DOI: 10.1534/genetics.106.057034

Rapid Evolution of Major Histocompatibility Complex Class I Genes in Primates Generates New Disease Alleles in Humans via Hitchhiking Diversity

Takashi Shiina,*[,1] Masao Ota,[†,1] Sayoko Shimizu,* Yoshihiko Katsuyama,[‡] Nami Hashimoto,* Miwa Takasu,[§] Tatsuya Anzai,* Jerzy K. Kulski,*[,] Eri Kikkawa,* Taeko Naruse,* Natsuki Kimura,* Kazuyo Yanagiya,* Atsushi Watanabe,* Kazuyoshi Hosomichi,* Sakae Kohara,[††] Chie Iwamoto,[‡‡] Yumi Umehara,[‡‡] Alice Meyer,[§§] Valérie Wanner,[§§] Kazumi Sano,*[,§§] Cécile Macquin,[§§] Kazuho Ikeo,[‡‡] Katsushi Tokunaga,[§] Takashi Gojobori,[‡‡] Hidetoshi Inoko* and Seiamak Bahram[§§,2]**

**Department of Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Isehara, Kanagawa 259-1143, Japan, [†]Department of Legal Medicine, Shinshu University School of Medicine, Matsumoto, Nagano 390-8621, Japan, [‡]Department of Pharmacy, Shinshu University Hospital, Matsumoto, Nagano 390-8621, Japan, [§]Department of Human Genetics, Graduate School of Medicine, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan, **Centre for Bioinformatics and Biological Computing, Murdoch University, Murdoch, Western Australia 6150, Australia, [††]Pharmacokinetics and Bioanalysis Center, Shin Nippon Biomedical Laboratories, Kainan, Wakayama 642-0017, Japan, [‡‡]Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization of Information and Systems, Mishima, Shizuoka 411-8540, Japan and [§§]Human Molecular Immunogenetics, Centre de Recherche d'Immunologie et d'Hématologie, Faculté de Médecine, 67085 Strasbourg, France*

Manuscript received February 20, 2006
Accepted for publication May 2, 2006

ABSTRACT

A plausible explanation for many MHC-linked diseases is lacking. Sequencing of the MHC class I region (coding units or full contigs) in several human and nonhuman primate haplotypes allowed an analysis of single nucleotide variations (SNV) across this entire segment. This diversity was not evenly distributed. It was rather concentrated within two gene-rich clusters. These were each centered, but importantly not limited to, the antigen-presenting *HLA-A* and *HLA-B/-C* loci. Rapid evolution of MHC-I alleles, as evidenced by an unusually high number of haplotype-specific (hs) and hypervariable (hv) (which could not be traced to a single species or haplotype) SNVs within the classical MHC-I, seems to have not only hitchhiked alleles within nearby genes, but also hitchhiked deleterious mutations in these same unrelated loci. The overrepresentation of a fraction of these hvSNV (hv1SNV) along with hsSNV, as compared to those that appear to have been maintained throughout primate evolution (trans-species diversity; tsSNV; included within hv2SNV) tends to establish that the majority of the MHC polymorphism is *de novo* (species specific). This is most likely reminiscent of the fact that these hsSNV and hv1SNV have been selected in adaptation to the constantly evolving microbial antigenic repertoire.

THE human major histocompatibility complex (MHC; also known as HLA), a minute (4-Mb) segment of the genome, harbors the full range of challenges awaiting a genome-scale search for predisposing loci to complex disorders (Hirschhorn and Daly 2005). The MHC is characterized by a set of highly polymorphic (>1700 alleles) antigen-presenting HLA class I and II genes, embedded within well over 230 loci, collectively associated with >100 pathologies (*HLA 2004*). Remarkably, recent genomewide scans have shown that, for a majority of these diseases, the MHC remains the first and foremost genetic component to pathogenesis (Onengut-Gumuscu and Concannon 2002). However, to date, with few exceptions (see below), it has been extremely difficult to identify genuine mutations/polymorphisms at the origin of the observed associations. This in large part has been due to two facts. First, the lack (until recently) of anonymous markers (that is, in addition to the highly polymorphic MHC genes themselves), *i.e.*, microsatellites and SNPs. Second, the presence of a strong degree of linkage disequilibrium across the region, which yields to the existence of extended haplotypes (Ceppellini *et al.* 1955), a well-established fact that has gained recent momentum, given the initiation of the international HapMap project (Gabriel *et al.* 2002). To alleviate both these hurdles, it is necessary,

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AB088082–AB088115, AB103588–AB103621, AB110931–AB110940, AB201549–AB201552, and AB202079–AB202114 (human); AB210139–AB210212 (chimpanzee); AB128049, AB128833–ABAB128841, AB128843–AB128846, AB128848–AB128849, AB128852–AB128856, and AB128858–AB128860 (macaque).

[1]These authors contributed equally to this work.

[2]*Corresponding author:* Human Molecular Immunogenetics, Centre de Recherche d'Immunologie et d'Hématologie, Faculté de Médecine, 4 rue Kirschleger, 67085 Strasbourg, France.
E-mail: siamak@hemato-ulp.u-strasbg.fr

following the report of the first human MHC sequence (MHC Sequencing Consortium 1999), to sequence *in fine* significant numbers of single MHC haplotypes. As medically important as the identification of the molecular basis of HLA-disease association is, a more fundamental question is, why are so many diseases linked to the MHC in the first place? The simple answer resides "somewhere" in the fact that MHC is polymorphic, which immediately raises a second question: Through which mechanism(s) has this level of diversity been created and maintained within the MHC? The answer to the second question (and by inference to the first) has occupied the field for the last 30 years. Here we aim to capitalize on sequence analysis of both human and nonhuman primate MHC haplotypes to answer these important questions.

MATERIALS AND METHODS

Human: *Cell lines:* Genomic DNA was extracted from the HLA homozygous AKIBA (*HLA-A24*, *-B52*, *-DR15* haplotype, IHW number 9286), LKT3 (*HLA-A24*, *-B54*, *-DR4* haplotype, IHW number 9107), and JPKO *(HLA-A33*, *-B44*, *-DR13* haplotype) cell lines (kindly provided by F. Numano, T. Kaneko, and Y. Ishikawa) (http://www.ecacc.org.uk/), representing the highest (8.2%), fourth highest (2.3%), and second highest (5.2%) population frequency, respectively, within the Japanese population (haplotype frequency data of the 11th International Histocompatibility Workshop; http://www.ihwg.org).

Long-range PCR amplifications: Ninety pairs of primers were designed with the assistance of Primer Express software (Applied Biosystems, Foster City, CA) for long-range PCR (LR-PCR) amplification of 55 expressed or potentially coding genes embedded within the 1.9-Mb HLA class I region and representing the entire coding content of the region linking the centromeric *LTB* to the telomeric *HLA-F* (Table S1 at http://www.genetics.org/supplemental/; Figure 1). In brief, the 50-μl amplification reaction contained 500 ng of genomic DNA, 2.5 units of TaKaRa long amplified (LA) Taq polymerase (TaKaRa Shuzo, Othu, Shiga, Japan), 1× PCR buffer, 400 μM of each dNTP, and 0.2 μM of each primer. The cycling parameters were as followes: an initial denaturation of 98°/5 min followed by 30 cycles of 98°/10 sec and 60°, 63°, or 68°/10 min, followed by a final cycle of 72° for 10 min. Two LR-PCR products of the *HCG2P7* and *HCG4P6* regions were not amplified, confirming the deletion of these two genes in the *HLA-A24* haplotype. The LR-PCR size is 5.7 kb on average and ranges from 1757 to 9448 bp (Table S1 at http://www.genetics.org/supplemental/). The entire nucleotide sequence of each gene, including the 5′-flanking region, promoter/enhancer region, exons, introns, and the 3′-flanking region, was determined by direct sequencing with 4403 sequencing primers (Table S2 at http://www.genetics.org/supplemental/). However, the exact sequence length of 55 poly(A) or poly(T) stretches of >10 nucleotides in length could not be determined accurately for 37 LR-PCR fragments amplified from *HLA-G*, *ZNRD1*, *TRIM40-1*, *TRIM40-2*, *TRIM10-1*, *TRIM15-1*, *TRIM26-1*, *TRIM26-6*, *TRIM39-1*, *TRIM39-2*, *TRIM39-3*, *RANP1*, *HLA-E*, *GNL1-1*, *ABCF1-1*, *ABCF1-2*, *ABCF1-3*, *PPP1R10-1*, *PPP1R10-2*, *PPP1R10-3*, *MRPS18B-1*, *MRPS18-2*, *DHX16-1*, *DHX16-2*, *MDC1-1*, *MDC1-2*, *FLOT1/IER3-1*, *FLOT1/IER3-2*, *FLOT1/IER3-3*, *GTF2H4*, *CDSN-3*, *TCF19-1*, *POU5F1-2*, *BAT1-1*, *NFKB/ATP6G-1*, *NFKB/ATP6G-3*, and *NFKB/ATP6G-4*. In addition, the nucleotide sequences of 15 short genomic regions (40–120 bp) could not be determined by direct sequencing for 13 LA PCR fragments amplified from *HLA-F*, *TRIM10-2*, *TRIM26-1*, *TRIM26-4*, *CAT75X*, *ABCF1-3*, *PPP1R10-1*, *MDC1-2*, *FLOT1/IER3-2*, *CDSN-3*, *C6otf18-2*, *MICA*, and *NFKB/ATP6G-1*. This sequencing difficulty was caused mainly by repetitive elements such as Alu and LINE sequences, and these regions, therefore, were further determined after sequencing cloned material. The total lengths of the sequenced nucleotides, including overlaps between long-range PCR products, were 535,285, 535,086, and 535,711 bp for LKT3, AKIBA, and JPKO cell lines, respectively. When the overlaps were excluded, the nonredundant nucleotide lengths in the LKT3, AKIBA, and JPKO cell lines were 475,879, 475,686, and 488,433 bp, respectively. The 59.4 kb of overlapping sequence did not have any nucleotide differences due to PCR and/or assembly errors, establishing high-quality sequence data and confirming the homozygous nature of cell lines.

Direct sequencing strategy, assembly, and analyses: Direct sequencing was performed using the ABI PRISM BigDye terminator cycle sequencing kit with AmpliTaq DNA polymerase (Applied Biosystems) and 4403 custom-designed primers (Table S2 at http://www.genetics.org/supplemental/). Gaps or areas of ambiguity were resolved after sequencing subcloned material (TA cloning kit, Invitrogen, Groningen, the Netherlands). Reactions were run on ABI 377 and 3100 sequencing systems. Assembly and database analyses were performed manually and using computer software following previously established procedures (Shiina *et al.* 1999).

