Associations between allelic variation at major histocompatibility complex I and inbreeding and fitness components in the house sparrow (*Passer domesticus*)

Petra Nieminen

MSc Thesis

University of Oulu

Faculty of Science

Department of Biology

June 2021

Table of contents

Abstract

Major histocompatibility complex (MHC) is a widely studied multigene family that is found in all vertebrates. The main role of MHC genes is in the adaptive immune system, but they are also known to have a role in other processes, for example in mate choice. In this thesis the main goal is to find out how inbreeding affects the number of MHC class I alleles (exon 3 sequences) in house sparrow (*Passer domesticus*) and whether individual allelic variation at MHC is associated to components of individual fitness. The data used in this thesis was collected between the years 1998 and 2013 in six house sparrow populations in an archipelago of northern Norway. Associations between the nucleotide and amino acid sequence numbers and inbreeding, lifespan and lifetime reproductive success were studied using regression analyses. In total of 127 unique nucleotide and 113 unique amino acid sequences (i.e., alleles) were found at MHC class I exon 3, with on average 12.6 (SD ± 3.5) nucleotide and 11.4 (SD ± 2.9) amino acid alleles per individual. Positive associations were found between the number of nucleotide alleles per individual and lifetime reproductive success, but there was only weak evidence for the association between number of amino acid alleles and reproductive success. There was no evidence for associations between the individual numbers of either type of alleles and inbreeding or lifespan. The positive association found between the number of MHC alleles and reproductive success could be explained by ecological traits that may be positively affected by a higher number of MHC alleles; for example, the ability to defend a wider range of pathogens and thus produce more and healthier offspring. It is also possible that there could be associations between MHC alleles and other fitness-components that are not detected using my analysis methods, for example if specific MHC alleles are advantageous or an intermediate number of alleles is the most beneficial.

1 Introduction

Genetic diversity may decrease if populations are fragmented and population sizes stay small for several generations (Keller & Waller, 2002). Small populations are threatened by genetic problems caused by inbreeding, because mating possibilities are limited, and genetic drift can lead to fixation or loss of alleles from populations faster than in large populations. Selection is also weaker in small populations, which increases the risk that deleterious mutations become more common in the population by chance. Increasing rate of inbreeding can lead to inbreeding depression, i.e., reduction in fitness (Keller & Waller, 2002).

Human activity, like urbanization, is known to cause fragmentation of populations, disrupt movement and hence reduce geneflow between subpopulations (Liu *et al.* 2020; Richardson *et al.* 2020). Preserving genetic variation in small populations is important, because it helps them to adapt to environmental changes. Adaptation increases the fitness and survival of individuals and therefore the survival of populations (Frankham, 2005).

Preserving genetic variation in immune loci, such as the major histocompatibility complex (MHC) genes, is likely to be important since they have crucial role in the immune system (Stervander *et al.* 2020). Little is still known about how inbreeding affects MHC variation and how MHC variation affects fitness in wild populations.

1.1 MHC genes

The major histocompatibility complex is a diverse multigene family found in all vertebrates (Bernatchez & Landry, 2003). Different variants of MHC genes are found to be associated to multiple biological traits in different species, for example immune recognition, autoimmune diseases, mate choice and individual odour (Jordan & Bruford 1998; Radwan *et al.* 2008; Wieczorek *et al.* 2017). The classical MHC family consists of two subfamilies, MHC class I and class II genes, that are linked together in one gene complex in mammals and birds (Bernatchez & Landry, 2003).

1.1.1. MHC class I and II

The MHC class I and II proteins are essential for the adaptive immune system since they introduce peptides for T cells to be recognized (Wieczorek *et al.* 2017). Class I genes are expressed on the cell surface of all nucleated cells, and they help the immune system to recognize infected cells by introducing peptide fragments from intracellular proteins on the cell surface (Bernatchez & Landry, 2003; Hewitt, 2003). Normally, the peptides presented on the cell surface are from the cell's own proteins. If peptides from pathogenic proteins, for example viral proteins, are presented, cytotoxic T lymphocytes (CTL) recognize them and eliminate the cells that are infected (Hewitt, 2003).

Class II genes are expressed on cells of the immune system that present antigens, like macrophages and B cells. Class II genes also monitor the extracellular environment and present peptides from extracellular proteins to helper T cells (Bernatchez & Landry, 2003). Associations between pathogen infections and MHC class I and II genes have been studied in different systems and are also found in house sparrow (Hedrick, 2002; Loiseau *et al.* 2011; Borg *et al.* 2011). For example, several MHC class I alleles are found to affect the probability of getting an avian malaria parasite (*Plasmodium relictum*) infection in house sparrow (Loiseau *et al.* 2011).

1.1.2 MHC class I antigen presentation pathway and exon 3

The antigen presentation pathway of the MHC class I molecules begins with proteolysis, in which proteasome degrades proteins to peptides (Hewitt 2003). Peptide fragments are associated to transporter associated with antigen processing (TAP) protein that transfers peptide fragments into endoplasmic reticulum (ER). In ER lumen, chaperone proteins help MHC class I molecules to fold and assemble, and a complex of MHC class I molecule and chaperones ERP57 and calreticulum is attached to TAP. Tapasin assists to bind peptide fragments to MHC class I molecules, after which the TAP separates and MHC class I molecules are transported in vesicles at first to the Golgi apparatus and from there to the plasma membrane.

MHC class I exon 3 sequences encode parts of the peptide binding region in MHC molecules (Richardson & Westerdahl, 2003). The diversity of MHC molecules is important because each

molecule recognizes only peptides that can bind into its peptide-binding groove (Woelfing *et al.* 2009).

1.2 Balancing selection and MHC diversity

Balancing selection works to preserve genetic diversity in populations through adaptive forces (Koenig *et al.* 2019). Balancing selection may enable to maintain MHC gene diversity even in populations that are very restricted (Richardson & Westerdahl, 2003). Balancing selection is suggested to occur via pathogen resistance and reproductive mechanisms by improving the immune response and decreasing inbreeding in offspring (Westerdahl *et al.* 1999). Parts of MHC genes that code for peptide-binding region (PBR) are very variable sites and they are thought to undergo balancing selection (Westerdahl *et al.* 1999). MHC diversity determines in part how individuals' immune system can recognize pathogens and defend against them. The optimal number of expressed MHC alleles should probably not be too high or too low, since high diversity can predispose to self-reactivity, but low diversity limits the number of presented antigens (O'Connor & Westerdahl 2021).

Pathogen-driven selection and mate choice are the main hypotheses explaining the maintenance of MHC diversity (Loiseau *et al.* 2009). In pathogen-driven selection the MHC diversity can be maintained through three main mechanisms. In frequency-dependent selection pathogens are likely to adapt to the most common alleles, so there is supposed to be selective advantage for individuals with rare alleles (Clarke & Kirby, 1966; Hedrick, 2002). In heterozygote advantage, heterozygous individuals have different alleles at a given locus, whereas homozygous individuals have two similar copies of the same allele, and diverse alleles provide wider or more efficient protection against the pathogens (Carrington *et al.* 1999; Stear *et al.* 2005). A third way maintaining MHC diversity in pathogen-driven selection is diversifying selection, in which gene-pathogen interactions are likely to create local adaptation in space, time or both (Hedrick, 2002; Loiseau *et al.* 2009). This can be seen for example so that specific MHC alleles provide protection against a pathogen infection in a certain population but not in another (Loiseau *et al.* 2009). Loiseau and others (2011) found evidence that MHC I variation is most probably maintained in French house sparrow populations through diversifying selection. They found that selection for MHC alleles is local, occurring at population level (Loiseau *et al.* 2011).

According to Hedrick (2002), it is possible that these types of selection are not mutually exclusive.

Also mate choice can preserve MHC diversity through several mechanisms (Stervander *et al.* 2020). In the frequency-dependent way, mating partner has certain MHC alleles that are advantageous, but the advantage these alleles provide changes in time. In the second way, mating partner has a diverse set of MHC alleles, which ensures that the partner is healthy. In the third way, the MHC allele sets carried by two individuals can vary from very similar to highly different, and mate choice ensures that the set of MHC alleles is compatible in both partners. (Stervander *et al.* 2020). In MHC-based mate choice, individuals generally prefer mating partners that carry an MHC allele set different to their own, which enhances the heterozygosity of offspring and may increase their fitness (Ekblom *et al.* 2004).

1.3 Inbreeding and its consequences

Inbreeding means a situation where mating partners are related, i.e., they share common ancestors. Inbred individuals have higher probability of carrying two copies of the same allele in their genome, which increases risk of getting two copies of a deleterious allele and predisposes to inbreeding depression, that is a state where inbreeding causes reduction in fitness (Keller & Waller, 2002; Charlesworth & Willis, 2009). Inbreeding depression can manifest as lower survival and reduced offspring production in wild populations (Niskanen *et. al* 2020; Zilko *et al.* 2020). Inbreeding depression is a threat especially for small and fragmented populations that are more prone to changes in environmental conditions and fluctuations in demographic stochasticity (Keller & Waller, 2002). If gene flow is restricted or alleles affecting phenotypic features are lost via genetic drift, possibilities to adapt to varying environmental conditions are limited. These are threats that can affect the severity of inbreeding depression. If fitness is strongly decreased by inbreeding depression, it may increase the risk of extinction (Keller & Waller, 2002; Zilko *et al.* 2020).

Inbreeding avoidance may have an important role in species suffering from inbreeding depression (e.g., Rymešová *et al.* 2017). To avoid inbreeding, individuals may favour mating partners that have dissimilar set of MHC alleles. For example, Rymešová and others (2017) found that grey partridge (*Perdix perdix*) formed pairs with more dissimilar MHC class IIB loci than expected by random mate choice.

Inbreeding coefficient of an individual can be determined from pedigree information or from genetic data, preferably using genome-wide genetic markers (Keller & Waller, 2002; Kardos *et al.* 2016; Zilko *et al.* 2020). The higher the value of inbreeding coefficient is, the higher the rate of inbreeding is, for example value 0 is considered outbred, whereas value 0.125 can result from mating between half-siblings and 0.25 can result from mating between full-siblings (Huisman *et al.* 2016). Inbreeding coefficients are relative measures of inbreeding and are estimated compared to a reference population that is often the focal study population (Keller & Waller, 2002).

1.4 Study species *Passer domesticus*

The house sparrow (*Passer domesticus*) is a small passerine bird widely spread in Eurasia, America, Australia and Africa. Even if the current population trend for house sparrow is decreasing, the evaluation for the species is Least Concern due to a large range size and a population size of approximately one billion mature individuals world-wide. The house sparrow is mostly a resident species. It lives close to human impact, and it is found from rural areas to suburban areas and city centres (BirdLife International, 2021).

House sparrows eat different parts of plants, for example seed of grasses, cereals, buds and berries, but also arthropods during summer months (BirdLife International, 2021). In western Europe, the decline of the population size is associated to changes in cultivation practices. Both the use of pesticides and herbicides and sowing the cereal in the fall have reduced the invertebrate and vegetable food available.

The house sparrow is a colonial breeding species that prefers to build its nest into a hole, for example in a nest box or under a roof and in other crevices in buildings (BirdLife International, 2021). It lays on average five eggs and has from one to three clutches per nesting season (Ringsby *et al.* 2002; Husby *et al.* 2006). Chicks leave the nest by the age of about two weeks, but only 16-26% of hatched individuals survive to their first breeding year (Ringsby *et al.* 1998; Cleasby *et al.* 2010). Approximately 50-70% of adult house sparrows survive to the following year, but there is both temporal and spatial variation in survival (Sæther *et al.* 1999).

1.5 Aims of the thesis

In my thesis I will use individual data on house sparrows to study the possible connections between the number of MHC class I alleles (different exon 3 sequences) and inbreeding and fitness components. Specifically, the aims of this study are to 1) examine if inbreeding affects the number of different MHC alleles, and 2) test if the number of MHC alleles is associated to lifespan and offspring production.

My hypotheses are that inbreeding lowers MHC allele diversity due to increased genome-wide homozygosity. If parents are closely related to each other, it is very possible that their offspring will get the same copy of an allele from both parents which reduces their genome-wide level of variation. Because MHC alleles code proteins that present antigens for T cells (Wieczorek *et al.* 2017), it seems probable that individuals with many MHC alleles can recognize different pathogens better and therefore have higher survival probabilities and produce more offspring (i.e., have a selective advantage).

2 Materials and methods

2.1 Study system

This thesis is part of a long-term research project on wild house sparrow populations in northern Norway. The project has been ongoing since year 1993 on 18 research islands in an archipelago on the Helgeland coast (Ringsby *et al.* 2002; Jensen *et al.* 2013; Kvalnes *et al.* 2017; Niskanen *et al.* 2020). Approximately 90% of hatched house sparrows are ringed every year during the field work period from May to August. Coloured rings enable recognition of each bird also without being captured. Morphological measures, blood samples and faecal samples are taken from nestlings and adults when captured. Blood samples are collected from the brachial vein. The study system has been described in earlier studies (Jensen *et al.* 2003; Jensen *et al.* 2008; Kvalnes *et al.* 2017). The research database contains information on more than 20,000 individuals, and of these, 3116 individuals have been genotyped at a custom house sparrow Axiom 200K single nucleotide polymorphism (SNP) array (Lundregan *et al.* 2018; Niskanen *et al.* 2020). Ecological and genome-wide data of house sparrows are collected by the house sparrow group from Centre for Biodiversity Dynamics, NTNU, Norway. Data from six of their research islands were included in this study: Aldra, Gjerøy, Indre Kvarøy, Hestmannøy, Selvær and Træna (Figure 1). On farm islands Gjerøy, Indre Kvarøy, Hestmannøy and Aldra sparrows live on dairy farms, whereas on nonfarm islands Selvær and Træna the sparrows live in gardens in small villages (Niskanen *et al.* 2020). Population sizes of the smallest and the largest populations are approximately 30 and 140 adult individuals, respectively. The number of individuals from each population included in this study are listed in table 1, including the annual mean adult population sizes.

