The DNA sequence of human chromosome 22

I. Dunham, N. Shimizu, B. A. Roe, S. Chissoe *et al.*†

† *A full list of authors appears at the end of this paper*

Knowledge of the complete genomic DNA sequence of an organism allows a systematic approach to defining its genetic components. The genomic sequence provides access to the complete structures of all genes, including those without known function, their control elements, and, by inference, the proteins they encode, as well as all other biologically important sequences. Furthermore, the sequence is a rich and permanent source of information for the design of further biological studies of the organism and for the study of evolution through cross-species sequence comparison. The power of this approach has been amply demonstrated by the determination of the sequences of a number of microbial and model organisms. The next step is to obtain the complete sequence of the entire human genome. Here we report the sequence of the euchromatic part of human chromosome 22. The sequence obtained consists of 12 contiguous segments spanning 33.4 megabases, contains at least 545 genes and 134 pseudogenes, and provides the first view of the complex chromosomal landscapes that will be found in the rest of the genome.

Two alternative approaches have been proposed to determine the human genome sequence. In the clone by clone approach, a map of the genome is constructed using clones of a suitable size (for example, 100–200 kilobases (kb)), and then the sequence is determined for each of a representative set of clones that completely covers the map[1]. Alternatively, a whole genome shotgun[2] requires the sequencing of unmapped genomic clones, typically in a size range of 2–10 kb, followed by a monolithic assembly to produce the entire sequence. Although the merits of these two strategies continue to be debated[3], the public domain human genome sequencing project is following the clone by clone approach[4] because it is modular, allows efficient organization of distributed resources and sequencing capacities, avoids problems arising from distant repeats and results in early completion of significant units of the genome. Here we report the first sequencing landmark of the human genome project, the operationally complete sequence of the euchromatic portion of a human chromosome.

Chromosome 22 is the second smallest of the human autosomes, comprising 1.6–1.8% of the genomic DNA[5]. It is one of five human acrocentric chromosomes, each of which shares substantial sequence similarity in the short arm, which encodes the tandemly repeated ribosomal RNA genes and a series of other tandem repeat sequence arrays. There is no evidence to indicate the presence of any protein coding genes on the short arm of chromosome 22 (22p). In contrast, direct[6] and indirect[7,8] mapping methods suggest that the long arm of the chromosome (22q) is rich in genes compared with other chromosomes. The relatively small size and the existence of a high-resolution framework map of the chromosome[9] suggested to us that sequencing human chromosome 22 would provide an excellent opportunity to show the feasibility of completing the sequence of a substantial unit of the human genome. In addition, alteration of gene dosage on part of 22q is responsible for the aetiology of a number of human congenital anomaly disorders including cat eye syndrome (CES, Mendelian Inheritance in Man (MIM) 115470, http:// www.ncbi.nlm.nih.gov/omim/) and velocardiofacial/DiGeorge syndrome (VCFS, MIM 192430; DGS, MIM 188400). Other regions associated with human disease are the schizophrenia susceptibility locus[10,11], and the sequences involved in spinocerebellar ataxia 10 (SCA10)[12]. Making the sequence of human chromosome 22 freely available to the community early in the data collection phase has benefited studies of disease-related and other genes associated with this human chromosome[13–19].

Genomic sequencing

To identify genomic clones as the substrate for sequencing chromosome 22, extensive clone maps of the chromosome were constructed using cosmids, fosmids, bacterial artificial chromosomes (BACs) and P1-derived artificial chromosomes (PACs). Clones representing parts of chromosome 22 were identified by screening BAC and PAC libraries representing more than 20 genome equivalents using sequence tagged site (STS) markers known to be derived from the chromosome, or by using cosmid and fosmid libraries derived from flow-sorted DNA from chromosome 22. Overlapping clone contigs were assembled on the basis of restriction enzyme fingerprints and STS-content data, and ordered relative to each other using the established framework map of the chromosome[9]. The resulting nascent contigs were extended and joined by iterative cycles of chromosome walking using sequences from the end of each contig. In two places, yeast artificial chromosome (YAC) clones were used to join or extend contigs (AL049708, AL049760). The sequence-ready map covers 22q in 11 clone contigs with 10 gaps and stretches from sequences containing known chromosome 22 centromeric tandem repeats to the 22q telomere[20].

In the final sequence, one additional gap that was intractable to sequencing is found 234 kb from the centromere (see below). The gaps between the clone contigs are located at the two ends of the map, in the 4.3 Mb adjacent to the centromere and in 7.3 Mb at the telomeric end. These regions are separated by a central contig of 23 Mb. We have concluded that the gaps contain sequences that are unclonable with the available host-vector systems, as we were unable to detect clones containing the sequences in these gaps by screening more than 20 genome equivalents of bacterial clones using sequences adjacent to the contig ends.

