The DNA sequence and analysis of human chromosome 13

A. Dunham[1], L. H. Matthews[1], J. Burton[1], J. L. Ashurst[1], K. L. Howe[1], K. J. Ashcroft[1], D. M. Beare[1], D. C. Burford[1], S. E. Hunt[1], S. Griffiths-Jones[1], M. C. Jones[1], S. J. Keenan[1], K. Oliver[1], C. E. Scott[1], R. Ainscough[1], J. P. Almeida[1], K. D. Ambrose[1], D. T. Andrews[1], R. I. S. Ashwell[1], A. K. Babbage[1], C. L. Bagguley[1], J. Bailey[1], R. Bannerjee[1], K. F. Barlow[1], K. Bates[1], H. Beasley[1], C. P. Bird[1], S. Bray-Allen[1], A. J. Brown[1], J. Y. Brown[1], W. Burrill[1], C. Carder[1], N. P. Carter[1], J. C. Chapman[1], M. E. Clamp[1], S. Y. Clark[1], G. Clarke[1], C. M. Clee[1], S. C. M. Clegg[1], V. Cobley[1], J. E. Collins[1], N. Corby[1], G. J. Coville[1], P. Deloukas[1], P. Dhami[1], I. Dunham[1], M. Dunn[1], M. E. Earthrowl[1], A. G. Ellington[1], L. Faulkner[1], A. G. Frankish[1], J. Frankland[1], L. French[1], P. Garner[1], J. Garnett[1], J. G. R. Gilbert[1], C. J. Gilson[1], J. Ghori[1], D. V. Grafham[1], S. M. Gribble[1], C. Griffiths[1], R. E. Hall[1], S. Hammond[1], J. L. Harley[1], E. A. Hart[1], P. D. Heath[1], P. J. Howden[1], E. J. Huckle[1], P. J. Hunt[1], A. R. Hunt[1], C. Johnson[1], D. Johnson[1], M. Kay[1], A. M. Kimberley[1], A. King[1], G. K. Laird[1], C. J. Langford[1], S. Lawlor[1], D. A. Leongamornlert[1], D. M. Lloyd[1], C. Lloyd[1], J. E. Loveland[1], J. Lovell[1], S. Martin[1], M. Mashreghi-Mohammadi[1], S. J. McLaren[1], A. McMurray[1], S. Milne[1], M. J. F. Moore[1], T. Nickerson[1], S. A. Palmer[1], A. V. Pearce[1], A. I. Peck[1], S. Pelan[1], B. Phillimore[1], K. M. Porter[1], C. M. Rice[1], S. Searle[1], H. K. Sehra[1], R. Shownkeen[1], C. D. Skuce[1], M. Smith[1], C. A. Steward[1], N. Sycamore[1], J. Tester[1], D. W. Thomas[1], A. Tracey[1], A. Tromans[1], B. Tubby[1], M. Wall[1], J. M. Wallis[1], A. P. West[1], S. L. Whitehead[1], D. L. Willey[1], L. Wilming[1], P. W. Wray[1], M. W. Wright[2], L. Young[1], A. Coulson[1], R. Durbin[1], T. Hubbard[1], J. E. Sulston[1], S. Beck[1], D. R. Bentley[1], J. Rogers[1] & M. T. Ross[1]

[1]*The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK*
[2]*HUGO Gene Nomenclature Committee, The Galton Laboratory, Wolfson House, 4 Stephenson Way, London, NW1 2HE, UK*

Chromosome 13 is the largest acrocentric human chromosome. It carries genes involved in cancer including the breast cancer type 2 (*BRCA2*) and retinoblastoma (*RB1*) genes, is frequently rearranged in B-cell chronic lymphocytic leukaemia, and contains the *DAOA* locus associated with bipolar disorder and schizophrenia. We describe completion and analysis of 95.5 megabases (Mb) of sequence from chromosome 13, which contains 633 genes and 296 pseudogenes. We estimate that more than 95.4% of the protein-coding genes of this chromosome have been identified, on the basis of comparison with other vertebrate genome sequences. Additionally, 105 putative non-coding RNA genes were found. Chromosome 13 has one of the lowest gene densities (6.5 genes per Mb) among human chromosomes, and contains a central region of 38 Mb where the gene density drops to only 3.1 genes per Mb.

The draft sequence of the human genome[1] has provided the basis for a systematic effort to finish each chromosome[2–5] in order to produce an accurate and detailed description of the entire genome. In common with the other acrocentric autosomes (14, 15, 21, and 22) the short arm of chromosome 13 is heterochromatic and contains families of repeated sequences, including the ribosomal RNA gene arrays[6]. The long arm is euchromatic and contains most or all of the protein-coding genes of the chromosome. We have completed the euchromatic sequence and examined the characteristics of this gene-poor autosome in relation to other human chromosomes, and to the corresponding sequence in other species. The annotation of the sequence is available via the Vertebrate Genome Annotation (VEGA) database (http://vega.sanger.ac.uk/Homo_sapiens), and will provide a platform for the continued study of medical genetics, genome instability and evolution of human chromosomes.

Clone map and finished sequence

The physical map of the long arm of chromosome 13 comprises five contigs. A minimally overlapping set of 863 clones (the 'tilepath') was selected that contained bacterial artificial chromosome (BAC) clones, supplemented with three yeast artificial chromosome (YAC) clones where no bacterial clones could be found (see Supplementary Table S1). The sequence derived from the tilepath clones comprises 95,564,076 base pairs (bp), determined to an accuracy well above 99.99% using procedures described previously[7]. At the proximal end, the finished sequence stretches into the pericentromeric region. Here the sequence is highly similar (93.9% sequence identity) to the pericentromeric region of chromosome 21. As a result, we have not so far been able to extend the map further towards the centromere. At the distal end, the finished sequence extends to within 15 kilobases (kb) of the TTAGGG telomeric repeats (http://www.wistar.upenn.edu/Riethman). Six gaps remain in the tilepath despite our screening of genomic libraries containing a combined 87-fold coverage of the chromosome. All the gap sizes were estimated by fluorescence *in situ* hybridization (FISH) analysis and represent a combined total of 1,665 kb. The long arm of chromosome 13 therefore measures 97.2 Mb, and the finished sequence covers 98.3% of the total (see http://www.sanger.ac.uk/HGP/Chr13 and Supplementary Table S2).