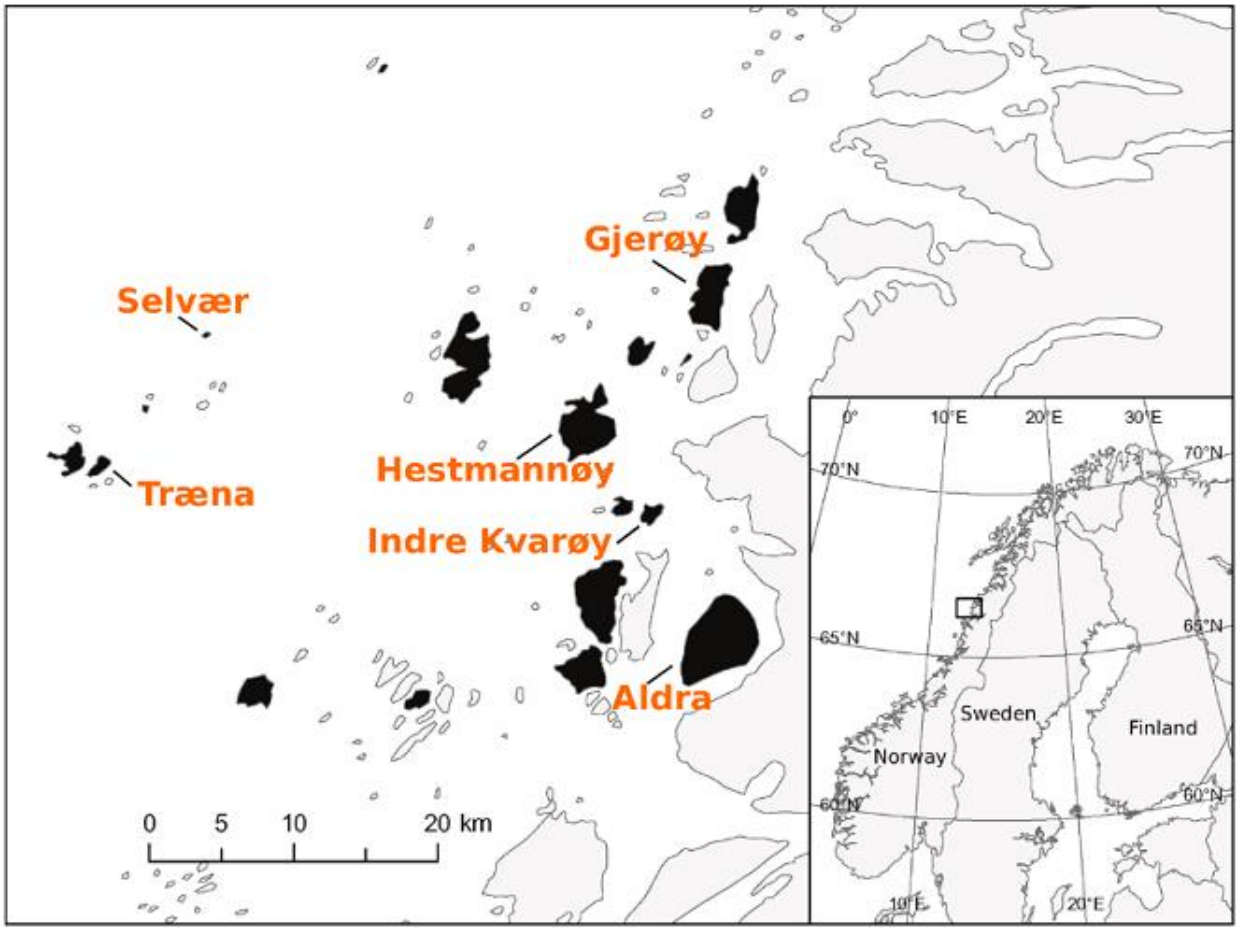

Figure 1. Research islands. All islands belonging in the long-term research project are coloured black. Named islands are the ones where the data was used in this study (map by Alina Niskanen).

2.2 Materials

Every bird included in this study were genotyped at a SNP panel that reflects genome-wide genetic variation. The SNP panel was custom developed for the house sparrow and contains 183 000 high-quality SNPs (Lundregan *et al.* 2018) that have been used to determine for example the inbreeding rate and genome-wide heterozygosity (Niskanen *et al.* 2020). All 3116 SNP-genotyped sparrows were used to construct a pedigree, using the 605 most informative SNPs that were also in low linkage disequilibrium (Niskanen *et al.* 2020). This subset of SNPs was selected based on PLINK (Purcell *et al.* 2007). Pedigree construction is described in the supplementary information of Niskanen *et al.* (2020).

Inbreeding coefficient calculations were performed by Alina Niskanen in the house sparrow group at Centre for Biodiversity Dynamics, NTNU, Norway. FGRM has been estimated using individual genotype data on 118 810 SNPs for 3116 individuals. The estimation has been done

against the allele frequencies of all genotyped individuals in the study system, using genome-wide complex trait analysis (GCTA) software (Yang *et al.* 2011) for the whole population simultaneously (described in the supplementary information of Niskanen *et al.* 2020). This has enabled a more representative set of individuals when allele frequencies have been estimated. The inbreeding coefficient used in this thesis, FGRM, is a genome-wide estimate that takes account of SNP homozygosity at every locus (Yang *et al.* 2011).

Individual lifetime reproductive success (LRS) was estimated only from offspring produced that survived to adulthood, i.e., lived at least to one year old and hence recruited to the breeding population. The number of recruited offspring is based on the pedigree, which enabled connecting offspring to their genetic parents. (Niskanen *et al.* 2020).

In my thesis I included 501 adult house sparrows (Table 1). These house sparrows hatched during years 1998-2011 and were followed during their whole life.

Table 1. Annual mean adult population size (*N*) and standard deviation (*sd*) of each sub population during years 1998-2013 and the number of house sparrow individuals from each population in this study before and after data cleaning.

	mean population size	*Number of individuals*	
Research island	*N (sd)*	*(before data cleaning)*	*(after data cleaning)*
Træna	57.3 (21.1)	39	36
Selvær	60.6 (18.1)	45	42
Gjerøy	83.8 (36.7)	111	111
Hestmannøy	135.8 (45.9)	162	161
Indre Kvarøy	44.3 (11.2)	62	62
Aldra	28.6 (15.1)	82	80
Total		**501**	**492**

2.3 Bioinformatic methods

2.3.1 Laboratory methods

DNA was extracted by the laborants of NTNU (Norway). Later steps were performed in Lund University (Sweden), mostly by Alina Niskanen. Samples were pooled by Kelly Hultman Skogsmyr. Sequencing was performed in spring 2019 in Lund university, hosted by the research group of associate professor Helena Westerdahl.

DNA was extracted from blood samples with ReliaPrep Blood gDNA Miniprep System (Promega Corporation) according to manufacturer's protocol by using a Beckman Coulter NXp pipetting robot. Exon 3 of MHC class I loci were amplified by using forward and reverse primers HNalla: `5′TCCCCACAGGTCTCCACAC` and spspr2: `5′TTGCGCTCCAGCTCCYTCT`, respectively. PCR was performed according to the protocol described in Attachments, Table A1. Agencourt AMPure XP-PCR Purification kit (Beckman Coulter) was used to purify the PCR products according to manufacturer's protocol with the modifications: A ratio used between PCR product and beads was 1:0.8, beads were cleaned by using 80% EtOH, 43 ul of double distilled water was used for elution and incubation was done in room temperature (RT) for two minutes.

Index PCR was performed to attach individual index sequences into each individual's MHC PCR products to separate the individuals from the pooled sequence data. PCR was performed according to the protocol described in Attachments, Table A2. Agencourt AMPure XP-PCR Purification kit (Beckman Coulter) was used to clean the index PCR products. Manufacturer's protocol was followed with the next modifications: A ratio used between PCR product and beads was 1:1.12, beads were cleaned by using 80% EtOH, 43 ul of double distilled water was used for elution and incubation was done in RT for two minutes. DNA concentration measurements were performed with PicoGreen, samples were pooled for sequencing and sequencing was performed with NGS Illumina MiSeq. Replicates were run for 42 samples.

2.3.2 Quality check

At first the primer sequences were removed by using program Cutadapt 2.7 (Martin 2011) with Python 3.7.3. (Van Rossum & Drake 2009). Quality-cutoff (-q) was set to 15 to remove all sequences with low quality. Because both primers were 19 bases long and reverse primer spspr2

included a degenerated base, three errors were allowed in the primer sequences. Therefore, the allowed error rate (-e) was set to 0.16 (3/19 = 0.16). The remaining sequence length (-l) was set to 245, because that is the maximum expected length for the primer combination HNalla and spspr2.

After primer removal, sequence quality was checked by examining trimmed data with FastQC 0.11.8 (Andrews, 2010) and then combining the FastQC results into one report with MultiQC 1.8 (Ewels *et al.* 2016). Steps from primer removal to quality check were performed in computing environment provided by CSC – IT Center for Science, Finland.

2.3.3 Construction of amplicon sequence variant table

After quality check, the data was further filtered and analysed with Dada2 (Callahan *et al.* 2016) in R 3.6.2 (R Core Team 2020) to create the amplicon sequence variant (AVI) table.

With Dada2, it is possible to correct amplicon errors, compare sequences and recognize differences at the exactitude of one nucleotide (Callahan *et al.* 2016). Amplicon means a piece of nucleic acid that is multiplied in replication event. Because with next-generation sequencing (NGS) it is possible to produce high coverage of amplicons for the region of interest, amplicon sequencing allows to detect sequence variants at low frequencies (Eurofins Genomics, 2021).

The Dada2 pipeline was performed with 42 replicate pairs of samples to find out the best filtering level for the whole dataset and calculate the minimum per amplicon frequency, which defines a good cutting point to include only reliable sequences (Callahan *et al.* 2016). At the beginning of the Dada2 pipeline the quality profiles of the first four samples (ring numbers: 8118430, 8309128, 8309135 and 8309141) were examined to estimate the point where to cut the forward and reverse reads of the data set. Reads were truncated with command truncLen at the point where their quality scores dropped below 30 with short extra of 10 bp for forward and 30 pb for reverse reads to avoid possible errors that can arise among the last few nucleotides (Callahan *et al.* 2016). The reads were filtered with maximum expected errors allowed in the read (maxEE) of two for forward reads and three for reverse reads. The criteria for the filtering settings allowed minimized sequence loss with good sequence quality. Error rates of every possible base transition were checked to ensure the quality trend of the sequences. In Dada2, forward and reverse sequences were merged with the mergePairs command and reads that did

not overlap exactly were removed. Chimeric sequences (sequences that are mix of two or more biological sequences) were removed with removeBimeraDenovo command. (Callahan *et al.* 2016).

Minimum per amplicon frequency was defined from the AVI table of 42 duplicates. Read number 1000 was required from all sequences that were preserved in the AVI table to ensure that they were true alleles. At first, one sample (ring number: 8L19656) was removed from the table since one of the replicates had very low sequence content and was not reliable. After that, the number of sequences in each sample was turned into percentages. Sequences that occurred only in the other replicate of the duplicate pairs were removed. In this step also another sample (ring number: 8L57916) was removed since its assumed replicates were not similar, which was probably due to a sample mix-up error in the laboratory. After all the pairless sequences were removed, the lowest percentage among the remaining ones was used as a starting point for determination of minimum per amplicon frequency. The percentage was increased step by step, always beginning from the lowest frequency among the remaining sequences and moving on to the second lowest, to find a frequency where all the sequences removed below the chosen percentage lead to as many as possible perfectly matching replicate pairs. The minimum per amplicon frequency among the sequences was set to 1.06% and all the sequences with a lower frequency were removed from the duplicate data.

After defining the minimum per amplicon frequency, the data set without replicates was analyzed with Dada2 using the same settings as for the duplicate data. Sequences were saved into FASTA format. One of the common sequences was searched using Basic Local Alignment Search Tool (BLAST) provided by NCBI (Johnson *et al.* 2008) to find a reference sequence and the right reading frame. The reference sample was aligned with all the allele sequences in MegaX (Kumar *et al.* 2018) to detect the non-functional sequences that included stop codon(s). Sequences that were not 245, 242 or 239 bp long were considered as wrong length, since the expected length for the primer pair used is 245 bp, but sequences might also be shorter depending on the possible deletions of one or two amino acids (3 or 6 base pairs). In house sparrow, the total length of exon 3 of MHC class I is 268, 271 or 274 bp, depending on the possible 3 or 6 nt deletions (Bonneaud *et al.* 2004a), so with the primer pair used approximately 30 bp of the total length of exon 3 was missed.

After the non-functional sequences were identified, the original output of Dada2 was run again in R according to the pipeline described by Roved *et al.* (2020) and filtered so that all the

sequences that were either non-functional, wrong length or existed at a frequency less than the minimum per amplicon frequency (1.06%) were removed.

2.3.4 MHC allele data examination

All the nucleotide sequences (alleles) were searched against the NCBI database to find the alleles that were already named. To ensure that sequences were the same alleles as the ones in the NCBI database, identity and query cover were both required to be 100%.

