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

Roger Horton · Richard Gibson · Penny Coggill · Marcos Miretti · Richard J. Alcock · 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
© The Author(s) 2007

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-
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
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 Fgenes (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 Poly-morphisme Humain (CEPH) trios (de Bakker et al. 2006). A ~214 kb segment spanning the _HLA–DRB1–DQB1_ genes
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 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*010201 | 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*02010101] | 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.
Table 3 Distribution of substitutions and indels within different sequence regions amongst haplotypes

| Sequence region | Base pairs | COX | QBL | SSTO | APD | DBB | MANN | MCF |
|-----------------|-----------|-----|-----|------|-----|-----|------|-----|
| Coding          | 247,505   | 353 | 8   | 503  | 19  | 380 | 2    |     |
| UTR             | 155,960   | 382 | 34  | 438  | 59  | 331 | 35   |     |
| Intronic        | 1,283,472 | 3,141| 571 | 3,135| 590 | 2,658| 505  |     |
| Total intragenic| 1,686,937 | 3,876| 613 | 4,076| 668 | 3,369| 542  |     |
| Pseudogenic     | 57,223    | 235 | 15  | 226  | 21  | 227 | 19   |     |
| Pseudogenic intron| 63,108   | 507 | 54  | 220  | 27  | 215 | 18   |     |
| Transcript exon | 78,092    | 190 | 30  | 207  | 33  | 119 | 22   |     |
| Transcript intron| 332,705 | 1,243| 197 | 1,186| 216 | 1,053| 155  |     |
| REPEATS:        |           |     |     |      |     |     |      |     |
| LINEs           | 608,429   | 2,110| 221 | 2,015| 240 | 2,388| 255  |     |
| SINEs           | 428,567   | 1,381| 428 | 1,316| 401 | 1,311| 385  |     |
| Other repeats   | 487,863   | 2,605| 207 | 2,518| 229 | 2,514| 207  |     |
| Total in repeats| 1,524,859| 6,096| 856 | 5,849| 870 | 6,213| 847  |     |
| Microsatellite  | 15,185    | 186 | 168 | 95   | 85  | 222 | 198  |     |
| All above       | 4,754,829 | 15,967| 2,393| 15,282| 2,360| 14,982| 2,300 |     |

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
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.
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–DRB4, 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.

**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 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
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 (C4AL, C4BL) or short (C4AS, C4BS) 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 C4BS and QBL has a single module with C4AS (Traherne et al. 2006). For the new haplotype sequences reported in this paper, SSTO was bimodular with two copies of C4BL, whereas DBB was bimodular with C4AL followed by C4BS. 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 C4AL. The RCCX region in the APD and MANN haplotypes was incomplete because of sequence gaps.

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         | 1,267 | 1,191 | 294 | 281  | 166 | 281 | 264  | 261 |
| Total variants     | 1,058 | 568  | 1,138 | 960  | 264 | 261 |

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-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
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).

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    | tgg       | rs28400887       | cDNA          | NC  | NC  | NC  | C    | ND  | NC  | NC   | C   |
| TRIM31 | 5       | 2/3            | ggt    | tgg       | rs28400887       | EST           | NC  | NC  | NC  | C    | ND  | NC  | NC   | C   |
| C4B    | 7       | 3/4            | ggt    | cgg       | –                | EST           | NC  | ND  | NC  | C    | ND  | ND  | ND   |    |
| C4A    | 7       | 3/4            | ggt    | cgg       | –                | EST           | NC  | NC  | ND  | C    | ND  | ND  | NC   |    |
| HLA-DQA1| 4      | 4/5            | ggt    | cgg       | rs707947         | cDNA          | C   | C   | C   | NC   | ND  | NC  | NC   | NC  |
| HLA-DQA1| 5      | 4/5            | ggt    | taa/caa   | rs3667           | cDNA          | NC  | NC  | NC  | C    | ND  | C   | C    | C   |
| HLA-DRB1| 2      | 2/3            | gat    | cag       | rs9271083        | EST           | NC  | C   | C   | C    | ND  | C   | ND   | 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).
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, C6orf136 and HLA-DRB2 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.

