Immunogenetics (2008) 60:1–18
DOI 10.1007/s00251-007-0262-2

ORIGINAL PAPER

Variation analysis and gene annotation of eight MHC haplotypes: The MHC Haplotype Project

Roger Horton · Richard Gibson · Penny Coggill · Marcos Miretti · Richard J. Allcock · Jeff Almeida · Simon Forbes · James G. R. Gilbert · Karen Halls · Jennifer L. Harrow · Elizabeth Hart · Kevin Howe · David K. Jackson · Sophie Palmer · Anne N. Roberts · Sarah Sims · C. Andrew Stewart · James A. Traherne · Steve Trevanion · Laurens Wilming · Jane Rogers · Pieter J. de Jong · John F. Elliott · Stephen Sawcer · John A. Todd · John Trowsdale · Stephan Beck

Received: 28 August 2007 / Accepted: 29 October 2007 / Published online: 10 January 2008

Abstract The human major histocompatibility complex (MHC) is contained within about 4 Mb on the short arm of chromosome 6 and is recognised as the most variable region in the human genome. The primary aim of the MHC Haplotype Project was to provide a comprehensively annotated reference sequence of a single, human leukocyte antigen-homozygous MHC haplotype and to use it as a basis against which variations could be assessed from seven other similarly homozygous cell lines, representative of the most common MHC haplotypes in the European population. Comparison of the haplotype sequences, including four haplotypes not previously analysed, resulted in the identification of >44,000 variations, both substitutions and indels (insertions and deletions), which have been submit-

Horton and Gibson contributed equally to this work.

R. Horton · R. Gibson · P. Coggill · M. Miretti · J. Almeida · S. Forbes · J. G. R. Gilbert · J. L. Harrow · E. Hart · D. K. Jackson · S. Palmer · S. Sims · S. Trevanion · L. Wilming · J. Rogers · S. Beck
Wellcome Trust Sanger Institute, Genome Campus,
Hinxton, Cambridge CB10 1SA, UK

K. Howe
CRUK Cambridge Research Institute,
Robinson Way,
Cambridge CB2 0RE, UK

R. J. Allcock
School of Surgery and Pathology,
University of Western Australia,
Nedlands 6009 WA, Australia

K. Halls
The Wellcome Trust/Cancer Research UK Gurdon Institute,
The Henry Wellcome Building of Cancer and Developmental Biology, University of Cambridge,
Tennis Court Road,
Cambridge CB2 1QN, UK

C. A. Stewart
National Cancer Institute,
P.O. Box B., 567/206, Frederick, MD 21702, USA

P. J. de Jong
Children's Hospital Oakland Research Institute,
Oakland, CA 94609-1673, USA

J. F. Elliott
Alberta Diabetes Institute (ADI), Department of Medical Microbiology and Immunology, Division of Dermatology and Cutaneous Sciences, University of Alberta,
Edmonton AB T6G 2H7, Canada

S. Sawcer
Department of Clinical Neurosciences, University of Cambridge,
Addenbrooke's Hospital, Hills Road,
Cambridge CB2 2QQ, UK

A. N. Roberts · J. A. Todd
Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics,
Cambridge Institute for Medical Research,
University of Cambridge, Wellcome Trust/MRC Building,
Addenbrooke's Hospital,
Cambridge CB2 0XY, UK

ted to the dbSNP database. The gene annotation uncovered haplotype-specific differences and confirmed the presence of more than 300 loci, including over 160 protein-coding genes. Combined analysis of the variation and annotation datasets revealed 122 gene loci with coding substitutions of which 97 were non-synonymous. The haplotype (A3-B7-DR15; PGF cell line) designated as the new MHC reference sequence, has been incorporated into the human genome assembly (NCBI35 and subsequent builds), and constitutes the largest single-haplotype sequence of the human genome to date. The extensive variation and annotation data derived from the analysis of seven further haplotypes have been made publicly available and provide a framework and resource for future association studies of all MHC-associated diseases and transplant medicine.

Keywords Major histocompatibility complex · Haplotype · Polymorphism · Retroelement · Genetic predisposition to disease · Population genetics

Introduction

The MHC has long been believed to be the most important region in the human genome with respect to infection, inflammation, autoimmunity and transplant medicine (Lechler and Warrens 2000). This was recently confirmed by the largest genome-wide association study carried out to date for seven common diseases, including two autoimmune diseases (type 1 diabetes and rheumatoid arthritis) and one inflammatory disease (Crohn’s disease). The highest associations were found between the MHC and these two autoimmune diseases (The Wellcome Trust Case Control Consortium 2007). The complex aetiology of MHC-associated disease coupled with high density, polymorphism, linkage disequilibrium (LD) and frequent non-Mendelian inheritance of gene loci have made it challenging to identify variations that cause or contribute to disease phenotypes. Additional limiting factors have been our incomplete knowledge of the allelic variation of genes and regions flanking the nine classical human leukocyte antigen (HLA) loci and the lack of a single haplotype reference sequence, the original reference sequence being a composite of multiple MHC haplotypes (Mungall et al. 2003; The MHC Sequencing Consortium 1999).

Recognizing that the future identification of variants conferring susceptibility to common disease is critically dependent on fully informative polymorphism and haplotype maps, the MHC Haplotype Consortium formed in 2000 with the aim to generate these critical data and to make them publicly available as a general resource for MHC-linked disease studies. Similar efforts, but with different experimental approaches, were also carried out in Japan (Shiina et al. 2006) and the USA (Smith et al. 2006). To develop the resource, eight HLA-homozygous MHC haplotypes were selected on the basis of conferring either protection against or susceptibility to two autoimmune diseases, type 1 diabetes and multiple sclerosis, and that represented common haplotypes in European populations. In the subsequent years, incremental data, materials and tools comprising this resource have been released (Allcock et al. 2002; Horton et al. 2004; Stewart et al. 2004; Traherne et al. 2006) and have contributed towards the construction of a high-resolution LD map and a first generation of HLA tag single nucleotide polymorphisms (SNPs; de Bakker et al. 2006; Miretti et al. 2005) and the identification of a second MHC susceptibility locus for multiple sclerosis (The International Multiple Sclerosis Genetics Consortium; Yeo et al. 2007). In this paper, we report the final account of this international effort, including, analysis of the last four of the eight haplotypes, up-to-date variation statistics, gene annotation, population-specific aspects and a detailed description of the databases and tools for viewing and accessing the data in the context of existing genome annotation.

Materials and methods

Variation analysis

The method previously reported for comparison of MHC haplotype sequences (Stewart et al. 2004; Traherne et al. 2006) was extended to cover all eight haplotypes. Briefly, the most suitable method proved to be a clone by clone comparison using the discrepancy-list option of the cross_match program (Green, unpublished; http://www.phrap.org/), an implementation of the Smith–Waterman sequence alignment algorithm (Smith and Waterman 1981), using the alignment of a haplotype clone sequence with the appropriate overlapping reference sequence from a PGF clone or clones. All variations were submitted to dbSNP using the submitter handle SI_MHC_SNP and user identifiers of the form [PGF BAC clone sequence version]_[position in PGF BAC clone

J. A. Traherne · J. Trowsdale
Department of Pathology, Immunology Division,
University of Cambridge,
Cambridge CB2 1QP, UK

Present address:
S. Beck (✉)
UCL Cancer Institute, University College London
72 Huntley Street,
London WC1E 6BD, UK
e-mail: s.beck@ucl.ac.uk

sequence]_[variation change]. Thus, AL662890.3_6645_ TC indicates a substitution in which the base T at base position 6645 in AL662890.3 (PGF BAC 308K3) was substituted by C in the other haplotype. In the case of indels, the 'variation change' consists of 'i' or 'd' (for insertion or deletion), followed by a numerical value for the length of the indel, in turn followed by the inserted or deleted sequence if this were of 12 or fewer bases. For longer indels, an *X* value is given, which refers to a look-up table (http://www.sanger.ac.uk/HGP/Chr6/MHC/Xfile). Thus, AL662890.3_7470_d8TACACACA indicates a deletion in AL662890.3 after base 7470 of the eight bases 'TACACACA'. Further, AL662890.3_10559_i5A-TATT indicates an insertion in AL662890.3 starting after base 10559 of the five bases 'ATATT'. AL662890.3_7475_ d14X1 indicates a 14-base deletion after base 7475 in AL662890.3 of a sequence coded as X1 which is 'ATACACACACACAC'.

Major indel sequences, appearing as breaks in the cross_match discrepancy lists between two clones from difference haplotypes, were extracted and subjected to analysis by RepeatMasker to detect the presence of retrotransposible elements.

Gene annotation

The finished genomic sequence for each of the eight haplotypes was analysed using a modified Ensembl pipeline (Searle et al. 2004). CpG islands were predicted on unmasked sequence. Interspersed and tandem repeats were masked out by RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996–2004, http://www.repeatmasker.org) and Tandem Repeats Finder (TRF; Benson 1999), respectively. The sequence was then BLAST searched (BLAST, basic local alignment search tool; Altschul et al. 1990) using a vertebrate set of complementary DNAs (cDNAs) and expressed sequence tags (ESTs) from the European Molecular Biology Laboratory (EMBL) nucleotide database (Kulikova et al. 2007), followed by the re-alignment of significant hits. Non-redundant proteins were aligned similarly. Protein domain matches were provided through alignment of Pfam to the genomic sequence using Genewise (Birney et al. 2004), thereby providing protein domain data to the annotator. Ab initio gene predictions were performed by Genscan (Burge and Karlin 1997) and Fgenesh (Salamov and Solovyev 2000), and potential transcriptional start sites were predicted by Eponine (Down and Hubbard 2002). Analysis results were displayed, and annotation was performed through an in-house annotation software system.

Genes were manually annotated according to the human and vertebrate analysis and annotation (HAVANA) guidelines (http://www.sanger.ac.uk/HGP/havana/) using evidence based on comparison with external databases as of August 2005. All gene structures are supported by transcriptional evidence, either from cDNA, EST, or protein. In general, annotations are supported by best-in-genome evidence. Haplotype-specific evidence is assigned where possible. As with previous MHC annotation (Stewart et al. 2004; Traherne et al. 2006), some olfactory receptors have been built upon protein homology alone because of their restricted expression.

