Inferring the evolution of the major histocompatibility complex of wild pigs and peccaries using hybridisation DNA capture-based sequencing

Carol Lee, Marco Moroldo, Alvaro Perdomo-Sabogal, Núria Mach, Sylvain Marthey, Jerôme Lecardonnel, Per Wahlberg, Amanda Y. Chong, Jordi Estellé, Simon Y.W. Ho, et al.

Inferring the evolution of the major histocompatibility complex of wild pigs and peccaries using hybridisation DNA capture-based sequencing

Carol Lee[1], Marco Moroldo[2], Alvaro Perdomo-Sabogal[1,3], Nuria Mach[2], Sylvain Marthey[2], Jérôme Lecardonnel[2], Per Wahlberg[2], Amanda Y. Chong[1,4], Jordi Estellé[2], Simon Y. W. Ho[5], Claire Rogel-Gaillard[2], and Jaime Gongora[1,*]

[1] *The University of Sydney, Faculty of Science, Sydney School of Veterinary Science, Sydney, New South Wales 2006, Australia*

[2]*GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France*

[3]*Institute of Animal Science (460i), Department of Bioinformatics, University of Hohenheim, Stuttgart, Germany*

[4]*Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, United Kingdom*

[5] *The University of Sydney, Faculty of Science, School of Life and Environmental Science, Sydney, New South Wales 2006, Australia*

* Corresponding author. Mailing address: The University of Sydney, Faculty of Science, Sydney School of Veterinary Science, RMC Gunn Building B19, Sydney, New South Wales 2006, Australia. Phone: +61-2 9036 9348. Fax: +61-2 9351 3957. Email: jaime.gongora@sydney.edu.au

Abstract

The major histocompatibility complex (MHC) is a key genomic model region for understanding the evolution of gene families and the co-evolution between host and pathogen. To date, MHC studies have mostly focused on species from major vertebrate lineages. The evolution of MHC classical (Ia) and non-classical (Ib) genes in pigs has attracted interest because of their antigen presentation roles as part of the adaptive immune system. The pig family Suidae comprises over eighteen extant species (mostly wild), but only the domestic pig has been extensively sequenced and annotated. To address this, we used a DNA-capture approach, with probes designed from the domestic pig genome, to generate MHC data for 11 wild species of pigs and their closest living family, Tayassuidae. The approach showed good efficiency for wild pigs (~80% reads mapped, ~87× coverage), compared to tayassuids (~12% reads mapped, ~4× coverage). We retrieved 145 MHC loci across both families. Phylogenetic analyses show that the class Ia and Ib genes underwent multiple duplications and diversifications before suids and tayassuids diverged from their common ancestor. The histocompatibility genes mostly form orthologous groups and there is genetic differentiation for most of these genes between Eurasian and sub-Saharan African wild pigs. Tests of selection showed that the peptide-binding region of class Ib genes was under positive selection. These findings contribute to better understanding of the evolutionary history of the MHC, specifically, the class I genes, and provide useful data for investigating the immune response of wild populations against pathogens.

Key words Major histocompatibility complex, DNA sequence capture, Adaptive immunity, Pigs, Peccaries

Introduction

The family Suidae (Artiodactyla, Mammalia), commonly known as pigs or suids, diverged from their sister family Tayassuidae (peccaries or tayassuids) ~35 Ma (Gongora et al 2011). Most immunogenetic knowledge is limited to the domestic pig (*Sus scrofa*), and to a lesser extent to wild species of suids and tayassuids. These wild species play important roles in their natural environment, in agriculture, and in emerging, re-emerging, and zoonotic diseases (Al Dahouk et al. 2005; Meng 2012; Na Ayudhya et al. 2012). The lack of genetic resources from wild suids and tayassuids limits our understanding of the evolution of their adaptive immune responses, including that from the major histocompatibility complex (MHC).

The MHC is a multi-gene family (three subregions; Figure 1) that comprises immune (innate and adaptive) and non-immune genes and is important for understanding the development and regulation of immune responses in vertebrates (Penn and Ilmonen 2001). Comprehensive studies of major vertebrate groups, including mammals, have contributed to our knowledge about MHC function and diversity (Kulski et al. 2002; Frazer et al. 2003; Renard et al. 2006). The ecological adaptability and evolutionary success of suids and tayassuids in various environments (Meijaard et al. 2011; Taber et al. 2011) make them ideal for studying the genetic mechanisms behind the evolution of the MHC.

In the domestic pig, the MHC is located on chromosome 7 and is known as the swine leukocyte antigen (SLA), and the histocompatibility genes within the MHC region are referred to as SLA genes (Renard et al. 2006). The histocompatibility molecules (referred to as MHC class I and class II) are responsible for self/nonself recognition as part of the adaptive immune system (Borghans et al. 2004). In vertebrates, the classical (Ia) and non-classical (Ib) class I genes encode surface proteins expressed on nucleated cells. The former is highly polymorphic and expressed in most tissue types, whereas the latter have more limited diversity and expression. The MHC class II molecules are expressed on specialised antigen-presenting cells (e.g. macrophages) and have similar roles to class Ia molecules (Renard et al. 2001; Lunney et al. 2009). The antigen peptide-binding region (PBR) of the MHC class I and class II proteins are encoded by exons 2 to 3 and exon 2, and present intracellular and extracellular antigen peptides to T lymphocytes, respectively (Takahashi et al. 2000; Lunney et al. 2009). Most of the polymorphisms are within the PBR, contributing to the overall diversity of these genes (Piertney and Oliver 2006; Jaratlerdsiri et al. 2014). The class III region, also known as the inflammatory region, is comparatively well conserved among different vertebrate groups. Several immune-related genes, such as the tumour necrosis factor gene family (TNF) and complement proteins, are orthologous in Teleost fish, amphibians, mammals, and eutherians (Kelley et al. 2005; Deakin et al. 2006).

Comparative studies between model species such as humans and mice have broadened our understanding of adaptive evolution in the MHC (Carver and Stubbs 1997; Emes et al. 2003; Kelley et al. 2005). However, studies of non-model species have revealed considerable MHC variation between and within species (Janova et al. 2009; Kloch et al. 2010; Alcaide et al. 2014). This diversity is often explained by pathogen-driven selection (Kelley et al. 2005; Janova et al. 2009; Kloch et al. 2010), and mediated by balancing selection (Hughes and Yeager 1998) and/or the birth-and-death model (Nei et al. 1997). This model involves gene duplication leading to increased allelic diversity and gene loss occurring via the accumulation of deleterious mutations producing non-functional pseudogenes (Nei et al. 1997; Hughes and Yeager 1998; Barbisan et al. 2009). In mammals, a well-conserved set of anchor genes distributed across

the three subregions provides a framework for the histocompatibility genes to expand and diversify (Amadou 1999; Ando and Chardon 2006).

Orthologous MHC class II genes shared between distantly related mammals indicate that these loci were present before the radiation of major placental orders (Yeager and Hughes 1999). For example, the MHC class II DR and DQ loci are found in multiple mammalian species. Other loci, such as the DP and DY, have been lost (gene death) and gained (gene birth), respectively, in cows, sheep, and pigs (Kelley et al. 2005). In contrast, the MHC class I genes are not orthologous between mammalian orders, but some are orthologous within orders (Yeager and Hughes 1999), even though the positions of the anchor genes are identical (Kulski et al. 2002; Lunney et al. 2009). Differentiated blocks of genes, like the histocompatibility genes, have also been produced from rapid diversification after these taxa diverged from a common ancestor (Yeager and Hughes 1999; Kelley et al. 2005). In addition, the clustering of the class Ib genes with class Ia genes of taxa within the same order also suggests that the class Ib genes arose independently by gene duplication from the class Ia genes within orders or species (Rodgers and Cook 2005). The differences in MHC genes between mammalian orders and even species maintain the debate on what and how genetic mechanisms generate or maintain MHC diversity in wild populations, particularly in the MHC class I and class II genes (Piertney and Oliver 2006; Spurgin and Richardson 2010).

An early study of domestic pig MHC genes proposed that some class Ia genes emerged approximately 15 Ma, that the class Ib genes emerged after suids separated from other artiodactyls ~65 Ma, and that an ancestral gene of a class Ia pseudogene originated ~120 Ma (Renard et al. 2003). However, these hypotheses are yet to be tested in the context of the extant wild suids and tayassuids, including more recent estimates of divergence times between these taxa (Gongora et al. 2011). Furthermore, a study of MHC class II loci in domestic pig breeds and a limited number of wild suid and tayassuids showed that a few class II alleles are shared between these taxa (Luetkemeier et al. 2009). In the absence of species-specific data, the domestic pig MHC is, therefore, a valuable reference for investigating the retention and divergence patterns of these loci and the evolution of the MHC between extant taxa.

In this study, we generated MHC data for 11 wild species of suids and tayassuids using hybridisation DNA capture-based sequencing. This method is commonly used to resequence specific genomic loci in individuals belonging to the same species (homologous capture), but it can also be used to target distantly related species (heterologous capture) using a known sequence as the reference (Buckley 2007; Mamanova et al. 2010). We used the MHC Hp1a.1 haplotype of *S. scrofa* as a reference sequence for probe design (Renard et al. 2006; Stam et al. 2008; Groenen et al. 2012). Data from individuals within a species were combined to generate consensus sequences to infer the evolutionary relationships of genes within the MHC region and, in particular, the MHC class I genes. Our study provides the first MHC data for suids and tayassuids, laying the foundations for a better understanding of the diversity of this genomic region, and of host immune responses to environmental challenges at the species level.

Materials and methods

Sampling and sequence capture

Genomic DNA was extracted from 88 specimens, representing 9 out of 18 species (Meijaard et al. 2011) of Suidae (*n* = 69) from Eurasia and Africa and 2 out of 3 species (Taber et al. 2011) of Tayassuidae (*n* = 19) from the Americas (Table 1). These represent five suid genera (*Sus*, *Hylochoerus*, *Phacochoerus*, *Potamochoerus*, and *Babyrousa*) and two tayassuid genera (*Pecari* and *Tayassu)*. Samples were submitted to the Biosample Project PRJNA384704 (accession numbers SAMN07139417 to SAMN07139502). Two samples of the species *H. meinertzhageni* were not submitted because of the low quality of the sequences obtained downstream (see Results). For specific details of samples, see Online Resource 1.

The MHC sequence used for designing the capture array was obtained by merging the ~2.4 Mb sequence described by Renard et al. (2006) and the ~0.4 Mb sequence produced by Stam et al. (2008) (GenBank accession number: MF029693, Online Resource 2). At the time of experimentation, *SLA-12* (Tanaka-Matsuda et al. 2009) was not available, and therefore not included in the array design for retrieval. The custom 385K capture array was then designed by NimbleGen (Madison, WI, USA) using standard parameters, but increasing probe unicity from 1 to 25 due to the presence of multiple duplicated genes and repetitive elements in the MHC region. The final version of the design (101001_Sscrofa_INRA_SM_cap) covered 2,003,926 bp, approximately 72% of the initial region.

For each sample, 1.5 µg of genomic DNA was measured using Qubit fluorometer (Invitrogen, Carlsbad, CA, USA), resuspended in 130 µL of ddH_2O and fragmented using a Covaris S-2 instrument (Covaris, Woburn, MA, USA). The samples were purified with 1.8X AMPure Beads (Beckman Coulter Genomics, Brea, CA, USA) and resuspended in Resuspension buffer (Illumina, San Diego, CA, USA) to a final volume of 60 µL. DNA quality was checked using a DNA 1000 Bioanalyzer Chip (Agilent Technologies, Santa Clara, CA, USA). For each sample, the remaining DNA (59 µL) was used for library preparation. The TruSeq DNA Sample Preparation Kit (Illumina, San Diego, CA, USA) was used for end-repair, A-tailing, and adaptor ligation. Agarose gel size selection was omitted. After ligation, indexed samples were PCR amplified ('pre-capture enrichment') following the NimbleGen Array User's Guide Version 3.2 (NimbleGen), and the quality was assessed using a Qubit (Invitrogen) and a DNA 1000 Bioanalyzer Chip (Agilent Technologies).

