Spatial and temporal variation in *Mhc* class I genes in the house sparrow (*Passer domesticus*)

Dk Haslina Pg Razali

Thesis submitted for the degree of Doctor of Philosophy

Department of Animal and Plant Sciences
The University of Sheffield

March 2016

Abstract

The genes of the major histocompatibility complex (*Mhc*) are known to be highly polymorphic in songbirds, in part due to gene duplication. The *Mhc* class I exon 3 of the house sparrow (*Passer domesticus*) is investigated in this study as it has been characterized previously. We expected next-generation sequencing to enable *Mhc* genotyping to be achieved more accurately than traditional methods with lower resolution. We first compared genotyping performance of two next-generation sequencing techniques (454 amplicon and Illumina MiSeq sequencing). Illumina MiSeq sequencing was more successful than 454 amplicon sequencing when sequencing individuals with a high number of *Mhc* alleles because of the higher read-depth coverage that could be achieved.

Pathogens are thought to maintain the high polymorphism at *Mhc*. We investigated the pattern of differentiation at the *Mhc* in relation to differences in the prevalence and infection with *Plasmodium*, also assessed using new-generation sequencing, in multiple house sparrow populations in New Zealand. The study provided evidence for geographical differentiation at the *Mhc* class I in New Zealand. There was no evidence of a relationship, however, between *Mhc* diversity and *Plasmodium* infection.

Studies on *Mhc*-related fitness have found a relationship between *Mhc* diversity and fitness traits, such as reproductive fitness and survival. Moreover, empirical evidence has suggested an association between specific *Mhc* alleles and fitness traits. We therefore examined the fitness consequences (reproductive fitness and survival) of *Mhc* across cohorts within an intensively monitored long-term study of a house sparrow population on Lundy Island. We found two *Mhc* alleles to be associated with reproductive fitness (number of offspring produced in their lifetime).

Acknowledgements

This PhD has been an eventful journey and I would not have gone through the whole process without the support from many people. First of all, my deepest gratitude goes to my supervisor, Professor Terry Burke, for his endless encouragement, guidance, support and patience throughout the course of my study.

I am very grateful for the extra guidance, support and encouragement from Dr Helena Westerdahl, Dr Emily O'Connor and Dr Julia Schroeder. I am also grateful to Dr Mirre JP Simons for his invaluable statistical help. This project would not have been possible without the support and encouragement of Dr Shinichi Nakagawa, who also provided the blood samples from New Zealand house sparrows. I would also like to thank my PhD predecessors and those involved in monitoring and collecting blood samples from the Lundy Island house sparrows.

Many thanks to Andy Krupa, Maria-Elena Mannarelli and Dr Gavin Horsburgh for their advice and lab support, Dr Helen Hipperson, who set up the *Mhc* work flow in Iceberg, Dr Deborah Dawson who provided advice on primer design, Dr Kang-Wook Kim who repaired the initial MiSeq sequences, Gillian Durieux for her hard work in running the PCR for the TLR samples and Celine Pagnier, for organising Terry's schedule which allowed me to have my meetings with Terry. I really appreciated the help and support of Susana De Freitas, Óscar Mira, Hannah Edwards and Michela Busana, all of whom contributed time to help me run genetic programs and statistical analysis.

A loud shout out to my dear friends from back home for their unending support: Madeeha Samad, Jackie Liew, Chun-Yen Wee, Hadizah Latiff, Khadizah Latiff and Samia Samad. I am also thankful for my friends and colleagues from the UK and Lund for being unreservedly positive: Issie Winney, Rachel Tucker, Steeves Buckland, Hanyuan Zhang, Mauricio

Montano Rendon, Mar Marzo Llorca, Pragya Chaube, Anna Drews, Elin Videvall, Anna-lise Liabot, Jenny Armstrong, Joe Gallagher, Mohammad Awad, Romain Villoutreix and Natalie dos Remedios. To my friends at the GRC, who've been really awesome and a bunch of cheerful people, many thanks to Chanda N Tembo, Fatma Nur Aycicek, Majid Alshahwan, Jawad K Abaies, Sillih Warni and Talia Garza. Many more friends have lent me their support, be it in the form of tea, biscuit, candy and even a small pat on the back, I thank you.

This goes without saying, I would like to thank my family for their unwavering support emotionally, mentally and spiritually. My parents, for being there in my moments of weakness. My brother Lanny, for tolerating the endless complaints and being the voice of calmness in the void of darkness. The last and of course the 'bestest' sister in the world, I especially thank my beloved sister Aisah, who has become the rock in my life, the forever optimist, and of course, for putting up with the ungodly hour calls by the older sister.

I thank the Ministry of Education Brunei Darussalam for providing financial support during my PhD study.

Contents

Abstract I

Acknowledgement II

Contents IV

CHAPTER 1: General introduction 1–19

1.0 The major histocompatibility complex (*Mhc*) 1

1.1 Structure and function of the *Mhc* 2

1.2 Selection on *Mhc* genes 3

1.3 *Mhc* screening methods 5

1.4 Study species: the house sparrow (*Passer domesticus*) 7

1.4.1 New Zealand house sparrows 8

1.4.2 Lundy Island house sparrows 8

1.5 Aims of this study 10

1.6 Hypotheses and predictions of this study 11

1.7 References 13

CHAPTER 2: A qualitative and quantitative comparison of Illumina MiSeq and 454 amplicon sequencing for genotyping major histocompatibility complex (*Mhc*) genetic diversity in house sparrows 20–54

2.0 Abstract 20

2.1 Introduction 21

2.2 Methods 24

2.2.1 Samples and molecular methods 24

2.2.2 Sequence preprocessing 25

2.2.3 Genotyping of MiSeq and 454 data 25

2.2.4 Genotype match between replicates and methods 33

2.2.5 Genotyping efficiency in individuals with different number of alleles.34

2.2.6 Statistical analysis 35

2.3 Results 36

2.3.1 Sequencing depths from the MiSeq and 454 techniques before error correction 36

2.3.2 Distinguishing real putative and artefactual alleles using the MiSeq and 454 techniques 36

2.3.3 Repeatability of MiSeq and 454 genotypes 38

2.3.4 Separating alleles from artefacts in amplicons with different numbers of putative alleles 38

2.3.5 Comparison of genotyping success given the number of putative alleles present 41

2.4 Discussion 43

2.5 References 48

CHAPTER 3: Malaria and its association with *Mhc* in the introduced house sparrow in New Zealand 55–102

3.0 Abstract 55

3.1 Introduction 56

3.2 Methods 60

3.2.1 Study sites 60

3.2.2 Malarial screening 60

3.2.2.1 Polymerase chain reaction (PCR) gel-based detection method for avian malaria detection 61

3.2.2.2 Illumina method for avian malaria detection 62

3.2.3 *Mhc* variation screening 63

3.2.4 Preprocessing sequence data for both malaria and *Mhc* 64

3.2.5 Microsatellite sequencing and genotyping 64

3.2.6 Intrapopulation genetic variability 66

3.2.6.1 Microsatellites 66

3.2.6.2 *Mhc* 66

3.2.7 Population differentiation 67

3.2.7.1 Microsatellite 67

3.2.7.2 *Mhc* 68

3.2.8 Statistical analysis 68

3.2.8.1 Repeatability and malaria infection 68

3.2.8.2 Prevalence of malaria infection 68

3.2.8.3 Genetic variation at *Mhc* and microsatellite 69

3.2.8.4 Isolation by distance test and malaria diversity index 69

3.2.8.5 Association between malaria and *Mhc* 70

3.3 Results 72

3.3.1 Comparison between gel-based and sequencing methods for detecting avian malaria 72

3.3.2 Malarial prevalence and diversity with next-generation sequencing in each location 73

3.3.2.1 Prevalence of infected individuals in each location 73

3.3.2.2 Malaria diversity and prevalence 75

3.3.2.3 Prevalence of each malarial lineage in each house sparrow population 77

3.3.3 Intrapopulation genetic diversity 78

3.3.3.1 *Mhc* diversity 78

3.3.3.2 Microsatellite diversity 81

3.3.4 Population differentiation and isolation by distance 82

3.3.5 Population differentiation and prevalence of malaria 85

3.3.6 Associations between malaria and *Mhc* variation 86

3.4 Discussion 88

3.5 References 94

CHAPTER 4: Temporal variation and survival analysis in an isolated house sparrow (*Passer domesticus*) population 103–142

4.0 Abstract 103

4.1 Introduction 104

4.2 Methods 108

4.2.1 Study sites 108

4.2.2 *Mhc* screening 108

4.2.3 Preprocessing sequence data and *Mhc* genotyping 109

4.2.4 Microsatellite genotyping 110

4.2.5 Testing for temporal variation in *Mhc* allele frequency and individual diversity 111

4.2.6 Survival analysis 112

4.2.6.1 Association between *Mhc* diversity and survival in terms of lifespan 113

4.2.6.2 Association between specific *Mhc* alleles and survival in terms of lifespan 114

4.2.6.3 Association between fitness and variation at control microsatellite variation 116

4.3 Results 116

4.3.1 *Mhc* class I exon 3 genotyping 116

4.3.2 Temporal patterns of *Mhc* diversity 118

4.3.3 Association between survival and *Mhc* diversity 123

4.3.4 Association between survival and specific *Mhc* alleles 127

4.3.5 Association between survival and microsatellite variation 129

4.4 Discussion 133

4.5 References 137

CHAPTER 5: General discussion.. 143–162

5.0 Summary of key findings 143

5.0.1 Next generation typing of the *Mhc* 143

5.0.2 Malaria and its association with *Mhc* 144

5.0.3 Temporal variation and survival analysis in the Lundy population .... .145

5.1 Improving the quality of amplification of *Mhc* alleles for Illumina sequencing 146

5.2 Multiple-parasite infections 147

5.3 Disentangling the forces to explain variation at *Mhc*........................... 148

5.4 Future research.. 153

5.4.1 Malaria from here onwards 153

5.4.2 *Mhc*-dependent patterns of mate choice and extra-pair fertilization 153

5.5 References 156

CHAPTER 1: General introduction

1.0 The major histocompatibility complex (*Mhc*)

The genes of the major histocompatibility complex (*Mhc*), which contain loci coding for peptides involved in antigen presentation, are highly polymorphic. These peptides are responsible for distinguishing an individual organism's own healthy cells from pathogen-infected cells. Besides their role in detecting pathogens, the diverse genes of the *Mhc* are important because they code for different types of antibodies that are responsible for neutralizing the pathogen and stopping it from further invasion (Janeway *et al.* 2005).

Understanding the selection and maintenance driving the variability of *Mhc* alleles has been considered to be a key goal in evolutionary biology (Piertney & Oliver 2006). The *Mhc* genes are directly linked to individual fitness, and it appears to be adaptively beneficial across many taxa to maintain high variability in the *Mhc* genes. One theory to explain the high polymorphism in the *Mhc* genes is that this is a result of selection pressure exerted by pathogens on a population (Doherty & Zinkernagel 1975; Slade & McCallum 1992; Hughes & Yeager 1998; Hedrick 2002). Another theory states that *Mhc*-based mate choice has a role in maintaining diversity in *Mhc* genes (Hamilton & Zuk 1982; Penn 2002).

Several studies especially in rodents have attempted to link the *Mhc* and behaviour, this was evidenced by urinary chemosignals that mediate *Mhc*-mate choice preferences (Egid & Brown 1989; Potts *et al.* 1991; Janssen & Zavazava 1999). Other hypotheses that have been proposed to explain the maintenance of diversity in the *Mhc* genes include mate choice via good genes, genetic compatibility and inbreeding avoidance (Hamilton & Zuk 1982; Nowak *et al.* 1992; Zeh & Zeh 1996; Brown 1997; Jordan & Bruford 1998; Janssen & Zavazava 1999; Penn *et al.* 1999).

1.1 Structure and function of the *Mhc*

The genes of the *Mhc* are a multigene family, which codes for molecules that bind to peptides that act as markers for the immune cell to recognize whether a cell is infected or not (Hughes & Yeager 1998; Westerdahl *et al.* 2005; Spurgin & Richardson 2010). These peptides are synthesized inside the cell. When a cell is infected by a virus, the virus will order the cell to synthesize peptides that are different from the cell's own peptides (self peptides) (Janeway *et al.* 2005). For example, when the *Mhc* molecules bind to self-peptides, the immune cells such as cytotoxic T lymphocytes (CTL) encounter the self-peptide and the CTL avoids destroying the cell because the CTL recognizes that this is an uninfected cell (Trowsdale 1993; Alberts *et al.* 2002; Janeway *et al.* 2005). On the other hand, when a virus infects a cell, the infected cell will synthesizes non-self peptides that are not recognized by CTL and this will prompt the CTL to destroy the infected cell, because the *Mhc* has marked it as such (Bjorkman *et al.* 1987; Alberts *et al.* 2002; Janeway *et al.* 2005).

Two main classes of *Mhc* have been studied in birds: the *Mhc* class I and *Mhc* class II. The class I *Mhc* molecules are synthesized into glucoproteins that are expressed on the surface of nucleated cells (Alberts *et al.* 2002; Janeway *et al.* 2005). These molecules present peptides to the CTL (Trowsdale 1993; Hughes & Yeager 1998; Alberts *et al.* 2002). On the other hand, class II *Mhc* molecules interact with helper T cells, which release cytokines and trigger other cells such as the B-cells to produce antibodies, macrophages and CTL to destroy infected cells (Trowsdale 1993; Alberts *et al.* 2002).

1.2 Selection on *Mhc* genes

Several hypotheses have been proposed to explain the mechanism that maintains the high diversity of the *Mhc* (Penn *et al.* 1999). Pathogen-mediated selection was at first used to explain *Mhc* diversity, where one might expect a close association between specific *Mhc* alleles and the susceptibility to disease (Westerdahl *et al.* 2005; Bonneaud *et al.* 2006). Across the vertebrate spectrum, there is evidence for the role of parasites in *Mhc* polymorphism, for example in mammals (Paterson *et al.* 1998; Oliver *et al.* 2009; Zhang & He 2013; Scherman 2015), fish (Grimholt *et al.* 2003; Wegner *et al.* 2003; Šimková *et al.* 2006; Kalbe *et al.* 2009; Smith *et al.* 2011), reptiles (Olsson *et al.* 2005) and birds (Westerdahl *et al.* 2005; Bonneaud *et al.* 2006; Brouwer *et al.* 2010; Loiseau *et al.* 2011; Spurgin *et al.* 2011).

Mechanisms have been proposed to explain pathogen-driven *Mhc* polymorphism including heterozygote advantage, frequency-dependent selection and fluctuating selection. The heterozygote advantage hypothesis states that mechanisms thought to maintain the high level of *Mhc* polymorphism are the result of heterozygous individuals being more resistant to parasites than homozygotes (Doherty & Zinkernagel 1975; Penn *et al.* 2002). Heterozygotes, who will carry a higher diversity of *Mhc* alleles, might be able to present a wider range of antigens and would thus be able to resist a broader range of pathogens (Doherty & Zinkernagel 1975; Hughes & Yeager 1998; Penn *et al.* 1999).

The rare-allele advantage hypothesis suggests that a rare allele will have selective advantage when pathogens evolve and overcome the resistance conferred by the most common alleles (Hamilton 1980; Takahata & Nei 1990; Slade & McCallum 1992; Penn *et al.* 1999; Zelano & Edwards 2002; Spurgin & Richardson 2010). The rare allele increases in frequency and becomes more common. It will, in turn, lose its advantage due to selection pressure on the pathogen to adapt to the now common allele, thus leading to a continuous evolutionary cycle. The hypothesis mentioned predict a race between the rise of new strains of pathogen and the evolution or spread of new *Mhc* genotypes (Zelano & Edwards 2002). Therefore there is selection pressure on the *Mhc* to be polymorphic in order to combat a wide range of pathogens.

The fluctuating selection hypothesis implies that there will be selection for different alleles at different times and locations as a result of selection pressure determined by pathogen distribution and abundance (Hedrick 2002; Bernatchez & Landry 2003). Significant associations have been found between specific *Mhc* alleles and parasite diversity and abundance (Westerdahl *et al.* 2005; Bonneaud *et al.* 2006; Loiseau *et al.* 2011; Smith *et al.* 2011; Zhang & He 2013; Bichet *et al.* 2015). However, it is not easy to distinguish between the predictions of the rare-allele advantage and fluctuating selection hypotheses. Detecting temporal changes in both parasite resistance alleles and the spatial-temporal variation in parasite strains and their abundance would require long-term studies of multiple populations (Spurgin & Richardson 2010).

Some studies have suggested that high *Mhc* diversity is regulated by pathogen-mediated selection and, in turn, is associated with fitness benefits, such as a relationship between survival and *Mhc* variation (von Schantz *et al.* 1996; Kalbe *et al.* 2009; Brouwer *et al.* 2010; Sepil *et al.* 2012a; Karlsson *et al.* 2015).

1.3 *Mhc* screening methods

The complexity of the *Mhc* in songbirds (passerines) is a result of gene duplication, which has led to the presence of high variation in copy number and the presence of pseudogenes (Balakrishnan *et al.* 2010). Several studies have confirmed that the *Mhc* contains multiple loci and high allelic diversity. In the great reed warbler (*Acrocephalus arundinaceus*), as many as eight *Mhc*-I sequences were detected in an individual and it was suggested that there were at least four *Mhc*-I loci (Westerdahl *et al.* 1999). In the house sparrow (*Passer domesticus*), the number of *Mhc*-I loci found was at least 12 (Karlsson & Westerdahl 2013). The number of *Mhc*-I loci found in blue tits (*Cyanistes caeruleus*) was at least four (Schut *et al.* 2011). The great tits (*Parus major*) had at least 16 functional *Mhc*-I loci (Sepil *et al.* 2012b).

In an earlier house sparrow study using an RFLP technique, Bonneaud *et al.* (2004) found that the *Mhc* class I exon 3 alleles contain both with and without a 6-bp deletion (short and long sequences, respectively). This was further supported in a separate study using 454 amplicon sequencing by Karlsson *et al.* (2013), who found both short and long sequences present in an insular house sparrow population. In the study reported in this thesis, only long sequences that did not have the 6-bp deletion were used (Bonneaud *et al.* 2004; Karlsson & Westerdahl 2013). The long sequences are considered to have classical *Mhc* characteristics. These classical *Mhc* characteristics include high allelic diversity (high nucleotide diversity) and protein-binding sites subjected to positive selection (Aoyagi *et al.* 2002; Karlsson & Westerdahl 2013).

Previously, the *Mhc* was screened using methods such as DGGE (denaturing gradient electrophoreses), SSCP (single strand conformation analysis) and RSCA (reference-strand conformation analysis). These methods were not only time-consuming but, also, did not provide sequence information (Karlsson 2013). With the arrival of new-generation sequencing technologies, such as 454 amplicon sequencing, high resolution and large-scale *Mhc* genotyping became possible (e.g. in passerine birds, Radwan *et al.* 2010; Sommer *et al.* 2013). 454 pyrosequencing recently became obsolete, effectively replaced by Illumina sequencing. Lighten *et al.* (2014) successfully genotyped the targeted *Mhc* class II in fish and were able to distinguish between the alleles considered to be real and those due to PCR or sequencing artefacts. In this thesis, the Illumina MiSeq sequencing technique will be used to screen and genotype *Mhc* alleles.

1.4 Study species: the house sparrow (*Passer domesticus*)

The house sparrow has been used to study the role of pathogen-mediated selection on *Mhc* polymorphism (Bonneaud *et al.* 2006; Loiseau *et al.* 2011; Bichet *et al.* 2015). The natural distribution of the house sparrow extends from the British Isles, throughout Europe, and across north Africa, Arabia and northern Asia (Summers-Smith 1963; Long 1981; Anderson 2006). The house sparrow is a commensal of human societies and this vast distributional range has been attributed to the expansion of agriculture (Summers-Smith 1963; Anderson 2006). The ability of house sparrows to adapt to both agricultural and urban environments has allowed them to thrive, often following introduction, in most locations where humans are present (Summers-Smith 1963). The latter include north and south America, the Caribbean islands, southern and eastern Africa, and many locations in the Pacific including Australia and New Zealand. Many introductions were deliberately made by colonists trying to recreate a European-like environment (Long 1981). In many cases the house sparrow adapted to the new ecological niche with little resistance (Summers-Smith 1963).

The house sparrow is a relatively sedentary, sexually dimorphic and socially monogamous species. Sparrows do not usually migrate once they have become established in a suitable locality (Summers-Smith 1963). They nest in crevices in buildings, holes in trees, in dense bushes and man-made nestboxes (Karlsson 2013). Pair-bonds in house sparrows last throughout the breeding season and sometimes continue between years, until the partner is lost by means of death or divorce (Summers-Smith 1963). In this species, both males and females participate in nest building and nest defence, incubate the eggs and provide food for young (Nakagawa *et al.* 2007). House sparrows are multi-brooded (Anderson 2006) and typically lay four or five eggs in a clutch, in some cases up to six or seven (Murphy 1978).

House sparrows were sampled from two different locations: New Zealand and Lundy Island, England. In this thesis, different questions are addressed in different house sparrow populations.

1.4.1 New Zealand house sparrows

Both introduced and endemic bird species in New Zealand are infected with haemosporidian parasites (Tompkins & Gleeson 2006; Sturrock & Tompkins 2008; Howe *et al.* 2012). This malarial parasite has been spread by the expansion in the range of an exotic mosquito vector, *Culex quinquefasciatus* (Tompkins & Gleeson 2006). There is a gradual decrease of malaria infection with latitude from the North Island to the South Island (Tompkins & Gleeson 2006). The northern part of North Island has a sub-tropical climate, which becomes increasingly temperate towards the south of the South Island. Blood samples were collected from house sparrows in multiple locations in both the North and South Islands. These blood samples were collected and supplied by Dr Shinichi Nakagawa and his team from the University of Otago.

The aim of this part of the study was to investigate whether there is a difference in the frequency of malaria in the different locations and whether this influences *Mhc* diversity and allele frequencies.

1.4.2 Lundy Island house sparrows

Lundy Island is located in the Bristol Channel, 18 km off the north coast of Devon, England (51° 10’ N, 4° 40’ W) (Schroeder *et al.* 2011). This population has very little to no immigration and so is relatively isolated. The island therefore acts as a “natural laboratory”, especially because this population of house sparrows has been closely monitored since 2000 (e.g. Griffith *et al.* 1999; Nakagawa & Burke 2008; Nakagawa *et al.* 2008; Ockendon *et al.* 2009; Cleasby *et al.* 2011; Schroeder *et al.* 2011, 2012;

Karlsson & Westerdahl 2013; Hsu *et al.* 2014; Simons *et al.* 2015; Winney *et al.* 2015; Karlsson *et al.* 2015). During the breeding season (April to August), all occupied nests are observed frequently and details of adult reproductive success and chick or egg survival recorded. Both nestlings and fledglings are marked with a unique combination of plastic coloured rings and a metal ring provided by the British Trust for Ornithology (BTO) (Ockendon *et al.* 2009). Blood samples were taken from individuals and stored in 99% ethanol.

Another feature that makes this study population unique and special is that, unlike other wild bird population studies, a virtually complete pedigree is available for the Lundy sparrow population (Schroeder *et al.* 2011). This feature is especially useful in this study because it enables the inheritance of *Mhc* alleles to be traced back to both genetic parents.

1.5 Aims of this study

The house sparrow was used as a model to investigate the role of pathogens on the maintenance of genetic variation at the *Mhc*. The first objective of the thesis is to confirm that *Mhc* alleles are screened correctly by comparing two next-generation techniques (454 amplicon sequencing and Illumina MiSeq sequencing) (**Chapter 2**). This is followed by an investigation of how genetic variation at *Mhc* is maintained. There are two separate studies: The first tested for an association between *Mhc* and the prevalence of a widespread parasite shown elsewhere to be related to *Mhc* variation – infection with malaria – and addressed this by studying a house sparrow population in which the characteristics of the malarial infection are also investigated (**Chapter 3**). The second tested for an association between fitness and *Mhc* diversity in a long-term intensively monitored house sparrow population in which the pathogens have not been identified (**Chapter 4**).

1.6 Hypotheses and predictions of this study

Chapter 2: A qualitative and quantitative comparison of Illumina MiSeq and 454 amplicon sequencing for genotyping major histocompatibility complex (*Mhc*) genetic diversity in house sparrows

The aim of this chapter is to evaluate the performance of 454 amplicon sequencing and Illumina MiSeq in genotyping *Mhc* class I alleles. It was predicted that Illumina MiSeq sequencing would have greater accuracy due to its higher read depth.

Chapter 3: Malaria and its association with *Mhc* in the introduced house sparrow in New Zealand

The first aim in this chapter is to investigate whether is there any selection on *Mhc* by comparing patterns of differentiation between functional (*Mhc*) and neutral (microsatellite) markers. The hypothesis is that spatial differences in selection will lead to greater among-population differences in *Mhc* diversity than at neutral marker loci. The second aim is to investigate whether population differentiation at the *Mhc* leads to a population-specific pattern of association between *Mhc* alleles and malarial infection.

Chapter 4: Temporal variation and survival analysis in an isolated house sparrow (*Passer domesticus*) population

The first aim is to test for significant heterogeneity in *Mhc* allele frequencies across a 13-year study of a single population, as would be expected if the *Mhc* responds to rapid changes in selection pressure due to rapidly evolving pathogens. Neutral microsatellite loci are again used as a comparator. It is hypothesized that *Mhc* allele frequencies will vary more than expected under random expectation*.* Second, it is possible that some alleles will confer fitness benefits by comparing resistance to current pathogenic challenges. This is examined by testing for an association between fitness (e.g. survival and number of offsprings) and *Mhc* genotype. Finally, previous studies have shown a benefit of having an intermediate rather than extremely low or high degree of *Mhc* allelic diversity within an individual (Kalbe *et al.* 2009). This hypothesis is also tested.

1.7 References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) *Molecular Biology of the Cell*. Garland Science, New York.

Anderson TR (2006) *Biology of the Ubiquitous House Sparrow : From Genes to Populations*. Oxford University Press, New York.

Aoyagi K, Dijkstra JM, Xia C, Denda I, Ototake M, Hashimoto K, Nakanishi T (2002) Classical MHC class I genes composed of highly divergent sequence lineages share a single locus in rainbow trout (*Oncorhynchus mykiss*). *Journal of Immunology*, **168**, 260–273.

Balakrishnan CN, Ekblom R, Völker M, Westerdahl H, Godinez R, Kotkiewicz H, Burt DW, Graves T, Griffin DK, Warren WC, Edwards SV (2010) Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology*, **8**, 1–19.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bichet C, Moodley Y, Penn DJ, Sorci G, Garnier S (2015) Genetic structure in insular and mainland populations of house sparrows (*Passer domesticus*) and their hemosporidian parasites. *Ecology and Evolution*, **5**, 1639–1652.

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, **329**, 512–518.

Bonneaud C, Pérez-Tris J, Federici P, Chastel O, Sorci G (2006) Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, **60**, 383–389.

Bonneaud C, Sorci G, Morin V, Westerdahl H, Zoorob R, Wittzell H (2004) Diversity of Mhc class I and IIB genes in house sparrows (*Passer domesticus*). *Immunogenetics*, **55**, 855–865.

Brouwer L, Barr I, Van De Pol M, Burke T, Komdeur J, Richardson DS (2010) MHC-dependent survival in a wild population: Evidence for hidden genetic benefits gained through extra-pair fertilizations.

Molecular Ecology, **19**, 3444–3455.

Brown JL (1997) A theory of mate choice based on heterozygosity. *Behavioral Ecology*, **8**, 60–65.

Cleasby IR, Burke T, Schroeder J, Nakagawa S (2011) Food supplements increase adult tarsus length, but not growth rate, in an island population of house sparrows (*Passer domesticus*). *BMC Research Notes*, **4**, 431.

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, **256**, 50–52.

Egid K, Brown JL (1989) The major histocompatibility complex and female mating preferences in mice. *Animal Behaviour*, **38**, 548–549.

Griffith SC, Stewart IRK, Dawson DA, Owens IPF, Burke T (1999) Contrasting levels of extra-pair paternity in mainland and island populations of the house sparrow (*Passer domesticus*): is there an “island effect”? *Biological Journal of the Linnean Society*, **68**, 303–316.

Grimholt U, Larsen S, Nordmo R, Midtlyng P, Kjoeglum S, Storset A, Saebø S, Stet RJM (2003) MHC polymorphism and disease resistance in Atlantic salmon (*Salmo salar*); facing pathogens with single expressed major histocompatibility class I and class II loci. *Immunogenetics*, **55**, 210–219.

Hamilton WD (1980) Sex versus Non-Sex versus Parasite. *Oikos*, **35**, 282–290.

Hamilton WD, Zuk M (1982) Heritable true fitness and bright birds: a role for parasites? *Science*, **218**, 384–387.

Hedrick PW (2002) Pathogen Resistance and Genetic Variation at MHC Loci. *Evolution*, **56**, 1902–1908.

Howe L, Castro IC, Schoener ER, Hunter S, Barraclough RK, Alley MR (2012) Malaria parasites (*Plasmodium* spp.) infecting introduced, native and endemic New Zealand birds. *Parasitology Research*, **110**, 913–923.

Hsu YH, Schroeder J, Winney I, Burke T, Nakagawa S (2014) Costly infidelity: Low lifetime fitness of extra-pair offspring in a passerine bird. *Evolution*, **68**, 2873–2884.

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. *Annual Review of Genetics*, **32**, 415–435.

Janeway C, Travers P, Walport M, Shlomchik M (2005) *Immunobiology: The*

Immune System in Health and Disease. Garland Science Publishing.

Janssen E, Zavazava N (1999) How Does the Major Histocompatibility Complex Influence Behavior? *Archivum Immunologiae et Therapiae Experimentalis*, **47**, 139–142.

Jordan WC, Bruford MW (1998) New perspectives on mate choice and the MHC. *Heredity*, **81**, 239–245.

Kalbe M, Eizaguirre C, Dankert I, Reusch TBH, Sommerfeld RD, Wegner KM, Milinski M (2009) Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 925–934.

Karlsson M (2013) The MHC genes: variation and impact on life-history traits in house sparrows. *Lund University*, PhD thesis.

Karlsson M, Schroeder J, Nakagawa S, Smith HG, Burke T, Westerdahl H (2015) House sparrow *Passer domesticus* survival is not associated with MHC-I diversity, but possibly with specific MHC-I alleles. *Journal of Avian Biology*, **46**, 167–174.

Karlsson M, 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.

Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) *Plasmodium relictum* infection and MHC diversity in the house sparrow (*Passer domesticus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **278**, 1264–1272.

Long JL (1981) *Introduced birds of the world: the worldwide history, distribution and influence of birds introduced to new environments*. David & Charles, London.

Murphy EC (1978) Seasonal Variation in Reproductive Output of House Sparrows: The Determination of Clutch Size. *Ecology*, **59**, 1189–1199.

Nakagawa S, Burke T (2008) The mask of seniority? A neglected age indicator in house sparrows *Passer domesticus*. *Journal of Avian Biology*, **39**, 222–225.

Nakagawa S, Gillespie DOS, Hatchwell BJ, Burke T (2007) Predictable males and unpredictable females: sex difference in repeatability of parental care in a wild bird population. *Journal of Evolutionary Biology*, **20**, 1674–1681.

Nakagawa S, Lee J-W, Woodward BK, Hatchwell BJ, Burke T (2008) Differential selection according to the degree of cheating in a status signal. *Biology Letters*, **4**, 667–669.

Nowak MA, Tarczy-Hornoch K, Austyn JM (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**, 10896–10899.

Ockendon N, Griffith SC, Burke T (2009) Extrapair paternity in an insular population of house sparrows after the experimental introduction of individuals from the mainland. *Behavioral Ecology*, **20**, 305–312.

Oliver MK, Telfer S, Piertney SB (2009) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 1119–1128.

Olsson M, Madsen T, Wapstra E, Silverin B, Ujvari B, Wittzell H (2005) MHC, health, color, and reproductive success in sand lizards. *Behavioral Ecology and Sociobiology*, **58**, 289–294.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3714–3719.

Penn DJ (2002) The scent of genetic compatibility: sexual selection and the major histocompatibility complex. *Ethology*, **108**, 1–21.

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11260–11264.

Penn DJ, Potts WK, Palumbi AESR (1999) The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. *The American Naturalist*, **153**, 145–164.

Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**, 7–21.

Potts WK, Manning CJ, Wakeland EK (1991) Mating patterns in seminatural populations of mice influenced by MHC genotype. *Nature*, **352**, 619–621.

Radwan J, Biedrzycka A, Babik W (2010) Does reduced MHC diversity decrease viability of vertebrate populations? *Biological Conservation*, **143**, 537–544.

von Schantz T, Wittzell H, Goransson G, Grahn M, Persson K (1996) MHC genotype and male ornamentation: genetic evidence for the Hamilton-Zuk Model. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 265–271.

Scherman K (2015) MHC polymorphism and host-pathogen interactions: The case of Borrelia in its reservoir host, the bank vole *Myodes glareolus*. *Lund University*.

Schroeder J, Burke T, Mannarelli M-E, Dawson DA, Nakagawa S (2012) Maternal effects and heritability of annual productivity. *Journal of Evolutionary Biology*, **25**, 149–156.

Schroeder J, Cleasby IR, Nakagawa S, Ockendon N, Burke T (2011) No evidence for adverse effects on fitness of fitting passive integrated transponders (PITs) in wild house sparrows *Passer domesticus*. *Journal of Avian Biology*, **42**, 271–275.

Schut E, Aguilar JR, Merino S, Magrath MJL, Komdeur J, Westerdahl H (2011) Characterization of MHC-I in the blue tit (*Cyanistes caeruleus*) reveals low levels of genetic diversity and trans-population evolution across European populations. *Immunogenetics*, **63**, 531–542.

Sepil I, Lachish S, Sheldon BC (2012a) *Mhc*-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular Ecology*, **22**, 384–396.

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012b) 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**, 1–19.

Šimková A, Ottová E, Morand S (2006) MHC variability, life-traits and parasite diversity of European cyprinid fish. *Evolutionary Ecology*, **20**, 465–477.

Simons MJP, Winney I, Nakagawa S, Burke T, Schroeder J (2015) Limited catching bias in a wild population of birds with near-complete census information. *Ecology and Evolution*, **5**, 3500–3506.

Slade RW, McCallum HI (1992) Overdominant Vs. Frequency-Dependent Selection at Mhc Loci. *Genetics*, **132**, 861–862.

Smith C, Ondračková M, Spence R, Adams S, Betts DS, Mallon E (2011) Pathogen-mediated selection for MHC variability in wild zebrafish. *Evolutionary Ecology Research*, **13**, 589–605.

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, **14**, 1–17.

Spurgin LG, Van Oosterhout C, Illera JC, Bridgett S, Gharbi K, Emerson BC, Richardson DS (2011) Gene conversion rapidly generates major histocompatibility complex diversity in recently founded bird populations. *Molecular Ecology*, **20**, 5213–5225.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **277**, 979–988.

Sturrock HJW, Tompkins DM (2008) Avian malaria parasites (*Plasmodium* spp.) in Dunedin and on the Otago Peninsula, southern New Zealand. *New Zealand Journal of Ecology*, **32**, 98–102.

Summers-Smith D (1963) *The House Sparrow*. Collins, London.

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, **124**, 967–978.

Tompkins DM, Gleeson DM (2006) Relationship between avian malaria distribution and an exotic invasive mosquito in New Zealand. *Journal of the Royal Society of New Zealand*, **36**, 51–62.

Trowsdale J (1993) Genomic structure and function in the MHC. *Trends in Genetics*, **9**, 117–122.

Wegner KM, Reusch TBH, Kalbe M (2003) Multiple infections drive major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology*, **16**, 224–232.

Westerdahl H, Waldenström J, Hansson B, Hasselquist D, von Schantz T, Bensch S (2005) Associations between malaria and MHC genes in a migratory songbird. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **272**, 1511–1518.

Westerdahl H, Wittzell H, von Schantz T (1999) Polymorphism and transcription of Mhc class I genes in a passerine bird, the great reed warbler. *Immunogenetics*, **49**, 158–170.

Winney I, Nakagawa S, Hsu YH, Burke T, Schroeder J (2015) Troubleshooting the potential pitfalls of cross-fostering. *Methods in Ecology and Evolution*, **6**, 584–592.

Zeh JA, Zeh DW (1996) The Evolution of Polyandry I: Intragenomic Conflict and Genetic Incompatibility. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 1711–1717.

Zelano B, Edwards S V (2002) An Mhc component to kin recognition and mate choice in birds: predictions, progress, and prospects. *The American Naturalist*, **160**, S225–S237.

Zhang M, He H (2013) Parasite-mediated selection of major histocompatibility complex variability in wild brandt's voles (*Lasiopodomys brandtii*) from Inner Mongolia, China. *BMC Evolutionary Biology*, **13**, 1–15.