The size of the seven gaps in the telomeric region has been estimated by DNA fibre fluorescence *in situ* hybridization (FISH). No gap in this region is judged to be larger than ~150 kb. For three of these gaps, a number of BAC and PAC clones that contain STSs on either side of the gap were shown to be deleted for at least a minimal core region by DNA fibre FISH. As these clones come from multiple donor DNA sources, these results are unlikely to be due to deletion in the DNA used to make the libraries. Furthermore, the same result was observed for the gap at 32,600 kb from the centromeric end of the sequence, when the DNA fibre FISH experiments were performed on DNA from two different lymphoblastoid cell lines. One possible explanation for this observation is that DNA fragments containing the gap sequences are initially cloned in the BAC library but clones that delete these sequences have a significant selective advantage as the library is propagated. As the observed size range of the cloned inserts in the BAC libraries ranges from 100 kb to more than 230 kb (http://bacpac.med.buffalo.edu/), such deletion events

are not distinguishable on the basis of size from undeleted BACs. Additional analysis of the distribution of BAC end sequences from dbGSS (http://www.ncbi.nlm.nih.gov/dbGSS/index.html) suggests that the BAC coverage is sparser closer to the gaps and that this analysis did not identify any BACs spanning the gaps. The three remaining clone-map gaps in the proximal region of the long arm are in regions that may contain segments of previously characterized low-copy repeats[21]. These gaps could not be sized by DNA fibre FISH because of the extensive intra- and interchromosomal repeat sequences (see below) but were amenable to long-range restriction mapping. The gap between AP000529 and AP000530 was estimated to be shorter than 150 kb by comparison with a previously established long-range restriction map[22]. The gap closest to the centromere, which is less than 2 kb in size, could not be sequenced despite BAC clone coverage as it was unrepresented in plasmid or M13 libraries, and was intractable to all sequencing strategies applied. Detailed descriptions of several of the clone contigs have been published[21,23,24] or will be published elsewhere.

Each sequencing group took responsibility for completion of adjacent areas of the sequence as illustrated in Fig. 1. A set of minimally overlapping clones (the 'tile path') was chosen from the physical map and sequenced using a combination of a random shotgun assembly, followed by directed sequencing to close gaps and resolve ambiguities ('finishing'). The major problems encountered during completion of the sequence in the directed sequencing phase were CpG islands, tandem repeats and apparent cloning biases. Directed sequencing using oligonucleotide primers, very short insert plasmid libraries, or identification of bridging clones by screening high complexity plasmid or M13 libraries solved these problems.

The completed sequence covers 33.4 Mb of 22q with 11 gaps and has been estimated to be accurate to less than 1 error in 50,000 bases, by internal and external checking exercises[25]. The order and size of each of the contiguous pieces of sequence is detailed in Table 1. The largest contiguous segment stretches over 23 Mb. From our gap-size estimates, we calculate that we have completed 33,464 kb of a total region spanning 34,491 kb and that therefore the sequence is complete to 97% coverage of 22q. The complete sequence and analysis is available on the internet (http://www.sanger.ac.uk/HGP/Chr22 and http://www.genome.ou.edu/Chr22.html).

Sequence analysis and gene content

Analysis of the genomic sequence of the model organisms has made extensive use of predictive computational analysis to identify genes[26–28]. In human DNA, identification of genes by these methods is more difficult because of extensive splicing, lower density of exons and the high proportion of interspersed repetitive sequences. The accuracy of *ab initio* gene prediction on vertebrate genomic sequence has been difficult to determine because of the lack of sequence that has been completely annotated by experiment. To determine the degree of overprediction made by such algorithms, all genes within a region need to be experimentally identified and annotated, however it is virtually impossible to know when this job is complete. A 1.4-Mb region of human genomic sequence around the BRCA2 locus has been subjected to extensive experimental investigation, and it is believed that the 170 exons identified is close to the total number expressed in the region.

The most recent calibration of *ab initio* methods against this region (R.B.S.K. and T.H., manuscript in preparation) shows that with the best methods[29,30] more than 30% of exon predictions do not overlap any experimental exons, in other words, they are over-predictions. Furthermore, having now applied this analysis to larger amounts of data (more than 15 Mb from the Sanger Annotated Genome Sequence Repository which can be obtained as part of the Genesafe collection (http://www.hgmp.mrc.ac.uk/Genesafe/)), it is confirmed that prediction accuracy also varies considerably between different regions of sequence. It was hoped that these calibration efforts would lead to rules for reliable gene prediction based on *ab initio* methods alone, perhaps on the basis of combining several different methods, GC content and so on. However, so far this has not been possible. The same analysis also shows that although 95% of genes are at least partially predicted by *ab initio* methods, few gene structures are completely correct (none in BRCA2) and more than 20% of experimental exons are not predicted at all. The comparison of *ab initio* predictions and the annotated gene structures (see below) in the chromosome 22 sequence is consistent with this, with 94% of annotated genes at least partially detected by a Genscan gene prediction, but only 20% of annotated genes having all exons predicted exactly. Sixteen per cent of all the exons in annotated genes were not predicted at all, although this is only 10% for internal exons (that is, not 5′ and 3′ ends). As a result, we do not consider that *ab initio* gene prediction software can currently be used directly to reliably annotate genes in human sequence, although it is useful when combined with other evidence (see below), for example, to define splice-site boundaries, and as a starting point for experimental studies.

