ΔΔG kcal/mol	Predicted (SVM3) stabilising/ destabilising/ neutral	Predicted (SVM2) stabilising/ destabilising
Q28R	−0.0458	Destabilising	−0.04	Neutral	Destabilising
S68G	−1.96	Destabilising	−0.64	Neutral	Destabilising
S77T	−1.1936	Destabilising	−0.59	Neutral	Destabilising
H112R	−1.1814	Destabilising	−0.29	Neutral	Destabilising
G143S	−1.0663	Stabilising	−0.72	Neutral	Destabilising
D178G	−0.0163	Destabilising	−1.21	Destabilising	Destabilising
P185S	−1.0663	Destabilising	−1.33	Neutral	Destabilising
S469G	−2.3400	Destabilising	−1.16	Destabilising	Destabilising

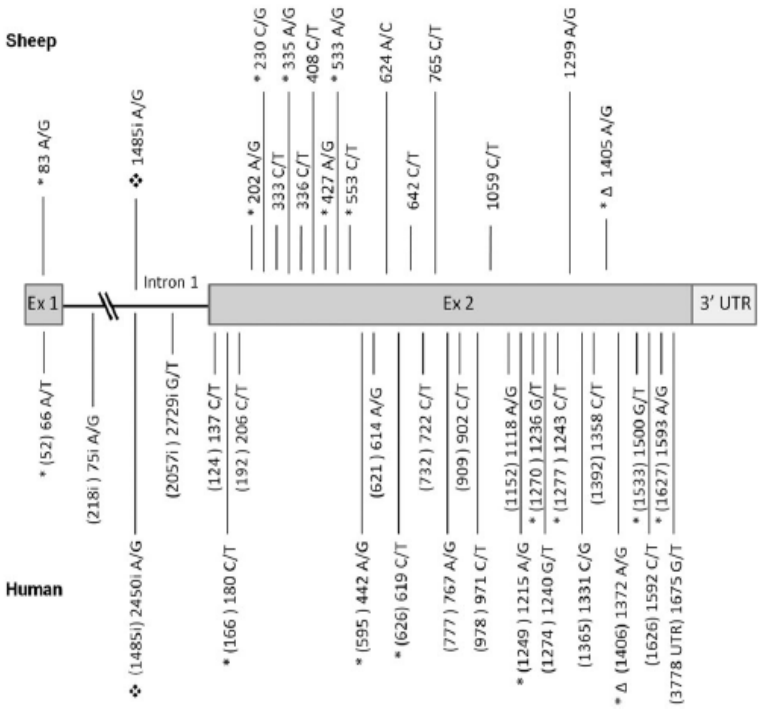


Figure 2 Single nucleotide polymorphisms (SNPs) identified within sheep (this study) and human *CDSN* (Guerrin *et al.* 2001). *Non-synonymous substitution. ❖Identical SNP identified in both human and sheep sequence. () Relative location of human SNP in the sheep coding sequence. Δ SNP located next to each other in coding sequence that effect the same amino acid. Location of SNPs in the intron region is based on the genomic sequence.

mutations that result in accumulation of truncated corneodesmosin aggregates in the superficial dermis and at the periphery of hair follicles have been identified in the *CDSN* gene in three families suffering from hypotrichosis of the scalp (Levy-Nissenbaum *et al.* 2003). It has been suggested that the accumulation of abnormal corneodesmosin aggregates is toxic to the hair follicle cells and that hypotrichosis simplex of the scalp is a disease associated with protein misfolding (Kalinin *et al.* 2001; McGrath & Wessagowit 2005).