The nucleotide sequences were translated to amino acid sequences in MegaX (Kumar *et al.* 2018), to determine the unique protein sequences. Numbers of unique nucleotide (nt) and amino acid (aa) sequences were identified for each individual and island population. The mean number of alleles and standard deviation of the mean were calculated for both sequence types and for each island.

2.4 Statistical analyses

Before beginning the statistical analyses, histograms of fitness components (lifespan and lifetime reproductive success, LRS) and the inbreeding coefficient (FGRM) were drawn to examine their distribution in the metapopulation. Boxplots of lifespan, LRS and FGRM were drawn to compare these components among island populations.

Statistical analyses were performed with R 3.6.2 (R Core Team 2020). All the analyses were run separately for both the number of nucleotide alleles (unique nucleotide (nt) sequences) and amino acid alleles (unique amino acid (aa) sequences). FGRM was standardized so that the variance equalled to 1 and was centred to the mean of the whole data set. The mean and variance of FGRM was 0.0313 and 0.0026, respectively before the standardization. Explanatory variables were included in the models to see how much they affect the variation in the response and also enabled more reliable test by accounting for potential confounding effects. Hatch year and island were included as factors in some of the models to explain the variation, because ecological conditions, for example weather conditions, infectious diseases and competition (e.g., because of differences in density) may vary among islands and years (Ringsby *et al.* 2002; Holand *et al.* 2013; Baalsrud *et al.* 2014). Island was also included because the inbreeding level

has been shown to vary among the islands (Niskanen *et al.* 2020). Hatching year and timing of breeding are found to affect the juvenile survival and individual's recruitment to the breeding population (Ringsby *et al.* 2002; Cleasby *et al.* 2010), and different habitat types of the islands may affect survival (Pärn *et al.* 2012). Also, house sparrow is a resident species, so there is supposed to be limited gene flow between the island populations. Individual's sex was chosen to explain variation, because sex is found to affect the survival probability and the probability of being recruited into breeding population the next year (Cleasby *et al.* 2010).

A linear mixed-effects model was fitted to study the associations between number of nucleotide alleles and amino acid alleles and FGRM, using the lmer function from the R package 'lme4' (Bates *et al.* 2015). FGRM was included as the mean of standardized FGRM per island (meanFGRMz) and the deviation of each individual's standardized FGRM from the island mean (devFGRMz). MeanFGRMz was included in the model to prevent the differences in mean inbreeding levels among islands to bias the results. Island was included as a random factor, where Hestmannøy was set as a reference island because it has the largest population size.

A generalized linear mixed-effect model was fitted to study the effect of number of nucleotide alleles and amino acid alleles on LRS using the glmer function from the R package 'lme4' (Bates *et al.* 2015). Standardised FGRM, island and sex were fitted as fixed effects (island and sex as factors) and hatch year as a random factor. Hestmannøy was set as a reference island and year 2008 was set as a reference hatch year since the second highest number of individuals had hatched that year and it was not too close to the end of the study period, so there were also older (the oldest ones from year 2008 had a lifespan of 7 years) individuals in the data. Since FGRM was not found to correlate with number of nucleotide alleles or amino acid alleles, it was possible to include it as an explanatory variable into the model. The family used was Poisson with a logarithmic link function since the data included count variables. Two models were constructed, one with and one without the island factor to find the model that explained the variation best. Models were compared with AIC using the R package AICcmodavg (Mazerolle, 2020). After studying the association between the explanatory variables and reproductive success, plots of effect sizes were made using the package sjPlot in R (Lüdecke, 2021). The R package effects (Fox & Weisberg, 2019) was used to visualize the effect size estimates. The data points and the predicted line from the regression model. 95% confidence interval were plotted using R package ggplot2 (Wickham, 2016).

Cox Proportional-Hazards Model (coxph) (Cox, 1972) from the R package survival (Therneau, 2021) was used to examine how the number of unique nucleotide alleles and amino acid alleles affected lifespan. Covariables included to explain their effect on the variation of lifespan were standardised FGRM, island, sex and hatch year. Hestmannøy was set as a reference island and year 2008 was set as a reference hatch year.

3 Results

3.1 Data processing

3.1.1 MHC sequence quality and error rates

Mean phred scores were relatively good for forward reads (> 30 on average) and 114 reverse reads. All reads with only reasonable (174 reads) or poor (255 reads) quality were reverse sequences and their quality seemed to collapse after around 200 base pairs (Figure 2). No positions had a content of uncalled nucleotides (N) above one percent. (Attachments, Figure A1a). The N nucleotides were introduced in a situation when the sequencer has not been able to call a base with sufficient confidence and substitutes the real base with N (Babraham Institute). When searching for adapter content, there was a low Illumina universal adapter contamination, raising from 0.4 to 1.21 percent in only one reverse read of one sample (ring number: 8943766), beginning from base position 137 (Attachments, Figure A1b).

In the forward reads, the quality scores stayed relatively good (>30) for the whole expected length of the reads (245 bp). In the reverse reads, the quality scores dropped below 30 at approximately 180 base pairs (Figure 3). Based on these results, the cutting point for the forward reads was set to 235 base pairs and for the reverse reads to 140 base pairs.

Error frequencies of both the forward and the reverse reads decreased as their quality scores increased. Error rates for every possible base transition were studied from plot tables after the filtering step (Attachments, Figure A2).

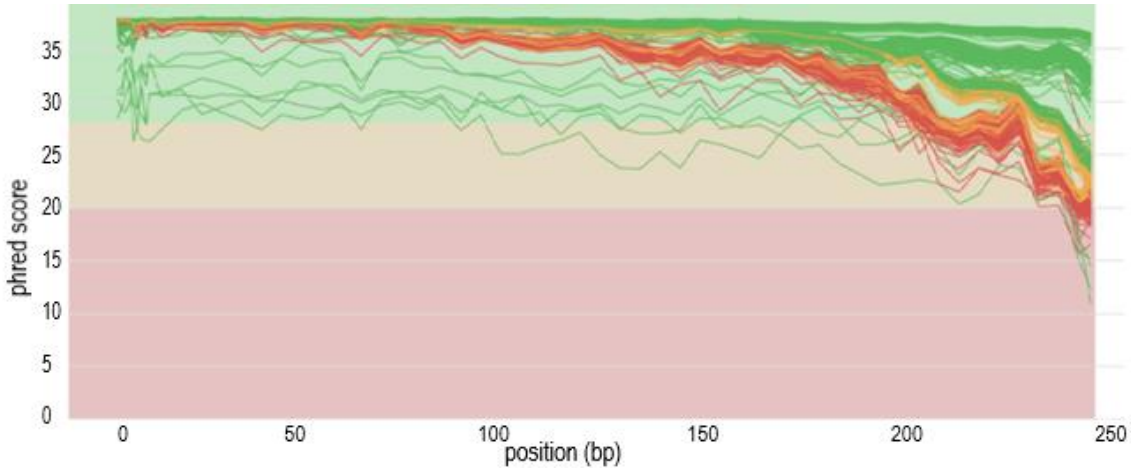

Figure 2. The mean values for phred quality scores across each base position in the reads. Plotted are all the forward and reverse reads of the samples (501 samples) and their replicates (42 samples). Quality scores are shown on the y-axis and base pair positions of the sequences are shown on the x-axis. Reads are coloured with green colour when they have good quality (n = 657), with yellow colour when they have reasonable quality (n=174) and with red colour (n = 255) when they have poor quality.

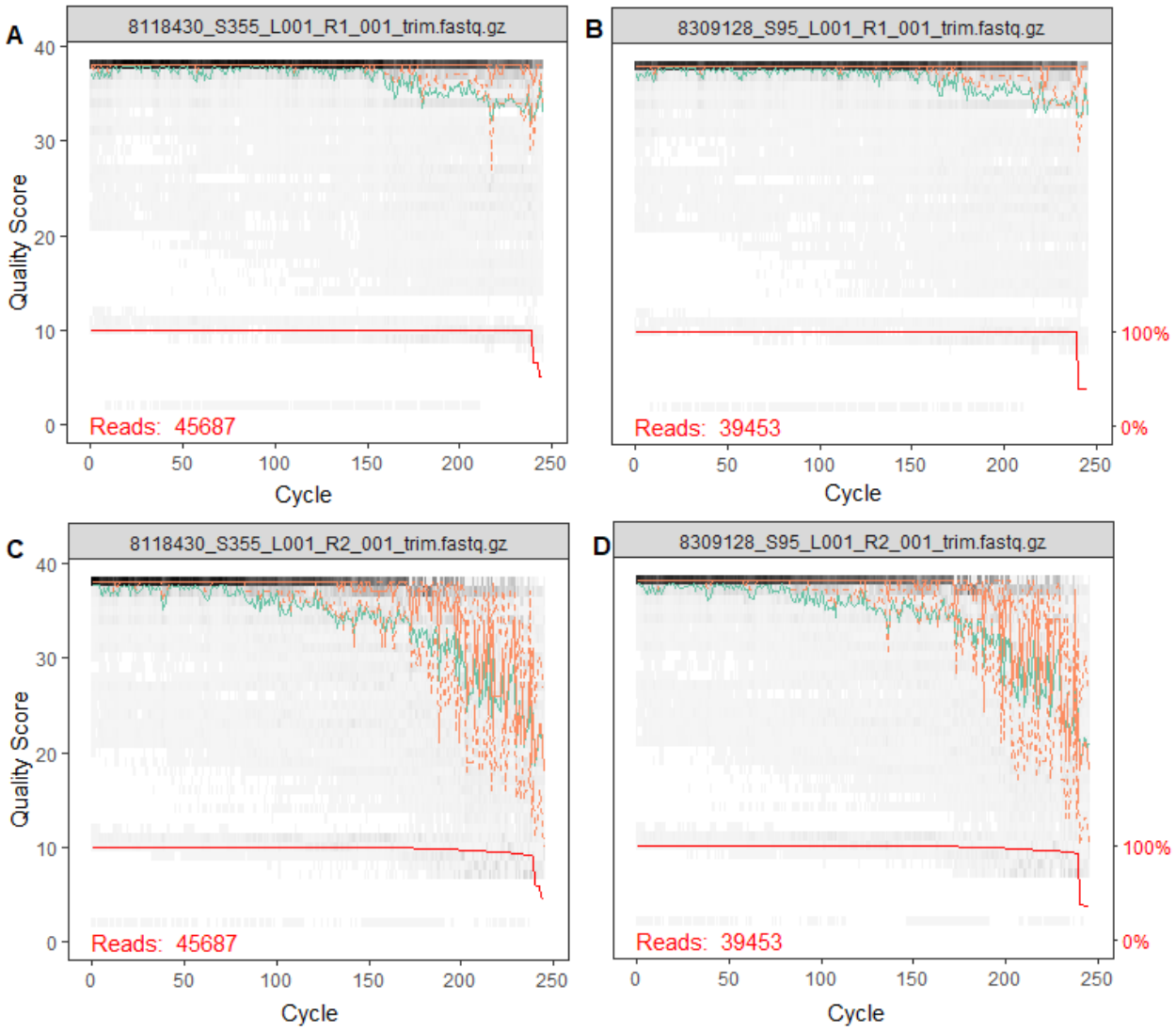

Figure 3. Quality scores of two random samples. Quality scores of forward reads of two individuals are presented on the upper row (A and B) and their reverse reads on the lower row (C and D), respectively. Frequencies of the quality scores (left y-axis) at each base position are shown with gray. The green line shows the mean quality score of each base position and the orange lines shows the quartiles of the quality score distribution. The scaled proportion of the reads (right y-axis) extending to each base position (x-axis) is showed with the red line (Callahan *et al.* 2016).

3.1.2. Defining minimum per amplicon frequency

The minimum per amplicon frequency was defined from the duplicate data and set to 1.06%. This limit was selected since there were only three replicate pairs left that had a mismatch due to one of their sequences occurring at a proportion lower than 1.06%. All sequences that remained in the data set after this point were considered as true alleles.

3.1.3 Sequence alignment and final dataset

The best matching BLAST result to find the reference sequence was KC585633.1 (Alcaide et. al 2013) with E-value 3e-120 and percent identity of 98.74%. The reference sequence was used to define the correct reading frame that was started from the second nucleotide in all of the sequences. This means that the first whole codon that codes an amino acid begins from the second base of the nucleotide sequence. After adjusting the sequences into right reading frame and aligning them, all the sequences that were 239 bp long were found to have a two-codon deletion at the amino acid sites 54 and 55, all the sequences that were 242 bp long had a one-codon deletion at the amino acid site 56 and none of the 245 bp long sequences had any deletions compared to the reference sequence (Attachments, Table A4). Stop codon was found in nine of the aligned sequences and these sequences were removed from the data set.

A total of nine samples were removed from the original data set: 8L19656 and 8L57916, since the expected replicate samples did not match, and 8934215, 8934484, 8934547, 8943766, 8L48457, 8L64169 and 8N05796, since their read content was below 1000. This step resulted a data set of 492 samples within total 127 unique nucleotide alleles.

3.2 Data examination

3.2.1 Number of nucleotide and amino acid sequences

After finding the synonymous mutations in MegaX (Kumar *et al.* 2018), a total of 113 unique MHC amino acid sequences (amino acid alleles) were identified among the 127 unique nucleotide sequences (nucleotide alleles). 80 nucleotide alleles had 100% identity when compared to sequences in the NCBI database, but since none of them had 100% query coverage, all nucleotide alleles were considered as novel alleles.

