# Supplementary Information for

## Lack of evidence for selection favouring MHC haplotypes that combine high functional diversity

Arnaud Gaigher[1], Alexandre Roulin[1], Walid H. Gharib[2], Pierre Taberlet[3,4], Reto Burri[5*] and Luca Fumagalli[1*]

[1] Laboratory for Conservation Biology, Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland

[2] Interfaculty Bioinformatics Unit, University of Bern, CH-3012 Bern, Switzerland

[3] CNRS, Laboratoire d'Ecologie Alpine (LECA), 38000 Grenoble, France

[4] Univ. Grenoble Alpes, Laboratoire d'Ecologie Alpine (LECA), 38000 Grenoble, France

[5] Department of Population Ecology, Institute of Ecology & Evolution, Friedrich Schiller University Jena, Dornburger Strasse 159, D-07743 Jena, Germany

* Joint senior authors

Corresponding author: Arnaud Gaigher, Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland. Fax: +4121 692 42 65; E-mail: arnaud.gaigher@gmail.com

**Supplementary Methods**

*MHC-IIB: primer development, PCR amplification and 454 sequencing preparation*

Barn owl's MHC class IIβ (MHC-IIB) genes DAB1 and DAB2 were previously isolated and characterized (Burri *et al.*, 2008). Forward primers Tyal-int1F (Burri *et al.*, 2008) and Tyal-DAB2-int1F (5'-CTCCCCGTGTCTGCCTGTGC-3') are situated in the region of intron 1 divergent between DAB1 and DAB2 and together with the single reverse primer Tyal-int2R (5'-GACGCGCTGCCACGCACTC-3') allow for the specific amplification of the species' polymorphic exon 2.

PCR reactions were carried out on Biometra T3000 thermocycler in a final volume of 25 µl containing approximately 10ng DNA, 1× buffer Gold, 2.0 mM MgCl2, 1x Q solution (Qiagen), 0.2 mM dNTPs, 0.25 µM each primer, and 1 U AmpliTaq Gold (Applied Biosystems, Switzerland). PCR conditions included an initial denaturation step at 95°C for 10 min, 30 cycles of denaturation at 95°C for 30 sec, primer annealing at 60°C (DAB1 and DAB2) for 45 sec, and primer extension at 72°C for 45 sec. A final step at 72°C for 7 min was used to complete primer extension.

We chose the 454 pyrosequencing protocol (Roche, Basel, Switzerland) to sequence efficiently MHC-IIB genes for the barn owl. One full run divided in eight different regions within the PicoTiterPlate was used for the parallel sequencing of the Swiss population. In order to identify individuals within region, PCR primers were tagged in 5' using 128 different tags of seven bp.

PCR purification prior to sequencing was carried out using the QIAquick PCR purification kit by pooling eight PCR products of similar amplification strength simultaneously on a single column. In order to equilibrate DNA volumes among PCRs, DNA concentrations of target amplicons were previously quantified either visually on

agarose gels or using the QIAxcel screening kit (Qiagen) on an eGene HDA-GT12™ machine. Prior to multiplexing of PCR products per sequencing region, DNA concentrations of purified PCR products were quantified using a Nanodrop ND-1000 spectrophotometer, in order to multiplex equal DNA volumes per individual and locus.

*MHC-IIB: raw data processing and genotyping*

Contrasting with most studies, here we took advantage of (i) the independent amplification of both MHC-IIB loci, with the expectation of a maximum of two alleles per sample, and (ii) the previously known allelic characterization (Burri *et al.*, 2008). The 454 technology used to sequence MHC-IIB loci resulted in an average coverage of 78 sequences per amplicon. However, preliminary view of the data showed that many samples have a low sequence count and/or high proportion of artifacts (mainly indels). Consequently, we deployed sequence similarity-based clustering approach to cluster true alleles with their potential artifacts. As a result, the sequence number of true alleles increases, facilitating their identification.

The processing phase is composed of three main steps. During these steps we tried to keep the same line of reasoning of previous studies using a high-throughput sequencing approach (Galan *et al.*, 2010; Sommer *et al.*, 2013; Lighten *et al.*, 2014; Stutz and Bolnick, 2014; Sebastian *et al.*, 2016), but adapted for our data. Whereas the two first steps are common to the MHC genotyping area, the last step aims to generate clusters encompassing true alleles with their artifacts. All these steps are described below and illustrated by a flow chart (Figure A).

The first processing phase intends to conserve the best quality sequences from the 454 raw data. Data were filtered to keep only sequences with a maximum of two errors within primers, and none within tags. Then, sequences longer than 200 bp and

with a count greater than one were retained (=singletons removed). Finally, identical sequences were grouped according to individual barcodes.

The second processing phase aims to reduce the number of variants on the whole data set by removing rare artifactual variants and samples with low coverage. Each variant with a maximum sequence count per individual lower than three were discarded. Due to low amplification intensity, only samples covered by less than 10 sequences were excluded from the analyses.

The third processing phase aims to cluster true alleles with their potential artifacts. Before starting the procedure, data were organized as follows: (i) a table containing the sequence count for each variant and sample (variants in rows and samples in columns) and (ii) variants were sorted according to sequence count (from largest to smallest). The clustering procedure uses a top-down variant similarity comparison at two different scales, first at a whole dataset scale, and then at the individual scale to group artifacts with their true alleles. The procedure is based on three assumptions: (i) at the whole dataset , true alleles should be found at higher frequency than their own artifacts, (ii) artifacts should be similar to true alleles, differentiating only by 1 or 2 indels (especially in homopolymer regions) and/or substitutions, and (iii) artifacts have to always co-occur with their own true alleles in the individual amplicon (for artifacts arising during PCR).

