Haplotyping MHC class IIa by high throughput screening in an isolated sheep population

Author names and addresses:

Dicks, K.L. [1*†], Pemberton, J.M. [1], Ballingall, K.T. [2], and Johnston, S.E. [1]

[1] Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, EH9 3FL, UK

[2] Moredun Research Institute, Edinburgh, EH26 0PZ, UK

[*] Corresponding author: kara.dicks@gmail.com

[†] Present address: Royal Zoological Society of Scotland, Edinburgh, EH12 6TS, UK

Key words

Major histocompatibility complex (MHC); Soay sheep; Single Nucleotide Polymorphisms; haplotypes; balancing selection

Abstract

Investigating the current evolutionary processes acting on a highly polymorphic gene region, such as the major histocompatibility complex (MHC), requires extensive population data for both genotypes and phenotypes. The MHC consists of several tightly linked loci with both allelic and gene content variation, making it challenging to genotype. Eight class IIa haplotypes have previously been identified in the Soay sheep (*Ovis aries*) of St. Kilda using Sanger sequencing and cloning, but no single locus is representative of all haplotypes. Here, we exploit the closed nature of the island population of Soay sheep and its limited haplotypic variation to identify a panel of SNPs that enable imputation of MHC haplotypes. We compared MHC class IIa haplotypes determined by Sanger sequence-based genotyping of 135 individuals to their SNP profiles generated using the Ovine Infinium HD BeadChip. A panel of 11 SNPs could reliably determine MHC diplotypes, and two additional SNPs within the *DQA1* gene enabled detection of a recombinant SNP haplotype. The panel of 13 SNPs was genotyped in 5951 sheep, of which 5349 passed quality control. Using the Soay sheep pedigree, we were able to trace the origin and inheritance of a recombinant SNP haplotype.

Introduction

The most highly polymorphic region of the vertebrate genome is the Major Histocompatibility Complex, and yet the evolutionary processes driving this diversity remain much debated (Bernatchez and Landry 2003; Spurgin and Richardson 2010). MHC genes encode proteins that present pathogen-derived peptides to T-cells, thus enabling an individual to mount an immune response (Klein 1986). Pathogen-mediated selection (PMS) is strongly implicated in driving diversity at the MHC (Bernatchez and Landry 2003; Piertney and Oliver 2006; Spurgin and Richardson 2010). The key PMS mechanisms involved, which are not necessarily mutually exclusive, are heterozygote advantage (Doherty and Zinkernagel 1975; Slade and McCallum 1992) and variation in selection pressure through either negative frequency-dependent selection (Apanius et al. 1997) or fluctuating selection (Hill 1991; Hedrick 2002). Disentangling these mechanisms is challenging in experimental systems which may be incapable of replicating the wide array of pathogens and parasites that occur within a wild host (although see Kalbe *et al.* 2009 and Eizaguirre *et al.* 2012), and within wild systems which may be swamped with variation in individual genetic backgrounds, pathogen diversity and environmental diversity.

A major impediment to understanding current selection patterns on the MHC in natural populations is our inability to accurately characterise variation at the level of the individual, and to do this at scale across many individuals. Developing genotyping methods which attribute alleles to loci and loci to haplotypes are key technical challenges when studying the MHC in non-model species (Bernatchez and Landry 2003; Spurgin and Richardson 2010). Large sample sizes are necessary to generate sufficient power in statistical analyses of highly polymorphic loci, such as the MHC. Rare alleles, which are of particular interest in studies of the MHC as they may confer increased fitness under negative frequency-dependent selection or in heterozygote form under heterozygote advantage (Apanius et al. 1997; Hill 1998), enhance the need for large sample sizes. Characterisation of each locus individually is required to partition variation amongst loci and determine an individual's true genotype. Using pooled diversity measures generated by co-amplification of multiple indistinguishable loci results prohibits assessment of the heterozygous state of each locus, and therefore the ability to effectively assess heterozygous advantage (Spurgin and Richardson 2010). Locus-based genotyping of the MHC, particularly at the haplotype level, is often expensive, time consuming or not possible due to gene duplication, gene conversion and variation in gene number which makes designing locus-specific primers difficult or impossible (Westerdahl 2007; Babik 2010; Whittaker et al. 2012, but see Worley et al. 2008). Moreover, MHC loci exist in haplotypes and, in the short-term, selection will act on the combination of loci present in a haplotype. Genotyping only a single locus may not capture the full

diversity of the MHC molecules generated by single haplotype (e.g. Dicks et al. 2019). An effective and reliable genotyping method should, therefore, be able to capture variation at individual loci and across multiple expressed loci.

The Soay sheep (*Ovis aries*) population of St. Kilda, UK provides an excellent study system in which to look for signatures of the evolutionary mechanisms operating to maintain diversity at the MHC. Soay sheep are descendants of an early domestication have been unmanaged on the island of Soay for probably thousands of years, and since the translocation of 107 sheep to the island of Hirta in 1932 (Clutton-Brock et al. 2004). The population has been the subject of a long-term study since 1985 (Clutton-Brock et al. 2004). Multiple population censuses per year and the sampling of lambs in spring and of sheep of all age categories in summer has provided a wealth of data on both phenotypic and fitness traits relevant to understanding signatures of evolutionary processes acting on the MHC, as well as a deep understanding of their ecology and demography. Driven primarily by climate and food limitation at high density, the population undergoes large fluctuations in population density (Boyd et al. 1964), varying between approximately 400 and 2000 sheep on the island of Hirta. However, the effective population size was estimated at 194 sheep (Kijas et al. 2012) and the population has generally low genetic diversity compared to other breeds (Lawson Handley et al. 2007; Kijas et al. 2009). This intensively studied island population which is genetically limited and subjected almost exclusively to natural evolutionary processes since their early domestication lends itself to investigating MHC evolution. Indeed, a previous study found that allele frequencies of two MHC-linked microsatellite loci were more even that would be expected under neutral processes, implicating balancing selection (Paterson 1998), and associations have been detected between the MHC-linked microsatellites and parasite resistance (faecal egg counts) and fitness traits (survival in lambs and yearlings) (Paterson et al. 1998).

Sequence-based genotyping methods have been developed for multiple ovine MHC class IIa loci (Ballingall and Tassi 2010; Ballingall et al. 2015, 2018), and variation at class IIa loci was previously characterised in the Soay sheep population using these methods (Dicks et al. 2019). Genotyping of the *DRB1*, *DQA1, 2* and *2-like*, as well as *DQB1, 2* and *2-like* loci revealed eight haplotypes, with some allele sharing amongst haplotypes (but not between loci; see Figure S1). No single locus captures all MHC class IIa variation in the Soay sheep population. In addition, sequence-based genotyping of the *DQ* loci, especially the *DQB* loci, was complicated by cross-amplification amongst loci and variation in locus configuration (Ballingall et al. 2015, 2018; Ali et al. 2016; Dicks et al. 2019). Sanger sequencing of multiple loci in large numbers of individuals would be costly and slow, yet such haplotypic data is

necessary to carry out analyses of the evolutionary processes underlying the high variability in this region.

SNP-based genotyping methods are easily scalable, and human population studies have used selected SNPs which are in linkage disequilibrium with specific MHC haplotypes to impute MHC haplotypes (Dilthey et al. 2011; de Bakker and Raychaudhuri 2012; Zheng et al. 2014). High linkage disequilibrium between loci within subregions is characteristic of the mammalian MHC (Dawkins et al. 1999), including the ovine MHC (Lee et al. 2012). However, high levels of polymorphism and sequence similarity between alleles at different loci within the class IIa genes limits the ability to select SNPs from within coding regions and sequence data for introns and other non-coding regions of the ovine class IIa is generally lacking. As a consequence, the Illumina Ovine 50K SNP beadchip has just one SNP in the putative MHC class IIa region and the Illumina Ovine Infinium HD BeadChip, with an attempted 606,066 SNPs, has just 27 (Figure 1).