The number of unique MHC nucleotide alleles within an individual varied from 3 to 29 and the number of unique MHC amino acid (aa) alleles varied from 3 to 22. The mean (± standard deviation, SD) number of unique nucleotide alleles per individual in the whole data set were 12.6 (± 3.5) and mean number of unique aa alleles 11.4 (± 2.9) (Table 2). The lowest number of unique nucleotide and aa alleles were found in the Aldra population: 69 and 50, respectively, and the highest number in Hestmannøy population: 100 and 76, respectively (Table 2). The lowest number of nucleotide alleles and aa alleles that individuals carry on average were in the

Gjerøy population: 11.9 (± 3.2) and 10.9 (± 2.7), respectively, and the highest numbers were in the Indre Kvarøy population: 13.7 (± 3.1) and 12.4 (± 2.6), respectively (Table 2). The distributions of nucleotide and aa MHC alleles were illustrated with histograms. Most individuals had intermediate numbers of alleles, even though histogram of allele numbers was slightly skewed towards the higher allele numbers (Figure 4). When individual unique MHC allele and aa sequence numbers in each island population were compared with boxplots (Figure 5), the number of nucleotide alleles seemed to be more evenly distributed than the number of unique aa alleles, even though the distributions of either type of alleles among the island populations did not vary greatly. The allele number range seemed to be quite similar on all islands. The notably wider distribution of the aa alleles was in the Hestmannøy population, which had the lowest minimum and the highest maximum for the individual number of alleles compared to the other islands. However, the median was at approximately the same level as in the other island populations.

Table 2. Total number of unique nucleotide alleles and amino acid alleles in each island population (n) and the mean number of alleles that individuals carry in each population (mean ± sd).

	Unique alleles		Unique amino acid sequences	
Island	*n*	*mean ± sd*	*n*	*mean ± sd*
Træna	85	13.2 ± 3.5	62	11.7 ± 2.8
Selvær	90	12.5 ± 3.1	67	11.1 ± 2.5
Gjerøy	89	11.9 ± 3.2	67	10.9 ± 2.7
Hestmannøy	100	12.5 ± 3.5	76	11.4 ± 2.9
Indre Kvarøy	88	13.7 ± 3.1	65	12.4 ± 2.6
Aldra	69	12.7 ± 3.9	50	11.6 ± 3.4
All individuals	127	12.6 ± 3.5	113	11.4 ± 2.9

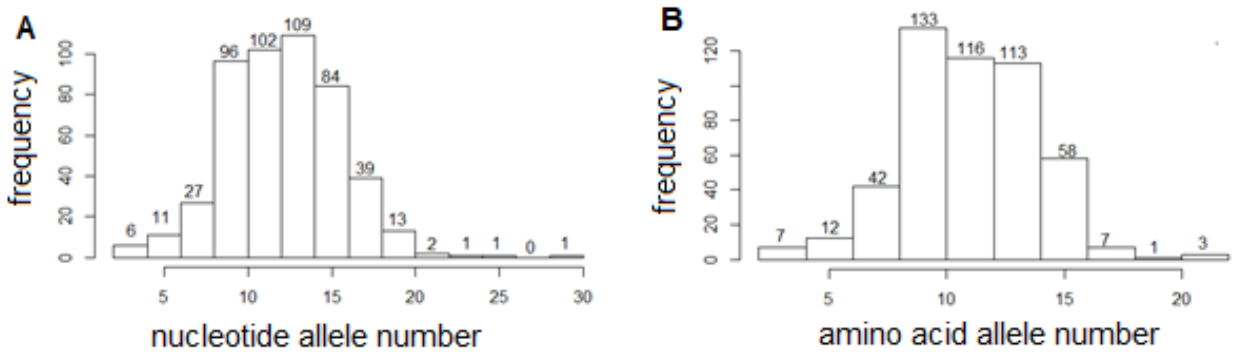

Figure 4. Histograms (A and B) of unique nucleotide alleles and amino acid alleles per individual on all study islands. Above the bars are marked the number of individuals (frequency) carrying the specific number of sequences (sequence number). In both histograms the width of the bars is two and the first interval covers sequence numbers 2 and 3.

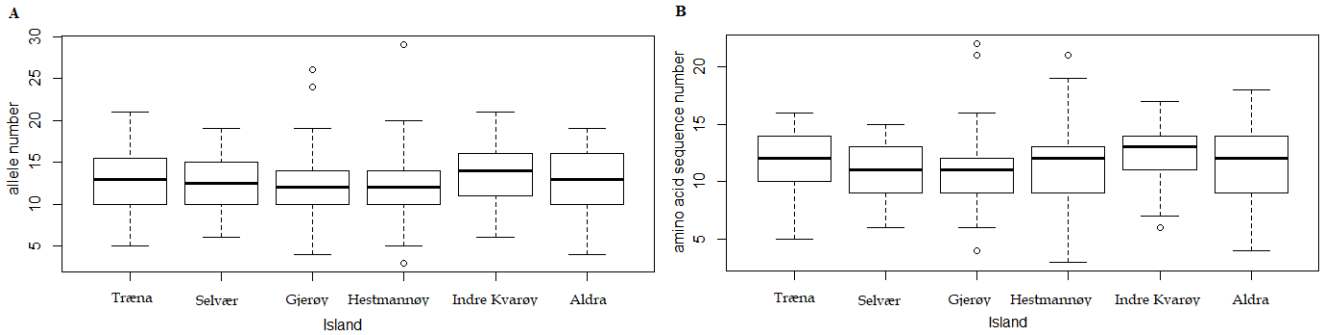

Figure 5. Unique number of nucleotide alleles (A) and amino acid alleles (B) in the island population as boxplots. The horizontal thick line shows the median, lower and upper edges of the boxes are 25% and 75% quartiles respectively (Q1 and Q3) and the lowest and highest lines at the ends of the dashed lines are the minimum (Q1–1,5*IQR) and the maximum (Q3+1,5*IQR), where interquartile range (IQR) is the height of the box. Separate dots are outliers.

3.2.2 Fitness components

Most individuals had a lifespan of 1 year and the distribution of lifespan in the metapopulation was strongly skewed towards shorter values (Figure 6, A1). Among the island populations, birds with longest lifespans were in the Gjerøy and Aldra populations, but the median lifespan was two in every island population (Figure 6, A2). The lifetime reproductive success was strongly skewed to none or only few recruiting offspring produced at the metapopulation level (Figure 6, B1). The lowest reproductive success was in the Aldra population, where the median was zero (Figure 6, B2). At the metapopulation level, most values for the inbreeding coefficient were from -0.08 to 0.06, indicating that the most birds were relatively outbred or inbred at low level (Huisman *et al.* 2016), but there were also more highly inbred than outbred individuals (Figure 6, C1). In island specific comparison, Aldra was the most inbred population (Figure 6,

C2). Hestmannøy and Træna were most outbred populations, with FGRM means of 0.015 and 0.021, respectively. Island specific mean values for FGRM and their standard deviations are presented in Table 3.

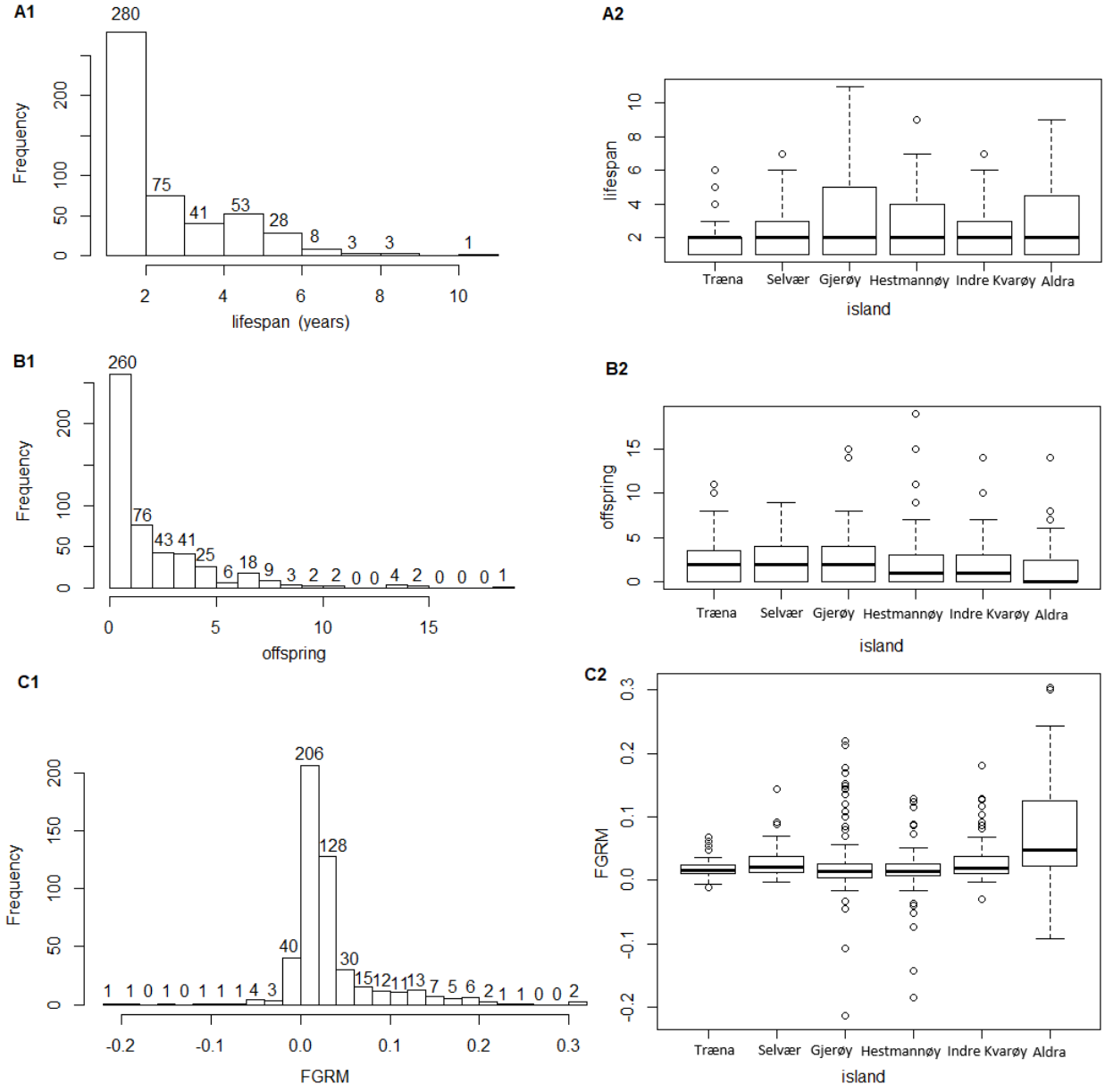

Figure 6. Histograms showing the distributions of lifespan (A1), lifetime reproductive success (LRS) (B1) and FGRM (C1) in the metapopulation, and box plots showing variation in lifespan (A2), lifetime reproductive success (LRS) (B2) and FGRM (C2) among island populations. In the histograms, number of individuals are marked above the bars (frequency). Lifespan is represented in years. For the lifespan histogram, the width of the bars is one and the first interval covers lifespan of one year. LRS is represented as number of recruited offspring an individual has produced during its lifetime. For the histogram of LRS, the width of the bars is one and the first interval covers offspring number 0. In the FGRM histogram the width of the bars is 0.020 and the first interval is from -0.220 to -0.201

Table 3. Island specific mean and standard deviation from the mean (SD) for FGRM.

Island	*mean FGRM*	*SD*
Træna	0.021	0.017
Selvær	0.030	0.029
Gjerøy	0.025	0.054
Hestmannøy	0.015	0.031
Indre Kvarøy	0.032	0.038
Aldra	0.077	0.076

3.2.3 Relationship between MHC allele variation and inbreeding

Inbreeding was not found to affect either the MHC allele numbers or amino acid sequence numbers (Table 4). In the model explaining the individual number of nucleotide alleles, for mean FGRM (meanFGRMz) β = 0.120 with 95% CI from 1.191 to 1.406, and for individual deviation (devFGRMz) β = -0.116 with 95% CI from -0.450 to 0.218. For individual number of amino acid alleles, for meanFGRMz β = 0.251 with 95% CI from -0.834 to 1.326, and for devFGRMz β = -0.124 with 95% CI from -0.401 to 0.152. Because zero was included in the 95 percent confidence intervals of both mean FGRMz and dev FGRMz (Table 4), we can conclude that there was no evidence that they explained a relevant amount of variation in allele or aa sequence numbers.

Table 4. Results of linear mixed-effect models estimating the effect of inbreeding on number of nucleotide and amino acid (AA) alleles. Island was included as a random effect. MeanFGRMz is the standardized island-specific mean of inbreeding coefficient. DevFGRMz is the deviation from the standardized island specific mean FGRM of each individual. SD means standard deviation from the mean, SE means the standard error of the estimate and 95% CI is the 95% confidence interval.

	Allele number			AA sequence number		
Fixed effects:	*Estimate*	*SE*	*95% CI*	*Estimate*	*SE*	*95% CI*
(Intercept)	12.710	0.285	(12.184, 13.255)	11.495	0.237	(11.055, 11.945)
meanFGRMz	0.120	0.689	(-1.191, 1.406)	0.251	0.574	(-0.834, 1.326)
devFGRMz	-0.116	0.170	(-0.450, 0.218)	-0.124	0.141	(-0.401, 0.152)
Random effects:	*Variance*	*SD*		*Variance*	*SD*	
island (intercept)	0.311	0.557		0.217	0.466	
Residual	11.813	3.437		8.097	2.856	

3.2.4 The effect of MHC allele diversity on lifetime reproductive success

AICc values for the models fitted for LRS including MHC allele number as explanatory variable were 2105.9 with island and 2119.3 without island (Table 5). AICc values for the models with MHC amino acid sequence number were 2106.9 with island and 2120.6 without island. Since the AICc differences were considerable (ΔAICc > 2), both analyses were performed with the model that has island as an explanatory factor variable. There was then evidence for a positive association (effect size $\beta = 0.019$, $P = 0.013$) between the number of MHC nucleotide alleles and lifetime reproductive success. Furthermore, there was weak evidence for a positive

association between unique MHC amino acid alleles and lifetime reproductive success (β = 0.021, P = 0.063; Table 6, Figures 7and 8). Effect sizes of the model variables and their 95% confidence intervals are presented visually in Figure 7. Figure 8 visualizes the positive association found between the number of unique alleles and lifetime reproductive success.