Locus and variant types were annotated according to established standards (Harrow et al. 2006), with the modification that, within the MHC region, the artefact locus has been used to tag historically annotated structures that are no longer deemed valid.

HLA-A, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQA1*, and *HLA-DQB1* allele types were assessed by comparison against the IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla/; Marsh et al. 2005).

Annotation status of haplotypes

The PGF, COX, and QBL haplotypes have already been annotated in detail (Stewart et al. 2004; Traherne et al. 2006). It was decided, however, to re-annotate and update this annotation to maintain consistency between all eight haplotypes with the current supporting evidence and pipeline analyses. The SSTO haplotype was manually annotated de novo. The new annotation from the PGF haplotype was projected through a DNA–DNA alignment to each of the remaining haplotypes (APD, DBB, MANN and MCF) where possible. This projection was checked thoroughly and non-alignable regions were manually adjusted (including the C4 and HLA-DRB1 hypervariable regions). Polyadenylation sites and signals were not annotated for haplotypes APD, DBB, MANN and MCF because of time constraints. In the main, however, these features may be assumed to correspond to the same positions as in the first four haplotypes.

Combination of variation and annotation data

By employing a series of Perl scripts, the array of haplotype variation was combined with the annotation of gene loci, repeat elements and microsatellites, extracted from the Vertebrate Genome Annotation (VEGA) database in general feature format (GFF; http://www.sanger.ac.uk/Software/formats/GFF/), to determine the variation status of all loci.

Distribution of sequenced HLA haplotypes in Europeans

To assess the distribution of sequenced haplotypes at the population level, 180 founder haplotypes were reconstructed using genotypic data from Centre d'Etude Polymorphisme Humain (CEPH) trios (de Bakker et al. 2006). A ~214 kb segment spanning the *HLA–DRB1–DQB1* genes

Table 1 Haplotype sequence contig length, number of gaps and HLA allele types

Haplotype	Length (bp)	Gaps	HLA-A	HLA-B	HLA-C	HLA-DQA1	HLA-DQB1	HLA-DRB1
PGF	4754829	0	A*03010101	B*070201	Cw*07020103	DQA1*010201	DQB1*0602	DRB1*150101
COX	4731878	0	A*01010101	B*080101	Cw*070101	DQA1*050101	DQB1*020101	DRB1*030101
QBL	4249272	5	A*260101	B*180101	Cw*050101	DQA1*050101	DQB1*020101	DRB1*030101
APD	4160965	16	A*01010101	–	–	–	–	–
DBB	2330101	28	A*02010101	–	Cw*06020101	DQA1*0201	DQB1*030302	DRB1*070101
MANN	4191014	10	A*290201	B*440301	Cw*160101	DQA1*0201	DQB1*0202	DRB1*070101
MCF	4087413	15	[A*020101]	B*15010101	Cw*030401	DQA1*0303	DQB1*030101	–
SSTO	3704249	22	A*320101	B*44020101	Cw*050101	DQA1*030101	DQB1*030501	DRB1*040301

Sequence length (bp) and number of gaps in each haplotype sequence, together with the HLA gene types obtained by BLAST against the IMGT/HLA database. Dashes or data in square brackets indicate the absence or the partial presence, respectively, of a gene owing to a sequence gap

was selected for the analyses. This segment, represented by 54 SNPs, is delimited by rs2187823 and rs2856691, with NCBI build 36 chromosome 6 coordinates 32547486 and 32761413, respectively. Phased haplotypes with known *HLA–DRB1–DQB1* alleles were then used to construct a neighbor-joining tree (Kumar et al. 2001) and a phylogenetic network (Bandelt et al. 1999).

Resources

All sequences presented in this paper have been submitted to the EMBL/GenBank/DNA Data Bank of Japan (DDBJ) database and allocated accession numbers. For clarity, all bacterial artificial chromosome (BAC) clones are referred to using their accession numbers. The annotation of each haplotype has been entered in the VEGA database and is accessible through its browser (http://www.VEGA.sanger.ac.uk). All variations from the study were submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP) using the submitter handle SI_MHC_SNP.

BAC clones from the CHORI-501 (PGF) and CHORI-502 (COX) libraries can be requested from BACPAC resources (http://www.bacpac.chori.org/). Clones from the other libraries can be requested from john.elliott@ualberta.ca.

The web site for the MHC Haplotype Project provides links to various data resources (http://www.sanger.ac.uk/HGP/Chr6/MHC/).

DAS sources for all substitutions and indels are available from http://www.das.ensembl.org/das as follows:

- ens_35_COX_SNP ens_35_COX_DIP
- ens_35_QBL_SNP ens_35_QBL_DIP
- ens_35_SSTO_SNP ens_35_SSTO_DIP
- ens_35_APD_SNP ens_35_APD_DIP
- ens_35_DBB_SNP ens_35_DBB_DIP
- ens_35_MANN_SNP ens_35_MANN_DIP
- ens_35_MCF_SNP ens_35_MCF_DIP

These can be accessed via the VEGA browser.

Results and discussion

Variation analysis

One of the main aims of the MHC Haplotype Project was to generate a comprehensive variation map of this most variable region of the human genome. To achieve this, eight haplotypes were sequenced and subjected to variation analysis. Table 1 details the lengths of the sequence contigs, the number of sequence gaps and the allelic types of major HLA loci for each haplotype. Of the eight haplotypes sequenced, three have already been described: PGF and Cox (Stewart et al. 2004) both of which formed single contigs of approximately 4.7 Mb, and QBL (Traherne et al. 2006), of approximately 4.2 Mb but with five gaps. The remaining haplotypes sequenced all contained gaps, their coverage ranging from 2.33 Mb (DBB with 28 gaps) to 4.19 Mb (MANN with 10 gaps).

For the variation analysis, each of the above haplotypes was compared with the PGF reference sequence, resulting in the identification of 44,544 variations (37,451 substitutions and 7,093 indels, Table 2), which have all been submitted to dbSNP. The success of this exercise is illustrated by the fact that examination of this public

Table 2 Distribution of substitutions and indels amongst haplotypes

Haplotype	Substitutions	Indels	ALL
COX	15,967	2,393	18,360
QBL	15,282	2,360	17,642
SSTO	14,982	2,300	17,282
APD	4,230	683	4,913
DBB	14,255	1,975	16,230
MANN	12,102	1,654	13,756
MCF	10,790	1,545	12,335
Overall	37,451	7,093	44,544

Number of variations found by comparing the PGF haplotype sequence with each of the other haplotype sequences in turn

Table 3 Distribution of substitutions and indels within different sequence regions amongst haplotypes

Sequence region	Base pairs	COX		QBL		SSTO		APD		DBB		MANN		MCF	
		S	ID	S	ID	S	ID	S	ID	S	ID	S	ID	S	ID
Coding	247,505	353	8	503	19	380	2	74	0	351	6	401	9	348	2
UTR	155,960	382	34	438	59	331	35	38	9	326	39	303	35	309	31
Intronic	1,283,472	3,141	571	3,135	590	2,658	505	602	147	2,897	509	2,185	393	2,126	404
Total intragenic	1,686,937	3,876	613	4,076	668	3,369	542	714	156	3,574	554	2,889	437	2,783	437
Pseudogenic	57,223	235	15	226	21	227	19	101	8	191	10	109	6	113	10
Pseudogenic intron	63,108	507	54	220	27	215	18	158	20	258	22	98	13	179	13
Transcript exon	78,092	190	30	207	33	119	22	71	8	136	17	88	16	70	15
Transcript intron	332,705	1,243	197	1,186	216	1,053	155	85	29	1,245	192	1,081	161	268	53
REPEATS:															
LINEs	608,429	2,110	221	2,015	240	2,388	255	755	93	2,097	217	2,084	193	1,530	164
SINEs	428,567	1,381	428	1,316	401	1,311	385	346	134	1,229	318	928	241	936	271
Other repeats	487,863	2,605	207	2,518	229	2,514	207	925	56	2,748	199	2,198	177	2,170	169
Total in repeats	1,524,859	6,096	856	5,849	870	6,213	847	2,026	283	6,074	734	5,210	611	4,636	604
Microsatellite	15,185	186	168	95	85	222	198	14	29	60	76	61	71	90	68
All above	3,297,590	12,333	1,933	11,859	1,920	11,418	1,801	3,169	533	11,538	1,605	9,536	1,315	8,139	1,200
Other intergenic	996,720	3,634	460	3,423	440	3,564	499	1,061	150	2,717	370	2,566	339	2,651	345
Total	4,754,829	15,967	2,393	15,282	2,360	14,982	2,300	4,230	683	14,255	1,975	12,102	1,654	10,790	1,545

Variations shown in Table 2 ascribed to sequence regions identified during annotation. These included exonic, UTR and intronic regions of coding; pseudogenic and transcript loci; repeat elements, microsatellites and other intergenic regions

S Substitution, *ID* indel

database (NCBI dbSNP build 127, March 2007) showed that there were only a further 19,598 variations, submitted by other laboratories, in this region which were not identified by this project. In accordance with the annotation that we also generated for each haplotype (see below), the variations shown in Table 2 were further classified as untranslated region (UTR), exonic, intronic, intergenic and eight more sub-categories (Table 3). Coding substitutions, which are of particular interest with respect to altered functionality, were further classified as synonymous, non-synonymous conservative, or non-synonymous non-conservative and grouped depending on whether they affected HLA or other genes (Table 4). The actual variations and affected amino acids can be viewed using the VEGA browser as illustrated in Fig. 1 and described in the corresponding section later on. In addition, we have analysed all haplotype sequences for inversions, which represent another important variation category that has been linked to genomic disorders (Shaw and Lupski 2004). Using Ssaha2 (Ning et al. 2001), we found no evidence of any inversion polymorphism within the generated sequences but could not exclude large-scale (e.g. involving entire MHC) inversions with breakpoints outside the MHC regions sequenced here.