A 385K array (NimbleGen) was used to hybridise 12 uniquely indexed and multiplexed libraries in parallel (Online Resource 3). The libraries were pooled in equimolar ratios to obtain a final amount of 5 µg of DNA (416 ng DNA/library). To improve hybridisation, 100 µg of *S. scrofa* Cot-1 DNA (Applied Genetics Laboratories, Melbourne, FL, USA) and 10 µL of six different 100 µM blocking oligonucleotides (Eurofins MWG Operon, Ebersberg, Germany) were added. These blockers (named BO1-O6) were described by Meyer and Kircher (2010) and were used to avoid hybridisation to the adaptors. The hybridisation mix was dried in a SpeedVac (Thermo Scientific, Waltham, MA, USA) at 60 °C. Hybridisation was performed following the manufacturer's instructions. Eluted pooled libraries were amplified by PCR ('post-capture enrichment') using the protocol recommended by Illumina, with the following modifications to cycles (cycle: number): L1–L2: 18, L3–L4: 17, and L5–L8: 15. A final quality check was performed and each library was quantified using Qubit (Invitrogen) and DNA 1000 Bioanalyzer Chip (Agilent Technologies).

Sequencing of each multiplexed captured library was performed on a lane of HiSeq 2000 (Illumina) as paired-end 101 bp reads with the TruSeq v3 Kit (Illumina). Raw image analysis and base calling were performed using the Illumina data analysis pipeline (Illumina 2009). Custom Perl scripts were used to successively trim raw reads for low-

quality bases at the 3' end, until finding a base with a phred quality score of >10 or until the read length became less than 40 bp. Sequences with Q<10 were removed. These steps were performed to increase the number of reads available for mapping. Reads were then mapped to the customised pig genome sequence (Online Resource 2) using BWA (Li and Durbin 2009). PCR duplicates were removed using rmdup from SAMtools (Li et al. 2009).

The efficiency of the DNA capture approach was measured using seven parameters: i) % of total reads mapped onto the reference; ii) specificity, given as the % of reads mapping on targeted regions (100 bp downstream and upstream from each contiguous group of capture probes); iii) evenness of coverage, or 'E score', as defined by Mokry et al. (2010); iv) coverage within target region; v) % of duplicated reads; vi) C15 score, % bases with at least 15× coverage; and vii) level of enrichment of the coverage in the targeted regions compared with the rest of the genome. An analysis of variance (ANOVA) was used to evaluate the significance of the capture parameters. This was compared between species, genera, and families using a mixed-effects ANOVA *F*-test in R 3.0.1 (R Development Core Team 2008) (covariate: library, random effect: sample, fixed effect: species/genus/family). When differences between species, genera, or families were significant ($P < 0.05$), Tukey's test was used to compare the means of the parameters. An unsupervised, two-way hierarchical analysis was performed using the FactoMineR package (Lê et al. 2008) to visualise clusters of individuals based on their variance-covariance structure and to examine the similarities of the capture array efficiency between the studied species. The resulting .bam files were submitted to the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) under the study SRP108721.

Generation of MHC consensus sequence and bioinformatics

Given the variability of coverage between conspecific individuals and genetic distance between species, consensus sequences were generated for each species using SAMtools (Li et al. 2009). This was to illustrate the degree of sequence conservation and divergence at a species level (Choo et al. 1991; Day and McMorris 1992) and for downstream analyses. All samples from the same species were merged to increase the number of reads mapped against the reference sequence. Variant calling was performed using bcftools (Li et al. 2009) and a 10× depth coverage was used as the initial cut-off value. A minimum phred-scaled probability score of 10 was used for SNP call quality to retrieve the conserved and divergent nucleotide. A final consensus was obtained for each species (Online Resource 4) using the pileup and variant call format (VCF) outputs. Automatic annotation was performed by aligning similar DNA regions between the *S. scrofa* reference and each consensus sequence using Advanced PipMaker (Schwartz et al. 2000). Interspersed repeats were detected using RepeatMasker (http://www.repeatmasker.org) and libraries available for mammals. This program treats artiodactyl queries best when compared with other species of the same family.

Phylogenetic analysis of MHC genes

We performed phylogenetic analyses of i) all genes retrieved in the MHC region independently (including 57 class I, 30 class II, and 58 class III genes) to infer the orthologous relationships of genes within the MHC between species of suids and tayassuids; and ii) 10 MHC class Ia (*SLA-1*, *SLA-2*, *SLA-3*, *SLA-4*, *SLA-5*, *SLA-9*, and *SLA-11*) and Ib (*SLA-6*, *SLA-7*, and *SLA-8*) genes to infer the evolutionary relationships between the MHC class I genes. The protein-coding sequences were used because they are likely to reflect the evolutionary forces acting on functional genes. For

pseudogenes, we included the whole exonic region. Custom Perl scripts were used to retrieve exonic regions of MHC genes and validated against existing databases including Ensembl (http://www.ensemble.org, Aken et al. 2016) and the Vertebrate Genome Annotation (VEGA, http://vega.sanger.ac.uk, Wilming et al. 2007) (Online Resource 5). All alignments were performed using MAFFT-E-INS-i (Katoh and Standley 2013).

For the analysis of dataset i), MEGA v7 (Kumar et al. 2016) was used to find the best-fitting substitution model, based on the Bayesian information criterion. Maximum likelihood was used to infer the phylogeny and node support was estimated using 1,000 bootstrap replicates. To infer the placement of the root, we treated the tayassuids as the outgroup. Details of the alignments are available in Online Resource 6. Gaps were partially deleted to retain as many informative sites as possible without removing the whole site. We tested for substitution saturation at each codon position using Xia's test of saturation in DAMBE version 6.4.42 (Xia 2017). For this analysis, none of our histocompatibility gene alignments shows evidence of saturation (data not shown), except for the first codon positions of *SLA-5*. Thus, we inferred the phylogenetic tree without the first codon position.

For the dataset with the 10 MHC class I genes, we performed Bayesian phylogenetic analyses in MrBayes v3.2 (Ronquist et al. 2012) to gain insight into the phylogenetic relationships of these genes. The best-fitting nucleotide substitution model, HKY+G (Hasegawa et al. 1985), was selected using Modelgenerator (Keane et al. 2006) based on the Bayesian information criterion. Posterior distributions of all parameters, including the tree, were estimated using Markov Chain Monte Carlo (MCMC) sampling. The MCMC analysis was run until the standard deviation of split frequencies between two independent MCMC runs fell below 0.01, with the first 25% of steps discarded as burn-in. As suggested by Renard et al. (2003), *SLA-11*, classified as a pseudogene, was used as an outgroup as the precursor of other SLA genes. Although unpublished RNA-sequencing data of the domestic pig suggest that *SLA-11* could be a functional gene (Rogel-Gaillard, in preparation), it has been shown to be orthologous to a relic segment of a human class I fossil gene (Renard et al. 2003). Therefore, this does not preclude the use of this sequence as an outgroup. Further studies will be required to clarify its pseudogene/functional gene status. As a means of comparison, we also inferred the phylogeny using maximum likelihood in MEGA v7 (Kumar et al. 2016), using the same model and all sites, with 1,000 bootstrap replicates.

Selection tests

To gain insight into the diversity and conservation and the evolutionary forces that might be acting on genes within the MHC between species, we tested for selection in i) 145 genes separately. We also compared the available transcripts of genes, where possible, with the online databases Ensembl and VEGA using MAFFT-E-INS-i (Katoh and Standley 2013). We then tested for selection in ii) exons 2 to 3, the PBR, and non-peptide binding region (non-PBR) of the class Ib histocompatibility genes (*SLA-6*, *SLA-7* and *SLA*-8). We excluded the MHC class Ia and II genes due to the large number of gaps (>30 bp) and low coverage in the PBR (Online Resource 9b) which are not reliable for a selection test. Tests of selection compared the rate of nonsynonymous substitutions (d_N) with the rate of synonymous substitutions (d_S). The d_N/d_S ratio provides an indication of the direction of selection acting on a gene: where $d_N/d_S < 1$ indicates negative selection, $d_N/d_S > 1$ indicates positive selection, and $d_N/d_S \sim 1$ indicates the absence of selection (Nei and Gojobori 1986). Here, we wished to detect whether genes in the MHC region were under positive or negative selection for future studies on specific selection mechanisms. The tests performed can indicate whether

changes are radical or conserved between species/family and to observe whether there are increased rates of nonsynonymous substitutions than under neutral evolution.

The mean rate of synonymous and nonsynonymous substitutions per site was estimated using the modified Nei & Gojobori (1986) method with Jukes-Cantor correction to account for multiple substitutions at the same site. Estimates of standard errors were obtained using 1,000 bootstrap replicates, and gaps were subject to pairwise deletion. Statistical significance was evaluated using codon-based Z-tests and testing the null hypothesis ($d_N=d_S$) against the alternative hypotheses, $d_N>d_S$ and $d_N<d_S$. We only considered the rates and ratios that were non-zero. All tests were performed in MEGA v7 (Kumar et al. 2016).

Results

Sequence capture efficiency and output

The capture parameters varied greatly across taxa (Table 2) and all other capture parameters between library and genus (Online Resource 7). For statistical analysis, we removed two outliers from *H. meinertzhageni* with low coverage (<1×). The highest percentages of mapped reads were found in *S. scrofa* (84.5%) as expected, followed by other related species of the genus S*us* (77.0–82.3%) and the related sub-Saharan genera (*Hylochoerus*, *Phacochoerus*, and *Potamochoerus*; 76.0–80.6%), and *Babyrousa* (73.3%). There were significantly lower percentages of mapped reads in tayassuids (11.7–12.2%). This pattern follows the taxonomic relationships and genetic distances between the different species studied here in relation to the *S. scrofa* used for designing the capture array. There were no clear patterns in most of the remaining parameters (i.e. parameters ii-vii), except for significant differences in the level of efficiency between suids and tayassuids.

Three parameters were used to cluster the data obtained by sequence capture: the percentage of total mapped reads, the specificity, and the total coverage. The resulting tree based on the percentage of total mapped reads (Online Resource 8a) clustered individuals according to their family and genus, with a few minor exceptions. This is similar to the taxonomic relationship (Figure 2), where tayassuids are the sister group to suids, followed by the South East Asian *Babyrousa* as sister taxon to the Eurasian (*Sus*) and sub-Saharan African suids (*Potamochoerus*, *Phacochoerus*, and *Hylochoerus*), and *Potamochoerus* as the sister taxon to *Hylochoerus* and *Phacochoerus* (Gongora et al. 2011). Although the status of *S. scrofa* still needs clarification due to the recent taxonomic update of some subspecies to species, the sampling used in Figure 2 does not reflect this (Gongora et al. 2017). The tree obtained using the total coverage (Online Resource 8b) followed a similar pattern to the total mapped reads, but there was a slightly weaker correspondence with the taxonomic relationships. In contrast, the tree based on specificity (Online Resource 8c) did not reproduce the same pattern as the total reads mapped or coverage, except for the clear split between suids and tayassuids. Therefore, the results of the clustering analyses closely reflected the structure of the data as presented in Table 2 when the same capture-efficiency parameters were considered.

The average coverage of the MHC region in suids is mostly even (Online Resource 9) and the lower coverage seen in tayassuids (<6×) reflects the lower efficiency in capture output. Some regions had extremely high coverage (>1000×), these regions correspond to approximately 890,000–896,000 bp, 4,607,500–4,615,000 bp and 4,720,000–4,807,000

bp of the MHC region (indicated by the star in Online Resource 9a). Comparison of gap regions (>30 bp) consistent between species contained the MHC class Ia genes (~327,000–497,006 bp) and class II genes (~4,900,000–5,270,000 bp) as indicated by the black boxes labelled 1 and 4 respectively. Gaps indicated by box 2 (~4,620,000-4,720,000) and 3 (~4,805,000-4,845,000) (Online Resource 9b) does not contain any genes, the former are made up of tandem repeats including 'paramyosin-like' and PolyA stretches and the latter contain PolyA stretches. Both these regions have a low probe coverage (~50%) compared to ~94% in other regions. Online Resource 9b shows the overall coverage per species in *SLA-1*, *SLA-6*, and *SLA-DQB1* genes for comparison. Particularly lower coverage was found in the PBR of the class Ia (exons 2 and 3) and class II (exon 2) genes. The MHC Ib genes did not display this pattern of low coverage or gap.