Fortunately, a vast resource of experimental data on human genes in the form of complementary DNA and protein sequences and expressed sequence tags (ESTs) is available which can be used to identify genes within genomic DNA. Furthermore about 60% of human genes have distinctive CpG island sequences at their 5′ ends[31] which can also be used to identify potential genes. Thus, the approach we have taken to annotating genes in the chromosome 22 sequence relies on a combination of similarity searches against all available DNA and protein databases, as well as a series of *ab initio* predictions. Upon completion of the sequence of each clone in the tile path, the sequence was subjected to extensive computational analysis using a suite of similarity searches and prediction tools. Briefly, the sequences were analysed for repetitive sequence content, and the repeats were masked using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html). Masked sequence was compared to public domain DNA and protein databases by similarity searches using the blast family of programs[32]. Unmasked sequence was analysed for C + G content and used to predict the presence of CpG islands, tandem repeat sequences, tRNA genes and exons. The completed analysis was assembled into contigs and visualized using implementations of ACEDB (http:// www.sanger.ac.uk/Software/Acedb/). In addition, the contiguous masked sequence was analysed using gene prediction software[29,30].

Figure 1 The sequence of human chromosome 22. Coloured boxes depict the annotated features of the sequence of human chromosome 22, with the centromere to the left and the telomere to the right. Coordinates are in kilobases. Vertical yellow blocks indicate the positions of the gaps in the sequence and are proportional in size to the estimated size of each gap. From bottom to top the following features are displayed: positions of interspersed repetitive sequences including tandem repeats categorized by nucleotide repeat unit length (at this resolutrion *Alu* repeats are not visibly separated in some regions); the positions of the microsatellite markers in the genetic map of Dib *et al.*[36]; the tiling path of genomic clones used to determine the sequence labelled by their Genbank/EMBL/DDBJ accession number and coloured according to the source of the sequence; and the annotated gene, pseudogene and CpG island content of the sequence. Transcripts and pseudogenes oriented 5′ to 3′ on the DNA strand from centromere to telomere are designated '+', those on the opposite strand '−'. In the transcript rows, the annotated genes are subdivided by colour according to the criteria in the text. Annotated genes with approved gene symbols from the HUGO nomenclature committee are labelled. For details of all the genes with their positions in the reference sequence, see Supplementary Information, Table 1. In the case of the immunoglobulin variable region, the entire locus has been drawn as a single block; in reality, this is a complex of variable chain genes (see ref. 27). At the top is a graphical plot of the repeat density for the common interspersed repeats *Alu* and Line1, and the C + G base frequency across the sequence. Each is calculated as a percentage of the sequence using a sliding 100-kb window moved in 50-kb iterations. Since the production of Fig. 1, six accession codes have been updated. The new codes are AL050347 (for Z73987), AL096754 (for Z68686), AL049749 (for Z82197), Z75892 (for Z75891), AL078611 (for Z79997) and AL023733 (for AL023593).

Gene features were identified by a combination of human inspection and software procedures. Figure 1 shows the 679 gene sequences annotated across 22q. They were grouped according to the evidence that was used to identify them as follows: genes identical to known human gene or protein sequences, referred to as 'known genes' (247); genes homologous, or containing a region of similarity, to gene or protein sequences from human or other species, referred to as 'related genes' (150); sequences homologous to only ESTs, referred to as 'predicted genes' (148); and sequences homologous to a known gene or protein, but with a disrupted open reading frame, referred to 'pseudogenes' (134). (See Supplementary Information, Table 1, for details of these genes.) The *ab initio* gene prediction program, Genscan, predicted 817 genes (6,684 exons) in the contiguous sequence, of which 325 do not form part of the annotated genes categorized above. Given the calibration of *ab initio* prediction methods discussed above, we estimate that of the order of 100 of these will represent parts of 'real' genes for which there is currently no supporting evidence in any sequence database, and that the remainder are likely to be false positives.

The total length of the sequence occupied by the annotated genes, including their introns, is 13.0 Mb (39% of the total sequence). Of this, only 204 kb contain pseudogenes. About 3% of the total sequence is occupied by the exons of these annotated genes. This contrasts sharply with the 41.9% of the sequence that represents tandem and interspersed repeat sequences. There is no significant bias towards genes encoded on one strand at the 5% level ($\chi^2 = 3.83$).

A striking feature of the genes detected is their variety in terms of both identity and structure. There are several gene families that appear to have arisen by tandem duplication. The immunoglobulin λ locus is a well-known example, but there also are other immunoglobulin-related genes on the chromosome outside the immunoglobulin λ region. These include the three genes of the immunoglobulin λ-like (IGLL) family plus a fourth possible member of the family (AC007050.7). There are five clustered immunoglobulin κ variable region pseudogenes in AC006548, and an immunoglobulin variable-related sequence (VpreB3) in AP000348. Much further away from the λ genes is a variable region pseudogene, 123 kb telomeric of IGLL3 in sequence AL008721 (coordinates 9,420–9,530 kb from the centromeric end of the sequence), and a cluster of two λ constant region pseudogenes and a variable region pseudogene in sequences AL008723/AL021937 (coordinates 16,060–16,390 kb from the centromeric end).

Human chromosome 22 also contains other duplicated gene families that encode glutathione *S*-transferases, Ret-finger-like proteins[19], phorbolins or APOBECs, apolipoproteins and β-crystallins. In addition, there are families of genes that are interspersed among other genes and distributed over large chromosomal regions. The γ-glutamyl transferase genes represent a family that appears to have been duplicated in tandem along with other gene families, for instance the BCR-like genes, that span the 22q11 region and together form the well-known LCR22 (low-copy repeat 22) repeats (see below).