The S68G, D178G, and P185S substitutions change the physicochemical properties of the residue at those particular positions in the sequence and may therefore influence function. However, no obvious phenotypic alterations were observed in the sheep that carry these mutations. Nevertheless, the absence of obvious phenotypic changes does not mean absence of a functional, or structural, alteration in the protein. Unfortunately, the thermodynamic approach using iPTree-STAB and iMutant appeared to produce contradictory results, with only the D178G and S469G substitutions being classed as 'destabilizing' by both methods. Interestingly, the S469G mutation is caused by a nucleotide substitution in the first position of the codon. A similar mutation in this codon occurs in human *CDSN* at position 1372 (A/G) in the second nucleotide position (shown in Fig. 2) (Guerrin *et al.* 2001; Hui *et al.* 2002), resulting in an asparagine for serine substitution. This human *CDSN* mutation, when analyzed using these *in silico* methods, is also shown to have a destabilizing effect on the protein. Interestingly, the mutation at position 1372 in humans is found on the human 1.11 haplotype, a haplotype commonly associated with psoriasis (Guerrin *et al.* 2001; Hui *et al.* 2002). Furthermore, we also observed an SNP in the first intron at position 1485 (A/G) in sheep *CDSN*, which is identical to the human SNP at this position, suggesting an ancient origin for this mutation.

Fleece rot and dermatophilosis are examples of bacterial infection in sheep associated with the degradation of a protective skin barrier (Norris *et al.* 2008). A previous study has shown that deletion of *CDSN* in adult knockout mice resulted in a chronic defect in the epidermal layer, suggesting that *CDSN* is essential in the preservation of the structural integrity of skin through maintenance of the desmosome (Leclerc *et al.* 2009). In neonates, a deficiency of *CDSN* results in reduced mechanical resistance and a lethal barrier defect in the skin due to mechanical stresses encountered after birth (Leclerc *et al.* 2009). A mutation in human *CDSN* resulting in a L59X mutation has been suggested to be the cause of the peeling skin disease in one family (Oji *et al.* 2010). It has also been suggested that abnormalities in corneodesmosome structures

could modify their availability to extracellular proteases and thereby lead to persistence of the desmosome in psoriasis (Guerrin *et al.* 2001). Therefore, increased susceptibility to proteolytic degradation of the desmosome may ultimately lead to a reduction in the protective skin barrier. Conversely, in callus formation there is an increased expression of *CDSN* and other adhesion molecules (Kim *et al.* 2010).

In cattle, a strong association between *BoLA-DRB3-DQB* class II haplotype and susceptibility to dermatophilosis has been observed (Maillard *et al.* 2003). It is also known that *CDSN* in cattle is located within the class I region of the MHC. It is possible, although untested, that the *BoLA-DRB3-DQB* association observed in these animals, rather than conferring susceptibility *per se*, may be part of an extended haplotype which harbors variations in *CDSN* that result in increased susceptibility to proteolytic degradation of the skin desmosomes.

Conclusion

The SNPs identified within sheep *CDSN* in this study will be useful in future investigations into wool production traits in sheep, given the location of this gene within the MHC and its known function in wool physiology. In a previous study, associations were observed between several MHC markers and wool production traits such as greasy and clean fleece weight (Bot 2000).

The information obtained from this study will also be used in future investigations into the effects of these mutations on skin integrity and physiology. In addition, the characterization and identification of SNPs within and close to the sheep *CDSN* gene will also provide an essential resource for the future identification of extended sheep MHC haplotypes, particularly in the class I region. These haplotypes will be useful for linkage and association studies and will confirm/disprove a direct or indirect role of *CDSN* polymorphisms in skin-related infections and wool production.

ACKNOWLEDGMENTS

PhD Studentship support for NSS from WABRI, CHIRI and Curtin University.

REFERENCES

Allen MH, Veal C, Faassen A, Powis SH, Vaughan RW, Trembath RC, Barker JNWN. 1999. A non-HLA gene within the MHC in psoriasis. *The Lancet* **353,** 1589–1590.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403–410.

Bot J. 2000. MHC studies relating to parasite resistance in merino sheep. 177. PhD Thesis. School of Biomedical Sciences, Curtin University of Technology, Perth, Western Australia.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268,** 78–94.

Capon F, Toal IK, Evans JC, Allen MH, Patel S, Tillman D, Burden D, Barker JNWN, Trembath RC. 2003. Haplotype analysis of distantly related populations implicates corneodesmosin in psoriasis susceptibility. *Journal of Medical Genetics* **40,** 447–452.

Capriotti E, Fariselli P, Casadio R. 2004. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics (Oxford, England)* **20** (Suppl 1), i63–i68.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research* **33,** W306–W310.

Capriotti E, Fariselli P, Rossi I, Casadio R. 2008. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9** (Suppl 2), S6.

