The DNA sequence and analysis of human chromosome 6

A. J. Mungall*, S. A. Palmer, S. K. Sims, C. A. Edwards, J. L. Ashurst, L. Wilming, M. C. Jones, R. Horton, S. E. Hunt, C. E. Scott, J. G. R. Gilbert, M. E. Clamp, G. Bethel, S. Milne, R. Ainscough, J. P. Almeida, K. D. Ambrose, T. D. Andrews, R. I. S. Ashwell, A. K. Babbage, C. L. Bagguley, J. Bailey, R. Banerjee, D. J. Barker, K. F. Barlow, K. Bates, D. M. Beare, H. Beasley, O. Beasley, C. P. Bird, S. Blakey, S. Bray-Allen, J. Brook, A. J. Brown, J. Y. Brown, D. C. Burford, W. Burrill, J. Burton, C. Carder, N. P. Carter, J. C. Chapman, S. Y. Clark, G. Clark, C. M. Clee, S. Clegg, V. Cobley, R. E. Collier, J. E. Collins, L. K. Colman, N. R. Corby, G. J. Coville, K. M. Culley, P. Dhami, J. Davies, M. Dunn, M. E. Earthrowl, A. E. Ellington, K. A. Evans, L. Faulkner, M. D. Francis, A. Frankish, J. Frankland, L. French, P. Garner, J. Garnett, M. J. R. Ghori, L. M. Gilby, C. J. Gillson, R. J. Glithero, D. V. Grafham, M. Grant, S. Gribble, C. Griffiths, M. Griffiths, R. Hall, K. S. Halls, S. Hammond, J. L. Harley, E. A. Hart, P. D. Heath, R. Heathcott, S. J. Holmes, P. J. Howden, K. L. Howe, G. R. Howell, E. Huckle, S. J. Humphray, M. D. Humphries, A. R. Hunt, C. M. Johnson, A. A. Joy, M. Kay, S. J. Keenan, A. M. Kimberley, A. King, G. K. Laird, C. Langford, S. Lawlor, D. A. Leongamornlert, M. Leversha, C. R. Lloyd, D. M. Lloyd, J. E. Loveland, J. Lovell, S. Martin, M. Mashreghi-Mohammadi, G. L. Maslen, L. Matthews, O. T. McCann, S. J. McLaren, K. McLay, A. McMurray, M. J. F. Moore, J. C. Mullikin, D. Niblett, T. Nickerson, K. L. Novik, K. Oliver, E. K. Overton-Larty, A. Parker, R. Patel, A. V. Pearce, A. I. Peck, B. Phillimore, S. Phillips, R. W. Plumb, K. M. Porter, Y. Ramsey, S. A. Ranby, C. M. Rice, M. T. Ross, S. M. Searle, H. K. Sehra, E. Sheridan, C. D. Skuce, S. Smith, M. Smith, L. Spraggon, S. L. Squares, C. A. Steward, N. Sycamore, G. Tamlyn-Hall, J. Tester, A. J. Theaker, D. W. Thomas, A. Thorpe, A. Tracey, A. Tromans, B. Tubby, M. Wall, J. M. Wallis, A. P. West, S. S. White, S. L. Whitehead, H. Whittaker, A. Wild, D. J. Willey, T. E. Wilmer, J. M. Wood, P. W. Wray, J. C. Wyatt, L. Young, R. M. Younger, D. R. Bentley, A. Coulson, R. Durbin, T. Hubbard, J. E. Sulston, I. Dunham, J. Rogers & S. Beck*

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

* These authors contributed equally to this work

Chromosome 6 is a metacentric chromosome that constitutes about 6% of the human genome. The finished sequence comprises 166,880,988 base pairs, representing the largest chromosome sequenced so far. The entire sequence has been subjected to high-quality manual annotation, resulting in the evidence-supported identification of 1,557 genes and 633 pseudogenes. Here we report that at least 96% of the protein-coding genes have been identified, as assessed by multi-species comparative sequence analysis, and provide evidence for the presence of further, otherwise unsupported exons/genes. Among these are genes directly implicated in cancer, schizophrenia, autoimmunity and many other diseases. Chromosome 6 harbours the largest transfer RNA gene cluster in the genome; we show that this cluster co-localizes with a region of high transcriptional activity. Within the essential immune loci of the major histocompatibility complex, we find *HLA-B* to be the most polymorphic gene on chromosome 6 and in the human genome.

Following the announcement of the completion of the human genome project on 14 April 2003, we present here our findings on the mapping, sequencing and analysis of chromosome 6. Chromosome 6 was best known for the major histocompatibility complex (MHC), a region of 3.6 megabases (Mb) on band 6p21.3 of the short arm. The MHC has an essential role in the innate and adaptive immune system, and is characterized by high gene density, high polymorphism and high linkage disequilibrium. Much of what we know today about genetic variation and the organization of haplotypes was first discovered from studies of this region. At a time when genetic variation was assessed by serology rather than sequence, the term 'haplotype' was first introduced to describe "the combination of individual antigenic [MHC] determinants that are positively controlled by an allele"[1]. Because of its crucial role in immunity and its association with many common diseases, the MHC was sequenced well ahead of the rest of chromosome 6 (ref. 2).

Particular care was taken to ensure that the highest quality was achieved for the sequence, analysis and annotation of chromosome 6. The annotation of all gene structures was manually checked and, in some cases, led to the correction of known reference genes. In addition to the genome sequences of *Mus musculus* and *Tetraodon nigroviridis*, the comparative analysis was enhanced by the inclusion (for the first time in the analysis of human chromosomes) of the recently assembled genomes of *Rattus norvegicus*, *Fugu rubripes* and *Danio rerio*. Our analysis is available through the new vertebrate genome annotation (VEGA) database (http://vega.sanger.ac.uk/), making the chromosome 6 annotation a high-quality and instantly available resource.