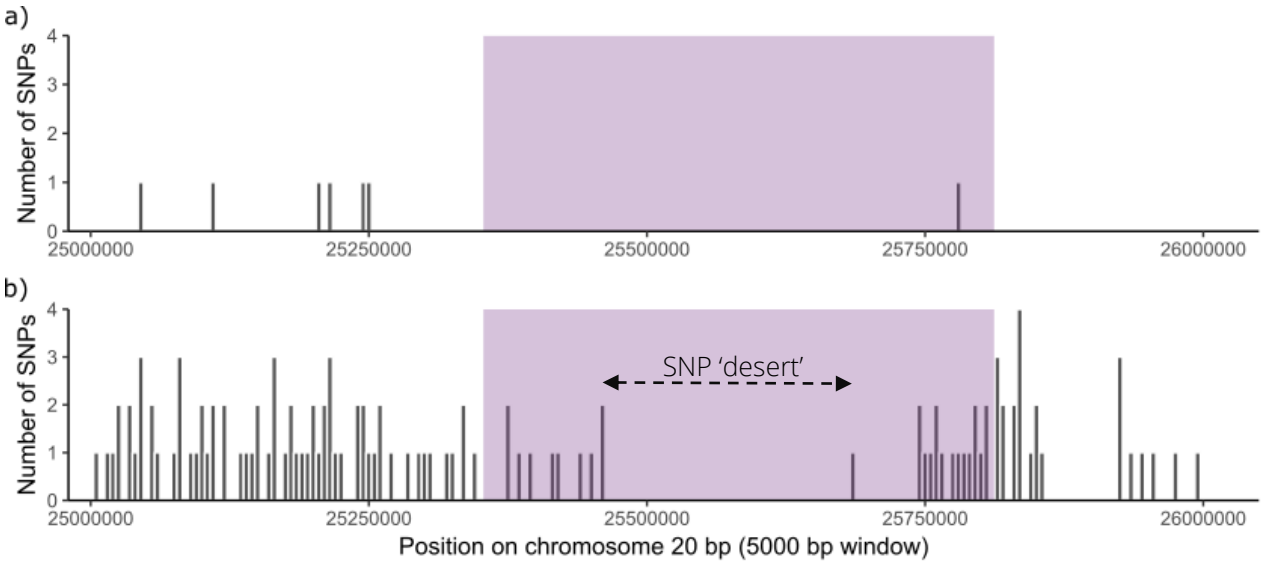


Figure 1. Number of SNPs within each 5000 bp window on chromosome 20 between 25 and 26 Mb on a) the 50K SNP chip and b) the HD SNP chip. The purple block represents the class IIa gene region between 25.353 to 25.812 Mb (determined here – see results). A SNP 'desert', an extended region lacking SNPs on the HD SNP chip, is shown by the dashed arrow.

Here, we develop a high throughput haplotyping system for the Soay sheep MHC class IIa. We hypothesised that on the Illumina Ovine Infinium HD BeadChip there might be a sufficient number of SNPs in the flanking regions that were in linkage disequilibrium with the class IIa region to enable imputation of the Soay haplotypes. We then aimed to identify a panel of SNPs capable of imputing the Soay sheep haplotypes, and to genotype them in a large number of individuals spanning the 30-year study period for which DNA was stored.

121 ## Materials and Methods

122 **Study system and DNA sampling.** The Soay sheep living in the Village Bay are of Hirta, St. Kilda, UK have
123 been intensively monitored since 1985 (Clutton-Brock et al. 2004). Lambs within the study area of
124 Village Bay are caught, ear-tagged and sampled for genetic analysis soon after birth. Individuals which
125 move into the study area as adults are sampled when possible, during spring, August catch-up or the
126 autumn rut. Individual life-histories are determined through lambing observations, censuses of the
127 Village Bay population during spring, August and autumn, and regular mortality checks. The pedigree
128 of Soay sheep has previously been defined using 315 SNPs selected from the 50K SNP chip combined
129 with mother-offspring relationships observed during lambing (Bérénos et al. 2014).

130 **Ovine Infinium HD SNP BeadChip and MHC class IIa genotyping.** Johnston *et al.* previously genotyped
131 188 sheep at 606 066 SNPs on the Ovine Infinium HD SNP BeadChip (2016). These sheep were
132 selected from the pedigree in order to maximise the genetic diversity captured by the fewest
133 individuals.

134 Soay sheep MHC class IIa haplotypes were characterised by Dicks *et al.* (2019) using Sanger sequence-
135 based genotyping of individual loci. We sequenced the *DRB1* and *DQA1, 2* and *2-like* loci in 94 of the
136 sheep that were also genotyped on the Ovine Infinium HD SNP BeadChip, and we sequenced *DRB1*
137 and *DQA1* in the remaining 94 individuals. *DQA2*, *DQA2-like* and *DQB* loci were not genotyped due
138 technical challenges with co-amplification of multiple loci, so it was not possible to distinguish between
139 haplotypes A and G.

140 **Identification of SNPs linked to MHC class IIa haplotypes.** The exact locations of the MHC genes on
141 chromosome 20 are poorly defined, so we first estimated the MHC class IIa gene region on the *O. aries*
142 genome (Oar_v3.1, GCA_000298735.1) on Ensembl (Yates et al. 2016). We searched for protein families
143 including the words "class II histocompatibility" in their description, limiting results to chromosome 20
144 and unmapped scaffolds, and using the single-copy gene *BTNL2* to define the q-arm telomeric end of
145 the class IIa region (Gao et al. 2010; Liu et al. 2011; Lee et al. 2012). The Soay sheep Ovine Infinium HD
146 SNP BeadChip genotypes were then filtered using PLINK v1.90 (Purcell et al. 2007) to select
147 polymorphic SNPs within the identified MHC class IIa region with a genotyping rate > 95%, and a minor
148 allele frequency < 1 %.

149 Pairwise linkage disequilibrium (LD) between SNPs within 500 kb of each other was calculated in
150 Haploview (Barrett et al. 2005) as both r^2 and D'. LD blocks were identified using the Four Gamete Rule
151 (Wang et al. 2002). Blocks were numbered according to their proximity to the SNP 'desert' in the centre
152 of the class II region (**Figure 1**), with those upstream (i.e. closer to the centromere) being negative and

153 those downstream (i.e. closer to the q-arm telomere) being positive. The seven SNPs within block –1
154 were phased using BEAGLE (Browning and Browning 2007) with standard settings and compared to
155 the class IIa haplotypes. Upstream (more negative) SNPs were then added sequentially to block –1
156 SNP haplotypes and phased using BEAGLE after each SNP addition. This identified if and where the
157 SNP and class IIa haplotypes matched, and where LD degraded, i.e. where a single class IIa haplotype
158 was represented by multiple SNP haplotypes. The process was repeated for SNPs downstream of the
159 SNP desert.

160 A recombination between haplotypes B and H was identified (see results), such that the SNP profile
161 matched haplotype H but the class IIa haplotype was haplotype B. To enable identification of this
162 recombinant haplotype using SNPs alone, the second exons of class IIa genes were searched to
163 identify SNPs suitable for a genotyping assay – biallelic across all haplotypes with a minimum of 50 bp
164 of flanking sequence in each direction containing limited variation.

165 KASP SNP panel selection and genotyping. Candidate SNPs were validated *in silico* using Kraken
166 software by LGC Genomics (Hoddesdon) for Kompetitive allele specific PCR (KASP). Of the approved
167 SNPs, alleles and allelic combinations were identified which were unique to each class IIa Sanger
168 haplotype, preferentially selecting those closest to the SNP 'desert' in the centre of the class IIa region.
169 A minimum subset of SNPs was identified which were able to impute the class IIa diplotype of an
170 individual with degeneracy. The final panel of SNPs used for KASP genotyping included 11 SNPs
171 selected from the Ovine Infinium HD BeadChip and two intragenic *DQA1* SNPs.

172 DNA was extracted from tissues (ear punch or *post-mortem*) or buffy coat samples for a previous
173 study using the Illumina Ovine50K SNP chip (Bérénos et al. 2014). The Qiagen DNeasy 96 Blood and
174 Tissue Kit was used following the manufacturer's protocol, with the exception that the final elution was
175 performed using two 50 µL elutions (Bérénos et al. 2014). DNA was normalised to 50 ng/µL for 5951
176 individuals and distributed across 66 96-well plates. A single sheep – the 'golden sheep' (ID 7568) –
177 was included on 61 plates to confirm genotyping consistency and a no template control (NTC) was
178 included on all plates. An additional 200 individuals were genotyped more than once for reasons
179 unrelated to this work (for example when DNA was used for genotyping on the Ovine50K SNP chip,
180 the DNA failed at the first attempt and so was re-attempted on a subsequent plate – note that
181 differences in genotyping technologies between Illumina arrays and KASP may mean that a sample
182 could be successful with one method but not the other). LGC Genomics performed KASP genotyping
183 and automatic genotype calling for all 66 plates.