CHAPTER 2: A qualitative and quantitative comparison of Illumina MiSeq and 454 amplicon sequencing for genotyping major histocompatibility complex (*Mhc*) genetic diversity in house sparrows

Availability of sequence data

The sequence data generated and analysed during this study will be submitted to an online repository.

2.0 Abstract

The passerine *Mhc* loci are polymorphic and polygenic as a result of gene duplication. Previous genotyping methods were of low resolution, underestimating the actual number of *Mhc* alleles in an amplicon. However, with the advent of next-generation sequencing, accurate and precise multiplexed genotyping has become possible. The aim of this study was to evaluate and compare the performance of 454 amplicon sequencing and Illumina MiSeq amplicon sequencing in genotyping *Mhc* class I alleles. The results showed that, even at low read depth, the 454 was normally successful in genotyping the *Mhc;* however, there was an increased failure rate when genotyping a higher number of *Mhc* alleles within an individual. There was a lower failure rate for *Mhc* genotyping using Illumina MiSeq, but this methodology is very sensitive to contamination between samples and appropriate methodological precautions are required. Overall, both next-generation sequencing techniques were consistently successful at assessing *Mhc* genetic diversity, provided that sequencing coverage was adequate.

2.1 Introduction

The major histocompatibility complex (*Mhc*) plays a key role in adaptive immunity by presenting antigens to the immune system for elimination. Across all known vertebrates, the genetic region encoding *Mhc* is the most polymorphic to have been described (Edwards & Hedrick 1998; Janeway *et al.* 2005). The *Mhc* is highly polymorphic and this was thought to be maintained by selection on multiple *Mhc* alleles that confers with pathogen recognition (Janeway *et al.* 2005; Sommer 2005; Spurgin & Richardson 2010). Within the field of ecology and evolutionary biology, *Mhc* genes have attracted a great deal of research attention, in part due to their association with fitness traits (e.g. survival, lifetime reproductive success, disease resistance and fecundity) (Paterson *et al.* 1998; Wegner *et al.* 2003; Bonneaud *et al.* 2004a; Kalbe *et al.* 2009; Sepil *et al.* 2012a). However, the polymorphic and polygenic nature of *Mhc* presents a challenge when sequencing these genes (Zagalska-Neubauer *et al.* 2010; Sepil *et al.* 2012b). Next-generation sequencing (NGS) technologies offer an excellent opportunity for high-throughput deep sequencing at relatively low cost. Thus, high-throughput PCR-based techniques have made sequencing all the variants in a given sample affordable and practical (Thomas *et al.* 2006; Babik *et al.* 2009). Consequently, NGS is becoming the standard approach for sequencing *Mhc* genes in non-model systems (e.g. Oomen *et al.* 2013; Sommer *et al.* 2013; Lighten *et al.* 2014; Bolnick *et al.* 2014).

NGS comes at the cost of ‘noisy’ data: the high read number obtained through NGS is associated with a substantial number of artefactual reads (Margulies *et al.* 2005; Gomez-Alvarez *et al.* 2009; Quince *et al.* 2011). An enduring challenge when working with NGS data is accurately separating true allelic reads from artefacts. This can be particularly difficult when working with a multi-gene family, such as the *Mhc*, and particularly so in species with a high number of similar genes that cannot be amplified separately. In part, this is because distinguishing artefacts from true alleles typically involves comparing the relative frequencies of all the alternative sequences detected within an amplicon (Babik *et al.* 2009; Galan *et al.* 2010; Lighten *et al.* 2014); as the number of alleles per amplicon increases the read depth per allele will inevitably decrease, and the signal of genuine alleles becomes harder to distinguish from the "noise" due to the artefacts.

454 sequencing of amplicons has been a popular choice in the study of *Mhc* (e.g. Babik *et al.* 2009; Galan *et al.* 2010; Oomen *et al.* 2013). However, there has been a shift towards sequencing using the Illumina MiSeq in place of the 454. The latter has now been discontinued (Anon 2013). The MiSeq platform offers greater sequence coverage at a lower per-base cost than the 454 and generates substantially fewer errors (Loman *et al.* 2012). A lower error-rate could be particularly beneficial in study systems where individuals possess many *Mhc* alleles and *Mhc* alleles with high similarity. Several examples of species with many *Mhc* alleles can be found among the birds of the order Passeriformes, as most species within this order have multiple *Mhc* genes, i.e. are highly polygenic (e.g. Schut *et al.* 2011; Alcaide *et al.* 2013; Karlsson & Westerdahl 2013). To date, the focal exons in NGS studies of *Mhc* in avian non-model organisms have been *Mhc* class I exon 3 and *Mhc* class IIB exon 2 (Bonneaud *et al.* 2004b; Westerdahl *et al.* 2004; Schut *et al.* 2011)

Genotyping *Mhc* class I alleles in songbirds provides a challenge but also an excellent opportunity to test quantitatively the level of polygenicity at which the MiSeq outperforms 454 sequencing. The MiSeq significantly improves our ability to discern true alleles from artefacts in the case of species with extremely high numbers of *Mhc* genes (Biedrzycka *et al.* submitted). In our study, we have chosen to work with *Mhc* class I in house sparrows, *Passer domesticus*, since house sparrows have a moderate number of *Mhc* class I genes (Bonneaud *et al.* 2004b; Karlsson & Westerdahl 2013). Moreover, *Mhc* class I has been partly characterized in house sparrows, enabling us to use appropriate primers (Bonneaud *et al.* 2004b; Karlsson & Westerdahl 2013). Furthermore, several studies have already investigated *Mhc* class I exon 3 in house sparrows over the last ten years using different molecular genetic techniques (Bonneaud *et al.* 2004b; Borg *et al.* 2011; Loiseau *et al.*

2011). This information provides us with prior expectations on allelic lengths and on the number of *Mhc* class I alleles per individual.

The aim of this study is to evaluate and compare, qualitatively and quantitatively, the performance of 454 amplicon sequencing in relation to Illumina MiSeq amplicon sequencing for genotyping *Mhc* class I alleles. We will quantify the extent to which true alleles can be distinguished from artefacts by comparing data from these two different sequencing methods. Our dataset comprises Illumina MiSeq and 454 data for all the individuals comprising 11 single-generation house sparrow families, plus 15% replicated samples. We have prior knowledge of primer performance, as well as the advantage of being able to use heritability within families to aid our assessment of the performance of the alternative techniques in *Mhc* genotyping.

2.2 Methods

2.2.1 Samples and molecular methods

Blood samples were taken from 81 house sparrow individuals belonging to 11 families: 11 adult males, 10 adult females and 60 nestlings (including offspring combined from successive broods belonging to the same pair). There was a minimum of three offspring in each family. Twelve individuals were picked at random and duplicated. The sparrow samples were obtained from a population inhabiting Lundy Island, located in the Bristol Channel (51°10° N, 4°40° W, UK) (Schroeder *et al.* 2012).

Genomic DNA was extracted using a salt extraction (Bruford *et al.* 1998). The DNA concentration was then standardized to 20–25 ng/µl. The forward primer HNalla 5’-TCCCCACAGGTCTCCACAC-3’ and the reverse primer rv3 5’-TGCGCTCCAGCTCCYTCTGCC-3’ were used to amplify a 219–225-bp long fragment of *Mhc* class I exon 3 that contains the most variable portion of the peptide binding region (Westerdahl *et al.* 2004). So that we could subsequently identify and separate the amplicons from each individual, the forward and reverse primers were each tagged with a unique 6-bp sequence combination (Kloch *et al.* 2010). Polymerase chain reactions were performed in 15-µl volumes containing QIAGEN Multiplex MasterMix, 10–20 ng DNA and 0.2 µM of each primer. PCRs were performed in Tetrad PTC-225 Thermal Cycler (MJ RESEARCH, Waltham, Massachusetts) using the following settings: 95°C at 15 min, then 30 cycles of 95°C for 30s, 65°C for 60s and 72°C for 60s, followed by a final extension at 72°C for 10 min. PCR products were separated on a 1.5% agarose gel stained with Syber Safe (Invitrogen). This was to test whether PCR products were present. Amplified DNA from sets of eight individuals was pooled and purified using a MinElute PCR purification kit (QIAGEN) according to the manufacturer’s instructions. These samples were then further pooled and prepared for 250-bp paired-end Illumina MiSeq sequencing (Illumina, Inc., San Diego, CA, USA) and 454 pyro-sequencing (Roche, Branford, CT, USA).

2.2.2 Sequence preprocessing

In the case of the MiSeq data, the sequences were assembled, based on ≥100-bp overlaps, using FLASH (Magoč & Salzberg 2011). Next, PRINSEQ was used to remove any sequences with a Phred quality score below Q30 (Schmieder & Edwards 2011). Finally, sequences were demultiplexed, trimmed of their tags and primer sequences, then summarised in a table listing the read number (read depth) of each sequence in each amplicon using jMHC (Stuglik *et al.* 2011). In the case of the 454 data, the raw fasta file was processed by jMHC in the same manner.

2.2.3 Genotyping of MiSeq and 454 data

Only variants between 239–242-bp in length were retained in the dataset in order to eliminate the 'short' (236-bp) *Mhc* class I alleles, which are co-amplified, from analysis. 'Short' *Mhc* class I alleles in house sparrows contain a 6-bp deletion in exon 3 and it is thought that they might represent non-classical *Mhc* genes (Karlsson & Westerdahl 2013). In the present study, we will focus only on classical (i.e. 'Long') *Mhc* alleles, which are characterized by high nucleotide diversity and positively selected protein-binding sites (Allen & Hogan 2001; Aoyagi *et al.* 2002; Karlsson & Westerdahl 2013).

The degree of change (DOC) method described by Lighten *et al.* (2014) was used for genotyping both MiSeq and 454 data. This method uses the relative read depths of variants to distinguish true alleles from artefactual sequences, as well as estimating copy number variation from the relative read depths of putative alleles. There are two steps: (1) error correction and (2) genotype estimation.

The error correction entails assigning reads arising from artefacts, caused by sequencing error, to the true alleles from which they arose, i.e. the parent variant sequence for a subset of the data. The parent variant sequence is defined as the sequence that had a high number of reads and that has been assigned as a real putative allele for each individual. This ‘cleaning’ step increases the read depth of true variants relative to artefactual variants. The error correction step was performed separately and sequentially for each amplicon. To identify variants arising from the true alleles, the 12 variants with the highest read depths were aligned for each amplicon separately and neighbour-joining trees produced in CodonCode aligner 5.0.2 (CodonCode Corporation). This approach enabled the identification of artefacts, such as variants that differed by fewer than three nucleotides and homopolymer errors. In the case of variants which differed by fewer than three nucleotides, the read depths of the two variants were checked and if the read depth of one of the variants was less than 50% of the other variant, then the variant with the lower read depth was considered a possible artefactual variant.

In the case of the 454 data, we checked whether the same possible artefact occurred in any other amplicons. If it did not, it was considered an artefact and thus deleted and the reads added to the parent variant. In 454 sequencing many nucleotide substitution errors occur during the PCR as a result of Taq polymerase action. The positions of these substitutions are believed to be more or less random making the probability of a substitution occurring twice in the same position twice small (Bracho *et al.* 1998). Thus we considered that if a variant differed from another by three or fewer nucleotides and these collectively occurred in more than one amplicon in the 454 data, then it was more likely to be a true variant than an artefact (**Figure 2.1**).

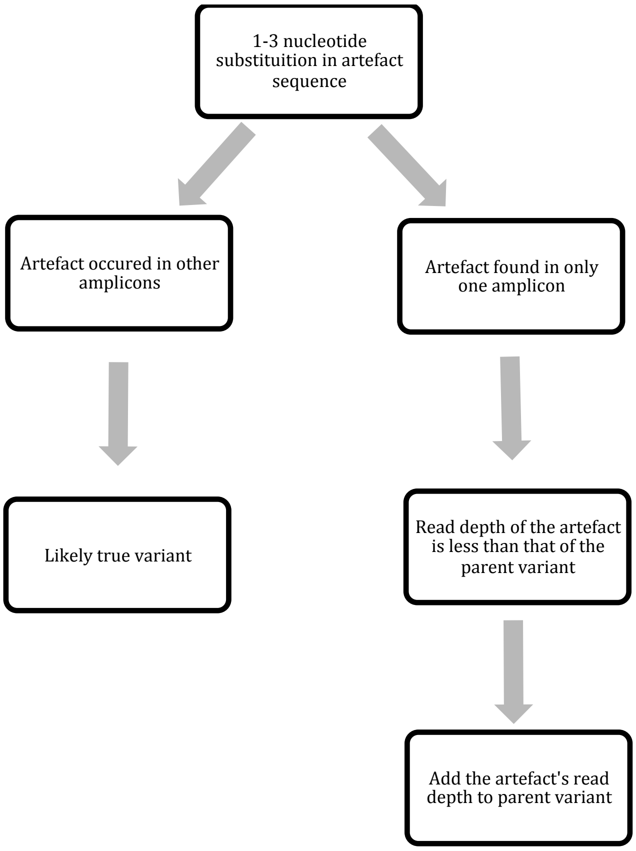

1-3 nucleotide substituition in artefact sequence
Artefact occured in other amplicons
Artefact found in only one amplicon
Likely true variant
Read depth of the artefact is less than that of the parent variant
Add the artefact's read depth to parent variant

Figure 2.1 Flow chart for error correction for 454 data.

This rule was different for variants differing by 3 or fewer nucleotides in the MiSeq data (Lighten *et al.* 2014), as the Illumina sequencing process is more prone to generating repeatable nucleotide substitution errors, *i.e.* miscalling the same nucleotide repeatedly, depending on the flanking nucleotide sequence (Nakamura *et al.* 2011; Schirmer *et al.* 2015). Thus, if two variants differed by fewer than 3 nucleotides in the Illumina data, and one of these variants always occurred at less than 50% of the read depth of the other, as well as never occurring in an amplicon without the parent variant, then this was treated as an artefact. In this case, the artefactual reads are added to the parent variant (**Figure 2.2**).

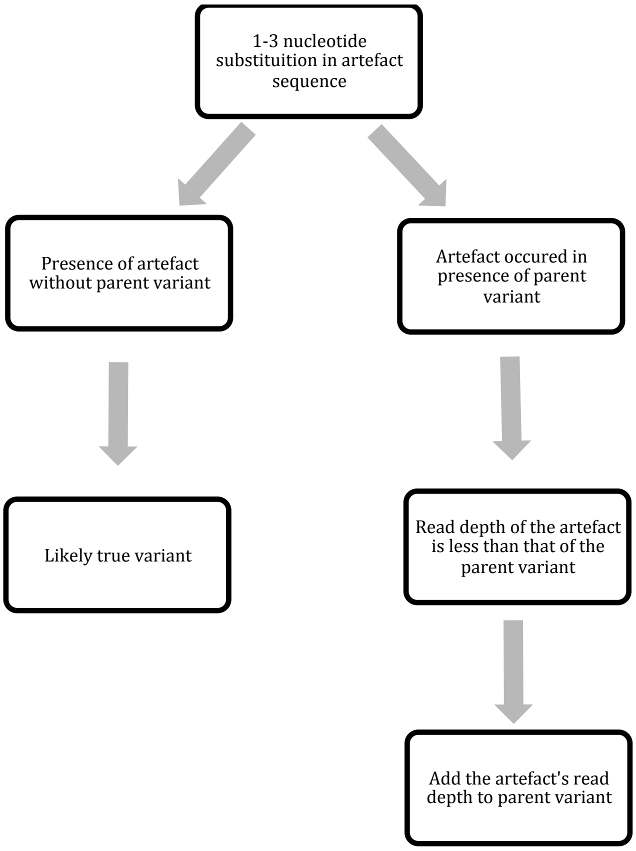

1-3 nucleotide substituition in artefact sequence
Presence of artefact without parent variant
Artefact occured in presence of parent variant
Likely true variant
Read depth of the artefact is less than that of the parent variant
Add the artefact's read depth to parent variant

Figure 2.2 Flow chart for error correction of Illumina data.

In the case of artefacts arising from homopolymer sequencing errors, as occur commonly in 454 data, the artefact was deleted and the reads were added to the parent variant, as long as (1) the homopolymer error variant had lower read depth than the parent variant and (2) the homopolymer error variant did not occur in any other amplicon without the parent variant. Correction of variants arising from homopolymer errors was undertaken across the whole dataset (**Figure 2.3**).

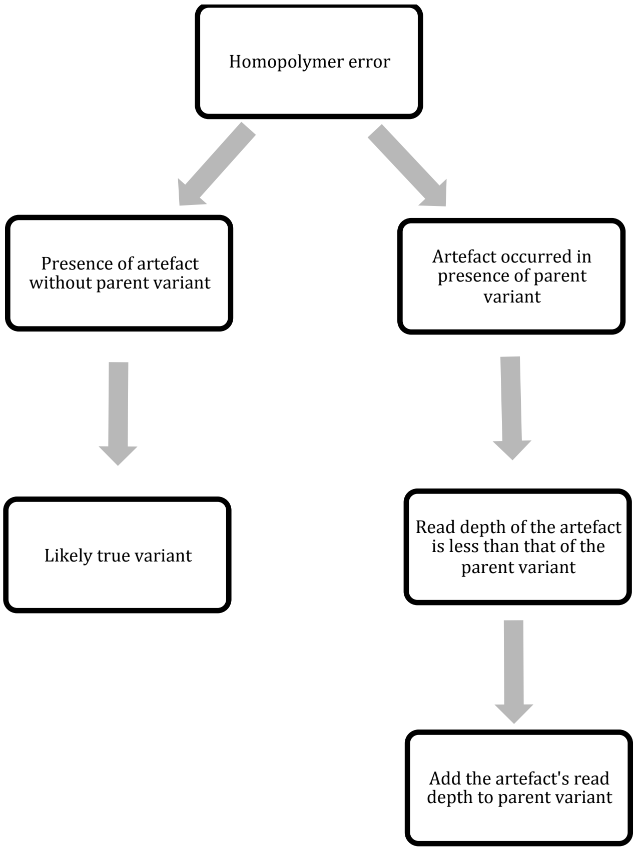

Homopolymer error
Presence of artefact without parent variant
Artefact occurred in presence of parent variant
Likely true variant
Read depth of the artefact is less than that of the parent variant
Add the artefact's read depth to parent variant

Figure 2.3 Flow chart for homopolymer correction of 454 data.

After error correction, the degree of change (DOC) value was calculated as detailed in Lighten *et al.* (2014). One of the key assumptions underpinning Lighten *et al.'s* DOC protocol is that real alleles will be amplified at significantly higher sequencing depths than artefacts, and that there should be a clear difference in the rate of change (ROC) in the cumulative sequencing depth between the true allele with the lowest sequencing depth and the artefact with the highest sequencing depth (See **Figure 2.4** for the flowchart for genotyping *Mhc* using the DOC method). The ROC value was calculated as the difference in sequencing depth between each variant when ordered in relation to read depth (e.g. depth of variant sequence 1 – depth of variant sequence 2 = ROC 1; depth of variant sequence 2 – depth of variant sequence 3 = ROC 2; calculation continues until determining value for ROC 9). Further calculations described by Lighten *et al.* (2014) enable the DOC around each variant to be calculated as a proportion of the total change (e.g.

ROC 1 divided by the sum of the ROC values) – in this study using the top 10 variants per individual, on the assumption that up to 8 true variants exist: this assumption was appropriate for our data, as house sparrows are thought to possess a maximum of eight classical *Mhc* alleles per individual (Borg *et al.* 2011; Karlsson & Westerdahl 2013). The variant with the largest DOC value is considered to be the least-amplified true allele, after which the DOC decreases substantially due to the lower sequencing depth of artefacts (**Figure 2.4**).

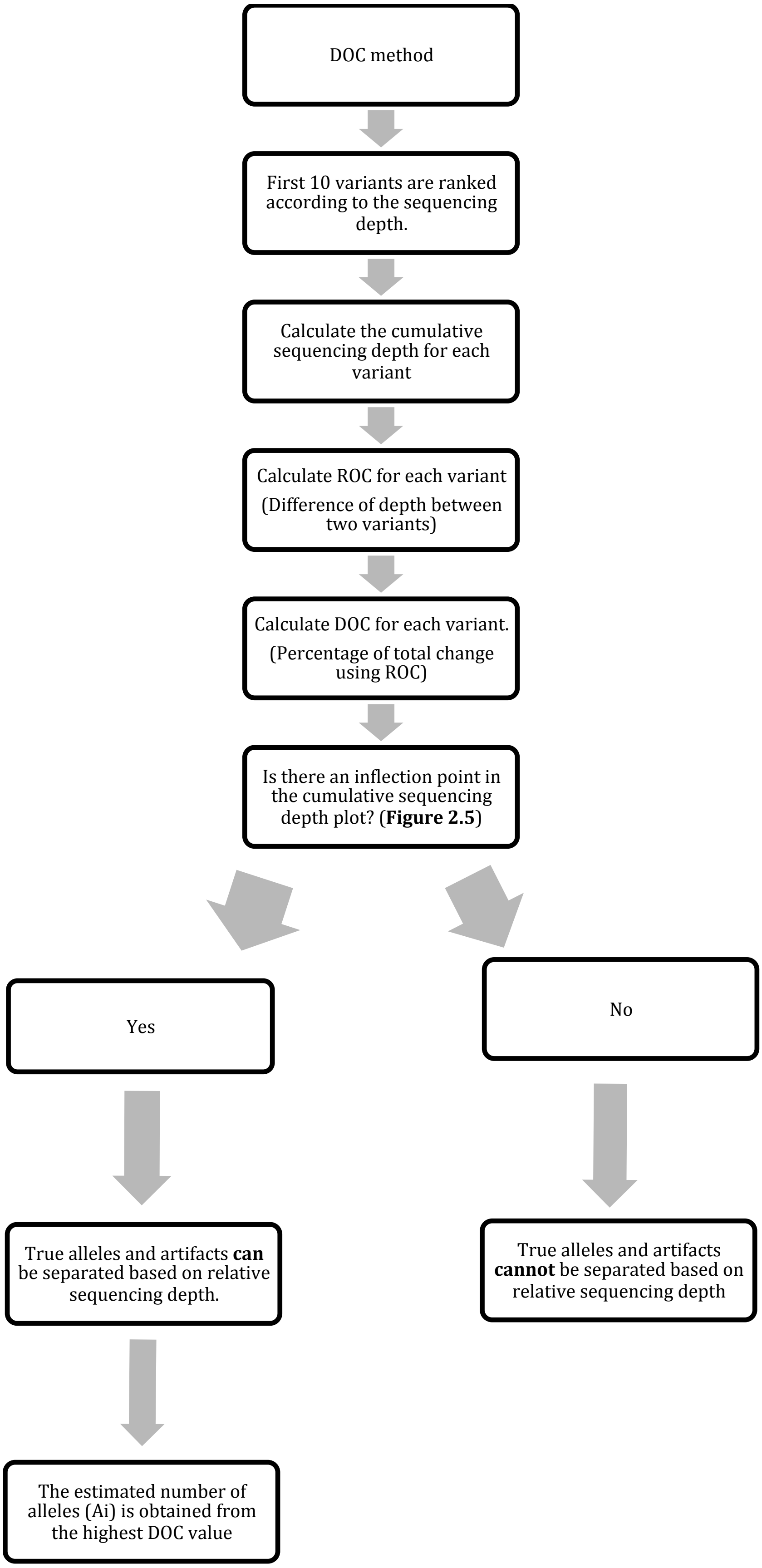
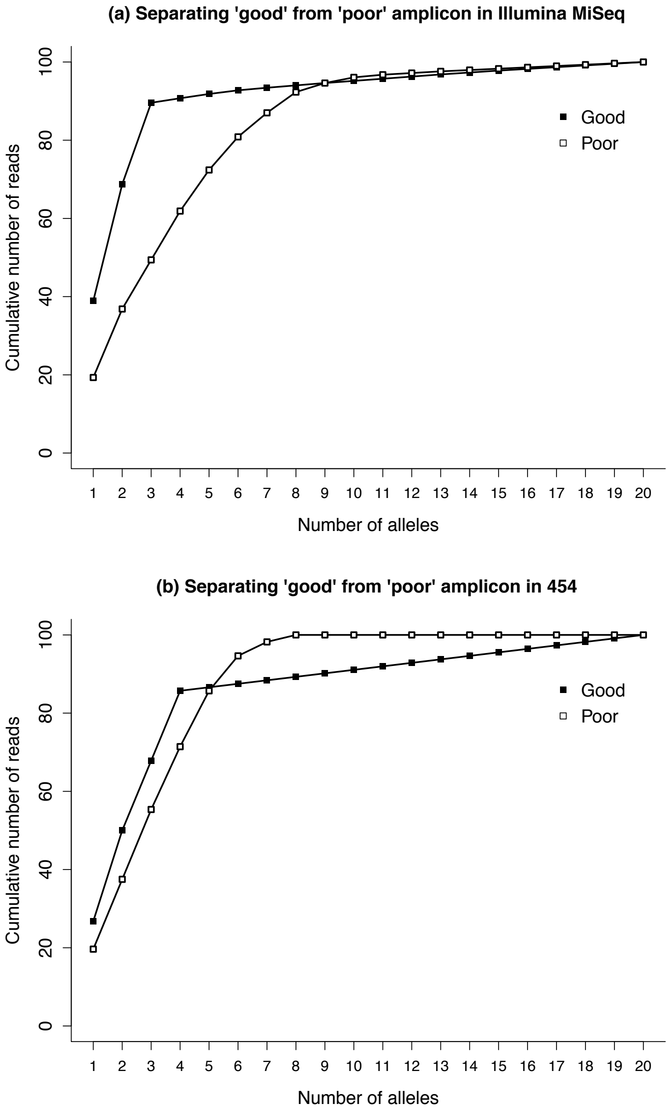

DOC method
First 10 variants are ranked according to the sequencing depth.
Calculate the cumulative sequencing depth for each variant
Calculate ROC for each variant
(Difference of depth between two variants)
Calculate DOC for each variant.
(Percentage of total change using ROC)
Is there an inflection point in the cumulative sequencing depth plot? (**Figure 2.5**)
Yes
No
True alleles and artifacts **can** be separated based on relative sequencing depth.
True alleles and artifacts **cannot** be separated based on relative sequencing depth
The estimated number of alleles (Ai) is obtained from the highest DOC value

Figure 2.4 Flow chart for estimating the number of *Mhc* alleles using the DOC method.

The DOC method enables an estimation of the number of alleles (A_i) present in each individual (**Figure 2.4**). Lighten *et al.* (2014) proposed that the difference in the ROC between true alleles and artefacts should be clearly visible as an inflection point in a linear plot of cumulative sequencing depth; it is therefore not possible to accurately genotype with this method any amplicon in which a clear inflection point is lacking (**Figure 2.5**). Cumulative depth graphs were plotted in Microsoft Excel for each amplicon to enable the identification of genotypes with, and without, clear inflection points. Three independent evaluations were made by three different researchers to classify each amplicon either as a ‘good amplicon’, with a clear inflection point, or a ‘poor amplicon’, without; the latter were excluded from further analysis.


(a) Separating 'good' from 'poor' amplicon in Illumina MiSeq

(b) Separating 'good' from 'poor' amplicon in 454

Figure 2.5 Examples of 'good' and 'poor' amplicons for MiSeq and 454 data, respectively. 'Good' (filled) amplicons have a clear difference in the sequencing depth of true alleles and artefact variants. In 'poor' amplicons (open) this difference is less distinct. (a) MiSeq amplicons: examples of a 'good' amplicon with three real putative alleles and a 'poor' amplicon. (b) 454 amplicons: examples of a 'good' amplicon with four real putative alleles and a 'poor' amplicon.

2.2.4 Genotypic match between replicates and methods

The proportional genotypic match for the 12 replicated sample pairs was calculated separately for the MiSeq and 454 data. Also, the number of alleles that the two replicates had in common was divided by the total number of alleles in the replicate pair, and then expressed as percentage match between the two methods for that individual. The consensus genotype of each replicated pair was determined and used to compare the two techniques. Then the match between the MiSeq and 454 data for the 81 non-replicated individuals was calculated using the same approach.

With the aim of further verifying the reliability of the genotypes, including the samples that were not replicated, the genotypes of the chicks were compared to those of their parents. This was to confirm that no artefactual alleles, which would not have been expected to be present in both parents and offspring, had been assigned as real.

2.2.5 Genotyping efficiency in individuals with different numbers of alleles

The mean read depths of true alleles (A_i, as calculated from the DOC method) and artefacts per amplicon were calculated for different A_i values. The relative success rate ('good amplicons') was estimated by summarising the numbers of amplicons detected using both techniques, MiSeq and 454, and then looking at each technique separately.

Genotyping success rates were assessed separately for genotypes comprising different numbers of alleles when comparing the techniques. For each technique, genotyping success at each *Mhc* diversity level (i.e. the number of putative alleles) was determined as the proportion of individuals successfully genotyped by that technique divided by the total number of individuals successfully genotyped using both techniques.

2.2.6 Statistical analysis

We used Mann-Whitney *U*-test to examine whether there is a difference in read depth between 454 and Illumina MiSeq amplicon sequencing. A χ^2 test was used to investigate whether the frequency of 'good' amplicons differed between the two techniques (Crawley 2005). A Mann-Whitney *U*-test was also used to examine whether the read depths between 'good' and 'bad' amplicons differed within each technique (Crawley 2005).

All statistical analysis and plots were constructed using R (R Development Core Team 2015) or otherwise stated.

2.3 Results

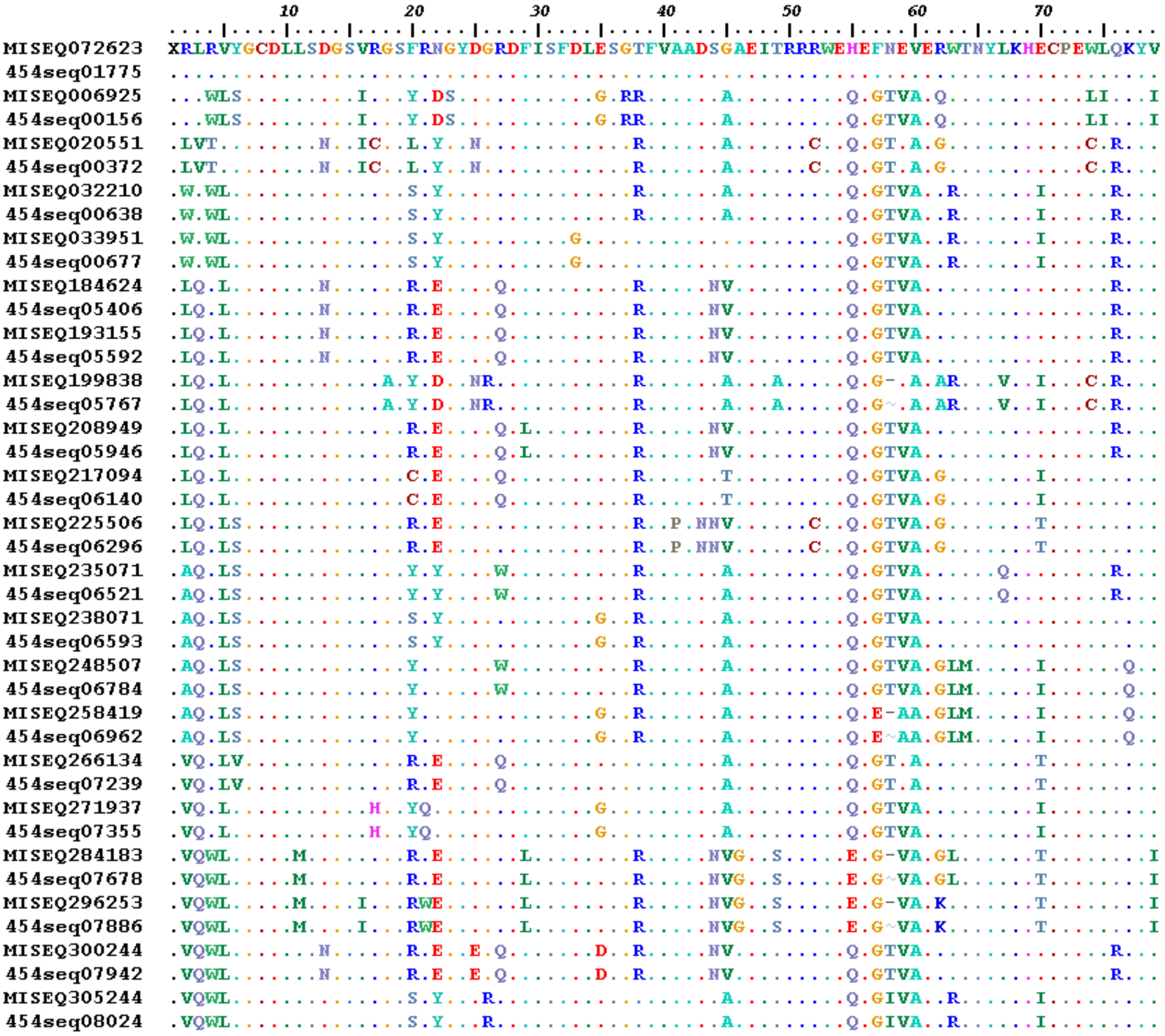
2.3.1 Sequencing depths from the MiSeq and 454 techniques before error correction

The same dataset, consisting of 81 individuals (parents and offspring from 11 families), and 12 replicates (in total 93 amplicons), was sequenced using amplicon sequencing with both the Illumina MiSeq (MiSeq) and 454 pyro-sequencing (454) techniques. As expected, considerably more reads were obtained from the MiSeq sequencing run than the 454 run (total number of reads for sequences with complete tags and primers, MiSeq: 727,913, 454: 17,687; mean ± se number of reads per amplicon for classical and non-classical *Mhc* class I variants combined, MiSeq: 4,923 ± 99, 454: 126 ± 3; mean ± se number of reads per amplicon for classical *Mhc* class I variants only, MiSeq: 2,747 ± 58, 454: 66 ± 1.6). Ninety-three and 92 amplicons were successfully sequenced in the MiSeq and 454 runs, respectively.

2.3.2 Distinguishing real from artefactual alleles using the MiSeq and 454 techniques

After error correction, 24 'poor' amplicons were discarded from the 454 dataset (*n* = 68 remaining amplicons), whereas 17 'poor' amplicons were removed from the MiSeq dataset (n = 76 remaining amplicons). No homopolymers, chimaeras or single base-pair mismatch sequences were identified among the real alleles identified by genotyping using the DOC method in either the MiSeq or the 454 data.