The finished sequence contains all markers previously mapped to chromosome 13 in the deCODE genetic map[8], 98% of those in the Marshfield genetic map[9] and 98% of those in the Genemap '99 radiation hybrid map[10]. Furthermore, there is excellent concordance between marker order on the maps and in the sequence. An additional check of the integrity of the finished sequence was obtained by examining the alignment of paired end-sequences of fosmid or BAC inserts in the finished sequence for evidence of deletions or mis-assemblies (D. Jaffe, personal communication). These alignments were consistent throughout the euchromatic sequence, with a single exception. One BAC (AL355611) was found to be ~20 kb longer than three fosmids aligned to that region. Further examination of the physical map in this region confirmed that the fingerprint patterns of the other BACs in the map are entirely consistent with the sequence of AL355611. We therefore conclude that there is a length variation between the DNAs represented in the BAC versus the fosmid libraries.

Global analysis of the sequence was performed to provide data on distribution of genes, repeats, G+C content, CpG islands, single nucleotide polymorphisms (SNPs), and recombination rates. The complete analysis of the sequence can be seen in Supplementary Fig. 1. Some specific features are highlighted below.

Table 1 **Summary of the annotated genes on chromosome 13, excluding predicted ncRNA genes**

Category	Number of genes	Total gene length (bp)	Mean gene length (bp)	Mean exon length (bp)	Mean exons per gene
Known genes	231	24,563,589	106,731	306	12.1
Novel genes	97	5,552,022	57,261	319	6.2
Novel transcripts	145	4,362,536	30,138	222	3.7
Sum of known and novel genes and novel transcripts	473	33,559,841	73,106	296	8.3
Putative genes	160	1,543,286	9,654	276	2.1
Sum of non-pseudogenes	633	34,617,102	57,068	295	6.7
Processed pseudogenes	268	252,294	946	946	1.0
Unprocessed pseudogenes	28	916,219	32,998	139	8.7
Sum of all genes	929	35,388,461	40,152	320	5.2

The rows in bold are running totals of the preceding categories.

Gene index

Gene structures were curated manually following analysis of the finished sequence by alignment to all publicly available expressed sequences and application of gene prediction algorithms. A total of 929 gene structures were identified and classified (according to the definitions described in Methods) as known genes (231), novel genes (97), novel transcripts (145), putative genes (160) and pseudogenes (296, of which 268 are processed and 28 are unprocessed) (see Table 1 and Supplementary Table S3). At the time of this analysis, one gene that was previously assigned to chromosome 13 (*RHOK*) was missing from the analysed sequence, and two other genes were only partially represented (*ATP4B* in AL4421245 and *RASA3* in AL161774). The missing parts of *ATP4B* and *RASA3* have since been identified in the sequences of two recently finished clones, BX537316 and BX537329, respectively. The sequence of BX537316, which lies adjacent to gap 4 in the chromosome map (Supplementary Table S2), also contains the first two exons of *RHOK*.

The largest gene on chromosome 13 spans 1,468,562 bp (*GPC5*). The largest exon is the single exon of the Spastic ataxia gene (*SACS*), which measures 12,865 bp. The average number of exons per gene is 5.19, ranging from single-exon genes, of which there are 44, to a gene containing 84 exons (*MYCBP2*). Predicted proteins were classified using the Interpro database (http://www.ebi.ac.uk/interpro/). The most populous protein families on chromosome 13 (Supplementary Table S4) broadly reflect those that are most populous in the genome as a whole[1]. We searched for CpG islands 5 kb upstream and 1 kb downstream of each annotated gene (see Methods), and found that 67% of known and novel genes are associated with a CpG island, which is consistent with previous reports[2,5,11].

Owing to the absence of coding potential and the lack of primary sequence conservation, non-coding RNA (ncRNA) genes pose a significant challenge for computational prediction. Furthermore, the discrimination between genes and pseudogenes by computational methods has not been possible for most ncRNA genes. However, recent software advances and the development of a database of structural RNA alignments (Rfam)[12] now permit a more comprehensive search for ncRNA genes (see Supplementary Table S5). Chromosome 13 contains only five of the 616 transfer RNA genes that have been found in the human genome. In addition, two tRNA pseudogenes were found. Rfam analysis identifies 98 additional putative ncRNA genes on chromosome 13, including 37 spliceosomal RNAs and 20 Y RNAs (components of the ribonucleoprotein particle Ro). Two of these genes, a U6 and a U2 spliceosomal RNA, are confirmed on the basis of near identity to previously published sequences.

MicroRNAs (miRNAs) are approximately 21-bp double-stranded products excised from a hairpin precursor, which regulate the expression of other genes by complementary binding to untranslated regions of a messenger RNA target. The miRNA Registry (http://www.sanger.ac.uk/Software/Rfam/mirna/) lists 147 identified human miRNAs, eight of which are found in two clusters in the chromosome 13 sequence. miR-15 and miR-16 locate