The size of individual genes encoded on this chromosome varies over a wide range. The analysis is incomplete as not all 5′ ends have been defined. However, the smallest complete genes are only of the order of 1 kb in length (for example, HMG1L10 is 1.13 kb), whereas the largest single gene (LARGE[15]) stretches over 583 kb. The mean genomic size of the genes is 19.2 kb (median 3.7 kb). Some complete gene structures appear to contain only single exons, whereas the largest number of exons in a gene (PIK4CA) is 54. The mean exon number is 5.4 (median 3). The mean exon size is 266 bp (median 135 bp). The smallest complete exon we have identified is 8 bp in the PITPNB gene. The largest single exon is 7.6 kb in the PKDREJ, which is an intron-less gene with a 6.7-kb open reading frame. In addition, two genes occur within the introns of other expressed genes. The 61-kb TIMP3 gene, which is involved in Sorsby fundus macular degeneration, lies within a 268-kb intron of the large SYN3 gene, and the 8.5 kb HCF2 gene lies within a 27.5-b intron of the PIK4CA gene. In each case, the genes within genes are oriented in the opposite transcriptional orientation to the outer gene. We also observe pseudogenes frequently lying within the introns of other functional genes.

Peptide sequences for the 482 annotated full-length and partial genes with an open reading frame of greater than or equal to 50 amino acids were analysed against the protein family (PFAM)[33], Prosite[34] and SWISS-PROT[35] databases. These data were processed and displayed in an implementation of ACEDB. Overall, 240 (50%) predicted proteins had matching domains in the PFAM database encompassing a total of 164 different PFAM domains. Of the residues making up these 482 proteins, 25% were part of a PFAM domain. This compares with PFAM's residue coverage of SWISS-PROT/TrEMBL, which is more than 45% and indicates that the human genome is enriched in new protein sequences. Sixty-two PFAM domains were found to match more than one protein, including ten predicted proteins containing the eukaryotic protein kinase domain (PF00069), nine matching the Src homology domain 3 (PF00018) and eight matching the RhoGAP domain (PF00620). Fourteen predicted proteins contain zinc-finger domains (See

Table 1 Sequence contigs on chromosome 22

Contig*	Size (kb)	
AP000522–AP000529	234	
gap		**1.9**
AP000530–AP000542	406	
gap		**150**
AP000543–AC006285	1,394	
gap		**150**
AC008101–AC007663	1,790	
gap		**100**
AC007731–AL049708	23,006	
gap		**50**
AL118498–AL022339	767	
gap†		**50–100**
Z85994–AL049811	1,528	
gap		**150**
AL049853–AL096853	2,485	
gap‡		**50**
AL096843–AL078607	190	
gap†		**100**
AL078613–AL117328	993	
gap		**100**
AL080240–AL022328	291	
gap†		**100**
AL096767–AC002055	380	
Total sequence length	33,464	
Total length of 22q	34,491	

*Contigs are indicated by the first and last sequence in the orientation centromere to telomere, and are named by their Genbank/EMBL/DDBJ accession numbers.
†These gaps are spanned by BAC and/or PAC clones with deletions.
‡This gap shows a complex duplication of AL096853 in DNA fibre FISH.

Table 2 The interspersed repeat content of human chromosome 22

Repeat type	Total number	Coverage (bp)	Coverage (%)
Alu	20,188	5,621,998	16.80
HERV	255	160,697	0.48
Line1	8,043	3,256,913	9.73
Line2	6,381	1,273,571	3.81
LTR	848	256,412	0.77
MER	3,757	763,390	2.28
MIR	8,426	1,063,419	3.18
MLT	2,483	605,813	1.81
THE	304	93,159	0.28
Other	2,313	625,562	1.87
Dinucleotide	1,775	133,765	0.40
Trinucleotide	166	18,410	0.06
Quadranucleotide	404	47,691	0.14
Pentanucleotide	16	1,612	0.0048
Other tandem	305	102,245	0.31
Total	55,664	14,024,657	41.91

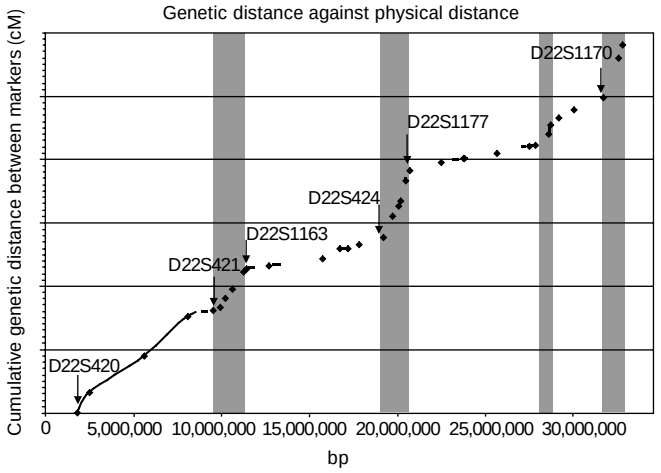

Figure 2 The relationship between physical and genetic distance. The sex-averaged genetic distances of Dib *et al.*[36] were obtained from ftp://ftp.genethon.fr/pub/Gmap/Nature-1995/ and the cumulative intermarker distances for unambiguously ordered markers (in cM) were plotted against the positions of the microsatellite markers in the genomic sequence. It should be stressed that the *y* axis does not represent the true genetic distance between distant markers but the sum of the local intermarker distances. The positions of selected genetic markers are labelled. Grey regions are indicative of areas of relatively increased recombination per unit physical distance.