MHC genomic sequences from three Caucasian cell lines with different HLA haplotypes: The gene sequences from the Caucasian cell lines COX (IHW 9022, South African Caucasoid consanguineous with *A1*, *B8*, *Cw7*, *DR3* haplotype), PGF (IHW 9318; European Caucasoid consanguineous with the *A3*, *B7*, *Cw7*, *DR15* haplotype), and QBL (IHW 9020; European Caucasoid consanguineous *A26*, *B18*, *DR3* haplotype) were obtained from the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/) (Stewart *et al.* 2004).

Rhesus macaque (*Macaca mulatta*): *Bacterial artificial chromosome clones, contig map, and sequencing of the macaque MHC class I region:* One bacterial artificial chromosome (BAC) library, CHORI-250, constructed from white blood cells of the male rhesus macaque was obtained from the BACPAC Resource Center at the Children's Hospital Oakland Research Institute (Oakland, CA). Hybridization screenings were performed following the recommended protocols. Hybridization probes, ~1 kb in length, were PCR generated from the macaque MHC class I and *MIC* genes (exons 2–4) as well as from 15 unique non-MHC genes—*LTB*, *POU5F1*, *HCR*, *CDSN*, *GTF2H4*, *DDR1*, *FLOT1*, *DHX16*, *ABCF1*, *CAT75X*, *TRIM26*, *TRIM10*, *TRIM31*, *C6orf12*, and *ETF1P1*—and seven MHC-based sequence-tagged sites (STS), using cloned macaque genomic DNA as template. The final contig map was constructed by comparison with the complete sequence of the human MHC (MHC Sequencing Consortium 1999). A total of 122 BACs were thus isolated and assembled into a single contig after Southern hybridizations with clone-derived PCR products and *Eco*RI fragments. Of these, 25 BACs defining a minimal tiling path linking both ends of the contig were selected (Figure S1 at http://www.genetics.org/supplemental/).

Figure 1.—An operational map of long-range PCR regions spanning from the *HLA-F* to *LTB*. Solid and open boxes indicate expressed genes and potential coding regions or pseudogenes, respectively. Striped lines around the *HLA-A* gene indicate the deleted segment of LKT3 and AKIBA haplotypes.

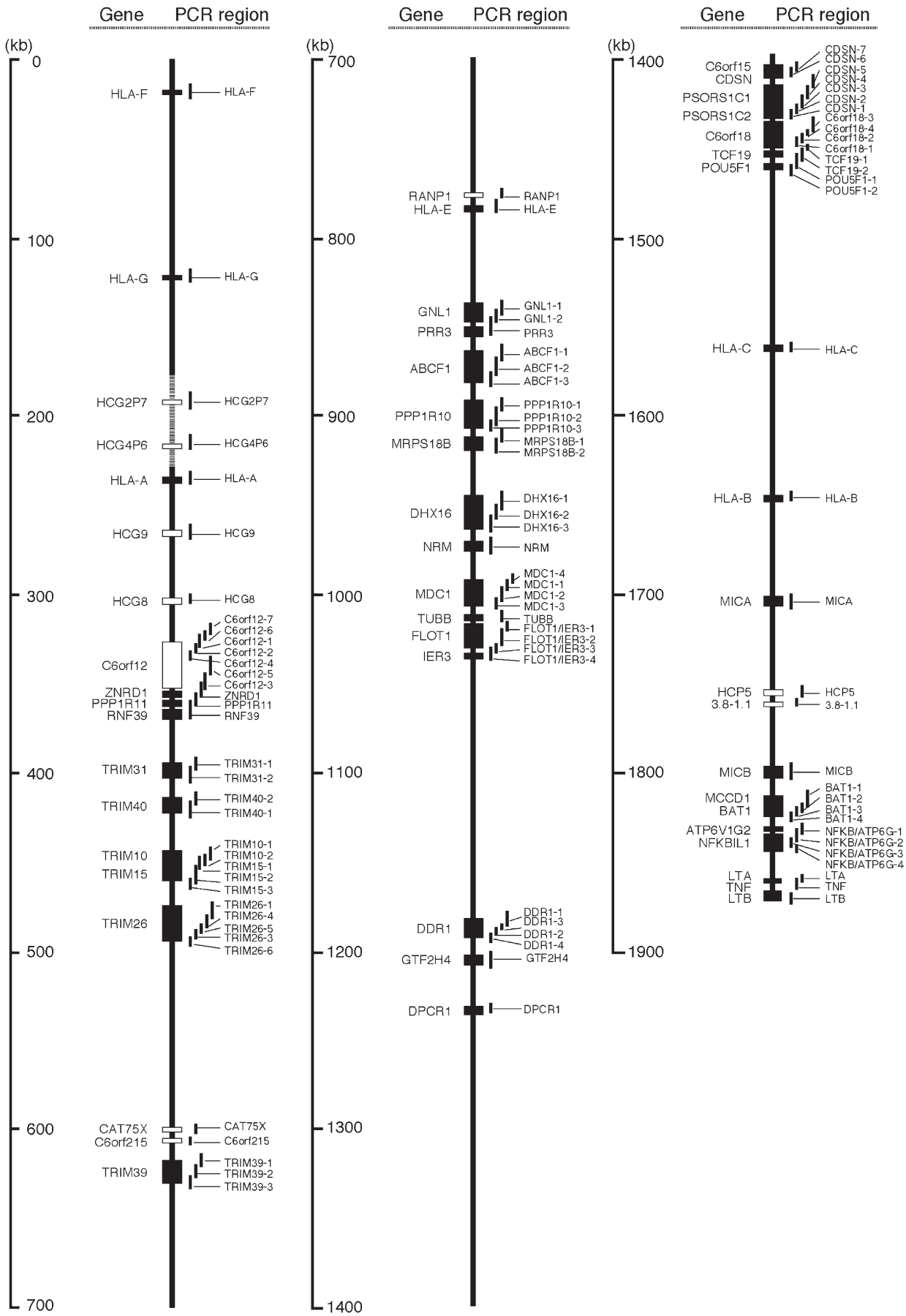

TABLE 1

SNV analysis of the MHC class I region genes in six human haplotypes

Gene or possible coding region	Annotated sequencing accession no.	Aligned sequencing length (bp)	Total SNV	SNV %	*P*-value	SNV in non-CDR	SNV in CDR	Synonymous SNV	Non-synonymous SNV	d_N (%)	d_S (%)	d_N/d_S
HLA-F	NM_018950	6,376	44	0.69 (0.31)	2.43E-03	41	3	2	1	0.04	0.33	0.12
HLA-G	NM_002127	5,512	62	1.12 (0.42)	1.39E-06	55	7	6	1	0.08	0.82	0.10
HCG2P7	X81001	8,489	122	1.44 (0.89)	4.67E-16	122	—	—	—	—	—	—
HCG4P6	X81005	3,615	71	1.97 (1.08)	2.02E-18	71	—	—	—	—	—	—
HLA-A	NM_002116	5,439	283	5.20 (2.46)	—	203	80	29	51	3.78	3.11	1.21
HCG9	NM_005844	7,308	115	1.57 (0.68)	1.15E-12	115	—	—	—	—	—	—
HCG8	X92110	4,195	35	0.83 (0.31)	2.43E-03	35	—	—	—	—	—	—
RNF39 and PPP1R11 and	NM_025236									0.20	0.39	0.51
ZNRD1 and *C6orf12*	NM_021959	45,589	226	0.50 (0.21)	NS	216	10	6	4	0	0.49	0
	NM_170783									0	0.32	0
	AF032110									—	—	—
TRIM31	NM_007028	11,762	63	0.54 (0.23)	NS	55	8	5	3	0.11	0.60	0.19
TRIM40	NM_138700	11,876	15	0.13 (0.05)	9.38E-07	15	0	0	0	0	0	—
TRIM10 and TRIM15	NM_006778	21,810	35	0.16 (0.07)	2.97E-05	33	2	1	1	0	0.08	0
	NM_033229									0.05	0	—
TRIM26	NM_003449	22,648	65	0.29 (0.11)	5.82E-03	63	2	2	0	0	0.14	0
CAT75X	L16951	3,369	11	0.33 (0.14)	8.64E-02	11	—	—	—	—	—	—
C6orf215	X90534	2,408	8	0.33 (0.14)	8.64E-02	8	—	—	—	—	—	—
TRIM39	NM_021253	18,070	30	0.17 (0.07)	2.97E-05	26	4	3	1	0.03	0.30	0.10
RANP1	XM_165758	3,603	11	0.31 (0.11)	5.82E-03	11	—	—	—	—	—	—
HLA-E	NM_005516	6,021	9	0.15 (0.06)	5.71E-06	7	2	1	1	0.08	0.11	0.74
PRR3 and GNL1	NM_025263	19,037	18	0.09 (0.04)	1.28E-07	15	3	3	0	0	0.02	0
	NM_005275									0	0.16	0
ABCF1	NM_001090	21,838	27	0.12 (0.05)	9.38E-07	27	0	0	0	0	0	—
PPP1R10 and MRPS18B	NM_002714	26,893	28	0.10 (0.04)	1.28E-07	26	2	2	0	0	0.08	0
	NM_014046									0	0	—
DHX16	NM_003587	21,033	33	0.16 (0.05)	9.38E-07	32	1	0	1	0.02	0	—
NRM	NM_007243	4,012	10	0.25 (0.09)	5.34E-04	10	0	0	0	0	0	—
MDC1	NM_014641	19,386	46	0.24 (0.08)	1.34E-04	31	15	6	9	0.09	0.09	0.96
TUBB	NM_178014	5,580	16	0.29 (0.10)	1.87E-03	15	1	1	0	0	0.09	0
IER3 and FLOT1	NM_003897	19,438	40	0.21 (0.08)	1.34E-04	38	2	1	1	0.10	0	—
	NM_005803									0	0.09	0
DDR1	NM_013993	14,655	38	0.26 (0.12)	1.61E-02	33	5	5	0	0	0.32	0
GTF2H4	NM_001517	7,851	28	0.36 (0.16)	NS	27	1	1	0	0	0.13	0
DPCR1	NM_080870	3,168	8	0.25 (0.10)	1.87E-03	6	2	0	2	0.13	0	—
CDSN and C6orf15 and	NM_001264									0.25	1.01	0.25
PSORS1C1 and PSORS1C2	NM_014070	30,877	312	1.01 (0.44)	3.82E-07	279	33	15	18	0.53	0.53	1.00
	NM_014068									0.21	0	—
	NM_014069									0.12	0	—

(*continued*)

TABLE 1

(Continued)

