Diversity and Evolution in the Avian Major Histocompatibility Complex

Rebecca Martin

St. Catharine's College

Department of Pathology

University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

March 2021

Declaration

I hereby declare that my dissertation entitled

Diversity and Evolution in the Avian Major Histocompatibility Complex

- Is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the acknowledgements and specified in the text.
- Is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.
- Does not exceed the prescribed word limit for the Biology Degree Committee.

31st March 2021

Abstract

Diversity and Evolution in the Avian Major Histocompatibility Complex

Rebecca Martin

Outside the well-studied mammalian models of human and mouse, the chicken has one of the best-understood immune systems, largely due to the interests and investment of the commercial poultry industry. The chicken Major Histocompatibility Complex (MHC) has been particularly well-studied, since variation at the chicken MHC is strongly associated with resistance and susceptibility to disease, particularly infectious pathogens. Furthermore, its 'minimal essential' gene content and organisation makes the chicken MHC a useful model for immunologists.

The vast majority of work on the chicken MHC has used highly inbred experimental lines, derived from commercial flocks, which contain a few standard MHC haplotypes. This thesis aims to understand the extent to which insights developed in these lines can be applied to other populations, species and broader taxonomic groups. The examination of features of the adaptive immune systems in non-model populations and species also provides insights into the fundamental evolutionary flexibility of this system, and could help to inform zoonotic disease surveillance, rural poultry development initiatives, conservation and commercial poultry breeding strategies.

First, diversity in the classical MHC class I and II genes in 'fancy' breeds and free-range local, or 'village', chickens was compared to that previously observed in commercial flocks. The expanded database of chicken MHC diversity resulting from examination of these populations is a valuable resource for both the scientific and agricultural communities and describes a largely untapped pool of potentially useful genetic diversity. Second, three defining features of the chicken MHC were examined in three passerine bird species. The results show deviation from the 'chicken-type' MHC in all three species, as well as variation between the passerine species themselves, highlighting substantial evolutionary flexibility in the system. Finally, a family of peptide editing genes, of which one, TAPBPL, is notably absent in placental mammals, was investigated using genomic databases to assess phylogenetic distribution and biochemical methods to investigate function and regulation.

At all scales, significant variation in antigen processing and presentation is reported, challenging some of the assumptions made in models of both chicken and mammalian MHC evolution. While the need to apply insights with care across lineages is highlighted, a better understanding of the similarities and differences present at different scales can also help appropriate insights to be applied with confidence in important inter-disciplinary contexts.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Jim Kaufman, without whose vision and guidance this project would not have been possible. I am also particularly grateful to Dr. Clive Tregaskes who developed the chicken MHC typing programme and has answered innumerable questions on every possible aspect of lab technique. Thanks also to all the members of the Kaufman lab who have provided support, advice and entertainment over the last four years. Specific contributions of lab members to the work described are detailed in the text.

I also acknowledge the contributions of a number of internal and external collaborators, which are described further in the text. I am grateful to all those mentioned for their generosity with their time, knowledge and resources.

- Dr. Michal Vinkler and Zuzana Świderská (Charles University, Prague), Prof. Helena Westerdahl and Dr. Anna Drews (University of Lund) and Dr. Jacqueline Smith (Roslin Institute, Edinburgh) who welcomed me to their labs, shared samples and showed me around their beautiful cities.
- Dr. Steffen Weigend (Friedrich-Loeffler-Institut, Griefswald), Dr. Susan Lamont (Iowa State University) and Dr. Huaijun Zhou (University of California, Davis) who provided samples and data for MHC typing.
- Dr. Mark Ravinet (University of Nottingham) who shared an unpublished sparrow genome.
- Dr. Gabriele Sorci (University of Burgundy) and Dr. Sally Rogers (University of Exeter) who provided passerine bird samples.
- Dr. Anton Enright and Albert Kang (University of Cambridge, Department of Pathology) who helped design, sequence and process data for the Oxford Nanopore library.
- Dr. Karsten Skøjdt (University of Southern Denmark) who produced monoclonal antibodies.
- Dr. Mike Deery and collegues at Cambridge Centre for Proteomics who advised on sample preparation for proteomics and performed LC-MS/MS analysis.

Contents

List of Figures

List of Tables

List of abbreviations

Abbreviation	Definition
AEBSF	4-benzenesulfonyl fluoride hydrochloride
AIDS	Acquired immune deficiency syndrome
AIV	Avian Influenza Virus
AMM	Apparent molecular mass
APC	Antigen presenting cell
ATP	Adenosine triphosphate
BME	2-Mercaptoethanol
cDNA	Complementary DNA
CNV	Copy number variation
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleoside triphosphate
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
EM	Electron microscopy
ER	Endoplasmic reticulum
ERAAP	Endoplasmic reticulum aminopeptidase associated with antigen processing
ERAD	Endoplasmic reticulum associated protein degradation
FBS	Fetal bovine serum
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
gDNA	Genomic DNA
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HRP	Horseradish peroxidase
IFN	Interferon
Ig	Immunoglobulin
Ii	Invariant chain
ILRI	International Livestock Research Institute
IP	Immunoprecipitation
LC-MS/MS	liquid chromatography-tandem mass spectrometry

LRT	Likelihood ratio test
mAb	Monoclonal antibody
MCS	Multiple cloning site
MHC	Major Histocompatibility Complex
MHCI	MHC class I
MW	Molecular weight
NBD	Nucleotide binding domain
NDV	Newcastle Disease Virus
NK	Natural killer
NTC	No template control
PAGE	Polyacrylamide gel electrophoresis
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate-buffered saline
PCA	Principle component analysis
PCR	Polymerase chain reaction
PLC	Peptide loading complex
PSS	Positively selected site
qPCR	Quantitative PCR
RPMI	Roswell Park Memorial Institute
SDS	Sodium dodecyl sulfate
SIV	Simian immunodeficiency virus
SNP	Single nucleotide polymorphism
TAP	Transporter associated with antigen processing
TCR	T-cell receptor
UGT1	UDP-glucose:glycoprotein glucosyltransferase 1
TAPBP	TAP binding protein (tapasin)
TAPBPR	TAP binding protein-related
TAPBPL	TAP binding protein-like
WGS	Whole genome sequencing

Chapter 1

Introduction

1.1 The importance of avian responses to pathogens

While human responses to pathogens, and the mouse models that simulate them, have been understandably well-studied, there is still much that is unknown about disease and immunity in non-model systems. The study of livestock animals has benefited from industry interest and investment, meaning that they are perhaps the next best understood group after humans and their mouse and primate models. Nonetheless, a huge amount remains to be understood, especially in contexts outside of western commercial production systems. As human population density increases, livestock are being kept in larger and larger numbers, and in increasingly close proximity to humans. Coupled with the destruction of natural habitats, which is forcing wildlife into contact with domestic animals, these trends suggest that the risk of a pathogen-borne disease that either devastates a major food source or becomes zoonotic is increasing.

Chickens are the most numerous and fastest expanding livestock animal (figure 1.1), but there seem to be substantial differences between the immune systems of mammals and birds, so chicken breeders and vets cannot benefit so directly from knowledge developed in mammalian models as breeders of mammalian livestock can. Thus, there is an urgent need to improve our understanding of the chicken's immune system.

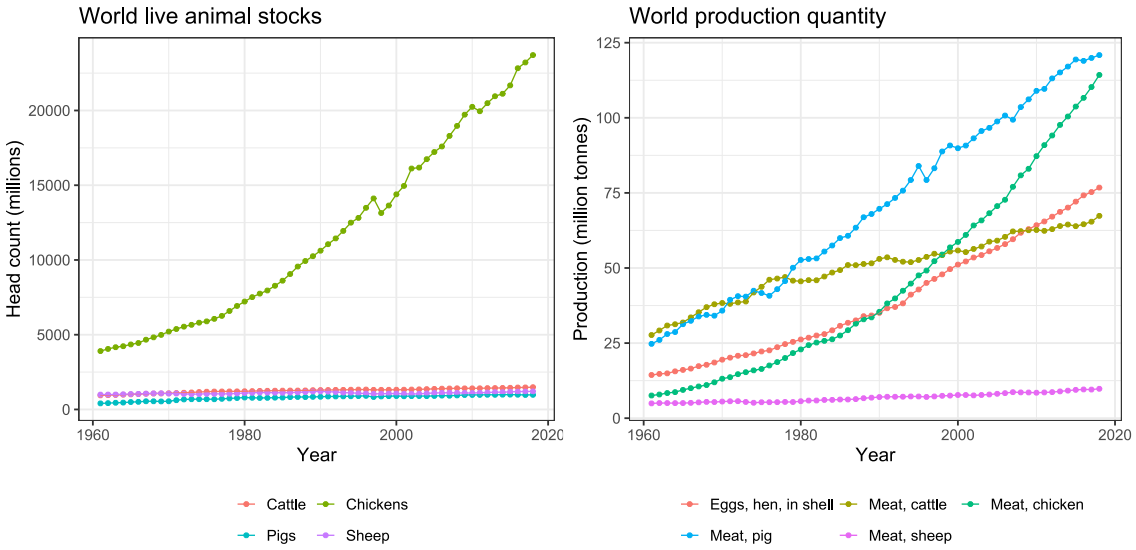


Figure 1.1: **Livestock numbers and meat/egg production from 1961 until 2018** Data from FAO (2020*a*)

Beyond the direct application of insights from research on chickens to food production, the chicken is also used as a model organism for other birds. Thus, a better understanding of the chicken system, coupled with knowledge of any characteristics that are *not* conserved across the group to which the model is being applied, also allows more informed management of bird species in ecology and conservation.

1.1.1 Public Health

Zoonotic pathogens pose a severe threat to public health, as recently highlighted by the COVID-19 pandemic. While COVID-19 has not been directly linked to poultry farming (indeed pigs and poultry do not seem to be susceptible to either SARS or SARS-CoV-2 (Weingart et al., 2004; Friedrich-Loeffler-Institut, 2020)), they do carry several other potentially high-risk pathogens. A highly pathogenic and transmissible influenza virus has long been considered one of the most likely pathogens to cause a devastating pandemic in humans. Birds contain a huge reservoir of Avian Influenza Viruses (AIVs) which periodically cross the species barrier into humans but have been controlled, in recent outbreaks, by their low human-to-human transmissibility (WHO, 2020). Nonetheless, the viruses involved in these recent outbreaks (H5N1 from birds and H1N1 from swine) are still circulating in animals, and a mutation that allows human-to-human transmission is probably eventually inevitable. Of the 19 influenza strains assessed by the Centre for Disease Control and Prevention using their Influenza Risk Assessment Tool, 15 are currently circulating in wild or domestic birds, highlighting the importance of understanding disease responses in both livestock and non-livestock bird species (figure 1.2, Centre for Disease Control and Prevention (2020)).

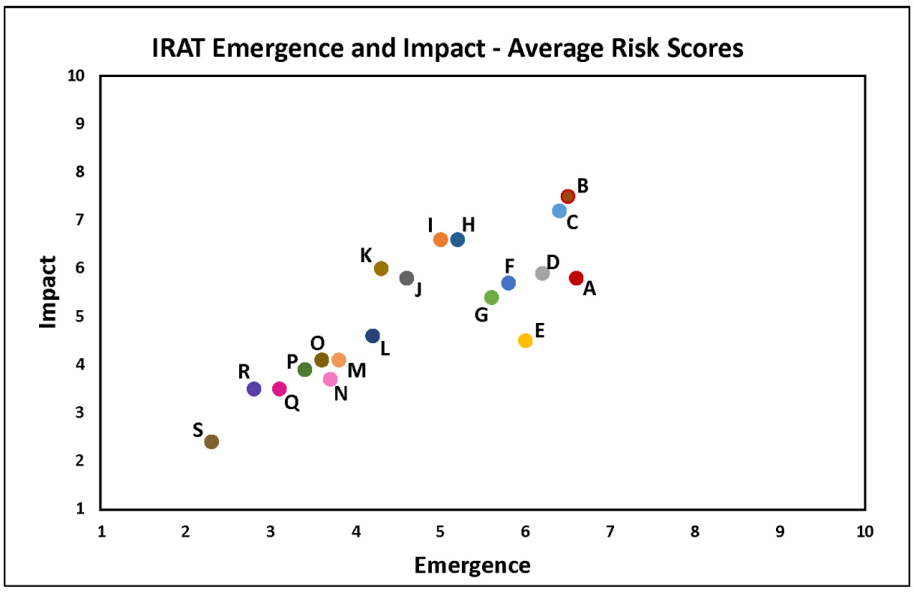


Figure 1.2: **Data from the CDC's Influenza Risk Assessment Tool.** Strains A, E and F are circulating in swine, N in canines and the remainder in birds. Bats carry additional strains with unique haemagglutinin subtypes H17 and H18, which are not assessed here. 'Emergence' refers to the risk of a novel (i.e., new in humans) influenza virus acquiring the ability to spread easily and efficiently in people. 'Impact' refers to the potential severity of human disease caused by the virus (e.g., deaths and hospitalizations) as well as the burden on society (e.g., missed workdays, strain on hospital capacity and resources, and interruption of basic public services) if a novel influenza virus were to begin spreading efficiently and sustainably among people (Centre for Disease Control and Prevention, 2020).

Live bird markets, common in East and South Asia, pose a particular risk due to the density and variety of avian hosts, which supports the spread of AIVs, and the frequency of inter-species transmission opportunities (Fournié and Pfeiffer, 2014).

1.1.2 Food Security and sustainability

High-income countries

In high-income countries, diets generally include large amounts of meat, increasingly poultry, and a poorly-managed novel disease could threaten a food source that many rely on. While poultry is not a necessary part of a healthy diet, a sudden reduction in supply would likely increase the price of meat products generally, leaving lower-income households in particular to suddenly make major adjustments to their diets.

While an overall reduction in consumption of animal products in high-income countries is generally considered to be necessary in the medium to long term, there is a place for efficient poultry production in a sustainable future food system. Chickens produce fewer greenhouse gases than

ruminants and can be reared on feed made from food waste without the risk of Bovine Spongiform Encephalopathy (Mad Cow Disease) that would be associated with feeding waste-derived products to cows or sheep. Such a waste-recycling system has been successfully implemented in Japan since 2001, with government policy both supporting and regulating the industry (Maeda, 2008). This system reduces the poultry industry's consumption of corn and soy, which could be more efficiently consumed by humans directly, but does carry a slightly increased risk of novel pathogen exposure (although waste is carefully treated during processing).

Low-income countries

80% of rural households in developing countries rear poultry, generally in simple night shelters, with minimal management, disease prevention inputs or supplemental feeding. The chickens are effective foragers, and make a significant contribution to the food security and protein intake of the communities who rear them. Nonetheless, the efficiency of these systems and the indigenous birds reared within them is generally low (FAO, 2020*b*).

Development of these local poultry systems has the potential to mitigate ongoing problems with protein and micronutrient malnutrition associated with low-variety vegetarian diets (Müller and Krawinkel, 2005; FAO et al., 2015; Semba, 2016), as well as boost rural economies. Many farmers currently have to choose between eating or selling the products from their flock, meaning that either they, or nearby communities unable to rear livestock, do not have access to the nutritional benefits. Improving the production efficiency of rural production systems (through vaccination or breeding for improved traits) gives smallholders the opportunity to sell the excess, while still benefiting themselves from their products (Gates, 2016). However, almost everything currently known about chicken genetics and immunology is based on experimental lines derived from western commercial chickens. There is a clear need to understand the evolutionary pressures affecting local chicken breeds before deciding on interventions.

1.1.3 Conservation

The Major Histocompatibility Complex (MHC) locus is often used as a high-resolution marker of overall genetic diversity in a population, which can be an important indicator of vulnerability to the general effects of inbreeding depression as well as limited potential to respond to changing environmental, including pathogen, pressures.

More specifically, understanding the evolution of the MHC in a non-model species of interest can help to identify specific pathogen-mediated threats that require intervention, possibly through breeding programmes. In small or fragmented populations, which are becoming increasingly com-

mon due to habitat loss, the presence or absence of specific alleles in subpopulations can have a significant impact on their survival (Rüdel, 2004; Schad et al., 2005). This is more likely to be the case in populations where MHC genotype is more tightly linked to disease outcome (as is seen in chickens). Study of specific MHC alleles may also allow the identification of so-called 'generalist' alleles (Kaufman, 2018) which should ideally be preserved for future adaptive potential as species are increasingly forced into new environments by climate and land use change.

1.2 Structure and function of the major histocompatibility complex (MHC)

1.2.1 The MHC region

MHC regions are genomic regions present in jawed vertebrates (gnathostomes) which contain a high density of immune-related genes, particularly classical MHC genes and genes for the antigen processing machinery with which they are functionally associated. The MHC region was first identified as the locus responsible for tumour graft rejection in mice (Little and Tyzzer, 1916; Gorer et al., 1948), although it was not until later that the rejection was attributed to an immune response against the graft (Medawar, 1944, 1945).

In 1999, the first fully sequenced and annotated MHC region for humans was published (The MHC sequencing consortium, 1999). The sequence was 3.6 Mb in length and contained 224 identified gene loci. The paper appeared back to back with the complete sequence of the chicken MHC which was just 92 kb in length and contained 19 identified genes (Kaufman et al., 1999).

1.2.2 Classical MHC molecules

The discovery of MHC molecules specifically as "genetically determined structures on the cell surface that regulate immunological reactions" is attributed to the work of George Snell, Jean Dausset and Baruj Benacerraf, for which they received the Nobel Prize in Physiology or Medicine in 1980.

It is now known that classical MHC molecules play key roles in adaptive immunity, primarily by binding and presenting antigenic peptides to T cells. As will be discussed in section 1.2.4, a rigid definition of classical MHC molecules is becoming increasingly difficult as the complexity of the functions of these genes is uncovered, however their role in antigen presentation remains crucial to the disease response phenotypes with which they are associated.

Two classes of MHC genes are distinguished: class I and class II. The relative numbers of genes coding for molecules of each class can vary widely and while almost all species carry genes of both

classes, loss of class II has been reported in specific species of the genus *Syngnathus* (pipefish) (Haase et al., 2013; Small et al., 2016), order Gadiformes (cod-like fish) (Star et al., 2011; Malmstrøm et al., 2016) and order Lophiiformes (anglerfish) (Dubin et al., 2019). Both classes function primarily to present antigenic peptides to T cells, although the origins of the peptides and T cell subsets involved differ between the class I and class II systems.

Structure

Classical MHC class I molecules are heterodimeric and composed of a membrane-anchored α chain and separate β_2-microglobulin (figure 1.3). The β_2-microglobulin gene is normally found outside the MHC region, and the protein also associates with some non-classical class I molecules such as CD1. The α chain, which is encoded by classical MHC class I genes in the MHC region, has a membrane-proximal immunoglobulin-like domain (α3) and two membrane-distal domains (α1 and α2) which fold together to form a peptide-binding groove comprised of two α-helices on top of eight anti-parallel β-sheets. This groove accommodates antigenic peptides, whose side chains and backbone interact with the exposed residues on the class I molecule. Particularly important are the so-called 'anchor' residues (generally the second (P2) and last (PΩ) amino acids of the peptide), whose side-chains are required to 'fit' into pockets in the groove of a given class I molecule in order for that peptide to be bound with high affinity and presented to T cells.

Class II molecules are also heterodimeric but the two chains, α and β, are both membrane-anchored (figure 1.3). The genes encoding the two proteins may be found together in the MHC region, as they are in humans, or the gene for the α chain may be outside the MHC region, as in chickens. As for class I molecules, the two membrane-proximal domains are immunoglobulin-like, and the two membrane-distal domains fold together into a peptide binding groove, although in the case of class II the two distal domains are part of independent proteins rather than being a single chain. Anchor residues are also present in class II, although they are more likely to be at peptide positions P1, P4, P6 and/or P9 (Stern et al., 1994; Wieczorek et al., 2017). Notably, PΩ is not a typical anchor residue in class II; indeed the class II binding groove is open at both ends, permitting the binding of peptides of more variable length than class I, typically 13-25 amino acids.

Function

The primary function of classical class I and II molecules is to present antigenic peptides to T cells. Recognition of a peptide-MHC complex by a T cell receptor (TCR) triggers a range of downstream effector functions. While class I molecules are constitutively expressed on the surface of almost all nucleated cells, constitutive class II expression is restricted to professional antigen presenting cells (APCs); expression of class II can be induced in non-professional APCs by a variety of immune regulators.

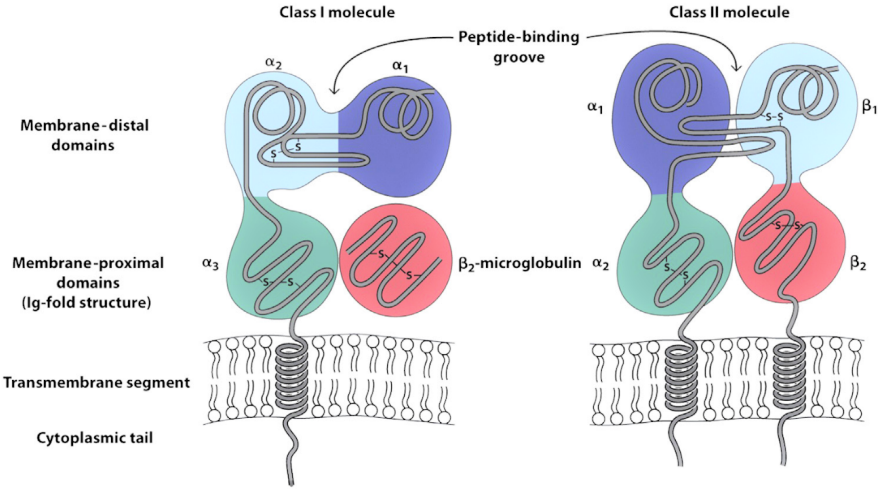


Figure 1.3: **Structure of MHC class I and II molecules.** From *The major histocompatibility complex and antigen presentation* (2013).

Class I molecules primarily present peptides derived from intracellular proteins, including those derived from viruses. The class I-peptide complex is recognised by the TCR on $CD8^+$ 'killer' T cells which are subsequently activated and induce apoptosis of the presenting cell either through the release of perforin and granzymes or the expression of FAS ligand, both of which trigger a caspase cascade leading to cell death. This response, although clearly damaging to host cells, is necessary since it allows the identification and destruction of cells which would otherwise release many more virions to infect neighbouring cells.

Classical class I molecules also act as ligands for inhibitory natural killer (NK) cell receptors such that the absence of class I also causes NK cell-mediated cell death. This is a response to down-regulation of class I expression by some pathogens, which would otherwise make the presence of the pathogen invisible to NK cells.

Classical class II molecules primarily present peptides from extracellular pathogens including bacteria and helminths. Peptides are presented to $CD4^+$ 'helper' T cells, which subsequently differentiate into various cell subsets (including the T helper (T_h) 1, T_h2, T_h17, regulatory T (T_{reg}) and T follicular helper (T_{FH}) lineages) depending, predominantly, on the cytokine environment. Effector cells then release cytokines which active other immune cells, for example T_h1 primarily activates $CD8^+$ T cells and macrophages (cell-mediated response) and T_h2 primarily activates B-cells, eosinophils and mast cells (humoral response).

MHC molecules have also been implicated in mate choice, with evidence that many species can detect the absolute and/or relative MHC diversity of a potential mate through olfactory signals,

and will often chose a more dissimilar mate if a choice is presented (for examples across fish, amphibians, birds, reptiles and mammals see: Yamazaki et al. (1978); Landry et al. (2001); Olsson et al. (2003); Milinski et al. (2005); Bonneaud et al. (2006); Santos et al. (2016) and a quantitative meta-analysis by Kamiya et al. (2014)). Indeed, it has been proposed that the MHC may have had a role in social signalling before it became central to adaptive immunity (Ruff et al., 2012). While this system is beneficial in that offspring will benefit from increased MHC diversity, the MHC, if detectable, also provides a useful marker that can be used by species to avoid the more general effects of inbreeding depression (Potts et al., 1994). The mechanism by which MHC-type can be detected remains controversial. In mammals, MHC-binding peptides appear to be present in the urine and other secretions and can be detected by the vomeronasal organ (Leinders-Zufall et al., 2004, 2009), but the volatility of these peptides is generally low, and it is not clear exactly how or why these peptides come to be enriched in secretions, although there have been reports of soluble HLA isoforms in tears, saliva and sweat (Aultman et al., 1999). In other animals, especially birds, whose olfactory capabilities have been traditionally thought to be limited, the mechanisms are even less clear, although the general phenomenon itself is well-supported. As summarised by Bernatchez and Landry (2003) and Kaufman (2021), the associations between reproduction and MHC diversity seem to be complex, variable and context-based, even between closely related species. Variable associations could possibly result from sexual conflicts between males and females each trying to control the outcome of reproduction, or from the potentially opposing pressures of pathogen resistance and mate choice (Kaufman, 2021). In humans, HLA-C variability has a profound effect on placentation, highlighting yet another potential trade-off (Parham and Moffett, 2013). The genetic architecture of the MHC, degree of population inbreeding and local pathogen abundance or diversity may all influence the outcomes of these conflicts and trade-offs (Jordan and Bruford, 1998; Bernatchez and Landry, 2003), resulting in the highly variable associations observed in nature.

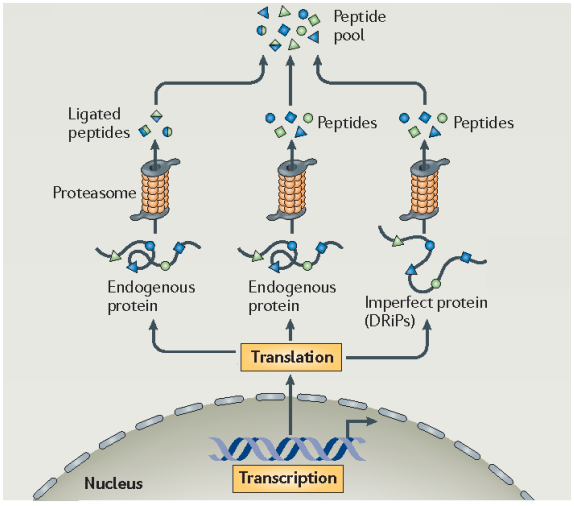
1.2.3 Antigen processing

This section provides a generalised description of the functions of key components of the antigen processing and presentations pathways. Variation between taxonomic groups is discussed later.

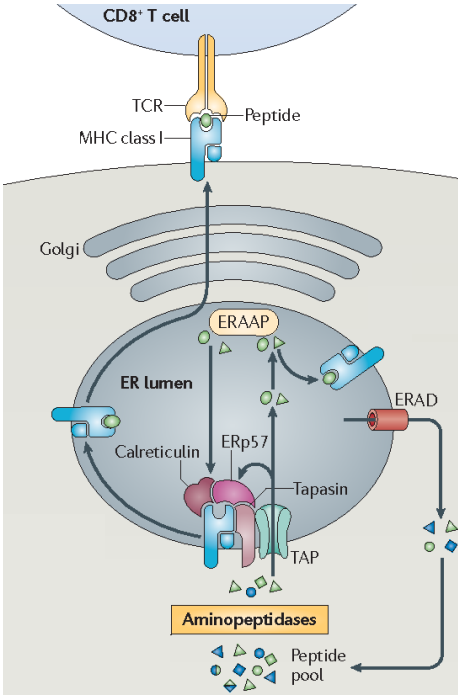
Class I

The classical model for processing pathogen-derived proteins into peptides loaded on classical class I molecules is outlined in figure 1.4. Class I peptides are primarily derived from proteins in the cytosol, including both self-peptides and those synthesised from viral nucleic acids. These proteins are degraded by the proteasome and aminopeptidases into short peptides which are transported into the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP), where they may be trimmed further by ERAAP aminopeptidases. The peptides are finally loaded

onto the class I heterodimer with the help of the chaperone and peptide editor tapasin and the complex is transported to the cell surface.



(a) Generation of a peptide pool from cytosolic proteins



(b) Loading and presentation of peptides on MHC class I

Figure 1.4: **Summary of class I antigen processing and presentation.** Panel (a) shows the proteasome functioning to degrade proteins produced by endogenous transcription and also to join peptide fragments to produce new antigens in so-called 'peptide splicing'. Panel (b) shows these peptides being further trimmed by aminopeptidases and transported into the ER by the TAP transporter as part of the PLC. Peptides can either be loaded directly or trimmed further by ERAAP aminopeptidases in the ER before loading. The endoplasmic-reticulum-associated protein degradation (ERAD) pathway transports peptides which do not bind to class I back into the cytosol where they are eventually fully degraded by the proteasome and aminopeptidases. Modified from Neefjes et al. (2011)

The pathway in figure 1.4 is not the only mechanism of class I loading and the complexities of cross-presentation, peptide splicing and endocytic recycling of class I molecules are still not fully understood. Nonetheless, the major function of class I molecules can be explained according to this mechanism.

Peptide editing is a key part of both the class I and class II systems, allowing the preferential binding and presentation of high-affinity immunodominant epitopes. For class I, the best studied peptide editor is tapasin. In the peptide loading complex (PLC), tapasin connects the class I molecule to the TAP transporter, although it is notable that while both TAP1 and TAP2 have tapasin binding domains in humans, this region is missing from TAP1 in chickens (Walker et al., 2005).

Tapasin has been shown to act by stabilising the class I molecule in an 'open' or peptide-receptive conformation, thus catalysing the release of low-affinity peptides. Once a high-affinity antigen binds, the forces acting to close the peptide binding groove are able to overcome those exerted by tapasin to open it, resulting in a peptide-bound equilibrium state where tapasin's affinity for class I is low and the complex is primed for dissociation from the PLC (Fisette et al., 2016). Different class I alleles are known to be differently tapasin-dependant (Peh et al., 1998; Williams et al., 2002; Lewis et al., 1998; Park et al., 2003), a characteristic that is discussed further in the introduction to Chapter 4.

Additional peptide editing processes are likely to occur downstream of peptide loading in the ER, mediated by proteins such as TAPBPR (Boyle et al., 2013) and possibly, in non-mammals, TAPBPL (Grimholt, 2018). These molecules, both of which show a degree of sequence homology to tapasin, are discussed in detail in the introduction to Chapter 4.

Class II

Class II molecules primarily present peptides that are endocytosed from the extracellular environment and degraded by proteases in the early endosomes. The α and β chains are assembled in complex with the invariant chain (Ii) in the ER and then transported to a multivesicular body known as the MIIC. Here, Ii is cleaved to a short peptide known as CLIP, which can then be exchanged for a high affinity antigenic peptide, usually in the presence of the chaperone DM, which facilitates the selection of these high affinity peptides. The complex then travels to the cell surface and is recognised by $CD4^+$ T cells (figure 1.5).

Peptide editing in class II is catalysed by the non-classical class II molecule DM. As with class I, class II alleles exhibit a range of DM dependencies in mammals (Stebbins et al., 1995) and chickens (Wise, 2019).

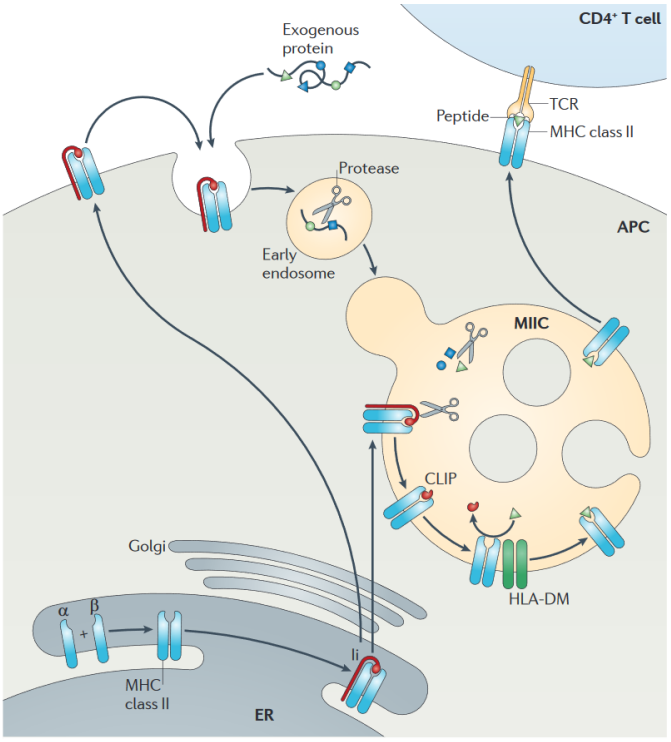


Figure 1.5: **Summary of class II antigen processing and presentation.** Neefjes et al. (2011).

1.2.4 Non-classical MHC molecules

Alongside their sets of classical class I and class II genes, all gnathostomes appear to have non-classical class I and II genes. These genes code for molecules which, despite having relatively low amino acid sequence conservation in some cases, all have roughly the same protein structure as classical MHC molecules, most with a ligand binding groove at the membrane-distal end of the molecule (D'Souza et al., 2019). Some appeared early in evolution and form a distinct clade across different species, that is, the gene in one species is more closely related to the homologous gene in another species, than it is to the other classical or non-classical genes in its own species. Others seem to arise in specific lineages, probably by duplication and neofunctionalisation of existing class I genes. These genes will be more closely related to others in the same organism than they are to anything in a different species or lineage (Rodgers and Cook, 2005; Dijkstra et al., 2018).

Defining classical and non-classical is difficult. In general, classical MHC molecules are defined as showing high polymorphism, wide and high tissue expression and the ability to bind peptides and present them to T cells. Any MHC molecule which does not meet all three of these criteria is generally considered non-classical. Increasingly, however, these criteria are being interpreted on a spectrum rather than within a binary framework; HLA-C for example is considered a human classical class I molecule but has noticeably lower polymorphism and expression than HLA-A or -B (Snary et al., 1977; McCutcheon et al., 1995), and is a dominant ligand for KIR on NK cells

(although it does also present peptides to T cells (Wesley et al., 1993)).

Non-classical class I genes

Many non-classical class I genes are involved in immunity, with all but HFE, FcRn and ZAG known to interact with either T cells or NK cells in humans (D'Souza et al., 2019), either as direct ligands for immune cell receptors or by presenting a bound antigen. Antigens are not restricted to peptides as they are for classical class I molecules and include, for example, lipids (CD1) and vitamin metabolites (MR1). The majority form heterodimers with β_2-microglobulin.

Although not all species necessarily have homologues of all the known 'ancient' non-classicals, the majority seem to have at least a subset of 'ancient' genes in addition to their 'young' non-classicals. Examples of 'ancient' non-classicals include CD1 and UT (which was later lost in eutherian mammals), both of which apparently predate the emergence of tetrapods (Dijkstra et al., 2018). In humans, the young genes are HLA-G and -F, which are thought to have arisen within the primate lineage (Otting and Bontrop, 1993; Langat and Hunt, 2002), while HLA-E appears to have a homologue in mice (Qa1), and is therefore thought to have arisen early in the mammalian expansion.

Non-classical class II genes

The conventional non-classical class II molecules DM and DO play a role in peptide editing but do not, themselves, present antigens. DM is highly conserved as far back as amphibians (Ohta et al., 2006) and even lungfish (Yamaguchi and Dijkstra, 2019), but appears to be absent in teleosts (Dijkstra et al., 2013), while DO may be a more recent development in mammals (NCBI Homologene).

Non-conventional non-classical class II genes, which can be identified as lacking one of the three characteristics of classical MHC genes include DQ2 in primates, H2-E2 in mice, RT1-D2 in rats and the YLB genes in chickens.

1.3 The MHC, coevolution and disease associations

Two of the best-studied MHCs, those of the human and the chicken, illustrate two key models of MHC structure and function. The mouse has also been well-studied, but its MHC largely resembles that of the human in terms of the key features discussed in this section. Variation on a finer taxonomic scale is discussed in more detail in section 1.5.2.

1.3.1 The human MHC

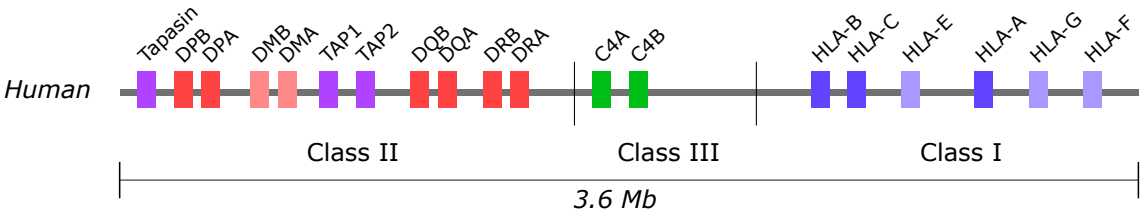


Figure 1.6: **Simplified structure of the human MHC.** The class III region containing C4 genes (green) is in between class I and class II. In the class I region there are three classical (dark blue) and three non-classical (light blue) class I loci, while the class II region contains three sets of classical class II genes (dark red, note that most haplotypes have two DRB-like genes), a single pair of non classical class II genes (light red) and genes related to class I antigen processing (TAP and tapasin, purple).

The human MHC is approximately 3.6 Mb in length and has the class III region flanked by the class I and class II regions (figure 1.6). The class I region contains three classical (HLA-A, B and C) and three non-classical (HLA-E, F and G) class I genes. As discussed in section 1.2.4, the distinction between classical and non-classical is increasingly unclear, however this division between A/B/C and E/F/G is used widely in the field. The class III region contains immune genes such as the complement components C2, C4 and factor B, and cytokines of the TNF family as well as the non-immune related chaperone HSP70 among others. The class II region contains genes for classical and non-classical class II heterodimers and also the TAP and tapasin genes, which are associated with class I antigen processing.

The physical and recombinational distance between the TAP genes and the classical class I genes in the human MHC requires that the human TAP transporter (hTAP) be almost completely functionally monomorphic and highly permissive in its peptide transport repertoire, since any functional polymorphism would risk an incompatible TAP-class I combination arising through recombination (Kaufman, 2015). Some preference for hydrophobic and positively charged residues at the C-terminus has been reported, reflecting one of the key MHC class I (MHCI) anchor residues (Rammensee et al., 1995) and while some hTAP selectivity can be seen at residues 1-3 of 9-mer peptides, residues 5-8, which are key for TCR interactions, do not seem to be restricted at all (Garboczi et al., 1996).

A permissive TAP transporter in turn allows the maintenance of a multi-gene family of classical class I genes, all of which will be able to receive appropriate peptides, regardless of their peptide binding repertoires. Expansion of the number of class I genes seems to be adaptive to a point, providing a greater number of 'chances' to find and present an immunogenic peptide. Beyond this theoretical optimum, it has been proposed that further expansion of the multi-gene family would become deleterious, with too many T cells potentially being removed during negative selection in

the thymus (Nowak et al., 1992; Migalska et al., 2019).

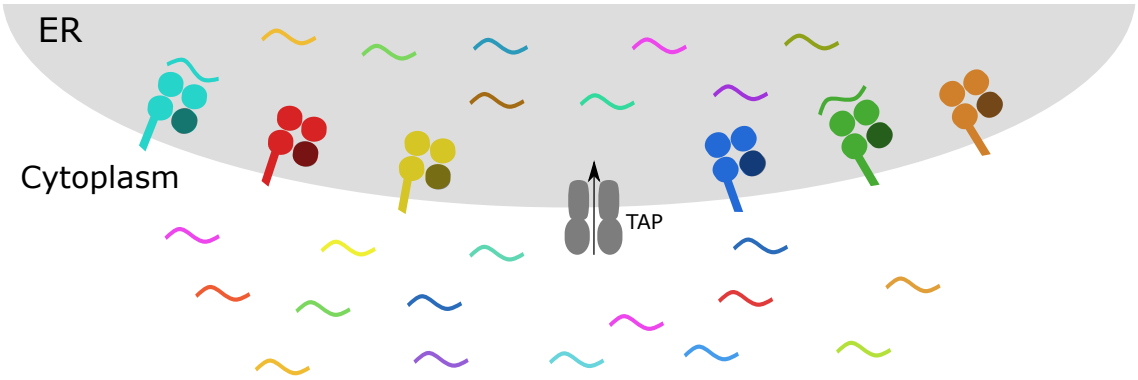


Figure 1.7: **Human TAP transport and peptide binding.** Most humans express six allelic variants of the class I molecule and therefore have six 'chances' to find a peptide from the pool produced by the proteasome. The TAP transporter has wide peptide transport capabilities and moves a range of peptides which the class I molecules can sample.

The multi-gene family of class I molecules in humans results in only weak genetic associations being observed between MHC haplotype and the outcome of infectious disease. Since most humans are heterozygous at all three classical class I loci, they have six 'chances' to find and present a protective peptide from any given pathogen (figure 1.7). This seems to be sufficient to allow the vast majority of humans to mount an immune response against the vast majority of pathogens. As a result, the outcome of infection by a given pathogen is more likely to be determined by other factors or comorbidities, than by the specific combination of MHC alleles carried by the individual (Kaufman, 2000).

Nonetheless, some associations between human class I alleles and infectious disease outcomes do exist. The best-studied class I association is with HIV progression to AIDS, with HLA-B*57 and HLA-B*27 consistently associated with slower disease progression and HLA-B*35 associated with faster disease progression (Kaslow et al., 1996; Carrington et al., 1999). Associations have also been reported with other infectious diseases including hepatitis B and C, leprosy, tuberculosis, malaria, leishmaniasis and schistosomiasis (reviewed in Blackwell et al. (2009)). Particular class I alleles have also been associated with predisposition to autoimmune conditions including ankylosing spondylitis, Graves' disease, multiple sclerosis and Type 1 diabetes (reviewed in Gough and Simmonds (2007)). Indeed, autoimmune conditions generally show stronger associations with HLA genotypes, both class I and class II, than do infectious diseases.

Associations with class II are more widely reported for both infectious diseases (Blackwell et al., 2009) and autoimmune conditions (Gough and Simmonds, 2007) in humans, although some linkage

disequilibrium across the region makes it challenging to detect associations with class I and class II independently.

1.3.2 The chicken MHC

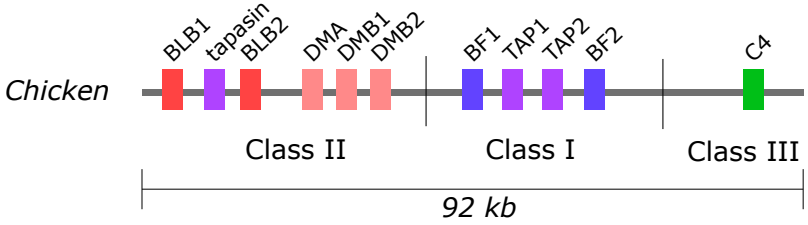


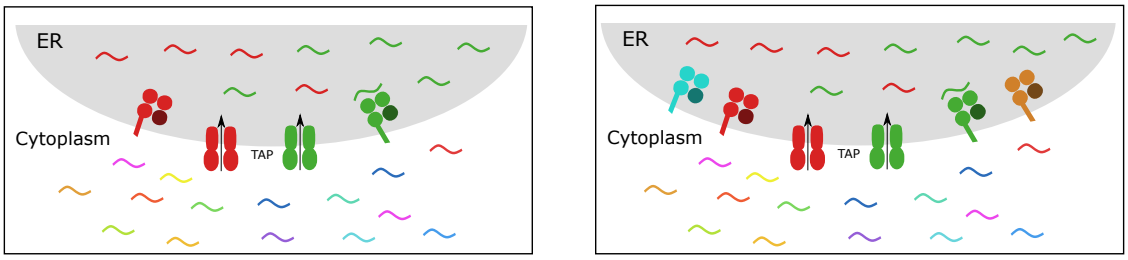
Figure 1.8: **Simplified structure of the chicken MHC** with the class I region situated in between class II and class III. The class I region contains the two classical class I genes, BF1 and BF2, which flank the TAP genes. The class II region contains two classical class II genes, BLB1 and BLB2, along with the non-classical DM genes and tapasin. The class II α chain is not in the MHC region in chickens.

The chicken MHC is much smaller than the human MHC, just 92 kb in length (Kaufman et al., 1999), and almost no recombination occurs across it, meaning that alleles of the genes present are inherited together as stable haplotypes. The class I, II and III regions are still present but the class I region is in between classes II and III, unlike in the human (figure 1.8). This region is sometimes referred to as the 'core' MHC, while the TRIM/Blec and BG regions upstream and the CD1 region downstream are considered part of the 242 kb 'extended' MHC-B locus (Shiina et al., 2007).

There are two classical class I (BF1 and BF2) and two classical class IIB (BLB1 and BLB2) genes, but both β_2-microglobulin and the class II α chain, BLA, are located on different chromosomes. For both class I and class II, only one of the two classical genes is expressed on the cell surface at high levels in most tissues - BF2 and BLB2. The minor class I gene, BF1, has been shown to be transcribed at a significantly lower level than BF2 in many chicken haplotypes (Shaw et al., 2007; O'Neill et al., 2009). This effect is seen across a variety of tissues (Wallny et al., 2006; Shaw et al., 2007), and while not all tissue types have been examined to date, the mechanisms by which BF1 expression is restricted (enhancer A deletion, enhancer A divergence and transcription start site deletion and rearrangement/deletion leading to pseudogenes (Shaw et al. (2007), F. Coulter, J. Kaufman, unpublished)) would appear to preclude the possibility of higher expression elsewhere in the body. For class II, dominant expression of BLB2 at the cDNA level has been shown in a wide range of tissues, but notably not in intestinal epithelial cells, where BLB1 is expressed at a higher level (Parker, 2012). This is concordant with minimal observable difference in the proximal promoters of BLB1 and BLB2, and their abilities to drive expression of a reporter gene to roughly the same extent (J. Jacob and J. Kaufman, unpublished, referenced in Parker and Kaufman (2017)).

The only non-classical genes in the chicken MHC are those for two DMβ and one DMα chains. It was proposed that BLB1 and DMB1 might function together in a intestine-specific system, while BLB2 and DMB2 formed a more general system (Parker and Kaufman, 2017). However, recent work has suggested that the interactions of these molecules are more complex than this, with both DM heterodimers plausibly contributing to peptide editing for both BLB molecules (Wise, 2019).

The chicken class I genes flank the TAP genes, unlike in the human where the TAPs are in the class II region. The tight linkage between TAP and class I in chickens has facilitated coevolution between these genes, since allele pairs can be functionally optimised through evolution without the risk of deleterious combinations arising through recombination. Walker et al. (2011) and Tregaskes et al. (2015) demonstrated that TAP transport specificity matched the peptide binding motif of the BF2 allele from the same haplotype and proposed this as the mechanism by which a single dominantly expressed class I is maintained in chickens. Under a system where TAPs are specialised for a particular class I molecule, expression of additional class I molecules with different peptide binding repertoires would be costly in terms of cell resources and would be unlikely to allow a greater variety of peptides to be presented on the cell surface (figure 1.9). As such, BF1 (even in haplotypes where it is expressed) is unlikely to play an important role in the presentation of antigenic peptides to T cells and may have been evolutionarily 'allowed' to degrade, since there would be little purifying selection for antigen presentation functionality.



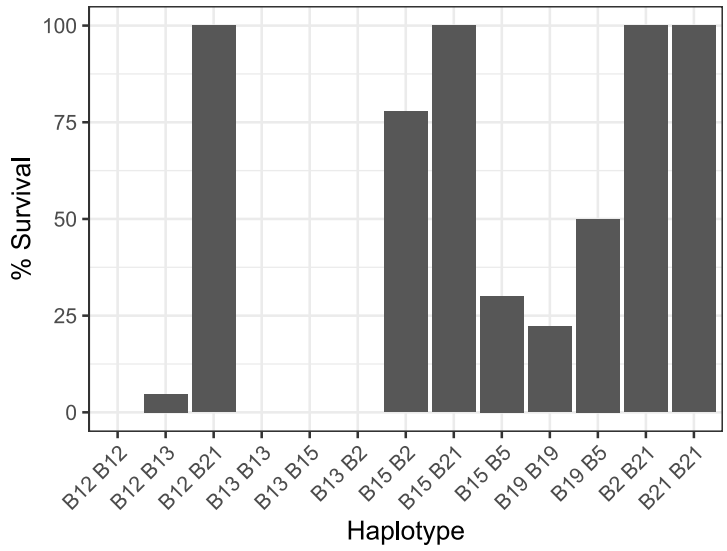
(a) (b)

Figure 1.9: **Chicken TAP transport and peptide binding.** Chickens have polymorphic TAP genes which coevolve with BF2 (red and green class I molecules) such that only peptides appropriate for the binding motif of the BF2 molecule are transported into the ER. (a) Some haplotypes have a deletion of BF1, while (b) in haplotypes where BF1 (turquoise and orange) is expressed at an appreciable level (although still less than BF2), very few of the peptides available in the ER through TAP transport are likely to bing to BF1 with high affinity.

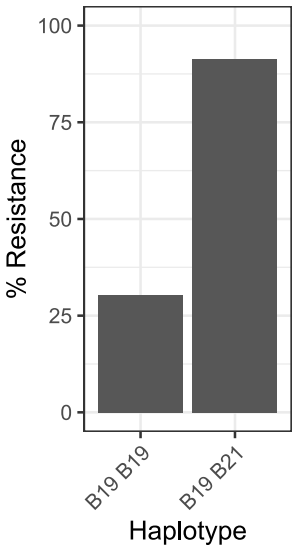
Nonetheless, BF1 has been maintained in most haplotypes, and may well have played a role in an ancestral haplotype, before being subject to degradation. It has been suggested that where BF1 is expressed it acts as an NK cell receptor, first based on a conserved motif that both resembles and aligns with the KIR binding site on HLA-C (Ewald and Livant, 2004) and later from functional studies showing that expression of BF1*021:01 (the BF1 molecule encoded by the B21 haplotype) inhibited NK cell lysis of target cells (Kim et al., 2018).

As in the human, tapasin is in the class II region in chickens, but the physical and recombinational distance between tapasin and the class I genes is sufficiently small that coevolution between tapasin and BF is also thought to occur (van Hateren et al., 2013). However, the exact characteristics that are coevolving with class I are less clear for tapasin than they are for TAP.

The presence of a single dominantly expressed class I molecule has been proposed as the reason for the strong genetic associations between MHC haplotype and the outcome of infectious disease observed in chickens (Kaufman, 2000, 2014). Whether or not BF1 is expressed at an appreciable level, specific TAP transport means that chickens only have two 'chances' to find a protective peptide when they get an infection, compared to the six 'chances' of humans (figure 1.9). The result of this is thought to be a reasonable probability that an individual chicken is simply unable to find a protective peptide from a given pathogen that it is able to present, which reads out as high mortality rates in chickens with particular MHC haplotypes. Key examples of this in diseases of public health (Avian influenza) and economic importance (Marek's disease) respectively are the results of Boonyanuwat et al. (2006) and Briles et al. (1977) (later expanded on and extended to vaccine efficacy by Bacon et al. (1981) and Bacon and Witter (1994)) (figure 1.10). Many other examples of MHC disease associations in chickens are reviewed in Miller and Taylor (2016).



(a) Data from Boonyanuwat et al. (2006) showing the haplotypes of chickens at a field site in Thailand after an outbreak of Avian influenza. Both dead and live chickens were typed, and the proportions of different haplotypes found in the two groups were used to calculate the survival rate for each haplotype. B21 shows a dominant protective effect, while extremely high mortality was associated with B12 and B13.



(b) Data from Briles et al. (1977) showing the protective effect of B21 in Marek's disease.

Figure 1.10: **MHC haplotype associations with infectious disease in chickens**

It is important to note that the minimal recombination across the chicken MHC makes it difficult to assign phenotypes to the action of a particular gene within the MHC region. A particular example is the mapping of the Marek's disease resistance trait to the butyrophilin-like BG1 gene by Goto et al. (2009), which suggests that at least part of the genetic influence of MHC haplotype on Marek's disease resistance can be attributed to BG1. Furthermore, especially in light of the greater number of class II than class I associations reported with infectious disease in humans (Blackwell et al., 2009), it is plausible that class II, and especially having a single major class II molecule, could also influence these associations.

The MHC and vaccine efficacy

Poultry vaccination is an incredibly important defence against the threat of avian diseases to food security and public health. Studies have shown that MHC haplotype has a major influence on the efficacy of various vaccines (Bacon and Witter, 1993, 1995; Butter et al., 2013) (figure 1.11). It is therefore possible that the success of vaccination programmes, especially in non-commercial settings where MHC genetics are not well controlled, could be significantly improved by more thoughtful matching of livestock to vaccines.

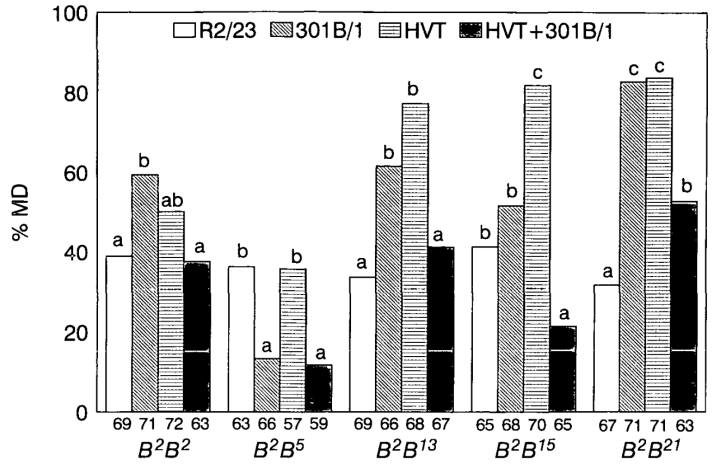


Figure 1.11: **Chicken MHC haplotype associations with Marek's disease vaccine efficacy.** Chicks from one homozygous and four heterozygous chicken lines were vaccinated with one of four Marek's Disease vaccines and challenged with very virulent Md5 at 8 days of age. The number of chicks in each experimental group is indicated at the base of each bar. The proportion of chicks with gross MD tumors, nerve lesions at death or necropsy at 10 weeks of age is shown as '% MD'. Within a chicken line, shared letters on top of bars indicate $p > 0.05$ (no significant difference) in a pairwise t test. No single vaccine strain provides the best protection for all 5 lines. Figure from Bacon and Witter (1995).

1.4 Generalists and Specialists

In both humans and chickens, it has been observed that different class I alleles have peptide binding repertoires of different breadths (Koch et al., 2007; Paul et al., 2013; Chappell et al., 2015). Those with wide peptide binding repertoires are frequently described as having a 'promiscuous' peptide-binding phenotype and acting as 'generalists', while those with narrow repertoires are described as being 'fastidious' in their binding or acting as 'specialists' (Kaufman, 2018). It should be noted that while these terms are often used interchangeably, the terms 'promiscuous' and 'fastidious' refer to observed difference in peptide repertoire, measured by mass spectrometry of peptides eluted from the class I molecule for example, while 'generalist' and 'specialist' refer to the hypothesised adaptive function of alleles with wide and narrow peptide binding repertoires respectively.

There are various strategies through which a class I molecule can accommodate a wide range of peptides. The best studied generalist molecule, BF2*021:01, was shown by x-ray crystallography to have a particularly wide binding groove due to the presence of small residues at positions 69 and 97 in the chicken α chain (equivalent to positions 70 and 99 in the human molecule). Furthermore, this allele's arginine residue at position 9 has significant conformational flexibility which allows the groove to be 'remodelled' to accommodate peptides with various residues (Koch et al., 2007; Chappell et al., 2015). A counterbalancing charge at position 24 (asparagine) allows charge transfer which further increases the range of permissible anchor residue combinations (Koch et al.,

2007). The result is that there is no clear motif for peptides eluted from BF2*021:01. Similarly, peptides eluted from BF2*002:01 and BF2*014:01 have no obvious motif due to broad hydrophobic pockets which are able to accommodate a wide range of peptides (Chappell et al., 2015).

The vast majority of chicken class I molecules also have an arginine at position 83 (which is almost invariant across non-mammalian vertebrates (Kaufman et al., 1994)), in contrast to the almost invariant tyrosine at the equivalent position 84 in mammals. Arginine is also seen in this position in mammalian class II molecules, which allow peptide C-termini to hang out of the peptide binding groove. Indeed, Xiao et al. (2018) demonstrated that BF2*012:01 bound peptides which extended beyond the F pocket, and attributed this to the conformational flexibility of Arg83 by mutating the residue to tyrosine and abolishing the ability of the overhanging peptide to bind. Despite the presence of Arg84 in all chicken MHC class I proteins suggesting that they all have the capacity to bind peptide which extend beyond the F pocket, very few alleles seem to do so with significant frequency. One exception is the 31-31-31-31 haplotype found in commercial chickens (Potts, 2016); peptides eluted from BF2*031:01 conformed to a relatively strict motif over their first eight amino acids but the length distribution included significant numbers of longer sequences than had been observed in peptides eluted from other class I molecules (N. Ternette and J. Kaufman, unpublished).

In contrast, fastidious molecules have narrow binding grooves with bulky residues that fill or overhang the groove to some extent (Wallny et al., 2006; Shaw et al., 2007; Zhang et al., 2012). BF2*004:01 also has an extremely positively charged binding groove with four inward-pointing arginines, which results in the specific binding motif DE2-DE5-E8 being clearly visible within eluted peptides (Wallny et al., 2006).

Repertoire breadth appears to be correlated with a suite of associated properties of class I molecules (Kaufman, 2017). Chappell et al. (2015) showed that, for both humans and chickens, peptide repertoire was inversely correlated with cell surface expression of the class I molecule. This relationship can be used to predict the peptide binding promiscuity of an allele based on flow cytometric analysis of surface class I on homozygous cells expressing that allele. The relationship apparently holds whether promiscuity is derived from a weak binding motif (such as in B21) or from other binding characteristics, such as the ability to allow peptides to extend beyond the F pocket (B31, E. Meziane and J. Kaufman, unpublished). Importantly, it has been shown in chickens that the difference in surface expression is not reflected in the overall rate of production of the class I protein in the cell (Tregaskes et al., 2015). The post-translational mechanism by which surface expression varies in chickens is still not fully understood. Tregaskes et al. (2015) also showed that high-expressing class I molecules with fastidious peptide binding motifs were more thermostable than

those with promiscuous motifs, suggesting enrichment of high-affinity peptides, possibly through increased peptide editing by molecules such as tapasin.

For human HLA-B alleles, predicted peptide binding repertoire (Košmrlj et al., 2010; Paul et al., 2013), tapasin dependence (Rizvi et al., 2014) and cell surface expression (Chappell et al., 2015) are roughly correlated (Kaufman, 2017). Reanalysis of the 11 alleles which had been included in both the studies by Paul et al. (2013) and Rizvi et al. (2014), confirmed that the correlation of predicted peptide repertoire and tapasin dependence was significant (Spearman's rho, $p < 0.05$), consistent with the later results of Bashirova et al. (2020).

Chickens have polymorphic tapasin genes and therefore may exhibit a more complex relationship. van Hateren et al. (2013) showed that BF2*019:01 was more tapasin dependant than BF2*015:01, despite both exhibiting high cell surface expression (Chappell et al., 2015) and similar, highly fastidious, peptide binding motifs (Kaufman et al., 1995; Wallny et al., 2006). Further work is needed to understand this relationship in chickens.

It seems likely that promiscuous alleles act as generalists, providing resistance to a wide range of pathogens (Chappell et al., 2015; Kaufman, 2018). The repertoire of the most promiscuous chicken alleles is greater than that of the most promiscuous human alleles (Koch et al., 2007) and thus a chicken generalist may have the ability to provide wide-ranging protection in a manner similar to that of a multi-gene family of class I genes in humans. The haplotypes containing the promiscuous alleles BF2*021:01 and BF2*002:01 have indeed been seen to associate with resistance to several important diseases including Marek's disease (Bacon et al., 1981; Simonsen, 1987), Rous sarcoma virus (Collins et al., 1977; Bacon et al., 1981), avian leukosis virus (Bacon et al., 1981), avian pathogenic *Escherichia coli* (Cavero et al., 2009), avian influenza (Boonyanuwat et al., 2006), avian coronavirus (Banat et al., 2013) and northern fowl mite (Owen et al., 2008). Resistance to Marek's disease in particular has been correlated with peptide binding repertoire across a number of haplotypes (Chappell et al., 2015) although the mechanism by which this protection is conferred remains unclear. Chappell et al. (2015) suggest that promiscuous peptide binding would result in activation of a larger number of T cell clones, which might provide a more effective immune response. Work by Mwangi et al. (2011) showing a limited repertoire of CD8 T cell clones infiltrating tumours in Marek's disease-susceptible chickens carrying the fastidious BF2*019:01 allele may support this hypothesis. Alternatively, promiscuous alleles may simply have a higher probability of being able to present the few truly protective peptides during an infection (Kaufman, 2018).

Fastidious alleles may provide resistance to specific pathogens that are, or were, particularly dangerous, although this has not been explicitly proven (Kaufman, 2018). They may be retained in the

population even once the specific threat has gone, because their specialist phenotype is recessive to the generalist phenotype of a promiscuous allele. This means that they are only strongly selected against in homozygotes, which get increasingly rare as the allele frequency decreases. There are no clear examples of fastidious alleles providing resistance to particular pathogens in chickens, however the HLA-B*27:05 and HLA-B*57:01 alleles, which are known to be associated with slower disease progression after HIV infection in humans, are high-expressing fastidious alleles (Košmrlj et al., 2010; Chappell et al., 2015) and a very similar motif is observed for the rhesus macaque alleles Mamu-B*008 and Mamu-B*017, which confer resistance to simian immunodeficiency virus (SIV) (Mothé et al., 2002; Loffredo et al., 2009). It is of note that the high tapasin dependence of HLA-B*27:01 and HLA-B*57:01 is actually protective, since rigorous editing in these cases ensures that the highly-conserved, protective epitopes that these alleles can present are loaded efficiently. Excluding these specific protective alleles, HLA molecules with promiscuous (as assessed by the number of HIV peptides eliciting response in *ex vivo* CD8 T cells from infected individuals), tapasin-independent peptide binding are associated with slower disease progression from HIV to AIDS (Bashirova et al., 2020), suggesting that resistance could be conferred by both mechanisms described above (activation of many T cell clones or presentation of key protective peptides) within a single system, by 'generalists' and 'specialists' respectively.

In chickens, coevolution between BF2 and the TAP transporter means that promiscuous class I molecules are inherited in haplotypes with TAP transporters with similarly broad peptide transport repertoires. Nonetheless, in heterozygous cells, fastidious class I molecules were inferred to present a wider range of peptides if a promiscuous haplotype was present on the other chromosome, suggesting that the TAP genes play an active role in peptide selection within some fastidious haplotypes (Tregaskes et al., 2015).

1.5 Origins and Evolution of the MHC

The sudden emergence of a seemingly fully-functional MHC at the base of the jawed vertebrates (with resolution limited by the many gaps left in the phylogeny by lineages which are now completely extinct) raises interesting questions about its origins. The evolution of novel pathways was discussed and theorized over at length in the mid-20th century, in the absence of much, if any, direct experimental evidence for any of the models proposed. The systems considered were almost exclusively metabolic pathways and three major hypotheses were eventually distilled: patchwork evolution, retrograde evolution and forward evolution (Fani and Fondi, 2009). The patchwork hypothesis proposed by Yčas (1974) and Jensen (1976) allows for the assembly of novel pathways from existing components with broad substrate specificity, which later specialise by gene duplication and subfunctionalisation. Several metabolic pathways have been invoked as examples of

this process including the degradation of toxic pentachlorophenol by *Sphingomonas chlorophenolica*. This pathway, proposed to have recruited enzymes from two existing degradation pathways, is thought to have evolved recently and its low efficiency is taken as evidence of the ongoing process of specialisation (Copley, 2000). Klein et al. (1993) proposed a similar 'composite origin' for the MHC.

Subsequent evolution of the MHC can be considered on two main scales, the relatively short-term and ongoing evolution of classical MHC genes within a species, and long-term, historic changes in structure and function within evolutionary lineages, which have led to observable variation in MHC region structure in extant species.

1.5.1 Generation and maintenance of diversity in MHC molecules

Classical MHC class I and II molecules are among the most diverse in the genome, both in terms of allelic polymorphism (the number of alleles) and sequence diversity (differences between alleles). The generation and maintenance of such diversity has several possible non-exclusive explanations, many of which are likely to play a role to a greater or lesser extent (Nei and Hughes, 1991; Spurgin and Richardson, 2010).

New alleles arise by mutation at random in the MHC, as in every other gene. There is no evidence that MHC genes are subject to a higher mutation rate than other genes in the genome (Hayashida and Miyata, 1983), so the accumulation of polymorphism seems to be more dependant on the maintenance than the generation of diversity.

Mechanisms by which MHC diversity can be maintained are often considered in two main categories: pathogen-mediated selection and sexual selection (reviewed in Bernatchez and Landry (2003)). Maternal-foetal interactions such as the role of HLA-C in human placentation (Moffett and Loke, 2006), may also play a role in the outcomes of adaptive trade-offs in MHC evolution (reviewed in Radwan et al. (2020)).

Pathogen-mediated selective processes include heterozygote advantage, negative frequency-dependent selection and fluctuating selection. The heterozygote advantage hypothesis posits that a greater diversity of MHC alleles allows an individual to present peptides from a wider variety of pathogens, and therefore confers resistance to more diseases. It was first proposed as an explanation for observed MHC diversity by Doherty and Zinkernagel (1975), and later supported by evidence that heterozygous individuals in populations show higher resistance to pathogens than do homozygous individuals (Kekäläinen et al., 2009; Oliver et al., 2009) and that populations with low overall MHC diversity are highly susceptible to certain pathogens (O'Brien et al., 1985).

Negative frequency-dependant selection proposes that pathogens are under selection to overcome the resistance of the most common host genotypes (in this case, MHC alleles). New alleles, or old, rare alleles, are therefore likely to confer a selective advantage in hosts (Takahata and Nei, 1990; Slade and McCallum, 1992). The inherent time lag between the evolution of a pathogen and the response of its host population results in cyclical dynamics of allele fitness and the maintenance of diversity via this dynamic process.

The third mechanism of pathogen-mediated selection is fluctuating selection, which describes the effect of spatio-temporal heterogeneity in type and abundance of pathogens on selection acting on the MHC (Hill, 1991). This differs from negative frequency-dependent selection in that the variation in pathogen pressure results from changes in the biotic or abiotic environment, or stochastic pathogen population dynamics, rather than from selection imposed by host defences on the pathogen. A model by Hedrick (2002) indicated that diversity at the MHC could be maintained by fluctuating selection even in the absence of either of the other pathogen-mediated mechanisms of selection.

Linked, in some ways, to heterozygote advantage is the role of sexual selection in the maintenance of diversity (Penn, 2002). In one model, sexual selection acts in either an absolute or relative way to give a reproductive advantage to those individuals who carry more diverse alleles or alleles more dissimilar to those of the 'chooser'. Indeed, individuals of many species have been seen to prefer mates who will increase the MHC diversity of their offspring (see references in section 1.2.2).

Alternatively, Potts and Wakeland (1990) proposed that MHC-disassortative mating was a mechanism of general inbreeding avoidance, by which individuals use detection of allelic variation at the highly polymorphic MHC locus to identify related individuals and thus avoid the problems associated with inbreeding such as the expression of recessive deleterious mutations. Jordan and Bruford (1998) suggested that populations at higher risk of inbreeding would therefore be more likely to exhibit MHC-disassortative mating.

1.5.2 Variation in MHC structure and evolution between taxonomic groups

The MHC is present in all jawed vertebrates (gnathostomes) but not in their extant sister group, the jawless fish (agnathans or cyclostomes), or any invertebrates (figure 1.12). However, whether or not all gnathostomes have an MHC region that can be effectively modelled by the human or chicken is still unclear.

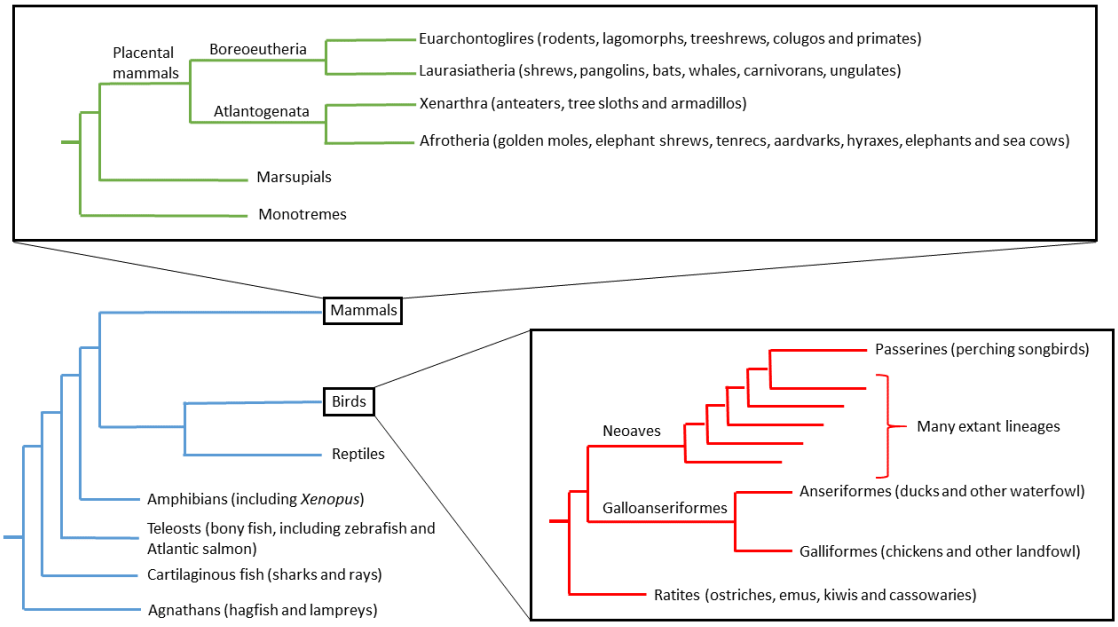


Figure 1.12: **Phylogeny of vertebrates** with expansions of the classes Mammalia (mammals) and Aves (birds).

It is necessary to infer the evolutionary trajectory of MHC regions and antigen processing and presentation systems from comparative analysis of extant species. The chicken and human systems are noticeably different, but determining where and when the significant evolutionary changes occurred is by no means straightforward and is confounded by significant within-group variation. Comparative genomics of gene networks is limited by the availability of high quality genome sequences for the species of interest and this has been a particular challenge for the MHC, which tends to be unusually difficult to sequence for various reasons including high polymorphism and the presence of clusters of related genes. Nonetheless, dedicated work by numerous groups has yielded information about the genetic architecture of a range of species form across the gnathostome phylogeny, particularly animals produced for food and model organisms for which some genomic and molecular resources were already available.

A useful framework is based on defining a 'chicken-type' system as tightly linked (i.e. low recombinational distance between genes), with coevolution between polymorphic TAPs and a single well expressed MHCI and a 'human-type' system as unlinked, with permissive and monomorphic TAPs allowing peptide loading onto multiple expressed MHCI molecules. Based on this framework, we can look for diagnostic features of each of the systems in a range of species and identify phylogenetic groups which contain these features, even if it is impossible to assess the full range of features for an individual species.

Fish

In the earliest-diverging lineage, the cartilaginous fish, linkage between TAPs and MHCI has been observed in the nurse shark. Further complexity could be contributed by variation in inducible proteasome components; only some sharks carry functional LMP7, LMP7-like or LMP2 genes and it has been suggested that peptide generation could thus be variable between haplotypes and possibly co-evolving with other parts of the pathway (Ohta et al., 2002).

Evolutionary flexibility in the adaptive immune system more generally is demonstrated by the debated absence or significant divergence of the canonical CD4 co-receptor and some associated transcription factors, cytokines and cytokine receptors in cartilaginous fish, despite the presence of polymorphic major histocompatibility complex class II molecules (Venkatesh et al., 2014; Dijkstra, 2014). The extent to which cartilaginous fish differ from other lineages in the presence or absence of particular cell systems, or whether similar functionality is achieved despite divergent sequences, remains the subject of debate.

The teleosts (bony fish) are an incredibly diverse group, with divisions between clades occurring deep in evolutionary time. Nonetheless, similarities can be observed between species, including maintenance of the ancient paralogue of TAP2, TAP2t. In zebrafish, TAP2 has at least two highly divergent lineages which share just 50% sequence identity and has been shown to be linked to a class I locus (McConnell et al., 2014). Furthermore, variation is observed at residues known to confer functional variation in rat TAP2 (Deverson et al., 1998; Joly et al., 1998; McConnell et al., 2016), suggesting that zebrafish TAP2 lineages are functionally distinct.

Recent work on Atlantic salmon has revealed a system highly reminiscent of that seen in the chicken. Two MHC regions have been identified but only one contains a classical, expressed MHCI gene (the other contains the non-classical UDA gene and two pseudogenes both identified as UCA). This single expressed class I is flanked by tapasin, proteasome components and TAP2, as are the non-classical and non-expressed class I genes in the paralogous region (Grimholt, 2018). Salmonid TAP2 does not seem to exhibit polymorphism which would affect peptide transport, although a sequence variant TAP2a_#C differs at many of the first 31 amino acids of the N-terminal region and may therefore differ in its interaction with tapasin (Koch et al., 2006; Grimholt, 2018).

Within lobe-finned bony fish, variation in the adaptive immune system more generally is seen in the coelacanth, which has been reported to lack IgM genes; two IgW genes are present, which could serve a compensatory function (Amemiya et al., 2013). Perhaps even more striking is the apparent loss of the class II system at least three independent times within the ray-finned fish.

The absence of MHC II α/β genes as well as associated receptors including CD4 has been reported in Gadiformes (cod-like fish) (Star et al., 2011), the pipefish *Syngnathus typhle* (Haase et al., 2013) and the anglerfish *Lophius piscatorius* (Dubin et al., 2019). Interestingly, pipefish exhibit male pregnancy, with the female laying unfertilised eggs into the male brood pouch, while anglerfish exhibit sexual parasitism, with the male and female circulatory systems fusing completely in some species. It may be the case that the derived immune system observed in these species is related to tolerance of these unusual reproductive behaviours.

Amphibians

The amphibian genus *Xenopus* is characterised by high variation in ploidy, yet functional diploidisation (Du Pasquier et al., 1977; Kobel and Du Pasquier, 1986) seems to maintain the single dominantly expressed class I which is expected under a chicken-type model (Flajnik et al., 1999; Shum et al., 1993). As in chicken, *Xenopus* TAP genes are adjacent and in opposite transcriptional orientations and linkage over a region containing class I, TAP and inducible proteasome component genes maintains stable haplotypes within biallelic lineages of these genes which seem to evolve trans-specifically (Ohta et al., 2003).

Sauropsids (Reptiles and Birds)

The chicken tends to be used as a model for birds, despite the Galloanserae (chicken and duck) lineage diverging from the Neoaves (all other birds except ratites) early in the evolution of the class. As expected, close relatives of the chicken within the Galliforme (landfowl) order (for example quail and guineafowl) share many features of their genomic architecture in the MHC region with the chicken. The sister group to the Galliformes, the Anseriformes (largely waterfowl), is typified by the duck which has a single dominantly expressed class I gene adjacent to a polymorphic TAP2 gene although further complexity in this system has recently been proposed in that a 'minor' class I in the duck could be regulated by the Let-7 microRNA and expressed at a higher level on infection (Chan et al., 2016). Within the Neoaves, there is remarkably little convincing data. For the crown group, the passerines, class I copy number is notably high and variable (O'Connor et al., 2016; Minias et al., 2019). A single dominantly expressed class I has been reported in the sparrow (Drews et al., 2017) but does not seem to be present, at least at the RNA level, in the siskin (Drews and Westerdahl, 2019). In zebra finch, Balakrishnan et al. (2010) used genomic analysis and FISH mapping to suggest that a single classical class I gene was present but that the TAP and class I genes were not syntenic, while Ekblom et al. (2011) were unable to detect sufficient polymorphism in the TAP genes to carry out single nucleotide polymorphism (SNP) mapping.

Limited work has been published on the structure of the reptile MHC, although the saltwater crocodile has been reported to have its TAP2 gene within a class I region (like other non-mammals)

rather than within a class II region (like humans) (Jaratlerdsiri et al., 2014). Examination of genome assemblies available in the NCBI genome database (Green sea turtle, (rCheMyd1.pri, GCF_015237465.1); Abingdon Island giant tortoise (ASM359739v1, GCF_003597395.1); Western terrestrial garter snake (rThaEle1.pri, GCF_009769535.1); Viviparous lizard (UG_Zviv_1, GCF_011800845.1)) suggests that reptiles are largely chicken-like in that the TAP genes are flanked on one or both sides by proximal MHC class I genes, however the TAP genes appear to be in the same transcriptional orientation, as they are in mammals, rather than opposite as they are in birds.

Mammals

Among mammals, examples of chicken-type MHC arrangements can still be observed. The earliest diverging mammals, the monotremes (typified by the echidna and platypus) resemble other non-mammalian vertebrates in that their class I and class II regions are contiguous on one of the five pairs of sex chromosomes, while a region with syteny to the human class III was mapped to a different pair of sex chromosomes. TAP genes are linked to a pair of MHCI genes (Dohm et al. (2007) and platypus genome assembly mOrnAna1.p.v1 (GCF_004115215.1)), for which expression has not been investigated.

The MHC of the sister group to the placental mammals, the marsupials, also resembles the non-mammalian MHC in its organisation, if not its size (Belov et al., 2006). The class I and class II regions are adjacent and somewhat interspersed, with the TAP genes closely linked to a single dominantly expressed MHCI gene in the opossum (Miska and Miller, 1999). The Tamar wallaby, conversely, does not have linkage between TAPs and MHCI but this has been attributed to convergent evolution facilitated by retroviral elements, rather than being related by descent to the lack of linkage seen in placental mammals (Siddle et al., 2011). Marsupial genome assemblies (Tasmanian devil (mSarHar1.11, GCF_902635505.1); koala (phaCin_unsw_v4.1, GCF_002099425.1); common wombat (bare-nosed wombat genome assembly, GCF_900497805.2)); common brushtail (mTriVul1.pri, GCF_011100635.1)) for the MHC resemble those of reptiles, with the TAP genes in the same transcriptional orientation but adjacent to MHC class I loci.

Unlike in monotremes and marsupials, features of the human-type system are seen in many placental mammals. However, resources have historically been extremely limited for species of the Atlantogenata, which contains the Xenarthra (armadillos, tree sloths and anteaters) and Afrotheria (golden moles, elephant shrews, tenrecs, aardvarks, hyraxes, elephants, sea cows). The division between the Atlantogenata and their sister group, the Boreoeutheria is a deep evolutionary branch point within the placental mammals and thus it is important to investigate MHC structure in Atlantogenata before confidently generalising across all placental mammals. Available genome assem-

blies for the southern two-toed sloth (mChoDid1.pri GCF_015220235.1), nine-banded armadillo (Dasnov3.0, GCF_000208655.1), African savanna elephant (Loxafr3.0, GCF_000001905.1), aardvark (OryAfe1.0, GCF_000298275.1), Cape elephant shrew (EleEdw1.0, GCF_000299155.1), small Madagascar hedgehog (ASM31398v2, GCF_000313985.2) and Florida manatee (TriManLat1.0, GCF_000243295.1) place the TAP genes in a class II region in Atlantogenata, as they are in the Boreoeutheria, suggesting a consistent MHC arrangement across the placental mammals.

The pig MHC, although smaller than that of the human, closely resembles it in terms of gene content and order. Three classical MHC class I genes sit in the class I region, interspersed with two classical MHCI pseudogenes and one possible classical MHCI pseudogene. Adjacent to the class I region is the class III region as in humans, and the class II region, which contains the TAP genes, is separated from the class III region by the centromere. The total distance between the pig classical MHCI and TAP genes is just under 2 Mb (Renard et al., 2006).

In the Yangtze finless porpoise, a relatively close relative of the pig, TAP genes have also been identified within the class II region, along with the inducible proteasome components PSMB8 and PSMB9 (Ruan et al., 2016). The same gene cluster is seen in the dog, cat, horse, cow and mouse (Viļuma et al., 2017). Class I tends to be less well characterised in mammals but multiple expressed classical class I genes have been localised to the expected class I region in species including the horse (Tallmadge et al., 2005) and cat (Okano et al., 2020).

Closely related to humans, as members of the Euarchontoglires (rodents, lagomorphs, treeshrews, colugos and primates), are mice and rats. The MHCs of these species have been well studied and close investigation has revealed features of their structure and function which differ from humans.

In both mouse and rat, the main class I, class III, class II region structure is conserved but there is a second class I region centromeric to the class II region. In the rat, the many class I genes in the telomeric class I region are non-classical, meaning that the only expressed, classical MHCI molecules come from the centromeric class I region. There appears to be copy number variation between haplotypes (between one and three loci), but the number of loci which are expressed at any time is unclear (Joly et al., 1996). The TAPs, located in the class II region as in humans, are therefore fairly close to the class I genes in rat and are able to co-evolve with them to some extent. Two major lineages of TAP genes have been seen, one with a highly permissive repertoire similar to human TAP (TAP-A) and one which cannot transport peptides with basic C-termini (TAP-B) (Deverson et al., 1998). As in chickens, although to a less specific extent, the peptide motif of the restrictive TAP-B heterodimer matches the binding motif of the RT1-A alleles found in the same haplotypic lineage; RT1-A alleles which co-segregate with TAP-B tend to have more

basic F pockets (Joly et al., 1998).

In mice, the two class I genes in the telomeric class I region are classical and expressed (although in some haplotypes one of the genes, H2-L, is deleted), as is the single class I in the centromeric class I region (H2-K). TAP polymorphism is observed to be equal to or slightly greater than that in the human (Marusina et al., 1997), but significantly less than that in the rat, with functional consequences for peptide translocation unlikely (Obst et al., 1995). Mouse TAPs are not, however, as permissive as human TAPs and show peptide specificity similar to rat TAP-B. This transport selectivity is reflected in the binding repertoires of mouse class I molecules, which tend to have hydrophobic F pockets (Rammensee et al., 1995).

Primates are also well-understood models. The well-studied rhesus macaque has multiple copies of the equivalents of HLA-A and -B, with variable numbers of genes and pseudogenes present in different haplotypes (Otting et al., 2005). Beyond this expansion of the class I *A* and *B* regions, which brings the total size of the rhesus macaque MHC to 4.7 Mb, the structure is very similar to that of the human MHC (Shiina et al., 2017). It is, therefore, somewhat surprising that the macaque TAP genes, like over 40 of the 60 immune-related genes in the MHC identified by Daza-Vamenta et al. (2004), are significantly more polymorphic in macaques than they are in humans, although the functional significance of this remains unclear.

Some of the additional class I gene copies in the macaque share characteristics with the non-classical human genes HLA-E, -F and -G, including low polymorphism (Otting et al., 2007). Alleles of the Mamu-A2 locus exhibit high allelic polymorphism but significant conservation of the peptide-binding cleft, resulting in a conserved peptide binding motif similar to that of HLA-B*27:01 and Mamu-B*008, which confer resistance to HIV and SIV respectively (Mothé et al., 2002; Loffredo et al., 2009). In contrast to HLA-B*27:01, however, Mamu-A2*05 presents predominantly 8mers and is only expressed on the cell surface at a low level under normal conditions, possibly because high affinity peptides are very rare. Indeed, de Groot et al. (2017) proposed that the Mamu-A2 locus has evolved a specialised role in scanning the activation of retroviruses. Such examples further highlight the complexity of strategies present in different species, which needs to be considered when attempting to describe concepts such as classical/non-classical or generalist/specialist alleles.

1.5.3 The placental inversion

By comparative genomics, the division between the chicken- and human-type systems was localised to the lineage leading to placental mammals. A model for the genomic change responsible was proposed in which an ancestral chicken-type system underwent an inversion with a breakpoint

between the single well expressed MHCI and the TAP genes. The inversion is proposed to have swung the class III region, which is on the outside of the class I region in chicken, into the middle and the expressed MHCI gene to the outside, leaving the TAP genes in what would become known as the human-type class II region (Kaufman (2011), figure 1.13). Such an inversion broke the linkage between MHCI and TAP, requiring TAPs to suddenly become sufficiently permissive to function with whichever MHCI alleles happened to be inherited. Once the TAPs had become non-specific, expansion of the MHCI gene family could occur, with a possibly optimal number eventually being fixed in each descendant species (Nowak et al., 1992; Bentkowski and Radwan, 2019)

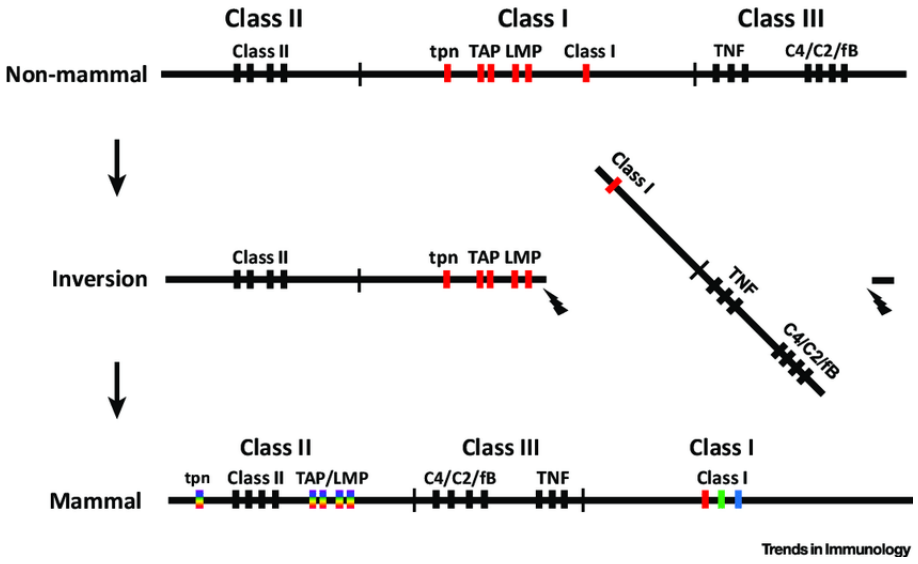


Figure 1.13: **The placental inversion.** An inversion on the lineage leading to placental mammals could have swung the class I region to the 'outside' of the MHC region, breaking the linkage between TAP and MHC class I and forcing class I antigen processing genes (TAP and tapasin) to become universal, since recombination could occur between them and the class I genes. The class I genes could then expand into a multi-gene family. Figure from Kaufman (2018).

1.5.4 Summary

The examples described in this section illustrate the evolutionary flexibility of the MHC region and the potential for immunological strategies to vary considerably even within relatively small groups of species. It is therefore important to apply insights from model organisms with care, and to investigate non-model organisms in order to understand the extent of potential deviations from models within a lineage. Furthermore, insights from chicken immunogenetics have proven valuable beyond the field of avian immunology and the discovery of new strategies could similarly reveal fundamental features of MHC biology with far-reaching impact.

1.6 Sequencing technologies

1.6.1 '3 generations' of progress in sequencing

Progress in comparative genetics and genomics has been tightly coupled to developments in DNA sequencing technologies. Nonetheless, it is not the case that the most recent technology is the best for every experiment and researchers must consider factors including read length, throughput, cost, speed and accuracy when deciding on a method (table 1.1). Modern sequencing technologies are generally considered with a '3 generations' framework, with an underlying technology fundamental to each generation but various platforms available. This project makes use of sequencing technologies from all three generations. The principles of each are presented here, rather than being introduced in the relevant chapter, to facilitate comparison.

	First generation	**Second generation**	**Third generation**
Example platforms	ABI Sanger	454, Illumina, Ion Torrent	PacBio, Oxford Nanopore
Read accuracy	High	Moderate-High	Low (but improving)
Read length	Moderate (800-1000 bp)	Low (<300 bp/end)	High (>100 kb)
Throughput	Low	High	High
Cost per run	Low	High	High
Cost per base	High	Low	Low

Table 1.1: **Summary of key features of sequencing technologies**

1.6.2 Sanger sequencing

Sanger sequencing is appropriate for low-throughput work because the cost per run is low. In this project it is used for verification and follow-up on specific samples after a high-throughput method and for sequencing experiments where the number of samples is low, particularly in cases where the sequence of one amplicon guides primer design for subsequent amplicons.

Sequencing is based on the incorporation of chain-terminating radioactively or fluorescently labelled dideoxynucleoside triphosphates by DNA polymerases during *in vitro* DNA replication (Sanger et al., 1977). Amplicons, beginning at the same sequencing primer but terminated at different points in the sequence, are then separated by length and the fluorescent signal is detected (figure 1.14). Samples submitted for sequencing should be single DNA species since the output is a single chromatogram representing the proportion of each base called at each position. It is therefore extremely difficult to distinguish individual sequences if the sample is found to be a mixture.

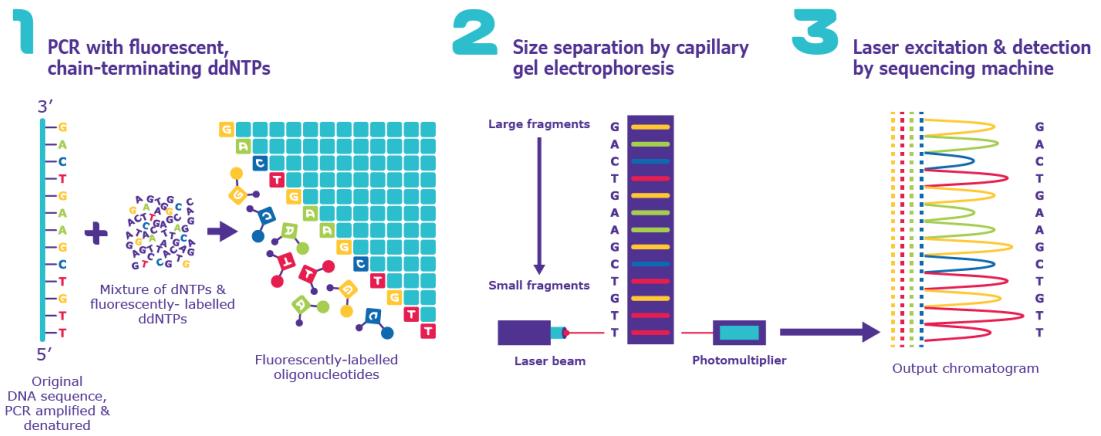


Figure 1.14: **Principles of Sanger sequencing.** Image from *https://www.sigmaaldrich.com/technical-documents/articles/biology/sanger-sequencing.html*

1.6.3 Illumina MiSeq

Illumina MiSeq is a second generation platform, allowing high-throughput sequencing of a DNA library containing multiple DNA species. Each output sequence corresponds to a cluster of DNA strands amplified from a single molecule, so mixtures of DNA species can be separated during analysis. Furthermore, DNA barcodes can be used to multiplex many samples on a single library. The maximum read length supported by the platform is 300 bp, although paired-end sequencing allows 300 bp to be read from each end of the DNA fragment for a total of 600 bp, although the sequence may be non-contiguous.

The platform requires ligation of Y-shaped adapters onto the DNA fragments, which correspond to oligonucleotides immobilised on the sequencing chip. This allows binding of the library to the chip and also facilitates the crucial 'bridge amplification' stage (figure 1.15).