184 **Quality control and SNP haplotyping.** Quality control was carried out excluding *DQA1* intragenic SNPs
185 as they carry an expected null allele. PLINK v1.90 (Chang et al. 2015) was used to exclude individuals
186 with more than 50% missing genotypes. Genotypes of the 'golden sheep' (ID 7568) were compared
187 across all plates to identify any systematic errors. For other duplicated individuals, DNA was expected
188 to be of poor quality and therefore more susceptible to allelic dropout, and so the genotyping attempt
189 with the highest match to parental genotypes was included. For individuals genotyped in triplicate, the
190 genotyping attempt with fewest mismatches to alternative attempts was retained. Assessments of the
191 multiple genotyping attempts of the 'golden sheep' and 201 repeated individuals, as well as Mendelian
192 error checks between offspring and parents indicated a high genotyping error rate on DNA plates 1-
193 3, the very oldest plates, and so all genotypes from these plates were discounted.

194 Individuals with more than two (of 11) missing genotypes were excluded, and SNPs were phased using
195 BEAGLE v4.0 (Browning and Browning 2007). Phased SNP haplotypes were assigned to their
196 corresponding class IIa haplotypes. Twenty-six individuals with novel haplotypes were excluded as
197 these were presumed genotyping or phasing errors (maximum novel haplotype frequency was four).
198 A Mendelian inheritance check was carried out on the inferred haplotypes for each individual using
199 the Soay sheep pedigree (Bérénos et al. 2014). Three individuals whose phased MHC diplotypes
200 mismatched the diplotype of at least one parent were excluded from further analysis. 5349 sheep
201 passed quality control measures.

202 **Detection of B-H recombinants.** The class IIa gene *DQA1* is not present in all haplotypes (Ballingall et
203 al. 2018; Dicks et al. 2019) and consequently the *DQA1* SNPs are effectively tri-allelic – they have a
204 major allele, a minor allele, and a null allele where the DQA1 locus is absent from the haplotype. A
205 *DQA1* genotype failure can, therefore, represent both true failures and the diplotype of an individual
206 carrying two *DQA1*null* haplotypes. For all sheep identified as carrying haplotype H, the *DQA1* SNP
207 genotypes were assessed to identify B–H recombinants.

208 Results

209 **Identification of SNPs linked to MHC class IIa haplotypes.** Full class IIa diplotypes were determined by
210 sequence-based genotyping for 135 out of the 188 individuals genotyped on the Ovine Infinium HD
211 BeadChip. Partial diplotype information was determined for an additional 50 individuals which carried
212 haplotypes A or G, but these haplotypes were not differentiated (using only *DRB1* and *DQA1* loci).
213 Three individuals failed sequence-based genotyping.

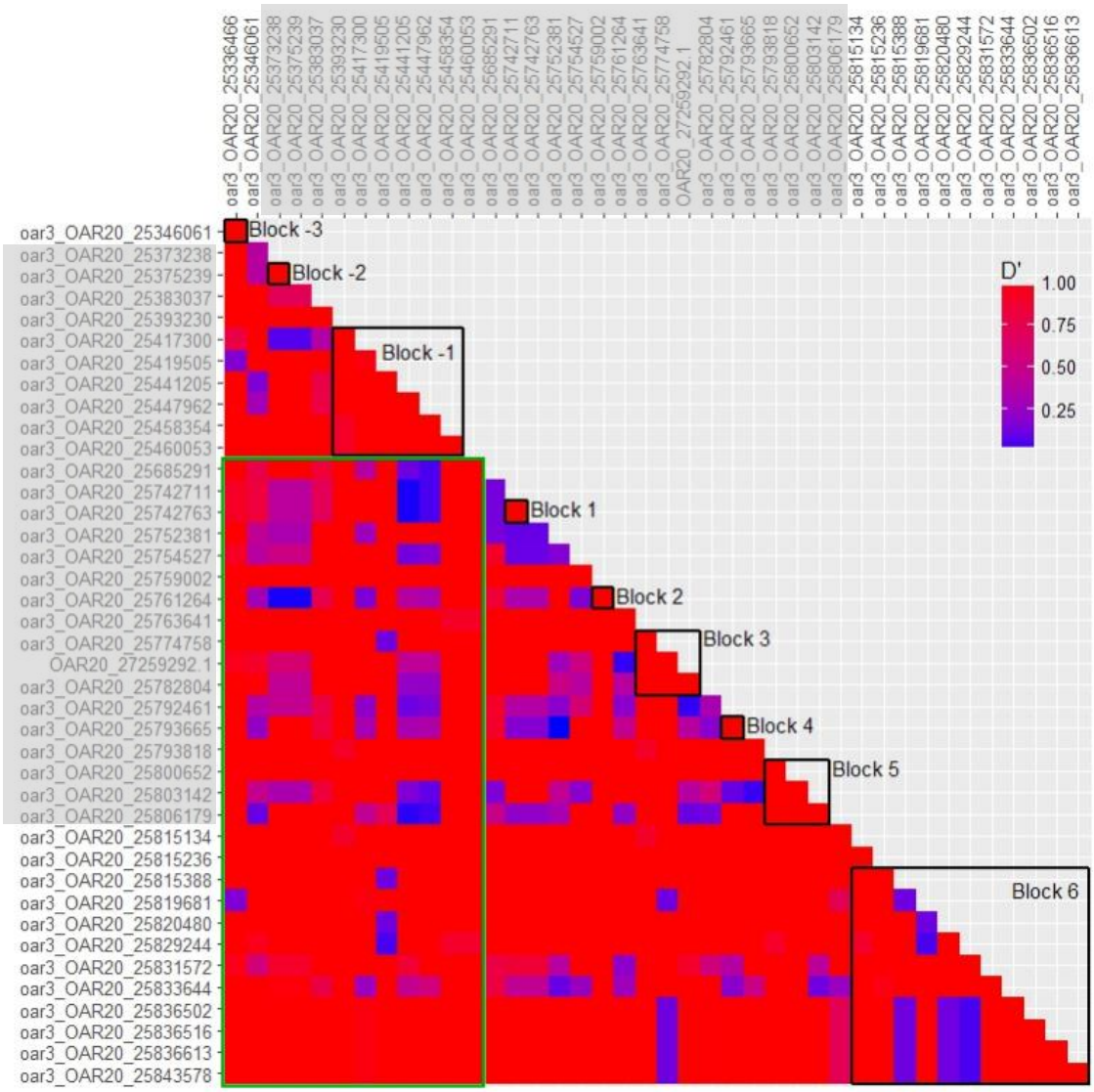
214 Nine protein families represented by 22 transcripts on the Oar_v3.1 genome were identified with the
215 term "class II histocompatibility" in their descriptions, as well as *BTNL2* (Table S1). Five transcripts were

located on unmapped scaffolds. The remaining 17 transcripts were clustered in two regions of chromosome 20, 7.164 – 7.426 Mb and 25.353 – 25.812 Mb, with the former including primarily class IIb gene annotations (*DO* and *DM* genes) and the latter class IIa gene annotations (*DR* and *DQ* genes). *DQA* gene annotations were found in both class II regions, which likely reflects poor sequence assembly or gene annotation in this highly repetitive and polymorphic region. Previous assemblies of this region using BAC cloning did not identify *DQ* loci outside of the class IIa region (Liu et al. 2006; Gao et al. 2010). The classical class IIa genes are therefore expected to be located between 25.353 and 25.812 Mb.

On the Ovine Infinium HD BeadChip, there were 11,757 HD SNPs located on chromosome 20, of which 8,420 SNPs passed quality control in the Soay sheep sample. Twenty-one SNPs were located in the class IIa region predicted above (Figure 1b).

Pairwise LD estimates of D' for the SNPs between 25.353 and 25.812 Mb were typically very high (Figure 2). There was no obvious pattern to the degradation of LD, which would be suggestive of a recombination hotspot. On the other hand, r^2 estimates, were typically very low (Figure S2) and were also uninformative in identifying linkage regions. Given the lack of obvious LD pattern, linkage blocks were identified using the Four Gamete Rule in Haploview (shown in Figure 2).