First, variants sorted according to their sequence count were aligned using MAFFT (Katoh and Standley, 2013). Paired comparisons were carried out by a top-down analysis (Figure B). When the top variant did not possess ambiguous sites (i.e. "n") and had the expected length of 279 and 270 bp respectively for DAB1 and DAB2, we performed a comparison with the second variant of the list. If only one or two indels of divergence were observed between both sequences, the second variant was accepted as

an artifact of the top variant. All the information related to the second variant (i.e. sequence count occurring in each individual) were added to the top one. When the first variant was compared with all other variants, we moved to the next one, and re-performed the procedure. At the end, we obtained a reduced table, in which artifactual variant sequence counts were added to their cluster variant. During the procedure we control that, (i) the top variants receiving new sequences should occur within samples, and (ii) each variant should not be clustered in several cluster variants.

Second, we performed a procedure similar to the previous one, but here at the individual scale. Variants were aligned independently within each individual using MAFFT. Then, paired comparisons were carried out based on the following scoring method within each individual:

Score:

Substitution in non-conserved region: -12

Substitution in conserved region: -4

Indel: -1

Match: 0

Threshold: -11

Delimitations between conserved and non-conserved regions across the nucleotide sequence have been based on the peptide-binding region (PBR) sites and previous knowledge on polymorphic sites. If comparison between variants resulted in a score lower than -12, then the second variant was added to the top cluster variant. When the top variant was compared with all variants, we moved to the second variant, and re-performed the procedure. After the top-down procedure was done for each sample independently, we checked that each variant was not found in several cluster variants. Then, we generated a file gathering all individuals with their cluster variants (which are

potential true alleles). Finally, artifactual clusters resulting from chimeras or substitution errors were discarded based on cluster features (i.e. low number of sequences in this cluster variant and the number of individuals that possess this cluster variant), and on a meticulous observation of the nucleotide sequence. Retained clusters were used to define MHC-IIB genotypes.

Reliability of the MHC-IIB DAB1 and DAB2 genotyping was evaluated with segregation patterns within families. Concretely, after the genotyping we checked that sequences attributed as true allele in offspring were also found and attributed as true allele in one of the parents. Over our 140 families, we found almost 100% matches. In addition, around 100 individuals were also genotyped using cloning/Sanger method, and showed congruent results with the 454 sequencing.

**Figure A:** Flow chart of the MHC-IIB genotyping procedure

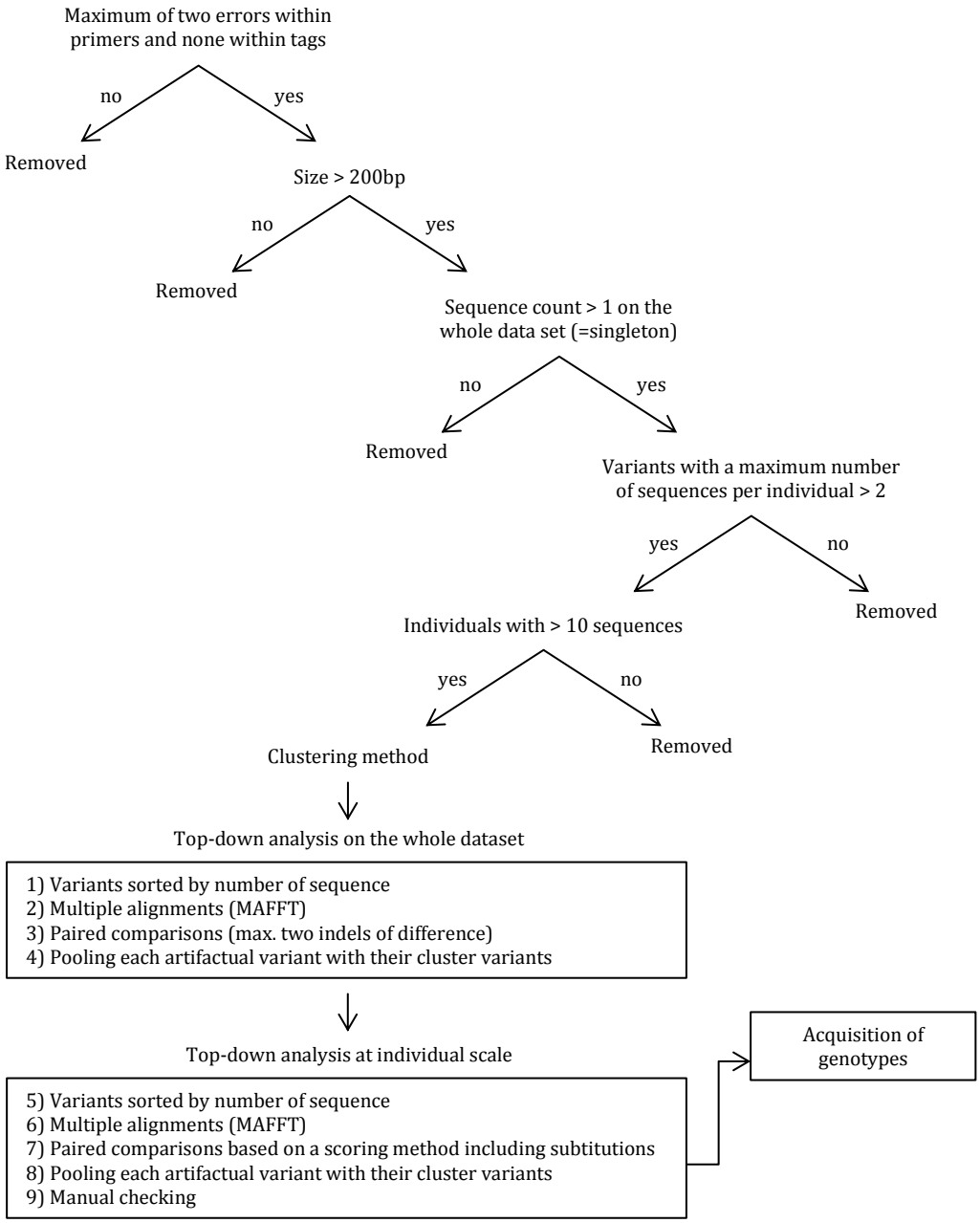
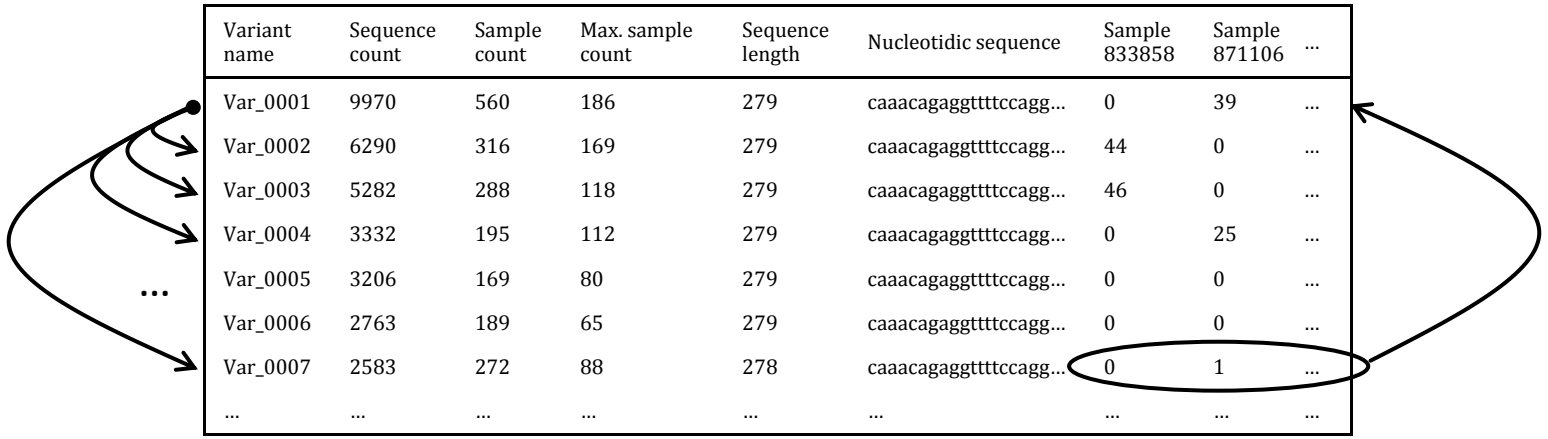