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-DBP2 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.

Other non-canonical splice sites

Eight variants within six loci were shown to exhibit haplotype 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 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<sup>a</sup>         | AGER                     | BAT2                     |
| AGPAT1     | BAT5                     | BRD2<sup>a</sup>         | BAT3                     |
| AIF1       | C2                       | BTNL2                    | BAT4                     |
| APOM       | CREBL1                   | C6orf21I                 | C4A                      |
| ATP6V1G2   | DAXX                     | C6orf27                  | C4B                      |
| B3GALT4    | DDR1<sup>a</sup>         | CFB                      | C6orf10                  |
| C6orf134   | GNL1                     | DOM3Z                    | C6orf100                 |
| C6orf136<sup>a</sup> | GPSM3                   | DPCR1                    | C6orf15                  |
| C6orf26    | GTF2H4                   | EGFL8                    | C6orf205                 |
| C6orf48    | HLA-DOA<sup>a</sup>      | EHMT2                    | C6orf25                  |
| CLIC1      | HSPA1B                   | FKBP1                    | C6orf47                  |
| CSNK2B     | LY6G6C                   | GABBR1                   | CCHCR1                   |
| CUTA       | MSH5                     | HLA-DMa                  | CDSN                     |
| CYP21A2    | PBX2                     | HLA-DOB                  | COL11A2                  |
| DDAH2      | POUSF1                   | HLA-DQB2                 | DHX16                    |
| FLOT1      | PPP1R11I                 | HLA-DRA                  | HLA-A                    |
| HLA-DPA1   | PRR3                     | HSPA1A                   | HLA-B                    |
| HLA-DRB5   | RING1                    | LY6G6D                   | HLA-C                    |
| HSD17B8    | RNF5                     | MCCD1                    | HLA-DMB                  |
| KIFC1<sup>a</sup> | RXRB            | MOG<sup>a</sup>          | HLA-DBP1                 |
| LSM2       | SYNGAP1                  | OR11A1                   | HLA-DBP2                 |
| LST1       | TRIM10                   | OR2H2                    | HLA-DQA1                 |
| LTB        | TRIM26                   | OR2H2I                   | HLA-DQA2                 |
| LY6G5C     | TRIM27                   | OR2H2I                   | HLA-DQA2                 |
| LY6G6E     | TRIM39<sup>a</sup>       | OR2H2I                   | HLA-DRB1                 |
| MAS1L      | VPS52                    | PHF1                     | HLA-E                    |
| MRPS18B    | ZBTB12                   | PSMB9                    | HLA-F                    |
| NCR3       | ZBTB9                    | RPP21                    | HLA-G                    |
| NEU1       | ZNDR1                    | SFTPG                    | HSPA1L                   |
| NRM        | SKIV2L                   | IER3                     | KIAA1949                 |
| OR2B3      | SLC44A4                  | TAP2                     | LTA                      |
| OR2H1      | TRIM15                   | LY6G5B                   | MDC1                     |
| OR2W1      | TRIM15                   | WDR46                    | MICA                     |
| PFDN6      | TRIM15                   | ZBTB22                   | MICB                     |
| PPP1R10    | ZBTB22                   | ZNF311                   | NFKB1L                   |
| PRRT1      | PSORS1C1                 | NOTCH4                   | OR10C1                   |
| PSMB8<sup>b</sup> | RDN19         | OR12C1                   | OR12D2                   |
| RDBP       | RGL2                     | OR12D3                   | OR5U1                    |
| RPS18      | SLC39A7                  | PPT2                     | PSORS1C1                 |
| STK19      | TNF                      | TAP1                     | PSORS1C2                 |
| TUBB       | TCF19                    | TAPBP                    | RNF39                    |
| ZFP57      |                        |                          |                          |