Clone map and sequence map

Bacterial clone contigs were assembled using restriction enzyme fingerprinting and sequence-tagged site (STS) content analysis of the clones, anchored to a radiation hybrid (RH) map with a marker density of 16 per Mb. A tiling path of 1,797 clones and polymerase chain reaction (PCR) fragments (see Supplementary Table S1) were selected for sequencing spanning the chromosome in nine contigs separated by gaps of 50–200 kilobases (kb), as estimated by DNA fibre fluorescence *in situ* hybridization (FISH) (see Supplementary Table S2). All but two gaps (gaps 2 and 6) reside in the pericentromeric or sub-telomeric chromosomal regions. We assessed the chromosome coverage in several ways. First, 38% of the clones selected for sequencing were hybridized to metaphase chromosomes using FISH. This provided independent support of the map construction and also highlighted the presence of intra- and inter-chromosomal repeats. Next we identified known chromosome 6 markers in both genetic (deCODE[3] and Marshfield comprehensive genetic maps[4]) and RH maps ($n = 3{,}036$). D6S1694 was the only genetic marker found to be absent from the sequence. The position of D6S1694 on these maps indicates that it is likely to reside within gap 6, between the sequences AL135906 and AL731777. We also accounted for all RefSeq genes mapping to chromosome 6. In the final sequence, no RefSeq gene was entirely missing. Three RefSeq

genes were found to be only partially present in the sequence (T. Furey, personal communication). Two of these, *LATS1* (NM_004690) and *C6orf35* (NM_018452), extended into sequence gaps for which there is now unfinished sequence that is confirmed to contain them: BX322632 and BX284653, respectively. Alignment of the *ASF1A* (NM_014034) messenger RNA to the genomic sequence indicates a potential deletion/polymorphism event in the P1-derived artificial chromosome (PAC) RP3-329L24 (AL132874.30), which requires further investigation. Finally, as a further check, we examined fosmid and bacterial artificial chromosome (BAC) end-sequences aligned to the chromosome 6 sequence in the University of California, Santa Cruz Genome Browser (http://genome.cse.ucsc.edu/) to verify the final sequence assembly.

Of the 1,797 tiling path clones, 1,739 were sequenced and finished at the Sanger Institute. For a small number of clones, the initial draft sequence was carried out by others who are credited in the corresponding EMBL (European Molecular Biology Laboratory Nucleotide Sequence Database)/GenBank/DDBJ (DNA Data Bank of Japan) database submissions. The remaining 58 clone sequences have all been published previously (see Supplementary Table S3). At the telomeres, half-YAC (yeast artificial chromosome) analysis determined that the sequence extends to within 5 kb and 3 kb of the $(TTAGGG)_n$ telomeric repeat motif for the shortest known allele of the p and q arms, respectively (H. Riethman, personal communication; http://www.wistar.upenn.edu/Riethman/). We detected alpha satellite repeat sequences flanking either side of the centromere, but the higher-order alpha satellite repeat structure characteristic of the core centromere has only been reached for the long arm. The finished chromosome sequence is 166,880,988 base pairs (bp) in length, and we estimate that this represents 99.5% of the euchromatin. Independent quality assessment[5] confirmed the sequence to be at least 99.99% accurate. The current sequence assembly (v1.4) and that used in the analysis presented in this study (v1.3) are available at http://www.sanger. ac.uk/HGP/Chr6/.

Gene index

Crucial to the utility of a reference sequence is the associated annotation. On the basis of human interpretation of supporting evidence, we annotated 2,190 gene structures on the finished sequence (http://vega.sanger.ac.uk/). These gene loci were classified as previously defined[6]. Briefly, we identified 772 'known genes', identical to known human complementary DNAs or protein sequences; 287 'novel genes', with an open reading frame (ORF) and identity to spliced expressed sequence tags (ESTs); 213 'novel transcripts', as for novel genes but with an ambiguous ORF; 285 'putative genes', with identity to spliced ESTs not containing an ORF; and 633 pseudogenes, similar to known proteins but containing frameshifts and/or stop codons that disrupt the ORF. We searched for CpG islands 2 kb upstream and 1 kb downstream of the 5′ and 3′ ends of each annotated gene, and found that 61% of known genes were associated with a CpG island, consistent with previous estimates.

Table 1 summarizes the main features of the annotation. Excluding pseudogenes, the overall chromosome 6 gene density is 9.2 genes per Mb; after accounting for the gene-rich MHC (43 genes per Mb), it represents a relatively gene-poor chromosome. The length of sequence occupied by annotated genes (including introns but excluding pseudogenes) is 70,396,075 bp or 42.2% (mean gene length is 31,195 bp). This is comparable to previous estimates for chromosomes 7, 14, 20 and 22 (46.5%, 43.6%, 42.4% and 51%, respectively). The chromosome sequence occupied by exons is only 2.2% with a mean exon length of 281 bp. The longest coding exon (9,114 bp) belongs to the known gene *ZNF451*; the *BPAG1* gene has the most exons with 101. The longest intron is the first intron of the *TCBA1* gene, spanning 479 kb. The largest gene is the *PARK2* gene at 6q24, which spans almost 1.4 Mb, has 12 exons and is mutated in patients with juvenile onset parkinsonism[7]. Finally, the mean number of transcripts annotated per gene is 2.34 (excluding putative genes), and the *FYN* oncogene has the most with 16 annotated transcripts.