**Figure B:** Illustration of the top-down analysis at the whole dataset (MHC-IIB DAB1 example). Data were sorted according to the sequence count. Variant 0001 represent the top variant (i.e. with the maximum sequence count). The only difference between 0001 and 0007 variants is one indel, consequently sequence count at each sample of the variant 0007 are pooled with data of the variant 0001.



| Variant name | Sequence count | Sample count | Max. sample count | Sequence length | Nucleotidic sequence | Sample 833858 | Sample 871106 | … |
|---|---|---|---|---|---|---|---|---|
| Var_0001 | 9970 | 560 | 186 | 279 | caaacagaggttttccagg… | 0 | 39 | … |
| Var_0002 | 6290 | 316 | 169 | 279 | caaacagaggttttccagg… | 44 | 0 | … |
| Var_0003 | 5282 | 288 | 118 | 279 | caaacagaggttttccagg… | 46 | 0 | … |
| Var_0004 | 3332 | 195 | 112 | 279 | caaacagaggttttccagg… | 0 | 25 | … |
| Var_0005 | 3206 | 169 | 80 | 279 | caaacagaggttttccagg… | 0 | 0 | … |
| Var_0006 | 2763 | 189 | 65 | 279 | caaacagaggttttccagg… | 0 | 0 | … |
| Var_0007 | 2583 | 272 | 88 | 278 | caaacagaggttttccagg… | 0 | 1 | … |
| … | … | … | … | … | … | … | … | … |

## Supplementary Results

*MHC-I and MHC-IIB characterization*

Positive selection and recombination (*sensu lato*) were shown to play an important role in shaping MHC diversity at both classes. Accordingly, branch-site tests of positive selection revealed that M2a and M8 were best-fit models for both MHC classes. For both models, nine positively selected sites (PSS) were found at MHC-I (Figure 1). For MHC-IIB combined, a total of 9 sites was identified as PSS. More than double the number of PSS were identified in MHC-IIB DAB1 compared to DAB2 in both models. In total 15 and 6 PSS were detected with M8 for DAB1 and DAB2 respectively, whereas based on M2a, 12 and 4 sites were conserved as PSS (Figure 1). In more than 70 percent of cases, sites detected under positive selection were located within the PBR (Figure 1). Finally, based on a set of methods we detected evidence for recombination (*sensu lato*) for all MHC loci (Table S2). Although, statistical analysis performed with RDP4 or Geneconv failed to detected recombination events in MHC-I, some of the applied tests are known to have limited power when gene conversion is too frequent (Mansai and Innan, 2010).