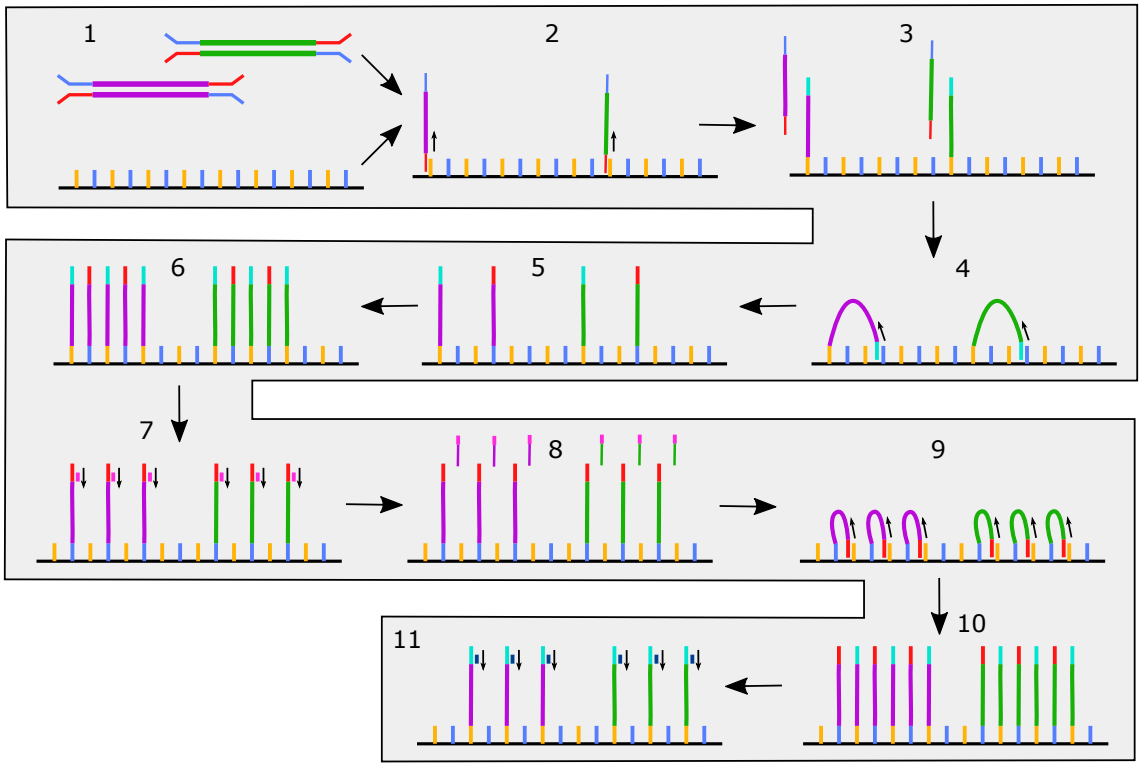


Figure 1.15: **Principles of paired-end Illumina MiSeq sequencing.** 1) Adapters (with tail sequences ***5*** (red) and ***7*** (blue)) are ligated to amplicons in the library. 2) Double-stranded DNA is denatured and the ***5*** adapter sequence binds to the reverse complement oligonucleotide (***5'*** (orange)) which is immobilised on the sequencing chip. The immobilised oligonucleotide acts as a primer for synthesis of the complementary strand by PCR. 3) The new DNA duplex is denatured and the original strand is washed off. 4) The tethered strand forms a bridge, with the ***7'*** (pale blue) sequence on the synthesised strand binding to the immobilised ***7*** oligonucleotide. The complementary strand is synthesised. 5) The bridge duplex is denatured. 6) The bridge amplification process is repeated 35 times until a cluster of strands is formed from each originally molecule bound in stage 2. 7) The strands tethered by ***7*** are cleaved and washed off, leaving just forward strands. A sequencing primer binds to the ***7'*** and the strands are sequenced by addition of reversible dye terminators which are incorporated into a synthesised strands by DNA polymerase. After each PCR cycle, the chip is imaged. The predominant colour that has been added to the strands in a given cluster during that cycle is visible and is interpreted as the next base in the sequence of that cluster. 8) The strands formed during sequencing are washed off. At this point, if indexed adapters have been used, an additional sequencing read can be performed. 9-10) A single cycle of bridge amplification regenerates the reverse strands. 11) The forward strands, tethered by ***5'***, are cleaved and washed off, a primer complementary to ***5*** is added, and the second sequencing read is performed.

1.6.4 Oxford Nanopore

Oxford Nanopore is a third generation platform which allows high-throughput, long-read sequencing with simple library preparation and real-time data reporting. The main trade-off is read accuracy, which is much lower than for Sanger or Illumina sequencing, although this has been steadily improving and now stands at 95% for raw reads (R10.3 chemistry, Oxford Nanopore). The high-throughput nature of the sequencing means that random errors should not be a major issue when a consensus is drawn from many sequences. However, nanopore sequencing also suffers

from non-random errors, particularly related to homopolymers (several of the same base in a row) in the DNA sequence, which are harder to detect and correct. The other major third generation platform, Pacific Biosciences (PacBio), has similar characteristics but is based on a single-molecule sequencing-by-synthesis strategy and does not use nanopores.

The concept of sequencing DNA through nanopores has existed since the 1980s (reviewed in Deamer et al. (2016)). Nonetheless, it was not commercialised until the Oxford Nanopore MinION was made available through the pre-release MinION Access Programme in April 2014 (Leggett and Clark, 2017).

The platform requires that adapters with motor proteins are ligated to the molecules in the DNA library. These adapters 'dock' onto biological nanopores inserted into an electrically resistant membrane. A current applied across the membrane is disrupted as the DNA strand passes through the nanopore, with the resulting trace indicating the sequence of bases in the specific DNA molecule that passed through the pore (figure 1.16). The flow cell has 512 sensors ('channels'), each connected to four nanopores (of which only one is active at any one time), meaning that the flow cell can generate up to 512 simultaneous traces, depending on the state of each of the respective channels.

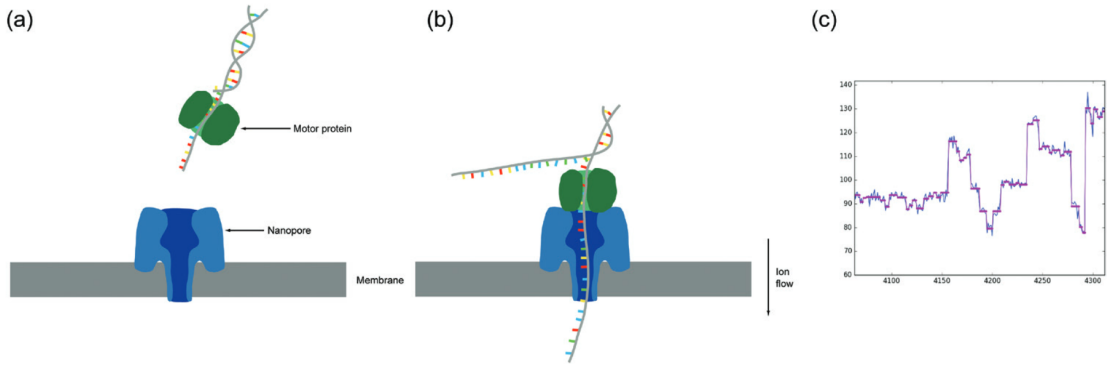


Figure 1.16: **Principles of Oxford Nanopore technology.** Figure from Leggett and Clark (2017).

1.7 Summary and Aims

The study of the chicken MHC has been important in three main areas:

- In livestock production and animal husbandry directly, to improve food security and public health
- As a model organism for non-placental vertebrates, providing a framework within which species-specific characteristics can be investigated for conservation purposes

- As a 'minimal essential' model of adaptive immunity more generally, providing insights into fundamental properties and functions of components of the adaptive immune system that can have implications in both human and veterinary medicine

However, there is still much that is unknown about the variation in MHC structure and function on taxonomic scales ranging from intra-species to inter-class. Current literature increasingly suggests that the MHC is highly flexible and adaptable, and it is important that this variation is understood if insights from decades of excellent work on experimental chicken lines are to continue to be applied effectively.

The aims of this project are therefore:

- To compare MHC class I and II diversity in a range of non-commercial chicken populations and understand the evolution of the classical class I and II genes with a larger dataset
- To assess whether key features of the 'chicken-type' MHC are conserved across birds, particularly the passerine crown group
- To investigate a novel component of the chicken antigen processing pathway, TAPBPL, and to assess the phylogenetic distribution of likely peptide editing components more generally

Each aim will be expanded upon in the introduction to the relevant section.

Chapter 2

MHC class I and II diversity in chicken populations worldwide

2.1 Introduction

2.1.1 Nomenclature

The nomenclature used to refer to the haplotypes and alleles of the MHC in chickens has evolved as the knowledge of their characteristics, and later sequences, has expanded.

The chicken MHC was originally described as the *B* blood group system, which determined erythroid alloantigens (Briles et al., 1950). During the 1970s and 1980s, the role of the *B* system in acute allograft rejection, mixed lymphocyte reactions and the immune response to antigens was gradually elucidated, followed by discovery of associations between *B* system genes and disease resistance, vaccine efficacy and autoimmunity (Miller et al. (2004) and references therein).

By 1981, the chicken MHC region was known to consist of three loci, referred to as B-F, B-G and B-L, which were polymorphic and controlled expression of the corresponding B-F, B-G and B-L antigens. A workshop was held to "compare the MHC haplotypes present in chicken flocks in different parts of Europe and North America, and to establish a uniform nomenclature for chicken MHC genes and their products" (Briles et al., 1982). A total of 27 serologically distinct *B* haplotypes were identified, with a single chicken strain designated as the reference population for each haplotype. Haplotypes were named based on the numerical system first used by Briles et al. (1957), with a capital B and a superscript number indicating the allele. It was also known that the B-F and B-G loci underwent a small amount of recombination and a nomenclature was proposed in which these recombinant haplotypes were named, for example, B^{4r1} for the first identified haplotype with

the same B-F and B-L regions as B^4, but a different B-G region. Finally, it was noted that alleles of individual loci could be referred to, for example, as B-F^1 for the BF locus allele in the B^1 haplotype.

A refined version of the nomenclature was developed during a series of meetings of the International Society of Animal Genetics and Avian Immunology Research Group in the early 2000s and published by Miller et al. (2004). By this time the full sequence of the chicken MHC and its component genes was known (Kaufman et al., 1999), although no new serologically distinct haplotypes had been discovered and the highest *B* haplotype number was still 29. The historic *B* numbers for haplotypes were retained, but were now written without superscript, and the names of component alleles within haplotypes referred to the specific gene followed by an asterisk, for example BF1*12 was the BF1 allele in the B12 haplotype.

Recombinant haplotypes were described with a base haplotype number derived from the BF genes followed by 'R' and a number to distinguish independent recombination events from the same base haplotype. For example, B4R2 was the second recombinant haplotype discovered to have the same BF component as B4 but BL or BG components from other haplotypes. The second haplotype involved in the recombination was not defined in the name, but described in the haplotype annotation. It was also acknowledged that some serologically defined standard haplotypes (for example B4 and B13) actually shared their BF and BL alleles, and were likely the product of historical recombination. It was decided that these haplotypes would continue to be referred to by their previously-assigned standard name.

It was also becoming obvious that there was allelic variation on finer scales (due to mutation or gene conversion) than could be described using single serologically defined haplotype numbers (Simonsen et al., 1982; Hunt et al., 1994). Sequences from different lines supposedly containing the same serological haplotype were distinguished using 'v' and a number; for example B19 and B19v1 (Miller et al., 2004).

Finally, the 2004 review of chicken MHC nomenclature (Miller et al., 2004) acknowledged the influence of the human HLA allele nomenclature which had, since 1987, implemented a system by which alleles were named according to their locus followed by 4 digits, the first two of which indicated the most closely associated serological group and the second two the specific sequence (Bodmer et al., 1989). Subsequent reviews of the human nomenclature added additional digits to describe alleles differing in specific ways, eventually separating these into distinct fields using colons to distinguish digits describing allele family (often corresponding to serological antigen), non-synonymous changes, synonymous changes in coding sequences and changes in introns or untranslated regions respectively (Marsh et al., 2010).

The nomenclature used throughout this project is described in Afrache et al. (2020) and closely resembles the human nomenclature described above. Allele names have a locus followed by an asterisk, with up to three subsequent numerical fields indicating allele groups, non-synonymous variants and synonymous variants respectively (e.g. BF2*084:01:02). The third numerical field is not shown, implying ':01', if there is only one known nucleotide sequence for a given protein sequence. BLB sequences which could not be assigned to a particular locus with confidence have the locus descriptor 'BLB'.

As described in Afrache et al. (2020), alleles were considered to be part of the same allele group (indicated by the same number in the first numerical field) if they differed at fewer than four amino acids for the BLB exon 2 sequences or eight amino acids (with no more than four per exon) for the BF exon 2-3 sequences. This was based on the BF2 sequences from the standard B15 and B19 haplotypes which differ at seven amino acids and are known to be derived from the same original B15 haplotype.

Since allele naming in this project refers only to the peptide-binding exons, in some cases a range will be stated in one of the fields (e.g. BF1*004:01-02) which indicates that multiple variants of the allele are known but they differ outside of the peptide-binding region and therefore cannot be distinguished with just the sequence data obtained in this project.

Full allele names also include a reference after the numerical fields, indicating the accession number of the allele if it has been previously reported in the literature, or the reference of the sequencing run in which it was first observed, if it was discovered during the project (e.g. BF2*021:01_AM282697 or BF2*081:01_run10_unk211). The reference is not usually included in figures or descriptions of results to improve readability.

With the BG region increasingly considered separately from the BF-BL region (Salomonsen et al., 2014), a 'haplotype' here is defined by a unique combination of BLB1, BLB2, BF1 and BF2 sequences, considering just the polymorphic peptide binding domains (exons 2 and 3 for BF and exon 2 for BLB). While the standard haplotypes are still commonly referred to by their 'B number' names, even when only the BF and BLB peptide binding exons are known, Afrache et al. (2020) proposed the use of 'Bfbl numbers' for combinations of BF (exon 2-3) and BLB (exon 2) sequences that are known to be derived from the same haplotype but where no further sequence information about the remainder of the linkage group is available.

Individual alleles are no longer necessarily named after their haplotype as they were in the 2004

chicken nomenclature, since it has become increasingly obvious that allele sequences can occur in multiple combinations. Instead, new haplotypes may also be referred to by abbreviations derived from the names of their component alleles. In this thesis, any following ':01's are not shown in the haplotype abbreviation, such that the haplotype BLB1*004:02_AB426152.1 - BLB2*021:01_AB426152.1 - BF1*004:01-03_AM279337 - BF2*021:01_AM282697 is abbreviated 4:02-21-4:01_03-21 (with an underscore in the range to distinguish it from the hyphens between alleles).

BLB alleles which could not be assigned to a locus (with names beginning BLB*) are indicated by brackets around the allele number in the four-field format of haplotype names. BLB alleles with a particular locus specified in their name which subsequently occur in a haplotype with another BLB allele assigned to the same locus are indicated by an asterisk against the number of the allele in the 'wrong' position such that BLB2*039:03_run13_unk24 - BLB2*005:02_run9_unk24 - BF1*006:01_AB426143 - BF2*111:01_run13_unk419 becomes 39:03*-5:02-6-111. Haplotypes known or suspected to be missing a locus are indicated with a question mark (?) in place of the null allele (e.g. 15-15-?-15).

2.1.2 MHC typing

The MHC is an important but notoriously difficult locus to type. Typing of the human MHC is crucial for successful transplant surgery, and this application has driven much of the progress in typing strategies. Still, the typical methods using sequence-specific primers, sequence-specific oligonucleotide probes, and Sanger sequencing-based typing are limited by time-consuming iterative protocols, low throughput, unphased data and ambiguity (Wittig et al., 2015). Classical MHC genes are characterised by dense SNPs, which rarely have sufficiently conserved adjacent sequences to allow oligonucleotide binding for high-throughput SNP typing. SNP typing in flanking regions combined with imputation algorithms can predict HLA alleles based on linkage disequilibrium with known SNPs, however the accuracy is relatively low, especially in populations where reference data sets are less complete (Karnes et al., 2017; Pappas et al., 2018). Next- and third-generation sequencing are becoming increasingly important in clinical contexts (Wittig et al., 2015; Mayor et al., 2015).

In non-model species, copy number variation and co-amplification of multiple loci make MHC typing for research purposes even more challenging (Babik, 2010). However, although transplantation is rarely the end goal, typing the MHC is a useful and important exercise in conservation, ecology, livestock management and public health due to close association of variation at this locus with disease responses (discussed in section 1.3). The chicken MHC is well-characterised and has

a small and conserved number of loci in all populations so far examined (Kaufman et al., 1999; Hosomichi et al., 2008) making it relatively amenable to typing. Anonymous markers, including microsatellites (Fulton et al., 2006) and an MHC-focussed SNP panel (which does not include SNPs in the classical MHC genes themselves due to high polymorphism preventing identification of conserved probe sites) (Fulton et al., 2016a) have been used to assess diversity and uncover genetic associations in various chicken populations (McElroy et al., 2005; Fulton et al., 2016b; Nguyen-Phuc et al., 2016). However, beyond the detection of markers associated with the few well-characterised haplotypes in experimental chicken lines (Hosomichi et al., 2008; Walker et al., 2011), they provide no information about the MHC alleles themselves.

Potts et al. (2019) established a method based on restriction strand-mediated conformation analysis (RSCA) to accurately type the chicken MHC based on classical MHC allele sequences directly. This was successful, and provided the first indication that commercial flocks might have unusually low MHC diversity (Potts, 2016). However, the process was time-consuming, and required manual cloning and sequencing of new alleles.

2.1.3 PCR-NGS typing of commercial chickens

To examine large numbers of MHC alleles from non-standard populations in a fast, cheap and high-throughput manner, a pipeline was developed to amplify the variable peptide-binding regions from the classical class I and class II genes of chickens and sequence them using the Illumina MiSeq platform. Development of the method was led by Dr. Clive Tregaskes and Prof. Jim Kaufman (University of Cambridge). The pipeline methodology is discussed in detail in section 2.2.

Original reference lists

The data analysis pipeline (section 2.2.4) detects allele sequences present in each bird sample in the library and assigns them names based on comparisons to a list of known reference alleles that is provided to the program. If a perfect match is detected between a sequence found in a bird and a sequence on the reference list, that allele is said to have been 'called' in that bird, and the name of the allele is reported in the output file.

Initial reference lists of alleles were compiled by Prof. Jim Kaufman (class I) and Dr. Hassnae Afrache (class II). The lists included any BF or BLB sequence that was reported in the published literature, as well as direct submissions to the GenBank databases from trusted labs. Sequences were given allele numbers according to the standard nomenclature if they were reported as having been seen in two independent polymerase chain reactions (PCRs). Those sequences which were not considered to have been fully verified in their original publication were listed with just their

locus and accession number. If they were later verified by PCR-NGS (polymerase chain reaction - next generation sequencing; the term used for the Kaufman lab typing program methodology), the original accession number was retained in the allele name, and an allele number was assigned in addition.

Results from runs 1-7

Runs 1-7 typed commercial broiler and layer lines from several breeding companies, including historic lines which were removed from the main breeding stock at various points and maintained without further selection. Runs 1 and 2 included samples that had been previously typed by RSCA (Potts, 2016) in order to verify the accuracy of the pipelines (L. Elder, C. Tregaskes and J. Kaufman, unpublished).

The results of the MHC typing were concordant with the findings of Muir et al. (2008), who reported low diversity at 2551 non-MHC SNPs in 1440 commercial birds relative to the ancestral breeds. However, by analysing MHC class I expression on the surface of red blood cells by flow cytometry, it was further shown that the few MHC haplotypes present in commercial breeds were predominantly those that expressed class I at a relatively low level on the cell surface (E. Meziane, F. Coulter, S. Hilton, E. Doran and J. Kaufman, unpublished) suggesting a wide peptide binding repertoire and a 'generalist' disease resistance phenotype (Chappell et al., 2015; Kaufman, 2018). A novel low-expressing haplotype not previously identified from experimental lines, now referred to as B31, was seen at over 50% frequency in some flocks (Potts, 2016), alongside more familiar low-expressing haplotypes such as B21 and B2.

2.1.4 History of domestic chickens

The genetics of domesticated animals are strongly influenced by the artificial selection and reproductive isolation imposed on them by humans as well as by natural selection in the new environments in which they have been placed. An understanding of the ancestry of modern populations can help to explain genetic patterns observed, and also influence the ways in which the outcomes of research are applied, for example by suggesting how widely conclusions drawn from a particular breed or population can be generalised across other populations.

Domestication and spread

It has been widely reported that there were probably several independent domestication events across India and South-East Asia (Liu et al., 2006; Kanginakudru et al., 2008; Miao et al., 2013), although recent whole-genome sequencing data shows domestic flocks as a largely monophyletic clade derived from a single subspecies, *Gallus gallus spadiceus*, whose modern range covers south-

western China, northern Thailand and Myanmar (Wang et al., 2020). The red junglefowl (*Gallus gallus*) was the primary ancestor, however the genomes of modern domestic chickens show signals of extensive introgression from the closely related Grey junglefowl (*Gallus sonneratii*) and (to a lesser extent) the slightly more distant Green and Ceylon junglefowl (*Gallus varius* and *Gallus lafayettii*) (Lawal et al., 2020), as well as other *G. gallus* subspecies. A key example of the complex ancestry of modern chickens is the yellow skin phenotype, which was shown to be due to an allele of *BCDO2* originating from the Grey junglefowl (Eriksson et al., 2008), which has a restricted range on the Indian sub-continent. Indeed, there were probably two independent introgression events related to this gene, since the *BCDO2* sequences of the White Leghorn and Chinese Shek-ki breeds are on different branches of the phylogenetic tree of this gene (Eriksson et al., 2008).

The date of chicken domestication is still unclear. Mitochondrial DNA extracted from chicken fossils in Northern China suggested that the earliest domestic chickens could have been part of a mixed farming system in the region as early as 8000 BC (Xiang et al., 2014), however these data, their interpretation and the conclusions drawn are still controversial. It seems fairly clear that chickens had been domesticated by at least 2500 BC, with seals depicting fighting cocks, clay figurines of chickens and chicken bones larger than those of the wild junglefowl found at the Mohenjo-daro city site in Pakistan. Leisure and game were probably the first motivation for chicken domestication, which must have taken place in agricultural regions where feed could be made available for chickens without competing with humans (Tixier-Boichard et al., 2011). A recent molecular clock analysis suggested that domestic chickens diverged from *G. g. spadiceus* 9500 ± 3300 years ago (Wang et al., 2020), however this node does not necessarily reflect the date of domestication; the same divergence analysis performed on whole genomes of dogs and wolves dates the divergence 15,000 years earlier than the accepted archaeological evidence (Wang et al., 2016).

By around 1000 BC domestic chickens from Southeast Asia were present on the Pacific island of Vanuatu (Storey et al., 2010). Slightly more controversial is the theory that chickens were transported much further across the Pacific via Polynesia into pre-Colombian South America (Storey et al., 2007; Gongora et al., 2008). Fitzpatrick and Callaghan (2009) showed that travel via this route would have had up to a 40% chance of success at certain times of year, so it it certainly feasible that chickens could have been introduced in this way. Furthermore, the blue egg shell phenotype is present in the South American Araucana and various Chinese breeds but not in European breeds, supporting a pre-Colombian introduction of chickens. DNA evidence, however, has been used to both support and oppose the theory (Storey et al., 2007; Gongora et al., 2008; Luzuriaga-Neira et al., 2017; Herrera et al., 2020), with the historic genetic composition of the Easter Island chicken population (an intermediate point on the route from Polynesia to South

America) a key point of contention. Further ancient DNA, radiocarbon and stable isotype analysis will be required to confirm or refute this idea definitively.

There appear to have been two independent routes by which chickens reached Europe, although the exact geography of the routes is still debated. By 700 BC domestic chickens were established in Greece and were later developed as a source of food, alongside their traditional uses in leisure, religion and divination, by the Romans (Tixier-Boichard et al., 2011). A second wave of imported Asian breeds during the 19th century led to the development of a 'younger' set of European breeds, which remain genetically distinct from the descendants of the early European chickens (Lyimo et al., 2014).

Global genetic diversity

The earliest large scale assessments of chicken diversity used microsatellites. The resolution of these studies was relatively low, and the markers were anonymous and therefore difficult to relate to specific traits. Nonetheless, microsatellite analysis allows the calculation of population-level measures of diversity (such as F-statistics) and can be used to detect admixture or reproductive isolation between different populations. Early studies revealed low diversity in commercial chickens and a high degree of population stratification in European chicken breeds (Hillel et al., 2003). Later studies, using slightly expanded sets of markers and more samples, confirmed these findings and additionally reported minimal population stratification in African and Asian breeds (Lyimo et al., 2014).

Other key analyses of global diversity have been based on sequencing of the mitochondrial DNA hypervariable segment I (HVS-I). Liu et al. (2006) found nine divergent mtDNA clades (A-I) in 834 Eurasian domestic chickens and 66 Southeast Asian and Chinese Red Junglefowl, of which A, B and E were distributed ubiquitously, C was mainly distributed in Japan and Southeast China, F and G were exclusive to Yunnan, China and D was closely related to the distribution of cockfighting (China, Japan and Madagascar). H and I were rare and the distribution reported in this first paper is likely to have been an artefact of the small sample size. Miao et al. (2013) expanded the analysis to include chickens from Europe, Africa, South America and the Pacific, detecting the same nine major clades, as well as an additional four red junglefowl-specific clades (W-Z). It was proposed that these clades represented different origins (and subsequent expansions) for domestic chickens across China, Southeast Asia and the Indian subcontinent and they have been used as a framework for many subsequent studies looking at the origins and ancestry of chicken populations worldwide.

More recently, SNP panels have become popular for high-resolution assessments of diversity. Early

studies used panels of a few thousand SNPs and further confirmed the loss of diversity in commercial stocks relative to ancestral breeds (Muir et al., 2008). Wider and more detailed analyses are now possible, with a high density 600K SNP panel (Kranis et al., 2013) used in a global analysis of chicken diversity through the SYNBREED project (Malomane et al., 2019), in which genetic relationships corresponding to origin, geographic expansion, selection and different management practises were observed. In general, increased genetic distance from the wild ancestors was associated with lower genetic diversity in populations, attributed both to the natural bottlenecks of migration and also to human management practises (Malomane et al., 2020).

There is still, however, little known about the functionality of this variation, although some attempts have been made to identify gene classes and functions that seem to have been under selection from SNP data (Fleming et al., 2016; Malomane et al., 2020). The MHC region in particular is subject to strong selective pressures, meaning that it can evolve very differently to neutral markers.

Whole-genome sequencing (WGS) is increasingly becoming affordable on high-throughput scales, and has recently been used in a study looking at 863 whole chicken genomes, of which 787 were sequenced specifically for this comparative work (Wang et al., 2020). As well as developing insights into to the ancestry of domestic chickens and subsequent interbreeding between domestic and wild populations, the authors were able to detect important signals of positive selection on genes involved in behaviour, growth and reproduction including including GNRH-I (gonadotropin-releasing hormone 1) and KIF18A (kinesin family member 18A). While the MHC regions of these genomes have not yet been specifically analysed, it can be difficult to accurately assemble such highly polymorphic regions and any insights related to MHC variation in these samples would need to be verified using additional techniques.

Village chickens

Free-range, scavenging village chickens are now found in all agroecological zones across Africa, from villages in the humid and subhumid tropical rain forests of West and Central Africa to the temperate highlands of East Africa and the arid and semi-arid regions of the Sahel and Kalahari deserts (DAGRIS, 2007; Mwacharo et al., 2013). They have high sociocultural significance in many modern African societies.

The history of chickens in Africa is likely to have been complex and characterised by multiple independent introductions via both terrestrial and maritime routes. On introduction, chickens may have been preferable to the native guineafowl (which are still kept in conjunction with the chicken) because they are easier to control and less prone to becoming feral (MacDonald, 1992). Three key routes of introduction were likely to have been 1) overland from the north, through

Egypt and Sudan 2) via the Indian Ocean trading network between Southeast Asia and East Africa and 3) later, through the terrestrial and maritime expansion of European empires. More recent introductions of exotic breeds for higher productivity may also have impacted local breeds through intentional and non-intentional cross-breeding (reviewed in Mwacharo et al. (2013)). The yellow skin phenotype is widespread in Africa, and has been used to support a predominant role for the Indian Ocean maritime trading network in the establishment of domestic chickens in Africa (Eriksson et al., 2008; Mwacharo et al., 2013).

Understanding the genetics and population genetics of village chickens is extremely important for the success of poultry development and conservation initiatives, and could provide opportunities for Western commercial breeding to benefit from new genetic resources. The links between possible observed genetic characteristics of a population and considerations for interventions are summarised in figures 2.1 and 2.2. Interventions could include such strategies as the introduction of new breeds, traits or breeding programmes, vaccination, changes in management practises or changes to feeding among others.

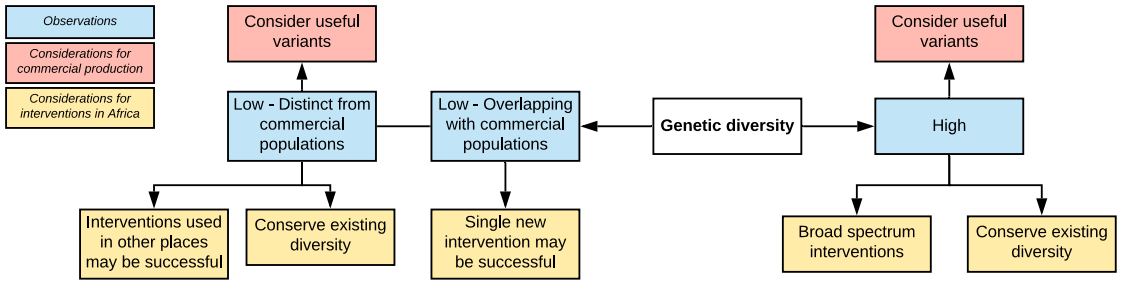


Figure 2.1: **Potential interpretations of data and applications of insights from genetic diversity studies in African free-range local chickens.** Genetic diversity in a population of African chickens may be high, low but distinct from commercial populations or low and overlapping with commercial populations. Given the low diversity observed in commercial flocks, a high diversity population is assumed to contain novel genotypes. Diversity that is not present in commercial flocks could include genetic resources that could be usefully incorporated into commercial breeding strategies if fully understood. The diversity observed in the population of interest can also indicate whether interventions used in commercial production could be successfully implemented in the population, or whether, for example, multiple vaccine strains might be necessary given that MHC haplotype is known to influence vaccine efficacy. Populations with low diversity are often priorities for conservation efforts, since low genetic diversity reduces the ability of a population to respond to change, and makes them vulnerable to the effects of inbreeding depression. Populations with high diversity should also be monitored and their diversity maintained, since this helps to ensure adaptability and resilience in the future and provides a reservoir of traits that may be useful when they become better understood.

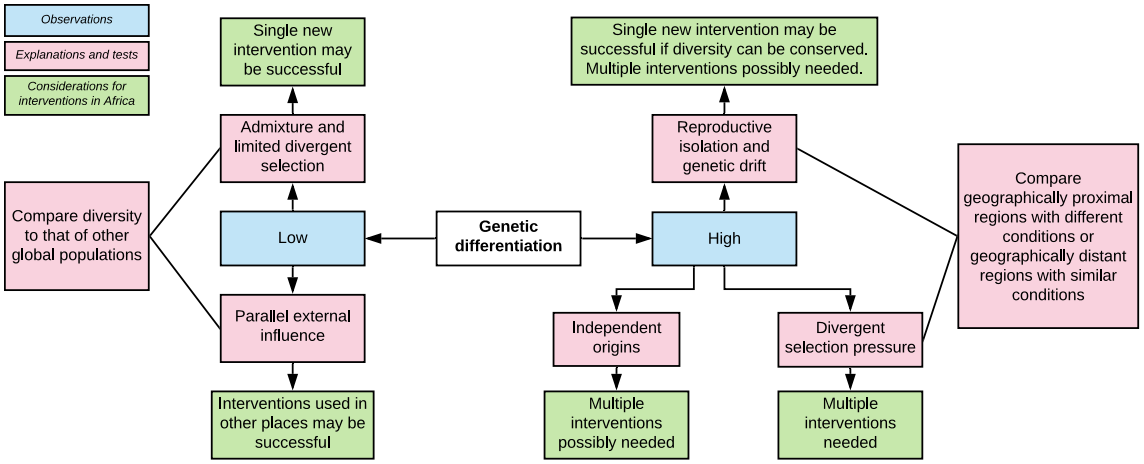


Figure 2.2: **Potential interpretations of data and applications of insights from genetic differentiation studies in African free-range local chickens.** Differentiation, or the degree of segregation of alleles within sub-populations, can vary on a spectrum from low to high across various geographic scales. Assuming a single origin, a high degree of differentiation could indicate reproductive isolation and genetic drift, divergent selection pressures, or a combination of multiple effects. Attempts to distinguish between these explanations can be made by comparing sub-populations in geographically proximal regions with different ecological conditions and sub-populations in geographically distant regions with similar ecological conditions. This distinction can suggest whether multiple interventions are necessary for the different sub-populations. It may be necessary for sub-populations affected purely by genetic drift to also be managed differently if a segregating trait affects the efficacy of a particular intervention. It is also important that a single intervention applied across multiple sub-populations doesn't lead to a major loss of genetic diversity, since birds in free-range local production systems need greater adaptability than commercial flocks. Differentiation could also be due to separate origins for populations, with differences maintained by selection or drift. Low genetic differentiation can be explained either by interbreeding between sub-populations and similar selection pressures acting on both sub-populations, or a single external influence, such as the introduction of a commercial production strain in multiple regions simultaneously (Leroy et al., 2012; Wimmers et al., 2000).

A number of studies have looked at the genetics of African village chickens using the methods described above. The rationale for the majority of these studies was to identify conservation priorities by assessing the genetic diversity present in a subset of populations from one or a few countries. However, their results also give insights into general patterns of diversity and differentiation, which can be considered according to the framework presented in figures 2.1 and 2.2, allowing informed design of interventions. Key relevant studies using microsatellite, mtDNA and SNP typing respectively are presented in tables 2.1, 2.2 and 2.3.

One additional study of interest used a single microsatellite locus, the MHC-linked LEI0258, rather than the typical panel of 20-30 microsatellites (Mwambene et al., 2019). While this reduces the validity of general conclusions about the diversity of the Tanzanian populations in question, LEI0258 alleles have been shown to correlate with with MHC haplotype in several commercial and indigenous lines (Lima-Rosa et al., 2005; Chazara et al., 2011), and thus LEI0258 diversity is a useful indicator of MHC diversity. Indigenous populations with less well-characterised haplotypes cannot,

at this stage, be accurately MHC haplotyped by LEI0258 alone since there is a lack of appropriate reference flocks in which both LEI0258 and MHC haplotype-specific typing have been performed. It is hoped that the MHC allele-level typing performed in this project can be linked to microsatellite variation, providing an easy, low-cost way for other groups to assess MHC diversity in similar populations.

General trends within Africa seem to include high overall diversity (see references in tables 2.1, 2.2 and 2.3), which has been reduced in some conservation flocks, possibly limiting the positive impact of these interventions. Differentiation between sub-populations on the scale of individual countries seems to be low, suggesting that there is reasonable gene flow, probably facilitated by local market networks. Geographical distance does seem to be correlated with genetic distance on international scales however, likely due to limits on gene flow and/or separate origins for flocks in different geographical regions. Low differentiation also suggests that the majority of the genome is not under diversifying selection in different agroecological zones in most countries, although this does not preclude important selective pressures acting on specific loci as evidenced by visible phenotypic differences between ecotypes. A few studies did report differentiation related to environmental factors, so this may be important across a larger proportion of the genome in some regions. Introgression from European commercial populations is clearly visible in many populations.

Fancy breeds

During the early- to mid-nineteenth century the collection of exotic breeds of chickens became incredibly fashionable, primarily inspired by the menagerie of Queen Victoria, who was provided with birds (and other animals) from wherever British explorers happened to be returning. A particular favourite was the Cochin from China, and eggs of this breed, along with other exotic breeds who were laying in the Royal aviaries, were sent across Europe to other Royal families, where chicken fancying became equally popular. 'Hen fever' in the US peaked equally fast, with the 'magnificent exhibition' of the 1849 Boston Poultry Show and subsequent regional shows promoting new and exotic breeds (Rude, 2015).

Modern fancy breeds, many of which have their origins as hobby birds during the peak of 'hen fever' are regulated by breed standards such as the British Poultry Standard or the American Standard of Perfection. These state the externally observable qualities such as appearance, movement and temperament, which must be exhibited by an animal to be considered a true representative of the breed. While there is no genetic definition of these breeds, outbreeding is rare since breeders want to preserve the standardised combinations of characteristics. Each breed can therefore be considered reproductively isolated. There is likely to be some sub-population structure as a result of regional communities of breeders who might periodically sell or exchange birds for breeding.

Study	Population of interest (N)	Control/comparison populations	Number of MS used	Observed heterozygosity in population of interest	Key relevant findings
Wimmers et al. (2000)	23 populations from Tanzania, Nigeria and Cameroon (405)	Germany, India, Bolivia, Experimental	22	0.53-0.67	African breeds similar to a commercial improver strain. Similarity between Tanzania and Nigeria possibly the result of crossing with the same exotic breeds rather than exchange between countries.
Muchadeyi et al. (2007)	5 ecotypes from Zimbabwe (238)	Malawi, Sudan, Commercial	29	0.59-0.625	No substructure found between agro-ecological zones in Zimbabwe but differentiation from Malawian and Sudanese chickens
Mwacharo et al. (2007)	15 populations from Kenya, Uganda, Ethiopia and Sudan (657)	Commercial	30	-	Significant positive correlation between genetic and geographical distance. Kenya/Uganda form a distinct cluster from Sudan/Ethiopia
Hassen et al. (2009)	7 populations from Ethiopia (147)	South Africa, Commercial	22	0.74-0.93	Ethiopian chickens not strongly influenced by commercial introgression. Phylogeny reflects population geography and marketing places.
Osei-Amponsah et al. (2010)	2 ecotypes from Ghana (114)	Commercial, RJF	22	0.539-0.596	Chickens in forest and savannah eco-zones from a single, randomly mating population
Mtileni et al. (2010)	4 populations from South Africa (128)	-	29	0.61-0.63	2 indigenous field populations form a single genetic cluster. Conservation flocks from the same founder populations have reduced diversity and increased differentiation.
Youssao et al. (2010)	2 ecotypes from Benin (121)	-	22	0.536-0.568	Significant phenotypic differences between forest and savannah ecotypes but low genetic differentiation
Goraga et al. (2011)	5 ecotypes from Ethiopia (155)	Commercial	26	0.53-0.57	Admixture between ecotypes but within 2 segregated clusters
Leroy et al. (2012)	23 populations from Cameroon, Benin, Ghana, Côte d'Ivoire, and Morocco (472)	Commercial	22	0.51-0.67	Evidence of gene flow between commercial and local populations in Morocco and in Cameroon,attributed to long-term improvement programs with the distribution of crossbred chicks. Limited impact of introgression, probably because of poor adaptation of exotic birds to village conditions, and because of the consumers' preference for local chickens. No such gene flow was observed in Benin, Ghana, and Cote d'Ivoire, where improvement programs are also less developed. Three genetically differentiated areas ($P<0.01$) were identified, matching with Major Farming Systems (namely Tree Crop, Cereal-Root Crop, and Root Crop) described by the FAO.
Lyimo et al. (2013)	5 ecotypes from Tanzania (196)		29 (+mtDNA)	0.56-0.67	Morogoro-medium and Ching'wekwe cluster together and separately from Kuchi in both analyses. Unique haplogroup in Kuchi suggests a recent introduction from Japanese game birds, reflected by morphology.
Berima et al. (2013)	6 breeds from Sudan (147)	Malawi, Zimbabwe, Commercial	29	0.461-0.581	No substructure within Sudan but differentiation from Malawian and Zimbabwean populations
Mwacharo et al. (2013)	15 populations from Kenya, Uganda, Ethiopia and Sudan (657)	-	30	0.45-0.61	Ethiopia/Sudan (gene pool I) distinct from Kenya/Uganda. Gradient of admixture between gene pools II and III east-west across Kenya and Uganda. Differentiation attributed to separate introductions not subsequent selection or introgression from exotic stocks.
Mahammi et al. (2016)	3 ecotypes from Algeria (233)	France, Commercial, RJF	23	0.55-0.62	Low levels of genetic differentiation attributed to exchange of birds between farmers at local markets. Small but significant influence of altitude/temperature on diversity detected in a supervised analysis.
Habimana et al. (2020)	5 ecotypes from Rwanda (313)	Commercial	28	0.598-0.626	Four gene pools, three related to genetic distance, one influenced by commercial introgression. One to two ecotypes per cluster, one ecotype is present in two clusters due to commercial introgression.

Table 2.1: **Key microsatellite (MS) typing studies in Africa.** N is the sample size of the African populations (excluding any control or comparison populations). Observed heterozygosity (H_O) is included as a measure of population diversity where available. Key findings relevant to this project and the framework presented in in figures 2.1 and 2.2 are briefly summarised.

Study	Population of interest (N)	Control/comparison populations	Haplotype diversity	Key relevant findings
Muchadeyi et al. (2008)	5 ecotypes from Zimbabwe (99)	Malawi, Sudan, Germany, Commercial	0.61-0.73	96.79% of variation was within ecotypes. 2 maternal lineages present but both in all 5 ecotypes (i.e. limited differentiation).
Mwacharo et al. (2011)	23 populations from Kenya, Ethiopia, Sudan and Uganda (512)	-	0.167-0.923	Results suggest two independent introductions from India and South East Asia followed by more recent introgression of commercial haplotypes.
Eltanany and Hemeda (2016)	3 breeds from Egypt (33)	Global wild and domestic dataset	0.556-0.709	No substructure within indigenous breeds. Single maternal lineage with subsequent commercial admixture in a single region.
Osman et al. (2016)	5 breeds from Egypt (123)	Global dataset	0.62-0.81	86.4% of variation is within breeds but separation between native breeds and improved native breeds. Analysis of global dataset suggests 2 routes of expansion 1) terrestrial from India, through the Middle East and into Europe and North and Central Africa and 2) maritime from South East Asia or Indonesia via Madagascar into East Africa (reflecting distribution of Austronesian languages).
Ajibike et al. (2017)	4 populations from Nigeria (171)	Commerical layers	0.25-0.59	Single maternal lineage from Indian subcontinent. 97.3% of diversity was within populations. Suggestion of purifying selection based on prevalence of negatively selected sites.
Al-Jumaili et al. (2020)	25 populations from Algeria, Libya and Ethiopia (322)	Pakistan, Iraq, Saudi Arabia	0-0.929	Indian subcontinent haplogroup D was present in all countries. All except Algeria and Libya also had haplogroup E, indicating independent role for Indian Ocean trading network. Very little structure phylogeographic structure observed within countries.

Table 2.2: **Key mtDNA typing studies in Africa.** *N* is the sample size of the African populations (excluding any control or comparison populations). Key findings relevant to this project and the framework presented in in figures 2.1 and 2.2 are briefly summarised.

Study	Population of interest (N)	Control/comparison populations	Size of SNP panel	Observed heterozygosity in population of interest	Key relevant findings
Khanyile et al. (2015)	7 populations from Malawi, Zimbabwe and South Africa (312)	-	60,000	0.62	All populations showed significant inbreeding, although this was most severe in conservation flocks. Clustering of the village chickens followed a geographic gradient whereby South African chickens were closer to those from Zimbabwe than Malawi.
Fleming et al. (2016)	9 ecotypes from Uganda and Rwanda, plus Kuroilers (196)	-	600,000	-	High admixture, especially within countries attributed to shared ancestry, trade and selection parameters used by farmers to purchase new birds. Signatures of selection present in regions containing genes enriched for GO terms related to oxidative and metabolic stress response.
Malomane et al. (2019)	22 populations from 7 African countries (410)	Asian/South American/European local, Fancy breeds, Commercial, RJF	>580,000	approx. 0.23-0.32	Similarity between African chickens and commercial flocks attributed possibly to shared Asian ancestry. North and East African breeds (Egypt, Sudan, Ethiopia) clustered with Saudi Arabian and European breeds. Breeds from Uganda, Rwanda, Tanzania (partly), and Zimbabwe clustered together, suggesting gene flow between geographically proximal countries.

Table 2.3: **Key SNP typing studies in Africa.** *N* is the sample size of the African populations (excluding any control or comparison populations). Key findings relevant to this project and the framework presented in in figures 2.1 and 2.2 are briefly summarised.

Understanding the immunogenetics of these fancy or hobby breeds is important for the welfare and conservation of the breeds themselves and may have important implications for the control of the potential disease reservoirs they represent, which may become increasingly significant with the expansion of free range poultry production systems. Furthermore, disease resistance traits may exist in fancy breeds which could be beneficial if well-understood and bred back into commercial flocks. Indeed, several studies have investigated the resistance of Fayoumi birds to Newcastle disease (Deist et al., 2017; Schilling et al., 2019), Marek's disease (Lakshmanan et al., 1996), avian influenza (Wang et al., 2014), *Salmonella* (Cheeseman et al., 2007) and *Eimeria* (Pinard-van der Laan et al., 2009), relative to White Leghorn, a commercially important layer breed. Various immune and non-immune genes and QTLs were identified as potentially contributing to the resistance phenotype in Fayoumis but MHC class I and II have never been directly implicated, although sub-lines differing only at microchromosome 16 were observed to have differential responses to Newcastle disease virus which suggests a role for one of the many immune-linked genes in this region (Schilling et al., 2019).

The majority of studies of the genetics of fancy breeds focus on identification of conservation priorities, similarly to the African studies, but are able to relate patterns of diversity to specific breeding strategies and management practises more easily. Diversity is generally very low within breeds, particularly within a particular breeder's flock, but there is substantial diversity between breeds, with many genotypes unique to particular breeds. Genetic characteristics of specific populations are strongly influenced by the breeding and management strategies of the breeder, fancier or hobbyist. A summary of existing literature related to the genetic diversity present in fancy breeds is shown in table 2.4.

Study	Population of interest *(N)*	Comparison/control populations	Methodology	Key relevant findings
Osman et al. (2005)	8 Japanese breeds (213)	Rhode Island Red, White Leghorn	20 microsatellites	Expected heterozygosity of 0.33-0.566 in Japanese breeds. Phylogenetic distinction between fancy and utility breeds within a single prefecture.
Granevitze et al. (2007)	53 non-commercial chicken populations of various origins and breeding management histories (1000*)	Commercial, RJF	29 microsatellites	Management of standardised chicken breeds results in even lower heterozygosity and polymorphism than commercial flocks. Inbreeding coefficient and deviation from HWE was also high in the fancy breeds. Average observed and expected heterozygosity in the standardised (fancy) breeds were 0.41 and 0.48 respectively.
Oka et al. (2015)	Kurokashiwa (133)	-	29 microsatellites	Low allelic richness but the flocks are well managed to conserve diversity so there was no significant signature of inbreeding. Different breeders' flocks within prefectures showed limited differentiation, but populations in different prefectures were genetically distinct. Observed and expected heterozygosity ranges were 0.268-0.436 and 0.240-0.445 respectively.
Fulton et al. (2016b)	6 heritage broiler lines, 3 Barred Plymouth Rock, 2 New Hampshire and one each of Rhode Island Red, Light Sussex, White Leghorn, Dark Brown Leghorn, and 2 synthetic lines (1351)	-	101 SNPs in MHC-B region	1-11 haplotypes per line. 10/52 haplotypes were identical to serologically defined haplotypes. 9 recombinants with respective parental haplotypes.
Fulton et al. (2017)	Finnish Landrace (195)	-	90 SNPs in MHC-B region	20/36 haplotypes were novel. Some haplotypes were unique to specific populations of the breed.
Bortoluzzi et al. (2018)	37 traditional Dutch fancy breeds (480)	4 commercial white layer lines	60,000 genomic SNPs	Detected generation of novel diversity through neo-bantamisation (crossing large birds with true bantams or, later, other neo-bantams), but subsequent inbreeding for phenotype selection. Observed and expected heterozygosities were 0.116-0.327 and 0.108-0.335 respectively.
Guo et al. (2019)	White Plymouth Rock, Black Cochin, Buff Cochin, Partridge Cochin, Black Minorca, Black Java, Langshan, Light Brahma, and Dominique (150)	-	WGS	Genetics consistent with available historical records. Major contributors to White Plymouth Rock were Dominique males and Black Java/Cochin females.
Malomane et al. (2019)	100 fancy breeds (2000*)	Commercial, RJF	>580,000 genomic SNPs	Majority of Asian breeds clustered separately from European breeds, regardless of sampling location. Breeds from Eastern Europe and Finland have slightly higher heterozygosity and are slightly closer to Asian breeds. Bantams have lower diversity than equivalent standard populations. Breeds sampled in Germany had lower diversity than when sampled in the continent of origin.
Tarrant et al. (2020)	Silkie (101)	-	90 SNPs in MHC-B region	19/27 haplotypes were unique to this population (i.e. not reported in Fulton et al. (2016b)). 4/6 haplotypes at >5% frequency were unique. Up to 3 novel recombinant haplotypes detected
Berres et al. (2020)	Finnish Landrace (192)	-	101 genomic SNPs	Substantial differentiation between distinct populations of the breed, reflecting conservation programme structure. Observed heterozygosity was 0.1426-0.3260 in the various populations of the breed.

Table 2.4: **Key genetic diversity studies of fancy breeds.** *N* is the sample size of the fancy populations (excluding any control or comparison populations).

*In some studies a range of fancy and local breeds kept under different management systems were included and could not be specifically defined in order to give an exact sample size for the fancy breeds alone.

2.1.5 Specific aims

This work builds upon an existing chicken MHC typing programme, expanding the scope of the project to include non-commercial populations. Key aims include:

- Determine whether the low MHC diversity so far observed in commercial populations is a feature of chicken flocks worldwide, or whether there is a larger pool of genetic diversity in non-commercial populations
- Investigate genetic differentiation between populations on different geographic scales, and suggest implications for management strategies and interventions
- Integrate observed diversity in different populations with published phenotypic data and reports from breeders to investigate possible associations between MHC haplotypes and pathogen resistance or susceptibility

2.2 Materials and Methods

2.2.1 Project structure

The data presented here form part of a larger typing effort which includes populations from many commercial breeders among others, the results of which are discussed briefly in section 2.1. Illumina MiSeq runs were numbered sequentially as they were performed, with the data presented here primarily generated during runs 8-11, 13 and 14. Some results from runs 15 and 16 are also described but not all populations included on these runs have so far been analysed.

2.2.2 Samples

The samples that were used in the work described here are detailed in table 2.5. All samples were received as purified DNA, although samples from other populations studied in the wider project were often received as blood. Run 9 was constructed by Kerstin Wilhelm (visiting from University of Ulm) and Felicity Coulter (University of Cambridge, now Oregon Health and Science University). Runs 13 and 14 were constructed with the assistance of Catherine Toase (University of Bristol Veterinary School).

Population name	Source	Number of samples	Details	Run number	Library preparation	Data analysis
Vinkler fancy breeds	Dr. Michal Vinkler, Zuzana Świderská, Adéla Šmídová, Anežka Fabiánová (Charles University, Prague, Czech Republic)	576	45 fancy breeds from European breeders	8	Rebecca Martin	Rebecca Martin
Roslin Ethiopia 1-3	Dr. Jacqueline Smith (Roslin Institute, Edinburgh, UK), Dr. Olivier Hanotte (University of Nottingham, UK and ILRI, Addis Ababa, Ethiopia)	288	Free-range local 'village' chickens from Ethiopia	9	Kerstin Wilhelm, Felicity Coulter	Rebecca Martin
Weigend fancy breeds	Dr. Steffen Weigend (Friedrich-Loeffler-Institut, Greifswald, Germany)	1293	69 breeds, samples part of the SYNBREED chicken diversity panel (Malomane et al., 2019)	10 and 11	Rebecca Martin	Rebecca Martin
Roslin E&N	Dr. Jacqueline Smith (Roslin Institute, Edinburgh, UK)	192	Free-range local 'village' chickens from Ethiopia and Nigeria	11	Rebecca Martin	Rebecca Martin
UC Davis African	Dr. Huaijun Zhou (University of California, Davis, USA), Dr. Susan Lamont (Iowa State University, USA)	1680	Free-range local 'village' chickens from Tanzania (3 ecotypes) and Ghana	13 and 14	Catherine Toase, Rebecca Martin	Rebecca Martin
Roslin Oman	Dr. Jacqueline Smith (Roslin Institute, Edinburgh, UK)	27	Free-range local 'village' chickens from Oman	14	Catherine Toase, Rebecca Martin	Rebecca Martin
Roslin Various African	Dr. Jacqueline Smith (Roslin Institute, Edinburgh, UK)	124	Free-range local 'village' chickens from Africa (location not specified)	14	Catherine Toase, Rebecca Martin	Rebecca Martin
Commercial A1 and A2	Breeding company A	146	Commercial broilers	15	Eve Doran, Fabien Filaire	Dr. Clive Tregaskes
Commercial B1 and B2	Breeding company B	150	Commercial layers (B1) and dual-purpose birds (B2)	16	Undine-Sophie Deumer	Matt Jayasekara, Rebecca Martin
Commercial C	Breeding company C	215	Commercial broilers	14	Catherine Toase, Rebecca Martin	Rebecca Martin
Commercial D	Breeding company D	196	Commercial layers	14	Catherine Toase, Rebecca Martin	Rebecca Martin

Table 2.5: **Summary of key sample sets, relevant Illumina MiSeq runs and contributors to data presented here.**

2.2.3 PCR-NGS library construction

The procedure for preparing libraries to sequence class I (BF) and class II (BLB) genes from chickens was developed predominantly by Dr. Clive Tregaskes (Kaufman lab, University of Cambridge).

Each library could, in theory, contain BF and BLB amplicons from up to 1152 birds across twelve 96-well plates, although the number of birds from which amplicons were successfully obtained for each run was lower than this. The 'double-barcoded' multiplexing used to distinguish sequences that originated from different samples was suggested by Ed Farnell (Cambridge Genomic Services, now Illumina).

The processing of samples from receipt to sequencing is summarised in figure 2.3.

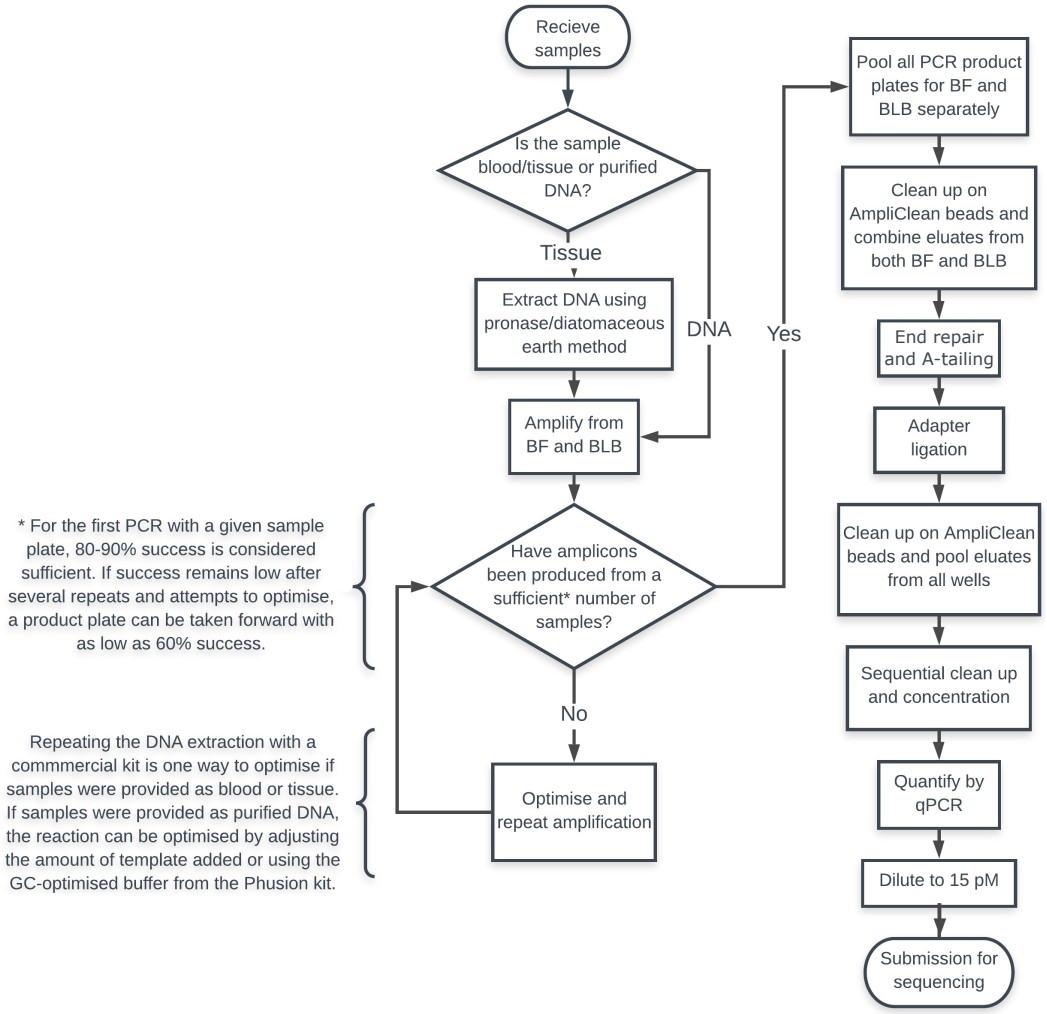


Figure 2.3: **Summary of library preparation protocol.**

DNA extraction

The DNA extraction module for processing samples that are received as blood is based on the pronase digestion method of Bailes et al. (2007) and the diatomaceous earth clean-up protocol of Carter and Milton (1993). In the runs described here, only the commercial samples analysed on run 9 were received as blood and required DNA extraction.

Amplification of variable peptide-binding domains from BF and BLB genes

PCRs were performed to amplify exons 2 and 3 from both BF and both BLB loci. While only exon 2 is a peptide binding domain for the class IIβ chain, including exon 3 in the amplicon means that the amplicon lengths for BF and BLB are roughly equivalent, which reduces the risk of substantial differences in amplification or sequencing efficiency. The locations of the primer binding sites are indicated in figure 2.4.

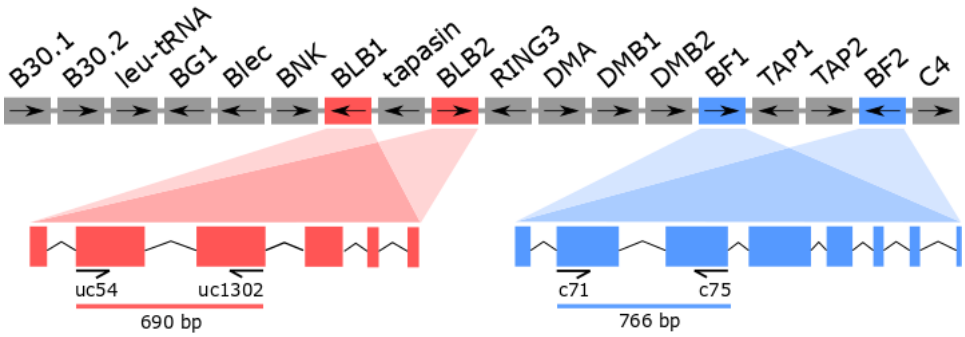


Figure 2.4: **Primers used to amplify exons 2 and 3 from the BF and BLB genes.** The relative positions of the genes of interest are shown in the context of the other genes in the chicken MHC. Amplicon lengths are indicated for each class.

Twelve pairs of primers were synthesised, each with a unique 5' 10 nt 'PCR barcode', which differed from all other barcodes at a minimum of two positions. Each 96-well plate of samples was assigned a barcode such that all amplicons from that plate had the same PCR barcode.

PCRs were performed in 10 µl total volume using the Phusion Hot Start Flex DNA polymerase kit (NEB) with final concentrations of reagents as follows: 1x reaction buffer (HF or GC), 200 µM each dNTP, 0.5 µM each primer and 0.2 units/10 µl reaction Phusion polymerase. Thermocycling was performed as follows for BF: 98 °C (30 s), 35 cycles of 98 °C (10 s), 68.4 °C (20 s) and 72 °C (30 s) and a final extension step of 72 °C (5 min) and BLB: 98 °C (30 s), 40 cycles of 98 °C (10 s), 64 °C (20 s) and 72 °C (30 s) and a final extension step of 72 °C (5 min). Both of these protocols required a ramp speed between temperatures of 0.8 °C s^{-1}. All thermocycling protocols are also tabulated in appendix A.2.

Several types of errors were seen to occur during PCR amplification of the genes of interest. It was not unusual for PCR chimeras to form during amplification due to the polymerase 'falling off' the template strand while extending through the GC-rich intron. The partial strand then separated from its template during the next denaturation step and acted as a long primer, annealing non-specifically to another amplicon and extending to form a product that had exons from two different alleles. These are referred to as inter-exon chimeras. Less frequently, extension would terminate during amplification of an exon, resulting in a partial exon sequence acting as a non-specific primer and leading to an intra-exon chimera being formed. Finally, as with any PCR, single nucleotide misincorporations were not uncommon, despite the proof-reading capacity of the Phusion polymerase, which is conferred by its 3'-5' exonuclease activity.

In early runs, PCR success was assessed by gel electrophoresis of 5 µl of each PCR product on an agarose gel (1% agarose w/v in TAE buffer (1 mM EDTA, 40 mM Tris, 20 mM acetic acid))

with 1:10000 SYBRsafe (Invitrogen), and visualisation under UV light using the GBOX Chemi XX6 (Syngene). In later runs, PCR products were separated and visualised using the QIAxcel Advanced System (Qiagen) for capillary electrophoresis.

Plates where PCR success was low were repeated, including re-extracting DNA from blood if applicable, such that some samples ended up being amplified with two different barcode combinations. Samples that were extracted using the diatomaceous earth protocol but consistently demonstrated poor amplification efficiency were re-extracted using the DNeasy 96 Blood and Tissue kit (Qiagen).

Library preparation

Once BF and BLB had been amplified from 12 plates of samples the plates of PCR products were pooled such that the amplicons in each well were combined with the amplicons in corresponding wells in the other 11 plates. Since the amplicons were barcoded by plate, pooling in this way did not combine any amplicons that could not later be distinguished. Short fragments such as primer dimers were removed from the pooled PCR products using AmpliClean magnetic beads (NimaGen) in a 1:1 ratio of sample:bead suspension. End repair and A-tailing of amplicons was performed using the KAPA Hyper Prep kit (Roche) according to the manufacturer's instructions. Finally, the same kit was used to ligate NextFlex-96 barcoded Illumina adapters (Bioo Scientific) to the ends of the amplicons. These adapters allowed binding of the amplicons to the sequencing chip and the NextFlex-96 module contained 96 uniquely barcoded variants of the adapter such that each well of pooled PCR products had a unique 'adapter barcode'. The result was that each sequence obtained could be traced back to the plate (PCR barcode) and well (adapter barcode) of the sample from which it was originally amplified.

AmpliClean beads in a 1:1 sample:bead ratio were used to remove unligated adapters and the ligated amplicons were then pooled into a single tube. Six 100 µl aliquots of the pooled library were taken and cleaned up sequentially using 1:0.6 sample:AmpliClean beads, with each aliquot eluted in the eluate from the previous tube.

The amount of adapted DNA in the library was quantified by quantitative PCR (qPCR) (Luna Universal qPCR Master Mix, NEB; Applied Biosystems 7500 platform) using primers specific to the Illumina adapter sequences and standards made from BLB amplicons diluted to specific concentrations. qPCR products were analysed by agarose gel electrophoresis (as described previously) to verify that the size of the fragments in the library correspond to the expected amplicon size.

Illumina MiSeq sequencing

Libraries were diluted to 15 pM and submitted to Cambridge Genomic Services for 600 bp paired-end Illumina MiSeq sequencing, where they were also spiked with 5% phiX174 viral DNA. Both the BF and BLB amplicons were longer than 600 bp, however 300 bp from each end of the amplicon covered both the exons, leaving just a portion of the intron unsequenced.

2.2.4 Data processing

Allele calling pipeline

A pipeline developed by Dr. Clive Tregaskes (Kaufman lab, University of Cambridge) converts the raw sequencing data into human-readable .csv files containing lists of the alleles called for each bird. An SQL database, also constructed by Dr. Tregaskes but populated by researchers as they prepare libraries, links the barcode sequences detected on the amplicons to the details of the particular sample from which that sequence was amplified.

The pipeline takes the 192 files generated by the MiSeq (read 1 and read 2 for each of the 96 Illumina barcodes), trims off poor quality sequence and then reads the PCR barcode with the three immediately following nucleotides from the primer itself. This is sufficient to determine which bird the sequence is from, whether the sequence was amplified from BF or BLB, and whether the exon 2 or exon 3 sequence is in read 1. Based on this information, sequences are sorted into four files for each bird, corresponding to 'forward' (exon 2 in read 1 and exon 3 in read 2) and 'reverse' (exon 3 in read 1 and exon 2 in read 2) sequences for each of BF and BLB. Using the positional information in the description of each sequence, reads 1 and 2 for each cluster are combined into a single sequence, with the two reads separated by a tab.

If a set of sequences (defined as all the sequences for a given bird that have a particular PCR primer in their read 1) has fewer than 40 total sequences, no further processing is performed. For those sets with 40 or more sequences, 'allele signatures' are defined as any 150 bp sequence that occurs at the start of at least 15% of the sequences in that set. In theory, each set of sequences should have four allele signatures (in a heterozygote) or two (in a homozygote). Once the signatures for the set are defined, sequences that match the allele signature are pooled and an 80% consensus is drawn for each. If the set contains more than 150 sequences, the match to the signature has to be 100%. If the set contains 40-150 sequences, a sequence has to match the allele signature at 148/150 positions to be included in the consensus unless two of the signatures in the set differ at fewer than two positions, in which case a perfect match is required.

For each bird there is now a set of allele sequences in both the forward and reverse directions

for each of BF and BLB. Since the read 2 sequences are always lower quality than the read 1 sequences, the pipeline then compares each of the poor quality exon 3s in the forward sequences to each of the good quality exon 3s in the reverse sequences for the same bird. The best match is identified and the good quality exon 3 from the reverse sequence is substituted for the poor quality exon 3 in the forward sequence to give a final sequence that has both exons covered by high quality sequencing reads. The same process is repeated comparing the poor quality exon 2s in the reverse sequences to the good quality exon 2s in the forward sequences. In theory, both should yield the same set of complete allele sequences, and duplicates are then discarded.

Finally, the complete allele sequences are compared to a reference list which contains sequences known from the literature and from previous runs in this project (see supplementary file). If a sequence matches a known allele completely, that allele is listed as a 'called allele' in that bird. If a sequence is not found in the reference list (i.e. it is 'unknown') it is assigned the name 'UNK_#', with the UNKs numbered in the order they are detected. Each UNK, as it is discovered, is added to the reference list for that run, such that subsequent occurrences of the same unknown sequence are labelled with the same number. It is important to note that unknown sequences are not used in the reference list for future runs until they have been confirmed to be real and assigned an allele name. Instead, unconfirmed UNKs are kept in a separate list and compared to UNKs on future runs, since some rare alleles never occur more than once per run but are present in two independent PCRs on different runs.

The key outputs from the pipeline are i) a file for each of BF and BLB, containing identifiers for each bird that had at least one allele called in that class, and the names of all the alleles that were identified for each of those birds and ii) for BF, a so-called 'chimera test' file, containing the number of times each UNK allele was called and whether either of the individual exons in each UNK matched any exon already known.

Analysis of unknown sequences

The UNK sequences were manually curated. Since allele sequences must be amplified in two independent PCRs to be considered real, in the first stage only UNKs that were called more than once in the run being analysed were considered. In the second stage, all the UNKs seen once in the current run were compared to all UNKs from previous runs, such that a single call in run x and a single call in run y could provide the necessary evidence that a sequence represented a real allele.

For each UNK amplified more than once, the results for all birds carrying that UNK were extracted. Several factors were considered in order to determine whether an UNK represented a real

allele or an artefact, of which three main types were commonly identified: inter-exon chimeras, intra-exon chimeras and single nucleotide misincorporations.

- **How many alleles do the birds have?** If the majority have a odd number of alleles called for either BF or BLB, it is likely that there is a consistent artefact occurring.
- **If the UNK contains two known exons, are the alleles in which those exons normally found both present in the birds that carry the UNK?** For example, if an UNK contains the same exon 2 sequence as BF1*009:01_AB426145 and the same exon 3 sequence as BF2*009:01_AB426145, and all the birds that carry that UNK are seen to have the 9-9-9-9 haplotype, it is likely that the UNK is an inter-exon chimera.
- **Is the UNK very closely related to another allele present in all the birds?** If an UNK is a single nucleotide different from another allele that occurs in all the birds extracted, it is likely to be a single nucleotide misincorporation. Sometimes, if the misincorporation cause the artefactual sequence to be preferentially amplified, the true sequence may not be present at the require threshold and may 'drop out'. In this case, other lines of evidence are used to determine whether the UNK is a real new variant of the known sequence, or an artefact created from a dropped-out allele.
- **Is there a plausible haplotype pair?** If many of the birds extracted contain, for example, a particular BF1 allele ('A') but not the BF2 allele from the same haplotype with which A would normally co-occur (its 'haplotype pair') it is possible that the UNK is the real haplotype pair for allele A in this set of birds and a new haplotype has been discovered. This evidence is strengthened if there is a consistently co-occurring unpaired BF1 allele, and the UNK clusters with BF2, for example. If the UNK is related to the usual haplotype pair for the unpaired allele, for example an UNK that clusters with the BF1*004 sequences and co-occurs with an unpaired BF2*021:01_AM282697 makes sense as a haplotype and is likely to be real, as long as it cannot be explained by chimerism or misincorporation from a known BF1*004 allele which may have dropped out.
- **Can the UNK be explained by intra-exon chimerism?** If there is no evidence that a given UNK is real, but it cannot be explained by inter-exon chimerism or misincorporation, all the alleles present in the birds extracted are aligned and assessed for intra-exon chimerism. In rare cases, more than one incidence of chimerism was observed in a single UNK, with the sequence 'jumping' from one known allele to another, and back again. Single incidences of intra-exon chimerism are indicated by one exon in the UNK matching an allele that consistently co-occurs with the UNK. If multiple mispriming events have occurred, both exons will be new, and the artefact can only be identified by studying the alignment.

Since only exon 2 sequences were produced as output from the pipeline for BLB alleles, some of these questions did not apply and the assessment was therefore simpler.

Based on consideration of these factors, a decision was made about whether the sequence represented a real allele or an artefact. If the sequence was thought to be real, it was assigned a name and added to the reference list for subsequent runs. Inevitably, some UNKs that were identified as real in later runs may have been present a single time in previous runs and discounted. As such, it was necessary to repeat the allele calling for all runs with the reference list that was compiled at the end of the analysis of run 14, before performing any comparative analysis between runs.

Haplotype calling

The second stage of the analysis begins by combining the BF and BLB results into a single file. It is vital that every bird has a unique identifier (the 'unique_ID'), since this is used to identify results that come from the same individual, and group them accordingly. Repeated samples are manually examined and the data reduced to a single row by removing whichever replicate has fewer allele calls. If an individual has two sets of calls which do not match, the bird is removed from the dataset, since this suggests that a mis-assignment has occurred.

Subsequent scripts take a reference list of known haplotypes (appendix A.5 and supplementary file), which is compiled manually by examining co-occurring combinations of alleles. Haplotypes that are known to have genes missing, such as the standard B14 haplotype which lacks a BF1 allele, or known to be affected by allele dropout, can be included alongside 'complete' haplotypes.

Step 1 assigns haplotypes where all component alleles have been successfully called within an individual sample. The abbreviation of the identified haplotype is appended to the end of the row in the results table in the first available 'haplotype' field.

Step 2 identifies homozygotes where a single haplotype was assigned in step 1 and there are no alleles called for the bird that are *not* part of the identified haplotype. If these conditions are met, the haplotype abbreviation is repeated in the next available haplotype field at the end of the row, with '(homo)' added. In this way, haplotype frequencies can later be calculated more accurately.

Step 3 takes the 'chimera test' file, uses it to convert all remaining UNK sequences into a string containing the names of alleles that match the exon 2 *or* exon 3 sequence of the UNK. This allows alleles that are not fully called but are partially present in inter-exon chimeras to be used in haplotype identification. A process similar to step 1 is then repeated, with any additional haplotypes found added to the next available haplotype field in that row with '(chimera)' appended. The

suffixes added to haplotype abbreviations in steps 2 and 3 allow these haplotype calls to be included or excluded from analysis, depending on the stringency requirements of the question. Step 3 currently only includes the first match for each of exon 2 and 3 when the reference list is searched in order, and may therefore omit or call additional haplotypes in a small subset of cases where multiple alleles share exon sequences.

Any birds where three or more haplotypes were called (54 out of 3598 total birds with both BF and BLB allele calls) were manually examined and corrected where erroneous haplotype calls were obvious. This also occurred in cases where alleles could be arranged into multiple haplotype combinations. For example, a bird carrying the 32-32-4:01_03-32 and 5:01_02-23-23:01and04-36 haplotypes would always have 32-32-23:01and04-36 called as a third haplotype. Haplotypes identified as errors had the suffix '(error)' manually added to the haplotype call in the full results table, and were excluded from analyses.

This method is able to successfully identify haplotypes in an automated way, within the constraints of sequencing quality, haplotype list completeness and a small number of idiosyncrasies of the process, such as BF1*004:01_03 and BF1*004:04 never being called together in the same bird because they share their first 150 nt and thus get assigned to the same allele signature. Populations where these effects were observed to be significantly affecting the haplotype assignment were assessed on a case-by-case basis and any additional haplotype assignments used in analyses were described when reporting results.

2.2.5 Software and packages

Analysis and visualisation of data was predominantly performed in R (R core team, 2013), using packages including ggplot2 (Wickham, 2016) and ggtree (Yu et al., 2017). Geneious Prime versions 11 to 2021.1 and Clustal Omega (Sievers et al., 2011) were used to generate alignments. Structural models were visualised and annotated in PyMol 2.3.4 (Schrodinger LLC, 2019). Analysis of positively selected sites was performed using CODEML from the PAML package (Yang et al., 2005; Yang, 2007). GenAlEx (Peakall and Smouse, 2006, 2012) was used for analysis of genetic differentiation between populations Polymorphism analysis was performed in DnaSP6 (Rozas et al., 2017). MEGA X (Kumar et al., 2018) was used to test multiple tree-drawing algorithms.

2.3 Results

Tables of allele and haplotype calls for all populations are provided in the supplementary file.

2.3.1 60 novel BF1 and 182 novel BF2 alleles were discovered

The number of alleles on the reference list increased from 23 BF1 and 46 BF2 after run 7 to 83 BF1 and 228 BF2 after run 14 (figure 2.5). Runs 15 and 16 are not included because the analysis of all populations on them has not been finalised, but no additional alleles appear to be present in these runs, which exclusively contain lines from commercial breeders. A large number of new alleles were discovered during run 8 (14 BF1/35 BF2), which was the first set of fancy breeds to be sequenced, and run 13 (19 BF1/76 BF2) which was the largest set of African samples.

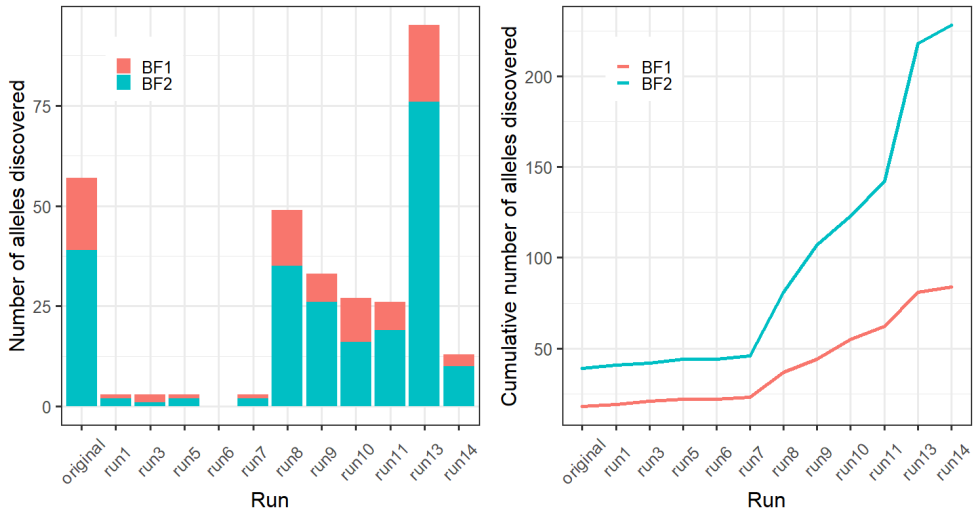


Figure 2.5: **Significant numbers of new alleles were discovered in runs containing non-commercial populations.** 'Original' refers to the original reference lists compiled from the literature as described in section 2.1.3. Four BF2 alleles on the original reference list have still not been verified.

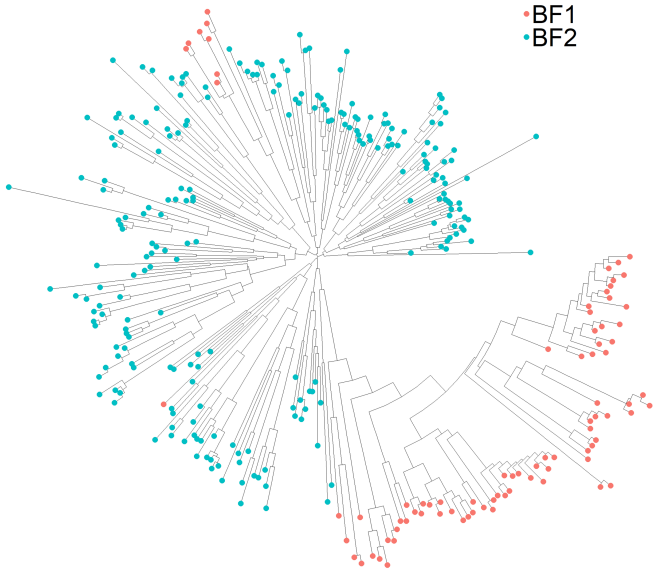


Figure 2.6: **BF alleles segregate almost completely by locus** Neighbour-joining phylogenetic tree of all BF alleles known after run14. A full version of this tree with branches labelled with full allele names is available in appendix A.4.1. Tree generated using Clustal Omega and visualised in R using the ggtree package.

BF1*071:01_run13_unk41 is the first clear example of gene conversion in chicken MHC class I

The BF sequences were all assigned to a locus, since there was sufficient evidence from similarity to known sequences and likely haplotype pairings to do so, given the simplicity of the phylogeny (figure 2.6). There are two major BF clades which largely separate BF1 and BF2, although the alleles of neither gene are truly monophyletic. Alleles assigned to BF1 are polyphyletic, with one major monophyletic clade and two smaller groups. The first of these two groups contains the BF1*002 and BF1*009 allele groups, including the sequences BF1*002:01_AM279336 (Shaw et al., 2007) and BF1*009:01_AB426145 (Hosomichi et al., 2008) from the standard haplotypes and several closely related sequences from the PCR-NGS project.

The second of the two BF1 groups within the paraphyletic BF2 clade contains a single sequence, BF1*071:01_run13_unk41. This sequence was present in two birds in run 13, both from Tanzania but in two different ecotypes, Kuchi (bird 2693) and Ching (bird 7206).

BF1*071:01 was assigned to BF1 because it pairs with BF2*115:01 (haplotype pair 71-115), which itself was assigned to BF2 both from its position in the phylogeny and its pairing with BF1*017:01 (haplotype pair 17-115). Table 2.6 lists all the birds which carried BF2*115:01. In all cases the alleles seen in addition to the haplotype pairs 71-115 or 17-115 make another known haplotype pair (23-36, 32-41, ?-66:02 and 32:02-52:03), so pairings here are described with a high level of confidence. run14_BF_UNK_56 is an inter-exon chimera of BF1*032:02 and BF2*052:03, and BF2*066:02 appeared without a plausible BF1 pair in the vast majority of the 81 birds it appeared in, suggesting either that this is a new BF1*NULL haplotype, or that the BF1 allele amplifies particularly poorly resulting in 'allele dropout'.

Bird ID	Plate	Clusters	allele1	allele2	allele3	allele4	allele5
2693	plate_95973	1730	BF1*023:01and04_AB426153	BF2*036:01_AY327147	BF2*115:01_run13_unk40	BF1*71:01_run13_unk41	
7206	plate_107642	851	BF1*032:01_Av3b	BF2*041:01_Av3a	BF2*115:01_run13_unk40	BF1*71:01_run13_unk41	
7946	plate_107642	861	BF1*017:01_AB426150	BF2*066:02_run13_unk25	BF2*115:01_run13_unk40		
8912	plate_107643	229	BF1*017:01_AB426150	BF2*066:02_run13_unk25	BF2*115:01_run13_unk40		
8812	plate_108916	286	BF1*017:01_AB426150	BF1*032:02_run9_unk84	BF2*052:03_run8_unk23	BF2*115:01_run13_unk40	
8515	plate_107381	1781	BF1*017:01_AB426150	BF1*032:02_run9_unk84	BF2*052:03_run8_unk23	BF2*115:01_run13_unk40	run14_BF_UNK_56
8519	plate_107381	2156	BF1*017:01_AB426150	BF1*032:02_run9_unk84	BF2*052:03_run8_unk23	BF2*115:01_run13_unk40	

Table 2.6: **Allele calls from all birds containing BF2*115:01_run13_unk40.**

Since BF1*071:01 was known to have the same exon 3 sequence as BF1*017:01, it was hypothesised that these birds might carry the more common haplotype 17-115, with BF1*071:01 simply a chimera which caused BF1*017:01 to drop out. Figure 2.7 shows the alignment of the three sequences. Based on the alignment, BF1*071:01 could plausibly be the product of two intra-exon chimerism events, although these are rare. BF1*071:01 matches BF1*017:01 over its first 150 nt, BF2*115:01 until positions 226/227 and then BF1*017:01 again over the remainder of the 494 nt.

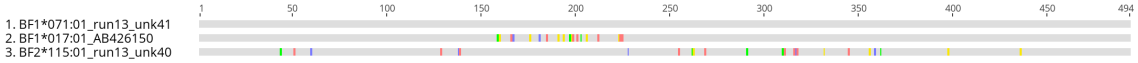


Figure 2.7: **Nucleotide alignment of BF1*071:01_run13_unk41 and its predicted 'parent' alleles, BF1*017:01_AB426150 and BF2*115:01_run13_unk40.** Coloured ticks indicate differences from BF1*071:01_run13_unk41 (red=A, blue=C, yellow=G, green=T). The sequence length corresponds to the exon2-exon 3 region amplified by the c71/c75 primer pair with the intron removed. Exon 3 begins at position 245 in this alignment.

However, even if BF1*017:01 had been present, it seems unlikely that the two intra-exon chimerism events would have occurred together in two independent birds. The windows in which the events would have needed to occur are just 19 and 2 nt long respectively, and no evidence of chimerism 'hotspots' within exons has been observed in other data.

Finally, if BF1*071:01 was truly a chimera, the parent sequence BF1*017:01 should be present in the raw sequence data, even at a very low level that would leave it below the threshold for an allele signature. This was not the case; neither bird had any reads which matched BF1*017:01.

Overall, the evidence strongly suggests that BF1*071:01_run13_unk41 is the first clear example of gene conversion between the BF1 and BF2 loci, derived from the known haplotype BF1*017:01_AB42615 - BF2*115:01_run13_unk40. The fact that the majority of the sequence is identical to a BF1 allele supports its assignment to that locus; in particular, the region containing the putative NK cell interaction site (Ewald and Livant, 2004; Kim et al., 2018) is derived from the BF1 allele. Nonetheless, the incorporation of the segment from BF2*115:01 causes it to cluster most closely with BF2 sequences in a phylogeny, regardless of the tree-drawing algorithm used (Neighbour-Joining, Maximum Likelihood, Maximum Parsimony, Minimum Evolution, UPGMA (unweighted pair group method with arithmetic mean)). The most closely related sequence overall, BF2*044:01_run5_unk22, differs from BF1*071:01_run13_unk41 at 21 nucleotide positions, corresponding to 17 amino acid differences across exons 2 and 3.

2.3.2 Analysis of positively selected sites within the expanded allele database supports divergent functions for BF1 and BF2

Random subsets of 50 sequences from each of BF1 and BF2 (using the reference list after run 14) were created and analysed in CODEML (Yang, 2007). A likelihood ratio test (LRT) was performed according to Yang et al. (2000), comparing models 7 (β) and 8 ($\beta+\omega > 1$). For both species, twice the log likelihood difference was compared to the χ^2 distribution with two degrees of freedom, since the difference in the number of parameters in these models is two. Both the BF1 and BF2 subsets contained positively selected sites with model 8 providing a significantly better fit to the data than model 7 ($p < 0.00001$ for both BF1 and BF2).

BF2 was observed to have more sites under positive selection than BF1, with 23 (BF2) compared to 15 (BF1) sites where $p(\omega > 1) > 0.95$ within the $\alpha 1$ and $\alpha 2$ domains (figure 2.8). Moreover, the locations of positively selected sites (PSS) unique to each gene were noticeably different, with PSS unique to BF2 generally having side chains protruding into the peptide binding groove, while those unique to BF1 had side chains pointing outwards from the peptide binding groove (figure 2.9).

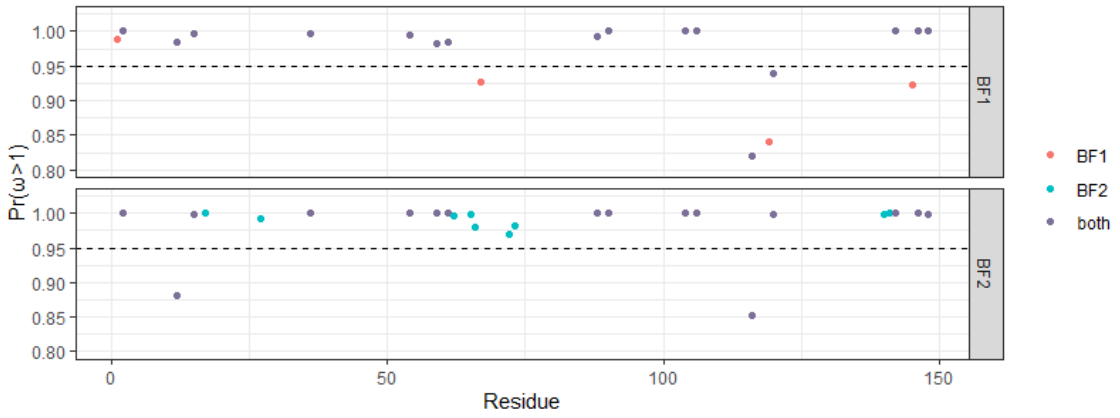


Figure 2.8: **BF2 alleles contain more positively selected sites than BF1 alleles.** PSS unique to one gene are indicated in red (BF1) or blue (BF2), while residues that are predicted to be under positive selection in both genes are indicated in purple. Residues correspond to the 164 complete codons in the exon2-exon3 amplicon.

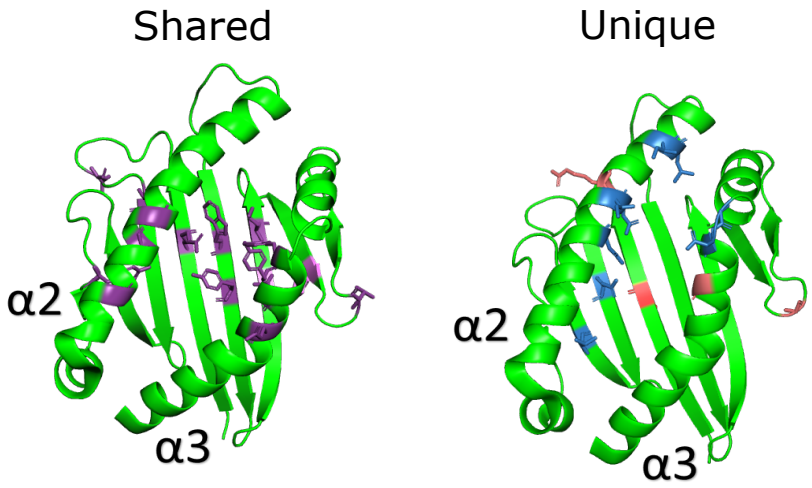


Figure 2.9: **Positively selected sites unique to BF2 have the potential to interact with antigenic peptides.** Locations of predicted positively selected sites in BF genes are shown on a structural model of BF2*002:01 (Chappell et al., 2015). Amino acids that are positively selected in both genes are shown in purple, with those only positively selected in BF1 or BF2 shown in red and blue respectively. The green cartoon shows the region of the protein corresponding to the BF amplicon. The $\alpha 2$ and $\alpha 3$ helices are indicated for orientation.

2.3.3 127 novel BLB alleles were discovered

The number of alleles on the reference list increased from 24 BLB1, 34 BLB2 and 57 BLB* after run 7 to 51 BLB1, 75 BLB2 and 116 BLB* after run 14 (figure 2.10). The sequences did not cluster phylogenetically by locus (figure 2.11).

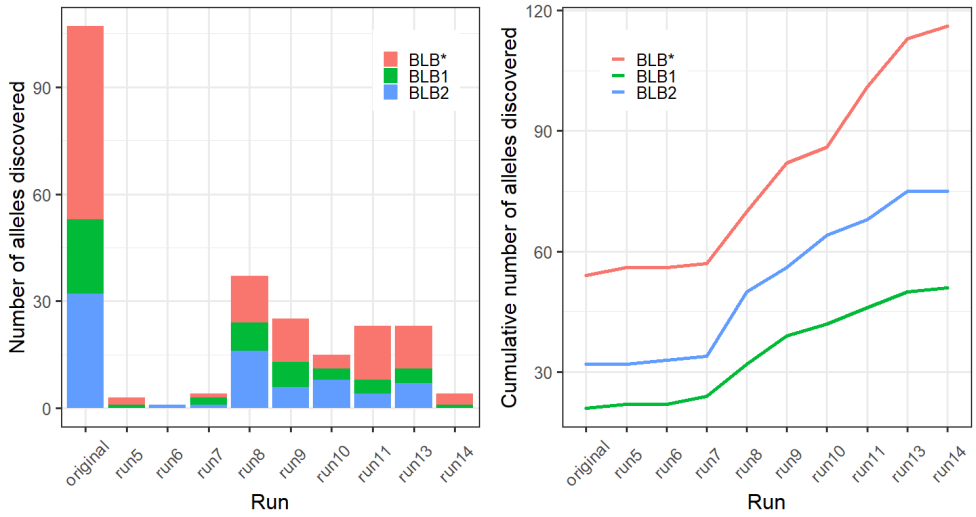


Figure 2.10: **Significant numbers of novel BLB alleles were discovered in runs containing non-commercial populations.** Four BLB1, 13 BLB2 and 49 BLB alleles on the original reference list from the literature have still not been fully verified by PCR-NGS.

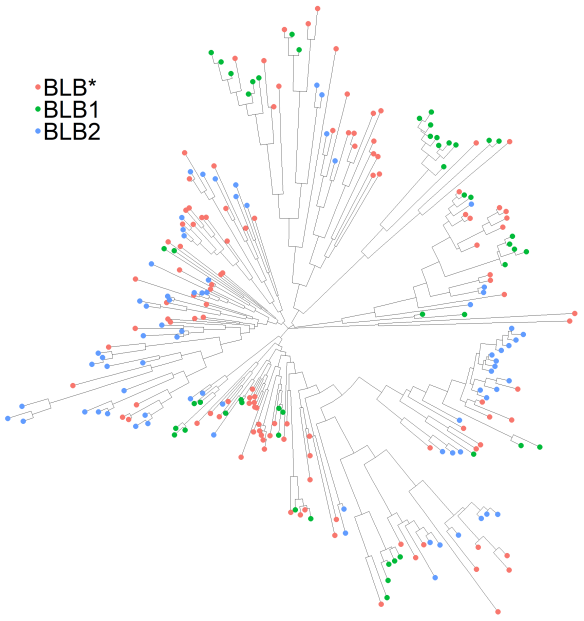


Figure 2.11: **There is no phylogenetic distinction between BLB1 and BLB2 alleles.** Neighbour-joining phylogenetic tree of all BLB alleles known after run14. A full version of this tree with branches labelled with full allele names is available in appendix A.4.2. Tree generated using Clustal Omega and visualised in R using the ggtree package.

2.3.4 BLB1 and BLB2 show highly similar patterns of positive selection

A random subset of 47 sequences was selected from the alleles assigned to each of BLB1, BLB2 and BLB* (since only 47 BLB1 sequences were known) and analysed in CODEML. The LRT showed that there were residues likely to be under positive selection in all three subsets. The majority of the PSS identified were shared between the three subsets. The BLB* subset had one PSS site that was seen in the BLB2 subset but not the BLB1 subset, probably because BLB2 is slightly more diverse (based on there being a greater number of sequences assigned to BLB2 than BLB1 in the current reference list) and therefore is likely to make up a higher proportion of the unassigned sequences (figure 2.12).

Each of BLB1 and BLB2 had three unique PSS, although only one and two sites for BLB1 and BLB2 respectively were predicted to be under positive selection with over 95% confidence. There was no noticeable difference between the orientations of the side chains relative to the peptide binding groove (figure 2.13).