Block –1, which is immediately upstream of the HD SNP gap, consisted of six SNPs and formed six haplotypes (Figure 3). Four of the six haplotypes in block –1 each corresponded to a single Sanger haplotype, and two each corresponded to a pair of haplotypes, A and G, or B and H (Figure 3). The addition of upstream SNPs to block –1 showed that linkage degraded immediately upstream of block –1 (at SNP oar3_OAR20_25383037) within haplotype E, but haplotypes A and G, as well as haplotypes B and H, only became differentiated at oar3_OAR20_25346061 (block –3). Downstream of the SNP 'desert', haplotypes B and H became differentiated at oar3_OAR20_25742763 (block 1), and A and G differentiated with the addition of oar3_OAR20_25752381 (adjacent to block 1). Linkage within haplotypes did not degrade until oar3_OAR20_25833644 (within block 6) for both haplotypes B and E (Figure 3). Thus, a continuous linkage region of 31 SNPs was identified between oar3_OAR20_25393230 and oar3_OAR20_25831572, which encompasses 438.34 kb.



243

244 Figure 2. D' linkage disequilibrium estimates for pairs of HD SNPs within and surrounding the class
245 IIa gene region, where values of 1 represents high LD, and 0 represents low LD. SNPs within the class
246 IIa gene region are highlighted by the grey shading. The SNP 'desert' is denoted by the green box,
247 and SNP pairs that fall within this green box represent pairs spanning this SNP 'desert' (i.e. that one
248 SNP is upstream and the other downstream). Open black boxes show linkage blocks identified within
249 Haploview using the Four Gamete Rule.

250

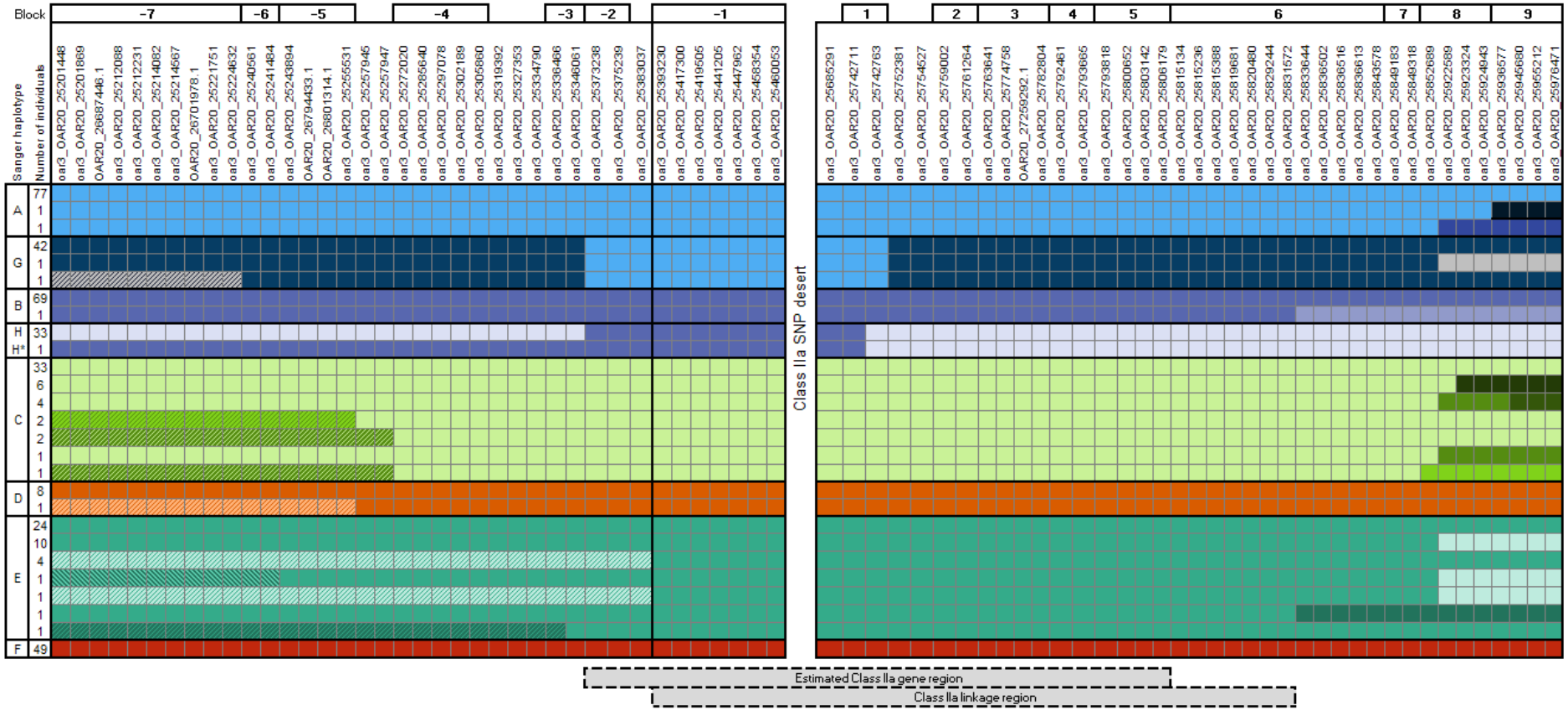


Figure 3. Representation of the phased SNP haplotypes (rows) formed by sequentially and cumulatively adding either upstream (left) or downstream (right) HD SNPs to block –1. Linkage blocks (Four Gamete Rule in Haploview) are indicated above the SNP names, and corresponding Sanger haplotypes in the left-most column. Moving outwards from block –1, a colour change along the row indicates that an additional haplotype is formed with the addition of that SNP. The estimated class IIa gene region (25.353 to 25.812 Mb) and the class IIa linkage region are indicated below the plot. H* is a recombinant haplotype detected in a single individual, for which the SNP haplotype within the class IIa linkage region matches haplotype H but class IIa Sanger sequencing is haplotype B.

258 **Haplotype B and H recombinant detection.** Following Sanger sequencing individual 4179 was found
259 to be heterozygous for haplotypes for B and G. However, the flanking SNPs revealed a recombination
260 event which did not affect the expressed haplotype but affected the haplotype that would be imputed
261 using the panel of 31 SNPs within the class IIa linkage region. Using only the seven SNPs upstream of
262 the SNP 'desert', 4179's SNP diplotype was consistent with haplotypes B/H (identical in this region) and
263 G (**Figure 3**, haplotype H*). Using only the 24 SNPs downstream of the SNP gap, 4179's SNP diplotype
264 was consistent with haplotypes H and G. Therefore, using the SNPs within the class IIa linkage region
265 (i.e. block –1 and downstream SNPs), the haplotype would be incorrectly imputed as H. This indicates
266 the existence of a recombination event between haplotypes B and H which has occurred downstream
267 of the class IIa genes included in the Sanger haplotype.

268 **KASP assay selection and genotyping.** All 31 SNPs within the class IIa linkage region that were selected
269 from the Ovine Infinium HD SNP BeadChip passed *in silico* validation for KASP. From this shortlist, we
270 selected 11 SNPs enabling to imputation the eight class II haplotypes whilst maintaining degeneracy
271 in case of genotyping failures (**Figure 4**; **Table 1**). To enable detection of the B-H recombinant
272 haplotype, two SNPs within exon 2 of the *DQA1* gene were identified (DQA1-171 and DQA1-195; Figure
273 S3) that met the requirements for KASP assay design. No suitable SNPs, which are biallelic, include
274 sufficient flanking sequence with limited variation could be identified in other MHC class IIa genes.
275 *DQA1* is not present on three Soay sheep MHC class IIa haplotypes, A, E and G, which instead carry
276 *DQA2-like* alleles. As such, *DQA1* SNPs are effectively tri-allelic and consist of a major allele, minor
277 allele, and null allele (see **Figure 4**). 5951 individuals were genotyped by LGC Genomics using KASP at
278 the final panel of 13 SNPs.