Gromiha MM. 2007. Prediction of protein stability upon point mutations. *Biochemistry Society Transactions* **35,** 1569–1573.

Guerrin M, Simon M, Montezin M, Haftek M, Vincent C, Serre G. 1998. Expression cloning of human corneodesmosin proves its identity with the product of the s gene and allows improved characterization of its processing during keratinocyte differentiation. *Journal of Biological Chemistry* **273,** 22640–22647.

Guerrin M, Vincent C, Simon M, Tazi Ahnini R, Fort M, Serre G. 2001. Identification of six novel polymorphisms in the human corneodesmosin gene. *Tissue Antigens* **57,** 32–38.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89,** 10915–10919.

Huang LT, Gromiha MM, Ho SY. 2007. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics (Oxford, England)* **23,** 1292–1293.

Hui J, Oka A, Tamiya G, Tomizawa M, Kulski JK, Penhale WJ, Tay GK, Lizuka M, Inoko H. 2002. Corneodesmosin DNA polymorphisms in MHC haplotypes and Japaneese patients with psoriasis. *Tissue Antigens* **20,** 77–83.

Ishihara M, Yamagata N, Ohno S, Naruse T, Ando A, Kawata H, Ozawa A, Ohkido M, Mizuki N, Shiina T, Ando H, Inoko H. 1996. Genetic polymorphisms in the keratin-like S gene within the human major histocompatibility complex and association analysis on the susceptibility to psoriasis vulgaris. *Tissue Antigens* **48,** 182–186.

Jenisch S, Koch S, Henseler T, Nair RP, Elder JT, Watts CE, Westphal E, Voorhees JJ, Christophers E, Krönke M. 1999. Corneodesmosin gene polymorphism demonstrates strong linkage disequilibrium with HLA and association with psoriasis vulgaris. *Tissue Antigens* **54,** 439–449.

Jonca N, Guerrin M, Hadjiolova K, Caubet CC, Gallinaro HN, Simon M, Serre G. 2002. Corneodesmosin, a component of epidermal corneocyte desmosomes, displays homophilic adhesive properties. *Journal of Biological Chemistry* **277,** 5024–5029.

Kalinin A, Marekov LN, Steinert PM. 2001. Assembly of the epidermal cornified cell envelope. *Journal of Cell Science* **114,** 3069–3070.

Kim SH, Kim S, Choi HI, Choi YJ, Lee YS, Sohn KC, Lee Y, Kim CD, Yoon TJ, Lee JH, Lee YH. 2010. Callus formation is associated with hyperproliferation and incomplete