Gene or possible coding region	Annotated sequencing accession no.	Aligned sequencing length (bp)	Total SNV	SNV %	*P*-value	SNV in non-CDR	SNV in CDR	Synonymous SNV	Non-synonymous SNV	d_N (%)	d_S (%)	d_N/d_S
C6orf18	NM_019052	15,756	118	0.75 (0.33)	6.25E-04	97	21	10	11	0.20	0.52	0.37
TCF19 and POU5F1	NM_007592	14,443	82	0.57 (0.27)	3.15E-02	75	7	5	2	0.13	0.23	0.55
	NM_002701									0	0.53	0
HLA-C	NM_002117	4,780	192	4.02 (1.94)	—	128	64	24	40	2.68	2.74	0.98
HLA-B	NM_005514	4,588	211	4.60 (2.17)	—	129	82	29	53	3.90	3.21	1.21
MICA	NM_000247	6,997	121	1.73 (0.85)	1.74E-15	100	21	8	13	0.88	1.19	0.74
HCP5	NM_006674	5,153	24	0.47 (0.31)	2.43E-03	24	—	—	—	—	—	—
3.8-1.1	L29376	2,233	18	0.81 (0.34)	3.14E-04	18	—	—	—	—	—	—
MICB	NM_005931	6,087	44	0.72 (0.26)	NS	38	6	1	5	0.22	0.11	2.03
NFKBIL1 and ATP6V1G2	NM_005007									0.04	0.10	0.41
and BAT1 and MCCD1	NM_130463	34,012	123	0.36 (0.14)	NS	118	5	3	2	0	0	—
	NM_004640									0	0.09	0
	NM_001011700									0.13	0.51	0.26
LTA and TNF	NM_000595	6,974	14	0.20 (0.09)	5.34E-04	12	2	0	2	0.20	0	—
	NM_000594									0	0	—
LTB	NM_002341	4,378	5	0.11 (0.04)	1.28E-07	5	0	0	0	0	0	—
Total		487,229	2761	0.57 (0.24)		2372	389	168	221			

Genes in italic indicate potential coding regions or pseudogenes. We calculated the number of synonymous or nonsynonymous SNVs of exonic regions using the mRNA sequences included in the "Annotated sequencing accession no." column. "SNV%" corresponds to the total number of substitutions divided by the length of the aligned sequence and multiplied by 100. Parentheses indicates normalized SNV%. Underlining indicates the genomic segments having significantly high SNV% ($P < 0.05$) and MHC gene segments. Dashes correspond, in all but three exceptions, to loci where it has not been possible to ascribe SNV in CDR (and subsequent calculations, *i.e.*, syn, non-syn, d_N, d_S, d_N/d_S), given that these loci are potential coding regions or pseudogenes. The three exceptions are HLA-A, -B, and -C. The *P*-value column values were removed for these loci (defined as hitchhiking attackers) to evaluate variation among hitchhiking receivers (all other genes except HLA-A, -B, and -C) independently. For further explanations, see text. CDR, coding region.

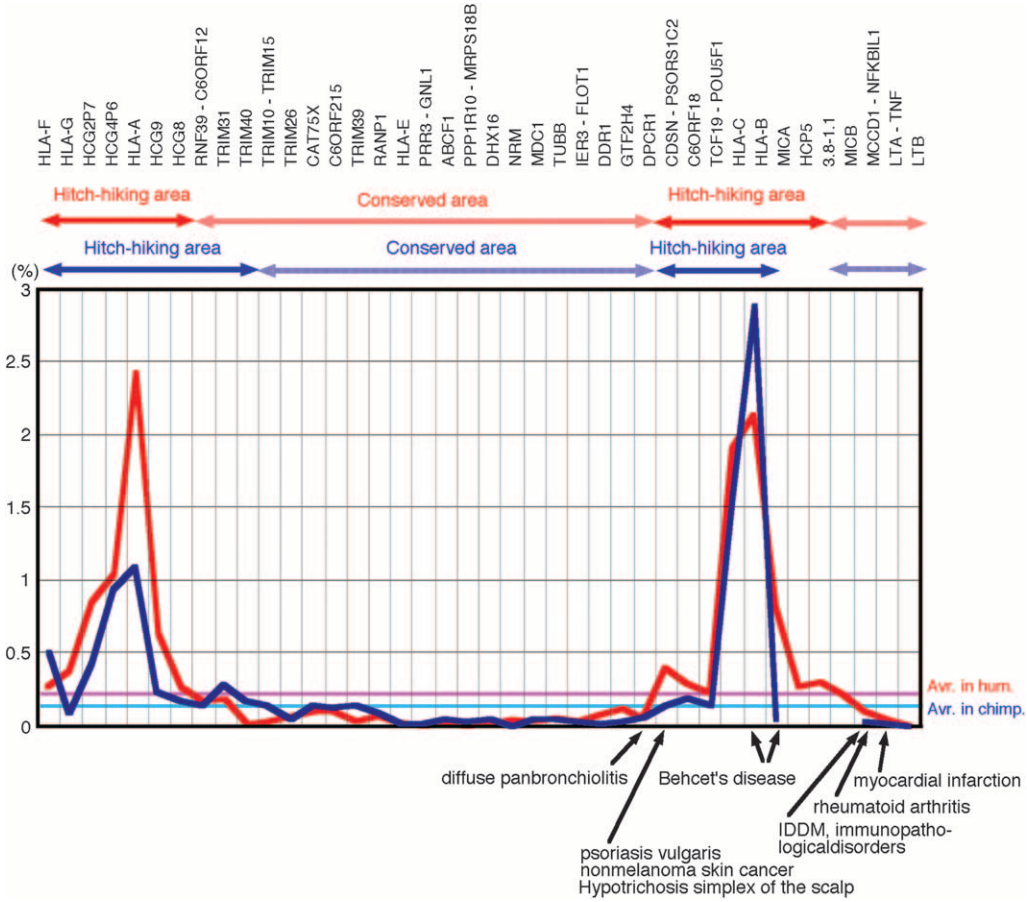

FIGURE 2.—Genomic diversity plot of the human and chimpanzee MHC class I region. Diversity plot drawn upon comparison of six human haplotypes (red) and four chimpanzee haplotypes (blue). Arrows below the plot depict the location of identified disease susceptibility genes. Arrows above the plot indicate "hitchhiked" regions for each species.

They were subjected to complete and bidirectional shotgun sequencing with an average 7.2× redundancy, which was sufficient for assembly and analysis of the entire sequence using previously established procedures (SHIINA *et al.* 1999) (Figure S1 at http://www.genetics.org/supplemental/). The total length of the contig, linking *BAT3* to *Mamu-F*, was established as 3,284,914 bp (Figure S1 and Table S3 at http://www.genetics.org/supplemental/). All clone overlaps were ascertained at the nucleotide level. The sequenced animal being MHC heterozygous, the reported sequence was derived from both chromosomes. BACs 251L06–144G11, covering 1600 kb from *BAT3* to *LOC285833*, and BACs 118F10–25A4, spanning 400 kb from *HCGII-16* to *RPN2P1*, were derived from one haplotype, whereas BACs 399F22–151J3, covering 400 kb linking *LOC285833* to *HCGII-16*, as well as BACs 164G20–48M22, covering 900 kb from *RPN2P1* to *P5-1-44*, were derived from the other haplotype. The genomic sequence was assembled into a single contig from 24 overlapping segments for the 25 BAC clones. Three overlaps located on haplotype boundaries (144G11/399F22, 151J3/118F10, and 25A4/164G20) were ascertained by significantly high nucleotide identities (>99.9% in at least 2 kb). On the other hand, the other 20 overlaps belonging to the same haplotypes were established through complete nucleotide identity. The obtained sequence was annotated using our previously published human and chimpanzee sequence data (SHIINA *et al.* 1999; ANZAI *et al.* 2003) as well as those publicly available at NCBI (http://www.ncbi.nlm.nih.gov/locuslink/). Sequence alignments were performed and homologies were determined using the programs contained within the Genetyx v11 (http://www.sdc.co.jp/genetyx). The calculation of nucleotide diversity was performed through pairwise sequence alignments by MAVID (http://baboon.math.berkeley.edu/mavid/) with three human (two Japanese and COX haplotypes) and chimpanzee sequences (ANZAI *et al.* 2003). The diversity profile was then drawn using the graphics output of Microsoft Excel. All insertion/deletions (indels) were removed from the alignments to standardize the number of nucleotides examined within each window. Well after our experimental work was finished, another macaque MHC genomic sequence was published. However, among other things, this sequence is not annotated (DAZA-VAMENTA *et al.* 2004) and hence our sequence is formally the first annotated macaque MHC sequence.

Chimpanzee (*Pan troglodytes*): *Materials:* Genomic DNA was extracted from the chimpanzee MHC heterozygous Ericka (*Patr-A0601/0901*, *-B0101/1701*, *-C0401/0601* haplotype) and Borie (*Patr-A0301/0401*, *-B0101/2401*, *-C0401/0901* haplotype) cell lines (kindly provided by Peter Parham at Stanford University). These individuals belong to *Pan troglodytes Verus* as ascertained through mitochondrial DNA D-loop region analysis (data not shown).

DNA typing of MHC class I genes: To determine the allelic sequences of Ericka and Borie *Patr-A*, *-B*, and *-C* loci, long-range PCR amplifications were performed (Table S4 at http://www.genetics.org/supplemental/). The products were subsequently subcloned by the TOPO XL PCR cloning system (Invitrogen), and eight clones for each allele were sequenced (Table S5 at http://www.genetics.org/supplemental/). The genomic sequences of *Patr-A*, *-B*, and *-C* genes in the Ericka and Borie cell lines matched perfectly those found in DNA databases.