The same 21 putative *Mhc*-I alleles were found in both the 454 and MiSeq datasets (**Figure 2.2**). There were between three and nine putative alleles per individual, indicating that the number of classical *Mhc*-I loci is probably between two and five per individual. As expected, there were substantially fewer reads for the 454 amplicons compared to the MiSeq amplicons after error correction (mean ± se reads per amplicon after error correction, MiSeq: 2,818 ± 62; 454: 63 ± 2); mean ± se reads per putative allele, MiSeq: 264 ± 7; 454: 11 ± 0.3; range of reads per putative allele, MiSeq: 65-768; 454: 3-36) (Mann-Whitney *U*-test, $Z = 22.6$, $P < 0.001$).



```
                    10        20        30        40        50        60        70
             ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
MISEQ072623  XRLRVYGCDLLSDGSVRGSFRNGYDGRDFISFDLESGTFVAADSGAEITRRRWEHEFNEVERWTNYLKHECPEWLQKYV
454seq01775  ...............................................................................
MISEQ006925  ...WLS.........I...Y.DS...........G.RR......A.........Q.GTVA.Q...........LI...I
454seq00156  ...WLS.........I...Y.DS...........G.RR......A.........Q.GTVA.Q...........LI...I
MISEQ020551  .LVT........N..IC..L.Y..N............R......A......C..Q.GT.A.G...........C.R...
454seq00372  .LVT........N..IC..L.Y..N............R......A......C..Q.GT.A.G...........C.R...
MISEQ032210  .W.WL..............S.Y...............R......A.........Q.GTVA..R......I.....R...
454seq00638  .W.WL..............S.Y...............R......A.........Q.GTVA..R......I.....R...
MISEQ033951  .W.WL..............S.Y..........G.....................Q.GTVA..R......I.....R...
454seq00677  .W.WL..............S.Y..........G.....................Q.GTVA..R......I.....R...
MISEQ184624  .LQ.L.......N......R.E....Q..........R.....NV.........Q.GTVA...............R...
454seq05406  .LQ.L.......N......R.E....Q..........R.....NV.........Q.GTVA...............R...
MISEQ193155  .LQ.L.......N......R.E....Q..........R.....NV.........Q.GTVA...............R...
454seq05592  .LQ.L.......N......R.E....Q..........R.....NV.........Q.GTVA...............R...
MISEQ199838  .LQ.L............A.Y.D..NR...........R......A...A.....Q.G-.A.AR...V..I...C.R...
454seq05767  .LQ.L............A.Y.D..NR...........R......A...A.....Q.G~.A.AR...V..I...C.R...
MISEQ208949  .LQ.L..............R.E....Q.L........R.....NV.........Q.GTVA...............R...
454seq05946  .LQ.L..............R.E....Q.L........R.....NV.........Q.GTVA...............R...
MISEQ217094  .LQ.L..............C.E....Q..........R......T.........Q.GTVA.G.......I.........
454seq06140  .LQ.L..............C.E....Q..........R......T.........Q.GTVA.G.......I.........
MISEQ225506  .LQ.LS.............R.E...............R..P.NNV......C..Q.GTVA.G.......T.........
454seq06296  .LQ.LS.............R.E...............R..P.NNV......C..Q.GTVA.G.......T.........
MISEQ235071  .AQ.LS.............Y.Y....W..........R......A.........Q.GTVA......Q........R...
454seq06521  .AQ.LS.............Y.Y....W..........R......A.........Q.GTVA......Q........R...
MISEQ238071  .AQ.LS.............S.Y............G..R......A.........Q.GTVA...................
454seq06593  .AQ.LS.............S.Y............G..R......A.........Q.GTVA...................
MISEQ248507  .AQ.LS.............Y......W..........R......A.........Q.GTVA.GLM.....I......Q..
454seq06784  .AQ.LS.............Y......W..........R......A.........Q.GTVA.GLM.....I......Q..
MISEQ258419  .AQ.LS.............Y..............G..R......A.........Q.E-AA.GLM.....I......Q..
454seq06962  .AQ.LS.............Y..............G..R......A.........Q.E~AA.GLM.....I......Q..
MISEQ266134  .VQ.LV.............R.E....Q.................A.........Q.GT.A.........T.........
454seq07239  .VQ.LV.............R.E....Q.................A.........Q.GT.A.........T.........
MISEQ271937  .VQ.L...........H..YQ.............G.........A.........Q.GTVA.........I.........
454seq07355  .VQ.L...........H..YQ.............G.........A.........Q.GTVA.........I.........
MISEQ284183  .VQWL.....M........R.E......L........R.....NVG..S.....E.G-VA.GL......T........I
454seq07678  .VQWL.....M........R.E......L........R.....NVG..S.....E.G~VA.GL......T........I
MISEQ296253  .VQWL.....M....I...RWE......L........R.....NVG..S.....E.G-VA.K.......T........I
454seq07886  .VQWL.....M....I...RWE......L........R.....NVG..S.....E.G~VA.K.......T........I
MISEQ300244  .VQWL.......N......R.E..E.Q.......D..R.....NV.........Q.GTVA...............R...
454seq07942  .VQWL.......N......R.E..E.Q.......D..R.....NV.........Q.GTVA...............R...
MISEQ305244  .VQWL..............S.Y...R..................A.........Q.GIVA..R......I.........
454seq08024  .VQWL..............S.Y...R..................A.........Q.GIVA..R......I.........
```

Figure 2.6 Alignment of house sparrow *Mhc* class I amino acid sequences deduced from MiSeq and 454 amplicon sequencing. The prefix of the sequence identifier indicates the sequencing techniques used. The inferred alleles had identical sequences identified with each method and all sequences were detected by both methods.

2.3.3 Repeatability of MiSeq and 454 genotypes

Among the 12 replicated amplicons, there were six and three ‘good' amplicons for the MiSeq and 454 data, respectively. There was a 100% match between the replicated MiSeq amplicons and a single discrepancy in the 454 data due to a missing allele in one individual with a high number of putative alleles (A_i = 9). The three replicated 454 amplicons had a 100% match with the MiSeq data when the consensus sequences of the 454 genotypes were used in the comparisons.

Among the 81 (MiSeq) and 80 (454) non-replicated amplicons, there were more ‘good’ amplicons in the MiSeq data than in the 454 data (69/81 *vs* 58/80). However, this was not significantly different between the techniques (χ^2 = 3.16, *d.f.* = 1, *P* = 0.075). Fifty-three individuals were successfully genotyped in both the MiSeq and 454 data, with a mean of 99% of genotypes matching across techniques. The 1% discrepancy was the result of a single allele detected in an individual’s MiSeq profile but not in the 454 amplicon.

Among the 11 families that were genotyped, seven and two families in the MiSeq and 454 data, respectively, had enough ‘good’ amplicons to enable successful genotyping of both parents and one or more chicks. There was 100% match between the *Mhc* genotypes of the parents and offspring in all these families for both the MiSeq and 454 data (MiSeq, 41 chicks; 454, 9 chicks). We confirmed that all alleles observed in the chicks were also detected in their parents, ratifying that no artefactual alleles were assigned as real.

2.3.4 Separating alleles from artefacts in amplicons with different numbers of putative alleles

Overall, there was a clear difference in the cumulative sequencing depth between putative alleles and artefacts after error correction for both the

MiSeq and 454 data (**Figure 2.7**). In the case of the MiSeq data, putative alleles occurred at relative sequencing depths between 3.3% and 19.1%, whereas artefacts were observed to have much lower depths (between 0.001% and 1.4%) (**Figure 2.7a**). In the 454 data, the putative alleles were observed at sequencing depths between 6.2–34.4% of total amplicon depth, whereas artefacts were observed at lower sequencing depths per amplicon (between 0.05%–3.0%) (**Figure 2.7b**). The difference between the cumulative sequencing depth of putative alleles and artefacts decreased as the number of alleles per amplicon increased in both MiSeq and 454 (**Figure 2.7**).

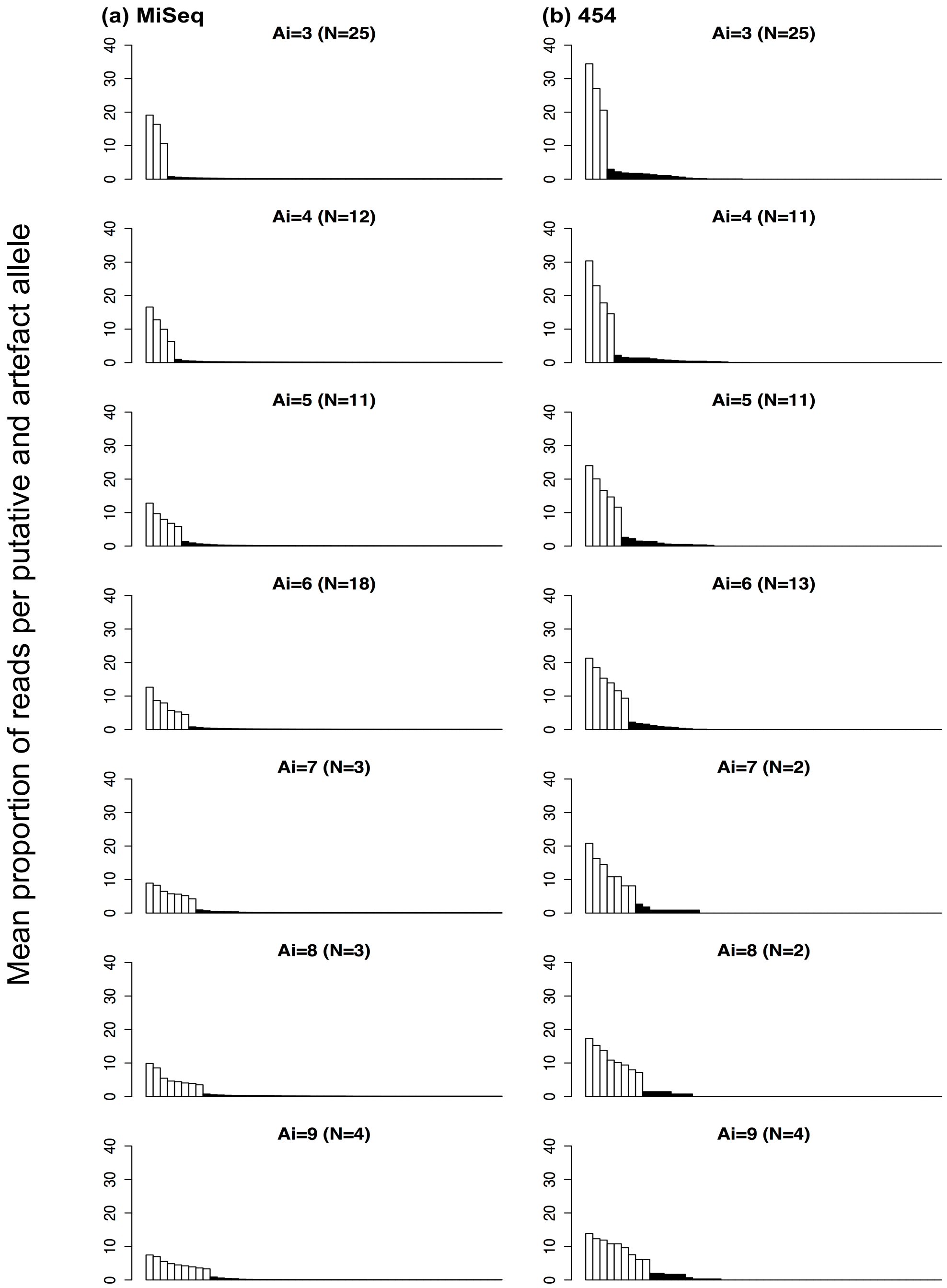


Figure 2.7 Mean sequencing depths of the first 50 variants in amplicons with different numbers of putative true alleles (Ai) for (a) the MiSeq and (b) the 454 data. The mean sequencing depth for each allelic level (*i.e.* putative alleles ordered by depth) was calculated as the total number of reads from all successfully genotyped amplicons per allelic level, divided by the total reads per amplicon. These calculations were performed separately on amplicons grouped by the number of putative alleles they possessed (*A*i = 3 to 9 alleles). Total numbers of amplicons: MiSeq – 76, 454 – 68. Grey bars show the sequencing depths of putative alleles, whereas black bars show the sequencing depths of artefacts.

2.3.5 Comparison of genotyping success given the number of putative alleles present

Table 2.1 Estimates of number of reads per amplicon (after error correction) that were retained as 'good' amplicons or discarded as 'poor' amplicons from MiSeq and 454.

	Mean (± *se*) reads per amplicon	
	MiSeq	**454**
Retained	2,819 (*62*)	63 (*2*)
Discarded	2,427 (*135*)	53 (*2*)

The 'poor' amplicons that were discarded from the dataset had lower read depths than the 'good' amplicons for both the MiSeq and the 454 data (Mann-Whitney *U*-test, Z_{MiSeq} = 2.82, $P < 0.005$, Z_{454} = 2.87, $P < 0.005$; **Table 2.1**). For both the MiSeq and the 454 data, the proportion of successfully genotyped amplicons was generally lower when the number of putative alleles was higher (**Figure 2.8**). It should be noted that the sample size decreased considerably for amplicons with more than six alleles.

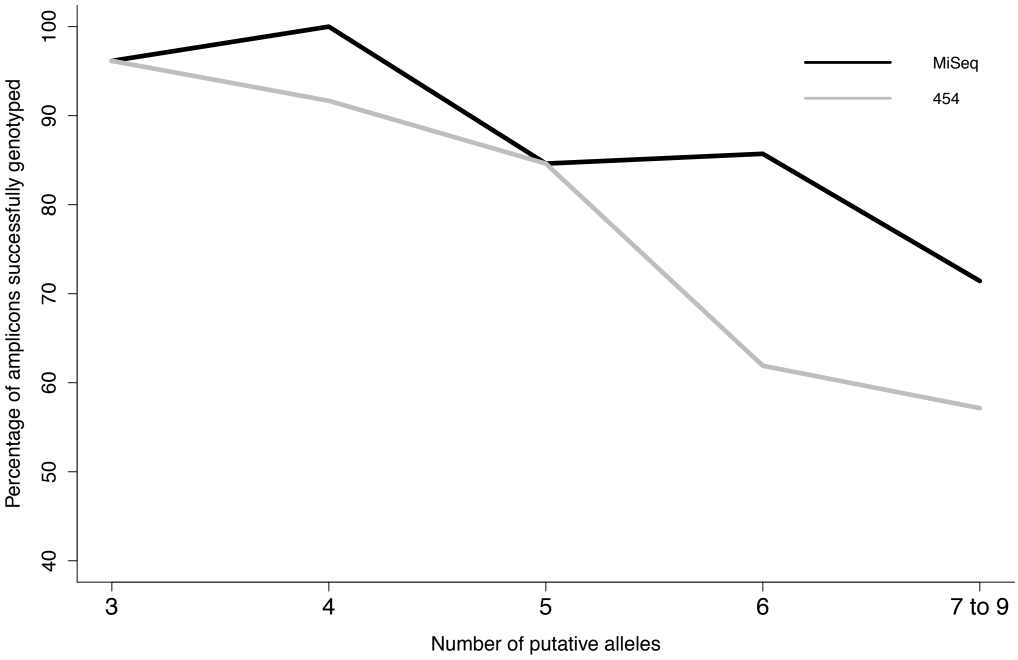


Figure 2.8 The percentages of amplicons that were successfully genotyped for each number of putative alleles, for each sequencing technique. The number of amplicons successfully genotyped are merged for numbers of alleles from 7 to 9.

2.4 Discussion

NGS based on amplicons has enabled a greatly improved resolution in measuring *Mhc* diversity within and between species. However, this resolution comes at a cost, since not only are true alleles, being amplified and sequenced but also artefactual variants (Babik *et al.* 2009; Babik 2010). In NGS studies of amplicons, there is a need to establish procedures that can be applied on a wide scale to identify all the true alleles, while effectively filtering out artefactual variants (Babik 2010). Moreover, the NGS technique must provide high coverage and the project must be designed so that all *Mhc* variants in an amplicon are accurately inferred, especially low-copy-number *Mhc* variants (Thomas *et al.* 2006).

Until very recently, 454 pyro-sequencing was the dominant NGS technique for genotyping *Mhc* diversity in non-model organisms (Zagalska-Neubauer *et al.* 2010; Radwan *et al.* 2012; Sepil *et al.* 2012b; Strandh *et al.* 2012; Karlsson & Westerdahl 2013; Oomen *et al.* 2013). However, in the last two years Illumina MiSeq has become the preferred method (Lighten *et al.* 2014). For a similar cost, MiSeq provides a much higher read depth per amplicon, *i.e.* a higher coverage, and is consequently now considered to be a more reliable method for genotyping *Mhc* in species with multiple *Mhc* genes. For a higher cost, 454 can provide similar read depth to MiSeq. In the present study, a qualitative and quantitative comparison of Illumina MiSeq and 454 amplicon sequencing was performed to discover if, and at what complexity, MiSeq outcompetes 454. The qualitative estimate investigated if complex genotypes could be correctly genotyped by comparing the two methods, and the quantitative estimate measured the proportion of successful amplicons obtained using each method.

Classical *Mhc* class I genes were used in this study, as they have been previously characterized in the house sparrow (Karlsson & Westerdahl 2013). Amplicons that had fewer than nine alleles were genotyped successfully and correctly with both MiSeq and 454 techniques. This suggests that the read depth per amplicon that we generated with 454 was adequate for amplicons, such as those in the house sparrow with moderate allelic diversity per individual, however, at higher allelic diversity, MiSeq might be more reliable. 454 tended to be quantitatively less successful throughout the comparison, *i.e.* a lower proportion of the amplicons were genotyped successfully and were considered ‘good amplicons’, though the difference was not significant.

Recent studies have clearly shown that new-generation sequencing of amplicons produces very large datasets and that processing these data in order to reduce the artefacts, *e.g.* mistakes related to base-reading errors, is a major challenge (Huse *et al.* 2007; Gomez-Alvarez *et al.* 2009; Quince *et al.* 2011). Artefactual sequences are unavoidable, and since they may originate early in the PCR process they are subsequently amplified and may be sequenced in high copy numbers, making them difficult to identify (Sommer *et al.* 2013). There are two general assumptions regarding read depths of putative alleles and artefactual variants: (1) true alleles will be more frequent than artefacts and (2) the artefacts originate from true alleles (Babik 2010; Galan *et al.* 2010; Zagalska-Neubauer *et al.* 2010). A range of different artefacts can occur, *e.g.* chimaeras, homopolymers and single base-pair mismatches. In the present study, we have used Lighten *et al.*’s (2014) DOC genotyping method, which assumes that the sequencing read depth of actual alleles is considerably higher than that of artefacts. The present study’s analyses showed that the DOC genotyping method accurately separated putative alleles from artefacts for both MiSeq and 454; the MiSeq had a slightly higher proportion of ´good´ amplicons than 454, though this difference was not significant. Amplicons that failed to be genotyped, *i.e.* where putative alleles cannot be distinguished from artefactual variants, may result from inadequate sequencing depth, poor DNA quality combined with PCR-carry-over contamination, or poorer amplification of some real alleles (Li & Stoneking 2012; Lighten *et al.* 2014). Eight amplicons out of 93 (92 for 454) were ‘poor’ in both the 454 and MiSeq datasets, and it is likely that these DNA samples were of poor quality prior to the NGS. Nine and 16 additional amplicons were classified as ‘poor’ during genotyping for MiSeq and 454, respectively. These poor quality sequences were associated with significantly lower read depth.

Some of the methodological problems faced in NGS are PCR biases and technical errors that occur before the actual sequencing step. PCR bias problems can be minimized by optimizing DNA extraction protocols (Tedersoo *et al.* 2010) and by reducing the number of PCR cycles to 20–25 (Kanagawa 2003; Medinger *et al.* 2010). It is well documented that PCR bias and artefact formation occur at a higher rate during the last few cycles of the reaction (Kanagawa 2003). Another artefact that can occur prior to the NGS is cross-contamination when setting up the PCR (Li & Stoneking 2012). Some amplicons included sequences that were similar to putative alleles in other samples, but were classified as artefacts due to low read-sequencing depth. This was probably due to cross-contamination between DNA samples, a common occurrence in large multiplexing studies (Gomez-Alvarez *et al.* 2009; Lighten *et al.* 2014). We used the same samples and repeated exactly the same PCR set-up prior to the MiSeq and 454 sequencing, and can therefore confirm that the final dataset did not include alleles due to contamination because both experiments identify the same 21 putative alleles. As mentioned above, a higher number of amplicons were classified as ‘poor’ amplicons in the 454 data than in the MiSeq data, and since the discrepancy in the success of the two methods increased with the number of alleles in the amplicons, read depth is probably the explanation.

Reads from next-generation sequencing approaches are short and also contain sequencing errors (Sims *et al.* 2014). The application of Illumina sequencing in this study produced hundreds to thousands of sequence reads per amplicon (Lighten *et al.* 2014). It is a challenge to separate artefacts generated during PCR and sequencing from true *Mhc* alleles. The method used to reliably genotype the *Mhc* depends on the number of reads produced and the assumption that real alleles have higher sequencing depths (Lighten *et al.* 2014). In our study, the differences in depth coverage between the two techniques meant that more amplicons, including a higher number of alleles, were successfully genotyped using Illumina compared to

454. This is consistent with the expectation that genotyping will be more successful at higher read depth. However, if the read quality is compromised, for example due to low DNA quality or cross-sample contamination, then this may lead to difficulties in separating true alleles from artefacts, despite having high read depth.

According to the birth-and-death model of evolution, multiple genes are found in the *Mhc* region as a result of repeated gene duplication and these genes are either maintained or deleted, or become non-functional, by deleterious mutation (Kasahara 1997; Nei *et al.* 1997). As an example, at least five haplotypes were found in human DRB class II, and the number of genes per haplotype varied between two and five (Trowsdale 1995). The *Mhc* loci in passerines are characterized by gene duplication, which resulted in variation in gene number, evidenced by studies on common yellowthroat (*Geothlypis trichas*; Bollmer *et al.* 2010), zebra finch (*Taeniopygia guttata*; Balakrishnan *et al.* 2010), New Zealand robin (*Petroica australis australis*; Miller & Lambert 2004), great reed warbler (*Acrocephalus arundinaceus*; Westerdahl *et al.* 2004) and house sparrow (*Passer domesticus*; Bonneaud *et al.* 2004b). Two processes explaining the high polymorphism found at the *Mhc* are (1) positive selection generated by nucleotide substitution at peptide-binding sites (the site is involved with pathogen-recognition) (Hughes & Nei 1988) and (2) occurrence of gene conversion that produces a random rearrangement of *Mhc* sequence between and within duplicated loci (Ohta 1991). These unique processes enhance the *Mhc* allelic diversity within a population and aid in recognizing a wide range of pathogens. The *Mhc* class I exon 3 in house sparrows encodes the protein-binding sites (PBS), the high variability of which is considered to be maintained by balancing selection (Bonneaud *et al.* 2004b). The diversity of the *Mhc* is thought to be maintained further by pathogen-driven selection, a co-evolutionary arms race between host and pathogens, and therefore predicted to result in a close association between specific *Mhc* alleles and resistance or susceptibility to infectious diseases (Bonneaud *et al.* 2006).

In conclusion, there was high agreement between the MiSeq and 454 methods for genotyping classical *Mhc* class I genes in house sparrows and, among the 53 amplicons that were genotyped with both techniques, the agreement was 99%. The small discrepancy was due to insufficient read number and therefore reduced coverage when *Mhc* allele diversity was high in an individual in the 454, which led to failure to detect the ninth allele. Our findings suggest that both MiSeq and 454 are reliable techniques for assessing *Mhc* genotypes when *Mhc* diversity is moderate, though MiSeq seemed to perform better at the highest diversity level. The amplicons that were successfully genotyped using both MiSeq and 454 techniques had higher read numbers than did failed amplicons; this study also suggests that reliable *Mhc* genotyping requires high read numbers.

2.5 References

Alcaide M, Liu M, Edwards SV (2013) Major histocompatibility complex class I evolution in songbirds: universal primers, rapid evolution and base compositional shifts in exon 3. *PeerJ*, **1**, e86.

Allen RL, Hogan L (2001) Non-Classical MHC Class I Molecules (MHC-Ib). *eLS*.

Anon (2013) Roche Shutting Down 454 Sequencing Business. *Retrieved from https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business*.

Aoyagi K, Dijkstra JM, Xia C, Denda I, Ototake M, Hashimoto K, Nakanishi T (2002) Classical MHC class I genes composed of highly divergent sequence lineages share a single locus in rainbow trout (*Oncorhynchus mykiss*). *Journal of Immunology*, **168**, 260–273.

Babik W (2010) Methods for MHC genotyping in non-model vertebrates. *Molecular Ecology Resources*, **10**, 237–251.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Balakrishnan CN, Ekblom R, Völker M, Westerdahl H, Godinez R, Kotkiewicz H, Burt DW, Graves T, Griffin DK, Warren WC, Edwards SV (2010) Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology*, **8**, 1–19.

Bollmer JL, Dunn PO, Whittingham LA, Wimpee C (2010) Extensive MHC class II B gene duplication in a passerine, the common yellowthroat (*Geothlypis trichas*). *Journal of Heredity*, **101**, 448–460.

Bolnick DI, Snowberg LK, Caporaso JG, Lauber C, Knight R, Stutz WE (2014) Major Histocompatibility Complex class IIb polymorphism influences gut microbiota composition and diversity. *Molecular Ecology*, **23**, 4831–4845.

Bonneaud C, Mazuc J, Chastel O, Westerdahl H, Sorci G (2004a) Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution*,

58, 2823–2830.

Bonneaud C, Pérez-Tris J, Federici P, Chastel O, Sorci G (2006) Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, **60**, 383–389.

Bonneaud C, Sorci G, Morin V, Westerdahl H, Zoorob R, Wittzell H (2004b) Diversity of Mhc class I and IIB genes in house sparrows (*Passer domesticus*). *Immunogenetics*, **55**, 855–865.

Borg ÅA, Pedersen SA, Jensen H, Westerdahl H (2011) Variation in MHC genotypes in two populations of house sparrow (*Passer domesticus*) with different population histories. *Ecology and Evolution*, **1**, 145–159.

Bracho MA, Moya A, Barrio E (1998) Contribution of Taq polymerase-induced errors to the estimation of RNA virus diversity. *The Journal of Veneral virology*, **79**, 2921–2928.

Bruford M, Hanotte O, Brookfield J, Burke T (1998) Multilocus and single-locus DNA fingerprinting. In: *Molecular Genetic Analysis of Populations: A Practical Approach* (ed Hoelzel A), pp. 287–336. IRL Press, Oxford.

Crawley MJ (2005) *Statistics: An Introduction using R*. John Wiley & Sons, Chichester.

Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: From genomics to sexual selection. *Trends in Ecology and Evolution*, **13**, 305–311.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson J-F (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. *ISME Journal*, **3**, 1314–1317.

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology*, **8**, 1–9.

Janeway C, Travers P, Walport M, Shlomchik M (2005) *Immunobiology: The*

Immune System in Health and Disease. Garland Science Publishing.

Kalbe M, Eizaguirre C, Dankert I, Reusch TBH, Sommerfeld RD, Wegner KM, Milinski M (2009) Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 925–934.

Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317–323.

Karlsson M, 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.

Kasahara M (1997) New insights into the genomic organization and origin of the major histocompatibility complex : Role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas*, **127**, 59–65.

Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Molecular Ecology*, **19**, 255–265.

Li M, Stoneking M (2012) A new approach for detecting low-level mutations in next-generation sequence data. *Genome Biology*, **13**, R34.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*, **14**, 1–15.

Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) *Plasmodium relictum* infection and MHC diversity in the house sparrow (*Passer domesticus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **278**, 1264–1272.

Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z, Dewell SB, Du L, Fierro JM, Gomes X V, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer MLI, Jarvie TP, Jirage KB, Kim J-B, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Medinger R, Nolte V, Pandey RAMV, Jost S (2010) Diversity in a hidden world : potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular Ecology*, **19**, 32–40.

Miller HC, Lambert DM (2004) Gene duplication and gene conversion in class II MHC genes of New Zealand robins (Petroicidae). *Immunogenetics*, **56**, 178–191.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Research*, **39**, e90–e90.

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 7799–7806.

Ohta T (1991) Role of Diversifying Selection and Gene Conversion in Evolution of Major Histocompatibility Complex Loci. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6716–6720.

Oomen RA, Gillett RM, Kyle CJ (2013) Comparison of 454 pyrosequencing methods for characterizing the major histocompatibility complex of nonmodel species and the advantages of ultra deep coverage. *Molecular Ecology Resources*, **13**, 103–116.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3714–3719.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing Noise From Pyrosequenced Amplicons. *BMC Bioinformatics*, **12**, 38.

R Development Core Team (2015) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.r-project.org/.

Radwan J, Zagalska-Neubauer M, Cichoń M, Sendecka J, Kulma K, Gustafsson L, Babik W (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology*, **21**, 2469–2479.

Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Research*.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schroeder J, Burke T, Mannarelli M-E, Dawson DA, Nakagawa S (2012) Maternal effects and heritability of annual productivity. *Journal of Evolutionary Biology*, **25**, 149–156.

Schut E, Aguilar JR, Merino S, Magrath MJL, Komdeur J, Westerdahl H (2011) Characterization of MHC-I in the blue tit (*Cyanistes caeruleus*) reveals low levels of genetic diversity and trans-population evolution across European populations. *Immunogenetics*, **63**, 531–542.

Sepil I, Lachish S, Sheldon BC (2012a) *Mhc*-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular Ecology*, **22**, 384–396.

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012b) 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**, 1–19.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, **15**, 121–32.

Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in zoology*, **2**, 16–34.

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, **14**, 1–17.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **277**, 979–988.

Strandh M, Westerdahl H, Pontarp M, Canbäck B, Dubois M-P, Miquel C, Taberlet P, 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 of London Series B: Biological Sciences*, **279**, 4457–4463.

Stuglik MT, Radwan J, Babik W (2011) jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, **11**, 739–742.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U (2010) 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist*, **188**, 291–301.

Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, Yuza Y, Garraway LA, LaFramboise T, Lee JC, Shah K, O'Neill K, Sasaki H, Lindeman N, Wong K-K, Borras AM, Gutmann EJ, Dragnev KH, DeBiasi R, Chen T-H, Glatt KA, Greulich H, Desany B, Lubeski CK, Brockman W, Alvarez P, Hutchison SK, Leamon JH, Ronan MT, Turenchalk GS, Egholm M, Sellers WR, Rothberg JM, Meyerson M (2006) Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Medicine*, **12**, 852–855.

Trowsdale J (1995) "Both man & bird & beast'': comparative organization of MHC genes. *Immunogenetics*, **41**, 1–17.

Wegner KM, Reusch TBH, Kalbe M (2003) Multiple infections drive major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology*, **16**, 224–232.

Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004) MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. *Heredity*, **92**, 534–42.

Zagalska-Neubauer M, Babik W, Stuglik M, Gustafsson L, Cichon M, Radwan J (2010) 454 sequencing reveals extreme complexity of the class II Major Histocompatibility Complex in the collared flycatcher. *BMC Evolutionary Biology*, **10**, 395.

CHAPTER 3: Malaria and its association with *Mhc* in the introduced house sparrow in New Zealand

Availability of sequence data

The sequence data generated and analysed during this study will be submitted to an online repository.

3.0 Abstract

In vertebrates, the most polymorphic loci known belong to the major histocompatibility complex (*Mhc*). The two hypotheses that are thought to explain the maintenance of genetic variation at the *Mhc* are pathogen-mediated selection (PMS) and *Mhc*-based mate choice selection. The mechanisms behind PMS, which involves the interaction between host and parasite, are heterozygote advantage, frequency-dependent selection and fluctuating selection. In this study, the presence of selection acting on *Mhc* class I exon 3 across spatially structured population was investigated. To test for the presence of selection, patterns of population differentiation at both *Mhc* and, presumably neutral, microsatellite loci were compared. There was significant spatial genetic differentiation at both the *Mhc* and microsatellite loci. The prevalence of malaria was higher in North Island than in South Island. There was no correlation between parasite diversity and genetic differentiation at the *Mhc*. In conclusion, this study provided evidence that, even at a small spatial scale, genetic differentiation at the *Mhc* is detectable, but this could not be attributed to parasite diversity.

3.1 Introduction

Mosquitoes belonging to the genus *Culex, Aedes*, *Culiseta* and *Anopheles* (Lapointe *et al.* 2012) are the known vectors of haemosporidian parasites. These mosquitoes are generalist blood feeders and are therefore likely to transfer the parasites to a wide range of birds. Haemosporidian parasites are apicomplexan protozoans that infect the blood of vertebrates (Pérez-Tris, *et al*. 2005). These parasites can be either specialist or generalist. A study on the evolutionary relationships of blood parasites revealed that several *Plasmodium* lineages exhibit a generalist host–parasite strategy, while other lineages are restricted to certain host families (Beadell *et al.* 2009).

The general life-cycle of a malarial parasite consist of two stages: the asexual and sexual reproductive stages (Valkiūnas 2005). When the parasite enters the vertebrate host via a bite from the mosquito vector, the sporozoites then undergo asexual divisions and mature into 'daughter' cells called merozoites (Aikawa 1977; Sherman 1979). The merozoites enter the red blood cells and continue to reproduce asexually, to increase in numbers and invade more red blood cells and liver (Manwell & Goldstein 1939). Some of merozoites differentiate into gametocytes. When a vector takes a blood meal from an infected individual, the male gametocytes will fertilize the female gametocytes in the gut of the mosquito. The resulting zygotes develop into oocytes, which release the sporozoites that are subsequently stored in the salivary glands of the mosquito. When the mosquito takes a blood meal, the sporozoites are inoculated into the bloodstream, thus continuing the life cycle of malaria (Sherman 1979; Valkiūnas 2005).

Haematozoan parasites have from little or no detectable effect on their hosts (Siikamaki *et al.* 1997; Hatchwell *et al.* 2000), to causing morbidity and mortality (Warner 1968; Atkinson *et al.* 1995, 2000, Palinauskas *et al.* 2008, 2009). Several studies have shown a heavy cost of being infected with malaria. For example, in an earlier experiment, iiwi (*Vestiaria coccinea*) were infected with *Plasmodium relictum* and it was found that this resulted in reduced food consumption, which led to a loss in body mass (Atkinson *et al.* 1995). In infected blue tits (*Parus caeruleus*), female parents reduced provisioning, a cost paid by offspring (Merino *et al.* 2000). However, with anti-malarial treatment the female parents had reduced parasitaemia, which was associated with increased fledging success (Merino *et al.* 2000). Another study on experimentally testing the effects of malaria parasites, such as *Plasmodium relictum*, on common avian hosts showed that the disease has different levels of pathogenicity in different avian host species (Palinauskas *et al.* 2008, 2009). It concluded that the malarial parasite (*Plasmodium relictum*) is pathogenic to European bird species (Palinauskas *et al.* 2008, 2009).

Passerines are hosts to a large number of haemosporidian parasites (Valkiūnas 2005). This class of malarial parasite is able to infect a broad range of hosts (Fallon *et al.* 2005; Hellgren *et al.* 2009). Infectious diseases such as avian malaria are therefore a considerable threat to endemic wildlife. For example, there was a devastating decimation of endemic avifauna in Hawai'i following exposure to novel parasites that were transmitted by an introduced mosquito species (*Culex pipens fatigan*) (Warner 1968). In New Zealand, a decreasing gradient in the prevalence of avian malarial infection from the North to the South Islands was associated with the distribution of an introduced vector (*Culex quinquefasciatus*) (Tompkins & Gleeson 2006).

Avian malaria provides a potential model for understanding the role of pathogens in influencing evolutionary processes in natural populations. Resistance genes, such as the major histocompatibility complex (*Mhc*), are good candidates for studying the interaction between infection and the evolution of a gene family. The *Mhc* is a multilocus family of genes and the recent development of next-generation sequencing is allowing easier identification of *Mhc* alleles. The *Mhc* genes encode surface proteins that bind peptides which aid immune cells such as the T-cells to distinguishing self peptides from antigenic peptides (Janeway *et al.* 2005). Activation of an immune response is triggered by displaying non-self peptides during antigen presentation (Janeway *et al.* 2005).

Mhc genes are highly polymorphic and theories have been proposed about how the genetic polymorphism is maintained, especially the pathogen-mediated selection (PMS) hypothesis (Doherty & Zinkernagel 1975). There are three main variants of the PMS hypothesis: heterozygote advantage (Doherty & Zinkernagel 1975), rare-allele advantage (Slade & McCallum 1992; Apanius *et al.* 1997) and fluctuating selection (Hedrick 1998, 2002).

The heterozygote advantage hypothesis states that being heterozygous at *Mhc* loci allows an individual to respond to wider range of pathogen peptides than if it were homozygous. The former will have higher relative fitness, leading to more *Mhc* alleles persisting in the population (Doherty & Zinkernagel 1975). Under rare-allele advantage, also called negative frequency-dependent selection (Slade & McCallum 1992; Apanius *et al.* 1997), it is hypothesized that pathogens are rapidly evolving, resulting in strong selection against common *Mhc* alleles that have lost their distinctiveness from pathogen antigens. Rare *Mhc* alleles will increase in frequency within populations because they will offer better protection against pathogens (Takahata & Nei 1990). The fluctuating or diversifying selection hypothesis states that variation in the abundance and diversity of parasites on both spatial and temporal scales maintains *Mhc* diversity (Hedrick 1998, 2002).

The heterogeneity in the distribution of avian malarial lineages exposes host bird species to spatially variable selection pressure (Olsson-Pons *et al.* 2015). Several studies have shown geographical variation in *Mhc* genes (Ekblom *et al.* 2007; Alcaide *et al.* 2008; Loiseau *et al.* 2009). These studies discovered that selection on the *Mhc* genes is weak, with evidence of balancing selection, or selection in a diversifying way as a result of heterogeneous distribution of parasite abundance and diversity (Ekblom *et al.* 2007; Alcaide *et al.* 2008; Loiseau *et al.* 2009). Significant associations were found between specific *Mhc* class I alleles and malaria in a house sparrow (*Passer domesticus*) study (Bonneaud *et al.* 2006; Loiseau *et al.* 2011).

The distribution of malaria is known to be heterogeneous in space, evidenced by differences in the prevalence of parasites and parasite diversity across populations (Loiseau *et al.* 2011; Bichet *et al.* 2015). Significant association were found between certain *Mhc* alleles and increased or decreased malaria infection in house sparrows (Bonneaud *et al.* 2006; Loiseau *et al.* 2011). The main aims of this chapter are: first, to investigate whether the pattern of differentiation at the *Mhc* class I exon 3 is more than expected based on neutral marker loci, i.e. microsatellite loci, and second, to investigate whether there is an association between malaria and variation in *Mhc* class I loci.

