Database, Vol. 2013, Article ID bat011, doi:10.1093/database/bat011

Original article

Sequencing and comparative analysis of the gorilla MHC genomic sequence

Laurens G. Wilming[1,*], Elizabeth A. Hart[1], Penny C. Coggill[1], Roger Horton[1], James G. R. Gilbert[1], Chris Clee[1], Matt Jones[1], Christine Lloyd[1], Sophie Palmer[1], Sarah Sims[1], Siobhan Whitehead[1], David Wiley[1], Stephan Beck[2] and Jennifer L. Harrow[1]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1HH, UK and [2]UCL Cancer Institute, University College London, London WC1E 6BT, UK

***Corresponding author:** Tel: 01223496843; Fax: 01223496802; Email: lw2@sanger.ac.uk

Submitted 30 November 2012; Revised 1 February 2013; Accepted 19 February 2013

Citation details: Wilming,L.G., Hart,E.A., Coggill,P.C., *et al*. Sequencing and comparative analysis of the gorilla MHC genomic sequence. *Database* (2013) Vol. 2013: article ID bat011; doi:10.1093/database/bat011

Major histocompatibility complex (MHC) genes play a critical role in vertebrate immune response and because the MHC is linked to a significant number of auto-immune and other diseases it is of great medical interest. Here we describe the clone-based sequencing and subsequent annotation of the MHC region of the gorilla genome. Because the MHC is subject to extensive variation, both structural and sequence-wise, it is not readily amenable to study in whole genome shotgun sequence such as the recently published gorilla genome. The variation of the MHC also makes it of evolutionary interest and therefore we analyse the sequence in the context of human and chimpanzee. In our comparisons with human and re-annotated chimpanzee MHC sequence we find that gorilla has a trimodular RCCX cluster, versus the reference human bimodular cluster, and additional copies of Class I (pseudo)genes between *Gogo-K* and *Gogo-A* (the orthologues of *HLA-K* and *-A*). We also find that *Gogo-H* (and *Patr-H*) is coding versus the *HLA-H* pseudogene and, conversely, there is a *Gogo-DQB2* pseudogene versus the *HLA-DQB2* coding gene. Our analysis, which is freely available through the VEGA genome browser, provides the research community with a comprehensive dataset for comparative and evolutionary research of the MHC.

Introduction

The major histocompatibility complex, MHC, is the multi-gene region of the genome crucial for the vertebrate immune response. It is the most gene-dense and polymorphic region of the mammalian genome (1, 2). In conjunction with those of the leucocyte receptor complex, its cell surface–expressed products are responsible for transplant effectiveness, reproductive success, autoimmunity and much of the resistance to infectious diseases. Recent studies on eight human haplotypes have provided a detailed picture of the haplotype structure and gene polymorphism of this region (3, 4). In addition, the Wellcome Trust Case Control Consortium genome-wide association study for seven common diseases (5) has found the highest associations for two autoimmune diseases—type I diabetes and rheumatoid arthritis—to be with the MHC. These combined findings offer, for the first time, a secure baseline from which to start evaluating the significance of the evolutionary development of the MHC as a whole.

The MHC region is divided into sub-regions according to the type or function of the genes they contain: extended class I, classical class I, classical class III, classical class II and extended class II (6). The extended class I region contains, for example, histone and olfactory receptor genes, the classical class III region contains the complement factor genes, and the primary role of the classical MHC class I and II genes is to encode key receptor molecules that recognise foreign peptides and present them to specialist immune cells to initiate an immune response. Each MHC molecule comprises a non-variable immunoglobulin 'stalk' that anchors the molecule to the cell surface and a basket-like receptor that forms the peptide-binding region. It is these peptide-binding regions that are highly polymorphic in all primates (7). Comparing the polymorphism in other great apes with that in humans may indicate how evolution is being driven. Analysis of rhesus macaque MHC genes has informed the study of the region as a whole, but the human genome, in general, is more similar to the chimpanzee genome, with which it shares 98.8% nucleotide and >99% amino acid identity (8). However, *de novo* or species-specific polymorphisms were implied by diversity centred largely on the *HLA-A* and *HLA-B/C* clusters (9). There are also a few specific gene regions where the similarity to gorilla is higher (10), so the study of the gorilla as an out-group between human and chimpanzee will add another layer to the understanding of the evolutionary changes that have occurred in humans in their ability to combat infection and mount appropriate immune responses.

In the gorilla, the MHC cluster is located on the equivalent to human chromosome 6 (where the human MHC gene cluster is located); depending on the nomenclature used, this is chromosome 5 (when numbering the chromosomes 1–23, where chromosomes 11 and 12 are equivalent to q and p arms of human chromosome 2, respectively) or chromosome 6 (when numbering the human chromosome 2 equivalents as 2p and 2q or 2a and 2b) (11–13). The bacterial artificial chromosome (BAC) library containing the region was made from a heterozygous gorilla and we have attempted to give contiguous single haplotype coverage of the whole 4-Mb region that includes MHC class I, III and II. But it is clear that, at least across the class II—*DRB*—region, the two haplotypes present are distinct. Here we provide the first complete reference sequence and gene map of the full MHC region in the gorilla. As this sequence is based on contiguous BAC clones, as opposed to whole-genome shotgun assembly used for the genome of female gorilla 'Kamilah' (14), we are able to provide comprehensive annotation of this region. We make direct comparisons with the reference human genome and the chimpanzee MHC sequence (9), both of which have been annotated by the HAVANA group and published in the Vertebrate Genome Annotation (VEGA) genome browser (15). Similarities and differences between orthologous genes in other primates are discussed.

Table 1. The clone names, versioned accession numbers and variation between the BACs that make up the MHC of gorilla 'Frank'

Clone name	ENA accession	Number of			Overlap length (bp)
		SNPs	Deletions	Insertions	
CH255-522B23	CU104671.1	0	0	0	2000
CH255-127C23	CU104653.1	0	0	0	2000
CH255-37P17	CU104661.1	0	0	0	2000
CH255-179G12	CT025620.2	0	0	0	2000
CH255-451A14	CU104667.1	0	0	0	2000
CH255-405H12	CU104665.1	0	0	0	2000
CH255-259J17	CU104658.1	0	0	0	2000
CH255-39I5	CU104664.1	0	0	0	2000
CH255-83O18	CU104675.1	0	0	0	2000
CH255-289P22	CU104659.1	0	0	0	2000
CH255-201E9	CU104656.1	0	0	0	2000
CH255-478L19	CU104669.1	0	0	0	2000
CH255-48E14	CU104670.1	0	0	0	65339
CH255-386C2	CU104662.1	0	0	0	2000
CH255-375N4	CU104660.1	0	0	0	37060
CH255-13G2	CU104654.2	8	4	3	24914
CH255-415I16	CU104666.2	98	20	22	103353
CH255-559J12	CU104673.1	0	0	0	2000
CH255-397I3	CU104663.1	0	0	0	2000
CH255-469C9	CU104668.1	0	0	0	2635
CH255-56N15	CU104674.1	4	2	1	32416
CH255-114D6	CU104652.1	0	0	0	2000
CH255-58L21	CU104676.1	0	0	0	2000
CH255-351B13	CT025711.1	0	0	0	29819
CH255-354J20	CT025621.2	0	0	0	2000
CH255-336G22	CT025558.1	0	0	0	2000
CH255-191J6	CU104655.1	0	0	0	2000
CH255-206J13	CU104657.1	101	14	16	59844
CH255-529K7	CU104672.1				

Variation shown is between the clone listed on that row and the next, in the length of the overlap between the two. Clones are listed in the order of contiguous overlap.
SNPs = single nucleotide polymorphisms.