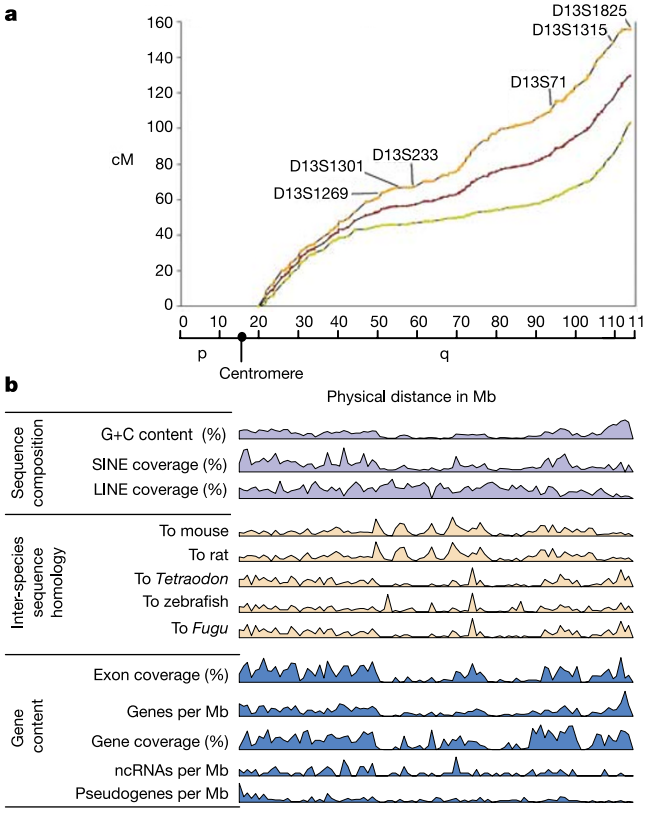
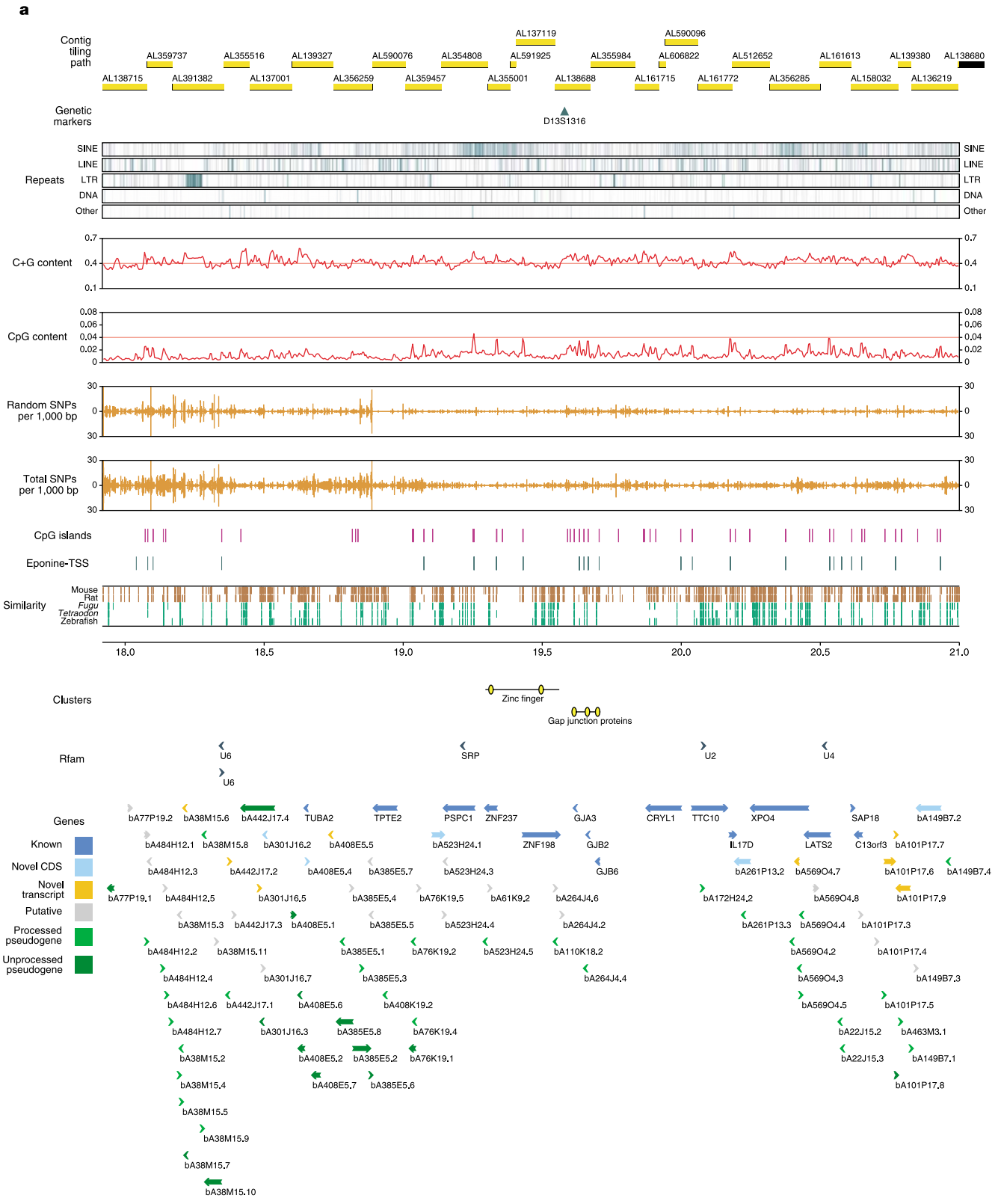


Figure 1 Genetic and physical characteristics of the chromosome 13 sequence. **a**, A comparison of physical and genetic distance along chromosome 13. Markers from the deCODE genetic map[8] were localized in the sequence. Their locations in the genetic map are plotted on the *y* axis and locations in the sequence are on the *x* axis (the sequence starts at position 17,918,001 to allow for the chromosome short arm and centromere). The female genetic map is shown in orange, the male map in green and the sex average map in red. The loci shown mark the extent of the region of low gene density (D13S1269–D13S71), the recombination jungle (D13S1315–D13S1825) and the recombination desert (D13S1301–D13S233). **b**, Variation in features along the chromosome. The sequence was divided into 1-Mb non-overlapping sections, and each section was studied for the features shown. The data for each feature were normalized to set the largest figure at 100% and the lowest at 0%. For the repeat coverage, the interspecies sequence homologies (defined in the Methods), and the exon and gene coverage, the percentage of each 1-Mb window is calculated. The range of data points for each plot are as follows: % G+C content (33.5–52.4); % SINE coverage (3.5–25); % LINE coverage (9.2–31.4); % with homology to mouse (1.1–14.2); % with homology to rat (0.9–11.9); % with homology to *Tetraodon* (0–1.8); % with homology to zebrafish (0–3); % with homology to *Fugu* (0–1.9); % exon coverage (0–4.1); genes per Mb (0–38); % gene coverage (0–100); ncRNAs per Mb (0–6); pseudogenes per Mb (0–24).