3.2 Methods

3.2.1 Study sites

A total of 400 house sparrow individuals was sampled across five latitudinally different populations in New Zealand (**Figure 3.1**; Auckland=100, Wellington=100, Christchurch=100, Dunedin=88, Fjordland=12) in 2005–2006. These blood samples were collected by Dr Shinichi Nakagawa and his team from the University of Otago. Birds were caught using mist nets, and 20–40 µl blood was taken from the brachial vein from each bird and stored in 95% ethanol. Genomic DNA was extracted using the salt ammonium acetate protocol (Bruford *et al.* 1998) and concentrations were standardized to 20–25 ng/µl.

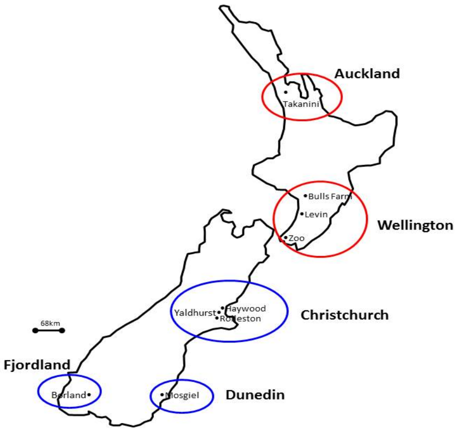


Figure 3.1 Sampling locations in New Zealand. Locations in the North Island are ringed in red, South Island in blue.

3.2.2 Malarial screening

Initial malaria screening was conducted using a polymerase chain reaction (PCR) gel-based detection method (see next paragraph for PCR details). The performance of each primer set was tested (**Table 3.1**). For every 48 blood samples, a negative control and positive control were included.

Positive controls were DNA from blackbird (*Turdus merula*) from New Zealand that was known to be infected. Ultraspure water was used instead of DNA in the negative control.

Table 3.1 Primers used to detect haemosporidian parasites.

Primer pairs (F/R)	Product sizes (bp)	Amplification success	Source of primer sequences
HaemNF1/HaemNR3 HaemF/HaemR2, HaemFL/HaemR2L	478 - 480	No	Hellgren *et al.* (2004)
213F/372R	160	No	Beadell & Fleischer (2005)
343F/496R	286	No	Fallon *et al.* (2003)
621F/983R	342	Yes	S. Fallon and R. Ricklefs, unpubl. data (cited in Richard *et al.* 2002)

F stands for forward primer
R stands for reverse primer

3.2.2.1 Polymerase chain reaction (PCR) gel-based detection method for avian malaria detection

PCR preparation was done according to a protocol for exrtracting ancient DNA to minimize DNA cross-contamination because of the high sensitivity of PCR-based malaria screening (Knapp *et al.* 2012). The total volume for PCR amplification was 25 µl, consisting of 1x PCR buffer, 3.0 mM $MgCl_2$, 400 µM of each dNTP, 0.2 µM of each primer, 0.625 units *Taq* polymerase (Thermoscientific) and 10–20 ng DNA. The initial PCR conditions consisted of heating the mixture for 90 s at 94°C, 40 s at 50°C and 70 s at 72°C, followed by 35 cycles of 30 s at 94°C, 40 s at 50°C and 70 s at 72°C, and a final elongation step of 30 s at 94°C, 40 sec at 50°C and 3 min at 72°C, performed in a Tetrad PTC-225 Thermal Cycler (MJ Research, Waltham, Massachusetts).

PCR products were run on 1.5% agarose, stained with SyberSafe (Invitrogen). Only one primer pair worked during initial malarial screening (621F/983R, **Table 3.1**). The presence of 342-bp PCR products on the gel indicated the presence of malaria parasites. Each sample was tested separately three times. Individuals were only considered infected when all three runs were found to be positive. The PCR products were sequenced and blasted against the NCBI database using a nucleotide BLAST (Basic Local Alignment Search Tool) to check for a match with malaria.

3.2.2.2 Illumina method for avian malaria detection

The forward primer, 621F 5’-AAAAATACCCTTCTATCCAAATCT-3’ and the reverse primer 983R 5’-CATCCAATCCATAATAAAGCAT-3’, amplified a 342-bp fragment of the *Plasmodium* cytochrome *b* (Richard *et al.* 2002). Each primer was uniquely modified at the 5$^{'}$ end with a 6-bp multiplex identifier (**Table 3.2**). Each sample within a batch was amplified using a unique combination of forward and reverse labelled primers (Galan *et al.* 2010; Kloch *et al.* 2010). Samples were amplified in 15-µl reactions consisting of QIAGEN Multiplex MasterMix (Qiagen), 10–20 ng DNA and 0.2 µM of each primer. PCRs were performed in a Tetrad PTC-225 Thermal Cycler (MJ Research, Waltham, Massachusetts) using the following settings: 95°C at 15 min, then 30 cycles of 95°C for 30 s, 50°C for 60 s and 72°C for 60 s. The final extension was at 72°C for 10 min.

Each set of 8 individuals was pooled and purified using the MinElute PCR purification kit (Qiagen), according to the manufacturer’s instructions. PCR amplicons were pooled and prepared for 250-bp paired-end Illumina MiSeq sequencing (Illumina, Inc., San Diego, CA, USA). All 400 samples were sequenced twice independently. Individuals were only identified as infected when found to be positive in both runs.

Table 3.2 List of unique 6-bp identifiers attached to the forward and reverse primers (5' end) for screening malaria and *Mhc* genotypes.

6-bp tags		
CCGGAA	CACAGT	ATATCG
AGTGTT	CAATCG	GAACCG
CCGCTG	CCGTCC	GTTAGT
AACGCG	AAGACA	CTGCCG
GGCTAC	GGTAAG	GCGAAC
TTCTCG	ATAATT	TACGGC
TCACTC	CGTCAC	TCGTTC
GAACTA	GCGCTA	CTCTCA

3.2.3 *Mhc* variant screening

The forward primers HNalla 5'-TCCCCACAGGTCTCCACAC-3' and the reverse primer rv3 5'-TGCGCTCCAGCTCCYTCTGCC-3' were used to amplify the *Mhc* class I exon 3 fragment size of 236–242-bp, which contained the most variable peptide binding region (Westerdahl *et al.* 2004). A 6-bp tag on each primer end was again used to distinguish individuals (**Table 3.2**). Polymerase chain reactions were performed in a total of 15 μl containing QIAGEN Multiplex MasterMix, 1 μl of DNA and 0.2 μM of each primer. PCRs were performed in a Tetrad PTC-225 Thermal Cycler (MJ Research, Waltham, Massachusetts) using the following settings: 95°C at 15 min, then 30 cycles of 95°C for 30s, 65°C for 60s and 72°C for 60s. The final extension was at 72°C for 10 min.

PCR products were run on a 1.5% agarose gel stained with SyberSafe (Invitrogen). Again, each set of 8 individuals was pooled and purified using a MinElute PCR purification kit (Qiagen) according to the manufacturer's instructions. PCR amplicons were pooled and prepared for 250-bp paired-end Illumina MiSeq sequencing (Illumina, Inc., San Diego, CA, USA).

3.2.4 Preprocessing sequence data for both malaria and *Mhc*

First, raw FASTQ paired-end sequence files were assembled based on 100-bp overlaps, using FLASH (Magoč & Salzberg 2011). Next, sequences with a Phred quality score below Q30 were removed with PRINSEQ (Schmieder & Edwards 2011). Finally, the primer and 6-bp tags were trimmed off the sequences and the data formatted into a table containing the read number of each sequence in each amplicon using jMHC (Stuglik *et al.* 2011).

A genotyping protocol was followed that made the assumption that real alleles are amplified at higher sequencing depth than artefacts (Lighten *et al.* 2014). First, all variants within each amplicon were checked for homopolymers and chimaeras. Then, a cumulative depth graph was plotted for each amplicon. The presence of an inflection point indicates that the amplicon is of a good quality because it enables separation between putative alleles and artefacts (refer to **Chapter 2**, **Figure 2.5**). Finally, the number of *Mhc* genotypes within an amplicon was estimated from the degree of change (DOC) value (refer to **Chapter 2,** section **2.2.3** for more detail, Macro sheet in Lighten *et al*. 2014).

The malarial sequences from this study were blasted against the NCBI database and sequences retrieved that had a 100% match. Both the new and retrieved sequences were aligned and a maximum-likelihood phylogenetic tree was plotted in MEGA 6.06 (Tamura *et al.* 2013). A mammalian malaria, *Plasmodium reichenowi* was used as an outgroup, to root the tree. The bootstrap value was calculated from 1000 iterations.

3.2.5 Microsatellite sequencing and genotyping

A total of 400 individuals was genotyped using thirteen microsatellite loci: *Ase18* (Richardson *et al.* 2000), *Pdo1* (Neuman & Wetton 1996), *Pdo3, Pdo5, Pdo6* (Griffith *et al.* 1999), *Pdo9, Pdo10* (Griffith *et al.* 2007), *Pdo16a*, *Pdo17, Pdo19, Pdo22,* and, *Pdo27, Pdo40* (Dawson *et al.* 2012). Two multiplex PCRs were performed, one including *Pdo1, Pdo3, Pdo5, Pdo6, Pdo9, Pdo10, Ase18* and the other for *Pdo19, Pdo17, Pdo16a, Pdo22, Pdo27, Pdo40* (see **Table 3.3** for primers concentrations).

Table 3.3 Concentrations of primers in each microsatellite multiplex mix

Multiplex set	Locus	Forward primer (nM)	Reverse primer (nM)
1	*Ase18*	300	300
	Pdo1	140	140
	Pdo3	300	300
	Pdo5	180	180
	Pdo6	300	300
	Pdo9	900	900
	Pdo10	100	100
2	*Pdo16a*	500	500
	Pdo17	450	450
	Pdo19	400	400
	Pdo22	500	500
	Pdo27	400	400
	Pdo40	400	400

Each multiplex was amplified in a final volume of 2 µl, including 20–25 ng of DNA, and 1 µl of QIAGEN Multiplex PCR mix and primers (see **Table 3.3**). The thermocycling program was set to initial denaturation at 95°C for 15 minutes, followed by 31 cycles of 94°C for 30 seconds, 57°C for 90 seconds, 72°C for 90 seconds and, finally, an extension phase at 72°C for 10 minutes. This was performed in a Tetrad PTC-225 Thermal Cycler (MJ Research, Waltham, Massachusetts). PCR products were diluted appropriately and separated on an ABI 3730 DNA Analyser. Alleles were scored using GENEMAPPER 3.0 (Applied Biosystems, Inc).

3.2.6 Intrapopulation genetic variability

3.2.6.1 Microsatellites

For microsatellites, in each population, the observed (H_O) and expected (H_E) heterozygosities were estimated using the program Genetix (Belkhir *et al.* 2004). Allelic richness was determined using the program Fstat 2.9.3 (Goudet 1995), which used a rarefaction index and was estimated for the smallest sample size (12 individuals from Fjordland). Deviations from Hardy–Weinberg equilibrium were tested for at each locus within each population using exact tests implemented in GenePop 4.0 (Rousset 2008). Linkage disequilibrium between pairs of loci within each population was also tested using exact tests implemented in GenePop 4.0 (Rousset 2008). Global tests for deviation from Hardy–Weinberg equilibrium across all loci were implemented for each population using Fisher's exact tests. To test whether any of the 13 microsatellite loci were under selection (Beaumont & Nichols 1996), we used the F_{st}-outlier detection method implemented in LOSITAN (Antao *et al.* 2008).

3.2.6.2 *Mhc*

For *Mhc*, the mean *Mhc* diversity in each population (*Mhc*/ind) was calculated as the mean number of *Mhc* alleles per individual. The index of allelic richness (*Theta K*) was estimated from the expected infinite-allele equilibrium given the observed number of alleles, the sample size and the population genetic parameter, Theta; *Theta K* was calculated using Arlequin 3.5 (Excoffier & Lischer 2010). The information from the *Mhc* data, such as the sequence name and the number of individuals having that sequence (*Mhc* allele), were entered as haplotypic data in Arlequin 3.5. The allele frequency was also calculated using Arlequin 3.5, as the number of individuals carrying a particular allele divided by the total allele count in the population. The total allele count is the sum of alleles found per individual in a population (Loiseau *et al.* 2011). Note that this way of calculating allele frequencies may result in the frequency of common alleles being underestimated and the frequency of rare alleles being overestimated (Ekblom *et al.* 2007).

3.2.7 Population differentiation

F_{st} was used to measure population differentiation and is defined as the proportion of genetic variation within a population relative to the genetic variation between populations (Hedrick 2011). F_{st} values range between 0 (not differentiated) and 1 (highly differentiated). Nine locations (subsampling locations within major locations (N = 5)) were used to calculate F_{st} values. Each pairwise F_{st} values were corrected for multiple-test using Bonferroni correction. We corrected the critical value by dividing with the total number of locations (0.05/9 = 0.006). Any *p*-value lower than 0.006 is considered as significant.

3.2.7.1 Microsatellite loci

Differentiation across all populations and between population pairs was tested using log-likelihood based G-tests (Goudet *et al.* 1996) in GenePop 4.2 (Rousset 2008). F_{st} was calculated using Arlequin 3.5 (Excoffier *et al.* 2005; Excoffier & Lischer 2010).

To infer population structure, Bayesian cluster analysis using the program STRUCTURE version 2.2 was used (Pritchard *et al.* 2000). A total of 10 independent runs was performed for each *K* (number of clusters) set in the range from 2 to 5, a burn-in of 3 x 10^4 iterations and 10^6 subsequent Markov Chain Monte Carlo Steps. Three models (no admixture, admixture and a combination of admixture and support with locations) were analysed. The value of *K* was evaluated using Δ*K* obtained from STRUCTURE HARVESTER version 0.6.8 (Earl & vonHoldt 2012). The results were further processed in CLUMPP (Jakobsson & Rosenberg 2007), which performed permutations for an optimal *q*-matrix output file, and finally plotted in R.

3.2.7.2 *Mhc*

For the *Mhc* class I marker, log-likelihood based G-tests in Arlequin 3.5 (Excoffier *et al.* 2005; Excoffier & Lischer 2010) were used to test for differentiation across all populations and between population pairs. F_{st} values were calculated with Arlequin 3.5 (Excoffier *et al.* 2005; Excoffier & Lischer 2010).

3.2.8 Statistical analysis

3.2.8.1 Repeatability and malaria detection

The repeatability of the gel-based method was estimated by running all 400 samples three times. The repeated measure of the Illumina-based sequencing method was obtained by running all samples twice. The ‘rptR’ package in R was used to measure repeatability for each technique (Nakagawa & Schielzeth 2010).

3.2.8.2 Prevalence of malaria infection

A χ^2 test was used to determine whether the two techniques (gel-based and MiSeq sequencing) consistently identified the same infected individuals. We also used a χ^2 test to determine whether the frequency of each malaria lineage differed between locations. A Fisher’s exact test was used to compare the prevalence of infection among the locations (Crawley 2005) in the “RVAideMemoire” package (Hervé 2016). The *p*-values were corrected for multiple-testing using Bonferroni error correction.

3.2.8.3 Genetic variation at *Mhc* and microsatellite.

We used the ‘kruskal.test’ function in R to compare *Mhc* diversity (i.e. *Mhc* alleles in an individual) across the different locations (Crawley 2005). This a non-parametric test comparing the medians of more than three independent groups (i.e. locations) (Fowler *et al.* 1998) .

Analysis of molecular variance (AMOVA) was used to determined the proportion of genetic variance partitioned among the different groupings using the program Arlequin 3.5 (Excoffier *et al.* 2005; Excoffier & Lischer 2010). This was to evaluate how much genetic variation contributes to the different hierarchical structures (among the islands, among the populations within islands and within populations). The *Mhc* data were entered as nucleotide sequences and the *Mhc* haplotype data as the numbers of individuals with a specific allele. The Euclidean squared distances were used in AMOVA to calculate the variance components and probabilities for the variation among groups, among populations within groups, and within populations based on Wright’s fixation indices (Excoffier *et al.* 2005; Excoffier & Lischer 2010). The significance of each level was tested using 10,000 permutations. This test was also used for the microsatellite dataset.

3.2.8.4 Isolation by distance test and malarial diversity index

Longitude and latitude were converted into a distance matrix with the function geodist in package ‘gmt’ (Magnusson 2014). Different packages in R were used for different analyses: Mantel test (ade4), partial Mantel test (vegan) and the Steinhaus similarity index (ecodist, 1 – value of Bray-Curtis index). For both Mantel and partial Mantel tests the number of permutations was set to 100,000 (Jackson & Somers 1989). The Mantel test (Mantel 1967) was used to assess the significance of association between patterns of differentiation (in terms of $F_{st}/(1- F_{st})$) for the different genetic markers (*Mhc* and microsatellite loci) (Rousset 1997) and geographical distance. The dependent variable consisted of F_{st} (*Mhc* or microsatellite) and the predictor variable was geographical distance, in each case consisting of a dissimilarity or distance matrix among the sampled locations. A partial Mantel test (Smouse *et al.* 1986) was used to assess the dependence of two distance matrices (F_{st} at *Mhc* and geographical distance among locations), while controlling for the effect of a third distance matrix (F_{st} at microsatellites). The Steinhaus similarity index quantifies the similarity of composition between two sites on a scale between 0 (dissimilar) and 1 (similar) (Magurran 2004; Kindt & Coe 2005).

3.2.8.5 Association between malaria and *Mhc*

The associations between the prevalence of malaria and the *Mhc* were tested using generalized linear models (GLM) with a quasi-binomial error distribution and a logit link function. The binomial variances were found to be over-dispersed and hence a quasi-binomial distribution was fitted into the GLM model. Twenty alleles were included in the analysis but the other thirty-four alleles were excluded because they had frequencies below 5%. The dependent variable is in binary form and consists of presence/absence of malaria infection. The explanatory variables comprised presence/absence of alleles (binary), sites (factor), allele x site interactions and *Mhc* diversity x site interactions. We compared the two models using a Chi-test in the “ANOVA” function. The explanatory variables in the first model consisted of alleles, sites, and all the interactions between site and allele, and between site and *Mhc* diversity. The second model consisted only of sites, alleles and no interactions added in this model.

The selected model was further simplified by removing variables in a reverse order of significance, sequentially removing non-significant terms. The significant effects of the variables were obtained from the analysis of deviance (i.e. *P*-values) generated from the “ANOVA” function. The *p*-values from the model were corrected for multiple-test error using a Bonferroni correction with the “p.adjust” function.

All statistical analysis and plots were constructed using R (R Development Core Team 2015) or as otherwise stated.

3.3 Results

3.3.1 Comparison between gel-based and sequencing methods for detecting avian malaria

A total of 170 individuals were found to be infected with malaria. Ten individuals were removed from the analysis because their infection status was not confirmed in both Illumina MiSeq runs. The number of individuals that were found to be positively infected in both MiSeq runs was 160 (40%); however, only 34 individuals (8.5%) were confirmed to be infected in all three gel-based assays.

Repeatability was higher in Illumina MiSeq (*R* = 0.872, CI: 0.768–0.892) than in gel-based (*R* = 0.352, CI: 0.271–0.43) detection. This showed that the sequencing technique had a higher success rate of detecting infected individuals than the gel-based method.

Thirty-two (20%) infected individuals were detected by both techniques. The observed prevalence of infection (N = 32) found in both techniques (gel-based and sequencing) exceeded random expectation (χ^2 = 7.87, *d.f.* = 1, $P < 0.05$). This meant that the majority of individuals found in the gel-based method (32 out of 34) were also found by the sequencing technique. Many more infected individuals were found only by sequencing (128 individuals, 79%) than by gel-based method (2 individuals, 1%).

3.3.2 Malarial prevalence and diversity detected by next-generation sequencing in each location

3.3.2.1 Prevalence of infected individuals in each location

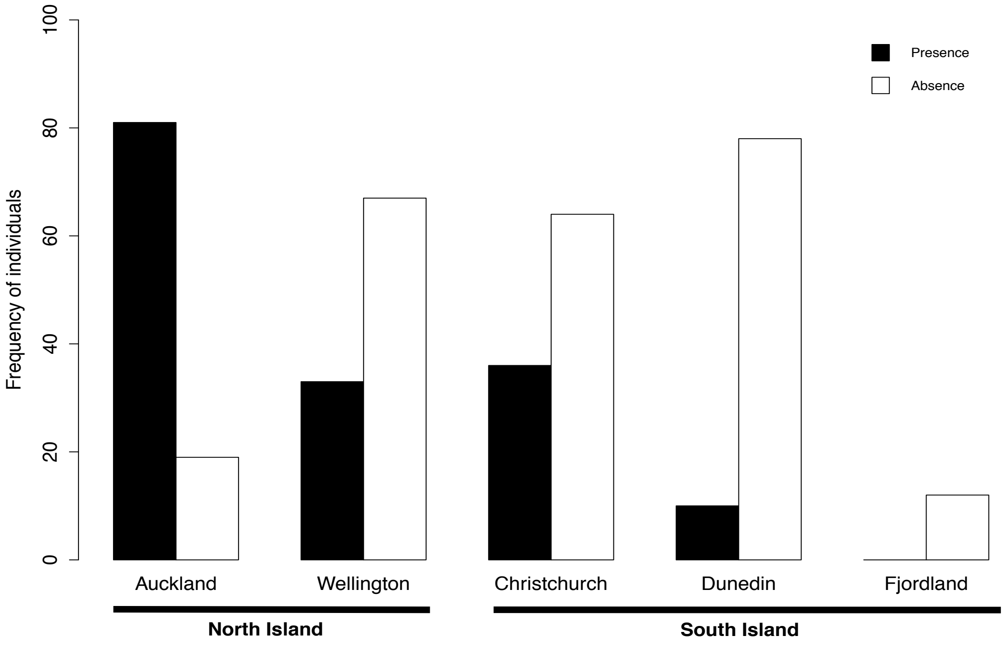


Figure 3.2 The frequency of individuals with or without malarial infection in each location.

The prevalence of infection varied significantly among locations (Fisher's exact test, *P* < 0.001). Auckland had the highest number of infected individuals (81/100, 81%, **Figure 3.2**). Prevalence of infection in Auckland (81/100, 81%) differed significantly from Wellington (33/100, 33%), Christchurch (36/100, 36%) and Dunedin (10/88) (**Table 3.4**). There was a gradual decrease in the proportion of infected individuals towards the southern latitudes, as seen in Wellington (33/100, 33%, **Figure 3.2**), Christchurch (36/100, 36%, **Figure 3.2**) and Dunedin (10/88, 11%, **Figure 3.2**), but the relationship with latitude could not be tested statistically. No infection was found in Fjordland (0/12, 0%, **Figure 3.2**). The frequency of infected individuals was not significantly different between Christchurch and

Wellington (**Table 3.4**). However, Wellington and Christchurch were each significantly different from Dunedin (**Table 3.4**).

Table 3.4 Pairwise Fisher's exact test for prevalence of infection between the different locations. *P*-values are corrected with Bonferroni error correction for multiple-test.

	North Island		South Island	
	Auckland	Wellington	Christchurch	Dunedin
Auckland	x			
Wellington	***P* < 0.001**	x		
Christchurch	***P* < 0.001**	*P* = 0.766	x	
Dunedin	***P* < 0.001**	***P* < 0.001**	***P* < 0.001**	x

Values in bold indicate significant difference in prevalence between locations.

3.3.2.2 Malaria diversity and prevalence

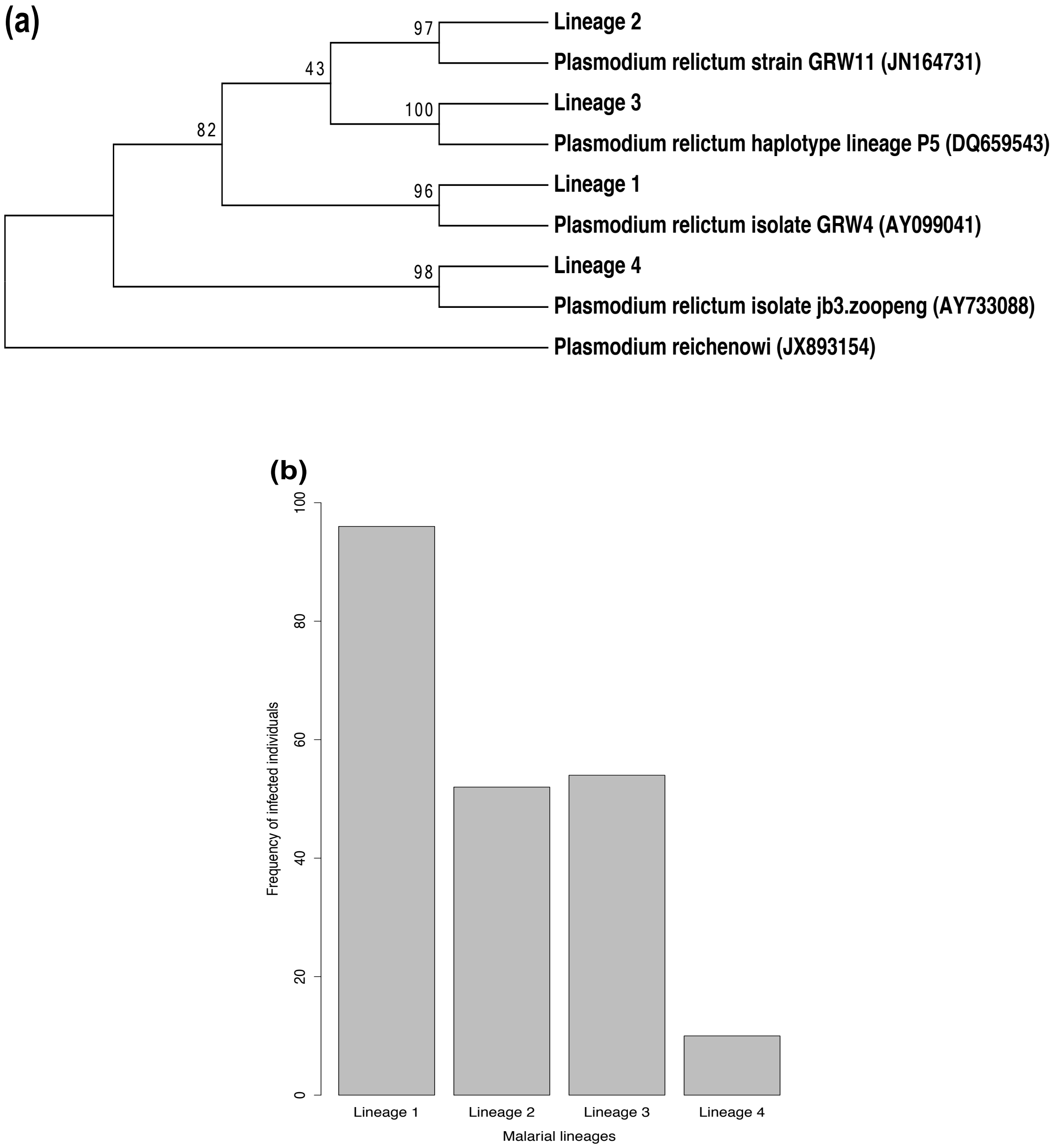


Figure 3.3 (a) Maximum-likelihood tree of the different malarial lineages in house sparrow. *Plasmodium reichenowi*, mammalian malaria, was selected as an outgroup. The bootstrap values (numbers at the nodes) illustrate how well the branches are supported. In parenthesis is the accession number from NCBI database. (b) The distribution of infected individuals carrying the different malarial strains. The total number of infected individuals was 160.

Four distinct haemosporidian parasite lineages were found (**Figure 3.3a**), belonging to the species, *Plasmodium relictum*. Different numbers of individuals were infected with each lineage (χ^2 = 69.8, *d.f.* = 3, *P* < 0.05). Lineage 1 (96) was the most common, followed by Lineages 3 (54) and 2 (52), with Lineage 4 (10) being the least commonly detected (**Figure 3.3b**).

3.3.2.3 Prevalence of each malarial lineage in each house sparrow population

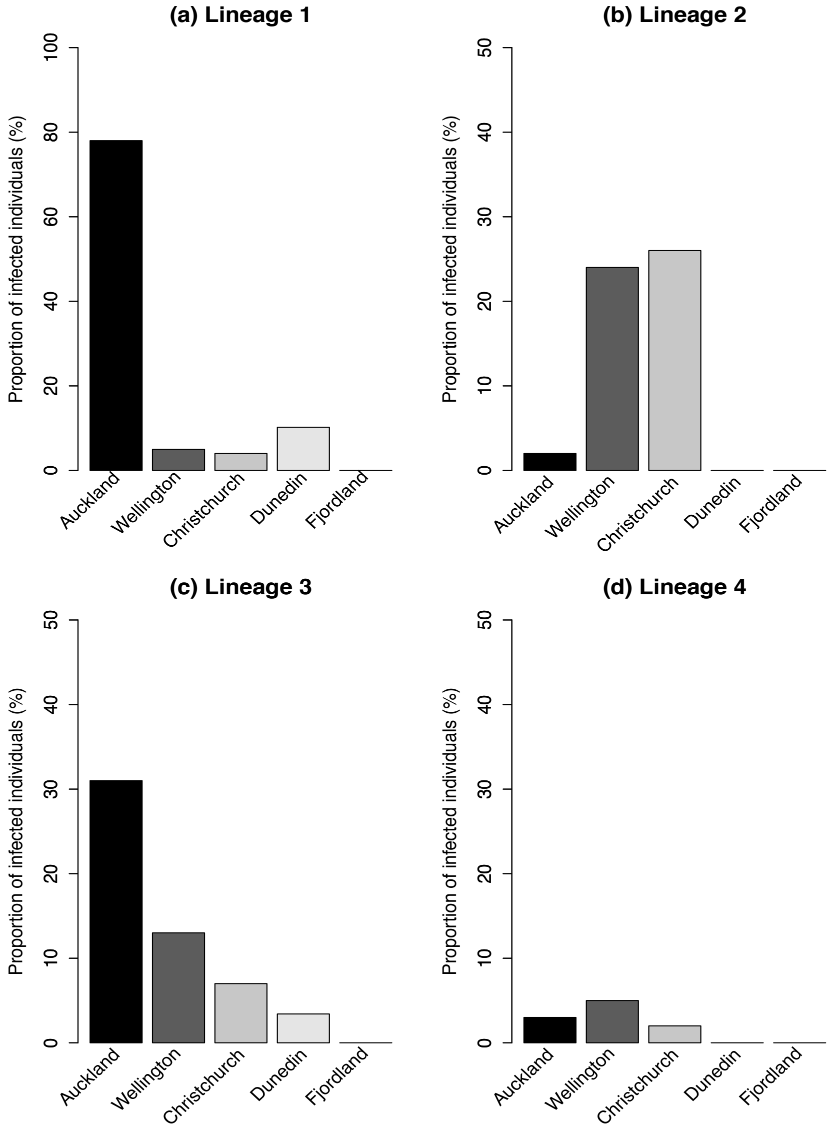


Figure 3.4 The proportion of infected individuals found in each location for each malarial strain. (a) Lineage 1 (b) Lineage 2 (c) Lineage 3 (d) Lineage 4. Note that some individuals were infected with more than one strain.

Among the 160 infected individuals, 70.6% (113) were infected once, 26.9% (43) were infected with two different lineages, 1.9% (3) were infected with three different lineages and only a single individual was found to be infected with all four lineages (0.6%).

The prevalence of the Lineage 1 infection differed between the five locations (χ^2 = 227.2, *d.f.* = 4, $P < 0.05$). Auckland had the highest infection frequency compared to the other locations (**Figure 3.4a**). This trend was similarly seen in Lineage 3 (**Figure 3.4c**). The prevalence of the Lineage 3 infection also differed between the locations (χ^2 = 56.0, *d.f.* = 4, $P < 0.05$). The prevalence of Lineage 2 also differed between the locations (χ^2 = 68.8, *d.f.* = 4, $P <$ 0.05); Wellington and Christchurch had the highest frequency of Lineage 2 infection (**Figure 3.4b**). Lineage 4 (**Figure 3.4d**) was not examined because the number of infected individuals is too small to allow the test.

3.3.3 Intrapopulation genetic diversity

3.3.3.1 *Mhc* diversity

A total of 301 individuals was screened for *Mhc* class I alleles. Fifty-four *Mhc* class I, exon 3 ‘long’ alleles were found among the five populations (**Table 3.5**). Ten alleles were found in all the locations, while nine alleles were specific to 4 locations (in parenthesis is the number of *Mhc* alleles specific to that location: Auckland (1), Wellington (4), Christchurch (1) and Dunedin (3)). Allele frequencies are reported in **Table 3.5**. The number of alleles per individual (*Mhc* diversity) differed among populations (Kruskal-Wallis test: *H* = 19.1, $P < 0.001$). This was due to the Dunedin population, which had fewer alleles per individual than the other populations (mean ± se, 3.95 ± 1.10, **Table 3.6**). When this population was removed, there was no significant difference in the number of alleles per individual among populations (Kruskal-Wallis: *H* = 5.24, *P* = 0.155). There were 2–9 *Mhc* alleles in an individual (**Figure 3.5**). This suggested that the number of loci amplified per individual in this study was between one and five.

Table 3.5 The frequencies of alleles in each population. Frequencies were calculated as the number of individuals carrying an allele divided by the sum of alleles per individual in each population. In parenthesis is the number of *Mhc*-genotyped individuals in each location. Alleles in bold were found in all five locations.

Sequence name	Auckland (72)	Wellington (67)	Christchurch (78)	Dunedin (73)	Fjordland(11)
nzmhcseq1	**0.115**	**0.121**	**0.196**	**0.191**	**0.204**
nzmhcseq2	0.084	0.062	0	0	0
nzmhcseq3	0.044	0.021	0.008	0.007	0
nzmhcseq4	0.016	0.018	0.041	0.069	0
nzmhcseq5	**0.044**	**0.015**	**0.024**	**0.017**	**0.037**
nzmhcseq6	0.037	0.012	0.033	0.021	0
nzmhcseq7	**0.047**	**0.024**	**0.022**	**0.024**	**0.037**
nzmhcseq8	0.016	0.003	0	0	0
nzmhcseq9	**0.006**	**0.006**	**0.016**	**0.007**	**0.019**
nzmhcseq10	0	0.006	0	0.01	0
nzmhcseq11	0	0.012	0.016	0	0.056
nzmhcseq12	**0.125**	**0.121**	**0.071**	**0.094**	**0.111**
nzmhcseq13	**0.016**	**0.012**	**0.019**	**0.021**	**0.037**
nzmhcseq14	0.016	0.012	0.014	0	0.037
nzmhcseq15	0	0	0.022	0.017	0
nzmhcseq16	0.016	0.024	0.008	0.017	0
nzmhcseq17	0	0.003	0.003	0.031	0.019
nzmhcseq18	0.003	0.003	0.003	0	0.019
nzmhcseq19	0	0.006	0.005	0.003	0
nzmhcseq20	0.006	0.003	0.022	0.014	0
nzmhcseq21	0	0.006	0.016	0.007	0.056
nzmhcseq22	0	0	0	0.01	0
nzmhcseq23	**0.103**	**0.112**	**0.071**	**0.09**	**0.074**
nzmhcseq24	0.012	0.003	0.03	0.035	0
nzmhcseq25	0.003	0.009	0.024	0.01	0
nzmhcseq26	0	0	0.024	0.007	0
nzmhcseq27	0.003	0.018	0.019	0	0
nzmhcseq28	0	0	0.008	0.017	0.056
nzmhcseq29	0.019	0.018	0	0	0
nzmhcseq31	0	0	0	0.003	0.056
nzmhcseq32	0.006	0.012	0	0	0
nzmhcseq33	0.006	0	0	0	0
nzmhcseq34	0.022	0.024	0.033	0.063	0
nzmhcseq35	0.019	0.012	0	0	0
nzmhcseq36	0.012	0.006	0	0	0
nzmhcseq37	0	0	0.014	0.007	0
nzmhcseq38	0.012	0.003	0.019	0.007	0
nzmhcseq39	0.016	0.003	0	0	0
nzmhcseq41	0.003	0.009	0.014	0.003	0
nzmhcseq43	**0.069**	**0.044**	**0.046**	**0.024**	**0.037**
nzmhcseq44	0.059	0.082	0.022	0.028	0
nzmhcseq45	0	0.024	0.043	0.052	0.037
nzmhcseq46	**0.028**	**0.053**	**0.022**	**0.028**	**0.019**
nzmhcseq47	**0.019**	**0.044**	**0.054**	**0.049**	**0.093**
nzmhcseq48	0	0.003	0.003	0.003	0
nzmhcseq49	0	0	0	0.003	0
nzmhcseq50	0	0	0.003	0.003	0
nzmhcseq51	0	0	0	0.003	0
nzmhcseq52	0	0.012	0.003	0	0
nzmhcseq53	0	0	0.011	0	0
nzmhcseq54	0	0.012	0	0	0
nzmhcseq55	0	0.009	0	0	0
nzmhcseq57	0	0.003	0	0	0
nzmhcseq58	0	0.003	0	0	0

Table 3.6 Genetic indices for each sampled location.