Compared with protein-coding genes, non-protein-coding RNA genes are difficult to predict and verify experimentally. They can be divided into a growing number of classes involved principally in structural, catalytic or regulatory function[8]. tRNA genes are one of the functionally best-understood classes and can be predicted with high confidence. We determined the distribution of 616 tRNA genes across the human genome (Fig. 1). The most striking cluster contains 157 tRNAs, including all major species except for Asn-tRNA and Cys-tRNAs, and is located on chromosome 6p within the extended MHC class I region (shown as a double peak in Fig. 1). Other notable clusters are located on chromosomes 1, 5, 7, 14, 16 and 17. Except for the cluster on chromosome 7 (90% Cys-tRNAs), all these clusters contain different mixtures of tRNA genes, excluding tandem duplication of individual tRNAs as the main mechanism for their origin. RNAs including tRNAs are required in enormous quantities in the cell, constituting about 80% of cellular transcription in eukaryotes[9]. We tested whether tRNA genes co-localize with chromosomal regions that have higher-than-average transcriptional activity (transcription hotspots) of other genes, by identifying putative transcription hotspots using a database of non-normalized human ESTs. Disregarding the effects of RNA turnover, this analysis revealed the major peaks of transcription shown in Fig. 1. Sixty per cent of the tRNA clusters co-localize with predicted transcription hotspots, including the MHC, and an exact randomization test[10] shows that this association is highly nonrandom ($P < 0.001$; see Methods). We postulate that selection-mediated recruitment and/or hitchhiking might be mechanism(s) responsible for the observed co-localization. This is another demonstration of the concept that chromosomal location matters, as has previously been emphasized by studies showing large-scale (multi-Mb) changes in chromatin organization following transcriptional activation[11].

Chromosome features

Figure 2 and Supplementary Fig. S1 summarize all the features identified on chromosome 6 in our analysis. Supplementary Fig. S1 provides a detailed view, including separate tracks for tiling path clones, genetic markers, various types of repeat, G+C content, CpG

Table 1 **Gene classification and statistics**

Category *	Number of genes	Total gene length (bp)	Mean gene length (bp)	Mean exon length (bp)	Mean transcripts per gene	Mean exons per gene
Known genes	772	50,364,469	65,239	303	2.78	9.95
Novel CDS	287	10,180,377	35,472	309	1.86	5.81
Novel transcripts	213	7,058,938	33,141	262	1.40	3.25
Known and novel genes	1,272	67,603,784	53,148	301	2.34	7.90
Putative genes	285	2,792,291	9,798	248	1.13	2.36
Non-pseudogenes	1,557	70,396,075	45,213	298	2.12	6.88
Pseudogenes	633	844,936	1,335	568	1.00	1.34
Total gene structures	2,190	71,241,011	32,530	318	1.79	5.28

*As defined by Deloukas *et al.*[6] and at http://vega.sanger.ac.uk/.

content, single-nucleotide polymorphisms (SNPs, random and total), CpG islands, transcription start sites, similarity to mouse, rat, *Fugu*, *Tetraodon* and zebrafish, gene clusters and genes/pseudogenes.

Repeat sequences are the most abundant features shaping the chromosomal landscape, providing a detailed 'fossil record' of the chromosome. The repeat content of chromosome 6 is 43.95%. Transposon-derived repeats such as long interspersed elements (LINEs), short interspersed elements (SINEs), long terminal repeat (LTR) retrotransposons and DNA transposons occupy 20.85%,