to 13q14 and are separated by only 100 bp. These genes are deleted or downregulated in patients with chronic lymphocytic leukaemia[13]. A second group of six miRNAs is clustered within an 800 bp region of 13q32. The clustering of these miRNAs suggests that each set may be processed from the same primary transcript[14]. Both miRNAs in the first cluster and five of the genes in the second cluster are conserved in order and orientation on mouse chromosome 14.

Chromosome landscape

Chromosome 13 shows striking features of low gene density compared to the other finished, annotated autosomes, and also a very variable gene distribution. The overall gene density is the lowest of all the sequenced autosomes (see Table 2), with an average (excluding pseudogenes and ncRNA genes) of 6.5 genes per Mb (refs 2–5,15,16). This analysis extends and confirms previous



observations[10,17,18]. Consistent with the low gene density, the G+C content of 38.5% and the predicted CpG island density of 5.4 Mb^{-1} are considerably below the genome averages (Table 2). Exon coverage of chromosome 13 sequence is also substantially lower (1.3%) than that of other autosomes, except chromosome 7. However, the average gene length on chromosome 13 is 57 kb, which is almost double that of other chromosomes (31 and 26 kb, for chromosomes 6 and 22, respectively). As a result, the 633 genes cover 37% of the sequence, which is only slightly lower than that of the other finished, annotated autosomes (Table 2).

Gene density varies considerably along the chromosome, as do other characteristics of the gene-rich and gene-poor regions (see

