nature ecology & evolution

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41559-018-0509-3

In the format provided by the authors and unedited.

The evolution of immunity in relation to colonization and migration

Emily A. O'Connor*, Charlie K. Cornwallis, Dennis Hasselquist, Jan-Åke Nilsson and Helena Westerdahl

Molecular Ecology and Evolution Laboratory, Lund University, Lund, Sweden. *e-mail: emily.o_connor@biol.lu.se

SUPPLEMENTARY METHODS

Extracting details of the distribution range of each species using BirdLife International data. The shapefiles provided by Birdlife International[1] were visualised using the software ArcMap V. 10.2.2. The polygon data for the distribution ranges of each species were annotated with the following information: ‘Presence’ (Extant, Possibly Extant, Probably Extant, Possibly Extinct, Extinct (post 1500) or Presence Uncertain), ‘Origin’ (Native, Introduced, Reintroduced, Vagrant or Origin Uncertain) and ‘Seasonality’ (‘Resident’, ‘Breeding Season’, ‘Non-breeding Season’, ‘Passage’ or ‘Seasonal occurrence uncertain’). The seasonality information enabled us to select the polygon data for the breeding and wintering ranges of the migratory species separately. The distribution ranges of interest for this study were never categorised as ‘Passage’ or ‘Seasonal occurrence uncertain’. When there was more than one distribution range within a species for the ‘seasonality’ we were interested in (for example if the species had been introduced elsewhere), we selected the distribution range that matched the site individuals were sampled from for this study. Within the ‘Attributes Table’ for the polygon describing the distribution range of interest we used the ‘calculate geometry’ option to calculate the centroid latitude of the distribution range. We used the latitude of the most northerly and southerly vertices of the polygon to calculate the number of degrees latitude spanned by the distribution range of each species. We cross-checked the distribution ranges for each species included in the study against those in the Handbook of Birds of the World Alive[2] for consistency. In all 32 species, there was a good match.

Sample sizes. We have shown in our previous work that sampling a comparably small number of individuals from a species results in similar MHC-I diversity estimates as those obtained from many more individuals[3]. This is further corroborated by the observation that Karlsson & Westerdahl[4] detected a mean of 14.4 MHC-I alleles per individual in a study of 45 *Passer domesticus* individuals and we obtained a highly similar estimates (15.3 MHC-I alleles per individual) using just two individuals. Furthermore, when we compared the number of MHC-I alleles per individual detected in the current study with those from a previous study using similar approaches (on different populations and individuals) we found highly comparable results (see Supplementary Fig. 7, r = 0.70). Therefore, the sample sizes in our study appear to give robust estimates of MHC-I diversity. However, it should be noted that the most important consideration for a comparative study is to have estimate that can be reliably compared across species, not to obtain completely accurate estimates of a given trait within a species. Given that the between-species variation in the number of MHC-I alleles detected per individual (standard deviation = 9.54 alleles/individual) was much greater than the within-species variation (standard deviation = 2.70 alleles/individual), we are confident that any noise around the within species MHC diversity estimates is minor compared to the magnitude of differences between species.

In some of the species, it was only possible to sample one or two individuals. These under-sampled species are split in a proportionately equal fashion across Palearctic residents (3/10 species), African residents (5/15 species) and migratory species (2/7 species), and therefore it is highly unlikely that this has any qualitative effect on our overall findings. To confirm this, we re-ran our main analyses excluding the species with fewer than two individuals, and get highly similar results (see Supplementary Tables 3 to 8).

Details of primers, PCR conditions and 454 sequencing. Three separate primer pairs were used to amplify overlapping fragments of MHC-I exon 3 in all species (Supplementary Fig.