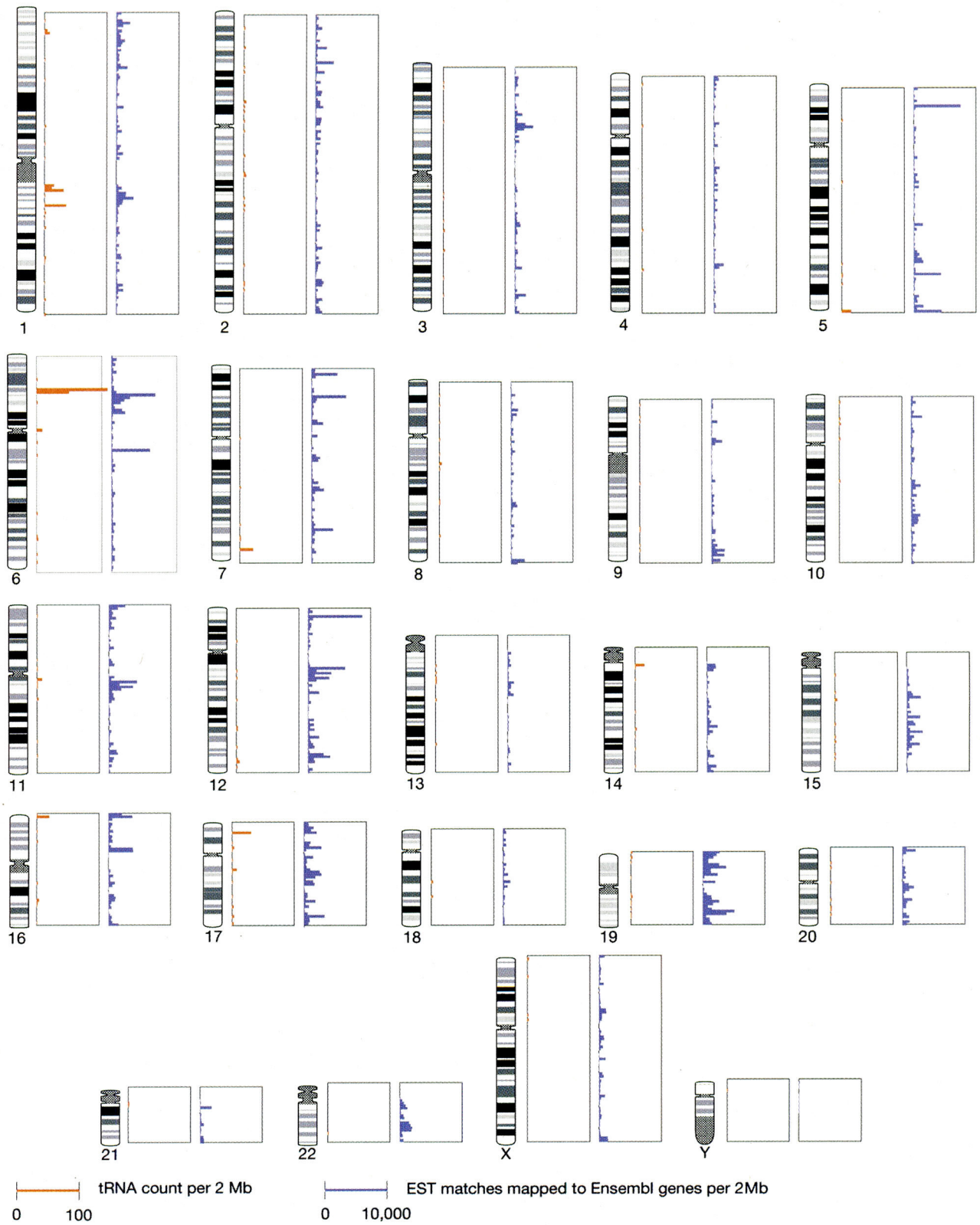

Figure 1 Distribution of tRNA genes and transcription hotspots in the human genome. The red bars to the right of the ideogram for each human chromosome represent the tRNA count per 2 Mb of sequence. The blue bars represent the number of non-normalized EST matches mapped to Ensembl genes in a 2 Mb window.

11.29%, 8.05% and 3.24% of the sequence, respectively. The repeat analysis also revealed an unusually striking structure for the human genome sequence, consisting of 17 tandemly arranged, adjoining *Alu* repeat fragments within the clone RP11-13J16 (AL499606) in 6pter, indicative of a complex series of duplications. There is mounting evidence that *Alu* repeat clusters are mediators of recurrent chromosomal aberrations[12], and it is interesting to note that there are a number of disease phenotypes involving chromosome rearrangements at 6pter, including glaucoma[13], orofacial cleft palate[14] and neoplasms (http://cgap.nci.nih.gov/Chromosomes/Mitelman/).

The sex-averaged genetic length of chromosome 6 in the deCODE map[3] is 189.60 cM, giving a chromosome average recombination rate of 1.11 cM per Mb. In a previous study, Yu and colleagues[15] identified only one recombination 'desert' on chromosome 6 (defined as a sex-averaged recombination rate of ≤0.3 cM per Mb for physical distances up to 5 Mb in length) between markers D6S1706 and D6S1608, and no recombination 'jungles' (>3 cM per Mb). However, analysis of the deCODE map, which offers a fivefold increase in resolution over previous genetic maps and high-resolution recombination maps in the MHC based on single-sperm typing[16,17], poses some questions about the desert and jungle concept. The classical MHC, a 3.6 Mb region at 6p21.3, has an overall low sex-averaged recombination rate across its length in the genetic map (0.49 cM per Mb; Fig. 3) and yet three hotspots of recombination are observed at high resolution[17] (Fig. 3, inset). The finding of recombination hotspots within a region of low long-range recombination rate illustrates that even the current genome-wide genetic map masks local recombination hotspots/rates, and hence the description of Mb regions as containing high or low recombination may be of limited value in understanding recombination rates. Despite this, we were able to identify several small regions of interest. We found that 116 marker intervals had recombination rates greater than the chromosome average (1.11 cM per Mb), ten of these showing greater than a fivefold increase and four intervals showing a tenfold increase in recombination rate (Table 2). In the four intervals with the highest recombination rates, the markers were within 20 kb of each other in the sequence. It could be that these intervals pin-point the location of recombination events and provide indicators of possible hotspot locations, although genotyping errors could also result in inflated recombination rates within these intervals.

Recent studies suggest that around 5% of the human genome has arisen from segmental duplications[18]. We initially identified large segmental duplications on chromosome 6 by the observed 'stacking' of restriction enzyme fingerprints in the clone assembly as a result of near-identical sequences (and therefore restriction sites). One such duplication lies on 6p, which harbours two pseudogenes of the ancestral β-glucuronidase (*GUSB*) gene on 7q11.21, one in the extended MHC at 6p21.31, and the second in the pericentromeric region 6p11.1. The spinal muscular atrophy (*SMA*) locus of chromosome 5q13 and three other chromosome 5 regions also contain paralogous *GUSB* sequences. Clones mapping to these regions by fingerprint and STS-content analysis (for example, RP11-239L20 and b55C20) were used in FISH to metaphase chromosomes and showed multiple signals on chromosomes 6, 5 and 22 (Supplementary Table S4), in agreement with recently published data[19].

Table 2 **Inter-marker recombination rates**

Marker left	Marker right	Interval size (bp)	Sex-averaged position in deCODE map (cM)	Recombination rate (cM per Mb)
D6S1038	D6S1722	6,245	132.25	59.25
D6S383	D6S970	6,284	145.75	33.42
D6S444	D6S417	3,677	99.71	19.04
D6S1571	D6S1545	19,075	49.26	12.58
D6S1588	D6S1029	32,586	44.86	10.74
D6S1578	D6S1559	132,974	33.27	10.08
D6S407	D6S1690	20,782	127.89	9.62
D6S464	D6S105	39,531	50.57	8.09
D6S942	D6S1600	65,330	0	6.89
D6S1697	D6S503	193,688	184.1	5.89

Both large- and small-scale duplications are required to explain the age distribution of human gene families[20]. To assess the effect of local duplications on the number of genes on chromosome 6, we carried out a gene cluster analysis (excluding pseudogenes) using BLASTP. For the purpose of this analysis, a cluster was defined as two or more paralogous genes (*e*-values $<1 \times 10^{-15}$) within 1 Mb of each other. Adjacent clusters of the same gene family were further grouped into superclusters. In this analysis, we found 65 clusters (Supplementary Fig. S1 and Supplementary Table S5), of which 25 could be grouped into six superclusters, resulting in 46 gene family clusters containing 2–55 paralogues per cluster. In all, 223 (14.3%) out of the 1,557 annotated genes on chromosome 6 seem to have arisen through local duplication. This number is likely to increase if global (genome-wide) duplication events are considered as well. Interestingly, all major gene clusters locate to the extended MHC region on the short arm of chromsome 6 (ref. 21), with the exception of the recently described *RAET* genes[22]. The remaining 39 clusters contain two to four paralogues each and are scattered randomly across chromosome 6. In addition, we carried out a cluster analysis based on known protein domains using InterProScan (http://www.ebi.ac.uk/interpro/scan.html), resulting in a similar, although not identical, ranking of top-scoring clusters (Supplementary Table S6).