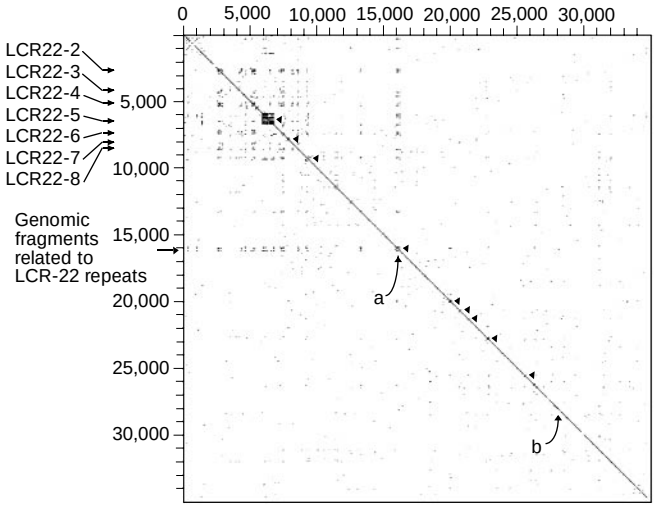

Figure 3 Intrachromosomal repeats on human chromosome 22. High- and medium-copy repeats and low complexity sequence were masked using RepeatMasker and Dust, and masked sequences were compared using Blastn. The results were filtered to identify regions of more than 50% identity to the query sequence, and were plotted in a 2D matrix with a line proportional to the size of the region of identity. Localized gene family repeats are indicated by arrowheads along the diagonal. From the top, these are the immunoglobulin λ locus, the glutathione *S*-transferase genes, the β-crystallin genes, the Ret-finger-protein-like genes, the apolipoprotein genes, the colony-stimulating factor receptor (CSF2RB) inverted partial duplication, the lectins LGALS1 and LGALS2, the APOBEC genes and the CYP2D genes. Two 60-kb regions of more than 90% homology are labelled 'a' (AL008723/AL021937) and 'b' (AL031595/AL022339). Seven low-copy repeat regions (LCR22) and a region containing related genomic fragments are indicated at the left margin.

Supplementary Information, Table 2, for details of the PFAM domains identified in the predicted proteins).

Nineteen per cent of the coding sequences identified were designated as pseudogenes because they had significant similarity to known genes or proteins but had disrupted protein coding reading frames. Because 82% of the pseudogenes contained single blocks of homology and lacked the characteristic intron–exon structure of the putative parent gene, they probably are processed pseudogenes. Of the remaining spliced pseudogenes, most represent segments of duplicated gene families such as the immunoglobulin κ variable genes, the β-crystallins, CYP2D7 and CYP2D8, and the GGT and BCR genes. The pseudogenes are distributed over the entire sequence, interspersed with and sometimes occurring within the introns of annotated expressed genes. However, there also is a dense cluster of 26 pseudogenes in the 1.5-Mb region immediately adjacent to the centromere; the significance of this cluster is currently unclear.

Given that the sequence of 33.4 Mb of chromosome 22q represents 1.1% of the genome and encodes 679 genes, then, if the distribution of genes on the other chromosomes is similar, the minimum number of genes in the entire human genome would be at least 61,000. Previous work has suggested that chromosome 22 is gene rich[6] by a factor of 1.38 (http://www.ncbi.nlm.nih.gov/genemap/page.cgi?F=GeneDistrib.html), which would reduce this estimate to 45,000 genes. It is important, however, to recognize that the analysis described here only provides a minimum estimate for the gene content of chromosome 22q, and that further studies will probably reveal additional coding sequences that could not be identified with the current approaches.

Two lines of evidence point to the existence of additional genes that are not detected in this analysis. First, the 553 predicted CpG islands, which typically lie at the true 5′ ends of about 60% of human genes[31], are in excess of 60% of the number of genes identified (60% = 327, excluding pseudogenes); 282 of the genes identified have CpG islands at or close to the 5′ end (within 5-kb upstream of the first exon, or 1-kb downstream). Thus, there could be up to 271 additional genes associated with CpG islands undetected in the sequence. Second, there are 325 putative genes predicted by the *ab initio* gene prediction program, Genscan, that are not in regions already containing annotated transcripts. We estimate (see above) that roughly 100 of these will represent parts of real genes. Identifying additional genes will require further computational and experimental studies. These studies are continuing and entail testing candidate sequences for possible messenger RNA expression, implementing new gene prediction software able to detect the regions around or near CpG islands that currently have no identified transcript, and further analysis of sequences that are conserved between human and mouse. Furthermore, full-length cDNA sequences that accumulate in the sequence databases of human and other species will be used to refine the gene structures.