LR-PCR, sequencing, and analysis: Among 98 pairs of human LR-PCR amplicons, 93 were well amplified in chimpanzees. The following 5 pairs, however, did not: *TRIM26-4*, *PPP1R10-2*,

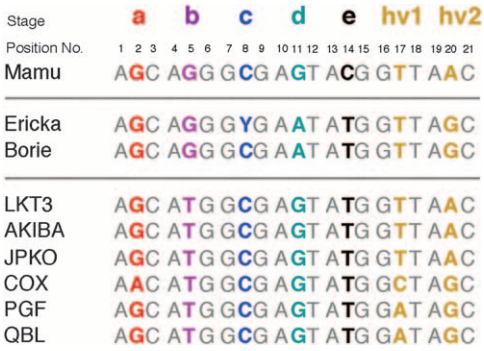

FIGURE 3.—Definition of SNV categories in a sample sequence. a, b, c, d, e, hv1, and hv2 indicate SNV generated after birth of human haplotypes; SNV generated in humans after speciation of humans and chimpanzees but before birth of human haplotypes; SNV generated in chimpanzees after speciation of humans and chimpanzees; SNV generated before speciation of human- and chimpanzee- or macaque-specific SNV and hypervariable SNV, respectively. A total of seven SNV are shown (position nos. 2, 5, 8, 11, 14, 17, and 20). There is one "a" stage SNV (position no. 2) (LKT3 = 0 AKIBA = 0 JPKO = 0 COX = 1 PGF = 0 QBL = 0); hence (0 + 0 + 0 + 1 + 0 + 0)/6 sequence = 0.17 average SNV; 0.17 SNV/21 bp × 100 = 0.81 SNV% (the DI). For "b," there is one location (position no. 5): 1 SNV/21 bp × 100 = 4.76% SNV (DI). For the "c" stage, there is one location (position no. 8) (Ericka = C/A Borie = C/C); hence (0 + 1 + 0 + 0)/4 sequence = 0.25 average SNV; 0.25 SNV/21 bp × 100 = 1.19 SNV% (DI) in stage c. For the "d" stage, there is also one location (position no. 11): 1 SNV/21 bp × 100 = 4.76% SNV (DI). For the "e" stage, there is also one variation (position no. 14) 1 SNV/21 bp × 100 = 4.76% SNV (DI). Finally, for hvSNV there are two locations (position nos. 17 and 20): 2 SNV/21 bp × 100 = 9.52% SNV (DI) at this level. To be comprehensive, the large-scale analysis of hvSNV depicted in Figure 6 was accomplished while dividing the observed hvSNV into two categories: hv1 and hv2. hv1 refers to situations similar to those seen in position 17 of this figure, where there is a patent case for *de novo* generation of the diversity in Homo, whereas for hv2 (equivalent to position 20 in this hypothetical sequence) the question (*i.e.*, the molecular genealogy of the diversity) can be definitely settled only once the syntenic segment in the most recent common ancestor of humans and chimpanzees has been sequenced.

TUBB, *POU5F1-1*, and *POU5F1-2*. These pairs were redesigned with the assistance of Primer Express software (Applied Biosystems) (Table S2 at http://www.genetics.org/supplemental/). The LR-PCR procedure was the same as for humans. The entire nucleotide sequence of each gene, including the 5′-flanking region, the promoter/enhancer region, exons, introns, and the 3′-flanking region, was determined by direct sequencing with 4174 (3987 human primers and 187 chimpanzee newly designed primers) sequencing primers (Tables S2 and S5 at http://www.genetics.org/supplemental/) as described above (*Human*). However, the exact sequence length of poly(A) or poly(T) stretches of >10 nucleotides in length, microsatellite repeats, as well that of the following 10 short genomic regions (50–400 bp) in seven LR-PCR fragments, amplified from *Patr-F*, *HCG4P6*, *C6orf12-3*, *TRIM26-4*, *CAT75X*, *MDC1-4*, and *POU5F1-2*. These sequencing difficulties, in addition to gaps or areas of ambiguity, were resolved after sequencing subcloned material (TA cloning kit, Invitrogen). The nonredundant nucleotide length in the Ericka and Borie cell lines was 472,506 and 472,528 bp, respectively.

Calculation of the statistical significance of SNP numbers in hitchhiked areas: Classical MHC class I genes such as *HLA-A*, *-B*, and *-C* in humans and *Patr-A*, *-B*, and *-C* in chimpanzees are designated as "hitchhiking attackers" (as they are the prime polymorphic loci) while the remainder of the loci are considered "hitchhiking receivers." "Hitchhiked" areas are therefore the result of hitchhiking receivers being "attacked" by hitchhiking attackers. This allowed us to perform hierarchical statistical analyses by considering the percentage of SNPs in hitchhiking receivers in each species. Significant differences were calculated by *t*-test between the average and region-specific percentage of SNPs (37 in humans and 34 in chimpanzees), excluding *HLA-A*, *-B*, and *-C* after arc sin transformation. Subsequently, 12 regions in humans (*HLA-F*, *HLA-G*, *HCG2P7*, *HCG4P6*, *HCG9*, *HCG8*, *CDSN*, *C6orf18*, *TCF19*, *MICA*, *HCP5*, and *3.8-1.1*) have a significantly increased percentage of SNPs ($P < 0.05$). Moreover, there was a significant difference ($P = 0.00001$) between these 12 gene regions and the other 25 gene regions. In chimpanzees, 11 gene regions (*Patr-F*, *HCG2P7*, *HCG4P6*, *HCG9*, *HCG8*, *RNF39*, *TRIM31*, *TRIM40*, *CDSN*, *C6orf18*, and *TCF19*) had a significantly high percentage of SNPs ($P < 0.05$). Finally, there was a significant difference ($P < 0.0001$) between these 11 gene regions and the other 23 gene regions.

RESULTS AND DISCUSSION

Our target is the telomeric half of the MHC. This 1.9-Mb segment links the centromeric *LTB* to the telomeric *HLA-F* in humans and the syntenic segments in chimpanzees and in the rhesus macaque. This segment is known to contain the most polymorphic genes of the MHC and therefore of vertebrate genomes (*HLA 2004*). Moreover, it is also the sole segment of the MHC subject to intense interspecies variability, perhaps reminiscent of the selective microbial pressure facing each of these species (KUMANOVICS *et al.* 2003). In humans, there are 55 expressed or potentially expressed genes in this area, including 8 HLA class I (*HLA-A/B/C/E/F/G*) and class I-related (*MICA/B*) genes (BAHRAM *et al.* 1994; SHIINA *et al.* 1999). Using LR-PCR on genomic DNA of three Japanese HLA homozygous individual (JPKO) or typing cell lines—AKIBA (*HLA-A24*, *-B52*, *-DR15*), LKT3 (*A24*, *B54*, *DR4*), and JPKO (*A33*, *B44*, *DR13*)—all these 55 loci (with the exception of 2 in LKT3 and AKIBA) were amplified within a set of 100 amplicons partitioned in 38 mini-contigs (40 in JPKO) (Figure 1; Table S1 at http://www.genetics.org/supplemental/). These were fully sequenced using 4403 primers (Table S2 at http://www.genetics.org/supplemental/) yielding nonredundant nucleotide lengths of 475,879, 475,686, and 488,433 bp for LKT3, AKIBA, and JPKO, respectively. The complete sequence of three other cell lines—COX (*A1*, *B8*, *Cw7*, *DR3*), PGF (*A3*, *B7*, *Cw7*, *DR15*), and QBL (*A26*, *B18*, *DR3*)—this time Caucasoid, was extracted from the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk/) (STEWART *et al.* 2004).

Cross-comparing all six cell lines using a 487,229-bp template (once indels had been excluded) (Table 1) unveiled 2761 single nucleotide variations (SNV) (1/176

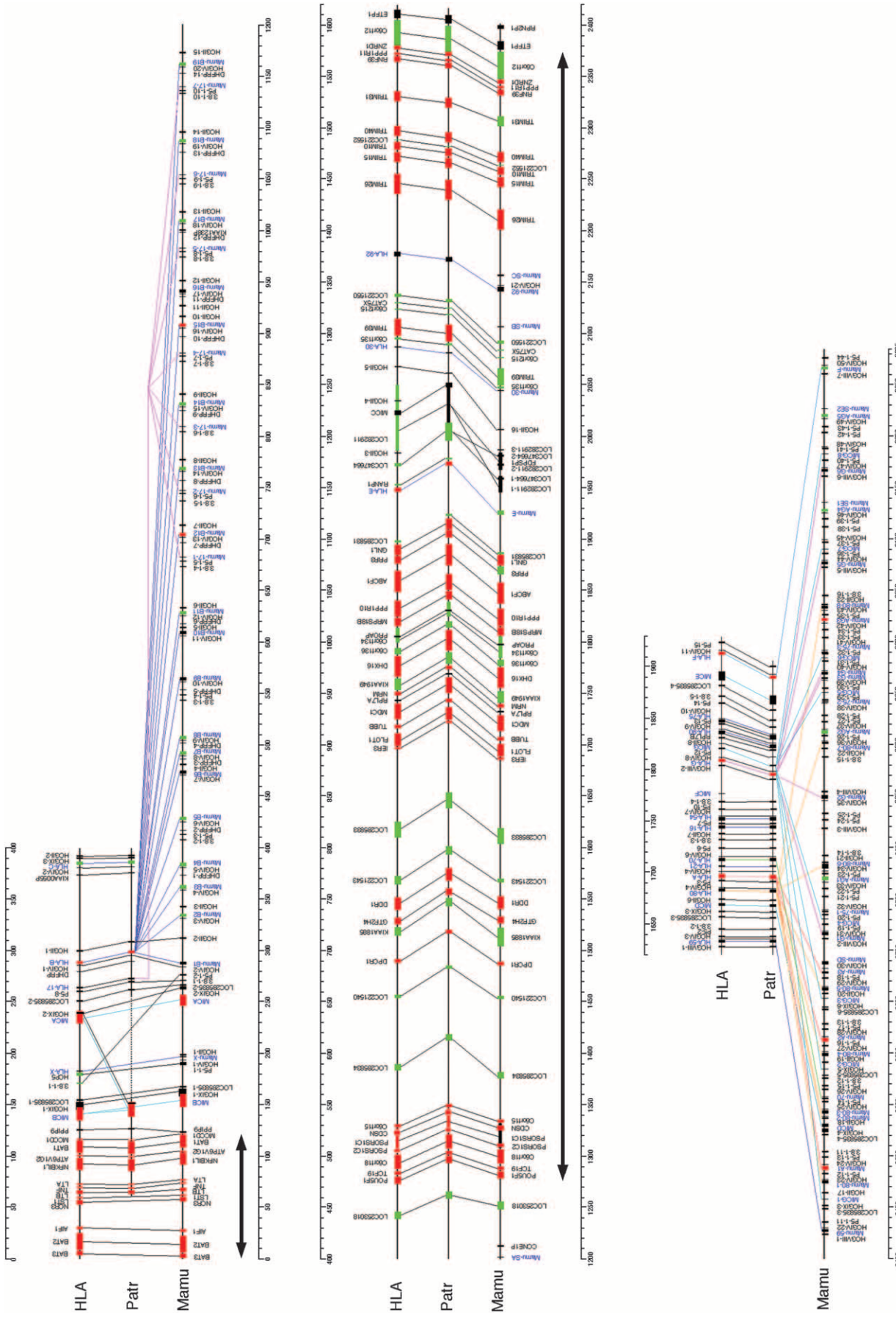

FIGURE 4.—Comparative genomic map of HLA, Patr, and Mamu class I regions. Lines show orthologous relationship. *Mamu-B* region from 300 to 1200 kb and *Mamu-A, -G, -F* region from 2400 to 3285 kb were shown only with respect to Mamu class I loci. Arrows show segments used for calculating genomic diversity analysis shown in Figure 5.