Comparative analysis

The recent accumulation of assembled genome sequences for mouse, rat, *Tetraodon*, *Fugu* and zebrafish has allowed us to evaluate our gene annotation. The results are summarized in Table 3 and Supplementary Table S7. Because comparative genome analysis was not used in the gene structure annotation, it is possible to use it to estimate the completeness of annotation by considering regions that are conserved in all six organisms (the assumption being that such completely conserved regions are unlikely to be the result of artefact). There are 5,409 such regions, of which 5,204 have overlap with 4,466 annotated exons, suggesting that 95.6% (4,466/(4,466 + 205)) of coding exons on chromosome 6 have been annotated. This estimate is likely to be low because not all sequence conservation will be confined to exons. The figure is, however, in agreement with similar estimates of annotation coverage made for chromosomes 20 (ref. 6) and 14 (ref. 23), and to a lesser extent chromosome 22 (ref. 24), although slightly different methods and resources were used in each case.

It is also interesting to consider the conservation between chromosome 6 and groups of species. For example, regions of the human sequence which are conserved in either rat, mouse, *Fugu*, zebrafish or *Tetraodon* account for more exons than regions of conservation with any single species in isolation: 84% of genes and 80% of exons (98% and 88% respectively if we consider only "Known genes") are supported by conservation with at least two

Figure 2 Features of the human chromosome 6 sequence. Feature tracks from left to right are as follows. (1) Sequence scale in megabases. (2) Extent of human chromosome 6 sequence (black) separated by gaps (grey). (3) Rodent synteny: mouse, left; rat, right. (4) Location of predicted CpG islands. (5) From left to right, the location of sequence homology to *Fugu*, zebrafish and *Tetraodon*. (6) The location of the 'known' and 'novel CDS' (novel coding sequence) annotated gene structures. Official gene symbols are used when available. Because of space limitations, the foldout shown in this figure had to be condensed for the printed version. Therefore, we strongly recommend the downloading of Supplementary Fig. S1 to follow points discussed in the main text. ►

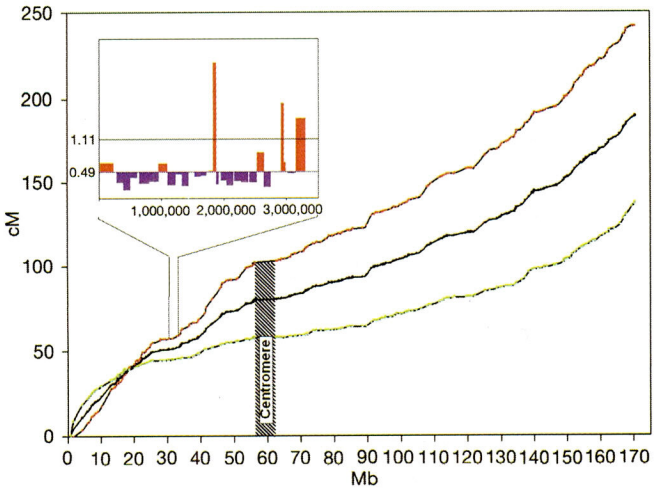

Figure 3 Alignment of the genetic markers in the deCODE linkage map[3] with the chromosome 6 sequence. The physical position of each genetic marker on the female, the male and the sex-averaged genetic map is indicated. Inset, high-resolution recombination intensity across the 3.6 Mb classical MHC region at 6p21.3 (ref. 17).

of the above species, 81% of genes and 77% of exons when considering the best single species comparison (mouse) (see Supplementary Information). Figure 4 illustrates this point and emphasizes the potential of multi-species comparative analysis[25] for the identification of evolutionarily conserved regions (ECRs). In this case, the three identified ECRs have an open reading frame and putative splice sites suggesting that they constitute part of a potential novel gene.

Sequence variation

The most common variations in the human genome are SNPs, and their exploration and use is expected to revolutionize our understanding of human disease and evolution[26]. Towards this end, we have mapped a total of 183,019 SNPs onto the finished sequence of chromosome 6. The non-normalized SNPs were taken from the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/), where they have been deposited by the International SNP Consortium (TSC)[27] and others. Of these SNPs, 113,284 (62%) were discovered as part of the chromosome 6 project itself by sequence comparison of clone overlaps using a modification of the SSAHA program[28]. After correlating the positions of all SNPs to the annotation presented here, chromosome 6 contains 2,761 SNPs in protein-coding exons. When applied genome-wide (using the Ensembl annotation), followed by further clustering of the SNPs according to gene structures and normalizing for coding-sequence length, this analysis provides an estimate of the most polymorphic genes in the genome. With 86 SNPs per kb mapping to *HLA-B* (human leukocyte antigen, class I, B), we determined this MHC class I gene to be the most polymorphic gene on chromosome 6 and in the human genome, closely followed by other class I and class II genes. This result is concordant with the HLA database (http://www.ebi.ac.uk/imgt/hla/stats.html), which lists *HLA-B* as the most polymorphic MHC gene, with 511 known alleles.