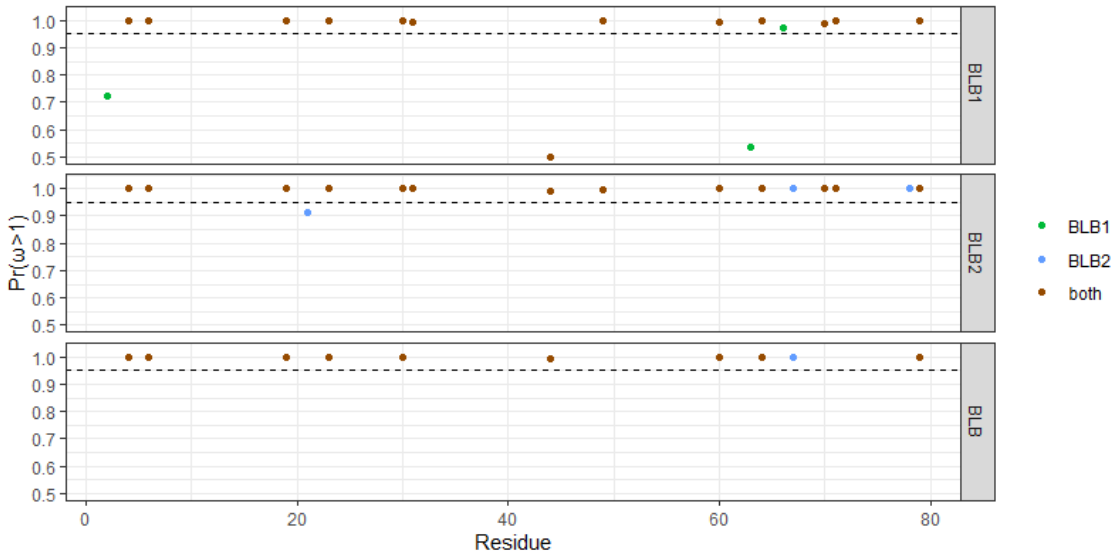


Figure 2.12: **BLB1, BLB2 and BLB* have similar numbers of predicted positively selected sites.** PSS unique to one locus assignment are indicated in green (BLB1) or blue (BLB2), while those common to both are indicated in brown. The residues correspond to the 87 complete codons in the exon 2 region of the amplicon, which is the variable peptide-binding domain. The dashed line indicates 95% confidence.

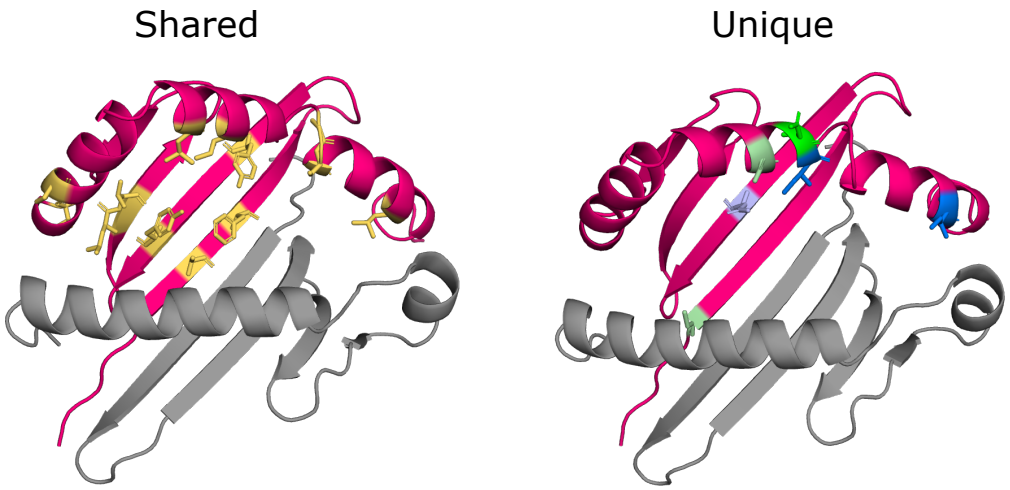


Figure 2.13: **Structural locations of predicted positively selected sites in BLB genes.** Amino acids that are positively selected in both genes are shown in yellow, with those only positively selected in BLB1 or BLB2 shown in green and blue respectively. Residues that are identified as being under positive selection with less than 95% confidence are in pale shades. The pink cartoon shows the region of the protein corresponding to the BLB amplicon. The grey region is the α1 domain from BLA which was not amplified in this study. The model used was BLB2*002:01 (S. Halabi and J. Kaufman, unpublished).

2.3.5 172 novel haplotypes were discovered

While a plausible haplotype was not required for a novel allele to be considered real, many new alleles could be seen to co-occur consistently with other alleles which did not otherwise form complete haplotype combinations. Furthermore, the presence of a plausible haplotype for a new allele could often be used as evidence that the unknown sequence was not simply an artefact. To be included on the reference haplotype list, an allele combination had to be seen in at least two birds where the presence of the allele combination could not otherwise be explained.

From the data obtained in runs 8-11, 13 and 14, 172 new haplotypes were identified. These were combined with 60 haplotypes known from the literature and runs 1-7 (that is, almost exclusively experimental and western commercial chicken lines) to create the current working haplotype list which includes 232 unique haplotypes (appendix A.5). 202 of these haplotypes have four component alleles (BF1, BF2 and two BLB) while 26 appear to have no BF1 allele, two have a single identified BLB allele and two have a single BLB and no identified BF1. Haplotypes with fewer than four component alleles identified will require further analysis to determine whether loci are genuinely missing (as seems to be the case in the standard B14 haplotype [Shaw et al. (2007) and E. Palmer, F. Coulter and Prof. J Kaufman, unpublished]), alleles at multiple loci are identical in the examined exons or certain alleles do not amplify efficiently with the current primers. Despite these possible inaccuracies in the specific descriptions of certain haplotypes, the combinations of co-occurring alleles described on the current reference haplotype list remain useful and biologically

relevant units of MHC diversity for the analysis of data in this project.

Haplotypes with 'missing' BLB alleles may provide important insights into class II evolution

While degradation or absence of the BF1 gene in some standard MHC haplotypes has been previously reported (Shaw et al., 2007), the standard chicken MHC haplotypes all have two reported BLB alleles. Work by Worley et al. (2008) suggested that a single BLB allele sequence could occur in both the BLB1 and BLB2 loci, but it was not clear whether there were cases where the same sequence was occurring in both loci within a single haplotype.

Multiple allele combinations where only one co-occurring BLB allele sequence was present were found in the dataset generated by this project. Long-range amplification or fragment capture and sequencing is the only way to confirm whether one BLB locus has been deleted in these haplotypes, or whether a single sequence is present at both loci, resulting in a single BLB allele call. A preliminary hypothesis can be generated by analysis of the numbers of individual sequencing reads which map to each allele present in birds carrying these haplotypes; a BLB allele present in both loci would contribute twice as much genomic template to the PCR as an allele sequence present at just one locus, which may read out as a higher number of reads corresponding to that sequence. It is, however, difficult to control for factors such as differing amplification efficiencies (figure 2.14).

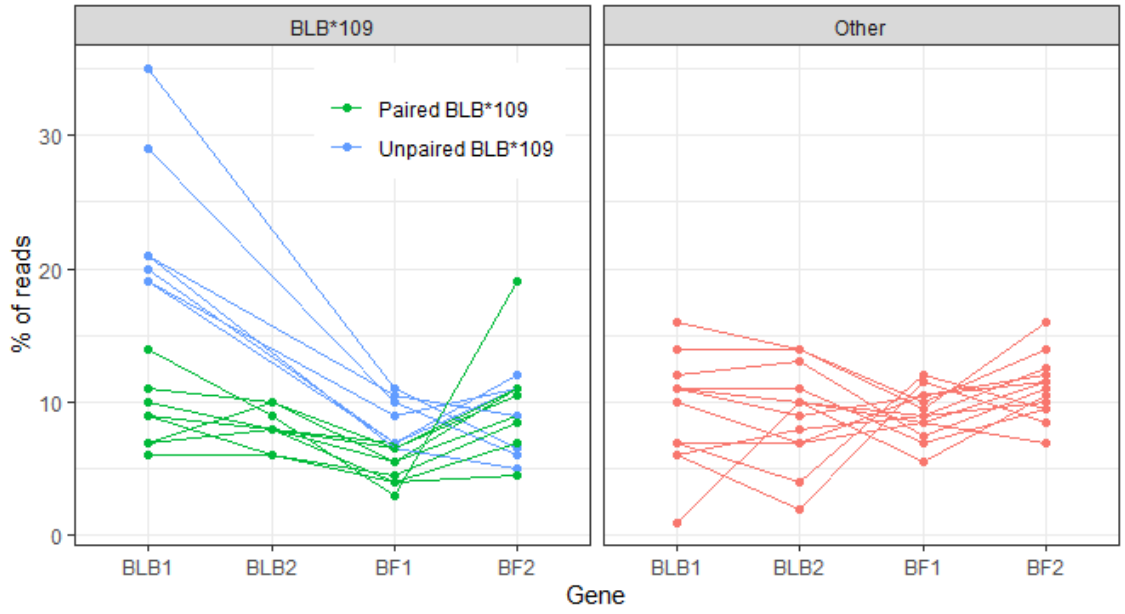


Figure 2.14: **Relative frequencies of reads matching particular allele sequences can indicate possible duplications of alleles in haplotypes**. Percentage of reads for either BF or BLB matching the sequences of alleles called in 16 heterozygous birds containing BLB*109 (reported in the BLB1 position in the haplotype) are shown. Left panel shows haplotypes containing BLB*109, both in haplotypes with [(109)-5:02-12-98] and without [(109)-?-23:02-38:02] a second BLB allele. Right panel shows the other haplotypes present in these heterozygote birds. There are more reads matching BLB*109 in haplotypes where no other BLB allele is called than in haplotypes where BLB*109 has a haplotype pair, suggesting that the allele might be present in both BLB loci in the unpaired haplotype.

2.3.6 Commercial chickens at the 'farm gate' contain limited diversity which segregates by production type

'Farm gate' level birds are those at the bottom of the breeding structure, which are provided directly as eggs or chicks to production farms. These birds contain a range of desired production traits which are bred into the final 'farm gate' birds from a hierarchy of 'elite' or high-level breeding lines, each of which contains a subset of the necessary traits and importantly retains physical and behavioural characteristics necessary for breeding, which the 'farm gate' birds do not. 'Farm gate' broilers and layers sourced from two different breeding companies (populations C and D) sequenced and analysed on run 14 were compared to 'farm gate' broilers, layers and dual-purpose birds from two further breeding companies (A1 and A2 are two broiler lines from one company; B1 and B2 are a layer and dual-purpose line respectively from a different company) which were sequenced on runs 15 and 16.

Populations B1 and D included heterozygote birds carrying both BLB1*004:01_03 and BLB1*004:04, which are never called together because they are identical over their first 150 nt. All birds from these two populations were manually examined and the results compared to those from an analysis

method based on hierarchical clustering (ESPRIT; Cai and Sun (2011), performed by Dr. Clive Tregaskes) which is able to call both BLB1 alleles. Haplotypes which were identified to have been missed due to the presence of two similar BLB1 alleles in a bird were manually added with the suffix '(manual)'.

A total of 23 unique haplotypes were identified across the six populations, which represent the 'farm gate' production lines of four major commercial breeders (figure 2.15). The total number of haplotypes per population varied from 5 to 13, with populations from companies A and B containing 5 or 6 haplotypes compared to 13 and 8 haplotypes for populations C and D respectively. Populations C and D included a greater number of very low-frequency haplotypes, which may not have been picked up in the smaller sample sets from companies A and B.

The low-expressing 4:02-21-4:01_03-21 haplotype, which is know to confer resistance to economically-important pathogens including Marek's disease, was found in all populations except B2, although never at especially high frequencies. The highest frequency haplotypes were 4:03-33-4:04-9:02 in layers and 31-31-31-31 in broilers. Both of these haplotypes are known to be low-expressing, although a broad peptide binding repertoire has only so far been proven by immunopeptidomics for 31-31-31-31 (N. Ternette and J. Kaufman, unpublished). The dual-purpose birds contained both 4:03-33-4:04-9:02 and 31-31-31-31 at reasonably high frequency. Other haplotypes also appeared to segregate by production type, regardless of breeding company. 9-9-9-9, 4-8-4:01_03-24 and 4-8-4:01_03-43 were all present in both layer lines, with the dual-purpose line also containing 9-9-9-9. Broilers from both companies contained the low-expressing 2-2-2-2 and 5:01_02-5:04-12-37:02, with the dual-purpose breed additionally sharing 2-8-6-6 and 30-30-6:02-30 with the broilers from company A.

Homozygotes were infrequent and were only observed for a few very low-expressing haplotypes, consistent with breeding structures being designed to maximise individual birds' peptide presentation capabilities.

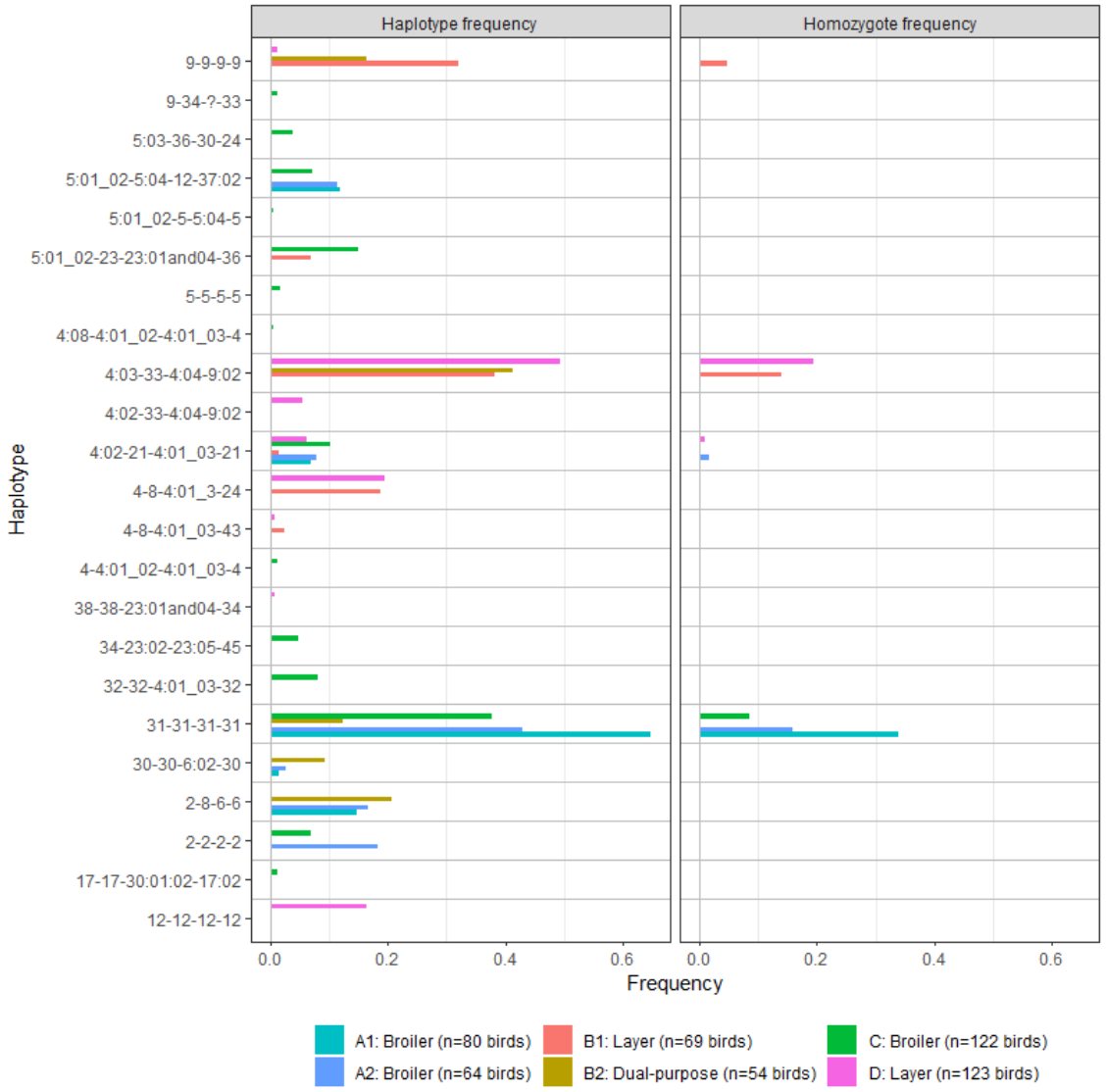


Figure 2.15: **Haplotypes in 'farm gate' commercial broilers and layers show limited diversity and segregate by production type.** Haplotype frequency was calculated as *occurrences in population/total number of haplotypes called in population*, with the assumption that missing data points occurred randomly. The maximum amount of missing data in a population was 10.2% (population B2). Homozygote frequency was calculated as *number of homozygote birds carrying haplotype x/(total number of haplotypes called in population/2)*.

2.3.7 Fancy breeds contain more diversity than commercial flocks

Of the 1869 birds in the Vinkler and SYNBREED sample sets, 1280 had at least one allele call for both BF and BLB. Twenty-four birds had more than two haplotypes called; errors were identified and corrected in twenty birds and the remaining four were removed from the dataset since sample contamination was indicated. 2257 haplotypes were called within the final dataset of 1276 birds, representing an 88.4% haplotype assignment rate (based on all 1276 birds being expected to carry two haplotypes). Missing haplotypes were generally caused by allele dropout preventing the assignment of one or more haplotypes, or additional sequences being erroneously called in likely

homozygotic birds, which prevented the automatic assignment of the second haplotype.

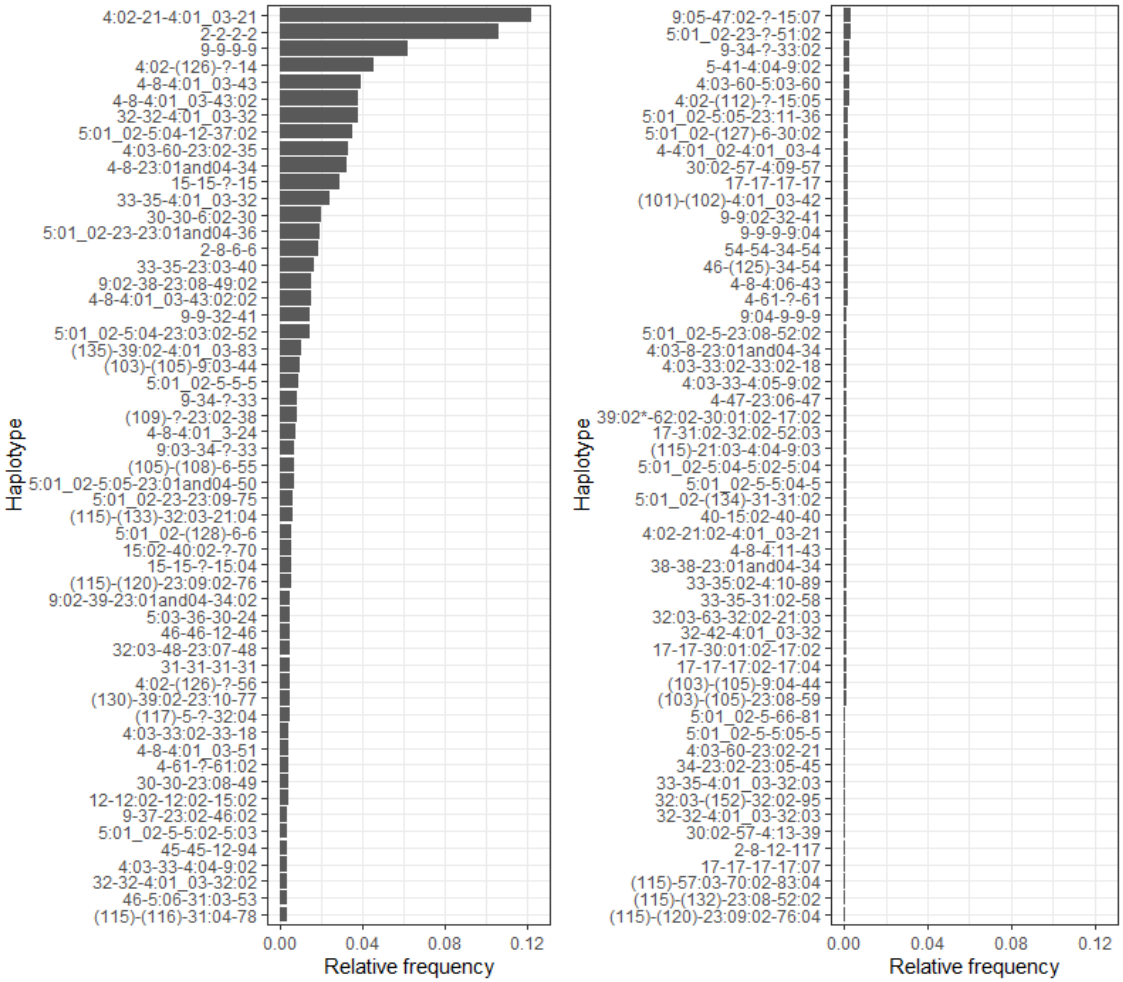


Figure 2.16: **Fancy breeds contain more diversity than commercial flocks.** All birds from the Vinkler and SYNBREED sample sets which had at least one allele call for each of BF and BLB were included in the dataset. Four birds with >2 haplotypes called were removed from the analysis. Relative frequency was calculated as *occurrences in population/total number of haplotypes called in population*, with the assumption that missing data (11.6%) were distributed randomly across birds carrying different haplotypes.

The most common haplotypes in fancy breeds reflect common haplotypes seen in commercial flocks

The most common haplotypes in the fancy breeds overall were 4:02-21-4:01_03-21 and 2-2-2-2 (figure 2.16), which are both seen commonly in commercial flocks and are known to contain BF2 alleles which express at a low level on the cell surface and have wide peptide binding repertoires. Other haplotypes seen commonly in commercial birds including 9-9-9-9, 4-8:4:01_03-43 and 32-32-4:01_03-32 are also present at high frequency.

BF1*NULL haplotypes are common in fancy breeds

Two haplotypes which appear to lack a BF1 allele altogether are also present at high frequency. 15-15-?-15 is the standard B15 haplotype, while 4:02-(126)-?-14 is likely to be related to the standard B14 haplotype, both of which contain a deletion of the BF1 gene (Wallny et al. (2006); Shaw et al. (2007); Afrache et al. (2020) and F. Coulter, E. Palmer and J. Kaufman, unpublished). 9-34-?-33, which occurs at a slightly lower frequency has also been reported as lacking BF1 (Afrache et al., 2020), which is consistent with the absence of alleles attributable to this locus in the sequences obtained from birds carrying this haplotype. For other haplotypes, such as 15:02-40:02-?-70, which were newly discovered in this project, the absence of evidence for BF1 is not necessarily evidence of absence, since non-universality of primers or poor amplification of specific sequences are also possible (albeit unlikely) explanations.

The high overall frequencies of BF1*NULL haplotypes are largely driven by high frequencies in breeds represented by large numbers of individuals (figure 2.17). 4:02-(126)-?-14 is extremely common in Czech Golden Pencilled birds, while 15-15-?-15 is common in Leghorns. In both these breeds, the BF1*NULL haplotype occurs in homozygotes at reasonable frequency, which could indicate inbreeding depression (if the absence of BF1 is detrimental) or might suggest that BF1*NULL haplotypes do not confer a significant disadvantage and are therefore subject to minimal negative selection. The Friesian, Malay, La Fleche and Minorca breeds also carry these haplotypes.

31-31-31-31 is rare in fancy breeds despite being common in commercial birds

Despite its prevalence in commercial broilers, 31-31-31-31 is only seen in ten fancy breed individuals: three Plymouth Rocks, six Albanische Krahers and one Cochin. The Plymouth Rock breed was a significant contributor to early commercial broiler development and it may be that commercial breeders would never have had access to the 31-31-31-31 haplotype had they not chosen this breed to provide genetic resources to early broiler flocks. The other breeds in which 31-31-31-31 is seen do not share a geographical origin or recent heritage with the Plymouth Rock, and it seems unlikely that 31-31-31-31 would have been selected *out* of almost all other breeds, so the distribution of this haplotype within fancy breeds remains unexplained. Typing more individuals from breeds currently only represented by a few birds might uncover 31-31-31-31 at low frequencies, and help to explain the distribution. Nonetheless, with other low-expressing, common commercial haplotypes such as 4:02-21-4:01_03-21 and 2-2-2-2 being clearly beneficial for fancy breeds, there may be potential for 31-31-31-31 to be actively bred into other breeds to improve health.

2.3.8 There is minimal population structure within fancy breeds

Fancy breed samples were obtained from two sources: the laboratory of Dr. Michal Vinkler and the SYNBREED chicken diversity panel (Malomane et al., 2019). Each of the two sample sets contained samples from multiple breeds, each sampled from multiple European breeders. Twenty out of 55 total breeds for which samples were obtained were included in both sample sets. In subsequent analysis, 'population' will be used to describe the birds of a given breed in a given sample set and 'breed' will describe all birds of a given breed pooled between sample sets (if applicable).

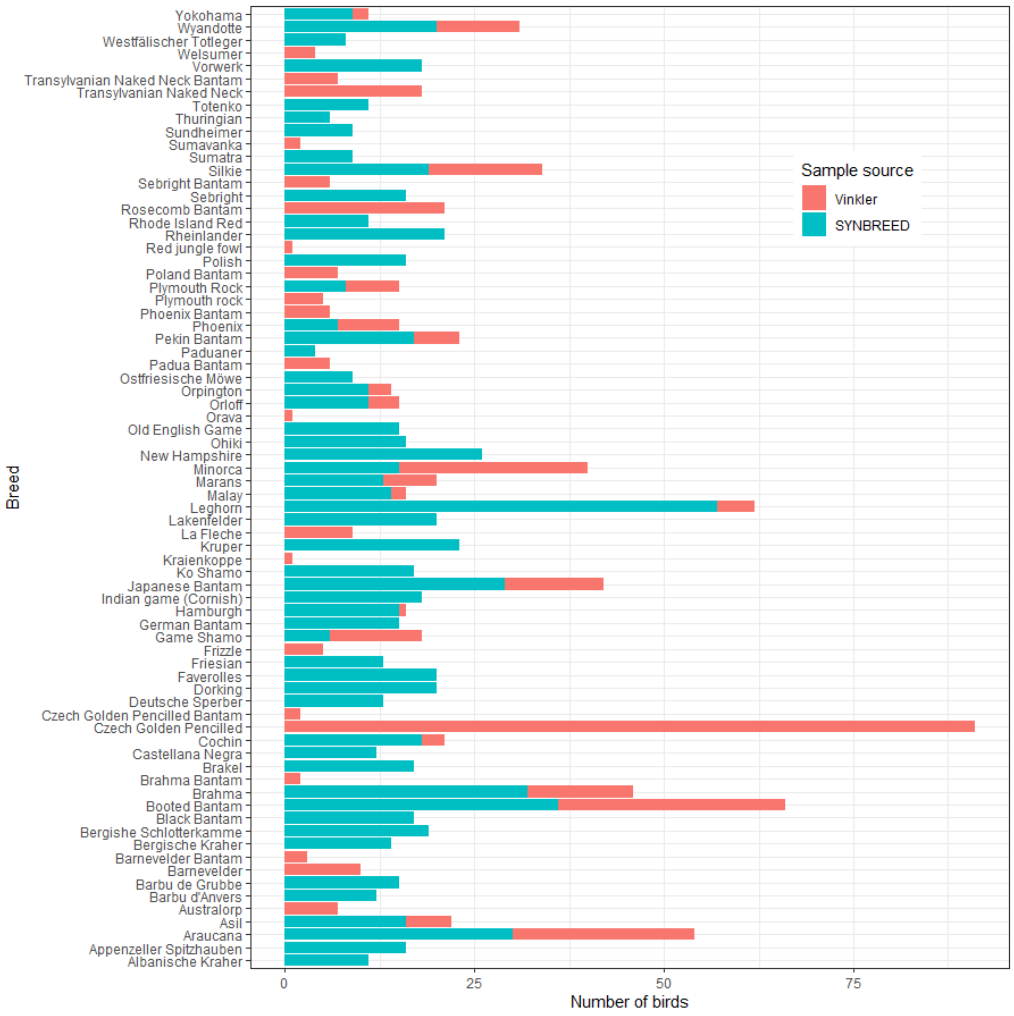


Figure 2.17: **Breed distribution of birds from the Vinkler and SYNBREED sample sets which at least one allele call for each of BF and BLB.** Birds are included in the dataset if at least one call was obtained for each of class I and class II.

Breeds represented by >25 individuals and where no more than 70% of individuals were from a single sample set (figure 2.17) were investigated to determine the influence of sampling on the haplotype frequencies observed in the breeds, given that samples were obtained from different groups of European breeders. The most common haplotypes in each of the breeds examined were

seen in both sample sets, with rarer haplotypes sometimes present in just one sample set, as would be expected given the observed frequency distributions and independent sampling (figure 2.18). Wyandotte birds had more and higher-frequency haplotypes which were restricted to a single sample set, indicating some additional population stratification in this breed. Few breeds shared high-frequency haplotypes with other breeds. The prevalence of missing data (both allele and haplotype calls) was variable between breeds, possibly due to variation in the quality of samples provided by different breeders or to the relative frequencies of haplotypes particularly susceptible to allele dropout. Non-random allele dropout was not seen to be a significant confounding factor, but manual examination of the allele calls in each individual bird may improve the accuracy of the stated haplotype frequencies. Overall, these data suggest that the majority of fancy breeds in Europe are genetically distinct from one another, but have little internal population structure.

Figure 2.18: **Haplotype frequencies in seven fancy breeds show limited inter-population variation.** The majority of high-frequency alleles in a given breed were detected in both sample sets, indicating limited variation between populations. Variation between breeds is noticeably higher. Haplotype '0' indicates missing haplotypes.

The observations of low intra-breed/inter-population variation and higher inter-breed variation were broadly reflected in an analysis of all populations represented by 10 or more birds. Genetic distance between all pairs of individuals was calculated and then averaged over pairs of populations to obtain a matrix of genetic distances between populations using GenAlEx (Peakall and Smouse, 2006, 2012). The dimensionality of this matrix was reduced in R and the resulting table was

visualised in R using ggplot2 (figure 2.19). Some features of figure 2.18, such as generally low genetic distance between different sample sets from a single breed, are also present in this analysis, however other details, such as Wyandottes having more differentiation between sample sets than other breeds in figure 2.18, are not obviously reflected in figure 2.19. With only 32.5% of the variation contained in the distance matrix represented in the two dimensional visualisation, this analysis suggested that there was no major underlying population structure in fancy breeds, and with so much variation not represented, some apparent disagreements with figure 2.18 are not necessarily unexpected. In particular, the dimensions on this plot (corresponding to the first two principal components or eigenvectors) do not obviously correlate with the geographic origins of the breeds, although some association between high V1/low V2 and Asian ancestry is possibly discernable (Ko Shamo, Asil, Cochin, Brahma, Malay, Game Shamo). It therefore seems unlikely that fancy breeds now being bred in Europe show significant signatures of their geographic origins in their MHC haplotypes, although analysis of individual breed combinations may reveal patterns in the presence or absence of particular alleles which reflect breed history.

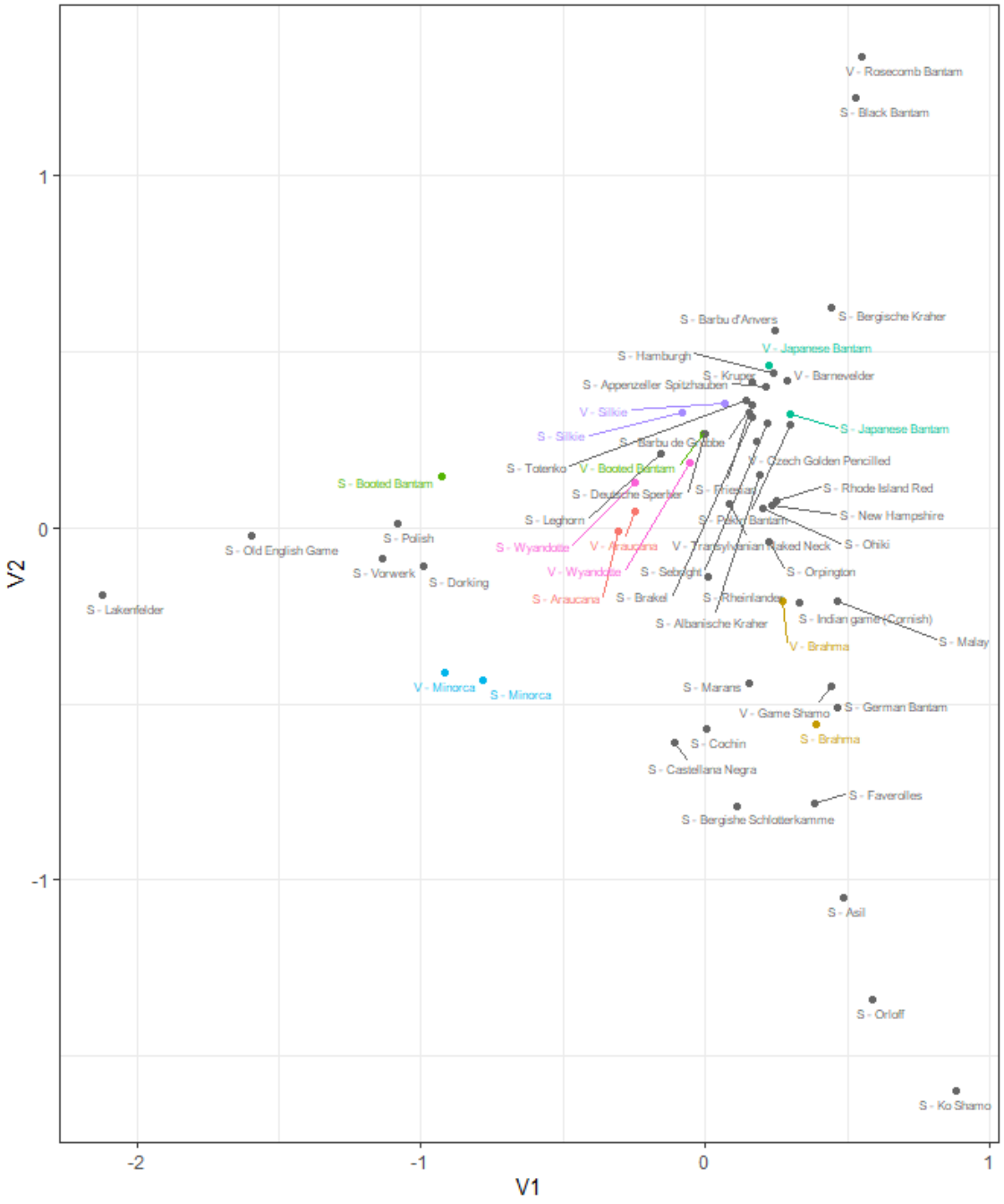


Figure 2.19: **Reduced-dimensionality visualisation of genetic distance between 55 fancy breed populations.** Breeds present in both sample sets and analysed in figure 2.18 are coloured to highlight proximity. 'S' and 'V' before breed names indicate birds from the SYNBREED and Vinkler sample sets respectively.

Breeds associated with a proposed 'Pacific expansion' do not carry different MHC haplotypes to those associated with the 'European expansion'

It was hypothesised that if two major geographic expansions of domestic chickens from Asia had occurred then breeds associated with the geographies of the two expansions might differ in their haplotype frequencies, either as a result of genetic bottlenecks, genetic drift, natural/artificial selection or most likely a combination of these factors. In this project, a potential expansion of chickens

into South America via South-East Asia and the Pacific islands was represented by the Indonesian Sumatra breed and the Chilean Araucana. Sumatra chickens (n=9) contained 4:02-21-4:01_03-21, 4-8-23:01and04-34 and 30-30-6:02-30, all haplotypes seen commonly in European/North American breeds and in the commercial lines derived from them. Araucanas (n=54) contained one unique haplotype, 5:01_02-(127)-6-30:02, which was seen in five individuals, but the predominant haplotypes 2-2-2-2 (18 birds), 5:01_02-23-23:01and04-36 (15 birds), 4-8-4:01_03-43 (13 birds) and 9-9-9-9 (8 birds) are all shared with a number of breeds of wide-ranging origin. There was therefore no evidence from the MHC haplotypes present in these breeds to support this proposed route for the expansion of domestic chickens, although neutral markers, and examination of a larger number of loci may yet provide genetic support. Furthermore, there was little evidence that the Sumatra and Araucana breeds, which would be predicted under the hypothesis of two separate expansions to share a relatively ancient common ancestor with the majority of modern production birds, contain additional variation at the classical MHC loci which could be usefully incorporated into breeding strategies.

Disease resistance in Transylvanian Naked Necks cannot be attributed to breed-specific MHC alleles

Transylvanian Naked Neck birds are anecdotally resistant to a large number of poultry diseases, however they carry no unique MHC haplotypes and their highest-frequency haplotype, (103)-(105)-9:03-44, is also present in Plymouth Rock, Minorca and Wyandotte birds, none of which are described as having the same disease-resistant characteristics. It is therefore unlikely that this apparently generalised resistance is MHC-linked and further work would be required to determine the genetic basis of the trait if it is to be used to improve the health of other populations.

Susceptibility to Marek's disease in Silkies and Sebrights may be associated with a lack of resistant MHC haplotypes at high frequencies

Although no formal studies have been conducted, Silkies and Sebrights are commonly reported by vets and breeders to be particularly susceptible to Marek's disease (Roberts, 2009; Wallner-Pendleton, 2019). MHC typing of both breeds (Silkies n=34, Sebrights n=16) suggests that known Marek's disease resistant haplotypes 4:02-21:4:01_03-21 and 2-2-2-2 are present in the breeds but at very low frequencies (figure 2.20). Low frequencies may be the result of vaccination of these breeds against Marek's disease, which would reduce the strength of selection for resistant haplotypes provided the vaccine was effective in most haplotypes. This hypothesis, however, would require that these resistant haplotypes were only introduced into the breeds recently, and may suggest that genetic resistance could become more common in subsequent generations. If the haplotype frequencies in these flocks are representative of the breed, it suggests that haplotypes such as 9:02-38-23:08-49:02, 5:01_02-5:04-12-37:02, 9-34-?-33 and 4-8-4:01_03-43 are unlikely to

provide protection from Marek's disease. Surprisingly, 4-8-4:01_03-43 is relatively common in commercial layers (N. Jackson, C. Tregaskes and J. Kaufman, unpublished), and has been observed in homozygotes within these populations, which may suggest a strong commercial reliance on effective vaccination.

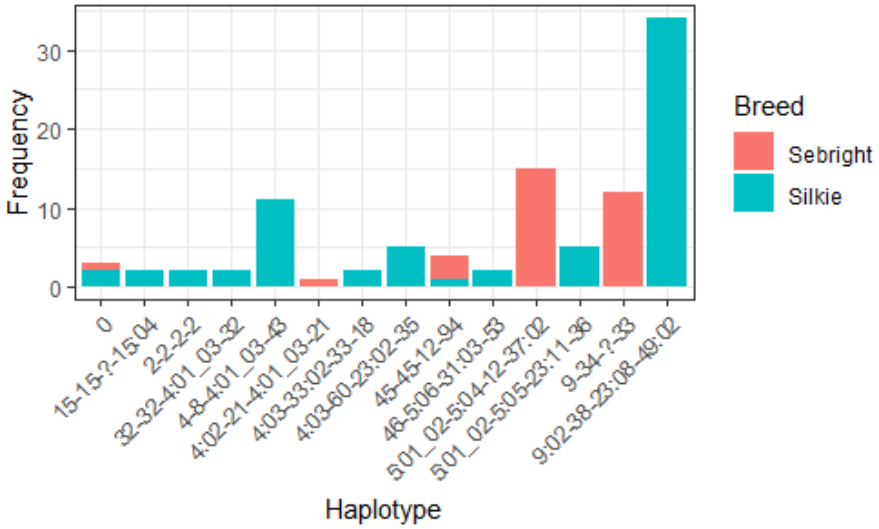


Figure 2.20: **Silkies and Sebrights, which are susceptible to Marek's disease, carry resistant haplotypes at very low frequencies.**

2.3.9 African village chickens contain very high haplotype diversity

A total of 1383 chickens from the Roslin Ethiopia/Nigeria, Ethiopia and Various African, and UC Davis African populations had at least one allele called for both BF and BLB. Four birds with more than two reported haplotypes and where the calls could not be corrected on manual examination were removed from this dataset. It is possible that additional loci could be present in groups of related individuals, but the distribution of these four birds across the various sample sets suggested that this was unlikely to be the case. The total proportion of possible haplotype calls which were missing was 26.7%, attributed to both low-quality samples and sequencing, and the presence of a number of novel low-frequency allele combinations for which insufficient evidence was available to define new haplotypes. Missing data was unevenly distributed across the sample sets, reflecting variation in sample, library preparation and sequencing quality. Proportions of missing data were 37.9%, 55.6%, 17.7% and 25.0% in the Roslin Ethiopia, Ethiopia/Nigeria, Various African and UC Davis African populations respectively and the typing programme will aim to repeat all birds which were not fully typed at a later date.

Some key commercial haplotypes are also present in African flocks

Common haplotypes from commercial and fancy breeds (4:02-21-4:01_03-21, 4-8-4:01_03-24, 9-9-9-9, 4:03-33-4:04-9:02 etc.) remain common in the African birds albeit at lower frequencies; in the

dataset overall, no individual haplotype was present at >5%, whereas both 4:02-21-4:01_03-21 and 2-2-2-2 were present at >10% overall in the fancy breeds (figure 2.21). 31-31-31-31 was present in five individuals from the Roslin Various African and UC Davis sample sets (0.25% overall relative frequency), perhaps due to introgression from commercial broilers introduced for development initiatives.

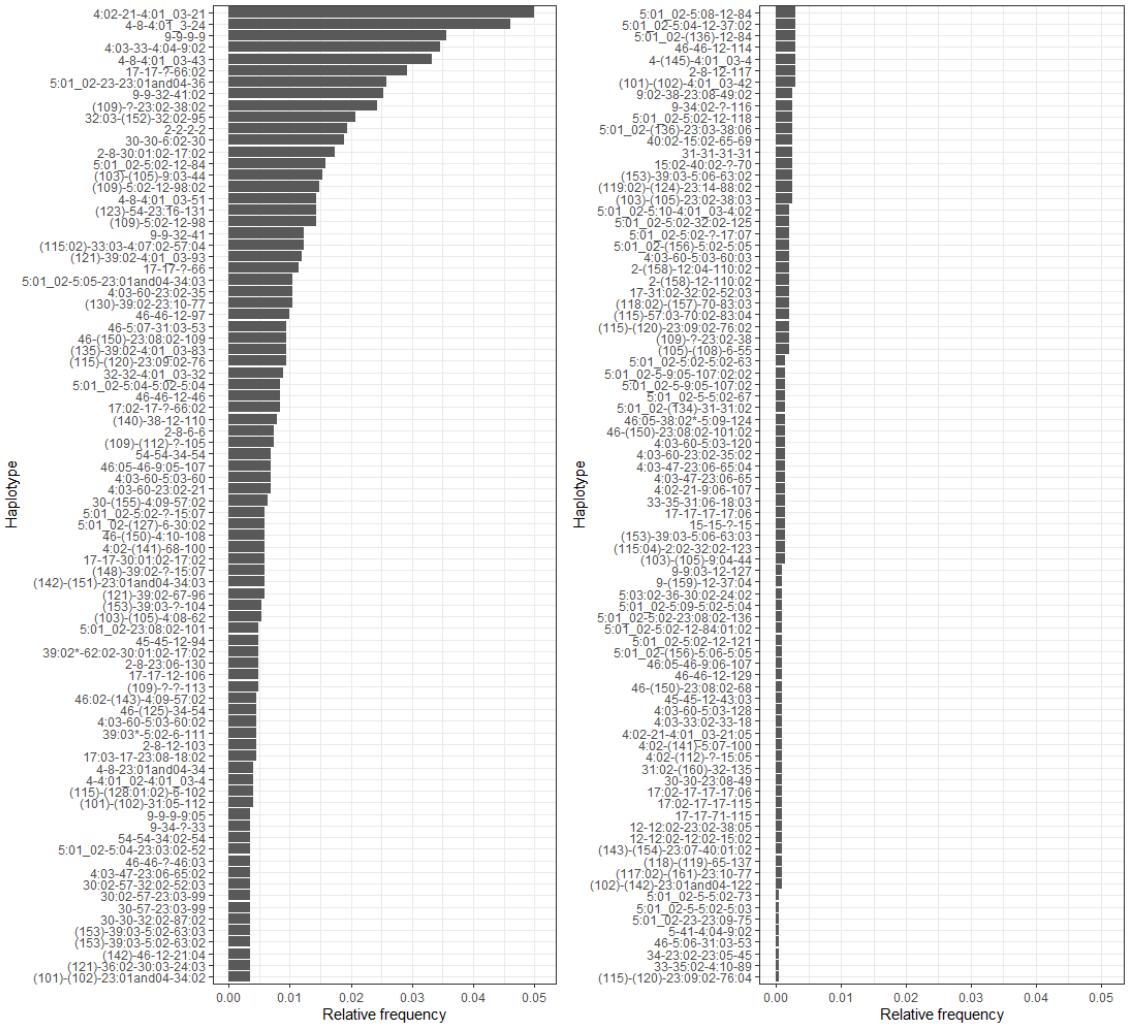


Figure 2.21: **Haplotype frequencies in African village chickens.** All birds from the Roslin and UC Davis African sample sets which had at least one allele call for each of BF and BLB were included in the dataset. Four birds with >2 haplotypes called and where the true haplotypes could not be resolved were removed from the analysis. Relative frequency was calculated as *occurrences in population/total number of haplotypes called in population*, with the assumption that missing data (26.8%) were distributed randomly across birds carrying different haplotypes. Missing haplotypes may also correspond to low-frequency novel haplotypes for which insufficient evidence was present.

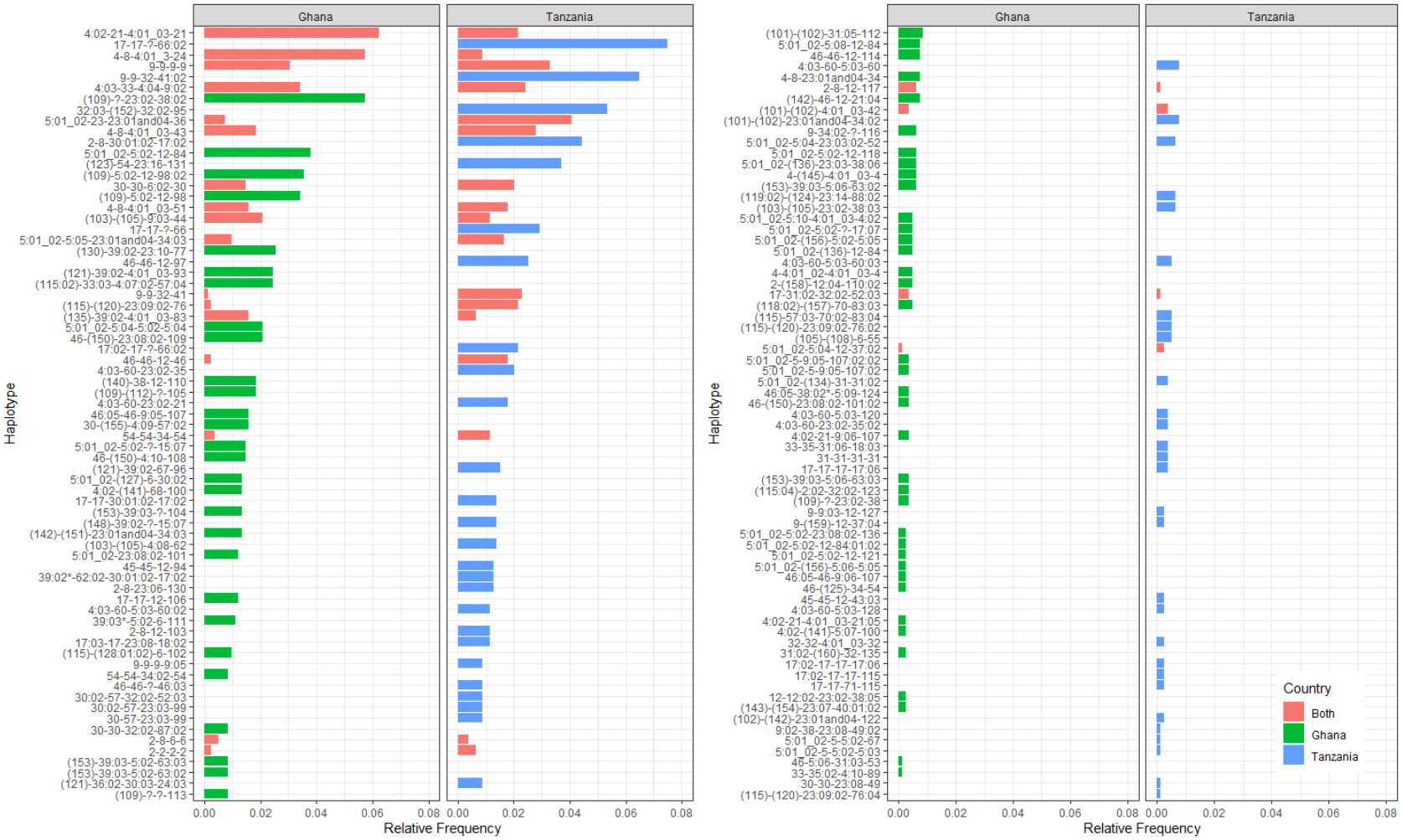


Figure 2.22: **Relative haplotype frequencies in Ghanaian and Tanzanian chickens.** Relative frequencies calculated as *number of times called/total haplotypes called* per population, assuming that missing values occurred randomly. The dataset included 538 Ghanaian and 530 Tanzanian chickens, with 257 (23.8%) and 273 (25.8%) missing haplotypes in the two populations respectively. Haplotypes are shown in order of total frequency across the two populations.

2.3.10 The majority of MHC haplotypes present in Ghanaian and Tanzanian chickens are unique to one country

Of the 142 total haplotypes found in the UC Davis samples from Ghana and Tanzania, only 21 are found in both countries (figure 2.22). These shared haplotypes are marginally more common than haplotypes that are unique to each country; the average relative frequency of a shared haplotype is 0.015 in both populations, compared to 0.010 for each of the 66 unique Ghanaian haplotypes and 0.012 for each of the 55 unique Tanzanian haplotypes. Several known low-expressing haplotypes (4:02-21-4:01_03-21, 4:03-33-4:04-9:02 and 4-8-4:01_03-43) are all within the six most prevalent shared haplotypes. The presence of these important commercial haplotypes in both African populations could mean that the flocks have some European ancestry, or that these haplotypes were important in Red Jungle Fowl and have been maintained by selection in a huge range of domestic and semi-domestic environments.

The fact that the majority of haplotypes are unique to a particular country raises interesting questions about the relative roles of genetic drift and divergent selection in these two populations. There is no evidence that either country's unique haplotypes more closely resemble the haplotypes seen in commercial flocks, so neither seems to have been affected by immigration and introgression by European birds significantly more than the other.

2.3.11 Within Tanzania, different ecotypes share the majority of their MHC haplotypes

The Tanzanian birds used in this study include three popular ecotypes, Ching'wekwe (Ching, n=176), Kuchi (n=103), and Morogoro Medium (MoroMid, n=251). Twenty-one haplotypes are shared between all three populations, and these haplotypes occur at higher average relative frequencies in the populations than those haplotypes which are restricted to particular ecotypes (figure 2.23). The low-expressing haplotypes 4:02-21-4:01_03-21, 2-2-2-2 and 4-8-4:01_03-43 are all present in this group, as is 17-17-?-66:02, the most common Tanzanian haplotype overall. Since the sample sizes in this analysis are relatively small considering the diversity of haplotypes present, it is likely that a proportion of low-frequency haplotypes currently represented in one or two ecotypes are present in other populations but were not detected due to sampling effects. These results should therefore be considered a minimum for the amount of shared diversity between these three ecotypes.

These results are somewhat contrasting to the results of Walugembe et al. (2019) who modelled the relationships between these three ecotypes using a 600K SNP array and found that a 2-subpopulation model in which Ching and MoroMid formed a single admixed population, separate from Kuchi, gave the optimum fit for the data. The MHC typing suggests that while Kuchi

birds may be genetically distinct from Ching and MoroMid, indicating a degree of reproductive isolation, they have maintained a largely similar distribution of MHC haplotypes, perhaps due to similar pathogen pressures in their respective habitats. The haplotypes present in Ching and MoroMid but not in Kuchi may have been lost by drift or may not have been detected in the relatively small Kuchi sample. The haplotypes unique to the Ching and MoroMid populations are more difficult to explain, since these ecotypes were sampled from neighbouring regions where natural admixture is likely to be facilitated by trading and market chains. Furthermore, Ching and MoroMid birds are highly similar in appearance and birds can sometimes be wrongly assigned, further increasing any apparent similarities between these ecotypes. A degree of differentiation could be maintained by birds preferentially mating with their own ecotype, even when potential mates from both ecotypes are present.

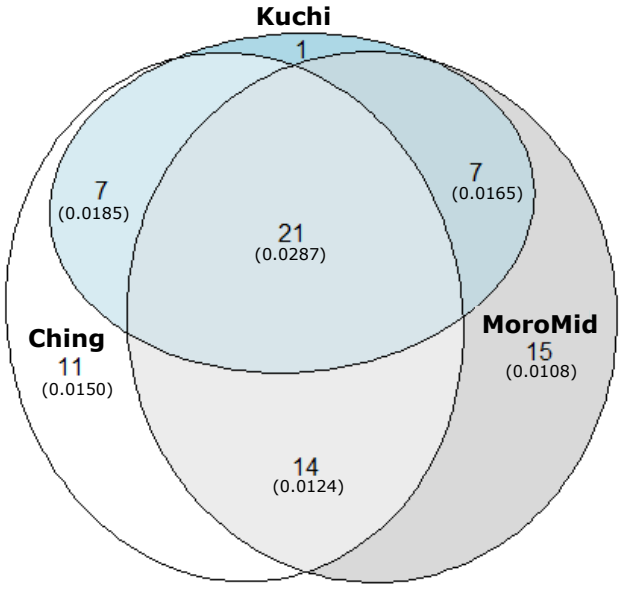


Figure 2.23: **Tanzanian ecotypes share the majority of their MHC haplotypes with other ecotypes.** Numbers of haplotypes shared between all combinations of ecotypes are shown, with the average relative frequency of the haplotypes in that subset across the three populations indicated in brackets.

2.3.12 Several common haplotypes in Ghana are associated with faster clearance of Newcastle disease virus, while common haplotypes in Tanzania are associated with slower clearance

Data from Walugembe et al. (2019, 2020) allowed associations between particular MHC haplotypes and Newcastle disease virus (NDV) response phenotypes to be investigated. Birds were challenged via the oculo-nasal route with 10^7 50% embryonic infectious dose (10^7EID_{50}) of a live attenuated type B1 LaSota lentogenic NDV strain at 28 days of age and monitored for 10 days (Walugembe et al., 2019). Viral load and NDV antibody titer were measured by qPCR from lachrymal fluid and ELISA from blood respectively (Rowland et al., 2018; Walugembe et al., 2019). Since the propor-

tion of homozygotes in the population was low, the analysis utilised both homo- and heterozygous birds that carried particular haplotypes. It was therefore largely restricted to the detection of dominant traits and limited by the variation in phenotype which is expected in groups of heterozygotes sharing a single haplotype.

Overall, while the majority of the variance for all traits occurred within, rather than between, haplotype groups, some groups did exhibit significant differences in trait values relative to the overall population (figure 2.24). The three most common haplotypes in Tanzania, 17-17-?-66:02, 9-9-32-41:02 and 32:03-(152)-32:02-95, are all associated with higher than average viral loads at 6dpi, and, in the cases of 17-17-?-66:02 and 32:03-(152)-32:02-95, lower than average viral clearance. Birds carrying 17-17-?-66:02 also grew significantly less over the course of the experiment, while those carrying 9-9-32-41:02 had higher than average viral loads at 2dpi, possibly indicating the involvement of an innate immune-related susceptibility phenotype.

Conversely, two of the three most common haplotypes in Ghana, 4-8-4:01_03-24 and (109)-?-23:02-38:02, both appear to confer significant resistance to Newcastle disease virus, with birds carrying these haplotypes clearing the virus significantly faster and exhibiting a lower viral load at 6dpi. Taken together, these results suggest that evolution of the MHC is not being affected by Newcastle disease virus in the same way in these two countries, perhaps because of differences in viral exposure, virulence or access to vaccines which might reduce the strength of selection imposed by the virus.

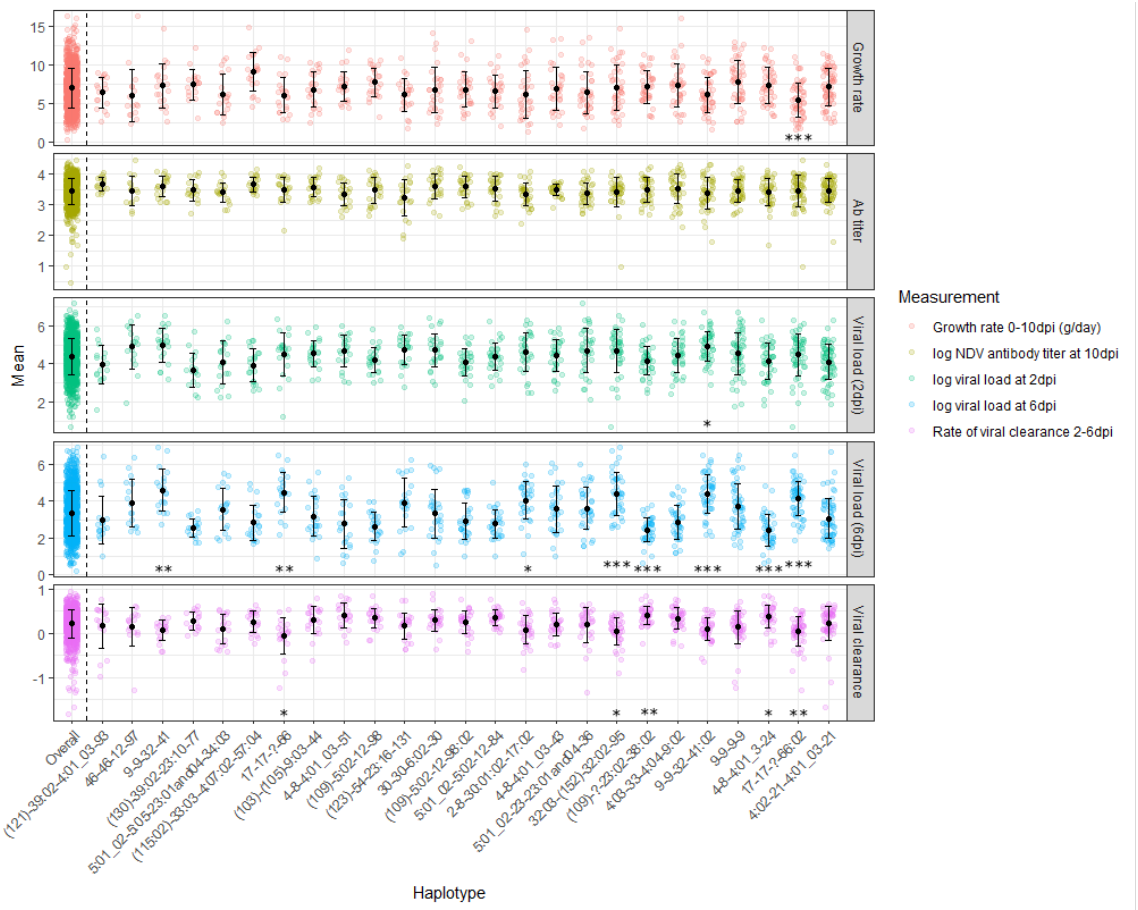


Figure 2.24: **Newcastle disease virus responses in Tanzanian and Ghanaian chickens differ between populations carrying different common haplotypes.** Means and standard deviations of measurements taken following exposure to an infected individual. Since very few homozygotes were present in the population, birds were included in the dataset for a given haplotype if they carried that haplotype on either or both chromosomes; visible effects were therefore limited to dominant phenotypes. The 25 most common haplotypes across Ghana and Tanzania were included in the analysis and are presented R-L from most to least common. Antibody titer and viral load were both $\log_{10}$ transformed. Viral clearance was calculated as the difference in viral load at 2 and 6dpi divided by the viral load at 2dpi. Significant differences from the overall population average (Holm-corrected Mann-Whitney U test) are indicated by asterisks below the data: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

2.4 Discussion

In this work, the database of known chicken MHC alleles and haplotypes was significantly expanded and populations outside commercial and experimental lines were sequenced in detail at the classical MHC loci for the first time. Comparisons between populations at various geographic scales illustrated the need to carefully consider the utility of interventions developed in one population in another. In particular, free-range local or 'village' chickens contain a huge amount of diversity, much of which is restricted to particular geographies and may have evolved in response to particular pathogen pressures.

The expanded database also highlighted areas where current approaches did not account for particular types of variation in the chicken MHC. In particular, locus designations for both BF and BLB alleles were shown to be more complex than previously thought.

Analysis of positively selected sites in random subsets of sequences from the expanded database showed results consistent with BF1 and BF2 having distinct functions. In particular, BF2 appeared to be under greater selection for diversity in the peptide binding groove, consistent with BF1 playing a lesser role in presentation of pathogenic peptides to T cells and perhaps acting as an NK cell ligand (Ewald and Livant, 2004; Kim et al., 2018). Sequences assigned to BLB1 and BLB2 did not obviously differ in their PSS, consistent with these genes having similar functions, perhaps in different tissues or in response to different immune stimuli. It is, however, becoming clear that BLB allele sequences cannot be easily assigned to a particular locus, especially outside the context of a defined haplotype, and thus the subsets examined may not be particularly biologically meaningful.

2.4.1 Issues with current nomenclature

The development of a significantly expanded database of chicken MHC alleles highlighted the challenges of establishing a nomenclature which incorporates useful information about the alleles into their names while being sufficiently flexible and robust to be applicable to future work. The nomenclature used in this project was described by Afrache et al. (2020), who also acknowledged and discussed several of the limitations of the system discussed below, which were becoming clear as PCR-NGS typing data was being generated at the time of writing.

It should be noted that the system used for naming new alleles was developed throughout this project, as the scope of the complexities of the system were gradually understood. There are therefore some inconsistencies in how alleles were named between runs, but it was considered preferable to avoid renaming alleles and creating version incompatibilities until a robust nomenclature had been agreed and data could be re-analysed *en masse*.

MHC allele locus designations may be more complex than anticipated

The first field in the current nomenclature refers to the locus at which the allele is found. It was thought new BF alleles could be reliably assigned to a locus according to their position on the phylogeny, but we now know of at least two independent events which have caused BF1 and BF2 alleles to become poly- and paraphyletic respectively. It therefore seems important to establish a nomenclature which is, as far as possible, robust to the discovery of further, similar cases of evolutionary idiosyncrasy.

For class II, without a simple phylogeny to support locus assignments, and with existing evidence that identical BLB exon 2 sequences can be amplified from both BLB loci (Worley et al., 2008), some non-locus-specific naming was already applied in this project. In earlier runs, before the scope of the problem was fully understood, some sequences were assigned to a locus based on the presence of a consistently co-occurring allele (the presumed 'haplotype pair') which was assigned to the other locus. In later runs, however, this was considered weak evidence, and sequences were more likely to be assigned a locus-neutral name.

Allele 'groups' based on sequence similarity thresholds cause conflict in class II and eventually overlap in both classes

The similarity-threshold rule was broadly applied in earlier runs, resulting in BLB2*039:02 being named based on similarity to the BLB2 gene in the 9:02-39-23:01and04-34:02 haplotype and BLB2*060:02 being named based on similarity to BLB2 in 4:03-60-5:03-60, despite these two alleles occurring together in the haplotype 39:02*-62:02-30:01:02-17:02 (with an asterisk indicating that 39:01 is in the BLB1 position despite its name referring to the BLB2 locus).

There are additional limitations of the use of a threshold sequence similarity to assign closely-related alleles to groups. Inevitably, as the database expands, conflicts arise; if BLB alleles A and B differ at two amino acids they might be assigned BLB*50:01 and BLB*50:02, but if allele C differs from A at two amino acids and B at four amino acids, it is not clear whether C should be BLB*50:03 or not. Importantly, if the discovery of new alleles suggested that a cluster of closely related alleles existed around A and C, it would be tempting to rename B in order to have allele group 50 include only sequences which were below the similarity threshold with respect to all other members of the group. This would lead to the same allele sequence having different names in different versions of datasets. Alternatively, if 'chains' of allele names were allowed to arise (for example, A, B and C all being considered part of allele group 50), eventually a new allele would be below the threshold similarity level with respect to alleles from two different allele groups.

2.4.2 Proposed alternative nomenclature

A possible alternative nomenclature (figure 2.25) should avoid the two key points of conflict and inaccuracy, namely the use of similarity thresholds for assignment to allele groups and the use of locus designations in allele names.

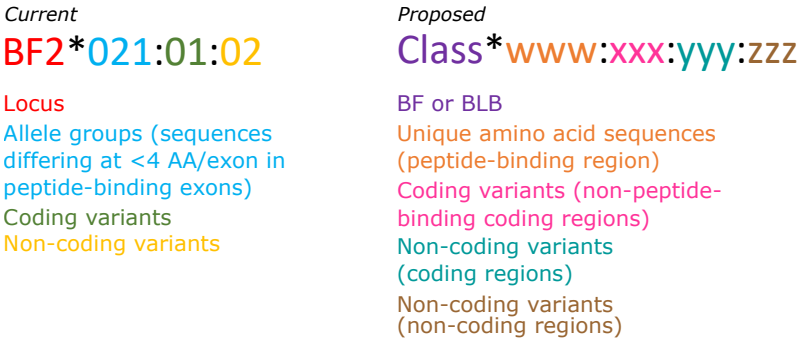


Figure 2.25: **Current vs. proposed chicken MHC nomenclature.** Descriptions refer to the types of sequence variation distinguished by each field.

Locus assignments

In class II in particular, there does not seem to be anything inherent in exon 2 sequences that restrict them to either BLB1 or BLB2, so including these assignments in sequence names is almost inevitably going to cause conflicts as more birds are examined. For class I, although locus designations seem to be more predictable, there is clearly potential for rare recombination events to generate unexpected tree structures, and the benefits of using the same simple system for both BF and BLB seem to outweigh the benefits of distilling locus information into class I allele sequence names.

The proposed nomenclature would therefore assign novel sequences to a class (in this case based on the primer pair with which they were amplified) but not to a specific locus. Alleles can, however, be linked to a particular locus by their position in *haplotype* nomenclature, since locus assignments are meaningful in the context of a given haplotype. The existing BLB1-BLB2-BF1-BF2 structure would still be applicable, with allele numbers inserted at the most likely locus and uncertainty indicated by brackets until the structure had been confirmed. A long-amplicon sequencing methodology to investigate the locus assignments of novel haplotypes is currently in development (T. Tan and J. Kaufman, University of Edinburgh). This methodology will also be key to the investigation of haplotypes which appear to either be lacking alleles or contain identical allele sequences at multiple loci.

Allele naming

The use of similarity thresholds to assign alleles to groups raised several conflicts during the project. While identification of related sequences can be useful, for example when assessing likely haplotype combinations, there is no distinction made between functional and non-functional variation, and thus these names could be misleading when considering the disease-response phenotypes associated with different alleles or haplotypes. An alternative suggested by Afrache et al. (2020) is

to use a descriptor of peptide-binding repertoire to assign alleles to functionally related groups, however there is as yet no consensus on the best way to assign alleles to so-called 'supertypes' based on sequence alone, even for the well-studied HLA system (Sidney et al., 2008; Greenbaum et al., 2011). Robinson et al. (2017) identified 42 'core' HLA alleles, which could not be derived by simple recombination events from other alleles, from a database of well over 10,000 alleles. A similar retrospective approach to grouping alleles would be inappropriate when the chicken MHC database is still relatively small and expanding so rapidly, and it is still not clear whether these core alleles and the groups of alleles related to them are have distinct functionalities.

A system to support the expansion of the chicken MHC database going forward, in the absence of a reliable metric by which to group alleles, is likely to assign arbitrary numeric identifiers to each unique sequence found. Since most studies examine the variable peptide-binding exons (exons 2 and 3 for BF and exon 2 for BLB), and sequences identical at the amino acid level over these regions are likely to have very similar binding repertoires, it seems appropriate that allele sequences are defined by these regions in the first level of the classification.

Variation outside of the variable peptide-binding exons is unlikely to affect peptide binding repertoire, and is often not detected by studies, which typically examine only the peptide-binding exons. In order to ensure that the first numeric field reflects functional variation (to the extent this is currently possible to define), and to minimise problems associated with naming partial sequences, non-synonymous variation outside these regions would be reported in the second field. For the majority of current studies which only examine peptide binding exons, the second field would be '0' to indicate the absence of sequence information.

Sequences which differ in the variable peptide-binding exons only at synonymous positions are likely to have very similar phenotypes, and it therefore seems sensible to distinguish non-coding variants in a later field, allowing these sequences to be grouped together in the first field.

Finally, variation in non-coding regions (introns, 5' UTR etc.) would be distinguished in a fourth field.

Standard haplotypes

It does not seem to be in the interests of the field to make significant changes to the names of well-studied standard haplotypes. It is important that existing literature is still accessible and can be integrated with ongoing work, and unrealistic to expect researchers to 'jump ship', when the existing nomenclature works well for low-diversity experimental or commercial populations. A possible solution would be to maintain locus-specific naming for the standard haplotypes, which

would allow the vast majority of standard alleles to keep their current numbers. If the new structure for numeric fields was applied to the standard haplotypes, a few changes would be necessary, for example BF2*005:02 (the BF2 gene in the standard B8 haplotype) would become BF2*008:01, since this sequence is not identical at the amino acid level to BF2*005:01 over exons 2 and 3.

Partial sequences

The proposed nomenclature deals well with the potential for new alleles to be discovered as either full-length or peptide-binding exon-only sequences with respect to coding changes but does not currently make this distinction for non-coding changes. Manual curation of the database would therefore still be required to monitor the exchange of incomplete sequences for more complete ones as they were obtained.

2.4.3 Implications for poultry management

Commercial flocks

While commercial poultry breeding strategies are closely-guarded secrets, it can be assumed that the disease- and vaccine-response phenotypes of these birds have been carefully optimised for performance in production systems. High-frequency haplotypes in fancy breeds resemble high-frequency haplotypes in commercial flocks, suggesting that there is limited variation that could be introduced from fancy breeds into commercial flocks to confer a significant benefit. Free-range local birds in Tanzania however do contain previously unreported haplotypes at high frequencies. While the strongest selection pressures on these birds are likely to differ significantly from those on commercial flocks, the phenotype of 17-17-?-66:02 would definitely be interesting to investigate in more detail.

Common pathogens are carefully controlled by vaccination in commercial production systems, so the biggest threats are likely to come from novel or significantly mutated pathogens. 'Generalist' haplotypes, which have been shown to be common in all commercial flocks, are predicted to confer a degree of resistance to most pathogens, but conservation of a pool of genetic resources in the wider chicken population, and an understanding of how this diversity is distributed, could be crucial if genetic resistance to a new pathogen is suddenly required. Furthermore, the relatively high frequency of 4-8-4:01_03-43 (which is predicted to confer a generalist phenotype based on low cell surface expression of MHC class I in homozygotes (E. Meziane and J. Kaufman, unpublished)) in Marek's disease-susceptible Silkies suggests that even MHC class I molecules with wide peptide binding repertoires could have 'gaps' in their repertoire which may preclude the presentation of key antigenic peptides from particular pathogens. This remains a hypothesis in the absence of infection or peptidomic studies of 4-8-4:01_03-43, but would be an interesting exception to the

correlation between cell surface expression and resistance to Marek's disease reported by Kaufman et al. (1995) and Chappell et al. (2015) if proven.

Differences between broilers and layers are also likely to be the result of intentional enrichment of specific traits in the different lines, as opposed to an unintentional deficiency in one production type relative to the other. Differences in immune responses between broilers and layers, admittedly not controlling for the effects of differences between breeding companies, have previously been reported and linked to the different lifespans of the two production types. Layers produced long-term IgG humoral responses and strong cellular responses to immunization with trinitrophenyl-conjugated keyhole limpet hemocyanin, while broilers produced an IgM-dominant response (Koenen et al., 2002). Having said that, the presence of 31-31-31-31 in broilers is likely to be the result of the inclusion of Plymouth Rock alongside Cornish birds in the original hybrids (Clauer, 2012a), and layers may not have had access to this haplotype due to their original breed makeup which was predominantly Rhode Island Red, New Hampshire Red and Leghorn (Clauer, 2012b). Breeding companies (such as B) with dual-purpose birds containing 31-31-31-13, however, may have had the opportunity to cross this haplotype into their layers if it was seen to confer an advantage.

Fancy breeds

Fancy breeds are anecdotally known to differ in their 'hardiness' or overall susceptibility to infectious disease but few formal comparative studies on specific pathogens have been conducted. There may be opportunities for the welfare of disease-susceptible breeds such as Silkies to be improved by higher frequencies of haplotypes known to confer resistance to key pathogens including Marek's disease; however if the low frequencies of resistant haplotypes detected in this project are representative of the population, resistance could increase naturally over the next few generations. Sufficient populations of unvaccinated 'backyard' chickens are likely to exist for natural selection to act on resistance traits in these breeds.

Free-range local

The greatest potential for this work to improve poultry management may relate to development initiatives which aim to support the establishment of productive local poultry industries in rural communities. The socio-economics of these initiatives is largely beyond the scope of this project, with debates ongoing with respect to the merits of a Brazilian-style contract-based system for local producers where large corporations provide birds, feed and medication to smallholders (FAO, 2018) vs. the local market model favoured by Bill Gates where smallholders have control of their own businesses (Gates, 2016). Nonetheless, key to the success of either of these models is increasing profitability, by improving feed conversion and survival rates without incurring high feed, veterinary or processing costs. Combining traits such as disease resistance and heat tolerance from

local breeds with the productivity of commercial production lines is likely to be the most effective solution, and such improvements have contributed significantly to the establishment of a highly efficient poultry industry in Brazil over the past few decades (Patricio et al., 2012). With MHC diversity directly influencing key production traits, understanding this topic has huge potential to facilitate the development of profitable chicken lines.

Newcastle disease is a major threat to local chicken flocks in Africa and while vaccines are effective they are not always physically or financially accessible. The standard B13 haplotype has been suggested to associate with stronger immune responses to NDV (Dunnington et al., 1992; Norup et al., 2011) and may be beneficial in African free-range local poultry systems. The PCR-NGS system is unable to distinguish between the standard B4 and B13 haplotypes, since their allele sequences are identical in the peptide-binding exons examined, but 4-4:01_02-4:01_03-4 (which refers to both both B4 and B13) was entirely absent in Tanzanian chickens and present at very low frequency in Ghanaian chickens. It may, therefore, be beneficial to breed this MHC haplotype into a local ecotype background to promote resistance. Care would need to be taken, however, not to offset the potential benefits with an associated loss of resistance to other local pathogens. Indeed, homozygote birds carrying 4-4:01_02-4:01_03-4 have relatively high cell surface class I (E. Meziane and J. Kaufman, unpublished) meaning this could possibly be an example of a NDV specialist allele. In the work presented here, (109)-?-23:02-38:02 and 4-8-4:01_03-24 were seen to associate with faster viral clearance and have been selected in free-range local poultry systems in at least one country. These haplotypes may, therefore, be a more appropriate choice for a breeding programme looking to improve NDV resistance in Africa and highlight the importance of considering the broader context when looking to apply insights developed under experimental conditions in the field.

The differences detected in this project between Tanzanian and Ghanaian chickens suggest that interventions may need to be country-specific. The prevalence of country-specific haplotypes may reflect distinct pathogen pressures which would need to be taken into account; for example 17-17-?-66:02 may be conferring resistance to a key pathogen in Tanzania which is not present in Ghana. If selection is shaping the diversity in this way, country-specific chicken lines will be necessary, and any new vaccines should be tested for efficacy with the local haplotypes. Alternatively, reproductive isolation and somewhat stochastic haplotype frequency dynamics could have resulted in the irreversible loss of this beneficial haplotype in Ghana, and re-introduction of the haplotype as part of a new line used in multiple countries could confer health benefits. Identification of key poultry pathogens in the two regions would be the first step to determining whether these differences are likely to be the results of selection or drift. Further infection trials using birds carrying haplotypes of interest, or immunopeptidomic analysis of the peptide binding repertoires would also help to

determine the resistance phenotypes associated with particular MHC haplotypes.

The commercial haplotypes seen in African flocks could have various origins. Analysis of Red Jungle Fowl diversity is a key next step in this project, since this will help to discern ancient haplotypes which have been maintained by selection in a range of habitats from haplotypes which are present as a result of introgression from poorly-adapted European birds. This would help to understand the extent to which birds optimised for western commercial production would need to be modified (at least in terms of the MHC) to be successful in alternative pathogen environments. Intermediate stages in the development of domestic chickens, such as free-range local breeds from Asia, would also help to reconstruct the historic relationships of populations and the haplotypes within them.

Somewhat reassuringly, different ecotypes within Tanzania appear to have limited differentiation at the MHC, which is possibly dependant on the presence of substantial trade links between regions. This suggests that as far as MHC-determined pathogen or vaccine responses are concerned, strategies applied on a whole-country scale have a good chance of being successful. The observations of high inter- and low intra-country MHC variation are consistent with trends in overall genetic diversity reported in the literature (tables 2.1, 2.2 and 2.3).

2.4.4 Integration of an expanded MHC database with microsatellite typing

Despite the relatively low cost (~£4/bird) of the PCR-NGS protocol described here, price or access to NGS equipment could still be prohibitive for researchers who could otherwise benefit from the expanded chicken MHC database. It is therefore intended that the database will be integrated with information on MHC-linked microsatellites, particularly the commonly-used LEI0258 (McConnell et al., 1999; Fulton et al., 2006) which is located between the BG and BF regions. This locus includes variable numbers of repeats of two sequences, 12 bp and 13 bp respectively in length. The length of these repeated sequences means that differences in the total lengths of different LEI0258 alleles are generally visible by agarose gel electrophoresis.

The applicability of LEI0258 typing to a given population of interest needs to be carefully assessed before it is used. Not all haplotypes can be distinguished by variation at this locus; for example, the standard B2/B15 and B13/B17 haplotypes share the same allele size. These can often be distinguished by additional analysis of the MCW0371 locus (Buitenhuis et al., 2003), but since alleles at this locus generally differ by 1 bp in length, they are more difficult to distinguish without specialist techniques such as capillary electrophoresis. It is therefore important to have a

basic understanding of the variation present in a population before performing microsatellite typing to ensure that all meaningful variation can be discerned. Highly diverse populations are more likely to contain multiple MHC haplotypes with the same LEI0258 allele length, and are therefore more difficult to accurately type without sequencing the MHC loci themselves. Furthermore, a linked marker will always be vulnerable to the effects of recombination and it is therefore possible that different populations or lines might have different associations between MHC haplotype and microsatellite length. Separate tables of microsatellite-MHC haplotype associations may, therefore, need to be produced for each common experimental or commercial line.

The same idea could be extended to SNPs, which are commonly used in genome-wide association studies (GWAS). For the human MHC, HLA alleles at specific loci can be imputed from intergenic flanking SNPs within the MHC region, and subsequently associated with immune phenotypes (Dilthey et al., 2011; Goyette et al., 2015). For the 'minimal essential' chicken MHC, SNP typing with gene-level resolution is extremely difficult due to the density of the polymorphism in the MHC, however SNPs flanking the entire region could be associated with haplotypes which could then be incorporated into GWAS studies in chickens (Fulton et al., 2016a; Psifidi et al., 2016).

Chapter 3

Diversity in class I antigen processing and presentation components in passerine birds

3.1 Introduction

3.1.1 Avian MHCs

While much is known about the chicken MHC, there are over 10,000 other avian species (Billerman et al., 2020) for which limited data is available. Being able to apply insights from chickens to other species would be extremely useful for conservationists and those monitoring avian diseases, however the most recent common ancestor of chickens and Neoaves (which includes all bird lineages other than the Palaeognathae (ostriches, emus and others) and Galloanserae (chickens, ducks and others)) is dated approximately 70 million years ago, allowing plenty of time for divergence of these lineages. Of particular note is the observation of a major radiation of birds following the Cretaceous–Palaeogene (K–Pg) mass extinction, which occurred after the separation of the Galloanserae (chickens, ducks and others) from the lineage that would become the Neoaves (Prum et al., 2015).

Many features of the chicken MHC are shared with other species in their order, the Galliformes. Turkeys, grouse and pheasants all have equivalents of the chicken B (classical) and Y (non-classical) loci, with BG genes also reported in turkey and pheasant (Jarvi et al., 1996; Bauer and Reed, 2011; Wang et al., 2012). Quails appear to have undergone several rounds of gene cassette duplication leading to more genes, including copy number variation (CNV) for class II B genes, but retain a similar structure and gene composition to the chicken as well as single dominantly expressed class I and IIB genes (Shiina et al., 1999, 2004). TAP polymorphism has not been specifically

investigated, despite the potential of peptide transport specificity to interact with class I peptide repertoire, especially in heterozygotes (Tregaskes et al., 2015).

Outside of the Galliformes, where research has historically been concentrated for reasons associated with commercial agriculture or sport, work on avian MHCs has typically focused on the assessment of CNV or allelic diversity of class I and II genes, and their associations with life history traits, mate choice or pathogen resistance, particularly to easily quantifiable parasites such as avian malaria. A large number of these studies are reviewed by Kaufman (2021), including a ten-year study of Seychelles warblers where class I diversity correlated with juvenile survival and one particular allele conferred a five-fold increase in lifespan (Brouwer et al., 2010). As should be expected when investigating highly polygenic life history and disease resistance traits, the observed effects of MHC diversity vary significantly between populations and studies. For example, neither class I genotype nor diversity was found to correlate with survival in the Raso lark (Stervander et al., 2020). MHC diversity has also been implicated in mate choice and sexual selection, although the mechanisms by which these genetic characteristics are detected remains controversial. Hypotheses have included olfactory signals, with blue petrels exhibiting MHC class II-dissasortative mating in the wild (Strandh et al., 2012) and males preferentially approaching the odour of a more MHC-dissimilar female in a Y-maze (Leclaire et al., 2017).

There appears to be a high degree of evolutionary flexibility in classical and non-classical class I and class II copy number, with different lineages expanding particular groups of genes but not others. A general trend appears to be that passerine birds have more class I and II genes than non-passerines (Minias et al., 2019; He et al., 2020), although there is significant variation in copy number between passerine species (O'Connor et al., 2016; Minias et al., 2019), and in the majority of cases there is still no distinction made between classical and non-classical, or well and poorly expressed genes. Where this distinction has been investigated, further variation appears; a single dominantly expressed classical class I gene has been reported in sparrows (Drews et al., 2017) and zebra finch (Balakrishnan et al., 2010) but siskins appear to have multiple highly expressed class I genes, at least at the RNA level (Drews and Westerdahl, 2019). Class I allele phylogenies are similarly variable between species; house sparrows have been reported to have a distinct cluster of putative non-classical class I alleles, defined by a 6 bp deletion and low nucleotide diversity (Bonneaud et al., 2004; Karlsson and Westerdahl, 2013), while great tits appear to have multiple clusters of monomorphic class I sequences (without a deletion) distributed throughout an otherwise highly diverse phylogeny (Sepil et al., 2012). These differences suggest variation in the 'ages' of non-classical genes in different species, which can make them extremely difficult to identify from sequence data alone. The differences also indicate potential roles for genetic drift and/or divergent selection acting on the genetic structures of different species' immune systems, indicating a high

degree of evolutionary flexibility.

Balakrishnan et al. (2010) reported that MHC class I and TAP genes did not co-localise to a single chromosome in zebra finch, apparently precluding the possibility of TAP-MHCI co-evolution and raising questions about the mechanism by which a single classical class I is maintained. TAPs which are not in linkage disequilibrium with class I should be required to be widely permissive, which might be expected to allow the expansion of a multigene family of class I genes, allowing the presentation of more pathogenic peptides which should be an advantageous trait (up to a theoretical optimum number of loci). Evidence that zebra finch TAPs were minimally polymorphic and therefore likely to be permissive (as in humans) came from Ekblom et al. (2011) who were unable to find polymorphisms in the TAP genes when they attempted to design a SNP mapping assay.

There is very little other published work on TAP genes specifically in any avian species, although a few studies have looked at the genetic structure of MHC regions in other non-passerine birds and identified the TAP genes within genomic regions that resemble the chicken MHC. In particular, the class I and II regions appear to be relatively conserved between Galliformes and the crested ibis (Chen et al., 2015) and Oriental stock (Tsuji et al., 2017), although tapasin is conspicuously absent. With long-read sequencing data increasingly being used for genome assemblies, it is becoming easier to look for features such as physical linkage between genes of interest, but further improvements in chromosome level assembly will be needed to fully understand the scope of interspecific variation in the architecture of the MHC region, and the underlying evolutionary processes (He et al., 2020). However, genomics alone will not be sufficient to understand the extent to which the antigen presentation system varies functionally among birds. Even the limited existing data suggests that these systems can be highly flexible, and identification of one 'chicken-type' or 'human-type' feature may not necessarily imply the presence of others. Furthermore, the adaptive function and relative expression of various genes is extremely difficult to infer (beyond the identification of transcription factor binding sites, for example) from sequence alone.

While class I has been investigated to some extent in several species, the TAP genes, which can play a key role in shaping the peptide repertoire, have rarely been discussed. This chapter assesses the genomic architecture of the MHC in passerine birds in order to determine the extent to which the chicken-type MHC is conserved across birds, and therefore the extent to which insights from chickens can be generalised to other species.

3.1.2 Structure and function of the TAP heterodimer

The TAP transporter is a heterodimer of the TAP1 and TAP2 monomers, each of which comprise a series of transmembrane helices and a C-terminal nucleotide binding domain (NBD). The most conserved part, both between the loci and between species, is the nucleotide binding domain, specifically motifs such as the Walker A and B sequences, C signature motif and Q loop which characterise the NBDs of ABC transporters.

Upstream of the NBD, each gene encodes six transmembrane helices which make up a 'core' domain, resulting in a 12-helix bundle in the heterodimer. Four additional helices (absent in TAP1 of birds, see Walker et al. (2005)) at the N-terminal are involved in membrane targeting and interactions with other components of the peptide loading complex (PLC) such as tapasin (Hulpke et al., 2012; Hinz et al., 2014).

The 11-exon structure of both TAP1 and TAP2 seems to be relatively conserved across gnathostomes, although in chickens (and at least some related birds) exon 1 of TAP1 is truncated to remove the tapasin binding region and TAP2 is reduced to 9 exons by fusion of the exons equivalent to human exons 1/2 and 7/8 (Beck et al., 1992; Walker et al., 2005).

As discussed in section 1.2.3, the TAP heterodimer primarily functions as part of the PLC to transport antigenic peptides into the ER for loading onto class I molecules. It can also play a role in cross-presentation of exogenous peptides on class I by transporting peptides from the cytosol into endosomes (Burgdorf et al., 2008), however a key part of this process, the prior export of endocytosed antigens to the cytosol, remains poorly understood (Gros and Amigorena, 2019).

3.1.3 Specific aims

Here, comparisons are made between features of the chicken MHC and those observed in other birds, particularly passerines. Three defining features of the chicken MHC, namely a single dominantly expressed class I gene, polymorphic TAP genes and linkage between the TAPs and the dominantly expressed class I, are used as a framework against which non-model species can be compared. Few of these features can be effectively assessed with just genome assemblies, even with high-quality, long read data. Assessing polymorphism requires multiple individuals from a population to be sequenced, which is still uncommon in whole genome sequencing of non-model species. Expression is very difficult to assess from sequence data alone, with promoter regions potentially able to identify pseudogenes but rarely able to distinguish low expressers from high expressers. Finally, physical distance and recombinational distance are not always proportional, and while physical linkage can be observed in genomes (assuming the assembly is correct, which is

not always the case, particularly for MHC regions), this does not necessarily imply that the genes in question have the potential to co-evolve.

The aims of this project were therefore to assess as many of these three features as possible in several passerine species.

3.2 Materials and Methods

3.2.1 Samples

European Pied Flycatcher (*Ficedula hypoleuca*)

DNA samples from 18 pied flycatcher chicks were provided by Dr. Sally Rogers (University of Gloucestershire, now University of Exeter). The birds were part of a constant effort ringing programme in woods around Hertfordshire, UK, and had been found dead in their nest boxes. Sample names include nest box number followed by chick number, with chicks found in the same nest box assumed to be siblings (although extra-pair matings are not uncommon in pied flycatchers). Cause of death was unknown and the chicks were frozen at −20 °C on discovery. Dr. Sally Rogers extracted heart tissue and isolated genomic DNA using the PureLink Genomic DNA kit (Invitrogen). After initial PCR tests were unsuccessful, all samples were cleaned up using GeneReleaser (BioVentures) according to the manufacturer's protocol prior to use in any amplification reactions for sequencing.

In addition, a single DNA sample was provided from a chick that had died in the hand while being ringed by Dr. David Canal (Doñana Biological Station, Spain). This bird had been immediately frozen, and genomic DNA extracted in the same way as for the UK samples.

House Sparrow (*Passer domesticus*)

Blood from three house sparrows was received from Dr. Gabriele Sorci (Université de Bourgogne) in September 2010 and stored at −70 °C. Genomic DNA was extracted for this project using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma-Aldrich) according to the manufacturer's instructions.

A further five house sparrow cDNA samples were obtained from the lab of Dr. Helena Westerdahl (University of Lund). These samples were taken from the same individuals for which MHC class I was investigated in Drews et al. (2017). Blood was collected in autumn 2012 from birds caught in mist nets in Löberöd, Skåne, Sweden. cDNA was synthesised during a visit to the University of Lund using the RETROscript kit (Life Technologies) from existing RNA samples which

had previously been extracted by Anna Drews using a combination of the TRIzol LS protocol (Life Technologies) and the RNeasy Mini kit (QIAGEN).

Zebra Finch (*Taeniopygia guttata*)

Muscle and spleen samples from 24 zebra finch individuals, stored in ethanol or RNAlater (Sigma Aldrich) respectively at −80 °C, were provided by Dr. Michal Vinkler (Charles University, Prague) in December 2018. The birds were all adult males, bought randomly from multiple hobby breeders, and represent an approximately random sample of the domestic population. Half the birds had been stimulated with LPS injection for a different study, and all were culled and dissected in November 2018. DNA and RNA were extracted during a visit to Charles University using the DNeasy 96 Blood and Tissue kit (Qiagen) and High Pure RNA Tissue Kit (Roche) respectively, according to the manufacturer's instructions. cDNA was produced from RNA using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems).

3.2.2 Primer design and screening

TAP primers were initially designed to provide coverage over as large a proportion of the gene as possible. Specific regions of interest were those that aligned to the regions on human TAP1 and TAP2 reported to be involved in peptide binding (Nijenhuis and Hammerling, 1996) or to the residues on rat TAP2 reported to be involved in determining peptide specificity (Momburg et al., 1996; Deverson et al., 1998). PCRs were specifically optimised later in the project to cover these regions if they had not already been covered by PCRs that were successful in initial screens. The locations of these sequence features are indicated in figures 3.1 and 3.2.

Primers were designed manually, avoiding runs of more than three identical nucleotides or regions of particularly high GC-richness, and requiring a 3' GC clamp. T_m was checked using the NEB T_m calculator and the primer length adjusted if necessary to ensure consistent annealing temperatures. The OligoEvaluator tool (Sigma-Aldrich) was used to identify primers with secondary structure or the potential to form primer dimers. In the majority of cases primers with these characteristics were not synthesised or screened. In a few cases primers with theoretically non-optimal properties were screened if there were few appropriate alternatives that would provide coverage of a particular sequence.

Sparrow

Primers for amplifying sparrow TAP genes from genomic DNA were designed based on conserved regions in alignments of all available passerine TAP sequences, since there was no sparrow genome available at the time (table 3.1). All primers were designed with a T_m of approximately 68 °C so

that most pairs would be compatible and could be screened in a high-throughput manner using a single thermocycling protocol (SP_TP2_1; details in appendix A.2). Primer combinations that were incompatible because the forward primer binding site was downstream of the reverse were not screened.