**Supporting Data**

**Table S1:** Genetic diversity at barn owl MHC-I and MHC-IIB genes.

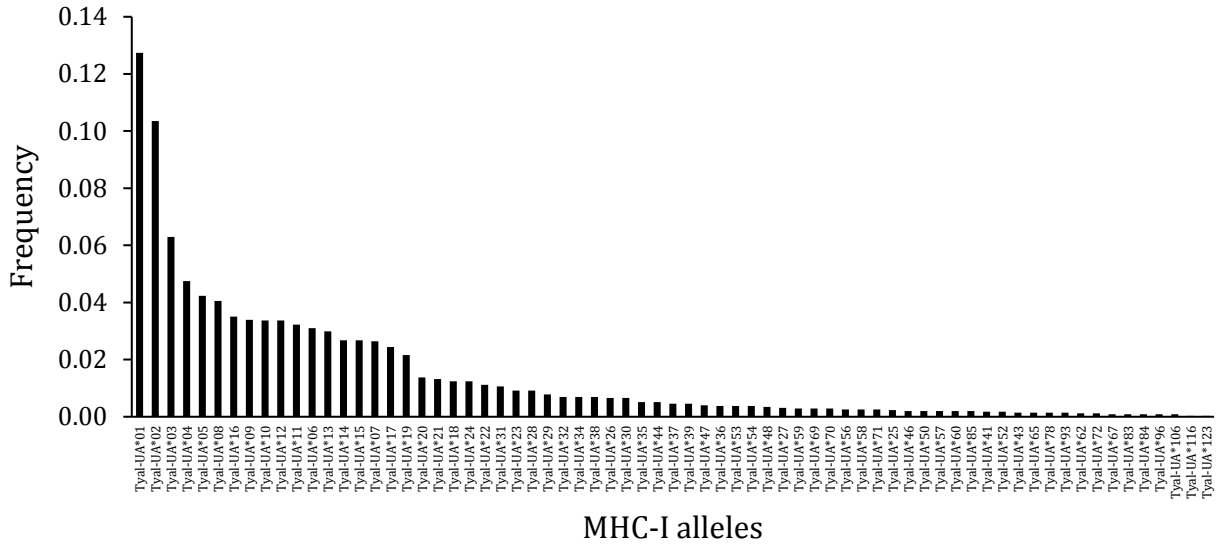
| | Number of alleles | Number of sites | Number of codons | S | k | π (S.D.) | AA distance (S.E.) |
|---|---|---|---|---|---|---|---|
| MHC-I | | | | | | | |
| All | 69 | 276 | 91 | 39 | 10.14 | 0.036 (0.001) | 0.075 (0.019) |
| PBR | 49 | 39 | 13 | 17 | 6.32 | 0.162 (0.006) | 0.336 (0.079) |
| Non-PBR | 43 | 237 | 78 | 22 | 3.82 | 0.016 (0.001) | 0.031 (0.011) |
| MHC-IIB combined | | | | | | | |
| All | 42 | 270 | 90 | 86 | 32.04 | 0.119 (0.004) | 0.208 (0.026) |
| PBR | 41 | 72 | 24 | 41 | 16.54 | 0.230 (0.005) | 0.378 (0.055) |
| Non-PBR | 33 | 198 | 66 | 45 | 15.49 | 0.078 (0.004) | 0.146 (0.026) |
| MHC-IIB DAB1 | | | | | | | |
| All | 25 | 270 | 90 | 57 | 19.08 | 0.071 (0.004) | 0.138 (0.023) |
| PBR | 25 | 72 | 24 | 32 | 11.74 | 0.163 (0.007) | 0.309 (0.051) |
| Non-PBR | 19 | 198 | 66 | 25 | 7.34 | 0.033 (0.004) | 0.075 (0.019) |
| MHC-IIB DAB2 | | | | | | | |
| All | 17 | 270 | 90 | 41 | 14.07 | 0.052 (0.005) | 0.099 (0.018) |
| PBR | 16 | 72 | 24 | 19 | 7.56 | 0.105 (0.008) | 0.175 (0.048) |
| Non-PBR | 14 | 198 | 66 | 22 | 6.51 | 0.033 (0.004) | 0.072 (0.018) |

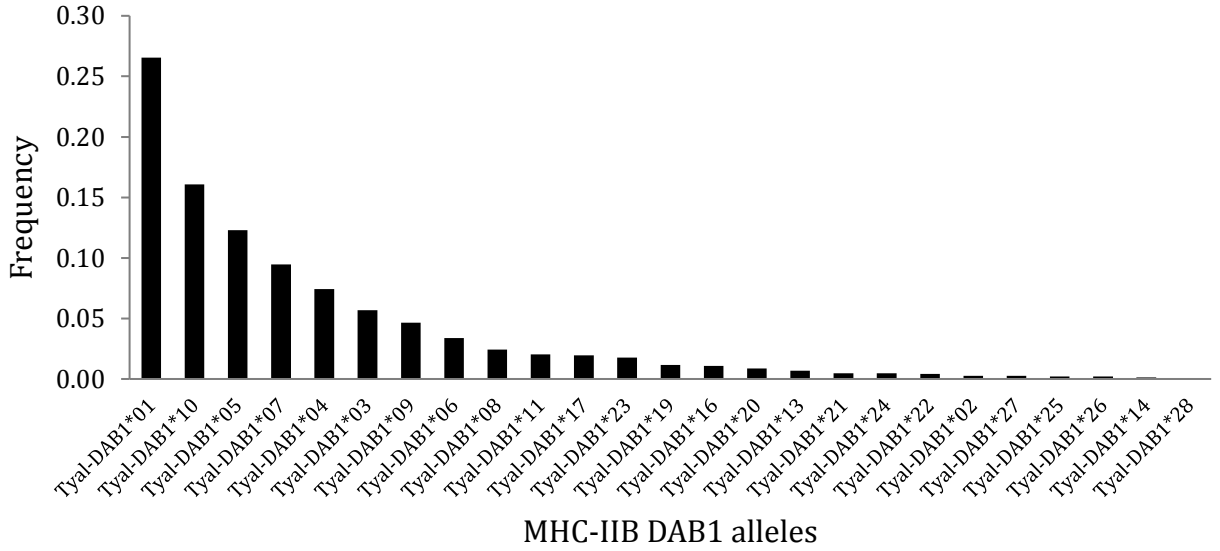
S, number of polymorphic sites; ps, proportion of segregating sites; k, average number of nucleotide differences; π, average number of pairwise differences per base pair; AA distance, amino acid pairwise distance.