4). The first pair of primers (PP1) consisted of the forward primer ‘HNalla’ 5’-TCCCCACAGGTCTCCACAC-3’ and the reverse primer ‘RVS3’ 5’-GGCAGACGTGCTYCWRGTAATT-3’. PP1 amplifies a fragment of exon 3 roughly 190 bp long, depending on the presence of codon deletions, which encompasses amino acids from position 5 to 67 (see Supplementary Fig. 5). The second pair of primers (PP2) consisted of the forward primer ‘FWD3’ 5’-TGGTTGCGAGTTTACGGYTRTG-3’ and the reverse primer ‘RVS’ 5’- TGCGCTCCAGCTCCYTCTGCC -3’. PP2 amplifies a fragment around 219 bp long covering amino acids from position 13 to 84 (Supplementary Fig. 5). The third primer pair (PP3) was the forward primer ‘FWD5’ 5’- GAYGGGYRGGATTTCATCTCC -3’ with the reverse primer ‘RVS4’ 5’- TATYTCYGGAGCCATTCYGGGCA -3’. PP3 amplifies a shorter fragment of roughly 117 bp (amino acids 35 to 73, Supplementary Fig. 5). There was overlap in the fragments amplified by each of the three primer pairs (PP1 & PP2 167 bp, PP1 & PP3 99 bp and PP2 & PP3 117 bp). All primers were GS FLX Titanium Fusion Primers with one of fifteen possible individual 6 bp tags added to the 5’ end of the original primers. Unique combinations of forward and reverse tags in the primer pairs for each sample enabled sequences to be re-assigned to samples after sequencing.

Standard PCRs were performed on DNA from all individuals using each primer pair. For every species PCRs were performed twice on each individual to provide technical replicates ($n_{\text{amplicons}}$ = 162 per PP). Each 15 µl PCR reaction contained 7.5 µl of Multiplex PCR Master Mix (QIAGEN), 5.3 µl of double-distilled water, 0.6 µl each of the Forward and Reverse primer (5 µM) and 1 µl DNA (25 ng/µl). The PCR reactions were run in a thermal cycler GeneAmp PCR System 9700 starting at 95ºC for 15 min, followed by 25-35 cycles of 95ºC for 30 s, 60-65ºC for 90 s and 72 ºC for 60 s (the number of cycles and annealing temperature depended on the PP, see S3 for details). Finally, a 10 min extension phase at 60ºC was applied. PCR products were run on agarose gels (1.5%) and the strength of bands was examined to facilitate pooling samples into approximately equimolar quantities. These pooled samples were then purified using the MinElute PCR Purification kit (QIAGEN). Following purification, the concentration of each pool was measured using a Nanodrop spectrophotometer and the pools were combined in equimolar quantities. Bi-directional pyrosequencing on the 454 GS FLX system by 454/Roche was performed Lund University Sequencing Facility (Faculty of Science). Sequencing of samples was spread across three separate 454 runs using two regions of a 4-region 454 pico titre plate for the first run and three regions in each of the second and third runs (see Supplementary Table 18 for details of the number of samples per region in each 454 run and associated read depths per samples).

Primer performance. As the three primer pairs used in this study (PP1, PP2 and PP3) were designed based upon sequence similarity across species representing a phylogenetic spread across the parvorder Passerida[5], we anticipated that the primer pairs (PPs) should perform universally across Passerida. All three PPs amplified fragments of MHC-I exon 3 in every species reported in this study (Supplementary Fig. 6). Full details of the number of different alleles detected in each individual for each PP can be found in Supplementary Table 10. On average, the highest proportion of alleles per individual were detected by PP1 (0.60 ± 0.02). PP3 detected a slightly lower proportion of the alleles per individual (0.53 ± 0.03), followed by PP2 (0.33 ± 0.02).