Results and discussion

Gorilla MHC gene map

The 29 BAC clones that comprise the gorilla MHC span a region of 4.64-Mb and are listed in Table 1, together with the variations found within the 2000 base-pair or longer overlaps between clones. A map of the entire region is shown in Figure 1. Within this contiguous region, 305 gene loci have been identified and annotated: 155 of these are known genes that are predicted to be functional, 5 loci are classified as novel CDS (see 'Material and Methods: Sequence annotation' section for definitions), 15 are known transcripts, 5 are novel transcripts and 20 are putative transcripts. The remaining 105 loci have been classified as pseudogenes: 44 as processed pseudogenes, 59 as unprocessed pseudogenes, 1 as a polymorphic pseudogene and 1 as a transcribed pseudogene. The full annotation is available online at the VEGA database

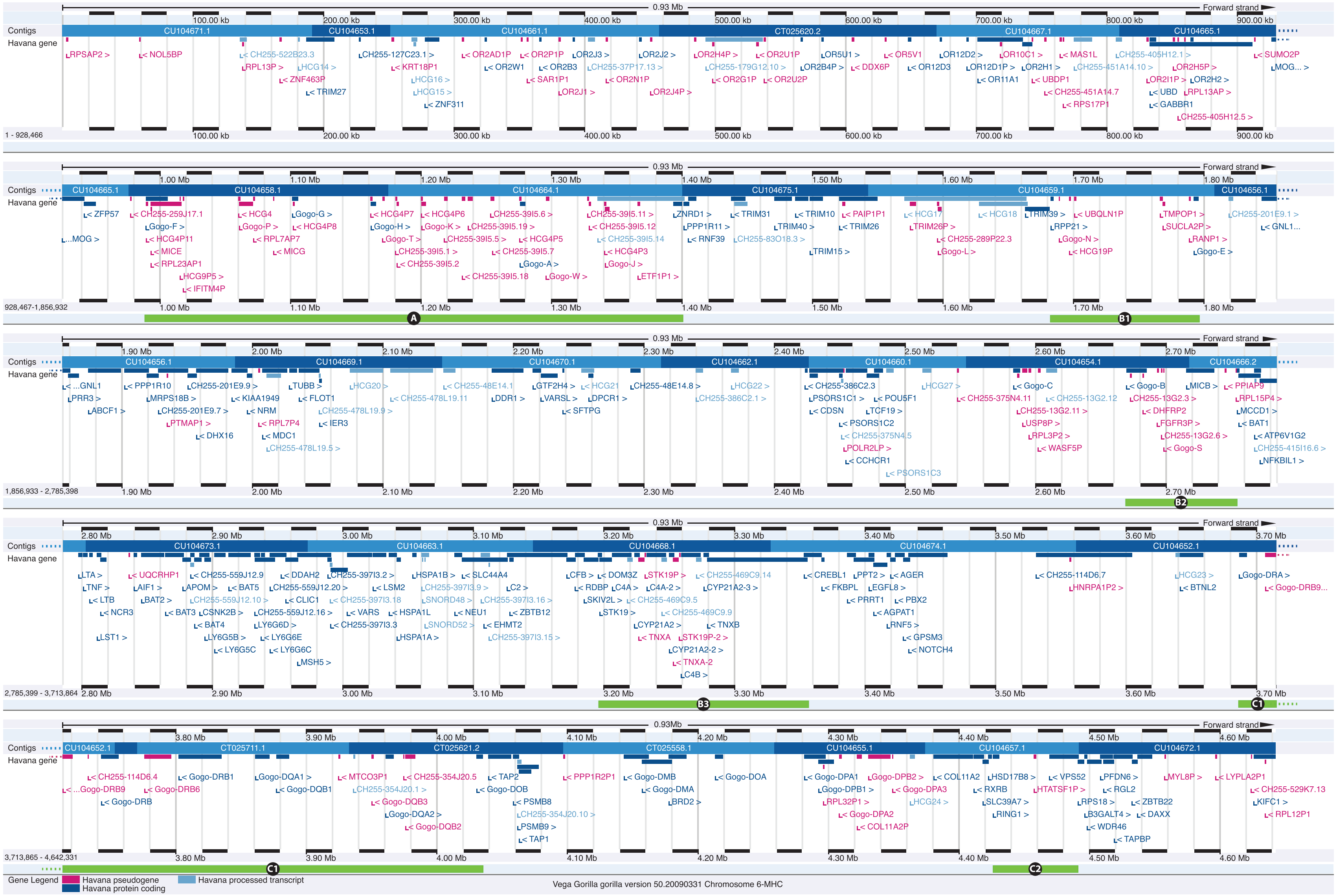

Figure 1. Feature map of the gorilla MHC, modified from the VEGA browser (release 50, December 2012). Each locus is labelled with a name, coloured according to type (see legend at bottom) and with indication of orientation (angle bracket before or after the name) and position within the region. The tiling path of the sequenced BACs is shown at the top of each panel (labelled contigs), with clones in alternating dark and light blue and, space permitting, with accession numbers. At the top and bottom of each panel, a size scale is shown. The regions highlighted in Figure 2 are marked with green bars at the bottom of a panel and labelled with the figure section identifier.

(15, 16). It should be noted that virtually all gorilla transcript models have been built on the basis of homology support by human transcripts or proteins. Notable exceptions are the following transcripts, which are supported by native locus-specific mRNAs: variant 001 of *TAP1* (*TAP1*-001), *TAP2*-001, *Gogo-A*-006, *Gogo-C*-001 and *Gogo-DMB*-001.

Comparative analysis

Overall, a high level of conserved synteny is observed between the entire gorilla MHC and the human equivalent as annotated for the PGF cell line (which is the MHC reference haplotype). There are, however, regions where notable differences between the human and gorilla MHC have been identified. These regions are highlighted in Figure 2, which will be referred to during the discussion that follows. In addition to the conservation of functional loci across the region, many processed pseudogenes are also conserved between gorilla and human.

Extended class I. There are very few differences between gorilla and human in the extended class I region. Moving along the chromosome from the telomeric end of the p arm towards the centromere, we find two olfactory receptor loci (*OR5V1* and *OR10C1*) within the olfactory receptor gene cluster that have been classified as coding loci in human but appear to be pseudogenes in gorilla. Conversely, there are two olfactory loci (*OR2B4P* and *OR12D1P*) that are classified as pseudogenes in human but have the potential to encode a protein in gorilla (*OR2B4* and *OR12D1*). The total number of olfactory receptor loci within this region, 25, is the same in human and gorilla. This is in broad agreement with the findings of others that the deterioration of the olfactory repertoire of Old World monkeys compared with mouse, New World monkeys (except the howler monkey) and the lemur (a prosimian) has occurred concomitant with the acquisition of full trichromatic colour vision in primates (17). The human olfactory region has evolved by copy number variation through duplication and deletion, and other findings revealed that all deletion alleles were human derived when compared with the chimpanzee reference genome (18). More recently, gorillas have been found to have a functioning olfactory sense that they use to investigate their environment. Thus, olfaction may not be as irrelevant in great apes as has been suggested (19).