Gene annotation

There have been several previous annotations of the gene content of the MHC (Horton et al. 2004; Mungall et al. 2003; Stewart et al. 2004; The MHC Sequencing Consortium 1999; Traherne et al. 2006). The maximum region annotated in this study extends from the telomeric *ZNF452* gene in the MHC extended class I region (COX haplotype) to the centromeric *ZBTB9* gene just telomeric of the MHC extended class II region (PGF and SSTO haplotypes). The PGF haplotype (Stewart et al. 2004) remains the longest complete MHC haplotype, encompassing 320 annotated loci with 1,267 variants. The number of variants ascribed to each locus-type is listed in Table 5. A comparison of the statistics for loci in each haplotype is shown in Table 6.

VEGA database and browser

The VEGA database provides access to gene annotation of the eight MHC haplotype sequences, a valuable public resource and a means of integrating annotation and variation data. The VEGA database also provides the facility to download nucleotide or peptide sequences for genes of interest, by selecting ‘export cDNA’ or ‘export peptide’ from the menu obtained by clicking on gene cartoons in the VEGA ‘detailed view’ or ‘basepair view’ window. From these, any desired alignments can be made. Variation data may be viewed in the browser linked to a distributed annotation system (DAS) source of any given

Table 4 Codon variation caused by substitutions in HLA and other gene loci

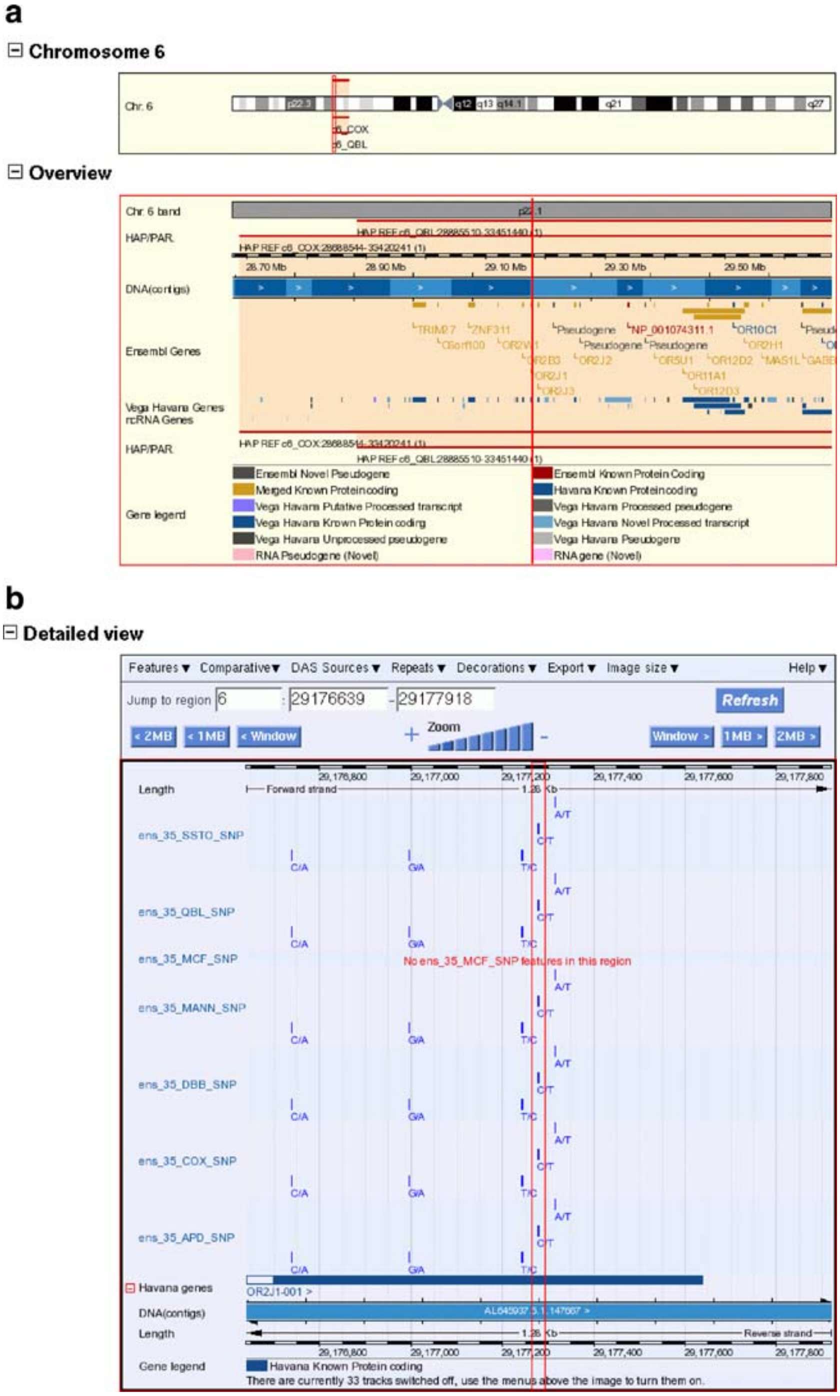
Codons variation by virtue of substitutions		COX			QBL			SSTO			APD			DBB			MANN			MCF		
		HLA	Other	Total	HLA	Other	Total	HLA	Other	Total	HLA	Other	Total	HLA	Other	Total	HLA	Other	Total	HLA	Other	Total
Synonymous		49	81	130	71	106	177	72	57	129	1	24	25	66	69	135	59	79	138	80	52	132
Non-synonymous	Total	125	76	201	184	121	305	164	72	236	19	27	46	120	76	196	144	91	235	147	56	203
	Conservative	68	42	110	102	72	174	92	39	131	11	18	29	67	40	107	77	60	137	82	35	117
	Non-conservative	57	34	91	82	49	131	72	33	105	8	9	17	53	36	89	67	31	98	65	21	86
Total		174	157	331	255	227	482	236	129	365	20	51	71	186	145	331	203	170	373	227	108	335

Coding substitutions analysed for their effects on protein sequences and listed in by haplotype for HLA genes (*HLA-A HLA-B HLA-C HLA-DRB1 HLA-DRA HLA-DQA1 HLA-DQB1 HLA-DPA1 HLA-DPB1*) and for all other genes according to the changes they induced in codons as either synonymous, non-synonymous conservative, or non-synonymous non-conservative changes variation (see “Materials and methods”). This is illustrated An example of the use of this browser to view a C to T substitution is illustrated for the OR2J1 locus (Fig. 1). An overview of the genomic environment is given in Fig. 1a, showing the gene within a cluster of olfactory gene loci on chromosome 6. The detailed view (Fig. 1b) shows *OR2J1* with associated variations in all haplotypes. The basepair view (Fig. 1c) illustrates the presence of the C/T substitution in all haplotypes except MCF, and its positioning above the translated sequence, at the first position of a CAG codon, indicating the presence of a stop codon instead of glutamine.

Annotation changes

In addition to loci annotated in the previous studies, newly recognised with official Hugo Gene Nomenclature Committee (HGNC) symbols have also been annotated. These

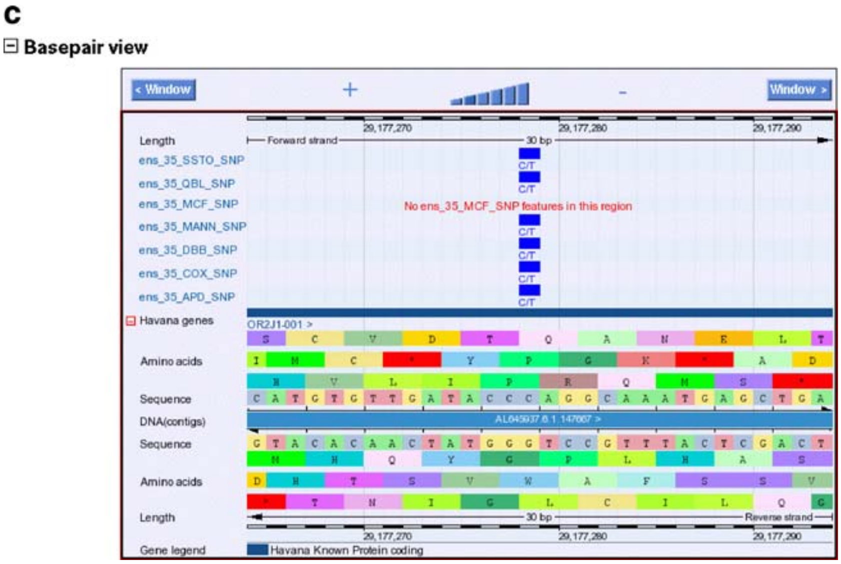
Fig. 1 Annotation and variation data in VEGA. VEGA ‘overview’ (**a**), ‘detailed view’ (**b**) and ‘basepair view’ (**c**) example of the variation in the *OR2J1* locus in which a STOP codon is present in all haplotypes except MCF



No ens_35_MCF_SNP features in this region

There are currently 33 tracks switched off, use the menus above the image to turn them on.

Fig. 1 (continued)



have included the mitochondrial coiled–coil domain protein 1 gene *MCCD1* (Semple et al. 2003) and the related unprocessed pseudogenes *MCCD1P1* and *MCCD1P2*, as well as the zinc-finger and BTB domain-containing protein gene *ZBTB9*, annotated at the very centromeric boundary of the sequenced region.

The *C6orf21* gene (De Vet et al. 2003; XXbac–BPG32J3.17-001) of the MHC class III region was annotated as a separate locus from the adjacent centromeric locus *LY6G6D* (splice variants XXbac–BPG32J3.4-001 and XXbac–BPG32J3.4-002). There was, however, a further coding splice variant of *LY6G6D* (XXbac–BPG32J3.4-004), which spanned not only the other *LY6G6D* splice variants but also *C6orf21*, suggesting that this is a possible so-called chimeric transcript (Parra et al. 2006).