Location	*Mhc* class I				Microsatellites			
	N	*h*	*Theta k*	*Mhc/ind ± se*	*N*	*A*	H_O	H_E
Auckland	72	32	9.04	4.46 ± 1.14	100	7.98	0.81	0.80
Wellington	67	43	12.47	5.07 ± 1.22	100	8.38	0.81	0.81
Christchurch	78	37	9.79	4.72 ± 1.26	100	8.57	0.83	0.83
Dunedin	73	36	10.31	3.95 ± 1.10	88	8.69	0.84	0.84
Fjordland	12	18	11.38	4.91 ± 0.60	12	7.62	0.82	0.79

N indicates the sample size for each marker. *Mhc*: number of haplotypes (*h*), index of allelic richness (*Theta k*), the mean number of *Mhc* alleles per individual ± standard error (*Mhc*/ind ± se). Microsatellites: allelic richness (*A*), observed (H_O) and expected (H_E) heterozygosities.

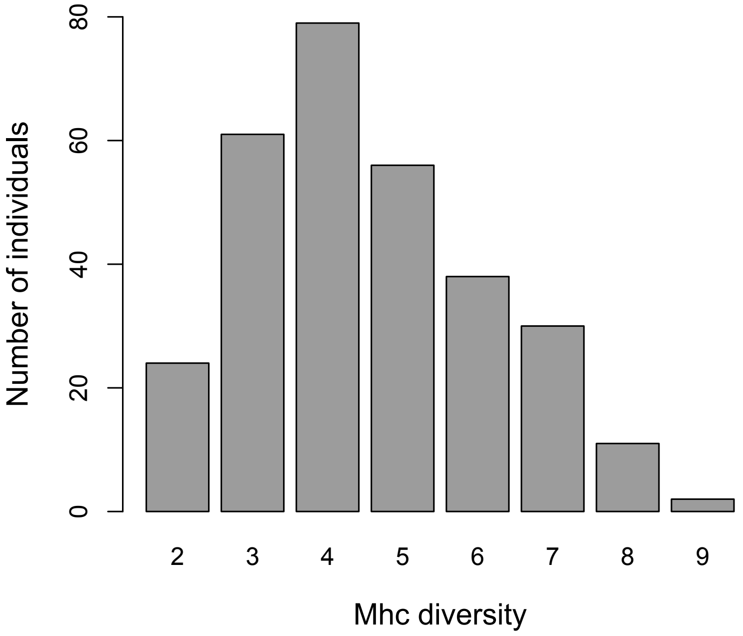


Figure 3.5 The distribution of the number of *Mhc* alleles (*Mhc* diversity) within an individual in all the population combined.

3.3.3.2 Microsatellite genetic diversity

A total of 254 alleles were found across all 5 populations combined and the number of microsatellite alleles per population varied between 3 and 93. Linkage disequilibrium was assessed in each population between all pairs of loci. Out of the 390 exact tests performed, 29 showed significant linkage disequilibrium at the *P* = 0.05 level (7.44%). After correcting for the False Discovery Rate (FDR), only three remained significant. These significant pairs of loci were each detected in just one location. In conclusion, the microsatellite loci can be considered to be statistically independent.

In each population, Hardy–Weinberg equilibrium (HWE) was tested at each locus. One of 65 population–locus tests was found to be significant at the 0.05 level. After correcting with FDR, it did not remain significant. However, one locus was found to be more differentiated than expected by chance in an outlier analysis (**Figure 3.6**); this locus, *Ase18*, was therefore removed from further analysis.

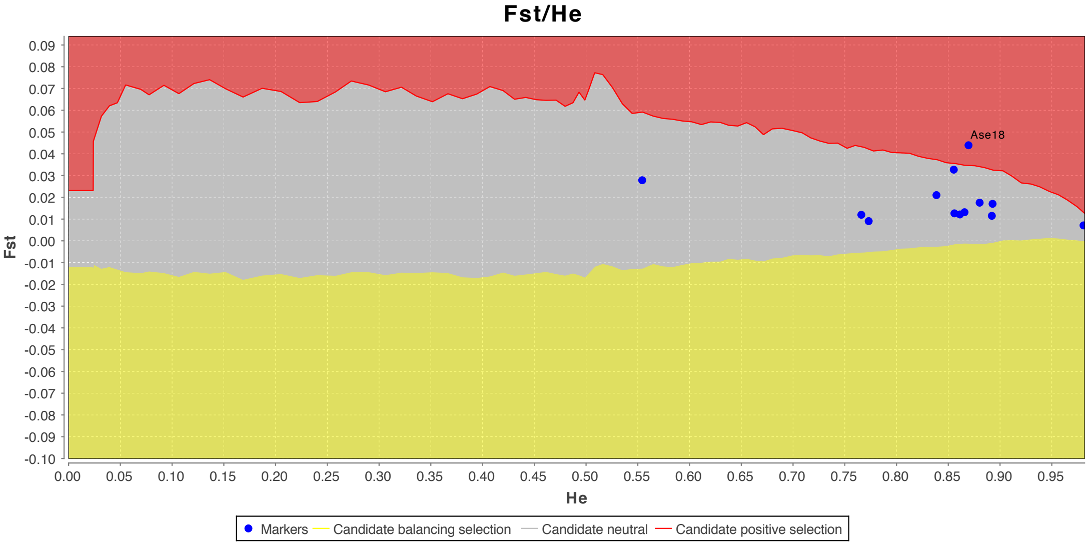


Figure 3.6 A graphic visualization of the distribution of the various microsatellite markers. The red and yellow shaded areas correspond to the 95% confidence level thresholds for neutral expectation. Ase18 showed higher differentiation than expected by chance and was thus removed from further analysis.

3.3.4 Population differentiation and isolation by distance

Most of the variance at the microsatellite loci was contributed by variation within samples (97.5%, **Table 3.7a**). Weak but significant genetic differentiation was also observed between islands (2.1% of variance) and among populations within islands (0.4% of variance) (**Table 3.7a**). Pairwise F_{st} values for microsatellite loci varied between –0.006 (between Rolleston and Borland) and 0.033 (between Levin and Rolleston) (**Table 3.8**). In total, 50% (18/36) of pairwise F_{st} values were significantly positive. There was a significant degree of isolation by distance at the microsatellite loci (Mantel test, r = 0.608, P = 0.003, **Figure 3.7a**). However, there was no significant isolation by distance within North Island (Mantel test, r = 0.791, P = 0.17) or within South Island (Mantel test, r = 0.028, P = 0.57).

Table 3.7 Results from the analysis of molecular variance (AMOVA) of microsatellites (a) and *Mhc* class I exon 3 (b) in house sparrows. Significant values are from 10,000 permutations. The islands consisted of North and South Islands.

(a)

Source of variation	*d.f.*	SS	Variance components	Percentage of variation	*P* - value
Among islands	1	49.75	0.11	2.08	< 0.05
Among populations within islands	7	44.35	0.02	0.39	< 0.05
Within populations	791	3888.94	4.92	97.53	

(b)

Source of variation	*d.f.*	SS	Variance components	Percentage of variation	*P* - value
Among islands	1	107.02	0.14	1.06	< 0.05
Among populations within islands	7	90.39	0.0004	0.00	> 0.05
Within populations	1362	17525.46	12.87	98.94	

Table 3.8 Genetic differentiation among populations (F_{st}) at microsatellite and *Mhc* Class I loci.

	North Island				South Island				
	Auckland	*Wellington*			*Christchurch*			*Dunedin*	*Fjordland*
	Takanini	Bulls	Levin	Zoo	Haywood	Yaldhurst	Rolleston	Mosgiel	Borland
Takanini	0	**0.005**	**0.011**	**0.009**	**0.027**	**0.026**	0.022	**0.029**	**0.022**
Bulls Farm	-0.002	0	0.003	0.001	0.019	**0.019**	0.016	**0.023**	**0.020**
Levin	-0.002	-0.002	0	-0.002	**0.031**	**0.029**	**0.033**	**0.030**	**0.032**
Zoo	0.005	0.005	-0.005	0	0.021	**0.018**	0.019	**0.020**	0.017
Haywood	0.006	0.009	-0.012	-0.021	0	-0.003	0.000	-0.002	0.002
Yaldhurst	**0.014**	**0.016**	0.006	0.005	-0.010	0	0.002	**0.003**	0.003
Rolleston	0.006	0.006	-0.006	-0.012	-0.021	-0.006	0	-0.001	-0.006
Mosgiel	**0.012**	**0.014**	0.000	-0.001	-0.012	0.000	-0.012	0	0.002
Borland	0.012	0.012	0.000	0.006	-0.003	0.010	-0.009	0.005	0

Top half matrix contains the F_{st} estimated with microsatellite loci. Bottom half matrix contains F_{st} estimated at the *Mhc* class I gene. F_{st} values in bold represent significant differentiation tests after Bonferroni correction.

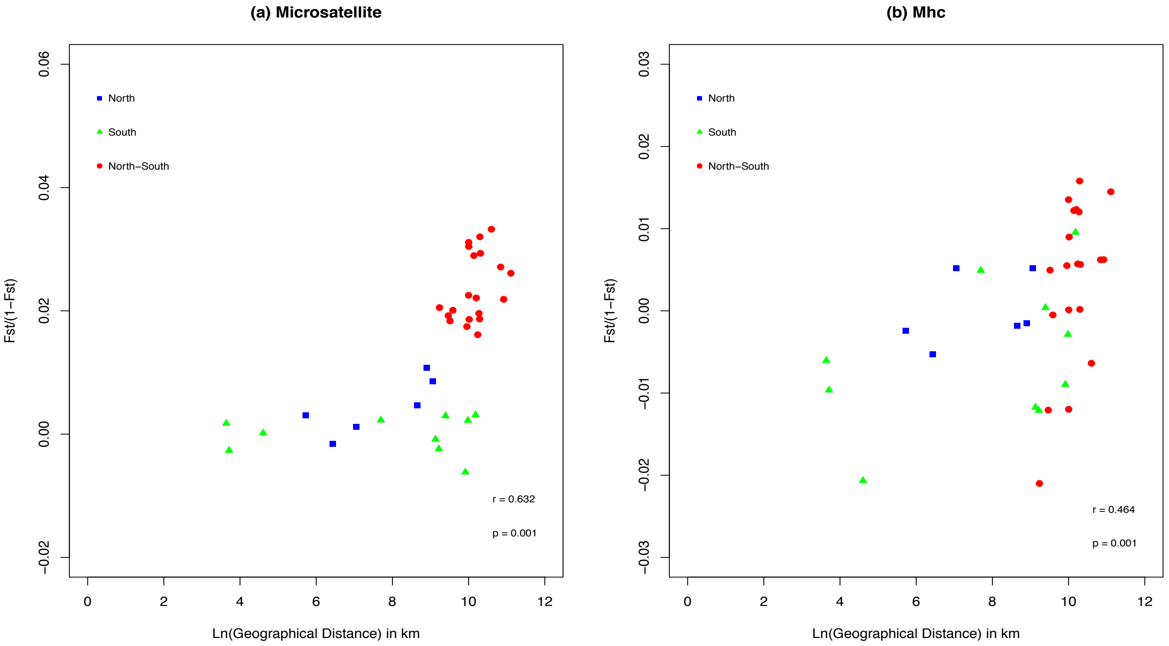


Figure 3.7 Isolation by distance analysis. Correlation between pairwise $F_{st}/(1- F_{st})$ and the natural logarithm of geographical distance for the 9 different populations: (a) for the 12 microsatellite markers, (b) for the *Mhc* class I exon 3 gene. Blue symbols indicate pairwise F_{st} estimated between locations within the North Island. Green symbols indicate pairwise F_{st} estimated between locations within the South

Island. Red symbols indicate pairwise F_{st} estimated between North Island and South Island locations.

The assignment analysis implemented in the STRUCTURE program indicates that there are two populations (**Figure 3.8**) based on microsatellite markers. The North Island, consisting of Auckland and Wellington, is distinct from the South Island consisting of Christchurch, Dunedin and Fjordland. Although the pairwise F_{st} showed a difference within North Island (**Table 3.8**), this differentiation was not detected by Structure (**Figure 3.8**).

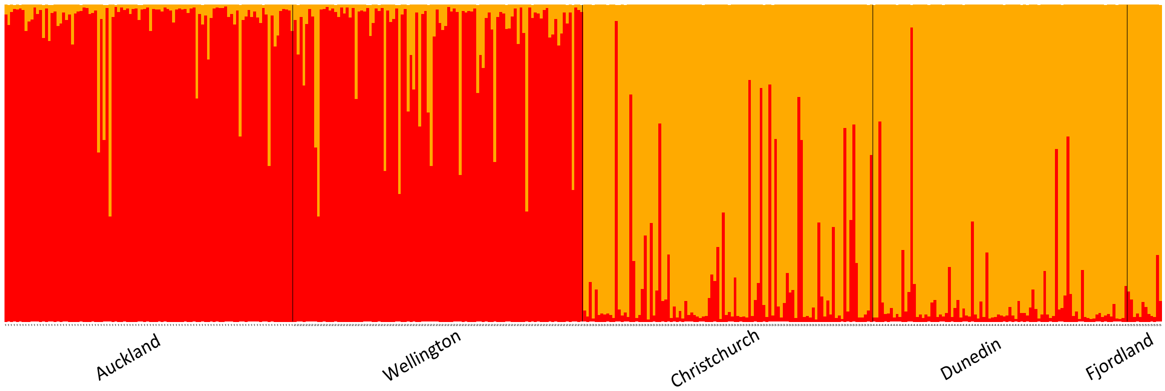


Figure 3.8 Population structure in New Zealand house sparrows. Genetic structure of house sparrow populations as defined by STRUCTURE (K=2 for all sampled house sparrow in all populations) using prior information of population origin. Each vertical bar represents a single individual. The height of each colour represents the probability of assignment to that cluster.

The majority of the genetic variation at the *Mhc* was also observed within populations (98.8%, **Table 3.7b**). There was weak but significant differentiation between islands (1.1% of variance) but not among populations within islands (**Table 3.7b**). Pairwise *Mhc* F_{st} values varied between -0.021 (between Haywood and Zoo) and 0.016 (between Yaldhurst and Bulls) (**Table 3.8**). Only 11% (4/36) of the pairwise F_{st} values were significant, and these were all between the North and South Islands comparison (**Table 3.8**). A similarly significant isolation by distance was also seen for the *Mhc* as for the microsatellites (Mantel test, $r = 0.464$, $P = 0.001$, **Figure 3.7b**). However, no significant isolation by distance was found within North Island (Mantel test, $r = 0.369$, $P = 0.294$) or South Island (Mantel test, $r = 0.427$, $P = 0.149$). *Mhc* pairwise F_{st} was positively correlated with microsatellite pairwise F_{st} (Mantel test, $r = 0.310$, $P = 0.039$). Taking the differentiation at microsatellites as a measure of isolation by distance under neutrality, we assessed *Mhc* divergence controlling for microsatellite F_{st}: significant differentiation at *Mhc* in relation to geographical distance remained (partial Mantel: $r = 0.327$, $P = 0.026$).

3.3.5 Population differentiation and malaria diversity

Four malaria lineages belonging to *Plasmodium relictum* were found in the infected locations (Auckland, Wellington, Christchurch and Dunedin). Some individuals were infected with more than one malaria lineage and, therefore, in order to quantify the selection pressure exerted by malaria diversity on local hosts, a Steinhaus similarity index was generated for each pair of infected populations to quantify malaria diversity (**Table 3.9**) There was no significance correlation between the parasite similarity index and geographical distance between populations ($r = -0.197$, $P = 0.84$). Also, there was no correlation between microsatellite or *Mhc* F_{st} and the parasite similarity index (microsatellite: $r = 0.155$, $P = 0.173$, *Mhc*: $r = -0.261$, $P = 0.86$).

Table 3.9 Similarity in malaria infections among locations. Pairwise Steinhaus similarity index between each pair of populations. Second row consist of the sites from which the blood samples were taken.

	North Island				South Island				
	Auckland	*Wellington*			*Christchurch*			*Dunedin*	*Fjordland*
	Takanini	Bulls	Levin	Zoo	Haywood	Yaldhurst	Rolleston	Mosgiel	Borland
Takanini	-								
Bulls Farm	0.242	-							
Levin	0.068	0.205	-						
Zoo	0.098	0.372	0.667	-					
Haywood	0.068	0.205	0.75	0.667	-				
Yaldhurst	0.176	0.783	0.211	0.381	0.211	-			
Rolleston	0.017	0.056	0.4	0.222	0.4	0.057	-		
Mosgiel	0.19	0.213	0.25	0.3	0.125	0.304	0	-	
Borland	0	0	0	0	0	0	0	0	-

Values close to 1 indicate that the malaria composition is similar between a pair of populations. Values close to or equal to 0 indicate that the malaria lineages are very different or that one population does not have any infection, as seen in Fjordland.

3.3.6 Associations between malaria and *Mhc* variation

Two models were compared to investigate the relationship between prevalence of infection and variation in the *Mhc*. The first model contained the variables: site, allele and all the interactions between site and each allele. The second model consists of the variables: sites and alleles. There was a significant difference between the two models (ANOVA: χ^2 = 88.7, *d.f.* = 66, *P* < 0.05). This suggests that the first model containing all the variables and interactions is a better model for further investigation. Season was not included in the analysis because most samples were taken during winter.

The model containing all the alleles and interactions was simplified by removing variables in the order of non-significance; the final model was fitted to a quasi-binomial error structure and defined as follows: Infection status ~ Sites + nzmhcseq12*Sites + nzmhcseq23*Sites + nzmhcseq34*Sites. Analysis of the model revealed that the probability of infection differed over sites. The significance remained the same even after it was corrected for multiple tests using Bonferroni error correction (deviance = 87.2, *d.f.* = 4, $P < 0.005$). The model also revealed that there was a significant overall effect of site x allele interactions for three alleles (nzmhcseq12, nzseqmhc23 and nzseqmhc34). However, the significance disappeared when it was adjusted using Bonferroni correction for multiple tests. This showed that there were no effects of specific alleles on the probability of infection across sites.

3.4 Discussion

This study found: (1) that Illumina next-generation sequencing was more efficient and repeatable than the gel-based assay method at detecting malaria infection. (2) As expected, malaria was more prevalent at higher and warmer latitudes (i.e. North Island), with North Island having a higher prevalence of malaria, while prevalence decreased as latitude decreased in South Island. (3) There was genetic differentiation in both microsatellite and *Mhc* markers. (4) No correlation between malaria parasite composition and population differentiation at the *Mhc* class I locus. (5) There was no association between prevalence of malaria and *Mhc* diversity and/or specific *Mhc* alleles.

Haemosporidian parasite identification was originally based on morphological traits using light microscopy (Valkiūnas 2005), then shifted more recently to the PCR-based identification technique, which is faster, cheaper and more sensitive in determining malarial prevalence (Jarvi *et al.* 2002; Richard *et al.* 2002; Hellgren *et al.* 2004) . In this study, Illumina-based sequencing had a higher detection rate than the gel-based technique. The sequences that were concluded to be real malarial sequences using our genotyping criteria (i.e. under the assumption that real sequences have higher read numbers compared to the artefacts) all matched perfectly with sequences that were already present in the NCBI database. These malaria strains have been found to infect various birds across Europe, Africa and have also been documented in New Zealand and Hawai’i (Perkins & Schall 2002; Bonneaud *et al.* 2006; Tompkins & Gleeson 2006; Beadell *et al.* 2006, 2009; Loiseau *et al.* 2012).

We showed that there were latitudinal differences in the prevalence of malaria between North Island and South Island. Auckland had the highest incidence of infection, followed by Wellington, Christchurch, Dunedin and, finally, Fjordland, where there was no detected infection, albeit in a small sample. An earlier study conducted by Tompkins & Gleeson (2006) also showed a strong pattern of decrease in the frequency of *Plasmodium relictum* infection from north to south in several bird species (E.g. *Turdus merula*, *Sturnus vulgaris* and *Passer domesticus*), and attributed this to the spread of its exotic mosquito vector (*Culex quinquefasciatus*). This exotic mosquito was originally found in Auckland, where an outbreak of avian malaria was documented in Auckland Zoo (Tompkins & Gleeson 2006). It was hypothesized that the colder temperatures in the south were limiting the abundance and transmission activity of mosquitoes (Patz & Reisen 2001; Bødker *et al.* 2003). The prevalence of avian malaria has been found to be elevated during spring, as expected given annual fluctuations in the populations of the vector and in the presence of immunologically naïve juveniles in the house sparrow host population, and to decline during winter (Applegate 1971). The decline in winter has been attributed to the decrease in vector activity and disappearance of malaria parasites from the blood (Cosgrove *et al.* 2008). However, a study on *Plasmodium circumflexen* in blue tits showed an increase in prevalence in autumn, which was the result of spring-infected individuals relapsing and a drop in prevalence over winter (Cosgrove *et al.* 2008). These authors also showed that there were seasonal patterns of different malaria morphospecies being prevalent in different age groups, indicating complexity in host–vector and vector–parasite interactions. In another study on house sparrows, the prevalence of *Plasmodium* infection was higher in the winter season; this was due to spring relapse of infected birds (Loiseau *et al.* 2011). While we detected seasonal variation in malaria prevalence, this might be a result of unequal sampling in different seasons (infected spring samples: 2/40, infected winter samples: 158/360; $\chi^2 = 21.1$, *d.f.* = 1, $P < 0.0005$) and, also, some sites not being sampled in both seasons (Auckland, Christchurch and Fjordland). In this study, an individual is scored as either infected or not infected (a measure of prevalence) and the question remained whether the scoring type has importance in understanding the impact of malarial parasites on host (Bentz *et al.* 2006). Measures of prevalence are affected by sample size, for example the accuracy of prevalence lowers with smaller sample size (Jovani & Tella 2006). There is a possibility that a small proportion of samples remained undetected, however, the advantage of this protocol is to reliably detect and distinguish malarial infection.

There were two distinct house sparrow populations, split between the North and South Islands (**Figure 3.8**). The house sparrow has a wide natural distribution, including the British Isles, all of Europe, across central Siberia, north Africa, across Arabia and across central Asia to south Asia and Myanmar (Long 1981). It was successfully introduced, deliberately or accidentally, into many additional locations including New Zealand (Long 1981). The latter introduction was during the 1860s, when the colonists were settling in an unfamiliar environment and acclimatisation societies were formed to enrich the local fauna with familiar species from Europe. House sparrows were among the many imported animals that caused destruction to the local fauna. The sparrows were imported and released into multiple locations in both North and South Islands of New Zealand in multiple years. They survived and spread from several locations and rapidly increased in number. By the 1930s, house sparrows were well distributed across both the North and South Islands of New Zealand (Long 1981). The differentiation of microsatellite markers between two populations may be a result of past history, reflecting different source populations and founding events, and there is no or little evidence of gene flow having occurred between the islands. A study on house sparrow populations in the Faroe Islands found that house sparrows are reluctant to cross water bodies (Bengtson *et al.* 2004; Magnussen & Jensen 2009).

Neutral markers such as microsatellites are used to infer population history or demographic events, such as genetic drift and gene flow. These markers are often preferred for this purpose because of their high variability, and they are generally considered to be unaffected by selection (Boyce *et al.* 1997; Ekblom *et al.* 2007). Gene flow and drift will affect all loci similarly, while selection is more likely to be locus-specific (Lewontin & Krakauer 1973); any difference in differentiation among the markers can therefore be taken as indicative of selection (Spitze 1993). When controlling for microsatellite variability, the *Mhc* in this house sparrow population still exhibited a significant pattern of differentiation. The genetic differentiation at the *Mhc* loci was lower than microsatellite markers. Bichet *et al*. (2015), also found weaker genetic differentiation at *Mhc* loci than microsatellite markers and suggested that balancing selection overrides divergent selection in the *Mhc* class I genes.

Comparison between the two marker classes potentially provides insight into the selection acting on the *Mhc* (Spurgin & Richardson 2010). For microsatellite loci, the assumption of neutrality is violated if the loci are linked to genes under selection (Larsson *et al.* 2007). All twelve microsatellite markers that were used in the analysis were not detectably under selection using the F_{st} outlier approach in LOSITAN (Antao *et al.* 2008), except for one marker that was found outside the 95% limits of a randomly generated F_{st} distribution, and so was removed from the analysis. Direct information on selective processes involving individuals and their environment cannot be provided by variation at neutral loci (Meyers & Bull 2002). However, adaptive processes within and between populations are expected to be reflected in the variability at the *Mhc* (Schwensow *et al.* 2007).

However, direct comparison of F_{st} between the two markers may not be appropriate because different classes of marker have different mutation rates, which may bias estimates of population differentiation, so that they may be affected differentially by evolutionary processes (Scribner *et al.* 1994). To date, many studies have made comparisons between neutral and functional markers, such as the *Mhc* (Boyce *et al.* 1997; Landry & Bernatchez 2001; Cohen 2002; Campos *et al.* 2006; Ekblom *et al.* 2007; Loiseau *et al.* 2011). For example, a study of estuarine fish, *Fundulus heteroclitus*, found no signal of selection at the neutral marker locus (hypervariable control region of mitochondria) during the separation of populations but there was a signal of selection seen at the *Mhc*. The evidence of selection on the *Mhc* locus came from elevated rates of amino-acid replacement in the protein-binding region (PBR) and population-specific effects on amino-acid substitution in the PBR (Cohen 2002). Detecting the selection acting on a population is a hit or miss kind of situation, according to

Garrigan & Hedrick (2002): a signal of selection will only be detectable in a generation that is experiencing a selection event and testing for this event depends on the sample size, the number of alleles and the selection coefficient (i.e. the measure of relative fitness of a phenotype in a population).

In this study, both *Mhc* class I and microsatellite markers were found to be positively correlated with distance, similar to the patterns reported by Loiseau *et al.* (2009) and Bichet *et al*. (2015). A partial Mantel test was used to estimate the correlation between population pairwise differentiation at *Mhc* and geographical distance, while controlling for differentiation at the microsatellite loci. This test, which was also used in the study of great snipes (*Gallinago media*), indicated that there was a pattern of *Mhc* differentiation due to isolation by distance (IBD), independent of the microsatellite IBD, meaning that the increasing differentiation at *Mhc* markers with distance was not attributable to neutral and/or demographic factors (Ekblom *et al.* 2007).

There are differences in malaria prevalence between the North Island and South Island house sparrow populations. At small spatial scales, selection pressure is presumed to be similar, a result of similar parasite abundance and diversity; thus selection should reduce between-population diversity (Bernatchez & Landry 2003). At larger geographical scales the population is highly likely to experience different parasite-selected pressure. Therefore, with increased geographical distance there is expected to be an increase in genetic differentiation of the parasites. However, this trend was not seen in this study. Previous studies also showed a correlation between *Plasmodium* infection and specific *Mhc* class I alleles in European house sparrow populations (Bonneaud *et al.* 2006; Loiseau *et al.* 2011), which resulted in local adaptation at *Mhc* loci. For example: Bonneaud *et al*. (2006) found that three *Mhc* class I alleles (a151, a172 and a161) had population-specific effects such as increased resistance or susceptibility to a single malarial strain (SGS1). Loiseau *et al*. (2011) also found population-specific effects of *Mhc* class I alleles (*pado83*, *pado109*, and *pado133*) on the risk of malarial infection. Such a pattern was not seen in this study. The lack of detectable selection (e.g. balancing selection, diversifying selection) at the *Mhc* may be due to weak selection, small sample sizes and high levels of heterozygosity (Garrigan *et al.* 2003). Compared to this study, which only included nine populations, Loiseau *et al*. (2009) and Bichet *et al*. (2015) had larger house sparrow populations, 13 and 12 populations, respectively, and spread over a larger geographical scale, making them more powerful.

In conclusion, four malaria strains were found to be present in the population. Even though there was no evidence of a relationship between parasite diversity and population differentiation at *Mhc* class I, there was significant population differentiation at the markers (*Mhc* and microsatellites) between the North and South islands. This may be largely due to the dataset containing two distinct populations of house sparrow. The inability to detect any signal of selection in this study may have been due to the relatively small sample size. More extensive sampling separately in each of the North and South Islands would therefore be desirable to test for finer-scale geographical effects.

3.5 References

Aikawa M (1977) Variations in structure and function during the life cycle of malarial parasites. *Bulletin of the World Health Organization*, **55**, 137–156.

Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). *Molecular Ecology*, **17**, 2652–2665.

Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: A workbench to detect molecular adaptation based on a Fst-outlier method. *BMC Bioinformatics*, **9**, 1–5.

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology*, **17**, 179–224.

Applegate JE (1971) Spring relapse of *Plasmodium relictum* infections in an experimental field population of English sparrows (*Passer domesticus*). *Journal of Wildlife Diseases*, **7**, 37–42.

Atkinson CT, Dusek RJ, Woods KL, Iko WM (2000) Pathogenicity of avian malaria in experimentally-infected Hawaii Amakihi. *Journal of Wildlife Diseases*, **36**, 197–204.

Atkinson CT, Woods KL, Dusek RJ, Sileo LS, Iko WM (1995) Wildlife disease and conservation in Hawaii: Pathogenicity of avian malaria (*Plasmodium relictum*) in experimentally infected Iiwi (*Vestiaria coccinea*). *Parasitology*, **111**, S59–S69.

Beadell JS, Covas R, Gebhard C, Ishtiaq F, Melo M, Schmidt BK, Perkins SL, Graves GR, Fleischer RC (2009) Host associations and evolutionary relationships of avian blood parasites from West Africa. *International Journal for Parasitology*, **39**, 257–266.

Beadell JS, Ishtiaq F, Covas R, Melo M, Warren BH, Atkinson CT, Bensch S, Graves GR, Jhala Y V, Peirce MA, Rahmani AR, Fonseca DM, Fleischer RC (2006) Global phylogeographic limits of Hawaii's avian malaria. *Proceedings of the Royal Society of London Series B:*

Biological Sciences, **273**, 2935–2944.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 1619–1626.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (1996) GENETIX 4.05, Logiciel sous Windows TM pour la génétique des populations. *Laboratoire Génome, Populations, Interactions*, **5000**, 1996–2000.

Bengtson S-A, Eliasen K, Jacobsen LM, Magnussen E (2004) A history of colonization and current status of the house sparrow (*Passer domesticus*) in the Faroe Islands. *Fróðskaparrit*, **51**, 237–251.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bichet C, Moodley Y, Penn DJ, Sorci G, Garnier S (2015) Genetic structure in insular and mainland populations of house sparrows (*Passer domesticus*) and their hemosporidian parasites. *Ecology and Evolution*, **5**, 1639–1652.

Bødker R, Akida J, Shayo D, Kisinza W, Msangeni HA, Pedersen EM, Lindsay SW (2003) Relationship Between Altitude and Intensity of Malaria Transmission in the Usambara Mountains, Tanzania. *Journal of Medical Entomology*, **40**, 706–717.

Bonneaud C, Pérez-Tris J, Federici P, Chastel O, Sorci G (2006) Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, **60**, 383–389.

Boyce WM, Hedrick PW, Muggli-Cockett NE, Kalinowski S, Penedo MCT, Ramey RR (1997) Genetic Variation of Major Histocompatibility Complex and Microsatellite Loci: A Comparison in Bighorn Sheep. *Genetics*, **145**, 421–433.

Bruford M, Hanotte O, Brookfield J, Burke T (1998) Multilocus and single-locus DNA fingerprinting. In: *Molecular Genetic Analysis of Populations: A Practical Approach* (ed Hoelzel A), pp. 287–336. IRL Press, Oxford.

Campos JL, Posada D, Moran P (2006) Genetic variation at MHC, mitochondrial and microsatellite loci in isolated populations of Brown trout (*Salmo trutta*). *Conservation Genetics*, **7**, 515–530.

Cohen S (2002) Strong positive selection and habitat-specific amino acid substitution patterns in MHC from an estuarine fish under intense pollution stress. *Molecular Biology and Evolution*, **19**, 1870–1880.

Cosgrove CL, Wood MJ, Day KP, Sheldon BC (2008) Seasonal variation in *Plasmodium* prevalence in a population of blue tits *Cyanistes caeruleus*. *Journal of Animal Ecology*, **77**, 540–548.

Crawley MJ (2005) *Statistics: An Introduction using R*. John Wiley & Sons, Chichester.

Dawson DA, Horsburgh GJ, Krupa AP, Stewart IRK, Skjelseth S, Jensen H, Ball AD, Spurgin LG, Mannarelli M-E, Nakagawa S, Schroeder J, Vangestel C, Hinten GN, Burke T (2012) Microsatellite resources for Passeridae species: a predicted microsatellite map of the house sparrow *Passer domesticus*. *Molecular Ecology Resources*, **12**, 501–523.

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, **256**, 50–52.

Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Ekblom R, Sæther SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kålas JA, Höglund J (2007) Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology*, **16**, 1439–1451.

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics*, **1**, 47–50.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Fallon SM, Bermingham E, Ricklefs RE (2005) Host specialization and geographic localization of avian malaria parasites: a regional analysis in the Lesser Antilles. *The American naturalist*, **165**, 466–480.

Fowler J, Cohen L, Jarvis P (1998) *Practical Statistics for Field Biology*. John Wiley & Sons, Chichester.

Galan M, Guivier E, Caraux G, Charbonnel N, Cosson J-F (2010) A 454 multiplex sequencing method for rapid and reliable genotyping of highly polymorphic genes in large-scale studies. *BMC Genomics*, **11**, 296.

Garrigan D, Hedrick PW, Mitton J (2003) Perspective: Detecting adaptive molecular polymorphism: Lessons from the MHC. *Evolution*, **57**, 1707–1722.

Goudet J (1995) FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. *Journal of Heredity*, **86**, 485–486.

Goudet J, Raymond M, de Meeüs T, Rousset F (1996) Testing Differentiation in Diploid Populations. *Genetics*, **144**, 1933–1940.

Griffith SC, Dawson DA, Jensen H, Ockendon N, Greig C, Neumann K, Burke T (2007) Fourteen polymorphic microsatellite loci characterized in the house sparrow *Passer domesticus* (Passeridae, Aves). *Molecular Ecology Notes*, **7**, 333–336.

Griffith SC, Stewart IRK, Dawson DA, Owens IPF, Burke T (1999) Contrasting levels of extra-pair paternity in mainland and island populations of the house sparrow (*Passer domesticus*): is there an “island effect”? *Biological Journal of the Linnean Society*, **68**, 303–316.

Hatchwell BJ, Wood MJ, Anwar M, Perrins CM (2000) The prevalence and ecology of the haematozoan parasites of European blackbirds, *Turdus merula*. *Canadian Journal of Zoology*, **78**, 684–687.

Hedrick P (1998) Balancing selection and MHC. *Genetica*, **104**, 207–214.

Hedrick PW (2002) Pathogen Resistance and Genetic Variation at MHC Loci. *Evolution*, **56**, 1902–1908.

Hedrick PW (2011) *Genetics of Populations*. Jones & Bartlett Publishers, MA.

Hellgren O, Pérez-Tris J, Bensch S (2009) A Jack-of-All-Trades and Still a Master of Some: Prevalence and Host Range in Avian Malaria and Related Blood Parasites. *Ecology*, **90**, 2840–2849.

Hellgren O, Waldenström J, Bensch S (2004) A new PCR assay for simultaneous studies of *Leucocytozoon*, *Plasmodium*, and *Haemoproteus* from avian blood. *The Journal of parasitology*, **90**, 797–802.

Hervé M (2016) RVAideMemoire: Diverse Basic Statistical and Graphical Functions. *R package 0.9-57*.

Jackson DA, Somers KM (1989) Are probability estimates from the permutation model of Mantel's test stable? *Canadian Journal of Zoology*, **67**, 766–769.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Janeway C, Travers P, Walport M, Shlomchik M (2005) *Immunobiology: The Immune System in Health and Disease*. Garland Science Publishing.

Jarvi SI, Schultz JJ, Atkinson CT (2002) PCR Diagnostics Underestimate the Prevalence of Avian Malaria (*Plasmodium relictum*) in Experimentally-Infected Passerines. *The Journal of Parasitology*, **88**, 153–158.

Kindt R, Coe R (2005) *Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies*. World Agroforestry Centre (ICRAF), Nairobi.

Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the load of gut parasites in the bank vole *Myodes glareolus*. *Molecular Ecology*, **19**, 255–265.

Knapp M, Clarke AC, Horsburgh KA, Matisoo-Smith EA (2012) Setting the stage – Building and working in an ancient DNA laboratory. *Annals of Anatomy*, **194**, 3–6.

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **10**, 2525–2539.