Near the boundary with the classical class I region, the human genome carries the single-exon *MAS1LP* pseudogene as well as *MAS1L*, a coding gene for a Mas-related multi-pass membrane G-protein–coupled receptor. In gorilla, where *MAS1LP* is present too, the *MAS1L* gene encodes, at best, an N-terminally truncated peptide (317 versus 378-amino acids), or is, at worst, a pseudogene. The truncation is caused by a single nucleotide insertion in the gorilla genome in codon 6, just downstream of the equivalent of the human ATG start codon. The next available ATG is 61 codons downstream, with two further ATGs 2 and 4 codons from there; the latter has good Kozak sequence, on par with the upstream human ATG.

Class I—classical: HLA genes and orthologues. The class I genes in gorilla show a high level of conserved synteny with human. The notable differences are confined mainly to small sub-regions, as illustrated in Figure 2A, and will be highlighted as they are discussed.

As expected from the literature, the three classical class I gorilla orthologues of the human *HLA-A*, *-B* and *-C* genes—*Gogo-A*, *-B* and *-C*—are present and appear to be expressed. In our analysis of the chimpanzee BAC sequences from Anzai *et al*., we confirm that they are expressed there as well (20–22).

Class I—non-classical: HLA genes and orthologues. The human *HLA-E*, *-F* and *-G* genes have protein-coding gorilla and chimpanzee counterparts. The remaining *HLA-A* paralogues in human are unprocessed pseudogenes, often polymorphic and in varying states of degradation: from full length with eight exons in *HLA-H*, *-K* and *-J* to only single exons in *-N* and *-U* (3). Interestingly, in contrast to human, the gorilla and chimpanzee orthologues of *HLA-H*—*Gogo-H* and *Patr-H*—have the potential to encode a full-length 362-amino acid peptide (Figure 2A). *HLA-H* is a pseudogene in the human reference genome (haplotype PGF) and in haplotypes APD, COX, DBB, MANN, MCF, QBL and SSTO. Pseudogenisation is caused by different frameshift-causing single nucleotide polymorphisms in different haplotypes: in PGF, it is an insertion just after the splice acceptor site of exon 4, and in the other haplotypes, it is a deletion in the middle of exon 4. The pattern of variation within human *HLA-H* is consistent with it being in neutral linkage with the neighbouring *HLA-A* paralogues as a result of balancing selection acting on the neighbouring functional *HLA* loci (23). A *Gogo-H*-specific transcript has been identified in gorilla, suggesting this locus is indeed transcribed, translated and functional. This gene has been found by others to be expressed in both chimpanzee and gorilla (7, 24). However, Anzai *et al*. (22) found it to be a pseudogene in chimpanzee, so the locus appears to be haplotypic in chimpanzees. The human orthologue of *Gogo-P*—*HLA-P*—has been defined as a transcribed pseudogene in the *HLA* family, but as yet, there is no evidence that it is transcribed in gorilla. The orthologue has been lost in rhesus macaque (25), but we annotated an orthologue in chimpanzee along with *Patr-T* and *Patr-W* (26, 27), as indicated in Figure 2A.

In gorilla, the region between *Gogo-K* and *Gogo-A* is of interest in that it is 72.6-kb in length, some six times the 12-kb distance between *HLA-K* and *HLA-A* (see Figure 2A). This

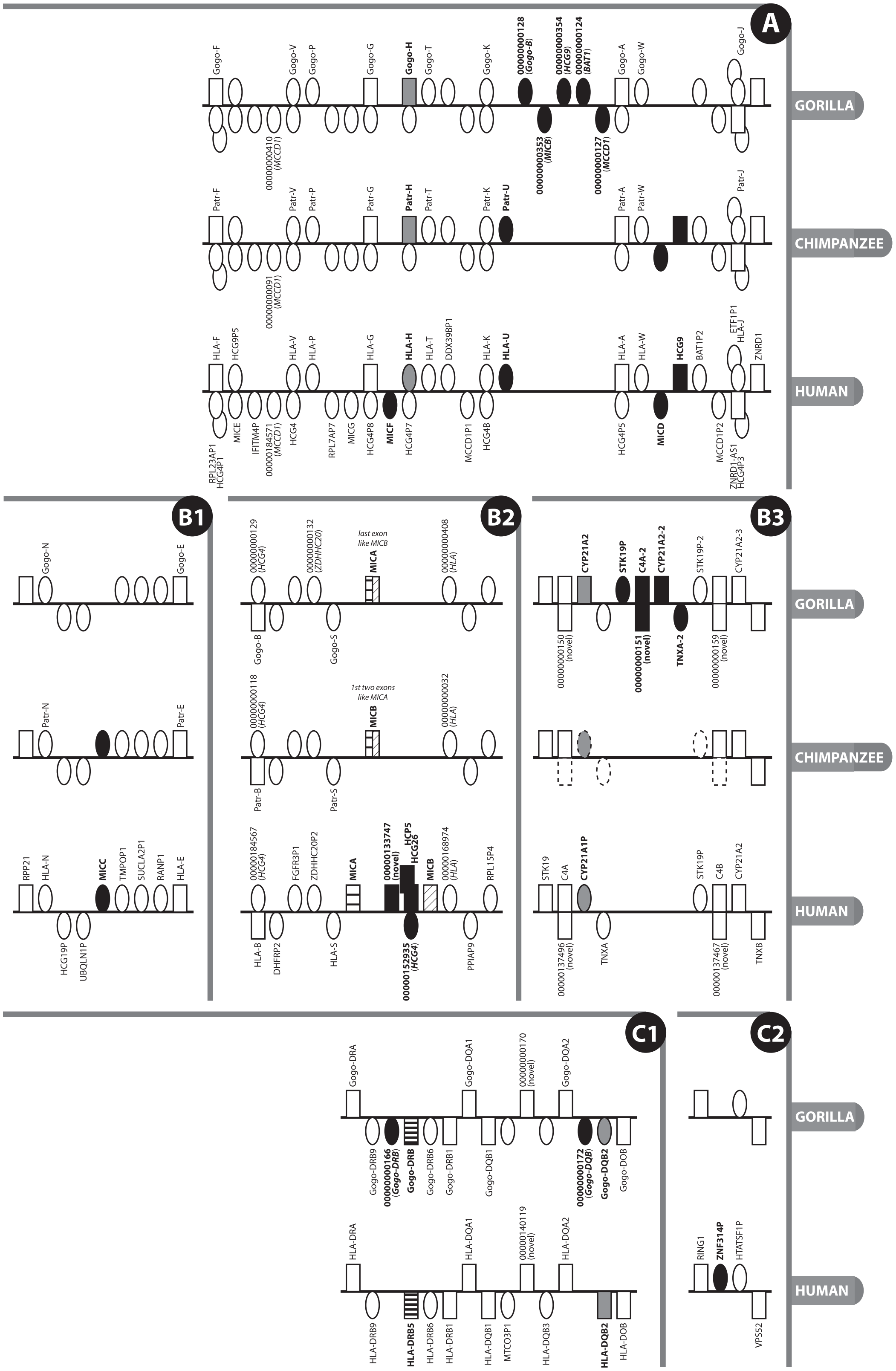

Figure 2. Detailed view of regions of the MHC where there is a difference in gene content or type between gorilla, human reference and chimpanzee. Figure is not to scale. Rectangle = gene; oval = pseudogene; grey fill = type difference (pseudogene versus gene); black fill = gene absent/present in at least one species and not another; black and white striped = not direct