Supplementary Fig. S1 shows the distribution of SNPs across chromosome 6. The lower track shows the total SNP density but is likely to be skewed by local SNP-discovery efforts. The upper SNP track shows the distribution of a random set of SNPs generated by the TSC and is, therefore, a better indicator of regions of high and low sequence variation. As expected from previous studies[2], the most variable regions map to segments containing MHC class I and class II genes. Other notable regions of high sequence variation are located at 26.7–27.0 Mb, 57.4–57.5 Mb and 58.2–58.5 Mb on the short arm of the chromosome. Interestingly, all three regions contain segmental duplications, supporting the hypothesis that paralogous sequence variants may have been falsely identified as SNPs in dbSNP[18], thereby increasing the apparent density of 'SNPs'.

Medical implications

At the time of writing, there are 130 genes mapping to chromosome 6 that cause, predispose to or protect from disease (Supplementary Table S8). Of these, 84 (65%) have already been cloned and include the well-studied MHC class I-like gene, *HFE*[29], mutated in hereditary haemochromatosis. With respect to autoimmunity, the MHC is the most important genetic region in the human genome. Well over 100 diseases have been linked to the MHC, including most if not all autoimmune diseases[30]. Autoimmune disorders are common, complex, and involve genetic and environmental factors[31]. They affect about 4% of the population, and include type 1 diabetes, rheumatoid arthritis and multiple sclerosis, to name but a few.

Genes with protein products that affect the brain or neural tissue are of particular interest in the study of neurological or psychiatric disorders. Examples of chromosome 6 genes that have been cloned include *SCA1*, which can cause spinocerebellar ataxia when mutated; *EPM2A*, a gene mutated in patients with Lafora's myoclonus epilepsy[32]; and the *PARK2* gene, point mutations or deletions in which are responsible for a juvenile-onset autosomal-recessive form of parkinsonism[7]. The application of genetic mapping to human disease gene identification has led to the definition of many linkages for, or associations with, complex traits on chromosome 6. One such trait is schizophrenia. Recent studies[33,34] found family-based association with the dysbindin (*DTNBP1*) gene at 6p22. Dysbindin is a promising candidate because it is localized to presynaptic terminals, and could be involved in the formation and/or maintenance of synapses and in signal transduction. *HTR1E*, at 6q14.2, and *B3GAT1*, at 11q25, have recently been identified as candidate genes for a schizophrenia-like psychosis through molecular analysis of a balanced translocation t(6:11)(q14.2:q25)[35]. Sequence analysis of the translocation breakpoints reveals that neither gene is disrupted; however, the expression of these genes could potentially be altered by a positional effect due to alteration of the chromatin environment as a consequence of the translocation.

The complete sequence of chromosome 6 not only allows us to look at genetic conditions, but also provides a tool with which to study epigenetic changes resulting in disease[36]. One such disease is transient neonatal diabetes mellitus (TNDM), which is the result of overexpression of an imprinted and paternally expressed gene(s) at 6q24. Two imprinted genes at 6q24, *PLAGL1* and *HYMAI*, have been identified as potential candidates for TNDM through studies of paternal uniparental isodisomy of chromosome 6 (UPD6), paternally inherited duplication of 6q24, and a methylation defect at a

Table 3 **Comparative sequence analysis of the annotation set**

Species	Known		Novel CDS		Novel transcript		Pseudogene		Putative		Total	
	G	E	G	E	G	E	G	E	G	E	G	E
Rodents and fish	545 (0.71)	3,495 (0.33)	163 (0.57)	567 (0.26)	15 (0.07)	27 (0.03)	319 (0.50)	345 (0.41)	9 (0.03)	32 (0.04)	1,051 (0.48)	4,466 (0.30)
Rodents or fish	757 (0.98)	9,214 (0.88)	275 (0.96)	1,776 (0.82)	108 (0.51)	226 (0.26)	592 (0.94)	719 (0.85)	111 (0.39)	175 (0.24)	1,843 (0.84)	12,110 (0.80)

The number of annotated genes and exons in each category supported by an overlapping conserved region. A more detailed breakdown of the comparative sequence analysis by species is given in Supplementary Information. G, Genes; E, Exons.

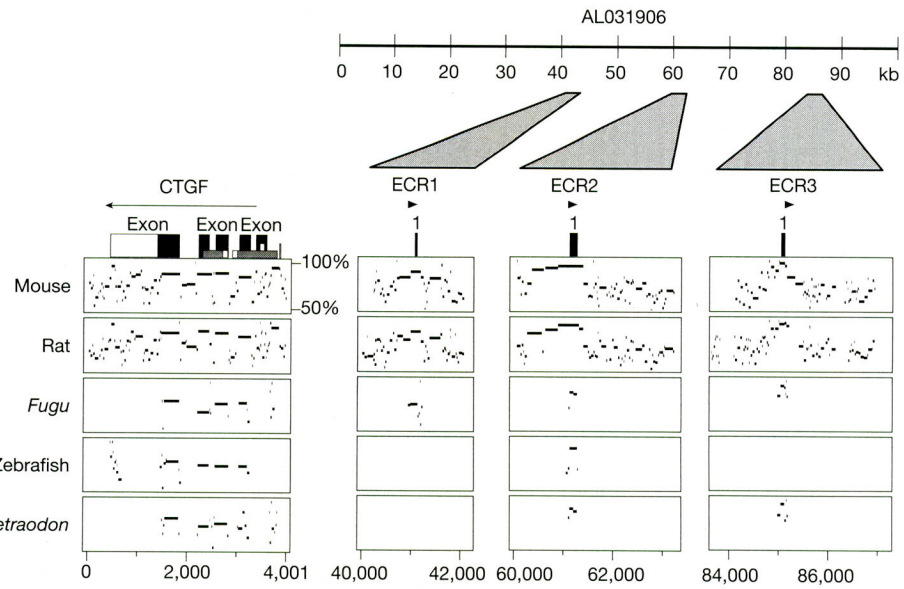

Figure 4 Identification of evolutionarily conserved regions using multi-species comparative analysis. Pair-wise alignments between the chromosome 6 sequence and the draft sequences of each indicated species were generated and displayed using MultipPipMaker[49]. The percentage identity of each alignment is indicated. The first plot shows the known connective tissue growth factor (*CTGF*) gene aligned to syntenic draft sequences of rodents and fish. Three ECRs identified from the comparative sequence analysis were found within 50 kb of sequence AL031906.