Lapointe DA, Atkinson CT, Samuel MD (2012) Ecology and conservation biology of avian malaria. *Annals of the New York Academy of Sciences*, **1249**, 211–226.

Larsson LC, Laikre L, Palm S, André C, Carvalho GR, Ryman N (2007) Concordance of allozyme and microsatellite differentiation in a marine fish, but evidence of selection at a microsatellite locus. *Molecular Ecology*, **16**, 1135–1147.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*, **14**, 1–15.

Loiseau C, Harrigan RJ, Robert A, Bowie RCK, Thomassen HA, Smith TB, Sehgal RNM (2012) Host and habitat specialization of avian malaria in Africa. *Molecular Ecology*, **21**, 431–441.

Loiseau C, Richard M, Garnier S, Chastel O, Julliard R, Zoorob R, Sorci G (2009) Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology*, **18**, 1331–1340.

Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) *Plasmodium relictum* infection and MHC diversity in the house sparrow (*Passer domesticus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **278**, 1264–1272.

Long JL (1981) *Introduced birds of the world: the worldwide history, distribution and influence of birds introduced to new environments*. David & Charles, London.

Magnussen E, Jensen J-K (2009) Ringing recoveries of house sparrow (*Passer domesticus*) in the Faroe Islands during the years 1963-2007. *Fróðskaparrit*, **57**, 182–189.

Magnusson A (2014) gmt: Interface between GMT Map-Making Software and R. *R package 1.2-0*, http://cran.r-project.org/web/packages/gmt.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Magurran AE (2004) *Measuring Biological Diversity*. Blackwell Science, Oxford.

Mantel N (1967) The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, **27**, 209–220.

Manwell RD, Goldstein F (1939) The asexual life cycle of the avian malaria parasite, *Plasmodium circumflexum*. *Science*, **89**, 131–132.

Merino S, Moreno J, Sanz JJ, Arriero E (2000) Are avian blood parasites pathogenic in the wild? A medication experiment in blue tits (*Parus caeruleus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **267**, 2507–2510.

Meyers LA, Bull JJ (2002) Fighting change with change: adaptive variation in an uncertain world. *Trends in Ecology & Evolution*, **17**, 551–557.

Nakagawa S, Schielzeth H (2010) Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, **85**, 935–956.

Neuman K, Wetton JH (1996) Highly polymorphic microsatellites in the house sparrow *Passer domesticus*. *Molecular Ecology*, **5**, 307–309.

Olsson-Pons S, Clark NJ, Ishtiaq F, Clegg SM (2015) Differences in host species relationships and biogeographic influences produce contrasting patterns of prevalence, community composition and genetic structure in two genera of avian malaria parasites in southern Melanesia. *Journal of Animal Ecology*, **84**, 985–998.

Palinauskas V, Valkiunas G, Bolshakov C V, Bensch S (2008) *Plasmodium relictum* (lineage P-SGS1): Effects on experimentally infected passerine birds. *Experimental Parasitology*, **120**, 372–380.

Palinauskas V, Valkiunas G, Križanauskiene A, Bensch S, Bolshakov C V. (2009) *Plasmodium relictum* (lineage P-SGS1): Further observation of effects on experimentally infected passeriform birds, with remarks on treatment with Malarone^{TM}. *Experimental Parasitology*, **123**, 134–139.

Patz JA, Reisen WK (2001) Immunology, climate change and vector-borne diseases. *Trends in Immunology*, **22**, 171–172.

Perkins SL, Schall J (2002) A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology*, **88**, 972–978.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Development Core Team (2015) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.r-project.org/.

Richard FA, Sehgal RNM, Jones H. I, Smith TB (2002) A Comparative Analysis of PCR-Based Detection Methods for Avian Malaria. *The Journal of Parasitology*, **88**, 819–822.

Richardson DS, Jury FL, Dawson DA, Salgueiro P, Komdeur J, Burke T (2000) Fifty Seychelles warbler (*Acrocephalus sechellensis*) microsatellite loci polymorphic in Sylviidae species and their cross-species amplification in other passerine birds. *Molecular Ecology*, **9**, 2225–2230.

Rousset F (1997) Genetic Differentiation and Estimation of Gene Flow from F-Statistics under Isolation by Distance. *Genetics*, **145**, 1219–1228.

Rousset F (2008) GENEPOP’007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schwensow N, Fietz J, Dausmann KH, Sommer S (2007) Neutral versus adaptive genetic variation in parasite resistance: importance of major histocompatibility complex supertypes in a free-ranging primate. *Heredity*, **99**, 265–277.

Scribner KT, Arntzen JW, Burke T (1994) Comparative analysis of intra- and interpopulation genetic diversity in *Bufo bufo*, using allozyme, single-locus microsatellite, minisatellite, and multilocus minisatellite data. *Molecular Biology and Evolution*, **11**, 737–748.

Sherman IW (1979) Biochemistry of *Plasmodium* (malarial parasites). *Microbiological Reviews*, **43**, 453–495.

Siikamaki P, Ratti O, Hovi M, Bennett GF (1997) Association between haematozoan infections and reproduction in the Pied Flycatcher. *Functional Ecology*, **11**, 176–183.

Slade RW, McCallum HI (1992) Overdominant Vs. Frequency-Dependent Selection at Mhc Loci. *Genetics*, **132**, 861–862.

Smouse PE, Long JC, Sokal RR (1986) Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence. *Systematic Zoology*, **35**, 627–632.

Spitze K (1993) Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics*, **135**, 367–374.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **277**, 979–988.

Stuglik MT, Radwan J, Babik W (2011) jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, **11**, 739–742.

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, **124**, 967–978.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, **30**, 2725–2729.

Tompkins DM, Gleeson DM (2006) Relationship between avian malaria distribution and an exotic invasive mosquito in New Zealand. *Journal of the Royal Society of New Zealand*, **36**, 51–62.

Valkiūnas G (2005) *Avian Malaria Parasites and other Haemosporidia*. CRC Press, Florida.

Warner RE (1968) The role of introduced diseases in the extinction of the endemic Hawaiian avifauna. *The Condor*, **70**, 101–120.

Westerdahl H, Wittzell H, von Schantz T, Bensch S (2004) MHC class I typing in a songbird with numerous loci and high polymorphism using motif-specific PCR and DGGE. *Heredity*, **92**, 534–42.

CHAPTER 4: Temporal variation and survival analysis in an isolated house sparrow (*Passer domesticus*) population

Availability of sequence data

The sequence data generated and analysed during this study will be submitted to an online repository.

4.0 Abstract

The extreme polymorphism at major histocompatibility complex (*Mhc*) genes is presumed to be maintained by selection pressure from pathogens. Testing this hypothesis requires long-term individual monitoring data to measure fitness traits such as survival and lifetime reproductive success. This study was therefore conducted on a house sparrow (*Passer domesticus*) population that was intensively monitored between 2000 to 2012. We predicted that an intermediate number of *Mhc* alleles within an individual would be associated with higher fitness, as has been shown in a previous study on three-spined sticklebacks (*Gasterosteus aculeatus*). In contrast to our expectation, we found that higher survival was associated with extreme (low or high) numbers of *Mhc* alleles. Two *Mhc* alleles were found to be significantly associated with the number of offspring produced in an individual's lifetime. Even at a small temporal scale, there was evidence of associations between *Mhc* and fitness traits such as lifespan and number of offspring produced.

4.1 Introduction

The major histocompatibility complex (*Mhc*) plays an important role in the adaptive immune system. The *Mhc* genes code for *Mhc* molecules, whose primary function is to display 'self' and 'non-self' peptides to the T-cells. An immune reaction by the T-cells is triggered when a non-self peptide is displayed on the cell's surface, indicating that the cell has been compromised by pathogens (Janeway *et al.* 2005). The *Mhc* is highly polymorphic and this diversity is thought to be maintained by balancing selection (Hedrick & Thomson 1983; Hedrick 1998). The two mechanisms that have been shown to drive the high polymorphism at *Mhc* genes are (1) frequency-dependent selection and (2) heterozygote advantage. In frequency-dependent selection, rare alleles are advantageous because pathogens have not adapted to them. Such rare alleles will increase in frequency and become common, so exerting selection on the pathogen to adapt to them. This will, in turn, select for new, initially rare alleles for immunity (Bodmer 1972). The alternative, heterozygote advantage, hypothesis states that individuals heterozygous for the *Mhc* will be able to present a wider range of foreign peptides to T-cells than homozygotes, and so confer greater immunity (Nei & Hughes 1991).

Parasites and pathogens seem to be the main factor driving the maintenance of genetic variation at the *Mhc* loci in natural populations (Spurgin & Richardson 2010). Nowak *et al.* (1992) hypothesized that having an intermediate number of *Mhc* alleles (i.e. optimal *Mhc* diversity) is more beneficial than having maximum individual *Mhc* diversity. An empirical study on sticklebacks (*Gasterosteus aculeatus*) showed that intermediate numbers of *Mhc* alleles were associated with lower parasite load and maximum lifetime fitness (Wegner *et al.* 2003; Kalbe *et al.* 2009). Maximal individual *Mhc* diversity may not be beneficial. High diversity at the *Mhc* genes should lead to a wider presentation of antigens, which might be expected to improve immune function, but this can have a negative consequence. Too many *Mhc* variants is expected to result in presentation of too many self-peptides, which will lead to the lost of T-cell diversity due to the elimination of T-cells that react with the self-peptide *Mhc* molecule combination (Nowak *et al.* 1992; Mason 2001), so reducing immune function.

Parasites such as malaria are not homogeneous in space (see study on willow warblers (*Phylloscopus trochilus*) (Bensch & Åkesson 2003)) and their frequency varies; thus, birds may encounter different parasites in different years. A study of the great reed warbler (*Acrocephalus arundinaceus*) demonstrated that certain *Mhc* alleles showed significant frequency changes between cohorts (Westerdahl *et al.* 2004). The authors attributed this to fluctuating selection pressure from parasites, caused by temporal variation in parasite abundance between years on both the breeding and wintering grounds (Westerdahl *et al.* 2004). *Mhc* allele frequency variation among cohorts can result from different alleles being selected in different years (Westerdahl *et al.* 2004). However, demographic events (e.g. presence of predators, seasonal changes and changes in food abundance) can also produce such changes. It has been suggested that (1) the variation in *Mhc* allele frequencies between cohorts is likely to be the result of demographic events if the allele frequency changes between cohorts at the *Mhc* loci are similar to those at neutral (e.g. microsatellite) markers and (2) if there is selection acting on *Mhc* genes, then the between-cohort variation will be greater at the *Mhc* than at the “random” level observed at comparable neutral loci (Westerdahl *et al.* 2004). Genetic differentiation at *Mhc* loci has been observed even at small spatial scales with a low prevalence of malaria (Bichet *et al.* 2015).

Several studies have shown an association between *Mhc* genotypes and pathogen resistance (Paterson *et al.* 1998; Schad *et al.* 2005; Westerdahl *et al.* 2005; Bonneaud *et al.* 2006; Oliver *et al.* 2009; Loiseau *et al.* 2011). This leads to the question of what are the fitness consequences of *Mhc*-related pathogen resistance? Studies of the relationship between survival and *Mhc* genotype in natural populations have not produced consistent results (Bonneaud *et al.* 2004; Hansson *et al.* 2004; Kalbe *et al.* 2009; Brouwer *et al.* 2010; Sepil *et al.* 2012). The earliest study to show an effect of the *Mhc* on survival was in wild ring-necked pheasants (*Phasianus colchicus*) (von Schantz *et al.* 1996), which found an association between annual male survival and *Mhc* genotype (both *Mhc* classes I and II). In the Seychelles warbler (*Acrocephalus sechellensis*), females gained extra-pair paternity with *Mhc* diverse males, which may have provided indirect benefits; for example, juvenile survival was found to be positively associated with *Mhc* diversity and a specific *Mhc* allele (Richardson *et al.* 2005; Brouwer *et al.* 2010). A study of a wild population of Soay sheep (*Ovis aries*) found an association between certain *Mhc* alleles and both juvenile survival and resistance to intestinal parasite (Paterson *et al.* 1998).

In an experimental study on house sparrows, Bonneaud *et al.* (2004a) investigated the effect on reproduction of inoculating the individuals with Newcastle disease virus. They found that nestling survival was positively correlated with female *Mhc* class I diversity (independent of treatment) (Bonneaud *et al.* 2004). In another study on house sparrows, in an isolated wild population, Karlsson *et al.* (2015) found no relationship between nestling survival, fledging or recruitment with *Mhc* diversity. However, there was an association between fledgling survival and a specific *Mhc* allele (allele 1105) (Karlsson *et al.* 2015). On the other hand, a study of collared flycatchers (*Ficedula albicollis*) found no evidence for an association between lifetime reproductive success and *Mhc* class II diversity (Radwan *et al.* 2012). A study of a wild population of great tits (*Parus major*) found no evidence that either intermediate or maximal *Mhc* diversity affected fitness, but specific *Mhc* types may have conferred better survival (Sepil *et al.* 2012).

Previous work on *Mhc* class I in Lundy house sparrow was conducted with Reference Strand-mediated Conformation Analysis (RSCA) (Karlsson *et al.* 2015). Here, a more advanced, next-generation sequencing-based genotyping method, with higher resolution, was used to screen for *Mhc* class I alleles (**Chapter 2**). The aims of this chapter were (1) to investigate whether there was variation in *Mhc* allele frequencies across 13 cohorts of Lundy house sparrows; and (2) to test whether is there any survival advantage associated with intermediate *Mhc* diversity and/or allele specific *Mhc.*

4.2 Methods

4.2.1 Study sites

The house sparrows belonged to a population inhabiting Lundy Island located in the Bristol Channel (51°10° N, 4°40° W, UK) (Nakagawa & Burke 2008; Schroeder *et al.* 2011, 2012). A total of 884 individuals were selected but only 584 were successfully *Mhc* typed in 2000–13 (2000, 43; 2001, 37; 2002, 50; 2003, 71; 2004, 55; 2005, 50; 2006, 30; 2007, 12; 2008, 26; 2009, 23; 2010, 62; 2011, 32; 2012, 64; 2013, 29). Individuals that had reached adulthood were selected based on fledgling capture, resighting and breeding data.

4.2.2 *Mhc* screening

Degenerate primers HNallaForward 5’-TCCCCACAGGTCTCCACAC-3’ and HNAlla reverse- 5’-TGCGCTCCAGCTCCYTCTGCC-3’ were used to amplify the *Mhc* class I exon 3 fragment size of 221–224 bp, which contains the peptide binding region (Karlsson & Westerdahl 2013). A 6-bp tag was added to the 5’ end of each primer in order to distinguish amplicons obtained from individuals. A total of 442 unique 6-bp tag combinations was prepared for each library (See Chapter 3, **Table 3.2** for the list of 6-bp identifiers). Each run consisted of a library of 442 pooled individuals. Two runs were prepared so covering all 884 individuals.

Polymerase chain reactions (PCRs) were performed in 15-μl reactions containing QIAGEN Multiplex MasterMix, 10–20 ng DNA and 0.2 μM of each primer. PCRs were performed in Tetrad PTC-225 Thermal Cycler (MJ Research, Waltham, Massachusetts) using the following program: 95°C at 15 min, then 30 cycles of 95°C for 30s, 65°C for 60s and 72°C for 60s. The reaction was completed with a final extension for 10 min at 72°C.

PCR products were run on a 1.5% agarose gel stained with SyberSafe (Invitrogen). Each set of 8 individuals was pooled and purified using the MinElute PCR purification kit (Qiagen). PCR amplicons were pooled and prepared for 250-bp paired-end Illumina sequencing on a MiSeq (Illumina, Inc., San Diego, CA, USA).

4.2.3 Preprocessing sequence data and *Mhc* genotyping

Forward and reverse sequences were aligned using FLASH with an overlap of 100-bp (Magoč & Salzberg 2011). The aligned sequences were then assessed in PRINSEQ and removal was initiated if the Phred quality score was below Q30 (Schmieder & Edwards 2011). The last step consisted of removing the unique 6-bp identifier barcodes and priming sequences in jMHC (Stuglik *et al.* 2011). This program not only trimmed the 6-bp unique identifier but also formatted the sequences and amplicon details into a table containing read numbers of each sequence for each amplicon.

The assumption was made that real alleles are amplified at higher sequencing depth than artefacts (Lighten *et al.* 2014). First, all variants in each amplicon were checked for the presence of homopolymers and chimaeras. A cumulative depth graph for each amplicon was then plotted (See **Chapter 2,** section **2.2.3** for clarification). The presence of an inflection point indicates that the amplicon is of a good quality because it enables separation between actual and artefactual allele sequences. Finally, the number of *Mhc* genotypes for each individual was estimated from the DOC value (see Macro sheet in Lighten *et al.* 2014). The number of *Mhc* alleles in an individual is also known as *Mhc* diversity. A Mann-Whitney *U*-test in R was used to compare the read depth between amplicons that had ‘good’ and ‘poor’ quality.

Haplotypes consisting of combinations of alleles that were consistently coinherited were identified (Janeway *et al.* 2005). Therefore, to account for the true number of *Mhc* alleles in the population, these haplotypes were treated as single *Mhc* alleles. Haplotypes were identified by a Spearman's correlation test in R for the presence of each pairwise combination of alleles.

4.2.4 Microsatellite genotyping

Variation at neutral markers should reflect stochastic and demographic events. Comparison between neutral markers such as microsatellites and a functional marker such as *Mhc* can provide useful information about any non-random polymorphism at the *Mhc* (Sommer 2005). A total of 884 individuals were genotyped using thirteen microsatellite loci: *Ase18* (Richardson *et al.* 2000), *Pdo1* (Neuman & Wetton 1996), *Pdo3, Pdo5, Pdo6* (Griffith *et al.* 1999), *Pdo9, Pdo10* (Griffith *et al.* 2007), *Pdo16a*, *Pdo17, Pdo19, Pdo22,* and, *Pdo27, Pdo40* (Dawson *et al.* 2012). PCRs were performed in two multiplexes, one including *Pdo1*, *Pdo3*, *Pdo5*, *Pdo6*, *Pdo9*, *Pdo10* and *Ase18*, and the other including *Pdo19*, *Pdo17*, *Pdo16a*, *Pdo22, Pdo27* and *Pdo40* (see **Chapter 3, Table 3.3** for primer concentrations).

The microsatellite markers were tested for deviations from Hardy–Weinberg equilibrium and linkage disequilibrium in GenePop 4.0 (Rousset 2008). Each locus was tested for Hardy–Weinberg equilibrium (HWE) using an exact test. Each pair of loci was tested for linkage equilibrium (LD) with an exact test, as implemented in GenePop 4.0 (Rousset 2008). Since this study was conducted in a monitored island population with a known pedigree, we removed related individuals that were half or full siblings. Individuals that had missing parental information were also removed from the analysis. A total of 143 individuals remained. None of the 13 microsatellite loci was found to deviate significant from Hardy–Weinberg equilibrium. Four pairs of loci were found to show linkage disequilibrium even after Bonferroni correction for multiple tests. One locus from each of these pairs was therefore omitted from further analyses (*Ase18*, *Pdo1*, *Pdo40* and *Pdo17*). We also tested for any significant deviation from Hardy–Weinberg equilibrium and linkage disequilibrium at each microsatellite locus within each cohort year (2000–

2011). We again found no evidence of microsatellite loci deviating from Hardy–Weinberg equilibrium, and no evidence of any loci in linkage disequilibrium.

Fourteen microsatellite alleles were selected, based on the frequency of each allele in the population, to provide a range of frequencies comparable to those of alleles at the *Mhc* class I loci, for use in further tests.

4.2.5 Testing for temporal variation in *Mhc* allele frequency and individual diversity

First, to examine whether there was more variation among cohorts than expected from random, the overall genetic differentiation among cohorts of recruits based on all alleles (total of 32 alleles: including both common and rare alleles) and just common alleles (14 alleles that were found in at least 30 of the 584 individuals) in *Mhc*, and also 14 microsatellite alleles, selected from the four loci, were analysed using structural analysis (AMOVA) with a permutation of 10,100 in Arlequin 3.5 (Excoffier & Lischer 2010). The data were input as haplotype data, with each sequence defined as an allele, along with the total number of individuals having that particular allele (Ekblom *et al.* 2007; Loiseau *et al.* 2009). Allele frequency was calculated as the number of individuals having a particular allele in a cohort divided by the total number of alleles among all the individuals in that cohort. Note that this way of calculating allele frequencies may result in the frequency of common alleles being underestimated and the frequency of rare alleles being overestimated (Ekblom *et al.* 2007).

Second, we tested whether *Mhc* and microsatellite allele frequencies varied among cohorts using a χ^2 test of homogeneity. The association between *Mhc* alleles and fitness measures, such as the number of offspring, was investigated. The lme4 package with the glmer function was used to run the generalized linear mixed effects model (GLMM) (Bates *et al.* 2014). The GLMM was fitted with a log link function and Poisson distribution structure.

The response variable was the number of offspring. The explanatory variables (or fixed effects) consisted of each *Mhc* allele (all 14 *Mhc* alleles). Cohort (2000–2010) was added as a random effect, which corrects the model for the confounding effects of cohort. To correct for overdispersion, an observation-level factor was added as a random effect in the model (Harrison 2014). Lifespan was incorporated as an offset to correct for the fact that individuals that lived longer might have had more opportunities, and to control for the expectation that number of offspring is proportional to lifespan. The 'drop1' function with a χ^2 test was used to obtain the minimal adequate model that best explains the variation in number of offspring (Zuur *et al.* 2009).

Third, an assessment was made of whether the total number of *Mhc* alleles found in an individual (*Mhc* diversity) varied among cohorts. An ANOVA test was conducted with *Mhc* diversity as the dependent variable and cohort as the independent variable.

4.2.6 Survival analysis

Survival analysis uses the duration of time until an event happens, for example the time until death (Cox & Oakes 1984). Because Lundy is isolated, reducing immigration and emigration to very low numbers, and the population is so well monitored, we can use our detailed sighting records (yearly resighting of 96%, Simons *et al.* 2015) to construct individual records of adult lifespan. Individuals that were still alive at the time of censusing were right-censored. In this study, a long-term data set between cohorts of 2000 and 2010 was used because individuals between these years had sufficient re-sighting data and highly resolved pedigree.

Data were analysed in R. The survival package has two functions: the "Surv" function, provided an estimate of survival, and the "coxph" function, which fits a Cox proportional hazard regression model to the data containing survival estimates and *Mhc* genotype (Therneau 2015a).

4.2.6.1 Association between *Mhc* diversity and survival in terms of lifespan

This section of this study was to investigate whether survival was associated with *Mhc* diversity and/or specific *Mhc* alleles. Survival in this case is the lifespan of each individual. The first section explored the relationship between the hazard of death (predicted hazard) and *Mhc* diversity (the number of unique *Mhc* alleles) by fitting a quadratic Cox proportional hazard regression model using the function “coxph” (Therneau 2015a). An ANOVA test was used to compare between fitting a linear and quadratic Cox proportional hazard regression model. The relationship between *Mhc* diversity and survival was inferred from the model that had a better fit. We illustrate the relationship between *Mhc* diversity and lifespan by plotting a quadratic curve. A Kaplan–Meier curve was plotted to illustrate the changes in survival across the lifespan.

To correct for potential confounding effects of cohort on this relationship, the “coxme” package in R that allows the addition of random terms (Therneau 2015b) was used. The relationship between *Mhc* diversity and survival was interpreted from the models that included the random term for cohort and corrected for the confounding effects of cohort.

A potential risk with fitting polynomial models, such as quadratic relationships, is that they are sensitive to outliers. The sensitivity analysis consisted of testing whether there was evidence for a true continuous relationship. Therefore, *Mhc* diversity was fitted as a factor in the model and we compared the independent estimates of risk for each *Mhc* diversity category to the quadratic model estimate. Cox proportional hazard models assume that hazards are proportional across independent variables, and the possible violation of this assumption was tested using the “cox.zph” function. There were no violations detected, as interpreted from the global test across all the included independent variables (**Table 4.1**).

Table 4.1 Test of the proportional hazard assumption for a Cox model. Global χ^2 and p-values (*P*). The non-significant values indicate that the predictors (e.g. survival function, which is the predicted hazard, versus *Mhc* diversity/and or the presence of specific *Mhc* alleles) were proportional across time.

	Global	
	χ^2	*P*
Mhc diversity	5.47	0.071
Specific *Mhc* alleles	0.211	0.573

4.2.6.2 Association between specific *Mhc* alleles and survival in terms of lifespan

In the second section, the relationship between predicted hazard (death) and specific *Mhc* alleles was investigated by fitting a Cox proportional hazard regression model that included a categorical factor for specific *Mhc* alleles; the latter were added as explanatory variables in the model. The *Mhc* alleles that occurred in at least 30 individuals were included in this model, to aid model convergence (14 *Mhc* alleles, Lseq01 to Lseq09, Lseq11, Lseq12, Lseq13, Lseq15, Lseq16). This factor was added to a model that contained *Mhc* diversity as a quadratic function (as we found this was important see **4.3.3**).

4.2.6.3 Association between fitness and variation at control microsatellite loci

In this section we investigate whether there is an association between survival and microsatellite heterozygosity, and/or specific microsatellite alleles. Individual heterozygosity was calculated as the number of heterozygous loci divided by the total number of loci. We used the function "GENHET" in R to calculate the individual heterozygosity for each of the 584 individuals (Coulon 2010). To investigate the relationship between survival and microsatellite heterozygosity, we fitted a quadratic Cox proportional hazard model, which consists of the hazard of death (dependent variable) and microsatellite heterozygosity as the explanatory variable. An ANOVA test was used to compare the fit of linear and quadratic models. To investigate the relationship between survival and specific microsatellite alleles, we fitted a Cox proportional hazard model consisting of hazard as the dependent variable with specific microsatellite alleles as the explanatory variable**.** We also fitted another model correcting for confounding effects of cohort using “coxme” package (Therneau 2015b).

All statistical analyses were conducted using R (R Development Core Team 2015) or as otherwise stated.

4.3 Results

4.3.1 *Mhc* class I exon 3 genotyping

The total number of reads produced was 4,644,142 (for 884 individuals). Sequences that were present as only a single copy were removed, leaving 4,523,080 reads. The mean number of reads per individual was 5,116 ± 83 (se). Among these samples, the total number of amplicons (= the number of individuals) successfully genotyped was 584 (3,491,639 reads, mean reads per individual = 5,978 ± 121 (se)). About one-third of amplicons failed the genotyping step (*N* = 300, mean reads per individual = 5,884 ± 205 (se)). There were no differences in the total read depth between 'good' and 'poor' amplicons (Mann-Whitney *U*-test, *Z* = 1.534, *P* = 0.125).

The total number of putative alleles found was 34. The alleles were checked for possible haplotypes using a correlation analysis. All individuals (122) that had Lseq13 also had Lseq14 (**Table 4.2**). Both alleles were always inherited together in those cases that had pedigree information (11 families). These two alleles were subsequently considered to belong to a single haplotype (entered as Lseq13). Lseq09 and Lseq10 were found to be highly correlated (0.938); 127/140 individuals were found to have both Lseq09 and Lseq10 alleles. Just thirteen individuals had only Lseq09. We therefore considered these two alleles to be non-independent, and usually belonging to a single haplotype, because nine families showed that both alleles were inherited from a single parent. In addition, we could not show that any individuals with only Lseq09 inherited it from a parent with both alleles, because of genetic parental information was not available. We therefore disregarded Lseq10. The total number of observed independent alleles was therefore 32 (see **Table 4.3**).

Table 4.2 Correlation matrix between *Mhc* alleles within individuals.

	Lseq01	Lseq02	Lseq03	Lseq04	Lseq05	Lseq06	Lseq07	Lseq08	Lseq09	Lseq10	Lseq11	Lseq12	Lseq13	Lseq14	Lseq15	Lseq16	Lseq17	Lseq18	Lseq19	Lseq20	Lseq21	Lseq22	Lseq23	Lseq24	Lseq25	Lseq26	Lseq27	Lseq28	Lseq29	Lseq30	Lseq31	Lseq32	Lseq33	Lseq34
Lseq01	-																																	
Lseq02	-0.329	-																																
Lseq03	-0.130	-0.337	-																															
Lseq04	-0.262	0.512	-0.110	-																														
Lseq05	-0.064	0.051	0.186	-0.235	-																													
Lseq06	-0.267	0.634	-0.285	0.822	-0.211	-																												
Lseq07	0.371	-0.236	-0.184	-0.209	-0.122	-0.174	-																											
Lseq08	0.251	0.059	-0.217	-0.153	-0.194	-0.119	-0.167	-																										
Lseq09	0.153	-0.202	0.244	0.080	-0.155	-0.186	-0.188	-0.118	-																									
Lseq10	0.161	-0.188	0.189	0.117	-0.156	-0.163	-0.195	-0.112	**0.938**	-																								
Lseq11	0.295	-0.081	-0.228	-0.086	-0.147	-0.052	-0.178	0.714	-0.103	-0.083	-																							
Lseq12	-0.128	-0.024	0.482	-0.182	0.320	-0.131	-0.098	-0.089	-0.091	-0.112	-0.063	-																						
Lseq13	0.354	-0.074	-0.196	-0.088	-0.151	-0.057	-0.212	0.837	-0.109	-0.089	0.866	-0.080	-																					
Lseq14	0.354	-0.074	-0.196	-0.088	-0.151	-0.057	-0.212	0.837	-0.109	-0.089	0.866	-0.080	**1**	-																				
Lseq15	-0.195	-0.124	-0.100	-0.045	-0.162	-0.037	-0.110	-0.053	-0.067	-0.054	-0.077	-0.126	-0.046	-0.046	-																			
Lseq16	-0.064	-0.218	0.238	-0.232	0.359	-0.178	-0.001	-0.125	-0.116	-0.133	0.055	0.495	-0.101	-0.101	-0.086	-																		
Lseq17	0.052	-0.241	0.228	-0.201	0.165	-0.190	-0.075	-0.119	-0.070	-0.057	-0.079	-0.086	-0.077	-0.077	-0.020	-0.057	-																	
Lseq18	-0.064	0.276	-0.123	-0.070	-0.098	-0.127	-0.030	0.388	-0.070	-0.088	-0.084	-0.063	-0.052	-0.052	-0.042	-0.085	-0.104	-																
Lseq19	-0.220	0.006	0.337	-0.017	-0.064	0.005	-0.112	-0.088	-0.038	-0.026	-0.085	0.567	-0.099	-0.099	-0.074	-0.038	-0.075	-0.018	-															
Lseq20	-0.198	-0.004	-0.076	-0.057	0.005	-0.060	-0.088	-0.035	-0.078	-0.071	-0.068	-0.106	-0.067	-0.067	-0.043	-0.047	-0.072	0.026	-0.060	-														
Lseq21	0.156	-0.185	0.279	-0.069	-0.087	-0.077	-0.082	-0.087	-0.002	0.010	-0.129	-0.089	-0.056	-0.056	-0.059	-0.050	0.036	-0.076	-0.071	0.020	-													
Lseq22	0.035	-0.060	0.054	-0.089	0.111	-0.089	-0.043	-0.084	-0.069	-0.061	0.197	-0.013	-0.076	-0.076	-0.049	0.249	-0.023	-0.040	-0.034	0.032	-0.056	-												
Lseq23	-0.126	-0.108	0.130	-0.084	0.165	-0.070	-0.066	-0.068	-0.025	-0.020	-0.057	-0.055	-0.057	-0.057	0.068	-0.049	-0.037	-0.034	-0.031	0.132	-0.027	-0.025	-											
Lseq24	-0.050	-0.072	-0.094	-0.057	-0.061	-0.055	0.001	-0.075	-0.040	-0.034	0.313	-0.028	-0.057	-0.057	-0.054	0.368	-0.055	-0.050	-0.005	0.073	-0.040	0.724	-0.018	-										
Lseq25	-0.039	-0.111	-0.015	-0.106	-0.065	-0.088	0.232	-0.029	-0.077	-0.073	-0.010	-0.070	-0.009	-0.009	-0.046	-0.061	-0.047	-0.042	-0.040	-0.029	-0.034	-0.031	-0.015	-0.022	-									
Lseq26	0.084	-0.019	0.162	-0.074	0.211	-0.050	-0.068	-0.046	-0.061	-0.056	-0.054	0.005	-0.053	-0.053	-0.015	-0.039	0.021	-0.008	-0.045	-0.033	0.009	0.646	-0.017	-0.025	-0.022	-								
Lseq27	0.104	-0.009	0.178	-0.068	0.226	-0.070	-0.064	-0.041	-0.057	-0.052	-0.050	0.011	-0.049	-0.049	-0.012	-0.035	0.025	-0.005	-0.043	-0.032	-0.038	0.672	-0.017	-0.024	-0.021	**0.963**	-							
Lseq28	-0.003	0.072	-0.050	-0.067	0.187	-0.049	0.020	-0.077	-0.037	-0.033	-0.065	-0.028	-0.064	-0.064	-0.042	-0.017	0.049	-0.038	0.017	0.042	-0.031	-0.028	-0.014	-0.020	-0.017	-0.020	-0.019	-						
Lseq29	0.081	0.032	-0.100	0.154	0.017	-0.010	-0.037	-0.072	-0.066	-0.062	-0.061	-0.059	-0.061	-0.061	-0.040	-0.052	-0.040	0.385	-0.034	-0.025	-0.029	-0.026	-0.013	-0.019	-0.016	-0.018	-0.018	-0.015	-					
Lseq30	0.064	-0.054	-0.003	-0.071	0.018	-0.059	-0.013	-0.057	0.035	-0.049	-0.048	-0.047	-0.048	-0.048	-0.031	-0.041	0.030	-0.028	-0.027	-0.020	-0.023	-0.021	-0.010	-0.015	-0.013	-0.015	-0.014	-0.012	-0.011	-				
Lseq31	-0.060	-0.041	-0.035	0.054	-0.028	0.065	-0.025	-0.025	-0.023	-0.022	-0.022	-0.021	-0.021	-0.021	-0.014	-0.018	-0.014	-0.013	-0.012	-0.009	-0.010	-0.009	-0.005	-0.007	-0.006	-0.006	-0.006	-0.005	-0.005	-0.004	-			
Lseq32	-0.060	-0.041	-0.035	-0.032	-0.028	0.065	-0.025	-0.025	-0.023	-0.022	-0.022	-0.021	-0.021	-0.021	-0.014	-0.018	-0.014	-0.013	-0.012	-0.009	0.167	-0.009	-0.005	-0.007	-0.006	0.264	-0.006	-0.005	-0.005	-0.004	-0.002	-		
Lseq33	-0.060	-0.041	-0.035	0.054	-0.028	0.065	-0.025	-0.025	-0.023	-0.022	-0.022	-0.021	-0.021	-0.021	-0.014	-0.018	-0.014	-0.013	-0.012	-0.009	-0.010	-0.009	-0.005	-0.007	-0.006	-0.006	-0.006	-0.005	-0.005	-0.004	**1**	-0.002	-	
Lseq34	0.050	-0.070	-0.061	-0.055	-0.048	-0.046	0.120	0.117	-0.040	-0.038	-0.037	-0.036	-0.037	-0.037	-0.024	-0.032	-0.024	-0.022	-0.021	-0.015	-0.018	-0.016	-0.008	-0.012	-0.010	-0.011	-0.011	-0.009	-0.008	-0.007	-0.003	-0.003	-0.003	-

Values in bold indicate that all the individuals have a pairwise combination of alleles. E.g. Lseq13 and Lseq14 have a correlation of 1, which means that all the individuals that had Lseq13 also had Lseq14

4.3.2 Temporal patterns of *Mhc* diversity

There was a slight but significant difference in the frequency of *Mhc* alleles among cohorts (32 *Mhc* alleles, AMOVA $_{13,583}$, $F_{\mathrm{st}} = 0.0017$, $P(F_{\mathrm{st}}) < 0.05$); 14 *Mhc* alleles, AMOVA $_{13,583}$, $F_{\mathrm{st}} = 0.0016$, $P(F_{\mathrm{st}}) < 0.05$), **Table 4.3**). However, in contrast, there was no significant difference in the frequency of the selected representative microsatellite alleles among cohorts (14 microsatellite alleles, AMOVA $_{13,583}$, $F_{\mathrm{st}} = 0.008$, $P(F_{\mathrm{st}}) > 0.05$, n.s.).

Table 4.3 The frequency of each allele across cohorts. This was calculated as the number of times an allele was observed in each cohort divided by the total number of all alleles observed in that cohort (2000–2013).