TABLE 2

SNV analysis of the MHC class I region genes in four chimpanzee haplotypes

Gene or possible coding region	Human ortholog sequencing accession no.	Aligned sequencing length (bp)	Total SNV	SNV%	*P*-value	SNV in non-CDR	SNV in CDR	Synonymous SNV	Nonsynonymous SNV	d_N (%)	d_S (%)	d_N/d_S
Patr-F	NM_018950	6,462	61	0.94 (0.52)	2.24E-13	47	14	8	6	0.48	1.09	0.44
Patr-G	NM_002127	5,533	10	0.18 (0.10)	NS	10	0	0	0	0	0	—
HCG2P7	X81001	8,490	54	0.64 (0.42)	3.46E-11	54	—	—	—	—	—	—
HCG4P6	X81005	3,602	67	1.86 (0.93)	3.51E-19	67	—	—	—	—	—	—
Patr-A	NM_002116	5,452	116	2.13 (1.09)	—	85	31	11	20	1.44	1.46	0.99
HCG9	NM_005844	7,376	32	0.43 (0.23)	2.68E-05	32	—	—	—	—	—	—
HCG8	X92110	4,216	12	0.28 (0.18)	2.39E-03	12	—	—	—	—	—	—
RNF39 and PPP1R11 and	NM_025236									0.06	0.43	0.14
ZNRD1 and *C6orf12*	NM_021959	46,836	110	0.23 (0.15)	3.25E-02	106	4	3	1	0	0	—
	NM_170783									0	0	—
	AF032110									—	—	—
TRIM31	NM_007028	11,776	56	0.48 (0.29)	1.87E-07	53	3	0	3	0.19	0.13	1.46
TRIM40	NM_138700	11,873	39	0.33 (0.18)	2.39E-03	38	1	0	1	0.14	0	—
TRIM10 and TRIM15	NM_006778	21,121	60	0.28 (0.14)	NS	53	7	4	3	0.10	0.24	0.42
	NM_033229									0	3.68	0
TRIM26	NM_003449	21,546	15	0.07 (0.04)	3.68E-03	15	0	0	0	0	0	—
CAT75X	L16951	3,346	10	0.30 (0.15)	3.25E-02	10	—	—	—	—	—	—
C6orf215	X90534	2,411	6	0.25 (0.13)	NS	6	—	—	—	—	—	—
TRIM39	NM_021253	18,095	48	0.27 (0.14)	NS	47	1	0	1	0.05	0	—
RANP1	XM_165758	3,694	5	0.14 (0.09)	NS	5	—	—	—	—	—	—
Patr-E	NM_005516	6,063	2	0.03 (0.02)	3.31E-05	2	0	0	0	0	0	—
PRR3 and GNL1	NM_025263	19,063	4	0.02 (0.01)	1.01E-06	4	0	0	0	0	0	—
	NM_005275									0	0	—
ABCF1	NM_001090	21,541	12	0.06 (0.04)	3.68E-03	12	0	0	0	0	0	—
PPP1R10 and MRPS18B	NM_002714	26,950	15	0.06 (0.03)	4.56E-04	14	1	1	0	0	0.08	0
	NM_014046									0	0	—
DHX16	NM_003587	21,145	14	0.07 (0.04)	3.68E-03	14	0	0	0	0	0	—
NRM	NM_007243	4,016	0	0 (0)	3.19E-10	0	0	0	0	0	0	—
MDC1	NM_014641	19,659	14	0.07 (0.04)	3.68E-03	8	6	4	2	0.03	0.12	0.25
TUBB	NM_178014	5,582	4	0.07 (0.04)	3.68E-03	3	1	1	0	0	0.18	0
IER3 and FLOT1	NM_003897	19,405	11	0.06 (0.03)	4.56E-04	11	0	0	0	0	0	—
	NM_005803									0	0	—
DDR1	NM_013993	14,676	4	0.03 (0.01)	1.01E-06	4	0	0	0	0	0	—
GTF2H4	NM_001517	7,860	4	0.05 (0.03)	4.56E-04	4	0	0	0	0	0	—
DPCR1	NM_080870	3,170	3	0.09 (0.06)	NS	3	0	0	0	0	0	—
CDSN and C6orf15 and	NM_001264									0.14	0.35	0.40
PSORS1C1 and PSORS1C2	NM_014070	30,910	82	0.27 (0.15)	3.25E-02	73	9	4	5	0	0	—
	NM_014068									0.21	0	—
	NM_014069									0.23	0.42	0.55
C6orf18	NM_019052	15,840	57	0.36 (0.19)	9.69E-04	53	4	3	1	0.03	0.23	0.13

(*continued*)

TABLE 2

(Continued)

Gene or possible coding region	Human ortholog sequencing accession no.	Aligned sequencing length (bp)	Total SNV	SNV%	*P*-value	SNV in non-CDR	SNV in CDR	Synonymous SNV	Nonsynonymous SNV	d_N (%)	d_S (%)	d_N/d_S
TCF19 and POU5F1	NM_007592	13,815	37	0.27 (0.15)	3.25E-02	34	3	2	1	0.07	0.16	0.44
	NM_002701									0	0.15	0
Patr-C	NM_002117	4,790	152	3.17 (1.59)	—	108	44	18	26	1.72	2.63	0.65
Patr-B	NM_005514	4,583	246	5.37 (2.89)	—	154	92	33	59	4.16	5.64	0.74
MIC	NM_000247	6,096	6	0.10 (0.05)	1.97E-02	3	3	0	3	0.20	0	—
NFKBIL1 and ATP6V1G2	NM_005007									0	0	—
and BAT1 and MCCD1	NM_130463	34,056	23	0.07 (0.03)	4.56E-04	22	1	0	1	0.20	0	—
	NM_004640									0	0	—
	NM_001011700									0	0	—
LTA and TNF	NM_000595	6,979	1	0.01 (0.01)	1.01E-06	1	0	0	0	0	0	—
	NM_000594									0	0	—
LTB	NM_002341	4,375	0	0 (0)	3.19E-10	0	0	0	0	0	0	—
Total		472,403	1398	0.30 (0.16)		1173	225	92	133			

Genes in italic indicate potential coding regions or pseudogenes. We calculated the number of synonymous or nonsynonymous SNVs of exonic regions using the mRNA sequences included in the "Human ortholog sequencing accession no." column and annotation from the genomic sequence of chimpanzee MHC class I region (accession no. BA000041). SNV% corresponds to the total number of substitutions divided by the length of the aligned sequence and multiplied by 100. Parentheses indicate normalized SNV%. Underlining indicates the genomic segments having significant high SNV% ($P < 0.05$) and MHC gene segments. Dashes correspond, in all but three exceptions, to loci where it has not been possible to ascribe SNV in CDR (and subsequent calculations, *i.e.*, syn, non-syn, d_N, d_S, d_N/d_S), given that these loci are potential coding regions or pseudogenes. The three exceptions are Patr-A, -B, and -C. The *P*-value column values were removed for these loci (defined as hitchhiking attackers) to evaluate variation among hitchhiking receivers (all other genes except Patr-A, -B, and -C) independently. For further explanations, see text. CDR, coding region.

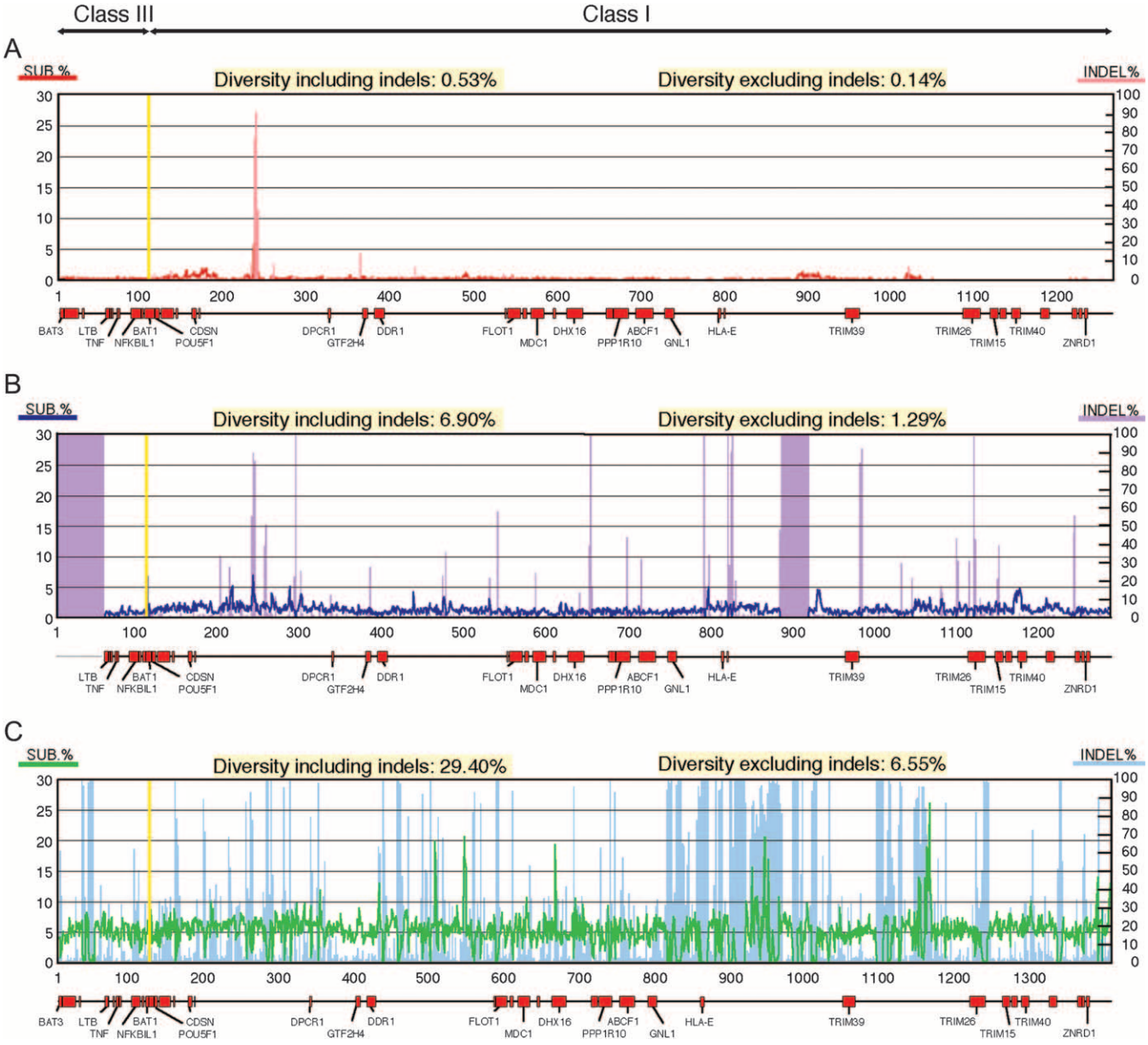

FIGURE 5.—SNV *vs.* indels in the MHC class I region. The aligned sequence (excluding indels) is shown along the horizontal axis and the percentage of nucleotide differences calculated per 1 kb of nonoverlapping windows is shown along the vertical axis. (A) Human *vs.* human (1.24 Mb). (B) Human *vs.* chimpanzee (1.26 Mb). (C) Human *vs.* macaque (1.38 Mb). Deep red, deep blue, and green lines show nucleotide substitution (%) whereas light red, light blue, and sky blue refer to indel percentages.