279

280

281

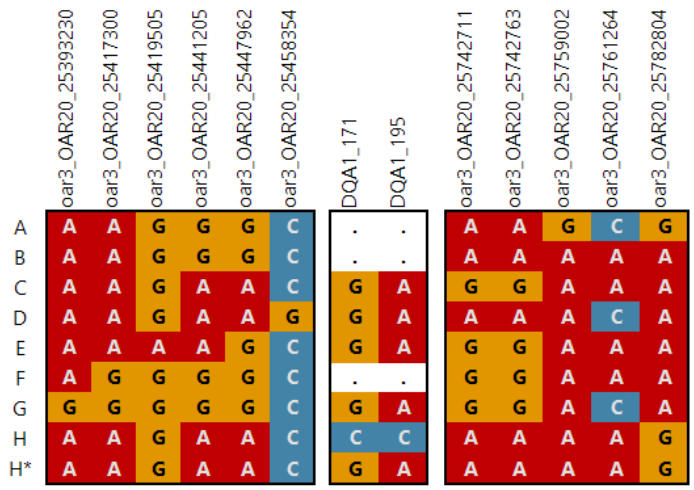


Figure 4. Alleles at KASP-genotyped SNPs (columns) within the MHC class IIa linkage region for each Sanger haplotype (rows). For DQA1 SNPs, dots correspond to DQA1*null alleles on haplotypes where the DQA1 locus is absent. DQA1*null alleles are therefore the absence of amplification. SNP loci are shown in order of their position on chromosome 20.

Quality control. The 'golden sheep' was KASP genotyped 57 times, with an overall genotyping error rate of 0.48% and a maximum error rate of 1.85% per locus. After excluding DNA plates 1-3 due to exceptionally high error rates, 5675 individuals were KASP genotyped for the panel of 13 SNPs. Quality control measures on the 11 flanking SNPs, and samples sizes at each stage are shown in Figure S4. The genotyping rate was below 50% for 195 sheep, which were excluded.

After assessing genotyping attempts for repeated individuals and excluding the least successful genotyping attempt, 102 individuals were removed due to having four or more missing genotypes. The 11 flanking SNPs were phased in the remaining 5378 individuals. This revealed eight frequent haplotypes which exactly matched the eight expected class IIa haplotypes. Twenty novel haplotypes were identified in 26 individuals (with a maximum frequency of four, see Table S2) and were excluded from further analysis as they were considered probably due to genotyping errors or extremely rare haplotypes. Mendelian error checks identified only three sheep whose haplotypes mismatched their parents'; the offspring were removed from the dataset. The final dataset consisted of diplotypes for 5349 individuals. In this sample, all SNPs were in HWE, except for the *DQA1* SNPs which were expected to deviate from HWE due to null alleles (**Table 1**).

305 **Table 1.** Locus information, genotyping success rate (%), MAF (minor allele frequency), H_O (observed
306 heterozygosity), H_E (expected heterozygosity), and Hardy-Weinberg Equilibrium (HWE) significance
307 values, for the SNPs selected for KASP genotyping. SNP statitistics were calculated for the 5349
308 individuals which passed quality control. Genome positions for DQA1 loci are unknown.

SNP ID	rs accession number [a]	Genome Position	Major Allele	Minor Allele	Genotyping success rate (%)	MAF	H_O	H_E	HWE p-value
oar3_OAR20_25759002	rs401547149	25759002	G	A	98.38	0.170	0.286	0.282	0.802
oar3_OAR20_25458354	rs403182119	25458354	C	G	99.38	0.123	0.205	0.215	0.327
oar3_OAR20_25447962	rs407228358	25447962	A	G	97.63	0.439	0.499	0.492	0.809
oar3_OAR20_25393230	rs408059130	25393230	G	A	99.25	0.145	0.254	0.248	0.636
oar3_OAR20_25742711	rs408458106	25742711	G	A	97.25	0.485	0.490	0.500	0.670
oar3_OAR20_25441205	rs412204478	25441205	T	C	98.38	0.473	0.517	0.499	0.476
oar3_OAR20_25417300	rs412489249	25417300	G	A	97.88	0.243	0.335	0.368	0.060
oar3_OAR20_25761264	rs414259673	25761264	C	A	98.25	0.442	0.502	0.493	0.738
oar3_OAR20_25419505	rs419818757	25419505	T	C	98.38	0.031	0.063	0.061	0.677
oar3_OAR20_25742763	rs421513037	25742763	C	T	97.75	0.490	0.494	0.500	0.813
oar3_OAR20_25782804	rs422511469	25782804	G	A	94.51	0.279	0.403	0.402	0.952
DQA1_171	NA	NA	C	G	83.52[b]	0.146	0.118	0.249	<0.0001
DQA1_195	NA	NA	C	A	82.77[b]	0.148	0.116	0.252	<0.0001

309 [a] Accession number in Ensembl http://www.ensembl.org/Ovis_aries/
310 [b] SNP carries known null-allele
311

312

313 ***DQA1* SNP genotypes.** *DQA1* SNP genotypes were compared to the diplotypes identified by the panel
314 of 11 flanking SNPs. The two selected *DQA1* SNPs, *DQA1*-171 and *DQA1*-195, are tri-allelic such that
315 haplotype H carries the minor allele (C for both loci), four haplotypes (B, C, D, and F) have the major
316 allele (G or A respectively), and three haplotypes (A, E and G) do not carry the locus and present as
317 null alleles. As such, it is possible to predict the expected genotypes for the *DQA1* SNPs from an
318 individual's diplotype. In a test sample of 801 sheep, five had unexpected *DQA1* genotypes, all of which
319 carried *DQA1**null haplotypes (see Figure S5), which is an error rate of 0.62%. Additionally, genotypes
320 classified as fail, which include true negatives, true failures to amplify, and homozygote null-alleles,
321 showed greater variance in fluorescence for the *DQA1* SNPs compared to the panel of 11 truly bi-
322 allelic SNPs. This suggests non-specific amplification may be occurring in the absence of the target
323 locus. Nevertheless, B–H recombinant haplotypes could be identified (see Figure S5).

324

In the full dataset, genotypes for the two *DQA1* SNPs were assessed only in individuals carrying haplotype H (N=906), identifying 56 recombinant individuals. After recoding recombinant haplotypes as R, all recombinant individuals were found to be descendants of the individual in which the recombinant haplotype was identified, male 4179 (Figure 5). Furthermore, the mother of 4179 carried both B and H haplotypes, and thus was the likely origin of the cross-over event.

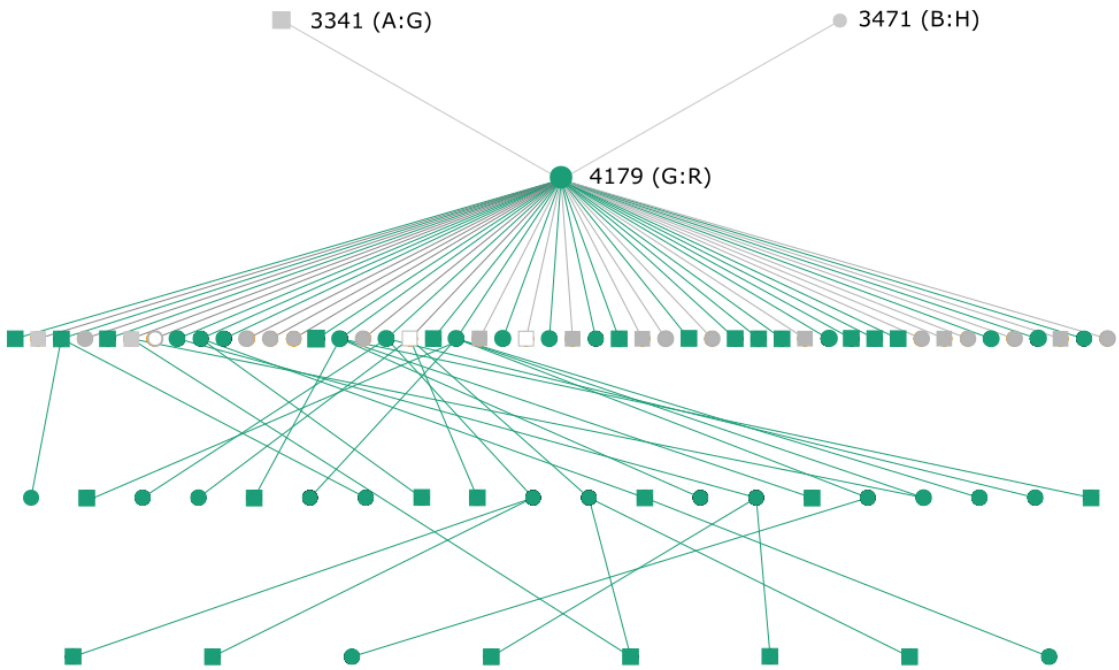


Figure 5. Pedigree of individual 4179, who carries the first identified B–H recombinant haplotype. Green represents individuals carrying the recombinant haplotype, grey not carrying the recombinant haplotype, and open represents individuals that have not been genotyped. Squares are males, circles are females. All offspring of ID 4179 are shown, but subsequent descendants are only shown if they carry the recombinant haplotype. Diplotypes of 4179 and his parents are shown in brackets.