CpG island. A second cluster of imprinted genes occurs at 6q26, a region actively studied for the presence of tumour-suppressor genes (TSGs).

Tumour-specific mutations that inactivate TSGs often involve large-scale changes, such as loss of the entire chromosome or large fragments of it containing the TSG. The long arm of chromosome 6 has been a focus of attention for cancer geneticists because it harbours genes of importance in tumour progression in a diverse set of solid and haematological malignancies (see ref. 37 and references therein). DNA, amplified by PCR, from the clones used in the sequencing of chromosome 6 have been arrayed for comparative genomic hybridization, and are currently being used to define commonly deleted and/or homozygous deletions in astrocytic gliomas, the results of which will be published elsewhere. The use of sequenced clones in the array provides a direct link back to the sequence and its features, and therefore provides a new and powerful tool in the search for TSGs. Furthermore, the identification of the genetic loci and associated polymorphisms responsible for complex traits should greatly benefit from the availability of the complete sequence. □

Methods

The strategy and method of construction of the bacterial clone physical map of chromosome 6 has been described elsewhere[38,39]. Briefly, 2,802 STSs on the radiation hybrid map of chromosome 6 (http://www.sanger.ac.uk/cgi-bin/rhtop?chr=6) were used to screen up to 87 genomic equivalents (87× coverage) of BAC, PAC, cosmid and YAC libraries. Identified clones were fluorescently fingerprinted and their landmark content established to generate clone contigs. End-sequencing of terminal clones in the contigs was performed and novel STSs were designed for further walking experiments. In regions devoid of bacterial clone coverage, STSs flanking the gaps were used to screen YAC libraries. Multiple checks were performed on each clone selected for sequencing from the map.

Random shotgun sequences of bacterial clones were generated from pUC plasmids with inserts of mainly 1.4–4 kb, which were sequenced from both ends using the dideoxy chain termination method with different versions of big dye terminator chemistry[40]. The resulting sequencing reactions were analysed on various models of ABI sequencing machines, and the data generated were processed by a suite of in-house programs (http://www.sanger.ac.uk/Software/sequencing/) before assembly with the PHRED and PHRAP (http://www.phrap.org/) algorithms. For the finishing phase, we used the GAP4 program[41] to help us to assess, edit and select reactions, to eliminate ambiguities and to close sequence gaps. All DNA sequence has been deposited in the EMBL, GenBank or DDBJ databases.

The finished genomic sequence was analysed using an automatic Ensembl pipeline[42] with modifications to aid the manual curation process. The identification of interspersed repeats using RepeatMasker; simple repeats using Tandem Repeat Finder[43]; matches to vertebrate cDNAs and ESTs using WU-BLASTN and EST_GENOME; and *ab initio* gene prediction using FGENESH and GENESCAN was as described previously[6]. A protein database combining non-redundant data from SwissProt and TrEMBL was also searched using WU-BLASTX. Using these pipeline results as a starting point, gene structures were manually annotated according to the human annotation workshop (HAWK) guidelines (http://www.sanger.ac.uk/HGP/havana/hawk.shtml) and, where possible, were given gene symbols approved by the HUGO and HLA Gene Nomenclature Committees[44,45]. The sequence was further analysed to identify putative tRNA genes using tRNAscan-SE[46].

For comparison with mouse and rat, the repeat-masked sequence of chromosome 6 was aligned to these genomes using BLASTZ[47]. The resulting matches were post-processed using the programs axtBest and subsetAxt, as previously described[47], to select the best match and to make the alignments relatively specific to exons by using a specific scoring matrix and threshold. For the comparative analysis with fish, genome sequences were aligned to the chromosome using WU-TBLASTX, with the scoring matrix, parameters and filtering strategy used in the Exofish (exon finding by sequence homology) procedure[48]. Overlapping alignments to different sequences were merged to produce contiguous regions of sequence conservation, analogous to the evolutionarily conserved regions, or 'Ecores', reported by Exofish.

SNPs determined as part of this sequencing project were identified from sequence overlaps using a modification of the SSAHA program[28]. Clone overlaps from available finished and unfinished public human genomic sequence were aligned, and high-quality base discrepancies ($Q \geq 23$) were identified as candidate SNPs provided that the total overlap was >6,000 bp. Candidate SNPs were rejected if any of its neighbouring five bases had Phrap quality values of <15 (bases of finished sequence were assumed to have a quality value of 40; that is, one error in 10,000 bp) or if fewer than nine of the ten neighbours matched. If the number of detected SNPs in one clique was greater than five in a 500 bp interval, then all SNPs were discarded for that interval. All SNPs for a clone overlap were similarly discarded if the SNP density for the entire overlap region was less than one SNP per 4,000 bp. In some overlap regions, more than two clones overlap. A SNP detected from multiple overlapping clones was merged and represented as one SNP.

For transcription hotspot analysis, human ESTs derived from non-normalized libraries were matched to the human genome sequence with BLAT using the Ensembl pipeline. Matches were accepted only if they showed >97% nucleotide identity over >90% of the EST sequence; if they were the single best identity match for an individual EST; and if they exhibited more than 80% coverage with an Ensembl gene. The null hypothesis that the association of the tRNA and expression peaks is random was then tested: with the genome divided into 2 Mb windows, peaks of expression and tRNA clusters were defined as those windows where the number of matches was greater than the mean value by more than one standard deviation. The positions of the resultant tRNA peaks were randomized, and a distribution of the associations with an expression peak (either coincident or within 2 Mb of the expression peak) was generated from 10,000 replicates. The observed association of peaks was compared to this distribution with a *P*-value calculated as the proportion of replicates where an equal or greater number of tRNA peaks was associated with an expression peak.