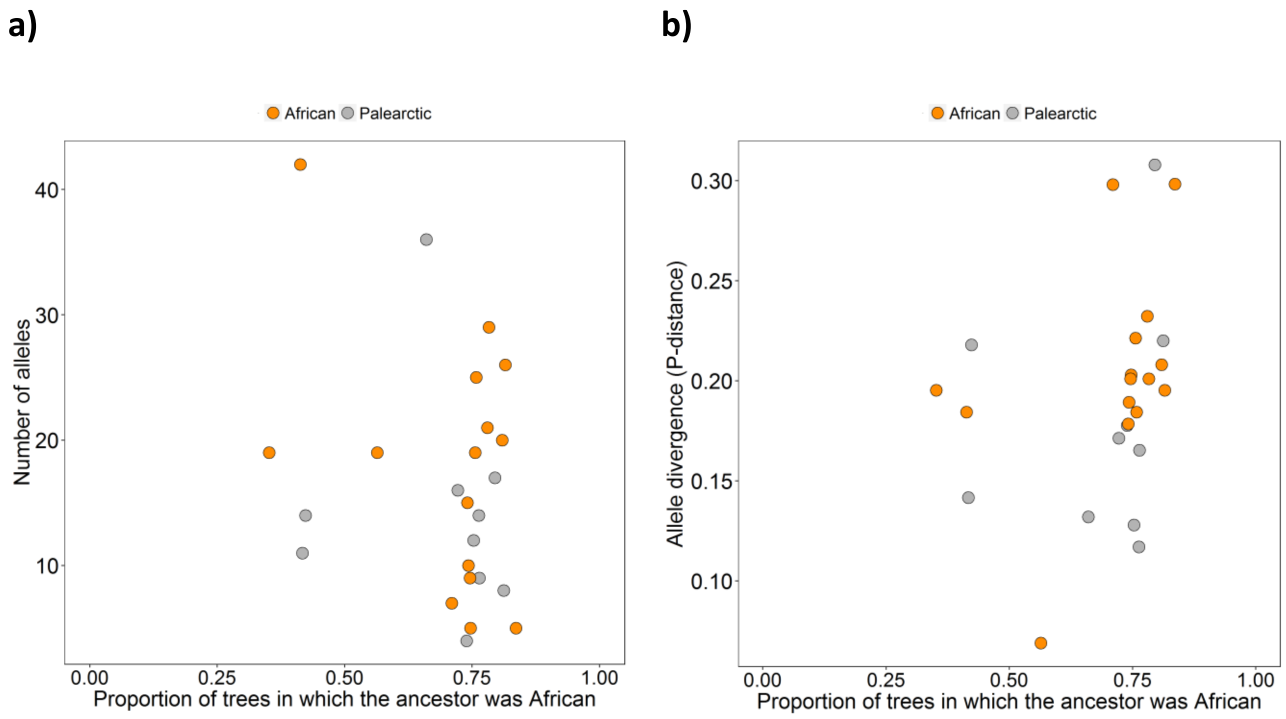
The proportion of the variance in the number of alleles detected on the level of species, i.e. variance in allelic number explained by species, was high for all three PPs (PP1 Posterior Mode (PM) = 0.73, Credible Intervals (CI) = 0.56 to 0.85; PP2 PM = 0.68, CI = 0.49 to 0.82; PP3 PM = 0.77, CI = 0.59 to 0.86). This indicates that the number of alleles detected in individuals of the same species was highly repeatable within all three primer pairs. Furthermore, this high degree of repeatability was not significantly different between any of the three PPs (PP1 vs PP2 PM = 0.048, CI = -0.176, to 0.265, pMCMC = 0.31; PP1 vs PP3 PM = -0.047, CI = -0.224 to 0.167, pMCMC = 0.43; PP2 vs PP3 PM = -0.044, CI = -0.288 to 0.127, pMCMC = 0.23). It is possible that each PP may have performed better in some species than others. However, the use of three different degenerate PPs in all species was intended to minimise the impact of any such variation on the final results.

De-multiplexing, clustering and filtering 454 data using AmpliSAS. Raw 454 reads were first assigned to amplicons based upon their primer and 6 bp tag sequences (i.e. de-multiplexed). The abundance of each unique sequence (variant) was also calculated. Next, the variants within each amplicon were assigned to clusters based upon sequence similarity and the relative abundance of variants within the amplicon. This clustering step is performed in a stepwise fashion starting with the most abundant sequence as the 'dominant sequence' or core of the first cluster. Every other variant in the amplicon is then compared to the dominant sequence to assess whether it should be clustered with the dominant sequence. Variants are clustered with a dominant sequence if they are likely to have arisen from a PCR or sequencing errors of the dominant sequence e.g. single base substitutions or homoploymer and non-homopolymer indels. Determining whether a variant is a likely error of the dominant sequence is based upon a number of user-defined thresholds. We implemented the recommended thresholds for 454 sequencing[6]. Therefore in order to join a cluster a variant had to have a substitution error rate below 0.5% (error rate = number of nucleotide substitutions x length of variant), a non-homopolymer indel rate below 1% (indel rate = number of indels x length of variant) and occur at a frequency of less than 25% of the dominant sequence. Any variants with indels in homopolymer regions of three or more consecutive identical nucleotides are clustered by default. Once every variant had been checked against the dominant sequence of this first cluster, the process begins again with the next most abundant variant as the dominant sequence in the next cluster. Only sequences that are in frame, given the expected length, could be the dominant sequence within a cluster. This clustering process continues until all possible clusters have been created based upon the aforementioned thresholds. Any single read sequences that are not assigned to clusters are deleted. The AmpliSAS software determines a consensus sequence for each cluster based upon the most frequent nucleotide in each position (this is usually identical to the dominant sequence).