Haplotype and diplotype frequencies.

Of 5349 individuals, 976 (17.9%) had homozygous diplotypes. Haplotype B was the most frequent, whilst haplotype D the rarest, with only six individuals homozygous for haplotype D (Figure 6b). All possible diplotypes were observed (Figure 6a). Haplotype frequencies for individuals born in 3-year periods from 1985 to 2012 (5137 individuals with known birth years) varied over time (Figure 7).

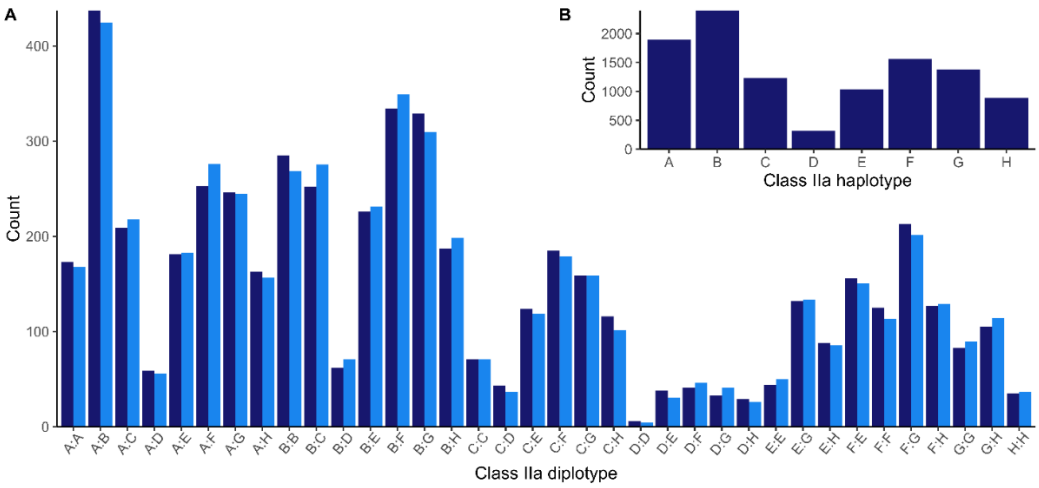


Figure 6. Observed (dark) and expected (light) diplotypes frequencies under HWE (A) and observed haplotype frequencies (B). N= 5349 individuals in both cases.

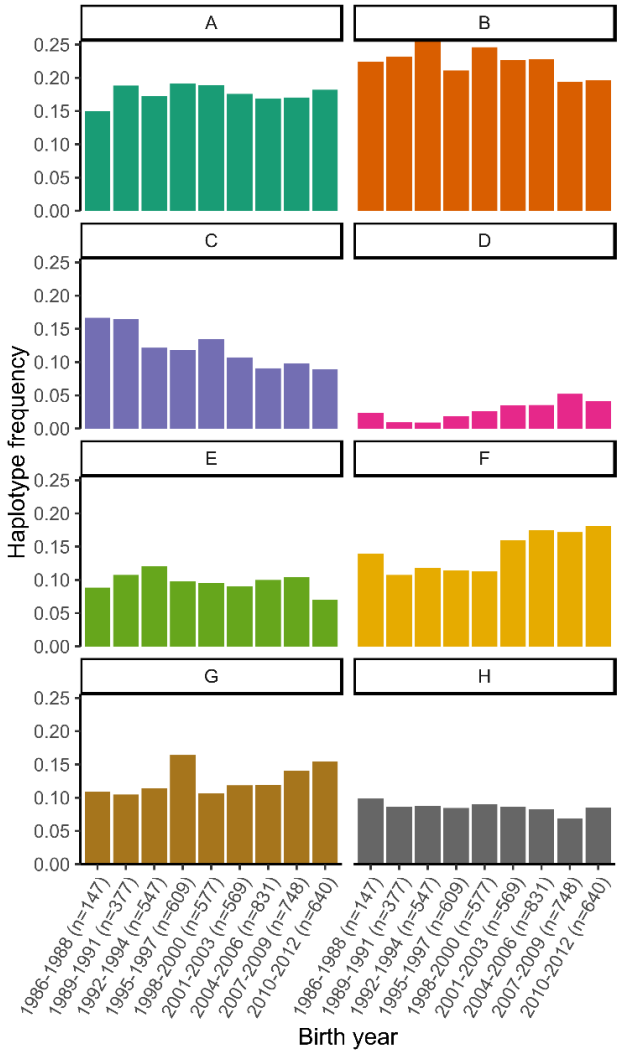


Figure 7. Haplotype frequencies in individuals born in each 3-year period. Sample sizes are shown beside the birth year category.

349 ## Discussion

350 Analyses of the evolutionary processes underlying variation at the MHC require population-scale data.
351 To achieve this, a SNP-based genotyping method was developed for rapid and accurate determination
352 of the class IIa diplotypes in Soay sheep. Using SNPs selected from the Ovine Infinium HD BeadChip,
353 a linkage region surrounding the class IIa genes was identified, and from it a minimal panel of 11 SNPs
354 was selected that would enable imputation of the class IIa haplotypes. Furthermore, a recombination
355 event between haplotypes B and H was detected, and two intragenic SNPs were selected to detect
356 such recombinant haplotypes. Finally, this panel of 13 SNPs was applied to 5349 Soay sheep using
357 KASP genotyping technology.

358 The classical MHC class IIa genes were estimated as occurring between 25.353 and 25.812 Mb on the
359 *O. aries* genome, although poor mapping within this region prevents identification of the exact
360 positions of MHC genes. Nevertheless, twenty-one SNPs on the Ovine Infinium BeadChip were located
361 within the identified class IIa gene region, with a 0.225 Mb 'SNP desert' located in the centre of this
362 region. This lack of SNPs is not a reflection of a lack of variable SNPs in this region but rather their
363 unsuitability for incorporation on a SNP chip due to the presence of low minor allele frequencies, and
364 variable flanking sequences that prevent the design of robust primers and probes.

365 Pairwise measures of LD, D' and r^2, showed opposing patterns across the MHC class IIa region.
366 Pairwise D' was typically very high, whilst pairwise r^2 was typically very low in this region. This opposing
367 pattern of D' and r^2 measures of LD was also found in the class IIa, IIb and III regions in Rylington
368 Merino sheep (Lee et al. 2012). VanLiere and Rosenberg (2008) showed that r^2 and D' are not always
369 directly related, and r^2 is sensitive to MAF, especially MAF < 0.3 which is not uncommon within the
370 MHC (see **Table 1**). Lee *et al.* (2012) showed that pairwise LD using D' in Rylington Merino sheep was
371 high within class IIa but decreased between the class IIa and class IIb regions, as well as between class
372 IIa and class III regions. This suggests recombination *within* the class IIa is reduced compared to
373 recombination between it and the neighbouring MHC subregions. Pairwise LD measures were
374 therefore perhaps unlikely to reveal a relationship between the class IIa haplotypes and flanking SNPs
375 on the Ovine Infinium HD BeadChip.

376 An alternative method of assessing the pattern of LD between the flanking SNPs and the class IIa
377 region, the Four Gamete Rule, was used to identify linkage blocks as a starting point to cumulatively
378 add SNPs to determine where SNP haplotypes degraded. This method detected a linkage region of 31
379 SNPs that matched the class IIa haplotypes. This class IIa linkage region extended between 25.39 –
380 25.83 Mb according to the allocated physical positions of the SNPs, which sits well within the estimated
381 class IIa gene region of 25.35 – 25.81 Mb. The downstream extent of the estimated class IIa gene

382 region is marked by the gene *BTNL2*, which, because it is a single copy gene, is more likely to be
383 accurately mapped than the other multi-copy genes in this region. The overlap between the estimated
384 class IIa gene region and the SNP linkage region suggests that at least 24 of the 31 SNPs are located
385 within the estimated class IIa region.