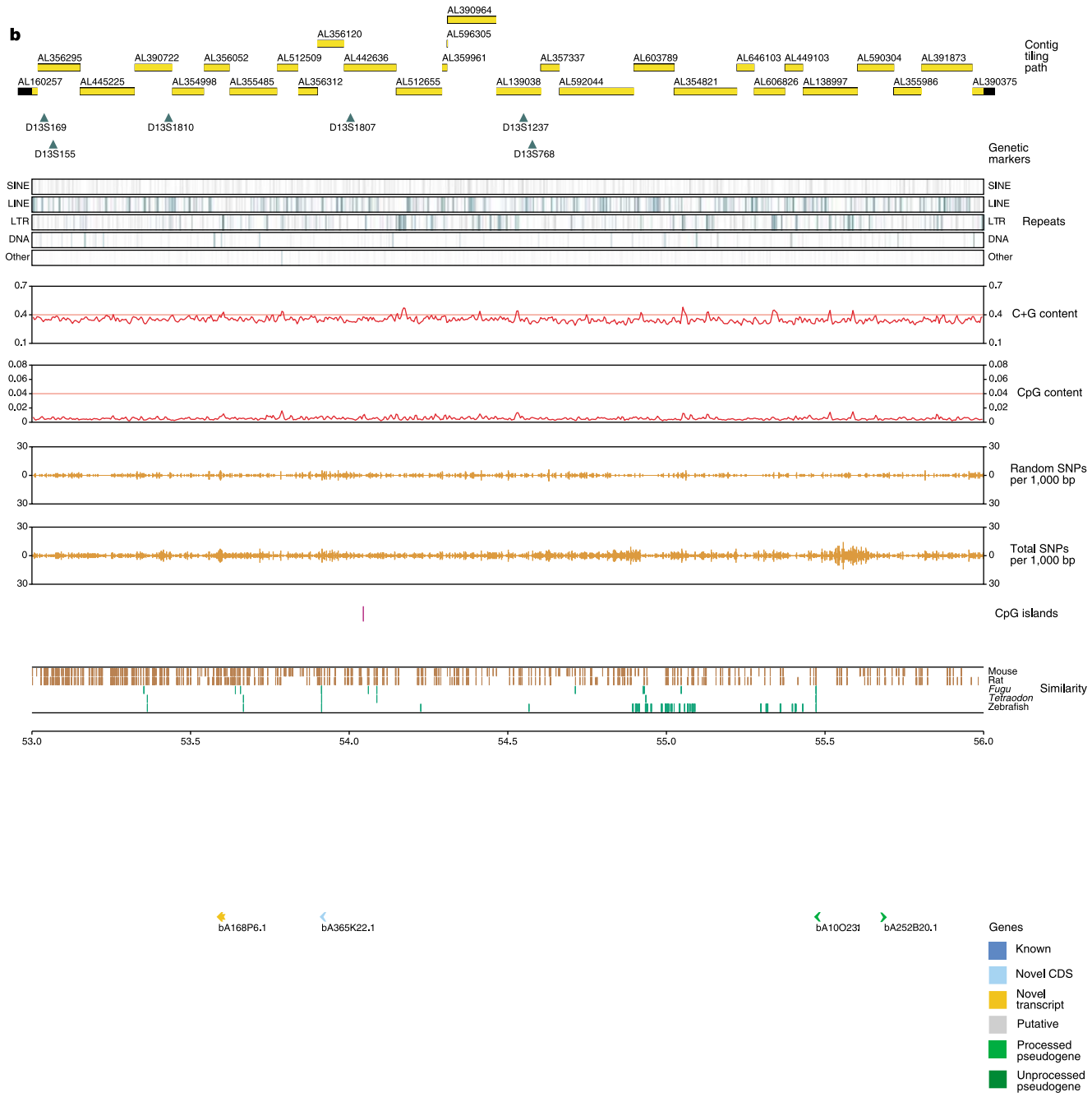


Figure 2 Characteristics of a gene-rich (**a**) and a gene-poor (**b**) region. The overlapping tilepath, labelled by accession number, is shown in yellow. Genetic markers from the deCODE map have been positioned on the sequence. Occurrences of repeats are shown as vertical turquoise bars. G+C content and CpG dinucleotide content are shown in overlapping windows of 8 kb, with adjacent windows overlapping by 4 kb. CpG islands are predicted using a modification of the CpG program developed by G. Micklem (personal communication). The positions of transcription start sites were predicted by the Eponine[49] program. Regions conserved in mouse and rat are shown by orange bars, and those conserved in *Fugu*, *Tetraodon* and zebrafish are shown by green bars. The Rfam track contains the predicted ncRNA genes. Annotated gene structures are shown, subdivided into categories by colour. The direction of transcription is indicated by the arrow. The SNP tracks indicate the number of SNPs per kb. The random SNPs were generated by the whole-genome shotgun approach and were obtained from dbSNP by querying for SNPs produced by The SNP Consortium[50]. The locations of clusters of two or more related genes within 1 Mb of each other are indicated. The scale shows the approximate Mb position along the chromosome. The chromosome view is available in Supplementary Fig. S1.

Fig. 1 and Supplementary Fig. 1). Here we define 'gene-rich' as containing more than 15 genes per Mb, and 'gene-poor' as containing fewer than 5 genes per Mb. A detailed picture of the characteristics of an example of each regional class is shown in Fig. 2. A comparison of the two regions can be seen in Table 3. There is a 37.8-Mb region, from 52.9–90.7 Mb, where the average gene density drops to 3 genes per Mb. This region actually comprises two gene-poor segments (52.9–71.9 Mb and 78.9–89.9 Mb) flanking a section with a gene density of 7 Mb^{-1} (Fig. 1). The first gene-poor section contains a 3-Mb region with no genes (53.9–56.9 Mb).

The two major gene-rich areas are at either end of the q arm, with 90 of the predicted coding genes lying within 3 Mb of either end of the euchromatic region. As observed with other completed human chromosomes, gene-poor regions have a low G+C content, a low SINE coverage and a high LINE coverage, relative to genome averages. These trends are reversed for the gene-rich regions. Figure 1 shows that between positions 90 and 100 Mb, although there are very few genes, a large percentage of the region is covered by transcribed gene structures (exons plus introns). This region contains the largest gene on the chromosome (*GPC5*) as well as two others (*GPC6* and *HS6ST3*) each covering over 500 kb. In sharp contrast to the protein-coding genes, the ncRNA genes are distributed evenly between the gene-rich and gene-poor regions (see Fig. 1).

A total of 96,894 SNPs, from the dbSNP database, were mapped onto the sequence. The coding regions of the annotated genes contain 654 SNPs (1 SNP per 1.6 kb). These can be subdivided into 345 synonymous and 309 non-synonymous cSNPs. To analyse the overall distribution along the chromosome, a subset (38,069) of SNPs identified previously by alignment of random shotgun sequence to the draft sequence were plotted separately (see Fig. 2a, b and Supplementary Fig. 1). From this distribution plot, there is no obvious difference in the variation rate between the gene-rich and gene-poor areas. There is one region between positions 18.0 and 18.4 Mb (Figs 1 and 2a), where the SNP density is substantially higher than the average, reaching one SNP per 0.3 kb (1,329 SNPs in 400 kb). There is a known duplication with chromosome 21 in this region, and it is possible that the apparently high SNP density is due to the presence of paralogous sequence variants, as has been suggested previously[16].

Around 5% of the human genome may be accounted for by segmental duplications, and this may play an important role in genetic disease and genome evolution[19,20]. A study by Cheung and colleagues[19], which classified regions as segmental duplications if they show at least 90% homology over a minimum of 5 kb, suggested that there is approximately 1.8 Mb of duplicated sequence on chromosome 13. This sequence comprises 0.9 Mb of intrachromosomal and 1.2 Mb of interchromosomal duplications[20]. This includes 0.3 Mb of sequence common to both categories. For example, the *TPIP* gene on chromosome 13 has undergone both inter- and intrachromosomal duplications. Guipponi and colleagues[21] described phylogenetic analysis of the *TPTE* gene family, of which *TPIP* is a member, and suggested that all family members originate from a common ancestor since the divergence of human and mouse lineages, because there is only a single *Tpte* gene in mouse. There are seven other genes surrounding the mouse *Tpte* gene on chromosome 8, each of which has a functional homologue on human chromosome 13. This observation and the fact that four of the genes have homology only to chromosome 13 suggest that the human orthologue of the *Tpte* gene lies on chromosome 13. There have been a number of duplication events resulting in one functional copy and four pseudogenes of *TPIP* on chromosome 13. In addition, there is another functional member of the gene family, *TPTE*, on chromosome 21 and a number of pseudogenes on chromosomes 3, 15, 22 and Y.

Comparison to the genetic map

Recombination rates in the human genome are higher on average in females than males, and vary considerably along each chromosome. The sex-average genetic length of 13q is 129.52 cM, equating to an average recombination rate of 1.3 cM Mb^{-1} (slightly higher than the genome average of 1.13 cM Mb^{-1}). Overall female and male rates for chromosome 13 are 1.6 and 1.1 cM Mb^{-1} respectively. Figure 1 illustrates male, female and sex-averaged recombination as a function of chromosome position, correlated with a range of other features of the sequence annotation described above. In the 2 Mb closest to the centromere, the recombination rate in females compared to males is high (4.4 versus 1.7 cM Mb^{-1}) (Supplementary Table S6). Near the telomere, by contrast, the male recombination rate is the higher of the two (4.5 compared to 1.9 cM Mb^{-1} in females).