**Table S2:** Results of recombination analysis. Values indicate the number of detected recombination event estimated by several methods.

| | Rm | Φw test | Geneconv | MaxChi | Chimerae | RDP |
|---|---|---|---|---|---|---|
| MHC-I | 11 | $p < 10^{-3}$ | 0 | 0 | 0 | 0 |
| MHC-IIB combined | 14 | $p < 10^{-3}$ | 24 | 3 | 2 | 1 |
| MHC-IIB DAB1 | 14 | $p < 10^{-3}$ | 6 | 2 | 2 | 1 |
| MHC-IIB DAB2 | 9 | $p = 0.437$ | 1 | 1 | 0 | 0 |

**Figure S1:** Allele frequencies at MHC-I, MHC-IIB DAB1 and MHC-IIB DAB2 genes.



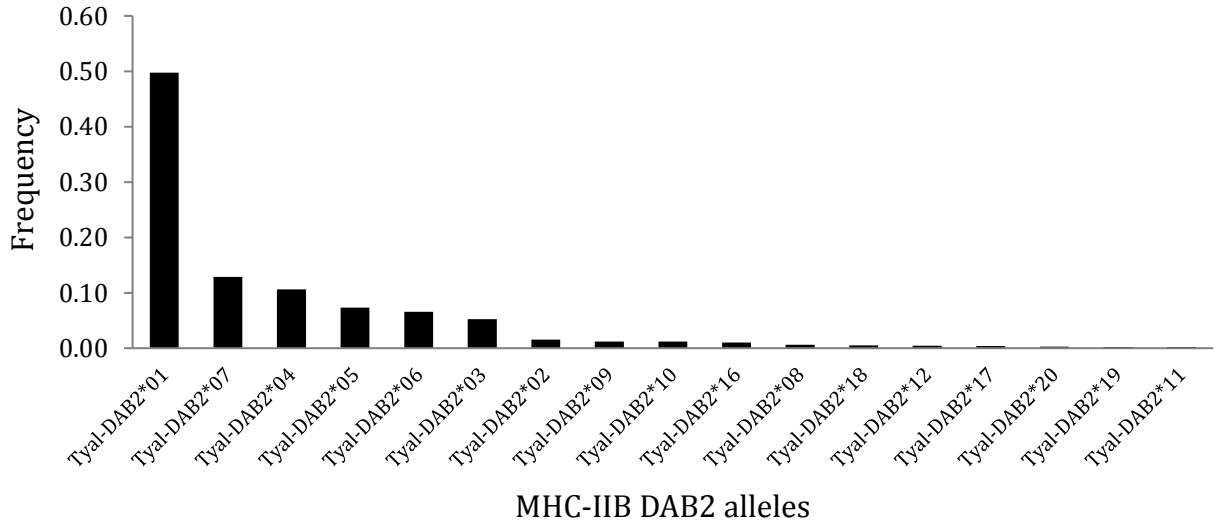
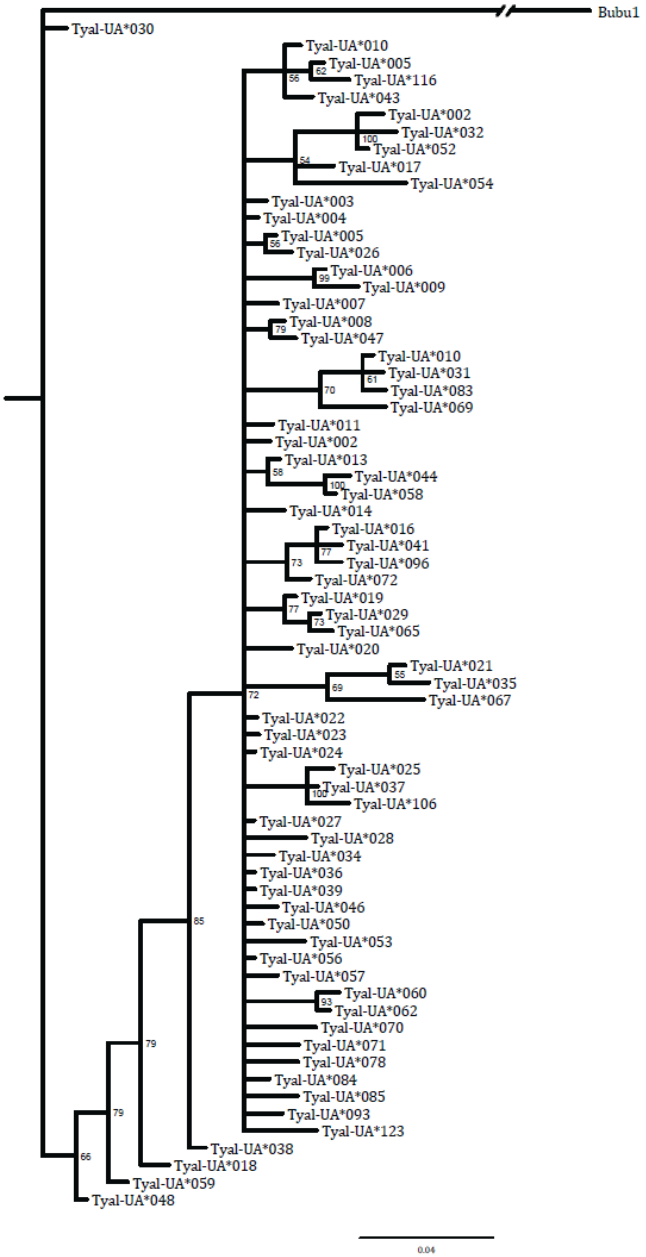



**Figure S2:** Bayesian phylogenetic trees of (a) MHC-I exon 3 and (b) MHC-IIB exon 2. Trees have been rooted with *Bubo bubo* sequences (accession numbers EU120697, EF641238, EF641236). Node values represent Bayesian posterior probability. Green: MHC-IIB DAB1, Blue: MHC-IIB DAB2.