The long-range chromosome landscape

Critical to the utility of the genomic sequence to genetic studies is the integration of established genetic maps. The positions of the commonly used microsatellite markers from the Genethon genetic map[36] are given in Fig. 1. The correlation of the order of markers between the genetic map and the sequence is good, within the limitations of genetic mapping. Only a single marker (D22S1175) is discrepant between the two data sets, and this lies in a sequence that is repeated twice on the chromosome (AL021937, see below). In the telomeric region, four of the Genethon markers must lie in our sequence gaps, and we were unable to identify clones from all libraries tested for these. Comparison of genetic distance against physical distance for all the microsatellites whose order is maintained between the datasets shows a mean value of 1.87 cM Mb^{-1}. However, the relationship between genetic and physical distance across the chromosome partitions into two types of region, areas of high and low recombination (Fig. 2). The areas of high recombination may represent recombinational hot spots, although we have not yet been able to identify any specific sequence characteristics common to these areas.

The mean G + C content of the sequence is 47.8%. This is significantly higher than the G + C content calculated for the

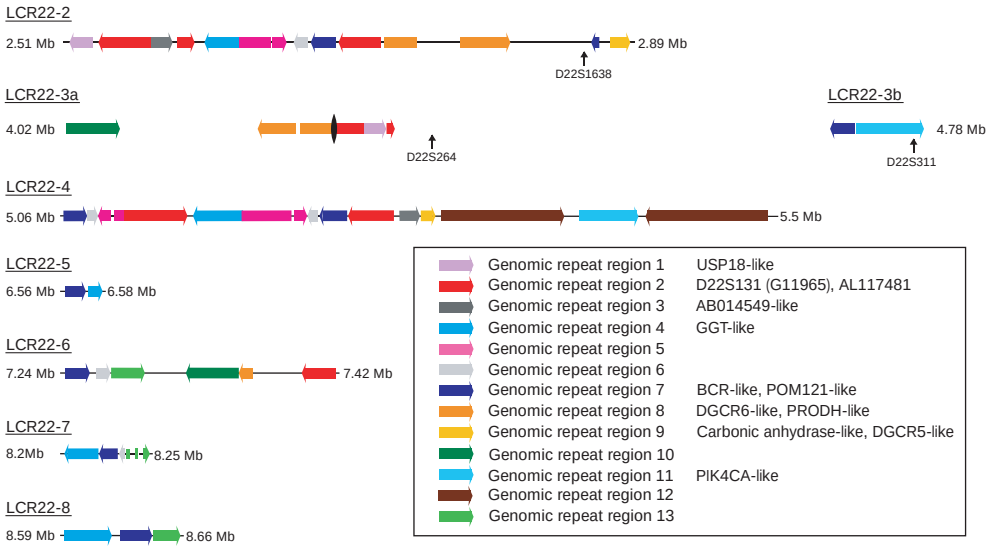

Figure 4 Sequence composition of the LCR22 repeats. Illustration of the sequence composition of seven LCR22 repeats. The span of each LCR22 region is shown in megabases from the centromere. Coloured arrows indicate the extent of one of the thirteen genomic repeat regions and the orientation of the repeat. The known gene and marker content of these genomic repeat regions is indicated in the key. The black oval indicates the position of the gap in the sequence in LCR22-3.

sum of all human genomic sequence determined so far (42%). Although this result was expected from previous indirect measurements of the G + C content of chromosome 22[7,8,37], the distribution is not uniform, but regionally segmented as illustrated in Fig. 1. There are clear fluctuations in the base content, resulting in areas that are relatively G + C rich and others that are relatively G + C poor. On chromosome 22 these regions stretch over several megabases. For example, the 2 Mb of sequence closest to the centromeric end of the sequence is relatively G + C poor, with the G + C content dropping below 40%. Similarly, the area between 16,000 and 18,800 kb from the centromeric end of the sequence is consistently below 45% G + C. The G + C rich regions often reach more than 55% G + C (for example, at 20,100–23,400 kb from the centromeric end of the sequence). This fluctuation appears to be consistent with previous observations that vertebrate genomes are segmented into 'isochores' of distinct G + C content[38] and is similar to the structure seen in the human major histocompatibility complex (MHC) sequence[39]. Isochores correlate with both genes and chromosome structure. The G + C rich isochores are rich in genes and *Alu* repeats, and are located in the G + C rich chromosomal R-bands, whereas the G + C poor isochores are relatively depleted in genes and *Alu* repeats, and are located in the G-bands[8,37,40]. The G + C poor regions of chromosome 22 are depleted in genes and relatively poor in *Alu* sequences. For example, the region between 16,000 and 18,800 kb from the centromeric end contains just three genes, two of which are greater than 400 kb in length. The G + C poor regions also are depleted in CpG islands, which are clustered in the gene-rich, G + C rich regions. Although it is tempting to correlate the sequence features that we see with the chromosome banding patterns, we believe that high-resolution mapping of the chromosome band boundaries will be required to assign definitively these to genomic sequence.