In total, 145 loci out of the 153 loci identified by Renard et al. (2006) were retrieved in the MHC region (See Online Resource 6 for gene details). MHC gene sequences (class I, II, and III) are available in Online Resource 10. MHC gene alignments include VEGA sequences for alignment purposes. The genes not retrieved were the olfactory receptors (*OLF42-1*, *OLF42-2*, and *OLF42-3*), *AFP*, *LST1*, *LY6G6E*, and *TNXB*. The olfactory receptors and TNXB are within highly repetitive LINE1 clusters and are challenging to sequence (Zozulya et al. 2001; Treangen and Salzberg 2011; Chiovaro et al. 2015). *AFP* is a possible pseudogene of *TRIM26* and is adjacent to *TRIM26* in the domestic pig genome (Renard et al. 2006) and may not have been sufficiently captured. The probe design for the region in which *LST1* was located might not have been optimal because it seems to be a highly diverse gene, with low sequence similarity between eutherian species (36% between human and mouse) and appears to be absent in marsupials (Deakin et al. 2006). Similarly, this is the case for LY6G6E, which is expressed in lineage-specific patterns and is widely used as a cell marker for leukocytes (Loughner et al. 2016).

The *SLA-DYB* gene was not retrieved in *P. tajacu*, *P. africanus*, and *P. larvatus*. *SLA-11* was translated due to its uncertain pseudogene/functional gene status in the domestic pig, and complete reading frames were identified in the domestic pig and the rest of the species, which has not been reported previously. We identified premature stop codons in *Hyme SLA-3* (residue 168), *Phaf SLA-7* (residue 386), *Peta SLA-7* (residue 386), *Sucel SLA-5* (residue 135) and *Hyme SLA-5* (residue 372). Stop codons were also found in the pseudogenes *SLA-4* and *SLA-9* in all species, as expected. Although the class II gene *SLA-DOB2* is classified as a pseudogene (Renard et al. 2006), we only identified stop codons in the tayassuids (residue 37).

Phylogenetic analysis of MHC genes

From our phylogenetic analyses of i) all genes in the MHC, we will focus on the histocompatibility genes (MHC class I and class II) and the anchor genes that were found in all of the species examined. Our results show that some MHC class I genes yielded similar tree topologies when tayassuids was used as the outgroup (Online Resource 11). Most genes are distinguished to the genus level (Online Resources 11a, e, g-j) with some exceptions, some to the family level (Online Resources 11c, d), and others not grouped according to any taxonomic relationship (Online Resources 11b, f) as illustrated in Figure 2 (Gongora et al. 2011).

Compared with the class Ia genes, the groupings of the class Ib genes closely followed the taxonomy of the species. Similar tree topologies are seen for the MHC class II genes (Online Resource 12). *SLA-DRA* and *SLA-DQB1* (Online

Resources 12a and d) show topologies more similar with the species tree, whereas *SLA-DQA* (Online Resource 12c) mostly presented family-level grouping and some genus-level grouping. For *SLA-DRB1* (Online Resource 12b), the tayassuids grouped with the rest of *Sus* and sub-Saharan African species, with no distinction in terms of genus or family. The anchor genes (Online Resource 13) displayed a similar variation in topology to the MHC class I and II genes. Genes that followed closely to the species taxonomy include *GNL* and *C4A* (Online Resource 13b and l), those by genus with some exceptions include *MOG* and *TCF19* (Online Resource 13a and d). One gene, *RXRB* (Online Resource 13p), showed no distinction between most of *Sus* and the sub-Saharan African suids but had a low bootstrap value (5).

For our analysis of the 10 MHC class I genes, we found complete putative protein sequences for *SLA-1* to *SLA-9* and *SLA-11* in all species of suids and tayassuids (Figure 3). The MHC class I genes are grouped in four main clades which are mostly orthologous (posterior probabilities, 0–1; and supporting bootstrap values, 0–100, are given in parentheses): 1) *SLA-3* genes (0.74); 2) *SLA-2* and some peccary class Ia genes (0.76); 3) *SLA-1* genes (0.63); *SLA-5* and *SLA-9* (0.56/6); and 4) *SLA-4* and the class Ib genes (0.98/43). Within clade 2, *SLA-1* to *SLA-3* from *T. pecari* and *SLA-2* and *SLA-3* from *P. tajacu* form a monophyletic group. Genes not within any main clade include *Baba SLA-3* from and *Popo SLA-5*, which forms a polytomy with the classical Ia genes in clades 2 and 3 respectively. The relationships of these genes between the species in clades 1 to 3 are partially resolved, where most branches show dichotomous branching, although they do not necessarily follow the species tree. The polytomy here might indicate the rapid differentiation of these genes before speciation. However, the low bootstrap values and inconsistencies between the Bayesian and maximum-likelihood trees for most of the classical class I genes (clades 1 to 3) suggested poor resolution of the relationships of these genes across wild suids and tayassuids.

Selection within the MHC region

Across the species studied here, we found variable rates of synonymous and nonsynonymous substitutions (Table 3). The average nonsynonymous substitution rates of the protein-coding MHC (histocompatibility) class I genes is higher (0.028) than that of class II (0.01976), as expected due to the polymorphic nature of the class Ia genes. The overall nonsynonymous substitution rates of class I genes is lower (0.01305) than expected compared with genes in the class II region (0.01487), but the values for coding genes (0.01272) in class I are slightly higher than for those in class II (0.0127). The slightly lower synonymous rate in all the class I protein-coding genes (0.03413) also inflated the overall d_N/d_S ratio (0.47622) compared with the class II region (0.39976). The synonymous substitution rates are more similar across different gene types and are also higher in the anchor genes compared with nonsynonymous substitutions. As expected, the anchor genes have very low nonsynonymous substitution rates compared with any other type of genes ($d_N < 0.007$, $d_S > 0.034$) but are similar across different type of genes (overall averages, non-protein coding genes, and histocompatibility genes). Overall, we see a large portion of synonymous substitutions within the genes in the MHC region regardless of their role, with variable rates of nonsynonymous substitutions between various genes in the region.

Tests that yielded significant evidence of positive selection included those conducted for *SBAB-499E6.10* ($P = 0.01$) mapping to the class I region (Online Resource 14a), but this gene contained multiple stop codons in all the species examined. Seven other genes (*C7H6orf12*, *ZNRD1*, *TRIM40*, *C7H6orf15*, *PSORS1C2*, *SLA-8*, and *SLA-7*) were found to be under negative selection. Ten class III genes (*MCCD1*, *APOM*, *LY6G6B*, *LY6G6C*, *LY6G6D*-005, *C7H6orf25*,

LSM2, *HSPA1A*, *GPSM2*, and *BTNL5*) showed evidence of negative selection and two genes (*LY6G6D*-004 and *C7H6orf31*) were found to be under positive selection (Online Resource 14b). These genes are all classified as protein-coding in *S. scrofa*. However, only partial sequences were retrieved for *HSPA1A* (5' end missing), a novel coding sequence classified by Renard et al. (2006), and contained stop codons. Stop codons were also present in *Tape MCCD1* (residue 117), and in *Phaf* and *Peta SLA-7* (residue 386). Six genes in the class II region, including histocompatibility and non-histocompatibility genes (*SLA-DRB1*, *SLA-DQA*, *SLA-DOB2*, *SLA-DRB5*-201, *SLA-DOA*, and *SLC39A7*-003) showed evidence of negative selection (Online Resource 14c). In contrast with the findings of Renard et al. (2006), the predicted protein-coding sequences of these *Susc* loci contained stop codons (except *SLA-DQA*, *Popo* and *Pola SLA-DRB1*).

Our analysis of the PBR and non-PBR (Online Resource 15) showed the MHC class Ib genes significant for positive selection ($P<0.05$). The d_N/d_S ratios of the PBR and non-PBR in the Ib genes (*SLA-6*, *SLA-7*, and *SLA-8*) are 1.85–2.91. 0.51–0.95 respectively. The higher nonsynonymous substitution rates and synonymous substitution rates in the PBR than in the non-PBR was expected.

Discussion

Sequence capture efficiency

Overall, our DNA capture approach was more efficient for suids than in tayassuids. The capture parameters used in this study provide some insight into the performance of the method. The effect of sequence divergence of the target from the reference sequence is important and can be assessed by the percentage of total reads mapped. Although the variation found within family or closely related species and individual parameters can be affected by technical aspects, this can be useful in determining improvements to the protocol.

A typical homologous capture experiment has an average specificity of about 60%, with values ranging from 20% to 70% (Hodges et al. 2009; Cummings et al. 2010; Hoppman-Chaney et al. 2010). The blockers used here were the only ones available at the time (Meyer and Kircher 2010). These blocker sequences influence how well the protocol prevents the cross-hybridisation of off-target fragments, such as adaptors (Burbano et al. 2010; Harakalova et al. 2011). The adaptors used were also considered longer than others and prone to off-target hybridisation (Nijman et al. 2010; Rohland and Reich 2012). In a later heterologous study using the same protocol in chickens (*Gallus gallus*), an increase in specificity of ~10% was achieved (42.2%) (Moroldo, unpublished data). The percentage of duplicated reads is similarly affected by technical issues. The two rounds of amplification (pre- and post-capture PCR) increased the likelihood of duplicates, commonly occurring during PCR. However, this can be amended by duplicate removal steps and does not affect subsequent analyses (Ebbert et al. 2016).

Specificity, coverage and the E score can be influenced by how polymorphic the region is, where less divergence between the target and probe sequence is easier to recover (Buckley 2007). The exceptionally high polymorphism in the class Ia and class II genes (Lunney et al. 2009) makes this challenging. In addition, species-specific and pig breed line-specific class Ia and class II haplotypes add to this challenge (Ho et al. 2010; Essler et al. 2013). On the other hand, regions that are highly conserved have better coverage. The first region of high coverage (>1200×; 890,000–

896,000) extends ~6000 bp and contains the *DPCR1* gene. This gene is classified as a pseudogene, but a complete open reading frame (ORF) was found in our consensus sequence. The *DPCR1* gene is also found in a region of high sequence similarity (~60%) to humans (Shigenari et al. 2004). The second region (4,607,500–4,615,000) also corresponds to a highly conserved region between mammals, the BTNL family proteins. However, the third region (~4,720,000–4,807,000) corresponds to an uncharacterised region containing no genes, but have homology to some segments of chromosome X, Y, 2, 5, 7, and 14. Further studies will be needed to determine the characteristics of the *DPCR1* and uncharacterised region in suids and tayassuids. The E scores within each family are mostly homogeneous and within benchmarked values (Mokry et al. 2010). However, the low levels obtained in tayassuids (33.4%) compared with suids (64.3%) likely reflect the high sequence divergence of the target from the reference sequence, similarly seen in specificity and other capture parameters.

Regarding the two gap regions corresponding to ~4,620,000-4,720,000 bp and ~4,805,000-4,845,000 bp, these contain many tandem repeats which are difficult to sequence. These sequences are paramyosin-like which contain repetitive elements similar to the polyA stretches (Mooseker and Cheney 1995; Treangen and Salzberg 2011). In addition to this, the probe coverage of both regions was 50% and 45% respectively, compared to the probe coverage where the MHC histocompatibility genes were located, was ~94% (data not shown). This resulted in almost 0x coverage in the region.

The degree of genetic divergence between the reference and target sequence also influences the capture efficiency. Its effect was clearly seen in the hierarchical clustering of the coverage and percentage of reads mapped (Online Resource 8a and b), with close intermingling of *Sus* species. These species diverged about ~1.3-3.7 Ma, which is much more recent than their divergence from sub-Saharan African suids ~10 Ma (Gongora et al. 2011). Our sampling of *Sus* also included feral pigs, which possibly consisted of populations Asian and European genetic origins (Gongora et al. 2004) and might be reflected by this clustering as some *S. scrofa* subspecies have been elevated to the species level (Gongora et al. 2017). The effect of divergence is also reflected by the species in each library (Online Resource 3). Library 5, which contained only *Sus* species, had a higher coverage. Library 6, which contained only *S. scrofa*, had the highest percentage of reads mapped. These values contrast with those of other libraries including taxa more divergent from the reference sequence, as seen in library 7 containing only tayassuids, for which all parameters were the lowest compared with the other libraries. The use of multiple, closely related species as a reference might improve the overall capture results (Peñalba et al. 2014).