(a)



(b)

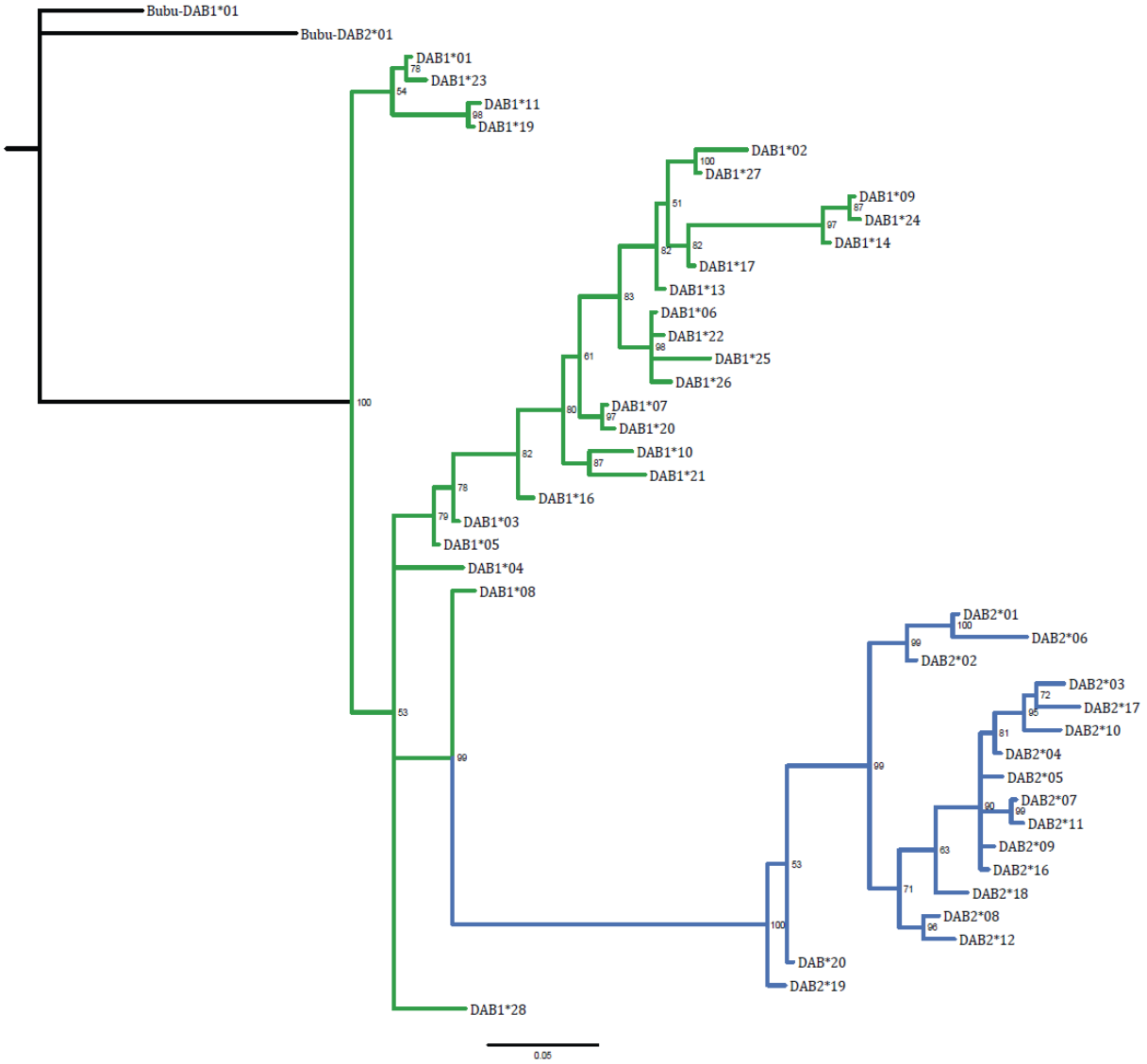
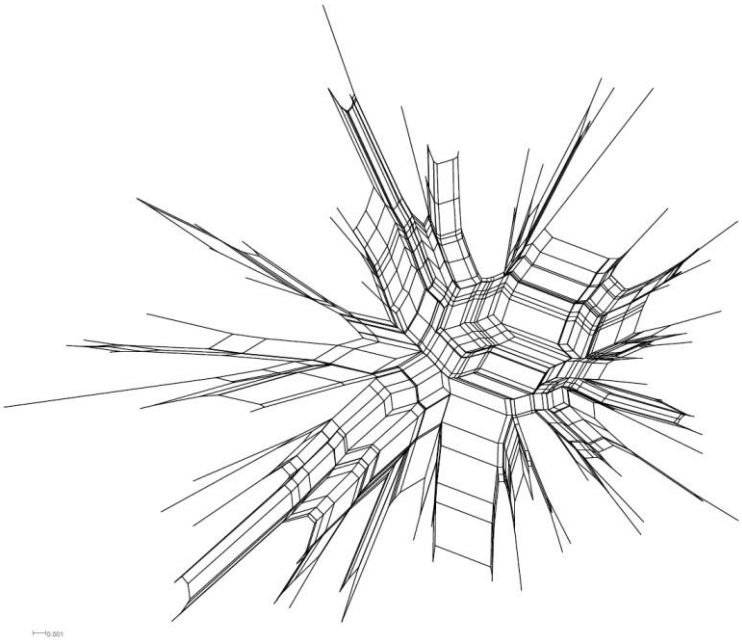