Table 9 (continued)

| Invariable | Synonymous variation only | Non-synonymous variation |
|------------|--------------------------|--------------------------|
|            |                          | Conservative variation   | Non-conservative variation |
| TRIM31     |                          |                          |                          |
| TRIM40     |                          |                          |                          |
| UBD        |                          |                          |                          |
| VARS       |                          |                          |                          |
| VARS<sup>L</sup> |                 |                          |                          |
| 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    | Complex region (A) |
| 29440424  | OR5V1:OR12D3   | ✓   | ✓   | ?    | ✓    | ?    | ?   | X    | X    | AluYa5  |
| 29784097  | C6orf40:HCPS15 | ✓   | X   | ✓    | ?    | X    | X   | X    | X    | AluYa5/8 175..304 |
| 29788451  | Within HCPS15  | X   | ✓   | X    | ?    | X    | X   | ✓    | ✓    | AluYa5/8 176..310 |
| 29794763  | HCPS15:HLA-F   | ✓   | X   | ✓    | ?    | X    | X   | X    | SVA_E plus simple rpt.s |
| 29922942  | HLA-G:MCF      | ✓   | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | L1ME3B 5940..6165 |
| 29954495  | MICF:HLA-H     | X   | X   | X    | ?    | X    | X   | X    | SVA E/F inserted in MER9 |
| 30008633  | HLA-K:HLA-21   | ✓   | X   | ✓    | X    | ✓    | ✓    | X    | X    | AluYb8  |
| 30106475  | HCG8:ETF1P1    | X   | ✓   | X    | ✓    | ✓    | ✓    | X    | X    | AluYa5 within more or less partial LTR12 |
| 30547387  | SUCLA2P:RANP1  | X   | X   | ✓    | ?    | ?    | ?   | X    | X    | Complex region (B) |
| 31079582  | C6orf205:HC22  | X   | X   | ✓    | X    | X    | ?   | ?    | X    | L1PA5 (part) and AluY (whole & part) |
| 31176638  | C6orf205:HC22  | ✓   | X   | ✓    | ?    | ✓    | ?   | ✓    | ✓    | L1PA4 (parts) |
| 31301931  | HCG27:HLA-C    | ✓   | ✓   | X    | ✓    | ?    | ✓    | ✓    | SVA F 349..850 plus GC rich rpt. |
| 31320352  | HCG27:HLA-C    | X   | X   | X    | ?    | X    | X   | ✓    | ✓    | AluYa5 35..306 |
| 31358220  | RPL3P2:WASF5P   | X   | X   | ✓    | ?    | X    | X   | X    | X    | AluSp plus L1PREC2 part (3205..4617) |
| 31405648  | WASF5P:HLA-B   | ✓   | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (C) |
| 31406458  | WASF5P:HLA-B   | ✓   | X   | ✓    | ?    | X    | X   | X    | L1P1/L1HS parts |
| 31418854  | WASF5P:HLA-B   | ✓   | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (D) |
| 31530995  | MICA:HCPS5     | ✓   | X   | ✓    | ?    | ?    | ?   | X    | ?    | Complex region (A) |
| 32241915  | within C6orf10 | ✓   | X   | X    | ✓    | X    | X   | ✓    | ✓    | AluYa5  |
| 32486228  | BTNL2:HLA-DRA  | ✓   | ✓   | ✓    | ✓    | ✓    | ✓    | L1HS 7..6032 |
| 32655545  | HLA-DRB1 intron 5 | ✓  | x   | x    | ✓    | ?    | ✓    | ?    | ✓    | Complex region (A) |
| 32660731  | HLA-DRB1 intron 1 | X/X | ✓/X | X/X  | ✓/✓ | ✓/✓ | ✓/✓ | ✓/✓ | ✓/✓ | Complex region (B) |
| 32661119  | HLA-DRB1 intron 1 | C   | C   | C    | C    | ?    | ?    | ✓    | ✓    | Complex region (C) |
| 32663167  | HLA-DRB1 intron 1 | X/X | ✓/X | X/X  | ✓/✓ | ✓/✓ | ✓/✓ | ✓/✓ | ✓/✓ | Complex region (D) |