There was strong evidence that inbreeding affected LRS negatively (β = -0.151, P < 0.001; Table 6, Figure 7). Furthermore, islands differed in their LRS levels such that sparrows at Træna and Gjerøy were found to have higher LRS compared to the reference island Hestmannøy (β = 0.364, P = 0.002 for Træna, and β = 0.370, P = 0.002 for Gjerøy; Table 6, Figure 7).

Table 5. AICc values for the generalized linear mixed-effect models with and without island as an explanatory factor. K is the number of independent explanatory parameters in a model that explain variation in the dependent variable. AICc is the Akaike information score adjusted for small sample sizes that is estimated for a model. Delta AICc (ΔAICc) is the difference of the AICc values between the one observed and the best model. AICcWt is the AICc weight of a model and it tells how much predictive power each model has; the likelihood of the model being best compared to other tested models given the data. The AICc table was constructed with R package AICcmodavg (Mazerolle, 2020). Models build for the number of nucleotide alleles (allelenumber) or amino acid alleles (aa_number), included standardized FGRM (FGRMz), island and sex as explanatory variables, and hatch year as random effect (1|hatchyear), and were compared to otherwise similar models but without island as explanatory variable.

Response: offspring	*K*	*AICc*	*ΔAICc*	*AICcWt*
Variables:				
allelenumber+FGRMz+island+sex+(1|hatchyear)	10	2105.93	0	1
allelenumber+FGRMz+sex+(1|hatchyear)	5	2119.25	12.33	0
aa_number+FGRMz+island+sex+(1|hatchyear)	10	2106.93	0	1
aa_number+FGRMz+sex+(1|hatchyear)	5	2120.61	13.68	0

Table 6. Results of generalized linear mixed-effect models estimating the association between explanatory variables (predictors) and lifetime reproductive success. FGRMz is the standardized inbreeding coefficient. SE means the standard error of the estimate and *P*-value (*P*) indicates the significance level of an association between the covariate and LRS.

	Number of nucleotide alleles			Number of amino acid alleles		
Predictors	*Estimate*	*SE*	*P*	*Estimate*	*SE*	*P*
(Intercept)	0.652	0.262	0.013	0.656	0.268	0.014
Number of alleles	0.019	0.009	0.034	0.021	0.011	0.063
FGRMz	-0.151	0.036	<0.001	-0.151	0.035	<0.001
Træna	0.364	0.119	0.002	0.370	0.119	0.002
Selvær	0.137	0.120	0.253	0.141	0.120	0.239
Gjerøy	0.283	0.085	0.001	0.283	0.085	0.001
Indre Kvarøy	0.186	0.0111	0.093	0.186	0.111	0.094
Aldra	-0.117	0.107	0.276	-0.119	0.107	0.266
Sex (male)	0.091	0.063	0.150	0.090	0.063	0.152

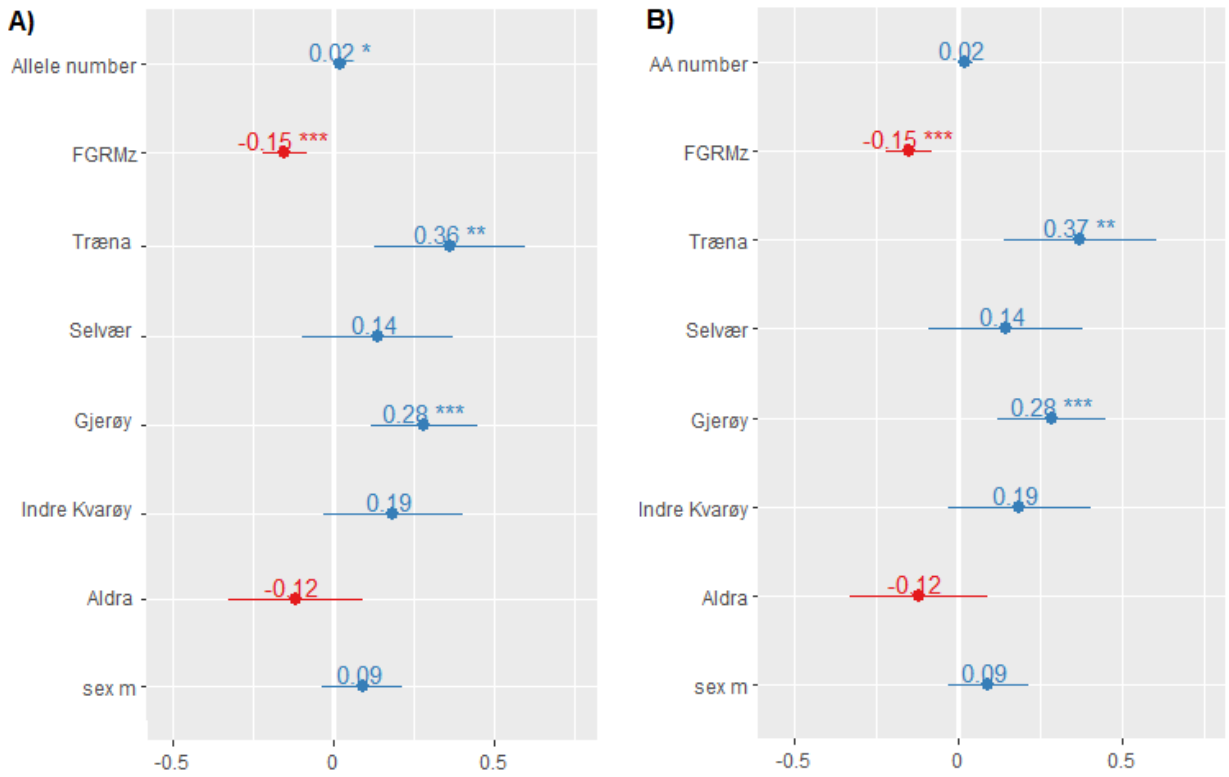

Figure 7. Effect sizes of explanatory variables on reproductive success. In figure A) the first predictor is number of nuclaotide alleles and in figure B) it is the number of amino acid alleles (AA number). Stars indicate the strength of the evidence for the association between predictor and response variable: * = P<0.01, ** = P<0.005, *** = P<0.001. The response variable is lifetime reproductive success; the number of recruited offspring one individual has produced during its whole lifetime. Plots were drawn with R package sjPlot (Lüdecke, 2021).

Predicted lifetime reproductive success for individuals carrying the lowest observed number of nucleotide MHC alleles (three alleles) was 2.36 with 95% CI from 1.46 to 3.81 (Table 7). For individuals carrying the mean number of nucleotide alleles (16 alleles), the predicted LRS was 3.0 with 95% CI from 1.93 to 4.75 and for the individuals carrying the highest nucleotide allele number (29 alleles) the predicted LRS was 3.89 with 95% CI from 2.28 to 6.62 (Table 7).

Table 7. Effects of number of unique nucleotide MHC alleles and amino acid alleles on lifetime reproductive success (LRS). The table shows predicted LRS that correspond to the number of MHC alleles (na) or amino acid alleles (naa), and the estimates' 95 percent confidence limits. LRS is measured in numbers of recruited offspring produced during an individual's lifetime. Predicted allele numbers were estimated with R package effects (Fox & Weisberg, 2019).

Effect of number of unique nucleotide alleles on LRS

n_a	*LRS*	*95% CI*
3	2.36	(1.46, 3.81)
9.5	2.67	(1.71, 4.19)
16	3.03	(1.93, 4.75)
22	3.40	(2.11, 5.47)
29	3.89	(2.28, 6.62)

Effect of number of unique amino acid alleles on LRS

n_{aa}	*LRS*	*95% CI*
3	2.38	(1.47, 3.86)
7.8	2.63	(1.67, 4.13)
12	2.87	(1.84, 4.48)
17	3.18	(2.01, 5.15)
22	3.53	(2.14, 5.83)

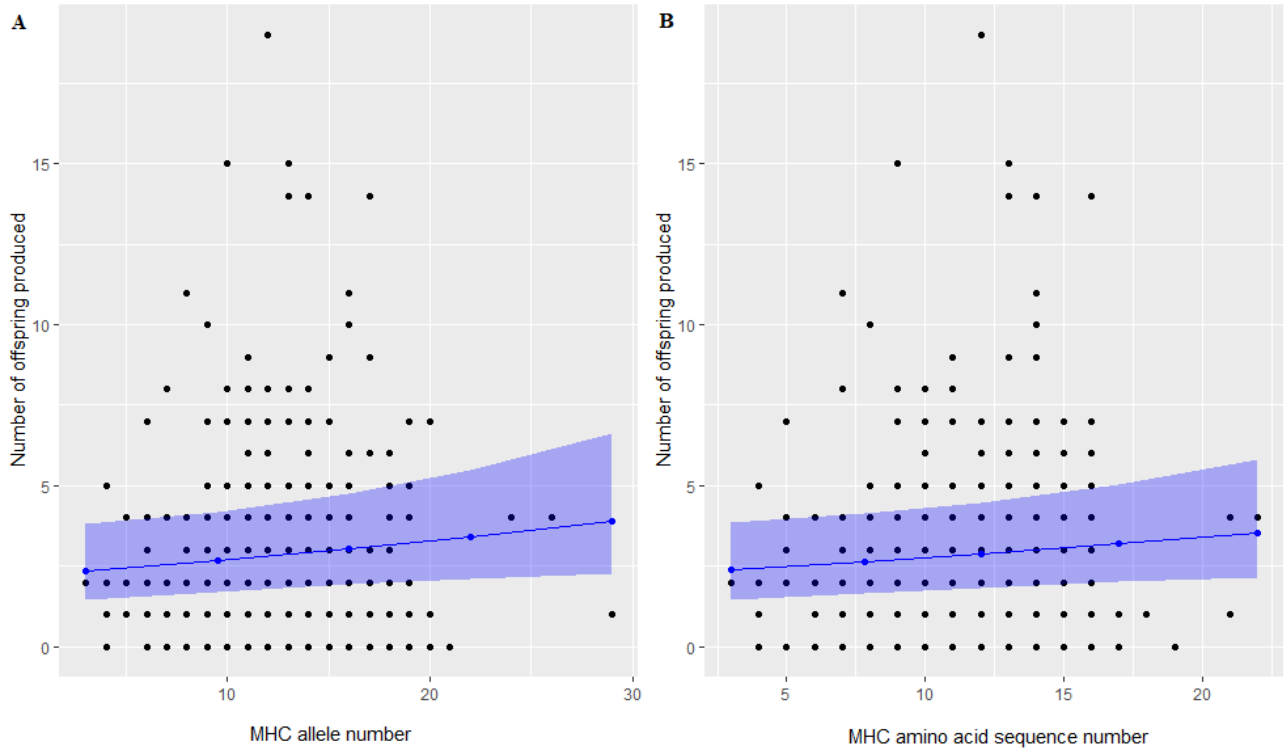

Figure 8. The effect of number of nucleotide MHC alleles (A) and number of unique amino acid alleles (B) on lifetime reproductive success. Each black dot represents one individual, and they are placed based on how many unique alleles they carry and how many recruited offspring they produced during their lifetime. The blue line indicates the model prediction and the coloured area indicates the 95% confidence interval. Blue dots are the predictions from Table 7. This figure was drawn with the R package ggplot2 (Wickham, 2016).

3.2.5 The effect of MHC allele diversity on lifespan

There was no evidence for an association between the number of nucleotide (β = 0.020, *P*= 0.143) or amino acid alleles (β = 0.026, *P* = 0.108; Table 8, Figure 9) and lifespan. There was no strong evidence found that FGRMz would increase the risk of dying. Risk for dying seemed to be highest on Træna and lowest in Selvær compared to Hestmannøy, but there was no evidence for strong association between the islands and lifespan. Compared to hatch year 2008, the risk for dying seemed to be lowest for those birds who had hatched during the earlier years and highest for those who had hatched during the most recent years. Results are shown in Table 8. Effect sizes of estimates with their 95% CI on lifespan are presented in Figure 9.

Table 8. Effect of predictors on lifespan (number of years the individual was alive). The table shows test results for the analyses that examined the effect of number of unique nucleotide and amino acid alleles on lifespan. Regression coefficients (coef) indicate the risk of death so that the higher the value is, the higher (positive value) or lower (negative value) than random the risk is. The hazard ratio (exp(coef)) explains the effect of the covariate on the risk; a value above 1 indicates a higher risk and a value below 1 indicates a lower risk for increasing values of the covariate, or for the fixed factor level compared to the intercept levels (i.e. the island Hestmannø and hatch year 2008). P-value (P) indicates the strength of the evidence for an association between the covariate and the risk of death, for example being born on year 2002 seems to reduce the risk by a factor of 0.52 = 48% compared to 2008. If the 95% confidence interval (L95%; U95%) for the hazard ratio includes 1, there is little evidence that covariate has an impact on hazard ratio. The concordance index is used to estimate how well the models are fit to the predictions. If overall tests have *P*-value lower than 0.05, they all indicate that the model is reliable.