HLA-DRB1 hypervariable region

Of the five newly annotated MHC haplotypes, APD alone exhibited the *HLA–DRBDR52* antigenic specificity found on DRB1*3, DRB1*05 (DRB1*11 and DRB1*012) and DR6 (DRB1*13 and DRB1*14) haplotypes and encoded by *HLA–DRB3*, whereas the remainder (SSTO, DBB, MANN and MCF) exhibited the *DR53* specificity, encoded by *HLA–DRB4*, here annotated for the first time in genomic sequence. The *HLA–DRB53* sequences included three known loci (*HLA–DRB4*, *HLA–DRB7* and *HLA–DRB8*), as well as three novel pseudogenes (*DASS–218M11.1*, *DASS–23B5.1* and *DASS–23B5.2*). *DASS–23B5.1* corresponds to a pseudogene derived from the gene for the protein kinase, interferon-inducible double-stranded RNA dependent activator (Chida et al. 2001) for which the symbol *PRKRAP1* has now been recognised. A further processed pseudogene, *FAM8A5P* (Jamain et al. 2001), was also annotated in the DR53 specificity.

Table 5 Splice-variant statistics for PGF annotation

Type	No.
Total splice variants	1,267
Coding	**523**
Unprocessed_pseudogene	50
Processed_pseudogene	41
Expressed_pseudogene	7
Transcript	271
Putative	71
Retained_intron	263
Nonsense_mediated_decay	30
Artefact	11
Total loci	320

Splice variants annotated in the PGF haplotype

HLA-V and HLA-P

Our analysis showed that the two unprocessed class I pseudogenes *HLA-V* and *HLA-P* (previously *HLA-75* and *HLA-90*, Geraghty et al. 1992) should in fact be merged together; individually they merely represented the 5′ and 3′ portions of a single unprocessed pseudogene, separated by repeat elements. According to our annotation guidelines (see “Materials and methods”), the newly merged locus was assigned the symbol from the 3′ component, in this case, *HLA-P*. Best-in-genome nucleotide evidence was found to support five transcript variants at the 5′ end, which, together with evidence for continued locus-transcription, led us to designate the locus as a transcribed pseudogene. Because transcription appears to still occur at this locus, it was, therefore, designated as a transcribed pseudogene. A further six expressed pseudogenes were identified in the MHC region (*HLA–DPB2, HLA-J, CYP21A1P, HLA–DRB6, HLA–L* and *PPP1R2P1*).

Table 6 Gene annotation statistics for eight MHC haplotypes

Locus type	PGF	COX	QBL	SSTO	APD	DBB	MANN	MCF
Coding	165	159	150	131	82	146	129	150
Transcript	28	28	26	26	19	26	27	22
Putative	18	18	15	15	6	16	12	14
Pseudogenes total	98	95	93	98	59	92	95	75
Unprocessed	50	48	48	53	36	52	53	42
Processed	41	42	40	39	19	34	37	28
Expressed	7	5	5	6	4	6	5	5
Artefact	11	11	10	11	0	0	0	0
Total loci	320	311	294	281	166	281	264	261
Total variants	1,267	1,191	1,155	1,058	568	1,138	960	1,115

Annotation statistics for loci in each haplotype. For definitions of locus types see "Materials and methods"

RCCX hypervariable region

This module within the MHC class III region, named for its gene content (*RP-C4A/B-CYP21-TNXB*), may be duplicated or triplicated (Chung et al. 2002), and the pseudogenes *CYP21A1P*, *TNXA* and *STK19P* contain the complement component gene, *C4*, in either or both of the two versions, *C4A* and *C4B* (Awdeh and Alper 1980). This gene may also be present in either long (*C4A*L, *C4B*L) or short (*C4A*S, *C4B*S) forms depending on the presence or absence of an inserted HERVC4 element in intron 9. Contrary to our previous annotation (Stewart et al. 2004) see also legend to (Fig. 2), the PGF haplotype now appears to possess an arrangement in which C4AL precedes C4BL, whereas COX has a single module with *C4B*S and QBL has a single module with *C4A*S (Traherne et al. 2006). For the new haplotype sequences reported in this paper, SSTO was bimodular with two copies of *C4B*L, whereas DBB was bimodular with *C4A*L followed by *C4B*S. Although a sequence gap was present in MCF, this haplotype appeared to be bimodular in that, although the telomeric copy of the *C4* gene could not be identified, there was evidence for the pseudogenes *CYP21A1P*, *TNXA* and *STK19P* in a telomeric module. The second centromeric module in MCF contained *C4A*L. The RCCX region in the APD and MANN haplotypes was incomplete because of sequence gaps.

C6orf205

Variability in the *C6orf205* gene has been reported to consist of extension of the minisatellite in exon 2 from 27 copies in PGF and COX to 31 copies in QBL (Traherne et al. 2006). In the newly annotated haplotypes, we found the minisatellite to extend to 29 in MANN. The APD, DBB and MCF possessed 27 copies. There was a sequence gap in this region in the SSTO haplotype.

Table 7 Other newly annotated loci

Locus	Locus type
XXbac-BCX196D17.5	Transcript
XXbac-BPG116M5.14	Putative
XXbac-BPG116M5.15	Putative
XXbac-BPG116M5.16	Putative
XXbac-BPG118E17.9	Putative
XXbac-BPG126D10.10	Processed pseudogene
XXbac-BPG126D10.11	Processed pseudogene
XXbac-BPG13B8.10	Transcript
XXbac-BPG13B8.9	Unprocessed pseudogene
XXbac-BPG154L12.4	Putative
XXbac-BPG181B23.4	Transcript
XXbac-BPG181M17.4	Putative
XXbac-BPG246D15.8	Transcript
XXbac-BPG248L24.10	Unprocessed pseudogene
XXbac-BPG248L24.9	Processed pseudogene
XXbac-BPG249D20.9	Putative
XXbac-BPG250I8.13	Transcript
XXbac-BPG254F23.5	Putative
XXbac-BPG254F23.6	Putative
XXbac-BPG254F23.7	Transcript
XXbac-BPG254F23.7	Putative
XXbac-BPG27H4.7	Transcript
XXbac-BPG27H4.8	Transcript
XXbac-BPG294E21.7	Processed pseudogene
XXbac-BPG296P20.14	Putative
XXbac-BPG296P20.15	Putative
XXbac-BPG299F13.14	Putative
XXbac-BPG308J9.3	Transcript
XXbac-BPG308K3.5	Putative
XXbac-BPG308K3.6	Transcript
XXbac-BPG309N1.15	Unprocessed pseudogene
XXbac-BPG32J3.18	Putative
XXbac-BPG8G10.2	Unprocessed pseudogene
DAQB-12N14.5	Transcript
DAQB-331I12.5	Putative
DAQB-335A13.8	Transcript

Newly annotated loci without HGNC symbols

Table 8 Haplotype variation at splice sites

Gene	Variant	Affected exons	Donor*	Acceptor*	dbSNP cluster ID	Best evidence	PGF	QBL	COX	SSTO	DBB	APD	MANN	MCF
TRIM31	2	3/4	ggt	t**g**g	rs28400887	cDNA	NC	NC	NC	C	ND	NC	NC	C
TRIM31	5	2/3	ggt	t**g**g	rs28400887	EST	NC	NC	NC	C	ND	NC	NC	C
C4B	7	3/4	ggt	cgg	–	EST	NC	ND	NC	C	NC	ND	ND	ND
C4A	7	3/4	ggt	cgg	–	EST	NC	NC	ND	C	NC	ND	ND	NC
HLA-DQA1	4	4/5	ggt	c**g**g	rs707947	cDNA	C	C	C	NC	NC	ND	NC	NC
HLA-DQA1	5	4/5	ggt	ta**a**/ca**a**	rs3667	cDNA	NC	NC	NC	C	C	ND	C	C
HLA-DRB1	2	2/3	g**a**t	cag	rs9271083	EST	NC	C	C	C	C	ND	C	ND

Gene loci and variants that are affected by disruptive variations at splice sites. *C* Canonical splice site (donor=ngt; acceptor=nag), *NC* non-canonical, and *ND* no data (gene absent or gap). Donor and acceptor variable nucleotides in **bold** with equivalent dbSNP cluster ID number given in column to right. The C4A and C4B genes are, for these purposes, effective duplicates of each other. The two TRIM31 variants share the same splice site (but differ elsewhere in structure). The two HLA–DQA variants share the same donor but have alternative acceptors. Note the mutually exclusivity of these variants amongst the haplotypes (Hoarau et al. 2004; Hoarau et al. 2005)

MICA

The known allelic polymorphism of *MICA* reported for the DRB1*03 QBL cell line sequence, in which a four-base insertion (GCGT) extended the open reading frame in coding exon 5 haplotype (Traherne et al. 2006), was also present in the DRB1*07 MANN haplotype. The insertion was absent from PGF, COX and SSTO. No sequence was available in APD, DBB and MCF for this gene.

PPP1R2P1

The intronless pseudogene *PPP1R2P1* reported to have a full-length open reading frame in the PGF, COX and QBL haplotypes (Stewart et al. 2004; Traherne et al. 2006) was found to have a similar open reading frame in the DBB and MANN haplotypes but to have the frameshift mutation seen in the original chromosome reference sequence (Mungall et al. 2003) in the SSTO, APD and MCF haplotypes.

PSORS1C1

The QBL haplotype remains the only one in which there was a single nucleotide deletion in a polyC tract of exon 5 (Traherne et al. 2006). DBB, MANN and MCF resembled PGF and COX. No sequence was available for this gene in SSTO or APD.

POU5F1

The PGF haplotype has been reported to have a disrupted start codon for alternative splice variant of *POU5F1* (Traherne et al. 2006). This disruption was not present in COX or QBL nor was it present in the further haplotypes reported in this paper, namely SSTO, DBB, MANN and MCF. APD had no sequence in this region.

OR2J1

This olfactory receptor *OR2J1* has been reported to have both functional and non-functional alleles (Ehlers et al. 2000), the latter the result of a premature stop codon at amino acid position 194 introduced by a substitution in the coding sequence. In our annotation, we found the PGF and MCF haplotypes to contain the full-coding sequence, whereas the COX, QBL SSTO, APD, DBB and MANN haplotypes to contain the truncated sequence as an unprocessed pseudogene (see above and Fig. 1).