386 Diplotypes for the 5349 Soay sheep which passed quality control can be considered highly accurate.
387 Indirect MHC typing methods, such as SNP genotyping, are faster and cheaper than direct sequencing
388 methods but may incur an increased error rate which must be minimised. Phasing SNPs to determine
389 the two SNP haplotypes carried by an individual, carried out here with BEAGLE (Browning and
390 Browning 2007), uses information from all individuals within the dataset to calculate the most likely
391 haplotypes. BEAGLE also uses this information to impute missing SNP genotypes. Therefore, phasing
392 errors are increased when there is a high genotyping error rate, and thus errors can be reduced by
393 using a more stringent genotyping rate cut-off across all SNPs. The repetition of a single individual on
394 almost all DNA plates was valuable in detecting systematic errors on three plates, probably due to the
395 DNA having been exhausted or degraded by freeze-thawing and revealed a genotyping error rate of
396 0.48%. The Mendelian inheritance check demonstrated that the phased haplotypes were consistent
397 between parents and offspring, with only three exceptions. Additionally, only 20 novel haplotypes were
398 identified in 26 individuals (0.5 % of haplotyped individuals), which were probably due to genotyping
399 errors, including allelic dropout. Whilst this suggests that genotyping errors can lead to incorrect
400 haplotype inferences, in this population, with good pedigree information, it is clearly possible to detect
401 such errors.

402 A recombination event between haplotypes B and H (Sanger haplotype B, SNP haplotype H) was
403 identified in one individual (ID 4179) genotyped on the Ovine Infinium HD BeadChip. The
404 recombination event must have occurred downstream of the expressed class IIa genes characterised
405 within the Sanger haplotypes (*DRB1, DQA* and *DQB*) but upstream of the SNP oar3_OAR20_25742763,
406 leaving the functional class IIa haplotype unaffected. Individuals were selected for genotyping on the
407 Ovine Infinium HD BeadChip because they were highly representative of the genetic variation within
408 the Soay sheep population (Johnston et al. 2016), and male 4179 was included because he was very
409 prolific with a large number of descendants. Using two SNPs located within the exons of the class IIa
410 locus *DQA1*, this recombination was detected in 54 of 4179's descendants and, because both of 4179's
411 parents were genotyped, we were able to detect the origin of the recombination event during
412 gametogenesis in his mother. The use of SNPs located within loci that are not present on every
413 haplotype, as for the two *DQA1* loci, was less than ideal due to the increased error rate caused by non-
414 specific amplification. Nevertheless, their inclusion on the KASP panel enabled the identification of a

415 substantial proportion of recombinant individuals with high accuracy, as apparent by a lack of parent-
416 offspring mismatches following recombinant calling. Were this recombination event not identified, and
417 therefore not detectable, a substantial number of haplotypes would have been incorrectly imputed,
418 which could have impacted subsequent analyses.

419 Whilst the *DQA1* SNPs were effective for detecting the identified downstream B to H recombination, it
420 is unlikely to be possible to detect novel recombination events involving upstream SNPs between
421 haplotypes A and G. The six SNPs located upstream of the class IIa genes and the *DQA1* SNPs were
422 identical for these haplotypes (see Figure 3 and Figure 4). All other haplotype pairs have unique
423 upstream and downstream SNP profiles.

424 All possible diplotypes were identified within the population, although the rarest, homozygote for D,
425 was only identified in six out of the 5349 individuals genotyped (0.11 %). Over the study period (1985
426 – 2012), 5137 haplotyped individuals had known birth years (or expected birth years in the case of
427 foetuses). There was no evidence of dramatic shifts in haplotype frequencies, but they were found to
428 vary over time. There may be a slight decrease in haplotype C and a slight increase in haplotype D
429 (Figure 7) which warrants further investigation to identify whether these shifts are significant, and if so,
430 if they are due to natural or selective forces.

431 Genotyping the MHC is notoriously challenging (Bernatchez and Landry 2003; Piertney and Oliver
432 2006; Babik 2010; Spurgin and Richardson 2010) but accurate, locus-specific genotypes are vital for
433 unravelling the modes of evolution operating to maintain diversity at the MHC. Previous work
434 identified eight MHC class II haplotypes in the island population of Soay sheep (Dicks et al. 2019) using
435 sequence-based genotyping, and here we have described a SNP-based (KASP) genotyping method
436 which enabled rapid genotyping of 5349 sheep (out of a total 5675 sheep attempted). A panel of SNPs
437 has the advantage of being relatively easy to genotype, suitable for high-throughput genotyping and
438 does not suffer from artefacts, but it is susceptible to incorrect haplotype phasing due to genotyping
439 errors, particularly allelic dropout which may be most prevalent in poor quality or quantity DNA (Gill
440 2001; Butler 2005). The stringent quality control measures applied here should minimise any effects
441 of genotyping and phasing errors, and the Soay sheep pedigree was particularly valuable here in
442 enabling the inclusion of a Mendelian inheritance check as a quality control measure. As a method for
443 genotyping MHC in a non-model organism, however, this panel of SNPs required high initial
444 optimisation effort in terms of preliminary class IIa sequencing and is not easily transferable to other
445 breeds of sheep or other species without extensive preliminary sequencing of haplotypes.
446 Nevertheless, this method has provided a valuable haplotype-level MHC data set that is accompanied

447 by extensive phenotypic and fitness data that will enable future studies on the evolutionary processes
448 acting to maintain diversity at the MHC in a natural environment.

449

450 ## Acknowledgements

451 We thank the National Trust for Scotland and the Scottish Natural Heritage for permission to work on
452 St. Kilda and QinetiQ for logistics and support. Many thanks to Jill Pilkington and all Soay sheep project
453 members and volunteers for collection of data and samples. Thanks to Phil Elli and Camillo Bérénos
454 who prepared DNA samples for Ovine SNP50 and HD BeadChip genotyping. KLD was supported by
455 the Biotechnology and Biological Sciences Research Council (BBSRC) (grant number BB/J01446X/1).
456 KTB receives funding from the European Union's Horizon 2020 research and innovation programme
457 under grant agreement no. 731014KB (VetBioNet) and acknowledges the support received from the
458 Scottish Government's strategic research programme. The Soay Sheep Project is funded by the UK
459 Natural Environment Research Council. There are no conflicts of interest.

460

461 ## References

462 Ali AOA, Stear A, Fairlie-Clarke K, et al (2016) The genetic architecture of the MHC class II region in
463 British Texel sheep. Immunogenetics 1–7. doi: 10.1007/s00251-016-0962-6

464 Apanius V, Penn D, Slev PR, et al (1997) The nature of selection on the major histocompatibility
465 complex. Crit Rev Immunol 17:179–224. doi: 10.1615/CritRevImmunol.v17.i2.40

466 Babik W (2010) Methods for MHC genotyping in non-model vertebrates. Mol Ecol Resour 10:237–251.
467 doi: 10.1111/j.1755-0998.2009.02788.x

468 Ballingall KT, Lantier I, Todd H, et al (2018) Structural and functional diversity arising from intra and
469 inter haplotype combinations of duplicated DQA and B loci within the Ovine MHC.
470 Immunogenetics 70:257–269. doi: 10.1007/s00251-017-1029-z

471 Ballingall KT, Steele P, Lantier I, et al (2015) An ancient interlocus recombination increases class II MHC
472 DQA diversity in sheep and other Bovidae. Anim Genet 46:333–336. doi: 10.1111/age.12290

473 Ballingall KT, Tassi R (2010) Sequence-based genotyping of the sheep MHC class II DRB1 locus.
474 Immunogenetics 62:31–39. doi: 10.1007/s00251-009-0410-y

475 Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: Analysis and visualization of LD and haplotype
476 maps. Bioinformatics 21:263–265. doi: 10.1093/bioinformatics/bth457