or 0.57/100 bp), with 389 embedded in coding regions among which a majority of 221 were nonsynonymous (Table 1). [Two types of nucleotide diversity are dealt with in this article: intra- and interspecies; while the former is typically called SNP, the usage of the same term to refer to the latter is incorrect as the "P" in SNP refers to polymorphism, which, by definition, is intraspecies. Hence, to avoid further confusion and achieve homogeneity within this work, SNV is used throughout to refer to both intra- and interspecies nucleotide diversity (for further subdivisions, see below)]. The SNV were not evenly distributed and showed a strong regional bias, which we dissect below. The greatest overall diversity was centered around two MHC-I-bearing islands, *i.e.*, the centromeric *HLA-B/-C/MICA* and the telomeric *HLA-A* islands where substitutions were significantly above the 0.57 (0.24) average diversity index (DI) (the number in parentheses refers to the normalized percentage of SNVs) (Figure 2; Table 1; see MATERIALS AND METHODS for calculation of statistical significance and Figure 3's legend for calculation of DI). Interestingly, diversity was not limited to these immune response genes (Figure 2; Table 1) (*HLA 2004*). For instance, while *HLA-A* [DI: 5.20 (2.46)] and *HLA-B* [DI: 4.60 (2.17)] represent the penultimate and ultimate diverse genes within the MHC class I region, respectively, 250-kb

TABLE 3

SNV analysis of 33 genomic regions within the MHC class I regions of humans, chimpanzee, and macaques

Gene region	Total SNV%	SNV% in intron	Subregion	SNV% in CDR	Synonymous SNV	Nonsynonymous SNV	SNV% in nonsynonymous
HLA-F	8.41 (7.50)	8.73 (7.80)	HLA-F	6.92 (6.14)	44	28	2.69 (2.32)
HLA-A	12.77 (10.55)	12.56 (10.00)	HLA-A	13.56 (10.11)	62 (44.25)	87 (66.75)	7.92 (6.01)
HCG9	9.43 (8.41)						
HCG8	8.22 (7.68)						
RNF39 and PPP1R11			RNF39	3.88 (3.29)	34 (29.67)	15 (11.92)	1.19 (0.94)
and ZNRD1	6.14 (5.76)	6.30 (5.91)	PPP1R11	1.31 (1.23)	5 (4.17)	0	0
			ZNRD1	1.57 (1.36)	3 (2.67)	3 (3.00)	0.79 (0.79)
TRIM31[a]	6.93 (6.38)						
TRIM40	6.68 (6.43)	6.82 (6.58)	TRIM40	4.28 (4.20)	18 (18.00)	10 (9.50)	1.53 (1.45)
TRIM10 and TRIM15	6.81 (6.60)	7.26 (7.02)	TRIM10	3.94 (3.75)	39 (37.17)	18 (17.00)	1.24 (1.18)
			TRIM15	4.47 (4.38)	34 (33.25)	33 (32.42)	2.20 (2.16)
TRIM26	6.52 (6.33)	6.84 (6.65)	TRIM26	2.78 (2.34)	41 (39.83)	4 (4.00)	0.25 (0.25)
CAT75X	6.45 (6.10)						
C6orf215	5.91 (5.58)						
TRIM39	5.04 (4.86)	5.38 (5.20)	TRIM39	1.67 (1.45)	21 (19.17)	5 (3.42)	0.32 (0.22)
RANP1	8.93 (8.62)						
HLA-E	7.65 (7.54)	7.69 (7.59)	HLA-E	7.46 (7.34)	37 (36.17)	44 (43.50)	4.05 (4.01)
PRR3 and GNL1	5.14 (5.08)	5.58 (5.52)	PRR3	2.47 (2.47)	12 (12.00)	2 (2.00)	0.36 (0.36)
			GNL1	1.97 (1.91)	34 (32.83)	2 (2.00)	0.11 (0.11)
ABCF1	6.08 (5.98)	6.59 (6.49)	ABCF1	2.28 (2.28)	47 (47.00)	8 (8.00)	0.33 (0.33)
PPP1R10 and	5.29 (5.20)	5.65 (5.55)	PPP1R10	2.80 (2.72)	70 (68.67)	8 (8.00)	0.28 (0.28)
MRPS18B			MRPS18B	3.86 (3.86)	13 (13.00)	17 (17.00)	0.13 (0.13)
DHX16	6.17 (6.03)	6.94 (6.79)	DHX16	1.92 (1.89)	54 (54.00)	6 (5.17)	0.19 (0.17)
NRM	4.91 (4.78)	5.44 (5.30)	NRM	2.66 (2.66)	14 (14.00)	7 (7.00)	0.89 (0.89)
MDC1	6.41 (6.28)	7.32 (7.19)	MDC1	3.83 (3.73)	74 (72.25)	105 (101.67)	2.25 (2.18)
TUBB	4.86 (4.63)	1.21 (0.95)	TUBB	2.17 (2.07)	29 (27.67)	0 (0)	0 (0)
IER3 and FLOT1	2.37 (6.18)	6.85 (6.69)	IER3	3.6 (3.43)	9 (9.00)	8 (7.17)	1.70 (1.52)
			FLOT1	1.88 (1.88)	21 (21.00)	3 (3.00)	0.23 (0.23)
DDR1	5.11 (4.94)	5.75 (5.58)	DDR1	2.20 (2.06)	52 (48.67)	6 (6.00)	0.23 (0.23)
GTF2H4	5.07 (4.82)	5.87 (5.60)	GTF2H4	1.30 (1.27)	18 (17.67)	0 (0)	1.30 (1.27)
DPCR1	6.33 (6.12)	6.67 (6.41)	DPCR1	5.22 (5.11)	10 (10.00)	27 (26.17)	3.81 (3.70)
CDSN and C6orf15			CDSN	5.64 (4.76)	46 (37.64)	42 (35.95)	2.69 (2.30)
and PSORS1C1	7.62 (6.91)	7.86 (7.14)	C6orf15	6.46 (5.89)	35 (31.17)	26 (23.00)	2.75 (2.43)
and PSORS1C2			PSORS1C2	4.12 (3.80)	7 (7.00)	10 (8.67)	2.43 (2.11)
C6orf18	7.17 (6.56)	7.71 (7.09)	C6orf18	4.23 (3.53)	59 (51.92)	37 (31.75)	1.63 (1.40)
TCF19 and POU5F1	6.06 (5.63)	6.69 (6.29)	TCF19	3.37 (3.15)	18 (17.75)	17 (14.92)	1.64 (1.44)
			POU5F1	2.12 (1.91)	17 (14.67)	6 (6.00)	0.55 (0.55)
HLA-B	13.47 (9.91)	12.85 (9.63)	HLA-B	15.47 (10.85)	72 (44.50)	96 (74.17)	8.82 (6.81)
MICA/B	9.89 (9.40)	9.17 (8.70)	MICA/B	13.14 (12.61)	44 (41.67)	97 (93.92)	9.04 (8.75)
NFKBIL1 and			NFKBIL1	2.19 (2.11)	20 (20.00)	5 (4.17)	0.44 (0.36)
ATP6V1G2	5.79 (5.53)	6.19 (5.91)	ATP6V1G2	2.24 (2.17)	6 (6.00)	2 (1.75)	0.56 (0.49)
and BAT1			BAT1	0.47 (0.40)	6 (6.00)	0 (0)	0 (0)
and MCCD1			MCCD1	6.39 (6.39)	13 (13.00)	10 (10)	2.78 (2.78)
LTA and TNF	4.59 (4.48)	5.08 (4.97)	LTA	1.93 (1.80)	7 (6.33)	5 (4.83)	0.81 (0.78)
			TNF	2.99 (2.99)	12 (12.00)	9 (9.00)	1.28 (1.28)
LTB	4.97 (4.88)	5.41 (5.29)	LTB	2.87 (2.87)	15 (15.00)	6 (6.00)	0.82 (0.82)
Total	6.44 (6.08)	6.72 (6.38)		3.85 (3.49)	1172 (1080.76)	814 (736.82)	1.58 (1.43)

Genes in italic indicate potential coding regions or pseudogenes. Parentheses shows normalized SNV%. CDR, coding region.
[a] Macaque TRIM31 is a pseudogene.

peri-*HLA-A* (linking *HLA-F* to *HCG8*) and 350-kb peri-*HLA-B/-C* (linking *CDSN* to *3.8-1.1*) segments displayed significant ($P < 0.05$) above-average DIs of 0.69 (0.31)–0.83 (0.31) and 1.01 (0.44)–0.81 (0.34), respectively (Figure 2; Table 1). Intervening loci displayed a lesser-than-average diversity of 0.50 (0.21)–0.27 (0.10) (Figure 2; Table 1). At the opposite end of the spectrum were a number of loci with almost no diversity (<0.26%) (Figure 2; Table 1). Therefore, unexpectedly, the data presented here unveil the remarkable fact that MHC diversity is not limited to antigen/T-cell receptor (TCR) interacting sites of the HLA class I molecules (Bjorkman and Parham 1990), but spreads to the surrounding loci. These data document, the long suspected, but never