Recombination rate appears to be correlated with gene density (Fig. 1). The recombination rate in male meiosis is particularly low (0.36 cM Mb^{-1}) in the central 42 Mb portion of the chromosome

Table 2 **Comparison of chromosome 13 features with those of other sequenced autosomes**

Category	Chromosome							
	6	7	13	14	20	21	22	Genome
% G+C content	40.0	41.0	38.5	40.9	44.1	40.9	47.8	41.0
% Repeat content	43.9	45.0	42.3	46.2	42	40.1	41.9	44.8
CpG island density (Mb^{-1})	13.1	9.5	5.4	20.2	11.1	-	16.5	10
% Gene coverage	42.2	36.5	37	43.6	42.4	31.0	50.0	-
% Exon coverage	2.2	1.4	1.3	2.3	2.4	-	5.0	-
Gene density (Mb^{-1})	9.2	7.5	6.5	10	12.2	6.7	16.3	~10
Sequence length (Mb)	166.8	153.8	95.5	87.4	59.2	33.5	33.5	2,862.7

The sequence length for the whole genome was derived from the Ensembl Genome Browser (http://www.ensembl.org/Homo_Sapiens). Other data were taken from the following references: chromosome 6 (ref. 5), chromosome 7 (ref. 4), chromosome 14 (ref. 3), chromosome 20 (ref. 2), chromosome 21 (ref. 16), chromosome 22 (ref. 11), genome (ref. 1).

Table 3 **Comparison of the gene-rich (17.9–21 Mb) and gene-poor (53–56 Mb) regions**

Feature	Gene-rich region	Gene-poor region
Gene density	16.5 genes Mb^{-1}	0.7 genes Mb^{-1}
G+C content	42.1%	34.9%
CpG island density	16.4 Mb^{-1}	0.3 Mb^{-1}
SINE coverage	19.5%	5.7%
LINE coverage	19.6%	25.2%

See Fig. 2a for gene-rich regions and Fig. 2b for gene-poor regions.

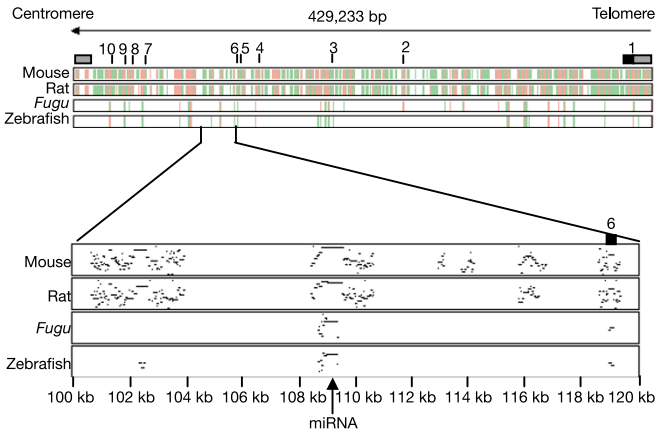


Figure 3 MultiPipMaker analysis[25] of the human *DACH1* gene showing conserved regions in mouse, rat, *Fugu* and zebrafish. **a**, The complete gene, with strongly aligned regions (at least 100 bp without a gap and with at least 70% nucleotide identity) in red. The green lines indicate regions conserved by local alignments. Exons are numbered. **b**, A more detailed picture of the area around exon 6. The arrow indicates a conserved intronic sequence which has 100% homology to a putative miRNA predicted using MiRscan[26].

between D13S1269 and D13S71, which corresponds to the region of lowest gene density on the chromosome. The female rate is 1.1 cM Mb^{-1} over the same region. Within this region there is also a noticeable increase in female recombination rate in the region of higher gene density (at position 72–80 Mb).

From analysis of the draft human genome sequence, Yu[22] predicted two recombination 'deserts' on chromosome 13, which have a sex-average recombination rate less than or equal to 0.3 cM Mb^{-1} for physical distances up to 5 Mb in length, D13S164–D13S1228 (0.12 cM Mb^{-1}) and D13S1301–D13S233 (0.22 cM Mb^{-1}). Re-analysis of these regions using the finished sequence indicates that the former has a sex-average recombination rate of 0.84 cM Mb^{-1}, and so cannot be classified as a recombination desert. The latter, however, was contained within a 3.7-Mb region with a recombination rate of 0.16 cM Mb^{-1}, and so meets the criteria for a recombination desert. In the same study[22], recombination 'jungles' were previously defined as having a sex-average recombination rate of 3 cM Mb^{-1} or more in regions up to 5 Mb, but none were predicted on chromosome 13. In our analysis there is one region of 4.1 Mb that meets the criteria for a jungle of high recombination, D13S1315–D13S1825 (3.2 cM Mb^{-1}).

Comparative analysis

The availability of assembled genome sequences for *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Tetraodon nigroviridis* (*Tetraodon*), *Fugu rubripes* (*Fugu*) and *Danio rerio* (zebrafish), allowed us to identify regions where there is conservation between them and human chromosome 13 (see Supplementary Table S7). In general, greater similarity is expected between human and mouse or rat compared to *Fugu, Tetraodon* and zebrafish. We observed that 96% of known and novel genes on human chromosome 13 have exons conserved in both rodents, whereas only 81% have exons conserved in all three fishes. Most (95%) of the regions conserved in humans and fish correspond to annotated exons. By contrast, only 25% of the regions conserved in both mouse and rat overlap annotated exons, and at least some of the remainder are expected to result from selection for functionally important non-coding regions. It is of interest that regions with the largest coverage of conserved sequence between humans and rodents are found in gene-poor areas of the chromosome (see mouse and rat homology tracks in Fig. 1).

To estimate the completeness of our annotation of exons in protein-coding genes, we adopted the approach of previous studies[2,5]. 2,553 regions on chromosome 13 were found to be conserved in all six reference genomes analysed here. Of these, 2,441 share overlap with 2,337 annotated exons on 13q, while the remaining 112 do not. On this basis, at least 95.4% (2,337/2,337 + 112) of exons on chromosome 13 are included in our annotation.