(continued)

human, and chimpanzee, 12-kb interval carries just the *HLA-A* paralogue *HLA-U*, an unprocessed pseudogene. In gorilla, this region contains five unprocessed pseudogenes—*CH255-39I5.5* (OTTGORG00000000128), *CH255-39I5.18* (OTTGORG00000000353), *CH255-39I5.19* (OTTGORG00000000354), *CH255-39I5.6* (OTTGORG00000000124) and *CH255-39I5.7* (OTTGORG00000000127)—that appear to be remnants of *Gogo-B*, *MICB*, *HCG9*, *BAT1* and *MCCD1*, respectively; gorilla lacks an *HLA-U* orthologue. A *MIC* pseudogene known as *MICD* and a non-coding locus known as *HCG9*, which lie between *HLA-W* and *HLA-J* in human, are also absent from the corresponding location in gorilla. Interestingly, what appears to be a remnant of *HCG9* (pseudogene CH255-39I5.19 or OTTGORG00000000354) is just discernible in gorilla in an upstream location close to *MIC* pseudogene CH255-39I5.18 between *Gogo-K* and *Gogo-A*. A dot plot of human versus gorilla genomic sequence from *HLA-F* to *HLA-J* (Supplementary Figure S1) suggests that stretches of sequence within this region have been subject to duplication and re-arrangement after the divergence of gorilla and human. The 48-kb distance between *HLA-W* and *HLA-J* in human and chimpanzee is ∼5-kb longer than the corresponding region in gorilla. Despite these differences, overall, the gorilla and human exhibit a high degree of similarity. This is unlike the situation in rhesus macaque where a large expansion of class I genes has occurred between *Mamu-G* and *Mamu-J*, the orthologues of *HLA-G* and *HLA-J* (28).

Class I: MIC. The human lineages studied to date carry two functional *MIC* genes—*MICA* and *MICB* (29)—and a number of unprocessed pseudogenes (*MICC, D, E, F* and *G*) (25, 30). *MICA* and *MICB* are thought to have diverged from each other 33-44 million years ago (31), before the divergence of chimpanzee and human. We have found a major difference between human and gorilla in the apparent fusion of the *MICA* and *MICB* genes into what we have called *MICA*, similar to what has been observed in chimpanzee (32) (see Figure 2B2). This single hybrid *MICA–MICB* gene appears to have originated as a result of a 95-kb deletion in the region compared with human (22, 33), and shows only limited polymorphism between different individuals (32). The fact that the gorilla coding exons 1-5 show greater sequence similarity to human *MICA* compared with the final 3′ coding exon, which shows greater similarity to *MICB*. In chimpanzee, it is a slightly different fusion: the first two exons appear to be derived from *MICA* and the remaining exons from *MICB*, supports the theory of a deletion between and through the genes, as opposed to deletion of a single gene. This apparent fusion of *MICA* and *MICB* in gorilla and chimpanzee is accompanied by the loss of orthologues of the intervening genes (*HCP5*, an *HCG4* pseudogene and *HCG26*). *MIC* pseudogene *MICC*, between *UBQLN1P* and *TMPOP1*, is also lost in gorilla, but it is present in chimpanzee (Figure 2B1). This region is clearly subject to both large and small deletions and subject to polymorphism as well. Ando *et al*. (34) found no linkage disequilibrium between *MICB* and *MICA*, but the human *MICA* and *MICB* genes exhibit extensive polymorphism (35) as evidenced by the fact that at the time of writing, the IMGT/HLA database (36) lists 84 *MICA* and 40 *MICB* alleles. Furthermore, human alleles DRB1*03 and DRB1*07 present with a polymorphism characterised by a 4-bp insertion in *MICB*, resulting in an extended and altered open reading frame (ORF) at the 3′ end (6). There are instances of Northeast Asian and Native American individuals with the *HLA-B*4801* allele haplotype who have lost the MICA locus owing to a large 100-kb genomic deletion. The genomic breakpoints are distinct from those observed in chimpanzee (37, 38). Anzai *et al*. note that this region is subject to deletion and, on the basis of sequence differences between human and chimpanzee, narrowed down a possible recombination breakpoint to a segment located between the ends of *MICA*'s second and *MICB*'s fourth introns. All in all, it is intriguing that a similar-sized deletion involving the same genes (*MICA* and *MICB*) has occurred independently in different primate species. The molecular basis of the apparently deletion-prone nature of the segment between *MICA* and *MICB* remains to be established, but the presence here of a HERV-L endogenous retrovirus sequence, which contains a 2.5-kb AT-rich insertion in its 5′ long terminal repeat that might serve as a recombination hot spot, could have some involvement (39).

MICF, one of five *MIC* pseudogenes, which lies between *HLA-G* and *HLA-H* in the human MHC, appears to be absent in gorilla (Figure 2A). This is consistent with the chimpanzee (22).

Figure 2 Continued
orthologue; above line = locus on forward strand (in reference to human chromosomes); below line = locus on reverse strand; stacked = genes overlap or are nested. Gene names are given where available and are only shown for gorilla and chimpanzee when different from human; locus names that appear as numbers with leading zeros are loci without approved nomenclature, with the numbers representing the numerical part of VEGA stable gene IDs (to obtain the full ID, the 11-digit number should be prepended with OTTGORG, OTTPANG or OTTHUMG for gorilla, chimpanzee and human, respectively). An italicised locus name between brackets for a pseudogene indicates the parent gene or gene family of that pseudogene. The loci on the chimpanzee contig in panel B3 are annotated by ENSEMBL (release 70, January 2013), with dotted outlined loci indicating manually determined genes not annotated by ENSEMBL. Section labels A, B and C have been added to allow for easier reference to this figure in the text.

Class I: other. The translation of the gorilla orthologue of *PSORS1C1* (psoriasis susceptibility 1 candidate 1) terminates prematurely relative to human and chimpanzee, both of which have genes encoding 152-amino acid proteins. A single nucleotide deletion in a polyC tract in the penultimate coding exon 5 (in codon 39 or 40 of the CDS) of gorilla *PSORS1C1* leads to a frameshift in the CDS. This introduces 24 novel C-terminal amino acids in the peptide before a premature termination codon in the last exon. In the genome of human haplotype QBL, the same deletion results in the same potential translation of 63-amino acids. This contrasts with haplotypes COX and PGF, which have *PSORS1C1* genes coding for a full-length 152-amino acid peptide. It would be of interest to investigate whether a similar discrepancy exists between different gorilla haplotypes. Further analysis is required to establish whether the gorilla *PSORS1C1* locus is functional population-wide or not.