	Effect of number of nucleotide alleles on lifespan						**Effect of amino acid alleles on lifespan**					
covariate	*coef*	*exp (coef)*	*se (coef)*	*P*	*L95%*	*U95%*	*coef*	*exp (coef)*	*se (coef)*	*P*	*L95%*	*U95%*
sequence number	0.020	1.020	0.013	0.143	0.993	1.047	0.026	1.026	0.016	0.108	0.994	1.060
FGRMz	0.064	1.066	0.053	0.225	0.961	1.183	0.067	1.069	0.053	0.209	0.963	1.186
Træna	0.347	1.416	0.188	0.065	0.979	2.048	0.353	1.424	0.188	0.060	0.985	2.059
Selvær	0.107	1.113	0.176	0.542	0.789	1.572	0.111	1.118	0.176	0.527	0.792	1.578
Gjerøy	-0.113	0.893	0.130	0.383	0.693	1.151	-0.109	0.896	0.130	0.400	0.695	1.156
Indre Kvarøy	-0.015	0.985	0.154	0.924	0.729	1.333	-0.014	0.986	0.154	0.926	0.729	1.333
Aldra	-0.015	0.985	0.153	0.920	0.729	1.330	-0.017	0.983	0.153	0.911	0.728	1.328
Male	-0.117	0.889	0.092	0.202	0.743	1.065	-0.115	0.891	0.092	0.210	0.744	1.067
1998	-1.893	0.151	0.731	0.010	0.036	0.631	-1.908	0.148	0.731	0.009	0.035	0.621
1999	-0.976	0.377	0.726	0.179	0.091	1.562	-1.000	0.368	0.726	0.168	0.089	1.526
2000	-1.257	0.284	0.477	0.008	0.112	0.725	-1.277	0.279	0.478	0.008	0.109	0.711
2001	-1.055	0.348	0.384	0.006	0.164	0.739	-1.072	0.342	0.385	0.005	0.161	0.728
2002	-0.652	0.521	0.239	0.006	0.326	0.833	-0.658	0.518	0.240	0.006	0.324	0.828
2003	-0.165	0.848	0.186	0.373	0.589	1.220	-0.158	0.854	0.186	0.394	0.593	1.229
2004	-0.715	0.489	0.380	0.060	0.232	1.030	-0.719	0.487	0.380	0.058	0.231	1.026
2005	-0.552	0.576	0.248	0.026	0.354	0.937	-0.546	0.579	0.248	0.028	0.356	0.942
2006	-0.089	0.915	0.186	0.634	0.635	1.319	-0.093	0.911	0.187	0.617	0.632	1.313
2007	-0.221	0.801	0.183	0.226	0.561	1.147	-0.232	0.793	0.183	0.205	0.554	1.135
2009	0.214	1.238	0.176	0.224	0.877	1.748	0.209	1.232	0.176	0.234	0.874	1.739
2010	0.450	1.569	0.165	0.006	1.135	2.169	0.437	1.548	0.165	0.008	1.121	2.138
2011	0.368	1.445	0.196	0.061	0.983	2.122	0.356	1.428	0.197	0.070	0.971	2.099

Concordance of the models

Concordance	*SE*	*Concordance*	*SE*
0.669	0.017	0.671	0.017

Overall tests

Likelihood ratio test	p=1e-10	p=1e-10
Wald test	p=1e-08	p=9e-09
Score (logrank) test	p=9e-10	p=8e-10

Figure 9. Effect sizes of estimates on lifespan. Effect sizes are from Cox proportional hazards model, and they therefore show the risk of dying before a given age in years (X-axis). In figure A) the first predictor is number of nucleotide alleles and in figure B) number of unique amino acid alleles (AA number). For islands, the parameter estimates are in comparison with Hestmannøy, and hatch years are compared to year 2008. Stars indicate the strength of the evidence for the association between predictor and response variable: * = P<0.01, ** = P<0.005, *** = P<0.001. Response variable is lifespan. Plots were drawn with R package sjPlot (Lüdecke, 2021).

4 Discussion

4.1 Sequence data quality check

In this thesis, a considerably large sample of wild house sparrows were analysed for their MHC allele diversity. After quality control, 492 of 501 individuals (98.2%) were retained in the data set, with 127 unique sequences from 188 amplicon sequence variants (67.6%). The sequences were of expected length for the primer pair used: 245, 242 or 239 base pairs. The sequence data had relatively good quality based on the MultiQC report, only the reverse reads had poorer quality, which is expected since the quality of reverse reads is normally lower toward the 3' end (Bolger *et al.* 2014). Poor-quality calls may have been caused by their long waiting time in the sequencing machine since they are sequenced after forward samples are ready at which point the reagents have waited for much longer in the machine.

4.2 Data analyses

4.2.1 MHC allele diversity

Even though all nucleotide alleles were considered as novel alleles, since their query coverage did not reach 100% compared to corresponding alleles in the NCBI database, most of these are probably found already in earlier studies: 80 of sequences were found to have 100% identity compared to sequences in the NCBI database. The large number of undescribed alleles is likely caused by the primer pair used in this study, which has produced sequences that begin from a different position than alleles in the NCBI database.

In a recent study, the genomic number of MHC class I alleles in the house sparrow was found to be 12.4 on average per individual, of which an average of 6.6 (58 %) were expressed (O'Connor & Westerdahl 2021). In my thesis, the difference between the individual mean numbers of both allele types did not differ much from the number of genomic alleles found by O'Connor & Westerdahl (2021): 12.6 for nucleotide alleles and 11.4 for amino acid alleles. On the contrary, in a previous study, house sparrow individuals were found to carry 4.2 functional MHC class I alleles on average, with the number of alleles ranging between one and eight (Lukasch *et al.* 2017). The difference between my thesis and the previous study by Lukasch *et al.* (2017) may result from the different primer pair used, if for example all alleles were not

recognized (e.g., shorter sequences), and also different methods used, and criteria to consider sequences as true alleles.

4.2.2 Inbreeding and MHC allele diversity

An interesting result was that inbreeding does not seem to explain the number of unique MHC nucleotide or amino acid alleles even though differences between the individual numbers were large: from 3 to 29 nucleotide alleles and from 3 to 22 amino acid alleles. Similar results have been found in an earlier study from the same house sparrow metapopulation, as Borg *et al.* (2011) found that allelic variation at MHC class I and IIB loci was similar in two populations, even though one population was inbred (Aldra) and the other outbred (Hestmannøy). Although their sample sizes were extremely small (N = 5 in each population), they found that the allelic variation at MHC was similar despite significantly higher allelic richness at microsatellite loci in the outbred population. Borg *et al.* (2011) also found that the individual mean number of MHC alleles was similar between the inbred and outbred populations. This was in concordance with the results of this thesis, where I had much larger samples sizes, since the mean sequence numbers between the populations were found to be close to each other (Table 2).

One reason for the absence of evidence for association between MHC allele numbers and inbreeding might be balancing selection that keeps the number of unique MHC sequences relatively high in inbred populations (Loiseau *et al.* 2009). If numbers of nucleotide and amino acid alleles are maintained at high levels, that might also explain why no association between the allele numbers and lifespan could be found. Numerous previous studies have suggested that either pathogen interactions or mechanisms that affect inbreeding avoidance, for example mate choice, might have caused balancing selection that helps to maintain MHC allele diversity even in inbred population (Ekblom *et al.* 2004; Loiseau *et al.* 2011; Borg *et al.* 2011; Rymešová *et al.* 2017).

One reason for not finding an effect of inbreeding on MHC allele numbers may be that the results do not show the true association between inbreeding and MHC alleles. We know the number of unique MHC alleles, but we do not know the exact number of MHC loci each individual carries. Because the exact locus number is unknown and we do not know the order of the loci, it is not possible to say if individuals have two copies of the same allele in one locus, and how many homozygous loci they have. This means that there could be an association

between the level of FGRM and number of homozygous or heterozygous loci, since the locus number alters among the individuals of the same species. On island population level, we do not know if individuals from the most inbred populations have more homozygous loci compared to individuals from less inbred populations. However, if locus number is not altering very widely within the species, results suggest that there are individuals who possess more homozygous than heterozygous loci of MHC I gene, because the lowest number of alleles found per individual was three and the highest number found was 29. This suggests that house sparrows have at least 15 MHC class I loci, which would mean that individuals with only a few different alleles have multiple homozygous loci. This is supported by Karlsson and Westerdahl (2013), who suggested that house sparrows carry at least 12 MHC class I loci. Thus, the estimate of MHC allele numbers would also reflect the level of homozygosity to some extent.

4.2.3 Lifespan and MHC allele diversity

I found no evidence for any associations between lifespan and number of MHC alleles. Similar results are found with extremely endangered Raso lark (*Alauda razae*) that has population size of 20 pairs (Stervander *et al.* 2020). Stervander and others (2020) did not find association between the MHC class I genotype or MHC I diversity and survival in Raso lark, but they found that MHC diversity was maintained in the population despite of severe bottlenecks.

One reason for not finding associations between MHC allele numbers and lifespan in this study could be that only about half of the adult house sparrows and about 20 percent of nestlings survive to the following year (Sæther *et al.* 1999; Ringsby *et al.* 1998). It is possible that the effect of MHC allele diversity on survival could not be seen in the population, if there were other things that affected lifespan more than number of unique MHC sequences. One possibility is that ecological reasons like climatic conditions, predation and competition for food resources affect lifespan (Sæther *et al.* 1999; Barnard, 1980) more than MHC sequence numbers and diversity, even though the lifespan would be affected by the MHC sequences for example by providing wider immunity against the pathogens. The effect of diverse alleles could then be covered under the other reasons.

Karlsson and others (2015) did not find evidence that number of unique MHC alleles affects survival or recruitment success of offspring, so it is unlikely that MHC diversity affects young birds more than adults and associations were not found because of that. In addition to this, even

though MHC I allele diversity is not found to affect survival after hatching, a possible connection is found between survival of young birds and three different MHC I alleles (Karlsson *et al.* 2015), so there could possibly be certain alleles affecting also adult birds. In the future, it would be interesting to study the association between the fitness-related features and unique nucleotide alleles, amino acid alleles, and nucleotide variants, in addition to the number of these alleles.

Note however, that the samples selected for MHC sequencing were collected in a way that using lifespan as a fitness measure was not optimal. The birds chosen for MHC sequencing mostly based on parasite information for years 2007–2012, but year 2004 was added for temporal aspect. This selection included adult birds selected relatively randomly among the ones present in the given islands and years, but it means that for example the birds born in 1998 needed to have survived already 6 years (e.g., 1998–2004) to be included in the sample. The way samples were selected for MHC sequencing is probably the reason why the risk of dying was lower for the earliest hatch years (i.e., <2004). For hatch years 2006–2011, the effect should be more or less correct, because birds that hatched these years should be more randomly selected for MHC sequencing, and they had the possibility to die before the end of the data collection. For more reliable results, one possibility would be to re-run the cox proportional hazard models for lifespan including only birds with hatch years 2006–2011. Such an analysis would avoid most of the bias, but also reduce sample size.

4.2.4 LRS and MHC allele diversity

I found evidence for an association between the number of unique MHC nucleotide alleles and lifetime reproductive success, and weaker evidence for the association between number of unique MHC amino acid alleles and LRS. Since there was no evidence for an association between the number of either type of alleles and lifespan, it is not likely that the association of MHC allele numbers and LRS was due to effects on LRS through longer lifespan. It seems more probable that individuals with higher numbers of unique alleles have a more diverse set of expressed alleles, which affects the production of more and healthier offspring, for example by increasing the ability to find a partner, or affecting body size or condition, fertility, clutch size, number of clutches, and caring for chicks. For example, the association between the body size and number of offspring is found earlier (Lukasch *et al.* 2017). Lukasch and others (2017) found in their study one MHC allele that was associated to reduced body mass and size and

therefore reduced survival of 6 days old house sparrow nestlings. They also found another MHC allele that was associated to longer tarsus height of 12 days old nestlings and might indicate higher resistance to certain pathogens (Lukasch *et al.* 2017). Also, infection of female house sparrows by gapeworm *Syngamus trachea* is found to affect reproductive success negatively (Holand *et al.* 2015), so if individuals with wider set of MHC alleles have a better defence against the infections, this could be seen as an association between the number of MHC alleles and reproductive success.

The evidence for a positive association between the number of nucleotide MHC alleles and lifetime reproductive success was stronger compared to the relationship between LRS and the number of unique amino acid alleles. This is probably explained by the fact that some nucleotide alleles are synonymous, i.e., they code the same amino acid sequence. Individuals that have a higher number of nucleotide alleles have more likely a higher number of synonymous sequences, compared to individuals that have only few nucleotide alleles, which weakens the association. It would be interesting to examine if the results were different, if effects of individual numbers of expressed amino acid sequences on LRS were studied instead of the total number of alleles, since all MHC nucleotide alleles are not expressed as amino acid sequences (O'Connor & Westerdahl, 2021), and the proportion of expressed amino acid alleles compared to the total number of nucleotide alleles can be relatively higher for individuals with lower nucleotide allele numbers. The results could be quite different, if the number of expressed amino acid alleles alter between individuals, since expressed amino acid sequences can be expected to have a stronger effect on LRS and other fitness components than total number of nucleotide alleles, even though individuals with higher numbers of nucleotide alleles have more probably a higher diversity among the MHC sequences. Furthermore, Bonneaud and others (2004b) found that females with intermediate MHC allele numbers had the largest clutch size, and it is possible that this could also be applied to MHC amino acid sequence numbers, so further analyses are needed to examine if individuals with a lower than maximum number of different MHC alleles have highest fitness also in this study system.