There are 112 conserved regions that do not correspond to annotated exons. Some or all of them might be exons that remain unconfirmed owing to a lack of transcriptional evidence, or they might be conserved regulatory regions. Recent reports suggest that non-exonic sequences conserved across species may be regulatory or structural elements[23,24]. Nobrega and colleagues showed that seven evolutionarily conserved sequences, from in and around the mouse *Dach* gene, enhanced expression of reporter genes in specific tissues[23]. These regions were also identified in our analysis. In fact, 28 of the 112 conserved regions are located in the introns of the *DACH1* gene. The MultiPipMaker program[25] was used to compare sequence from the *DACH1* gene in human, mouse, rat, *Fugu* and zebrafish (Fig. 3). One of the regions of conservation has a 100% match to a putative miRNA[26]. Further studies are required to discover whether other of the conserved intronic regions have any regulatory effect.

Medical implications

Forty-eight mendelian conditions listed in Online Mendelian Inheritance in Man (OMIM) have been linked to genes on chromosome 13 (Supplementary Table S8). Thirty-five of these genes have been cloned and can be positioned on the sequence. The map and sequence data for chromosome 13 have contributed to the identification of a number of these. Notably, an early phase of the sequencing led to characterization of the breast cancer type-2 gene *BRCA2* (ref. 27). More recently, sequence from the 13q34 region was used in the identification of the *DAOA* locus, which is associated with schizophrenia and bipolar disorder[28,29].

The sequence is being applied to search for genes that are implicated in human disease but are as yet unidentified. B-cell chronic lymphocytic leukaemia (CLL) is one of the most common leukaemias in the western world and approximately 10% of CLL patients have a homozygous deletion in 13q14.3 (ref. 30): the chromosome 13 data have allowed further refinement of the region of minimal deletion[30]. Follicular lymphoma (FL) accounts for approximately 25% of non-Hodgkin's lymphomas in the western world[31]. Using clones from the 13q32–q33 region of the tilepath, a region of recurrent amplification in FL patients has been identified[32]. Asthma and some forms of dermatitis are atopic or immunoglobulin E (IgE)-mediated diseases, which have shown consistent linkage to 13q14 (refs 33, 34). Zhang and colleagues have localized a quantitative trait locus influencing IgE levels and asthma using a dense SNP map of the region[35]. In the same study, an association with severe clinical asthma was found with several alleles of the *PHF11* gene.

In addition to *BRCA2*, a number of other genes involved in tumorigenesis have been identified on chromosome 13, including the retinoblastoma gene *RB1* and the alveolar rhabdomyosarcoma gene *FOXO1A*. As described in Supplementary Table S8, several other such genes have been linked to chromosome 13 but remain to be identified. In the Mitleman database of recurrent chromosome aberrations in cancer (http://cgap.nci.nih.gov/Chromosomes/Mitelman), more than 1,400 cases affect chromosome 13. The availability of a tilepath of sequence clones will allow techniques such as array comparative genomic hybridization to refine the regions of rearrangement in these cases. □

Methods

Using the Genemap '99 radiation hybrid map[10] as a starting point, a landmark mapping approach was used to seed and anchor BAC clone contigs[36]. A total of 863 markers (8.8 markers Mb^{-1}) were used to screen the RPCI-11 BAC library[37], resulting in the identification of 3,354 clones. Using the whole genome fingerprint data from the International Human Genome Mapping Consortium[38], the BACs were built into sequence-ready contigs using a combination of shared *Hin*dIII fingerprint bands and marker content. Gaps in the map were closed by directed screening of BAC, P1-derived artificial chromosome (PAC) and yeast artificial chromosome (YAC) libraries with probes derived from contig ends. Additionally, end-sequence data from the WIBR whole-genomic fosmid library were used in an effort to identify further contig extensions. All the gaps were sized using fluorescent *in situ* hybridization on DNA fibres, apart from gap 1 (Supplementary Table S2), which was sized on metaphase chromosomes.

All the tilepath clones processed at the Sanger Institute were sequenced using a shotgun approach and assembled as previously reported[5]. All finished clones meet or exceed the agreed international finishing standards of 99.99% sequence accuracy.

The finished genomic sequence was analysed using an automatic ENSEMBL[39] pipeline with modifications to aid the manual curation process. The G+C content of each clone sequence was analysed and putative CpG islands marked. CpG islands were predicted using a modification of the CpG program (G. Micklem, personal communication). The identification of interspersed repeats using RepeatMasker; simple repeats using Tandem Repeat Finder[40]; matches to vertebrate complementary DNAs and expressed sequence tags (ESTs) using WU-BLASTN (W. Gish 1996–2002, http://blast.wustl.edu) and EST_GENOME; and *ab initio* gene prediction using FGENESH and GENSCAN, were as described previously[2]. A protein database combining non-redundant data from SwissProt[41] and TrEMBL[41] was searched using WU-BLASTX. ENSEMBL gene predictions, including the EST gene build, were displayed on each clone present in the finished sequence assembly. The predicted gene structures were manually annotated according to the human annotation workshop (HAWK) guidelines (http://www.sanger.ac.uk/HGP/havana/hawk.shtml). The gene categories used were as described on the VEGA website. Known genes are identical to known human cDNAs or protein sequences and should have an entry in Locuslink (http://www.ncbi.nlm.nih.gov/LocusLink). Novel genes have an open reading frame (ORF), are identical to spliced ESTs or have some similarity to other genes or proteins. Novel transcripts are similar to novel genes, but the ORF cannot be determined with confidence. Putative genes are identical to spliced human ESTs, but do not contain an ORF. Processed pseudogenes are non-functional copies of genes that lack

introns. Unprocessed pseudogenes are non-functional copies of genes that contain introns. Where possible, gene symbols were approved by the HUGO Gene Nomenclature Committee[42].

tRNA genes were predicted using tRNAscan-SE v1.23 (ref. 43) in eukaryotic mode with default parameters and a threshold of 20 bits. Other ncRNAs were predicted by searching the chromosome sequence against the Rfam database of RNA families (version 4.1)[12]. Predicted ncRNA genes were compared by WU-BLASTN to reference sequences collected from the EMBL nucleotide sequence database[44], in an effort to discriminate between genes and pseudogenes. True ncRNA genes were operationally defined as BLAST hits with at least 95% sequence identity over 95% of the query length, as previously described[1]. Published human microRNA gene sequences were downloaded from the miRNA Registry v2.0 and mapped to the chromosome sequence using BLASTN.