Fig. 2 Variation and annotation map of eight MHC haplotypes. The ▶ map represents the complete reference sequence (*orange bar* split into three 1.6 Mb sections) labelled PGF and marked with a scale (Mb) and approximate megabase positions on the NCBI36 build of chromosome 6 (grey milestones). Below the reference sequence are *arrows* representing gene positions and orientations colour-coded for variation status (invariable, black; with synonymous variation only, green; with non-synonymous, conservative variation, red; with non-synonymous, non-conservative variation, purple; see Table 8) and their symbols on a band denoting MHC class (extended class I, green; class I, yellow; class III, pale orange; class II, light blue; extended class II, pink; outside MHC, pale grey). Above the reference sequence, coloured bands represent the sequences of the other seven haplotypes (COX, orange; QBL, mauve; APD, yellow; DBB, green; MANN, light blue; SSTO, dark blue; MCF, purple) with sequence gaps in *dark grey*; the RCCX hyper-variable region shown with *green* (C4A block) and/or *red* (C4B block) or *black* (block absent), and the HLA–DRB hyper-variable region in shades of *blue-green*. Above each haplotype bar, a bar-graph represents total variation between the haplotype and the reference sequence (total variations/10 kb) in *dark red*. Re-examination of the sequence AL645922 from the PGF haplotype, which contains the RCCX region, has shown that the original assembly was erroneous. Correction of these errors leads us now to the conclusion that the C4A gene precedes the C4B gene in this clone sequence. This new gene order is reflected in Fig. 2

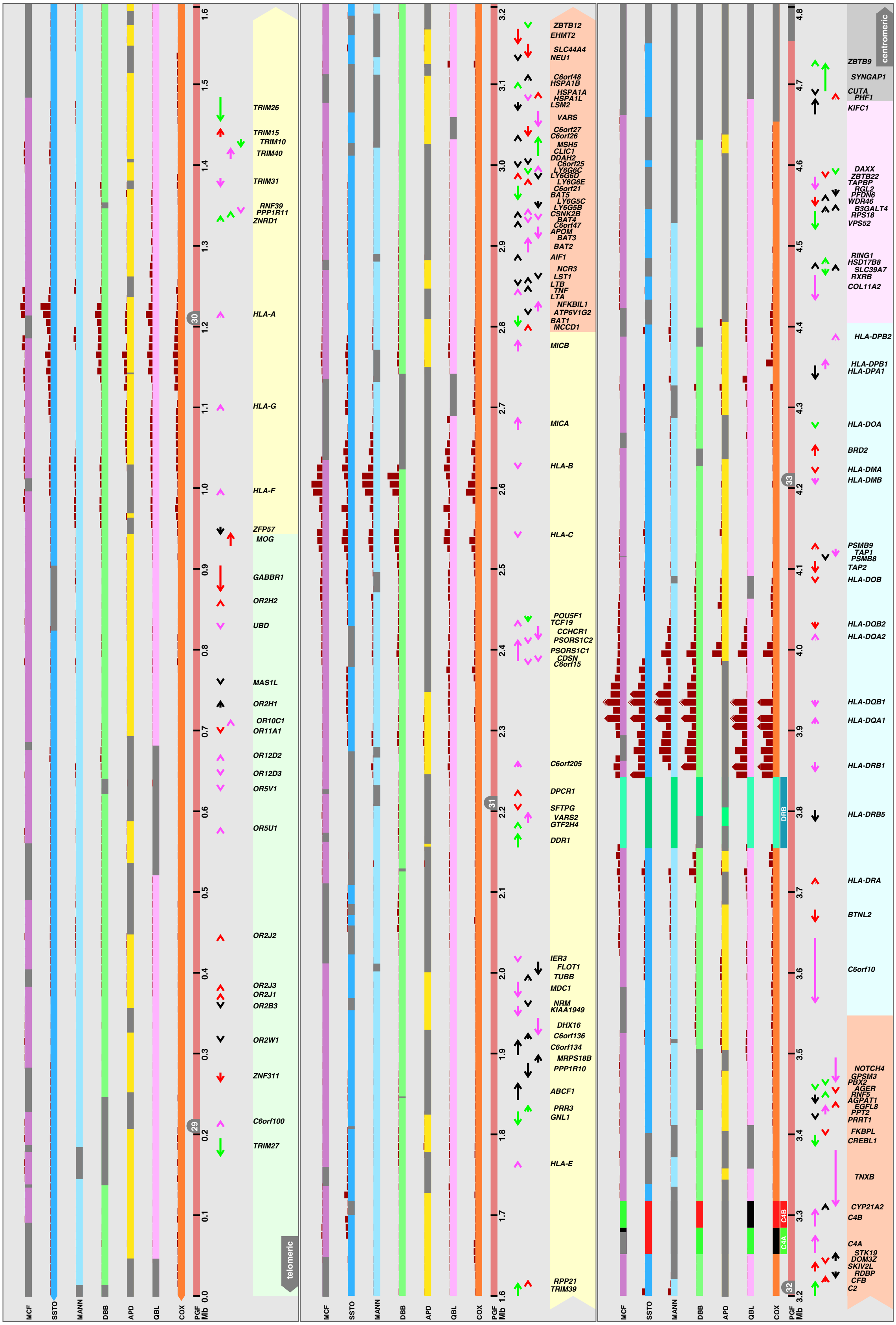

MCF
SSTO
MANN
DBB
APD
QBL
COX
PGF
Mb
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
29
30
telomeric
TRIM26
TRIM15
TRIM10
TRIM40
TRIM31
RNF39
PPP1R11
ZNRD1
HLA-A
HLA-G
HLA-F
ZFP57
MOG
GABBR1
OR2H2
UBD
MAS1L
OR2H1
OR10C1
OR11A1
OR12D2
OR12D3
OR5V1
OR5U1
OR2J2
OR2J3
OR2J1
OR2B3
OR2W1
ZNF311
C6orf100
TRIM27
MCF
SSTO
MANN
DBB
APD
QBL
COX
PGF
Mb
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
3.0
3.1
3.2
31
ZBTB12
EHMT2
SLC44A4
NEU1
C6orf48
HSPA1B
HSPA1A
HSPA1L
LSM2
VARS
C6orf27
C6orf26
MSH5
CLIC1
DDAH2
C6orf25
LY6G6C
LY6G6D
LY6G6E
C6orf21
BAT5
LY6G5C
LY6G5B
CSNK2B
BAT4
C6orf47
APOM
BAT3
BAT2
AIF1
NCR3
LST1
LTB
TNF
LTA
NFKBIL1
ATP6V1G2
BAT1
MCCD1
MICB
MICA
HLA-B
HLA-C
POU5F1
TCF19
CCHCR1
PSORS1C2
PSORS1C1
CDSN
C6orf15
C6orf205
DPCR1
SFTPG
VARS2
GTF2H4
DDR1
IER3
FLOT1
TUBB
MDC1
NRM
KIAA1949
DHX16
C6orf136
C6orf134
MRPS18B
PPP1R10
ABCF1
PRR3
GNL1
HLA-E
RPP21
TRIM39
MCF
SSTO
MANN
DBB
APD
QBL
COX
PGF
Mb
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
32
33
DRB
C4B
C4A
centromeric
ZBTB9
SYNGAP1
CUTA
PHF1
KIFC1
DAXX
ZBTB22
TAPBP
RGL2
PFDN6
WDR46
B3GALT4
RPS18
VPS52
RING1
HSD17B8
SLC39A7
RXRB
COL11A2
HLA-DPB2
HLA-DPB1
HLA-DPA1
HLA-DOA
BRD2
HLA-DMA
HLA-DMB
PSMB9
TAP1
PSMB8
TAP2
HLA-DOB
HLA-DQB2
HLA-DQA2
HLA-DQB1
HLA-DQA1
HLA-DRB1
HLA-DRB5
HLA-DRA
BTNL2
C6orf10
NOTCH4
GPSM3
PBX2
AGER
RNF5
AGPAT1
EGFL8
PPT2
PRRT1
FKBPL
CREBL1
TNXB
CYP21A2
C4B
C4A
STK19
DOM3Z
SKIV2L
RDBP
CFB
C2

Table 9 Variation status of the main coding variant of each gene in the PGF haplotype annotation

Invariable	Synonymous variation only	Non-synonymous variation	
		Conservative variation	Non-conservative variation
ABCF1	BAT1[a]	AGER	BAT2
AGPAT1	BAT5	BRD2[a]	BAT3
AIF1	C2	BTNL2	BAT4
APOM	CREBL1	C6orf21	C4A
ATP6V1G2	DAXX	C6orf27	C4B
B3GALT4	DDR1[a]	CFB	C6orf10
C6orf134	GNL1	DOM3Z	C6orf100
C6orf136[a]	GPSM3	DPCR1	C6orf15
C6orf26	GTF2H4	EGFL8	C6orf205
C6orf48	HLA-DOA[a]	EHMT2	C6orf25
CLIC1	HSPA1B	FKBPL	C6orf47
CSNK2B	LY6G6C	GABBR1	CCHCR1
CUTA	MSH5	HLA-DMA	CDSN
CYP21A2	PBX2	HLA-DOB	COL11A2
DDAH2	POU5F1	HLA-DQB2	DHX16
FLOT1	PPP1R11	HLA-DRA	HLA-A
HLA-DPA1	PRR3	HSPA1A	HLA-B
HLA-DRB5	RING1	LY6G6D	HLA-C
HSD17B8	RNF5	MCCD1	HLA-DMB
KIFC1[a]	RXRB	MOG[a]	HLA-DPB1
LSM2	SYNGAP1	OR11A1	HLA-DPB2
LST1	TRIM10	OR2H2	HLA-DQA1
LTB	TRIM26	OR2J1	HLA-DQA2
LY6G5C	TRIM27	OR2J2	HLA-DQB1
LY6G6E	TRIM39[a]	OR2J3	HLA-DRB1
MAS1L	VPS52	PHF1	HLA-E
MRPS18B	ZBTB12	PSMB9	HLA-F
NCR3	ZBTB9	RPP21	HLA-G
NEU1	ZNRD1	SFTPG	HSPA1L
NRM		SKIV2L	IER3
OR2B3		SLC44A4	KIAA1949
OR2H1		TAP2	LTA
OR2W1		TRIM15	LY6G5B
PFDN6		WDR46	MDC1
PPP1R10		ZBTB22	MICA
PRRT1		ZNF311	MICB
PSMB8[b]			NFKBIL1
RDBP			NOTCH4
RGL2			OR10C1
RPS18			OR12D2
SLC39A7			OR12D3
STK19			OR5U1
TNF			OR5V1
TUBB			PPT2
ZFP57			PSORS1C1
			PSORS1C2
			RNF39
			TAP1
			TAPBP
			TCF19
			TNXB
			TRIM31
			TRIM40
			UBD
			VARS
			VARSL