Class II. The class II region generally shows notable divergence between species (40, 41); however, a high degree of conservation exists between gorilla and human in the vicinity of the *DRB* loci. Only one of the two variants of the *C6orf10* gene that have been annotated in human could be annotated in gorilla. The second variant annotated in the human PGF assembly (OTTHUMT00000076177) contains three additional exons, one of which cannot be resolved in gorilla (between gorilla exons 19 and 20).

The gorilla class II region contains one *DRA* locus—*Gogo-DRA*—and five *DRB* loci: *Gogo-DRB9* (a pseudogene, as in human), *CH255-114D6.4* (OTTGORG00000000166, a small *DRB* pseudogene fragment not present in human), *Gogo-DRB* (known as *Gogo-DRBY*01* or *Gogo-DRB*W8:02*) (42), pseudogene *Gogo-DRB6* (*Gogo-DRB6*01:02N*) (43) and *Gogo-DRB1* (*Gogo-DRB1*10:01*) (44). Alleles mentioned above are from the Immuno Polymorphism Database, non-human primate MHC section (45, 46). In human, this region is subject to haplotypic variation (47–49), with different individuals having a different number and/or complement of genes out of the nine *DRB* genes. The *Gogo-DRB* gene is located in a position equivalent to *HLA-DRB5*, but it is not a direct orthologue: it is from a haplotype with a complement of genes slightly different from the human reference. *DQB3* exists as a pseudogene fragment, as is the case in human. The *Gogo-DQB2* locus terminates prematurely in gorilla compared to human owing to an out-of-frame tandem repeat expansion in exon 3 and has therefore been annotated as a pseudogene (see Figure 2C1). An additional *DQB* pseudogene fragment—*CH255-354J20.5* (OTTGORG00000000172)—lies immediately downstream of *Gogo-DQB2*. Except for the absence of zinc finger pseudogene *ZNF314P* (Figure 2C2) and an orthologue of human *HCG25*, the gene order and content of the remainder of the class II and extended class II in gorilla, from *TAP2* to *KIFC1*, is identical to that in human.

Based on the gene and exon arrangement of *DRB* genes in this gorilla, the DR type would appear to be DR51 (3). It is difficult to say much about this region in chimpanzee, as this falls outside the area for which clone-based sequence is available. Owing to the repetitive and variable nature of this region, the whole genome shotgun sequence that is available cannot be completely relied on, especially because there are currently many assembly gaps in this region. With this caveat in mind, however, it would appear that chimpanzee has a functional *DQB2* gene (ENSPTRG00000023404) like human.

Class III. The genes in the class III region of the gorilla were found to share a high degree of conserved synteny with the human class III genes, and this is in accordance with what others have found in pig, mouse and zebrafish (50–53). One location where differences do occur among the primates, however, is within the RCCX region (54). Early identification of the genetic heterogeneity of this *C4* to *CYP21* region in primates was made in 1991 (55), when the likely timing of the original duplication was established to be at >30 million years ago. Not only do there seem to have been duplications of this gene block (56) but also an insertion of a 6.5-kb endogenous retrovirus ERV-K into intron 9 of the *C4* gene, which has given rise to both short and long forms of this gene (57, 58). Incidentally, this ERV is one of the most biologically active of the endogenous retroviral insertion elements and is possibly still being actively transcribed at a low level in the population (59). The ancestral *C4* gene in primates was likely to have been the short *C4* gene (57) and because in some humans at least both the derived copies *C4A* and *C4B* are found in the long form (4, 60), one can surmise that this insertion pre-dates any duplication event. Results from the many studies of this region in both different primates (55) and in different individuals from the same species (61, 62) show that after the initial duplication, the region has undergone repeated and unequal homologous intragenic crossing over in all the primate species examined (63). No examples of chimpanzee or gorilla RCCX regions studied to date (56, 57, 64), however, have shown the presence of the ERV in any of the copies of the *C4* gene, and both species can carry multiple copies of the block (62). Because the ERV is found in African green monkeys, rhesus macaques (56), orangutans and humans in exactly the same intron of their *C4* genes (60), it seems more probable that the processes of duplication and homogenisation that have occurred in this region have resulted in the loss of the ERV [perhaps early in the lineages of these two species after the first duplication(s)], rather than that neither species ever carried them (57).

In our gorilla, there appear to be three copies of this module (trimodular structure): we have annotated three

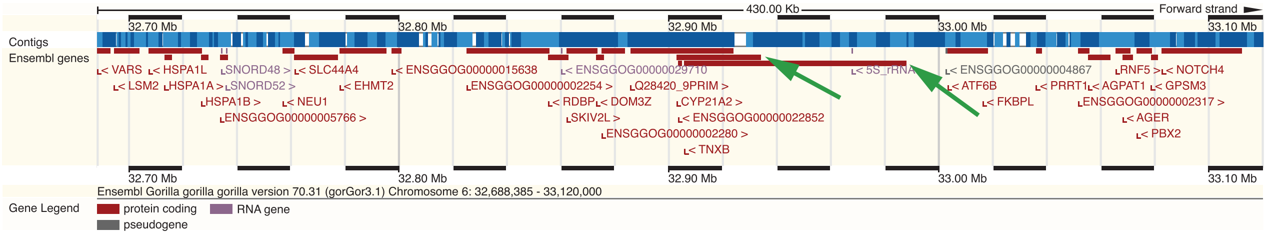

Figure 3. ENSEMBL browser view (release 70, January 2013) of the RCCX cluster and flanking regions of the genome of 'Kamilah' (whole-genome shotgun gorilla sequence) showing assembly gaps (white between the blue contigs) and gene models straddling assembly gaps and merging separate fragmented loci (green arrows). See Figure 1 legend for description of features.

C4 loci (*C4A*, *C4A-2* and *C4B*, all of the same short length), three full-length *CYP21A2* loci (*CYP21A2*, *CYP21A2-2*, *CYP21A2-3*), two copies of the *TNXA* pseudogene and two copies of *STK19P* (*STK19P* and *STK19P-2*) (Figure 2B3). All three copies of *CYP21A2* are full length, and thus have the potential for being functional. This is not the case with the human haplotypes studied, where just one copy of *CYP21A2* is functional and additional copies appear to be pseudogenes (4). Although in the human haplotypes we have studied we did not find a trimodular structure, it has been found in other individuals (61, 65). There is no chimpanzee BAC sequence for this region, but whole-genome shotgun sequence from the same chimpanzee ('Clint') shows it to have a bimodular structure like the human reference (Figure 2B3). Variation in the RCCX cluster has been linked to various autoimmune diseases, such as arthritis (66) and lupus (67).

Conclusion

We have, for the first time, presented the comprehensive annotation of the gorilla MHC genomic region (16). Sequencing was clone based, and therefore of the high quality required for regions of a highly variable and duplicated nature. As can be seen in Figure 3, which shows the RCCX region of the mostly Illumina *de novo*–assembled whole-genome shotgun sequenced 'Kamilah' genome (14) in the ENSEMBL browser, there are several assembly gaps. Also, some predicted gene models (*TNXB* for example) straddle the gaps and almost certainly merge parts of separate loci (*TNXA* and *TNXB* for example) that are not completely presented on the assembly because of mis-assembly and/or assembly gaps. Because of the assembly gap in the middle of the RCCX region, it is not possible to say whether 'Kamilah' has a bimodular or trimodular RCCX structure. Quantitative and qualitative module differences have been associated with disease in human (61, 66, 68, 69).