All markers from the deCODE genetic map were mapped back onto the finished sequence using a combination of electronic PCR[45] and SSAHA[46]. For the comparative analysis with other genomes the following methods were used. The mouse and rat sequence were compared to the human sequence using BLASTZ[47]. The programs axtBest and subsetAxt (W. J. Kent, http://www.soe.ucsc.edu/~kent/src) were used to post-process the resulting matches, as previously described[47], to select the best match and make the alignments relatively specific to exons by using a specific scoring matrix and threshold. The sequences from *Tetraodon, Fugu* and zebrafish were aligned to the chromosome using WU-TBLASTX, using the same scoring matrix, parameters and filtering strategy used in the Exofish procedure[48]. The resources used for these comparisons are shown in the Supplementary Information.

Received 5 December 2003; accepted 27 January 2004; doi:10.1038/nature02379.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
3. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421,** 601–607 (2003).
4. Hillier, L. W. *et al.* The DNA sequence of human chromosome 7. *Nature* **424,** 157–164 (2003).
5. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
6. Kehrer-Sawatzki, H. *et al.* Mapping of members of the low-copy-number repetitive DNA sequence family chAB4 within the p arms of human acrocentric chromosomes: characterization of Robertsonian translocations. *Chromosome Res.* **6,** 429–435 (1998).
7. 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).
8. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
9. 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).
10. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
11. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
12. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31,** 439–441 (2003).
13. Calin, G. A. *et al.* Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA* **99,** 15524–15529 (2002).
14. Lee, Y., Jeon, K., Lee, J. T., Kim, S. & Kim, V. N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.* **21,** 4663–4670 (2002).
15. Collins, J. E. *et al.* Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13,** 27–36 (2003).
16. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
17. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90,** 11929–11933 (1993).
18. Craig, J. M. & Bickmore, W. A. The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7,** 376–382 (1994).
19. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* [online] **4,** R25 (2003).
20. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).
21. Guipponi, M. *et al.* The murine orthologue of the Golgi-localized TPTE protein provides clues to the evolutionary history of the human TPTE gene family. *Hum. Genet.* **109,** 569–575 (2001).
22. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
23. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
24. Glazko, G. V., Koonin, E. V., Rogozin, I. B. & Shabalina, S. A. A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genet.* **19,** 119–124 (2003).
25. Schwartz, S. *et al.* MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31,** 3518–3524 (2003).
26. Lim, L. P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17,** 991–1008 (2003).
27. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–792 (1995).
28. Hattori, E. *et al.* Polymorphisms at the G72/G30 gene locus, on 13q33, are associated with bipolar disorder in two independent pedigree series. *Am. J. Hum. Genet.* **72,** 1131–1140 (2003).
29. Chumakov, I. *et al.* Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc. Natl Acad. Sci. USA* **99,** 13675–13680 (2002).
30. Rowntree, C. *et al.* Deletion analysis of chromosome 13q14.3 and characterisation of an alternative splice form of LEU1 in B cell chronic lymphocytic leukemia. *Leukemia* **16,** 1267–1275 (2002).
31. Armitage, J. O. & Weisenburger, D. D. New approach to classifying non-Hodgkin's lymphomas: clinical features of the major histologic subtypes. Non-Hodgkin's Lymphoma Classification Project. *J. Clin. Oncol.* **16,** 2780–2795 (1998).
32. Neat, M. J. *et al.* Localisation of a novel region of recurrent amplification in follicular lymphoma to an approximately 6.8 Mb region of 13q32–33. *Genes Chromosom. Cancer* **32,** 236–243 (2001).
33. Beyer, K. *et al.* Association and linkage of atopic dermatitis with chromosome 13q12–14 and 5q31–33 markers. *J. Invest. Dermatol.* **115,** 906–908 (2000).
34. Kimura, K. *et al.* Linkage and association of atopic asthma to markers on chromosome 13 in the Japanese population. *Hum. Mol. Genet.* **8,** 1487–1490 (1999).
35. Zhang, Y. *et al.* Positional cloning of a quantitative trait locus on chromosome 13q14 that influences immunoglobulin E levels and asthma. *Nature Genet.* **34,** 181–186 (2003).
36. 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).
37. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11,** 483–496 (2001).
38. International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
39. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
40. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
41. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370 (2003).
42. Wain, H. M., Lush, M., Ducluzeau, F. & Povey, S. Genew: the human gene nomenclature database. *Nucleic Acids Res.* **30,** 169–171 (2002).
43. 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).
44. Stoesser, G. *et al.* The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.* **31,** 17–22 (2003).
45. Schuler, G. D. Sequence mapping by electronic PCR. *Genome Res.* **7,** 541–550 (1997).
46. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
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. Down, T. A. & Hubbard, T. J. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12,** 458–461 (2002).
50. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).

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

Acknowledgements We thank H. M. Wain, R. C. Lovering, E. A. Bruford, M. J. Lush, V. K. Khodiyar, C. C. Talbot and S. Povey from the HUGO Gene Nomenclature Committee for official gene nomenclature, H. Reithman and R. Sudbrak for provision of clones, Y. Chen for assistance with the SNP mapping 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.D. (ad1@sanger.ac.uk).
Accession numbers for the sequences analysed for this paper can be found in Supplementary Fig. S1. All reported DNA sequences have been deposited in EMBL or GenBank.