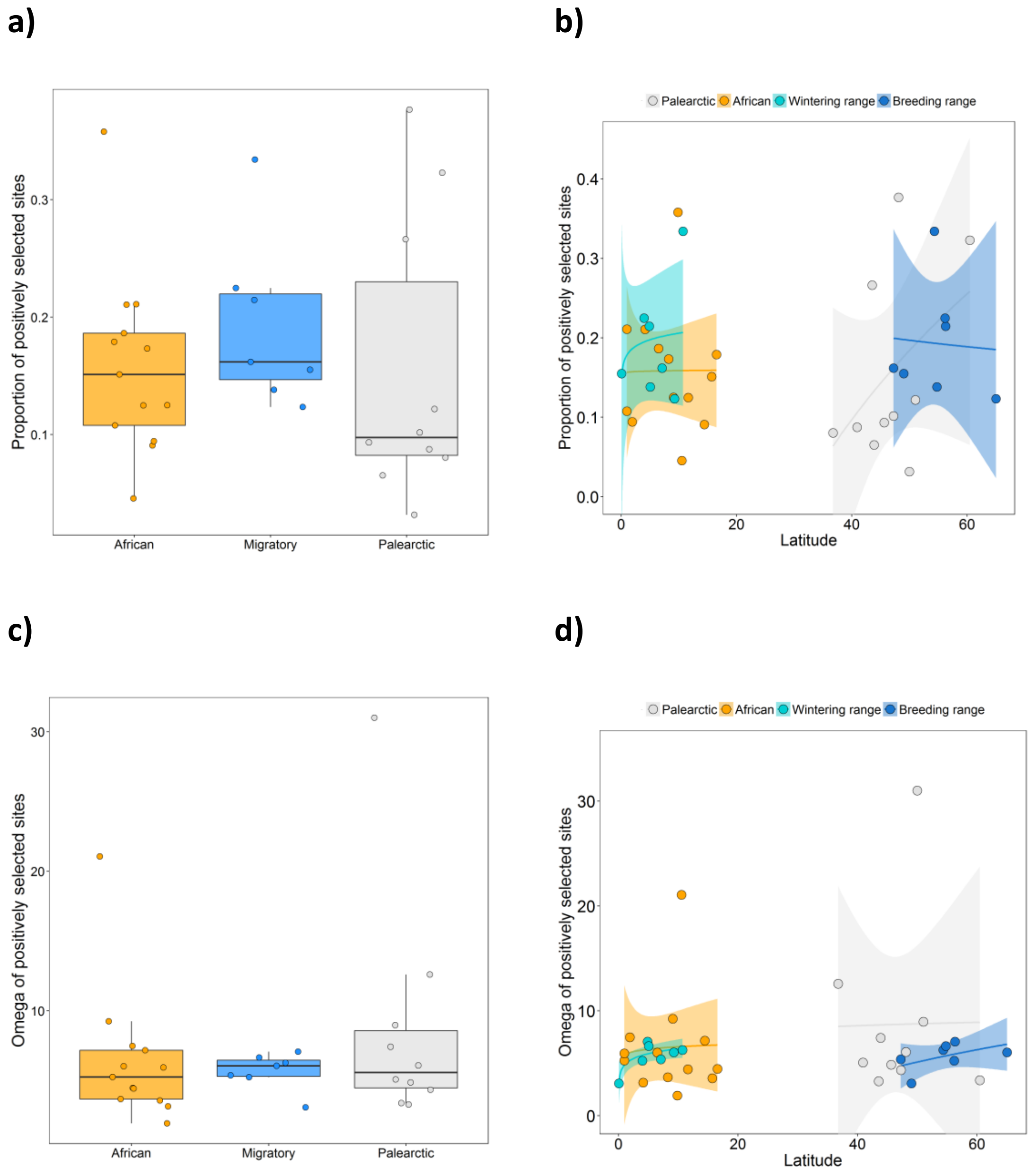
After clustering, each amplicon is then filtered using user-defined thresholds intended to discard any artefactual small clusters, low-depth non-clustered variants and PCR chimeras remaining in the data. Again user-defined parameters are used. The filtering parameters we defined through the AmpliSAS were that any amplicons with fewer than 10 reads should be discarded as should any variant or cluster with fewer than two reads as well as any amplicon with more than 80 clusters or variants. Though the later stipulation was never the case. We also selected the option to discard chimeras. Chimeras were defined as any variants or clusters that began identical to one variant or cluster and ended identical to another with no unique portion in the sequence (with the proviso that the portions identical to the other clusters/variants had to be longer than 10 nucleotides).