TABLE 4

Cross-species SNV analysis

Classification	a	b	c	d	e	hv1	hv2
			Total region (intron + CDR)				
Total SNV%	0.37 (0.12)	0.37	0.18 (0.08)	0.42	4.93	0.07	0.09
Non-MHC genes	0.31 (0.10)	0.37	0.14 (0.06)	0.43	4.88	0.05	0.05
MHC-A and MHC-B	2.71 (0.74)	0.05	1.84 (0.70)	0.07	5.78	0.90	1.74
MHC-E and MHC-F	0.34 (0.10)	0.50	0.42 (0.15)	0.53	6.04	0.10	0.09
			Intron region				
Total SNV%	0.37 (0.13)	0.41	0.18 (0.07)	0.48	5.46	0.06	0.08
Non-MHC genes	0.31 (0.10)	0.42	0.13 (0.06)	0.49	5.42	0.05	0.05
MHC-A and MHC-B	2.50 (0.99)	0.04	1.66 (0.66)	0.06	6.36	0.58	1.47
MHC-E and MHC-F	0.38 (0.12)	0.47	0.38 (0.13)	0.55	6.24	0.09	0.10
			Coding region				
Total SNV%	0.34 (0.10)	0.22	0.19 (0.07)	0.26	2.57	0.12	0.14
Non-MHC genes	0.20 (0.06)	0.21	0.06 (0.03)	0.26	2.40	0.03	0.02
MHC-A and MHC-B	3.43 (1.04)	0.09	2.47 (0.83)	0.09	3.70	2.01	2.70
MHC-E and MHC-F	0.14 (0.04)	0.66	0.56 (0.22)	0.47	5.17	0.14	0.05
			Nonsynonymous				
Total SNV%	0.15 (0.04)	0.08	0.09 (0.04)	0.11	0.97	0.09	0.08
Non-MHC genes	0.16 (0.02)	0.09	0.03 (0.02)	0.11	0.85	0.01	0.01
MHC-A and MHC-B	1.83 (0.60)	0	1.05 (0.38)	0.05	1.92	1.74	1.78
MHC-E and MHC-F	0.09 (0.03)	0.09	0.24 (0.09)	0.24	2.63	0.09	0

Average SNV content is expressed in percentages and is calculated respectively, for all loci and their subgroups and divided in to several categories with respect to coding/noncoding regions. "a," "b," "c," "d," and "e" correspond to the evolutionary stages in Figure 3. CDR, coding region.

proven, hitchhiking diversity within the MHC (Smith and Haigh 1974; Thomson 1977; Nei *et al.* 1997; Takahata and Satta 1998; Satta *et al.* 1999; Gaudieri *et al.* 2000; O'hUigin *et al.* 2000).

To further advance on the extent of this haplotypic diversity, we sequenced the equivalent region in the rhesus macaque, a primate species of prime biomedical relevance as well as the chimpanzee, our closest primate relative. The 3,284,914-bp (3.28-Mb) Mamu class I region, linking *BAT3* to *Mamu-F*, was significantly (1.3 and 1.45 Mb, respectively) larger than the equivalent segments in humans (1.9 Mb) (Shiina *et al.* 1999) and chimpanzees (1.75 Mb) (Anzai *et al.* 2003) (Figure S1 at http://www.genetics.org/supplemental/ and Figure 4) and contained a total of 312 genes (Table S3 at http://www.genetics.org/supplemental/)—45 expressed, 37 potentially expressed—and 230 pseudogenes (Figure S1 and Table S3 at http://www.genetics.org/supplemental/). The region contained a remarkable—at least 64—MHC-I, the reason behind its expansion (Figure 4) in comparison with 18 and 17 class I loci in HLA (humans) and Patr (chimpanzees) regions, respectively (Figure 4). Among the 64 MHC-I, 23 were expressed or putatively expressed, including 5 already known (*Mamu-A1*, *-A2*, *-B12*, *-B15*, and *-AG3*), as well as 18 that were previously unaccounted for, the remaining 41 being pseudogenes. These results are partially by recent independent efforts aimed at unraveling the complexity of macaque MHC (Daza-Vamenta *et al.* 2004; Otting *et al.* 2005). Among the 18 newly identified MHC-I, 12 were novel *Mamu-B*'s, two were assigned as *Mamu-E* and *-F*, and four as *Mamu-AG1*, *-AG2*, *-AG4*, and *-AG5* (Table S6 at http://www.genetics.org/supplemental/). No *-C* locus homolog was identified and, as previously reported, all *Mamu-G* were pseudogenes (Boyson *et al.* 1997) (Figure 4). Upon aligning putative peptide and/or TCR-binding sites of these new Mamu-B with HLA counterparts, one can conclude a major diversification of the peptide-binding repertoire of the species (Table S7 and supporting information at http:www.genetics.org/supplemental/), which will eventually widen the epitope-selection opportunities as related to development of simian immunodeficiency virus/simian–human immunodeficiency virus/human immunodeficiency virus vaccines and help to better understand the cytotoxic T-cell response in this important animal model (Yang 2004).

To further our knowledge of the evolutionary descent of MHC-I alleles, we established the sequence of four chimpanzee class I haplotypes (Anzai *et al.* 2003) (see materials and methods; Tables S4 and S5 at http://www.genetics.org/supplemental/; Table 2). Integration of these data allowed an initial assessment of cross-species SNV content in this immunologically crucial region. The average intrahuman, human–chimpanzee,

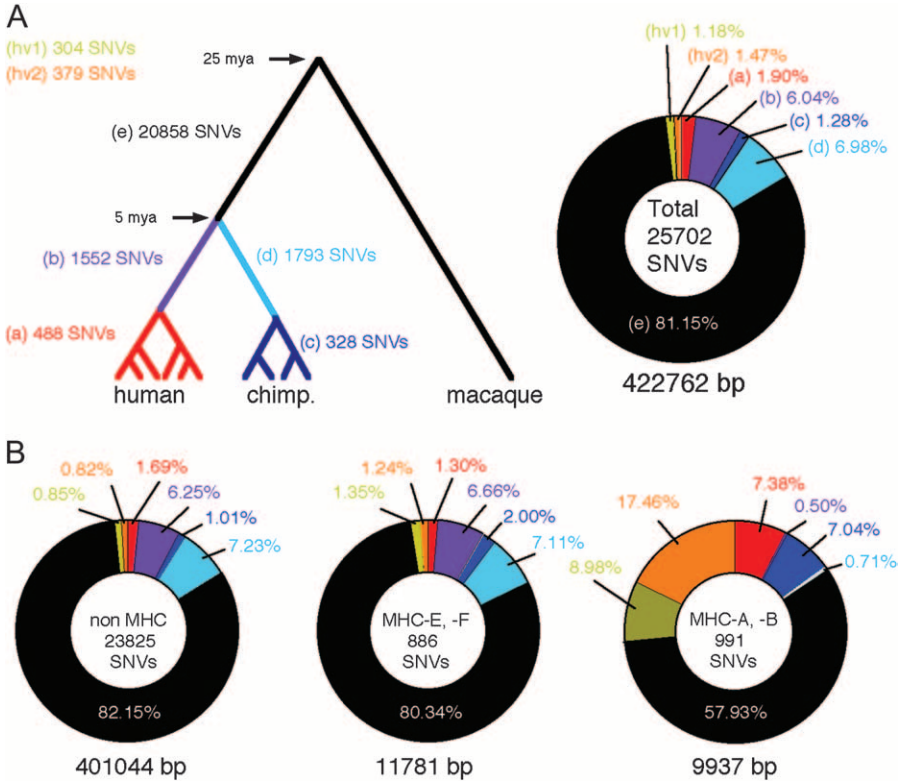

FIGURE 6.—Classification of SNV by species and gene category unveils the origin of MHC diversity. Red designates SNVs generated after birth of four human haplotypes; purple those generated after speciation of chimpanzees and humans but prior to divergence of the six human haplotypes; dark blue designates SNVs generated after birth of four chimpanzee haplotypes; light blue highlights the chimpanzee "counterparts" generated after human–chimpanzee speciation; and yellow and orange refer to hvSNV. (A) Total SNV content. (B) MHC (classical and nonclassical) *vs.* nonMHC SNVs. For definition of hv1 and hv2 SNV, see legend of Figure 3.

and human–macaque degrees of nucleotide differences were 0.53/0.14% (including/excluding indels), 6.90/1.29%, and 29.4/6.55%, respectively (Figures 4 and 5). As depicted in Figure 6A, Table 3, and Table 4, a total of 25,702 normalized SNV were uncovered following pairwise comparison of all 11 sequences (6 human, 4 chimpanzee, and 1 rhesus macaque) using a common template of 422,762 bp. Not surprisingly, the number of SNV correlated well with the evolutionary time of divergence among species (Figure 6A). Figure 6B shows a breakdown of the SNV count in three categories: non-MHC genes (left pie chart); the nonclassical *MHC-E* and *-F* genes (center pie chart, representing "internal controls"; *i.e.*, despite being MHC genes, with at least *MHC-E*-binding peptides, they show little, if any, diversity); and, finally, the classical, polymorphic, antigen-presenting *MHC-A* and *-B* loci (right pie chart). Whereas the left and center pie charts in Figure 6B are very similar, the right pie chart is radically different (Figure 6B; Tables 3 and 4). This is due to two overrepresented categories of SNV in the *MHC-A* and *-B* genes. While the first category, "a" (red) in humans and "c" (dark blue) in chimpanzees, is "haplotype specific," as it is new with respect to the virtual common HLA or Patr framework haplotypes, the second category (hv1 and hv2: yellow and green, respectively) corresponds here to "hypervariable SNVs" (hvSNVs) as they cannot be traced to any particular species and/or haplotype(s). hvSNV are further subdivided into hv1 and hv2. The former is defined by positions encompassing two or more nucleotides while invariant in chimpanzees and macaques and the latter encompasses the remainder (see Figure 3 for definitions). The hv2 fraction, therefore, theoretically includes the trans-species (ts) diversity. However, to formally calculate the number of tsSNV within hv2SNV, one needs to have access to the sequence of syntenic genes within our most recent common ancestor or, in more practical terms, sequence a larger number of haplotypes in all three species. In any event and whatever the fraction of tsSNV within hv2SNV, the totality of hv2SNV is still lesser than *de novo* polymorphism as defined by hsSNV (a + b) + hv1SNV; *i.e.*, 17.46 $<$ 23.4% in Figure 6. These data hence tend to settle the long-debated issue of the origin of MHC diversity where "trans-species polymorphism" (FIGUEROA *et al.* 1988; LAWLOR *et al.* 1988) has been opposed to species-specific *de novo* diversity (BERGSTROM *et al.* 1998). It is also notable that while hsSNV and hv1SNV are enriched within classical MHC-I genes with respect to non-MHC genes ($\times$4.4 for human hsSNV, *i.e.*, "a" in Figure 6; $\times$7 for chimpanzee hsSNV, *i.e.*, "b" in Figure 6; and $\times$10.6 for hv1SNV), framework SNV (those common to human or chimpanzee haplotypes) are diminished by a ratio of 12.1, showing that a great deal of functional MHC diversity is species specific and likely aimed at arming each species against the specific microbiological threat that it faces (Figure 6B). Finally, in humans and chimpanzees, SNV were further divided into those located within coding or noncoding regions. In the former, synonymous (d_S) *vs.* nonsynonymous (d_N) SNV were

further recognized as well as the resulting d_N/d_S ratios (Tables 1 and 2) (Nei and Gojobori 1986). These ratios, however, should be interpreted with caution, because of the weak number of haplotypes analyzed. The application of the McDonald and Kreitman test gave equally nonconclusive results (McDonald and Kreitman 1991).