| 32669534  | HLA-DRB1:HLA-DQA1 | C  | C   | C    | C    | ?    | ?    | C    | C    | Complex region (A) |
| 32679461  | HLA-DRB1:HLA-DQA1 | ✓  | X   | X    | ?    | X    | X    | X    | X    | Complex region (B) |
| 32692371  | HLA-DRB1:HLA-DQA1 | ✓  | ✓   | ✓    | ?    | X    | X    | X    | X    | Complex region (C) |
| 32697545  | HLA-DRB1:HLA-DQA1 | X  | X   | X    | ✓    | ✓    | ✓    | X    | X    | Complex region (D) |
| 32701428  | HLA-DRB1:HLA-DQA1 | ✓  | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (A) |
| 32728179  | HLA-DQA1:HLA-DQB1 | C  | C   | C    | ?    | C    | C    | C    | C    | Complex region (B) |
| 32739664  | within HLA-DQB1 | X   | X   | ✓    | X    | ?    | X    | ✓    | ✓    | Complex region (C) |
| 32743646  | HLA-DQB1:MTCO3P1 | X  | X   | X    | ✓    | ✓    | ✓    | X    | X    | Complex region (D) |
| 32746780  | HLA-DQB1:MTCO3P1 | X  | X   | X    | ?    | X    | ✓    | ✓    | ✓    | Complex region (A) |
| 32751442  | HLA-DQB1:MTCO3P1 | X  | X   | X    | ?    | X    | ✓    | ✓    | ✓    | Complex region (B) |
| 32753489  | HLA-DQB1:MTCO3P1 | ✓  | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (C) |
| 32756020  | HLA-DQB1:MTCO3P1 | X  | X   | X    | ✓    | X    | ✓    | X    | X    | Complex region (D) |
| 32764047  | HLA-DQB1:MTCO3P1 | ✓  | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (A) |
| 32765930  | HLA-DQB1:MTCO3P1 | X  | X   | X    | ?    | X    | ✓    | ✓    | ✓    | Complex region (B) |
| 32785062  | MTCO3P1:HLA-DQB3 | ✓  | ✓   | ✓    | ✓    | ✓    | ✓    | ✓    | ✓    | Complex region (C) |
| 32795150  | MTCO3P1:HLA-DQB3 | X  | X   | X    | X    | ✓    | ✓    | ✓    | ✓    | Complex region (D) |
| 32796573  | MTCO3P1:HLA-DQB3 | X  | X   | X    | X    | ✓    | ✓    | ✓    | ✓    | Complex region (A) |
| 32815974  | HLA-DQB3:HLA-DQA2 | X  | ✓   | X    | ✓    | X    | X    | X    | X    | Complex region (B) |
| 32857369  | HLA-DQB2:HLA-DOB | ✓  | ✓   | ✓    | X    | ✓    | ✓    | X    | X    | Complex region (C) |
| 32881426  | HLA-DQB2:HLA-DOB | X  | X   | ?    | ✓    | ✓    | X    | X    | X    | Complex region (D) |
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, HLA-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/MIRrb, 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.
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.

**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-DRB1 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.
It should also be noted that SNP haplotypes derived from CEPH pedigrees of European ancestry by no means represent an exhaustive sampling of European diversity. Nevertheless, 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.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

**References**

Alcock 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

Binney 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, Hira 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 genomeric patterns among Old World primates. Immunogenetics 40:425–436

de Bakker PJ, 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 Biol 55:726–744

Dunn DS, Inoko H, Kulski JK (2003) Dimeric 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
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