TAP1 NCBI reference	TAP2 NCBI reference	Species	Common name
	XM_017742928.1	*Corvus brachyrhynchos*	American crow
	XM_005431465.2	*Geospiza fortis*	Medium ground finch
	XM_017839388.1	*Lepidothrix coronata*	Blue-crowned manakin
XM_021549745.1	XM_021549744.1	*Lonchura striata domestica*	Bengalese finch
	XM_018069957.1	*Manacus vitellinus*	Golden-collared manakin
XM_015616525.1	XM_019005499.1	*Parus major*	Great tit
XM_005533821.2	XM_014262316.1	*Pseudopodoces humilis*	Tibetan ground tit
	XM_018925596.1	*Serinus canaria*	Atlantic canary
	XM_014894048.1	*Sturnus vulgaris*	Common starling
XM_005062210.2	XM_016305372.1	*Ficedula albicollis*	Collared flycatcher

Table 3.1: **Passerine TAP1 and TAP2 sequences used for primer design.** Sequences of passerine TAPs were obtained by BLAST with chicken TAP1 and TAP2 sequences against the passerine genome database. Conserved regions were identified and used to design universal passerine primers initially for use with sparrow gDNA. Several of these sequences have since been updated in light of new genome assemblies. At the time of the search, more predicted protein coding regions corresponding to TAP2 than TAP1 had been annotated in passerine genomes; fewer sequences were therefore included in the alignment for TAP1 on the assumption that a more detailed analysis of unannotated genomic regions could be performed to obtain more sequences if the initial primers designed were unsuccessful.

For TAP1, the initial screen included 14 forward and 15 reverse primers in 65 combinations. For TAP2, the initial screen included 14 forward and 13 reverse primers in 55 combinations. The primer combinations giving the longest amplicons of the expected size (based on predictions made using the alignments of passerine TAP genes) were selected and the specificity verified by cloning and Sanger sequencing.

Once a number of short amplicons had been generated it became clear that longer amplicons which would link these short fragments and allow polymorphisms to be phased into longer allelic sequences would be important. In many cases, longer fragments (amplified using different combinations from the same primer pool) had been amplified in the initial screen but were either amplifying with low efficiency or with additional non-specific products. Primer pairs which amplified long fragments over specific regions of interest but required optimisation were identified and the PCR improved by optimisation of buffer conditions (HF (high-fidelity) or GC (optimised for GC-rich templates)), annealing temperature (including the use of touchdown protocols), extension time or number of PCR cycles.

Later primers for amplification of sparrow TAP genes from cDNA were designed based on a PacBio house sparrow genome provided by Dr. Mark Ravinet (University of Nottingham). Scaffold 2491 was identified by BLAST using the sparrow TAP sequences that had previously been amplified and sequenced from genomic DNA. Primers were designed to cover as much of the total coding sequence as possible, with intron-exon boundaries predicted by alignment with LOC101809503 and LOC101809296 (flycatcher TAP1 and TAP2 genomic sequences) and by analysis of canonical splice sites in regions where the flycatcher sequences were of poor quality.

Flycatcher

Since the primers that had successfully been used to amplify fragments of TAP genes in sparrows were designed such that they should be universal for passerines, those that had been used successfully in sparrows were tested on flycatcher samples. Where these were unsuccessful in producing an amplicon of the expected length, primers were designed specifically for flycatchers using the *Ficedula albicollis* genome, FicAlb1.5 (GCA_000247815.2) which has annotated open reading frames LOC101809503 and LOC101809296 corresponding to TAP1 and TAP2 respectively, although they are not named as such. All primers were designed with a T_m of approximately 68 °C to allow flycatcher specific primers to be combined with universal passerine primers.

Zebra Finch

TAP1 and 2 sequences were inferred by alignment of the annotated flycatcher TAP genes against a region of the genome Tgut_diploid_1.0 (GCA_002008985.2) that had been identified using the flycatcher TAPs as queries in a BLAST search. Since cDNA was available, making it possible to cover many more exons in a single amplicon, primers were designed in the inferred exons 1 and 8 for TAP1 and 1, 8 and 9 for TAP2. The primers were screened in multiple combinations using a touchdown thermocycling protocol (ZF_CDNA).

The single primer pair for each of TAP1 and TAP2 which gave the longest amplicon with no non-specific amplification in zebra finch was re-synthesised as a set of 10 5'-barcoded primer pairs such that TAP1 primer pairs had barcodes 1-10 and TAP2 primers had barcodes 11-20. Barcodes were 8 nucleotides in length and were designed such that each was at least 3 nucleotides different from all the others. Zebra finch samples Tagu5 - Tagu9 were each assigned two barcodes for TAP1 and two barcodes for TAP2 (table 3.2). Barcodes 1, 6, 11 and 16 were originally assigned to Tagu4, however it was observed that the TAP2 PCR gave very low yields when used to amplify from Tagu4, so the barcodes were reassigned to Tagu9.

TAP1			TAP2		
Barcode number	Sequence	Sample	Barcode number	Sequence	Sample
1	CAAGACTA	Tagu9	11	GATAGACT	Tagu9
2	TGGAACAT	Tagu5	12	GCCACATA	Tagu5
3	GCTAACGT	Tagu6	13	AACGTGAT	Tagu6
4	GTCTGTCA	Tagu7	14	GAACAGGC	Tagu7
5	GAATCTGT	Tagu8	15	CCGACAAG	Tagu8
6	CCGTGAGA	Tagu9	16	ATAGCGAC	Tagu9
7	GGTGCGAT	Tagu5	17	AGATGTAG	Tagu5
8	AAGGTACA	Tagu6	18	CCATCCTC	Tagu6
9	AACGCTTA	Tagu7	19	ATCATTCG	Tagu7
10	TATCAGCA	Tagu8	20	AAGGACAC	Tagu8

Table 3.2: **Barcode sequences and sample assignments for Nanopore library construction.**

Primers were also designed to amplify exons 2 and 3 from MHC class I in zebra finch. An alignment was made of the putative single functional MHC class I gene in the zebra finch genome (ENSTGUG00000017273, Balakrishnan et al. (2010)) and BAC clone TGAC-102M22 (AC232985) which was found in a BLAST search using ENSTGUG00000017273 as the query. The exons were highly conserved in these two sequences, although the introns varied in both length and sequence. Since amplification would be performed on cDNA, uncertainty as to the intron sequences was not problematic. Two forward and two reverse primers were designed to cover the variable peptide binding domains of MHC class I and were screened in all four possible combinations (thermocycling protocol: ZF_MHCI). All four combinations successfully amplified fragments of the expected length. Two pairs which did not share either of their component primers were selected and used to amplify from Tagu5 - Tagu9. The use of two independent primer pairs for amplification of a potentially highly polymorphic locus reduced the probability that amplification of particular allele sequences would fail entirely due to the presence of SNPs in the primer binding sites.

3.2.3 PCR

Unless otherwise specified, all PCRs were performed using the Phusion Hot Start Flex DNA Polymerase kit (NEB). The final concentrations of reagents were 1x reaction buffer (HF or GC), 200 µM dNTPs, 0.5 µM each primer and 0.2 units/10 µl reaction Phusion polymerase. Approximately 10-100 ng DNA was added per 10 µl reaction volume. Initial PCRs to screen primers were performed in 10 µl total reaction volume and PCRs for sequencing were performed in 20-50 µl total reaction volume depending on the PCR efficiency, number of cycles, clean-up required (with or without gel extraction) and sequencing method. Reaction volumes larger than 10 µl were divided into 10 µl aliquots prior to thermocycling.

Thermocycling was performed in the Tetrad 2 Peltier Thermal Cycler (BioRad) and T100 Thermocycler (BioRad).

Amplification of passerine TAP1

Table 3.3 and figure 3.1 describe the amplicons generated from TAP1 genes in three passerine species.

Sample	Fragment	Length	Forward primer	Reverse primer	Thermocycling protocol	Buffer
Flycatcher gDNA	F2	725	TAP1_5F	TAP1_7aR	SP_TP2_1	HF
	F3	1524	TAP1_8F	TAP1_10R	SP_TP2_1	HF
	FGf	885	TAP1_F6_Fihy	TAP1_R8_Fihy	SP_TP2_1	HF
	FLf	1590	TAP1_2aF	TAP1_R5_Fihy_a	TP1_R5M1	GC
	SS1*	822	T1_E5FP2	T1_MIDRP2	TCD_SS1	GC
Sparrow gDNA	F1	493	TAP1_1F	TAP1_2R	SP_TP2_1	HF
	F2	725	TAP1_5F	TAP1_7aR	SP_TP2_1	HF
	F3	1524	TAP1_8F	TAP1_10R	SP_TP2_1	HF
	FL	2195	TAP1_2aF	TAP1_R5_mod1	TP1_R5M1	GC
	FG	1056	TAP1_F6_mod1	TAP1_R8_mod2	TD-T1FG	GC
	FH	2729	TAP1_1F	TAP1_7aR	SP_LONG	GC
Sparrow cDNA	Lund_1	1593	Pado_TAP1_ex1b_F	Pado_TAP1_ENDb_R	HS_CDNA	GC
Zebra Finch cDNA	Nanopore_1	964	ZF_TAP1_F2	ZF_TAP1_R3	ZF_CDNA	GC

Table 3.3: **Passerine TAP1 amplicons.** Primer sequences and details of thermocycling protocols can be found in appendices A.1 and A.2. The primers and thermocycling for SS1* were designed and optimised by Jimmy Xu (University of Cambridge), under supervision. Amplicon lengths are excluding primers.

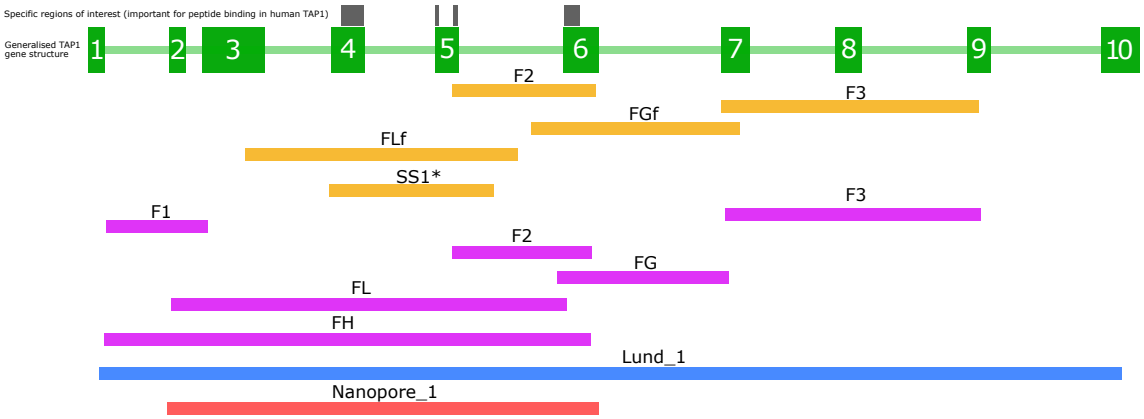


Figure 3.1: **TAP1 regions amplified by primer pairs described in table 3.3.** Orange, purple, blue and red bars represent fragments amplified from pied flycatcher gDNA, house sparrow gDNA, house sparrow cDNA and zebra finch cDNA respectively. Fragments are shown against the passerine intron-exon structure (exons as numbered blocks) based on sparrow, flycatcher and zebra finch TAP1 identified from genome sequences and annotations (collared flycatcher: LOC101809503 in FicAlb1.5 (NCBI Eukaryotic Genome Annotation Pipeline); zebra finch: ENSTGUG00000015337 in bTaeGut1_v1.p (Ensembl genebuild); house sparrow: genomic sequence provided by Dr. Mark Ravinet) and sequences obtained from gDNA and cDNA during this project. The structure reflects that of human and chicken TAP1, with the exception that exons 3 and 4 are fused into a single exon (3) in the passerine structure resulting in one fewer exons total and human TAP1 has a retained tapasin binding domain at the start of exon 1. The zebra finch genome bTaeGut1_v1.p shows a much longer intron between exons 6 and 7 which contains highly repetitive sequence elements. Specific regions and residues of interest are indicated above the gene structure as grey bars (regions which align to sequences important for peptide binding in human TAP1)

Amplification of passerine TAP2

Table 3.4 and figure 3.2 describe the amplicons generated from TAP2 genes in three passerine species.

Sample	Fragment	Length	Forward primer	Reverse primer	Thermocycling protocol	Buffer
Flycatcher gDNA	F1f	1010	TAP2_F4_Fihy_b	TAP2_6R	SP_TP2_1	HF
	F2	1550	TAP2_6F	TAP2_11R	SP_TP2_1	HF
	L1f	2498	TAP2_F4_Fihy_b	TAP2_11R	SP_LONG	GC
Sparrow gDNA	F1	829	TAP2_4aF	TAP2_6R	SP_TP2_1	HF
	F2	1657	TAP2_6F	TAP2_11R	SP_TP2_1	HF
	FS	983	TAP2_2F	TAP2_4R	SP_TP2_G	HF
	L1	3226	TAP2_2F	TAP2_11R	SP_LONG	GC
Sparrow cDNA	Lund_2.2	1107	TAP2_2F	TAP2_11R	SP_TP2_1	HF
	Lund_2.4	859	TAP2_4aF	TAP2_11R	SP_TPTD	HF
Zebra Finch cDNA	Nanopore_2	1500	ZF_TAP2_F5	ZF_TAP2_R1	ZF_CDNA	GC

Table 3.4: **Passerine TAP2 amplicons.** Primer sequences and details of thermocycling protocols can be found in appendices A.1 and A.2. Figure 3.2 shows the regions of the gene amplified by these primer pairs.

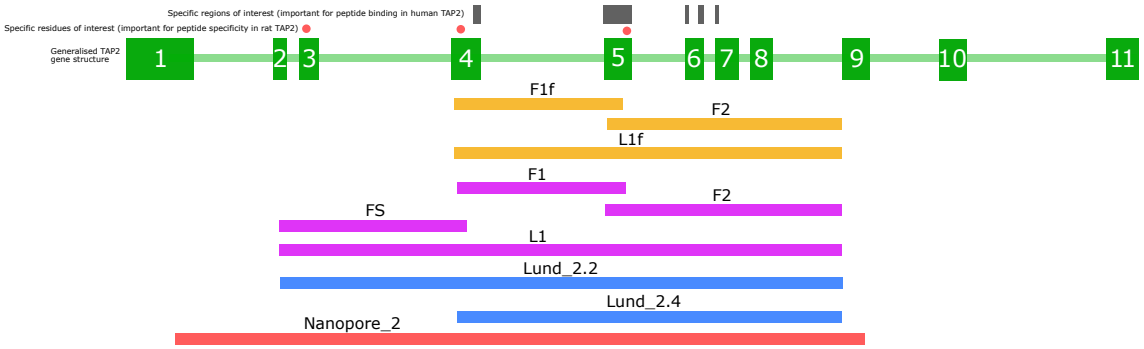


Figure 3.2: **TAP2 regions amplified by primer pairs described in table 3.4.** Orange, purple, blue and red bars represent fragments amplified from pied flycatcher gDNA, house sparrow gDNA, house sparrow cDNA and zebra finch cDNA respectively. Fragments are shown against a generalised intron-exon structure (exons as numbered blocks) based on conserved features of sparrow, flycatcher and zebra finch TAP2 identified from genome sequences and annotations (collared flycatcher: LOC101809296 in FicAlb1.5 (NCBI Eukaryotic Genome Annotation Pipeline); zebra finch: ENSTGUG00000020195 in bTaeGut1_v1.p (Ensembl genebuild); house sparrow: genomic sequence provided by Dr. Mark Ravinet) and sequences obtained from gDNA and cDNA during this project. The structure reflects that of human TAP2, but differs from chicken TAP2 as described by (Walker et al., 2005). Specific regions and residues of interest are indicated above the gene structure as grey bars (regions which align to sequences important for peptide binding in human TAP2) or red circles (residues which align to those important for peptide specificity in rat TAP2 [three circles indicate positions equivalent to residues 217/218, 262/265/266 and 374/380 in the rat TAP2 sequence]).

Amplification of Zebra Finch MHC class I

Table 3.5 describes the exon 2-exon 3 amplicons generated from MHC class I genes in zebra finches.

Sample	Fragment	Length	Forward primer	Reverse primer	Thermocycling protocol	Buffer
Zebra Finch cDNA	MHCI_pp1	544	ZF_ex2_F	ZF_ex4_R	ZF_MHCI	HF
	MHCI_pp2	481	ZF_ex2_F2	ZF_ex3_R	ZF_MHCI	HF

Table 3.5: **Zebra Finch MHC class I amplicons.** Primer sequences and details of thermocycling protocols can be found in appendices A.1 and A.2.

3.2.4 Cloning

PCR products were separated by gel electrophoresis on an agarose gel (1% agarose w/v in TAE buffer (1 mM EDTA, 40 mM Tris, 20 mM acetic acid)) with 1:10000 SYBRsafe (Invitrogen). Bands were visualised under UV light and the target band excised. Amplicons were purified from the excised gel fragment according to the QIAquick Gel Extraction kit protocol, using EconoSpin silica membrane mini spin columns (Epoch Life Science) and eluted in nuclease free water. The reaction to amplify zebra finch MHCI was sufficiently specific that the PCR product could be cleaned up directly according to the QIAquick PCR purification protocol using EconoSpin silica membrane mini spin columns (Epoch Life Science) and eluted in nuclease free water.

Purified amplicons were ligated into pJET1.2 blunt ended cloning vector using the CloneJET PCR cloning kit (ThermoFisher) according to the manufacturer's instructions. Chemically competent NEB5α cells were transformed with the vector and grown up overnight at 37 °C on LB-agar plates containing 100 µg/ml ampicillin.

Colonies were screened for the correct insert in a 'colony PCR' using primers specific to the vector either side of the multiple cloning site (MCS) (pJET_pcr_F/pJET_pcr_R; sequences in appendix A.1) and the COLONY thermocycling protocol (appendix A.2). Colonies positive for the expected insert size were grown at 37 °C overnight in a shaking incubator in 3 ml LB medium with 100 µg/ml ampicillin. Plasmid DNA was extracted by alkaline lysis (Birnboim and Doly, 1979) and purified on EconoSpin silica membrane mini spin columns (Epoch life sciences).

3.2.5 Sanger sequencing

Sanger sequencing was performed by Eurofins Genomics in early experiments and Genewiz in later experiments. Sequencing primers T7 and pJET1.2_rev, which are specific to the vector sequence either side of the MCS, were used as standard. In addition, internal sequencing primers were provided in cases where the insert exceeded 1.5kb in length, or where sequencing results were of poor quality and required additional primers to cover the entire length.

3.2.6 Oxford Nanopore sequencing

Library construction

PCRs were carried out as 3 x 10 µl reactions per barcode using the Phusion Hot Start Flex DNA Polymerase kit as previously detailed. Targets were amplified using a touchdown thermocycling protocol (ZF_MHCI).

Amplicons were purified by gel extraction and pooled in approximately equimolar amounts (80 ng purified DNA from each 1.1 kb TAP1 amplicon and 120 ng purified DNA from each 1.6 kb TAP2 amplicon). The pooled library was concentrated to 50 ng µl^{-1} by ethanol precipitation.

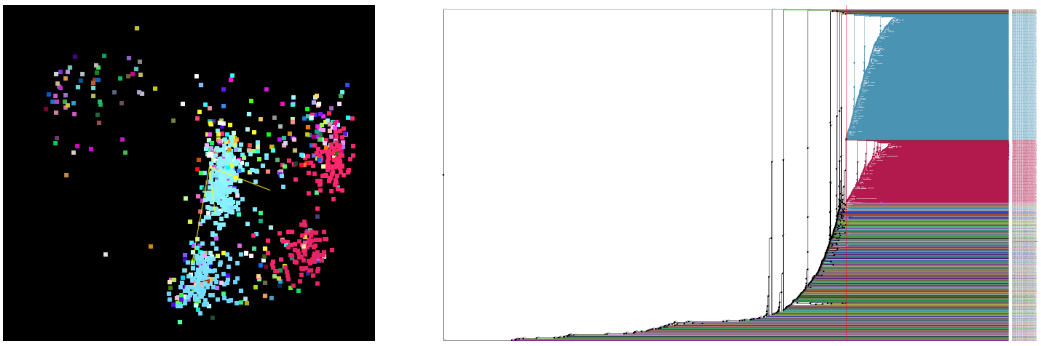
Ligation of adapters and sequencing

The library was prepared for sequencing using the Oxford Nanopore Ligation Sequencing kit, which includes DNA end repair and dA-tailing using the NEBNext End Repair/dA-tailing module, followed by ligation of sequencing adapters. The library was sequenced for 49 h on an Oxford Nanopore MinION, with additional library added after 20 h 35 min. Adapter ligation and sequencing was performed in collaboration with Dr. Anton Enright (University of Cambridge, Cambridge Genomic Services).

Data analysis

Initial data processing, including base-calling, de-multiplexing and filtering was performed by Dr. Anton Enright. The output was a fasta file for each barcode, containing all the sequences assigned to that barcode (allowing 2 mismatches between the reference barcode and the sequence, but requiring the same barcode to be identified on both ends of the sequence), with the primers and any non-amplicon sequence trimmed off.

The first 1000 sequences in each barcode file were aligned using MUSCLE (Edgar et al., 2004). From this alignment, subsets of related sequences within the file were identified using Principle Component Analysis (PCA) and Neighbour-Joining/Average Distance phylogenetic trees in Jalview 2.11.0 (Waterhouse et al., 2009) (figure 3.3).



(a) PCA

(b) Average distance tree

Figure 3.3: **Representative example of PCA and Average Distance phylogeny analysis of nanopore reads from barcode 03.** The blue and pink clusters represent the two alleles present in this individual. Each PCA cluster has two lobes because the 'forward' and 'reverse' reads cluster slightly separately. The blue and pink sequences would then be extracted as the two subsets for this barcode.

The subsets were extracted from the alignment and a 45% consensus was drawn from each using modules from Biopython (Cock et al., 2009). Each consensus was compared to the reference sequence (ENSTGUT00000015946 and ENSTGUT00000043148 for TAP1 and TAP2 respectively). Indels were identified and, if determined to be artefacts of the alignment, resolved manually by examination of the original alignment. Results from replicates were then compared to each other and disagreements examined manually at the level of the original alignment. Errors in the consensus primarily occurred due to inaccuracies in Nanopore sequencing of homopolymeric sequences, poor alignment of genuine polymorphisms and poor handling of 'noise' in the sequencing data by the alignment algorithm.

Since this was the first time this protocol had been used, the final sequences were verified by cloning and Sanger sequencing of the original amplicons that had been used in the construction of the Nanopore library (see sections 1.14 and 3.2.4). This allowed the accuracy of the Oxford

Nanopore consensus-based protocol to be compared to the accuracy of Sanger sequencing, which is higher per read but produces a single sequence.

3.2.7 Polymorphism analysis and data visualisation

In addition to the software and packages already described in relation to specific methodologies, Geneious Prime versions 11 to 2021.1 were used for analysis of Sanger sequencing chromatograms and for basic alignments of DNA and protein sequences (https://www.geneious.com). BLAST algorithms (Altschul et al., 1990) were accessed through the NCBI BLAST suite (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`). Pairwise K_a/K_s analysis was performed in DnaSP6 (Rozas et al., 2017). Results and phylogenetic trees were visualised in R (R core team, 2013) using ggplot2 (Wickham, 2016) and ggtree (Yu et al., 2017). Protein alignments were producing using Clustal Omega (Sievers et al., 2011) and structural models were visualised in PyMol 2.3.4 (Schrodinger LLC, 2019). Analysis of positively selected sites was performed using CODEML from the PAML package (Yang et al., 2005; Yang, 2007). The model of the human TAP heterodimer used here was produced by Megan O'Mara (University of Queensland) and Rachael Gaudet (Harvard University) and is based on a bacterial ABC transporter.

3.3 Results

3.3.1 TAP1 and TAP2 were sequenced from multiple individuals from three passerine species

Flycatcher gDNA

PCRs were optimised using the high-quality genomic DNA extracted from the Spanish sample (ESP). The quality of the UK genomic DNA received was poor and amplification was only successfully achieved from two UK samples (TP2-5 and MK70-5, figure 3.4), even after performing a GeneReleaser clean up on all samples and testing various dilutions of some samples. The two UK samples are from different woods and therefore all three birds are assumed to be unrelated.

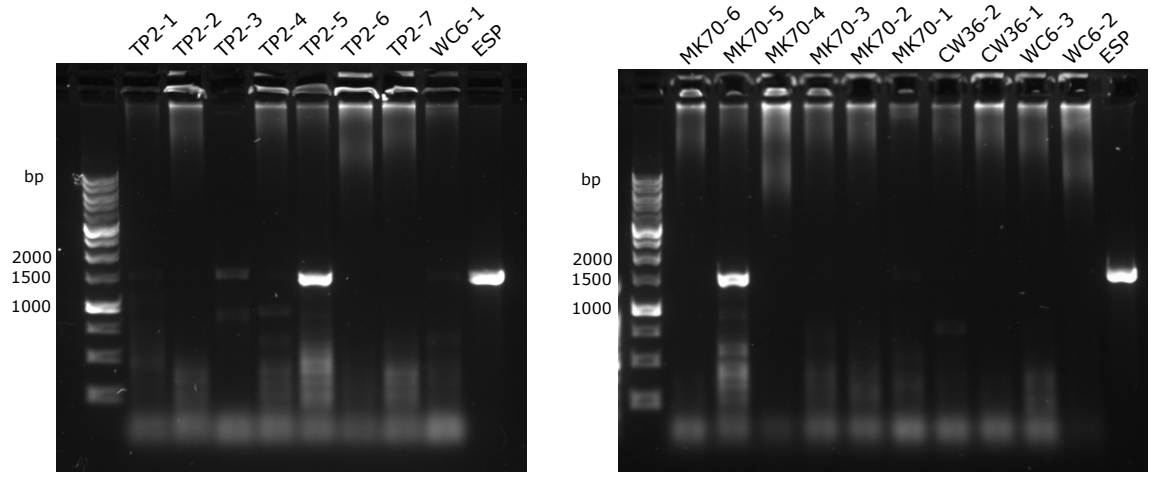


Figure 3.4: **Identification of TP2-5 and MK70-5 as suitable PCR templates.** After GeneReleaser cleanup all UK flycatcher samples were tested in the TAP1 F3 PCR which had been optimised and confirmed by sequencing in the ESP sample (amplicon size = 1524bp excluding primers). Further testing, using different PCRs and sample dilutions, confirmed that TP2-5 and MK70-5 were the only samples amenable to amplification without significant additional processing.

A summary of the amplicons sequenced from the three flycatcher samples is provided in table 3.6. Sequences from individual clones are summarised in appendices A.6.1 and A.6.2.

		ESP	TP2-5	MK70-5
TAP1	F2	5	3	6
	F3	5	4	6
	FGf	3	4	3
	FLf	2	0	0
	SS1*	0	5	6
TAP2	F1f	7	4	5
	F2	4	0	1
	L1f	8+3	5	7+3

Table 3.6: **Number of nucleotide sequences obtained from amplicons from flycatcher genomic DNA.** Partial sequences, where the majority of the sequence was obtained from an amplicon but sequencing quality precluded a full sequence being determined, are included after a '+' sign. SS1* amplicons were obtained by Jimmy Xu under supervision.

The Sanger sequencing of gene sequences as multiple overlapping fragments provided an overall indication of the level of polymorphism and the specific locations of some polymorphic residues. However, since the genes were shown to have low levels of polymorphism, there were rarely SNPs in the overlapping regions between fragments which could be used to phase the SNPs in separate amplicon sequences into complete allele sequences. Furthermore, without sequencing each fragment in significantly more depth, it was impossible to tell whether all possible polymorphism had been observed for each fragment. That is, the absence of polymorphism in a fragment could indicate

that the two alleles in a heterozygotic individual were identical in this region, or that only one allelic sequence had been obtained.

Sparrow gDNA

A summary of the amplicons sequenced from the three Sorci sparrow samples is provided in table 3.7. Sequences from individual clones are summarised in appendices A.6.3 and A.6.4.

		Pado1	Pado2	Pado3
TAP1	F1	3	5	5
	F2	3	5	4
	F3	3	5	5
	FL	2	2	3
	FG	0	0	5
	FH	1+7	2+5	1+6
TAP2	F1	3	5	6
	F2	5	5	5
	FS	7	6	4
	L1	7+1	8+2	9+3

Table 3.7: **Number of nucleotide sequences obtained from amplicons from house sparrow genomic DNA.** Partial sequences, where the majority of the sequence was obtained from an amplicon but sequencing quality precluded a full sequence being determined, are included after a '+' sign.

Longer fragments were successfully amplified from sparrow gDNA than flycatcher gDNA for both TAP1 and TAP2. This allowed SNPs to be phased, with shorter amplicons providing support for observed polymorphisms. Longer amplicons also allowed a more confident assessment of the level of polymorphism present in the genes; regions identical between alleles could be confirmed as such based on polymorphisms upstream or downstream, rather than potentially representing fragments where only one allele had been sequenced.

This experiment was still limited by difficulties in cloning and sequencing long fragments. Since intronic variation was not of interest in this study, cDNA was subsequently obtained to facilitate the cost-effective and efficient sequencing of several contiguous exons.

Sparrow cDNA

		HS1	HS2	HS3	HS4	HS5
TAP1	Lund_1	5	3+2	3+1	3+3	6+2
TAP2	Lund_2.2	4	4+1	1	7	6+1
	Lund_2.4	13+1	10+1	9+1	9	10

Table 3.8: **Number of nucleotide sequences obtained from amplicons from house sparrow cDNA.** Partial sequences, where the majority of the sequence was obtained from an amplicon but sequencing quality precluded a full sequence being determined, are included after a '+' sign.

Amplification from cDNA allowed significantly more efficient sequencing of coding sequences which facilitated the inclusion of more samples (HS1-HS5). The cloning efficiency of the Lund_1 amplicon was relatively low, precluding a complete assessment of TAP1 polymorphism in the available time. The influence of cloning efficiency on the quality of these results highlighted the potential for high-throughput sequencing methods to improve this study. Nonetheless, coverage of TAP1 (appendix A.6.5) was sufficient to make generalised comments about the density of SNPs. The shorter Lund_2.4 amplicon amplified and sequenced well, allowing allele sequences over the majority of the regions of specific interest to be confidently determined. Two major sequences were obtained from TAP2 in HS1, HS2 and HS3, while only one was obtained from each of HS4 and HS5, indicating that these birds were potential homozygotes. While TAP1, which is predicted to co-segregate with TAP2 due to physical proximity of the genes, was not sequenced in sufficient depth to fully confirm whether these birds were homo- or heterozygous for the TAPs, there was no evidence that any birds seen to carry one allele at the TAP2 locus carried two at the TAP1 locus. Additional low-frequency sequences seen in TAP2 could largely be explained by PCR misincorporations, random sequencing errors or PCR chimerism events.

The intron between TAP2 exons 7 and 8 was retained in seven clones: three from a single allele in HS4 and four from a different allele shared by HS1, HS2 and HS5. In all cases, the allele was also seen without the intron in the same birds (appendix A.6.6).

Zebra Finch cDNA

A cost assessment indicated that high-throughput sequencing using Oxford Nanopore technology was as, if not more, cost-effective than Sanger sequencing for this study and would overcome limitations associated with cloning and sequencing long amplicons. The Nanopore protocol was cost-effective even when sequencing just five birds and taking into account the cost of synthesising 40 barcoded primers. Future experiments would likely use a larger sample set and re-use the primers, reducing the cost per bird even further.

The Oxford Nanopore MinION run produced 10,823,997 reads, of which 9,737,578 were high-quality 'pass' reads. Both 'pass' and 'fail' reads were included in downstream analysis since failed reads are often flagged unnecessarily, and downstream filtering for the presence of matching barcode and primer sequences should be sufficiently stringent as quality control (Albert Kang, personal communication). The bimodal distribution of read lengths reflected the two amplicon lengths that were included in the library, and the overall read density was split evenly between the two amplicon lengths (figure 3.5).

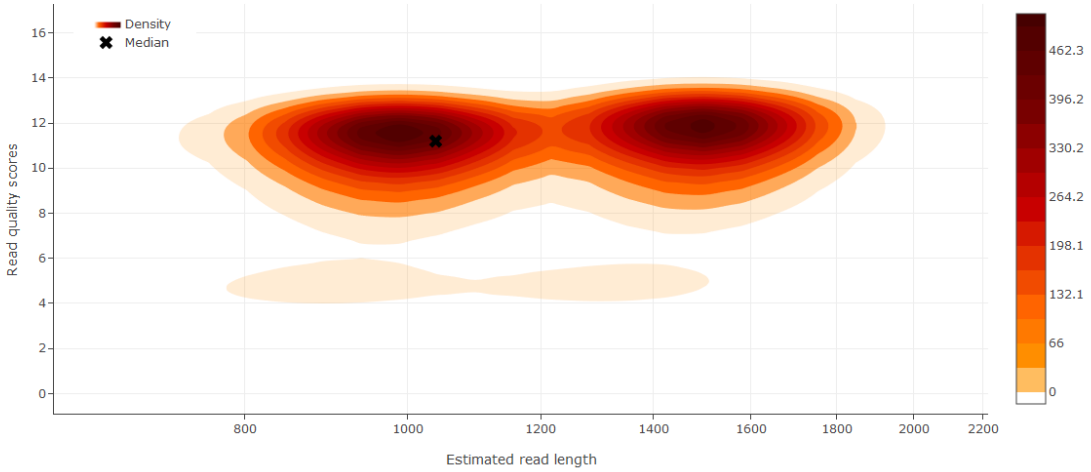


Figure 3.5: **Distribution of Nanopore read density, length and quality.** 'Pass' and 'fail' reads are not discriminated because all reads were used in the downstream filtering and demultiplexing process. Figure produced using pycoQC (Leger and Leonardi, 2019)

Between one and four apparent allele sequences were obtained per gene per bird. A subset of the PCR products used to construct the Nanopore library were subsequently cloned for Sanger sequencing and in all cases the Sanger results corroborated the Nanopore sequences. The results of both the Nanopore and Sanger sequencing are discussed in detail in section 3.3.7, with allele sequences available in appendix A.8.

3.3.2 Comparison of amino acid-level polymorphism in human, chicken and passerine TAPs

The patterns of polymorphism in each passerine species are discussed in detail in subsequent sections. The alignment of four out of the seven residues known to influence peptide specificity in rats with variable residues in chickens (which are also known to have functionally variable TAPs) in figure 3.7 is of note, since these two species share a significantly more distant common ancestor that chickens and passerines. As such, variation at these positions might be expected in passerine birds, should they have a chicken-type MHC. The locations of the residues identified in rats, which

do not all coincide with the regions identified in humans, demonstrate the complexity of functional variation in the TAP heterodimer and highlight the need to consider variation across as much of the genes as possible.

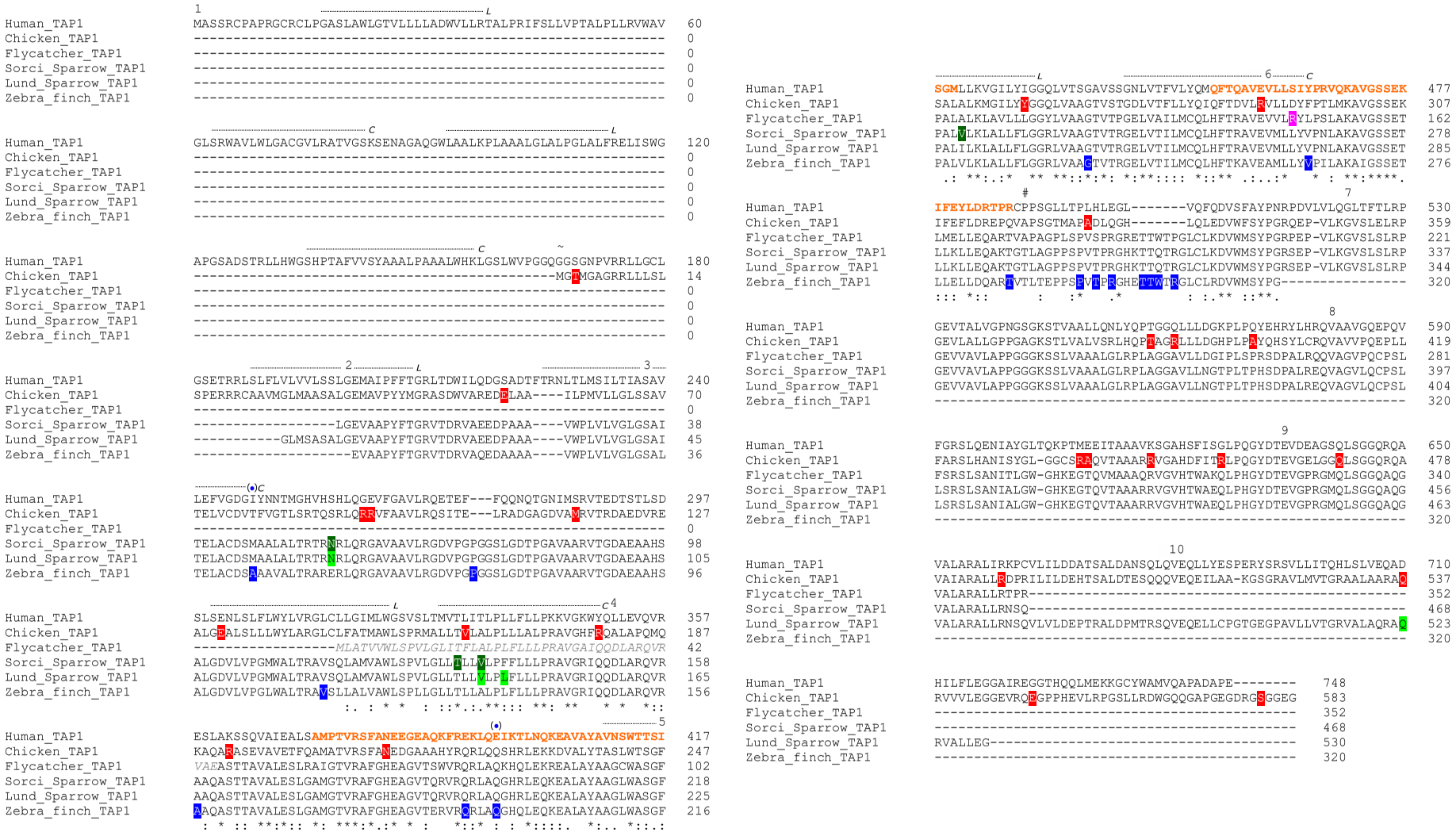


Figure 3.6: **Alignment of human, chicken and passerine TAP1 sequences.** Full caption overleaf.

Figure 3.6 caption cont.: Human (NCBI ref. NP_000584.3) and chicken (translated from the B2 haplotype sequence AB426141) TAP1 protein sequences were aligned to sequences obtained during this project using Clustal Omega. Intron/exon boundaries of passerine TAP1 (which appear to be conserved between the three passerine species) are indicated by numbers above the alignment over the first amino acid of the exon or the amino acid with the split codon. In human and chicken TAP1, the region equivalent to passerine exon 3 is present as two separate exons, resulting in one additional exon overall. A tilde (~) above the alignment indicates the last residue of the tapasin binding domain and a hash (#) indicates the first residue of the nucleotide binding domain. Dotted lines above the alignment indicate transmembrane regions, with 'C' or 'L' indicating subsequent cytoplasmic or lumenal loops according to Koch et al. (2004). Possible positively selected sites (despite a non-significant likelihood ratio test) in zebra finch TAP1 are indicated with a blue bullet in brackets above the alignment. Symbols below the alignment indicate conservation; residues conserved between all sequences are shown with an asterisk (*), residues where strongly similar amino acid properties are conserved are shown with a colon (:) and residues where weakly similar amino acid properties are conserved are shown with a dot (.). Regions shown to be important for peptide binding in humans (Nijenhuis and Hammerling, 1996) are indicated in bold orange text. Polymorphic residues in chicken, flycatcher, sparrow and zebra finch are highlighted in red, magenta, green and blue respectively, with the Sorci and Lund sparrow populations analysed separately and shown in dark and pale green respectively. Chicken SNPs were based on an alignment of the 14 unique TAP1 sequences from the standard serological haplotypes: B2, B4/B13, B5, B6, B8/B11, B9, B12, B15, B17, B19, B21, B23 and B24 from Hosomichi et al. (2008) and B14 from Walker et al. (2011) (GenBank accession numbers AB426141-AB426154 and JF794480). SNPs identified in Sanger sequencing results had to be present in at least two clones. The Sorci sparrow region in grey italic text was only covered by sequencing from two clones which was considered insufficient to draw conclusions.

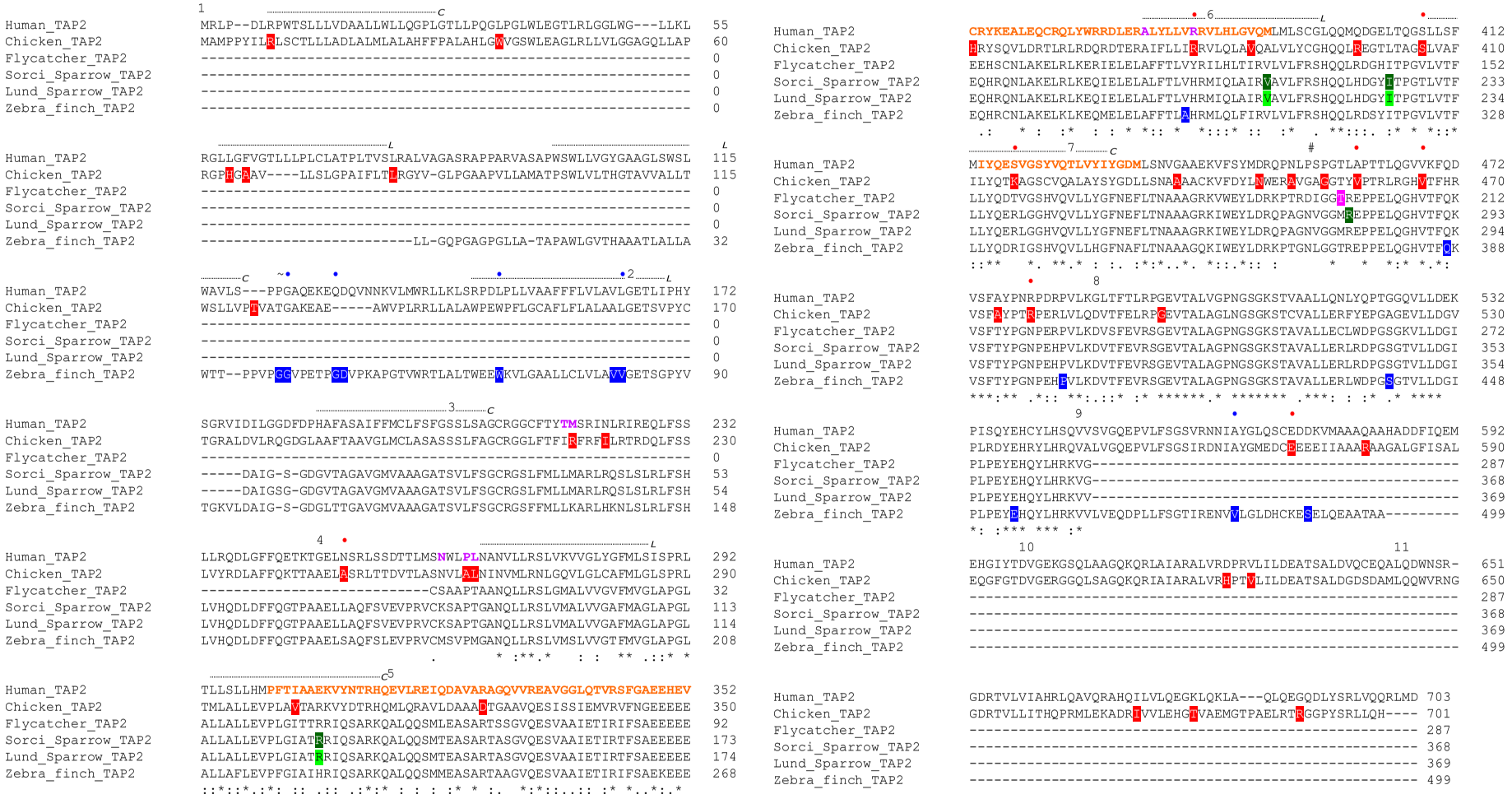


Figure 3.7: **Alignment of human, chicken and passerine TAP2 sequences.** Full caption overleaf.

Figure 3.7 caption cont.: Human (NCBI ref. NP_000535.3) and chicken (translated from the B2 haplotype sequence AB426141) TAP2 protein sequences were aligned to sequences obtained during this project using Clustal Omega. Intron/exon boundaries of passerine TAP2 are indicated by numbers above the alignment over the first amino acid of the exon or the amino acid with the split codon. Human TAP2 has the same intron/exon structure, while chicken TAP2 differs in the ways described by Walker et al. (2005), namely fusion of exons 1/2 and 7/8. A tilde (~) above the alignment indicates the last residue of the tapasin binding domain and a hash (#) indicates the first residue of the nucleotide binding domain. Dotted lines above the alignment indicate transmembrane regions, with 'C' or 'L' indicating subsequent cytoplasmic or lumenal loops according to Koch et al. (2004). Red and blue bullets above the alignment indicate positively selected sites (PSS) predicted at $p < 0.05$ in chicken and zebra finch respectively, analysed over the range of the Nanopore_2 amplicon. Symbols below the alignment indicate conservation; residues conserved between all sequences are shown with an asterisk (*), residues where strongly similar amino acid properties are conserved are shown with a colon (:) and residues where weakly similar amino acid properties are conserved are shown with a dot (.). Regions shown to be important for peptide binding in humans (Nijenhuis and Hammerling, 1996) are indicated in bold orange text and residues which align to those known to be important for determining peptide specificity in rat TAP2 (Deverson et al., 1998; Momburg et al., 1996) are shown in bold magenta text. Polymorphic residues in chicken, flycatcher, sparrow and zebra finch are highlighted in red, magenta, green and blue respectively, with the Sorci and Lund sparrow populations analysed separately and shown in dark and pale green respectively. Chicken SNPs were based on an alignment of the 15 unique TAP2 alleles from the standard serological haplotypes: B2, B5, B6, B8/B11, B9, B12, B13, B15, B17, B19, B21, B23, and B24 from Hosomichi et al. (2008) and B4 and B14 from Walker et al. (2011) (GenBank accession numbers AB426141-AB426154, JF794485 and JF794487). SNPs identified in Sanger sequencing results had to be present in at least two clones. The SNP at position 279 in the Sorci sparrows was observed in two clones from independent PCRs but in both cases is inconsistent with the two likely allele sequences in the birds (based on observation of two high-frequency sequences in each bird which do no include this SNP; see appendix A.6.4) so is likely to be artefactual.

3.3.3 Pied flycatchers have low TAP polymorphism

The preliminary sequencing of TAP1 and TAP2 in three pied flycatcher individuals aimed to provide an initial qualitative assessment of polymorphism in these genes in a passerine species. The genes were sequenced as multiple overlapping fragments but the depth of sequencing from each fragment was insufficient to fully determine the polymorphisms present.

The sequenced amplicons (excluding the region covered by just two FLf clones indicated by grey italic text in figure 3.6) covered the majority of passerine exons 4-9, including all regions that aligned to sequences known to be important for peptide binding in human TAP1 (Nijenhuis and Hammerling, 1996). A single polymorphic residue was found in each of TAP1 and TAP2, with three additional synonymous changes per gene in protein coding regions (appendices A.6.1 and A.6.2). The variable residue in TAP1 is within the predicted peptide binding region (by homology with the human sequence; figure 3.6) and it is therefore plausible that it could interact with bound peptides. The single variable residue detected in flycatcher TAP2 is situated at the start of the nucleotide binding domain (figure 3.7), where it is unlikely to play a role in peptide transport specificity.

While SNPs could not be reliably phased from overlapping amplicons into contiguous allele sequences, there was some evidence of heterozygosity at both loci in all three birds. Thus, the low polymorphism observed in these genes does not seem to be attributable to low allelic diversity or inbreeding.

3.3.4 Collared flycatchers may contain TAPs in linkage disequilibrium with a class I locus

The *F. albicollis* genome was used as a model for *F. hypoleuca* throughout, since the two species are from the same genus. The TAP genes in *F. albicollis* sit on an unplaced scaffold of approximately 180 Kb with a class I locus approximately 51 Kb away (figure 3.8). A tenascin X-like gene is also present, which is found in the class III region of the chicken MHC. No other typical MHC-region genes are currently annotated on this scaffold. The physical distance between the TAPs and class I gene on this scaffold, which is here considered a rough proxy for recombinational distance, is significantly more than in chickens, where BF2 and TAP2 are separated by just 2 Kb (GRCg6a chicken genome assembly (GCF_000002315.6)), but less than in rats, where TAP1 and RT1-A3 are separated by 179 Kb but are still able to coevolve to some extent (Joly et al., 1996).

Two other class I loci (figure 3.9) were identified in the *F. albicollis* genome by BLAST search with XP_016160862.1 (the protein encoded by LOC101808190, the class I sequence on scaffold

NW_004775961.1) and BF2. The three protein sequences were aligned and eight residues identified by Kaufman et al. (1994) as being conserved in classical MHC class I molecules throughout vertebrates were examined (table 3.9). These residues are involved in binding the ends of peptides; an arginine at position 84 in chickens and other non-mammals (in contrast to the tyrosine at the same position in mammals) allows peptides to overhang the class I binding groove in these species (Xiao et al., 2018).

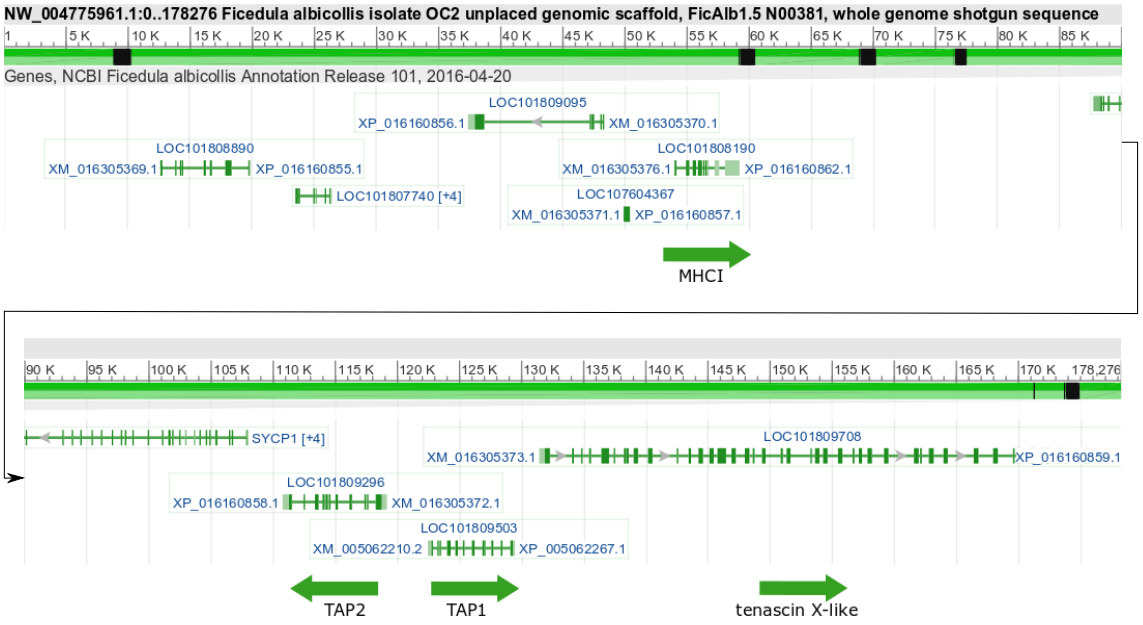


Figure 3.8: **Scaffold NW_004775961.1 from the assembly FicAlb1.5 (GCF_000247815.1).** *F. albicollis* has a class I locus approximately 51 Kb away from its TAP genes.

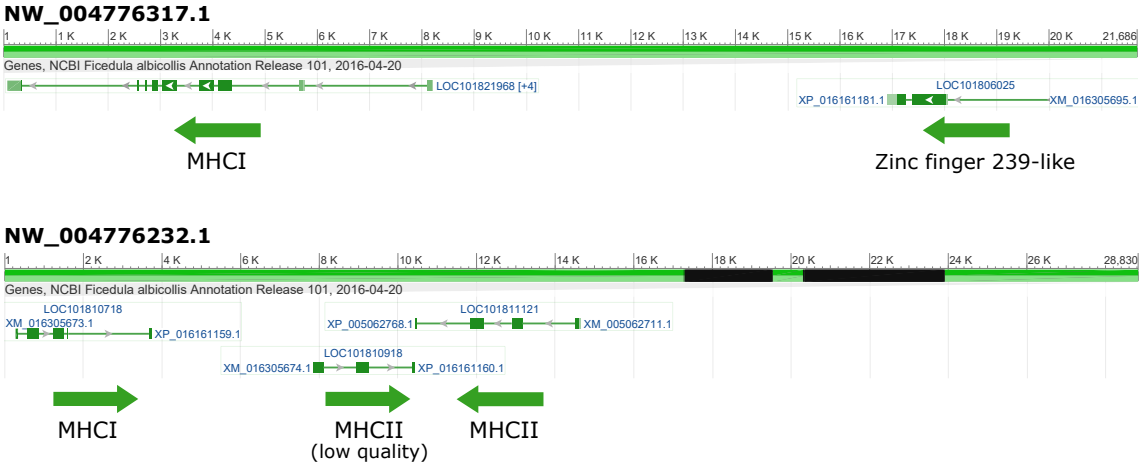


Figure 3.9: **Scaffolds NW_004776317.1 and NW_004776232.1 from the assembly FicAlb1.5 (GCF_000247815.1).** Two additional MHC class I loci in *F. albicollis* were identified, one on a scaffold with MHC class IIβ genes. The gene labelled 'low quality' is annotated as a coding sequence but the sequence of the model RefSeq protein was modified (1 bp insertion) relative to this genomic sequence to represent the inferred CDS, so this may be a pseudogene.

Protein ref.	Genes nearby	Conserved residues							
		Y7	Y59	Y/R84	T143	K146	W147	Y159	Y171
XP_016160862.1	TAP1, TAP2	C	Y	R	T	R	W	Y	Y
XP_016161180.1	-	Y	Y	R	T	R	W	Y	Y
XP_016161159.1	MHC class IIB	-	-	V	T	R	W	Y	Y

Table 3.9: **Presence of conserved residues in MHC class I sequences in *F. albicollis*.** Positions are numbered according to HLA-A2 after Kaufman et al. (1994). Position 84 is a conserved tyrosine in mammals but is an arginine in non-mammals. XP_016161159.1 is truncated by the end of the scaffold so the first two positions could not be assessed. 'Genes nearby' indicates MHC-related genes on the same scaffold as the class I sequence. All three sequences appear on unplaced genomic scaffolds.

None of the three sequences carry a lysine residue equivalent to K146, which might suggest that none of these are classical class I molecules. However, the single class I gene identified in the zebra finch genome (Balakrishnan et al. (2010) and confirmed by BLASTP against the up-to-date NCBI zebra finch database) also carries an arginine at this position, suggesting that this could be 'permitted' in passerine classical class I sequences, despite being an indicator of non-classical sequences in the chicken, where R146 is common in YF genes. Putative classical sequences obtained by Drews et al. (2017) also had R146 (or occasionally C146), with putative non-classicals characterised by H146. Only XP_016161180.1, which appears on a short 21 Kb scaffold with a zinc finger protein 239-like locus has all the other expected residues. Despite their proximity to other MHC-related genes, there is therefore little evidence that either XP_016160862.1 or XP_016161159.1 are classical class I molecules.

3.3.5 House sparrows have low TAP polymorphism

Sorci population

Preliminary sequencing of three house sparrows from Dr. Gabriele Sorci revealed four potential variable amino acid residues in each of TAP1 and TAP2. One variable residue in TAP2 (position 279 in the Sorci_Sparrow sequence in figure 3.7) was seen in two clones from independent PCRs but was not consistent with the two likely allele sequences (based on observation of two high-frequency sequences in each bird which did not include this SNP) in each of the two birds in which it appeared, so may not represent a true polymorphism. Unlike in the flycatcher, where synonymous nucleotide variation in coding regions was more common than non-synonymous variation, only one synonymous nucleotide variant was found in TAP1 in the Sorci sparrows. TAP2 had three synonymous nucleotide substitutions (appendices A.6.3 and A.6.4).

Pado1 and Pado2 were clearly heterozygous at both loci, although complete allele sequences could not be fully constructed. In Pado3 TAP2, 26/27 clones appeared to be amplicons from a single

allele sequence, suggesting that this could be a homozygous locus, although TAP1 in this bird was clearly heterozygous.

In TAP1, none of the observed polymorphisms were in regions predicted to be important for peptide binding, however two of the TAP2 polymorphisms were in the likely peptide-binding sequences. None of the polymorphisms aligned with residues known to contribute to peptide specificity in the rat (figures 3.6 and 3.7). There was therefore little evidence for functional TAP polymorphism in this population.

Lund population

In contrast to the preliminary studies on the flycatchers and Sorci sparrows, TAP genes in the Lund population were sequenced from more individuals and as continuous coding sequence amplified from cDNA (as opposed to overlapping fragments amplified from gDNA). This facilitated more quantitative analysis of polymorphism, specifically in TAP2, where the depth of sequencing was sufficient to define complete allele sequences across the length of the amplicon with relative confidence.

Five unique allele sequences (over the range of the Lund_2.4 amplicon) were obtained from TAP2 in the Lund sparrows (table 3.10 and appendix A.6.6). There was evidence of alleles being shared between individuals and some indication that individuals might be homozygous at this locus, which together suggest relatively low allelic polymorphism (the number of alleles present in a population). While there is no directly comparable data from chicken, we can assume that TAP allelic polymorphism in an equivalent wild population of chickens would reflect BF2 allelic polymorphism, which would be predicted to be high.

Bird	TAP2 alleles present
HS1	A, D
HS2	B, C, E*
HS3	A, C
HS4	A
HS5	B

Table 3.10: **TAP2 alleles present in each Lund sparrow sample.** Allele E* may or may not be artefactual.

HS2 appeared to contain three well-supported sequences; allele E* was obtained from four clones (compared to seven and five clones for alleles B and C respectively) and contained only SNPs which were seen in either B nor C, suggesting that it could plausibly be an artefactual PCR chimera. However, E* appeared in the products of three independent PCRs. PCR chimeras do not seem

to be a major factor in the PCR products from any of the other samples, which have very similar TAP sequences, so it seems unlikely that there is a region in one or both of the other alleles in HS2 which is particularly difficult for the polymerase to extend through, and therefore particularly prone to the formation of chimeras. It is therefore difficult to understand how E* could appear so consistently if it was artefactual. Since E* could not be confidently characterised as either real or artefactual, subsequent analysis of sequence diversity was repeated both with and without allele E*.

The average numbers of synonymous changes per synonymous site (K_s) and non-synomymous changes per non-synonymous site (K_a) were averaged over all possible pairwise comparisons of allele sequences in both sparrow and chicken, with all nucleotide sequences from both species trimmed such that their translations covered the same region of the protein alignment. Of the fifteen unique chicken TAP2 alleles (unique nucleotide sequences within coding regions) from the standard serological haplotypes: B2, B5, B6, B8/B11, B9, B12, B13, B15, B17, B19, B21, B23, and B24 from Hosomichi et al. (2008) and B4 and B14 from Walker et al. (2011) (GenBank accession numbers AB426141-AB426154, JF794485 and JF794487) there were 13 unique sequences over this range. B4 and B5 were excluded since they were identical to B13 and B8/B11 respectively. Since the sparrow alleles could also differ outside the sequenced region, it would have been inappropriate to retain these duplicated sequences in the chicken dataset. The results of the analysis describe variation in sequence diversity (differences between alleles) and can be considered alongside the description of allelic polymorphism to obtain a fuller picture of overall diversity at this locus. Averaging over all pairwise comparisons corrected for the variable numbers of sequences available from the two species, although the power of the analysis increases with additional sequences.

	Unique alleles	**Total coding SNPs**	**Total non-coding SNPs**	**Mean K_s**	**Mean K_a**
Chicken	13	20	22	0.0257	0.0103
Lund (inc. E)	5	3	2	0.00525	0.00224
Lund (exc. E)	4	3	2	0.00583	0.00267

Table 3.11: **Sequence diversity of Chicken and Sparrow TAP2 alleles.** The average numbers of synonymous changes per synonymous site (K_s) and non-synomymous changes per synonymous site (K_a) were averaged over all possible pairwise comparisons of allele sequences in both sparrow and chicken to give 'Mean K_s' and 'Mean K_a' which represent the average number of differences between any pair of randomly selected alleles in that species. This strategy accounts for differences in the numbers of alleles included from each species when calculating a measure of diversity. All nucleotide sequences from both species were trimmed such that their translations covered the same region of the protein alignment.

On average, a pair of chicken alleles had 3.9-4.6 times as many non-synonymous differences per non-synonymous site as a pair of sparrow alleles (Mean K_a=0.00224 or 0.00267 for sparrow and Mean K_a=0.0103 for chicken), depending on whether or not sparrow allele E was included in the analysis (table 3.11). Chicken alleles therefore contain noticeably more protein-level sequence diversity, which has the potential to result in functional diversity, than sparrow alleles, even once

the smaller number of alleles included in the analysis is taken into account.

The sequencing of TAP1 was insufficiently deep to perform a similar analysis on this locus, however the preliminary results are sufficient to conclude that TAP1 also contains little polymorphism and is extremely unlikely to contain sufficient functional diversity to confer variable peptide transport specificity on the heterodimer as a whole (figure 3.6 and appendix A.6.5).

The locations of variable residues in the TAP heterodimer were predicted based on a model of the human TAP heterodimer (figure 3.10). Of all the residues identified as variable in the sparrow populations, only one in TAP2 was on an internal surface which would be likely to directly interact with a peptide binding in the transmembrane regions of the heterodimer. Furthermore, a cluster of rat and chicken variable residues was identified in TAP2 which might represent an important region for the functional diversity observed in the TAPs of these species. No residues in this region were seen to be variable in sparrows. The presence of this cluster of rat and chicken variable residues outside the regions predicted to be peptide binding by homology with the human sequence further highlights the complexity of peptide binding and specificity and the need to consider all potential sources of functional variation. While this structure is only a model, and care should be taken when generalising between species, the analysis provides no evidence to support functional variation in sparrow TAPs.

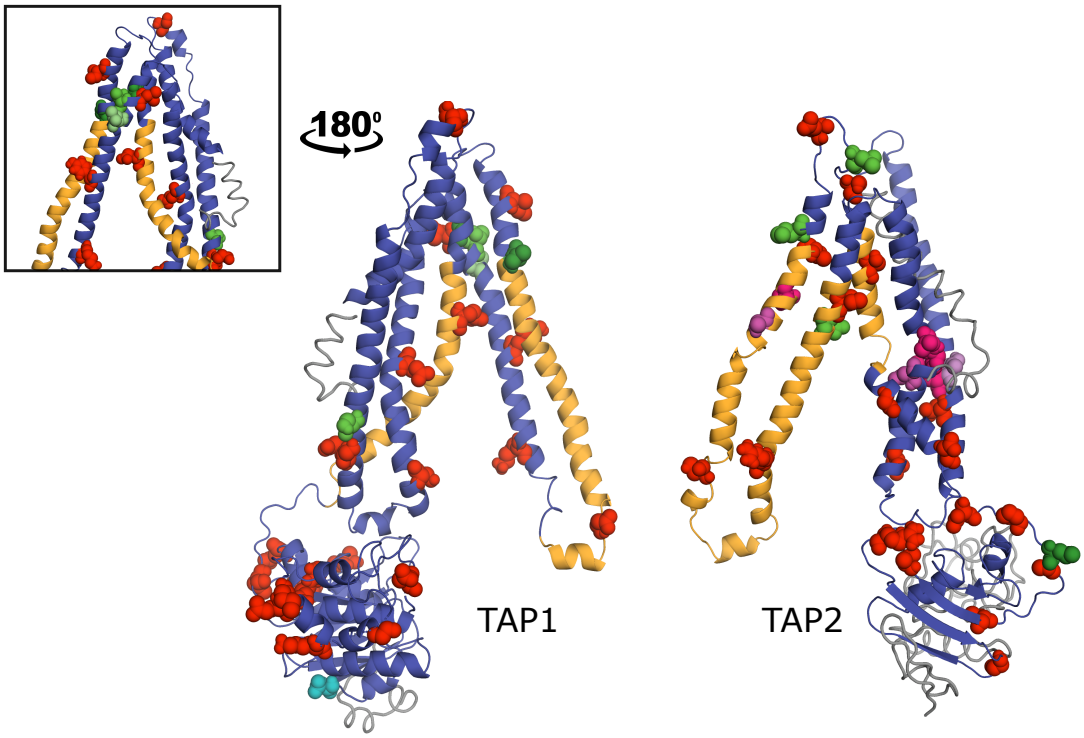


Figure 3.10: **Locations of variable residues in chicken and sparrow TAP1 and TAP2 on structural models of human TAPs**. Residues shown in red and green spheres indicate variable amino acids in chicken and sparrow respectively. Light, dark and mid green distinguish residues variable in the Sorci, Lund or both populations respectively. One TAP1 residue which is variable in both chickens and the Lund sparrow population is shown in teal. Residues known to affect peptide binding specificity in rat TAP2 are shown in purple spheres; the subset of these residues which are also variable in chickens are shown in pink. The human peptide binding domain is indicated in orange and the region not covered by the amplicon is shown as grey ribbon. Figure produced in PyMOL2.

3.3.6 House sparrows have a single dominantly expressed classical MHC class I gene but there is little evidence for TAP-MHCI linkage

MHC class I exon 3 alleles were amplified and sequenced from cDNA from the three Sorci sparrows previously described and an additional two samples from the same population by members of the Kaufman lab (K. Chen, J. M. Sanchez, D. B. Li, H. Siddle and C. Tregaskes, unpublished; sequences available in appendix A.7). Analysis of these sequences as a combined dataset showed 1-2 major allele sequences per bird, consistent with a single dominantly expressed class I gene (figure 3.11). Sequences obtained at low frequency can largely be explained as PCR misincorporations or sequencing errors, since they generally cluster phylogenetically around a major sequence from the same bird. Sequences W, V and k, which are low frequency but do not cluster with a major sequence in Pado1 are missing the conserved residue R146 and cannot be explained by PCR chimerism, suggesting that they could represent a genuine non-classical class I sequence. Other sequences seen multiple times which are not related to a major sequence (such as J) could be alleles

from low-expressing classical loci, consistent with the findings of Drews et al. (2017).

Pado1 and Pado2 appear to share a highly expressed class I allele but without being able to accurately phase SNPs in the TAP data, it is difficult to determine whether shared class I sequences are associated with shared TAP sequences in this data. SNPs shared between Pado1 and Pado2 are present in both TAP1 and TAP2 so there is currently insufficient data to refute the hypothesis that TAPs and class I are co-segregating in the Sorci sparrow population.

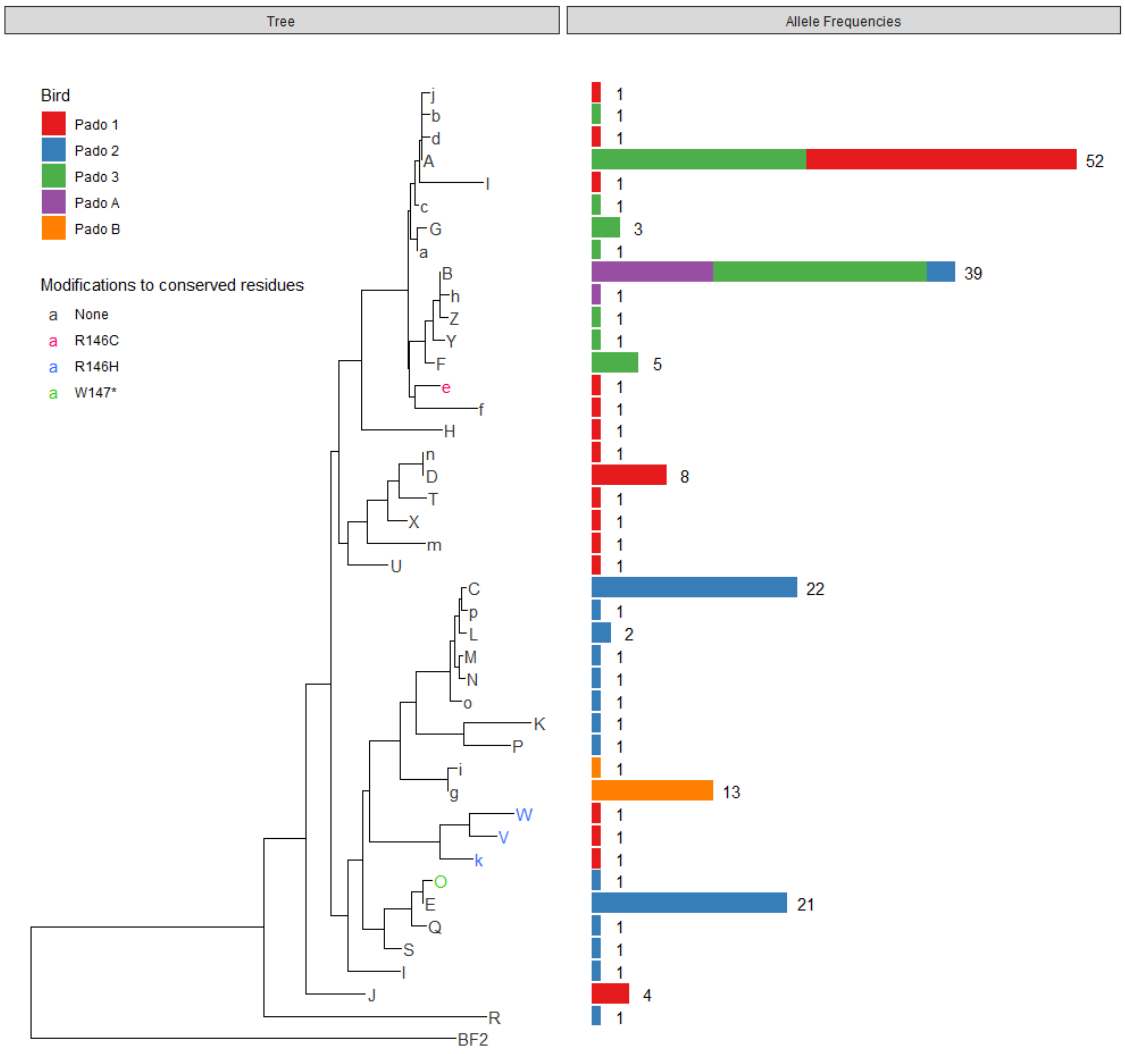


Figure 3.11: **Class I exon 3 sequences amplified from five Sorci sparrow cDNA samples.** Each lettered branch represents a unique sequence obtained from amplification of cDNA. Horizontal bars indicate the number of clones from which each sequence was obtained, with more clones indicating higher expression at the RNA level. The translations of the sequences were examined for the presence of classical conserved residues as reported by Kaufman et al. (1994) and sequences which did not contain the expected residues at conserved sites are indicated by coloured text at branch tips. Other sequences may lack the classical residues outside the range of the amplicon sequenced in this experiment so sequences represented in grey text cannot necessarily be assumed to be classical.

Of the five Lund sparrows from which TAPs were amplified, class I exon 3 sequences were ob-

tained from gDNA and cDNA from HS2, HS3 and HS4 by Drews et al. (2017). The study found evidence for a single dominantly expressed class I gene in house sparrows, and reported that HS2 and HS3 shared their two highest-expressing class I alleles (Pado-UA_392 and Pado-UA_393, see Drews et al. (2017) fig. S4). The TAP2 sequencing performed in this project showed that HS2 and HS3 share a maximum of one TAP2 allele, with potential additional variation outside the range sequenced in this project. TAP1 was sequenced in insufficient depth to make strong conclusions, however one SNP seen in three out of four HS3 clones was entirely absent from the five HS2 clones, providing some preliminary evidence that these birds do not share both TAP1 alleles as would be expected under a chicken-type model with two shared class I alleles.

This indication that sparrows do not have linked TAP and MHC class I loci was supported by genomic analysis using BLASTN to detect TAP and class I loci in a new assembly of the house sparrow genome produced using PacBio sequencing technology (M. Ravinet, University of Nottingham). TAP1 and TAP2 were detected together in opposite transcriptional orientations within a region of slightly more than 10 Kb. The contig on which the TAP genes were found extended approximately 80 Kb upstream and 110 Kb downstream of the TAPs, and contained no sequence with significant similarity to any class I query (classical and non-classical sequences from Drews et al. (2017); sequence A found in Sorci sparrows 1 and 3; zebra finch class I). Class I hits were found on two other contigs; the same two loci were identified with all four class I queries.

3.3.7 Zebra finch and chicken TAPs are similarly polymorphic

Nine unique TAP1 allele sequences were obtained from the five zebra finch individuals (figure 3.12). Tagu5 appears to be a homozygote, with just a single allele sequence identified in each replicate. The fact that TAP2 sequencing also showed a single sequence present in this bird supports the hypothesis that Tagu5 bird is indeed homozygous, rather than the primers being non-universal. Tagu7 and Tagu9 are clear heterozygotes, each with two matching alleles detected in each of their replicates.

Tagu9 had four putative TAP1 alleles detected in both replicates (barcodes 01 and 06). These two barcodes represent independent PCRs, so this is unlikely to be a PCR artefact. Tagu6 had two alleles in one replicate and four alleles in the other replicate. Interestingly, the two alleles that were only seen in one replicate match two of the alleles seen in both replicates of Tagu9. It is plausible that these two sequences represent a contaminating or artefactual sequence, since discounting them from both birds would leave the expected two sequences per barcode. All these alleles were retained in subsequent analyses since there was no conclusive evidence for their being erroneous in both birds at this stage. Two other TAP1 sequences appeared in multiple birds: one

allele was shared between Tagu7 and Tagu8, and another between Tagu5 and Tagu9.

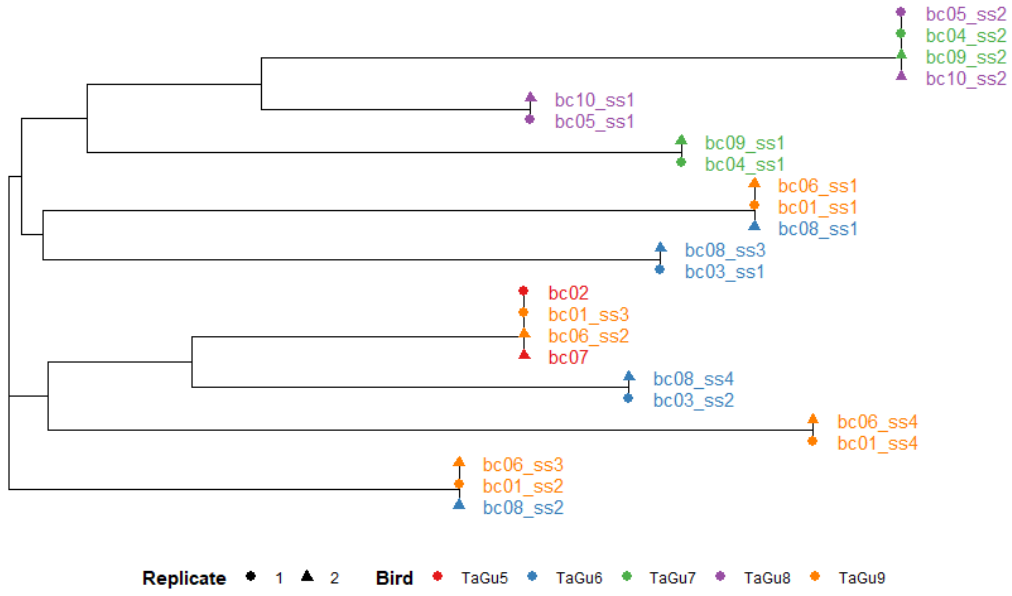


Figure 3.12: **Unrooted phylogenetic tree of TAP1 alleles identified in five zebra finch individuals.** 'bc' indicates the barcode with which the sequence was identifies, with two barcodes per bird representing the two technical replicates. Distinct allele sequences identified within the reads for a given barcode are distinguished by 'ss', with each 'ss' representing a subset of the total reads which were identified as being closely related and from which a consensus was drawn.

Seven unique TAP2 allele sequences were obtained from the five zebra finch individuals. In all cases, the replicates were consistent with each other and no bird had more than two alleles amplified. As for TAP1, Tagu5 was observed to be homozygous and to share its allele with Tagu9, and Tagu7 and Tagu8 were also seen to share an allele (figure 3.13).

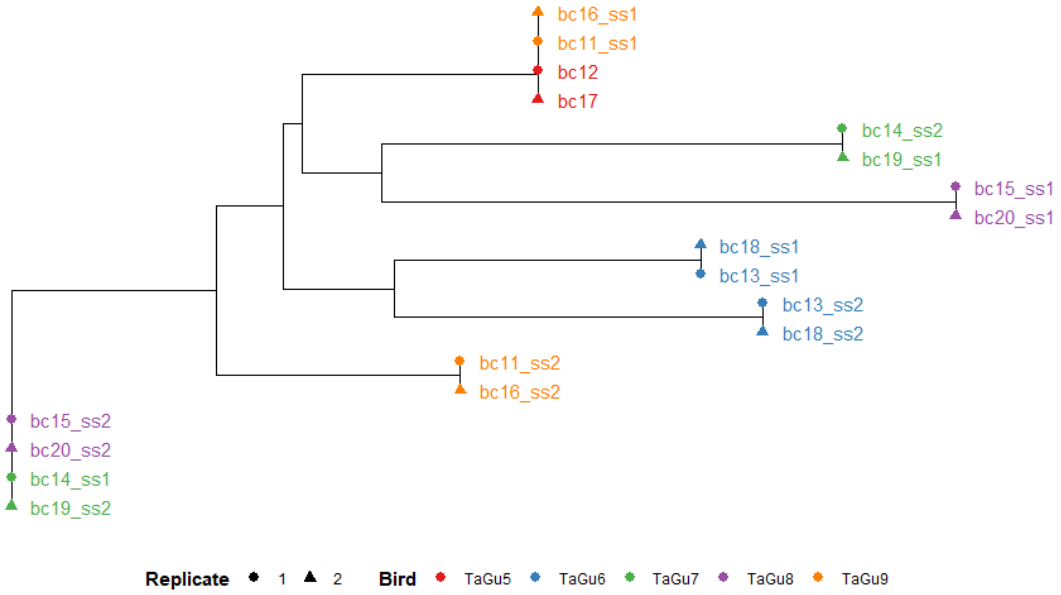


Figure 3.13: **Unrooted phylogenetic tree of TAP2 alleles identified in five zebra finch individuals.** 'bc' indicates the barcode with which the sequence was identifies, with two barcodes per bird representing the two technical replicates. Distinct allele sequences identified within the reads for a given barcode are distinguished by 'ss', with each 'ss' representing a subset of the total reads which were identified as being closely related and from which a consensus was drawn.

To confirm the accuracy of sequences obtained by Nanopore sequencing and investigate the extra TAP1 sequences amplified from Tagu6 and Tagu9 several of the PCR products from which the Nanopore library was constructed were cloned and Sanger sequenced. PCR products amplified with barcodes 1, 3, 4, 5, 6, 7, 8, 11 and 15 were all assessed in this way. Any differences from the Nanopore consensus sequences were only seen in single clones, suggesting that differences observed were attributable to Sanger sequencing errors or low-frequency PCR misincorporations which did not affect the Nanopore consensus. It would therefore be appropriate to use Nanopore sequencing alone in subsequent experiments.

Seven clones were sequenced from each of barcodes 1 and 6 (replicates of Tagu9 TAP1). All sequences reported in the Nanopore results were confirmed by Sanger sequencing except bc01_ss3/bc06_ss2. Confidence in the genuine presence of this sequence in Tagu9 is nonetheless high, since this sequence is shared with Tagu5, a feature which is also observed in the TAP2 phylogenetic tree. The absence of this sequence in the Sanger results is therefore attributed to the relatively small number of clones sequenced from each PCR product.

Five and 10 clones were sequenced from barcodes 3 and 8 (replicates of Tagu6 TAP1) respectively. The sequences obtained reflected the Nanopore results, with all Nanopore sequences observed in the Sanger results. There was no evidence from Sanger sequencing that the sequences identified in

the Nanopore data in barcode 8 but not barcode 3 (bc08_ss1 and bc08_ss2) were present in the barcode 3 PCR product.

Very little is known about the ancestry of these individuals, however Tagu7 and Tagu8 were both reported to be 'white spotted' variants, while Tagu5 and Tagu9 were 'black', so it is plausible that these individuals could have been purchased from the same breeder and could be related. Tagu6 was a 'pale' variant. If these pairs of birds were not related then these shared alleles, combined with the apparent TAP homozygosity of Tagu5, might suggest relatively low allelic diversity of zebra finch TAPs, which would appear to be inconsistent with a model in which these genes co-evolved with a class I locus.

Sequence diversity in the zebra finch alleles was compared to the unique chicken allele sequences over the same range. The 14 unique TAP1 sequences from the standard serological haplotypes: B2, B4/B13, B5, B6, B8/B11, B9, B12, B15, B17, B19, B21, B23 and B24 from Hosomichi et al. (2008) and B14 from Walker et al. (2011) (GenBank accession numbers AB426141-AB426154 and JF794480) were all unique in this range and were all included in the analysis. Of the fifteen unique chicken TAP2 alleles described in section 3.3.5, only the B5 and B8 haplotype sequences were identical over this range and 14 unique sequences were subsequently included in the analysis. There is no evidence that zebra finch TAP alleles have lower sequence diversity than chicken TAP alleles (table 3.12). Indeed, over the range covered by the Nanopore_2 amplicon, zebra finch TAP1 allele pairs have almost twice as many non-synonymous changes per non-synonymous site on average than chicken allele pairs. In TAP2, the zebra finch alleles have marginally fewer non-synonymous changes than chicken alleles.

		Unique alleles	Mean Ks	Mean Ka
TAP1	Chicken	14	0.041107	0.004921
	Zebra finch	9	0.043306	0.008542
TAP2	Chicken	14	0.023782	0.007876
	Zebra finch	7	0.030967	0.005476

Table 3.12: **Sequence diversity of Chicken and Zebra finch TAP alleles**

3.3.8 TAP2, but not TAP1, shows signals of positive selection in both chicken and zebra finch

Neither chicken nor zebra finch TAP1 had strong evidence of sites under positive selection. A likelihood ratio test (LRT) was performed according to Yang et al. (2000), comparing models 7 (β) and 8 ($\beta + \omega > 1$). For both species, twice the log likelihood difference was compared to the χ^2 distribution with two degrees of freedom, since the the difference in the number of parameters

in these models is two. There was no evidence that model 8, which allows for positively selected sites, fit the sequence data for either species better than model 7, which does not allow for sites to be under positive selection (chicken: $p = 0.210$, zebra finch: $p = 0.073$). However, the LRT is known to be particularly conservative with short, relatively similar sequences, or alignments with relatively few sequences, so it it likely that the test statistic for one or both of these species was underestimated (Anisimova et al., 2001). A Bayes Empirical Bayes analysis of both species was performed despite the result of the LRT. No individual sites in chicken TAP1 were identified to be under positive selection, supporting the outcome of the LRT. Two sites were identified as being positively selected in zebra finch TAP1; this result can be cautiously considered real, given that $0.05 < p < 0.1$ for the zebra finch LRT, and the test statistic is likely to have been at least slightly underestimated. The two residues are indicated on figure 3.6.

Positively selected sites (PSS) were shown to be present in the TAP2 genes of both species, according to the LRT (chicken: $p < 0.00001$, zebra finch: $p = 0.00012$). Eight sites were identified as being positively selected ($p < 0.05$) in the chicken according to a Bayes Empirical Bayes analysis performed in CODEML (Yang et al., 2005; Yang, 2007), compared to five in the zebra finch. None of the sites were identified as being positively selected in both species (figure 3.7).

3.3.9 The locations of variable and positively selected sites differ between chicken and zebra finch TAP sequences

The positions of polymorphic residues in the chicken and zebra finch were compared to the positions of residues known to be important for TAP2 peptide transport specificity in the rat, and those shown to be important for peptide binding in humans (figures 3.7 and 3.14). While the specific positions and orientations of side-chains is difficult to accurately predict when using a model to generalise across species, there seems to be approximately similar numbers of variable residues in potentially peptide-binding transmembrane helices in TAP1 in both chicken and zebra finch. The excess of variable residues in zebra finch TAP1 relative to chicken TAP1 is influenced by a highly variable disordered loop at the start of the nucleotide binding domain, some of which is apparently deleted in both human and chicken.

The variable residues in zebra finch TAP2 are almost entirely restricted to the fourth transmembrane helix (the first after the tapasin-binding domain, corresponding to the first fully resolved helix in the structural model), the preceding cytoplasmic loop, and the nucleotide binding domain (based on annotations of TAP sequences in Walker et al. (2005)). This variable fourth helix would likely be buried between the tapasin-binding domains (not modelled) and the peptide-binding domains, but it is impossible to know whether variation in the fourth helix would affect either of

these functional regions. No variation was observed in this helix in the chicken. There is one additional variable residue in zebra finch TAP2 which aligns just one residue away from a shared chicken and rat residue, amongst a cluster of residues that were identified as being functionally important in the rat and within the region that aligns to the sequence shown to be important for peptide binding in humans (figure 3.14).

Four out of five PSS identified in zebra finch TAP2 are in the third loop and fourth helix (end of exon 1; the structural model starts at the very end of the third loop and thus the variable residues corresponding to the first two PSS are not shown in figure 3.14), with a single PSS in the nucleotide binding domain. Chicken TAP2 has four PSS in the nucleotide binding domain and a further four in helices six to nine, which contain all the predicted human peptide binding regions.

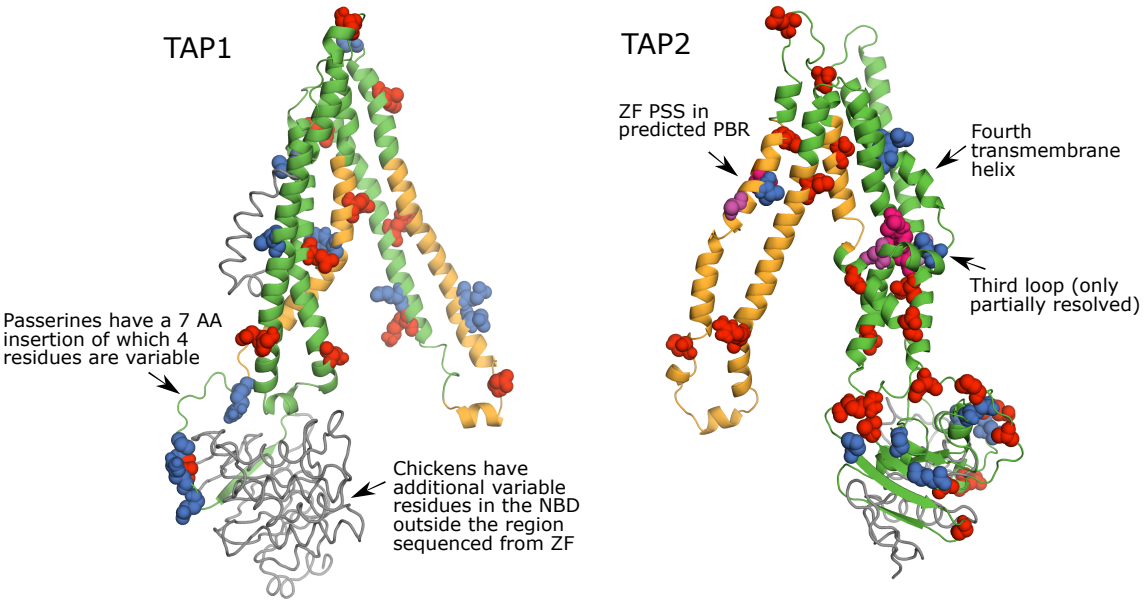


Figure 3.14: **Locations of variable residues in chicken and zebra finch TAP1 and TAP2 on structural models of human TAPs**. Residues shown in red and blue spheres indicate variable amino acids in chicken and zebra finch respectively. Residues known to affect peptide binding specificity in rat TAP2 are shown in purple spheres; the subset of these residues which are also variable in chickens are shown in pink. The human peptide binding domain is indicated in orange and the region not covered by the amplicon is shown as grey ribbon. The TAP2 model begins at the end of the cytoplasmic third loop, although the zebra finch amplicon additionally covers the remainder of the third loop, the third helix and part of the lumenal second loop (not shown). The region of the third loop not covered by the structural model contains an additional four zebra finch variable residues, of which two are PSS, outside the range shown in the model. Additional variable residues are known to exist in chickens outside of the region equivalent to that amplified from the zebra finch (Walker et al., 2011), but these are not shown to avoid biasing the comparison. Four additional variable sites were identified in zebra finch TAP1 in a seven-amino acid insertion on the disordered region at the start of the nucleotide binding domain. This insertion appears to be conserved in passerines but is absent from both humans and chickens.

3.3.10 TAP and MHC class I alleles do not co-segregate in zebra finches

Multiple expressed class I sequences were obtained from all five zebra finch individuals (figure 3.15). Only Tagu6_pp2_h was missing any of the seven conserved classical residues covered by the amplicon. No sequence that appeared in more than one clone could be explained by PCR chimerism and every sequence that appeared in more than two clones was amplified in two independent PCRs.

Despite the zebra finch genome containing only a single annotated class I locus, at least three of the zebra finch samples appeared to contain three alleles expressed to roughly similar levels. None of these sequences could be obviously attributed to PCR artefacts, since they were obtained in multiple independent reactions and none look obviously non-classical; all sequences obtained from more than one clone had the conserved classical residues and there is no obvious clade characterised by shorter branch lengths which might represent alleles amplified from a non-classical locus, which might be expected to be under weaker or no diversifying selection.

Two pairs of zebra finch samples were shown to share TAP alleles: Tagu 7/8 and Tagu 5/9. Neither of these pairs of birds share any MHCI sequences, indicating that these loci are not in linkage disequilibrium in the zebra finch. Furthermore, Tagu 5 was apparently homozygous at the TAP loci but showed no evidence of lower class I allelic diversity than any other bird.