Mhc alleles	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Lseq01*	**0.176**	**0.162**	**0.143**	**0.138**	**0.135**	**0.155**	**0.126**	**0.172**	**0.110**	**0.092**	**0.098**	**0.092**	**0.111**	**0.129**
Lseq02*	**0.083**	**0.063**	**0.085**	**0.091**	**0.069**	**0.092**	**0.088**	**0.069**	**0.067**	**0.143**	**0.127**	**0.118**	**0.093**	**0.129**
Lseq03	**0.098**	**0.110**	**0.081**	**0.066**	**0.083**	**0.077**	**0.094**	**0.086**	**0.074**	**0.076**	**0.063**	**0.066**	**0.102**	**0.048**
Lseq04*	**0.052**	**0.037**	**0.085**	**0.077**	**0.052**	**0.052**	**0.082**	**0.052**	**0.049**	**0.084**	**0.092**	**0.066**	**0.090**	**0.095**
Lseq05	**0.093**	**0.089**	**0.062**	**0.044**	**0.052**	**0.044**	**0.063**	**0.034**	**0.049**	**0.050**	**0.061**	**0.072**	**0.051**	**0.075**
Lseq06*	**0.036**	**0.026**	**0.047**	**0.064**	**0.038**	**0.044**	**0.057**	**0.034**	**0.025**	**0.076**	**0.081**	**0.072**	**0.069**	**0.082**
Lseq07*	**0.083**	**0.079**	**0.066**	**0.077**	**0.042**	**0.063**	**0.057**	**0.086**	**0.018**	**0.034**	**0.014**	**0.039**	**0.027**	**0.054**
Lseq08*	**0.021**	**0.052**	**0.054**	**0.052**	**0.063**	**0.052**	**0.025**	**0.052**	**0.086**	**0.042**	**0.066**	**0.046**	**0.048**	**0.061**
Lseq09	**0.026**	**0.037**	**0.054**	**0.039**	**0.056**	**0.055**	**0.057**	**0.086**	**0.055**	**0.050**	**0.043**	**0.033**	**0.051**	**0.014**
Lseq10*	**0.016**	**0.037**	**0.031**	**0.041**	**0.056**	**0.041**	**0.019**	**0.034**	**0.080**	**0.042**	**0.052**	**0.039**	**0.036**	**0.034**
Lseq11	**0.031**	**0.037**	**0.016**	**0.033**	**0.042**	**0.055**	**0.038**	**0.034**	**0.043**	**0.042**	**0.040**	**0.046**	**0.051**	**0.027**
Lseq12*	**0.016**	**0.037**	**0.039**	**0.033**	**0.056**	**0.041**	**0.019**	**0.034**	**0.086**	**0.025**	**0.052**	**0.026**	**0.039**	**0.041**
Lseq13	0.010	0	0.008	0.019	0.021	0.007	0.025	0.034	0.006	0.025	0.026	0.033	0.033	0.034
Lseq14	**0.041**	**0.042**	**0.023**	**0.039**	**0.035**	**0.026**	**0.044**	**0.017**	**0.025**	**0.025**	**0.032**	**0.039**	**0.024**	**0.014**
Lseq15*	0.047	0.058	0.019	0.014	0.021	0.033	0.025	0.034	0	0	0.003	0.013	0.015	0.007
Lseq16	**0.021**	**0.010**	**0.027**	**0.019**	**0.007**	**0.018**	**0.006**	**0.017**	**0.006**	**0.025**	**0.020**	**0.020**	**0.009**	**0.027**
Lseq17	0	0.005	0.004	0.019	0.021	0.007	0.013	0	0.031	0.025	0.012	0.020	0.027	0.007
Lseq18	0.005	0.005	0.008	0.011	0.003	0.004	0.013	0	0.012	0.034	0.006	0.013	0.009	0
Lseq19	0.026	0.037	0.012	0.011	0.017	0.007	0	0	0	0	0.003	0.007	0.015	0.007
Lseq20	0.010	0.005	0.012	0.011	0.003	0.007	0.025	0	0.012	0.017	0.006	0.020	0.003	0.007
Lseq21	0	0	0	0	0	0	0	0	0.006	0.008	0	0	0.009	0.014
Lseq22	0.005	0.005	0.004	0.011	0	0	0.006	0	0.012	0.017	0.003	0.013	0	0
Lseq23	0.016	0	0.008	0.008	0.003	0	0	0	0	0	0	0.007	0	0.007
Lseq24	0.005	0	0.008	0	0.003	0.007	0.019	0	0	0	0.003	0.013	0.003	0.007
Lseq25	0.005	0	0.008	0	0.003	0.007	0.019	0	0	0	0.003	0.007	0.003	0.007
Lseq26	0.005	0	0.004	0.006	0.007	0	0.006	0	0.006	0	0	0.007	0	0
Lseq27	0.005	0	0.008	0.003	0	0.007	0	0	0	0	0	0	0	0.014
Lseq28	0.016	0.005	0	0	0	0	0	0	0	0	0	0	0.003	0
Lseq29	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0
Lseq30	0	0	0	0	0	0	0	0	0	0	0	0.007	0	0
Lseq31	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0
Lseq32	0	0.005	0	0.003	0	0	0	0	0	0	0	0	0	0.007

**Mhc* alleles whose frequency differed significantly among cohorts.

Alleles present across all cohorts are highlighted in bold.

In the analysis of each *Mhc* allele separately, nine *Mhc* allele frequencies were found to differ significantly among cohorts (**Table 4.4**, **Figure 4.1**). However, none of the representative microsatellite alleles showed among-year variation (**Table 4.5**).

Table 4.4 *Mhc* frequency variation across cohorts. χ^2 test of homogeneity, *d.f.* = 13, Total number of individuals = 584.

Mhc alleles	N	χ^2	*P*
Lseq01	193	36.4	**0.0005**
Lseq02	129	34.7	**0.001**
Lseq03	130	16.4	0.231
Lseq04	108	27	**0.012**
Lseq05	87	13.2	0.429
Lseq06	79	28.6	**0.007**
Lseq07	77	41.3	**0.0001**
Lseq08	70	2.1	**0.023**
Lseq09	67	19.1	0.121
Lseq11	56	25.6	**0.019**
Lseq12	61	13.5	0.409
Lseq13	56	30.4	**0.004**
Lseq15	51	7.6	0.867
Lseq16	29	40.4	**0.0001**

N, Number of individuals with the allele.

χ^2 Chi-square value.

P, P-value from Chi-square test. Significant results (in bold) showed that there is variation in allele frequencies between the cohorts.

Figure 4.1 Frequencies of the nine *Mhc* alleles that showed significant variation in frequency among cohorts.

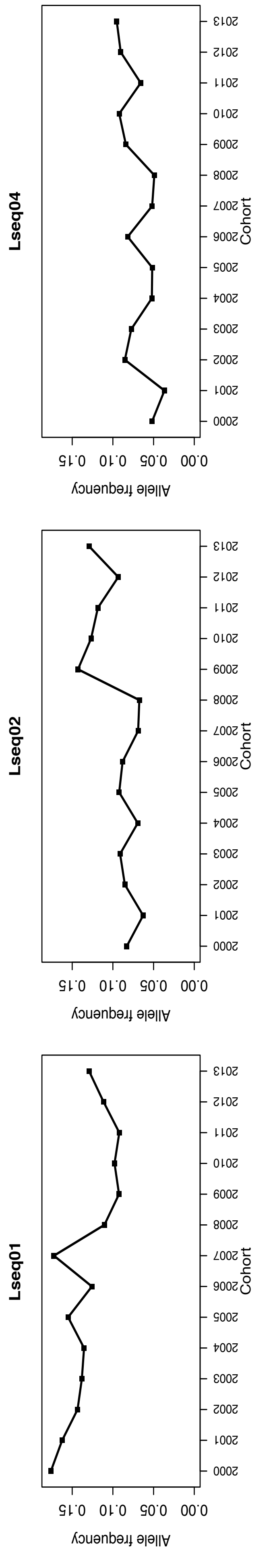


Table 4.5 Results from χ^2 test of homogeneity (*d.f.* = 13) for presence/absence of representative example microsatellite alleles in the fourteen cohorts of house sparrows (*N* = 584 individuals).

Microsatellite locus	Number of alleles	He	Allele	N	χ^2	*P*
Pdo5	10	0.821	P5A248	199	14.56	0.336
			P5A252	108	13.96	0.377
			P5A254	55	21.46	0.064
			P5A256	29	18.95	0.125
Pdo10	11	0.668	P10A135	37	22.28	0.051
			P10A139	78	14.94	0.311
			P10A143	99	14.47	0.342
Pdo19	4	0.662	P19A174	338	10.58	0.646
			P19A182	137	11.38	0.579
			P19A184	386	14.84	0.318
			P19A188	109	19.77	0.101
Pdo27	11	0.751	P27A230	186	19.96	0.096
			P27A236	225	16.08	0.245
			P27A256	55	10.60	0.644

H_e, Expected heterozygosity.

N, Number of individuals with the allele.

χ^2, Chi-square value.

P, P-value from Chi-test.

Out of fourteen *Mhc* alleles tested for an association with fitness, i.e. number of offspring produced, only two *Mhc* alleles (Lseq01 and Lseq02) were found to have a significant effect (**Table 4.6**). The number of *Mhc* alleles per individual (*Mhc* diversity) varied between 2 and 9 alleles (mean 5.0 ± 0.06 (se)). This suggested that the number of loci in an individual was between 1 and 5. A previous house sparrow study that used 454 amplicon sequencing showed that the number of *Mhc* class I exon 3 ‘long’ alleles found within an individual was between 3 and 8, translating to between 2 and 4 loci (sample size was 45 individuals) (Karlsson & Westerdahl 2013). There was no significant difference in the total number of *Mhc* alleles per individual between cohorts ($F_{13,\,570}$ = 1.579, $P > 0.05$, n.s., **Figure 4.2**).

Table 4.6 The number of offspring produced by Lundy house sparrows in relation to specific *Mhc* alleles. Parameter estimates and standard errors from a generalized linear mixed effects model (GLMM) with Poisson distribution. Statistically significant fixed effects are indicated in bold. (*N* = 459 individuals, mean number of offspring = 5.0 ± 0.31 (SE)).

Fixed	Offspring
Intercept	**-1.41 (-1.58 – -1.24)**
Lseq01	**0.29 (0.16 – 0.43)**
Lseq02	**0.27 (0.14 – 0.40)**
Random	Variance
Cohort	0.032

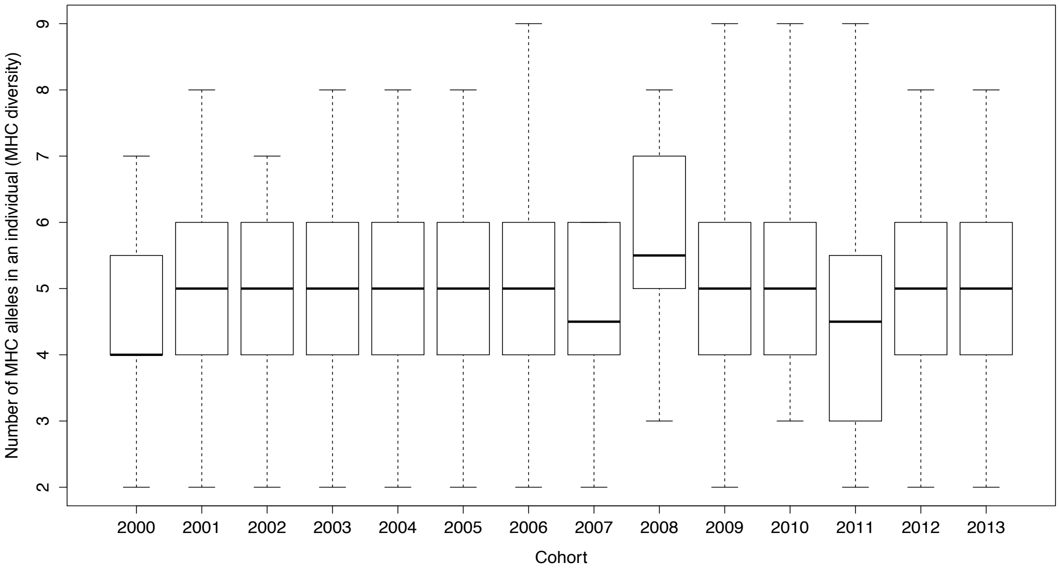


Figure 4.2 Boxplot of the distribution of the number of *Mhc* alleles in an individual (*Mhc* diversity) across the cohorts. The median *Mhc* diversity was 5. There was no significant variation in *Mhc* diversity among cohorts.

4.3.3 Association between survival and *Mhc* diversity

Two models were fitted to investigate the relationship between predicted hazard and *Mhc* diversity. Predicted hazard is the estimation of hazard (or risk) or death of a group (comprised of all individuals belonging to different groups, e.g. *Mhc* diversity). The linear model (survival ~ *Mhc* diversity) did not detect any effect on survival (**Table 4.7a**). However, the quadratic model (survival ~ *Mhc* diversity + (*Mhc* diversity)^2) was statistically significant (**Table 4.7b**). The two models were significantly different from one another (ANOVA, χ^2 = 4.28, *d.f.* = 1, P < 0.05) indicating that the quadratic model was a better model. The results indicate that survival is lowest for individuals with an intermediate number of alleles (**Figure 4.3a**). Individuals with the lowest and highest numbers of alleles (extreme ends of the *Mhc* diversity spectrum) were linked with higher survival (**Figure 4.3a**). A similar trend is also seen in lifespan, where an intermediate number of alleles was associated with a shorter lifespan, while the lowest and highest numbers of alleles were associated with a longer lifespan (**Figure 4.3b**).

Table 4.7 Cox proportional hazard analyses of survival. Both linear and quadratic model contains only one variable, individual *Mhc* diversity.

(a) Linear

Variable	Coef	se	*P*-value
Mhc diversity	0.023	0.031	0.275

(b) Quadratic

Variable	Coef	se	*P*-value
Mhc diversity	0.40	0.19	0.03
(*Mhc* diversity)^2	-0.04	0.018	0.04

Coef, estimated coefficient for survival from the hazard model
se, standard error
P-value from the Cox proportional hazard regression model

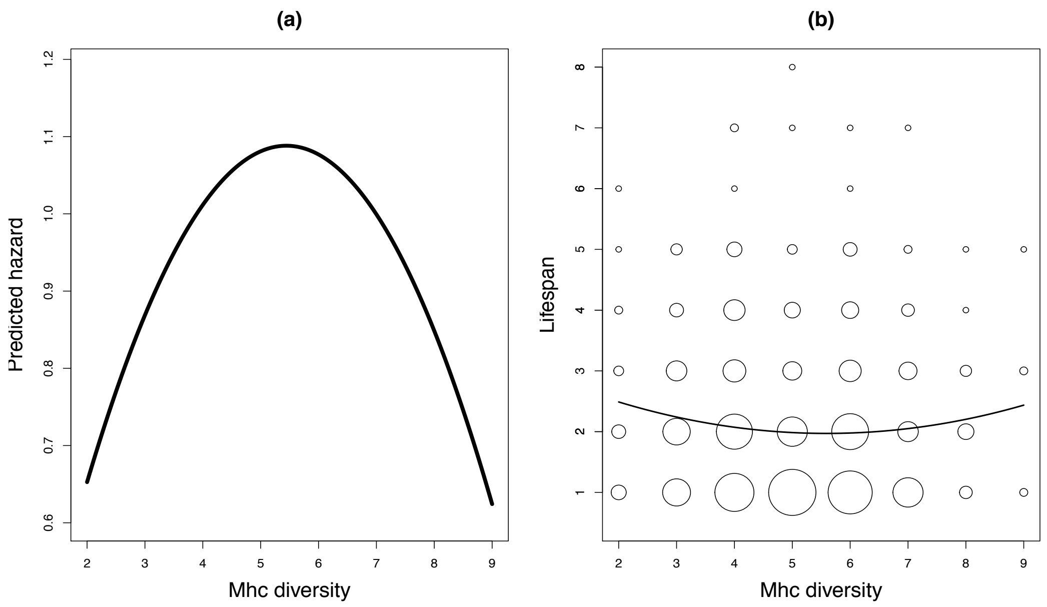


Figure 4.3 (**a**) Quadratic relationship between predicted hazard and *Mhc* diversity from a proportional hazard models. *Mhc* diversity was entered as a continuous variable. Intermediate *Mhc* diversity was associated with lower survival. (**b**) Quadratic relationship between *Mhc* diversity and lifespan $(y = 0.032x + 0.047x^2 + 5.58 \times 10^{-10})$. Intermediate *Mhc* diversity was associated with lower lifespan. The data contained 555 individuals.

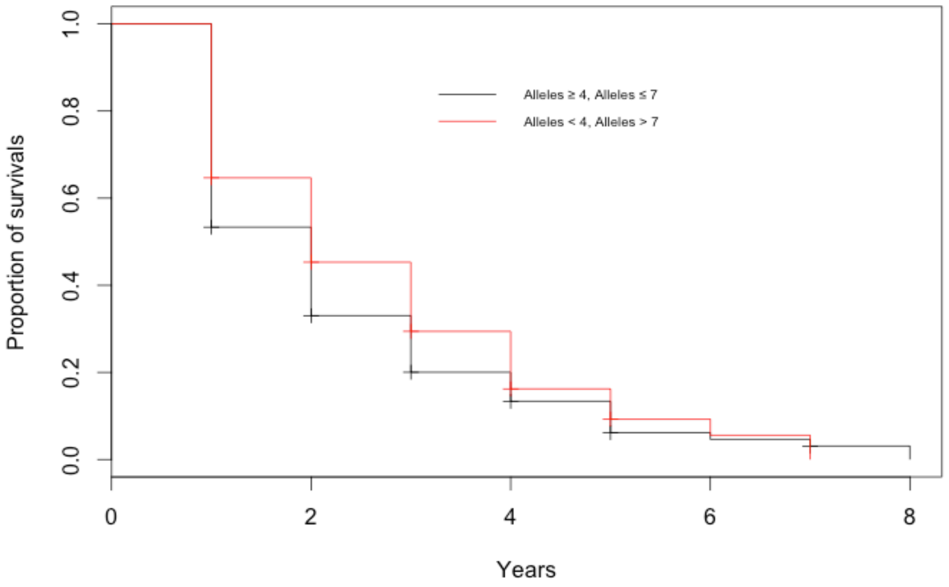


Figure 4.4 Kaplan–Meier survival curves in relation to intermediate (number of alleles between 4 and 7 – black line) and divergent *Mhc* diversity (consisting of the lowest [2 and 3] and highest [8 and 9] numbers of alleles – red line).

There was a statistically significant relationship between predicted hazard and *Mhc* diversity (**Figure 4.3a**); the quadratic relationship was not a result of bias in death among the different *Mhc* diversity levels (**Table 4.8**). There was an increase in proportion of death as the *Mhc* diversity increased from 3 to 5, and a decrease in the proportion of death as the *Mhc* diversity increased from 7 to 9 (**Table 4.8**).

The trend was further supported by plotting a Kaplan-Meier survival curve. The survival plot illustrated that survival was much lower at intermediate *Mhc* diversity (between 4 and 7 alleles, black line, **Figure 4.4**). However, survival was much higher for more divergent degrees of *Mhc* diversity (i.e. at the lowest (2 and 3) and highest numbers of *Mhc* alleles (8 and 9) (red line, **Figure 4.4**). When both linear and quadratic models were corrected for cohort effects (as a random effect), the models became non-significant (**Table 4.9**). This showed that when we control for temporal variation by cohorts, the significant relationship between survival and *Mhc* diversity disappeared.

Table 4.8 Cox proportional hazard regression model for different levels of *Mhc* diversity. Each coefficient is estimated relative to an *Mhc* diversity level of 2.

Mhc diversity levels	Coefficient	Standard error
3	-0.112	0.281
4	-0.017	0.262
5	0.274	0.263
6	0.150	0.261
7	0.065	0.284
8	-0.087	0.361
9	-0.524	0.627

Table 4.9 Cox proportional hazards analyses of survival. Both linear and quadratic models were corrected for cohort effects and with one variable: individual *Mhc* diversity.

(a) Linear

Variable	Coef	se	*P*-value
Mhc diversity	-0.001	0.032	0.99

(b) Quadratic

Variable	Coef	se	*P*-value
Mhc diversity	0.35	0.19	0.06
(*Mhc* diversity)^2	-0.03	0.018	0.06

Coef, estimated coefficient for survival from the hazard model

se, standard error

P-value from the Cox proportional hazard regression model

Table 4.10 Results from Cox proportional hazard regression model of *Mhc* diversity (as a quadratic function) and the commonest 14 *Mhc* alleles in 11 cohorts (2000–2010). The data contained 555 individuals (114 censored).

	Coef	se	*P-value*
Mhc diversity	0.51	0.29	0.07
(*Mhc* diversity)^2	-0.04	0.02	0.03*
Lseq01	-0.16	0.22	0.47
Lseq02	0.15	0.19	0.43
Lseq03	-0.26	0.22	0.23
Lseq04	0.04	0.24	0.87
Lseq05	-0.62	0.19	0.38
Lseq06	-0.54	0.29	0.06
Lseq07	0.22	0.19	0.25
Lseq08	-0.13	0.35	0.72
Lseq09	-0.44	0.38	0.24
Lseq11	-0.22	0.35	0.54
Lseq12	-0.16	0.22	0.47
Lseq13	0.19	0.43	0.66
Lseq15	0.37	0.21	0.08
Lseq16	0.21	0.21	0.30

Coef, estimated coefficient for survival from the hazard model

se, standard error

P-value from the Cox proportional hazard regression model

* significant at P < 0.05

4.3.4 Association between survival and specific *Mhc* alleles

The model, which included *Mhc* diversity as a quadratic function and 14 selected *Mhc* alleles (model: survival ~ *Mhc* diversity + (*Mhc* diversity)^2 + Lseq01 + Lseq02 + Lseq03 + Lseq04 + Lseq05 + Lseq06 + Lseq07 + Lseq08 + Lseq09 + Lseq11 + Lseq12 + Lseq13 + Lseq15 + Lseq16), showed that none of the alleles had a significant effect on survival (**Table 4.10**). However, a significant relationship was found between survival and *Mhc* diversity (**Table 4.10**). When this model was corrected for cohort effects, there was no statistically significant relationship found for either *Mhc* measure (diversity and specific alleles) with survival (**Table 4.11**).

Table 4.11 Results from a Cox proportional hazard regression model of *Mhc* diversity (as a quadratic function) and the commonest 14 *Mhc* alleles in 11 cohorts (2000 – 2010). In this model, cohort was added as a random effect. The estimate is the estimated coefficient for survival from the hazard model. The data contained 555 individuals (114 censored).

	Coef	se	*P-value*
Mhc diversity	0.44	0.26	0.084
(*Mhc* diversity)^2	-0.03	0.02	0.12
Lseq01	-0.11	0.23	0.65
Lseq02	0.16	0.21	0.43
Lseq03	-0.23	0.23	0.30
Lseq04	0.12	0.25	0.64
Lseq05	-0.01	0.19	0.95
Lseq06	-0.46	0.29	0.12
Lseq07	0.19	0.20	0.34
Lseq08	-0.28	0.36	0.44
Lseq09	0.00	0.45	0.99
Lseq11	-0.07	0.37	0.85
Lseq12	-0.25	0.23	0.26
Lseq13	0.09	0.44	0.85
Lseq15	0.35	0.21	0.094
Lseq16	0.20	0.22	0.36

Coef, estimated coefficient for survival from the hazard model

se, standard error

P-value from the Cox proportional hazard regression model

* significant at $P < 0.05$

4.3.5 Association between survival and microsatellite variation

The relationship between predicted hazard and microsatellite heterozygosity was investigated by fitting two models: linear (model: survival ~ heterozygosity) and quadratic (model: survival ~ heterozygosity + heterozygosity^2). Neither model detected an effect of heterozygosity on predicted hazard (**Table 4.12**). Both models remained non-significant even after cohort was added as a random effect (**Table 4.13**).

Table 4.12 Cox proportional hazard analyses of survival. Both linear and quadratic model contains only one variable, individual heterozygosity.

(a) Linear

Variable	Coef	se	*P*-value
heterozygosity	-0.32	0.35	0.37

(b) Quadratic

Variable	Coef	se	*P*-value
heterozygosity	-2.36	4.01	0.56
heterozygosity^2	1.05	2.06	0.61

Coef, estimated coefficient for survival from model

se, standard error

P-value from the Cox proportional hazard regression model

Table 4.13 Cox proportional hazards analyses of survival. Both linear and quadratic models were corrected for cohort effects and with one variable: individual heterozygosity.

(a) Linear

Variable	Coef	se	*P*-value
heterozygosity	-0.40	0.27	0.36

(b) Quadratic

Variable	Coef	se	*P*-value
heterozygosity	-2.36	4.07	0.56
heterozygosity^2	1.01	2.10	0.63

Coef, estimated coefficient for survival from the hazard model
se, standard error
P-value from the Cox proportional hazard regression model

The relationship between predicted hazard and specific microsatellite alleles was investigated by fitting Cox proportional models with specific microsatellite alleles as the explanatory variables (model: survival ~ P10A135 + P10A139 + P10A143 + P19A174 + P19A182 + P19A188 + P27A230 + P27A236 + P27A256 + P5A248 + P5A250 + P5A252 + P5A254 + P5A256). There were no significant effects of specific microsatellite alleles on survival (**Table 4.14**). No significance was found in the model when cohort was added as a random effect (**Table 4.15**).

Table 4.14 Cox proportional hazards analyses of survival. The dependent variable consists of individual heterozygosity or the presence of specific microsatellite alleles.

Variables	Coef	se	*P*-value
P10A135	-0.739	0.255	0.060
P10A139	-0.072	0.149	0.628
P10A143	-0.056	0.134	0.675
P19A174	-0.107	0.135	0.428
P19A182	0.131	0.143	0.357
P19A188	0.113	0.147	0.443
P27A230	-0.154	0.110	0.162
P27A236	0.111	0.104	0.288
P27A256	-0.283	0.187	0.129
P5A248	0.084	0.112	0.452
P5A250	0.003	0.114	0.977
P5A252	-0.022	0.135	0.870
P5A254	0.084	0.167	0.612
P5A256	0.064	0.241	0.789

Coef, estimated coefficient for survival from the hazard model
se, standard error
P-value from the Cox proportional hazard regression model

Table 4.15 Cox proportional hazards analyses of survival. Both linear and quadratic models were corrected for cohort effects and with one variable: individual heterozygosity.

Variables	Coef	se	*P*-value
P10A135	-0.345	0.261	0.190
P10A139	-0.054	0.153	0.720
P10A143	-0.141	0.138	0.300
P19A174	0.017	0.139	0.900
P19A182	0.120	0.143	0.400
P19A184	0.159	0.146	0.280
P19A188	0.104	0.153	0.500
P27A230	-0.214	0.113	0.059
P27A236	-0.011	0.106	0.920
P27A256	-0.258	0.194	0.190
P5A248	0.170	0.117	0.150
P5A250	0.057	0.114	0.620
P5A252	0.119	0.139	0.390
P5A254	0.090	0.172	0.600
P5A256	0.162	0.247	0.510

Coef, estimated coefficient for survival from the hazard model
se, standard error
P-value from the Cox proportional hazard regression model

4.4 Discussion

Out of 34 alleles that could be confidently genotyped as *Mhc* alleles, two alleles were exclusively associated with a single haplotype (Lseq13 and Lseq14, **Table 4.2**). All 112 individuals with one of these two alleles also had the other one of the pair, indicating tight linkage. Another pair of alleles (Lseq09 and Lseq10, **Table 4.2**) was also associated with a single haplotype.

The mean number of *Mhc* alleles per individual using Illumina sequencing was 5.0 ± 0.06 (se). In a previous study of the Lundy population using 454 pyrosequencing found a similar mean number of *Mhc* alleles per individual (mean ± se: 5.0 ± 1.6) (Karlsson & Westerdahl 2013). However, the resolution was much lower when using RSCA (mean ± se: 3.6 ± 1.1) (Karlsson *et al.* 2015).

Here, we hypothesised that balancing selection would maintain a high level of variation in *Mhc* genes. Direct comparison between *Mhc* alleles and presumably neutral microsatellite loci can be used to validate whether the variation in allele frequency at the *Mhc* is not due simply to demographic events (Westerdahl *et al.* 2004). The allele frequency variation among cohorts was significantly different at the *Mhc* than at the microsatellites. A similar pattern was also seen in great reed warblers (Westerdahl *et al.* 2004).

A study on great reed warblers (Westerdahl *et al.* 2004) showed that a difference in the frequency of one allele (B4b) among cohorts was associated with selection pressure exerted by parasites. It was suggested that this allele provided resistance against the malarial parasite in juvenile birds. In addition, temporal variation in the type and abundance of parasites might lead to selection favouring alternative *Mhc* alleles in different years. This will be especially true for long-distance migrant birds (such as great reed warblers but not house sparrows) that have different wintering and breeding grounds. However, we have no direct evidence of such effects in our non-migratory population; in a previous study on Lundy house sparrows, avian malaria was screened for using blood smears and PCR-based methods and it was concluded that malaria is rare or absent in this population (H. Westerdahl & T. Burke, unpubl. data). This could be further investigated with the use to next-generation sequencing, in which a low frequency of malarial parasites is more easily detected than by traditional PCR-based detection techniques (see **Chapter 3**). There is also the possibility of other intercellular pathogens also affecting the frequency of *Mhc* alleles (Karlsson *et al.* 2015).

In the study of sticklebacks (*Gasterosteus aculeatus*), intermediate *Mhc* diversity was associated with higher lifetime reproductive success (Kalbe *et al.* 2009). In contrast, this study found that the risk of death increased in individuals with an intermediate number of *Mhc* alleles (**Figure 4.3a**). Survival (lifespan) increased in those individuals with either a low or high number of *Mhc* alleles (**Figure 4.3a** and **Figure 4.4**). However, the significance disappeared when the model was correct for cohort. A previous study on Lundy house sparrows found no association between reproductive success (e.g. nestling success, fledging success or recruitment) and *Mhc* diversity (Karlsson *et al.* 2015). Studies on passerines have shown mixed results when investigating the relationship between survival and *Mhc*. In a study on collared flycatchers (*Ficedula albicollis*), there was no relationship between lifetime reproductive fitness and *Mhc* diversity (Radwan *et al.* 2012). In a Seychelles warbler population with relatively low genetic diversity, juvenile survival increased with high *Mhc* diversity (Brouwer *et al.* 2010). In addition, survivors tended to have higher *Mhc* diversity than non-survivors in great reed warblers (Westerdahl *et al.* 2004).

The study on collared flycatchers found an association (1) between the allele, *Fial-DNB*022* and lifetime reproductive success and (2) between the allele, *Fial-DNB*63* and susceptibility to infection, but these effects were non-significant after correction for false discovery rates. This suggests that different *Mhc* alleles might be associated with different fitness traits.

Evidence from the study of wild Soay sheep (*Ovis aries*) showed that *Mhc* alleles at the OLADRB locus had the strongest associations with juvenile survival (two age classes: lamb and yearling). They found allele *OLADRB257* to be significantly associated with both decreased parasite resistance and decreased survival in lambs, while allele *OLADRB263* was associated with both increased parasite resistance and increase survival in yearlings. These findings provided solid evidence that *Mhc* alleles have different associations with survival at different stages of age (Paterson *et al.* 1998), reflecting the complex interplay between between parasites and host immune system. No association was found between lifespan and specific *Mhc* alleles in our study. However, there was a positive association between the *Mhc* alleles (Lseq01 and Lseq02) and a reproductive measure of fitness, i.e. number of offspring produced in its lifetime. The study by Karlsson *et al.* (2015), on the same population, found that one *Mhc* allele (allele 1105) was positively associated with juvenile survival. We did not establish the relationship between the allele detected by RSCA by Karlsson *et al.* (2015) and those detected by sequencing in this study.

Evidence from recent studies has shown an association between survival and specific *Mhc* alleles (e.g. Seychelles warbler, Brouwer *et al.* 2010; Soay sheep, Paterson *et al.* 1998). The significant association between specific *Mhc* alleles and the fitness trait, lifetime reproductive fitness, in the collared flycatcher was lost when correcting the model for the false discovery rate. This was attributed to the low power of the statistical test, as it included many predictor variables (*Mhc* alleles). A solution would be to group these *Mhc* alleles into supertypes based on their antigen binding motifs, the sequences for which are under positive selection (Doytchinova & Flower 2005). A computational *Mhc* study showed that genetically divergent allele pairs will have less overlap in the antigen-binding region and will therefore be able to recognize a wider range of antigens (Lenz 2011). In the study of great tits (*Parus major*), 755 *Mhc* alleles were grouped into 17 supertypes. They found that (1) individuals having *Mhc* supertype 3 experienced higher survival, (2) individuals with *Mhc* supertype 6 experienced higher lifetime reproductive fitness and (3) individuals with *Mhc* supertype 5 had reduced lifetime fitness (Sepil *et al.* 2012).

In conclusion, this study suggested that there was selection acting on *Mhc* alleles based on the fact that the frequency of *Mhc* alleles varied non-randomly among cohorts spanning 13 years. We did not find any evidence for an optimal number of *Mhc* alleles in the form of a relationship between fitness, measured as survival, and the number of *Mhc* alleles, but instead we found the number of offspring produced was positively associated with the presence of two alleles (Lseq01 and Lseq02). In summary, this study of an insular population of house sparrows suggests that there is a relationship between fitness traits, such as the number of offspring, and the presence of specific *Mhc* alleles.

4.5 References

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, **67**, 1–48.

Bensch S, Åkesson S (2003) Temporal and Spatial Variation of Hematozoans in Scandinavian Willow Warblers. *Journal of Parasitology*, **89**, 388–391.

Bichet C, Moodley Y, Penn DJ, Sorci G, Garnier S (2015) Genetic structure in insular and mainland populations of house sparrows (*Passer domesticus*) and their hemosporidian parasites. *Ecology and Evolution*, **5**, 1639–1652.

Bodmer WF (1972) Evolutionary significance of the HL-A system. *Nature*, **237**, 139–145.

Bonneaud C, Mazuc J, Chastel O, Westerdahl H, Sorci G (2004) Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. *Evolution*, **58**, 2823–2830.

Bonneaud C, Pérez-Tris J, Federici P, Chastel O, Sorci G (2006) Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, **60**, 383–389.

Brouwer L, Barr I, Van De Pol M, Burke T, Komdeur J, Richardson DS (2010) MHC-dependent survival in a wild population: Evidence for hidden genetic benefits gained through extra-pair fertilizations. *Molecular Ecology*, **19**, 3444–3455.

Coulon A (2010) genhet: an easy-to-use R function to estimate individual heterozygosity. *Molecular Ecology Resources*, **10**, 167–169.

Cox DR, Oakes D (1984) *Analysis of Survival Data*. Chapman & Hall, London.

Dawson DA, Horsburgh GJ, Krupa AP, Stewart IRK, Skjelseth S, Jensen H, Ball AD, Spurgin LG, Mannarelli M-E, Nakagawa S, Schroeder J, Vangestel C, Hinten GN, Burke T (2012) Microsatellite resources for Passeridae species: a predicted microsatellite map of the house sparrow *Passer domesticus*. *Molecular Ecology Resources*, **12**, 501–

523.

Doytchinova IA, Flower DR (2005) In Silico Identification of Supertypes for Class II MHCs. *The Journal of Immunology*, **174**, 7085–7095.

Ekblom R, Sæther SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kålas JA, Höglund J (2007) Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology*, **16**, 1439–1451.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Griffith SC, Dawson DA, Jensen H, Ockendon N, Greig C, Neumann K, Burke T (2007) Fourteen polymorphic microsatellite loci characterized in the house sparrow *Passer domesticus* (Passeridae, Aves). *Molecular Ecology Notes*, **7**, 333–336.

Griffith SC, Stewart IRK, Dawson DA, Owens IPF, Burke T (1999) Contrasting levels of extra-pair paternity in mainland and island populations of the house sparrow (*Passer domesticus*): is there an “island effect”? *Biological Journal of the Linnean Society*, **68**, 303–316.

Hansson B, Westerdahl H, Hasselquist D, Åkesson M, Bensch S (2004) Does Linkage Disequilibrium Generate Heterozygosity-Fitness Correlations in Great Reed Warblers? *Evolution*, **58**, 870–879.

Harrison XA (2014) Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, **2**, e616.

Hedrick P (1998) Balancing selection and MHC. *Genetica*, **104**, 207–214.

Hedrick PW, Thomson G (1983) Evidence for Balancing Selection at HLA. *Genetics*, **104**, 449–456.

Janeway C, Travers P, Walport M, Shlomchik M (2005) *Immunobiology: The Immune System in Health and Disease*. Garland Science Publishing.

Kalbe M, Eizaguirre C, Dankert I, Reusch TBH, Sommerfeld RD, Wegner KM, Milinski M (2009) Lifetime reproductive success is maximized with optimal major histocompatibility complex diversity. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 925–934.

Karlsson M, Schroeder J, Nakagawa S, Smith HG, Burke T, Westerdahl H (2015) House sparrow *Passer domesticus* survival is not associated with MHC-I diversity, but possibly with specific MHC-I alleles. *Journal of*

Avian Biology, **46**, 167–174.