differentiation of keratinocytes, and increased expression of adhesion molecules. *The British Journal of Dermatology* **163,** 495–501.
Korber B. 2000. HIV signature and sequence variation analysis. In: Rodrigo AG, Learn GH (eds), *Computational Analysis of HIV Molecular Sequences*, Chapter 4, pp. 55–72. Kluwer Academic Publishers, Dordrecht, Netherlands.
Leclerc EA, Huchenq A, Mattiuzzo NR, Metzger D, Chambon P, Ghyselinck NB, Serre G, Jonca N, Guerrin M. 2009. Corneodesmosin gene ablation induces lethal skin-barrier disruption and hair-follicle degeneration related to desmosome dysfunction. *Journal of Cell Science* **122,** 2699–2709.
Levy-Nissenbaum E, Betz RC, Frydman M, Simon M, Lahat H, Bakhan T, Goldman B, Bygum A, Pierick M, Hillmer AM, Jonca N, Toribio J, Kruse R, Dewald G, Cichon S, Kubisch C, Guerrin M, Serre G, Nothen MM, Pras E. 2003. Hypotrichosis simplex of the scalp is associated with nonsense mutations in CDSN encoding corneodesmosin. *Nature Genetics* **34,** 151–153.
Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research* **33,** 6494–6506.
Maillard JC, Berthier D, Chantal I, Thevenon S, Sidibe I, Stachurski F, Belemsaga D, Razafindraibe H, Elsen JM. 2003. Selection assisted by a *BoLA-DR/DQ* haplotype against susceptibility to bovine dermatophilosis. *Genetic Selection and Evolution* **35** (Suppl 1), S193–S200.
Matsumoto M, Zhou Y, Matsuo S, Nakanishi H, Hirose K, Oura H, Arase S, Ishida-Yamamoto A, Bando Y, Izumi K, Kiyonari H, Oshima N, Nakayama R, Matsushima A, Hirota F, Mouri Y, Kuroda N, Sano S, Chaplin DD. 2008. Targeted deletion of the murine corneodesmosin gene delineates its essential role in skin and hair physiology. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 6720–6724.
McGrath JA, Wessagowit V. 2005. Human hair abnormalities resulting from inherited desmosome gene mutations. *Keio Journal of Medicine* **54,** 72–79.
Montezin M, Simon M, Guerrin M, Serre G. 1997. Corneodesmosin, a corneodesmosome-specific basic protein, is expressed in the cornified epithelia of the pig, guinea pig, rat, and mouse. *Experimental Cell Research* **231,** 132–140.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3,** 418–426.
Norris BJ, Colditz IG, Dixon TJ. 2008. Fleece rot and dermatophilosis in sheep. *Veterinary Microbiology* **128,** 217–230.
Oji V, Eckl KM, Aufenvenne K, Natebus M, Tarinski T, Ackermann K, Seller N, Metze D, Nurnberg G, Folster-Holst R, Schafer-Korting M, Hausser I, Traupe H, Hennies HC. 2010. Loss of corneodesmosin leads to severe skin barrier defect, pruritus, and atopy: unraveling the peeling skin disease. *American Journal of Human Genetics* **87,** 274–281.
Orru S, Giuressi E, Carcassi C, Casula M, Contu L. 2005. Mapping of the major psoriasis-susceptibility locus (*PSORS1*) in a 70-Kb interval around the corneodesmosin gene (*CDSN*). *The American Journal of Human Genetics* **76,** 164–171.
Rufaut NW, Pearson AJ, Nixon AJ, Wheeler TT, Wilkins RJ. 1999. Identification of differentially expressed genes during a wool follicle growth cycle induced by prolactin. *Journal of Investigative Dermatology* **113,** 865–872.
Schmitt-Egenolf M, Windemuth C, Hennies HC, Albis-Camps M, Engelhardt BV, Wienker T, Reis A, Traupe H, Blasczyk R. 2001. Comparative association analysis reveals that corneodesmosin is more closely associated with psoriasis than HLA-Cw*0602-B*5701 in German families. *Tissue Antigens* **57,** 440–446.
Simon M, Jonca N, Guerrin M, Haftek M, Bernard D, Caubet CC, Egelrud TR, Schmidt R, Serre G. 2001. Refined characterization of corneodesmosin proteolysis during terminal differentiation of human epidermis and its relationship to desquamation. *Journal of Biological Chemistry* **276,** 20292–20299.
Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7,** 62.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* **56,** 564–577.
Tazi Ahnini R, Camp NJ, Cork MJ, Mee JB, Keohane SG, Duff GW, di Giovine FS. 1999. Novel genetic association between the corneodesmosin (MHC S) gene and susceptibility to psoriasis. *Human Molecular Genetics* **8,** 1135–1140.
Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22,** 4673–4680.
Zhou Y, Chaplin DD. 1993. Identification in the HLA class I region of a gene expressed late in keratinocyte differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **90,** 9470–9474.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1 Nucleotide sequence of sheep CDSN is available as NCBI GenBank Accession no. GU591411.
Figure S2 Predicted peptide sequence of sheep CDSN is available as NCBI GenBank Accession no. ADD84518.1.
Table S1 Details of primers used to amplify sheep CDSN sequence.
Table S2 SNAP Analysis. Averages of all pairwise comparisons: ds = 1.9007, dn = 1.4149, ds/dn = 1.4426, ps/pn = 1.1720.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Appendix C

SNP data of sheep population used for association study

Homozygous to allele 1 is represented by purple colour. Homozygous to allele 2 is represented by green colour. Heterozygous is represented by grey colour.