Bubu-DAB1*01
Bubu-DAB2*01
DAB1*01
DAB1*23
DAB1*11
DAB1*19
DAB1*02
DAB1*27
DAB1*09
DAB1*24
DAB1*14
DAB1*17
DAB1*13
DAB1*06
DAB1*22
DAB1*25
DAB1*26
DAB1*07
DAB1*20
DAB1*10
DAB1*21
DAB1*16
DAB1*03
DAB1*05
DAB1*04
DAB1*08
DAB1*28
DAB2*01
DAB2*06
DAB2*02
DAB2*03
DAB2*17
DAB2*10
DAB2*04
DAB2*05
DAB2*07
DAB2*11
DAB2*09
DAB2*16
DAB2*18
DAB2*08
DAB2*12
DAB*20
DAB2*19
0.05

**Figure S3:** Neighbor-net networks of (a) MHC-I exon 3 and (b) MHC-IIB exon 2 alleles. Networks were built with uncorrected p-distances using Splitstree 4 (Huson and Bryant, 2006).

(a)



(b)

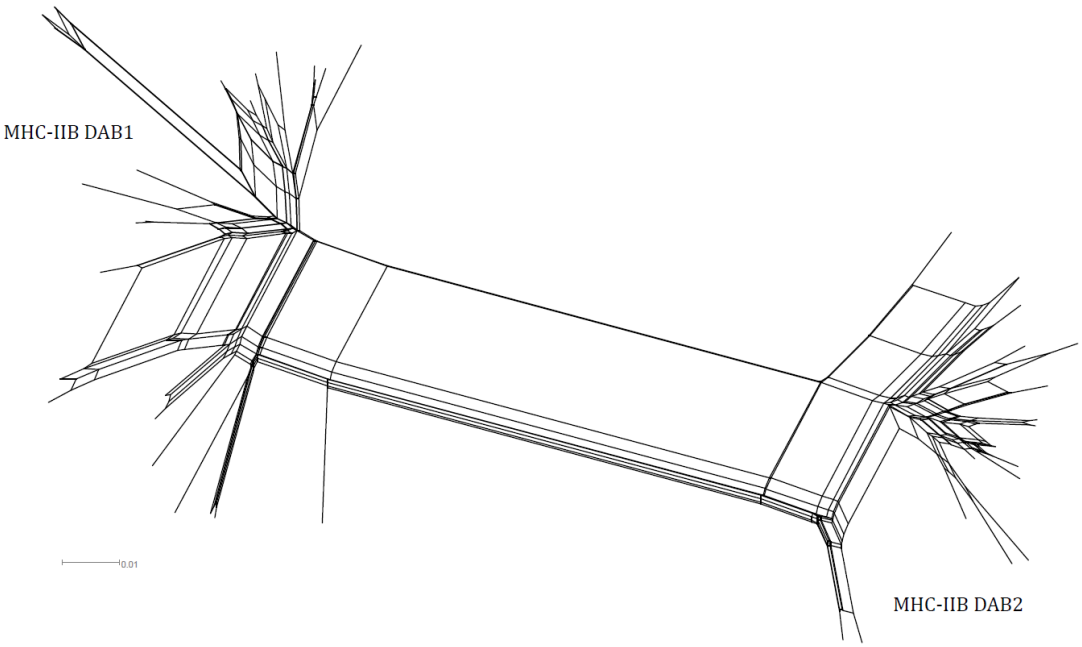
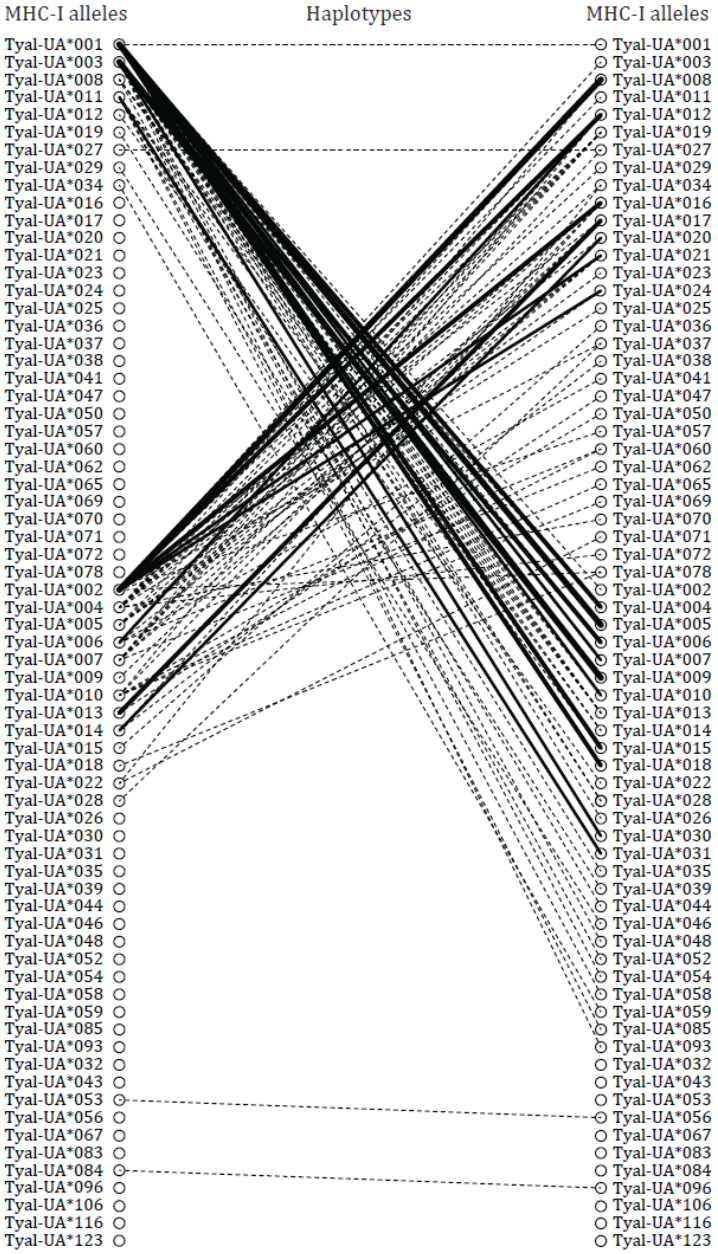
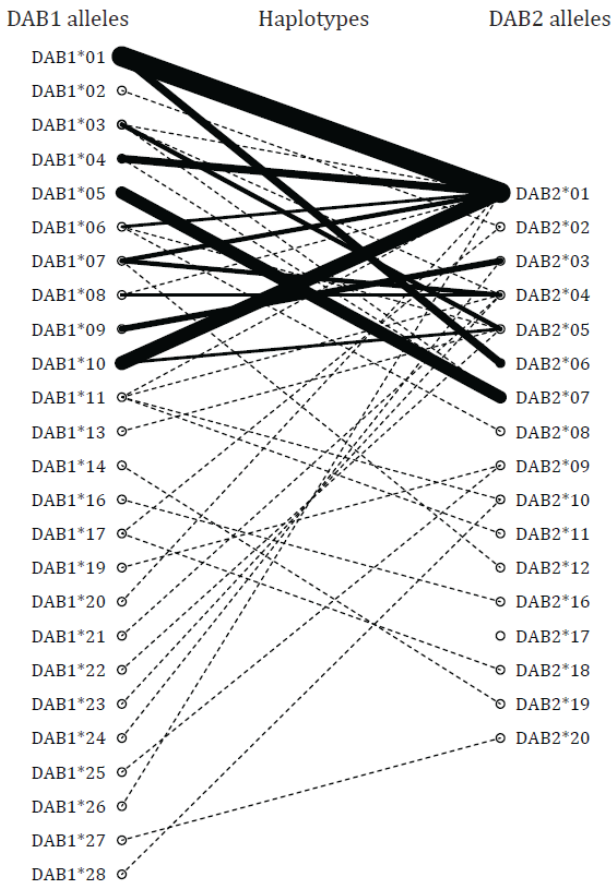