Karlsson M, 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.

Lenz TL (2011) Computational prediction of MHC II-antigen binding supports divergent allele advantage and explains trans-species polymorphism. *Evolution*, **65**, 2380–2390.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*, **14**, 1–15.

Loiseau C, Richard M, Garnier S, Chastel O, Julliard R, Zoorob R, Sorci G (2009) Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology*, **18**, 1331–1340.

Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) *Plasmodium relictum* infection and MHC diversity in the house sparrow (*Passer domesticus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **278**, 1264–1272.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Mason D (2001) Some quantitative aspects of T-cell repertoire selection: the requirement for regulatory T cells. *Immunological reviews*, **182**, 80–88.

Nakagawa S, Burke T (2008) The mask of seniority? A neglected age indicator in house sparrows *Passer domesticus*. *Journal of Avian Biology*, **39**, 222–225.

Nei M, Hughes AL (1991) *Polymorphism and evolution of the major histocompatibility complex loci in mammals* (R Selander, A Clark, T Whittam, Eds,). Sinauer Sunderland, MA.

Neuman K, Wetton JH (1996) Highly polymorphic microsatellites in the house sparrow *Passer domesticus*. *Molecular Ecology*, **5**, 307–309.

Nowak MA, Tarczy-Hornoch K, Austyn JM (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**, 10896–10899.

Oliver MK, Telfer S, Piertney SB (2009) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 1119–1128.

Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with juvenile survival and parasite resistance in a large unmanaged ungulate population. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 3714–3719.

R Development Core Team (2015) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.r-project.org/.

Radwan J, Zagalska-Neubauer M, Cichoń M, Sendecka J, Kulma K, Gustafsson L, Babik W (2012) MHC diversity, malaria and lifetime reproductive success in collared flycatchers. *Molecular Ecology*, **21**, 2469–2479.

Richardson DS, Jury FL, Dawson DA, Salgueiro P, Komdeur J, Burke T (2000) Fifty Seychelles warbler (*Acrocephalus sechellensis*) microsatellite loci polymorphic in Sylviidae species and their cross-species amplification in other passerine birds. *Molecular Ecology*, **9**, 2225–2230.

Richardson DS, Komdeur J, Burke T, von Schantz T (2005) MHC-based patterns of social and extra-pair mate choice in the Seychelles warbler. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **272**, 759–767.

Rousset F (2008) GENEPOP’007: A complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Schad J, Ganzhorn JU, Sommer S (2005) Parasite burden and constitution of major histocompatibility complex in the Malagasy mouse lemur, *Microcebus murinus*. *Evolution*, **59**, 439–450.

von Schantz T, Wittzell H, Goransson G, Grahn M, Persson K (1996) MHC genotype and male ornamentation: genetic evidence for the Hamilton-

Zuk Model. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 265–271.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.

Schroeder J, Burke T, Mannarelli M-E, Dawson DA, Nakagawa S (2012) Maternal effects and heritability of annual productivity. *Journal of Evolutionary Biology*, **25**, 149–156.

Schroeder J, Cleasby IR, Nakagawa S, Ockendon N, Burke T (2011) No evidence for adverse effects on fitness of fitting passive integrated transponders (PITs) in wild house sparrows *Passer domesticus*. *Journal of Avian Biology*, **42**, 271–275.

Sepil I, Lachish S, Sheldon BC (2012) *Mhc*-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular Ecology*, **22**, 384–396.

Simons MJP, Winney I, Nakagawa S, Burke T, Schroeder J (2015) Limited catching bias in a wild population of birds with near-complete census information. *Ecology and Evolution*, **5**, 3500–3506.

Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in zoology*, **2**, 16–34.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **277**, 979–988.

Stuglik MT, Radwan J, Babik W (2011) jMHC: software assistant for multilocus genotyping of gene families using next-generation amplicon sequencing. *Molecular Ecology Resources*, **11**, 739–742.

Therneau TM (2015a) Survival Analysis. *R package 2.38*-3, https://cran.r-project.org/web/packages/survival.

Therneau TM (2015b) Mixed Effects Cox Models. *R package 2.2-5*, https://cran.r-project.org/web/packages/coxme.

Wegner KM, Reusch TBH, Kalbe M (2003) Multiple infections drive major histocompatibility complex polymorphism in the wild. *Journal of Evolutionary Biology*, **16**, 224–232.

Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology*, **17**, 485–492.

Westerdahl H, Waldenström J, Hansson B, Hasselquist D, von Schantz T, Bensch S (2005) Associations between malaria and MHC genes in a migratory songbird. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **272**, 1511–1518.

Zuur A, Ieno E, Walker NJ, Saveliev AA, Smith GM (2009) *Mixed effects models and extensions in ecology with R*. Springer Science, New York.

CHAPTER 5: General discussion

5.0 Summary of key findings

The main aims of this study were to optimise *Mhc* class I exon 3 genotyping using Illumina MiSeq next-generation sequencing in house sparrows (**Chapter 2**), to examine the selection pressure on *Mhc class I exon 3* genes in different populations of house sparrows and any association between variation at the locus and *Plasmodium* parasites (**Chapter 3**), and to investigate temporal variation at the *Mhc* and its association with fitness in a long-term house sparrow population study (**Chapter 4**).

5.0.1 Next-generation typing of the *Mhc*

Screening for complex polymorphic and polygenic avian *Mhc* alleles has been made easier by the use of next-generation sequencing techniques (Babik *et al.* 2009; Sepil *et al.* 2012b; Sommer *et al.* 2013). One side effect of these sequencing techniques is the generation of artefacts, which are either the result of PCR-biased reactions or the chemistry of the machinery itself. We showed that characterizing the number of *Mhc* alleles per individual was more stable at low *Mhc* number for both of the next-generation sequencing techniques examined. In contrast, there was inconsistency in replicated individuals with higher numbers of *Mhc* alleles in 454 amplicon sequencing, which was attributed to inadequate read depth. There was very high consistency in the genotyping of replicates in Illumina MiSeq sequencing. In this house sparrow study, *Mhc* class I exon 3 genotyping using the Illumina MiSeq platform combined with conservative *Mhc* genotyping (Lighten *et al.* 2014) was successful.

5.0.2 Malaria and its association with *Mhc*

Several avian studies have provided evidence of population differentiation at varying geographical scales as a result of either balancing selection or diversifying selection on *Mhc* class I exon 3 (Bonneaud *et al.* 2006b; Ekblom *et al.* 2007; Alcaide *et al.* 2008; Loiseau *et al.* 2009; Bichet *et al.* 2015). In the study presented here, of a set of naturalized house sparrow populations in New Zealand, we found evidence of selection acting on *Mhc* class I exon 3. Consistent with a previous study on malaria (Tompkins & Gleeson 2006), there was a gradual decrease in prevalence of malaria between the North and South Islands of New Zealand (from high to low prevalence). The northern part of the North Island (Auckland) had the highest number of infected individuals, attributed to the warm sub-tropical climate providing ideal conditions for mosquitoes (Laird 1995). However, there was no evidence of selection on specific *Mhc* alleles in the highly infected sparrow population, nor was there any association between the presence of low-frequency *Mhc* alleles and malaria. This population study suggested that balancing selection is the main force maintaining the variation at the *Mhc* class I exon 3.

Studies in other house sparrow populations found an association between infection and specific *Mhc* class I alleles (Bonneaud *et al.* 2006b; Loiseau *et al.* 2011). However, the methods used in these studies did not specify whether both ‘long’ (without 6-bp deletion) and ‘short’ (with 6-bp deletion) alleles were used. Loiseau *et al.* (2011) identified 2–10 *Mhc* class I alleles, while in this study we found 2–9 *Mhc* class I ‘long’ alleles. It is not surprising to find fewer alleles as the New Zealand population was founded from a small subsample of the European population (Long 1981), so that some alleles would be expected to have been lost. Additionally, the sample sizes and geographical scale in this study were smaller than those used by Loiseau *et. al.* (2011).

5.0.3 Temporal variation and survival analysis in the Lundy population

The presence of temporal variation in the frequencies of *Mhc* class I exon 3 alleles was previously discovered in great reed warblers (*Acrocephalus arundinaceus)* (Westerdahl *et al.* 2004); however, it was hypothesized that this could be attributed to selection pressure imposed by malarial parasites on the wintering grounds. Despite the apparent lack of malaria in the Lundy population (Westerdahl, Burke and Karlsson, unpubl.), in this study there were significant differences in *Mhc* allele frequencies among cohorts. Individuals that either had extremely low or high numbers of *Mhc* alleles had higher survival compared to individuals that had an intermediate number of *Mhc* alleles. High number of *Mhc* alleles supposedly increase the number of *Mhc* proteins to allow a wide range of antigens to be recognised (Nowak *et al.* 1992). In a temporal scale of 10 years (2000-2010), two *Mhc* alleles (Lseq01 and Lseq02) were found significantly associated with a fitness trait i.e. the number of offspring produced.

5.1 Improving the quality of amplification of *Mhc* alleles for Illumina sequencing

Simultaneously genotyping multiple loci has been made possible by next-generation sequencing. The method proved to be less time-consuming compared to traditional methods, such as cloning and sequencing (Wegner 2009). However, there are problems associated with the data analysis. For example, some amplicons were removed from further analysis when they could not be clearly separated from artefacts. This may be the result of either low DNA quality and/or contamination, such as carry-over of DNA or amplified PCR products between samples during DNA preparation, PCR or sequencing (Champlot *et al.* 2010; Knapp *et al.* 2012). Protocols and infrastructure that have been developed for analysis of Ancient DNA – where contamination is a serious risk – may provide the key to reducing contamination in samples; for example, by using a separate location for DNA extraction and the pre-PCR setup, isolated from the post-PCR laboratory (Knapp *et al.* 2012).

PCR techniques for minimizing artefacts include reducing the number of PCR cycles (Meyerhans *et al.* 1990; Judo *et al.* 1998; Lenz & Becker 2008). Another method that can further reduce artefact formation is to dilute the initial PCR product then run a second PCR with a small number of cycles (Thompson *et al.* 2002; Kanagawa 2003).

There was a need to assess genotyping error by sequencing a fraction of the PCR amplicons in an independent run. Sometimes individuals replicated in additional runs are found to have additional alleles that were discarded in the original run due to low sequencing depth (Lighten *et al.* 2014). The method used to distinguish actual alleles from artefacts depends on their relative sequencing depths. The weakness of this method therefore lies in the risk of discarding real alleles as contaminants if the sequencing depth is too low (Lighten *et al.* 2014).

5.2 Multiple-parasite infections

This study demonstrated successful malarial screening using Illumina MiSeq sequencing. Illumina MiSeq sequencing detected higher numbers of infected individuals than the gel-based method. In spite of that, the primer pair used in this study only amplified *Plasmodium spp.* Wild birds have been shown to harbour multiple malarial parasites (Pérez-Tris & Bensch 2005). Evidence was found in a previous study of infected blackcaps (*Sylvia atricapilla*) and garden warblers (*Sylvia borin*) that most individuals with multiple infection were infected with two strains; however some individuals are infected with three or four strains (Pérez-Tris & Bensch 2005). We found four different lineages belonging to *Plasmodium relictum* in this house sparrow population. Individuals were infected with between one and four (rarely only one).

Island populations usually harbour fewer parasites than their mainland counterparts (Poulin 1997). Previous work on the Lundy house sparrows did not find any malarial infections in this population using a PCR-based detection method (Westerdahl, Burke and Karlsson, unpubl.). Even though malaria was not found in this population, other pathogens might mediate selection on the *Mhc*. For example, parasites belonging to isosporoid coccidians (Schrenzel *et al.* 2005) have never been investigated in this population. Mites are the main carriers of this parasite. The parasite enters the blood, colonizes the intestines and spreads further via the blood-circulating leukocytes to other organs. Isosporoid coccidians have different life stages in different affected tissues; for example, the intestines contain sexual and asexual forms, so that the faeces contain isosporoid oocysts, and extraintestinal regions contain merozoites (Schrenzel *et al.* 2005). An easy way to test whether a bird is infected is to check for the parasite's oocysts in the faeces. In the study of montane voles (*Microtus montanus*) found an association between an intestinal parasite (cestode) intensity and a specific *Mhc* allele (i.e. Mimo-DRB*07) (Winternitz *et al.* 2014).

5.3 Disentangling the forces to explain variation at *Mhc*

How genetic variation is maintained in the wild has been a major question in evolutionary biology. The genes of the *Mhc* have been used to study the different types of selection operating to maintain genetic variation in populations. Variation at the *Mhc* loci is considered to be maintained mainly by pathogen-mediated selection (PMS) (Doherty & Zinkernagel 1975; Bernatchez & Landry 2003). Three mechanisms of PMS have been proposed: heterozygote advantage (Doherty & Zinkernagel 1975), rare-allele advantage (Hamilton 1980; Slade & McCallum 1992) and fluctuating selection (Hedrick 2002; Bernatchez & Landry 2003). However, it is not an easy matter to identify and differentiate between the mechanisms empirically, especially in wild populations (Bernatchez & Landry 2003; Piertney & Oliver 2005; Spurgin & Richardson 2010). Spurgin and Richardson (2010) provided insights into possible empirical studies of selection on *Mhc* genes. Studies have attempted to differentiate the mechanisms to explain *Mhc* diversity, but with limited success.

Rare-allele advantage is a process by which, as the name implies, rare alleles within a population have a selective advantage over common alleles to provide resistance against pathogens (Takahata & Nei 1990). There is pressure on pathogens to evolve to overcome the immunity provided by the common *Mhc* alleles. These common *Mhc* alleles were initially rare alleles but have increased in frequency due to their initial selective advantage, but will in turn decrease in frequency as the pathogens evolve resistance against them after they become more common (Takahata & Nei 1990; Slade & McCallum 1992). Eventually, the common *Mhc* allele becomes rare again but does not disappear from the population (Slade & McCallum 1992). There is, then, an evolutionary arms race between pathogens and *Mhc* alleles – a dynamic process maintaining *Mhc* diversity (Slade & McCallum 1992). As an example, two *Mhc* alleles (*a151* and *a172*) from two different French house sparrow (*Passer domesticus*) population were associated with resistance against the same malarial strain (SGS1) (Bonneaud *et al.* 2006b). However, the other two hypotheses (heterozygote advantage and fluctuating selection) were not ruled out.

The fluctuating selection hypothesis proposes that the variation in pathogen abundance and diversity across space and time drives the maintenance of *Mhc* diversity (Hedrick 2002). This has resulted in different *Mhc* alleles being selected for at different points in space and/or time due to a fluctuating pathogen regime (Spurgin & Richardson 2010). The selection is directional and the fluctuation in pathogen types is not due to coevolution between host and pathogen, but from variation in the biotic or abiotic environment, dispersal by chance, and extinction events (Spurgin & Richardson 2010). Another study on thirteen house sparrow populations across France showed that, in five populations, several *Mhc* alleles were associated with the prevalence of *Plasmodium* infection (Loiseau *et al.* 2011). The authors concluded that fluctuating selection drives the patterns of local adaptation of these *Mhc* alleles (Loiseau *et al.* 2011). However, the pattern of selection on different *Mhc* alleles due to the prevalence of one *Plasmodium* type is also consistent with the predictions of the rare-allele advantage hypothesis.

Being heterozygous at the polymorphic *Mhc* loci confers enhanced resistance to infections by increasing the diversity of antigens presented to the T lymphocytes (Penn *et al.* 2002). Heterozygote advantage is defined as heterozygotes, on average, having higher fitness than homozygotes (Doherty & Zinkernagel 1975; Penn *et al.* 2002; Oliver *et al.* 2009). For example, water voles (*Arvicola terrestris*) that had only two *Mhc* alleles (three possible genotypes) in the population showed a significant association between individual *Mhc* genotype and parasite burden (gamasid mites, *Megabothris walker* fleas and *Ixodes ricinius* tick nymphs) (Oliver *et al.* 2009). In addition, the study also showed that, during co-infection, *Mhc* heterozygotes had fewer parasites compared to homozygotes (Oliver *et al.* 2009). This study showed evidence of different modes of selection (heterozygote superiority) acting simultaneously on the *Mhc*; however, the factor that maintained the genetic diversity in water voles was greater fitness of heterozygotes than homozygotes, i.e. heterozygote superiority. To date, this is the only study that has demonstrated *Mhc* heterozygote superiority against multiple infections in a natural population. However, a significant association was found between a specific *Mhc* allele (Arte-DRB*05) and gamasid mite infection (Oliver *et al.* 2009), indicating that rare-allele advantage and/or fluctuating selection may also be operating (Spurgin & Richardson 2010).

These hypotheses are not mutually exclusive from each other. For example, the frequency-dependent component is also found within heterozygote superiority, where rare alleles occur disproportionately in heterozygous genotypes, meaning that heterozygotes may be selected for because they carry rare, disease-resistant alleles (Apanius *et al.* 1997; Penn 2002). There are also situations where *Mhc* alleles under heterozygote or rare-allele advantage may also vary in frequency in space and time due to fluctuating selection (Spurgin & Richardson 2010). In any system where multiple selective forces are acting on allele frequencies, the selective forces may act in opposite directions, thus masking each other’s effects and resulting in patterns that do not deviate from neutral expectation (Apanius *et al.* 1997; Spurgin & Richardson 2010).

Distinguishing rare-allele advantage and fluctuating selection would require studying the *Mhc* and neutral variation in relation to parasite load over a long period of time across multiple replicate populations (Spurgin & Richardson 2010). Under the rare-allele advantage, I would expect different alleles to confer resistance against the same pathogen found in different locations. The resistance will change in time so different alleles would, in turn, be associated with the resistance against the same pathogen (Spurgin & Richardson 2010). With fluctuating selection, I would expect differences in the spatial and temporal variation of pathogen abundance, due to external abiotic and/or biotic forces, to lead to different alleles being selected for in different populations and/or at different periods of time (Spurgin & Richardson 2010). Appropriate population studies are not only difficult to establish but conducting them over a long period of time in multiple locations is also costly.

According to Spurgin & Richardson (2010), distinguishing heterozygote advantage requires examining the association between pathogens and both genotypes and specific alleles. The appropriate *Mhc* screening method is a single-locus type amplification. The lack of evidence in the literature reflects the difficulty of identifying appropriate study systems (but see Oliver *et al.* 2009).

Detecting parasite-driven selection on *Mhc* genes might also be difficult because the selection might be weak, requiring an enormous sample size (Apanius *et al.* 1997). In addition, multiple genes of the *Mhc* are a result of repeated gene duplication (Kasahara 1997; Nei *et al.* 1997); the variation is generated by positive selection on nucleotide mutations in the peptide-binding region (Hughes & Nei 1988), and gene conversion between duplicate loci and within loci (Ohta 1991). This degree of complexity makes it impossible to take a locus-specific approach to characterizing the *Mhc* (Hess & Edwards 2002; Spurgin & Richardson 2010). In species with many *Mhc* loci, the genes are usually tightly linked to one another and the *Mhc* alleles are shared among the loci (Sepil *et al.* 2012a). The question has therefore been raised of what constitutes a functionally important *Mhc* allele (Spurgin & Richardson 2010)? This has been tackled by grouping the alleles into functional supertypes (Doytchinova & Flower 2005); as seen in great tits, certain *Mhc* supertypes confer resistance against malaria parasites (Sepil *et al.* 2013). An alternative to analysing large numbers of *Mhc* alleles is to identify and group alleles into the supertypes.

In this study of New Zealand house sparrows, it was predicted that rare-allele advantage and/or fluctuating selection might be observed in locations with malaria present; however, no significant associations were found between malaria and *Mhc* alleles. We found no association between *Mhc* diversity (measure of heterozygosity in *Mhc*) and the prevalence of malaria. We did not investigate the effects of different malaria strains in this population, unlike the study of great reed warblers in which a positive association was found between a specific malaria parasite (GRW2) and *Mhc* diversity, leading to the conclusion that a large number of *Mhc* alleles conferred protection against the parasite (Westerdahl *et al.* 2005). This was because the large number of *Mhc* alleles in the Lundy population (54) made it difficult statistically to detect the effects of heterozygosity (as almost individuals are heterozygous). The large number of *Mhc* alleles does suggest the presence of a stable polymorphism, which may be the result of heterozygote advantage (Slade & McCallum 1992).

In this intensively monitored (since 2000) Lundy house sparrow population, we found significant genetic differentiation at the *Mhc* between the cohorts. However, there was no association between specific *Mhc* alleles/and or *Mhc* diversity and survival, but an association was found between two *Mhc* alleles (Lseq01 and Lseq02) and reproductive fitness (i.e. number of offspring produced). In addition, frequency of some of *Mhc* alleles varied more between cohorts than expected, suggesting that selection favours different *Mhc* alleles in different years. This led to the suggestion that there might be both frequency-dependent and fluctuating selection within this island population.

5.4 Future research

5.4.1 Malaria from here onwards

This study showed that haemosporidian parasites are easily screened using Illumina MiSeq sequencing. The advantage of this method is that it can detect low-intensity infections. This can be a tool for initial screening and as a check of whether individuals are infected with malaria or not. To shed some light on the dynamics among the different co-infected malarial lineages, specific primer designs are needed to quantify infection intensity (parasitaemia level) using qPCR. This could provide a better insight into the interaction between malarial infection and the *Mhc*.

Multiple infections were found in house sparrows using Illumina MiSeq sequencing, revealing that different parasites are present in this host species. This technique of identifying malarial parasites can be used to address questions on host–parasite interactions, such as how many malarial lineages can infect a host, or how do coexisting lineages affect virulence of different parasites (Poulin 1997; Pérez-Tris & Bensch 2005)?

5.4.2 *Mhc*-dependent patterns of mate choice and extra-pair fertilization

Mammals are able to discriminate, using olfactory cues, among individuals with different *Mhc* genotypes, and this results in non-random mate choice based on *Mhc* genotype, mainly driven by inbreeding avoidance (Penn *et al.* 1999). For example, female house mice (*Mus mus domesticus*) avoid mating with males that are genetically similar at the *Mhc* (Egid & Brown 1989; Penn & Potts 1998). However, the question remains whether females can discriminate *Mhc*-similar males directly or if this is the result of familial imprinting (Penn *et al.* 1999). A cross-fostering experiment in house mice showed that cross-fostered females were more likely to mate with *Mhc*-similar males than *Mhc*-dissimilar males that were identical to the female's foster families. The study provided evidence that familial imprinting, but not olfactory cues, is an effective mechanism for avoiding mating with *Mhc*-similar mates.

Birds are usually thought to be anosmic, i.e. lacking the ability to perceive odour (Zelano & Edwards 2002). However, several studies have revealed that some avian species do use odours, for instance during foraging (Nevitt 1999), selection of nesting materials (Petit *et al.* 2002) and to discriminate among conspecifics (Hagelin *et al.* 2003; Bonadonna & Nevitt 2004). In addition, a recent experimental study demonstrated that blue tits could detect changes in aromatic odour composition when different aromatic leaves were added to their nests (Mennerat 2008). In a study of blue petrels (*Halobaena caerulea*), a bird species that relies on odour cues in foraging, they were found to choose mates with functionally dissimilar *Mhc* class II genes (Strandh *et al.* 2012).

Mhc genes not only play an important role in the immune system but they have also been shown to affect mate-choice preferences (Penn 2002). Variability at the *Mhc* is maintained via mate-choice patterns, such as preferring heterozygotes or avoiding partners with similar *Mhc* alleles (Bonneaud *et al.* 2006a). *Mhc* genes may therefore provide an important mechanistic link between mate choice and the indirect genetic benefits that drive mate choice (Brouwer *et al.* 2010). For example, indirect benefits are proposed to include the increased survival of offspring when females mate with specific male genotypes. The earliest study on *Mhc*-dependent mate choice was on laboratory mice, which showed that female mice preferred to mate with males with dissimilar *Mhc* haplotypes from their own (Egid & Brown 1989). Such *Mhc*-disassortative mating may be involved in maintaining *Mhc* diversity, and a general preference for dissimilar mates (Jordan & Bruford 1998). Hypotheses have been proposed to explain an association between diverse *Mhc* genes and mate choice, such as good genes, genetic compatibility, inbreeding and outbreeding avoidance (Bonneaud *et al.* 2006a).

In mate choice situations where direct genetic benefits fail as an explanation, it has instead been proposed that individuals can gain indirect genetic benefits through mate choice (Charmantier & Sheldon 2006). By participating in extra-pair copulation, for example, females might gain benefits through improved fitness of offspring due to the superior genetic properties of the extra-pair males (Griffith *et al.* 2002; Kempenaers 2007). Indirect benefits can be gained by acquiring good paternal genes from the extra-pair sire (Petrie 1994) or from the enhanced genetic compatibility of both maternal and paternal genomes (Zeh & Zeh 1996).

Evidence of *Mhc*-based extra-pair mating comes from a study on savannah sparrows (*Passerculus sandwichensis*) where it was concluded, by detecting extra-pair mate choice, that females prefer genetically dissimilar males (Freeman-Gallant *et al.* 2003). Young females preferred *Mhc*-dissimilar males and the tendency of females to produce extra-pair offspring increased with increased genetic similarity (Freeman-Gallant *et al.* 2003). Another study, on Seychelles warblers (*Acrocephalus sechellensis*), did not find evidence that individuals maximized their offsprings' *Mhc* diversity by mating disassortatively. Instead, it showed that females were more likely to gain extra-pair paternity when the social male had low *Mhc* diversity (Richardson *et al.* 2005). The study also suggested that extra-pair paternity appeared to be influenced by male *Mhc* diversity (Richardson *et al.* 2005). A study on the Lundy house sparrows found, however, that there was a fitness cost associated with having extra-pair offspring (Hsu *et al.* 2014). Given the availability of both *Mhc* and paternity data in this population, the potential for *Mhc*–driven mate choice patterns should now be investigated.

5.5 References

Alcaide M, Edwards SV, Negro JJ, Serrano D, Tella JL (2008) Extensive polymorphism and geographical variation at a positively selected MHC class II B gene of the lesser kestrel (*Falco naumanni*). *Molecular Ecology*, **17**, 2652–2665.

Apanius V, Penn D, Slev PR, Ruff LR, Potts WK (1997) The nature of selection on the major histocompatibility complex. *Critical Reviews in Immunology*, **17**, 179–224.

Babik W, Taberlet P, Ejsmond MJ, Radwan J (2009) New generation sequencers as a tool for genotyping of highly polymorphic multilocus MHC system. *Molecular Ecology Resources*, **9**, 713–719.

Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, **16**, 363–377.

Bichet C, Moodley Y, Penn DJ, Sorci G, Garnier S (2015) Genetic structure in insular and mainland populations of house sparrows (*Passer domesticus*) and their hemosporidian parasites. *Ecology and Evolution*, **5**, 1639–1652.

Bonadonna F, Nevitt GA (2004) Partner-Specific Odor Recognition in an Antarctic Seabird. *Science*, **306**, 835.

Bonneaud C, Chastel O, Federici P, Westerdahl H, Sorci G (2006a) Complex Mhc-based mate choice in a wild passerine. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **273**, 1111–1116.

Bonneaud C, Pérez-Tris J, Federici P, Chastel O, Sorci G (2006b) Major histocompatibility alleles associated with local resistance to malaria in a passerine. *Evolution*, **60**, 383–389.

Brouwer L, Barr I, Van De Pol M, Burke T, Komdeur J, Richardson DS (2010) MHC-dependent survival in a wild population: Evidence for hidden genetic benefits gained through extra-pair fertilizations. *Molecular Ecology*, **19**, 3444–3455.

Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M

(2010) An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. *PLoS ONE*, **5**, e13042.

Charmantier A, Sheldon BC (2006) Testing genetic models of mate choice evolution in the wild. *Trends in Ecology & Evolution*, **21**, 417–419.

Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, **256**, 50–52.

Doytchinova IA, Flower DR (2005) In Silico Identification of Supertypes for Class II MHCs. *The Journal of Immunology*, **174**, 7085–7095.

Egid K, Brown JL (1989) The major histocompatibility complex and female mating preferences in mice. *Animal Behaviour*, **38**, 548–549.

Ekblom R, Sæther SA, Jacobsson P, Fiske P, Sahlman T, Grahn M, Kålas JA, Höglund J (2007) Spatial pattern of MHC class II variation in the great snipe (*Gallinago media*). *Molecular Ecology*, **16**, 1439–1451.

Freeman-Gallant CR, Meguerdichian M, Wheelwright NT, Sollecito S V (2003) Social pairing and female mating fidelity predicted by restriction fragment length polymorphism similarity at the major histocompatibility complex in a songbird. *Molecular Ecology*, **12**, 3077–3083.

Griffith SC, Owens IPF, Thuman KA (2002) Extra pair paternity in birds: a review of interspecific variation and adaptive function. *Molecular Ecology*, **11**, 2195–2212.

Hagelin JC, Jones IL, Rasmussen LEL (2003) A tangerine-scented social odour in a monogamous seabird. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **270**, 1323–1329.

Hamilton WD (1980) Sex versus Non-Sex versus Parasite. *Oikos*, **35**, 282–290.

Hedrick PW (2002) Pathogen Resistance and Genetic Variation at MHC Loci. *Evolution*, **56**, 1902–1908.

Hess CM, Edwards SV (2002) The Evolution of the Major Histocompatibility Complex in Birds: Scaling up and taking a genomic approach to the major histocompatibilty complex (MHC) of birds reveals surprising departures from generalities found in mammals in both large-scale structure. *BioScience*, **52**, 423–431.

Hsu YH, Schroeder J, Winney I, Burke T, Nakagawa S (2014) Costly infidelity: Low lifetime fitness of extra-pair offspring in a passerine bird. *Evolution*, **68**, 2873–2884.

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, **335**, 167–170.

Jordan WC, Bruford MW (1998) New perspectives on mate choice and the MHC. *Heredity*, **81**, 239–245.

Judo MS, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. *Nucleic Acids Research*, **26**, 1819–1825.

Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317–323.

Kasahara M (1997) New insights into the genomic organization and origin of the major histocompatibility complex : Role of chromosomal (genome) duplication in the emergence of the adaptive immune system. *Hereditas*, **127**, 59–65.

Kempenaers B (2007) Mate choice and genetic quality: A review of the heterozygosity theory. *Advances in the Study of Behavior*, **37**, 189–278.

Knapp M, Clarke AC, Horsburgh KA, Matisoo-Smith EA (2012) Setting the stage – Building and working in an ancient DNA laboratory. *Annals of Anatomy*, **194**, 3–6.

Laird M (1995) Background and findings of the 1993-94 New Zealand mosquito survey. *New Zealand Entomologist*, **18**, 77–90.

Lenz TL, Becker S (2008) Simple approach to reduce PCR artefact formation leads to reliable genotyping of MHC and other highly polymorphic loci - implications for evolutionary analysis. *Gene*, **427**, 117–123.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Molecular Ecology Resources*, **14**, 1–15.

Loiseau C, Richard M, Garnier S, Chastel O, Julliard R, Zoorob R, Sorci G (2009) Diversifying selection on MHC class I in the house sparrow (*Passer domesticus*). *Molecular Ecology*, **18**, 1331–1340.

Loiseau C, Zoorob R, Robert A, Chastel O, Julliard R, Sorci G (2011) *Plasmodium relictum* infection and MHC diversity in the house sparrow (*Passer domesticus*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **278**, 1264–1272.

Long JL (1981) *Introduced birds of the world: the worldwide history, distribution and influence of birds introduced to new environments*. David & Charles, London.

Mennerat A (2008) Blue tits (*Cyanistes caeruleus*) respond to an experimental change in the aromatic plant odour composition of their nest. *Behavioural processes*, **79**, 189–191.

Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. *Nucleic Acids Research*, **18**, 1687–1691.

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 7799–7806.

Nevitt G (1999) Foraging by Seabirds on an Olfactory Landscape: The seemingly featureless ocean surface may present olfactory cues that help the wide-ranging petrels and albatrosses pinpoint food sources. *American Scientist*, **87**, 46–53.

Nowak MA, Tarczy-Hornoch K, Austyn JM (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**, 10896–10899.

Ohta T (1991) Role of Diversifying Selection and Gene Conversion in Evolution of Major Histocompatibility Complex Loci. *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 6716–6720.

Oliver MK, Telfer S, Piertney SB (2009) Major histocompatibility complex (MHC) heterozygote superiority to natural multi-parasite infections in the water vole (*Arvicola terrestris*). *Proceedings of the Royal Society of London Series B: Biological Sciences*, **276**, 1119–1128.

Penn DJ (2002) The scent of genetic compatibility: sexual selection and the major histocompatibility complex. *Ethology*, **108**, 1–21.

Penn DJ, Damjanovich K, Potts WK (2002) MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11260–11264.

Penn D, Potts W (1998) MHC-disassortative mating preferences reversed by cross-fostering. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **265**, 1299–1306.

Penn DJ, Potts WK, Palumbi AESR (1999) The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. *The American Naturalist*, **153**, 145–164.

Pérez-Tris J, Bensch S (2005) Diagnosing genetically diverse avian malarial infections using mixed-sequence analysis and TA-cloning. *Parasitology*, **131**, 15–23.

Petit C, Hossaert-McKey M, Perret P, Blondel J, Lambrechts MM (2002) Blue tits use selected plants and olfaction to maintain an aromatic environment for nestlings. *Ecology Letters*, **5**, 585–589.

Petrie M (1994) Improved growth and survival of offspring of peacocks with more elaborate trains. *Nature*, **371**, 598–599.

Piertney SB, Oliver MK (2005) The evolutionary ecology of the major histocompatibility complex. *Heredity*, **96**, 7–21.

Poulin R (1997) Species richness of parasite assemlages: evolution and patterns. *Annual Review of Ecology and Systematics*, **28**, 341–358.

Richardson DS, Komdeur J, Burke T, von Schantz T (2005) MHC-based patterns of social and extra-pair mate choice in the Seychelles warbler. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **272**, 759–767.

Schrenzel MD, Mallouf GA, Gaffney PM, Tokarz D, Keener LL, McClure D, Griffey S, McAloose D, Rideout BA (2005) Molecular characterization of *Isosporoid coccidia* (*Isospora* and *Atoxoplasma* spp.) in passerine birds. *The Journal of Parasitology*, **91**, 635–647.

Sepil I, Lachish S, Hinks AE, Sheldon BC (2013) *Mhc* supertypes confer both qualitative and quantitative resistance to avian malaria infections in a wild bird population. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **280**, 1–8.

Sepil I, Lachish S, Sheldon BC (2012a) *Mhc*-linked survival and lifetime reproductive success in a wild population of great tits. *Molecular Ecology*, **22**, 384–396.

Sepil I, Moghadam HK, Huchard E, Sheldon BC (2012b) 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**, 1–19.

Slade RW, McCallum HI (1992) Overdominant Vs. Frequency-Dependent Selection at Mhc Loci. *Genetics*, **132**, 861–862.

Sommer S, Courtiol A, Mazzoni CJ (2013) MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics*, **14**, 1–17.

Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **277**, 979–988.

Strandh M, Westerdahl H, Pontarp M, Canbäck B, Dubois M-P, Miquel C, Taberlet P, 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 of London Series B: Biological Sciences*, **279**, 4457–4463.

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, **124**, 967–978.

Thompson JR, Marcelino LA, Polz MF (2002) Heteroduplexes in mixed-template amplifications: formation, consequence and elimination by “reconditioning PCR.” *Nucleic Acids Research*, **30**, 2083–2088.

Tompkins DM, Gleeson DM (2006) Relationship between avian malaria distribution and an exotic invasive mosquito in New Zealand. *Journal of the Royal Society of New Zealand*, **36**, 51–62.

Wegner KM (2009) Massive parallel MHC genotyping: titanium that shines. *Molecular Ecology*, **18**, 1818–1820.

Westerdahl H, Hansson B, Bensch S, Hasselquist D (2004) Between-year variation of MHC allele frequencies in great reed warblers: selection or drift? *Journal of Evolutionary Biology*, **17**, 485–492.

Westerdahl H, Waldenström J, Hansson B, Hasselquist D, von Schantz T, Bensch S (2005) Associations between malaria and MHC genes in a migratory songbird. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **272**, 1511–1518.

Winternitz JC, Wares JP, Yabsley MJ, Altizer S (2014) Wild cyclic voles maintain high neutral and MHC diversity without strong evidence for parasite-mediated selection. *Evolutionary ecology*, **28**, 957–975.

Zeh JA, Zeh DW (1996) The Evolution of Polyandry I: Intragenomic Conflict and Genetic Incompatibility. *Proceedings of the Royal Society of London Series B: Biological Sciences*, **263**, 1711–1717.

Zelano B, Edwards S V (2002) An Mhc component to kin recognition and mate choice in birds: predictions, progress, and prospects. *The American Naturalist*, **160**, S225–S237.