In contrast to Renard et al. (2003), our analyses reveal that *SLA-11* encoded complete ORFs in all our species and is potentially a functional gene. This finding is consistent with unpublished RNA (Rogel-Gaillard, in preparation) data that show this to be expressed in domestic pigs. In addition to the findings of previous studies (Renard et al. 2003; Renard et al. 2006), the presence of both functional and null alleles found for *SLA-3*, *SLA-5*, and *SLA-7* in both domestic pigs and some wild suids and tayassuids (premature stop codon in *Hyme SLA-3; Hyme* and *Sucel SLA-5; Phaf* and *Peta SLA-7*), and *SLA-DOB2* in tayassuids, indicates the extent of gene duplication, loss, and copy number variation that is occurring in the MHC. Null (or pseudogenes) genes arise as a result of causes such as the loss of promoters or mutations that produce premature stop codons (Pink et al. 2011). The protein-coding genes here have maintained correct ORFs and only one stop codon was identified in *Hyme SLA-3*, and *Peta* and *Phaf SLA-7*, in comparison with the classified pseudogenes where multiple stop codons were identified (~7 positions in *SLA-4* and

~5 in *SLA-9*). Stop codons were not identified in all species for *SLA-9*, but missing data present in the known stop codon positions prevent us from determining whether this locus is functional or not. As DNA sequence analysis was used to infer the functional status here, RNA studies are required to corroborate this beyond the domestic pig.

Evolutionary history of the MHC class I genes

Our Bayesian phylogenetic analyses of the 10 MHC class Ia and Ib genes have provided the first comprehensive insight into the evolution of these genes in wild pigs and tayassuids (Figure 3). The evolution of the class I genes has been difficult to assess because they are rarely orthologous between species of different orders or family. For example, there are 10 or 11 MHC class I genes in the chimpanzee and human (Primates, Hominidae) (Anzai et al. 2003) that are orthologous and 28 in the rhesus macaque (Primates, Cercopithecidae) (Kulski 2004). In our study, we found that MHC class I genes are orthologous to the suborder level (Suina) which include both suids and tayssuids. The presence of class Ia and Ib genes in all the species studied indicates these loci emerged and underwent a series of duplications before these families diverged from the common ancestor over 35 Ma (Gongora et al. 2011). This contrasts with a previous hypothesis that suggested that the class Ia genes emerged ~15 Ma (Renard et al. 2003). Although the number of MHC class I genes per species can potentially differ between haplotypes, our capture protocol could only capture genes that share homology with the reference sequenced used. The class Ib genes show greater genetic differentiation between the Eurasian, sub-Saharan African suids, and tayassuids compared with the class Ia genes. This further supports the idea that these genes have more species-specific roles than the class Ia genes (Kusza et al. 2011).

The roles of the non-classical genes in the domestic pig are still unclear, despite various studies on the level and localisation of their expression (Crew et al. 2004; Kusza et al. 2011). However, the non-classical genes in humans, including HLA-E, HLA-F, and HLA-G have been well studied and have been shown to bridge the innate and adaptive immune response (Rodgers and Cook 2005; Allen and Hogan 2013). The most transcribed non-classical gene is SLA-8 and it has been suggested to be a functional homologue to HLA-E because of their similar range and level of expression in tissue (Kusza et al. 2011). The high level of nonsynonymous substitutions found in *SLA-8* may indicate a role in presenting species-specific self and nonself peptides that are unique to the environment in which the individuals are found. HLA-E acts as a self-immune surveillance to protect against potential invasion from pathogens by regularly surveying the number of peptides derived from MHC Ia molecules as well as nonself peptides. A lack of MHC class Ia peptide presentation activates natural killer cell-mediated lysis as an indication of MHC down regulation and the cells/host are potentially immunocompromised (Joosten et al. 2016). In our study, we identified a predicted *SLA-7* transcript with seven exons for all species, but stop codons were present in *P.africanus* and *P.tajacu* (residue 386). An RNA study in domestic pigs for *SLA-7* identified a transcript with eight exons (Hu et al. 2011), differing from the current annotation of seven exons. Its complex transcription pattern reveals a need for further investigation of this, and the remaining non-classical genes, to address questions related to their role and species-specific variants.

Our individual phylogenetic analyses of the MHC class I, class II, and anchor genes (Online Resources 11-13) provided some insight into the evolution of specific genes within Suidae and Tayassuidae through the topology produced. The genes that show the closest topology to the species tree (Figure 2) are protein-coding; *SLA-1*, *SLA-2*, *SLA-6*, *SLA-7*, *SLA-8*, and *SLA-11*. However, our Bayesian analysis (Figure 3) showed less distinction between Eurasian, sub-Saharan African suids, and tayassuids in the class Ia genes. This suggests that the class Ia genes are

highly homologous to each other, but still contain species-specific differences, especially between *SLA-1*, *SLA-2*, and *SLA-3*. Furthermore, the class Ib genes are not clustered with the class Ia genes and follow the species tree closely, indicating that they might have more species-specific roles (Kusza et al. 2011).

The class II *SLA-DRA* and *SLA-DQB1* genes show more distinction between species than *SLA-DQA* and *SLA-DRB1*. Compared with the class I genes, for which orthology between different orders is rare, the class II genes have a different evolutionary history where orthologous class II genes occur between different mammals (e.g., human class II is orthologous to the mouse E region) (Yeager and Hughes 1999). This might be because the class II genes form a heterodimer with the α and β chains and any variation in either can cause disability in the class II molecule. This is also apparent in the extremely high synonymous substitution rate in *SLA-DQB1* non-PBR (0.076). Our phylogenetic analyses are in contrast with some MHC class II studies which show more shared alleles between donkey and horse *DRA* and a more variable *DQB*, whereas our *SLA-DRA* and *SLA-DQB1* showed more distinction between species than *SLA-DQA* and *SLA-DRB1* (Janova et al. 2009).

Within the anchor genes, GNL1 and C4A both have highly specific roles. GNL1 is a G protein-like receptor that has a role in cell signalling and binding hormones, neurotransmitters, ions, and other stimuli (Rosenbaum et al. 2009). This is consistent with the different social organisation and variety of scent glands between different species of suids and tayassuids. C4A is a part of the complement cascade and is a highly polymorphic complement gene, where diversity can be advantageous for combating a range of pathogens (Castley and Martinez 2012). The less differentiated genes, MOG, TCF19, and RXRB, are related to highly conserved basic biological functions in mammals, such as nerve myelination, transcription factor, and regulation of cellular growth, respectively (Krishnan et al. 1995; Nagata et al. 1995; Kersten et al. 1997; Johns and Bernard 1999; Castley and Martinez 2012).

Selection in MHC genes

Of the genes under positive selection (excess of nonsynonymous substitutions), *SBAB-499E6.10* (class I subregion) is suggested to be a transcribed pseudogene that overlaps the *CDSN* pseudogene (Renard et al. 2006). Transcribed pseudogenes may have roles in regulating their 'parent genes' (such as interfering RNA). CDSN regulates the formation of the cornified envelope of the skin protecting the internal body from the environment against physical, chemical, and microbial agents (Simon et al. 1997; Matsumoto et al. 2008). Strong positive selection in FLG, a human skin barrier gene, has been linked to some common loss-of-function alleles, but non-dysfunctional heterozygotes can also present benefits without being deleterious (Irvine and Irwin McLean 2006). It is important to investigate whether skin barrier type coding genes play a role in producing low-level exposure to pathogens that can promote local adaptation or 'natural vaccination' (Irvine and Irwin McLean 2006). The high tolerance of bushpigs (*Potamochoerus*) and warthogs (*P. africanus*) to ASFV in sub-Saharan Africa where it is endemic, is of particular interest (Costard et al. 2009).

Within class III, one of the loci under positive selection was *C7H6orf31*. This gene is orthologous to the human chromosome 6 open reading frame 31, but its function is unclear in pigs (Renard et al. 2006). This locus plays a role in synaptic transmission for AMPA receptors (Kirk et al. 2016) and learning and memory in rats (Brinton et al. 2008).

The cognitive abilities of *S. scrofa* have been shown in experimental and natural settings; these have included the social recognition of familiar and unfamiliar conspecifics (Gieling et al. 2011). Likewise, wild suids and tayassuids stay together in familial herds and are rarely accepted between herds (Fowler 1996; Taber et al. 2011). The white-lipped peccary (*T. pecari*) has also been recorded to counterattack predators and disperse into smaller groups for foraging to avoid repetitive foraging in recently foraged areas (Taber et al. 2011).

The *LY6G6D* gene (class III) codes for the lymphocyte antigen 6 complex, locus G6D, and is a part of the *LY-6* gene family. Studies have shown that this family of genes have diverse functions depending on the cell type(s) on which they are expressed (Mallya et al. 2006). Its role in detecting chemoattractant gradients can control the movement of macrophages and other immune cells to the required site of response (Rodríguez-Fernández and Cabañas 2013). Some diseases in wild suids have been shown to modulate or subvert the host immune response, such as Porcine reproductive and respiratory syndrome virus (Na Ayudhya et al. 2012) and ASFV (Dixon et al. 2004). It is possible that this positive selection seen in LY6G6D has a role in processing certain pathogens in the environment of these species. Alternatively, its potential role in hematopoietic cell differentiation allows monocytes/macrophages to be distinguished into different subpopulations, where some subpopulations of monocytes have been implicated with infection of African swine fever virus (Sánchez-Torres et al. 2003).

We found many genes to be under negative selection, which might reflect conservation of sequence function within wild suids and tayassuids. This form of selection tends to keep new or radical changes to the gene at low frequencies (Fay et al. 2001; Mukherjee et al. 2009). It is also interesting to note that if genes evolve under the birth-and-death model under strong negative selection (Nei and Hughes 1992), it is more likely that pseudogenes are generated (Piontkivska et al. 2002). This might explain the numerous MHC class II pseudogenes (*SLA-DRB2* to *SLA-DRB5*, *SLA-DQB2*) that are seen in suids and tayassuids, with the protein-coding gene being under negative selection (*SLA-DRB1*).

A high diversity of MHC alleles is generally recognised to confer protection to a range of pathogens, and rare alleles might offer better protection (van Oosterhout 2009). This is due to negative frequency-dependent selection, whereby pathogens evolve to avoid common MHC variants. This host-pathogen co-evolution results in a range of alleles and genes that give resistance to different pathogens, and variants are rarely fixed (van Oosterhout 2009). In turn, this leads to an abundance of similar genes, as seen here within the classes Ia and Ib and in *S. scrofa* (wild boar) other species such as crocodiles and birds (Barbisan et al. 2009; Jaratlerdsiri et al. 2014; Alcaide et al. 2014).

Our analysis of the PBR and non-PBR in the MHC class Ib presented significant for positive selection which was expected for antigen-presenting genes. The positive selection indicates that a diversity of antigen presenting peptides might provide an advantage in combating different infectious diseases (Sommer 2005; Kloch et al. 2010). The higher nonsynonymous substitution rates of the class Ib genes between species further suggests some species-specific differences, as previously reported (Tennant et al. 2007; Kusza et al. 2011). This is also interesting because the non-classical genes are usually described as oligomorphic (Lunney et al. 2009). In contrast, the non-PBR are usually conserved and contain the leader peptides, transmembrane domains, and cytoplasmic tail, and are also responsible for regulation and facilitating gene expression (Ballingall and McKeever 2005; Drake et al. 2006; Barrett et al. 2013). This is evident in the higher synonymous substitution rates in the MHC class Ib non-PBR. Within the domestic pig,

there are over 200 SLA classical class I alleles, 18 non-classical class I alleles, and 212 SLA class II alleles (https://www.ebi.ac.uk/ipd/mhc/group/SLA). Our protocol was therefore unable to obtain sufficient data of the class Ia and II genes due to to high divergence in this region (Burri et al. 2008; Jaratlerdsiri et al. 2012; Moutou et al. 2013; Alcaide et al. 2014), and a comparison of diversity in these genes between wild suids and tayassuids cannot be made. However, future work using the data from this study can be used to pursue more efficient sequencing of the PBR in the MHC class Ia and class II genes. This will allow us to perform more comprehensive diversity studies within and between species, and detect genes that might be maintained due to trans-species polymorphisms (shared alleles between long-diverged species) or other genetic mechanisms (van Oosterhout 2009).