477 Bérénos C, Ellis PA, Pilkington JG, Pemberton JM (2014) Estimating quantitative genetic parameters in
478 wild populations: A comparison of pedigree and genomic approaches. Mol Ecol 23:3434–3451.
479 doi: 10.1111/mec.12827

480 Bernatchez L, Landry C (2003) MHC studies in nonmodel vertebrates: What have we learned about
481 natural selection in 15 years? J Evol Biol 16:363–377. doi: 10.1046/j.1420-9101.2003.00531.x

482 Boyd JM, Doney JM, Gunn RG, Jewell PA (1964) The soay sheep of the island of Hirta, St. Kilda. A study
483 of a feral population. Proc Zool Soc London 142:129–164. doi: 10.1111/j.1469-
484 7998.1964.tb05159.x

485 Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference
486 for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum
487 Genet 81:1084–1097. doi: 10.1086/521987

488 Butler JM (2005) Forensic DNA typing: biology, technology, and genetics of STR markers, 2nd edn.
489 Elsevier, Burlington

490 Chang CC, Chow CC, Tellier LCAM, et al (2015) Second-generation PLINK: rising to the challenge of
491 larger and richer datasets. Gigascience 4:7. doi: 10.1093/bioinformatics/btu495

492 Clutton-Brock TH, Pemberton JM, Coulson T (2004) The sheep of St Kilda. In: Clutton-Brock TH,
493 Pemberton JM (eds) Soay Sheep: Dynamics and selection in an island population. Cambridge
494 University Press, Cambridge, pp 17–50

495 Dawkins R, Leelayuwat C, Gaudieri S, et al (1999) Genomics of the major histocompatibility complex:
496 haplotypes, duplication, retroviruses and disease. Immunol Rev 167:275–304. doi:
497 10.1111/j.1600-065X.1999.tb01399.x

498 de Bakker PIW, Raychaudhuri S (2012) Interrogating the major histocompatibility complex with high-
499 throughput genomics. Hum Mol Genet 21:R29–R36. doi: 10.1093/hmg/dds384

500 Dicks KL, Pemberton JM, Ballingall KT (2019) Characterisation of major histocompatibility complex class
501 IIa haplotypes in an island sheep population. Immunogenetics 1–11. doi: 10.1007/s00251-019-
502 01109-w

503 Dilthey AT, Moutsianas L, Leslie S, McVean G (2011) HLA*IMP--an integrated framework for imputing
504 classical HLA alleles from SNP genotypes. Bioinformatics 27:968–972. doi:
505 10.1093/bioinformatics/btr061

506 Doherty PC, Zinkernagel RM (1975) Enhanced immunological surveillance in mice heterozygous at the
507 H-2 gene complex. Nature 256:50–52. doi: 10.1038/256050a0

508 Eizaguirre C, Lenz TL, Kalbe M, Milinski M (2012) Divergent selection on locally adapted major
509 histocompatibility complex immune genes experimentally proven in the field. Ecol Lett 15:723–
510 731. doi: 10.1111/j.1461-0248.2012.01791.x

511 Gao J, Liu K, Liu H, et al (2010) A complete DNA sequence map of the ovine major histocompatibility
512 complex. BMC Genomics 11:466. doi: 10.1186/1471-2164-11-466

513 Gill P (2001) Application of low copy number DNA profiling. Croat Med J 42:229–232

514 Hedrick PW (2002) Pathogen resistance and genetic variation at MHC loci. Evolution (N Y) 56:1902–
515 1908. doi: 10.1111/j.0014-3820.2002.tb00116.x

516 Hill AVS (1991) HLA associations with malaria in Africa: some implications for MHC evolution. In: Klein
517 J, Klein D (eds) Molecular evolution of the major histocompatibility complex. Springer, Berlin, pp
518 403–419

519 Hill AVS (1998) The immunogenetics of human infectious diseases. Annu Rev Immunol 16:593–617

520 Johnston SE, Bérénos C, Slate J, Pemberton JM (2016) Conserved Genetic Architecture Underlying
521 Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries). Genetics
522 203:

523 Kalbe M, Eizaguirre C, Dankert I, et al (2009) Lifetime reproductive success is maximized with optimal
524 major histocompatibility complex diversity. Proc R Soc London B Biol Sci 276:. doi: DOI:
525 10.1098/rspb.2008.1466

526 Kijas JW, Lenstra JA, Hayes B, et al (2012) Genome-Wide Analysis of the World's Sheep Breeds Reveals
527 High Levels of Historic Mixture and Strong Recent Selection. PLoS Biol 10:e1001258. doi:
528 10.1371/journal.pbio.1001258

529 Kijas JW, Townley D, Dalrymple BP, et al (2009) A Genome Wide Survey of SNP Variation Reveals the
530 Genetic Structure of Sheep Breeds. PLoS One 4:e4668. doi: 10.1371/journal.pone.0004668

531 Klein J (1986) The Natural History of the Major Histocompatability Complex. Wiley, New York, NY

532 Lawson Handley LJ, Byrne K, Santucci F, et al (2007) Genetic structure of European sheep breeds.
533 Heredity (Edinb) 99:620–631. doi: 10.1038/sj.hdy.6801039

534 Lee CY, Qin J, Munyard KA, et al (2012) Conserved haplotype blocks within the sheep MHC and low
535 SNP heterozygosity in the Class IIa subregion. Anim Genet 43:429–437. doi: 10.1111/j.1365-
536 2052.2011.02268.x

537 Liu H, Liu K, Wang J, Ma RZ (2006) A BAC clone-based physical map of ovine major histocompatibility
538 complex. Genomics 88:88–95. doi: 10.1016/j.ygeno.2006.02.006

539 Liu K, Zhang P, Gao J, et al (2011) Closing a gap in the physical map of the ovine major histocompatibility
540 complex. Anim Genet 42:204–207. doi: 10.1111/j.1365-2052.2010.02083.x

541 Paterson S (1998) Evidence for balancing selection at the major histocompatibility complex in a free-
542 living ruminant. J Hered 89:289–294. doi: 10.1093/jhered/89.4.289

543 Paterson S, Wilson K, Pemberton JM (1998) Major histocompatibility complex variation associated with
544 juvenile survival and parasite resistance in a large unmanaged ungulate population. PNAS
545 95:3714–3719. doi: 10.1073/pnas.95.7.3714

546 Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex.
547 Heredity (Edinb) 96:7–21. doi: 10.1038/sj.hdy.6800724

548 Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: a tool set for whole-genome association and
549 population-based linkage analyses. Am J Hum Genet 81:559–575. doi: 10.1086/519795

550 Slade RW, McCallum HI (1992) Frequency-Dependent Selection at MHC Loci. Genetics 132:861–862

551 Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and
552 misunderstandings. Proc R Soc London B Biol Sci 277:979–988. doi: 10.1098/rspb.2009.2084

553 VanLiere JM, Rosenberg NA (2008) Mathematical properties of the measure of linkage disequilibrium.
554 Theor Popul Biol 74:130–137. doi: 10.1016/j.tpb.2008.05.006

555 Wang N, Akey JM, Zhang K, et al (2002) Distribution of recombination crossovers and the origin of
556 haplotype blocks: the interplay of population history, recombination, and mutation. Am J Hum
557 Genet 1227–34. doi: 10.1086/344398

558 Westerdahl H (2007) Passerine MHC: genetic variation and disease resistance in the wild. J Ornithol
559 148:469. doi: 10.1007/s10336-007-0230-5

560 Whittaker DJ, Dapper AL, Peterson MP, et al (2012) Maintenance of MHC Class IIB diversity in a recently
561 established songbird population. J Avian Biol 43:109–118. doi: 10.1111/j.1600-
562 048X.2012.05504.x

563 Worley K, Gillingham MA., Jensen P, et al (2008) Single locus typing of MHC class I and class II B loci in
564 a population of red jungle fowl. Immunogenetics 60:233–247. doi: 10.1007/s00251-008-0288-0

565 Yates A, Akanni W, Amode MR, et al (2016) Ensembl 2016. Nucleic Acids Res 44:D710–D716. doi:
566 10.1093/nar/gkv1157

567 Zheng X, Shen J, Cox C, et al (2014) HIBAG—HLA genotype imputation with attribute bagging.
568 Pharmacogenomics J 14:192–200. doi: 10.1038/tpj.2013.18

569