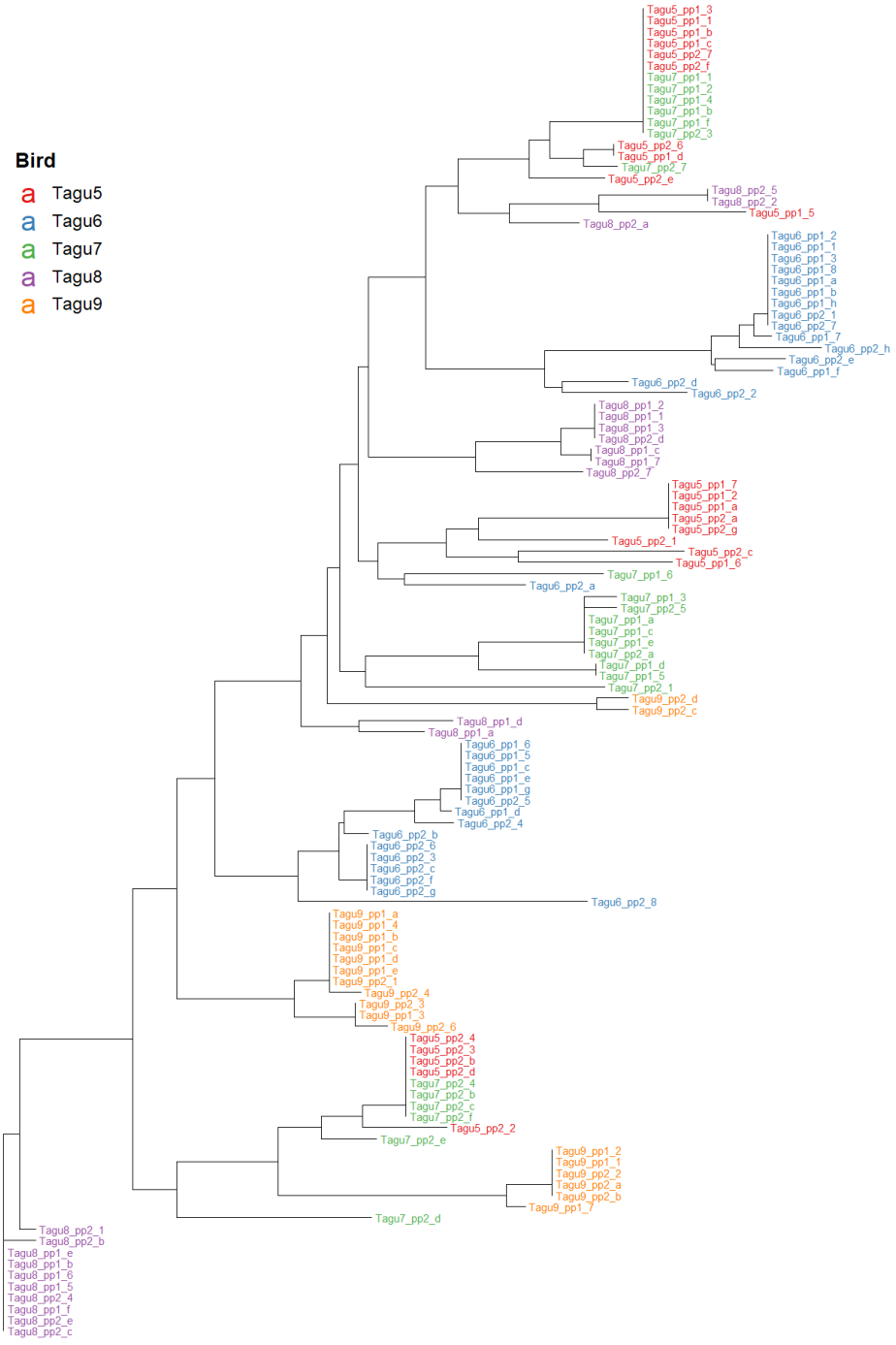


Figure 3.15: **Unrooted phylogenetic tree of MHC class I exon 2-3 sequences amplified from five zebra finch cDNA samples.** Each branch tip represents a single sequenced clone. Colours indicate the bird from which the cloned amplicon was obtained. "pp1"/"pp2" in sequence names denote the primer pair used to amplify the sequence, with the pp2 primers amplifying a slightly shorter amplicon. All sequences were trimmed to the same length prior to analysis. Numbered and lettered clones for a given bird and primer pair represent products of two independent PCRs.

3.4 Discussion

This work aimed to assess whether features of the 'chicken-type' MHC, namely a single dominantly expressed MHC class I gene, polymorphic TAPs and linkage between class I and TAPs, are present in Passeriformes, the crown group of the avian phylogeny. This would help to determine the extent to which insights developed in chickens are applicable to passerine birds. MHC diversity is often used as an indicator of conservation priority, but this metric has different implications in a 'chicken-type' system, where disease outcome is tightly linked to MHC haplotype, and a 'human-type' system, where MHC diversity is a good general measure of genetic diversity, but is unlikely to determine vulnerability to disease so directly. Furthermore, many passerine birds will be forced to adapt their migration routes in response to climate change, and an understanding of their immune system can help us to predict whether these birds are likely to be robust in a new pathogen environment. More generally, it is important to understand the extent to which adaptive immune systems can vary in structure and function, since this can help to better frame future experiments on non-model organisms.

As discussed in the introduction to this chapter, a high degree of evolutionary flexibility for passerine class I systems has been indicated by variable numbers of class I loci (O'Connor et al., 2016; Minias et al., 2019), variable expression patterns for class I genes (Balakrishnan et al., 2010; Drews et al., 2017; Drews and Westerdahl, 2019) and variable evolutionary relationships between classical and non-classical genes (Bonneaud et al., 2004; Sepil et al., 2012; Karlsson and Westerdahl, 2013). It is perhaps, therefore, unsurprising that variation is also observed in the TAP genes and their likely evolutionary histories and relationships with class I genes.

An interesting observation was that none of the likely classical passerine class I sequences had a lysine residue at position 146 in the class I protein sequence, which has been reported to be highly conserved in vertebrate classical class I protein sequences (Kaufman et al., 1994). An arginine at this position, which is seen in non-classical chicken YF genes, was present in all three collared flycatcher genomic sequences, the single zebra finch classical class I genomic sequence reported by Balakrishnan et al. (2010); Ekblom et al. (2011), all zebra finch class I sequences obtained in this project, almost all house sparrow class I sequences obtained in this project and almost all putative classical house and tree sparrow class I sequences described by Drews et al. (2017). The conserved residues identified by Kaufman et al. (1994) are involved in binding the N and C termini of peptides, and in non-mammalian vertebrates an arginine (as opposed to the mammalian tyrosine) at the equivalent of position 84 allows peptides to hang out of the peptide binding groove at the C-terminus (Xiao et al., 2018). This is reminiscent of the binding of peptides to classical class II molecules, which also carry an arginine residue at the equivalent position.

Despite all having the apparent ability to bind long peptides, most chicken class I molecules so far examined by immunopeptidomics seem to do so rarely, with the exception of BF2*031:01 (N. Ternette and J. Kaufman, unpublished), an observation likely related to TAP transport specificity. Position 146, which appears to differ in passerines, also helps to coordinate C-terminal peptide binding and could potentially be involved in allowing passerine class I molecules to effectively bind long peptides if they were available. It would be interesting to investigate the TAP transport specificity of the chicken B31 haplotype, as well as that of passerine birds.

3.4.1 Flycatchers

Limited sequencing of flycatcher TAP genes showed no evidence of significant polymorphism, which suggests that the 'chicken-type' system is not ubiquitous in birds. Under the hypothesis that the chicken and human define the two possible models we would expect flycatchers to have a 'human-type' family of expressed classical class I genes, unlinked to the TAP genes. Limited data is available to evaluate these characteristics, particularly with respect to classical class I gene expression. A class I locus potentially linked to the TAPs in the collared flycatcher does not contain the Y7 residue which would normally be expected in a classical gene. Flycatchers therefore do not neatly fit either model, and work on this species raised hypotheses related to the potential evolutionary flexibility of the class I antigen processing system in passerine birds.

The scaffold identified in figure 3.8 which contains the flycatcher TAP genes and a possible class I locus does not obviously resemble any particular region in the chicken, but does resemble regions in many other passerine species which also have SYCP1 and tenascin-X-like sequences proximal to their TAPs (including Swainson's thrush, Bengalese finch, White wagtail, New Caledonian crow, hooded crow, Eurasian tree sparrow, Great tit and Barn swallow). In species where this gene cluster is included on longer scaffolds, other nearby genes shared between species include QPRT, a cytochrome P450 4F3-like gene and a mitochondrial medium-chain specific acyl-CoA dehydrogenase. In the hooded crow, New Caledonian crow and great tit, some evidence of an MHC class I sequence is present, similarly to the collared flycatcher. However, in the New Caledonian crow the sequence is low quality (one base inserted and one base deleted from the genomic sequence to create the predicted CDS) and in the great tit the annotated CDS includes just two exons which cover the equivalents of exons 3-6 in the chicken BF2 gene. It therefore seems unlikely that an expressed classical class I gene in linkage disequilibrium with the TAPs is a feature of at least the vast majority of passerine genomes. The presence of class I-like sequences close to the TAPs in some species does, however, raise interesting questions about the processes by which the genetic architecture might have diverged in the passerine lineage.

3.4.2 Sparrows

No evidence was found for significant TAP polymorphism or TAP-MHCI linkage in house sparrows in this project, again indicating divergence from a 'chicken-type' system which is predicted to be ancestral. However, the characteristics of the 'human-type' system do not entirely explain the data either, with a single dominantly expressed class I locus shown to be present in sparrows in both this project and work by Drews et al. (2017). The existence of a single dominantly expressed class I in chickens has been related to co-evolution with the TAP genes (Walker et al., 2011), such that additional loci would confer minimal benefit due to incompatibility with the TAP transport repertoire. This model is difficult to reconcile with the largely monomorphic TAPs and single dominant class I seen in sparrows, since limited TAP polymorphism precludes TAP-class I co-evolution.

There are, however, aspects of the co-evolution model in chickens which are still unresolved. Some BF2 molecules bind extremely wide peptide repertoires and are therefore associated with very permissive TAPs which may transport almost as wide a range of peptides as the human TAP heterodimer. Why, then, has a multigene family evolved in humans but not chickens? The same mechanism by which class I copy number is restricted in chickens, which may be more complex than TAP co-evolution alone, may explain the single dominantly expressed class I in sparrows, which do not exhibit features of TAP-class I co-evolution.

The numbers of class I loci in genomes, while variable between species, are typically orders of magnitude lower than the number of potentially adaptive alleles segregating in the population. This suggests that trade-offs exist between the adaptive benefit of being able to present more pathogen-derived peptides and one or more costs of expressing additional alleles. The effects of these costs are especially visible in polyploid frogs of the genus *Xenopus*, which have actively silenced duplicated class I loci while retaining active duplicates of many other genes (Kobel and Du Pasquier, 1986; Flajnik, 1996).

Woelfing et al. (2009) reviewed possible costs associated with high intra-individual class I diversity of which the best-supported was T-cell repertoire depletion (Vidovic and Matzinger, 1988; Lawlor et al., 1990). Higher numbers of class I molecules seem to increase negative selection in the thymus more substantially than positive selection, resulting in a reduced repertoire with possible 'holes' (Migalska et al., 2019). In species where peptide binding repertoires can be very broad, such as chickens and possibly other non-mammalian vertebrates with the capacity to present long peptides, selection against increased copy number may be particularly strong. Such arguments for copy number optimisation may help to explain why sparrows have a single dominantly expressed class I despite no apparent co-evolution with the TAP genes, especially if Arg146 in passerine class

I molecules does indeed facilitate presentation of a greater diversity of peptide lengths.

Class I copy number does, however, seem to be variable in passerine birds, which may preclude the broad application of a model based on the presence of highly promiscuous class I alleles in birds. While most studies fail to distinguish between classical and non-classical genes, or high/low expression (O'Connor et al., 2016; Minias et al., 2019), and therefore cannot rule out the presence of a single dominantly expressed classical class I, Drews and Westerdahl (2019) detected multiple putatively classical genes in the siskin which were well-expressed at the RNA level.

3.4.3 Zebra Finch

Unlike sparrows and flycatchers, zebra finches exhibited similar levels of TAP polymorphism to chickens. This observation, combined with reports of a single dominantly expressed class I by Balakrishnan et al. (2010) and Ekblom et al. (2011), suggests that zebra finches may have a 'chicken-type' MHC. However, no evidence for the linkage disequilibrium which would be required for TAP-class I co-evolution has been found. Zebra finch individuals which share TAP alleles do not share class I alleles and vice versa, and genomics indicates that the TAP genes and single expressed classical class I locus are on different chromosomes. Furthermore, while sequence diversity was similar between chickens and zebra finch, allelic polymorphism could be lower in zebra finches (depending on the relatedness of the individuals in this study).

Without functional mutation studies it is difficult to know whether the polymorphisms in zebra finch TAP genes affect peptide transport to the extent seen in chickens or even rats. It is difficult to understand how a restricted transport repertoire would be adaptive in a system where advantageous combinations of TAP and class I repertoire could not be reliably inherited together unless all class I genes were highly promiscuous and peptide repertoire was instead defined by the TAPs. Alternatively, the polymorphisms could be a response to viral immune evasion strategies targeting the TAP transporter (reviewed in Verweij et al. (2015)) and may not affect peptide transport. The zebra finch TAP genes (like the chicken TAPs) do seem to have accumulated a number of mutations on the outside of the nucleotide binding domain, which are unlikely to affect peptide binding but may be a response to viral proteins which inhibit ATP binding to TAP, such as human cytomegalovirus US6 (Hewitt et al., 2001) or Epstein-Barr virus BNLF2a (Hislop et al., 2007).

Further work is required to investigate the presence of more TAP1 allele sequences than expected in zebra finch samples 6 and 9, and more class I allele sequences than expected in samples 5, 6, 7 and possibly 9. In all cases except sample 6 TAP1, all sequences under consideration were amplified in two independent PCRs; it is therefore necessary to consider contamination of the samples them-

selves, which seems unlikely given the observed patterns of shared and unique sequences amplified from individuals, or the presence of additional expressed genes with high sequence similarity to either TAP1 or class I in certain individuals.

Chapter 4

Diversity in tapasin homologues

4.1 Introduction

4.1.1 Tapasin

Tapasin, sometimes referred to as TAPBP, tsn or tpn, is a peptide editor and chaperone which forms the physical link between MHC class I and TAP in the PLC. It was first identified when an unknown 48 kDa protein was found to co-immunoprecipitate with TAP in addition to MHC class I, and was hypothesised to mediate the interaction between TAP and class I (Ortmann et al., 1994). The antigen processing mutant cell line 721.220 was key to the understanding of tapasin's function; it was shown that only some HLA alleles suffered impaired expression when transfected into .220 cells (Greenwood et al., 1994) and that the same alleles failed to associate with TAP in .220 but not wild-type cells (Grandea et al., 1995). The missing factor causing the link between failure to associate with TAP and lack of class I expression in .220 cells was identified as tapasin (Sadasivan et al., 1996), and reintroduction of the tapasin gene into .220 cells restored both class I expression and antigen presentation (Ortmann et al., 1997).

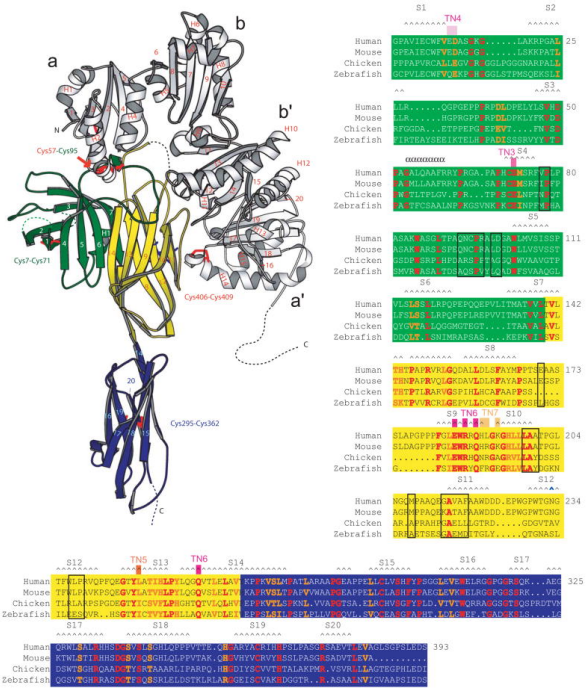
Structure

Tapasin consists of a signal sequence, three extracellular domains, a transmembrane region and a cytoplasmic tail. The N-terminal domains are a fusion of a 7-stranded β-barrel (residues 1–147) and an Ig-like fold (residues 150–269), while the C-terminal domain forms a more typical Ig-like fold (Dong et al., 2009). It forms a highly stable, disulphide-linked heterodimer with the thiol oxidoreductase ERp57 (Peaper et al., 2005), which subsequently recruits MHC class I molecules and the chaperone calreticulin to the PLC (Wearsch and Cresswell, 2007).

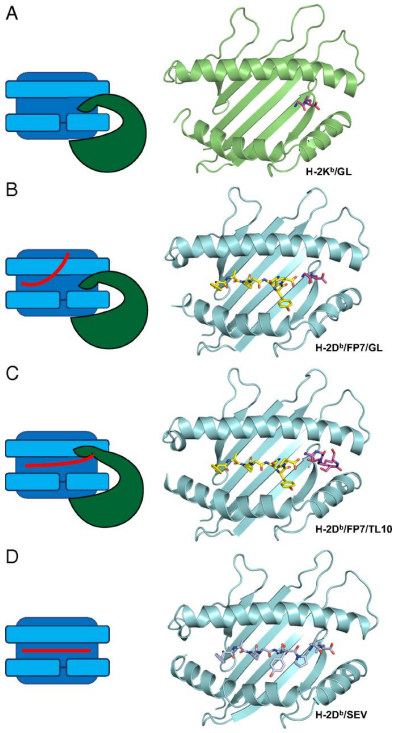
Interactions between components of the PLC had been deduced from mutational and functional studies but have recently been understood in more detail with the publication of a low-resolution

cryo-electron microscopy (cryo-EM) structure of the human PLC (Blees et al., 2017). This structure is likely to be largely conserved across species which express all the relevant components, although the loss of the tapasin binding domain in the TAP1 protein of birds presumably prevents the assembly of two editing modules (tapasin, calreticulin, ERp57, and MHCI), one on each side of the TAP heterodimer, as is seen in humans (Walker et al., 2005; Blees et al., 2017). Whether the avian PLC is 'one-sided' or includes an alternative protein complex on the TAP1 side is unknown.

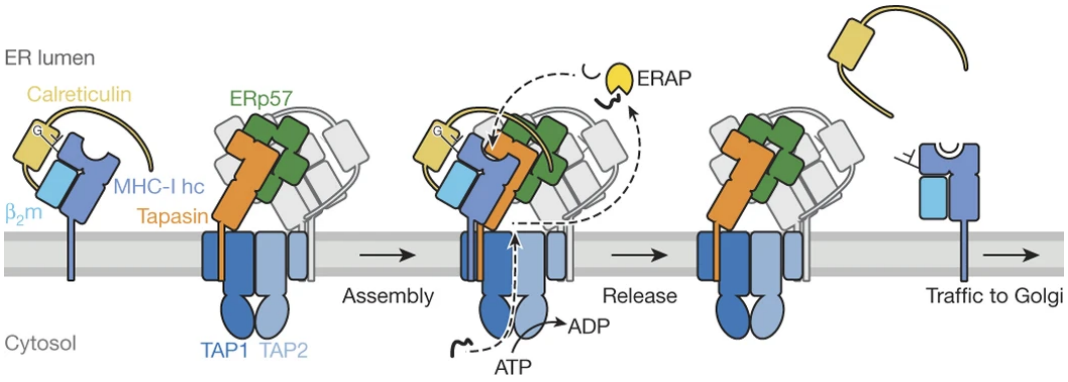
Higher resolution crystal structures of the murine MHC class I allele H-2K^{b} with the dipeptide GL, which is present at the tip of a disordered loop in tapasin, confirmed that this dipeptide could stabilise the peptide binding cleft in a state that renders it available for peptide binding (Hafstrand et al., 2019). While this loop was not resolved in the crystal structure of Dong et al. (2009) or the cryo-EM structure of Blees et al. (2017), modelling by Hafstrand et al. (2019) (taking into account the N-terminal β-barrel mobility reported by Blees et al. (2017)) indicated that the loop could be located close enough to the F-pocket of MHC class I to allow binding to occur.



(a) Crystal structure of tapasin (coloured) in complex with ERp57 (grey) with alignment of tapasin sequences from human, mouse, chicken and zebrafish. Figure from Dong et al. (2009)



(b) Proposed mechanism by which the 'scoop loop' of tapasin stabilises class I until a high affinity peptide is bound. Figure from Hafstrand et al. (2019)



(c) Assembly and disassembly of the PLC. Figure from Blees et al. (2017)

Figure 4.1: **Structure and molecular interactions of tapasin.**

Function

The tapasin-ERp57 heterodimer is required for recruitment and assembly of PLC components and for effective peptide loading and editing (Wearsch and Cresswell, 2007). Mutation of the so-called 'scoop loop' (a name first used in reference to the equivalent structure in TAPBPR) significantly reduced surface expression of class I, supporting its specific role in assembly of stable class I-peptide complexes (Hafstrand et al., 2019). This led to the hypothesis that the N-terminus of an MHC class I-restricted peptide binds to the peptide binding cleft first and that the C-terminus is subsequently tested for its ability to dislodge the tapasin loop, or more specifically the GL dipeptide, from the

F pocket.

Tapasin's role varies between MHC haplotypes. It has been shown that in both mammals and birds MHC class I alleles vary in their tapasin dependence, that is, their ability to load peptides in the absence of tapasin (Peh et al., 1998; Williams et al., 2002; Lewis et al., 1998; Park et al., 2003; van Hateren et al., 2013). Various specific residues in the human class I molecule have been implicated in this phenotype including 114 (H114E was sufficient to change the HLA-B*2705 allele from tapasin-independent to tapasin-dependent (Park et al., 2003)) and 116 (a single natural polymorphism permitted tapasin-independent loading of HLA-B*4405 (116Y) but not HLA-B*4402 (116D) (Williams et al., 2002)).

As described previously, tapasin dependence appears to correlate with a suite of other variable characteristics (Kaufman, 2017). More tapasin-dependant HLA alleles have higher cell surface expression (Rizvi et al., 2014; Chappell et al., 2015) and smaller predicted peptide binding repertoires (Košmrlj et al., 2010; Kaufman, 2017; Bashirova et al., 2020). This can cause them to be associated with poorer disease outcomes, for example in the case of HIV where tapasin-dependant MHC alleles were generally associated with faster progression to AIDS, attributed to the smaller range of HIV peptides that these alleles could present (Bashirova et al., 2020). However, it is important to note that a few specific tapasin-dependant HLA alleles such as B*57:01, which are able to bind highly conserved, immunodominant HIV epitopes, actually confer extremely strong resistance to disease progression (Bashirova et al., 2011; Chappell et al., 2015).

In chickens, the situation is more complicated, likely due to the presence of highly polymorphic tapasin genes. While tapasin dependence has not been explicitly tested in many haplotypes, the same suite of other correlated characteristics is generally present in chickens. It is well established that surface expression of chicken class I molecules is inversely correlated with peptide binding repertoire (Chappell et al., 2015) but positively correlated with thermal stability (Tregaskes et al., 2015), suggesting that high-expressing alleles present a few very high affinity peptides. It makes sense that such alleles would rely more heavily on a chaperone and editor to 'fine-tune' their repertoire than low expressing 'generalist' alleles which will successfully form stable complexes with a wider range of peptides of lower average affinity. Having said that, van Hateren et al. (2013) showed different degrees of tapasin dependence in BF2*019:01 and BF2*015:01, both of which are high expressing (Chappell et al., 2015) and have highly fastidious peptide binding motifs (Kaufman et al., 1995; Wallny et al., 2006). Furthermore, the polymorphism in chicken tapasin was seen to be co-evolving with the polymorphic BF2 gene; BF2*019:01 expressed in B15 cells (containing tapasin*015) matured very inefficiently, while mutating two key tapasin contact residues (126 and 220) in BF2*019:01 so that they resembled BF2*015:01 restored the majority of the maturation

efficiency (van Hateren et al., 2013).

Chickens generally show much stronger associations between MHC haplotype and infectious disease outcome than humans, with low expressing, generalist, presumably tapasin-independent haplotypes such as BF2*002:01 and BF2*021:01 being associated with protection from Marek's disease (Bacon et al., 1981; Simonsen, 1987), Rous sarcoma virus (Collins et al., 1977; Bacon et al., 1981), avian leukosis virus (Bacon et al., 1981), avian pathogenic *Escherichia coli* (Cavero et al., 2009), avian influenza (Boonyanuwat et al., 2006), avian coronavirus (Banat et al., 2013) and northern fowl mite (Owen et al., 2008). This reflects the trend observed in HIV infection in humans, although as yet no 'specialist' alleles (equivalent to HLA-B*57:01 in HIV) have been seen to provide unusually strong protection against a particular disease in chickens.

4.1.2 TAPBPR

TAPBPR (tapasin-related) was identified near an MHC paralogous region at human chromosome position 12p13.3 (Teng et al., 2002). Its amino acid sequence is approximately 22% identical to tapasin and it is found throughout the major groups of vertebrates, suggesting a conserved function. Boyle et al. (2013) found TAPBPR to be an additional component of the MHC class I presentation pathway which acts in both the ER and cis-Golgi, and independently of the PLC. Like tapasin, it is widely expressed, inducible with IFN-γ and binds a heterodimer of MHC class I and β_2m.

Structure

Two crystal structures of TAPBPR in complex with MHC class I revealed similarities to tapasin, with a composite N/IgV domain consisting of 19 β-sheets followed by an IgC structure (Jiang et al., 2017; Thomas and Tampé, 2017). Despite protein sequence identity being relatively low between tapasin and TAPBPR, residues that interact with MHC class I are broadly conserved and the TAPBPR-MHCI binding site identified in the crystal structures overlap with the regions identified as being important for tapasin-MHCI interactions by mutational analysis and molecular dynamics simulations (Simone et al., 2012; Hermann et al., 2013).

As in tapasin, a 'scoop loop' and 'jack hairpin' provide specific points of interaction between TAPBPR and the α1/2 domains of MHC class I but there are also significant interfaces between the concave surface of TAPBPR's N-terminal domain and the class I heavy chain and between the C-terminal domain of TAPBPR and the α3 domain of the heavy chain/β_2m, which bury a large amount of the surface area of the class I molecule (Thomas and Tampé 2017, figure 4.2).

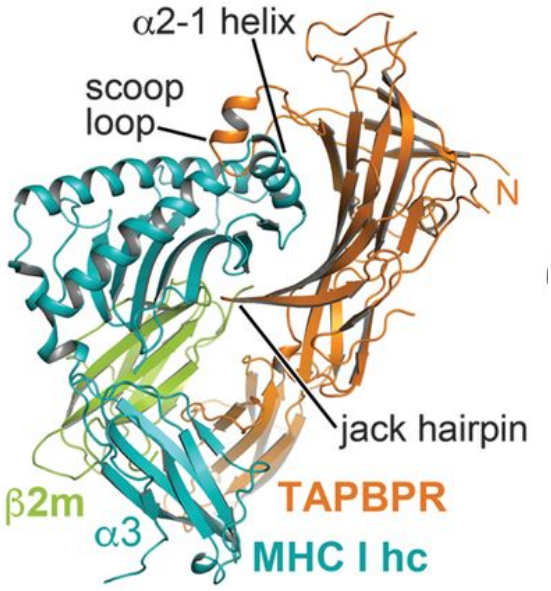


Figure 4.2: **Crystal structure of TAPBPR in complex with MHC class I.** Figure from Thomas and Tampé (2017).

Alternative splicing and the expression of multiple isoforms have been reported to be features of TAPBPR in trout and humans (Landis et al., 2006; Porter et al., 2014). In particular, the human β (additional 40 amino acids (exon 7a) retained in the cytoplasmic tail) and γ (membrane-proximal IgC domain (exon 5) lacking) transcripts were found to be abundant in peripheral blood mononuclear cells (PBMCs) and dendritic cells (DCs). TAPBPR γ was unable to associate with MHC class I however TAPBPR β did form this association and overexpression of the β isoform reduced surface expression of MHC class I to 33% of wild type (compared to 20% of wild type when TAPBPR α is expressed at the same level) (Porter et al., 2014). Porter et al. (2014) speculate that the predominant expression of TAPBPR β in cells with cross-presentation capabilities might indicate a role for this isoform in cross-presentation. While alternative tapasin transcripts have also been reported (Gao et al., 2004; Beutler et al., 2013), they seem to be uncommon in normal cells and are unlikely to play an important role in cells with unimpaired expression of the normal tapasin transcript.

Function

Like tapasin, TAPBPR can be broadly described as a peptide editor and will catalyse peptide exchange *in vitro*, narrowing the peptide repertoire and increasing the average affinity of bound peptides (Hermann et al., 2015; Morozov et al., 2016). However, within a cellular environment tapasin and TAPBPR can have distinct effects on both cell surface expression of class I molecules and class I peptide binding repertoire (Hermann et al., 2015). Different MHC class I alleles vary in their dependence on tapasin as previously discussed (section 4.1.1), but also in their binding

affinity for TAPBPR, which correlates with the degree of peptide exchange activity observed *in vitro* (Ilca et al., 2019). As a result, it can be difficult to describe the effects of these editors (or, indeed, the effects of knocking them out) in general terms, since the effects will vary depending on the MHC haplotype of the cells used. However, the observation that TAPBPR depletion does not significantly reduce cell surface class I expression (Boyle et al., 2013; Hermann et al., 2015), as well as the fact that TAPBPR appears to act more distally from the TAP transporters (that is, later in the antigen presentation pathway) (Boyle et al., 2013), has led to TAPBPR often being described as having a 'fine-tuning' or refining effect on the peptide repertoire (Ilca and Boyle, 2020).

As with tapasin, a disordered loop which interacts with the F pocket of the class I molecule is key to TAPBPR's peptide-editing activity. Indeed, the 'scoop loop' or 'leucine lever' and its critical role in class I binding and peptide exchange was described in TAPBPR (Ilca et al., 2018) before a similar structure was shown to confer the editing capabilities of tapasin (Hafstrand et al., 2019). TAPBPR stabilises peptide-receptive class I molecules through the binding of its loop, which acts as a peptide surrogate, in the F pocket. As such, TAPBPR will bind empty class I molecules or those where the bound peptide is of sufficiently low affinity to be dissociated and replaced by the loop. Rebinding of disassociated, low-affinity peptides is also inhibited by the presence of the loop, while high-affinity peptides would be able to subsequently displace the loop and cause the class I molecule to disassociate from TAPBPR and traffic to the cell surface (Ilca et al., 2018; Sagert et al., 2020). Ilca et al. (2018) do, however, note that TAPBPR is still able to facilitate peptide dissociation to a small extent in a loop-independent manner. This may occur only with peptide-MHCI complexes that are poorly loaded and therefore intrinsically prone to dissociation. Other regions of TAPBPR such as the 'jack hairpin' (Thomas and Tampé, 2017) could be involved in this loop-independent functionality, as could mechanisms such as the negative allosteric release cycle proposed by McShan et al. (2018).

In addition to being a peptide editor, TAPBPR was shown to promote optimal peptide loading by facilitating the return of empty or poorly-loaded class I molecules (with which it preferentially associates) to the PLC. Neerincx et al. (2017) demonstrated that TAPBPR interacts with UDP-glucose:glycoprotein glucosyltransferase 1 (UGT1), which is known to regenerate the $Glc_1Man_9GlcNAc_2$ moiety on glycoproteins. In the context of antigen processing, TAPBPR appears to recruit UGT1 to reglucosylate the glycan on class I molecules, promoting their recognition by calreticulin and subsequent recycling back into the PLC.

Not all MHC class I allotypes undergo TAPBPR-mediated peptide editing to the same extent. Ilca et al. (2019) demonstrated a preference of TAPBPR for HLA-A over HLA-B or -C (although not all HLA-A molecules exhibited stronger binding to TAPBPR than HLA-B or -C alleles), with

members of the A2 and A24 supertypes binding particularly strongly. Binding affinity was shown to be proportional to peptide exchange activity. Of the six alleles identified to have particularly high affinity for TAPBPR (Ilca et al., 2019), five were included in an analysis by Bashirova et al. (2020), who showed low tapasin dependence in all five cases. This suggests an inverse correlation for the importance of tapasin and TAPBPR for the loading of peptides on different HLA alleles.

4.1.3 TAPBPL

Grimholt (2018) described a 'previously undefined...TAPBP-like [TAPBPL]' gene which showed a similar level of sequence similarity with both tapasin and TAPBPR as is seen between tapasin and TAPBPR (approximately 22%). It was first identified in Atlantic salmon, where there are three copies of the gene, but was also found as a single copy in zebrafish, spotted gar, frogs, turtles, alligators, birds and marsupials (Grimholt, 2018), suggesting both that the duplications of this gene are unique to salmonids and that it was lost on the lineage leading to placental mammals. It is likely that the TAPBPL1a and TAPBPL1b genes, which share 88% amino acid sequence identity, are the result of the salmonid genome duplication event, but the more divergent TAPBPL2 appears to have a more ancient origin (Grimholt, 2018).

Many of the tapasin and TAPBPR residues known to be involved in MHC class I interactions were conserved in the TAPBPL sequences from salmon, zebrafish, medaka, gar, frog, chicken, kiwi, turtle and opossum but few other features of tapasin or TAPBPR were strongly conserved across the TAPBPL sequences (Grimholt, 2018). Many of the TAPBPL sequences had C-terminal ER retention motifs (KKXX or XKXX) like tapasin but none contained the conserved lysine (arginine in chicken, position 408 in human tapasin) thought to facilitate the TAP-tapasin interaction (Petersen et al., 2005). Neither the human tapasin C95 residue known to bind ERp57 (Dong et al., 2009) or the human TAPBPR C94 known to interact with UGT1 (Neerincx et al., 2017) are conserved in TAPBPL, suggesting a distinct function. TAPBPL was shown to be expressed in a range of tissues, with a distribution similar to that of tapasin and TAPBPR (Grimholt, 2018).

Taken together, these data suggest that TAPBPL is likely to bind MHC class I and play a role in antigen presentation, but with a function that is distinct from that of tapasin or TAPBPL. It is tempting to associate its absence in placental mammals with the other changes to the architecture of the MHC that seem to have occurred on this lineage, but at this stage this is no more than speculation. Depending on its role, which could plausibly influence peptide repertoire, the presence or absence of TAPBPL could be a key consideration when trying to apply insights (relating to the role of 'generalist' and 'specialist' class I alleles, for example) from placental mammals to non-placental vertebrates, and vice versa.

It has been hypothesised that TAPBPL could be involved in the chicken PLC (J. Kaufman, personal communication). With no tapasin binding domain on chicken TAP1, it seems unlikely that the chicken PLC has two symmetrical tapasin-ERp57-calreticulin-class I editing modules as is seen in humans, however it is plausible that TAPBPL could interact with TAP1 via an alternative binding surface and thus recruit a second class I molecule to the PLC (assuming class I binding is conserved, as is predicted from the sequence). The absence of an equivalent to K408 in TAPBPL does not necessarily preclude this, since the binding of TAPBPL to chicken TAP1 (which lacks a tapasin binding domain) would not be expected to closely resemble the binding of tapasin to human TAP1 on the scale of individual residues.

4.1.4 Specific aims

This work aims to investigate diversity in these peptide editors to determine the extent to which this variation needs to be considered when applying insights across lineages. It begins to investigate the role of TAPBPL specifically in order to assess whether the presence or absence of this protein in certain lineages might result in subtle or significant differences between the antigen processing and presentation systems of mammals and non-mammals. Key questions include:

- What is the phylogenetic distribution of tapasin, TAPBPR and TAPBPL? Do different lineages vary in the editors they retain or lose?
- Does the expression of peptide editors vary between different chicken haplotypes which differ in their class I peptide binding repertoires?
- What is the distribution of tapasin, TAPBPR and TAPBPL in chicken tissues?
- Is TAPBPL part of the chicken PLC?

4.2 Materials and Methods

4.2.1 Genomic resources and analysis

Genomic resources were accessed through Ensembl release 99 (Cunningham et al., 2019) and GenBank (Clark et al., 2016), with specific regions of interest identified using BLAST algorithms (Altschul et al., 1990) accessed through the NCBI BLAST suite (`https://blast.ncbi.nlm.nih.gov/Blast.cgi`). Alignments were created using Clustal Omega (Sievers et al., 2011) and Geneious Prime versions 11 to 2021.1 (`https://www.geneious.com`). Maximum-likelihood phylogenetic trees were produced in MEGA X (Kumar et al., 2018) . Manipulation of sequence files was done using python 3 (Van Rossum and Drake, 2009) with modules from biopython (Cock et al., 2009).

Data was visualised in R (R core team, 2013) using ggplot2 (Wickham, 2016) and ggtree (Yu et al., 2017).

4.2.2 Cell lines

Five cell lines representative of five MHC haplotypes common in commercial flocks were maintained in Roswell Park Memorial Institute (RPMI) 1640 medium supplemented with 10% foetal bovine serum (FBS), 0.3 g/l L-glutamine and 10 U/ml (1x) penicillin/streptomycin. The MHC haplotypes of the cell lines were confirmed by amplification and Sanger sequencing of TAP2 exon 6.

Lines TG12, TG15 and TG21 were produced by Dr. Thomas Göbel (Basel Institute for Immunology and Ludwig-Maximilians-Universität, Munich) and are homozygous for the the B12, B15 and B21 MHC haplotypes respectively. TG15 and TG21 were made from the H.B15 and H.B21 lines which were developed from Scandinavian White Leghorns while TG12 came from the highly inbred CB line developed from Reaseheath line C. Cell lines IS2 and IS19 were made by Iain Shaw (Institute for Animal Health, Compton, UK) from the 6_1 and P2a chicken lines. All lines were immortalised by infection with the highly transforming T strain of reticuloendotheliosis virus.

4.2.3 Antibodies

Monoclonal antibodies (mAb) to chicken class I heavy chain (F21-2) were raised against chicken erythrocytes and lymphocytes (Crone et al., 1985). F1-3 is a mAb raised against the C-terminal peptide of chicken TAP2 (LRTRGGPYSRLLQH) and was produced for this project by growing the hybridoma cells in CD Hybridoma medium (Gibco) supplemented with 0.3 g/l L-glutamine and 10 U/ml (1x) penicillin/streptomycin and then filter sterilising the supernatant. MAbs with names starting 11-46- were raised against the extracellular domain of chicken tapasin. MAbs with names starting 19-52, 19-53 and 19-54 were raised against the C-terminal peptides of chicken tapasin (AARPKEETKKSQ), TAPBPR (AAEPKPEQLLTASE) and TAPBPL (QVRSTTKPKPY) respectively. All antibodies and hybridomas were provided by Dr. Karsten Skjødt (University of Southern Denmark) with the immunogen for the tapasin extracellular domain provided by Dr. Andy van Hateren (University of Southampton).

4.2.4 Membrane-enriched cell lysates

Cell lines

Cell lysates were prepared according to a protocol optimised by Walker et al. (2011).

Live cell density was estimated by mixing 10 µl cells with 10 µl Trypan Blue solution (Sigma-Aldrich) and counting the live cells in a Neubauer haemocytometer (Thermo Scientific). 50 ml

cultured cells were spun down (400 xg, 5 min, 4 °C) and the pellet resuspended in 500 µl 'Freeze-Thaw' buffer (1 mM $MgCl_2$ in phosphate-buffered saline (PBS), 0.1 mM 4-benzenesulfonyl fluoride hydrochloride (AEBSF)). This was frozen on dry ice and thawed at room temperature twice and then centrifuged (17 000 xg, 45 min, 4 °C). The pellets were solubilised with three cycles of 20 s agitation and 5 min incubation on ice in a volume of digitonin lysis buffer (150 mM NaCl, 1 mM $MgCl_2$, 10 mM TrisCl, pH 8 with 1% digitonin (Sigma Aldrich) and 0.1 mM AEBSF) such that the final concentration was 6×10^5 cell equivalents/µl. The resuspended lysate was centrifuged again (17 000 xg, 20 min, 4 °C) and the supernatant used immediately or stored at −80 °C.

***Ex vivo* tissue**

Tissues (brain, thymus, bursa, gut, liver, spleen, lung) were extracted from line 0 chickens homozygous for the B21 haplotype following intraperitoneal pentobarbital injection, and split between tubes of filter sterilised PBS and RNAlater. 0.5 g of tissue in PBS was diced in a petri dish on ice and then incubated for 6 min at 37 °C with 2 ml Liberase TH digestion buffer (Liberase stock solution at 2.5 mg/ml in water diluted 1:25 in RPMI-1640 medium for 0.1 mg/ml final concentration). The tissue was agitated for 30 s using a wide-bore pipette tip and incubated at 37 °C for another 6 min. The supernatant from each digest was then split between two tubes each containing 200 µl FACS buffer (10% FBS in PBS). The tubes were spun at 17 000 xg, 4 °C for 45 min and the pellets solubilised with three cycles of 20 s agitation and 5 min incubation on ice in 80 µl digitonin lysis buffer. The tubes were then centrifuged again (17 000 xg, 20 min, 4 °C). The two supernatants from each tissue were pooled and used immediately or stored at −80 °C.

In later experiments, red blood cells were removed from highly vascular tissue samples (liver, spleen, lung) after the digested tissue was added to the stopping buffer. The samples were washed in PBS and each diluted to a total volume of 9 ml in PBS in a 15 ml tube. Approximately 5 ml Histopaque-1077 (Sigma-Aldrich) was slowly pipetted underneath the sample and the tubes were centrifuged (3000 rpm, 15 min, 4 °C, slow ramp). Cells from the interface of the two layers were isolated, washed in PBS and re-entered the normal protocol at the 45 min centrifugation step.

4.2.5 SDS-PAGE

Gels for protein electrophoresis were prepared as 4% stacking/12% separating polyacrylamide/bis-acrylamide (37.5:1). Standards were MagicMark XP Western Protein Standard (Thermo Fisher) and BLUeye Prestained Protein Ladder (GeneDireX). Samples were mixed 1:1 with 2x Laemmli sample buffer (125 mM TrisCl pH 6.8, 4% SDS, 20% glycerol, 0.02% w/v bromophenol blue ± 5% β-mercaptoethanol (BME)), heated at 92 °C for 2 min if applicable and loaded into wells. Lysates were boiled in all cases except where TAP2 was the target protein, since boiling can cause

aggregation of transmembrane proteins, and has been shown to do so in the case of chicken TAP2. Electrophoresis was performed at 200 V for 1 h.

4.2.6 Coomassie staining

For immediate visualisation of proteins, gels were incubated rocking at room temperature in 45% methanol, 45% water, 10% acetic acid with Coomassie Brilliant Blue (ThermoFisher Scientific) until the gel was dark blue and then washed repeatedly in destaining solution (45% methanol, 45% water, 10% acetic acid) to remove the background stain. InstantBlue (Expedeon) was used for higher sensitivity visualisation of total protein.

4.2.7 Western blotting

Gels were washed on a rocking platform for 10 minutes in transfer buffer (25 mM Tris, 192 mM glycine, 20% methanol, 0.0375% SDS) and transferred to Immobilon-P polyvinylidene difluoride Transfer Membranes (Merck Millipore) using the Mini Trans-Blot Electrophoretic Transfer Cell (BioRad) at 100 V for 70 min. Membranes were then blocked in 5% milk powder solution made up in PBS-T (PBS with 0.05% Tween-20) on a rocking platform for at least 1 h and incubated rotating in primary antibody overnight at 4 °C. Membranes were then washed three times for 10 min in PBS-T, incubated in secondary antibody on a rocking platform for 1 h at room temperature and washed a further three times. Secondary antibodies were either i) polyclonal goat anti-mouse IgG conjugated to horseradish peroxidase (HRP) (Dako) used 1:4000 in 5% milk powder solution made up in PBS-T or ii) TrueBlot (Rockland), a monoclonal rat anti-mouse native IgG, conjugated to HRP and used at 1:1000 in 5% milk powder solution made up in PBS-T (for western blot after immunoprecipitation). After washing, membranes were incubated with Western Lightning ECL Plus reagents (Perkin Elmer) and imaged using the GBOX Chemi XX6 gel imaging system (Syngene).

For experiments where a loading control was required, membranes were washed three times in PBS-T after imaging and the re-incubated with 1:8000 anti-GAPDH (6C5, Invitrogen) or 1:4000 anti-β actin (Sigma Aldrich) for 1 h at room temperature. After incubation membranes were washed, stained with secondary antibody and imaged as before.

4.2.8 Antibody screening

When screening multiple antibodies, 90 µl membrane-enriched cell lysate (5.4×10^7 cell equivalents) was mixed with 90 µl 2x Laemmli sample buffer and loaded into a single wide SDS-PAGE well. After electrophoresis, transfer and blocking, the membranes were cut into up to 18 strips, each of which was labelled in pencil and separately incubated in a different primary antibody in a bijou tube. The first wash after incubation with the primary antibodies was done in separate

containers to reduce any transfer of antibodies between strips, but subsequent washes and incubation in the secondary antibody were performed with strips in a single container for efficiency. The strips were re-assembled in a plastic pocket between two strips of double-sided tape prior to addition of the ECL reagents.

Alternatively, some screens were performed using the Mini-PROTEAN II Multiscreen Apparatus (Bio-Rad) to apply various antibodies to a single membrane instead of cutting the membrane into strips.

4.2.9 Co-immunoprecipitation

Crosslinking antibody to Protein G beads

Crosslinking was performed according to a protocol by Dr. Nicola Ternette (University of Oxford) published in Purcell et al. (2019). At each stage, beads were pelleted with a 12-20 s pulse in a benchtop centrifuge.

Protein G Sepharose 4 Fast Flow resin (GE Healthcare), supplied in 20% ethanol, was washed three times in PBS and resuspended back to a 50% slurry in PBS. F1-3 (α-TAP2) was added to the beads and incubated rotating for at least 1 h at room temperature. The antibody was removed and retained. Beads were washed three times in PBS and three times in 0.2 M triethanolamine (pH 8.3), then incubated for 1 h rotating at room temperature in 10x packed bead volume of crosslinking solution (40 mM dimethylpimelimidate in 0.2 M triethanolamine (pH 8.3)). The crosslinking solution was removed and the reaction terminated with two washes in ice-cold 0.2 M TrisCl (pH 8). The beads were then incubated on ice in 10x packed bead volume 0.2 M TrisCl (pH 8) for 30 min, washed three times in PBS and resuspended back to 50% slurry in PBS. Beads with bound antibody were used immediately or stored at 4 °C for future use.

Immunoprecipitation (IP)

Immunoprecipitation was performed according to Walker et al. (2011).

Membrane-enriched cell lysate was added to the bead-antibody conjugate and incubated overnight at 4 °C. The beads were briefly spun down as before and the supernatant removed and retained for loading as a control. Any remaining lysate was removed with three washes in ice-cold digitonin IP wash buffer (1 volume digitonin lysis buffer: 9 volumes 150 mM NaCl, 50 mM TrisCl (pH 8)). Proteins were eluted from the beads in 2x Laemmli buffer (± 5% BME) by rotating at room temperature for 15 min and analysed by western blot, coomassie staining or proteomics.

4.2.10 Proteomics

Proteins pulled down during immunoprecipitation were analysed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) at the Cambridge Centre for Proteomics. The proteins were supplied either bound to the sepharose beads in 10 mM Tris or as coomassie-stained gel fragments in destaining solution (45% methanol, 45% water, 10% acetic acid). Proteins from gel fragments were run on the LC-MS/MS for 60 min while proteins bound to beads were run for 120 min following an on-bead trypsin digest.

Data was analysed using MASCOT (Matrix Science) with the chicken proteome and a custom reference list containing protein sequences of interest (BF1, BF2, TAP1, TAP2, tapasin, TAPBPR, TAPBPL, calnexin, ERp57) for various haplotypes. The gel fragment proteins were purified by immunoprecipitation from pooled lysates from multiple cell lines while the proteins supplied on beads were from TG21 cells only.

4.2.11 RNA and cDNA preparation

Cell lines

To generate RNA and cDNA from the five cell lines 3.6 - 8×10^6 cells per line were spun down and RNA prepared using the Monarch total RNA miniprep kit (NEB) according to the manufacturer's protocol, including the on-column gDNA digestion. RNA was synthesised using the Maxima H minus cDNA synthesis kit (ThermoFisher) starting with 1 µl RNA per cell line.

Ex vivo tissue

Tissue was stored in RNAlater (Sigma-Aldrich) after extraction as described in section 4.2.4. RNA was extracted from 20-30 mg tissue using the Nucleospin RNA Mini kit (Machery-Nagel) and 500 ng RNA was used in a cDNA synthesis reaction using the Maxima H minus cDNA synthesis kit (ThermoFisher).

4.2.12 Primer design, PCR, cloning and sequencing

Primer sequences and details of thermocycling protocols are in appendices A.1 and A.2.

Primers to amplify TAPBPR were designed to cover the majority of the coding sequence and were screened in multiple combinations using the thermocycling protocol PRL_CDNA, with the primer pair yielding the longest specific product selected. Primers for TAPBPL had previously been designed by Eve Doran (University of Cambridge) but a PCR had never been successfully optimised. The primers were re-screened using a touchdown protocol (BPL_TD40) and the primer

pair yielding the longest product which amplified with high efficiency was selected.

TAPBPR was amplified for sequencing from cDNA using the primers TAPBPR_F1/TAPBPR_R1 and the thermocycling protocol PRL_CDNA. TAPBPL was amplified using the primers CFC/VAS with the touchdown thermocycling protocol BPL_TD40.

Cloning and Sanger sequencing were performed as described in sections 3.2.4 and 3.2.5.

4.2.13 qPCR

Primers for qPCR were designed using the PrimerQuest tool from IDT. Three TAPBPR primer pairs and three TAPBPL primer pairs were screened, with each tested on two different cDNA samples from different birds and different tissues. One pair was selected for each gene based on the presence of highly specific, efficient amplification of the product from both samples as assessed by agarose gel electrophoresis.

The primers selected were TAPBPR_a3_F/TAPBPR_a3_R and TAPBPL_a9_F/TAPBPL_a9_R for TAPBPR and TAPBPL respectively. Both pairs span intron 2 in their respective genes to limit the effect of any possible gDNA contamination. B_actin_F/B_actin_R and GAPDH_F/GAPDH_R were used to amplify the housekeeping genes β-actin and GAPDH.

qPCR was performed using the Luna Universal qPCR master mix (NEB) with all reactions in triplicate. Thermocycling was performed according to the manufacturer's protocol on the Applied Biosystems ViiA 7 platform. LinRegPCR (Ruijter et al., 2009) was used to calculate PCR efficiencies and template concentrations.

4.3 Results

4.3.1 There is no evidence for the presence of tapasin in Anseriformes or Passeriformes

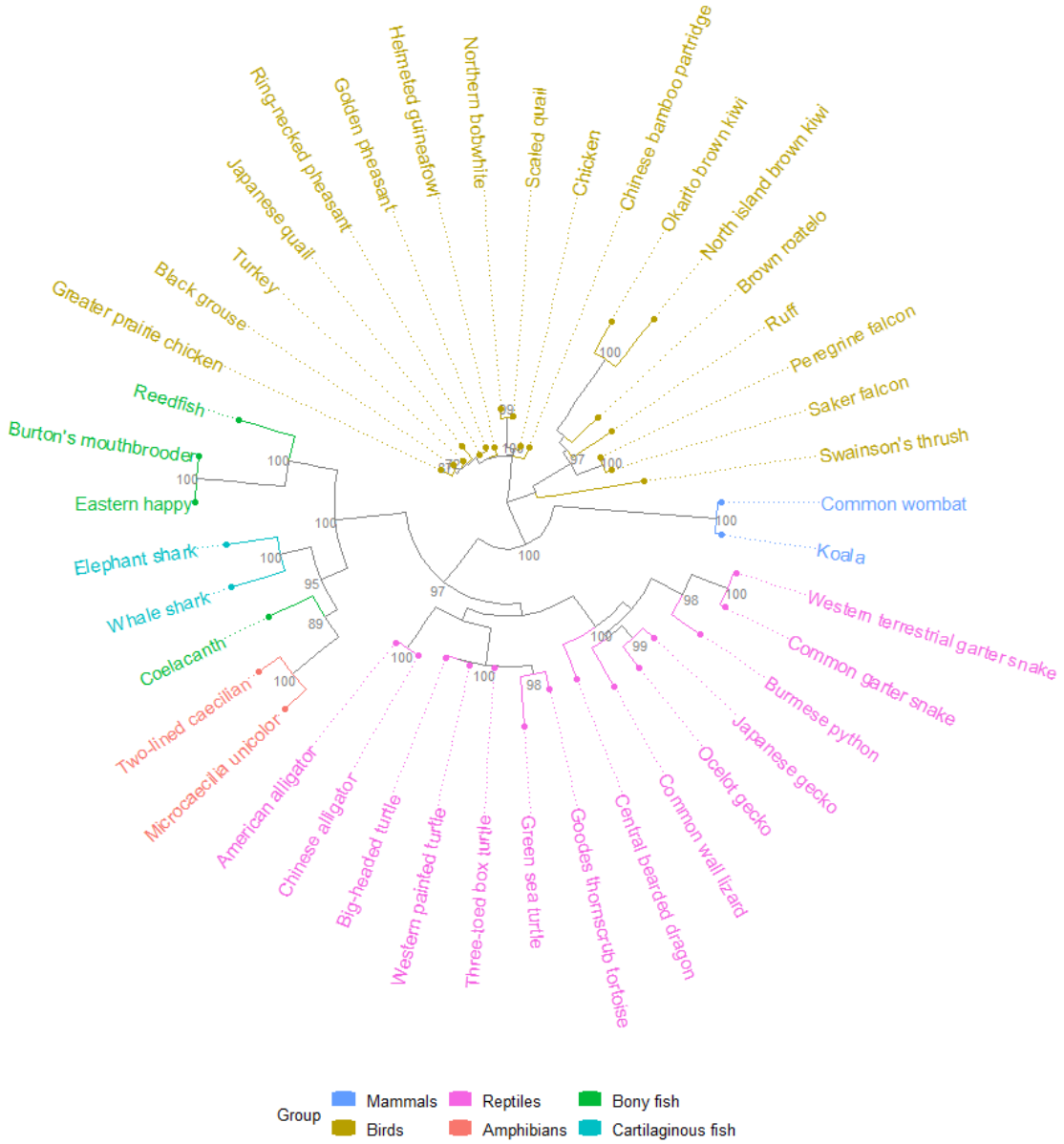


Figure 4.3: **Maximum-likelihood tree of sequences with homology to chicken tapasin as reported by BLASTP.** Maximum-likelihood tree based on the JTT matrix-based model of genome annotations retrieved by BLASTP with chicken tapasin. Bootstrap values after 100 replicates are displayed if the clade was supported in at least 70% of replicates. Tree constructed in MEGA X and visualised in R using the ggtree package.

BLASTP with chicken tapasin (BAG69418.1) against the complete GenBank database returned 41 genomic annotations from 28 different species and 59 direct submissions from 14 different species. A single sequence per species was collected, with direct submissions taken in preference to genome annotations. Tapasin has been well studied in Galliformes and thus a few species contributed a large

number of sequences to the 100 returned by the BLASTP search, leading to a smaller final dataset.

The tree in figure 4.3 corresponds roughly to the accepted phylogeny of the vertebrates, in that species within the same class are grouped together. Relationships between the classes are not entirely as expected, although this is not surprising given that the classes are not equally represented in the dataset and not every annotation or submission was full length. These trees are not intended to give a completely accurate indication of the evolution of these genes, but simply to visually represent the range of taxa in which they have been identified.

It should be noted that the absence of placental mammals in figure 4.3 does not contradict the well-known presence of tapasin in this lineage. The implication of the absence of placental mammals in the output from the BLASTP search (query=chicken tapasin; search set=all NCBI GenBank non-redundant protein sequences) is that the tapasin genes in placental mammals have diverged from that of the chicken sufficiently to not be within the top 100 most similar sequence records to chicken tapasin.

Notably absent from the tree are any species in the Anseriformes or Passeriformes. Anseriformes are closely related to Galliformes, of which the chicken is a member, so it would be surprising if tapasin was present in this group but not picked up in a BLASTP with chicken tapasin. A specific search within Anseriformes confirmed this observation, which was previously reported by Magor et al. (2013).

In order to investigate whether Passeriformes contained divergent tapasin gene sequences (like placental mammals), or no tapasin gene at all, BLASTP searches against Passeriformes with both chicken tapasin and the partial Peregrine falcon (*Falco peregrinus*) tapasin sequence (XP_027633192.1, chosen because falcons are the sister group to passerines) as queries were performed. Possible tapasin loci from the same five species (table 4.1) were returned in both searches. E values with the falcon sequences as the query are generally larger than those obtained with the chicken query because the sequence length is shorter. It is notable that none of the 100 sequences returned by BLASTP with chicken tapasin against the entire database had an E value greater than $3e^{-52}$, despite these results including sequences from all vertebrate classes. Only one of the potential passerine tapasin sequences has an E value below this, despite passerines being much more closely related to chickens than the majority of the species in the BLASTP results. None of these sequences are more similar to chicken TAPBPR or TAPBPL than they are to tapasin.

Name	Accession	Species	E value (Peregrine falcon)	E value (Chicken)
tapasin	XP_032942880.1	Swainson's thrush (*Catharus ustulatus*)	$4e^{-44}$	$1e^{-72}$
tapasin-like	XP_014118264.1	Tibetan ground tit (*Pseudopodoces humilis*)	$6e^{-17}$	$1e^{-46}$
tapasin-like	XP_027488771.1	White-ruffed manakin (*Corapipo altera*)	$8e^{-44}$	$1e^{-40}$
tapasin-like	XP_030089733.1	Atlantic canary (*Serinus canaria*)	$6e^{-31}$	$2e^{-27}$
tapasin-like (low quality protein)	XP_030825893.1	Small tree finch (*Camarhynchus parvulus*)	$2e^{-08}$	$4e^{-24}$

Table 4.1: **Top hits from BLASTP against Passeriformes with Peregrine falcon or Chicken tapasin as the query.** The low quality coding sequence annotated in the small tree finch had two bases inserted relative to the genomic sequence to keep the CDS in-frame; these could represent genuine frame-shift mutations rendering this locus a pseudogene.

Of these five loci, only those in the thrush and ground tit were on scaffolds with other genes. Both were flanked by SLC39A7 and PFDN6, which are found close to tapasin in the human MHC. This does not, however, exclude the possibility that the sequences returned by BLAST represent a degraded, non-functional region where the species' ancestors had a functional tapasin gene. If the gene became non-functional in passerines, selection to retain the structure of the protein would be eliminated and the region could diverge to the point where open reading frames were not detected in most species, and were highly divergent from the functional gene of chickens in others.

An alternative explanation is that tapasin is present in, or close to, MHC regions in passerines, but that these regions are poorly sequenced and so only a few genomes have annotated reading frames that correspond to tapasin. This seems unlikely given that the best-sequenced genomes, including the zebra finch, do not appear in the BLASTP hits.

A final explanation could be that some passerine species, including some or all of those listed in table 4.1, have a functional, if poorly conserved, tapasin gene, but that it has been lost in other lineages. However, if this was the case, it might be expected that small groups of closely related species would have a functional gene, rather than the pattern seen here where each species in table 4.1 is from a different taxonomic family and the families are distributed throughout the passerines.

It seems most likely that a functional tapasin gene is not present in passerines.

4.3.2 TAPBPR is present in most major lineages but not in Passeriformes

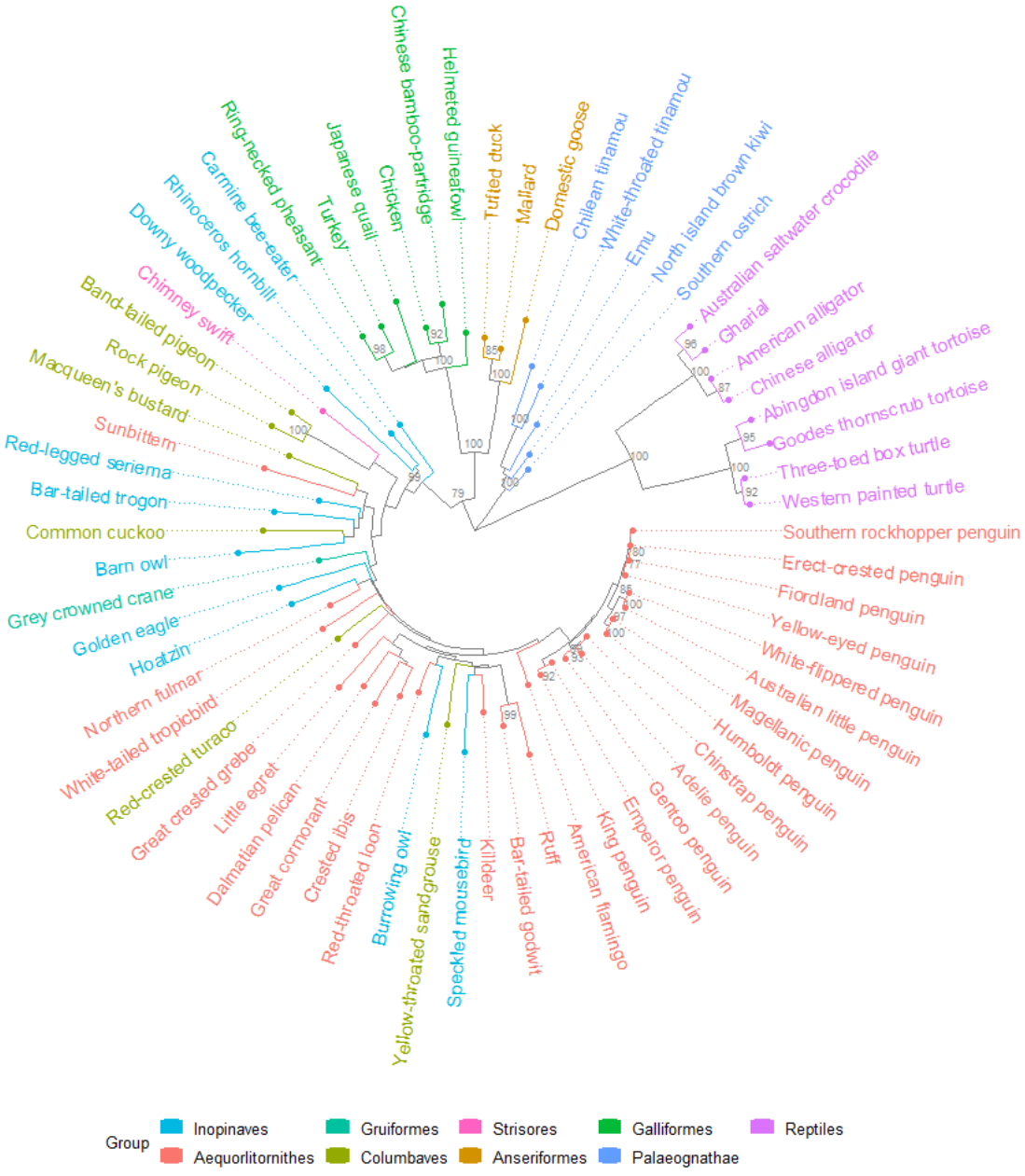


Figure 4.4: **Maximum-likelihood tree of sequences with homology to chicken TAPBPR as reported by BLASTP.** Maximum-likelihood tree based on the JTT matrix-based model of genome annotations retrieved by BLASTP with chicken TAPBPR. Bootstrap values after 100 replicates are displayed if the clade was supported in at least 70% of replicates. Tree constructed in MEGA X and visualised in R using the ggtree package.

BLASTP with chicken TAPBPR (NP_001026543.1) against the complete GenBank database returned 60 genomic annotations from 41 different species and 40 direct submissions from 25 different species. A single sequence per species was collected, with direct submissions taken in preference to genome annotations.

TAPBPR seems to be widespread and relatively well-annotated in birds, resulting in a list of hits from a BLASTP search with chicken TAPBPR which does not extend, phylogenetically, beyond the Sauropsida (BLASTP only returns the 100 highest-scoring hits). The maximum likelihood tree does not neatly reflect the taxonomic groups described in Prum et al. (2015), however all clades with bootstrap values > 70 (annotated clades in figure 4.4) do correspond to the expected groupings.

Likely TAPBPR sequences are present in all the major bird lineages described in Prum et al. (2015) (Palaeognathae, Galliformes, Anseriformes and five groups of neoaves, each forming successive sister groups to the rest of the neoaves). However, a notable absence is the Passeriformes, which make up a major clade within the Inopinaves, having undergone an adaptive radiation.

A specific BLASTP search against Passeriforme genomes, using chicken TAPBPR as the query, returned seven plausible sequences from six species (including two predicted isoforms at the same locus in the Ground tit). The sequences from four of the species were also returned in a search with chicken tapasin as the query, and match more closely to chicken tapasin than they do to chicken TAPBPR. The two top hits in this search, uncharacterised protein LOC116782524 (XP_032535166.1) from the Lance-tailed manakin (*Chiroxiphia lanceolata*) and tapasin-related protein-like (XP_027741005.1) from the Willow flycatcher (*Empidonax traillii*), do, however, match more closely to TAPBPR ($E = 2e^{-38}$ and $E = 6e^{-36}$ respectively). Both are annotated as 'low quality proteins' and have notes attached confirming that "the sequence of the model RefSeq protein was modified relative to its source genomic sequence to represent the inferred CDS". In the manakin sequence, two bases were inserted and one was substituted to 'correct' frameshifts and stop codons in what was thought to be a protein coding gene. In the flycatcher sequence, one base was inserted and one substituted relative to the genomic sequence.

Both of these annotations have flanking genes that closely resemble the genomic regions flanking known TAPBPR genes in other species. The sequence of genes seen in chickens (FAM162A, KPNA1, FBXO40, TAPBPR, SCNN1A, VAMP1, MRPL51, NCAPD2) is seen in both the manakin and flycatcher assemblies, suggesting that the low quality protein prediction is based on the locus where an ancestor of these species *did* have a TAPBPR gene, which has now degraded. Many of the same genes are seen to be flanking TAPBPR in humans.

4.3.3 TAPBPL is present in a range of taxa but not in Placentalia, Anseriformes or Passeriformes

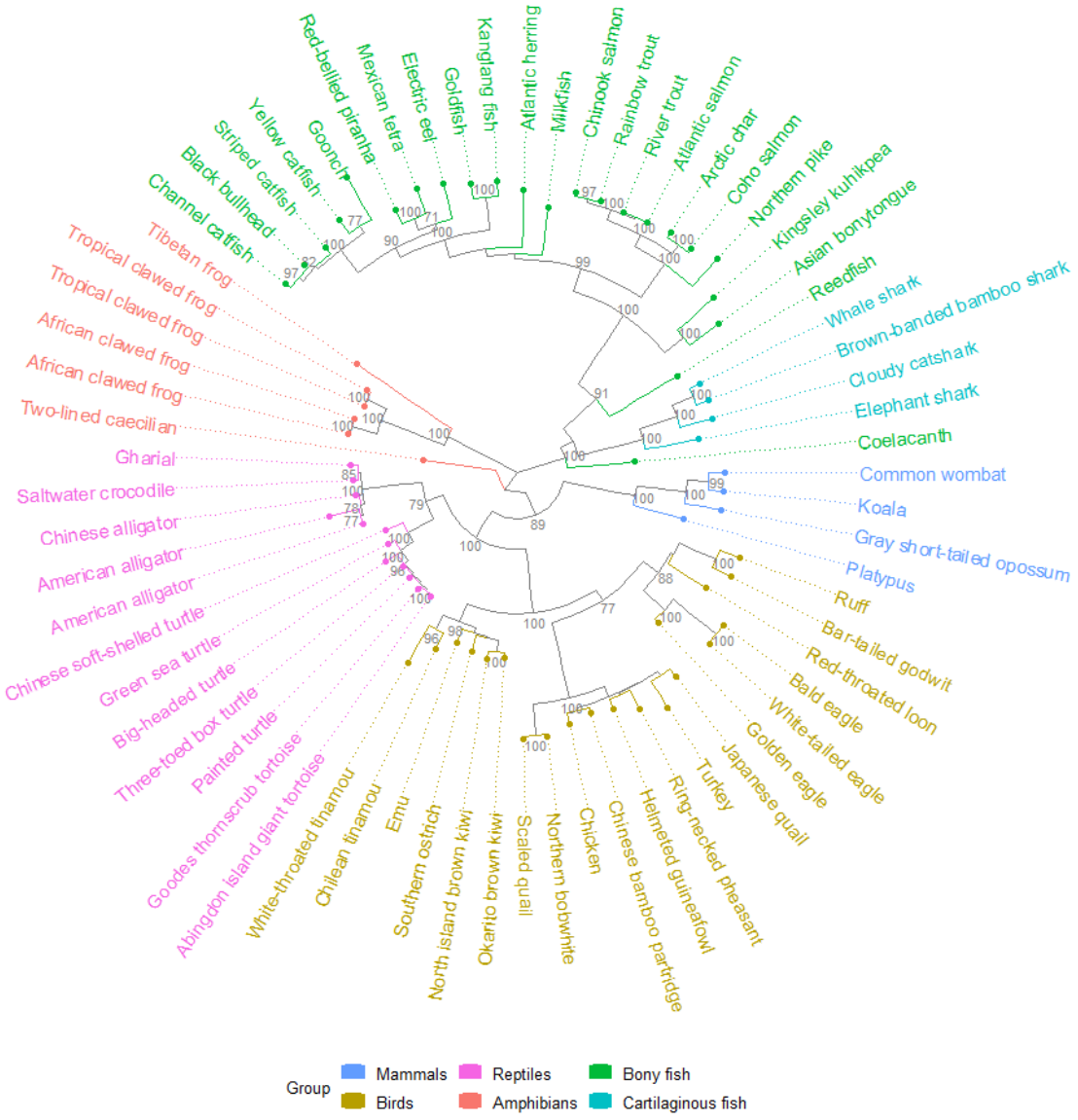


Figure 4.5: **Maximum-likelihood tree of sequences with homology to chicken TAPBPL as reported by BLASTP.** Maximum-likelihood tree based on the JTT matrix-based model of genome annotations retrieved by BLASTP with chicken TAPBPL. Bootstrap values after 100 replicates are displayed if the clade was supported in at least 70% of replicates. Tree constructed in MEGA X and visualised in R using the ggtree package.

BLASTP with chicken TAPBPL (XP_024997865.1) against the complete GenBank database returned 85 genomic annotations from 56 different species and 15 direct submissions from 13 different species. A single sequence per species was collected, with direct submissions taken in preference to genome annotations, including salmonids where duplications of TAPBPL have been reported (Grimholt, 2018).

The tree topology (figure 4.5) corresponds largely to the accepted phylogeny of the vertebrates, however the coelacanth would be expected to share a more recent common ancestor with the terrestrial tetrapods than the fish, and the cartilaginous fish should be basal to all the bony vertebrates. The monophyletic clade of fish TAPBPL genes seen here is well supported and is seen consistently with other tree-drawing algorithms.

Sequences with homology to chicken TAPBPL are found in all major gnathostome lineages. It is notable that while predicted TAPBPL sequences were found in monotremes and marsupials, the search didn't return any sequences in placental mammals as reported in Grimholt (2018). A further BLASTP search with koala TAPBPL (XP_020862690.1) against the full database returned a very similar list of monotremes, marsupials and non-mammalian vertebrates, but still no placental mammals. BLASTP with koala TAPBPL specifically against placental mammals returned sequences with no more than 32.6% identity to the query over >95% of the query length, compared to the 50% identity between koala TAPBPL and sauropsid TAPBPL sequences detected with the same query length coverage. A koala TAPBPR sequence was obtained by BLASTP against marsupial with human TAPBPR as the query. Using koala TAPBPR as the query in a BLASTP search against placental mammals returned a list which overlapped to a large extent with the list obtained in a search against placental mammals with koala TAPBPL as the query. However, the identities were 50-60% when the query was koala TAPBPR, compared to <33% with koala TAPBPL, indicating that TAPBPL is indeed absent from placental mammals while TAPBPR is widely present. The same comparison was made with BLASTP searches against placental mammals with chicken TAPBPR and TAPBPL, which again returned largely overlapping lists of hits, but with 40-42% identity when the query was TAPBPR and <30% when the query was TAPBPL.

There were also no TAPBPL sequences detected, during the initial search, in Passeriformes or Anseriformes, despite sequences being present in ratites, Galliformes and several groups of neoaves. A specific search using chicken TAPBPL as the query against Anseriformes returned TAPBPR sequences with <30% identity over <75% of the query length followed by other, less similar, proteins such as the immunoglobulin tumor antigen CD276. This suggests that TAPBPL is absent from Anseriformes. A search with chicken TAPBPL specifically within Passeriformes identified tapasin-like sequences in the White-ruffed manakin (*Corapipo altera*) (XP_027488771.1) and Atlantic canary (*Serinus canaria*) (XP_030089733.1), but with low sequence identity and query coverage (32% identity and 32% coverage, ($E = 1e^{-16}$) for the manakin; 25% identity and 30% coverage, ($E = 6e^{-9}$) for the canary). Both of these sequences were also returned by searches with with chicken TAPBPR and tapasin within Passeriformes, with the best match occurring when the query was chicken tapasin (manakin: 48% identity, 36% coverage, ($E = 1e^{-40}$); canary: 41% identity, 36% coverage, ($E = 2e^{-27}$), previously reported in table 4.1). There were no other relevant se-

quences returned from a search with chicken TAPBPL in passerines; the next closest hits were CD276 and V-set domain-containing T-cell activation inhibitor 1 in various species. Overall, there is no evidence of a functional TAPBPL protein in either Anseriformes or Passeriformes.

4.3.4 Chicken TAPBPR is less polymorphic than chicken tapasin in cell lines containing different MHC haplotypes

TAPBPR and TAPBPL were sequenced from cDNA extracted from the cell lines TG21, IS2, TG12, TG15 and IS19.

For TAPBPR, six to eight clones derived from an amplicon which covered the majority of the coding sequence were analysed per cell line, with each line found to contain a single allelic sequence. IS2 and TG12 contained identical sequences, with the TG15, IS19 and TG21 sequences differing from this sequence at four, six and seven nucleotide positions respectively, of which one, three and five were non-synonymous changes (appendix A.10). The degree of polymorphism was compared between chicken tapasin and TAPBPR using the mean pairwise number of synonymous changes per synonymous site (K_s) and mean pairwise number of non-synonymous changes per non-synonymous site (K_a), which corrected for the different numbers of sequences available. For tapasin, 15 unique sequences were identified from the standard serological haplotypes (GenBank accession numbers AB426141-AB426154, AM403067.1 and AM403069.1), with tapasin from the B8 haplotype excluded since it was identical to tapasin in B11.

	Unique alleles	Mean K_s	Mean K_a
tapasin (all)	15	0.019077	0.005814
tapasin (B2, B12, B15, B19, B21)	5	0.01608	0.00497
TAPBPR	4	0.00685	0.002833

Table 4.2: **Sequence diversity of Chicken tapasin and TAPBPR alleles.**

Pairs of chicken tapasin alleles differed at approximately twice as many coding and non-coding positions as chicken TAPBPR alleles in this analysis, which was limited in its power by the small number of TAPBPR alleles available. The results remained similar when only the tapasin genes present in the five cell lines were considered (table 4.2). It is therefore more likely that any variation in TAPBPR functionality or dependence between MHC haplotypes would be more significantly determined by variation in MHC class I than in TAPBPR. Since TAPBPR, unlike tapasin, is not in linkage disequilibrium with the classical MHC molecules, it was not predicted to exhibit significant polymorphism due to coevolution with BF2. Furthermore, since TAPBPR appears to have a more subtle, 'fine-tuning' role in antigen presentation and tapasin, it may not be a particularly effective target for viral immune evasion strategies. If TAPBPR is not being targeted by viruses it may be under less selective pressure to diversify than other components of the antigen

processing machinery which have more significant impacts if disrupted.

4.3.5 Chicken TAPBPL is almost monomorphic in cell lines containing different MHC haplotypes

For TAPBPL, four to six clones derived from an amplicon which covered the majority of the coding sequence were analysed per cell line, with each line found to contain a single allelic sequence (appendix A.11). IS2, TG15, IS19 and TG21 all contained the same TAPBPL sequence, while the sequence in TG12 differed at two synonymous and one non-synonymous positions. The TG12 allele also had a single codon deletion in a run of repeated CTG lysine codons resulting in six tandem CTG repeats rather than seven. The presence of a single allele in four out of the five cell lines, and minimal amino acid-level polymorphism in TG12, suggests that TAPBPL is not intrinsically functionally variable. This does not, however, preclude the possibility that its function varies between birds with different MHC haplotypes as a result of variation in the properties of different MHC alleles.

4.3.6 Screening of antibodies raised against the C-terminal peptide of chicken tapasin, TAPBPR and TAPBPL

Antibodies against the C-terminal peptide of tapasin were inferior to antibodies against the extracellular domain

Fifteen antibodies raised against the C-terminal peptide of tapasin (C-AARPKEETKKSQ) with names in the format 19-52-x were screened and compared to antibodies against the extracellular domain which had been used previously in the lab, with names in the format 11-46-x. Membrane-enriched cell lysate from the TG21 cell line was used in all antibody screens.

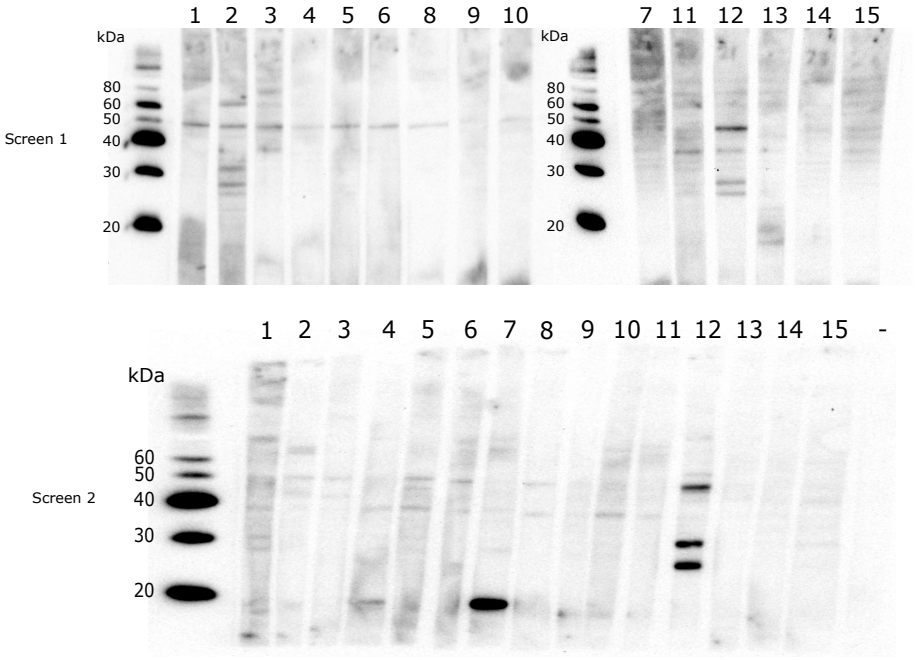


Figure 4.6: **Screening 19-52 antibodies.** The screen was performed twice to control for batch variation in lysates. All samples were reduced and boiled prior to loading. Labels above lanes correspond to the numeric name of the antibody in the format 19-52-x.

The expected size of the chicken tapasin protein is 46 kDa. Several antibodies stained a protein of approximately this size, but none stained in a particularly strong, consistent or specific manner (figure 4.6).

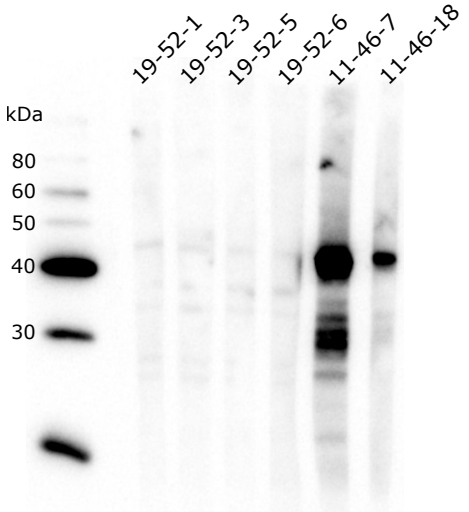


Figure 4.7: **11-46- antibodies against the tapasin extracellular domain stain more strongly than 19-52 antibodies against the C-terminal peptide.**

Two antibodies against the extracellular domain, 11-46-7 and 11-46-18, showed much stronger staining than four of the anti-C-terminal peptide antibodies (figure 4.7). 11-46-18 was used in subsequent experiments because of the large volume available and the specificity of staining relative to 11-46-7. This antibody has been used extensively in prior work which has established its

specificity for tapasin (C. Tregaskes, unpublished).

19-53 antibodies against the C-terminal peptide of chicken TAPBPR gave strong, consistent staining

Twenty-six antibodies raised against the C-terminal peptide of chicken TAPBPR (C-AAEPKPEQLLTASE) were screened against membrane-enriched cell lysate from the TG21 cell line. The antibodies were numbered with the prefix 19-53.

The majority of the 19-53 antibodies stained strongly at 62 kDa (figure 4.8). This protein was larger than the expected size of 48 kDa but was thought to be TAPBPR because the apparent molecular mass (AMM) of the stained protein was so consistent between the different antibody clones. Furthermore, a BLASTP analysis against the chicken protein database showed that there were no likely alternative candidate proteins containing a highly similar peptide; even those proteins containing a peptide with a lower degree of similarity had molecular weights significantly higher or lower than 62 kDa (table 4.3). Sequencing of TAPBPR cDNA from all five cell lines (section 4.3.4) showed no indication of intron retention relative to genomic annotations which could have potentially explained this observation.

It was therefore hypothesised that chicken TAPBPR might be post-translationally modified. The presence of ubiquitin moieties might be expected to give a 'ladder' of bands, which was not observed in the screen, although mono-ubiquitination might give an AMM shift of approximately the observed size. The observed shift in AMM from the expected size was larger than would normally be associated with phosphorylation of a protein, unless multiple sites were being phosphorylated simultaneously. Furthermore, highly phosphorylated cytoplasmic regions might be expected to stain more variably than was observed, given that the antibodies were raised against a non-phosphorylated peptide. Glycosylation therefore seemed the most likely modification to explain the observed staining.

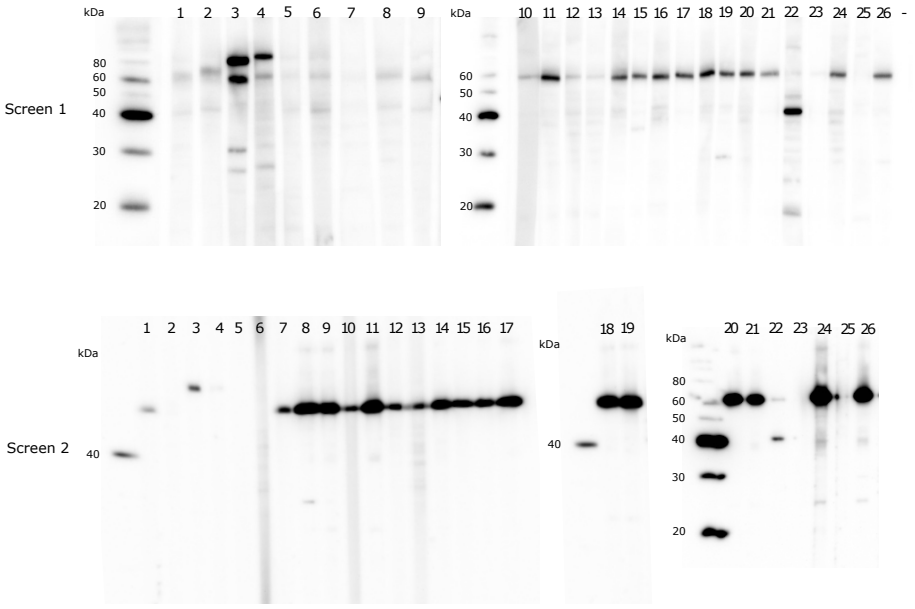


Figure 4.8: **Screening 19-53 antibodies.** The screen was performed twice to control for batch variation in lysates. Screen 2 was performed using the Mini-PROTEAN II Multiscreen Apparatus (Bio-Rad). All samples were reduced and boiled prior to loading. Labels above lanes correspond to the numeric name of the antibody in the format 19-53-x.

Gene	E value	Protein size (kDa)
tapasin-related protein	$7e^{-8}$	48
sperm flagellar protein 2	2.6	163
E3 ubiquitin-protein ligase MYCBP2	2.6	506
fibroblast growth factor-binding protein 3	3.7	18
serine/threonine-protein kinase greatwall	3.7	98
semaphorin-6A	5.3	107
E3 SUMO-protein ligase RanBP2-like	7.5	339

Table 4.3: **Results of BLASTP with TAPBPR C-terminal peptide.** No obvious candidates which would be likely to give and consistent band at the observed AMM were identified. A lower E value indicates higher sequence identity.

19-54 antibodies stain various combinations of bands at three distinct AMMs

Fifteen antibodies against the C-terminal peptide of TAPBPL (C-QVRSTTKPKPY) were screened. Staining was variable between antibodies and between screens performed with different methodologies and cell lysate batches, although the main patterns were conserved (figure 4.9). The majority of the antibodies stained a subset of three distinct protein species, all of which can be seen in lanes stained with 19-54-4 and 19-54-11. The band with the highest AMM was approximately 48 kDa which was the expected size of TAPBPL. Two bands with AMMs of approximately 40 and 44 kDa were also present. No alternative splicing, which could explain the presence of smaller protein

species, was observed in cDNA sequencing from the cell lines (appendix A.11).

While there were no proteins identified by BLASTP which contained a peptide with significant overall identity to C-QVRSTTKPKPY, several proteins, some of which were of similar sizes to TAPBPL, contained either a TTKPK or QVRST motif. However, while relative band intensity is variable between lanes stained with different antibodies, very few contain a single band. This observation is inconsistent with the hypothesis that some of the 19-53 antibodies are specific to the TTKPK or QVRST motifs only.

19-54-11 was selected for subsequent experiments, since there was insufficient evidence to determine which, if any, of the bands stained by the 19-54 antibodies were TAPBPL. 19-54-11 also provided the strongest consistent staining across both screens.

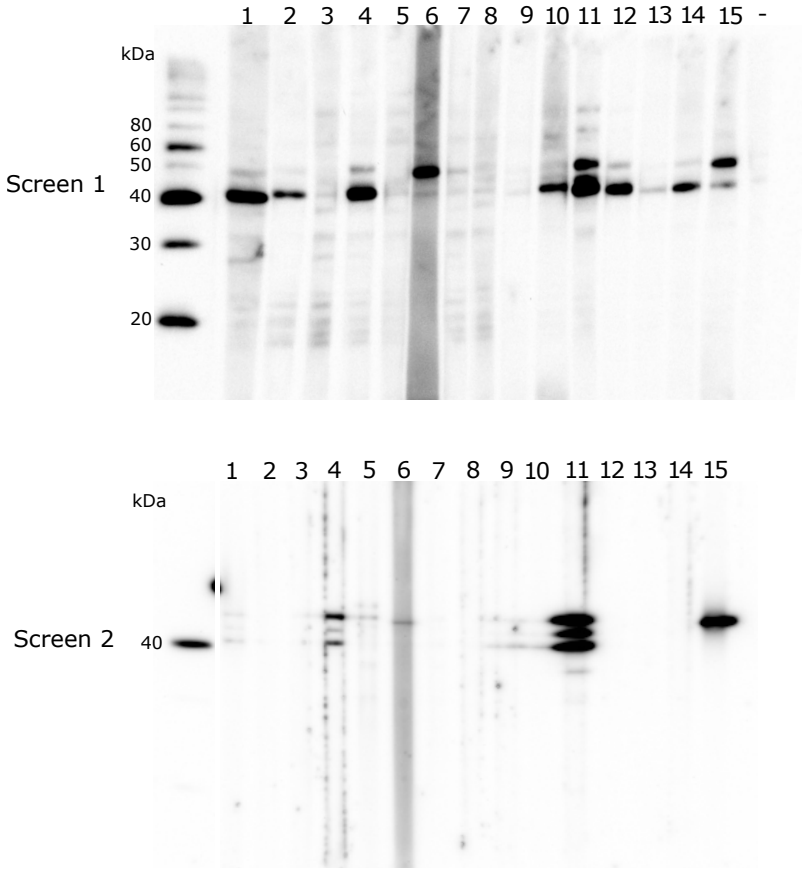


Figure 4.9: **Screening 19-54 antibodies.** The screen was performed twice to control for batch variation in lysates. Screen 2 was performed using the Mini-PROTEAN II Multiscreen Apparatus (Bio-Rad) and two samples between the ladder and 19-54-1 which were stained for a different target protein have been removed. All samples were reduced and boiled prior to loading. Labels above lanes correspond to the numeric name of the antibody in the format 19-54-x.

4.3.7 Chicken TAPBPR may be regulated by an N-linked glycan which is absent in human TAPBPR

It was hypothesised that glycosylation of TAPBPR could be causing the unexpectedly high AMM of the bands stained with 19-53 antibodies. While human TAPBPR has no canonical N-glycosylation sites (N-X-S/T) and is not Endo-H sensitive (Boyle et al., 2013; Porter et al., 2014), chicken TAPBPR contains a conserved NNST motif, which could be glycosylated on either of the asparagine residues.

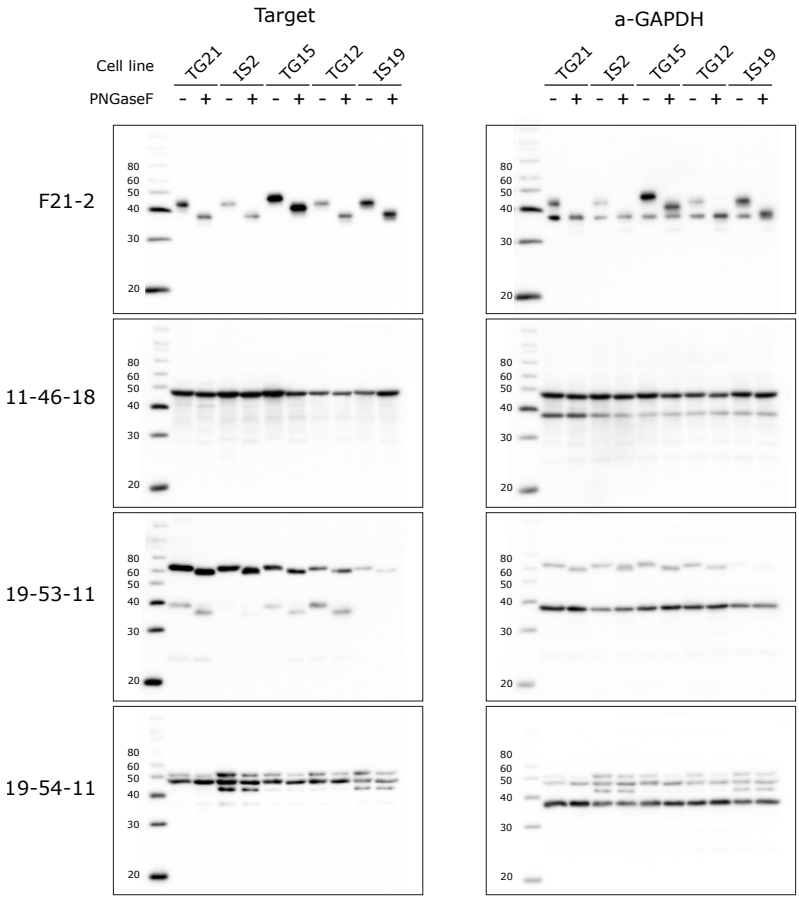


Figure 4.10: **F21-2 and 19-53-11 stain glycosylated protein species.** 1.5×10^6 cell equivalents membrane enriched lysate (boiled, reduced) per lane (1.5×10^5 for F21-2). After incubation with primary antibody against the protein of interest and imaging, the membranes were washed, re-incubated with anti-GAPDH and re-imaged.

MHC class I (F21-2), which is known to have an N-linked glycan, acted as a positive control for the action of PNGaseF in the membrane enriched lysates. Chicken tapasin (stained with 11-46-18) contains two possible N-linked glycosylation sites, but did not appear to be glycosylated in the cell lines. 19-53-11 (putative α-TAPBPR) stained a band at a lower AMM after PNGaseF treatment, suggesting that TAPBPR is glycosylated in chickens. The presence of this glycan may indicate a subtly different function or additional level of regulation for chicken TAPBPR relative to human TAPBPR but cannot explain the difference between the expected molecular mass of chicken TAPBPR (48 kDa) and the AMM of the deglycosylated protein product stained with 19-

53-11 (approximately 60 kDa).

4.3.8 19-54-11 may stain a protein species which is constitutively present in both glycosylated and non-glycosylated forms

It was also noted that some of the cell lines showed a reduction in band intensity of the highest AMM band stained with 19-54-11 (putative α-TAPBPL) after PNGaseF treatment (figure 4.10). This was consistent with a previous experiment which had showed a noticeable decrease in intensity of the highest AMM band and an increase in intensity of the mid-weight band after treatment with PNGaseF (figure 4.11), suggesting that these bands represent differentially glycosylated forms of the same protein species. The presence/absence of a glycan may be correlated, perhaps depending on the cellular localisation of individual molecules, with modification of the C-terminal peptide, since only a few of the 19-54 antibodies stained the mid-weight band. Differential modification of the protein in different cell compartments may explain the constitutive presence of both forms in the cell lines, in contrast to MHC class I and TAPBPR, which seem to be present in a predominantly glycosylated form. It could also indicate a more dynamic system of regulation for this protein. While phosphorylation of the C-terminal peptide was hypothesised to be a mechanism by which different forms of the protein could be differently detected by 19-54 antibodies, a preliminary experiment involving treatment of the cell lysates with Lambda protein phosphatase (NEB) showed no evidence for this (appendix A.17). Nonetheless, the absence of a suitable positive control for the activity of the phosphatase enzyme meant that phosphorylation could not be reliably reported to be uninvolved in regulation of these protein species. Other C-terminal modifications such as ubiquitination should also be considered.