Gene coding sequences may be invariable (no recorded variation), have synonymous variation only (variation at the nucleotide but not the peptide level) or have non-synonymous variation (variation at both the nucleotide and peptide level), which in turn, may be conservative or non-conservative variation according to the criteria of positive or negative values in the BLOSUM62 matrix. The main coding variant is that numbered 001 in the VEGA database except for *LY6G6E* and *HLA-DPB2* where the main variant is not coding. *C4A* and *C4B* were excluded from calculation of variation because the order of these genes in the PGF sequence precluded alignment with other haplotype sequences. Nevertheless, alignment of the coding sequences for each gene separately showed that there were non-synonymous, non-conservative variations. *HLA-DRB5* is present in this study only in the PGF haplotype and, therefore, here appears invariable

[a] Coding genes where the main variant does not harbour non-conservative, non-synonymous variation but other variants do (*BAT1 BRD2 DDR1 C6orf136 HLA-DOA MOG KIFC1* and *TRIM39*)

[b] Similarly, coding genes where the main variant does not harbour conservative non-synonymous variation but other variants do (*PSMB8*)

Other annotation differences

Other loci included in the current but not the previous PGF annotation were *HCG4P11*, *HCG4P8*, *HCG4P7*, *HCG4P5*, *HCG4P3* and the loci without symbols listed in Table 7. Previously annotated loci not annotated in this study or considered artefacts because they did not reach our current standards of annotation included *HLA-X, C6orf215*, *HCG2P7*, *HCG8*, *HCP5P2*, *HCP5P3*, *HCP5P6*, *HCP5P12*, *HCP5P13*, *HCP5P14*, *HCP5P15*, *HCG8* and *HCG26*.

Non-canonical splice sites

Eight variants within six loci were shown to exhibit haplotypic variation at their splice sites (canonical to non-canonical motif; Table 8). These variations may affect the gene expression at the post-transcriptional level. Hoarau et al. (2004, 2005) have already described the differential splicing within the *HLA–DQA1* locus, and this can clearly be seen by comparing the new *HLA–DQA1* annotation through the VEGA genome browser.

Table 10 Major indels in the form of retrotransposible elements

Chr6 pos'n	Flanking loci	Presence in haplotype								Details
		PGF	COX	QBL	SSTO	APD	DBB	MANN	MCF	
29002370	TRIM27:C6orf100	C	C	C	C	?	?	C	C	**Complex region (A)**
29440424	OR5V1:OR12D3	✓	✓	?	✓	?	?	X	X	AluYa5
29784097	C6orf40:HCP5P15	✓	X	✓	✓	?	X	X	X	AluYa5/8 175..304
29788451	Within HCP5P15	X	X	✓	X	?	✓	✓	X	AluYa5/8 176..310
29794763	HCP5P15:HLA-F	✓	X	X	✓	?	X	X	X	SVA_E plus simple rpt.s
29922942	HLA-G:MICF	✓	X	✓	✓	✓	✓	✓	✓	L1ME3B 5940..6165
29954495	MICF:HLA-H	✓	X	X	X	X	X	X	X	HERVK9 inserted in MER9
30008633	HLA-K:HLA-21	✓	X	X	✓	X	X	✓	?	SVA E/F plus simple rpt.
30106475	HCG8:ETF1P1	X	✓	X	X	✓	✓	X	X	AluYb8
30547387	SUCLA2P:RANP1	X	X	X	✓	?	X	X	?	AluJb 1..283 and parts of MLT1D/L1PBa
31079582	C6orf205:HCG22	X	X	✓	X	X	X	?	X	AluYb8 37..297
31117638	C6orf205:HCG22	✓	X	X	✓	✓	X	✓	✓	AluY (whole & part) and MER63 1017..1062
31301931	HCG27:HLA-C	✓	✓	X	✓	?	✓	✓	✓	HERV3 part (6489...7339)
31320352	HCG27:HLA-C	✓	X	X	X	?	X	X	X	SVA_F 349..850 plus GC rich rpt.
31358220	RPL3P2:WASF5P	X	X	✓	X	?	X	X	X	AluY 35..306
31400900	WASF5P:HLA-B	✓	✓	✓	✓	?	X	X	X	AluSp plus L1PREC2 part (3205...4617)
31405648	WASF5P:HLA-B	✓	X	✓	✓	?	X	x	x	HERVIP10F (part) and AluSg (only cf CX DB)
31418854	WASF5P:HLA-B	✓	✓	✓	✓	?	✓	✓	X	L1PA5 part (5503..5876)
31530995	MICA:HCP5	✓	X	?	✓	?	?	X	?	SVA B/F plus simple rpt.s
32421915	within C6orf10	✓	X	X	✓	X	X	✓	X	AluYb8
32486228	BTNL2:HLA-DRA	✓	✓	✓	✓	✓	X	X	X	L1P1/L1HS parts
32655545	HLA-DRB1 intron 5	✓	x	x	X	?	✓	✓	?	AluYa5 within more or less partial LTR12
32660731	HLA-DRB1 intron 1	X/X	✓/X	X/X	✓/✓	?	✓/✓	✓/✓	?	Tigger4/AluSx
32661119	HLA-DRB1 intron 1	C	C	C	C	?	C	C	?	**Complex region (B)**
32663167	HLA-DRB1 intron 1	X/✓	✓/✓	✓/✓	✓/X	?	✓/X	✓/X	?	AluSq/AluY
32669534	HLA-DRB1:HLA-DQA1	C	C	C	C	?	C	C	?	**Complex region (C)**
32679461	HLA-DRB1:HLA-DQA1	✓	X	X	X	?	X	X	?	AluY
32693271	HLA-DRB1:HLA-DQA1	✓	✓	✓	✓	?	X	✓	?	L1PA4 (parts)
32697545	HLA-DRB1:HLA-DQA1	X	X	X	X	?	✓	✓	?	L1HS 7..6032
32701428	HLA-DRB1:HLA-DQA1	✓	X	✓	✓	?	x	X	x	L1PA2 part and from CX: MER2B and AluY
32728179	HLA-DQA1: HLA-DQB1	C	C	C	C	?	C	C	C	**Complex region (D)**
32739664	within HLA-DQB1	X	X	✓	X	?	X	✓	X	AluY
32743646	HLA-DQB1: MTCO3P1	X	X	X	X	?	✓	X	X	LTR13
32746780	HLA-DQB1: MTCO3P1	X	X	X	X	?	✓	X	✓	L1PA4 (parts)
32751442	HLA-DQB1: MTCO3P1	X	X	X	X	?	X	✓	X	LTR5_Hs
32753489	HLA-DQB1: MTCO3P1	✓	✓	✓	✓	?	X	✓	X	L1PA10 268..4888 around L1PA4 (part)
32756020	HLA-DQB1: MTCO3P1	X	X	X	X	?	X	✓	X	LTR5_Hs
32764047	HLA-DQB1: MTCO3P1	✓	✓	✓	✓	?	X	✓	X	AluSx
32765930	HLA-DQB1: MTCO3P1	X	X	X	X	?	X	✓	X	AluYa5
32785062	MTCO3P1:HLA-DQB3	✓	✓	✓	✓	?	X	X	X	Tigger4 (Zombi)/L1HS (parts) and T-rich
32795150	MTCO3P1:HLA-DQB3	X	X	X	X	X	✓	X	✓	AluY
32796573	MTCO3P1:HLA-DQB3	X	X	X	X	X	✓	X	✓	AluY
32815974	HLA-DQB3: HLA-DQA2	X	✓	X	X	✓	X	X	X	AluYa5
32857369	HLA-DQB2:HLA-DOB	✓	X	✓	✓	X	X	✓	X	AluYg6
32881426	HLA-DQB2:HLA-DOB	X	X	?	✓	✓	X	X	?	AluYa5

Table 10 (continued)

Chr6 pos’n	Flanking loci	Presence in haplotype								Details
		PGF	COX	QBL	SSTO	APD	DBB	MANN	MCF	
32887265	HLA-DQB2:HLA-DOB	✓	X	?	X	X	X	✓	✓	LTR42 and parts of L1MC5 and AluSc 3..105
33201559	within HLA-DPB2	✓	X	X	X	✓	?	X	?	AluYb8
33234360	HCG24:COL11A2	✓	✓	✓	?	?	✓	✓	X	AluY (1..293) AluJb (26..306)

Where there was a break in the cross_match discrepancy list match between two clones, the inserted sequence was extracted and subjected to analysis by RepeatMasker to assess the number of major indels that were a result of retrotransposible elements. Chromosome 6 position (NCBI35/36) of the inserted sequence was that of the midpoint where the sequence was an insertion in PGF or the position before the deletion in PGF. Flanking loci were retrieved from the annotation. Insertion in a haplotypes is indicated by ‘✓’, deletion by ‘X’, complex regions by ‘C’. Where there is a sequence gap in a haplotype corresponding to the indel, this is shown by ‘?’. Four complex deletion/insertion events are listed: A, B, C and D; for details, see text

Combination of variation and annotation data

The data for sequence contig length, gaps, variation rate within haplotypes and PGF coding gene annotation have been combined in the map in Fig. 2. This illustrates the concentration of variation around the HLA gene loci, specifically in 3 areas: around *HLA-F*, *HLA-G* and *HLA-A*; around *HLA-C* and *HLA-B*; and around *HLA-DRB1*, *HLA-DQA1*, *HAL-DQB1*, *HLA-DQA2* and *HLA-DQB2*. The variation status of genes of the PGF haplotype is shown in Table 9.