Over 41.9% of the chromosome 22 sequence comprises interspersed and tandem repeat family sequences (Table 2). The density of repeats across the sequence is plotted in Fig. 1. There is variation in the density of *Alu* repeats and some of the regions with low *Alu* density correlate with the G + C poor regions, for example, in the region 16,000–18,800 kb from the centromeric end, and these data support the relationship of isochores with *Alu* distribution. However, in other areas the relationship is less clear. We provide a World-Wide Web interface to the long-range analyses presented here and to further analysis of the many other repeat types and features of the sequence at http://www.sanger.ac.uk/cgi-bin/cwa/22cwa.pl. The 1-Mb region closest to the centromere contains several interesting repeat sequence features that may be typical of other pericentromeric regions. In addition to the density of pseudogenes described above, there is a large 120-kb block of tandemly repeated satellite sequence (D22Z3) centred 500 kb from the centromeric sequence start (not shown in Fig. 1, but evident from the absence of *Alu* and LINE1 sequences at this point). There is also a cluster of satellite II repeats 80-kb telomeric of the D22Z3 sequences. Isolated alphoid satellite repeats are found closer to the centromeric end of the sequence. Furthermore, this pericentromeric 1 Mb closest to the centromere contains many sequences that are shared with a number of different chromosomes, particularly chromosomes 2 and 14. During map construction, 33 out of 37 STSs designed from sequence that was free of high-copy repeats amplified from more than one chromosome in somatic cell hybrid panel analysis.

Low-copy repeats on chromosome 22

To detect intra- and interchromosomal repeats, we compared the entire sequence of chromosome 22 to itself, and also to all other existing human genomic DNA sequence using Blastn[32] after masking high and medium frequency repeats. The results of the intrachromosomal sequence analysis were plotted as a dot matrix (Fig. 3) and reveal a series of interesting features. Locally duplicated gene families lie close to the diagonal axis of the plot. The most striking is the immunoglobulin λ locus that comprises a cluster of 36 potentially functional V-λ gene segments, 56 V-λ pseudogenes, and 27 partial V-λ pseudogenes ('relics'), together with 7 each of the J and C λ segments[24]. Other duplicated gene families that are visible from the dot matrix plot include the clustered genes for glutathione *S*-transferases, β-crystallins, apolipoproteins, phorbolins or APOBECs, the lectins LGALS1 and LGALS2 and the CYP2Ds. A partial inverted duplication of CSF2RB is also observed.

Much more striking are the long-range duplications, which are visible away from the diagonal axis. For example, a 60-kb segment of more than 90% similarity is seen between sequences AL008723/AL021937 (at 16,060–16,390 kb from the centromeric end) and AL031595/AL022339 (at 27,970–28,110 kb from the centromeric end) separated by almost 12 Mb. The 22q11 region is particularly rich in repeated clusters[41]. Previous work described a low-copy repeat family in 22q11 that might mediate recombination events leading to the chromosomal rearrangements seen in cat eye,

velocardiofacial and DiGeorge syndromes[21,42]. The availability of the entire DNA sequence allows detailed dissection of the molecular structure of these low-copy repeats (LCR22s). Edelmann *et al.* described eight LCR22 regions[21,42]. We were unable to find the LCR22 repeat closest to the centromere, but it may lie in the gap at 700 kb from the centromeric end of the sequence. The other LCR22 regions are distributed over 6.5 Mb of 22q11. Analysis of the sequence shows that each LCR22 contains a set of genes or pseudogenes (Fig. 4). For example, five of the LCR22s contain copies of the γ-glutamyl transferase genes and γ-glutamy-transferase-related genes. There is also evidence that a more distant sequence at 16,000 kb from the centromeric start of the genomic sequence shares certain sequences with the LCR22 repeats. This similarity involves related genomic fragments including parts of the Ret-finger-protein-like genes, and the IGLC and IGLV genes.

Regions of conserved synteny with the mouse

The genomic organization of different mammalian species is well known to be conserved[43]. Comparison of genetic and physical maps across species can aid in predicting gene locations in other species, identifying candidate disease genes[13], and revealing various other features relevant to the study of genome organization and evolution. For all the cross-species relationships, that between man and mouse has been most studied. We have examined the relationship of the human chromosome 22 genes to their mouse orthologues.

Of the 160 genes we identified in the human chromosome 22 sequence that have orthologues in mouse, 113 of the murine orthologues have known mouse chromosomal locations (data available at http://www.sanger.ac.uk/HGP/Chr22/Mouse/). Examination of these mouse chromosomal locations mapped onto the human chromosome 22 sequence confirms the conserved linkage groups corresponding to human chromosome 22 on mouse chromosomes 6, 16, 10, 5, 11, 8 and 15[18,44–46] (Fig. 5). Furthermore, these studies allow placement of the sites of evolutionary rearrangements that have disrupted the conservation of synteny more accurately at the DNA sequence scale. For example, the breakdown of synteny between the mouse 8C1 block and the mouse 15E block occurs between the equivalents of the human HMOX1 and MB genes, which are separated by less than 160 kb that also contains a conspicuous 41-copy 18-nucleotide tandem repeat. A clear prediction from these data is that, for the most part, the unmapped murine orthologues of the human genes lie within these established linkage groups, along with the orthologues of the human genes that currently lack mouse counterparts. Exploitation of the chromosome 22 sequence may hasten the determination of the mouse genomic sequence in these regions.