Received 8 June; accepted 11 September 2003; doi:10.1038/nature02055.

1. Cepellini, R. *et al.* in *Histocompatibility Testing* (eds Curtoni, E. S., Mattiuz, P. L. & Tosi, R. M.) 149–184 (Munksgaard, Copenhagen, 1967).
2. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401,** 921–923 (1999).
3. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
4. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
5. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).
6. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
7. Kitada, T. *et al.* Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392,** 605–608 (1998).
8. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2,** 919–929 (2001).
9. Paule, M. R. & White, R. J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* **28,** 1283–1298 (2000).
10. Sokal, R. R. & Rohlf, F. J. in *Biometry* 3rd edn 803–819 (Freeman and Company, New York, 1995).
11. Volpi, E. V. *et al.* Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.* **113,** 1565–1576 (2000).
12. Kolomietz, E., Meyn, M. S., Pandita, A. & Squire, J. A. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes Cancer* **35,** 97–112 (2002).
13. Lehmann, O. J. *et al.* Ocular developmental abnormalities and glaucoma associated with interstitial 6p25 duplications and deletions. *Invest. Ophthalmol. Vis. Sci.* **43,** 1843–1849 (2002).
14. Davies, A. F. *et al.* Evidence of a locus for orofacial clefting on human chromosome 6p24 and STS content map of the region. *Hum. Mol. Genet.* **4,** 121–128 (1995).
15. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
16. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29,** 217–222 (2001).
17. Cullen, M., Perfetto, S. P., Klitz, W., Nelson, G. & Carrington, M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Hum. Genet.* **71,** 759–776 (2002).
18. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
19. Courseaux, A. *et al.* Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. *Genome Res.* **13,** 369–381 (2003).
20. Gu, X., Wang, Y. & Gu, J. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nature Genet.* **31,** 205–209 (2002).
21. Trowsdale, J. The gentle art of gene arrangement: the meaning of gene clusters. *Genome Biol.* **3,** COMMENT2002 (2002).
22. Radosavljevic, M. *et al.* A cluster of ten novel MHC class I related genes on human chromosome 6q24.2–q25.3. *Genomics* **79,** 114–123 (2002).
23. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
24. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13,** 27–36 (2003).
25. Thomas, J. W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793 (2003).
26. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* **2,** 930–942 (2001).
27. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
28. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
29. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.* **13,** 399–408 (1996).
30. Tiwari, J. L. & Terasaki, P. I. *HLA and Disease Associations* (Springer Verlag, Berlin, 1985).
31. Vyse, T. J. & Todd, J. A. Genetic analysis of autoimmune disease. *Cell* **85,** 311–318 (1996).
32. Minassian, B. A. *et al.* Mutations in a gene encoding a novel protein tyrosine phosphatase cause progressive myoclonus epilepsy. *Nature Genet.* **20,** 171–174 (1998).
33. Straub, R. E. *et al.* Genetic variation in the 6p22.3 gene *DTNBP1*, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am. J. Hum. Genet.* **71,** 337–348 (2002).
34. Schwab, S. G. *et al.* Support for association of schizophrenia with genetic variation in the 6p22.3 gene, dysbindin, in sib-pair families with linkage and in an additional sample of triad families. *Am. J. Hum. Genet.* **72,** 185–190 (2003).
35. Jeffries, A. R. *et al.* β-1,3-Glucuronyltransferase-1 gene implicated as a candidate for a schizophrenia-like psychosis through molecular analysis of a balanced translocation. *Mol. Psychiatry* **8,** 654–663 (2003).
36. Novik, K. L. *et al.* Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* **4,** 111–128 (2002).
37. Acquati, F. *et al.* Cloning and characterization of a senescence inducing and class II tumor suppressor gene in ovarian carcinoma at chromosome region 6q27. *Oncogene* **20,** 980–988 (2001).
38. Mungall, A. J. *et al.* From long range mapping to sequence-ready contigs on human chromosome 6. *DNA Seq.* **8,** 151–154 (1997).
39. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409,** 942–943 (2001).
40. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
41. Bonfield, J. K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23,** 4992–4999 (1995).
42. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
43. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
44. Wain, H. M., Lush, M., Ducluzeau, F. & Povey, S. Genew: the human gene nomenclature database. *Nucleic Acids Res.* **30,** 169–171 (2002).
45. Robinson, J. *et al.* IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res.* **31,** 311–314 (2003).
46. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
47. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
48. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
49. Schwartz, S. *et al.* PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10,** 577–586 (2000).

Supplementary Information accompanies the paper on **www.nature.com/nature**.

Acknowledgements We thank H. Riethman and R. Sudbrak for the provision of clones, H. M. Wain, R. C. Lovering, E. A. Bruford, M. J. Lush, M. W. Wright, V. K. Khodiyar, C. C. Talbot and S. Povey from the HUGO Gene Nomenclature Committee for official gene nomenclature, S. G. Marsh and the WHO HLA Nomenclature Committee for HLA gene nomenclature, Y. Chen for assistance with the SNP mapping, the EMBL and Ensembl database teams at the European Bioinformatics Institute, the HUGO chromosome editors R. D. Campbell, H. M. Cann, E. Jazwinska, J. Ragoussis and A. Ziegler for support throughout the project, P. Deloukas for critical reading of the manuscript, and the Wellcome Trust for financial support.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.J.M. (ajm@sanger.ac.uk) or S.B. (beck@sanger.ac.uk). Accession numbers for the sequence analysed for this paper can be found in Supplementary Fig. S1. All reported DNA sequences have been deposited in EMBL, GenBank or DDBJ.