In contrast to the antibody screens, in which the membranes were cut into strips and then re-aligned by hand, this experiment was performed on intact membranes which allowed the AMM of the bands to be more accurately determined. The three bands were identified at approximately 51, 47 and 41 kDa.

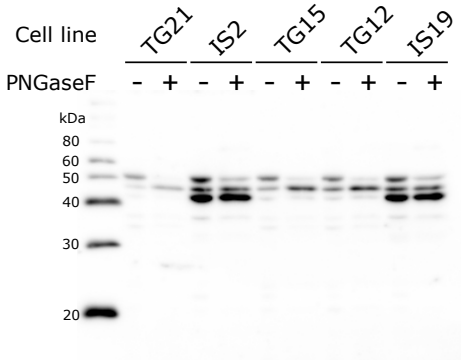


Figure 4.11: **The highest AMM band stained by 19-54-11 is PNGaseF-sensitive.** 1.5×10^6 cell equivalents membrane enriched lysate (boiled, reduced) per lane. After incubation with primary antibody against the protein of interest and imaging, the membranes were washed, re-incubated with anti-GAPDH and re-imaged to confirm that proteins in all lanes were running evenly.

4.3.9 Protein-level expression of tapasin is consistent across cell lines carrying variously tapasin-dependant MHC haplotypes

Since tapasin genes coevolve with class I in the tightly-linked chicken MHC, it was hypothesised that expression of tapasin in different cell lines homozygous at the MHC would reflect the tapasin dependence of the MHC class I alleles present. In multiple experiments tapasin was expressed relatively consistently across the five cell lines. In particular, there was no clear correlation of tapasin expression with the expected tapasin dependencies of the haplotypes in the cell lines (figure 4.12).

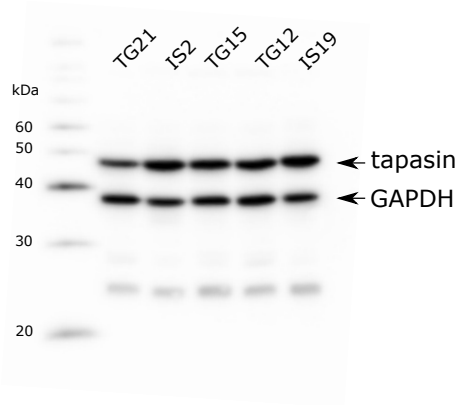


Figure 4.12: **Expression of tapasin is consistent between cell lines carrying different MHC haplotypes.** 1.5×10^6 cell equivalents membrane enriched lysate (boiled, reduced) per lane. After incubation with primary antibody against the protein of interest and imaging (not shown), the membranes were washed, re-incubated with anti-GAPDH and re-imaged to confirm equal loading of lanes. Cell lines are shown in order of predicted tapasin dependence (based on knowledge of correlated characteristics including cell surface expression of class I and peptide binding motif), with TG21 predicted to be the least tapasin dependant and IS19 the most.

4.3.10 Protein-level expression of TAPBPR is somewhat variable between cell lines

Some variation was observed in TAPBPR expression between cell lines (figure 4.13). Except IS19, which had somewhat consistently lower expression than the other lines, variation was not consistent between experiments and did not correlate with the breadth of the peptide binding repertoires presented by the class I genes in the various haplotypes. This suggests that the observed variation is unlikely to be related to inherent haplotype-specific functional variation and instead is dynamic and may be regulated by signals which report the overall condition or immune state of the cell.

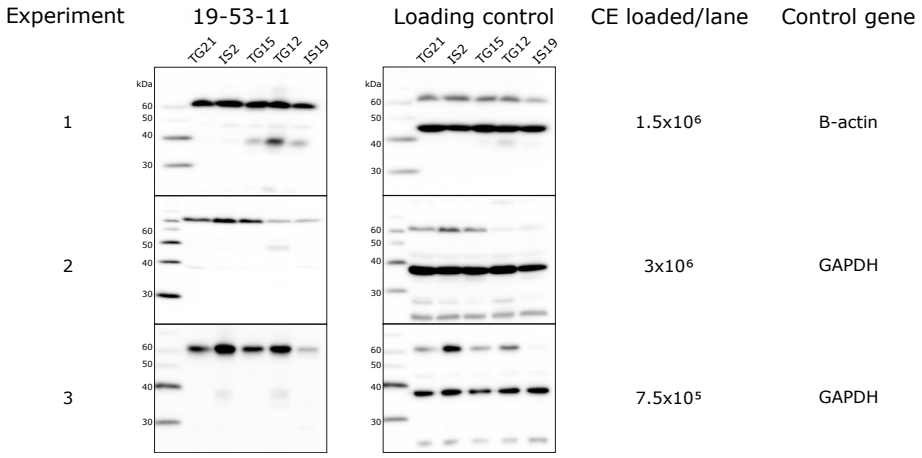


Figure 4.13: **Expression of TAPBPR is somewhat variable between cell lines carrying different MHC haplotypes.** Each experiment used a different batch of cell lysates and gel loading was optimised over the course of five experiments (CE loaded/lane = cell equivalents of lysate loaded per lane).

4.3.11 Protein-level expression of 19-54-11 antigens vary between cell lines

Expression of proteins which stained with 19-54-11 (raised against the C-terminal peptide of TAPBPL) was highly variable. While absolute quantities of protein in a given cell line are difficult to compare between experiments due to reported variation in loading volumes and unreported minor differences such as incubation times, relative expression in different lines is clearly variable between experiments. Furthermore, relative expression of the three protein species with different gel motilities in a given cell line is also variable between experiments. While cell density at harvest was controlled for in production of lysates to give consistent numbers of cell equivalents per microlitre, it is plausible that cells harvested at higher density were exhibiting generalised stress responses, which could be affecting regulation of these proteins. Batch-to-batch variation in media components could also have played an interacting role in determining the condition of the cells at harvest.

The effect of cell density at harvest on the strength of 19-54-11 staining was tested in preliminary work in the IS2 cell line with inconsistent results (described in appendix A.13), suggesting that multiple interacting factors may be involved in the regulation of these proteins.

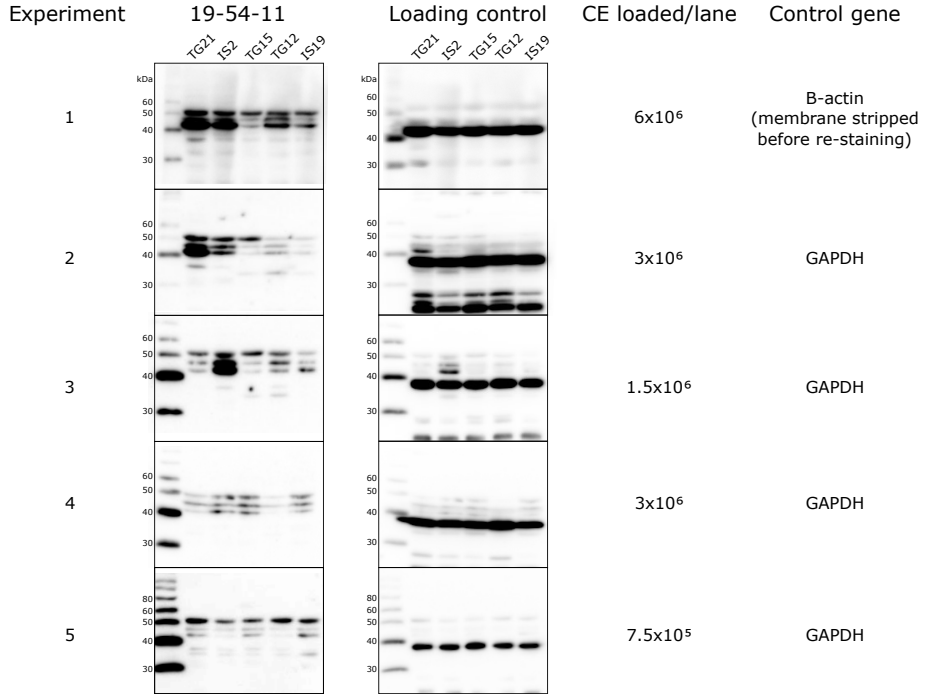


Figure 4.14: **Expression of 19-54-11 targets is highly variable.** Each experiment used a different batch of cell lysates and gel loading was optimised over the course of five experiments (CE loaded/lane = cell equivalents of lysate loaded per lane). Experiments 2-5 used α-GAPDH since the motility of GAPDH was sufficiently different from that of any other protein of interest that membrane stripping prior to re-staining was not required.

4.3.12 Expression of tapasin homologues at the RNA level is variable between tissues

Tissues were retrieved from two line 0 B21 homozygote birds and RNA was extracted from 20-30mg tissue. 500 ng RNA from each sample was subsequently used for cDNA synthesis. cDNA was amplified using primers specific to β-actin, GAPDH, TAPBPR and TAPBPL, which were estimated to have efficiencies of 73%, 89%, 88% and 82% respectively. These estimates were made by fitting a regression line to a subset of data points in the log-linear phase in LinRegPCR and subsequently used in estimates of starting concentrations for the cDNA templates (N_0) (Ruijter et al., 2009).

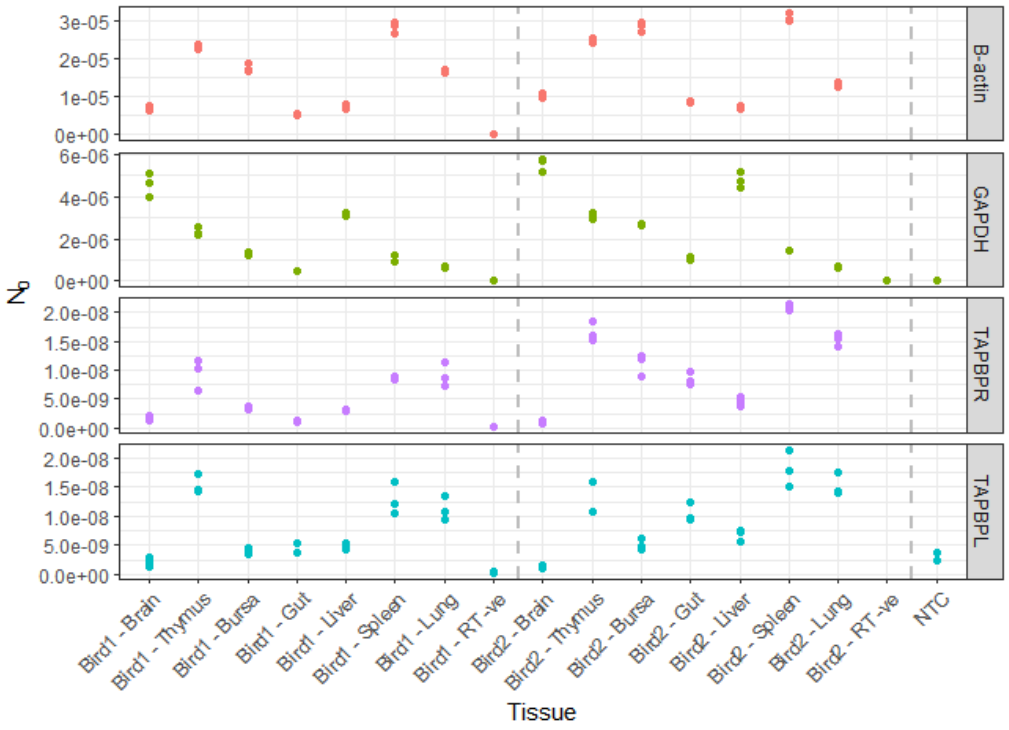


Figure 4.15: **RNA-level expression of TAPBPR and TAPBPL in tissues.** N_0 is an estimate of the original number of copies of the target in the sample. LinRegPCR was used to estimate PCR efficiency and baseline and subsequently calculate N_0. Dashed lines separate results for the two birds and the qPCR no template control (NTC). A reverse transcriptase negative (RT -ve) control for the cDNA synthesis reaction was included for each bird. All reactions were done in triplicate, with each point representing a single reaction well. Missing data points correspond to amplification curves which failed quality control checks, including a lack of amplification, plateau or accurate baseline.

All RT -ve and NTC reactions either gave extremely low N_0 values or failed to amplify sufficiently to calculate N_0 (figure 4.15). The housekeeping genes β-actin and GAPDH were included to ensure that differences observed between birds or tissues were not artefacts of unequal loading of total cDNA. Inconsistent expression of housekeeping genes is a common observation when comparing tissues by qPCR, however the absence of correlation between expression of β-actin and GAPDH indicates that this variation is due to genuine differences in expression of the housekeeping genes, rather than to varying amounts of cDNA present in the tissue samples. Furthermore, the patterns of high/low expression of β-actin and GAPDH between tissues are largely conserved between the two birds, indicating that samples had been prepared in a consistent manner. Triplicates clustered relatively closely together, suggesting that variation in reaction setup was not a major confounding factor. Figure 4.16 shows that the primers were not amplifying non-specific products.

Expression of TAPBPR and TAPBPL varied between tissues by approximately 5-10 fold, which was roughly equivalent to the variation seen in expression of the housekeeping genes. In all tissues, TAPBPR and TAPBPL transcripts were present at very similar levels to one another. While

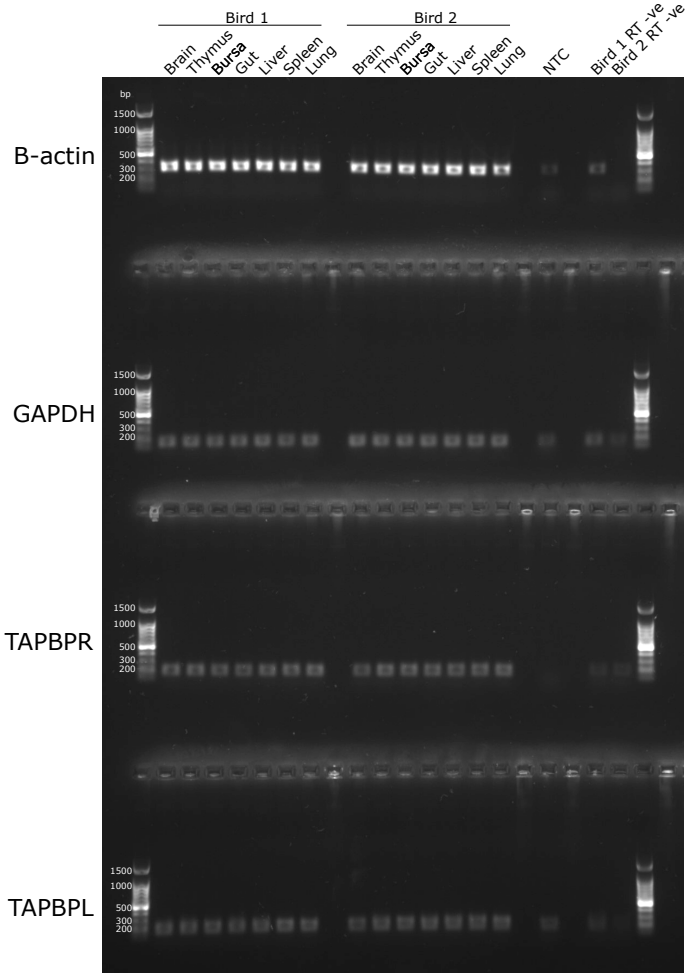


Figure 4.16: **Verification of qPCR primer specificity.** All primer pairs amplified products of the expected size.

expression was roughly correlated between TAPBPR and TAPBPL at the level of whole organs, restriction of either or both proteins to specific cell types could still be present. The brain consistently had the lowest expression of both TAPBPR and TAPBPL, followed by bursa, gut and liver which were all consistently lower than thymus, spleen and lung.

4.3.13 Tapasin homologues are expressed at varying levels between tissues and individuals at the protein level

A preliminary analysis of protein-level expression of MHC class I, tapasin, TAPBPR and TAPBPL in the same tissue samples revealed some patterns of variation which were conserved between the individuals, suggesting that they were attributable to inherent tissue-specific variation, and some patterns which were not conserved between individuals, suggesting dynamic regulation of expression. As described in section 4.2.4, the amount of tissue loaded in each lane was roughly normalised by starting with a consistent mass (0.5 g) of tissue for each sample, but samples could have been taken from different parts of the organ in each of the two birds, which could affect the proportions of different cell types present.

Class I expression in bird 1 could not be assessed, since staining with α-GAPDH indicated that proteins on this gel had not effectively transferred to the membrane. In bird 2, no class I was detected in the brain, and expression was lower in the liver than in other organs. Tapasin expression was consistently extremely low in the brain and relatively low in the liver, roughly reflecting the class I expression pattern.

19-53-11 (raised against the C-terminal peptide of chicken TAPBPR) stained strongly in the thymus and more weakly in the liver, spleen and lung of bird 1. In bird 2, only spleen and lung had significant staining at 62 kDa but a lower AMM band was present in thymus, gut, spleen and lung (and to a lesser extent bursa and liver). A band appears at a similar AMM and with a similar distribution on the 19-54-11 blot, and both patterns look highly similar to the pattern observed in the bird 2 11-46-18 blot, suggesting that a small amount of the high-affinity 11-46-11 antibody could have contaminated the other membranes. Variation in expression between the birds is consistent with dynamic regulation, which was suggested by the relative expression in the cell lines. Regulation would, however, be predicted to occur post-transcriptionally, since variation in RNA-level expression between tissues was conserved between birds.

The patterns of 19-54-11 staining in the tissues of each bird approximately mirror the patterns of 19-53-11 staining. Thymus, spleen and lung samples stained strongly at the expected size in bird 1 (confounded by smearing in the spleen sample attributed to the presence of a large number of red blood cells in the sample) and spleen and lung stained strongly in bird 2.

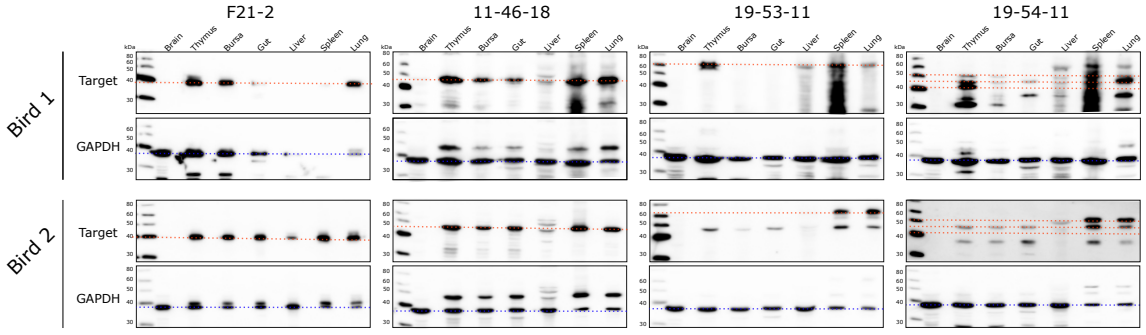


Figure 4.17: **Protein-level expression of chicken tapasin, TAPBPR and TAPBPL is variable between tissues.** The same tissue samples as analysed in section 4.3.12 were digested and used to produce membrane-enriched lysates for western blot analysis. Bird 2 liver, spleen and lung samples were additionally spun on a Histopaque gradient to remove red blood cells. After imaging, membranes were washed but not stripped, incubated with α-GAPDH and re-imaged. F21-2: α-MHC class I heavy chain, 11-46-18: α-tapasin extracellular domain, 19-53-11: α-TAPBPR C-terminal peptide, 19-54-11: α-TAPBPL C-terminal peptide. Red and blue dotted lines indicated expected band positions.

4.3.14 The chicken PLC contains proteins which stain with 19-54-11 but not 19-53-11

Protein G sepharose beads crosslinked to F1-3 (α-TAP2) were used to isolate components of the chicken PLC by immunoprecipitation. Both a 'generalist' (TG21) and 'specialist' (IS19) haplotype were used in case the relative dependence of different class I haplotypes on peptide editors was reflected in the composition of the PLC.

MHC class I (F21-2), TAP2 (F1-3) and tapasin (11-46-18) were all present in the eluate from the beads as expected (figure 4.18). A higher AMM band (approx. 62 kDa) was present in the eluate stained with 11-46-18 in addition to the expected band at approximately 48 kDa in multiple experiments. Assuming 19-53-11 stains TAPBPR, this protein is absent from the chicken PLC, as it is in humans.

Two of the three bands typically stained by 19-54-11 (putative α-TAPBPL) appear in the eluate. The higher AMM band is absent, but the mid-weight band appears to be more enriched in the eluate than any other examined protein, suggesting a significant restriction to the PLC. Since the high- and mid-weight bands are potentially the same protein species with and without glycosylation, it is plausible that the glycan is somehow associated with the ability of the protein to interact with the other components of the PLC. The lowest AMM band also appears in the eluate; it is not known whether this represents a protein which is related or unrelated to either of the other protein species.

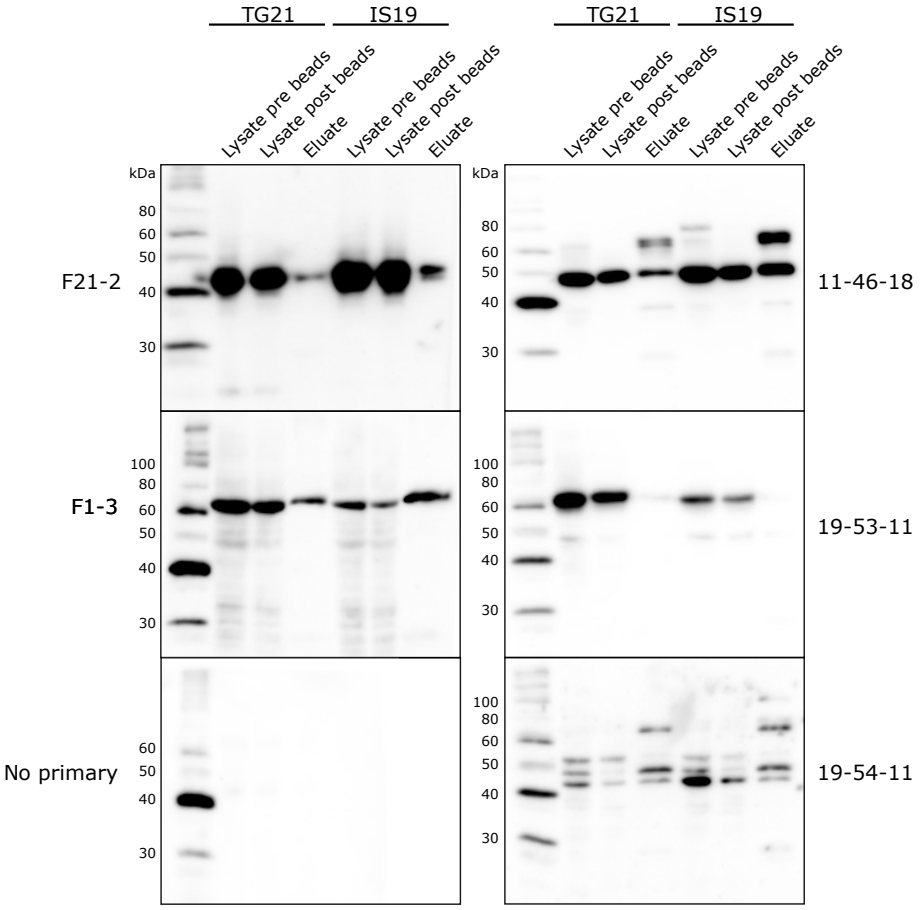


Figure 4.18: **Immunoprecipitation with α-TAP2 and western blot analysis staining for components of the PLC.** 1.5×10^6 cell equivalents were loaded in lysate lanes except when subsequent staining was for MHC class I or TAP2, where 3×10^5 or 3×10^6 cell equivalents were loaded per lane respectively. After elution in 60 µl 2x LB, 2 µl, 10 µl or 20 µl eluate were loaded on gels to be stained for class I, tapasin/TAPBPR/TAPBPL and TAP2 respectively.

4.3.15 Proteomic analysis provided no evidence for the presence of TAPBPL in the chicken MHC

Proteins which immunoprecipitated with α-TAP2 were also analysed by proteomics. Eluates from two independent IPs were analysed: the first used pooled lysates from multiple cell lines and batches to limit the impact of the observed variable expression of some proteins of interest, while the second used lysate from a single large TG21 culture split over multiple flasks. Eluate from the first IP was analysed by coomassie staining and western blot (figure 4.19), with the coomassie stained gel stored in destain solution for 6 weeks before the bands were excised and analysed by LC-MS/MS. Proteins in the second experiment were pulled-down on the same α-TAP2 sepharose beads but were not eluted and were provided to the facility for on-bead trypsin digest and analysis.

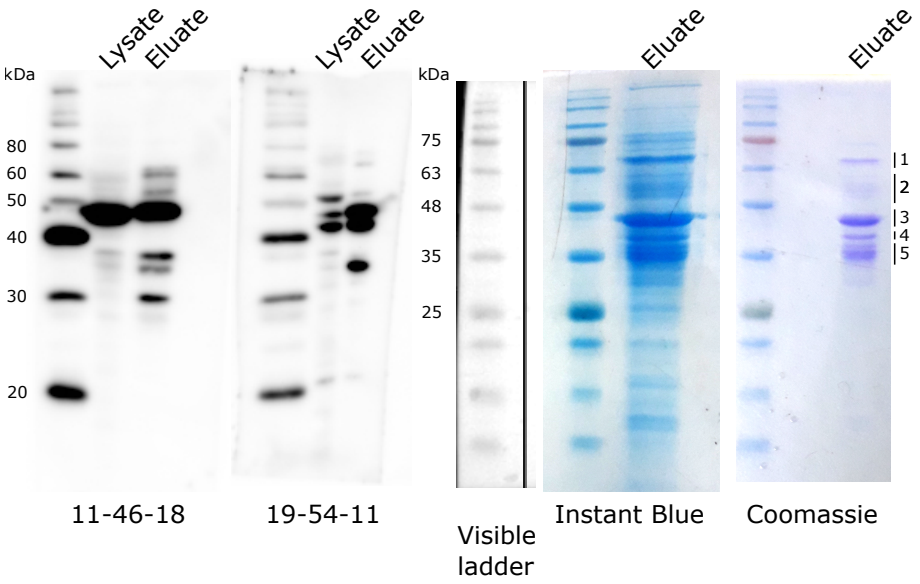


Figure 4.19: **Coomassie, InstantBlue and western blot analysis of eluate from IP with α-TAP2 for proteomics.** Western blotting with 11-46-18 and 19-54-11 confirmed the presence of proteins of interest in the sample that was analysed by LC-MS/MS. The gel for western blotting was run with both a prestained visible ladder and and IgG-binding ladder for chemiluminescent imaging, allowing the blots to be accurately aligned with the Coomassie or InstantBlue stained gels. The visible ladder shown was obtained from the membrane stained with 11-46-18. Gels not transferred to membranes were visualised with both InstantBlue and Coomassie Brilliant Blue since Coomassie-visible quantities of protein are required for LC-MS/MS but InstantBlue provides a more sensitive visualisation of the diversity of proteins present. Numbers alongside the Coomassie-stained gel indicate the gel fragments which were excised for analysis.

Some of the bands visible on the Coomassie-stained gel aligned approximately with bands obtained by western blot, and there was sufficient protein present in the size ranges of interest to proceed with LC-MS/MS.

In both the gel fragments and the bead digest sample all key components of the PLC (TAP1, TAP2, class I, tapasin, calreticulin and ERp57) were identified (table 4.4). There was no evidence of TAPBPR, consistent with the IP-western blot results (figure 4.18). There was also no evidence of TAPBPL in either of the pull-downs, despite visible staining with 19-54-11 when the same sample was analysed by western blot, suggesting that some or all of the bands stained by this antibody may not be specific for TAPBPL.

Gel fragment	Protein detected	Score
1	TAP2	1336
	calnexin	885
2	ERp57	2481
	calreticulin	1119
	TAP1	949
	tapasin	457
3	MHC class I α chain	675
	tapasin	382
4	MHC class I α chain	650
	TAP2	322
5	TAP1	127
Bead digest	calnexin	1028
	TAP2	852
	TAP1	508
	MHC class I α chain	450
	ERp57	404
	calreticulin	325
	tapasin	247

Table 4.4: **Results of LC-MS/MS analysis of eluate from IP with α-TAP2.** Gel fragment numbers correspond to those labelled in figure 4.19. Protein scores are derived from the scores of individual ions matched to that protein, where each ion score reflects the probability that the observed match is a random event. Higher scores therefore indicate higher confidence in the presence of the protein.

4.4 Discussion

This work illustrated the complex evolutionary dynamics of tapasin, TAPBPR and TAPBPL, and began to investigate the biochemistry of TAPBPR and TAPBPL in a non-mammalian system for the first time. While many questions about the specific role of TAPBPL are yet to be answered, the work indicates that it is widely expressed in tissues and has been retained in a wide range of taxonomic groups, suggesting a potentially important function in non-placental vertebrates.

4.4.1 Comparative genomics

Comparative genomics suggests that all three genes, tapasin, TAPBPR and TAPBPL, were present in early vertebrates, probably having arisen at least by the time of the divergence of bony fish from cartilaginous fish (Grimholt, 2018). In cartilaginous fish, sequences with up to 40% sequence similarity to chicken tapasin, TAPBPR and TAPBPL can be seen, but with few available genomes in which to investigate flaking genes and an extremely ancient common ancestor with species in which these genes are better characterised, allowing significant sequence divergence, it is very difficult to determine exactly when various members of this family may have evolved.

Individual lineages appear to have subsequently lost functional versions of one or more of these genes, suggesting a degree of redundancy in their functions. Variation is present at least on the scale of taxonomic orders in birds, with Anseriformes apparently lacking tapasin and TAPBPL, which are present in their sister group, the Galliformes, as well as many other non-passerine birds. The absence of tapasin in ducks, which has been previously reported (Magor et al., 2013), is supported by the fixation of residues associated with tapasin-independence in duck MHC class I molecules (Fleming-Canepa et al., 2016), namely D126 and Q221 (Q220 in chicken) as identified by van Hateren et al. (2013). Passerines, which also appear to lack tapasin, along with TAPBPR and TAPBPL, do not have these tapasin-independent residues. In house sparrow (Karlsson and Westerdahl, 2013), zebra finch (this project) and collared flycatcher (sequences from FicAlb1.5 assembly GCA_000247815.2) class I sequences, the equivalent of residue 126 is either glycine (as in the tapasin-dependant BF2*019:01) or glutamic acid. However, the associations of these residues with tapasin dependence or independence may not be conserved across the relatively large evolutionary distance between passerines and the Galloanserae (Galliformes and Anseriformes). Furthermore, examination of chicken BF2 sequences revealed that G126, associated with tapasin-dependence by van Hateren et al. (2013), is present in alleles ranging from BF2*021:01 and BF2*002:01, both of which have wide binding repertoires, low cell surface expression and are predicted to be tapasin-independent, to BF2*012:01 and BF2*015:02 (the BF2 allele in the standard B19 haplotype) which are thought to be tapasin-dependant. Q220, another residue associated with tapasin-independence and used as evidence for the absence of tapasin in duck, is not present in any BF2 allele other than that of the B19 haplotype within the standard haplotypes, despite a range of tapasin dependencies being present in these haplotypes. The associations may not, therefore, be as tight as is sometimes assumed.

Loss of antigen-processing proteins may be associated with escape from viral subversion of the pathway. For example, immune evasion via tapasin inhibition has been observed in adenoviruses and γ herpesviruses (Bennett et al., 1999; Lybarger et al., 2003). Whether protein functions, such as tapasin's role in assembly of the PLC, are lost with the gene or whether related proteins are able to perform these functions instead is still unclear. The duck PLC, or equivalent, may be able to provide important insights. Indeed, relatively high amino acid sequence conservation in the C-terminal peptide of chicken and duck TAP2 may allow immunoprecipitation and proteomic analysis of the duck PLC using the same methodology as in this project. Preliminary tests using F1-3 to stain duck spleen lysate by western blot were unsuccessful but many aspects of the experiment are yet to optimised.

None of tapasin, TAPBPR or TAPBPL were detected with any confidence in any passerine bird

species. This would be predicted to make functional compensation for missing genes by other members of the family impossible, and suggests that either other components of the antigen processing and presentation pathway are able to contribute to optimising the peptide repertoire or that class I molecules are loaded with low-affinity peptides at a much higher frequency in passerine birds than species with at least one member of the peptide-editing tapasin family. Development of an anti-MHC class I antibody for a passerine species would allow class I-peptide complexes to be purified and the peptides eluted and analysed by mass spectrometry. It can, however, be difficult to observe the presence or absence of low-affinity peptides on class I molecules, since weak interactions may not survive the purification process (Purcell et al., 2001). Without a peptide-editing control, against which the ratio of retrieved peptide:heavy chain or peptide:β_2-microglobulin, could be compared, this experiment would have to be carefully designed and performed.

Tapasin's other roles, as a chaperone and a bridge between TAP and MHC class I would also be interesting to investigate in passerines. The apparent absence of tapasin (as well as TAPBPR and TAPBPL, either of which could potentially evolve to compensate for its function) might suggest that passerines either do not assemble a PLC or assemble one with a unique composition or structure. Calreticulin and ERp57 are clearly present in many passerine genomes (BLASTP with chicken calreticulin and ERp57 against the passerine database returned genes with E values of 0 in 14 and 99 species respectively; a further 52 species had E values $<1 \times 10^{-80}$ in the results of the BLASTP with chicken calreticulin), so antibodies against either or both of these genes, in addition to the TAP molecules, might allow interaction partners to be elucidated by immunoprecipitation.

4.4.2 Expression of tapasin, TAPBPR and TAPBPL

Expression of human TAPBPR is not strongly tissue- or cell type-specific, although some epithelial cell types, particularly Paneth cells and enterocytes, both found in the gut, have noticeably higher levels of expression than average (Uhlén et al., 2015). Analysis of expression at the scale of whole tissues is a relatively imprecise method and is highly dependant on the relative numbers of various cell types in any sample taken, which cannot be easily controlled. Thus, while major differences can be detected and subsequently investigated in more detail, minor differences between tissues should not be over-interpreted. In this project, preliminary analysis of RNA-level expression of TAPBPR and TAPBPL on the scale of major tissues suggested that expression of these two genes was roughly correlated in various tissues, an observation that was also made in protein-level analysis using 19-53-11 and 19-54-11, supporting the hypothesis that 19-54-11 was specific for TAPBPL. Expression of both TAPBPR and TAPBPL was generally low in the brain, consistent with observations made in humans (TAPBPR only, Human Protein Atlas) and Atlantic salmon (TAPBPR and three TAPBPL genes, Grimholt (2018)). Generally high expression of TAPBPR observed here

in chicken spleen and lung is consistent with the results of Merkin et al. (2012) and reflects relatively high levels of expression of proteins associated with class I antigen processing in these tissues.

Expression of chicken TAPBPR was noticeably more variable between cell lines and between experiments than expression of chicken tapasin, which was consistently expressed between cell lines despite the MHC haplotypes exhibiting variable levels of tapasin-dependence. A likely N-linked glycan on chicken TAPBPR may have a regulatory role, although the protein was only detectable in the glycosylated form in cell lines.

The subcellular localisation of TAPBPL, which can be examined by immunofluorescence microscopy once a specific antibody has been confirmed, is another key issue which would help to shed light on the role of TAPBPL. While tapasin is restricted to the ER, reflecting its role in the assembly of the PLC, TAPBPR is present in the ER and Golgi (Boyle et al., 2013), reflecting its role in a later stage of antigen presentation.

4.4.3 19-54-11 and the chicken PLC

19-54-11 seemed likely to be specific to TAPBPL based on staining of proteins at roughly the expected AMM, staining of proteins at the same AMMs by other antibody clones raised against the C-terminal peptide of TAPBPL and similar patterns of tissue expression in TAPBPL RT-qPCR and western blot with 19-54-11. The results of the IP-western blot (figure 4.18) therefore suggested that TAPBPL might be a component of the chicken PLC.

No isotype control was used in the experiment, so the possibility that TAPBPL or a non-target antigen, which was not in fact part of the PLC, could be binding non-specifically to the protein G beads could not be ruled out. However, given that coomassie staining of the antibody solution after incubation with the beads showed that F1-3 (α-TAP2) was in excess, it seems unlikely that such a significant amount of protein could be non-specifically bound in this way.

Following proteomic analysis of eluates from two independent immunoprecipitations with F1-3, which showed no evidence of TAPBPL, alternative explanations for the IP-western blot results were considered.

Like 11-46-18 (α-tapasin), 19-54-11 stained a band at around 62 kDa in the eluate but not in the lysate. The possibility that the mid-weight band (47 kDa) stained by 19-54-11 was therefore non-target staining of tapasin (which appears at approximately the same AMM) was considered but the relative intensities of the bands in the lysate and eluate would seem to refute this. Further-

more, tapasin was expressed extremely consistently across all cell lines, whereas all three typical bands stained by 19-54-11 were variable between cell lines (figures 4.12 and 4.14).

Calreticulin, at 48 kDa, is an alternative candidate for the mid-weight band (which appears at an AMM of approximately 47 kDa) since it would be expected in the eluate of an immunoprecipitation of the PLC. It does have an N-glycosylation site and has been reported to exist in variable glycosylation states in response to cell stress (Jethmalani et al., 1994) in mammalian models, so a glycosylated form of calreticulin could also account for the band at 51 kDa if it could be shown that this form did not associate with the PLC. While chicken calreticulin contains no sequence with high overall identity to the TAPBPL C-terminal peptide (C-QVRSTTKPKPY), the motif QVKSGT in chicken calreticulin could be a potential site for recognition by antibodies raised against the peptide QVRSTTKPKPY. Indeed, the mid-weight band (47 kDa) was only present when staining with a small number of the 19-54 antibodies, suggesting that it could be a non-target antigen.

Since proteomic analysis indicated that chicken TAPBPL was not present in the eluate from the IP with α-TAP2, an alternative hypothesis is that the band at 51 kDa, which does not appear in the eluate, could be TAPBPL. The AMM of this band is slightly larger than the mass predicted for the TAPBPL protein based on the sequence (48 kDa), but motility in SDS-PAGE analysis is not an entirely accurate way to assess the true molecular mass of a protein. The 51 kDa band (which appears at a lower AMM in the manually aligned antibody screening blots, but corresponds to the band with the highest AMM in the screens) stains with multiple antibodies, suggesting the presence of a peptide with at least substantial sequence identity to the C-terminal peptide of TAPBPL. The protein sequence of chicken TAPBPL (XP_024997865.1) does contain two possible glycosylation sites (N-X-S/T), which might explain the PNGaseF sensitivity of this band. However, this is possibly inconsistent with the presence of the mid-weight band (47 kDa) in the WB analysis of the eluate from the IP, if the 51 and 47 kDa bands are predicted to be differentially glycosylated forms of the same protein.

The lowest AMM band stained by 19-54-11 (41 kDa) appears at approximately the same AMM as MHC class I, however the relative band intensities in the lysate and eluate do not seem to be equivalent in the F21-2 and 19-54-11 staining. Furthermore, the low AMM band seen with 19-54-11 was not PNGaseF-sensitive, unlike MHC class I. No other plausible alternatives were found for the lowest AMM band stained by 19-54-11, which appears at a slightly smaller AMM (41 kDa) than was predicted for TAPBPL (48 kDa). This band appeared commonly in 19-54 antibody screen 1, indicating that multiple antibodies were able to bind this protein species and it was therefore likely to contain a peptide with high sequence identity to TAPBPL. It was therefore possible that this band could represent TAPBPL, although this was apparently refuted by the proteomic analysis.

4.4.4 Future work on the role of TAPBPL

The role of TAPBPL remains unclear at the present time. Clearly, a priority is to determine the binding specificity of 19-54-11, either by immobilising this antibody on protein G-sepharose beads and performing immunoprecipitation and LC-MS/MS alongside an isotype control, or by creating a knockout cell line, which would facilitate further functional studies but not provide a positive identification of the antigen for 19-54-11 if it turned out not to be TAPBPL. Comparison of cell-surface MHC class I peptide binding repertoire by immunopeptidomics in wild type and TAPBPL knockout cells would revel whether TAPBPL has peptide-editing functionality and, if so, whether this functionality is a subtle 'fine-tuning' of the repertoire like that of TAPBPR (Boyle et al., 2013; Hermann et al., 2015) or a crucial part of the formation of stable class I-peptide complexes (at least in some haplotypes) like that of tapasin (Purcell et al., 2001; Williams et al., 2002).

TAPBPL may have different or additional functions which could be discovered through investigation of binding partners once a suitable anti-TAPBPL antibody was confirmed. In the case of TAPBPR, the discovery of an interaction with UGT1 revealed an important role in recycling of poorly-loaded class I molecules (Neerincx et al., 2017), and Ilca and Boyle (2020) speculate that interactions with additional co-factors could confer alternative functions, including in non-classical class I antigen presentation or cross-presentation.

Human tapasin and TAPBPR are both inducible with interferon γ (Boyle et al., 2013) and it would be interesting to investigate whether tapasin, TAPBPR and/or TAPBPL in chickens have this property. Preliminary tests in cell lines (which may exhibit unexpected immune gene regulation or response to stimulation due to transformation with reticuloendotheliosis virus) as well as *ex vivo* peripheral blood mononuclear cells and splenocytes were unable to achieve detectable immune activation and require further optimisation.

Chapter 5

Discussion

This thesis aimed to assess variation in the avian MHC and associated antigen processing pathways, with a view to assessing the extent to which insights developed in experimental chicken lines are applicable to other species. Three broad aims were defined to address this question:

- To compare MHC class I and II diversity in a range of non-commercial chicken populations and understand the evolution of the classical class I and II genes with a larger dataset
- To assess whether key features of the 'chicken-type' MHC are conserved across birds, particularly the passerine crown group
- To investigate a novel component of the chicken antigen processing pathway, TAPBPL, and to assess the phylogenetic distribution of likely peptide editing components more generally

In chapter 2, the wider global chicken population was shown to have significantly more MHC diversity than had previously been recorded in experimental and commercial chicken lines. Some haplotypes, particularly the so-called generalists 4:02-21-4:01_03-21 and 2-2-2-2, were common in almost all population examined, suggesting that interventions optimised for these haplotypes, such as vaccines, may be likely to have success in a wide range of populations. 31-31-31-31, which is extremely common in commercial flocks and is known to have a wide peptide binding repertoire facilitated by the ability to present long peptides (J. Kaufman, M. Harrison, N. Ternette, unpublished) was rare in non-commercial populations. While this haplotype is predicted to confer a 'generalist' disease response phenotype due to the low cell surface expression of its BF2 allele (E. Meziane and J. Kaufman, unpublished), it is not clear how its binding repertoire, which is characterised by accommodation of long peptides which extend beyond the F-pocket, relates to T cell activation, which is the key to mounting an effective adaptive immune response against a pathogen. This haplotype is clearly considered beneficial by commercial breeders, but an investigation of its associations with immune phenotypes will help to confirm whether this benefit is limited to the highly biosecure constitutions of commercial poultry production, or whether commercial breeders

have identified a potentially broadly adaptive haplotype which perhaps arose in Plymouth Rocks and has since been restricted to this and a small number of other breeds by the need to ensure ornamental birds remain 'pure-breeds'. Whether or not this haplotype is only adaptive in commercial production systems could inform development initiatives looking to cross commercial birds with local ecotypes to create hardy but productive lines for rural communities. 31-31-31-31 was detected at low levels in Africa; its current limited range could be related to one or more very recent introgressions from commercial birds or to negative selection acting on this haplotype in a non-commercial setting.

Chapter 3 illustrated variation in class I - TAP interactions between chickens and three passerine bird species. Deviation from the 'chicken-type' model of polymorphic TAPs, a single dominantly expressed MHC class I molecule and genetic linkage between the two suggested that insights from chickens need to be applied with care to studies in other birds and more widely across jawed vertebrates. While the apparent presence of a single dominantly expressed class I locus in passerines might suggest that species forced to change their ranges due to climate change or habitat loss could be vulnerable to novel pathogens, differences from chickens at the level of TAP and tapasin might mean that MHC class I associations with disease response are weaker in passerine birds. In particular, a hypothesised model in which passerine birds have extremely broad peptide binding repertoires is discussed in detail in section 5.1 below.

Chapter 4 investigated a group of homologous genes, tapasin, TAPBPR and TAPBPL. Despite evidence that all three genes appeared relatively early in vertebrate evolution, many lineages seem to be missing functional copies of one or more of these proteins, a feature which could be attributed either to selection, for example to overcome viral evasion strategies, or genetic drift if the functions of the genes are somewhat redundant. The extent to which the presence or absence of these proteins affects antigen presentation and therefore immune responses and disease outcomes remains unclear. The absence of TAPBPL in placental mammals does not, for example, prevent many relationships and correlations observed in chickens also being present in humans. Indeed, a suite of correlated properties of class I genes, namely peptide binding repertoire, cell surface expression, tapasin dependence and stability are at the heart of the 'generalist-specialist hypothesis' (Kaufman, 2017), which applies to both chickens and mammals.

The apparent absence of all three genes in passerine birds, or all but TAPBPR in ducks, however, may have more significant impacts on the antigen processing pathway, further highlighting the need to apply insights from chickens cautiously across birds.

5.1 New models of class I antigen presentation

Some observations made during this project suggest possible alternatives to the 'chicken-type' and 'human-type' models of class I antigen presentation. While tapasin, TAPBPR and TAPBPL are apparently absent in passerines, examination of predicted passerine TAP2 protein sequences in genomes suggests that they have retained their TAP2 tapasin binding domain. The TAP1 tapasin binding domain is absent in both chickens and passerines, suggesting a loss of this region prior to the split between the Galloanserae and the Neoaves, while Palaeognathae genomes are of insufficient quality to easily predict whether a TAP1 tapasin binding domain was present in the earliest birds or not. Furthermore, the region equivalent to the tapasin binding domain in passerine TAP2 is no more divergent between species than the remainder of the protein, possibly suggesting that it still has a function that is causing the sequence of this region to be similarly conserved to the rest of the gene. Possibly, recent loss of tapasin has not provided sufficient time for the TAP2 tapasin binding domains to accumulate noticeably more mutations. Indeed, falcons (part of the recently defined Austrolaves clade, which includes passerines, parrots and falcons (Prum et al., 2015)) appear to have retained a recognisable tapasin gene. However, the most recent common ancestor of falcons and passerines is still almost 60 million years ago (Prum et al., 2015), which would appear to provide ample time for degradation of a non-functional protein region unless it was critical to protein structure. Since the tapasin binding helices of TAP2 are on the outside of the protein, it seems unlikely that a structural role for the tapasin binding domain would be a significant evolutionary constraint. Alternatively, passerines may have a different chaperone which binds to this region of TAP to facilitate assembly of the PLC, if, indeed, passerines form a physically associated complex for loading peptides. If an alternative protein provides a physical link between TAP and class I molecules in passerines it would be interesting to know whether this had peptide editing capabilities, perhaps via a different mechanism to tapasin.

One possible interpretation of the observations made in passerine birds would be a strategy based on highly promiscuous peptide binding by class I molecules. Limited TAP polymorphism, at least in sparrows and flycatchers, combined with sequence variation in the class I genes suggests that the TAP transporter is likely to be highly permissive in these species so that all alleles receive appropriate peptides. Thus, there would be limited restriction on repertoire at the level of the TAPs. Absence of evidence of tapasin-related peptide editors in passerines is not necessarily evidence of absence of any proteins with editing capabilities, however it is plausible that passerines do not optimise the repertoire of peptides presented on class I in this way. The result of this system could be an extremely wide repertoire, unrestricted by either TAP or tapasin, assuming that the class I motifs themselves are not restrictive. The presence of Arg84 in passerine class I sequences, which has been associated with the ability to bind peptides which hang out of the groove at the

C-terminus (Xiao et al., 2018), may suggest a wide binding repertoire for many passerine class I alleles, despite only BF2*031:01 so far being seen to capitalise on this ability in chickens (Dr. N. Ternette and Prof. J. Kaufman, unpublished). Speculatively, Lys146, a second deviation in passerine class I sequences from the conserved residues reported by Kaufman et al. (1994), which is also involved in coordinating C-terminal binding, could contribute to passerine class I molecules having even more flexibility in peptide binding at the C-terminus.

If passerine class I molecules were shown to have wide peptide binding repertoires, this might provide an explanation for the presence of a single dominantly expressed class I molecule under the T cell depletion hypothesis, independent of co-evolution with TAPs which is important in the chicken model but apparently does not occur in passerines. Multiple loci may result in dangerously few T cell clones surviving negative selection in the thymus. This could depend on whether or not low-affinity peptides, which would be edited off by tapasin in chickens or mammals but are hypothesised to be presented in passerines, remain associated with class I for long enough to be presented on the cell surface and impact thymic selection.

5.2 Genotypes to phenotypes

This project has reported genetic variation in the MHC and antigen processing pathways of a range of species, and has raised many questions about the functional or phenotypic implications of this variation. Answering many of these questions now requires the combination of established methodologies such as immunopeptidomics, IP-LC-MS/MS, immunofluorescent microscopy, flow cytometry and experimental infection studies, with novel reagents and resources including antibodies against class I in additional non-model organisms, TAPBPL knockout cell lines and experimental chicken or cell lines carrying non-standard MHC haplotypes. Key questions which could be answered in work following this project include:

- Does the 31-31-31-31 haplotype confer resistance to important pathogens encountered by chickens in non-controlled environments?
- Are common haplotypes in village chickens generalists or specialists?
- Do passerine class I genes have a wider peptide binding repertoire than chickens due to their permissive TAPs and lack of peptide editing?
- Do passerines have a PLC?
- What is the impact of TAPBPL on peptide binding repertoire?
- Where is TAPBPL localised in the cell?

Other questions would require field studies, such as investigating pathogen diversity in village chickens in different locations, in order to frame hypotheses regarding the potential roles of haplotypes unique to particular geographies.

5.3 Concluding remarks

This thesis has illustrated diversity in MHC class I antigen processing and presentation, on taxonomic scales from intra-species to inter-class, which suggests greater flexibility in this system than had been previously acknowledged. Perhaps this is not surprising given the ongoing arms races between pathogens and hosts, and the wide-ranging strategies each are able to employ to escape the attacks of the other.

Model organisms (a term which is difficult to specifically define given the continuous spectrum of levels of detail in which various organisms have been studied) can be incredibly valuable, with the availability of research reagents, genomic resources, bioinformatic databases and prior literature facilitating the development of crucial insights. However, as this thesis has highlighted, these insights must be applied with care to less well-studied organisms, even from relatively closely related taxa. Having said that, understanding the similarities and differences present at different scales can also help appropriate insights to be applied with confidence in inter-disciplinary contexts where the necessary assumptions are met.

Appendix A

Appendix

A.1 Primer sequences

Archive reference corresponds to lab primer archive, currently at the University of Edinburgh. NA indicates that a primer is not present in the archive, either because it was provided as a stock sequence primer by sequencing services (NA†) or because it was synthesised after the physical archive moved to Edinburgh (NA*).

Primer name	Archive reference	Sequence
B_actin_F	uc268	ACGTCTCACTGGATTTCGAGCAGG
B_actin_R	uc269	TGCATCCTGTCAGCAATGCCAG
c71	c71	CGAGCTCCATACCCTGCGGTAC
c75	c75	CTCCTGCCCAGCTCAGCCTTC
CFC	uc1059	ACAAGGACCCCAAACAGCAT
GAPDH_F	uc270	GAGGGTAGTGAAGGCTGCTG
GAPDH_R	uc271	CATCAAAGGTGGAGGAATGG
Pado_TAP1_ENDb_R	NA*	CTCCAGCTCCCGCAGCTTTC
Pado_TAP1_ex1b_F	NA*	GCGCTGGACCGTGGTGG
pJET_pcr_F	uc764	ACACTTGTGCCTGAACACCA
pJET_pcr_R	uc765	TGAGAGTCGATTGCCAAGAA
pJET1.2_rev	NA†	AAGAACATCGATTTTCCATGGCAG
T1_E5FP2	uc1415	CAGGTGCGGGTGGCGGAG
T1_MIDRP2	uc1653	CGGGAATGTGGCCAAGAGAGCATG
T7	NA†	TAATACGACTCACTATAGGG
TAP1_10R	uc937	GCTCATCCAGCACCAGCAC
TAP1_1F	uc910	GGCTGATGTCGGCTTCTGC
TAP1_2aF	uc911	TACTTCACGGGCCGTGTCA
TAP1_2R	uc926	GTCACASGCCAGCTCGG

Continued on next page

Continued from previous page

Primer name	Archive reference	Sequence
TAP1_5F	uc916	CCATCCTCATGTGCCAGCT
TAP1_7aR	uc933	CTTGAGGACTGGCTCGGAC
TAP1_8F	uc921	GGGAAGAGCTCSCTGGTG
TAP1_F6_Fihy	uc1050	CCACGTCCCACTGTCTCATG
TAP1_F6_mod1	uc1041	CCTCCTTGTGCCACTGTCTC
TAP1_R5_Fihy_a	uc1047	ACATGGAGAGACACAGCCAG
TAP1_R5_mod1	uc1008	TCGGGACATACAGGAGCATGAC
TAP1_R8_Fihy	uc1049	ATCCCGTCCAGCAGCAC
TAP1_R8_mod2	uc1040	GTCCCGTTCAGCAGCAC
TAP2_11R	uc900	ABRGGATCCTGCTCCACC
TAP2_2F	uc876	TACGYCACYGGGAAGGTGCT
TAP2_4aF	uc880	AAGTGCCACGGGTGTGCA
TAP2_4R	uc893	CTCCAGCAGTGCCAGCAGT
TAP2_6F	uc883	GAAGCATCCGCCCGGAC
TAP2_6R	uc895	TCCTTCAGCCTCAGCTCCTTG
TAP2_F4_Fihy_b	uc1046	ACCGAAGTGCCTCAGGTG
TAPBPL_a9_F	NA*	GCTCAAAGGCTTCAGACAGA
TAPBPL_a9_R	NA*	GGCTGTAAAGTAAGGGCTGAT
TAPBPR_a3_F	NA*	TCCTCTCCTCCCATCATCTT
TAPBPR_a3_R	NA*	GCCCTCTTGCTGGAAACTAT
TAPBPR_F1	NA*	CCGGCAGTCTGGAGCAGG
TAPBPR_R1	NA*	CCTCCTGAGCACAATGAAGAGG
uc1302	uc1302	CCTTACCCCACGCCTGGCT
uc54	uc54	GTGCCCGCAGCGTTCTTC
VAS	uc1067	TGGTGGAGGTGATGGAGT
ZF_ex2_F	uc1401	GTTCTCCACTCCCTGCATTACCTGC
ZF_ex2_F2	uc1402	GCATTACCTGCACGTGGCGG
ZF_ex3_R	uc1403	TTGCGCTCCAGCTCCTTCTGC
ZF_ex4_R	uc1404	ACAGGATCAGCGTCCCGTATTCC
ZF_TAP1_F2	uc1160	GATGTCGGCTTCTGCTCTCG
ZF_TAP1_R3	uc1167	TTGAGGACTGGCTCGGACC
ZF_TAP2_F5	uc1151	GGTCTTCGGAACGTTCCAGC
ZF_TAP2_R1	uc1153	CTGGATGAAGTCCATGGCTCC

A.2 Thermocycling protocols

Protocol name	Initial denaturation		No. cycles	Denaturation		Annealing		Extension		Final extension	
	Temp. (°C)	Time (s)		Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)	Temp. (°C)	Time (s)
PHS_BF_G	98	30	35	98	10	68.4	20	72	30	72	300
PH_HS_BL	98	30	35	98	10	64	20	72	30	72	300
SP_TP2_1	98	30	30	98	10	68	10	72	60	72	300
TP1_R5M1	98	30	35	98	10	69	10	72	60	72	600
TCD_SS1	98	30	25	98	10	69.5 (-0.2/cycle)	20	72	60	72	600
			15	98	10	64.5	20	72	60		
TD_T1FG	98	30	10	98	10	72 (-1/cycle)	10	72	60	72	600
			20	98	10	61	10	72	60		
SP_LONG	98	30	33	98	10	68	10	72	120	72	600
HS_CDNA	98	30	25	98	10	69.5 (-0.2/cycle)	20	72	60	72	300
			10	98	10	64.5	20	72	60		
ZF_CDNA	98	30	25	98	10	68.5 (-0.1/cycle)	20	72	60	72	600
			5	98	10	66	20	72	60		
SP_TP2_G	98	30	28	98	10	72	10	72	60	72	300
SP_TPTD	98	30	25	98	10	71 (-0.2/cycle)	20	72	60	72	600
			7	98	10	66	20	72	60		
ZF_MHCI	98	30	35	98	10	66	20	72	30	72	300
COLONY	94	30	30	94	30	55	30	72	90	-	-
PRL_CDNA	98	30	29	98	10	65	20	72	60	72	300
BPL_TD40	98	30	30	98	10	65.5 (-0.2/cycle)	20	72	30	72	300
			11	98	10	59.5	20	72	30		

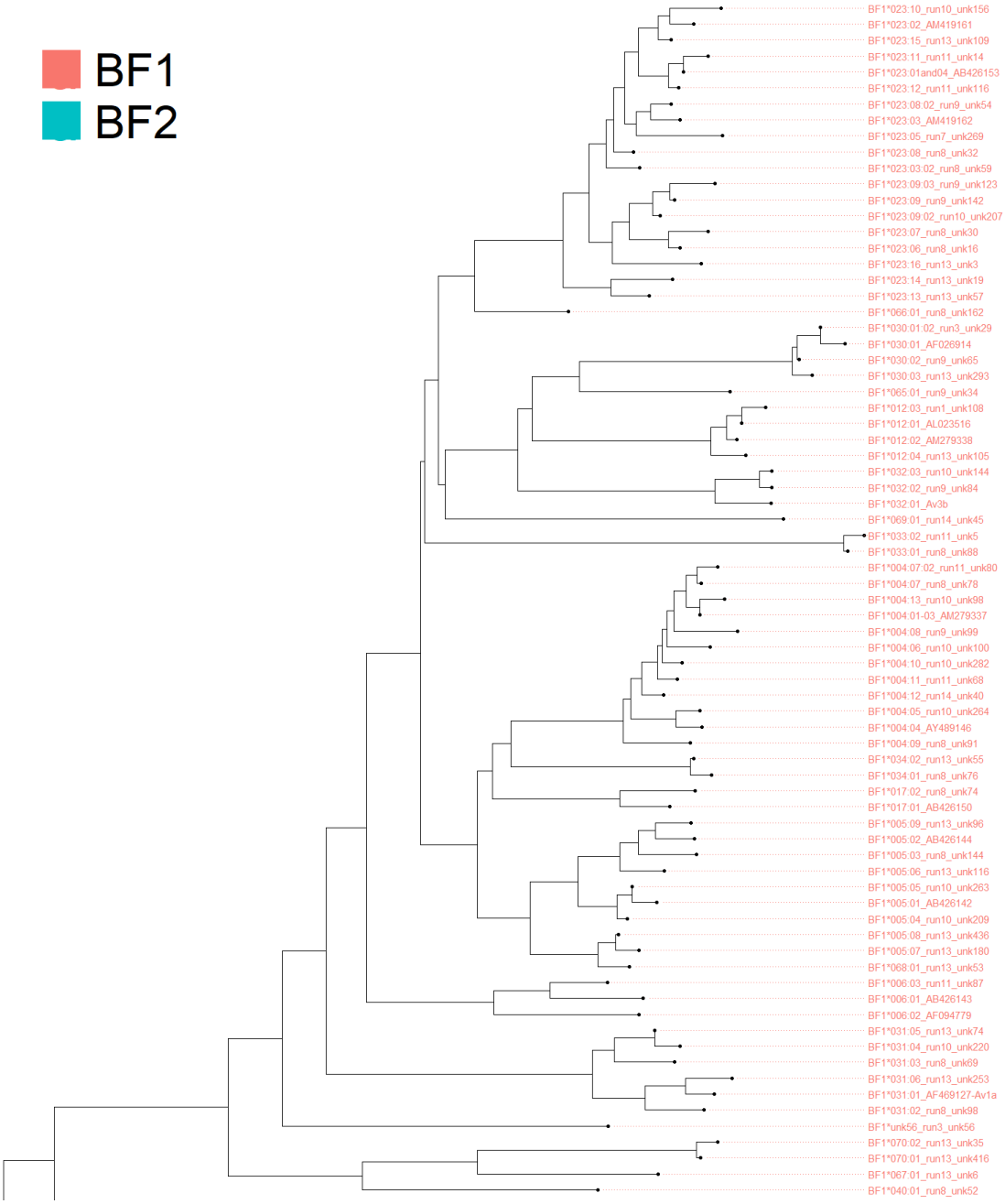
A.3 PCR barcodes used in NGS library preparation

PCR barcode	Sequence
R1	CTAAGGTAAC
R2	TAAGGAGAAC
R3	AAGAGGATTC
R4	TACCAAGATC
R5	CAGAAGGAAC
R6	CTGCAAGTTC
R7	TTCGTGATTC
R8	TTCCGATAAC
R9	TGAGCGGAAC
R10	CTGACCGAAC
R11	TCCTCGAATC
R12	TAGGTGGTTC

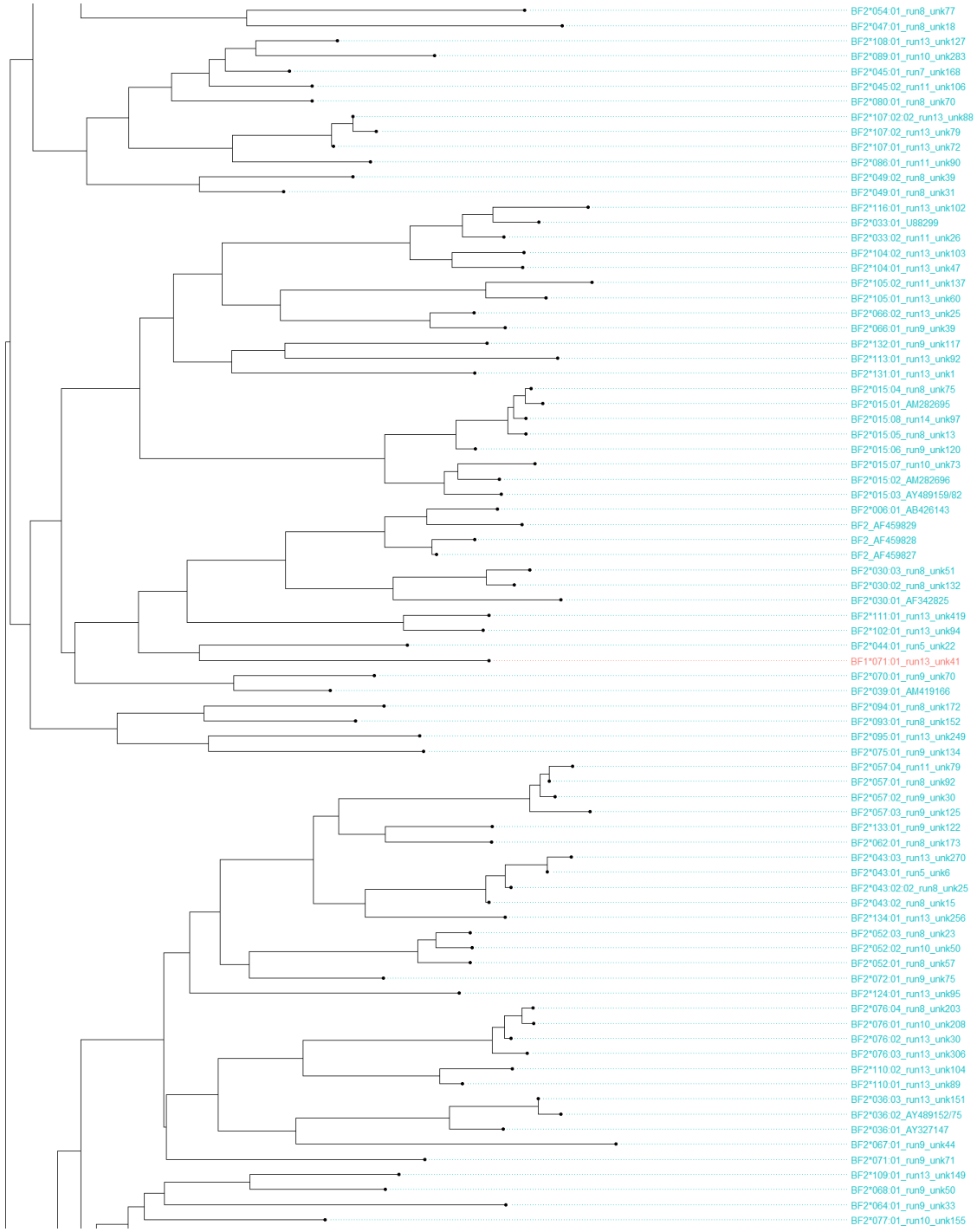
A.4 Phylogenetic trees of reference MHC class I and II sequences

Neighbour-joining phylogenetic trees were produced using Clustal Omega and visualised using the ggtree package in R. Sequence names are coloured according to locus assignment. BF sequences correspond to exons 2 and 3, BLB sequences correspond only to exon 2.

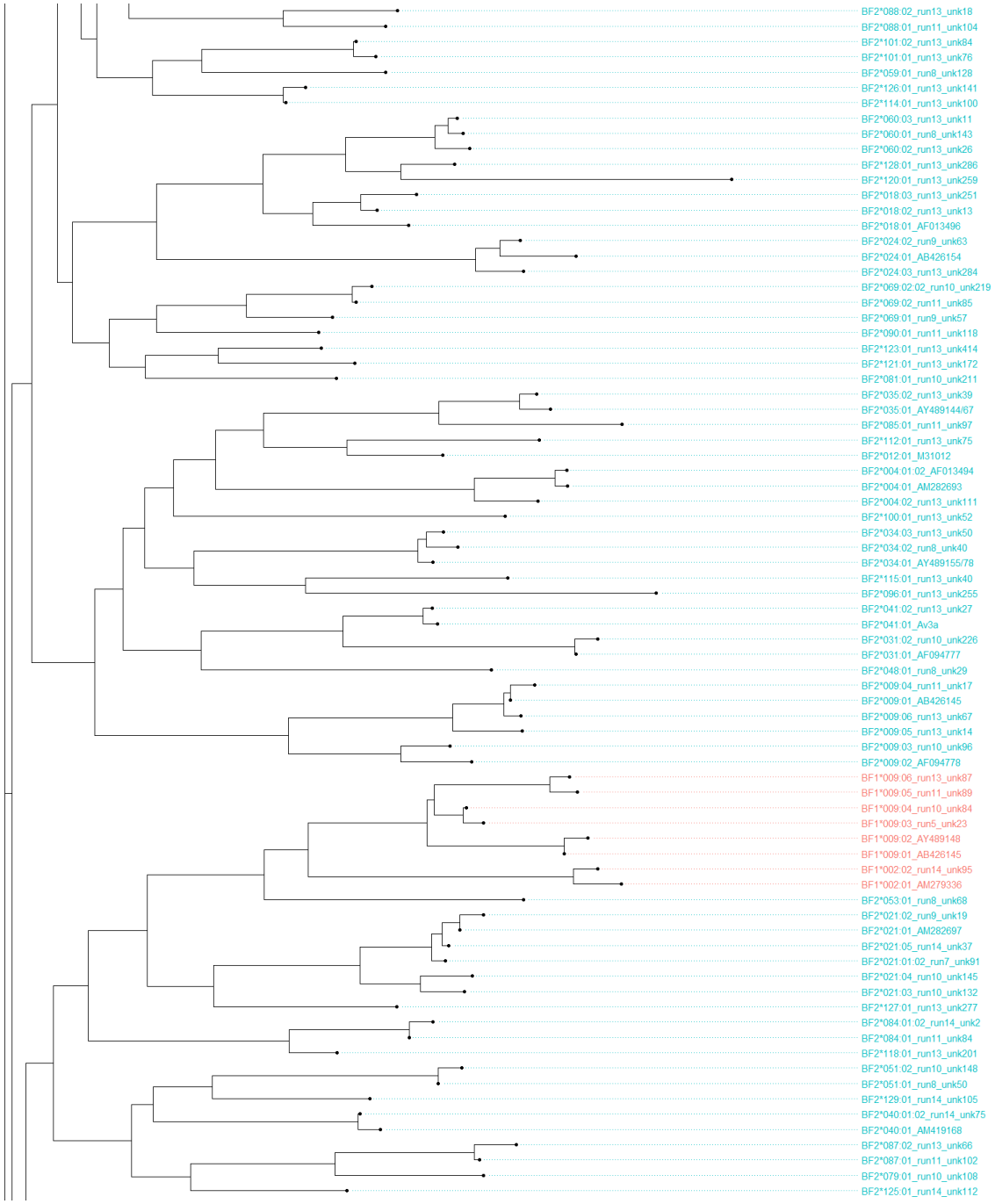
A.4.1 BF reference sequences



Continued on next page.



Continued on next page.



Continued on next page.

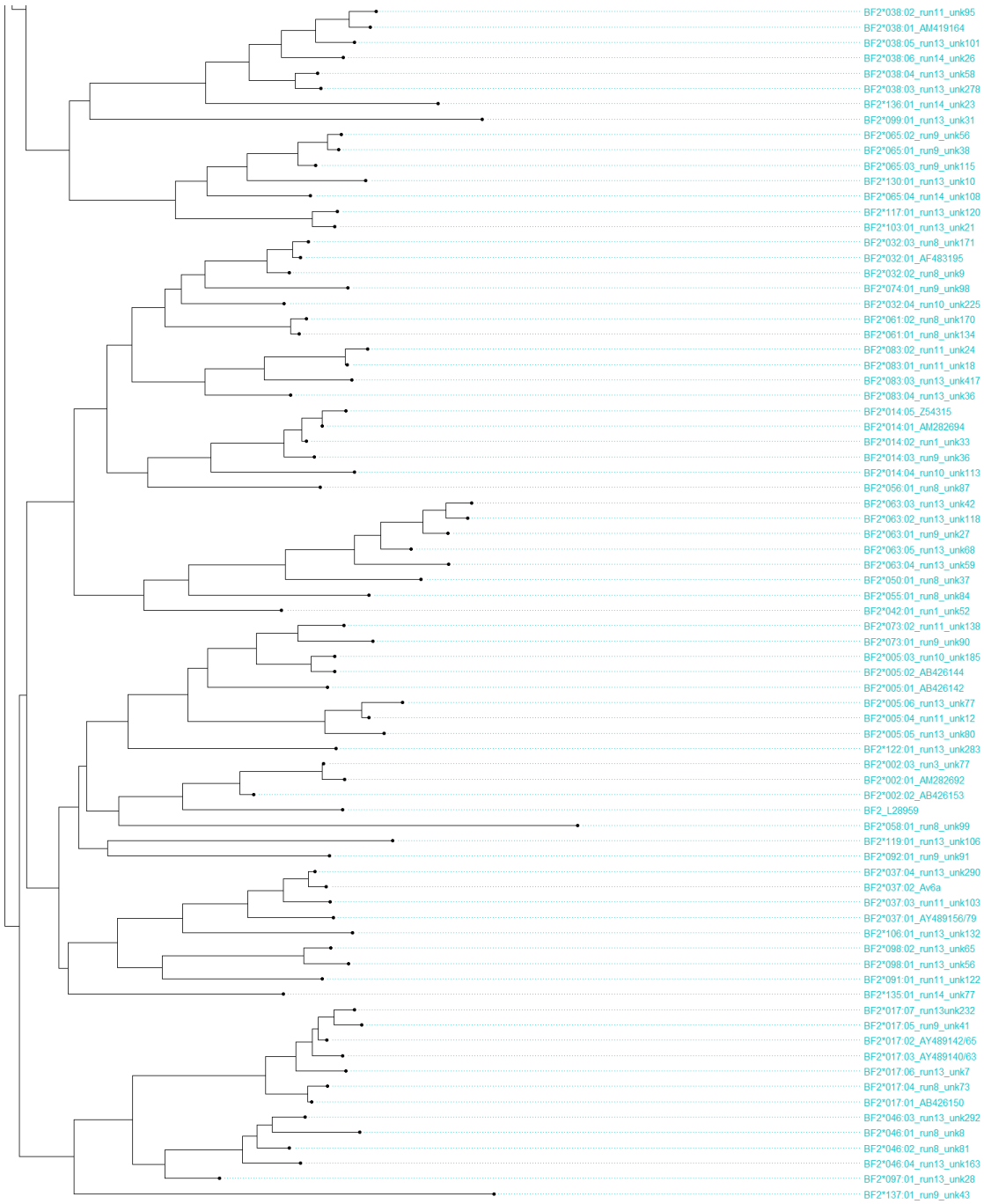
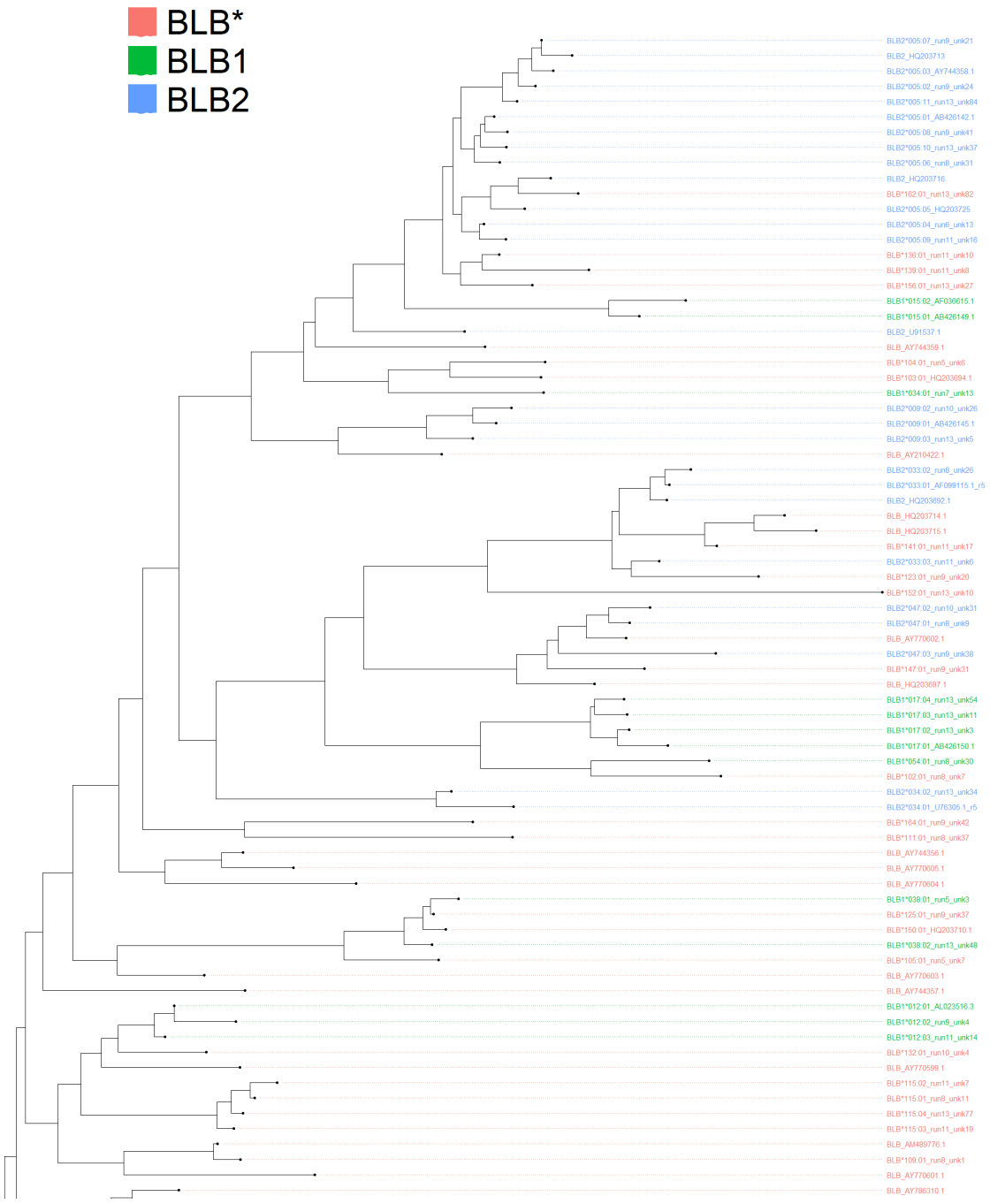
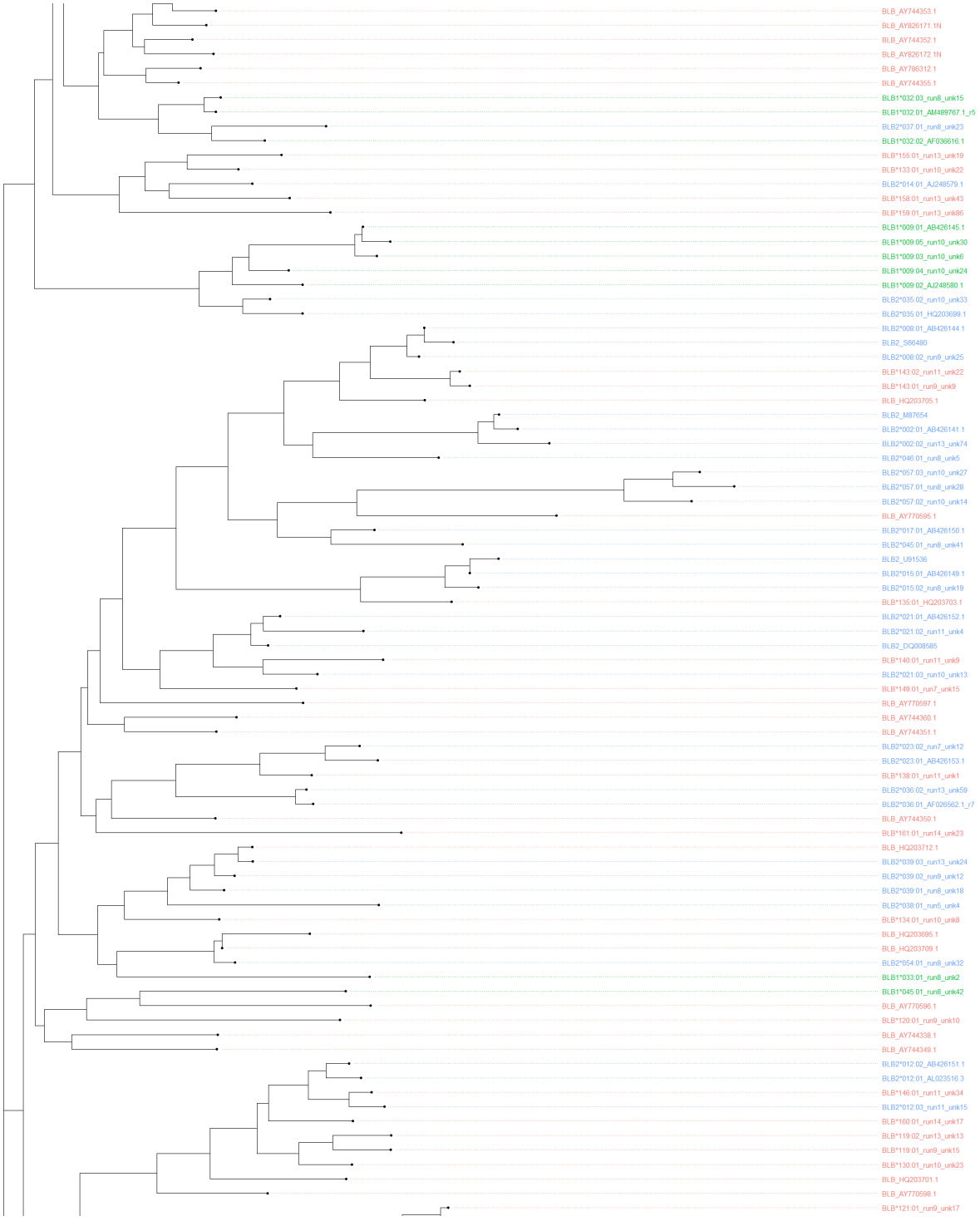


Figure A.1: **Neighbour-joining phylogenetic tree of all BF reference sequences.**

A.4.2 BLB reference sequences



Continued on next page.



Continued on next page.