5 Summary

Aim of this thesis was to find out if there are associations between the number of unique MHC class I nucleotide or amino acid alleles and inbreeding or fitness components. Research questions were: 1) does inbreeding decrease MHC allele diversity and 2) is there selective advantage for individuals with a wider range of MHC alleles. The questions were approached statistically by testing for associations between MHC allele numbers and inbreeding and fitness components. In this thesis, I found no evidence for strong associations between inbreeding and number of MHC alleles. However, there was evidence for a positive association between LRS and the number of nucleotide alleles, and weak evidence for an association between LRS and the number of amino acid alleles. Furthermore, there was no evidence for associations between the numbers of nucleotide or amino acid alleles and lifespan.

The results of my analyses answered the research questions set for this thesis, but much still remains unknown. For example, in addition to the total number of nucleotide and amino acid alleles, the associations between unique alleles and different fitness components should be studied. This thesis does not separate the effect of unique alleles on different fitness components, and it is possible that the effect of MHC does not show when using the total number of sequences. It is also possible that an intermediate number of nucleotide and amino acid alleles are the most beneficial for lifetime reproductive success and lifespan. Although the number of individuals included in the current study was relatively large, we should also keep in mind the possibility for the type II error since statistical analyses do not always have the power to show evidence for the biological processes examined. This is especially likely when focal relationships are affected by stochastic processes, which are likely to be especially important in smaller natural populations at higher latitudes. It would therefore be important to repeat the study in other populations, as well as maybe use different model species.

In future studies, the reason for the associations found between the number of alleles and lifetime reproductive success should be studied to identify the mechanisms that MHC allele numbers affect LRS. Also, effects of number of MHC alleles on annual reproductive success and annual survival of both fledglings and adults should be studied, along with timing of breeding, clutch size, number of clutches, nestling survival and fledgling size. In addition, the relative proportions of genomic and expressed MHC I alleles and their association to different fitness components within this species should be studied to better understand how the number of expressed MHC alleles affect, for example, individual reproductive success or survival.

6 Acknowledgments

I would like to thank very much my supervisors Alina Niskanen, Henrik Jensen (NTNU, Norway) and Stefanie Muff (NTNU, Norway). I would like to extend my gratitude to all persons who have worked with the data: Helena Westerdahl (Lund, Sweden), Sarah Lundregan (NTNU, Norway) and Anna Drews (Lund, Sweden). I want to thank Anna very much for all the help and advice she has given with bioinformatics. I wish to thank researchers and fieldworkers in the Centre for Biodiversity Dynamics (CBD) house sparrow group. I wish to thank Oulun Luonnonystäväin yhdistys ry for supporting my thesis project financially. I wish to thank CSC – IT Center for Science, Finland, for computational resources. I would like to thank Kimmo Mattila from CSC research support who was the most patient and helpful with my endless questions about data processing. I would like to thank Lauri Nieminen and Janne Kilponen for the help with Excel. I want to thank Juho Harmoinen and Timo Piepponen for all the days we have spent together writing our theses and helping each other. I want to thank my husband-to-be Mikko for being always supportive and encouraging. For the last but not the least I want to thank my fellow students and our biology guild Syntaksis ry who have provided company, peer support and balancing events among the studies.

7 References

Alcaide M, Liu M, Edwards SV. Major histocompatibility complex class I evolution in songbirds: universal primers, rapid evolution and base compositional shifts in exon 3. PeerJ 1:e86 (2013) Doi:10.7717/peerj.86

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Downloaded from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ on 13/10/2020

Baalsrud, H.T., Sæther, B.-E., Hagen, I. J *et al.* Effects of population characteristics and structure on estimates of effective population size in a house sparrow metapopulation. Molecular Ecology 23(11), 2653-2668 (2014). Doi: 10.1111/mec.12770

Babraham Institute, Bioinformatics group. Downloaded from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/6%20Per%20Base%20N%20Content.html on 28/04/2021

Barnard, C. J. Flock feeding and time budgets in the house sparrow (*Passer domesticus L.*). Animal Behaviour 28(1), 295-309 (1980). Doi: 10.1016/S0003-3472(80)80032-7

Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67(1), 1–48. Doi: 10.18637/jss.v067.i01.

Beckman Coulter (2016) Agencourt AMPure XP PCR Purification. Instructions For Use. B37419AB. USA. www.beckmancoulter.com

Bernatchez, L. and Landry, C. MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology16(3), 363-377 (2003). Doi: 10.1046/j.1420-9101.2003.00531.x

BirdLife International (2021) Species factsheet: Passer domesticus. Downloaded from http://www.birdlife.org on 29/04/2021

Bolger, A. M., Lohse, M. and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15) 2114–2120, 1 August 2014, Doi: 10.1093/bioinformatics/btu170

Bonneaud, C., Sorci G., Morin, V. *et al.* Diversity of Mhc class I and IIB genes in house sparrows (*Passer domesticus*). Immunogenetics 55, 855–865 (2004a). Doi: 10.1007/s00251-004-0648-3

Bonneaud, C., Mazuc, J., Chastel, O., Westerdahl, H. and Sorci, G. Terminal investment induced by immune challenge and fitness traits associated with major histocompatibility complex in the house sparrow. Evolution, 58(12), 2823-2830 (2004b). Doi: 10.1111/j.0014-3820.2004.tb01633.x

Borg, Å., Pedersen, S., Jensen, H. & Westerdahl, H. Variation in MHC genotypes in two populations of house sparrow (*Passer domesticus*) with different population histories. Ecology and evolution 1(2): 145–159 (2011). Doi: 10.1002/ece3.13

Callahan, B., McMurdie, P., Rosen, M. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13, 581-583 (2016). Doi: 10.1038/nmeth.3869

Carrington, M., Nelson, G. W., Martin, M. P. *et al.* HLA and HIV-1: Heterozygote Advantage and B*35-Cw*04 Disadvantage. Science 283(5408), 1748-1752 (1999). Doi: 10.1126/science.283.5408.1748

Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. Nature reviews genetics 10, 783-796 (2009). Doi: 10.1038/nrg2664

Clarke, B. & Kirby, D. R. S. Maintenance of histocompatibility polymorphism. Nature 211, 999-1000 (1966). Doi: 10.1038/211999a0

Cleasby, I. R., Nakagawa, S., Gillespie D. O. S. and Burke, T. The influence of sex and body size on nestling survival and recruitment in the house sparrow. Biological Journal of the Linnean Society 101(3) 680-688 (2010). Doi: 10.1111/j.1095-8312.2010.01515.x

Cox, D. R. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B (Methodological). 34(2), 187-202 (1972). Doi: 10.1111/j.2517-6161.1972.tb00899.x

Ekblom, R., Sæther, S. A., Grahn, M. *et al.* Major histocompatibility complex variation and mate choice in a lekking bird, the great snipe (*Gallinago media*). Molecular Ecology 13(12), 3821-3828 (2004). Doi: 10.1111/j.1365-294X.2004.02361.x

Eurofins Genomics (2021). Amplicon Sequencing. Downloaded from https://eurofinsgenomics.eu/en/eurofins-genomics/material-and-methods/amplicon-sequencing/ on 28/04/2021

Ewels, P., Magnusson, M., Lundin S. and Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics (2016). Doi: 10.1093/bioinformatics/btw354 PMID: 27312411

Fox J, Weisberg S (2019). An R Companion to Applied Regression, 3rd edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/index.html

Frankham, R. Genetics and extinction. Biological Conservation 126, 131-140 (2005). Doi: 10.1016/j.biocon.2005.05.002

Hedrick, P. W. Pathogen resistance and genetic variation at MHC loci. Evolution, 56(10), 1902-1908 (2002). Doi: 10.1111/j.0014-3820.2002.tb00116.x

Hewitt, E. W. The MHC class I antigen presentation pathway: strategies for viral immune evasion. Immunology 110(2): 163–169 (2003). Doi: 10.1046/j.1365-2567.2003.01738.x

Holand, A. M., Steinsland, I, Martino, S. and Jensen, H. Animal Models and Integrated Nested Laplace Approximations. Genetics society of America 3(8), 1241-1251 (2013). Doi: 10.1534/g3.113.006700

Holand, H., Jensen, H., Tufto, J. *et al.* Endoparasite Infection Has Both Short- and Long-Term Negative Effects on Reproductive Success of Female House Sparrows, as Revealed by Faecal Parasitic Egg Counts. Plos One (2015). Doi: 10.1371/journal.pone.0125773

Huisman, J., Kruuk, L. E. B., Ellis, P. A. *et al.* Inbreeding depression across the lifespan in a wild mammal population. PNAS 113(13), 3585-3590 (2016). Doi: 10.1073/pnas.1518046113

Husby, A., Sæther, B. E., Jensen, H. and Ringsby, T. H. Causes and consequences of adaptive seasonal sex ratio variation in house sparrows. Journal of Animal Ecology 75(5), 1128-1139 (2006). Doi: 10.1111/j.1365-2656.2006.01132.x

Jensen, H., Moe, R., Hagen, I. J. *et al.* Genetic variation and structure of house sparrow populations: is there an island effect? Molecular ecology 22(7), 1792-1805 (2013). Doi: 10.1111/mec.12226

Jensen, H., Steinsland, I., Ringsby, T.H. and Sæther, B.-E. Evolutionary dynamics of a sexual ornament in the house sparrow (*Passer domesticus*): The role of indirect selection within and between sexes. Evolution 62, 1275–1293 (2008). Doi: 10.1111/j.1558-5646.2008.00395.x

Jensen, H., Sæther B.-E., Ringsby T. H. *et al.* Sexual variation in heritability and genetic correlations of morphological traits in house sparrow (Passer domesticus). Journal of Evolutionary Biology 16, 1296–1307 (2003). Doi: 10.1046/j.1420-9101.2003.00614.x

Johnson, M., Zaretskaya, I., Raytselis, Y. *et al.* NCBI BLAST: a better web interface. Nucleic Acids Research, 36(2) W5–W9, 1 July 2008, Doi: 10.1093/nar/gkn201

Jordan, W. C. and Brufors, M. W. New perspectives on mate choice and the MHC. Heredity 81, 127–133 (1998). Doi: 10.1046/j.1365-2540.1998.00428.x

Kardos, M., Taylor, H. R., Ellegren, H. *et al.* Genomics advances the study of inbreeding depression in the wild. Evolutionary applications. 9(10), 1205-1218 (2016). Doi: 10.1111/eva.12414

Karlsson, M., Schroeder, J., Nakagawa, S. *et al.* House sparrow Passer domesticus survival is not associated with MHC-I diversity, but possibly with specific MHC-I alleles. Journal of Avian biology 45: 167–174 (2015) Doi: 10.1111/jav.00413

Karlsson, M. & Westerdahl, H. haracteristics 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 (2013). Doi: 10.1007/s00239-013-9575-y

Keller, L. F. and Waller, D. M. Inbreeding effects in wild populations. Trends in Ecology and Evolution 17(5), 230-241 (2002). Doi: 10.1016/S0169-5347(02)02489-8

Koenig, D., Hagmann, J., Li, R *et al.* Long-term balancing selection drives evolution of immunity genes in Capsella. Evolutionary Biology, Genetics and Genomics. (2009). Doi: 10.7554/eLife.43606

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35:1547-1549 (2018). Doi: 10.1093/molbev/msy096

Kvalnes, T., Ringsby, T. H., Jensen, H. *et al.* Reversal of response to artificial selection on body size in a wild passerine. Evolution 71(8), 2062-2079 (2017). Doi: 10.1111/evo.13277

Liu, L., Bosse, M., Megens, H.-J. *et al.* Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. Evolutionary Applications 2021(14), 710-720 (2020). Doi: 10.1111/eva.13150

Lukasch, B., Westerdahl, H., Strandh, M. *et al.* Major histocompatibility complex genes partly explain early survival in house sparrows. Scientific Reports 7, 6571 (2017). Doi: 10.1038/s41598-017-06631-z

Lundregan, S. L., Hage, I. J., Gohli, J. *et al.* Inferences of genetic architecture of bill morphology in house sparrow using a high-density SNP array point to a polygenic basis. Molecular ecology 27(17), 3498-3514 (2018). Doi: 10.1111/mec.14811

Lüdecke D (2021). sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.7, https://CRAN.R-project.org/package=sjPlot

Loiseau, C., Richard, M., Garnier, S. *et al.* Diversifying selection on MHC class I in the house sparrow (Passer domesticus). Molecular ecology 18(7), 1331-1340 (2009). Doi: 10.1111/j.1365-294X.2009.04105.x

Loiseau, C., Zoorob, R., Robert, A *et al.* Plasmodium relictum infection and MHC diversity in the house sparrow (*Passer domesticus*). Proceedings of the Royal Society B: Biological Sciences 278(1709), 1264–1272 (2011). Doi: 10.1098/rspb.2010.1968

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, [S.l.], 17(1): 10-12, ISSN 2226-6089 (2011). Doi: 10.14806/ej.17.1.200

Mazerolle, M. J. (2020). AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 2.3-1, https://cran.r-project.org/package=AICcmodavg

Niskanen, A. K., Billing, A. M., Holand, H. *et al.* Consistent scaling of inbreeding depression in space and time in a house sparrow metapopulation. PNAS 117(25), 14584-14592 (2020). Doi: 10.1073/pnas.1909599117

O'Connor & Westerdahl. Tradeoffs in expressed major histocompatibility complex diversity seen on a macro-evolutionary scale among songbirds. Evolution 05 March 2021. Doi: 10.1111/evo.14207