As well as the variations reported above, major indels revealed as breaks in cross_match discrepancy lists and analysed by RepeatMasker are given in Table 10. Many of these have been previously reported (Dangel et al. 1994; Dunn et al. 2003; Dunn et al. 2002; Gaudieri et al. 1999; Horton et al. 1998; Kulski and Dunn 2005; Stewart et al. 2004). These indels were most frequently but not exclusively associated with AluY elements.

Four of these major indels were complex and designated as complex regions A, B, C and D in Table 10. They include three known regions from the comparison of the PGF and COX haplotypes (Stewart et al. 2004). Complex region A (involving MIR, MER41B, MER115, *AluSx*, Flam_C, *AluSg*, *AluY*, *AluSx*, L2 and MER38 elements) maps between *TRIM27* and *C6orf100* and was found to be deleted in COX but present in PGF, QBL, SSTO, MANN and MCF. Complex region B (involving L2 and AluY elements) maps to intron 1 of *HLA–DRB1* and was also found by comparing PGF with COX, QBL, DBB, MANN and SSTO. Complex region C (SVA and low-complexity repeat elements) maps between *HLA–DRB1* and *HLA–DQA1* and was noted in COX as a deletion of the SVA and low-complexity repeats. Whereas, DBB, MANN and SSTO displayed the same deletion, as well as a telomeric deletion of *AluSx*/MIRb, QBL had both deletions plus that of an intervening 2.5 kb sequence containing *Alu*, L3 and MLT1A1 elements. Complex region D maps between *HLA−DQA1* and *HLA−DQB1* and is more complicated than previously reported. At the telomeric end, PGF lacks an L1PA4 fragment of >300 bp that is present in COX, QBL, SSTO and MCF and is also absent in DBB and MANN where it is interrupted by about 1.3 kb of SVA sequence. Centromeric to this PGF contains an *AluSx*, an *AluY* and an *AluYd2*, flanked by long interspersed nuclear element repeats, all deleted in the other haplotypes. Further towards the centromere there is an L1MA7 fragment, into which in PGF alone there are insertions of an *AluSx* followed by an *AluY*; a subsequent *AluSg* present in all haplotypes contains an insertion of 795 bp of SVA sequence in just COX and QBL. Finally, at the centromeric end of this region, PGF uniquely contains intact MER11C and LTR5 elements.

Representation of haplotypes within European populations

The eight haplotypes analysed in this study were selected on the basis of their association with type 1 diabetes and multiple sclerosis and their high population frequencies. To determine how representative these haplotypes are with respect to SNP haplotypic diversity in a population, we determined their distribution in the haplotypic tree space in the European population.

For this analysis, we selected a segment of ~214 kb, spanning the HLA–DRB1 and HLA–DQB1 genes in a population of European ancestry with known HLA allelic data (de Bakker et al. 2006). Phylogenetic analysis of 180 founder haplotypes derived from genotypic data (54 substitutions) shows that the eight haplotypes selected as part of the MHC Haplotype Project share identical HLA alleles over most of the tree space (Fig. 3a), representing almost the entire variation observed in the population

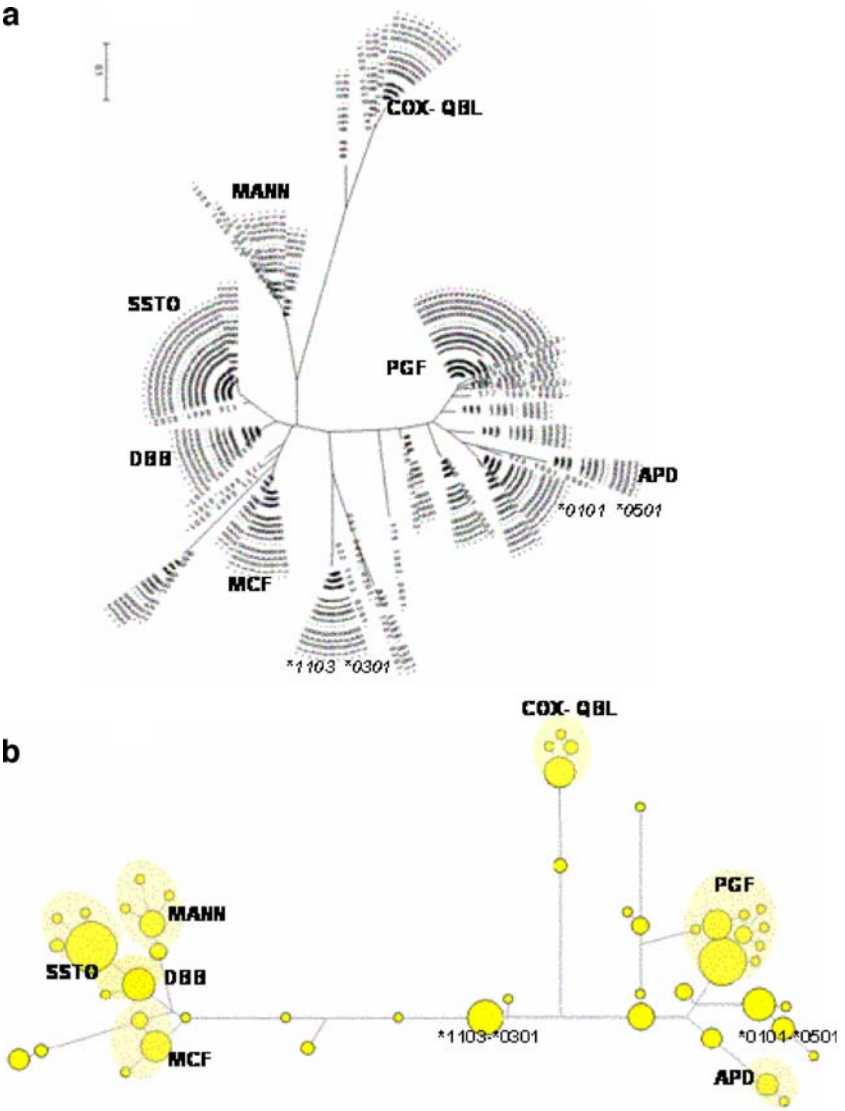


Fig. 3 Clusters of haplotypes in the European haplotypic diversity. Phylogenetic relationship of 180 founder SNP haplotypes from CEPH trios spanning a 214-kb segment of the MHC class II region, including the *HLA-DRB*1 and *HLA-DQB1* genes (54 substitutions from rs2187823 to rs2856691). **a** Sequenced haplotypes are widely distributed in this NJ tree and represent the vast majority of the variation in the population sampled. Four-digit alleles are indicated for the corresponding *DRB1* and *DQB1* genes in each haplotype ID label to highlight the HLA haplotypic distribution based on the underlying nucleotide variation. The NJ tree was constructed using pairwise genetic distances considering the Kimura 2-parameters model without correction for rate variation among sites as implemented in the MEGA2 software (Kumar et al. 2001). **b** Each haplotype sequenced is associated to a single haplotype cluster. This phylogenetic network (Bandelt et al. 1999) also shows that clusters (shaded area) are constituted by one central haplotype and its derivatives. *Circles* represent individual haplotypes, and the size of the circle is proportional to the haplotype frequency. The length of the lines connecting nodes is relative to the distance between them, e.g. distances within *shaded areas* (clusters) never exceed three mutation steps. Cluster of haplotypes sharing HLA alleles with sequenced cell lines are named accordingly: COX and QBL: DRB1*0301 DQB1*0201–PGF: DRB1*1501 DQB1*0602–APD: DRB1*1301 DQB1*0603–MCF: DRB1*0401 DQB1*0301–DBB: DRB1*0701 DQB1*0303–SSTO: DRB1*0403 DQB1*0302–MANN: DRB1*0701 DQB1*0202. HLA haplotypes DRB1*1103–DQB1*0301 and DRB1*0101–DQB1*0501 indicate the two major haplotype clusters not represented in the MHC haplotype project data

assayed with the exception of two branches (DRB1*1103–DQB1*0301 and DRB1*0101–DQB1*0501).

Haplotype diversity in this sub-population is restricted to relatively few haplotype clusters (Fig. 3b). Each cluster consists of a founder haplotype, depicted by the most frequent and centrally located haplotype within the cluster. Recently derived haplotypes show lower frequencies and are connected to the central haplotype by relatively few mutation steps (in this case, up to three). This phylogenetic network clearly shows that all the sequenced haplotypes occupy central positions in their respective haplotypic groups. Inferences about phylogenetic relationships between haplotype clusters are, however, only approximate as a consequence of recombination events.

It should also be noted that SNP haploptypes derived from CEPH pedigrees of European ancestry by no means represent an exhaustive sampling of European diversity. Nevertheless, the sampling has been shown to represent the European population in the UK reasonably well (Ke et al. 2005). In conclusion, our analysis demonstrates that the HLA haplotypes selected for the MHC Haplotype Project are ancestral haplotypes, representative of MHC diversity in the European population.

Conclusion and outlook

The MHC Haplotype Project has succeeded in providing a new public resource for immune-linked disease and population genetic studies. First reports from studies using the resource indicate that it adds significant power to the identification and fine-mapping of disease-associated variations (Yeo et al. 2007). The data have also contributed to the recent identification of a first set of HLA tag SNPs, which hold great promise for future applications in clinical settings, e.g. to complement or replace classical HLA-typing in transplant medicine (de Bakker et al. 2006). While costs and other limitations of the current (capillary) sequencing technology have restricted our study to only few (eight) MHC haplotypes, the number of new variations found, combined with the fact that no variation plateau has yet been reached, indicates that there are many more variations to be discovered. The recent introduction of several new and massively parallel sequencing platforms (for review, see Bentley 2006) has created the opportunity to do just that by re-sequencing haplotypes and, eventually, entire genomes at the population level and as integral part of case control studies. Because of its wide-ranging medical importance, the MHC can be expected to be among the first regions of the human genome to be sequenced in this way. Such sequencing will provide the critical, and until now missing, data to identify causal variations and their underlying mechanisms on an unprecedented scale.