Animal ID	New label	EBV for CFW	CD 1	Int 1	CD 2	CD 3	CD 4	CD 5	CD 6	CD 7	CD 8	CD 9	CD 10	CD 11	CD 12	CD 13	CD 14	CD 15	CD 16
			A/G	C/T	A/G	C/G	C/T	A/G	C/T	C/T	A/G	A/G	C/T	A/C	C/T	C/T	C/T	A/G	A/G
982000012173329	W2	0.169																	
982000012175894	W3	0.253																	
982000012175947	W4	0.753																	
982000012186364	W5	1.158																	
982000012186493	W6	0.298																	
982000012186494	W7	0.17																	
982000012186835	W8	0.271																	
982000012186885	W9	1.029																	
982000012186987	W10	-0.19																	
982000012186993	W11	-0.405																	
982000012187010	W12	0.247																	
982000012187042	W13	0.425																	
982000012187077	W14	0.957																	

982000012187143	W15	-0.046																	
982000012187158	W16	0.729																	
982000012187175	W17	1.161																	
982000012187356	W18	0.311																	
982000012187464	W19	0.94																	
982000012187475	W20	0.342																	
982000012187495	W21	0.537																	
982000012187514	W22	-0.896																	
982000012187683	W23	-0.255																	
982000012187715	W24	0.482																	
982000012187725	W25	0.057																	
982000012187800	W26	0.692																	
982000012187899	W27	-0.232																	
982000012187903	W28	-0.612																	
982000012187914	W29	0.57																	
982000012188029	W30	0.398																	
982000012188085	W31	1.11																	
982000012188100	W32	0.271																	
982000012188147	W33	-0.412																	
982000012188357	W34	1.154																	
982000012188439	W35	-0.33																	
982000012188458	W36	-0.519																	
982000012188476	W37	0.899																	
982000012188486	W38	0.245																	
982000012188519	W39	-0.36																	
982000012188526	W40	0.096																	
982000012188616	W41	0.564																	

982000012188630	W42	-0.508																
982000012188641	W43	0.616																
982000012188744	W44	-0.153																
982000012188938	W45	-0.477																
982000012189360	W46	-0.415																
982000012189367	W47	0.256																
982000012189410	W48	0.14																
982000012189533	W49	-0.272																
982000012189540	W50	0.649																
982000012191785	W51	0.699																
982000012202559	W52	-0.315																
982000012202573	W53	-0.438																
982000012202641	W54	-0.252																
982000012202973	W55	-0.589																
982000012203003	W56	0.675																
982000012203124	W57	-0.554																
982000012203147	W58	-0.115																
982000012203205	W59	-0.119																
982000012203349	W60	0.203																
982000012203354	W61	0.581																
982000012203380	W62	1.075																
982000012203403	W63	-0.639																
982000012203423	W64	-0.527																
982000012203429	W65	-0.462																
982000012203461	W66	-0.097																
982000012203482	W67	-0.249																
982000012203500	W68	-0.67																

982000012203504	W69	-0.205																
982000012203512	W70	0.171																
982000012203537	W71	-0.454																
982000012203590	W72	0.259																
982000012203682	W73	-0.27																
982000012203809	W74	0.546																
982000012203812	W75	0.399																
982000012203816	W76	-0.38																
982000012203873	W77	0.312																
982000012203905	W78	0.095																
982000012203913	W79	-0.295																
982000012204000	W80	-0.737																
982000012204073	W81	-0.58																
982000012204081	W82	0.075																
982000012204356	W83	0.649																
982000012204393	W84	-0.333																
982000012204395	W85	0.029																
982000012204402	W86	-0.151																
982000012204410	W87	0.719																
982000012204418	W88	-0.153																
982000012204466	W89	-0.482																
982000012204500	W90	0.443																
982000012204510	W91	-0.195																
982000012204534	W92	-0.256																
982000012204851	W93	-0.247																
982000012205053	W94	-0.333																
982000012205193	W95	-0.707																

982009103163990	W96	-0.473																	
982009103170127	W97	-0.762																	
982009103170414	W98	-0.52																	
982009103170627	W99	-0.212																	
982009103173794	W100	0.546																	
982009103174207	W101	-0.143																	
982009103175175	W102	-0.113																	
982009103178922	W103	-0.559																	
982009103217273	W104	-0.618																	
982009103307216	W105	-0.902																	
982009103171817	W106	0.181																	
982000012188873	W107	0.745																	
982000012203572	W108	-0.271																	