Purcell, S., Neale, B., Todd-Brown, K. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics 81(3), 559-375 (2007). Doi: 10.1086/519795

Promega Corporation (2012) ReliaPrep™ Blood gDNA Miniprep System. Technical manual. Part# TM330. USA. www.promega.com

Pärn, H., Ringsby, T. H., Jensen, H. and Sæther, B.-E. Spatial heterogeneity in the effects of climate and density-dependence on dispersal in a house sparrow metapopulation. Proc Biol Sci 279(1726), 144-152 (2012). Doi: 10.1098/rspb.2011.0673

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Radwan, Tkacz and Kloch. MHC and Preferences for Male Odour in the Bank Vole. Ethology 114(9), 827-833 (2008). Doi: 10.1111/j.1439-0310.2008.01528.x

Richardson, J. L., Michaelides, S., Combs, M. *et al.* Dispersal ability predicts spatial genetic structure in native mammals persisting across an urbanization gradient. Evolutionary applications 2021(14), 163-177 (2020). Doi: 10.1111/eva.13133

Richardson, D. S. & Westerdahl, H. MHC diversity in two Acrocephalus species: the outbred Great reed warbler and the inbred Seychelles warbler. Molecular ecology 12(12), 3523-3529 (2003). Doi: 10.1046/j.1365-294X.2003.02005.x

Ringsby, T. H., Sæther, B.-E. and Solberg, E. J. Factors affecting juvenile survival in House Sparrow *Passer domesticus*. Journal of Avian Biology 29: 241-247 (1998). Doi: 10.2307/3677106

Ringsby, T. H., Sæther, B.-E., Tufto, J. *et al.* Asynchronous spatiotemporal demography of a house sparrow metapopulation in a correlated environment. Ecology 82(2), 561-569 (2002). Doi: 10.1890/0012-9658(2002)083[0561:ASDOAH]2.0.CO;2

Roved, Hansson, Stervander, Hasselquist and Westerdahl. Non-random association of MHC-I alleles in favor of high diversity haplotypes in wild songbirds revealed by computer-assisted MHC haplotype inference using the R package MHCtools. BioRxiv 25/3/2020. Doi: 10.1101/2020.03.24.005207

Rymešová, D., Králová, T., Promerová, M. *et al.* Mate choice for major histocompatibility complex complementarity in a strictly monogamous bird, the grey partridge (*Perdix perdix*). Frontiers in Zoology 14(9) (2017). Doi: 10.1186/s12983-017-0194-0

Sæther, B.E., Ringsby, T. H., Bakke, Ø. and Solberg, E. J. Spatial and temporal variation in demography of a house sparrow metapopulation. Journal of Animal Ecology, 68(3), 628-637 (1999). 25/12/2001. Doi: 10.1046/j.1365-2656.1999.00314.x

Stear, M. J., Innocent, G. T., Buitkamp, J. The evolution and maintenance of polymorphism in the major histocompatibility complex. Veterinary Immunology and Immunopathology 108(1-2), 53-57 (2005). Doi: 10.1016/j.vetimm.2005.07.005

Stervander, M., Dierickx, E. G., Thorley, J. *et al.* High MHC gene copy number maintains diversity despite homozygosity in a Critically Endangered single-island endemic bird, but no evidence of MHC-based mate choice. Molecular ecology 29(19), 3578-3592 (2020). Doi: 10.1111/mec.15471

Therneau T (2021). A Package for Survival Analysis in R. R package version 3.2-11, https://CRAN.R-project.org/package=survival.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
Westerdahl, H., Wittzell, H. & von Schantz, T. Polymorphism and transcription of Mhc class I genes in a passerine bird, the great reed warbler. Immunogenetics 49, 158-170 (1999). Doi: 10.1007/s002510050477

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Wieczorek, M., Abualrous, E. T., Sticht, J *et al.* Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. Frontiers in Immunology 8:292 (2017). Doi: 10.3389/fimmu.2017.00292

Woelfing, B., Traulsen, A., Milinski, M. and Boehm, T. Does intra-individual major histocompatibility complex diversity keep a golden mean? Philos Trans R Soc Lond B Biol Sci 12; 364(1513), 117-128 (2009). Doi: 10.1098/rstb.2008.0174

Yang, J., Lee, S. H., Goddard, M., E. and Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics 88(1), 76-82 (2011). Doi: 10.1016/j.ajhg.2010.11.011

Zilko, J. P., Harley, D., Hansen, B. *et al.* Accounting for cryptic population substructure enhances detection of inbreeding depression with genomic inbreeding coefficients: an example from a critically endangered marsupial. Molecular Ecology 29(16), 2978-2993 (2020). Doi: 10.1111/mec.15540

8 Attachments

Table A1. First PCR to amplify MHC class I alleles.

PCR mix		
	1x	
Component	*Starting concentration*	*Vol (ul)*
Phusion master mix	2X	12,5
Primer “HNalla”	10 uM	1,25
Primer “spspr2”	10 uM	1,25
ddH2O		9
Sample DNA	25 ng/ul	1
total volume		25
PCR program		
°C	*time*	
98	30 s	
98	10 s	
66	10 s	25 x
72	10 s	
72	10 min	
4	hold	

Table A2. PCR to add individual index combination to the MHC sequences of each sample.

PCR mix

	1x	
Component	*Starting concentration*	*Vol (ul)*
Phusion master mix	2X	25
Primer S-series		5
Primer N-series		5
ddH2O		10
Sample DNA: 1st PCR*		5
total volume		50

*Either 5 or 15 ul, depending on the concentration that was estimated based on the strength of a band on 2% agarose gel

PCR program

°C	time	
98	30 s	
98	10 s	8 x
62	15 s	
72	15 s	
72	10 min	
4	hold	

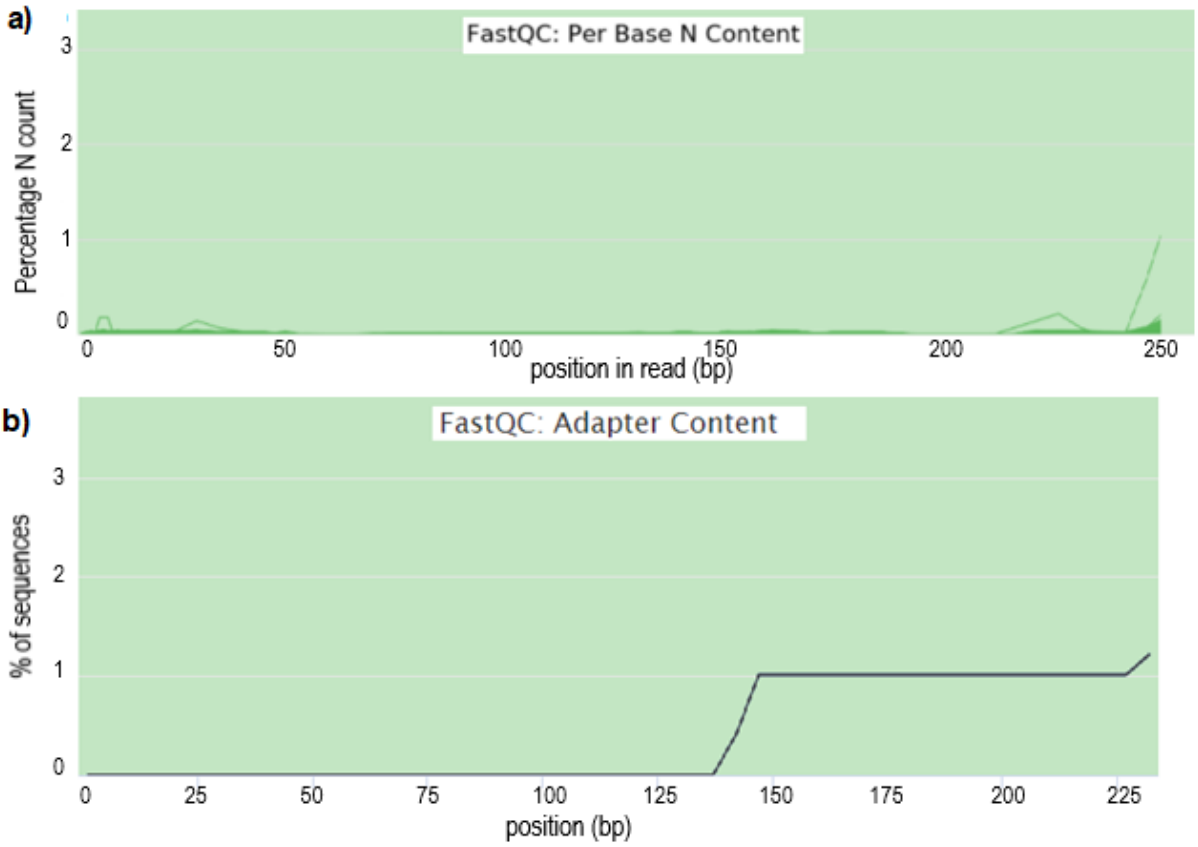

Figure A1. Per base N contents of reads (a) and cumulative proportions of adapter sequences at each position in sequence library (b). In a) the percentage of N-count is shown on the y-axis and the base pair positions of N contents in reads are shown on the x-axis. None of sequences exceeded N content of more than one percent in any base position. In b) samples with adapter contamination of ≥ 0.1% are shown. Adapter contamination was found in one sample.

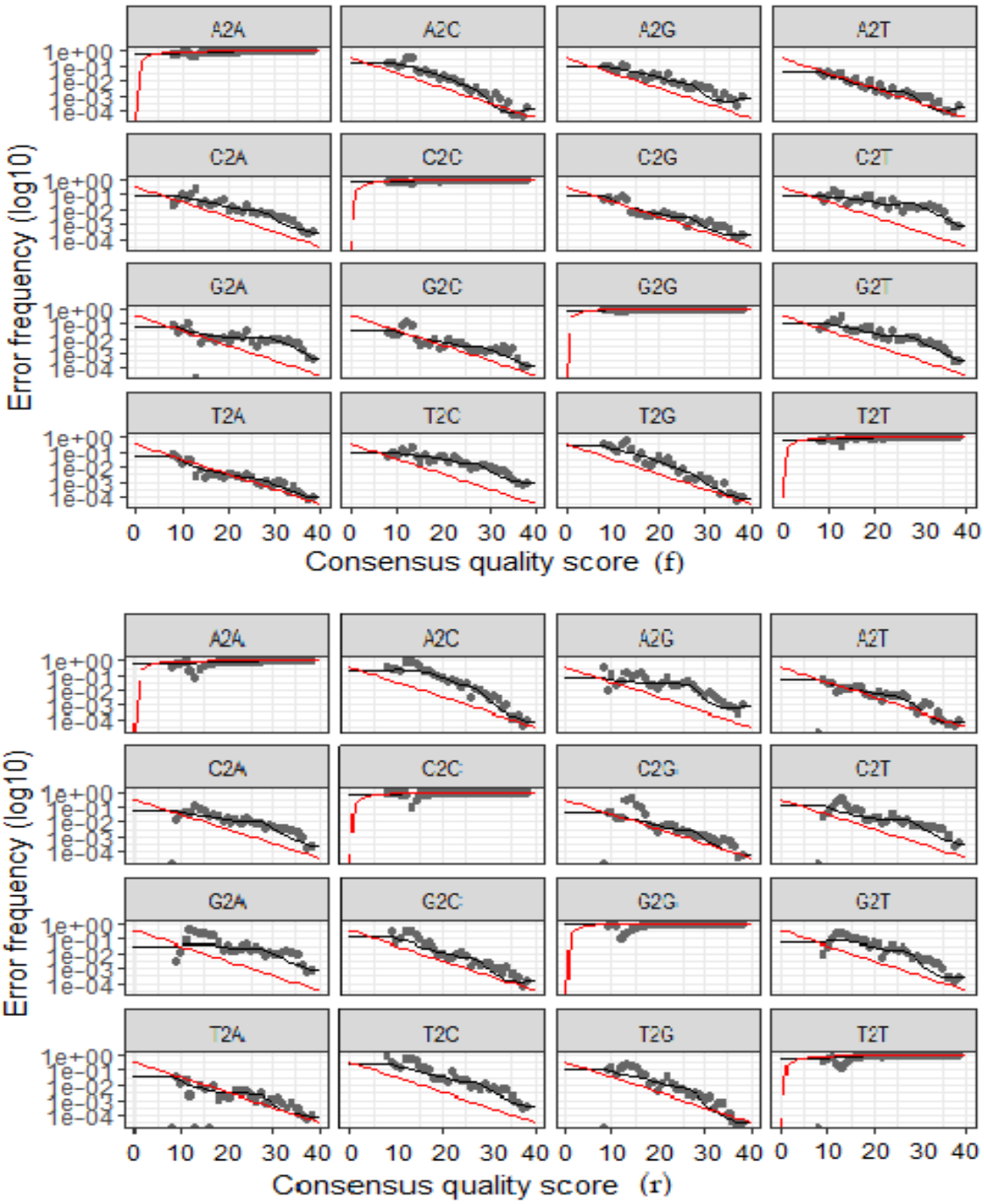

Figure A2. Learned error rates for every possible base transition. Error rates for the forward reads are shown on the left and for the reverse reads on the right. Observed error rates of consensus quality scores are shown as dots. The estimated error rates after algorithm convergence are shown with black lines and the expected error rates of the quality scores are shown with red lines. (Callahan *et al.* 2016).

Table A4. The deletion sites (amino acid site) and the number (n) of each sequence of certain length (in base pairs, bp) after the final trimming drive with Dada2.

Sequence length (pb)	deletion (amino acid site)	n after the final cleaning step
239	54,55	65
242	56	9
245	-	53