Each unique variant or consensus sequence from a cluster remaining in the dataset after clustering and filtering was considered a putative allele. We performed one further cleaning step by removing any putative alleles that did not occur in both replicates of each individual for each primer pair. This step may have discarded some true alleles, but as MHC-I alleles were amplified using three different primer pairs it is likely that any true alleles lost during this step for one primer pair were likely to be detected by one of the other two primer pairs. The alleles remaining after this cleaning step were considered verified alleles.

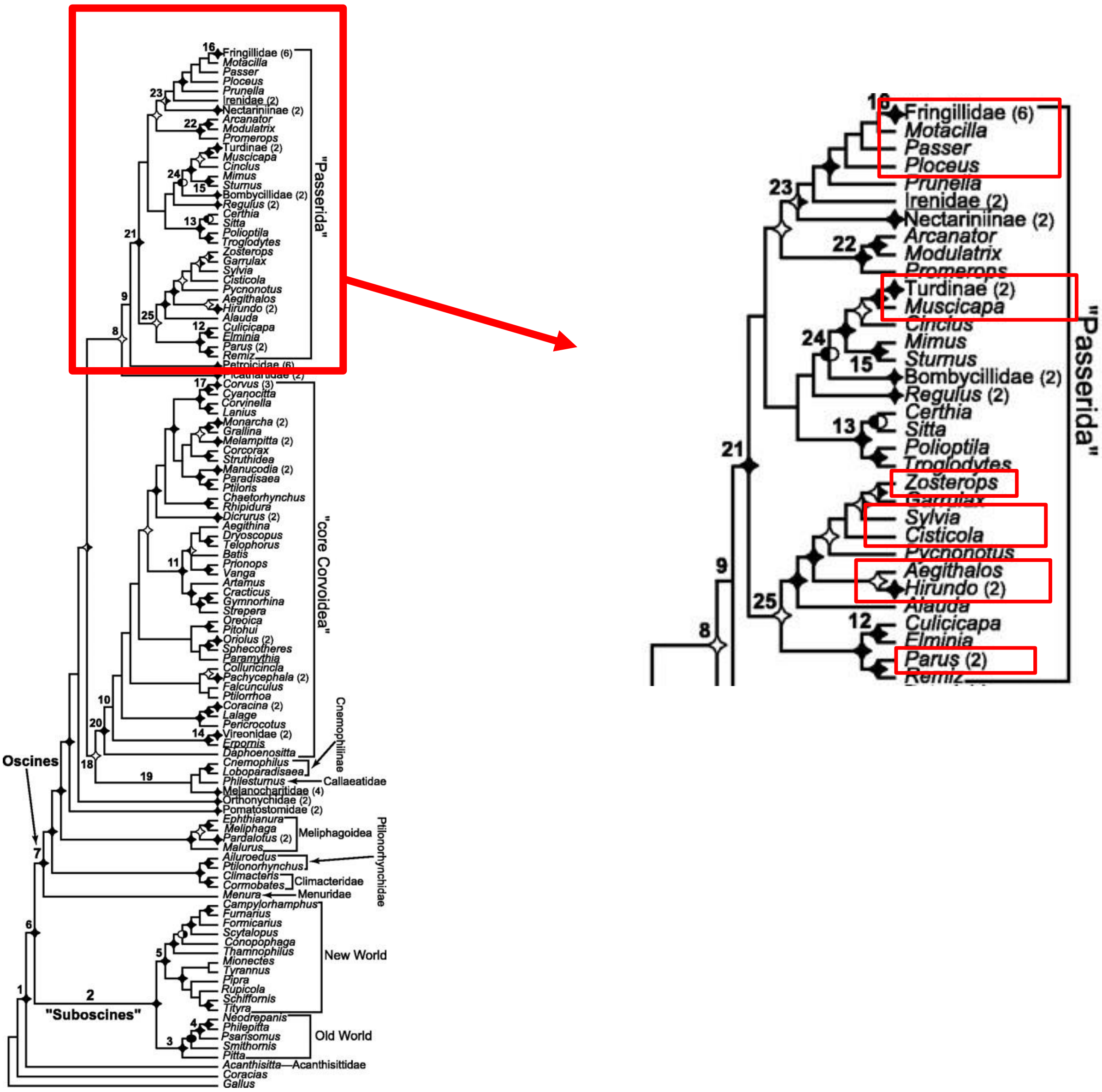
SUPPLEMENTARY FIGURES



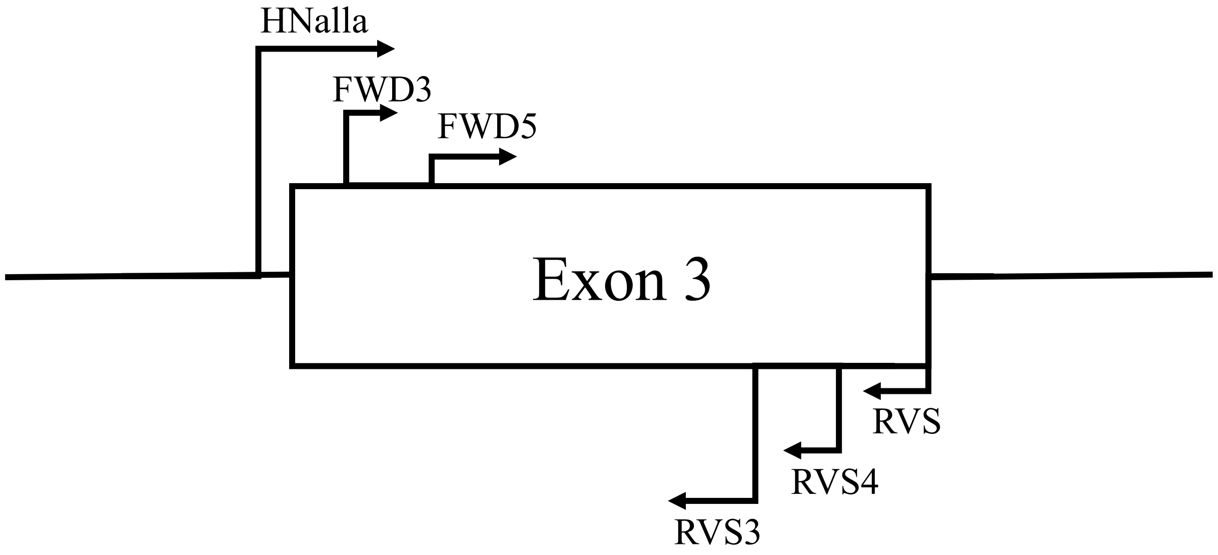
Supplementary Figure 1: MHC diversity and the predicted ancestry of African and Palearctic species. The relationship between the proportion of trees in which the ancestor of each species was assigned as African resident, using stochastic character mapping (SCM), and the number of alleles (a) and allele divergence (b) for African (n = 15) and Palearctic (n = 10) species.



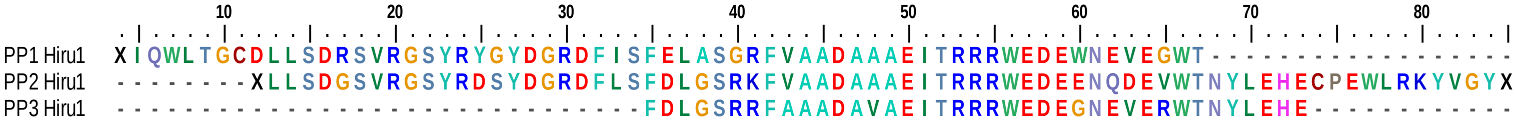
Supplementary Figure 2: The number of positively selected sites (PSS) and strength of positive selection (Omega) in relation to latitude and geographical distribution of residents and migrant species. a) Proportion of PSS in African residents (orange, n = 15 species), Palearctic residents (grey, n = 10 species) and migratory species (blue, n = 7 species) shown in box-plots (central line depicts the median, the lower and upper hinges correspond to the first and third quartiles and whiskers extend to the highest value within 1.5 * the inter-quartile range) overlaid with the individual data points for each species. b) Relationship between the centroid latitude of all species ranges and the proportion of PSS. Plotted line from linear regression of log proportion of PSS with the shaded area representing 95% confidence intervals. c) Omega value of the PSS in African, Palearctic and migratory species shown in box-plots (central line depicts the median, the lower and upper hinges correspond to the first and third quartiles and whiskers extend to the highest value within 1.5 * the inter-quartile range) overlaid with the individual data points for each species. d) The relationship between the centroid latitude of all species ranges and the Omega value of the PSS. Plotted line from linear regression of log Omega of PSS with the shaded area representing 95% confidence intervals.