Conclusions

Overall, our study has demonstrated that heterologous DNA capture is a useful approach for investigating the MHC of non-model species. This approach was more efficient in Suidae than Tayassuidae. Genetic distance is a major factor that influences the efficiency of this approach, but technical factors cannot be excluded. Approximately 145 MHC loci, including the histocompatibility genes, were characterised for each species.

We also reveal that the repertoire of classical and non-classical class I genes found in the domestic pig are present in both wild Suidae and Tayassuidae, indicating that these loci emerged prior to the divergence of these two groups. During this time, these genes underwent a series of duplications that generated 10 class I loci (including pseudogenes). Subsequently, genetic differentiation of these genes after speciation and pathogen-mediated selection might have contributed to the distinct genetic patterns of the classical and non-classical class I genes between extant Eurasian and sub-Saharan African suids. We detected positive selection in the peptide-binding regions of the non-classical genes, but additional studies for the classical class I and class II genes are needed.

Our findings lay the foundation for improving our understanding of the immunogenetics in Suidae and Tayassuidae such as conservation and diversity of the class I and II histocompatibility genes. For instance, our data can be used to investigate the immune response of histocompatibility genes in sub-Saharan African bushpigs to better understand their local adaptation to African swine fever. Similarly, our data can be used to assess the genetic diversity of the MHC among natural populations of wild suids and pecarries to identify the mechanisms of evolution and selection that have been shaping and maintaining variation. Finally, the DNA hybridisation-based method described here can be applied to study the MHC or other complex immune loci among closely related taxa and within species as a cost-effective method in non-model organisms.

Acknowledgements

We thank the Sydney School of Veterinary Science at the University of Sydney for providing research funding for sampling and financial support, allowing J. Gongora to undertake preliminary experiments on the MHC cross-species approach and a sabbatical period to generate data for this project at INRA. We thank INRA for making funding available from The French National Research Agency (PSC-08-GENO-CapSeqAn). All samples were provided by J. Gongora, collected by himself or accessed through Dr Stewart Lowden, Dr Joeke Nijboer (Rotterdam Zoo), or through collaboration with institutions in Eurasia, Africa, and the Americas. Many thanks to Bertrand Bed'hom for

constructive feedback on the manuscript and for advice on the MHC locus, and the INRA @BRIDGe platform where the hybridisation capture experiments were performed.

References

Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, Fernandez-Banet J, Billis K, Garcia-Giron C, Hourlier T, Howe KL, Kahari AK, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ, Fernandez Banet J, Billis K, García Girón C, Hourlier T, Howe KL, Kähäri A, Kokocinski F, Martin FJ, Murphy DN, Nag R, Ruffier M, Schuster M, Tang YA, Vogel J-H, White S, Zadissa A, Flicek P, Searle SMJ (2016) The Ensembl Gene Annotation System. Database (Oxford) 2016:baw093. doi: 10.1093/database/baw093

Al Dahouk S, Nöckler K, Tomaso H, Splettstoesser WD, Jungersen G, Riber U, Petry T, Hoffmann D, Scholz HC, Hensel A, Neubauer H (2005) Seroprevalence of brucellosis, tularemia, and yersiniosis in wild boars (Sus scrofa) from north-eastern Germany. J Vet Med B Infect Dis Vet Public Health 52:444–55. doi: 10.1111/j.1439-0450.2005.00898.x

Alcaide M, Muñoz J, Martínez-de la Puente J, Soriguer R, Figuerola J (2014) Extraordinary MHC class II B diversity in a non-passerine, wild bird: the Eurasian Coot Fulica atra (Aves: Rallidae). Ecol Evol 4:688–98. doi: 10.1002/ece3.974

Allen RL, Hogan L (2013) Non-Classical MHC Class I Molecules (MHC-Ib). In: eLS. John Wiley & Sons, Ltd, Chichester, UK, pp 1–12

Amadou C (1999) Evolution of the Mhc class I region: The framework hypothesis. Immunogenetics 49:362–367. doi: 10.1007/s002510050507

Ando A, Chardon P (2006) Gene organization and polymorphism of the swine major histocompatibility complex. Anim Sci J 77:127–137. doi: 10.1111/j.1740-0929.2006.00331.x

Anzai T, Shiina T, Kimura N, Yanagiya K, Kohara S, Shigenari A, Yamagata T, Kulski JK, Naruse TK, Fujimori Y, Fukuzumi Y, Yamazaki M, Tashiro H, Iwamoto C, Umehara Y, Imanishi T, Meyer A, Ikeo K, Gojobori T, Bahram S, Inoko H (2003) Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. Proc Natl Acad Sci U S A 100:7708–7713. doi: 10.1073/pnas.1230533100

Ballingall KT, McKeever DJ (2005) Conservation of promoter, coding and intronic regions of the nonclassical MHC class II DYA gene suggests evolution under functional constraints. Anim Genet 36:237–239. doi: 10.1111/j.1365-2052.2005.01281.x

Barbisan F, Savio C, Bertorelle G, Patarnello T, Congiu L (2009) Duplication polymorphism at MHC class II DRB1 locus in the wild boar (Sus scrofa). Immunogenetics 61:145–151. doi: 10.1007/s00251-008-0339-6

Barrett LW, Fletcher S, Wilton SD (2013) Untranslated Gene Regions and Other Non-coding Elements. 1–56. doi: 10.1007/978-3-0348-0679-4_1

Borghans JAM, Beltman JB, De Boer RJ (2004) MHC polymorphism under host-pathogen coevolution. Immunogenetics 55:732–9. doi: 10.1007/s00251-003-0630-5

Brinton RD, Thompson RF, Foy MR, Baudry M, Wang J, Finch CE, Morgan TE, Pike CJ, Mack WJ, Stanczyk FZ, Nilsen J (2008) Progesterone receptors: Form and function in brain. Front Neuroendocrinol 29:313–339. doi: 10.1016/j.yfrne.2008.02.001

Buckley BA (2007) Comparative environmental genomics in non-model species: using heterologous hybridization to DNA-based microarrays. J Exp Biol 210:1602–1606. doi: 10.1242/jeb.002402

Burbano HA, Hodges E, Green RE, Briggs AW, Krause J, Meyer M, Good JM, Maricic T, Johnson PLF, Xuan Z, Rooks M, Bhattacharjee A, Brizuela L, Albert FW, de la Rasilla M, Fortea J, Rosas A, Lachmann M, Hannon GJ, Pääbo S (2010) Targeted investigation of the Neandertal genome by array-based sequence capture. Science 328:723–5. doi: 10.1126/science.1188046

Burri R, Hirzel HN, Salamin N, Roulin A, Fumagalli L (2008) Evolutionary patterns of MHC class II B in owls and their implications for the understanding of avian MHC evolution. Mol Biol Evol 25:1180–1191. doi: 10.1093/molbev/msn065

Carver EA, Stubbs L (1997) Zooming in on the Human–Mouse Comparative Map: Genome Conservation Re-examined on a High-Resolution Scale. Genome Res 7:1123–1137. doi: 10.1101/gr.7.12.1123

Castley ASL, Martinez OP (2012) Molecular Analysis of Complement Component C4 Gene Copy Number. In: Christiansen FT, Tait BD (eds) Immunogenetics. Humana Press, Totowa, NJ, pp 159–171

Chiovaro F, Chiquet-ehrismann R, Chiquet M (2015) Transcriptional regulation of tenascin genes.pdf. 9:1–2.

Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res 19:1179–82.

Costard S, Wieland B, Glanville W De, Jori F, Rowlands R, Vosloo W, Roger F, Pfeiffer DU, Dixon LK, Royal T, College V, Lane H, Al H, de Glanville W, Jori F, Rowlands R, Vosloo W, Roger F, Pfeiffer DU, Dixon LK (2009) African swine fever: how can global spread be prevented? Philos Trans R Soc Lond B Biol Sci 364:2683–2696. doi: 10.1098/rstb.2009.0098

Crew MD, Phanavanh B, Garcia-Borges CN (2004) Sequence and mRNA expression of nonclassical SLA class I genes SLA-7 and SLA-8. Immunogenetics 56:111–4. doi: 10.1007/s00251-004-0676-z

Cummings N, King R, Rickers A, Kaspi A, Lunke S, Haviv I, Jowett JBM (2010) Combining target enrichment with barcode multiplexing for high throughput SNP discovery. BMC Genomics 11:641. doi: 10.1186/1471-2164-11-641

Day WH, McMorris FR (1992) Critical comparison of consensus methods for molecular sequences. Nucleic Acids Res 20:1093–9.

Deakin JE, Papenfuss AT, Belov K, Cross JGR, Coggill P, Palmer S, Sims S, Speed TP, Beck S, Graves J a M (2006) Evolution and comparative analysis of the MHC Class III inflammatory region. BMC Genomics 7:281. doi: 10.1186/1471-2164-7-281

Dixon LK, Abrams CC, Bowick G, Goatley LC, Kay-Jackson PC, Chapman D, Liverani E, Nix R, Silk R, Zhang F (2004) African swine fever virus proteins involved in evading host defence systems. Vet Immunol Immunopathol 100:117–134. doi: 10.1016/j.vetimm.2004.04.002

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN (2006) Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223–7. doi: 10.1038/ng1710

Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J, Duce J, Neuroimaging D, Kauwe JSK, Ridge PG (2016) Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. BMC Bioinformatics. doi: 10.1186/s12859-016-1097-3

Emes RD, Goodstadt L, Winter EE, Ponting CP (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet 12:701–709. doi: 10.1093/hmg/ddg078