What is the biological significance of this peri-HLA hitchhiking polymorphism? Examining the MHC class I region sequenced here is revealing. Disease association is not random as most, if not all, diseases mapped to this region are linked directly to *HLA-B/-C* loci and their surroundings (Figure 2). Although for a few of these, the incriminating loci seem to be MHC-I themselves (*e.g.*, ankylosing spondylitis and *HLA-B27*), for most it is becoming ever more clear that the hitchhiked area mentioned above is incriminated (Oka *et al.* 1999; Matsuzaka *et al.* 2002; Okamoto *et al.* 2003) (Figure 2). In contrast few, if any, disease(s) are found to be associated to the equally polymorphic *HLA-A*. In light of what has been presented here, this is no longer unexpected, as the *peri-HLA-A* region is considerably smaller in gene content. Indeed, the region contains only four genes (including an expressed pseudogene), as compared to 18 in the *HLA-B/-C* segment (Figures 1 and 2). This observation seems to remain valid for all the MHC as well. In fact, currently, three diseases have been unequivocally linked to a mutation or indel in the HLA region. These are (in chronological order of their identification): adrenal hyperplasia, which is due to genomic deletions of the MHC class III complement *C4*-linked *21-hydroxylase* gene (White *et al.* 1985); hypotrichosis simplex of the scalp, an autosomal dominant variant of alopecia, caused by nonsense mutations in the *CDSN* (corneodesmosin) gene (Levy-Nissenbaum *et al.* 2003); and the sarcoidosis' HLA-encoded component, which is due to a splice mutation leading to a premature stop codon in the *HLA-DRB*-linked butyrophilin-like 2 (*BTNL2*) (Valentonyte *et al.* 2005). What is the common denominator between these three unrelated mutations? A quick look at the MHC map reveals that they are all within genes located in the immediate vicinity of polymorphic MHC genes: *C4* is the single most polymorphic gene in the MHC class III region, *HLA-DRB1* has the highest level of diversity among MHC-II genes, and finally, *HLA-B/-C* tandem genes are the most polymorphic genes in the genome (MHC Sequencing Consortium 1999).

In summary, we demonstrate that the MHC-I diversity is not limited to the antigen/TCR-binding sites but spreads to surrounding segments. Corroborating data suggest that this hitchhiking diversity, otherwise eliminated by purifying selection in most other genomic sites, has perdured within the MHC perhaps because of the strong biological incentive for constant generation and maintenance of a species-specific diverse allelic repertoire (Hill *et al.* 1991; Kiepiela *et al.* 2004). This was evidenced by the existence of a large reservoir of hvSNV, reminiscent of the fact that most MHC diversity is *de novo* generated and not the result of trans-species inheritance as initially thought (Figueroa *et al.* 1988; Lawlor *et al.* 1988). This result finally puts the MHC in line with the bulk of population and evolutionary genetics data, which firmly conclude that a narrow bottleneck has occurred at the origin of our species (Cann *et al.* 1987; Hammer 1995), a fact inconsistent with massive flow of alleles from one species to the next as required by the trans-species postulate (Ayala *et al.* 1994). Moreover, this fitness in fighting infections seems to have taken its toll on neighboring loci, as the most gene-rich and polymorphic segment around *HLA-B* is also where most MHC-I diseases are mapped to. Finally, this observation is not limited to the MHC class I region as it extends to the other MHC-disease associations and perhaps to the wider genome.

We thank P. Parham, T. Kaneko, A. Kimura, Y. Ishikawa, and F. Numano for cell lines. PGF, COX, and QBL sequences were produced by "Team 50" at the Wellcome Trust Sanger Institute (ftp://ftp.sanger.ac.uk). This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan, Tokai University School of Medicine, "Séquençage à Grande Echelle" [Institut National de la Santé et de la Recherche Médicale (INSRM)/Centre National de la Recherche Scientifique/Ministère de la Recherche], Association de Recherche contre le Cancer, Agence de Biomédecine, as well as an INSRM/Japan Society for the Promotion of Science grant jointly awarded to S. Bahram and H. Inoko.

LITERATURE CITED

Anzai, T., T. Shiina, N. Kimura, K. Yanagiya, S. Kohara *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.

Ayala, F. J., A. Escalante, C. O'Huigin and J. Klein, 1994 Molecular genetics of speciation and human origins. Proc. Natl. Acad. Sci. USA **91:** 6787–6794.

Bahram, S., M. Bresnahan, D. E. Geraghty and T. Spies, 1994 A second lineage of mammalian major histocompatibility complex class I genes. Proc. Natl. Acad. Sci. USA **91:** 6259–6263.

Bergstrom, T. F., A. Josefsson, H. A. Erlich and U. Gyllensten, 1998 Recent origin of HLA-DRB1 alleles and implications for human evolution. Nat. Genet. **18:** 237–242.

Bjorkman, P. J., and P. Parham, 1990 Structure, function, and diversity of class I major histocompatibility complex molecules. Annu. Rev. Biochem. **59:** 253–288.

Boyson, J. E., K. K. Iwanaga, T. G. Golos and D. I. Watkins, 1997 Identification of a novel MHC class I gene, Mamu-AG, expressed in the placenta of a primate with an inactivated G locus. J. Immunol. **159:** 3311–3321.

Cann, R. L., M. Stoneking and A. C. Wilson, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

Ceppellini, R., M. Siniscalco and C. A. Smith, 1955 The estimation of gene frequencies in a random-mating population. Annu. Hum. Genet. **20:** 97–115.

Daza-Vamenta, R., G. Glusman, L. Rowen, B. Guthrie and D. E. Geraghty, 2004 Genetic divergence of the rhesus macaque major histocompatibility complex. Genome Res. **14:** 1501–1515.

Figueroa, F., E. Gunther and J. Klein, 1988 MHC polymorphism pre-dating speciation. Nature **335:** 265–267.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Gaudieri, S., R. L. Dawkins, K. Habara, J. K. Kulski and T. Gojobori, 2000 SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. Genome Res. **10:** 1579–1586.

Hammer, M. F., 1995 A recent common ancestry for human Y chromosomes. Nature **378:** 376–378.

Hill, A. V., C. E. Allsopp, D. Kwiatkowski, N. M. Anstey, P. Twumasi *et al.*, 1991 Common west African HLA antigens are associated with protection from severe malaria. Nature **352:** 595–600.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. **6:** 95–108.

HLA 2004: *Immunobiology of the Human MHC*, 2004 Proceedings of the 13th International Histocompatibility Workshop and Congress, edited by J.A. Hansen and B. Dupont. IHWG Press, Seattle.

Kiepiela, P., A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale *et al.*, 2004 Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. Nature **432:** 769–775.

Kumanovics, A., T. Takada and K. F. Lindahl, 2003 Genomic organization of the mammalian MHC. Annu. Rev. Immunol. **21:** 629–657.

Lawlor, D. A., F. E. Ward, P. D. Ennis, A. P. Jackson and P. Parham, 1988 HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. Nature **335:** 268–271.

Levy-Nissenbaum, E., R. C. Betz, M. Frydman, M. Simon, H. Lahat *et al.*, 2003 Hypotrichosis simplex of the scalp is associated with nonsense mutations in CDSN encoding corneodesmosin. Nat. Genet. **34:** 151–153.

Matsuzaka, Y., S. Makino, K. Okamoto, A. Oka, A. Tsujimura *et al.*, 2002 Susceptibility locus for non-obstructive azoospermia is localized within the HLA-DR/DQ subregion: primary role of DQB1*0604. Tissue Antigens **60:** 53–63.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

MHC Sequencing Consortium, 1999 Complete sequence and gene map of a human major histocompatibility complex. Nature **401:** 921–923.

Nei, M., and T. Gojobori, 1986 Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3:** 418–426.

Nei, M., X. Gu and T. Sitnikova, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94:** 7799–7806.

O'hUigin, C., Y. Satta, A. Hausmann, R. L. Dawkins and J. Klein, 2000 The implications of intergenic polymorphism for major histocompatibility complex evolution. Genetics **156:** 867–877.

Oka, A., G. Tamiya, M. Tomizawa, M. Ota, Y. Katsuyama *et al.*, 1999 Association analysis using refined microsatellite markers localizes a susceptibility locus for psoriasis vulgaris within a 111 kb segment telomeric to the HLA-C gene. Hum. Mol. Genet. **8:** 2165–2170.

Okamoto, K., S. Makino, Y. Yoshikawa, A. Takaki, Y. Nagatsuka *et al.*, 2003 Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. Am. J. Hum. Genet. **72:** 303–312.

Onengut-Gumuscu, S., and P. Concannon, 2002 Mapping genes for autoimmunity in humans: type 1 diabetes as a model. Immunol. Rev. **190:** 182–194.

Otting, N., C. M. Heijmans, R. C. Noort, N. G. de Groot, G. G. Doxiadis *et al.*, 2005 Unparalleled complexity of the MHC class I region in rhesus macaques. Proc. Natl. Acad. Sci. USA **102:** 1626–1631.

Satta, Y., H. Kupfermann, Y. J. Li and N. Takahata, 1999 Molecular clock and recombination in primate MHC genes. Immunol. Rev. **167:** 367–379.

Shiina, T., G. Tamiya, A. Oka, N. Takishima, T. Yamagata *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.

Smith, J. M., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

Stewart, C. A., R. Horton, R. J. Allcock, J. L. Ashurst, A. M. Atrazhev *et al.*, 2004 Complete MHC haplotype sequencing for common disease gene mapping. Genome Res. **14:** 1176–1187.

Takahata, N., and Y. Satta, 1998 Footprints of intragenic recombination at HLA loci. Immunogenetics **47:** 430–441.

Thomson, G., 1977 The effect of a selected locus on linked neutral loci. Genetics **85:** 753–788.

Valentonyte, R., J. Hampe, K. Huse, P. Rosenstiel, M. Albrecht *et al.*, 2005 Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat. Genet. **37:** 357–364.

White, P. C., D. Grossberger, B. J. Onufer, D. D. Chaplin, M. I. New *et al.*, 1985 Two genes encoding steroid 21-hydroxylase are located near the genes encoding the fourth component of complement in man. Proc. Natl. Acad. Sci. USA **82:** 1089–1093.

Yang, O. O., 2004 CTL ontogeny and viral escape: implications for HIV-1 vaccine design. Trends Immunol. **25:** 138–142.

Communicating editor: N. Takahata