Supplementary Figure 3: Phylogenetic distribution of the species genotyped for MHC-I. Distribution of the species in which we genotyped MHC-I in the current study across the Passerida phylogeny. The phylogeny is taken from the original Passerida phylogeny presented by Barker *et al.*[5]. Families that are represented in the current study are highlighted in red.

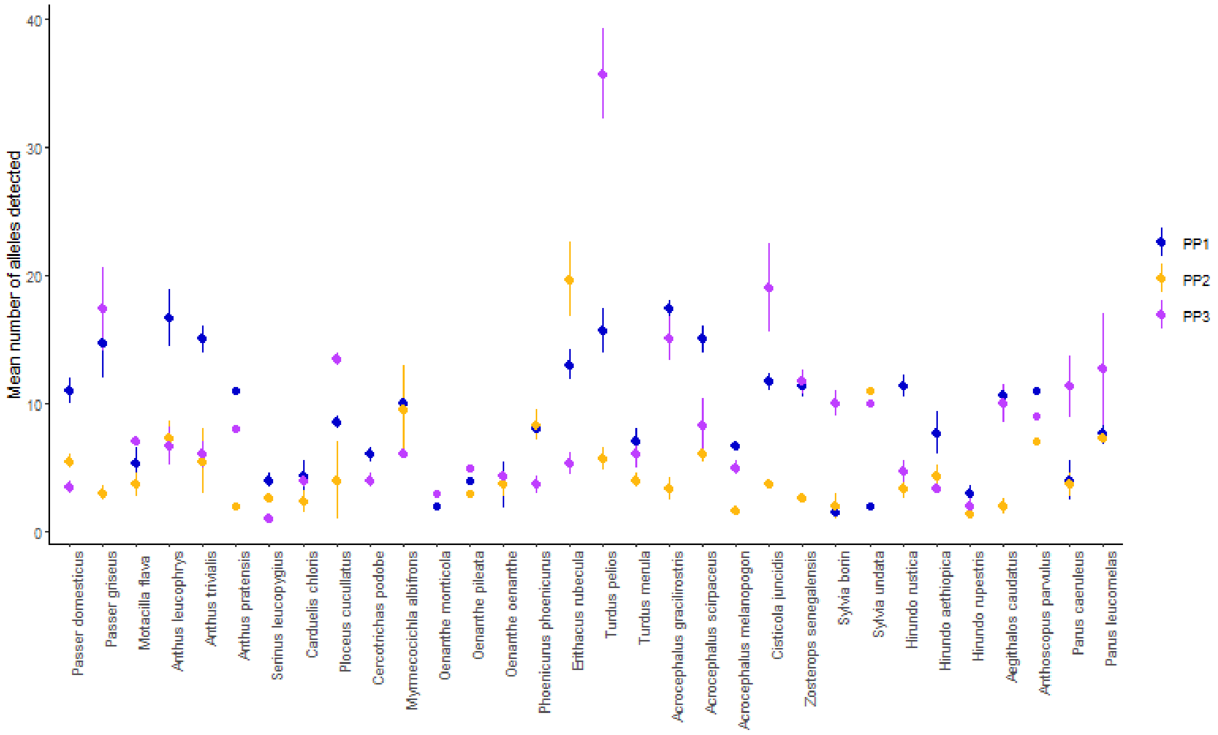


Supplementary Figure 4: Primer positions. Schematic representation of the position of the primers used in the present study. Forward and reverse primers shown above and below the exon respectively. The first pair of primers (PP1) consisted of the forward primer 'HNalla' and the reverse primer 'RVS3', the second pair of primers (PP2) was 'FWD3' with the reverse primer 'RVS' and the third primer pair (PP3) was the forward primer 'FWD5' with the reverse primer 'RVS4'. The primer 'HNalla' was positioned partially in the intron whereas the other primers were all within the exon.