690 Essler SE, Ertl W, Deutsch J, Ruetgen BC, Groiss S, Stadler M, Wysoudil B, Gerner W, Ho C-S, Saalmueller A
691 (2013) Molecular characterization of swine leukocyte antigen gene diversity in purebred Pietrain pigs. Anim
692 Genet 44:202–5. doi: 10.1111/j.1365-2052.2012.02375.x
693 Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158:1227–34.
694 doi: 10.1016/s0378-1119(99)00294-2
695 Fowler ME (1996) Husbandry and diseases of captive wild swine and peccaries. Rev Sci Tech 15:141–54.
696 Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-Species Sequence Comparisons : A
697 Review of Methods and Available Resources. Genome Res 1–12. doi: 10.1101/gr.222003
698 Gieling ET, Nordquist RE, van der Staay FJ (2011) Assessing learning and memory in pigs. Anim Cogn 14:151–
699 173. doi: 10.1007/s10071-010-0364-3
700 Gongora J, Cuddahee RE, Nascimento FF Do, Palgrave CJ, Lowden S, Ho SYW, Simond D, Damayanti CS, White
701 DJ, Tay WT, Randi E, Klingel H, Rodrigues-Zarate CJ, Allen K, Moran C, Larson G (2011) Rethinking the
702 evolution of extant sub-Saharan African suids (Suidae, Artiodactyla). Zool Scr 40:327–335. doi:
703 10.1111/j.1463-6409.2011.00480.x
704 Gongora J, Fleming P, Spencer PBS, Mason R, Garkavenko O, Meyer JN, Droegemueller C, Lee JH, Moran C
705 (2004) Phylogenetic relationships of Australian and New Zealand feral pigs assessed by mitochondrial control
706 region sequence and nuclear GPIP genotype. Mol Phylogenet Evol 33:339–348. doi:
707 10.1016/j.ympev.2004.06.004
708 Gongora J, Groves C, Meijaard E (2017) Evolutionary relationships and taxonomy of Suidae and Tayassuidae. In:
709 Melleti M, Meijaard E (eds) Ecology, Conservation and Management of Wild Pigs and Peccaries. Cambridge
710 University Press, p 480
711 Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C,
712 Milan D, Megens H-J, Li S, Larkin DM, Kim H, Frantz LAF, Caccamo M, Ahn H, Aken BL, Anselmo A,
713 Anthon C, Auvil L, Badaoui B, Beattie CW, Bendixen C, Berman D, Blecha F, Blomberg J, Bolund L, Bosse
714 M, Botti S, Bujie Z, Bystrom M, Capitanu B, Carvalho-Silva D, Chardon P, Chen C, Cheng R, Choi S-H,
715 Chow W, Clark RC, Clee C, Crooijmans RPMA, Dawson HD, Dehais P, De Sapio F, Dibbits B, Drou N, Du
716 Z-Q, Eversole K, Fadista J, Fairley S, Faraut T, Faulkner GJ, Fowler KE, Fredholm M, Fritz E, Gilbert JGR,
717 Giuffra E, Gorodkin J, Griffin DK, Harrow JL, Hayward A, Howe K, Hu Z-L, Humphray SJ, Hunt T,
718 Hornshøj H, Jeon J-T, Jern P, Jones M, Jurka J, Kanamori H, Kapetanovic R, Kim J, Kim J-H, Kim K-W,
719 Kim T-H, Larson G, Lee K, Lee K-T, Leggett R, Lewin HA, Li Y, Liu W, Loveland JE, Lu Y, Lunney JK,
720 Ma J, Madsen O, Mann K, Matthews L, McLaren S, Morozumi T, Murtaugh MP, Narayan J, Truong Nguyen
721 D, Ni P, Oh S-J, Onteru S, Panitz F, Park E-W, Park H-S, Pascal G, Paudel Y, Perez-Enciso M, Ramirez-
722 Gonzalez R, Reecy JM, Rodriguez-Zas S, Rohrer GA, Rund L, Sang Y, Schachtschneider K, Schraiber JG,
723 Schwartz J, Scobie L, Scott C, Searle S, Servin B, Southey BR, Sperber G, Stadler P, Sweedler J V., Tafer H,
724 Thomsen B, Wali R, Wang J, Wang J, White S, Xu X, Yerle M, Zhang G, Zhang J, Zhang J, Zhao S, Rogers
725 J, Churcher C, Schook LB (2012) Analyses of pig genomes provide insight into porcine demography and
726 evolution. Nature 491:393–398. doi: 10.1038/nature11622
727 Harakalova M, Mokry M, Hrdlickova B, Renkens I, Duran K, van Roekel H, Lansu N, van Roosmalen M, de Bruijn
728 E, Nijman IJ, Kloosterman WP, Cuppen E (2011) Multiplexed array-based and in-solution genomic
729 enrichment for flexible and cost-effective targeted next-generation sequencing. Nat Protoc 6:1870–1886. doi:
730 10.1038/nprot.2011.396

731 Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial
732 DNA. J Mol Evol 22:160–74.
733 Ho CS, Lunney JK, Lee JH, Franzo-Romain MH, Martens GW, Rowland RRR, Smith DM (2010) Molecular
734 characterization of swine leucocyte antigen class II genes in outbred pig populations. Anim Genet 41:428–
735 432. doi: 10.1111/j.1365-2052.2010.02019.x
736 Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, Brizuela L, Richard McCombie W, Hannon GJ
737 (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel
738 sequencing. Nat Protoc 4:960–974. doi: 10.1038/nprot.2009.68
739 Hoppman-Chaney N, Peterson LM, Klee EW, Middha S, Courteau LK, Ferber MJ (2010) Evaluation of
740 oligonucleotide sequence capture arrays and comparison of next-generation sequencing platforms for use in
741 molecular diagnostics. Clin Chem 56:1297–1306. doi: 10.1373/clinchem.2010.145441
742 Hu R, Lemonnier G, Bourneuf E, Vincent-Naulleau S, Rogel-Gaillard C (2011) Transcription variants of SLA-7, a
743 swine non classical MHC class I gene. BMC Proc 5:S10. doi: 10.1186/1753-6561-5-S4-S10
744 Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev
745 Genet 32:415–35. doi: 10.1146/annurev.genet.32.1.415
746 Illumina (2009) Sequencing Analysis Software - User Guide. 1–176.
747 Irvine AD, Irwin McLean WH (2006) Breaking the (Un)Sound Barrier: Filaggrin Is a Major Gene for Atopic
748 Dermatitis. J Invest Dermatol 126:1200–1202. doi: 10.1038/sj.jid.5700365
749 Janova E, Matiasovic J, Vahala J, Vodicka R, Van Dyk E, Horin P (2009) Polymorphism and selection in the major
750 histocompatibility complex DRA and DQA genes in the family equidae. Immunogenetics 61:513–527. doi:
751 10.1007/s00251-009-0380-0
752 Jaratlerdsiri W, Isberg SR, Higgins DP, Gongora J (2012) MHC class I of saltwater crocodiles (Crocodylus
753 porosus): polymorphism and balancing selection. Immunogenetics 64:825–38. doi: 10.1007/s00251-012-
754 0637-x
755 Jaratlerdsiri W, Isberg SR, Higgins DP, Ho SYW, Salomonsen J, Skjodt K, Miles LG, Gongora J (2014) Evolution
756 of MHC class I in the Order Crocodylia. Immunogenetics 66:53–65. doi: 10.1007/s00251-013-0746-1
757 Johns TG, Bernard CCA (1999) The structure and function of myelin oligodendrocyte glycoprotein. J Neurochem
758 72:1–9. doi: 10.1046/j.1471-4159.1999.0720001.x
759 Joosten SA, Sullivan LC, Ottenhoff THM (2016) Characteristics of HLA-E Restricted T-Cell Responses and Their
760 Role in Infectious Diseases. J Immunol Res 2016:1–11. doi: 10.1155/2016/2695396
761 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in
762 performance and usability. Mol Biol Evol 30:772–780. doi: 10.1093/molbev/mst010
763 Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO (2006) Assessment of methods for amino acid
764 matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not
765 justified. BMC Evol Biol 6:29. doi: 10.1186/1471-2148-6-29
766 Kelley J, Walter L, Trowsdale J (2005) Comparative genomics of major histocompatibility complexes.
767 Immunogenetics 56:683–695. doi: 10.1007/s00251-004-0717-7
768 Kersten S, Gronemeyer H, Noy N (1997) The DNA binding pattern of the retinoid X receptor is regulated by
769 ligand-dependent modulation of its oligomeric state. J Biol Chem 272:12771–12777. doi:
770 10.1074/jbc.272.19.12771
771 Kirk LM, Ti SW, Bishop HI, Orozco-Llamas M, Pham M, Trimmer JS, Díaz E (2016) Distribution of the

772 SynDIG4/proline-rich transmembrane protein 1 in rat brain. J Comp Neurol 524:2266–2280. doi:
773 10.1002/cne.23945
774 Kloch A, Babik W, Bajer A, Siński E, Radwan J (2010) Effects of an MHC-DRB genotype and allele number on the
775 load of gut parasites in the bank vole Myodes glareolus. Mol Ecol 19 Suppl 1:255–65. doi: 10.1111/j.1365-
776 294X.2009.04476.x
777 Krishnan BR, Jamry I, Chaplin DD (1995) Feature mapping of the HLA class I region: localization of the POU5F1
778 and TCF19 genes. Genomics 30:53–8. doi: 10.1006/geno.1995.0008
779 Kulski JK (2004) Rhesus Macaque Class I Duplicon Structures, Organization, and Evolution Within the Alpha
780 Block of the Major Histocompatibility Complex. Mol Biol Evol 21:2079–2091. doi: 10.1093/molbev/msh216
781 Kulski JK, Shiina T, Anzai T, Kohara S, Inoko H (2002) Comparative genomic analysis of the MHC: the evolution
782 of class I duplication blocks, diversity and complexity from shark to man. Immunol Rev 190:95–122.
783 Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger
784 Datasets. Mol Biol Evol 33:1870–1874. doi: 10.1093/molbev/msw054
785 Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open
786 software for comparing large genomes. Genome Biol 5:R12. doi: 10.1186/gb-2004-5-2-r12
787 Kusza S, Flori L, Gao Y, Teillaud A, Hu R, Lemonnier G, Bosze Z, Bourneuf E, Vincent-Naulleau S, Rogel-
788 Gaillard C (2011) Transcription specificity of the class Ib genes SLA-6, SLA-7 and SLA-8 of the swine major
789 histocompatibility complex and comparison with class Ia genes. Anim Genet 42:510–20. doi: 10.1111/j.1365-
790 2052.2010.02170.x
791 Lê S, Josse J, Husson F (2008) FactoMineR : An R Package for Multivariate Analysis. J Stat Softw 25:1–18. doi:
792 10.18637/jss.v025.i01
793 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics
794 25:1754–60. doi: 10.1093/bioinformatics/btp324
795 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence
796 Alignment/Map format and SAMtools. Bioinformatics 25:2078–9. doi: 10.1093/bioinformatics/btp352
797 Loughner CL, Bruford EA, McAndrews MS, Delp EE, Swamynathan SSK, Swamynathan SSK (2016)
798 Organization, evolution and functions of the human and mouse Ly6/uPAR family genes. Hum Genomics
799 10:10. doi: 10.1186/s40246-016-0074-2
800 Luetkemeier ES, Malhi RS, Beever JE, Schook LB (2009) Diversification of porcine MHC class II genes: evidence
801 for selective advantage. Immunogenetics 61:119–29. doi: 10.1007/s00251-008-0348-5
802 Lunney JK, Ho CS, Wysocki M, Smith DM (2009) Molecular genetics of the swine major histocompatibility
803 complex, the SLA complex. Dev Comp Immunol 33:362–374. doi: 10.1016/j.dci.2008.07.002
804 Mallya M, Campbell RD, Aguado B (2006) Characterization of the five novel Ly-6 superfamily members encoded
805 in the MHC, and detection of cells expressing their potential ligands. Protein Sci 15:2244–2256. doi:
806 10.1110/ps.062242606
807 Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010)
808 Target-enrichment strategies for next- generation sequencing. Nat Methods 7:111–118. doi:
809 10.1038/NMETH.1419
810 Matsumoto M, Zhou Y, Matsuo S, Nakanishi H, Hirose K, Oura H, Arase S, Ishida-Yamamoto A, Bando Y, Izumi
811 K, Kiyonari H, Oshima N, Nakayama R, Matsushima A, Hirota F, Mouri Y, Kuroda N, Sano S, Chaplin DD
812 (2008) Targeted deletion of the murine corneodesmosin gene delineates its essential role in skin and hair

physiology. Proc Natl Acad Sci 105:6720–6724. doi: 10.1073/pnas.0709345105
Meijaard E, D'Huart J-P, Oliver W (2011) Family Suidae (Pigs). In: Wilson DE, Mittermeier RA (eds) Handbook of the Mammals of the World, Vol. 2. Lynx Edicions, Barcelona, Spain, pp 248–291
Meng XJ (2012) Emerging and Re-emerging Swine Viruses. Transbound Emerg Dis. doi: 10.1111/j.1865-1682.2011.01291.x
Meyer M, Kircher M (2010) Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc 2010:pdb.prot5448-prot5448. doi: 10.1101/pdb.prot5448
Mokry M, Feitsma H, Nijman IJ, de Bruijn E, van der Zaag PJ, Guryev V, Cuppen E (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res 38:e116. doi: 10.1093/nar/gkq072
Mooseker MS, Cheney RE (1995) Unconventional Myosins. Annu Rev Cell Dev Biol 11:633–675. doi: 10.1146/annurev.cb.11.110195.003221
Moutou K a, Koutsogiannouli EA, Stamatis C, Billinis C, Kalbe C, Scandura M, Mamuris Z (2013) Domestication does not narrow MHC diversity in Sus scrofa. Immunogenetics 65:195–209. doi: 10.1007/s00251-012-0671-8
Mukherjee S, Sarkar-Roy N, Wagener DK, Majumder PP (2009) Signatures of natural selection are not uniform across genes of innate immune system, but purifying selection is the dominant signature. Proc Natl Acad Sci 106:7073–7078. doi: 10.1073/pnas.0811357106
Na Ayudhya SN, Assavacheep P, Thanawongnuwech R (2012) One World - One Health: The Threat of Emerging Swine Diseases. An Asian Perspective. Transbound Emerg Dis 59:9–17. doi: 10.1111/j.1865-1682.2011.01309.x
Nagata T, Weiss EH, Abe K, Kitagawa K, Ando A, Yara-Kikuti Y, Seldin MF, Ozato K, Inoko H, Taketo M (1995) Physical mapping of the retinoid X receptor B gene in mouse and human. Immunogenetics 41:83–90.
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–26.
Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94:7799–7806. doi: 10.1073/pnas.94.15.7799
Nei M, Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: K. Tsuji, M. Aizawa, and T. Sasazuki E (ed) 11th histocompatibility workshop and conference. Oxford University Press, Oxford, pp 27–38
Nijman IJ, Mokry M, van Boxtel R, Toonen P, de Bruijn E, Cuppen E (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods 7:913–915. doi: 10.1038/nmeth.1516
Peñalba J V., Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, McGuire JA, Bowie RCK, Moritz C (2014) Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol Ecol Resour 14:1000–10. doi: 10.1111/1755-0998.12249
Penn DJ, Ilmonen P (2001) Major Histocompatibility Complex (MHC). In: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, Chichester, pp 1–7
Piertney SB, Oliver MK (2006) The evolutionary ecology of the major histocompatibility complex. Heredity (Edinb) 96:7–21. doi: 10.1038/sj.hdy.6800724
Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Raul D, Carter F (2011) Pseudogenes : Pseudo-functional or key regulators in health and disease? Rna 17:792–798. doi: 10.1261/rna.2658311.transcription
Piontkivska H, Rooney AP, Nei M (2002) Purifying Selection and Birth-and-death Evolution in the Histone H4