For comparison purposes, we have presented updated annotation of the clone-based sequence of the chimpanzee class I MHC (70) sequenced by Anzai *et al*. (22). This means that at least for the class I MHC region, we have accurate annotation for three closely related primates—human, gorilla and chimpanzee—which allows researchers to compare MHC structure and evolution between, for example, Old World primates and New World monkeys, the latter of which seem to have a less diverse MHC (71).

The genes in the MHC region are of great medical interest (72) because they are critical for the vertebrate immune system. The region's evolution is driven by the capacity to combat infectious disease, and positive selection operates on MHC loci to maintain variation; hence, the greater the diversity of class I and II molecules (both qualitatively and quantitatively), the greater the possibility of survival of a species (73). Until now, in no other higher primate has the MHC been sequenced and analysed to the same depth as the human, so this study of such a close ancestor should prove a valuable resource that can be expected to advance our understanding of the structure, function, variation and evolution of this complex region in primates. It also adds to the growing body of data on MHC genes and regions (74) in vertebrates in general (75–78).

Materials and methods

Mapping

The BAC clones in this study were chosen from the CHORI-255 library (Children's Hospital Oakland Research Institute, Oakland, CA), in cloning vector pTARBAC2.1, from a blood sample of a heterozygous male Western lowland gorilla ('Frank', #327, *Gorilla gorilla gorilla*) housed at the Lincoln Park Zoo (Chicago) (79). Further details of the BAC library can be found at (80). Screening of the redundant set of library filters was carried out as described by Stewart *et al.* (81) for the human MHC haplotype study using the human-derived 'overgo pairs' probes. A total of 333 positive BAC clones were placed in a new array, which was used to generate 70 identical filters that were then probed individually with the full set of 'overgo pairs'. These clones were then BAC-end sequenced, restriction (HindIII) fingerprinted and mapped to a single contig using the methods described by Stewart *et al.* The high-quality BAC-end sequences were matched back to the human MHC in ENSEMBL with BLAST (Basic Local

Alignment Search Tool) to enable a minimally overlapping tile-path of 29 BAC clones to be selected for sequencing.

Sequencing

The entire gorilla MHC region constituting a tile-path of 29 BACs was sequenced at the Wellcome Trust Sanger Institute (United Kingdom). Sub-cloning, capillary sequencing and finishing were performed using the standard procedures in operation at the Sanger Institute at the time and as described in summary in the headers of the International Nucleotide Sequence Database submissions for these BAC sequences.

Sequence annotation and analysis

Manual annotation was uniformly performed on the entire gorilla MHC sequence by the Wellcome Trust Sanger Institute HAVANA team as follows. The finished gorilla genomic sequence was analysed using an automatic ENSEMBL pipeline (82) with modifications to aid the manual curation process. The G + C content of each clone sequence was analysed, and putative CpG islands marked. Interspersed repeats were detected by RepeatMasker (83) using the mammalian library along with primate-specific repeats submitted to the INSDC (International Sequence Database Collaboration of DNA Data Bank of Japan (DDBJ), European Nucleotide Archive (ENA) and GenBank) databases, and simple repeats were found using Tandem Repeats Finder (84). The combined repeat types were used to mask the full sequence. This masked sequence was searched for matches to vertebrate cDNAs and expressed sequence tags (ESTs) using WU-BLASTN (85), and matches were refined and cleaned up using EST2_GENOME (86). A protein database combining non-redundant data from SwissProt and TrEMBL was searched using WU-BLASTX (85). *Ab initio* gene structures were predicted using FGENESH (87) and GENSCAN (88). Based on the aforementioned analysis, gene or transcript models were manually annotated according to the ENCODE Genome Annotation Assessment Project (EGASP) guidelines (89). The gene categories used were as described on the VEGA website (15, 90) and the annotation guidelines available from the HAVANA website (91): known genes are identical to known gorilla cDNA or protein sequences or are orthologues of known human loci; novel CDS loci have an ORF and are identical to spliced ESTs and/or have some similarity to other genes or proteins; novel transcripts are similar to novel CDS loci but no ORF can be determined unambiguously; putative genes are identical to spliced ESTs, but do not contain an ORF; and pseudogenes are non-functional copies of known or novel loci with coding regions disrupted by premature stop codons and/or frameshifts.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

The author would like to thank the members of the sequencing, sequence finishing and HAVANA teams for their invaluable contribution.

Funding

This work was supported by the Wellcome Trust. The Wellcome Trust (098051).

Conflict of interest. None declared.

References

1. Allcock,R.J., Atrazhev,A.M., Beck,S. *et al.* (2002) The MHC haplotype project: a resource for HLA-linked association studies. *Tissue Antigens*, **59**, 520–521.
2. Mungall,A.J., Palmer,S.A., Sims,S.K. *et al.* (2003) The DNA sequence and analysis of human chromosome 6. *Nature*, **425**, 805–811.
3. Horton,R., Gibson,R., Coggill,P. *et al.* (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. *Immunogenetics*, **60**, 1–18.
4. Traherne,J.A., Horton,R., Roberts,A.N. *et al.* (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. *PLoS Genet.*, **2**, e9.
5. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
6. Horton,R., Wilming,L., Rand,V. *et al.* (2004) Gene map of the extended human MHC. *Nat. Rev. Genet.*, **5**, 889–899.
7. Adams,E.J. and Parham,P. (2001) Species-specific evolution of MHC class I genes in the higher primates. *Immunol. Rev.*, **183**, 41–64.
8. Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
9. Shiina,T., Ota,M., Shimizu,S. *et al.* (2006) Rapid evolution of major histocompatibility complex class I genes in primates generates new disease alleles in humans via hitchhiking diversity. *Genetics*, **173**, 1555–1570.
10. Hobolth,A., Christensen,O.F., Mailund,T. *et al.* (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.*, **3**, e7.
11. Yunis,J.J. and Prakash,O. (1982) The origin of man: a chromosomal pictorial legacy. *Science*, **215**, 1525–1530.
12. Jauch,A., Wienberg,J., Stanyon,R. *et al.* (1992) Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proc. Natl Acad. Sci. USA*, **89**, 8611–8615.
13. Mrasek,K., Heller,A., Rubtsov,N. *et al.* (2001) Reconstruction of the female Gorilla gorilla karyotype using 25-color FISH and multicolor banding (MCB). *Cytogenet. Cell Genet.*, **93**, 242–248.