Acknowledgements The authors thank C.-K. Chen and all the staff of the DNA Sequencing Division and the HAVANA group (http://www.sanger.ac.uk/HGP/havana) at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk) for their assistance. This work was supported by a joint grant (048880) from the Wellcome Trust to J.A. T., J.T., S.B. and S.S., and a Wellcome Trust/Juvenile Diabetes Research Foundation grant to J.A.T.

References

Allcock RJ, Atrazhev AM, Beck S, de Jong PJ, Elliott JF, Forbes S, Halls K, Horton R, Osoegawa K, Rogers J, Sawcer S, Todd JA, Trowsdale J, Wang Y, Williams S (2002) The MHC haplotype project: a resource for HLA-linked association studies. Tissue Antigens 59:520–521

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Awdeh ZL, Alper CA (1980) Inherited structural polymorphism of the fourth component of human complement. Proc Natl Acad Sci U S A 77:3576–3580

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16:545–552

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Chida S, Hohjoh H, Hirai M, Tokunaga K (2001) Haplotype-specific sequence encoding the protein kinase, interferon-inducible double-stranded RNA-dependent activator in the human leukocyte antigen class II region. Immunogenetics 52:186–194

Chung EK, Yang Y, Rennebohm RM, Lokki ML, Higgins GC, Jones KN, Zhou B, Blanchong CA, Yu CY (2002) Genetic sophistication of human complement components C4A and C4B and RP-C4-CYP21-TNX (RCCX) modules in the major histocompatibility complex. Am J Hum Genet 71:823–837

Dangel AW, Mendoza AR, Baker BJ, Daniel CM, Carroll MC, Wu LC, Yu CY (1994) The dichotomous size variation of human complement C4 genes is mediated by a novel family of endogenous retroviruses, which also establishes species-specific genomic patterns among Old World primates. Immunogenetics 40:425–436

de Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, Marchini J, Ke X, Monsuur AJ, Whittaker P, Delgado M, Morrison J, Richardson A, Walsh EC, Gao X, Galver L, Hart J, Hafler DA, Pericak-Vance M, Todd JA, Daly MJ, Trowsdale J, Wijmenga C, Vyse TJ, Beck S, Murray SS, Carrington M, Gregory S, Deloukas P, Rioux JD (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38:1166–1172

De Vet EC, Aguado B, Campbell RD (2003) Adaptor signalling proteins Grb2 and Grb7 are recruited by human G6f, a novel member of the immunoglobulin superfamily encoded in the MHC. Biochem J 375:207–213

Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res 12:458–461

Dunn DS, Naruse T, Inoko H, Kulski JK (2002) The association between HLA-A alleles and young Alu dimorphisms near the HLA-J, -H, and -F genes in workshop cell lines and Japanese and Australian populations. J Mol Evol 55:718–726

Dunn DS, Inoko H, Kulski JK (2003) Dimorphic Alu element located between the TFIIH and CDSN genes within the major histocompatibility complex. Electrophoresis 24:2740–2748

Ehlers A, Beck S, Forbes SA, Trowsdale J, Volz A, Younger R, Ziegler A (2000) MHC-linked olfactory receptor loci exhibit polymorphism and contribute to extended HLA/OR-haplotypes. Genome Res 10:1968–1978

Gaudieri S, Kulski JK, Dawkins RL, Gojobori T (1999) Different evolutionary histories in two subgenomic regions of the major histocompatibility complex. Genome Res 9:541–549

Geraghty DE, Koller BH, Hansen JA, Orr HT (1992) The HLA class I gene family includes at least six genes and twelve pseudogenes and gene fragments. J Immunol 149:1934–1946

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 (Suppl 1):1–9

Hoarau JJ, Cesari M, Caillens H, Cadet F, Pabion M (2004) HLA DQA1 genes generate multiple transcripts by alternative splicing and polyadenylation of the 3′ untranslated region. Tissue Antigens 63:58–71

Hoarau JJ, Festy F, Cesari M, Pabion M (2005) A new splicing acceptor site and poly(A) +sequence signal within DQA1*0401 and DQA1*0501 mRNA 3¢UTR contribute to increase the extraordinary diversity of mRNA isoforms. Immunogenetics 57:182–188

Horton R, Niblett D, Milne S, Palmer S, Tubby B, Trowsdale J, Beck S (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

Horton R, Wilming L, Rand V, Lovering RC, Bruford EA, Khodiyar VK, Lush MJ, Povey S, Talbot CC Jr, Wright MW, Wain HM, Trowsdale J, Ziegler A, Beck S (2004) Gene map of the extended human MHC. Nat Rev Genet 5:889–899

Jamain S, Girondot M, Leroy P, Clergue M, Quach H, Fellous M, Bourgeron T (2001) Transduction of the human gene FAM8A1 by endogenous retrovirus during primate evolution. Genomics 78:38–45

Ke X, Miretti MM, Broxholme J, Hunt S, Beck S, Bentley DR, Deloukas P, Cardon LR (2005) A comparison of tagging methods and their tagging space. Hum Mol Genet 14:2757–2767

Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Hoad G, Kanz C, Lee C, Leinonen R, Lin Q, Lombard V, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Nardone F, Pastor MP, Plaister S, Sobhany S, Stoehr P, Vaughan R, Wu D, Zhu W, Apweiler R (2007) EMBL nucleotide sequence database in 2006. Nucleic Acids Res 35:D16–D20

Kulski JK, Dunn DS (2005) Polymorphic Alu insertions within the major histocompatibility complex class I genomic region: a brief review. Cytogenet Genome Res 110:193–202

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17 (12):1244–1245

Lechler R, Warrens A (2000) HLA in health and disease. Academic, 2000

Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Hurley CK, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI, Trowsdale J (2005) Nomenclature for factors of the HLA system, 2004. Hum Immunol 66:571–636

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, Bentley DR, Rioux JD, Beck S, Deloukas P (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 76:634–646

Mungall AJ, Palmer SA, Sims SK, Edwards CA, Ashurst JL, Wilming L, Jones MC, Horton R, Hunt SE, Scott CE, Gilbert JG, Clamp ME, Bethel G, Milne S, Ainscough R, Almeida JP, Ambrose KD, Andrews TD, Ashwell RI, Babbage AK, Bagguley CL, Bailey J, Banerjee R, Barker DJ, Barlow KF, Bates K, Beare DM, Beasley H, Beasley O, Bird CP, Blakey S, Bray-Allen S, Brook J, Brown AJ, Brown JY, Burford DC, Burrill W, Burton J, Carder C, Carter NP, Chapman JC, Clark SY, Clark G, Clee CM, Clegg S, Cobley V, Collier RE, Collins JE, Colman LK, Corby NR, Coville GJ, Culley KM, Dhami P, Davies J, Dunn M, Earthrowl ME, Ellington AE, Evans KA, Faulkner L, Francis MD, Frankish A, Frankland J, French L, Garner P, Garnett J, Ghori MJ, Gilby LM, Gillson CJ, Glithero RJ, Grafham DV, Grant M, Gribble S, Griffiths C, Griffiths M, Hall R, Halls KS, Hammond S, Harley JL, Hart EA, Heath PD, Heathcott R, Holmes SJ, Howden PJ, Howe KL, Howell GR, Huckle E, Humphray SJ, Humphries MD, Hunt AR, Johnson CM, Joy AA, Kay M, Keenan SJ, Kimberley AM, King A, Laird GK, Langford C, Lawlor S, Leongamornlert DA, Leversha M et al (2003) The DNA sequence and analysis of human chromosome 6. Nature 425:805–811

Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigo R (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res 16:37–44

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516–522

Searle SM, Gilbert J, Iyer V, Clamp M (2004) The otter annotation system. Genome Res 14:963–970

Semple JI, Ribas G, Hillyard G, Brown SE, Sanderson CM, Campbell RD (2003) A novel gene encoding a coiled-coil mitochondrial protein located at the telomeric end of the human MHC Class III region. Gene 314:41–54

Shaw CJ, Lupski JR (2004) Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. Hum Mol Genet Special Issue 13(1):R57–R64

Shiina T, Ota M, Shimizu S, Katsuyama Y, Hashimoto N, Takasu M, Anzai T, Kulski JK, Kikkawa E, Naruse T, Kimura N, Yanagiya K, Watanabe A, Hosomichi K, Kohara S, Iwamoto C, Umehara Y, Meyer A, Wanner V, Sano K, Macquin C, Ikeo K, Tokunaga K, Gojobori T, Inoko H, Bahram S (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

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Smith WP, Vu Q, Li SS, Hansen JA, Zhao LP, Geraghty DE (2006) Toward understanding MHC disease associations: partial resequencing of 46 distinct HLA haplotypes. Genomics 87:561–571

Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, Forbes S, Halls K, Howson JM, Humphray SJ, Hunt S, Mungall AJ, Osoegawa K, Palmer S, Roberts AN, Rogers J, Sims S, Wang Y, Wilming LG, Elliott JF, de Jong PJ, Sawcer S, Todd JA, Trowsdale J, Beck S (2004) Complete MHC haplotype sequencing for common disease gene mapping. Genome Res 14:1176–1187

The MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature 401:921–923

The 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

Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, Ashurst JL, Atrazhev AM, Coggill P, Palmer S, Almeida J, Sims S, Wilming LG, Rogers J, de Jong PJ, Carrington M, Elliott

JF, Sawcer S, Todd JA, Trowsdale J, Beck S (2006) Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. PLoS Genet 2:e9

Yeo TW, De Jager PL, Gregory SG, Barcellos LF, Walton A, Goris A, Fenoglio C, Ban M, Taylor CJ, Goodman RS, Walsh E, Wolfish CS, Horton R, Traherne J, Beck S, Trowsdale J, Caillier SJ, Ivinson AJ, Green T, Pobywajlo S, Lander ES, Pericak-Vance MA, Haines JL, Daly MJ, Oksenberg JR, Hauser SL, Compston A, Hafler DA, Rioux JD, Sawcer S (2007) A second major histocompatibility complex susceptibility locus for multiple sclerosis. Ann Neurol 61:228–236