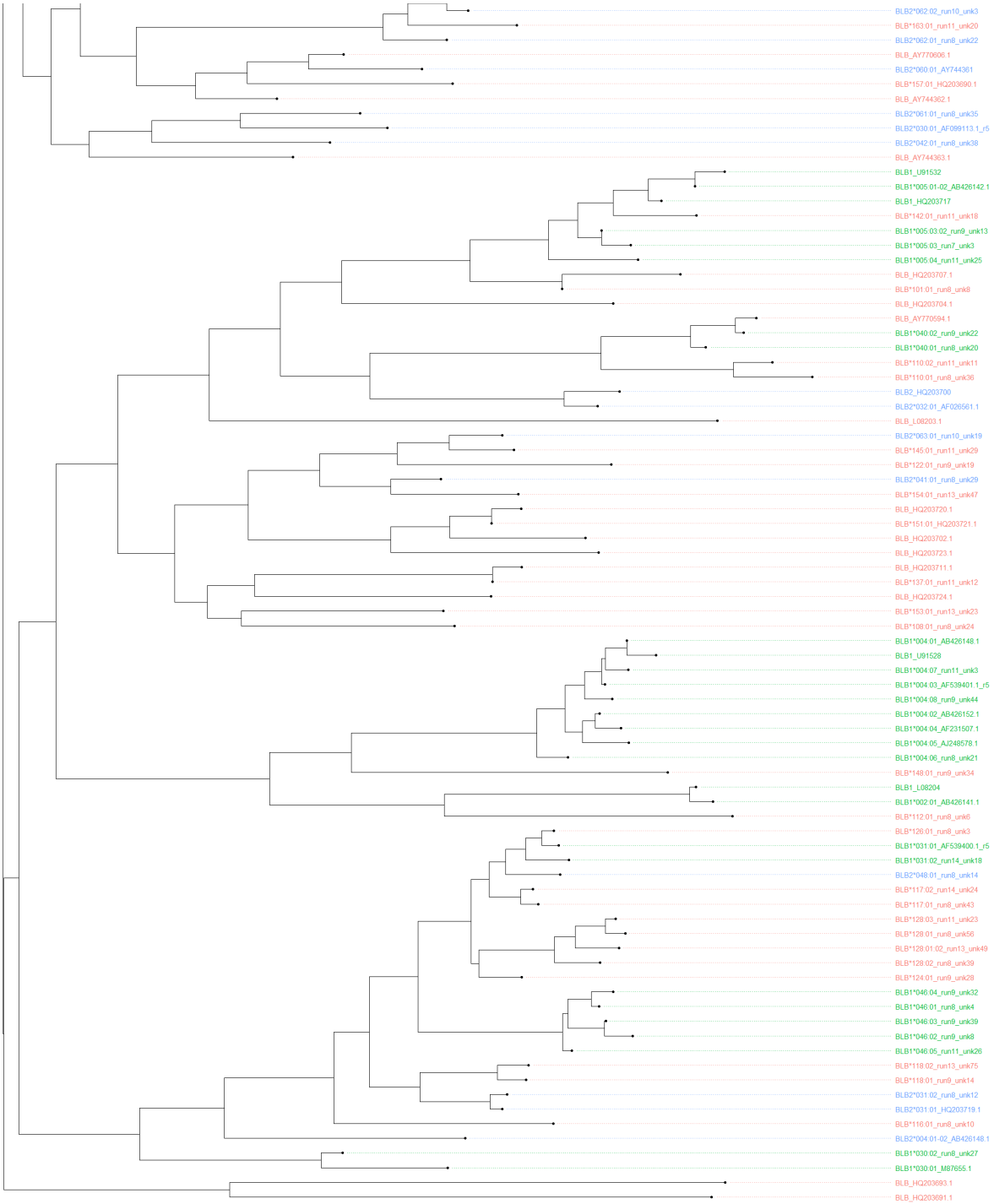


Figure A.2: **Neighbour-joining phylogenetic tree of all BLB reference sequences.**

A.5 Chicken MHC reference haplotype list

BLB1	BLB2	BF1	BF2	Haplotype
BLB1*002:01_AB426141.1	BLB2*002:01_AB426141.1	BF1*002:01_AM279336	BF2*002:01_AM282692	2-2-2-2
BLB1*002:01_AB426141.1	BLB2*002:01_AB426141.1	BF1*002:02_run14rpt_unk95	BF2*002:01_AM282692	2-2-2:02-2
BLB1*004:01_AB426148.1	BLB2*004:01-02_AB426148.1	BF1*004:01-03_AM279337	BF2*004:01_AM282693	4-4:01_02-4:01_03-4
BLB1*004:01_AB426148.1	BLB*145:01_run11_unk29	BF1*004:01-03_AM279337	BF2*004:01_AM282693	4-(145)-4:01_03-4
BLB1*004:08_run9_unk44	BLB2*004:01-02_AB426148.1	BF1*004:01-03_AM279337	BF2*004:01_AM282693	4:08-4:01_02-4:01_03-4
BLB1*005:01-02_AB426142.1	BLB2*005:10_run13_unk37	BF1*004:01-03_AM279337	BF2*004:02_run13_unk111	5:01_02-5:10-4:01_03-4:02
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:01_AB426142	BF2*005:01_AB426142	5-5-5-5
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*005:01_AB426142	BF2*005:01_AB426142	2-8-5-5
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:05_run10_unk263	BF2*005:01_AB426142	5:01_02-5-5:05-5
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:04_run10_unk209	BF2*005:01_AB426142	5:01_02-5-5:04-5
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:02_AB426144	BF2*005:03_run10_unk185	5:01_02-5-5:02-5:03
BLB1*005:01-02_AB426142.1	BLB2*005:04_r5	BF1*005:02_AB426144	BF2*005:04_run11_unk12	5:01_02-5:04-5:02-5:04
BLB1*005:01-02_AB426142.1	BLB2*005:09_run11_unk16	BF1*005:02_AB426144	BF2*005:04_run11_unk12	5:01_02-5:09-5:02-5:04
BLB1*005:01-02_AB426142.1	BLB*156:01_run13_unk27	BF1*005:06_run13_unk116	BF2*005:05_run13_unk80	5:01_02-(156)-5:06-5:05
BLB1*005:01-02_AB426142.1	BLB*156:01_run13_unk27	BF1*005:02_AB426144	BF2*005:05_run13_unk80	5:01_02-(156)-5:02-5:05
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*006:01_AB426143	BF2*006:01_AB426143	2-8-6-6
BLB1*005:01-02_AB426142.1	BLB*128:01_run8_unk56	BF1*006:01_AB426143	BF2*006:01_AB426143	5:01_02-(128)-6-6
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*009:01_AB426145	BF2*009:01_AB426145	9-9-9-9
BLB1*009:04_run10_unk24	BLB2*009:01_AB426145.1	BF1*009:01_AB426145	BF2*009:01_AB426145	9:04-9-9-9
BLB1*005:01-02_AB426142.1	BLB2*041:01_run8_unk29	BF1*004:04_AY489146	BF2*009:02_AF094778	5-41-4:04-9:02
BLB1*004:03_AF539401.1_r5	BLB2*033:01_AF099115.1_r5	BF1*004:04_AY489146	BF2*009:02_AF094778	4:03-33-4:04-9:02
BLB1*004:03_AF539401.1_r5	BLB2*033:01_AF099115.1_r5	BF1*004:05_run10_unk264	BF2*009:02_AF094778	4:03-33-4:05-9:02
BLB*115:01_run8_unk11	BLB2*021:03_run10_unk13	BF1*004:04_AY489146	BF2*009:03_run10_unk96	(115)-21:03-4:04-9:03
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*009:01_AB426145	BF2*009:04_run11_unk17	9-9-9-9:04
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*009:01_AB426145	BF2*009:05_run13_unk14	9-9-9-9:05
BLB1*012:01_AL023516.3	BLB2*012:01_AL023516.3	BF1*012:01_AL023516	BF2*012:01_M31012	12-12-12-12
BLB1*004:02_AB426152.1	BLB*126:01_run8_unk3		BF2*014:01_AM282694	4:02-(126)-?-14
BLB1*015:01_AB426149.1	BLB2*015:01_AB426149.1		BF2*015:01_AM282695	15-15-?-15
BLB1*012:01_AL023516.3	BLB2*012:02_AB426151.1	BF1*012:02_AM279338	BF2*015:02_AM282696	12-12:02-12:02-15:02
BLB1*015:01_AB426149.1	BLB2*015:01_AB426149.1		BF2*015:04_run8_unk75	15-15-?-15:04
BLB1*004:02_AB426152.1	BLB*112:01_run8_unk6		BF2*015:05_run8_unk13	4:02-(112)-?-15:05
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24		BF2*015:07_run10_unk73	5:01_02-5:02-?-15:07
BLB1*009:05_run10_unk30	BLB2*047:02_run10_unk31		BF2*015:07_run10_unk73	9:05-47:02-?-15:07
BLB*148:01_run9_unk34	BLB2*039:02_run9_unk12		BF2*015:07_run10_unk73	(148)-39:02-?-15:07
BLB1*031:01_AF539400.1_r5			BF2*015:08_run14rpt_unk97	31-?-?-15:08
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*017:01_AB426150	BF2*017:01_AB426150	17-17-17-17
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*030:01:02_run3_unk_29	BF2*017:02_AY489142/65	17-17-30:01:02-17:02
BLB2*039:02_run9_unk12	BLB2*062:02_run10_unk3	BF1*030:01:02_run3_unk_29	BF2*017:02_AY489142/65	39:02*-62:02-30:01:02-17:02
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*030:01:02_run3_unk_29	BF2*017:02_AY489142/65	2-8-30:01:02-17:02
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*017:02_run8_unk74	BF2*017:04_run8_unk73	17-17-17:02-17:04
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*017:01_AB426150	BF2*017:06_run13_unk7	17-17-17-17:06
BLB1*017:02_run13_unk3	BLB2*017:01_AB426150.1	BF1*017:01_AB426150	BF2*017:06_run13_unk7	17:02-17-17-17:06
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24		BF2*017:07_run13_BF_UNK_232	5:01_02-5:02-?-17:07
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*017:01_AB426150	BF2*017:07_run13_BF_UNK_232	17-17-17-17:07
BLB1*004:03_AF539401.1_r5	BLB2*033:02_run8_unk26	BF1*033:01_run8_unk88	BF2*018:01_AF013496	4:03-33:02-33-18
BLB1*004:03_AF539401.1_r5	BLB2*033:02_run8_unk26	BF1*033:02_run11_unk5	BF2*018:01_AF013496	4:03-33:02-33:02-18
BLB1*017:03_run13_unk11	BLB2*017:01_AB426150.1	BF1*023:08_run8_unk32	BF2*018:02_run13_unk13	17:03-17-23:08-18:02
BLB1*033:01_run8_unk2	BLB2*035:01_HQ203699.1	BF1*031:06_run13_unk253	BF2*018:03_run13_unk251	33-35-31:06-18:03
BLB1*004:02_AB426152.1	BLB2*021:01_AB426152.1	BF1*004:01-03_AM279337	BF2*021:01_AM282697	4:02-21-4:01_03-21
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*023:02_AM419161	BF2*021:01_AM282697	4:03-60-23:02-21
BLB1*004:02_AB426152.1	BLB2*021:02_run11_unk4	BF1*004:01-03_AM279337	BF2*021:01_AM282697	4:02-21:02-4:01_03-21
BLB1*032:03_run8_unk15	BLB2*063:01_run10_unk19	BF1*032:02_run9_unk84	BF2*021:03_run10_unk132	32:03-63-32:02-21:03
BLB*142:01_run11_unk18	BLB2*046:01_run8_unk5	BF1*012:01_AL023516	BF2*021:04_run10_unk145	(142)-46-12-21:04
BLB*115:01_run8_unk11	BLB*133:01_run10_unk22	BF1*032:03_run10_unk144	BF2*021:04_run10_unk145	(115)-(133)-32:03-21:04
BLB1*004:02_AB426152.1	BLB2*021:01_AB426152.1	BF1*004:01-03_AM279337	BF2*021:05_run14rpt_unk37	4:02-21-4:01_03-21:05
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:01-03_AM279337	BF2*024:01_AB426154	4-8-4:01_3-24

Continued on next page

Continued from previous page

BLB1	BLB2	BF1	BF2	Haplotype
BLB1*005:03_unk3_r7	BLB2*036:01_AF026562.1_r7	BF1*030:01_AF026914	BF2*024:01_AB426154	5:03-36-30-24
BLB1*005:03:02_run9_unk13	BLB2*036:01_AF026562.1_r7	BF1*030:02_run9_unk65	BF2*024:02_run9_unk63	5:03:02-36-30:02-24:02
BLB*121:01_run9_unk17	BLB2*036:02_run13_unk59	BF1*030:03_run13_unk293	BF2*024:03_run13_unk284	(121)-36:02-30:03-24:03
BLB1*030:01_M87655.1	BLB2*030:01_AF099113.1_r5	BF1*006:02_AF094779	BF2*030:01_AF342825	30-30-6:02-30
BLB1*005:01-02_AB426142.1	BLB*127:01_run8_unk39	BF1*006:01_AB426143	BF2*030:02_run8_unk132	5:01_02-(127)-6-30:02
BLB1*031:01_AF539400.1_r5	BLB2*031:01_HQ203719.1	BF1*031:01_AF469127-Av1a	BF2*031:01_AF094777	31-31-31-31
BLB1*005:01-02_AB426142.1	BLB*134:01_run10_unk8	BF1*031:01_AF469127-Av1a	BF2*031:02_run10_unk226	5:01_02-(134)-31-31:02
BLB1*032:01_AM489767.1_r5	BLB2*032:01_AF026561.1	BF1*004:01-03_AM279337	BF2*032:01_AF483195	32-32-4:01_03-32
BLB1*033:01_run8_unk2	BLB2*035:01_HQ203699.1	BF1*004:01-03_AM279337	BF2*032:01_AF483195	33-35-4:01_03-32
BLB1*032:01_AM489767.1_r5	BLB2*042:01_run8_unk38	BF1*004:01-03_AM279337	BF2*032:01_AF483195	32-42-4:01_03-32
BLB1*032:01_AM489767.1_r5	BLB2*032:01_AF026561.1	BF1*004:01-03_AM279337	BF2*032:02_run8_unk9	32-32-4:01_03-32:02
BLB1*033:01_run8_unk2	BLB2*035:01_HQ203699.1	BF1*004:01-03_AM279337	BF2*032:03_run8_unk171	33-35-4:01_03-32:03
BLB1*032:01_AM489767.1_r5	BLB2*032:01_AF026561.1	BF1*004:01-03_AM279337	BF2*032:03_run8_unk171	32-32-4:01_03-32:03
BLB*117:01_run8_unk43	BLB2*005:01_AB426142.1		BF2*032:04_run10_unk225	(117)-5-?-32:04
BLB1*009:01_AB426145.1	BLB2*034:01_U76305.1_r5		BF2*033:01_U88299	9-34-?-33
BLB1*009:03_run10_unk6	BLB2*034:01_U76305.1_r5		BF2*033:01_U88299	9:03-34-?-33
BLB1*009:01_AB426145.1	BLB2*034:01_U76305.1_r5		BF2*033:02_run11_unk26	9-34-?-33:02
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*023:01and04_AB426153	BF2*034:01_AY489155/78	4-8-23:01and04-34
BLB1*004:03_AF539401.1_r5	BLB2*008:01_AB426144.1	BF1*023:01and04_AB426153	BF2*034:01_AY489155/78	4:03-8-23:01and04-34
BLB1*038:01_run5_unk3	BLB2*038:01_run5_unk4	BF1*023:01and04_AB426153	BF2*034:01_AY489155/78	38-38-23:01and04-34
BLB1*009:02_AJ248580.1	BLB2*039:01_run8_unk18	BF1*023:01and04_AB426153	BF2*034:02_run8_unk40	9:02-39-23:01and04-34:02
BLB*101:01_run8_unk8	BLB*102:01_run8_unk7	BF1*023:01and04_AB426153	BF2*034:02_run8_unk40	(101)-(102)-23:01and04-34:02
BLB1*005:01-02_AB426142.1	BLB2*005:05_HQ203725	BF1*023:01and04_AB426153	BF2*034:03_run13_unk50	5:01_02-5:05-23:01and04-34:03
BLB*142:01_run11_unk18	BLB*151:01_HQ203721.1	BF1*023:01and04_AB426153	BF2*034:03_run13_unk50	(142)-(151)-23:01and04-34:03
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*023:02_AM419161	BF2*035:01_AY489144/67	4:03-60-23:02-35
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*023:02_AM419161	BF2*035:02_run13_unk39	4:03-60-23:02-35:02
BLB1*005:01-02_AB426142.1	BLB2*023:01_AB426153.1	BF1*023:01and04_AB426153	BF2*036:01_AY327147	5:01_02-23-23:01and04-36
BLB1*032:01_AM489767.1_r5	BLB2*032:01_AF026561.1	BF1*023:01and04_AB426153	BF2*036:01_AY327147	32-32-23:01and04-36
BLB1*005:01-02_AB426142.1	BLB2*005:05_HQ203725	BF1*023:11_run11_unk14	BF2*036:01_AY327147	5:01_02-5:05-23:11-36
BLB1*005:01-02_AB426142.1	BLB2*005:04_r5	BF1*012:01_AL023516	BF2*037:02_Av6a	5:01_02-5:04-12-37:02
BLB1*009:01_AB426145.1	BLB*159:01_run13_unk86	BF1*012:01_AL023516	BF2*037:04_run13_unk290	9-(159)-12-37:04
BLB*109:01_run8_unk1		BF1*023:02_AM419161	BF2*038:01_AM419164	(109)-?-23:02-38
BLB*109:01_run8_unk1		BF1*023:02_AM419161	BF2*038:02_run11_unk95	(109)-?-23:02-38:02
BLB*103:01_HQ203694.1	BLB*105:01_run5_unk7	BF1*023:02_AM419161	BF2*038:03_run13_unk278	(103)-(105)-23:02-38:03
BLB1*012:01_AL023516.3	BLB2*012:02_AB426151.1	BF1*023:02_AM419161	BF2*038:05_run13_unk101	12-12:02-23:02-38:05
BLB1*005:01-02_AB426142.1	BLB*136:01_run11_unk10	BF1*023:02_AM419161	BF2*038:06_run14rpt_unk26	5:01_02-(136)-23:03-38:06
BLB1*030:02_run8_unk27	BLB2*057:01_run8_unk28	BF1*004:13_run10_unk98	BF2*039:01_AM419166	30:02-57-4:13-39
BLB*143:01_run9_unk9	BLB*154:01_run13_unk47	BF1*023:07_run8_unk30	BF2*040:01:02_run14rpt_unk75	(143)-(154)-23:07-40:01:02
BLB1*033:01_run8_unk2	BLB2*035:01_HQ203699.1	BF1*023:03_AM419162	BF2*040:01_AM419168	33-35-23:03-40
BLB1*040:01_run8_unk20	BLB2*015:02_run8_unk19	BF1*040:01_run8_unk52	BF2*040:01_AM419168	40-15:02-40-40
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*023:03_AM419162	BF2*040:01_AM419168	9-9-23:03-40
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*032:01_Av3b	BF2*041:01_Av3a	9-9-32-41
BLB1*009:01_AB426145.1	BLB2*009:02_run10_unk26	BF1*032:01_Av3b	BF2*041:01_Av3a	9-9:02-32-41
BLB1*009:01_AB426145.1	BLB2*009:01_AB426145.1	BF1*032:01_Av3b	BF2*041:02_run13_unk27	9-9-32-41:02
BLB*101:01_run8_unk8	BLB*102:01_run8_unk7	BF1*004:01-03_AM279337	BF2*042:01_run1_unk_52	(101)-(102)-4:01_03-42
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:01-03_AM279337	BF2*043:01_run5_BF_unk_6	4-8-4:01_03-43
BLB1*004:01_AB426148.1	BLB2*008:02_run9_unk25	BF1*004:01-03_AM279337	BF2*043:01_run5_BF_unk_6	4-8:02-4:01_03-43
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:06_run10_unk100	BF2*043:01_run5_BF_unk_6	4-8-4:06-43
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:11_run11_unk68	BF2*043:01_run5_BF_unk_6	4-8-4:11-43
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:01-03_AM279337	BF2*043:02:02_run8_unk25	4-8-4:01_03-43:02:02
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:01-03_AM279337	BF2*043:02_run8_unk15	4-8-4:01_03-43:02
BLB1*045:01_run8_unk42	BLB2*045:01_run8_unk41	BF1*012:01_AL023516	BF2*043:03_run13_unk270	45-45-12-43:03
BLB*103:01_HQ203694.1	BLB*105:01_run5_unk7	BF1*009:03_run5_BF_unk_23	BF2*044:01_run5_BF_unk_22	(103)-(105)-9:03-44
BLB*103:01_HQ203694.1	BLB*105:01_run5_unk7	BF1*009:04_run10_unk84	BF2*044:01_run5_BF_unk_22	(103)-(105)-9:04-44
BLB1*034:01_r7	BLB2*023:02_r7	BF1*023:05_run7_BF_unk_269	BF2*045:01_run7_BF_unk_168	34-23:02-23:05-45
BLB1*046:01_run8_unk4	BLB2*046:01_run8_unk5	BF1*012:01_AL023516	BF2*046:01_run8_unk8	46-46-12-46
BLB1*009:01_AB426145.1	BLB2*037:01_run8_unk23	BF1*023:02_AM419161	BF2*046:02_run8_unk81	9-37-23:02-46:02
BLB1*046:01_run8_unk4	BLB2*046:01_run8_unk5		BF2*046:03_run13_unk292	46-46-?-46:03
BLB1*004:01_AB426148.1	BLB2*047:01_run8_unk9	BF1*023:06_run8_unk16	BF2*047:01_run8_unk18	4-47-23:06-47

Continued on next page

Continued from previous page

BLB1	BLB2	BF1	BF2	Haplotype
BLB1*032:03_run8_unk15	BLB2*048:01_run8_unk14	BF1*023:07_run8_unk30	BF2*048:01_run8_unk29	32:03-48-23:07-48
BLB1*030:01_M87655.1	BLB2*030:01_AF099113.1_r5	BF1*023:08_run8_unk32	BF2*049:01_run8_unk31	30-30-23:08-49
BLB1*009:02_AJ248580.1	BLB2*038:01_run5_unk4	BF1*023:08_run8_unk32	BF2*049:02_run8_unk39	9:02-38-23:08-49:02
BLB1*005:01-02_AB426142.1	BLB2*005:05_HQ203725	BF1*023:01and04_AB426153	BF2*050:01_run8_unk37	5:01_02-5:05-23:01and04-50
BLB1*004:01_AB426148.1	BLB2*008:01_AB426144.1	BF1*004:01-03_AM279337	BF2*051:01_run8_unk50	4-8-4:01_03-51
BLB1*005:01-02_AB426142.1	BLB2*023:01_AB426153.1		BF2*051:02_run10_unk148	5:01_02-23-?-51:02
BLB1*005:01-02_AB426142.1	BLB2*005:04_r5	BF1*023:03:02_run8_unk59	BF2*052:01_run8_unk57	5:01_02-5:04-23:03:02-52
BLB*115:01_run8_unk11	BLB*132:01_run10_unk4	BF1*023:08_run8_unk32	BF2*052:02_run10_unk50	(115)-(132)-23:08-52:02
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*023:08_run8_unk32	BF2*052:02_run10_unk50	5:01_02-5-23:08-52:02
BLB1*030:02_run8_unk27	BLB2*057:01_run8_unk28	BF1*032:02_run9_unk84	BF2*052:03_run8_unk23	30:02-57-32:02-52:03
BLB1*017:01_AB426150.1	BLB2*031:02_run8_unk12	BF1*032:02_run9_unk84	BF2*052:03_run8_unk23	17-31:02-32:02-52:03
BLB1*046:01_run8_unk4	BLB2*005:06_run8_unk31	BF1*031:03_run8_unk69	BF2*053:01_run8_unk68	46-5:06-31:03-53
BLB1*046:01_run8_unk4	BLB2*005:07_run9_unk21	BF1*031:03_run8_unk69	BF2*053:01_run8_unk68	46-5:07-31:03-53
BLB1*054:01_run8_unk30	BLB2*054:01_run8_unk32	BF1*034:01_run8_unk76	BF2*054:01_run8_unk77	54-54-34-54
BLB1*046:01_run8_unk4	BLB*125:01_run9_unk37	BF1*034:01_run8_unk76	BF2*054:01_run8_unk77	46-(125)-34-54
BLB1*054:01_run8_unk30	BLB2*054:01_run8_unk32	BF1*034:02_run13_unk55	BF2*054:01_run8_unk77	54-54-34:02-54
BLB*105:01_run5_unk7	BLB*108:01_run8_unk24	BF1*006:01_AB426143	BF2*055:01_run8_unk84	(105)-(108)-6-55
BLB1*004:02_AB426152.1	BLB*126:01_run8_unk3		BF2*056:01_run8_unk87	4:02-(126)-?-56
BLB1*030:02_run8_unk27	BLB2*057:01_run8_unk28	BF1*004:09_run8_unk91	BF2*057:01_run8_unk92	30:02-57-4:09-57
BLB1*046:02_run9_unk8	BLB*143:01_run9_unk9	BF1*004:09_run8_unk91	BF2*057:02_run9_unk30	46:02-(143)-4:09-57:02
BLB1*030:01_M87655.1	BLB*155:01_run13_unk19	BF1*004:09_run8_unk91	BF2*057:02_run9_unk30	30-(155)-4:09-57:02
BLB*115:02_run11_unk7	BLB2*033:03_run11_unk6	BF1*004:07:02_run11_unk80	BF2*057:04_run11_unk79	(115:02)-33:03-4:07:02-57:04
BLB1*033:01_run8_unk2	BLB2*035:01_HQ203699.1	BF1*031:02_run8_unk98	BF2*058:01_run8_unk99	33-35-31:02-58
BLB*103:01_HQ203694.1	BLB*105:01_run5_unk7	BF1*023:08_run8_unk32	BF2*059:01_run8_unk128	(103)-(105)-23:08-59
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*005:03_run8_unk144	BF2*060:01_run8_unk143	4:03-60-5:03-60
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*005:03_run8_unk144	BF2*060:02_run13_unk26	4:03-60-5:03-60:02
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*005:03_run8_unk144	BF2*060:03_run13_unk11	4:03-60-5:03-60:03
BLB1*004:01_AB426148.1	BLB2*061:01_run8_unk35		BF2*061:01_run8_unk134	4-61-?-61
BLB1*004:01_AB426148.1	BLB2*061:01_run8_unk35		BF2*061:02_run8_unk170	4-61-?-61:02
BLB*103:01_HQ203694.1	BLB*105:01_run5_unk7	BF1*004:08_run9_unk99	BF2*062:01_run8_unk173	(103)-(105)-4:08-62
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*005:02_AB426144	BF2*063:01_run9_unk27	5:01_02-5:02-5:02-63
BLB*153:01_run13_unk23	BLB2*039:03_run13_unk24	BF1*005:02_AB426144	BF2*063:02_run13_unk118	(153)-39:03-5:02-63:02
BLB*153:01_run13_unk23	BLB2*039:03_run13_unk24	BF1*005:06_run13_unk116	BF2*063:02_run13_unk118	(153)-39:03-5:06-63:02
BLB*153:01_run13_unk23	BLB2*039:03_run13_unk24	BF1*005:02_AB426144	BF2*063:03_run13_unk42	(153)-39:03-5:02-63:03
BLB*153:01_run13_unk23	BLB2*039:03_run13_unk24	BF1*005:06_run13_unk116	BF2*063:03_run13_unk42	(153)-39:03-5:06-63:03
BLB1*004:03_AF539401.1_r5	BLB2*047:01_run8_unk9	BF1*023:06_run8_unk16	BF2*065:01_run9_unk38	4:03-47-23:06-65
BLB1*004:03_AF539401.1_r5	BLB2*047:01_run8_unk9	BF1*023:06_run8_unk16	BF2*065:02_run9_unk56	4:03-47-23:06-65:02
BLB1*004:03_AF539401.1_r5	BLB2*047:01_run8_unk9	BF1*023:06_run8_unk16	BF2*065:04_run14rpt_unk108	4:03-47-23:06-65:04
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1		BF2*066:01_run9_unk39	17-17-?-66
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1		BF2*066:02_run13_unk25	17-17-?-66:02
BLB1*017:02_run13_unk3	BLB2*017:01_AB426150.1		BF2*066:02_run13_unk25	17:02-17-?-66:02
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:02_AB426144	BF2*067:01_run9_unk44	5:01_02-5-5:02-67
BLB1*046:01_run8_unk4	BLB*150:01_HQ203710.1	BF1*023:08:02_run9_unk54	BF2*068:01_run9_unk50	46-(150)-23:08:02-68
BLB1*040:02_run9_unk22	BLB2*015:02_run8_unk19	BF1*065:01_run9_unk34	BF2*069:01_run9_unk57	40:02-15:02-65-69
BLB1*040:02_run9_unk22	BLB2*015:02_run8_unk19		BF2*070:01_run9_unk70	40:02-15:02-?-70
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*005:02_AB426144	BF2*073:01_run9_unk90	5:01_02-5-5:02-73
BLB1*005:01-02_AB426142.1	BLB2*023:01_AB426153.1	BF1*023:09_run9_unk142	BF2*075:01_run9_unk134	5:01_02-23-23:09-75
BLB*115:01_run8_unk11	BLB*120:01_run9_unk10	BF1*023:09:02_run10_unk207	BF2*076:01_run10_unk208	(115)-(120)-23:09:02-76
BLB*115:01_run8_unk11	BLB*120:01_run9_unk10	BF1*023:09:02_run10_unk207	BF2*076:02_run13_unk30	(115)-(120)-23:09:02-76:02
BLB*115:01_run8_unk11	BLB*120:01_run9_unk10	BF1*023:09:02_run10_unk207	BF2*076:04_run8_unk203	(115)-(120)-23:09:02-76:04
BLB*130:01_run10_unk23	BLB2*039:02_run9_unk12	BF1*023:10_run10_unk156	BF2*077:01_run10_unk155	(130)-39:02-23:10-77
BLB*117:02_run14_unk24	BLB*161:01_run14_unk23	BF1*023:10_run10_unk156	BF2*077:01_run10_unk155	(117:02)-(161)-23:10-77
BLB*115:01_run8_unk11	BLB*116:01_run8_unk10	BF1*031:04_run10_unk220	BF2*078:01_run10_unk219	(115)-(116)-31:04-78
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*066:01_run8_unk162	BF2*081:01_run10_unk211	5:01_02-5-66-81
BLB*135:01_HQ203703.1	BLB2*039:02_run9_unk12	BF1*004:01-03_AM279337	BF2*083:01_run11_unk18	(135)-39:02-4:01_03-83
BLB*118:02_run13_unk75	BLB*157:01_HQ203690.1	BF1*070:01_run13_unk416	BF2*083:03_run13_unk417	(118:02)-(157)-70-83:03
BLB*115:01_run8_unk11	BLB2*057:03_run10_unk27	BF1*070:02_run13_unk35	BF2*083:04_run13_unk36	(115)-57:03-70:02-83:04
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*084:01:02_run14rpt_unk2	5:01_02-5:02-12-84:01:02
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*084:01_run11_unk84	5:01_02-5:02-12-84

Continued on next page

Continued from previous page

BLB1	BLB2	BF1	BF2	Haplotype
BLB1*005:01-02_AB426142.1	BLB2*005:08_run9_unk41	BF1*012:01_AL023516	BF2*084:01_run11_unk84	5:01_02-5:08-12-84
BLB1*005:01-02_AB426142.1	BLB*136:01_run11_unk10	BF1*012:01_AL023516	BF2*084:01_run11_unk84	5:01_02-(136)-12-84
BLB1*030:01_M87655.1	BLB2*030:01_AF099113.1_r5	BF1*032:02_run9_unk84	BF2*087:02_run13_unk66	30-30-32:02-87:02
BLB*119:02_run13_unk13	BLB*124:01_run9_unk28	BF1*023:14_run13_unk19	BF2*088:02_run13_unk18	(119:02)-(124)-23:14-88:02
BLB1*033:01_run8_unk2	BLB2*035:02_run10_unk33	BF1*004:10_run10_unk282	BF2*089:01_run10_unk283	33-35:02-4:10-89
BLB*121:01_run9_unk17	BLB2*039:02_run9_unk12	BF1*004:01-03_AM279337	BF2*093:01_run8_unk152	(121)-39:02-4:01_03-93
BLB1*045:01_run8_unk42	BLB2*045:01_run8_unk41	BF1*012:01_AL023516	BF2*094:01_run8_unk172	45-45-12-94
BLB1*032:03_run8_unk15	BLB*152:01_run13_unk10	BF1*032:02_run9_unk84	BF2*095:01_run13_unk249	32:03-(152)-32:02-95
BLB*121:01_run9_unk17	BLB2*039:02_run9_unk12	BF1*067:01_run13_unk6	BF2*096:01_run13_unk255	(121)-39:02-67-96
BLB1*046:01_run8_unk4	BLB2*046:01_run8_unk5	BF1*012:01_AL023516	BF2*097:01_run13_unk28	46-46-12-97
BLB*109:01_run8_unk1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*098:01_run13_unk56	(109)-5:02-12-98
BLB*109:01_run8_unk1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*098:02_run13_unk65	(109)-5:02-12-98:02
BLB1*030:02_run8_unk27	BLB2*057:01_run8_unk28	BF1*023:03_AM419162	BF2*099:01_run13_unk31	30:02-57-23:03-99
BLB1*030:01_M87655.1	BLB2*057:01_run8_unk28	BF1*023:03_AM419162	BF2*099:01_run13_unk31	30-57-23:03-99
BLB1*004:02_AB426152.1	BLB*141:01_run11_unk17	BF1*068:01_run13_unk53	BF2*100:01_run13_unk52	4:02-(141)-68-100
BLB1*004:02_AB426152.1	BLB*141:01_run11_unk17	BF1*005:07_run13_unk180	BF2*100:01_run13_unk52	4:02-(141)-5:07-100
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*023:08:02_run9_unk54	BF2*101:01_run13_unk76	5:01_02-23:08:02-101
BLB1*046:01_run8_unk4	BLB*150:01_HQ203710.1	BF1*023:08:02_run9_unk54	BF2*101:02_run13_unk84	46-(150)-23:08:02-101:02
BLB*115:01_run8_unk11	BLB*128:01:02_run13_unk49	BF1*006:01_AB426143	BF2*102:01_run13_unk94	(115)-(128:01:02)-6-102
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*012:01_AL023516	BF2*103:01_run13_unk21	2-8-12-103
BLB*153:01_run13_unk23	BLB2*039:03_run13_unk24		BF2*104:01_run13_unk47	(153)-39:03-?-104
BLB*109:01_run8_unk1	BLB*112:01_run8_unk6		BF2*105:01_run13_unk60	(109)-(112)-?-105
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*012:01_AL023516	BF2*106:01_run13_unk132	17-17-12-106
BLB1*046:05_run11_unk26	BLB2*046:01_run8_unk5	BF1*009:05_run11_unk89	BF2*107:01_run13_unk72	46:05-46-9:05-107
BLB1*046:05_run11_unk26	BLB2*046:01_run8_unk5	BF1*009:06_run13_unk87	BF2*107:01_run13_unk72	46:05-46-9:06-107
BLB1*004:02_AB426152.1	BLB2*021:01_AB426152.1	BF1*009:06_run13_unk87	BF2*107:01_run13_unk72	4:02-21-9:06-107
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*009:05_run11_unk89	BF2*107:02:02_run13_unk88	5:01_02-5-9:05-107:02:02
BLB1*005:01-02_AB426142.1	BLB2*005:01_AB426142.1	BF1*009:05_run11_unk89	BF2*107:02_run13_unk79	5:01_02-5-9:05-107:02
BLB1*046:01_run8_unk4	BLB*150:01_HQ203710.1	BF1*004:10_run10_unk282	BF2*108:01_run13_unk127	46-(150)-4:10-108
BLB1*046:01_run8_unk4	BLB*150:01_HQ203710.1	BF1*023:08:02_run9_unk54	BF2*109:01_run13_unk149	46-(150)-23:08:02-109
BLB*140:01_run11_unk9	BLB2*038:01_run5_unk4	BF1*012:01_AL023516	BF2*110:01_run13_unk89	(140)-38-12-110
BLB1*002:01_AB426141.1	BLB*158:01_run13_unk43	BF1*012:01_AL023516	BF2*110:02_run13_unk104	2-(158)-12-110:02
BLB1*002:01_AB426141.1	BLB*158:01_run13_unk43	BF1*012:04_run13_unk105	BF2*110:02_run13_unk104	2-(158)-12:04-110:02
BLB2*039:03_run13_unk24	BLB2*005:02_run9_unk24	BF1*006:01_AB426143	BF2*111:01_run13_unk419	39:03*-5:02-6-111
BLB*101:01_run8_unk8	BLB*102:01_run8_unk7	BF1*031:05_run13_unk74	BF2*112:01_run13_unk75	(101)-(102)-31:05-112
BLB*109:01_run8_unk1			BF2*113:01_run13_unk92	(109)-?-?-113
BLB1*046:01_run8_unk4	BLB2*046:01_run8_unk5	BF1*012:01_AL023516	BF2*114:01_run13_unk100	46-46-12-114
BLB1*017:02_run13_unk3	BLB2*017:01_AB426150.1	BF1*017:01_AB426150	BF2*115:01_run13_unk40	17:02-17-17-115
BLB1*017:01_AB426150.1	BLB2*017:01_AB426150.1	BF1*071:01_run13_unk41	BF2*115:01_run13_unk40	17-17-71-115
BLB1*009:01_AB426145.1	BLB2*034:02_run13_unk34		BF2*116:01_run13_unk102	9-34:02-?-116
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*012:01_AL023516	BF2*117:01_run13_unk120	2-8-12-117
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*118:01_run13_unk201	5:01_02-5:02-12-118
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*005:03_run8_unk144	BF2*120:01_run13_unk259	4:03-60-5:03-120
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*012:01_AL023516	BF2*121:01_run13_unk172	5:01_02-5:02-12-121
BLB*102:01_run8_unk7	BLB*142:01_run11_unk18	BF1*023:01and04_AB426153	BF2*122:01_run13_unk283	(102)-(142)-23:01and04-122
BLB*115:04_run13_unk77	BLB2*002:02_run13_unk74	BF1*032:02_run9_unk84	BF2*123:01_run13_unk414	(115:04)-2:02-32:02-123
BLB1*046:05_run11_unk26	BLB1*038:02_run13_unk48	BF1*005:09_run13_unk96	BF2*124:01_run13_unk95	46:05-38:02*-5:09-124
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*032:02_run9_unk84	BF2*125:01_run14rpt_unk112	5:01_02-5:02-32:02-125
BLB1*009:01_AB426145.1	BLB2*009:03_run13_unk5	BF1*012:01_AL023516	BF2*127:01_run13_unk277	9-9:03-12-127
BLB1*004:03_AF539401.1_r5	BLB2*060:01_AY744361	BF1*005:03_run8_unk144	BF2*128:01_run13_unk286	4:03-60-5:03-128
BLB1*046:01_run8_unk4	BLB2*046:01_run8_unk5	BF1*012:01_AL023516	BF2*129:01_run14rpt_unk105	46-46-12-129
BLB1*002:01_AB426141.1	BLB2*008:01_AB426144.1	BF1*023:06_run8_unk16	BF2*130:01_run13_UNK_10	2-8-23:06-130
BLB*123:01_run9_unk20	BLB2*054:01_run8_unk32	BF1*023:16_run13_unk3	BF2*131:01_run13_unk1	(123)-54-23:16-131
BLB1*031:02_run14_unk18	BLB*160:01_run14_unk17	BF1*032:01_Av3b	BF2*135:01_run14rpt_unk77	31:02-(160)-32-135
BLB1*005:01-02_AB426142.1	BLB2*005:02_run9_unk24	BF1*023:08:02_run9_unk54	BF2*136:01_run14rpt_unk23	5:01_02-5:02-23:08:02-136
BLB*118:01_run9_unk14	BLB*119:01_run9_unk15	BF1*065:01_run9_unk34	BF2*137:01_run9_unk43	(118)-(119)-65-137
BLB1*009:03_run10_unk6	BLB2*034:01_U76305.1_r5		BLB2*034:01_U76305.1_r5	9:03-34-?-34
BLB1*009:01_AB426145.1	BLB2*034:01_U76305.1_r5		BLB2*034:01_U76305.1_r5	9-34-?-34

A.6 Summaries of Sanger sequencing results obtained after amplifying and cloning passerine TAP1 and TAP2.

In figures A.6.1, A.6.2, A.6.3, A.6.4, A.6.5 and A.6.6 each grey bar represents the aligned nucleotide sequence obtained from a single clone. Black marks indicate deviations from the consensus and therefore show the locations of variable residues. SNPs in exons are highlighted in red if they are non-synonymous and in blue if they are synonymous. The conserved passerine exon structure is annotated in green on each alignment. Where sequence names are prefaced by 'Partial', gaps in the sequence are due to poor quality sequencing and not genuine deletions in the plasmid insert. Figures were generated in Geneious Prime.

A.6.1 Pied flycatcher TAP1 sequences

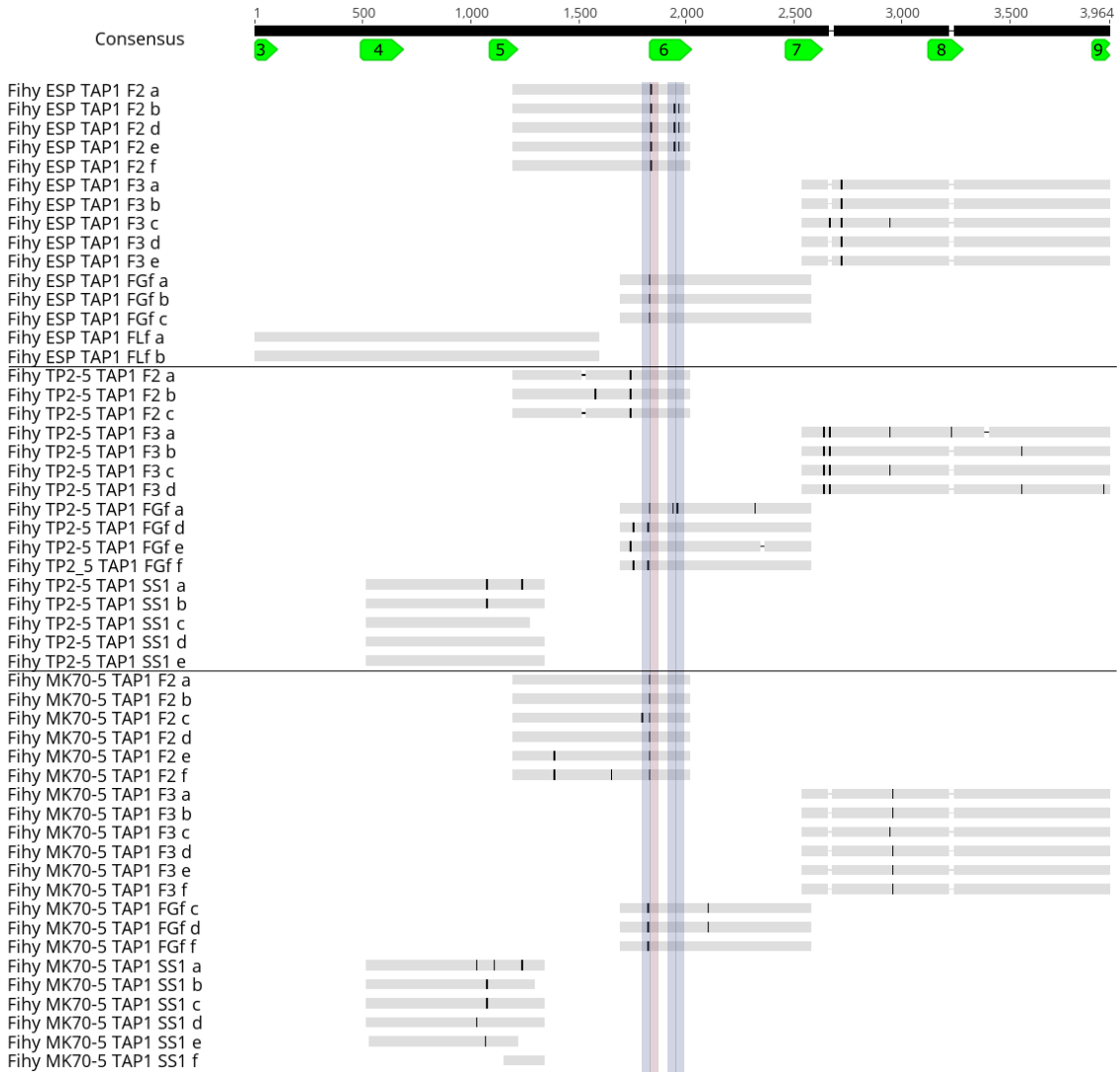


Figure A.3: **TAP1 sequences obtained from cloned PCR products following amplification of TAP1 from pied flycatcher gDNA.** Individual clones are distinguished by lower case letters. All SNPs present in multiple clones were amplified in multiple independant PCRs. All four TAP1 SNPs occur in exon 6; a non-coding and coding SNP occur very close together at the start of the exon and two non-coding SNPs occur very close toegther towards the end of the exon.

A.6.2 Pied flycatcher TAP2 sequences

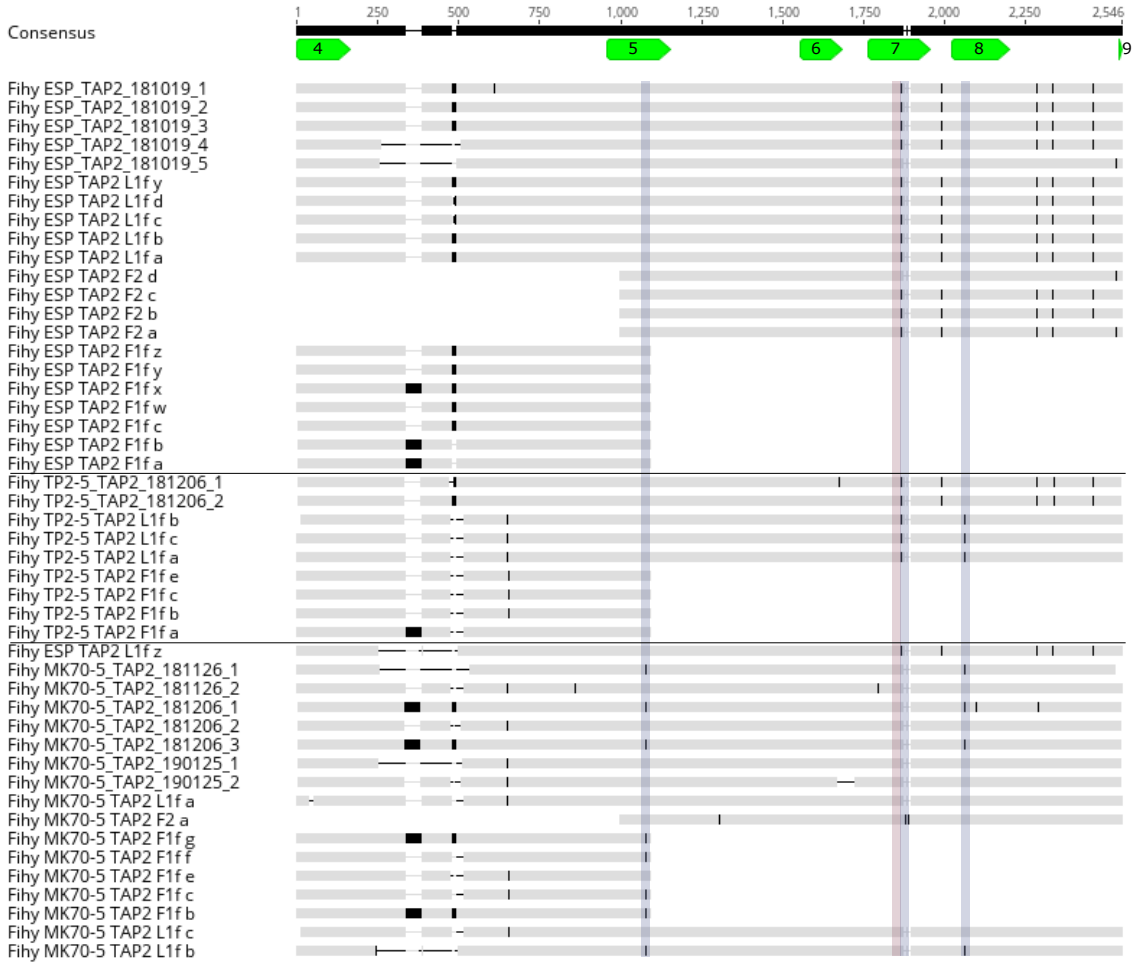


Figure A.4: **TAP2 sequences obtained from cloned PCR products following amplification of TAP2 from pied flycatcher gDNA.** Individual clones are distinguished by lower case letters and in cases where two independant PCRs were performed clones derived from the second amplification are lettered backwards from z. Clones numbered with a six-figure date followed by a clone number were produced and sequenced by George Farmer under supervision; dates represent independent amplifications of the sequence. Intron 4 is characterised by highly repetitive 6- and 5-nt sequence motifs and is therefore vulnerable to amplification, cloning and sequencing errors. Some of the apparent copy number variation in intron 4 correlates with variation elsewhere in the gene, consistent with this intron containing a microsatellite locus.

A.6.3 Sorci sparrow TAP1 sequences

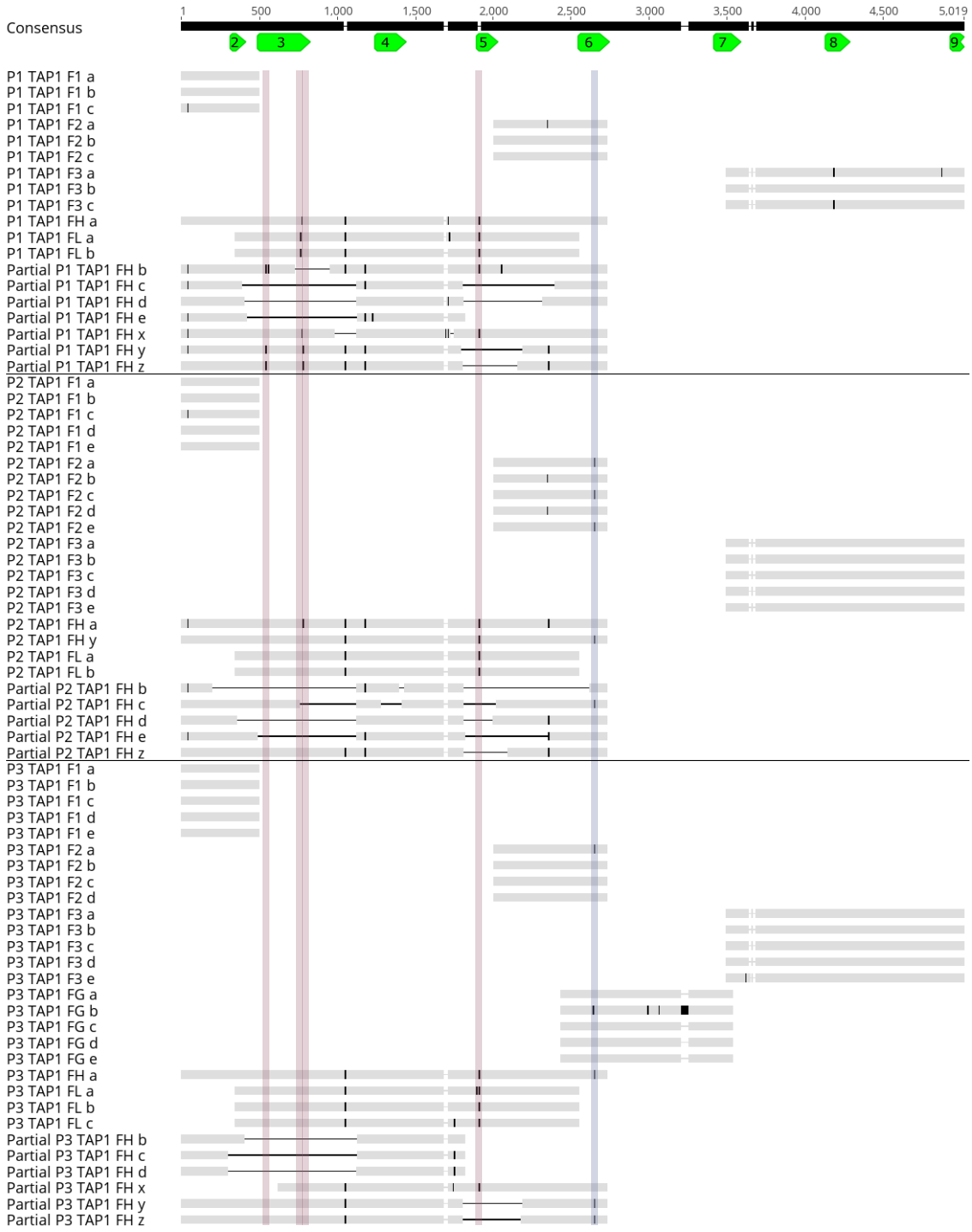


Figure A.5: **TAP1 sequences obtained from cloned PCR products following amplification of TAP1 from house sparrow gDNA.** P1-P3 indicates the sample from which the sequence was amplified. Clones are distinguished by lower case letters, with clones from the first PCR lettered forwards from a and clones from a second PCR (where applicable) lettered backwards from z. All SNPs seen in multiple clones were amplified in multiple independent PCRs.

A.6.4 Sorci sparrow TAP2 sequences

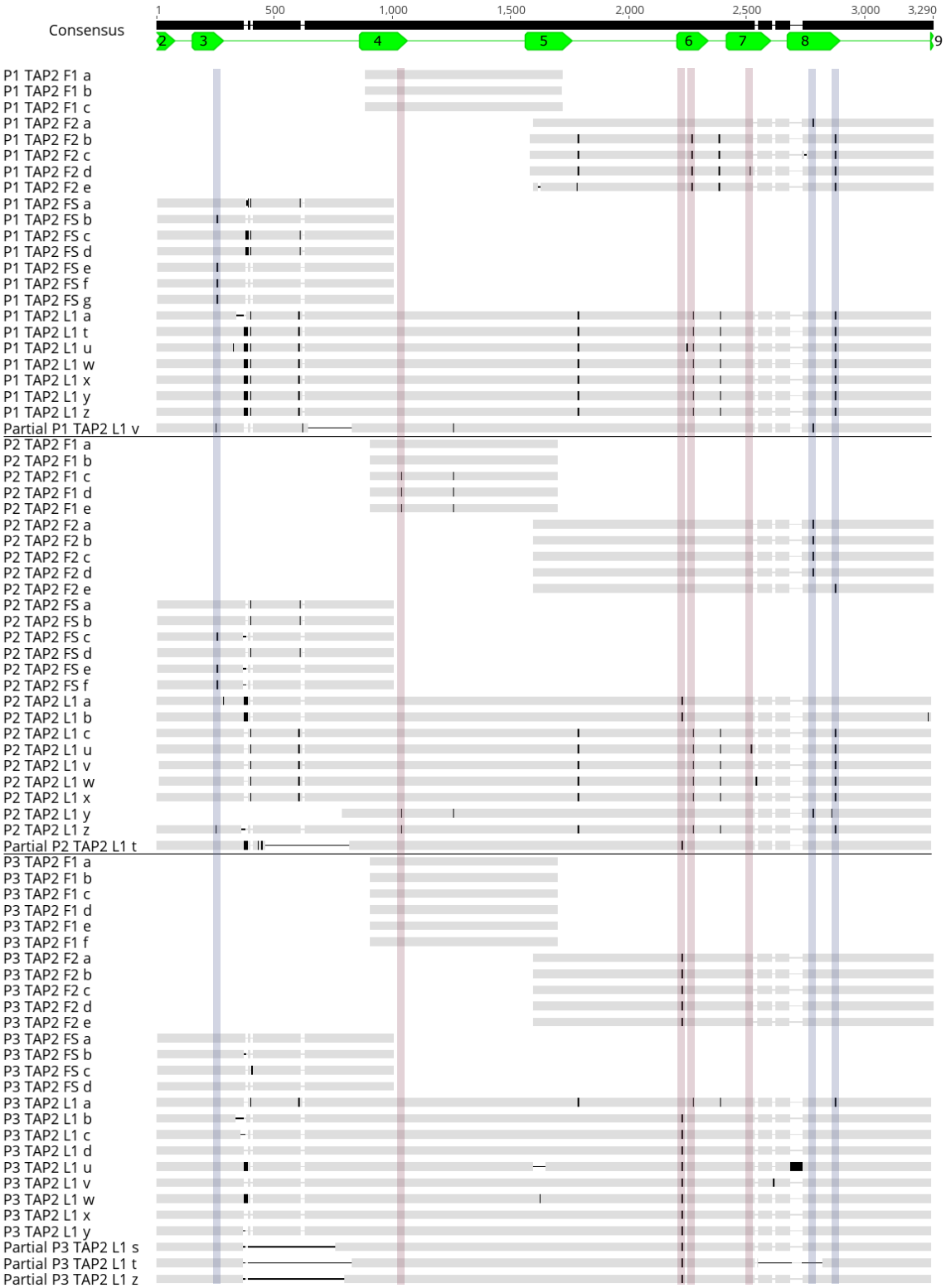


Figure A.6: **TAP2 sequences obtained from cloned PCR products following amplification of TAP2 from house sparrow gDNA.** P1-P3 indicates the sample from which the sequence was amplified. Clones are distinguished by lower case letters, with clones from the first PCR lettered forwards from a and clones from a second PCR (where applicable) lettered backwards from z. All SNPs seen in multiple clones were amplified in multiple independent PCRs. Intron 3 contains tandem CCCAT repeats which vary in copy number. Some of the observed copy number variation correlates with variation elsewhere in the gene and may represent a microsatellite locus.

A.6.5 Lund sparrow TAP1 sequences

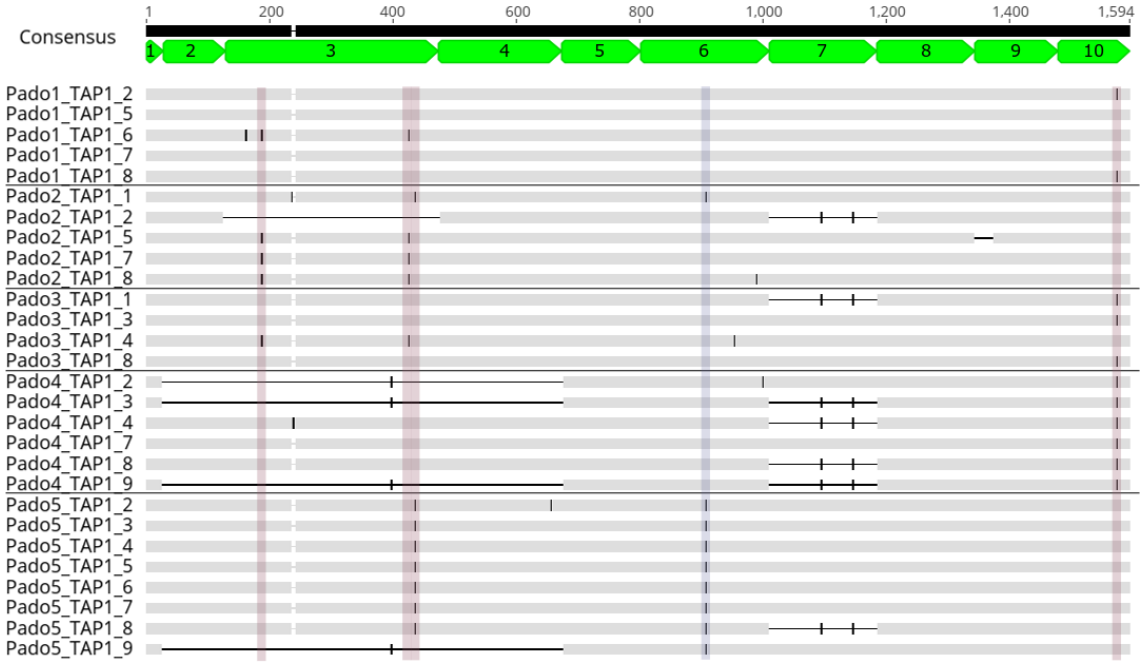


Figure A.7: **TAP1 sequences obtained from cloned PCR products following amplification of TAP1 from house sparrow cDNA.** All SNPs in more than one clone were also seen in more than one independent PCR. Sequences apparently missing exon 7 stay in frame and could represent an alternatively spliced protein product. Sequences missing exons 2-4 do not stay in frame.

A.6.6 Lund sparrow TAP2 sequences

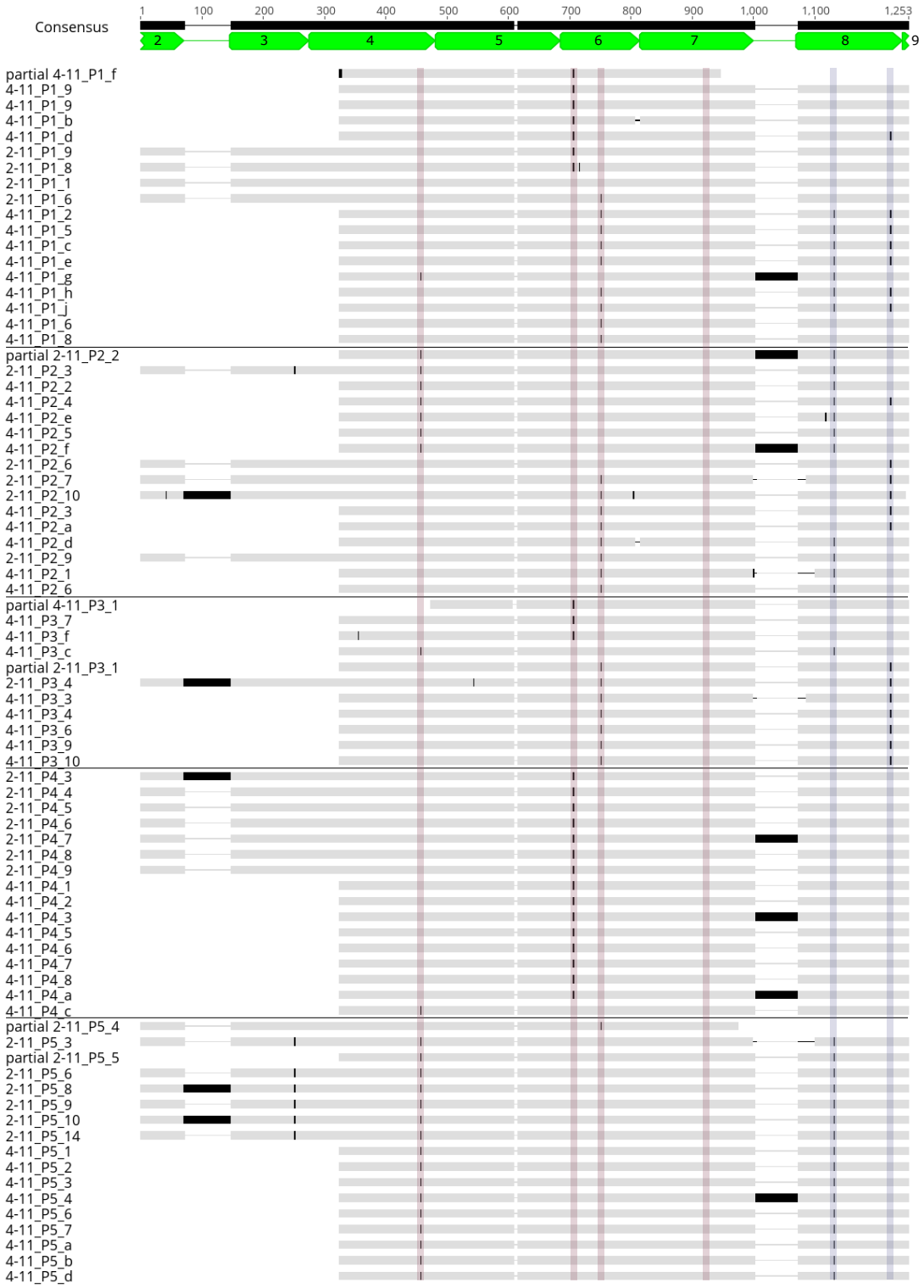


Figure A.8: **TAP2 sequences obtained from cloned PCR products following amplification of TAP2 from house sparrow cDNA.** Caption continued overleaf.

Figure A.6.6 caption cont. All SNPs in more than one clone were also seen in more than one independant PCR. '2-11' and '4-11' correspond to the Lund_2.2 and Lund_2.4 amplicons respectively. P1-P5 in sequence names refer to Lund Pado samples 1-5 and independent PCR reactions are indicated by the use of either numbers or letters in the final field of the sequence name. Neither apparent retained intron is likely to be a genuine splice variant; intron 2 would code for an in-frame stop codon and retention of intron 7 would lead to in-frame stop codons in exon 8.

A.7 Sparrow class I exon 3 sequences

Sequences were obtained by Sanger sequencing multiple clones by Kevin Chen, Josep Montserrat Sanchez, Derry Bo Li, Dr. Clive Tregaskes and Dr. Hannah Siddle (all University of Cambridge). Top alignment shows nucleotide sequences obtained from all clones. Bottom alignment shows nucleotide and amino acid sequences for the subset of alleles which occurred in more than one clone, as well as the three sequences W, V and k, which represent a potential non-classical locus in Pado 1. The first codon is equivalent to residue 124 in chicken BF2. Dots indicate identity to consensus.

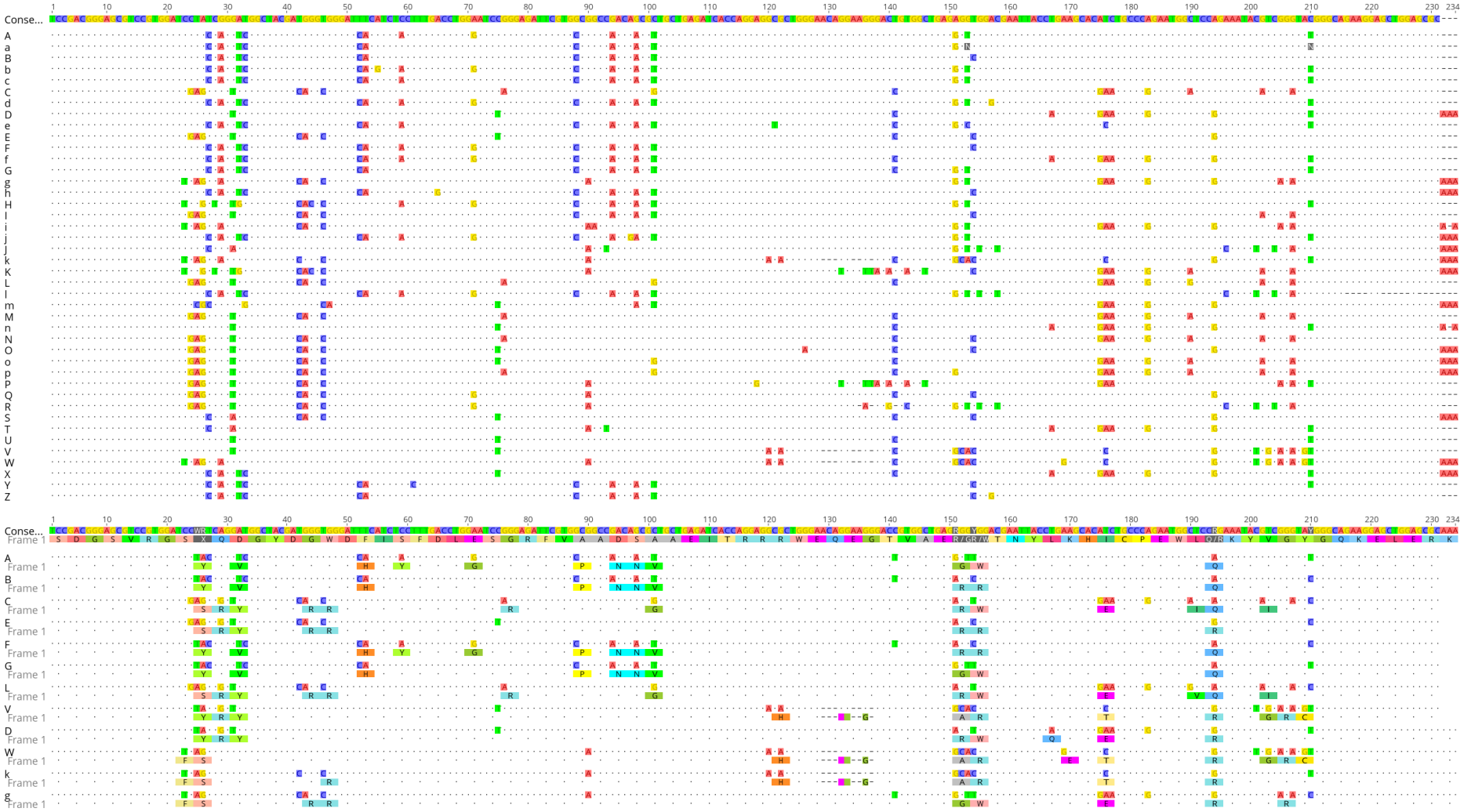


Figure A.9: **Sparrow class I exon 3 nucleotide sequences from five individuals.**

A.8 Zebra Finch TAP1 and TAP2 cDNA sequences

Nucleotide and amino acid sequences for each unique zebra finch TAP sequence identified in the Oxford Nanopore dataset are shown. Sequence names correspond to the numerically first occurrence of each particular sequence in the dataset. Dots indicate identity to the consensus. The passerine exon structure is annotated against the consensus with exons labelled at the start of the exon.

TAP1

Figure A.10: **Zebra Finch TAP1 allele sequences.**

TAP2

Figure continues on next page

Figure A.11: **Zebra Finch TAP2 allele sequences.**

A.9 Zebra Finch MHC class I exon 2 and 3 sequences

Sequences which were found in more than one clone are shown as nucleotide and translated amino acid sequences. Numbers in brackets after sequence names indicate the number of clones in which that sequence was found. Dots indicate identity to consensus. The first codon is equivalent to residue 34 in the chicken BF2 sequence.

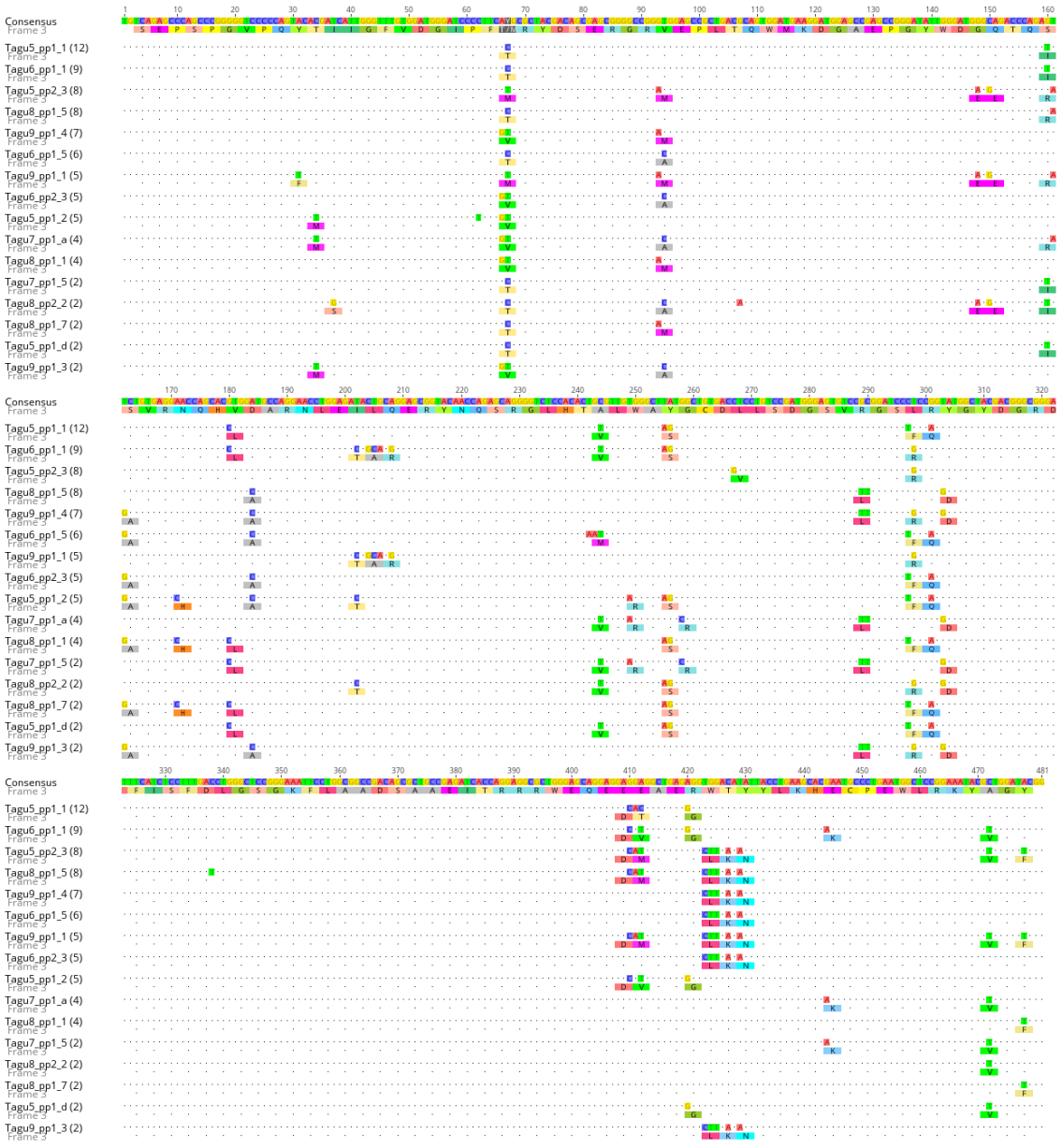


Figure A.12: **Zebra Finch MHC class I exon 2 and 3 sequences.**

A.10 Chicken cell line TAPBPR cDNA sequences

Sequences were obtained from multiple clones following amplification of TAPBPR from cDNA derived from five chicken cell lines. Figure A.13 shows a nucleotide alignment of all clone sequences which matched at least part of the TAPBPR gene. ENSGALT00000090321.2, the coding sequence for the TAPBPL gene annotated in the GRCg6a genome in the Ensembl database is also included. 'Partial' in a sequence name indicates that regions which are missing relative to other clones are the result of poor sequencing quality rather than genuine deletions in the sequence. Black marks indicate deviation from the consensus of the alignment and therefore represent positions of both coding and non-coding SNPs.

Figures A.13 and A.14 shows a nucleotide and protein sequence alignment of the consensus of clone sequences from each cell line.

Exon boundaries are annotated in red underneath the genomic sequence.

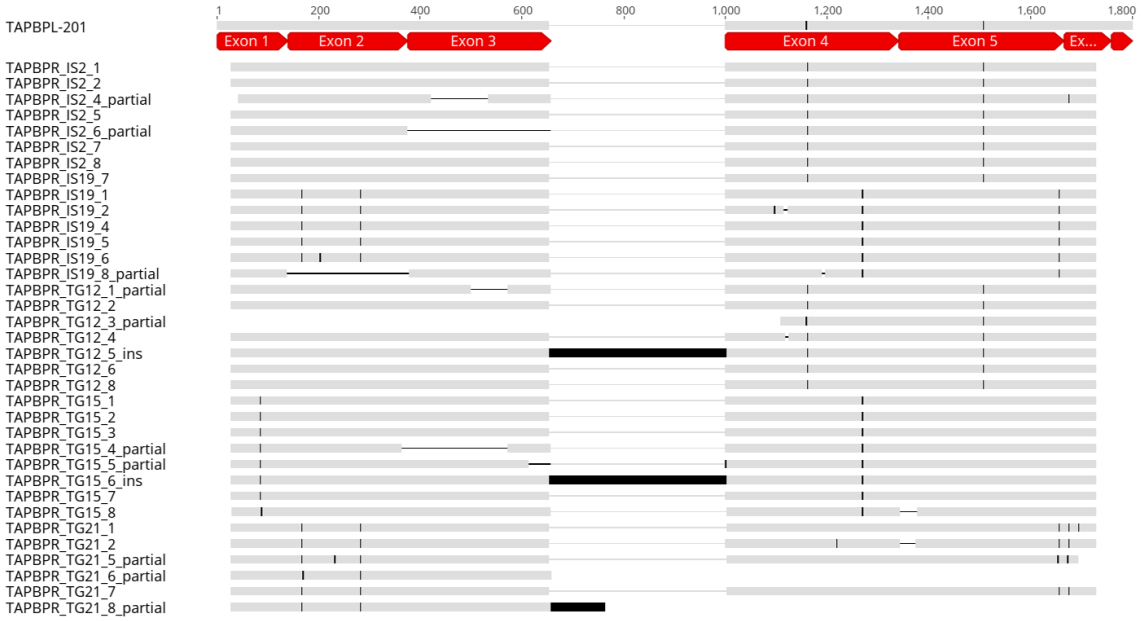


Figure A.13: **TAPBPR cDNA clone sequences obtained from five chicken cell lines.** Each line appeared to have a single dominant allele sequence present. In two sequences intron 3 was retained and in one additional sequence the start of intron 3 was retained but sequencing quality was too poor over the second half of the gene to resolve the sequence.

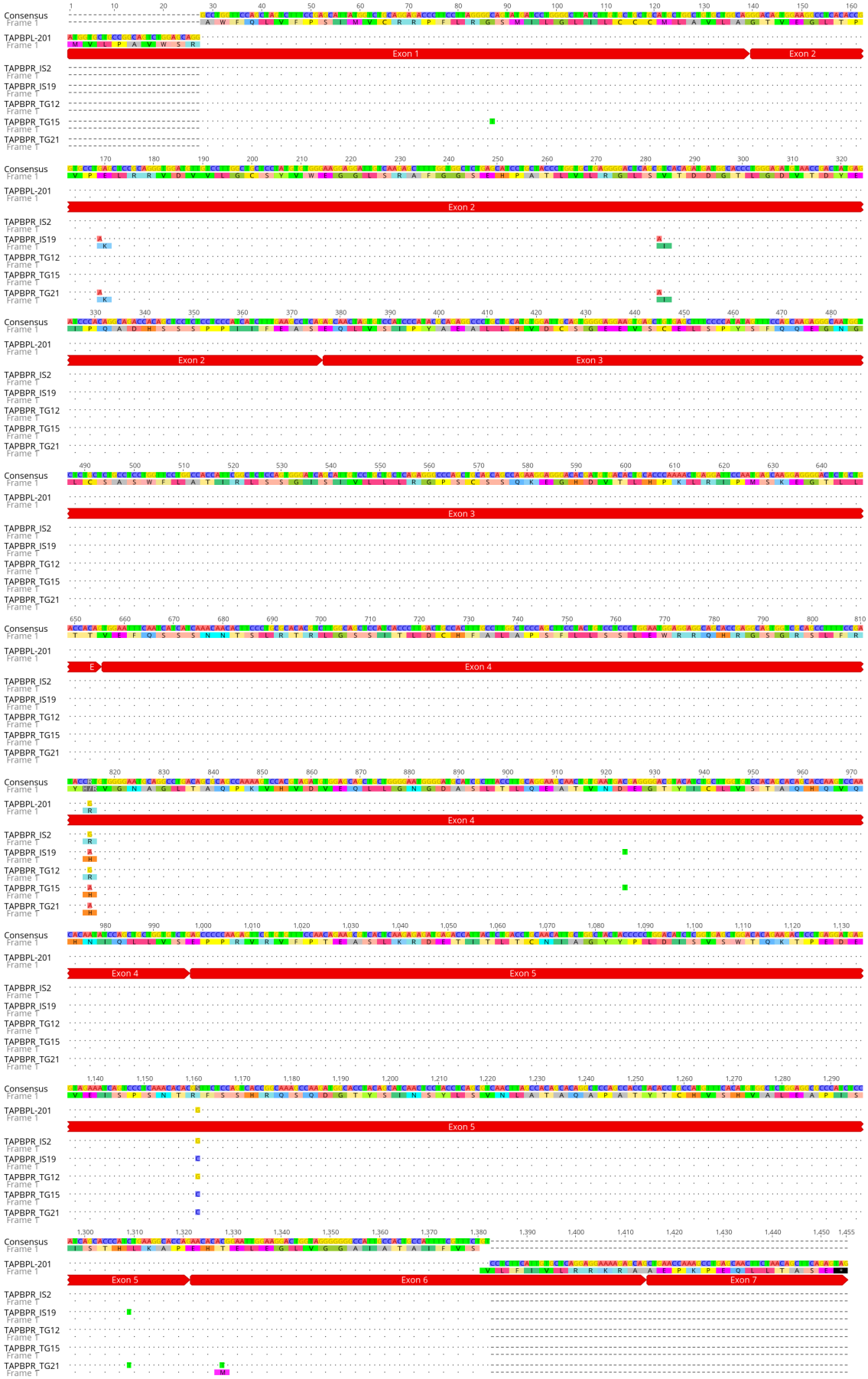


Figure A.14: **TAPBPR nucleotide and protein consensus sequences obtained from five chicken cell lines.**

A.11 Chicken cell line TAPBPL cDNA sequences

Sequences were obtained from multiple clones following amplification of TAPBPL from cDNA derived from five chicken cell lines. Figure A.15 shows a nucleotide alignment of all clone sequences which matched at least part of the TAPBPL gene. TAPBPL-201, the coding sequence for the TAPBPL gene annotated in the GRCg6a genome in the Ensembl database (ENSGALT00000023307.4) is also included. 'Partial' in a sequence name indicates that regions which are missing relative to other clones are the result of poor sequencing quality rather than genuine deletions in the sequence. Black marks indicate deviation from the consensus of the alignment and therefore represent positions of both coding and non-coding SNPs.

Figures A.15 and A.16 shows a nucleotide and protein sequence alignment of the consensus of clone sequences from each cell line. IS2, IS19, TG15 and TG21 had identical consensus sequences so only IS2 is shown.

Exon boundaries are annotated in red underneath the genomic sequence.

Figure A.15: **TAPBPL cDNA clone sequences obtained from five chicken cell lines.** Partial introns were retained in one (intron 3) and three (intron 5) sequences, but in no case was the full intron, as indicated by the genomic sequence, retained.

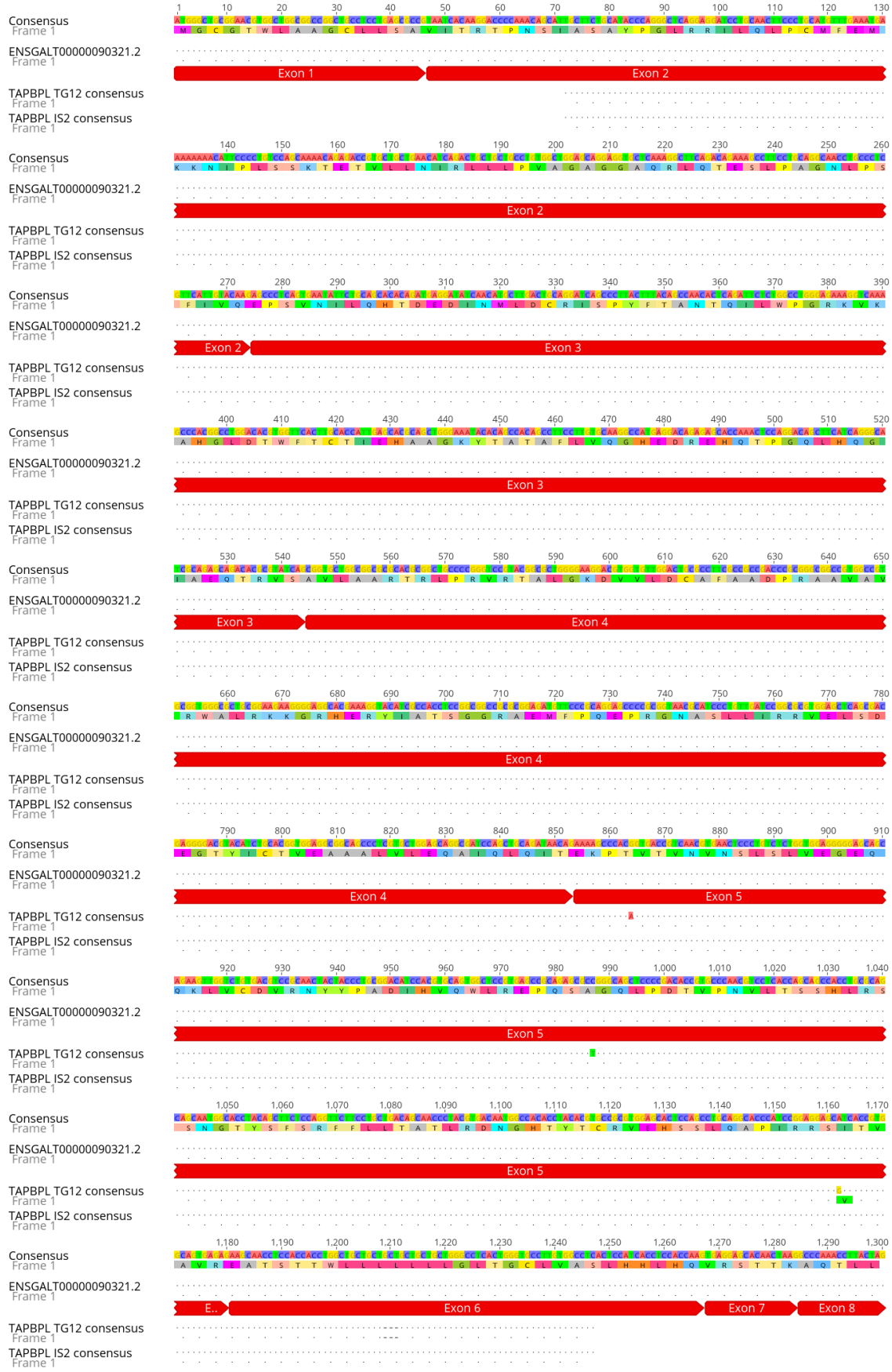


Figure A.16: **TAPBPL nucleotide and protein consensus sequences obtained from five chicken cell lines.**

A.12 No evidence for phosphorylation of any of the protein species stained with F21-2, 11-46-18. 19-53-11 or 19-54-11 was found

Cell lysates made from all five cell lines were treated with lambda protein phosphatase (lambda PP; NEB) according to the manufacturer's standard protocol. No antibodies against known phosphorylated chicken proteins were available to use as a positive control. 6×10^5 or 3×10^6 cell equivalents were loaded per lane for the F21-2 and 11-46-18/19-53-11/19-54-11 experiments respectively.

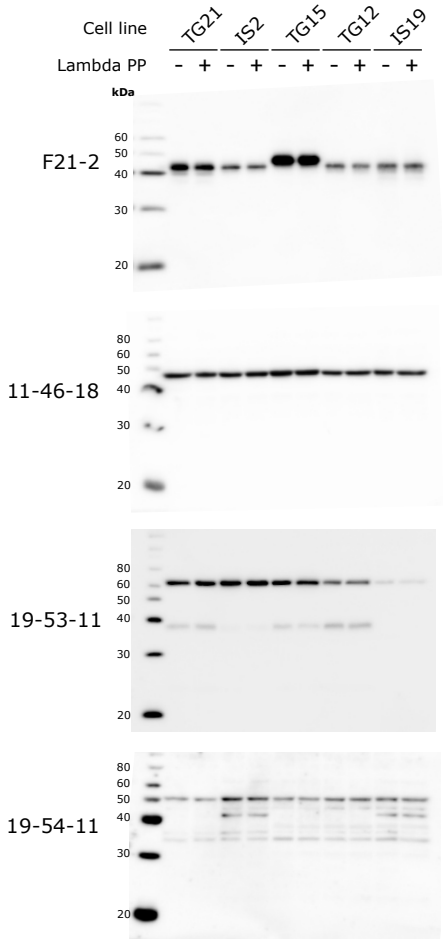


Figure A.17: **No evidence for phosphorylation of any of the protein species stained with F21-2, 11-46-18. 19-53-11 or 19-54-11 was found.** F21-2: α-MHC class I heavy chain, 11-46-18: α-tapasin extracellular domain, 19-53-11: α-TAPBPR C-terminal peptide, 19-54-11: α-TAPBPL C-terminal peptide

A.13 Initial studies suggested that variation in expression of proteins bound by 19-54-11 could relate to cell stress and/or immune stimulation

Experiments performed at roughly the same time as those presented in figures 4.12, 4.13 and 4.14 suggested that variation in expression of proteins which stained with 19-54-11 could be related to cell stress and/or activation of immune signalling pathways. However, later attempts to replicate the results presented here (figure A.18) were unsuccessful. At this later timepoint a repeat of the direct comparison between cell lines to determine if the same degree of variation in 19-54-11 staining occurred was not performed.

In the interferon gamma (IFN-γ) stimulation experiment, two dilutions of lab-produced chicken IFN-γ (of unknown protein concentration) and four concentrations of commercial chicken IFN-γ (Kingfisher Biotech) were added to 5×10^6 IS2 cells in 25 ml total cell culture. After 48 h membrane enriched cell lysates were produced and 1.5×10^6 cell equivalents of lysate was loaded in each well of an SDS-PAGE gel (3×10^5 for gels which would be stained with F21-2). Western blotting was performed as described in section 4.2.7. The blots showed variation in protein expression when stained with 19-54-11 but not when stained with any other antibody, despite upregulation of MHC class I, tapasin and TAPBPR (at least in humans) typically being detected following IFN-γ stimulation. There was a trend towards higher expression in flasks with more IFN-γ, possibly with no further induction above concentrations of 1 ng/ml.

A separate preliminary experiment tested cells grown at five different densities, after seeding various volumes of IS2 culture into new media such that the total volume was 25 ml. Cells were harvested and analysed as described above. There was some indication that cells grown at higher densities, predicted to be exhibiting generalised stress responses, were expressing lower levels of proteins which stained with 19-54-11.

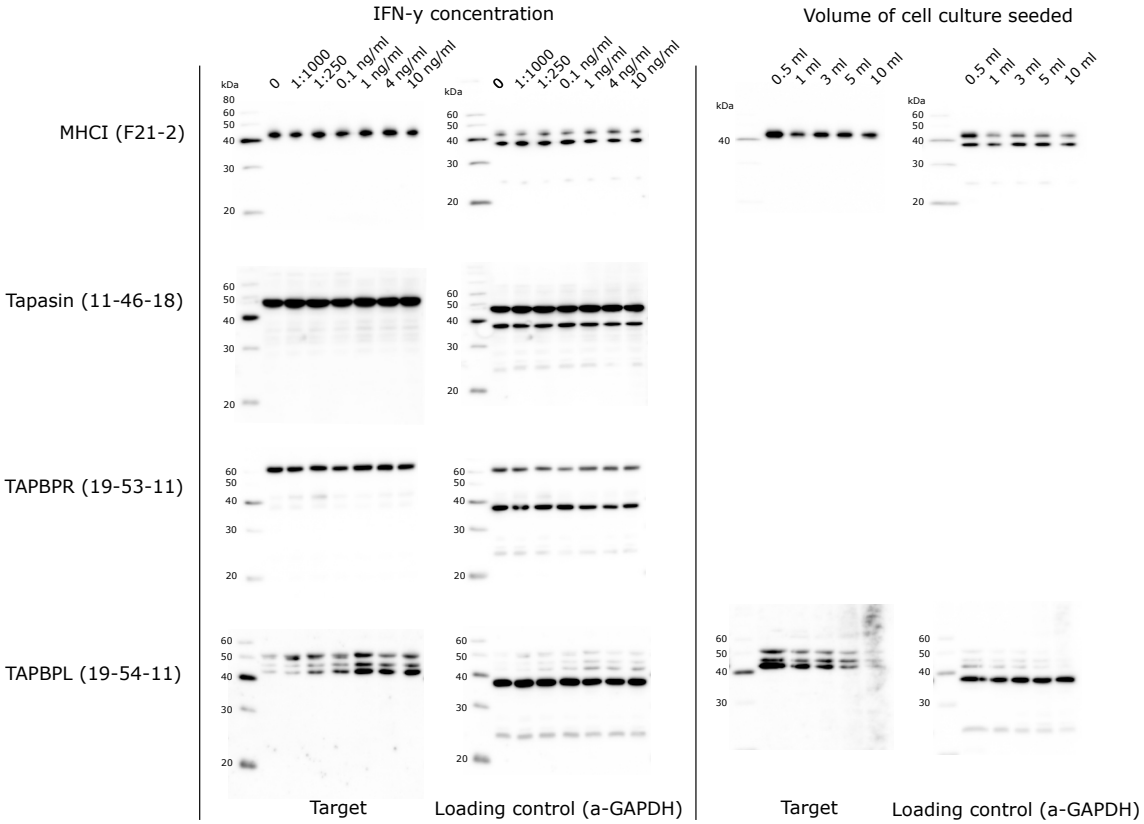


Figure A.18: **Variation in expression of proteins bound by 19-54-11 may relate to cell stress and/or immune stimulation.**

Bibliography

Afrache, H., Tregaskes, C. A. and Kaufman, J. (2020), 'A potential nomenclature for the Immuno Polymorphism Database (IPD) of chicken MHC genes: progress and problems', *Immunogenetics* **72**, 9–24.

Ajibike, A. B., Adeleye, O. O., Ilori, B. M., Osinbowale, D. A., Adeniyi, O. A., Durosaro, S. O., Sanda, A. J., Adebambo, O. A. and Adebambo, A. O. (2017), 'Genetic diversity, phylogeographic structure and effect of selection at the mitochondrial hypervariable region of Nigerian chicken populations', *Journal of Genetics* **96**(6), 959–968.

Al-Jumaili, A. S., Boudali, S. F., Kebede, A., Al-Bayatti, S. A., Essa, A. A., Ahbara, A., Aljumaah, R. S., Alatiyat, R. M., Mwacharo, J. M., Bjørnstad, G., Naqvi, A. N., Gaouar, S. B. S. and Hanotte, O. (2020), 'The maternal origin of indigenous domestic chicken from the Middle East, the north and the horn of Africa', *BMC Genetics* **21**(30), 1–16.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990), 'Basic local alignment search tool.', *Journal of molecular biology* **215**(3), 403–410.

Amemiya, C. T., Alfoldi, J., Lee, A. P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J. J., Robinson, M., Dorrington, R. A., Gerdol, M., Aken, B., Biscotti, M. A., Barucca, M., Baurain, D., Berlin, A. M., Blatch, G. L., Buonocore, F., Burmester, T., Campbell, M. S., Canapa, A., Cannon, J. P., Christoffels, A., De Moro, G., Edkins, A. L., Fan, L., Fausto, A. M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J. V., Haerty, W., Hahn, M. E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S. I., Kuraku, S., Lara, M., Levin, J. Z., Litman, G. W., Mauceli, E., Miyake, T., Mueller, M. G., Nelson, D. R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C. P., Prohaska, S. J., Przybylski, D., Saha, N. R., Ravi, V., Ribeiro, F. J., Sauka-Spengler, T., Scapigliati, G., Searle, S. M., Sharpe, T., Simakov, O., Stadler, P. F., Stegeman, J. J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., Van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J. N., Tabin, C. J., Shubin, N., Schartl, M., Jaffe, D. B., Postlethwait, J. H., Venkatesh, B., Di Palma, F., Lander, E. S., Meyer, A. and Lindblad-Toh, K. (2013), 'The African coelacanth genome provides insights into tetrapod evolution', *Nature* **496**(7445), 311–316.

Anisimova, M., Bielawski, J. P. and Yang, Z. (2001), 'Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution', *Molecular Biology and Evolution* **18**(8), 1585–1592.

Aultman, D., Adamashvili, I., Yaturu, K., Langford, M., Gelder, F., Gautreaux, M., Ghali, G. E. and McDonald, J. (1999), 'Soluble HLA in human body fluids', *Human Immunology* **60**, 239–244.

Babik, W. (2010), 'Methods for MHC genotyping in non-model vertebrates', *Molecular Ecology Resources* **10**(2), 237–251.

Bacon, L. D. and Witter, R. L. (1993), 'Influence of B-Haplotype on the Relative Efficacy of Marek's Disease Vaccines of Different Serotypes', *Avian Diseases* **37**(1), 53–59.

Bacon, L. D. and Witter, R. L. (1994), 'Serotype Specificity of B-Haplotype Influence on the Relative Efficacy of Marek 's Disease Vaccines', *Avian Diseases* **38**(1), 65–71.

Bacon, L. D. and Witter, R. L. (1995), 'Efficacy of Marek's Disease Vaccines in Mhc Heterozygous Chickens: Mhc Congenic × Inbred Line F1 Matings', *Journal of Heredity* **86**(4), 269–273.

Bacon, L. D., Witter, R. L., Crittenden, L. B., Fadly, A. and Motta, J. (1981), 'B-Haplotype Influence on Marek 's Disease, Rous Sarcoma, and Lymphoid Leukosis Virus-Induced Tumors in Chickens', *Poultry Science* **60**, 1132–1139.

Bailes, S. M., Devers, J. J., Kirby, J. D. and Rhoads, D. D. (2007), 'An inexpensive, simple protocol for DNA isolation from blood for high-throughput genotyping by polymerase chain reaction or restriction endonuclease digestion.', *Poultry Science* **86**, 102–106.

Balakrishnan, C. N., Ekblom, R., Völker, M., Westerdahl, H., Godinez, R., Kotkiewicz, H., Burt, D. W., Graves, T., Griffin, D. K., Warren, W. C. and Edwards, S. V. (2010), 'Gene duplication and fragmentation in the zebra finch major histocompatibility complex', *BMC Biology* **8**(29), 1–19.

Banat, G. R., Tkalcic, S., Dzielawa, J. A., Jackwood, M. W., Saggese, M. D., Yates, L., Kopulos, R., Briles, W. E. and Collisson, E. W. (2013), 'Association of the chicken MHC B haplotypes with resistance to avian coronavirus', *Developmental and Comparative Immunology* **39**, 430–437.

Bashirova, A. A., Thomas, R. and Carrington, M. (2011), 'HLA/KIR restraint of HIV: Surviving the fittest', *Annual Review of Immunology* **29**, 295–317.

Bashirova, A. A., Viard, M., Naranbhai, V., Grifoni, A., Garcia-Beltran, W., Akdag, M., Yuki, Y., Gao, X., O'hUigin, C., Raghavan, M., Wolinsky, S., Bream, J. H., Duggal, P., Martinson, J., Michael, N. L., Kirk, G. D., Buchbinder, S. P., Haas, D., Goedert, J. J., Deeks, S. G., Fellay, J., Walker, B., Goulder, P., Cresswell, P., Elliott, T., Sette, A., Carlson, J. and Carrington, M. (2020), 'HLA tapasin independence: broader peptide repertoire and HIV control', *Proceedings of the National Academy of Sciences of the United States of America* **117**(45), 28232–28238.

Bauer, M. M. and Reed, K. M. (2011), 'Extended sequence of the turkey MHC B-locus and sequence variation in the highly polymorphic B-G loci', *Immunogenetics* **63**(4), 209–221.

Beck, S., Kelly, A., Radley, E., Khurshid, F., Alderton, R. P. and Trowsdale, J. (1992), 'DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing', *Journal of Molecular Biology* **228**(2), 433–441.

Belov, K., Deakin, J. E., Papenfuss, A. T., Baker, M. L., Melman, S. D., Siddle, H. V., Gouin, N., Goode, D. L., Sargeant, T. J., Robinson, M. D., Wakefield, M. J., Mahony, S., Cross, J. G., Benos, P. V., Samollow, P. B., Speed, T. P., Marshall Graves, J. A. and Miller, R. D. (2006), 'Reconstructing an ancestral mammalian immune supercomplex from a marsupial major histocompatibility complex', *PLoS Biology* **4**(3), 0317–0328.

Bennett, E. M., Bennink, J. R., Yewdell, J. W. and Brodsky, F. M. (1999), 'Cutting edge: adenovirus E19 has two mechanisms for affecting class I MHC expression.', *The Journal of Immunology* **162**(9), 5049–52.

Bentkowski, P. and Radwan, J. (2019), 'Evolution of major histocompatibility complex gene copy number', *PLoS Computational Biology* **15**(5), 1–15.

Berima, M. e. A., Yousif, I. A., Eding, H., Weigend, S. and Musa, H. H. (2013), 'Population structure and genetic diversity of Sudanese native chickens', *African Journal of Biotechnology* **12**(45), 6424–6431.

Bernatchez, L. and Landry, C. (2003), 'MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years?', *Journal of Evolutionary Biology* **16**(3), 363–377.

Berres, M. E., Kantanen, J., Honkatukia, M., Wolc, A. and Fulton, J. E. (2020), 'Heritage Finnish Landrace chickens are genetically diverse and geographically structured', *Acta Agriculturae Scandinavica A: Animal Sciences* **69**(1-2), 81–94.

Beutler, N., Hauka, S., Niepel, A., Kowalewski, D. J., Uhlmann, J., Ghanem, E., Erkelenz, S., Wiek, C., Hanenberg, H., Schaal, H., Stevanović, S., Springer, S., Momburg, F., Hengel, H. and Halenius, A. (2013), 'A natural tapasin isoform lacking exon 3 modifies peptide loading complex function', *European Journal of Immunology* **43**(6), 1459–1469.

Billerman, S. M., Keeney, B. K., Rodewald, P. G. and Schulenberg, T. S. (2020), 'Birds of the World'. Cornell Laboratory of Ornithology, [online] accessed 06/01/2021.
URL: *https://birdsoftheworld.org/bow/home*

Birnboim, H. C. and Doly, J. (1979), 'A rapid alkaline extraction procedure for screening recombinant plasmid DNA', *Nucleic Acids Research* **7**(6), 1513–1523.

Blackwell, J. M., Jamieson, S. E. and Burgner, D. (2009), 'HLA and Infectious Diseases', *Clinical Microbiology Reviews* **22**(2), 370–385.

Blees, A., Januliene, D., Hofmann, T., Koller, N., Schmidt, C., Trowitzsch, S., Moeller, A. and Tampé, R. (2017), 'Structure of the human MHC-I peptide-loading complex', *Nature* **551**(7681), 525–528.

Bodmer, W. F., Albert, E., Bodmer, J. G., Dupont, B., Mach, B., Mayr, W. R., Sasazuki, T., Schreuder, G. M. T., Svejgaard, A. and Terasaki, P. I. (1989), *Nomenclature for factors of the HL-A system, 1987*, Vol. 1, Springer, New York.

Bonneaud, C., Chastel, O., Federici, P., Westerdahl, H. and Sorci, G. (2006), 'Complex Mhc-based mate choice in a wild passerine', *Proceedings of the Royal Society B: Biological Sciences* **273**, 1111–1116.

Bonneaud, C., Sorci, G., Morin, V., Westerdahl, H., Zoorob, R. and Wittzell, H. (2004), 'Diversity of Mhc class I and IIB genes in house sparrows (Passer domesticus)', *Immunogenetics* **55**(12), 855–865.

Boonyanuwat, K., Thummabutra, S., Sookmanee, N. and Vatchavalkhu, V. (2006), 'Influences of major histocompatibility complex class I haplotypes on avian influenza virus disease traits in Thai indigenous chickens', *Animal Science Journal* **77**, 285–289.

Bortoluzzi, C., Crooijmans, R. P., Bosse, M., Hiemstra, S. J., Groenen, M. A. and Megens, H. J. (2018), 'The effects of recent changes in breeding preferences on maintaining traditional Dutch chicken genomic diversity', *Heredity* **121**(6), 564–578.

Boyle, L. H., Hermann, C., Boname, J. M., Porter, K. M., Patel, P. A. and Burr, M. L. (2013), 'Tapasin-related protein TAPBPR is an additional component of the MHC class I presentation pathway', *Proceedings of the National Academy of Sciences of the United States of America* **110**(9), 3465–3470.

Briles, W. E., Allen, C. P. and W, M. T. (1957), 'The B Blood Group System of Chickens. I. Heterozygosity in Closed Populations', *Genetics* **42**(5), 631–648.

Briles, W. E., Bumstead, N., Ewert, D. L., Gilmour, D. G., Gogusev, J., Hála, K., Koch, C., Longenecker, B. M., Nordskog, A. W., Pink, J. R., Schierman, L. W., Simonsen, M., Toivanen, A., Toivanen, P., Vainio, O. and Wick, G. (1982), 'Nomenclature for chicken major histocompatibility (B) complex', *Immunogenetics* **15**(5), 441–447.