14. Scally,A., Dutheil,J.Y., Hillier,L.W. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.
15. Wilming,L.G., Gilbert,J.G., Howe,K. *et al.* (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
16. http://vega.sanger.ac.uk/Gorilla_gorilla/Location/Chromosome?r=6-MHC (26 February 2013, date last accessed).
17. Gilad,Y., Przeworski,M. and Lancet,D. (2004) Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol.*, **2**, E5.
18. Hasin,Y., Olender,T., Khen,M. *et al.* (2008) High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet.*, **4**, e1000249.
19. Hepper,P., Wells,D., McArdle,P. *et al.* (2008) Olfaction in the Gorilla. In: Hurst,J.L., Beynon,R.J., Roberts,S.C. *et al.* (eds), *Chemical Signals in Vertebrates 11*. Springer, New York.
20. Adams,E.J., Cooper,S., Thomson,G. *et al.* (2000) Common chimpanzees have greater diversity than humans at two of the three highly polymorphic MHC class I genes. *Immunogenetics*, **51**, 410–424.
21. Lawlor,D.A., Warren,E., Taylor,P. *et al.* (1991) Gorilla class I major histocompatibility complex alleles: comparison to human and chimpanzee class I. *J. Exp. Med.*, **174**, 1491–1509.
22. Anzai,T., Shiina,T., Kimura,N. *et al.* (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. USA*, **100**, 7708–7713.
23. Grimsley,C., Mather,K.A. and Ober,C. (1998) HLA-H: a pseudogene with increased variation due to balancing selection at neighboring loci. *Mol. Biol. Evol.*, **15**, 1581–1588.
24. Urvater,J.A. and Watkins,D.I. (2000) Isolation of the HLA-H orthologue in gorillas and chimpanzees. *Immunogenetics*, **51**, 69–74.
25. Kulski,J.K., Anzai,T., Shiina,T. *et al.* (2004) Rhesus macaque class I duplicon structures, organization, and evolution within the alpha block of the major histocompatibility complex. *Mol. Biol. Evol.*, **21**, 2079–2091.
26. Kulski,J.K., Gaudieri,S., Martin,A. *et al.* (1999) Coevolution of PERB11 (MIC) and HLA class I genes with HERV-16 and retroelements by extended genomic duplication. *J. Mol. Evol.*, **49**, 84–97.
27. Adams,E.J., Cooper,S. and Parham,P. (2001) A novel, nonclassical MHC class I molecule specific to the common chimpanzee. *J. Immunol.*, **167**, 3858–3869.
28. Daza-Vamenta,R., Glusman,G., Rowen,L. *et al.* (2004) Genetic divergence of the rhesus macaque major histocompatibility complex. *Genome Res.*, **14**, 1501–1515.
29. Bahram,S. (2000) MIC genes: from genetics to biology. *Adv. Immunol.*, **76**, 1–60.
30. Shiina,T., Tamiya,G., Oka,A. *et al.* (1999) Molecular dynamics of MHC genesis unraveled by sequence analysis of the 1,796,938-bp HLA class I region. *Proc. Natl Acad. Sci. USA*, **96**, 13282–13287.
31. Hughes,A.L., Yeager,M., Ten Elshof,A.E. *et al.* (1999) A new taxonomy of mammalian MHC class I molecules. *Immunol. Today*, **20**, 22–26.
32. de Groot,N.G., Garcia,C.A., Verschoor,E.J. *et al.* (2005) Reduced MIC gene repertoire variation in West African chimpanzees as compared to humans. *Mol. Biol. Evol.*, **22**, 1375–1385.
33. Kulski,J.K., Shiina,T., Anzai,T. *et al.* (2002) Comparative genomic analysis of the MHC: the evolution of class I duplication blocks, diversity and complexity from shark to man. *Immunol. Rev.*, **190**, 95–122.
34. Ando,H., Mizuki,N., Ota,M. *et al.* (1997) Allelic variants of the human MHC class I chain-related B gene (MICB). *Immunogenetics*, **46**, 499–508.
35. Fodil,N., Laloux,L., Wanner,V. *et al.* (1996) Allelic repertoire of the human MHC class I MICA gene. *Immunogenetics*, **44**, 351–357.
36. http://www.ebi.ac.uk/imgt/hla/stats.html (26 February 2013, date last accessed).
37. Komatsu-Wakui,M., Tokunaga,K., Ishikawa,Y. *et al.* (1999) MIC-A polymorphism in Japanese and a MIC-A-MIC-B null haplotype. *Immunogenetics*, **49**, 620–628.
38. Ota,M., Bahram,S., Katsuyama,Y. *et al.* (2000) On the MICA deleted-MICB null, HLA-B*4801 haplotype. *Tissue Antigens*, **56**, 268–271.
39. Shiina,T., Tamiya,G., Oka,A. *et al.* (1998) Nucleotide sequencing analysis of the 146-kilobase segment around the IkBL and MICA genes at the centromeric end of the HLA class I region. *Genomics*, **47**, 372–382.
40. Yuhki,N., Beck,T., Stephens,R.M. *et al.* (2003) Comparative genome organization of human, murine, and feline MHC class II region. *Genome Res.*, **13**, 1169–1179.
41. Wan,Q.H., Zeng,C.J., Ni,X.W. *et al.* (2009) Giant panda genomic data provide insight into the birth-and-death process of mammalian major histocompatibility complex class II genes. *PLoS One*, **4**, e4147.
42. Kupfermann,H., Mayer,W.E., O'HUigin,C. *et al.* (1992) Shared polymorphism between gorilla and human major histocompatibility complex DRB loci. *Hum. Immunol.*, **34**, 267–278.
43. Corell,A., Morales,P., Varela,P. *et al.* (1992) Allelic diversity at the primate major histocompatibility complex DRB6 locus. *Immunogenetics*, **36**, 33–38.
44. Kenter,M., Otting,N., de Weers,M. *et al.* (1993) Mhc-DRB and -DQA1 nucleotide sequences of three lowland gorillas. Implications for the evolution of primate Mhc class II haplotypes. *Hum. Immunol.*, **36**, 205–218.
45. http://www.ebi.ac.uk/ipd/mhc/nhp/index.html (26 February 2013, date last accessed).
46. de Groot,N.G., Otting,N., Robinson,J. *et al.* (2012) Nomenclature report on the major histocompatibility complex genes and alleles of Great Ape, Old and New World monkey species. *Immunogenetics*, **64**, 615–631.
47. Horton,R., Niblett,D., Milne,S. *et al.* (1998) Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.*, **282**, 71–97.
48. von Salome,J., Gyllensten,U. and Bergstrom,T.F. (2007) Full-length sequence analysis of the HLA-DRB1 locus suggests a recent origin of alleles. *Immunogenetics*, **59**, 261–271.
49. Andersson,G. (1998) Evolution of the human HLA-DR region. *Front. Biosci.*, **3**, d739–d745.
50. Trachtulec,Z. and Forejt,J. (2001) Synteny of orthologous genes conserved in mammals, snake, fly, nematode, and fission yeast. *Mamm. Genome*, **12**, 227–231.
51. Sultmann,H., Sato,A., Murray,B.W. *et al.* (2000) Conservation of Mhc class III region synteny between zebrafish and human as determined by radiation hybrid mapping. *J. Immunol.*, **165**, 6984–6993.
52. Peelman,L.J., Chardon,P., Vaiman,M. *et al.* (1996) A detailed physical map of the porcine major histocompatibility complex (MHC) class III region: comparison with human and mouse MHC class III regions. *Mamm. Genome*, **7**, 363–367.
53. Yung Yu,C., Yang,Z., Blanchong,C.A. *et al.* (2000) The human and mouse MHC class III region: a parade of 21 genes at the centromeric segment. *Immunol. Today*, **21**, 320–328.