Gene Family. Mol Biol Evol 19:689–697. doi: 10.1093/oxfordjournals.molbev.a004127
Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi: 10.1093/bioinformatics/btq033
R Development Core Team (2008) Computational Many-Particle Physics. Springer Berlin Heidelberg, Berlin, Heidelberg
Renard C, Chardon P, Vaiman M (2003) The Phylogenetic History of the MHC Class I Gene Families in Pig, Including a Fossil Gene Predating Mammalian Radiation. J Mol Evol 57:420–434. doi: 10.1007/s00239-003-2491-9
Renard C, Hart E, Sehra H, Beasley H, Coggill P, Howe K, Harrow J, Gilbert J, Sims S, Rogers J, Ando A, Shigenari A, Shiina T, Inoko H, Chardon P, Beck S (2006) The genomic sequence and analysis of the swine major histocompatibility complex. Genomics 88:96–110. doi: 10.1016/j.ygeno.2006.01.004
Renard C, Vaiman M, Chiannilkulchai N, Cattolico L, Robert C, Chardon P (2001) Sequence of the pig major histocompatibility region containing the classical class I genes. Immunogenetics 53:490–500. doi: 10.1007/s002510100348
Rodgers JR, Cook RG (2005) MHC class Ib molecules bridge innate and acquired immunity. Nat Rev Immunol 5:459–471. doi: 10.1038/nri1635
Rodríguez-Fernández JL, Cabañas LG (2013) Chemoattraction: Basic Concepts and Role in the Immune Response. eLS. doi: 10.1002/9780470015902.a0000507.pub3
Rohland N, Reich D (2012) Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. Genome Res 22:939–946. doi: 10.1101/gr.128124.111
Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. doi: 10.1093/sysbio/sys029
Rosenbaum DM, Rasmussen SGF, Kobilka BK (2009) The structure and function of G-protein-coupled receptors. Nature 459:356–363. doi: 10.1038/nature08144
Sánchez-Torres C, Gómez-Puertas P, Gómez-Del-Moral M, Alonso F, Escribano JM, Ezquerra A, Domínguez J (2003) Expression of porcine CD163 on monocytes/macrophages correlates with permissiveness to African swine fever infection. Arch Virol 148:2307–2323. doi: 10.1007/s00705-003-0188-4
Schwartz S, Zhang Z, Frazer K a, Smit a, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10:577–586. doi: 10.1101/gr.10.4.577
Shigenari A, Ando A, Renard C, Chardon P, Shiina T, Kulski JK, Yasue H, Inoko H (2004) Nucleotide sequencing analysis of the swine 433-kb genomic segment located between the non-classical and classical SLA class I gene clusters. Immunogenetics 55:695–705. doi: 10.1007/s00251-003-0627-0
Simon M, Montézin M, Guerrin M, Durieux JJ, Serre G (1997) Characterization and purification of human corneodesmosin, an epidermal basic glycoprotein associated with corneocyte-specific modified desmosomes. J Biol Chem 272:31770–6.
Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool 2:16. doi: 10.1186/1742-9994-2-16
Spurgin LG, Richardson DS (2010) How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc Biol Sci 277:979–988. doi: 10.1098/rspb.2009.2084
Stam M, Hayes H, Bertaud M, Teillaud A, Lemonnier G, Rogel-Gaillard C (2008) Centromeric/pericentromeric

junction within the MHC locus on chromosome 7 in pig. In: XXXI Conference of the International Society for Animal Genetics, Amsterdam, Netherlands (2008).

Taber A, Altrichter M, Beck H, Gongora J (2011) Family Tayassuidae (Peccaries). In: Wilson DE, Mittermeier RA (eds) Handbook of the Mammals of the World, Vol. 2. Lynx Edicions, Barcelona, Spain, pp 292–307

Takahashi K, Rooney AP, Nei M (2000) Origins and divergence times of mammalian class II MHC gene clusters. J Hered 91:198–204.

Tanaka-Matsuda M, Ando A, Rogel-Gaillard C, Chardon P, Uenishi H (2009) Difference in number of loci of swine leukocyte antigen classical class I genes among haplotypes. Genomics 93:261–273. doi: 10.1016/j.ygeno.2008.10.004

Tennant LM, Renard C, Chardon P, Powell PP (2007) Regulation of porcine classical and nonclassical MHC class I expression. Immunogenetics 59:377–89. doi: 10.1007/s00251-007-0206-x

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. doi: 10.1038/nrg3117

van Oosterhout C (2009) A new theory of MHC evolution: beyond selection on the immune genes. Proc Biol Sci 276:657–665. doi: 10.1098/rspb.2008.1299

Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL (2007) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36:D753–D760. doi: 10.1093/nar/gkm987

Xia X (2017) DAMBE6: New Tools for Microbial Genomics, Phylogenetics, and Molecular Evolution. J Hered. doi: 10.1093/jhered/esx033

Yeager M, Hughes a L (1999) Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunol Rev 167:45–58. doi: 10.1111/j.1600-065X.1999.tb01381.x

Zozulya S, Echeverri F, Nguyen T (2001) The human olfactory receptor repertoire. Genome Biol 2:research0018.1–12. doi: 10.1186/gb-2001-2-6-research0018

Table 1 Details of suids and tayassuids used in this study including number and sample location

Family	Species name[a]	Species distribution	*N*[b]	Source(s)
Suidae	*Sus scrofa* (Wild boar)	Eurasia	18	Yorkshire Farm (UK)
	Sus barbatus (Bornean bearded pig)	Southeast Asia	10	Singapore Zoo (Singapore); Zoological Society of London Animal Hospital (UK)
	Sus cebifrons (Visayan warty pig)	Southeast Asia	4*	Rotterdam Zoo (Netherlands)
	Sus celebensis (Sulawesi warty pig)	Southeast Asia	10*	Sulawesi mainland and Buton Island (Indonesia)
	Babyrousa babyrussa (Babirusa)	Southeast Asia	13	Surabaya Zoo (Indonesia); Marwell Zoo and Chester Zoo (UK); Essen Zoo (Germany), Copenhagen Zoo (Denmark)
	Hylochoerus meinertzhageni (Forest hog)	Sub-Saharan Africa	3	Uganda
	Phacochoerus africanus (Common warthog)	Sub-Saharan Africa	10	Windhoek (Namibia); Rotterdam Zoo (Netherlands); Iwaba (Zimbabwe)
	Potamochoerus larvatus (Bush pig)	Sub-Saharan Africa	4*	Natal, South Africa (Zimbabwe)
	Potamochoerus porcus (Red river hog)	Sub-Saharan Africa	1	Duisburg Zoo (Germany)
Tayassuidae	*Pecari tajacu* (Collared peccary)	South America	19*	Matecaña City Zoo, Sante Fe Zoo, Barranquilla and Macagual (Colombia); Marwell Zoo and Edinburgh Zoo (UK)
	Tayassu pecari (White-lipped peccary)	South America	4*	La Lagartija (Colombia); Antwerp Zoo (Belgium)

[a] Common names of species are given in parentheses

[b] The number of specimens from each species including duplicates (indicated by *)

919

Table 2 The effect of species and family on capture parameters. The models included DNA-sequencing library as a covariate, animal as a random effect, and family and species as fixed effect. The means of all capture parameters are represented.

Species	**Total reads***	**% mapped***	**Specificity***	**E score***	**Coverage (X)***	**Duplicates (%)***	**C15 (%)***	**Enrichment***
Sus scrofa[a]	27, 379, 402	84.49	30.38	62.58	53.47	60.7	72.6	390.34
Sus barbatus	26, 667, 397	82.31	30.74	65.35	94.31	45.43	80.69	396.41
Sus cebifrons	37, 700, 490	76.98	32.79	68.33	157.54	37.73	86.34	436.23
Sus celebensis	27, 159, 226	80.61	32.3	66.26	107.2	41.81	81.02	423.92
Hylochoerus meinertzhageni	69, 813, 982	69.66	33.31	65.40	164.85	46.75	84.06	443.28
Phacochoerus africanus	37, 880, 504	75.98	33.81	64.50	102.78	48.58	79.12	456.60
Potamochoerus larvatus	25, 108, 272	76.08	32.55	62.66	88.29	43.97	75.14	429.48
Potamochoerus porcus	12, 352, 464	75.62	32.63	61.81	29.01	51.32	60.17	429.75
Babyrousa babyrussa	31, 017, 878	73.34	33.69	63.57	82.80	47.58	76.08	455.53
Pecari tajacu	18, 175, 378	12.24	19.33	34.24	3.82	6.76	5.55	216.85
Tajacu pecari	23, 457, 748	11.66	22.34	29.50	5.55	6.28	7.95	263.21
Family								
Suidae	30, 922, 354	78.92	31.69	63.57	87.73	49.42	77.6	415.33
Tayassuidae	19, 094, 051	12.14	19.85	33.41	4.12	6.67	5.97	224.91

* $P<0.005$ in capture parameters between species, genus and family (for all significant values of each library, genus and family, see EMS_7-9).

[a] The *Sus scrofa* shown here includes wild boar and feral pig samples, see EMS_1 for specific details of each sample.

920

Table 3 The average rates of synonymous (d_N) and nonsynonymous substitution rates (d_S), and their d_N/d_S ratios reported below for genes across the class I, II and III regions. Average estimates of standard error (S.E) are shown in brackets. The genes included in each category are as follows; all genes: include all genes within each class respectively; all protein coding genes: all known genes including novel transcripts and novel CDS except for pseusdogenes as indicated in Online Resource 6; non-protein coding genes: pseudogenes only; protein-coding histocompatibility genes: *SLA-1*, *SLA-2*, *SLA-3*, *SLA-6*, *SLA-7*, *SLA-8*; all histocompatibility genes: *SLA-1* to *SLA-9*, *SLA-11*; anchor genes: as indicated in Online Resource 6

		d_N (S.E)	**d_S (S.E)**	**d_N/d_S ratio**
Class I	All genes	0.01305 (0.00239)	0.03441 (0.00639)	0.47248
	All protein coding genes	0.01272 (0.00230)	0.03413 (0.00654)	0.47622
	Non-protein coding genes	0.01663 (0.00249)	0.03648 (0.00585)	0.51116
	Protein-coding histocompatibility genes	0.02800 (0.00377)	0.03435 (0.00628)	0.80166
	All histocompatibility genes	0.02742 (0.00375)	0.03212 (0.00634)	0.87507
	Anchor genes	0.00394 (0.00098)	0.03372 (0.00637)	0.96606
Class II	All genes	0.01487 (0.00315)	0.03606 (0.00719)	0.48172
	Coding genes	0.01270 (0.00258)	0.03760 (0.00700)	0.39976
	Non-coding genes	0.02043 (0.00481)	0.03170 (0.00816)	0.70655
	Protein-coding histocompatibility genes	0.01976 (0.00347)	0.04398 (0.00841)	0.45199
	All histocompatibility genes	0.01873 (0.00388)	0.03303 (0.00785)	0.63131
	Anchor genes	0.00456 (0.00111)	0.03957 (0.00541)	0.10402
Class III	All genes	0.01065 (0.00219)	0.04116 (0.00714)	0.31770
	Coding genes	0.01040 (0.00215)	0.04128 (0.00711)	0.31327
	Non-coding genes	0.02001 (0.00438)	0.03405 (0.00842)	0.58772
	Anchor genes	0.00607 (0.00112)	0.04352 (0.00546)	0.13541

921

Figure Captions

Fig. 1 Diagram of the swine (*S. scrofa*) major histocompatibility complex (MHC). Class I is located on the p-arm of chromosome 7, followed by class III and II which is separated by the centromere (grey box) on the q-arm. The length of each region according to Lunney et al. (2009) is indicated underneath the corresponding regions.