**Figure S4:** MHC-I (a) and MHC-IIB (b) haplotype combinations in Swiss barn owls. Right and left columns represent alleles. Each line is an allelic combination between two loci. The lines' weight is proportional to the occurrence of the haplotype in the population. Dashed lines are rare haplotypes. (a) MHC-I haplotype combinations. Due to the co-amplification of both MHC-I loci, and the unknown origin of alleles, each column is composed of the same alleles. However, this figure show that the same panel of alleles combine to another different set of alleles, resulting to possible assignation of alleles to loci. (b) MHC-IIB haplotype combinations. Right and left columns represent MHC-IIB DAB1 and DAB2 alleles respectively.

(a)



(b)


DAB1 alleles
Haplotypes
DAB2 alleles
DAB1*01
DAB1*02
DAB1*03
DAB1*04
DAB1*05
DAB1*06
DAB1*07
DAB1*08
DAB1*09
DAB1*10
DAB1*11
DAB1*13
DAB1*14
DAB1*16
DAB1*17
DAB1*19
DAB1*20
DAB1*21
DAB1*22
DAB1*23
DAB1*24
DAB1*25
DAB1*26
DAB1*27
DAB1*28
DAB2*01
DAB2*02
DAB2*03
DAB2*04
DAB2*05
DAB2*06
DAB2*07
DAB2*08
DAB2*09
DAB2*10
DAB2*11
DAB2*12
DAB2*16
DAB2*17
DAB2*18
DAB2*19
DAB2*20

## References


Burri R, Niculita-Hirzel H, Roulin A, Fumagalli L (2008). Isolation and characterization of major histocompatibility complex (MHC) class IIB genes in the Barn owl (Aves: *Tyto alba*). *Immunogenetics* **60**: 543-550.

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

Huson DH, Bryant D (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254-267.

Katoh K, Standley DM (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014). Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol Ecol Res* **14**: 753-767.

Mansai SP, Innan H (2010). The power of the methods for detecting interlocus gene conversion. *Genetics* **184**: 517-527.

Sebastian A, Herdegen M, Migalska M, Radwan J (2016). Amplisas: a web server for multilocus genotyping using next-generation amplicon sequencing data. *Mol Ecol Res* **16**: 498-510.

Sommer S, Courtiol A, Mazzoni C (2013). MHC genotyping of non-model organisms using next-generation sequencing: a new methodology to deal with artefacts and allelic dropout. *BMC Genomics* **14**: 542.

Stutz WE, Bolnick DI (2014). Stepwise threshold clustering: a new method for genotyping MHC loci using next-generation sequencing technology. *PLoS One* **9**: e100587.