54. Blanchong,C.A., Chung,E.K., Rupert,K.L. *et al*. (2001) Genetic, structural and functional diversities of human complement components C4A and C4B and their mouse homologues, Slp and C4. *Int. Immunopharmacol.*, **1**, 365–392.
55. Bontrop,R.E., Broos,L.A., Otting,N. *et al*. (1991) Polymorphism of C4 and CYP21 genes in various primate species. *Tissue Antigens*, **37**, 145–151.
56. Bontrop,R.E. (2006) Comparative genetics of MHC polymorphisms in different primate species: duplications and deletions. *Hum. Immunol.*, **67**, 388–397.
57. Dangel,A.W., Baker,B.J., Mendoza,A.R. *et al*. (1995) Complement component C4 gene intron 9 as a phylogenetic marker for primates: long terminal repeats of the endogenous retrovirus ERV-K(C4) are a molecular clock of evolution. *Immunogenetics*, **42**, 41–52.
58. Patience,C., Wilkinson,D.A. and Weiss,R.A. (1997) Our retroviral heritage. *Trends Genet.*, **13**, 116–120.
59. Tonjes,R.R., Lower,R., Boller,K. *et al*. (1996) HERV-K: the biologically most active human endogenous retrovirus family. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.*, **13 (**Suppl. 1), S261–S267.
60. Horiuchi,Y., Kawaguchi,H., Figueroa,F. *et al*. (1993) Dating the primigenial C4-CYP21 duplication in primates. *Genetics*, **134**, 331–339.
61. Blanchong,C.A., Zhou,B., Rupert,K.L. *et al*. (2000) Deficiencies of human complement component C4A and C4B and heterozygosity in length variants of RP-C4-CYP21-TNX (RCCX) modules in caucasians. The load of RCCX genetic diversity on major histocompatibility complex-associated disease. *J. Exp. Med.*, **191**, 2183–2196.
62. Kawaguchi,H. and Klein,J. (1992) Organization of C4 and CYP21 loci in gorilla and orangutan. *Hum. Immunol.*, **33**, 153–162.
63. Kawaguchi,H., Zaleska-Rutczynska,Z., Figueroa,F. *et al.* (1992) C4 genes of the chimpanzee, gorilla, and orang-utan: evidence for extensive homogenization. *Immunogenetics*, **35**, 16–23.
64. Martinez,O.P., Longman-Jacobsen,N., Davies,R. *et al.* (2001) Genetics of human complement component C4 and evolution the central MHC. *Front. Biosci.*, **6**, D904–D913.
65. Chung,E.K., Yang,Y., Rupert,K.L. *et al*. (2002) Determining the one, two, three, or four long and short loci of human complement C4 in a major histocompatibility complex haplotype encoding C4A or C4B proteins. *Am. J. Hum. Genet.*, **71**, 810–822.
66. Rupert,K.L., Rennebohm,R.M. and Yu,C.Y. (1999) An unequal crossover between the RCCX modules of the human MHC leading to the presence of a CYP21B gene and a tenascin TNXB/TNXA-RP2 recombinant between C4A and C4B genes in a patient with juvenile rheumatoid arthritis. *Exp. Clin. Immunogenet*, **16**, 81–97.
67. Yu,C.Y. and Whitacre,C.C. (2004) Sex, MHC and complement C4 in autoimmune diseases. *Trends Immunol.*, **25**, 694–699.
68. Saxena,K., Kitzmiller,K.J., Wu,Y.L. *et al*. (2009) Great genotypic and phenotypic diversities associated with copy-number variations of complement C4 and RP-C4-CYP21-TNX (RCCX) modules: a comparison of Asian-Indian and European American populations. *Mol. Immunol.*, **46**, 1289–1303.
69. Yang,Z., Mendoza,A.R., Welch,T.R. *et al*. (1999) Modular variations of the human major histocompatibility complex class III genes for serine/threonine kinase RP, complement component C4, steroid 21-hydroxylase CYP21, and tenascin TNX (the RCCX module). A mechanism for gene deletions and disease associations. *J. Biol. Chem.*, **274**, 12147–12156.
70. http://vega.sanger.ac.uk/Pan_troglodytes/Info/Index (26 February 2013, date last accessed).
71. Ward,J.M. and Vallender,E.J. (2012) The resurgence and genetic implications of New World primates in biomedical research. *Trends Genet.*, **28**, 586–591.
72. Knight,J.C. (2012) Genomic modulators of the immune response. *Trends Genet.*, **29**, 74–83.
73. Piertney,S.B. and Oliver,M.K. (2006) The evolutionary ecology of the major histocompatibility complex. *Heredity (Edinb.)*, **96**, 7–21.
74. Li,G., Liu,K., Jiao,S. *et al*. (2012) A physical map of a BAC clone contig covering the entire autosome insertion between ovine MHC Class IIa and IIb. *BMC Genomics*, **13**, 398.
75. Yuhki,N., Beck,T., Stephens,R. *et al*. (2007) Comparative genomic structure of human, dog, and cat MHC: HLA, DLA, and FLA. *J. Hered.*, **98**, 390–399.
76. Siddle,H.V., Deakin,J.E., Coggill,P. *et al*. (2011) The tammar wallaby major histocompatibility complex shows evidence of past genomic instability. *BMC Genomics*, **12**, 421.
77. Renard,C., Hart,E., Sehra,H. *et al*. (2006) The genomic sequence and analysis of the swine major histocompatibility complex. *Genomics*, **88**, 96–110.
78. Debenham,S.L., Hart,E.A., Ashurst,J.L. *et al*. (2005) Genomic sequence of the class II region of the canine MHC: comparison with the MHC of other mammalian species. *Genomics*, **85**, 48–59.
79. http://www.lpzoo.com/ (26 February 2013, date last accessed).
80. http://bacpac.chori.org/library.php?id=127 (26 February 2013, date last accessed).
81. Stewart,C.A., Horton,R., Allcock,R.J. *et al*. (2004) Complete MHC haplotype sequencing for common disease gene mapping. *Genome Res.*, **14**, 1176–1187.
82. Potter,S.C., Clarke,L., Curwen,V. *et al*. (2004) The Ensembl analysis pipeline. *Genome Res.*, **14**, 934–941.
83. Smit,A.F.A., Hubley,R. and Green,P. (1996–2010) http://www.repeatmasker.org (26 February 2013, date last accessed). *RepeatMasker Open-3.0*.
84. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
85. http://blast.advbiocomp.com/ (26 February 2013, date last accessed).
86. Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.*, **13**, 477–478.
87. http://doc.bioperl.org/bioperl-live/Bio/Tools/Fgenesh.html (26 February 2013, date last accessed).
88. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
89. Guigo,R., Flicek,P., Abril,J.F. *et al.* (2006) EGASP: the human ENCODE Genome annotation assessment project. *Genome Biol.*, **7 (**Suppl. 1), S2.1–S31.
90. http://vega.sanger.ac.uk/ (26 February 2013, date last accessed).
91. http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/ (26 February 2013, date last accessed).