Fig. 2 Maximum likelihood tree of DNA sequences from eight nuclear (SINE, PRE-1, *P17*, *P207*, *P252*, *P408*, *GPIP* and the TNFα promoter) and 10 mitochondrial loci (cytb, 12S rRNA, 16S rRNA, ND1, ND2, tRNA- Leu, tRNA-Ile, tRNA-Gln, tRNA-Met and control region). The tree was produced in MEGA7 (Kumar et al. 2016) using data from Gongora et al. (2011). Bootstrap values are indicated on the respective branches. Coloured branches and symbols indicate the geographical region of Suidae and Tayassuidae species: South America (red branches; □); sub-Saharan (yellow branches; ○); Eurasia (green branches; ♦), south east Asia (Blue branches; ▲). *Hippopotamus amphibious* was used as an outgroup."

Fig. 3 Bayesian phylogenetic tree of the classical and non-classical MHC genes in wild species of suids and tayassuids. The dataset here represents 10 MHC class I genes (*SLA-1*, *SLA-2*, *SLA-3*, *SLA-4*, *SLA-5*, *SLA-6*, *SLA-7*, *SLA-8*, *SLA-9*, and *SLA-11*) in 11 species of suids and tayassuids, a total of 110 sequences. Abbreviations of species are as follows: *Sus scrofa* (*Susc*); *Sus barbatus* (*Suba*); *Sus cebifrons* (*Suce*); *Sus celebenis* (*Sucel*); *Hylochoerus meinertzhageni* (*Hyme*); *Phacochoerus africanus* (*Phaf*); *Potamochoerus lavartus* (*Pola*); *Potamochoerus porcus* (*Popo*); *Babyrousa babyrussa* (*Baba*); *Pecari tajacu* (*Peta*); and *Tayassu pecari* (*Tape*). The tree is rooted with *SLA-11* as the outgroup. Posterior probabilities (0–1) and likelihood bootstrap values (0–100%) for major branches are shown in boxes above the relevant branch (posterior probabilities/bootstrap values), the scale is indicated by the bar on the top left. Other posterior probabilities are indicated by colour; the scale is indicated by the bar on the top left.

Online Resource Captions

Online Resource 1 Voucher details of samples used in the study including the library, barcode within each library, project number, sample ID, species, natural distribution and sampling location

Online Resource 2 Schematic diagram describing the capture array design. **(a)** The blue bar represents the annotated sequence of the pig chromosome 7 outside the MHC locus, as retrieved from Ensembl (Sscrofa 10.2, (Groenen et al. 2012). The orange bar extending from base 24,614,801 to 29,807,435 corresponds to the MHC locus. This region was removed from the Ensembl sequence and then replaced with the VEGA chromosome 7-LW sequence (http://vega.sanger.ac.uk/Sus_scrofa/Info/Index). The chromosome 7-LW sequence only includes the MHC (5,406,156 bp) and was obtained by Renard et al. (2006). The quality of its assembly and annotation is considered of higher quality respect to the standard Ensembl version. **(b)** The pig MHC locus is made up by two regions (green and salmon coloured bars) separated by the centromeric region (in grey). The assembly obtained by Renard et al. (2006) enabled sequencing of 1,826,329 bp onto the p-arm and 579,827 bp, onto the q-arm, corresponding to a total 2,406,156 bp (green bars). The centromere, spanning roughly 3 Mb, was not used for the design (grey bar). A further contig produced by Stam et al. (2008) (Accession number: MF029693) was then merged to 7-LW to cover a further 394, 857 bp towards the centromere (salmon coloured bar) as a result of a small overlapping region onto the q-arm. This produced a target region of 2,801,013 bp. The NimbleGen probes covered 2,003,926 bp (~72% of the region). The merging and alignment among the different sequences was performed using MUMmer 3.0 (Kurtz et al. 2004) and masked using BEDtools 2.13.1 (Quinlan and Hall 2010) to avoid redundancy

Online Resource 3 Details on the species and number (*n*) for each library used in the array. Each library contains a total of 12 samples

Online Resource 5 Details on the extraction of coding sequences for downstream analyses

Online Resource 6 Details of genes retrieved and not retrieved (denoted with ^), including details for the alignments used for phylogenetic analyses. Information on the coding sequence (CDS) length is shown as base pairs (bp). Genes within the dotted lines indicate the framework anchor genes, and histocompatibility genes are denoted in bold. For pseudogenes, the whole exonic region was retrieved. The number of exons retrieved and the protein length (amino acids) for each alignment is shown, with figures in brackets indicating the amino acid length in the 10.2 annotation of the *S. scrofa* genome. Readings frames with stop codons in all species are indicated by the asterisk (*) except otherwise stated and the residue number in brackets. Locus type abbreviations are as follows: Known (K); Pseudogene (P); Novel Transcript (NT); Novel CDS (NCDS); Putative (PU). Known genes are identified to be functional by Renard et al. (2006)

Online Resource 7 The effect of DNA-sequencing on capture parameters. L indicates the library and N indicates the number of samples for each **(a)** library **(b)** genus and **(c)** family. The model included DNA-sequencing library as a covariate, genus as fixed effect, and animal as a random effect

Online Resource 8 Hierarchical clustering of **(a)** percentage of reads mapped by species **(b)** coverage (X) by species and **(c)** specificity (%) by species. The hierarchical cluster analysis was performed in FactoMineR package (Lê et al. 2008) using 'hclust' function with '1-cor (x) ' as distance and 'ward' as aggregation criterion. Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities. Each major cluster is indicated by different coloured branches and the geographical region of each species is categorised by the coloured box to the left: South America (red); sub-Saharan (yellow); Eurasia (Green), south east Asia (Blue)

Online Resource 9 Plot of the MHC region and the average coverage (X) of each species. In **(a)** The x-axis indicates the coverage and the y-axis indicate the position relative to the size of the MHC region. The centromere is indicated by the black arrow. Species and average coverage are as shown on the right (to the nearest integer), regions with high duplicates are indicated by a star (~890,000–896,000 bp, 4,607,500–4,615,000 bp and 4,720,000–4,807,000 bp) and regions with gaps consistent between species (over 30bp) are indicated by the black boxes numbered 1-4. These indicate the regions 1: ~327,000-497,006 bp; 2: ~4,620,000-4,720,000 bp; 3: ~4,805,000-4,845,000 bp; and 4: ~4,900,000-5,270,000 bp. Plots in **(b)** show an example of the coverage in one gene from class Ia (*SLA-1*), Ib (*SLA-6*) and class II (*SLA-DQB1*). The level of coverage (X) is shown on the x-axis, the y-axis indicates the length of the corresponding gene, the blue bar indicates regions covered by the probes, exon 2 (yellow boxes) and exon 3 (dark grey boxes) are indicated. Plots were produced using (R Development Core Team 2008)

Online Resource 11 Phylogenetic analyses of the histocompatibility genes, the classical and non-classical SLA class I genes ordered by their position in the genome **(a)** *SLA-1*; **(b)** *SLA-5*; **(c)** *SLA-9*; **(d)** *SLA-3*; **(e)** *SLA-2*; **(f)** *SLA-4*; **(g)** *SLA-11*; **(h)** *SLA-8*; **(i)** *SLA-7*; **(j)** *SLA-6*. Trees were produced by Maximum Likelihood in MEGA v7 (Kumar et al. 2016) based on the model-of-best-fit and a gamma category of 6 to account for evolutionary rate differences among sites. The branch lengths measured in the number of substitutions per site, positions with less than 95% site coverage were eliminated (>5% alignment gaps). Gaps were partially deleted. Different coloured branches and symbols indicate the geographical region of each species: South America (red branches; □); sub-Saharan (yellow branches; ○); Eurasia (green branches; ♦), south east Asia (Blue branches; ▲)

Online Resource 12 Phylogenetic analyses of the SLA class II genes ordered by their position in the **(a)** *SLA-DRA*; **(b)** *SLA-DRB1*; **(c)** *SLA-DQA*; **(d)** *SLA-DQB1*. Trees were produced by Maximum Likelihood in MEGA v7 (Kumar et al. 2016) based on the model-of-best-fit and a gamma category of 6 to account for evolutionary rate differences among sites. The branch lengths measured in the number of substitutions per site, positions with less than 95% site coverage were eliminated (>5% alignment gaps). Gaps were partially deleted. Different coloured branches and symbols indicate the geographical region of each species: South America (red branches; □); sub-Saharan (yellow branches; ○); Eurasia (green branches; ♦), south east Asia (Blue branches; ▲)

Online Resource 13 Phylogenetic analyses of the anchor genes ordered by their position in the genome as follows: class I **(a)** *MOG*; **(b)** *GNL1*; **(c)** *CCHCR1*; **(d)** *TCF19*; **(e)** *POU5F1*; class III **(f)** *BAT1*; **(g)** *NFKBIL1*; **(h)** *TNF*; **(i)** *LTB*; **(j)** *BAT2;* **(k)** *VARS2;* **(l)** *C4A;* **(m)** *CYP21A2;* **(n)** *PBX2;* class II **(o)** *COL11A2;* **(p)** *RXRB;* **(q)** *SLC39A7;*

(r) *HSD17B8;* **(s)** *RING1*. Trees were produced by Maximum Likelihood in MEGA v7 (Kumar et al. 2016) based on the model-of-best-fit and a gamma category of 6 to account for evolutionary rate differences among sites. The branch lengths measured in the number of substitutions per site, positions with less than 95% site coverage were eliminated (>5% alignment gaps). Different coloured branches and symbols indicate the geographical region of each species: South America (red branches; □); sub-Saharan (yellow branches; ○); Eurasia (green branches; ♦), south east Asia (Blue branches; ▲)

Online Resource 14 Graphs showing the synonymous (d_S) and nonsynonymous (d_N) substitution rates of **(a)** class I **(b)** class III and **(c)** class II genes in the MHC region. The left x-axis indicates the mean substitution rates and the right x-axis show the d_N/d_S ratios. Synonymous substitution rates are indicated in black striped columns, nonsynonymous substitution rates in solid black columns and the d_N/d_S ratios by yellow dots. Significant *P*-values for rejecting the null hypothesis (neutral selection, $d_N=d_S$) in favour of positive (green dots) or negative selection (red dots) are shown above the respective genes. Black triangles indicate the anchor genes and white triangles indicate the histocompatibility genes. All values were estimated in MEGA v7 (Kumar et al. 2016) using the modified Nei & Gojobori (1986) method with Jukes-Cantor correction to account for multiple substitutions at the same site. Gaps were treated with pairwise deletion. Genes are shown in order according to their position in the genome

Online Resource 15 Graph showing the synonymous (d_S) and nonsynonymous (d_N) substitution ratios of the histocompatibility genes in class Ib. These are shown as regions in the peptide binding region (PBR) encoded for by exon 2 and 3, and the non-PBR region. The left x-axis shows the d_N substitutions (solid black columns) and the d_S substitutions (striped column), and the right x-axis shows the d_S/d_N ratio (dots). Significant *P*-values for rejecting the null hypothesis (neutral selection, $d_N=d_S$) in favour of positive (green dots) or negative selection (red dots) are shown above the respective genes. All values were estimated in MEGA v7 (Kumar et al. 2016) using the modified Nei & Gojobori (1986) method with Jukes-Cantor correction to account for multiple substitutions at the same site. Gaps were treated with pairwise deletion