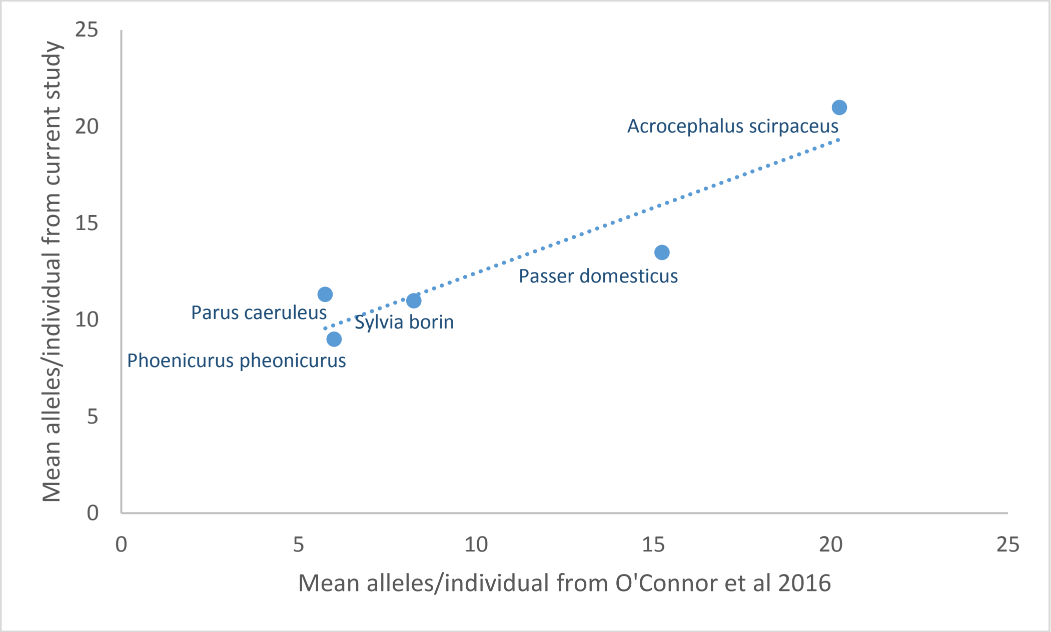


```
              10        20        30        40        50        60        70        80
          .|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
PP1 Hiru1 XIQWLTGCDLLSDRSVRGSYRYGYDGRDFISFELASGRFVAADAAAEITRRRWEDEWNEVEGWT------------------
PP2 Hiru1 --------XLLSDGSVRGSYRDSYDGRDFLSFDLGSRKFVAADAAAEITRRRWEDEENQDEVWTNYLEHECPEWLRKYVGYX
PP3 Hiru1 -------------------------------FDLGSRRFAAADAVAEITRRRWEDEGNEVERWTNYLEHE------------
```

Supplementary Figure 5: Overlap between primers. Alignments of example sequences from each of the three primer pairs used (PP1, PP2 and PP3) demonstrating the overlap between the fragments amplified by each primer pair. The numbering of the alignment reflects the position of amino acids within the full exon 3 sequence (92 amino acids). Examples provided are sequences from *Hirundo rustica*. The sites of the PBR, as inferred from HLA-A[7], are at positions 5, 7, 9, 23, 25, 52, 55, 56, 62, 65, 66, 69, 73, 77 and 81.



Supplementary Figure 6: Primer performance. The mean number of alleles (±SE) detected by each primer pair (PP1, PP2 and PP3) in each species (n = 32 species).



Supplementary Figure 7: Relationship between the mean number of MHC-I alleles detected in the current study and by O'Connor *et al.*[3] in the same species.

SUPPLEMENTARY REFERENCES

1. BirdLife International and NatureServe (2015) Bird species distribution maps of the world. Version 5.0. BirdLife International, Cambridge, UK and NatureServe, Arlington, USA.
2. del Hoyo, J., Elliott, A., Sargatal, J., Christie, D. A. & de Juana E. (Eds.) *Handbook of the Birds of the World Alive.* Lynx Edicions, Barcelona. Retrieved from http://www.hbw.com/ in October 2016.
3. O'Connor, E. A., Strandh, M., Hasselquist, D., Nilsson, J. Å. & Westerdahl, H. The evolution of highly variable immunity genes across a passerine bird radiation. *Mol. Ecol.* **25,** 977-989 (2016).
4. Karlsson, M. & Westerdahl, H. Characteristics of MHC class I genes in house sparrows Passer domesticus as revealed by long cDNA transcripts and amplicon sequencing. *J Mol. Evol.* **77,** 8-21 (2013).
5. Barker, F. K., Cibois, A., Schikler, P., Feinstein, J. & Cracraft, J. Phylogeny and diversification of the largest avian radiation. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 11040–11045 (2004).
6. Sebastian, A., Herdegen, M., Migalska, M. & Radwan, J. AmpliSAS: a web server for multilocus genotyping using next-generation amplicon sequencing data. *Mol. Ecol. Res.* **16,** 498-510 (2016).
7. Bjorkman, P. *et al.* The foreign antigen binding site and T cell recognition regions. *Nature* **329,** 512-518 (1987).