Briles, W. E., McGibbon, W. H. and Irwin, M. R. (1950), 'On multiple alleles effecting cellular antigens in the chicken.', *Genetics* **35**(6), 633–652.

Briles, W. E., Stone, H. A. and Cole, R. K. (1977), 'Marek 's Disease: Effects of B Histocompatibility Alloalleles in Resistant and Susceptible Chicken Lines', *Science* **195**(4274), 193–196.

Brouwer, L., Barr, I., van de Pol, M., Burke, T., Komdeur, J. and Richardson, D. S. (2010), 'MHC-dependent survival in a wild population: evidence for hidden genetic benefits gained through extra-pair fertilizations.', *Molecular Ecology* **19**(16), 3444–3455.

Buitenhuis, A. J., Rodenburg, T. B., Van Hierden, Y. M., Siwek, M., Cornelissen, S. J., Nieuwland, M. G., Crooijmans, R. P., Groenen, M. A., Koene, P., Korte, S. M., Bovenhuis, H. and Van Der Poel, J. J. (2003), 'Mapping quantitative trait loci affecting feather pecking behavior and stress response in laying hens', *Poultry Science* **82**(8), 1215–1222.

Burgdorf, S., Schölz, C., Kautz, A., Tampé, R. and Kurts, C. (2008), 'Spatial and mechanistic separation of cross-presentation and endogenous antigen presentation', *Nature Immunology* **9**(5), 558–566.

Butter, C., Staines, K., Van Hateren, A., Davison, T. F. and Kaufman, J. (2013), 'The peptide motif of the single dominantly expressed class i molecule of the chicken MHC can explain the response to a molecular defined vaccine of infectious bursal disease virus (IBDV)', *Immunogenetics* **65**(8), 609–618.

Cai, Y. and Sun, Y. (2011), 'ESPRIT-Tree: Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time', *Nucleic Acids Research* **39**(14), 1–10.

Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K. and Brien, S. J. O. (1999), 'HLA and HIV-1: Heterozygote Advantage and B835-Cw*04 Disadvantage', *Science* **283**, 1748–1753.

Carter, M. J. and Milton, I. D. (1993), 'An inexpensive and simple method for DNA purifications on silica particles', *Nucleic Acids Research* **21**(4), 1044.

Cavero, D., Schmutz, M., Philipp, H. C. and Preisinger, R. (2009), 'Breeding to reduce susceptibility to Escherichia coli in layers', *Poultry Science* **88**, 2063–2068.

Centre for Disease Control and Prevention (2020), 'Summary of Influenza Risk Assessment Tool (IRAT) Results'. [online] accessed 09/03/2021.
URL: *https://www.cdc.gov/flu/pandemic-resources/monitoring/irat-virus-summaries.htm*

Chan, W. F., Parks-Dely, J. A., Magor, B. G. and Magor, K. E. (2016), 'The Minor MHC Class I Gene UDA of Ducks Is Regulated by Let-7 MicroRNA', *The Journal of Immunology* **197**, 1212–1220.

Chappell, P. E., Meziane, E. K., Harrison, M., Magiera, Ł., Hermann, C., Mears, L., Wrobel, A. G., Durant, C., Nielsen, L. L., Buus, S., Ternette, N., Mwangi, W., Butter, C., Nair, V., Ahyee, T., Duggleby, R., Madrigal, A., Roversi, P., Lea, S. M. and Kaufman, J. (2015), 'Expression levels of MHC class I molecules are inversely correlated with promiscuity of peptide binding', *eLife* **4**, e05345.

Chazara, O., Juul-Madsen, H. R., Chang, C.-S., Tixier-Boichard, M. and Bed'hom, B. (2011), 'Correlation in chicken between the marker LEI0258 alleles and Major Histocompatibility Complex sequences', *BMC Proceedings* **5**(S4), 1–5.

Cheeseman, J. H., Kaiser, M. G., Ciraci, C., Kaiser, P. and Lamont, S. J. (2007), 'Breed effect on early cytokine mRNA expression in spleen and cecum of chickens with and without Salmonella enteritidis infection', *Developmental and Comparative Immunology* **31**(1), 52–60.

Chen, L. C., Lan, H., Sun, L., Deng, Y. L., Tang, K. Y. and Wan, Q. H. (2015), 'Genomic organization of the crested ibis MHC provides new insight into ancestral avian MHC structure', *Scientific Reports* **5**, 7963.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2016), 'GenBank', *Nucleic Acids Research* **44**, 67–72.

Clauer, P. (2012a), 'Modern Meat Chicken Industry'. PennState Extension [online] accessed 27/03/2021.
URL: *https://extension.psu.edu/modern-meat-chicken-industry*

Clauer, P. (2012b), 'Modern Egg Industry'. PennState Extension [online] accessed 27/03/2021.
URL: *https://extension.psu.edu/modern-egg-industry*

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T. and Kauff, F. (2009), 'Biopython : freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics* **25**(11), 1422–1423.

Collins, W. M., Briles, W. E., Zsigray, R. M., Dunlop, W. R., Corbett, A. C., Clark, K. K., Marks, J. L. and Mcgrail, T. P. (1977), 'The B Locus (MHC) in the Chicken: Association with the Fate of RSV-Induced Tumors', *Immunogenetics* **5**, 333–343.

Copley, S. D. (2000), 'Evolution of a metabolic pathway for degradation of a toxic xenobiotic: The patchwork approach', *Trends in Biochemical Sciences* **25**(6), 261–265.

Crone, M., Simonsen, M., Skjødt, K., Linnet, K. and Olsson, L. (1985), 'Mouse monoclonal antibodies to class I and class II antigens of the chicken MHC', *Immunogenetics* **21**(2), 181–187.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gil, L., Grego, T., Haggerty, L., Gall, A., Garc, C., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Laird, M. R., Lavidas, I., Liu, Z., Marug, C., Loveland, J. E., Maurel, T., Mcmahon, A. C., Moore, B., Morales, J., Mudge, J. M., Nuhn, M., Ogeh, D., Parker, A., Parton, A., Patricio, M., Abdul, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sparrow, H., Stapleton, E., Szuba, M., Taylor, K., Threadgold, G., Thormann, A., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Yates, A. D., Zerbino, D. R. and Flicek, P. (2019), 'Ensembl 2019', *Nucleic Acids Research* **47**, D745–D751.

DAGRIS (2007), 'Domestic Animal Genetic Resources Information System (DAGRIS)'. International Livestock Research Institute [online] accessed 12/06/2020.
URL: *http://dagris.ilri.cgiar.org/*

Daza-Vamenta, R., Glusman, G., Rowen, L., Guthrie, B. and Geraghty, D. E. (2004), 'Genetic divergence of the rhesus macaque major histocompatibility complex', *Genome Research* **14**(8), 1501–1515.

de Groot, N. G., Heijmans, C. M. C., de Ru, A. H., Janssen, G. M. C., Drijfhout, J. W., Otting, N., Vangenot, C., Doxiadis, G. G. M., Koning, F., van Veelen, P. A. and Bontrop, R. E. (2017), 'A Specialist Macaque MHC Class I Molecule with HLA-B*27–like Peptide-Binding Characteristics', *The Journal of Immunology* **199**(10), 3679–3690.

Deamer, D., Akeson, M. and Branton, D. (2016), 'Three decades of nanopore sequencing', *Nature Biotechnology* **34**(5), 518–524.

Deist, M. S., Gallardo, R. A., Bunn, D. A., Kelly, T. R., Dekkers, J. C., Zhou, H. and Lamonta, S. J. (2017), 'Novel mechanisms revealed in the trachea transcriptome of resistant and susceptible chicken lines following infection with newcastle disease virus', *Clinical and Vaccine Immunology* **24**(5), 1–17.

Deverson, E. V., Leong, L., Seelig, A., John, W., Tredgett, E. M., Butcher, G. W., Howard, C., Tredgett, E. M., Butcher, G. W. and Howard, J. C. (1998), 'Functional Analysis by Site-Directed Mutagenesis of the Complex Polymorphism in Rat Transporter Associated with Antigen Processing', *The Journal of Immunology* **160**, 2767–2779.

Dijkstra, J. M. (2014), 'TH 2 and Treg candidate genes in elephant shark', *Nature* **511**(7508), E7–E9.

Dijkstra, J. M., Grimholt, U., Leong, J., Koop, B. F. and Hashimoto, K. (2013), 'Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading DM system in a large part of vertebrates', *BMC Evolutionary Biology* **13**(1), 1–14.

Dijkstra, J. M., Yamaguchi, T. and Grimholt, U. (2018), 'Conservation of sequence motifs suggests that the non-classical MHC class I lineages CD1/PROCR and UT were established before the emergence of tetrapod species', *Immunogenetics* **70**(7), 459–476.

Dilthey, A. T., Moutsianas, L., Leslie, S. and McVean, G. (2011), 'HLA*IMP-an integrated framework for imputing classical HLA alleles from SNP genotypes', *Bioinformatics* **27**(7), 968–972.

Doherty, P. C. and Zinkernagel, R. M. (1975), 'A Biological Role for the Major Histocompatibility Antigens', *The Lancet* **305**(7922), 1406–1409.

Dohm, J. C., Tsend-Ayush, E., Reinhardt, R., Grützner, F. and Himmelbauer, H. (2007), 'Disruption and pseudoautosomal localization of the major histocompatibility complex in monotremes', *Genome Biology* **8**(8), R175.

Dong, G., Wearsch, P. A., Peaper, D. R., Cresswell, P. and Reinisch, K. M. (2009), 'Insights into MHC class I peptide loading from the structure of the tapasin/ERp57', *Immunity* **30**(1), 21–32.

Drews, A., Strandh, M., Råberg, L. and Westerdahl, H. (2017), 'Expression and phylogenetic analyses reveal paralogous lineages of putatively classical and non-classical MHC-I genes in three sparrow species (Passer)', *BMC Evolutionary Biology* **17**, 1–12.

Drews, A. and Westerdahl, H. (2019), 'Not all birds have a single dominantly expressed MHC-I gene: Transcription suggests that siskins have many highly expressed MHC-I genes', *Scientific Reports* **9**, 19506.

D'Souza, M. P., Adams, E., Altman, J. D., Birnbaum, M. E., Boggiano, C., Casorati, G., Chien, Y. H., Conley, A., Eckle, S. B. G., Früh, K., Gondré-Lewis, T., Hassan, N., Huang, H., Jayashankar, L., Kasmar, A. G., Kunwar, N., Lavelle, J., Lewinsohn, D. M., Moody, B., Picker, L., Ramachandra, L., Shastri, N., Parham, P., McMichael, A. J. and Yewdell, J. W. (2019), 'Casting a wider net: Immunosurveillance by nonclassical MHC molecules', *PLoS Pathogens* **15**(2), 1–15.

Du Pasquier, L., Miggiano, V. C., Kobel, H. R. and Fischberg, M. (1977), 'The genetic control of histocompatibility reactions in natural and laboratory-made polyploid individuals of the clawed toad Xenopus', *Immunogenetics* **5**(1), 129–141.

Dubin, A., Jørgensen, T. E., Moum, T., Johansen, S. D. and Jakt, L. M. (2019), 'Complete loss of the MHC II pathway in an anglerfish, Lophius piscatorius', *Biology Letters* **15**(20190594).

Dunnington, E. A., Larsen, C. T., Gross, W. B. and Siegel, P. B. (1992), 'Antibody responses to combinations of antigens in white Leghorn chickens of different background genomes and major histocompatibility complex genotypes.', *Poultry science* **71**(11), 1801–1806.

Edgar, R. C., Drive, R. M. and Valley, M. (2004), 'MUSCLE : multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research* **32**(5), 1792–1797.

Ekblom, R., Stapley, J., Ball, A. D., Birkhead, T., Burke, T. and Slate, J. (2011), 'Genetic mapping of the major histocompatibility complex in the zebra finch (Taeniopygia guttata)', *Immunogenetics* **63**(8), 523–530.

Eltanany, M. A. and Hemeda, S. A. (2016), 'Deeper insight into maternal genetic assessments and demographic history for Egyptian indigenous chicken populations using mtDNA analysis', *Journal of Advanced Research* **7**(5), 615–623.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E., Jensen, P. and Andersson, L. (2008), 'Identification of the Yellow skin gene reveals a hybrid origin of the domestic chicken', *PLoS Genetics* **4**(2), 1–8.

Ewald, S. J. and Livant, E. J. (2004), 'Distinctive Polymorphism of Chicken B-FI (Major Histocompatibility Complex Class I) Molecules', *Poultry Science* **83**(4), 600–605.

Fani, R. and Fondi, M. (2009), 'Origin and evolution of metabolic pathways', *Physics of Life Reviews* **6**(1), 23–52.

FAO (2018), 'Contract Farming in the Brazilian Chicken Industry: The Case of Pif Paf Alimentos'. FAO [online] accessed 28/02/2021.
URL: *http://www.fao.org/in-action/contract-farming/training/module-1/case-study-poultry-in-brazil/en/*

FAO (2020*a*), 'FAOSTAT statistical database'. [online] accessed 29/05/2020.
URL: *http://www.fao.org/faostat/en/#home*

FAO (2020*b*), 'Gateway to poultry production and products'. FAO [online] accessed 29/05/2020.
URL: *http://www.fao.org/poultry-production-products/production/production-systems/en/*

FAO, IFAD and WFP (2015), The State of Food Insecurity in the World 2015: Meeting the 2015 international hunger targets: taking stock of uneven progress., Technical report, FAO, Rome.
URL: *http://www.fao.org/3/a-i4646e.pdf*

Fisette, O., Wingbermühle, S., Tampé, R. and Schäfer, L. V. (2016), 'Molecular mechanism of peptide editing in the tapasin – MHC I complex', *Scientific Reports* **6**, 19085.

Fitzpatrick, S. M. and Callaghan, R. (2009), 'Examining dispersal mechanisms for the translocation of chicken (Gallus gallus) from Polynesia to South America', *Journal of Archaeological Science* **36**(2), 214–223.

Flajnik, M. F. (1996), 'The immune system of ectothermic vertebrates', *Veterinary Immunology and Immunopathology* **54**, 145–150.

Flajnik, M. F., Ohta, Y., Greenberg, A. S., Salter-Cid, L., Carrizosa, A., Du Pasquier, L. and Kasahara, M. (1999), 'Two ancient allelic lineages at the single classical class I locus in the Xenopus MHC.', *The Journal of Immunology* **163**, 3826–33.

Fleming-Canepa, X., Jensen, S. M., Mesa, C. M., Diaz-Satizabal, L., Roth, A. J., Parks-Dely, J. A., Moon, D. A., Wong, J. P., Evseev, D., Gossen, D. A., Tetrault, D. G. and Magor, K. E. (2016), 'Extensive Allelic Diversity of MHC Class I in Wild Mallard Ducks', *The Journal of Immunology* **197**(3), 783–794.

Fleming, D. S., Koltes, J. E., Markey, A. D., Schmidt, C. J., Ashwell, C. M., Rothschild, M. F., Persia, M. E., Reecy, J. M. and Lamont, S. J. (2016), 'Genomic analysis of Ugandan and Rwandan chicken ecotypes using a 600 k genotyping array', *BMC Genomics* **17**(1), 1–16.

Fournié, G. and Pfeiffer, D. U. (2014), 'Can closure of live poultry markets halt the spread of H7N9?', *The Lancet* **383**(9916), 496–497.

Friedrich-Loeffler-Institut (2020), Novel Coronavirus SARS-CoV-2: Fruit bats and ferrets are susceptible, pigs and chickens are not (Press Release), Technical report, Friedrich-Loeffler-Institut, Greifswald.
URL: *https://www.fli.de/en/press/press-releases/press-singleview/novel-coronavirus-sars-cov-2-fruit-bats-and-ferrets-are-susceptible-pigs-and-chickens-are-not/*

Fulton, J. E., Berres, M. E., Kantanen, J. and Honkatukia, M. (2017), 'MHC-B variability within the Finnish Landrace chicken conservation program', *Poultry Science* **96**(9), 3026–3030.

Fulton, J. E., Juul-Madsen, H. R., Ashwell, C. M., McCarron, A. M., Arthur, J. A., O'Sullivan, N. P. and Taylor, R. L. (2006), 'Molecular genotype identification of the Gallus gallus major histocompatibility complex', *Immunogenetics* **58**(5-6), 407–421.

Fulton, J. E., Lund, A. R., Mccarron, A. M., Pinegar, K. N., Korver, D. R., Classen, H. L., Aggrey, S., Utterbach, C., Anthony, N. B. and Berres, M. E. (2016b), 'MHC variability in heritage breeds of chickens', *Poultry Science* **95**(2), 393–399.

Fulton, J. E., McCarron, A. M., Lund, A. R., Pinegar, K. N., Wolc, A., Chazara, O., Bed'Hom, B., Berres, M. and Miller, M. M. (2016a), 'A high-density SNP panel reveals extensive diversity, frequent recombination and multiple recombination hotspots within the chicken major histocompatibility complex B region between BG2 and CD1A1', *Genetics Selection Evolution* **48**(1), 1–15.

Gao, B., Williams, A., Sewell, A. and Elliott, T. (2004), 'Generation of a functional, soluble tapasin protein from an alternatively spliced mRNA', *Genes and Immunity* **5**(2), 101–108.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. and Wiley, D. C. (1996), 'Structure of the complex between human T-cell receptor, viral peptide and HLA-A2', *Nature* **384**(6605), 134–141.

Gates, B. (2016), 'Why I would raise chickens'. Gates Notes [online] accessed 29/05/2020.
URL: *https://www.gatesnotes.com/Development/Why-I-Would-Raise-Chickens*

Gongora, J., Rawlence, N. J., Mobegi, V. A., Jianlin, H., Alcalde, J. A., Matus, J. T., Hanotte, O., Moran, C., Austin, J. J., Ulm, S., Anderson, A. J., Larson, G. and Cooper, A. (2008), 'Indo-European and Asian origins for Chilean and Pacific chickens revealed by mtDNA', *Proceedings of the National Academy of Sciences of the United States of America* **105**(30), 10308–10313.

Goraga, Z., Weigend, S. and Brockmann, G. (2011), 'Genetic diversity and population structure of five Ethiopian chicken ecotypes', *Animal Genetics* **43**(4), 454–457.

Gorer, P. A., Lyman, S. and Snell, G. D. (1948), 'Studies on the genetic and antigenic basis of tumour transplantation. Linkage between a histocompatibility gene and 'fused' in mice', *Proceedings of the Royal Society B: Biological Sciences* **135**(881), 499–505.

Goto, R. M., Wang, Y., Taylor, R. L., Wakenell, P. S., Hosomichi, K., Shiina, T., Blackmore, C. S., Briles, W. E. and Miller, M. M. (2009), 'BG1 has a major role in MHC-linked resistance to malignant lymphoma in the chicken', *Proceedings of the National Academy of Sciences of the United States of America* **106**(39), 16740–16745.

Gough, S. C. L. and Simmonds, M. J. (2007), 'The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action', *Current Genomics* **8**, 453–465.

Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E. S., Annese, V., Hauser, S. L., Oksenberg, J. R., Thomsen, I., Leslie, S., International Inflammatory Bowel Disease Genetics Consortium, Daly, M. J., Van Steen, K., Duerr, R. H., Barrett, J. C., McGovern, D. P., Schumm, L. P., Traherne, J. A., Carrington, M. N., Kosmoliaptsis, V., Karlsen, T. H., Franke, A. and Rioux, J. D. (2015),

'High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis', *Nature Genetics* **47**(2), 172–179.

Grandea, A. G., Androlewicz, M. J., Athwal, R. S., Geraghty, D. E. and Spies, T. (1995), 'Dependence of peptide binding by MHC class I molecules on their interaction with TAP', *Science* **270**(5233), 105–106.

Granevitze, Z., Hillel, J., Chen, G. H., Cuc, N. T., Feldman, M., Eding, H. and Weigend, S. (2007), 'Genetic diversity within chicken populations from different continents and management histories', *Animal Genetics* **38**(6), 576–583.

Greenbaum, J., Sidney, J., Chung, J., Brander, C., Peters, B. and Sette, A. (2011), 'Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes', *Immunogenetics* **63**(6), 325–335.

Greenwood, R., Shimizu, Y., Sekhon, G. S. and DeMars, R. (1994), 'Novel allele-specific, post-translational reduction in HLA class I surface expression in a mutant human B cell line.', *The Journal of Immunology* **153**(12), 5525–36.

Grimholt, U. (2018), 'Whole genome duplications have provided teleosts with many roads to peptide loaded MHC class I molecules', *BMC Evolutionary Biology* **18**(25).

Gros, M. and Amigorena, S. (2019), 'Regulation of antigen export to the cytosol during cross-presentation', *Frontiers in Immunology* **10**(41), 1–9.

Guo, Y., Lillie, M., Zan, Y., Beranger, J., Martin, A., Honaker, C. F., Siegel, P. B. and Carlborg (2019), 'A genomic inference of the White Plymouth Rock genealogy', *Poultry Science* **98**(11), 5272–5280.

Haase, D., Roth, O., Kalbe, M., Schmiedeskamp, G., Scharsack, J. P., Rosenstiel, P. and Reusch, T. B. (2013), 'Absence of major histocompatibility complex class II mediated immunity in pipefish, Syngnathus typhle: evidence from deep transcriptome sequencing.', *Biology Letters* **9**(20130044).

Habimana, R., Okeno, T. O., Ngeno, K., Mboumba, S., Assami, P., Gbotto, A. A., Keambou, C. T., Nishimwe, K., Mahoro, J. and Yao, N. (2020), 'Genetic diversity and population structure of indigenous chicken in Rwanda using microsatellite markers', *PLoS ONE* **15**(4), 1–17.

Hafstrand, I., Sayitoglu, E. C., Apavaloaei, A., Josey, B. J., Sun, R., Han, X., Pellegrino, S., Ozkazanc, D., Potens, R., Janssen, L., Nilvebrant, J., Nygren, P. Å., Sandalova, T., Springer, S., Georgoudaki, A. M., Duru, A. D. and Achour, A. (2019), 'Successive crystal structure snapshots suggest the basis for MHC class I peptide loading and editing by tapasin', *Proceedings of the National Academy of Sciences of the United States of America* **116**(11), 5055–5060.

Hassen, H., Neser, F. W., De Kock, A. and Van Marle-Köster, E. (2009), 'Study on the genetic diversity of native chickens in northwest Ethiopia using microsatellite markers', *African Journal of Biotechnology* **8**(7), 1347–1353.

Hayashida, H. and Miyata, T. (1983), 'Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex', *Proceedings of the National Academy of Sciences of the United States of America* **80**, 2671–2675.

He, K., Minias, P. and Dunn, P. O. (2020), 'Long-read genome assemblies reveal extraordinary variation in the number and structure of MHC loci in birds', *Genome Biology and Evolution* **13**(2), evaa270.

Hedrick, P. W. (2002), 'Pathogen resistance and genetic variation at MHC loci.', *Evolution* **56**(10), 1902–1908.

Hermann, C., Strittmatter, L. M., Deane, J. E. and Boyle, L. H. (2013), 'The Binding of TAPBPR and Tapasin to MHC Class I Is Mutually Exclusive', *The Journal of Immunology* **191**(11), 5743–5750.

Hermann, C., van Hateren, A., Trautwein, N., Neerincx, A., Duriez, P. J., Stevanović, S., Trowsdale, J., Deane, J. E., Elliott, T. and Boyle, L. H. (2015), 'TAPBPR alters MHC class I peptide presentation by functioning as a peptide exchange catalyst', *eLife* **4**, 1–22.

Herrera, M. B., Kraitsek, S., Alcalde, J. A., Quiroz, D., Revelo, H., Alvarez, L. A., Rosario, M. F., Thomson, V., Jianlin, H., Austin, J. J. and Gongora, J. (2020), 'European and Asian contribution to the genetic diversity of mainland South American chickens', *Royal Society Open Science* **7**(191558).

Hewitt, E. W., Gupta, S. S. and Lehner, P. J. (2001), 'The human cytomegalovirus gene product US6 inhibits ATP binding by TAP', *EMBO Journal* **20**(3), 387–396.

Hill, A. V. S. (1991), HLA Associations with Malaria in Africa: Some Implications for MHC Evolution, *in* J. Klein and D. Klein, eds, 'Molecular Evolution of the Major Histocompatibility Complex', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 403–420.

Hillel, J., Groenen, M. A. M., Tixier-Boichard, M., Korol, A. B., David, L., Kirzhner, V. M., Burke, T., Barre-Dirie, A., Crooijmans, R. P. M. A., Elo, K., Feldman, M. W., Freidlin, P. J., Mäki-Tanila, A., Oortwijn, M., Thomson, P., Vignal, A., Wimmers, K. and Weigend, S. (2003), 'Biodiversity of 52 chicken populations assessed by microsatellite typing of DNA pools.', *Genetics Selection Evolution* **35**(5), 533–557.

Hinz, A., Jedamzick, J., Herbring, V., Fischbach, H., Hartmann, J., Parcej, D., Koch, J. and Tampé, R. (2014), 'Assembly and function of the major histocompatibility complex (MHC) I peptide-loading complex are conserved across higher vertebrates.', *The Journal of Biological Chemistry* **289**(48), 33109–17.

Hislop, A. D., Ressing, M. E., Van Leeuwen, D., Pudney, V. A., Horst, D., Koppers-Lalic, D., Croft, N. P., Neefjes, J. J., Rickinson, A. B. and Wiertz, E. J. (2007), 'A CD8+ T cell immune evasion protein specific to Epstein-Barr virus and its close relatives in Old World primates', *Journal of Experimental Medicine* **204**(8), 1863–1873.

Hosomichi, K., Miller, M. M., Goto, R. M., Wang, Y., Suzuki, S., Kulski, J. K., Nishibori, M., Inoko, H., Hanzawa, K. and Shiina, T. (2008), 'Contribution of mutation, recombination, and gene conversion to chicken MHC-B haplotype diversity.', *The Journal of Immunology* **181**(5), 3393–3399.

Hulpke, S., Tomioka, M., Kremmer, E., Ueda, K., Abele, R. and Tampé, R. (2012), 'Direct evidence that the N-terminal extensions of the TAP complex act as autonomous interaction scaffolds for the assembly of the MHC I peptide-loading complex.', *Cellular and Molecular Life Sciences* **69**(19), 3317–27.

Hunt, H. D., Pharr, G. T. and Bacon, L. D. (1994), 'Molecular analysis reveals MHC class I intra-locus recombination in the chicken', *Immunogenetics* **40**(5), 370–375.

Ilca, F. T., Drexhage, L. Z., Brewin, G., Peacock, S. and Boyle, L. H. (2019), 'Distinct Polymorphisms in HLA Class I Molecules Govern Their Susceptibility to Peptide Editing by TAPBPR', *Cell Reports* **29**(6), 1621–1632.e3.

Ilca, F. T., Neerincx, A., Hermann, C., Marcu, A., Stevanovic, S., Deane, J. E. and Boyle, L. H. (2018), 'TAPBPR mediates peptide dissociation from MHC class I using a leucine lever', *eLife* **7**(e40126), 1–24.

Ilca, T. and Boyle, L. H. (2020), 'The Ins and Outs of TAPBPR', *Current Opinion in Immunology* **64**, 146–151.

Jaratlerdsiri, W., Deakin, J., Godinez, R. M., Shan, X., Peterson, D. G., Marthey, S., Lyons, E., McCarthy, F. M., Isberg, S. R., Higgins, D. P., Chong, A. Y., St John, J., Glenn, T. C., Ray, D. A. and Gongora, J. (2014), 'Comparative genome analyses reveal distinct structure in the saltwater crocodile MHC', *PLoS ONE* **9**(12), 1–33.

Jarvi, S. I., Goto, R. M., Briles, W. E. and Miller, M. M. (1996), 'Characterization of MHC genes in a multigenerational family of ring-necked pheasants', *Immunogenetics* **43**(3), 125–135.

Jensen, R. A. (1976), 'Enzyme Recruitment in Evolution of New Function', *Annual Review of Microbiology* **30**(1), 409–425.

Jethmalani, S. M., Henle, K. J. and Kaushal, G. P. (1994), 'Heat shock-induced prompt glycosylation. Identification of P-SG67 as calreticulin', *Journal of Biological Chemistry* **269**(38), 23603–23609.

Jiang, J., Natarajan, K., Boyd, L. F., Morozov, G. I., Mage, M. G. and Margulies, D. H. (2017), 'Crystal structure of a TAPBPR–MHC I complex reveals the mechanism of peptide editing in antigen presentation', *Science* **358**(6366), 1064–1068.

Joly, E., Le Rolle, A. F., Gonzàlez, A. L., Mehling, B., Stevens, J., Coadwell, W. J., Hünig, T., Howard, J. C. and Butcher, G. W. (1998), 'Co-evolution of rat TAP transporters and MHC class I RT1-A molecules', *Current Biology* **8**(3), 169–180.

Joly, E., Leong, L., Coadwell, W. J., Clarkson, C. and Butcher, G. W. (1996), 'The Rat MHC Haplotype RT1c Expresses Two Classical Class I Molecules', *Journal of Immunology* **157**(4), 1551–1558.

Jordan, W. C. and Bruford, M. W. (1998), 'New perspectives on mate choice and the MHC', *Heredity* **81**(3), 239–245.

Kamiya, T., O'Dwyer, K., Westerdahl, H., Senior, A. and Nakagawa, S. (2014), 'A quantitative review of MHC-based mating preference: The role of diversity and dissimilarity', *Molecular Ecology* **23**(21), 5151–5163.

Kanginakudru, S., Metta, M., Jakati, R. and Nagaraju, J. (2008), 'Genetic evidence from Indian red jungle fowl corroborates multiple domestication of modern day chicken', *BMC Evolutionary Biology* **8**(174).

Karlsson, M. and Westerdahl, H. (2013), 'Characteristics of MHC Class I Genes in House Sparrows Passer domesticus as Revealed by Long cDNA Transcripts and Amplicon Sequencing', *Journal of Molecular Evolution* **77**, 8–21.

Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., Mosley, J. D., Mallal, S., Denny, J. C., Phillips, E. J. and Roden, D. M. (2017), 'Comparison of HLA allelic imputation programs', *PLoS ONE* **12**(2), 1–12.

Kaslow, R. A., Carrington, M., Apple, R., Park, L., Munoz, A., Saah, A. J., Goedert, J. J., Winkler, C., O'Brien, S. J., Rinaldo, C., Detels, R., Blattner, W., Phair, J., Erlich, H. and Mann, D. L. (1996), 'Influence of combinations of human major histocompatibility complex genes on the course of HIV–1 infection', *Nature Medicine* **2**(4), 405–211.

Kaufman, J. (2000), 'The simple chicken major histocompatibility complex: Life and death in the face of pathogens and vaccines', *Philosophical Transactions of the Royal Society B: Biological Sciences* **355**(1400), 1077–1084.

Kaufman, J. (2011), The evolutionary origins of the adaptive immune system of jawed vertebrates, *in* S. H. E. Kaufmann, B. T. Rouse and D. L. Sachs, eds, 'The Immune Response to Infection', American Society of Microbiology Press, Washington DC, chapter 3, pp. 41–54.

Kaufman, J. (2014), The Avian MHC, *in* F. Davison, B. Kaspers and K. A. Schat, eds, 'Avian Immunology', 2 edn, Elsevier, Boston, chapter 8, pp. 159–182.

Kaufman, J. (2015), 'What chickens would tell you about the evolution of antigen processing and presentation', *Current Opinion in Immunology* **34**, 35–42.

Kaufman, J. (2017), A New View of How MHC Class I Molecules Fight Disease: Generalists and Specialists, *in* P. Pontarotti, ed., 'Evolutionary Biology: Self/Nonself Evolution, Species and Complex Traits Evolution, Methods and Concepts', Springer, Cham, pp. 3–25.

Kaufman, J. (2018), 'Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens', *Trends in Immunology* **39**(5), 367–379.

Kaufman, J. (2021), The Avian MHC, *in* B. Kaspers, K. A. Schat, T. Gobel and L. Vervelde, eds, 'Avian Immunology', 3 edn, Elsevier, Boston.

Kaufman, J., Milne, S., Göbel, T. W., Walker, B. A., Jacob, J. P., Auffray, C., Zoorob, R. and Beck, S. (1999), 'The chicken B locus is a minimal essential major histocompatibility complex', *Nature* .

Kaufman, J., Salomonsen, J. and Flajnik, M. (1994), 'Evolutionary conservation of MHC class I and class II molecules- different yet the same', *Seminars in Immunology* **6**(6), 411–424.

Kaufman, J., Volk, H. and Wallny, H.-J. (1995), 'A "Minimal Essential Mhc" and an "Unrecognized Mhc": Two Extremes in Selection for Polymorphism', *Immunological Reviews* **143**(1), 63–88.

Kekäläinen, J., Vallunen, J. A., Primmer, C. R., Rättyä, J. and Taskinen, J. (2009), 'Signals of major histocompatibility complex overdominance in a wild salmonid population', *Proceedings of the Royal Society B: Biological Sciences* **276**(1670), 3133–3140.

Khanyile, K. S., Dzomba, E. F. and Muchadeyi, F. C. (2015), 'Population genetic structure, linkage disequilibrium and effective population size of conserved and extensively raised village chicken populations of Southern Africa', *Frontiers in Genetics* **6**(13), 1–11.

Kim, T., Hunt, H. D., Parcells, M. S., Santen, V. V. and Ewald, S. J. (2018), 'Two class I genes of the chicken MHC have different functions : BF1 is recognized by NK cells while BF2 is recognized by CTLs', *Immunogenetics* **70**, 599–611.

Klein, J., Satta, Y. and O'hUigin, C. (1993), 'The Molecular Descent of the Major Histocompatibility Complex', *Annual Reviews of Immunology* **11**, 269–295.

Kobel, H. R. and Du Pasquier, L. (1986), 'Genetics of polyploid Xenopus', *Trends in Genetics* **2**, 310–315.

Koch, J., Guntrum, R., Heintke, S., Kyritsis, C. and Tampé, R. (2004), 'Functional Dissection of the Transmembrane Domains of the Transporter Associated with Antigen Processing (TAP)', *Journal of Biological Chemistry* **279**(11), 10142–10147.

Koch, J., Guntrum, R. and Tampé, R. (2006), 'The first N-terminal transmembrane helix of each subunit of the antigenic peptide transporter TAP is essential for independent tapasin binding', *FEBS Letters* **580**(17), 4091–4096.

Koch, M., Camp, S., Collen, T., Avila, D., Salomonsen, J., Wallny, H. J., van Hateren, A., Hunt, L., Jacob, J. P., Johnston, F., Marston, D. A., Shaw, I., Dunbar, P. R., Cerundolo, V., Jones, E. Y. and Kaufman, J. (2007), 'Structures of an MHC Class I Molecule from B21 Chickens Illustrate Promiscuous Peptide Binding', *Immunity* **27**(6), 885–899.

Koenen, M. E., Boonstra-Blom, A. G. and Jeurissen, S. H. (2002), 'Immunological differences between layer- and broiler-type chickens', *Veterinary Immunology and Immunopathology* **89**, 47–56.

Košmrlj, A., Read, E., Qi, Y. and Allen, T. (2010), 'Effects of thymic selection of the T-cell repertoire on HLA class I-associated control of HIV infection', *Nature* **465**(7296), 350–354.

Kranis, A., Gheyas, A. A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P. M., Fife, M., Salmon, N., Fulton, J., Strom, T. M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., Qanbari, S., Simianer, H., Watson, K. A., Woolliams, J. A. and Burt, D. W. (2013), 'Development of a high density 600K SNP genotyping array for chicken', *BMC Genomics* **14**(59).

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018), 'MEGA X : Molecular Evolutionary Genetics Analysis across Computing Platforms', *Molecular Biology and Evolution* **35**(6), 1547–1549.

Lakshmanan, N., Kaiser, M. G. and Lamont, S. J. (1996), Marek's disease resistance in MHC-congenic lines from leghorn and Fayoumi breeds, *in* R. F. Silva, H. H. Cheng, P. M. Coussens, L. F. Lee and L. F. Velicer, eds, 'Current research on Marek's disease', American Association of Avian Pathologists, Kennett Square, PA, pp. 57–62.

Landis, E. D., Palti, Y., Dekoning, J., Drew, R., Phillips, R. B. and Hansen, J. D. (2006), 'Identification and regulatory analysis of rainbow trout tapasin and tapasin-related genes', *Immunogenetics* **58**(1), 56–69.

Landry, C., Garant, D., Duchesne, P. and Bernatchez, L. (2001), ''Good genes as heterozygosity': The major histocompatibility complex and mate choice in Atlantic salmon (Salmo salar)', *Proceedings of the Royal Society B: Biological Sciences* **268**(1473), 1279–1285.

Langat, D. K. and Hunt, J. S. (2002), 'Do Nonhuman Primates Comprise Appropriate Experimental Models for Studying the Function of Human Leukocyte Antigen-G?', *Biology of Reproduction* **67**(5), 1367–1374.

Lawal, R. A., Martin, S. H., Vanmechelen, K., Vereijken, A., Silva, P., Al-Atiyat, R. M., Aljumaah, R. S., Mwacharo, J. M., Wu, D. D., Zhang, Y. P., Hocking, P. M., Smith, J., Wragg, D. and Hanotte, O. (2020), 'The wild species genome ancestry of domestic chickens', *BMC Biology* **18**(1), 1–18.

Lawlor, D. A., Zemmour, J., Ennis, P. D. and Parham, P. (1990), 'Evolution of class-I MHC genes and proteins: From natural selection to thymic selection', *Annual Review of Immunology* **8**(3), 23–63.

Leclaire, S., Strandh, M., Mardon, J., Westerdahl, H. and Bonadonna, F. (2017), 'Odour-based discrimination of similarity at the major histocompatibility complex in birds', *Proceedings of the Royal Society B: Biological Sciences* **284**(1846).

Leger, A. and Leonardi, T. (2019), 'pycoQC , interactive quality control for Oxford Nanopore Sequencing', *Journal of Open Source Software* **4**(34), 2–5.

Leggett, R. M. and Clark, M. D. (2017), 'A world of opportunities with nanopore sequencing', *Journal of Experimental Botany* **68**(20), 5419–5429.

Leinders-Zufall, T., Brennan, P., Widmayer, P., Chandramani S., P., Maul-Pavicic, A., Jäger, M., Li, X. H., Breer, H., Zufall, F. and Boehm, T. (2004), 'MHC class I peptides as chemosensory signals in the vomeronasal organ', *Science* **306**(5698), 1033–1037.

Leinders-Zufall, T., Ishii, T., Mombaerts, P., Zufall, F. and Boehm, T. (2009), 'Structural requirements for the activation of vomeronasal sensory neurons by MHC peptides', *Nature Neuroscience* **12**(12), 1551–1558.

Leroy, G., Kayang, B. B., Youssao, I. A., Yapi-Gnaoré, C. V., Osei-Amponsah, R., Loukou, N. E., Fotsa, J. C., Benabdeljelil, K., Bed'hom, B., Tixier-Boichard, M. and Rognon, X. (2012), 'Gene diversity, agroecological structure and introgression patterns among village chicken populations across North, West and Central Africa', *BMC Genetics* **13**(34).

Lewis, J. W., Sewell, A., Price, D. and Elliott, T. (1998), 'HLA-A*0201 presents TAP-dependent peptide epitopes to cytotoxic T lymphocytes in the absence of tapasin', *European Journal of Immunology* **28**(10), 3214–3220.

Lima-Rosa, C. A. d. V., Canal, C. W., Vargas Fallavena, P. R., de Freitas, L. B. and Salzano, F. M. (2005), 'LEI0258 microsatellite variability and its relationship to B-F haplotypes in Brazilian (blue-egg Caipira) chickens', *Genetics and Molecular Biology* **28**(3), 386–389.

Little, C. C. and Tyzzer, E. E. (1916), 'Further experimental studies on the inheritance of susceptibility to a Transplantable tumor, Carcinoma (J. W. A.) of the Japanese waltzing Mouse.', *The Journal of Medical Research* **33**(3), 393–453.

Liu, Y. P., Wu, G. S., Yao, Y. G., Miao, Y. W., Luikart, G., Baig, M., Beja-Pereira, A., Ding, Z. L., Palanichamy, M. G. and Zhang, Y. P. (2006), 'Multiple maternal origins of chickens: Out of the Asian jungles', *Molecular Phylogenetics and Evolution* **38**(1), 12–19.

Loffredo, J. T., Sidney, J., Bean, A. T., Beal, D. R., Bardet, W., Wahl, A., Hawkins, O. E., Piaskowski, S., Wilson, N. A., Hildebrand, W. H., Watkins, D. I. and Sette, A. (2009), 'Two MHC Class I Molecules Associated with Elite Control of Immunodeficiency Virus Replication, Mamu-B*08 and HLA-B*2705, Bind Peptides with Sequence Similarity', *The Journal of Immunology* **182**(12), 7763–7775.

Luzuriaga-Neira, A., Villacís-Rivas, G., Cueva-Castillo, F., Escudero-Sánchez, G., Ulloa-Nuñez, A., Rubilar-Quezada, M., Monteiro, R., Miller, M. R. and Beja-Pereira, A. (2017), 'On the origins and genetic diversity of South American chickens: one step closer', *Animal Genetics* **48**(3), 353–357.

Lybarger, L., Wang, X., Harris, M. R., Virgin IV, H. W. and Hansen, T. H. (2003), 'Virus subversion of the MHC class I peptide-loading complex', *Immunity* **18**(1), 121–130.

Lyimo, C. M., Weigend, A., Janßen-Tapken, U., Msoffe, P. L., Simianer, H. and Weigend, S. (2013), 'Assessing the genetic diversity of five Tanzanian chicken ecotypes using molecular tools', *South African Journal of Animal Sciences* **43**(4), 499–510.

Lyimo, C. M., Weigend, A., Msoffe, P. L., Eding, H., Simianer, H. and Weigend, S. (2014), 'Global diversity and genetic contributions of chicken populations from African, Asian and European regions', *Animal Genetics* **45**(6), 836–848.

MacDonald, K. C. (1992), 'The domestic chicken (Gallus gallus) in sub-Saharan Africa: A background to its introduction and its osteological differentiation from indigenous fowls (Numidinae and Francolinus sp.)', *Journal of Archaeological Science* **19**(3), 303–318.

Maeda, R. (2008), 'Japan feeds animals recycled leftovers'. Reuters [online] accessed 29/05/2020.
URL: *https://www.reuters.com/article/us-japan-food-recycled-idUST21465920080723*

Magor, K. E., Navarro, D. M., Barber, M. R. W., Petkau, K., Fleming-Canepa, X., Blyth, G. A. D. and Blaine, A. H. (2013), 'Defense genes missing from the flight division', *Developmental and Comparative Immunology* **41**, 377–388.

Mahammi, F. Z., Gaouar, S. B., Laloë, D., Faugeras, R., Tabet-Aoul, N., Rognon, X., Tixier-Boichard, M. and Saidi-Mehtar, N. (2016), 'A molecular analysis of the patterns of genetic diversity in local chickens from western Algeria in comparison with commercial lines and wild jungle fowls', *Journal of Animal Breeding and Genetics* **133**(1), 59–70.

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Baalsrud, H. T., Nederbragt, A. J., Hanel, R., Salzburger, W., Stenseth, N. C., Jakobsen, K. S. and Jentoft, S. (2016), 'Evolution of the immune system influences speciation rates in teleost fishes', *Nature Genetics* **48**(10), 1204–1210.

Malomane, D. K., Simianer, H., Weigend, A., Reimer, C., Schmitt, A. O. and Weigend, S. (2019), 'The SYNBREED chicken diversity panel: A global resource to assess chicken diversity at high genomic resolution', *BMC Genomics* **20**(345), 1–15.

Malomane, D. K., Weigend, S., Schmitt, A. O., Weigend, A., Reimer, C. and Simianer, H. (2020), 'Genetic diversity in global chicken breeds as a function of genetic distance to the wild populations', *bioRxiv Genetics* . https://www.biorxiv.org/content/early/2020/01/29/2020.01.29.924696.

Marsh, S. G., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernández-Viña, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., MacH, B., Maiers, M., Mayr, W. R., Müller, C. R., Parham, P., Petersdorf, E. W., Sasazuki, T., Strominger, J. L., Svejgaard, A., Terasaki, P. I., Tiercy, J. M. and Trowsdale, J. (2010), 'Nomenclature for factors of the HLA system, 2010', *Tissue Antigens* **75**(4), 291–455.

Marusina, K., Iyer, M. and Monaco, J. J. (1997), 'Allelic variation in the mouse Tap-1 and Tap-2 transporter genes.', *The Journal of Immunology* **158**(11), 5251–5256.

Mayor, N. P., Robinson, J., McWhinnie, A. J., Ranade, S., Eng, K., Midwinter, W., Bultitude, W. P., Chin, C. S., Bowman, B., Marks, P., Braund, H., Madrigal, J. A., Latham, K. and Marsh, S. G. (2015), 'HLA typing for the next generation', *PLoS ONE* **10**(5), 1–12.

McConnell, S. C., Hernandez, K. M., Wcisel, D. J., Kettleborough, R. N., Stemple, D. L., Yoder, J. A., Andrade, J. and De Jong, J. L. (2016), 'Alternative haplotypes of antigen processing genes in zebrafish diverged early in vertebrate evolution', *Proceedings of the National Academy of Sciences of the United States of America* **113**(34), E5014–E5023.

McConnell, S. C., Restaino, A. C. and De Jong, J. L. (2014), 'Multiple divergent haplotypes express completely distinct sets of class I MHC genes in zebrafish', *Immunogenetics* **66**(3), 199–213.

McConnell, S. K., Dawson, D. A., Wardle, A. and Burke, T. (1999), 'The isolation and mapping of 19 tetranucleotide microsatellite markers in the chicken', *Animal Genetics* **30**(3), 183–189.

McCutcheon, J. A., Gumperz, J., Smith, K. D., Lutz, C. T. and Parham, P. (1995), 'Low HLA-C expression at cell surfaces correlates with increased turnover of heavy chain mRNA', *Journal of Experimental Medicine* **181**(6), 2085–2095.

McElroy, J. P., Dekkers, J. C., Fulton, J. E., O'Sullivan, N. P., Soller, M., Lipkin, E., Zhang, W., Koehler, K. J., Lamont, S. J. and Cheng, H. H. (2005), 'Microsatellite markers associated with resistance to Marek's disease in commercial layer chickens', *Poultry Science* **84**(11), 1678–1688.

McShan, A. C., Natarajan, K., Kumirov, V. K., Flores-Solis, D., Jiang, J., Badstübner, M., Toor, J. S., Bagshaw, C. R., Kovrigin, E. L., Margulies, D. H. and Sgourakis, N. G. (2018), 'Peptide exchange on MHC-I by TAPBPR is driven by a negative allostery release cycle', *Nature Chemical Biology* **14**(8), 811–820.

Medawar, P. B. (1944), 'The Behaviour and Fate of Skin Autografts and Skin Homografts in Rabbits: A Report to the War Wounds Committee of the Medical Research Council', *Journal of Anatomy* **78**(5), 176–199.

Medawar, P. B. (1945), 'A second study of the behaviour and fate of skin homografts in rabbits: A Report to the War Wounds Committee of the Medical Research Council', *Journal of Anatomy* **79**(4), 157–176.

Merkin, J., Russell, C., Chen, P. and Burge, C. B. (2012), 'Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues', *Science* **338**, 1593–1599.

Miao, Y. W., Peng, M. S., Wu, G. S., Ouyang, Y. N., Yang, Z. Y., Yu, N., Liang, J. P., Pianchou, G., Beja-Pereira, A., Mitra, B., Palanichamy, M. G., Baig, M., Chaudhuri, T. K., Shen, Y. Y., Kong, Q. P., Murphy, R. W., Yao, Y. G. and Zhang, Y. P. (2013), 'Chicken domestication: An updated perspective based on mitochondrial genomes', *Heredity* **110**(3), 277–282.

Migalska, M., Sebastian, A. and Radwan, J. (2019), 'Major histocompatibility complex class i diversity limits the repertoire of T cell receptors', *Proceedings of the National Academy of Sciences of the United States of America* **116**(11), 5021–5026.

Milinski, M., Griffiths, S., Wegner, K. M., Reusch, T. B., Haas-Assenbaum, A. and Boehm, T. (2005), 'Mate choice decisions of stickleback females predictably modified by MHC peptide ligands', *Proceedings of the National Academy of Sciences of the United States of America* **102**(12), 4414–4418.

Miller, M. M., Bacon, L. D., Hala, K., Hunt, H. D., Ewald, S. J., Kaufman, J., Zoorob, R. and Briles, W. E. (2004), '2004 Nomenclature for the chicken major histocompatibility (B and Y) complex', *Immunogenetics* **56**(4), 261–279.

Miller, M. M. and Taylor, R. L. (2016), 'Brief review of the chicken Major Histocompatibility Complex : the genes, their distribution on chromosome 16, and their contributions to disease resistance', *Poultry Science* **95**, 375–392.

Minias, P., Pikus, E., Whittingham, L. A. and Dunn, P. O. (2019), 'Evolution of copy number at the MHC varies across the avian tree of life', *Genome Biology and Evolution* **11**(1), 17–28.

Miska, K. B. and Miller, R. D. (1999), 'Marsupial Mhc class I: Classical sequences from the opossum, monodelphis domestica', *Immunogenetics* **50**, 89–93.

Moffett, A. and Loke, C. (2006), 'Immunology of placentation in eutherian mammals', *Nature Reviews Immunology* **6**(8), 584–594.

Momburg, F., Armandola, E. A., Post, M. and Hammerling, G. J. (1996), 'Residues in TAP2 peptide transporters controlling substrate specificity.', *The Journal of Immunology* **156**(5), 1756–1763.

Morozov, G. I., Zhao, H., Mage, M. G., Boyd, L. F., Jiang, J., Dolan, M. A., Venna, R., Norcross, M. A., McMurtrey, C. P., Hildebrand, W., Schuck, P., Natarajan, K. and Margulies, D. H. (2016), 'Interaction of TAPBPR, a tapasin homolog, with MHC-I molecules promotes peptide editing', *Proceedings of the National Academy of Sciences of the United States of America* **113**(8), E1006–E1015.

Mothé, B. R., Sidney, J., Dzuris, J. L., Liebl, M. E., Fuenger, S., Watkins, D. I. and Sette, A. (2002), 'Characterization of the Peptide-Binding Specificity of Mamu-B*17 and Identification of Mamu-B*17-Restricted Epitopes Derived from Simian Immunodeficiency Virus Proteins', *The Journal of Immunology* **169**(1), 210–219.

Mtileni, B. J., Muchadeyi, F. C., Weigend, S., Maiwashe, A., Groeneveld, E., Groeneveld, L. F., Chimonyo, M. and Dzama, K. (2010), 'A comparison of genetic diversity between South African conserved and field chicken populations using microsatellite markers', *South African Journal of Animal Sciences* **40**(5), 462–466.

Muchadeyi, F. C., Eding, H., Simianer, H., Wollny, C. B., Groeneveld, E. and Weigend, S. (2008), 'Mitochondrial DNA D-loop sequences suggest a Southeast Asian and Indian origin of Zimbabwean village chickens', *Animal Genetics* **39**(6), 615–622.

Muchadeyi, F. C., Eding, H., Wollny, C. B., Groeneveld, E., Makuza, S. M., Shamseldin, R., Simianer, H. and Weigend, S. (2007), 'Absence of population substructuring in Zimbabwe chicken ecotypes inferred using microsatellite analysis', *Animal Genetics* **38**(4), 332–339.

Muir, W. M., Wong, G. K. S., Zhang, Y., Wang, J., Groenen, M. A., Crooijmans, R. P., Megens, H. J., Zhang, H., Okimoto, R., Vereijken, A., Jungerius, A., Albers, G. A., Lawley, C. T., Delany, M. E., MacEachern, S. and Cheng, H. H. (2008), 'Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds', *Proceedings of the National Academy of Sciences of the United States of America* **105**(45), 17312–17317.

Müller, O. and Krawinkel, M. (2005), 'Malnutrition And Health In Developing Countries.', *Canadian Medical Association journal* **173**(3), 279–86.

Mwacharo, J. M., Bjørnstad, G., Mobegi, V., Nomura, K., Hanada, H., Amano, T., Jianlin, H. and Hanotte, O. (2011), 'Mitochondrial DNA reveals multiple introductions of domestic chicken in East Africa', *Molecular Phylogenetics and Evolution* **58**(2), 374–382.

Mwacharo, J. M., Nomura, K., Hanada, H., Han, J. L., Amano, T. and Hanotte, O. (2013), 'Reconstructing the origin and dispersal patterns of village chickens across East Africa: Insights from autosomal markers', *Molecular Ecology* **22**(10), 2683–2697.

Mwacharo, J. M., Nomura, K., Hanada, H., Jianlin, H., Hanotte, O. and Amano, T. (2007), 'Genetic relationships among Kenyan and other East African indigenous chickens', *Animal Genetics* **38**(5), 485–490.

Mwambene, P. L., Kyallo, M., Machuka, E., Githae, D. and Pelle, R. (2019), 'Genetic diversity of 10 indigenous chicken ecotypes from Southern Highlands of Tanzania based on Major Histocompatibility Complex-linked microsatellite LEI0258 marker typing', *Poultry Science* **98**(7), 2734–2746.

Mwangi, W. N., Smith, L. P., Baigent, S. J., Beal, R. K., Nair, V. and Smith, A. L. (2011), 'Clonal Structure of Rapid-Onset MDV-Driven CD4+ Lymphomas and Responding CD8+ T Cells', *PLoS Pathogens* **7**(5), e1001337.

Neefjes, J., Jongsma, M. L. M., Paul, P. and Bakke, O. (2011), 'Towards a systems understanding of MHC class I and MHC class II antigen presentation', *Nature Reviews Immunology* **11**, 823–836.

Neerincx, A., Hermann, C., Antrobus, R., van Hateren, A., Cao, H., Trautwein, N., Stevanović, S., Elliott, T., Deane, J. E. and Boyle, L. H. (2017), 'TAPBPR bridges UDP-glucose: Glycoprotein glucosyltransferase 1 onto MHC class I to provide quality control in the antigen presentation pathway', *eLife* **6**, 1–25.

Nei, M. and Hughes, A. L. (1991), Polymorphism and evolution of the major histocompatibility complex loci in mammals, *in* R. Selander, A. Clark and T. Whittam, eds, 'Evolution at the Molecular Level', Sunderland, MA, chapter 11, pp. 222–247.

Nguyen-Phuc, H., Fulton, J. E. and Berres, M. E. (2016), 'Genetic variation of major histocompatibility complex (MHC) in wild Red Junglefowl (Gallus gallus)', *Poultry Science* **95**(2), 400–411.

Nijenhuis, M. and Hammerling, G. J. (1996), 'Multiple regions of the transporter associated with antigen processing (TAP) contribute to its peptide binding site.', *The Journal of Immunology* **157**, 5467–5477.

Norup, L. R., Dalgaard, T. S., Pedersen, A. R. and Juul-Madsen, H. R. (2011), 'Assessment of Newcastle disease-specific T cell proliferation in different inbred MHC chicken lines', *Scandinavian Journal of Immunology* **74**(1), 23–30.

Nowak, M. A., Tarczy-Hornoch, K. and Austyn, J. M. (1992), 'The optimal number of major histocompatibility complex molecules in an individual.', *Proceedings of the National Academy of Sciences of the United States of America* **89**, 10896–9.

O'Brien, S. J., Roelke, M. E., Marker, L., Newman, A., Winkler, C. A., Meltzer, D., Colly, L., Evermann, J. F., Bush, M. and Wildt, D. E. (1985), 'Genetic Basis for Species Vulnerability in the Cheetah', *Science* **227**, 1428–1434.

Obst, R., Armandola, E. A., Nijenhuis, M., Momburg, F. and Hämmerling, G. J. (1995), 'TAP polymorphism does not influence transport of peptide variants in mice and humans', *European Journal of Immunology* **25**(8), 2170–2176.

O'Connor, E. A., Strandh, M., Hasselquist, D., Nilsson, J. and Westerdahl, H. (2016), 'The evolution of highly variable immunity genes across a passerine bird radiation', *Molecular Ecology* **25**(4), 977–989.

Ohta, Y., Goetz, W., Hossain, M. Z., Nonaka, M. and Flajnik, M. F. (2006), 'Ancestral Organization of the MHC Revealed in the Amphibian Xenopus', *The Journal of Immunology* **176**(6), 3674–3685.

Ohta, Y., McKinney, E. C., Criscitiello, M. F. and Flajnik, M. F. (2002), 'Proteasome, Transporter Associated with Antigen Processing, and Class I Genes in the Nurse Shark Ginglymostoma cirratum : Evidence for a Stable Class I Region and MHC Haplotype Lineages', *The Journal of Immunology* **168**(2), 771–781.

Ohta, Y., Powis, S. J., Lohr, R. L., Nonaka, M., Du Pasquier, L. and Flajnik, M. F. (2003), 'Two highly divergent ancient allelic lineages of the transporter associated with antigen processing (TAP) gene in Xenopus: Further evidence for co-evolution among MHC class I region genes', *European Journal of Immunology* **33**(11), 3017–3027.

Oka, T., Ito, N., Sekiya, M., Kinoshita, K., Kawakami, S. I., Bungo, T. and Tsudzuki, M. (2015), 'Genetic differentiation among populations of the Kurokashiwa breed of indigenous Japanese chickens assessed by microsatellite DNA polymorphisms', *Journal of Poultry Science* **52**(2), 88–93.

Okano, M., Miyamae, J., Suzuki, S., Nishiya, K., Katakura, F., Kulski, J. K., Moritomo, T. and Shiina, T. (2020), 'Identification of Novel Alleles and Structural Haplotypes of Major Histocompatibility Complex Class I and DRB Genes in Domestic Cat (Felis catus) by a Newly Developed NGS-Based Genotyping Method', *Frontiers in Genetics* **11**(750).

Oliver, M. K., Telfer, S. and Piertney, S. B. (2009), 'Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (Arvicola terrestris)', *Proceedings of the Royal Society B: Biological Sciences* **276**(1659), 1119–1128.

Olsson, M., Madsen, T., Nordby, J., Wapstra, E., Ujvari, B. and Wittsell, H. (2003), 'Major histocompatibility complex and mate choice in sand lizards', *Proceedings of the Royal Society B: Biological Sciences* **270**(SUPPL. 2), 254–256.

O'Neill, A. M., Livant, E. J. and Ewald, S. J. (2009), 'The chicken BF1 (classical MHC class I) gene shows evidence of selection for diversity in expression and in promoter and signal peptide regions', *Immunogenetics* **61**, 289–302.

Ortmann, B., Androlewicz, M. J. and Cresswell, P. (1994), 'MHC class I/β2-microglobulin complexes associate with TAP transporters before peptide binding', *Nature* **368**(6474), 864–867.

Ortmann, B., Copeman, J., Lehner, P. J., Sadasivan, B., Herberg, J. A., Grandea, A. G., Riddell, S. R., Tampé, R., Spies, T., Trowsdale, J. and Cresswell, P. (1997), 'A critical role for tapasin in the assembly and function of multimeric MHC class I-TAP complexes', *Science* **277**(5330), 1306–1309.

Osei-Amponsah, R., Kayang, B. B., Naazie, A., Osei, Y. D., Youssao, I. A., Yapi-Gnaore, V. C., Tixier-Boichard, M. and Rognon, X. (2010), 'Genetic diversity of forest and savannah chicken populations of Ghana as estimated by microsatellite markers', *Animal Science Journal* **81**(3), 297–303.

Osman, S. A., Sekino, M., Nishibori, M., Yamamoto, Y. and Tsudzuki, M. (2005), 'Genetic variability and relationships of native Japanese chickens assessed by microsatellite DNA profiling - Focusing on the breeds established in Kochi Prefecture, Japan', *Asian-Australasian Journal of Animal Sciences* **18**(6), 755–761.

Osman, S. A., Yonezawa, T. and Nishibori, M. (2016), 'Origin and genetic diversity of Egyptian native chickens based on complete sequence of mitochondrial DNA D-loop region', *Poultry Science* **95**(6), 1248–1256.

Otting, N. and Bontrop, R. E. (1993), 'Characterization of the rhesus macaque (Macaca mulatta) equivalent of HLA-F', *Immunogenetics* **38**, 141–145.

Otting, N., De Vos-Rouweler, A. J., Heijmans, C. M., De Groot, N. G., Doxiadis, G. G. and Bontrop, R. E. (2007), 'MHC class I A region diversity and polymorphism in macaque species', *Immunogenetics* **59**(5), 367–375.

Otting, N., Heijmans, C. M., Noort, R. C., De Groot, N. G., Doxiadis, G. G., Van Rood, J. J., Watkins, D. I. and Bontrop, R. E. (2005), 'Unparalleled complexity of the MHC class I region in rhesus macaques', *Proceedings of the National Academy of Sciences of the United States of America* **102**(5), 1626–1631.

Owen, J. P., Delany, M. E. and Mullens, B. A. (2008), 'MHC haplotype involvement in avian resistance to an ectoparasite', *Immunogenetics* **60**, 621–631.

Pappas, D. J., Lizee, A., Paunic, V., Beutner, K. R., Motyer, A., Vukcevic, D., Leslie, S., Biesiada, J., Meller, J., Taylor, K. D., Zheng, X., Zhao, L. P., Gourraud, P. A., Hollenbach, J. A., Mack, S. J. and Maiers, M. (2018), 'Significant variation between SNP-based HLA imputations in diverse populations: The last mile is the hardest', *Pharmacogenomics Journal* **18**(3), 367–376.

Parham, P. and Moffett, A. (2013), 'How did variable NK-cell receptors and MHC class I ligands influence immunity, reproduction and human evolution?', *Nature Reviews Immunology* **13**(2), 133–144.

Park, B., Lee, S. and Kim, E. (2003), 'A Single Polymorphic Residue Within the Peptide-Binding Cleft of MHC Class I Molecules Determines Spectrum of Tapasin Dependence', *The Journal of Immunology* **170**, 961–968.

Parker, A. (2012), Classical and Non-classical Major Histocompatibility Complex Class II genes in the chicken, Phd thesis, University of Cambridge.

Parker, A. and Kaufman, J. (2017), 'What chickens might tell us about the MHC class II system', *Current Opinion in Immunology* **46**, 23–29.

Patricio, I. S., Mendes, A. A., Ramos, A. A. and Pereira, D. F. (2012), 'Overview on the performance of Brazilian broilers (1990 to 2009)', *Revista Brasileira de Ciencia Avicola* **14**(4), 233–238.

Paul, S., Weiskopf, D., Angelo, M. A., Sidney, J., Peters, B. and Sette, A. (2013), 'HLA Class I Alleles Are Associated with Peptide-Binding Repertoires of Different Size, Affinity, and Immunogenicity', *The Journal of Immunology* **191**(12), 5831–5839.

Peakall, R. and Smouse, P. E. (2006), 'GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research', *Molecular Ecology Notes* **6**, 288–295.

Peakall, R. and Smouse, P. E. (2012), 'GenALEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research-an update', *Bioinformatics* **28**(19), 2537–2539.

Peaper, D. R., Wearsch, P. A. and Cresswell, P. (2005), 'Tapasin and ERp57 form a stable disulfide-linked dimer within the MHC class I peptide-loading complex', *EMBO Journal* **24**(20), 3613–3623.

Peh, C. A., Burrows, S. R., Barnden, M., Khanna, R., Cresswell, P., Moss, D. J., Mccluskey, J. and Park, B. (1998), 'HLA-B27 – Restricted Antigen Presentation in the Absence of Tapasin Reveals Polymorphism in Mechanisms of HLA Class I Peptide Loading', *Immunity* **8**(5), 531–542.

Penn, D. J. (2002), 'The Scent of Genetic Compatibility: Sexual Selection and the Major Histocompatibility Complex', *Ethology* **108**(1), 1–21.

Petersen, J. L., Hickman-Miller, H. D., McIlhaney, M. M., Vargas, S. E., Purcell, A. W., Hildebrand, W. H. and Solheim, J. C. (2005), 'A Charged Amino Acid Residue in the Transmembrane/Cytoplasmic Region of Tapasin Influences MHC Class I Assembly and Maturation', *The Journal of Immunology* **174**(2), 962–969.

Pinard-van der Laan, M. H., Bed'hom, B., Coville, J. L., Pitel, F., Feve, K., Leroux, S., Legros, H., Thomas, A., Gourichon, D., Repérant, J. M. and Rault, P. (2009), 'Microsatellite mapping of QTLs affecting resistance to coccidiosis (Eimeria tenella) in a Fayoumi × White Leghorn cross', *BMC Genomics* **10**, 1–13.

Porter, K. M., Hermann, C., Traherne, J. A. and Boyle, L. H. (2014), 'TAPBPR isoforms exhibit altered association with MHC class I', *Immunology* **142**(2), 289–299.

Potts, N. (2016), Haplotype diversity and stability in the chicken major histocompatibility complex, PhD thesis, University of Cambridge.

Potts, N. D., Bichet, C., Merat, L., Guitton, E., Krupa, A. P., Burke, T. A., Kennedy, L. J., Sorci, G. and Kaufman, J. (2019), 'Development and optimization of a hybridization technique to type the classical class I and class II B genes of the chicken MHC', *Immunogenetics* **71**(10), 647–663.

Potts, W. K., Manning, C. J. and Wakeland, E. K. (1994), 'The role of infectious disease, inbreeding and mating preferences in maintaining MHC genetic diversity: an experimental test.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **346**, 369–378.

Potts, W. K. and Wakeland, E. K. (1990), 'Evolution of diversity at the major histocompatibility complex', *Trends in Ecology and Evolution* **5**(6), 181–187.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M. and Lemmon, A. R. (2015), 'A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing', *Nature* **526**, 569–573.

Psifidi, A., Banos, G., Matika, O., Desta, T. T., Bettridge, J., Hume, D. A., Dessie, T., Christley, R., Wigley, P., Hanotte, O. and Kaiser, P. (2016), 'Genome-wide association studies of immune, disease and production traits in indigenous chicken ecotypes', *Genetics Selection Evolution* **48**(74).

Purcell, A. W., Gorman, J. J., Garcia-Peydró, M., Paradela, A., Burrows, S. R., Talbo, G. H., Laham, N., Peh, C. A., Reynolds, E. C., López de Castro, J. A. and McCluskey, J. (2001), 'Quantitative and Qualitative Influences of Tapasin on the Class I Peptide Repertoire', *The Journal of Immunology* **166**(2), 1016–1027.

Purcell, A. W., Ramarathinam, S. H. and Ternette, N. (2019), 'Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics', *Nature Protocols* **14**(6), 1687–1707.

R core team (2013), 'R: A language and environment for statistical computing', *R Foundation for Statistical Computing, Vienna, Austria* .

Radwan, J., Babik, W., Kaufman, J., Lenz, T. L. and Winternitz, J. (2020), 'Advances in the Evolutionary Understanding of MHC Polymorphism', *Trends in Genetics* **36**(4), 298–311.

Rammensee, H. G., Friede, T. and Stevanović, S. (1995), 'MHC ligands and peptide motifs: first listing', *Immunogenetics* **41**(4), 178–228.

Renard, C., Hart, E., Sehra, H., Beasley, H., Coggill, P., Howe, K., Harrow, J., Gilbert, J., Sims, S., Rogers, J., Ando, A., Shigenari, A., Shiina, T., Inoko, H., Chardon, P. and Beck, S. (2006), 'The genomic sequence and analysis of the swine major histocompatibility complex', *Genomics* **88**(1), 96–110.

Rizvi, S. M., Salam, N., Geng, J., Qi, Y., Bream, J. H., Duggal, P., Hussain, S. K., Martinson, J., Wolinsky, S., Carrington, M. and Raghavan, M. (2014), 'Distinct assembly profiles of HLA-B molecules', *The Journal of Immunology* **192**(11), 4967–4976.

Roberts, V. (2009), 'Control of Marek's Disease and other tumours'. National Animal Disease Information Service [online] accessed 28/02/2021.
URL: *https://www.nadis.org.uk/disease-a-z/poultry/diseases-of-farmyard-poultry/part-2-control-of-mareks-disease-and-other-tumours/*

Robinson, J., Guethlein, L. A., Cereb, N., Yang, S. Y., Norman, P. J., Marsh, S. G. and Parham, P. (2017), 'Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles', *PLoS Genetics* **13**(6), 1–28.

Rodgers, J. R. and Cook, R. G. (2005), 'MHC class IB molecules bridge innate and acquired immunity', *Nature Reviews Immunology* **5**(6), 459–471.

Rowland, K., Wolc, A., Gallardo, R. A., Kelly, T., Zhou, H., Dekkers, J. C. and Lamont, S. J. (2018), 'Genetic analysis of a commercial egg laying line challenged with Newcastle disease virus', *Frontiers in Genetics* **9**(326).

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E. and Sanchez-Gracia, A. (2017), 'DnaSP 6: DNA sequence polymorphism analysis of large data sets', *Molecular Biology and Evolution* **34**(12), 3299–3302.

Ruan, R., Ruan, J., Wan, X. L., Zheng, Y., Chen, M. M., Zheng, J. S. and Wang, D. (2016), 'Organization and characteristics of the major histocompatibility complex class II region in the Yangtze finless porpoise (Neophocaena asiaeorientalis asiaeorientalis)', *Scientific Reports* **6**, 22471.

Rude, E. (2015), 'The Forgotten History of 'Hen Fever''. National Geographic [online] accessed 30/07/2020.
URL: *https://www.nationalgeographic.com/culture/food/the-plate/2015/08/05/the-forgotten-history-of-hen-fever/*

Rüdel, N. (2004), Consequences of degradation and fragmentation of Malagasy littoral rain forests on gray mouse lemur populations (Microcebus murinus), Diploma thesis, University Hamburg.

Ruff, J. S., Nelson, A. C., Kubinak, J. L. and Potts, W. K. (2012), MHC signaling during social communication, *in* C. López-Larrea, ed., 'Self and Nonself', Springer, New York, pp. 290–313.

Ruijter, J. M., Ramakers, C., Hoogaars, W. M., Karlen, Y., Bakker, O., Van den hoff, M. J. and Moorman, A. F. (2009), 'Amplification efficiency: Linking baseline and bias in the analysis of quantitative PCR data', *Nucleic Acids Research* **37**(6).

Sadasivan, B., Lehner, P. J., Ortmann, B., Spies, T. and Cresswell, P. (1996), 'Roles for calreticulin and a novel glycoprotein, tapasin, in the interaction of MHC class I molecules with TAP', *Immunity* **5**(2), 103–114.

Sagert, L., Hennig, F., Thomas, C. and Tampé, R. (2020), 'A loop structure allows TAPBPR to exert its dual function as MHC I chaperone and peptide editor', *eLife* **9**(e55326).

Salomonsen, J., Chattaway, J. A., Chan, A. C., Parker, A., Huguet, S., Marston, D. A., Rogers, S. L., Wu, Z., Smith, A. L., Staines, K., Butter, C., Riegert, P., Vainio, O., Nielsen, L., Kaspers, B., Griffin, D. K., Yang, F., Zoorob, R., Guillemot, F., Auffray, C., Beck, S., Skjødt, K. and Kaufman, J. (2014), 'Sequence of a Complete Chicken BG Haplotype Shows Dynamic Expansion and Contraction of Two Gene Lineages with Particular Expression Patterns', *PLoS Genetics* **10**(6).

Sanger, F., Nicklen, S. and Coulson, R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences of the United States of America* **74**(12), 5463–5467.

Santos, P. S., Courtiol, A., Heidel, A. J., Höner, O. P., Heckmann, I., Nagy, M., Mayer, F., Platzer, M., Voigt, C. C. and Sommer, S. (2016), 'MHC-dependent mate choice is linked to a trace-amine-associated receptor gene in a mammal', *Scientific Reports* **6**, 38490.

Schad, J., Ganzhorn, J. U. and Sommer, S. (2005), 'Parasite burden and constitution of major histocompatibility complex in the Malagasy mouse lemur, Microcebus murinus', *Evolution* **59**(2), 439–450.

Schilling, M. A., Memari, S., Cavanaugh, M., Katani, R., Deist, M. S., Radzio-Basu, J., Lamont, S. J., Buza, J. J. and Kapur, V. (2019), 'Conserved, breed-dependent, and subline-dependent innate immune responses of Fayoumi and Leghorn chicken embryos to Newcastle disease virus infection', *Scientific Reports* **9**, 7209.

Schrodinger LLC (2019), 'The PyMol Molecular Graphics System, version 2.3'. Available at https://pymol.org/2/.

Semba, R. D. (2016), 'The rise and fall of protein malnutrition in global health', *Annals of Nutrition and Metabolism* **69**(2), 79–88.

Sepil, I., Moghadam, H. K., Huchard, E. and Sheldon, B. C. (2012), 'Characterization and 454 pyrosequencing of Major Histocompatibility Complex class I genes in the great tit reveal complexity in a passerine system', *BMC Evolutionary Biology* **12**(1), 68.

Shaw, I., Powell, T. J., Marston, D. A., Baker, K., Hateren, A. V., Riegert, P., Wiles, M. V., Milne, S., Beck, S. and Kaufman, J. (2007), 'Different Evolutionary Histories of the Two Classical Class I Genes BF1 and BF2 Illustrate Drift and Selection within the Stable MHC Haplotypes of Chickens', *The Journal of Immunology* **178**(9), 5744–5752.

Shiina, T., Blancher, A., Inoko, H. and Kulski, J. K. (2017), 'Comparative genomics of the human, macaque and mouse major histocompatibility complex', *Immunology* **150**(2), 127–138.

Shiina, T., Briles, W. E., Goto, R. M., Hosomichi, K., Yanagiya, K., Shimizu, S., Inoko, H. and Miller, M. M. (2007), 'Extended Gene Map Reveals Tripartite Motif, C-Type Lectin, and Ig Superfamily Type Genes within a Subregion of the Chicken MHC - B Affecting Infectious Disease', *The Journal of Immunology* **178**(11), 7162–7172.

Shiina, T., Oka, A., Imanishi, T., Hanzawa, K., Gojobori, T., Watanabe, S. and Inoko, H. (1999), 'Multiple class I loci expressed by the quail Mhc', *Immunogenetics* **49**(5), 456–460.

Shiina, T., Shimizu, S., Hosomichi, K., Kohara, S., Watanabe, S., Hanzawa, K., Beck, S., Kulski, J. K. and Inoko, H. (2004), 'Comparative Genomic Analysis of Two Avian (Quail and Chicken) MHC Regions', *The Journal of Immunology* **172**(11), 6751–6763.

Shum, B. P., Avila, D., Du Pasquier, L., Kasahara, M. and Flajnik, M. F. (1993), 'Isolation of a classical MHC class I cDNA from an amphibian. Evidence for only one class I locus in the Xenopus MHC.', *The Journal of Immunology* **151**(10), 5376–5386.

Siddle, H. V., Deakin, J. E., Coggill, P., Wilming, L. G., Harrow, J., Kaufman, J., Beck, S. and Belov, K. (2011), 'The tammar wallaby major histocompatibility complex shows evidence of past genomic instability', *BMC Genomics* **12**(1), 421.

Sidney, J., Peters, B., Frahm, N., Brander, C. and Sette, A. (2008), 'HLA class I supertypes: A revised and updated classification', *BMC Immunology* **9**, 1–15.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., Thompson, J. D., Higgins, D. G., Mcwilliam, H., Remmert, M. and So, J. (2011), 'Fast , scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology* **7**(539).

Simone, L. C., Georgesen, C. J., Simone, P. D., Wang, X. and Solheim, J. C. (2012), 'Productive association between MHC class I and tapasin requires the tapasin transmembrane/cytosolic region and the tapasin C-terminal Ig-like domain.', *Molecular Immunology* **49**(4), 628–639.

Simonsen, M. (1987), 'The MHC Of The Chicken, Genomic Structure, Gene-Products, And Resistance To Oncogenic DNA And RNA Viruses', *Veterinary Immunology And Immunopathology* **17**, 243–253.

Simonsen, M., Crone, M., Koch, C. and Hála, K. (1982), 'The MHC haplotypes of the chicken', *Immunogenetics* **16**(6), 513–532.

Slade, R. W. and McCallum, H. I. (1992), 'Overdominant vs. frequency-dependent selection at MHC loci', *Genetics* **132**(3), 861–862.

Small, C. M., Bassham, S., Catchen, J., Amores, A., Fuiten, A. M., Brown, R. S., Jones, A. G. and Cresko, W. A. (2016), 'The genome of the Gulf pipefish enables understanding of evolutionary innovations', *Genome Biology* **17**(1), 1–23.

Snary, D., Barnstable, C. J., Bodmer, W. F. and Crumpton, M. J. (1977), 'Molecular structure of human histocompatibility antigens: the HLA[U+2010]C series', *European Journal of Immunology* **7**(8), 580–585.

Spurgin, L. G. and Richardson, D. S. (2010), 'How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings', *Proceedings of the Royal Society B: Biological Sciences* **277**(1684), 979–988.

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzén, A., Winer, R., Knight, J., Vogel, J. H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previti, C., Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S. D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S. W., Stenseth, N. C. and Jakobsen, K. S. (2011), 'The genome sequence of Atlantic cod reveals a unique immune system', *Nature* **477**(7363), 207–210.

Stebbins, C. C., Loss, G. E., Elias, C. G., Chervonsky, A. and Sant, A. J. (1995), 'The requirement for DM in class II-restricted antigen presentation and sds-stable dimer formation is allele and species dependent', *Journal of Experimental Medicine* **181**(1), 223–234.

Stern, L. J., Brown, J. H., Jardetzky, T. S., Gorga, J. C., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1994), 'Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide', *Nature* **368**(6468), 215–221.

Stervander, M., Dierickx, E. G., Thorley, J., Brooke, M. d. L. and Westerdahl, H. (2020), 'High MHC gene copy number maintains diversity despite homozygosity in a Critically Endangered single-island endemic bird, but no evidence of MHC-based mate choice', *Molecular Ecology* **29**(19), 3578–3592.

Storey, A. A., Ramírez, J. M., Quiroz, D., Burley, D. V., Addison, D. J., Walter, R., Anderson, A. J., Hunt, T. L., Athens, J. S., Huynen, L. and Matisoo-Smith, E. A. (2007), 'Radiocarbon and DNA evidence for a pre-Columbian introduction of Polynesian chickens to Chile', *Proceedings of the National Academy of Sciences of the United States of America* **104**(25), 10335–10339.

Storey, A. A., Spriggs, M., Bedford, S., Hawkins, S. C., Robins, J. H., Huynen, L. and Matisoo-Smith, E. (2010), 'Mitochondrial DNA from 3000-year old chickens at the Teouma site, Vanuatu', *Journal of Archaeological Science* **37**(10), 2459–2468.

Strandh, M., Westerdahl, H., Pontarp, M., Canbäck, B., Dubois, M. P., Miquel, C., Taberlet, P. and Bonadonna, F. (2012), 'Major histocompatibility complex class II compatibility, but not class I, predicts mate choice in a bird with highly developed olfaction', *Proceedings of the Royal Society B: Biological Sciences* **279**(1746), 4457–4463.

Takahata, N. and Nei, M. (1990), 'Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci', *Genetics* **124**(4), 967–978.

Tallmadge, R. L., Lear, T. L. and Antczak, D. F. (2005), 'Genomic characterization of MHC class I genes of the horse', *Immunogenetics* **57**(10), 763–774.

Tarrant, K. J., Lopez, R., Loper, M. and Fulton, J. E. (2020), 'Assessing MHC-B diversity in Silkie chickens', *Poultry Science* **99**(5), 2337–2341.

Teng, M., Stephens, R., Pasquier, L., Freeman, T., Lindquist, J. and Trowsdale, J. (2002), 'A human TAPBP (TAPASIN)-related gene, TAPBP-R', *European Journal of Immunology* **32**, 1059–1068.

The MHC sequencing consortium (1999), 'Complete sequence and gene map of a human major histocompatibility complex', *Nature* **401**, 921–923.

The major histocompatibility complex and antigen presentation (2013), *in* J. Owen, J. Punt and S. Stranford, eds, 'Kuby Immunology', 7 edn, W. H. Freeman, New York, chapter 8, pp. 261–296.

Thomas, C. and Tampé, R. (2017), 'Structure of the TAPBPR–MHC I complex defines the mechanism of peptide loading and editing', *Science* **358**(6366), 1060–1064.

Tixier-Boichard, M., Bed'Hom, B. and Rognon, X. (2011), 'Chicken domestication: From archeology to genomics', *Comptes Rendus - Biologies* **334**(3), 197–204.

Tregaskes, C. A., Harrison, M., Sowa, A. K., van Hateren, A., Hunt, L. G., Vainio, O. and Kaufman, J. (2015), 'Surface expression, peptide repertoire, and thermostability of chicken class I molecules correlate with peptide transporter specificity', *Proceedings of the National Academy of Sciences of the United States of America* **113**(3), 692–697.

Tsuji, H., Taniguchi, Y., Ishizuka, S., Matsuda, H., Yamada, T., Naito, K. and Iwaisaki, H. (2017), 'Structure and polymorphisms of the major histocompatibility complex in the Oriental stork, Ciconia boyciana', *Scientific Reports* **7**, 42864.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I. M., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A. K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Von Heijne, G., Nielsen, J. and Pontén, F. (2015), 'Tissue-based map of the human proteome', *Science* **347**(6220).

van Hateren, A., Carter, R., Bailey, A., Kontouli, N., Williams, A. P., Kaufman, J. and Elliott, T. (2013), 'A Mechanistic Basis for the Co-evolution of Chicken Tapasin and Major Histocompatibility Complex Class I (MHC I)', *The Journal of Biological Chemistry* **288**(45), 32797–32808.

Van Rossum, G. and Drake, F. L. (2009), *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

Venkatesh, B., Lee, A. P., Ravi, V., Maurya, A. K., Lian, M. M., Swann, J. B., Ohta, Y., Flajnik, M. F., Sutoh, Y., Kasahara, M., Hoon, S., Gangu, V., Roy, S. W., Irimia, M., Korzh, V., Kondrychyn, I., Lim, Z. W., Tay, B. H., Tohari, S., Kong, K. W., Ho, S., Lorente-Galdos, B., Quilez, J., Marques-Bonet, T., Raney, B. J., Ingham, P. W.,

Tay, A., Hillier, L. W., Minx, P., Boehm, T., Wilson, R. K., Brenner, S. and Warren, W. C. (2014), 'Elephant shark genome provides unique insights into gnathostome evolution', *Nature* **505**(7482), 174–179.

Verweij, M. C., Horst, D., Griffin, B. D., Luteijn, R. D., Davison, A. J., Ressing, M. E. and Wiertz, E. J. (2015), 'Viral Inhibition of the Transporter Associated with Antigen Processing (TAP): A Striking Example of Functional Convergent Evolution', *PLoS Pathogens* **11**(4), 1–19.

Vidovic, D. and Matzinger, P. (1988), 'Unresponsiveness to a foreign antigen can be caused by self-tolerance', *Nature* **336**, 222–225.

Viļuma, A., Mikko, S., Hahn, D., Skow, L., Andersson, G. and Bergström, T. F. (2017), 'Genomic structure of the horse major histocompatibility complex class II region resolved using PacBio long-read sequencing technology', *Scientific Reports* **7**, 45518.

Walker, B. A., Hunt, L. G., Sowa, A. K., Skjødt, K., Göbel, T. W., Lehner, P. J. and Kaufman, J. (2011), 'The dominantly expressed class I molecule of the chicken MHC is explained by coevolution with the polymorphic peptide transporter (TAP) genes', *Proceedings of the National Academy of Sciences of the United States of America* **108**(20), 8396–8401.

Walker, B. A., Van Hateren, A., Milne, S., Beck, S. and Kaufman, J. (2005), 'Chicken TAP genes differ from their human orthologues in locus organisation, size, sequence features and polymorphism', *Immunogenetics* **57**(3-4), 232–247.

Wallner-Pendleton, E. (2019), 'Marek's Disease in Chickens'. PennState Extension [online] accessed 28/02/2021. **URL:** *https://extension.psu.edu/mareks-disease-in-chickens*

Wallny, H.-J., Avila, D., Hunt, L. G., Powell, T. J., Riegert, P., Salomonsen, J., Skjødt, K., Vainio, O., Vilbois, F., Wiles, M. V. and Kaufman, J. (2006), 'Peptide motifs of the single dominantly expressed class I molecule explain the striking MHC-determined response to Rous sarcoma virus in chickens.', *Proceedings of the National Academy of Sciences of the United States of America* **103**(5), 1434–9.

Walugembe, M., Amuzu-Aweh, E. N., Botchway, P. K., Naazie, A., Aning, G., Wang, Y., Saelao, P., Kelly, T., Gallardo, R. A., Zhou, H., Lamont, S. J., Kayang, B. B. and Dekkers, J. C. (2020), 'Genetic Basis of Response of Ghanaian Local Chickens to Infection With a Lentogenic Newcastle Disease Virus', *Frontiers in Genetics* **11**(July), 1–16.

Walugembe, M., Mushi, J. R., Amuzu-Aweh, E. N., Chiwanga, G. H., Msoffe, P. L., Wang, Y., Saelao, P., Kelly, T., Gallardo, R. A., Zhou, H., Lamont, S. J., Muhairwa, A. P. and Dekkers, J. C. (2019), 'Genetic analyses of Tanzanian local chicken ecotypes challenged with newcastle disease virus', *Genes* **10**(7), 546.

Wang, B., Ekblom, R., Strand, T. M., Portela-Bens, S. and Höglund, J. (2012), 'Sequencing of the core MHC region of black grouse (Tetrao tetrix) and comparative genomics of the galliform MHC', *BMC Genomics* **13**(553), 1–10.

Wang, G. D., Zhai, W., Yang, H. C., Wang, L., Zhong, L., Liu, Y. H., Fan, R. X., Yin, T. T., Zhu, C. L., Poyarkov, A. D., Irwin, D. M., Hytönen, M. K., Lohi, H., Wu, C. I., Savolainen, P. and Zhang, Y. P. (2016), 'Out of southern East Asia: The natural history of domestic dogs across the world', *Cell Research* **26**, 21–33.

Wang, M.-s., Thakur, M., Peng, M.-s., Jiang, Y., Alain, L., Frantz, F., Li, M., Zhang, J.-j., Wang, S., Peters, J. and Otieno, N. (2020), '863 Genomes Reveal the Origin and Domestication of Chicken', *Cell Research* **30**, 693–710.

Wang, Y., Lupiani, B., Reddy, S. M., Lamont, S. J. and Zhou, H. (2014), 'RNA-seq analysis revealed novel genes and signaling pathway associated with disease resistance to avian influenza virus infection in chickens', *Poultry Science* **93**(2), 485–493.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. and Barton, G. J. (2009), 'Jalview Version 2 — a multiple sequence alignment editor and analysis workbench', *Bioinformatics* **25**(9), 1189–1191.

Wearsch, P. A. and Cresswell, P. (2007), 'Selective loading of high-affinity peptides onto major histocompatibility complex class I molecules by the tapasin-ERp57 heterodimer', *Nature Immunology* **8**(8), 873–881.

Weingart, H. M., Copps, J., Drebot, M. A., Marszal, P., Smith, G., Gren, J., Andonova, M., Pasick, J., Kitching, P. and Czub, M. (2004), 'Susceptibility of Pigs and Chickens to SARS Coronavirus', *Emerging Infectious Diseases* **10**(2), 179–184.

Wesley, P. K., Clayberger, C., cchen Lyu, S. and Krensky, A. M. (1993), 'The CD8 coreceptor interaction with the α3 domain of HLA class I is critical to the differentiation of human cytotoxic t-lymphocytes specific for HLA-A2 and HLA-Cw4', *Human Immunology* **36**(3), 149–155.

WHO (2020), 'Zoonotic influenza'. [online] accessed 29/05/2020.
URL: *https://www.who.int/influenza/spotlight/zoonotic-influenza*

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.
URL: *https://ggplot2.tidyverse.org*

Wieczorek, M., Abualrous, E. T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F. and Freund, C. (2017), 'Major histocompatibility complex (MHC) class I and MHC class II proteins: Conformational plasticity in antigen presentation', *Frontiers in Immunology* **8**(292).

Williams, A. P., Peh, C. A., Purcell, A. W., McCluskey, J. and Elliott, T. (2002), 'Optimization of the MHC class I peptide cargo is dependent on tapasin', *Immunity* **16**(4), 509–520.

Wimmers, K., Ponsuksili, S., Hardge, T., Valle-Zarate, A., Mathur, P. K. and Horst, P. (2000), 'Genetic distinctness of African, Asian and South American local chickens', *Animal Genetics* **31**(3), 159–165.

Wise, D. (2019), Understanding antigen processing in chickens using genome editing technology, Phd thesis, University of Cambridge.

Wittig, M., Anmarkrud, J. A., Kässens, J. C., Koch, S., Forster, M., Ellinghaus, E., Hov, J. R., Sauer, S., Schimmler, M., Ziemann, M., Görg, S., Jacob, F., Karlsen, T. H. and Franke, A. (2015), 'Development of a high-resolution NGS-based HLA-typing and analysis pipeline', *Nucleic Acids Research* **43**(11).

Woelfing, B., Traulsen, A., Milinski, M. and Boehm, T. (2009), 'Does intra-individual major histocompatibility complex diversity keep a golden mean?', *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**(1513), 117–128.

Worley, K., Gillingham, M., Jensen, P., Kennedy, L. J., Pizzari, T., Kaufman, J. and Richardson, D. S. (2008), 'Single locus typing of MHC class I and class II B loci in a population of red jungle fowl', *Immunogenetics* **60**(5), 233–247.

Xiang, H., Gao, J., Yu, B., Zhou, H., Cai, D., Zhang, Y., Chen, X., Wang, X., Hofreiter, M. and Zhao, X. (2014), 'Early Holocene chicken domestication in northern China', *Proceedings of the National Academy of Sciences of the United States of America* **111**(49), 17564–17569.

Xiao, J., Xiang, W., Zhang, Y., Peng, W., Zhao, M., Niu, L., Chai, Y., Qi, J., Wang, F., Qi, P., Pan, C., Han, L., Wang, M., Kaufman, J., Gao, G. F. and Liu, W. J. (2018), 'An Invariant Arginine in Common with MHC Class II Allows Extension at the C-Terminal End of Peptides Bound to Chicken MHC Class I', *The Journal of Immunology* **201**(10), 3084–3095.

Yamaguchi and Dijkstra (2019), 'Major Histocompatibility Complex (MHC) Genes and Disease Resistance in Fish', *Cells* **8**(4), 378.

Yamazaki, K., Yamaguchi, M., Andrews, P. W., Peake, B. and Boyse, E. A. (1978), 'Mating preferences of F2 segregants of crosses between MHC-congenic mouse strains', *Immunogenetics* **6**, 253–259.

Yang, Z. (2007), 'PAML 4 : Phylogenetic Analysis by Maximum Likelihood', *Molecular Biology and Evolution* **24**(8), 1586–1591.

Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A.-m. K. (2000), 'Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites', *Genetics* **155**, 431–449.

Yang, Z., Wong, W. S. W. and Nielsen, R. (2005), 'Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection', *Molecular Biology and Evolution* **22**(4), 1107–1118.

Yčas, M. (1974), 'On earlier states of the biochemical system', *Journal of Theoretical Biology* **44**(1), 145–160.

Youssao, I. A., Tobada, P. C., Koutinhouin, B. G., Dahouda, M., Idrissou, N. D., Bonou, G. A., Tougan, U. P., Ahounou, S., Yapi-Gnaoré, V., Kayang, B., Rognon, X. and Tixier-Boichard, M. (2010), 'Phenotypic characterisation and molecular polymorphism of indigenous poultry populations of the species Gallus gallus of Savannah and Forest ecotypes of Benin', *African Journal of Biotechnology* **9**(3), 369–381.

Yu, G., Smith, D. K., Zhu, H., Guan, Y. and Lam, T. T.-y. (2017), 'GGTREE : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data', *Methods in Ecology and Evolution* **8**, 28–36.

Zhang, J., Chen, Y., Qi, J., Gao, F., Liu, Y., Liu, J., Zhou, X., Kaufman, J., Xia, C. and Gao, G. F. (2012), 'Narrow Groove and Restricted Anchors of MHC Class I Molecule BF2*0401 Plus Peptide Transporter Restriction Can Explain Disease Susceptibility of B4 Chickens', *The Journal of Immunology* **189**(9), 4478–4487.