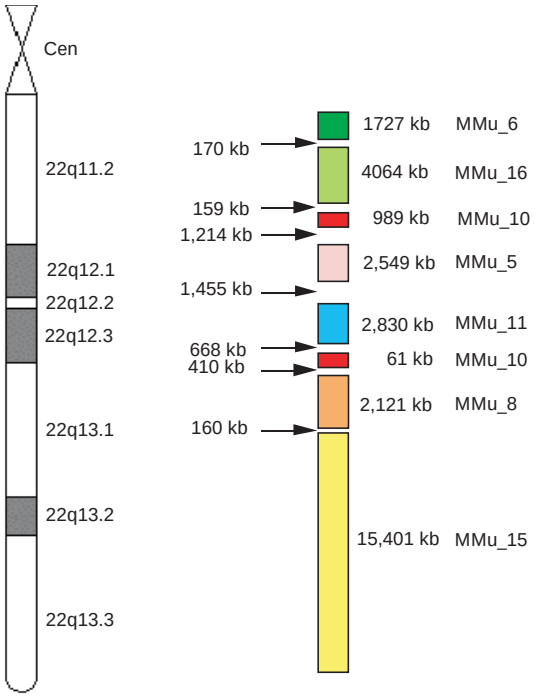

Figure 5 Regions of conserved synteny between human chromosome 22 and the mouse genome. Regions of mouse chromosomes with conserved synteny to human chromosome 22 are shown as adjacent coloured blocks, determined by the mouse map position of mouse orthologues to human chromosome 22 genes. The size of human chromosome 22 corresponding to each mouse chromosomal region is indicated in kb, as well as the size of the gap between the last orthologue in each conserved block. These data are available at http://www.sanger.ac.uk/Chr22/Mouse.

Conclusions

We have shown that the clone by clone strategy is capable of generating long-range continuity sufficient to establish the operationally complete genomic sequence of a chromosome. In doing so, we have generated the largest contiguous segment of DNA sequence to our knowledge to date. The analysis of the sequence gives a foretaste of the information that will be revealed from the remaining chromosomes.

We were unable to obtain sequence over 11 small gaps using the available cloning systems. It may be possible that additional approaches such as using combinations of cloning systems with small insert sizes and low-copy number could reduce the size of these gaps. Direct cloning of restriction fragments that cross these gaps into small insert plasmid or M13 libraries, or direct sequencing approaches might eventually provide access to all the sequence in the gaps. However, closing these gaps is certain to require considerable time and effort, and might be considered as a specialist activity outside the core genome-sequencing efforts. It also is probable that the sequence features responsible for several of these gaps are unlikely to be specific to chromosome 22. In the best case, similar unclonable sequences might be restricted to the centromeric and telomeric regions of the other chromosomes and areas with large tandem repeats, and it will be possible to obtain large contiguous segments for the bulk of the euchromatic genome.

Over the course of the project, the emerging sequence of chromosome 22 has been made available in advance of its final completion through the internet sites of the consortium groups and the public sequence databases[47]. The benefits of this policy can be seen in both the regular requests received from investigators for materials and information that arise as the result of sequence homology searches, and the publications that have used the data[14–19]. The genome project will continue to pursue this data release policy as we move closer to the anticipated completed sequence of humans, mice and other complex genomes[47,48]. □

Methods

The methods for construction of clone maps have been previously described[24,49,50] and can also be found at http://www.sanger.ac.uk/HGP/methods/. Details of sequencing methods and software are available at http://www.sanger.ac.uk/HGP/methods/, http://www.genome.ou.edu/proto.html, http://www-alis.tokyo.jst.go.jp/HGS/team_KU/team.html and in the literature[1,24].

Received 5 November; accepted 11 November 1999.

1. The Sanger Centre & The Genome Sequencing Centre. Toward a complete human genome sequence. *Genome Res.* **8,** 1097–1108 (1998).
2. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).
3. Green, P. Against a whole-genome shotgun. *Genome Res.* **7,** 410–417 (1997).
4. Collins, F. S. *et al.* New goals for the U.S. human genome project: 1998–2003. *Science* **282,** 682–689 (1998).
5. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88,** 7474–7476 (1991).
6. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
7. Craig, J. M. & Bickmore, W. A. The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7,** 376–382 (1994).
8. Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. & Bernardi, G. Identification of the gene-richest bands in human chromosomes. *Gene* **174,** 85–94 (1996).

9. Collins, J. E. *et al.* A high-density YAC contig map of human chromosome 22. *Nature* **377,** 367–379 (1995).
10. Pulver, A. E. *et al.* Psychotic illness in patients diagnosed with velo-cardio-facial syndrome and their relatives. *J. Nerv. Ment. Dis.* **182,** 476–478 (1994).
11. Gill, M. *et al.* A combined analysis of D22S278 marker alleles in affected sib-pairs: support for a susceptibility locus for schizophrenia at chromosome 22q12. Schizophrenia Collaborative Linkage Group (Chromosome 22). *Am. J. Med. Genet.* **67,** 40–45 (1996).
12. Zu, L., Figueroa, K. P., Grewal, R. & Pulst, S. M. Mapping of a new autosomal dominant spinocerebellar ataxia to chromosome 22. *Am. J. Hum. Genet.* **64,** 594–599 (1999).
13. Southard-Smith, E. M. *et al.* Comparative analyses of the dominant megacolon-SOX10 genomic interval in mouse and human. *Mamm. Genome* **10,** 744–749 (1999).
14. Nishino, I., Spinazzola, A. & Hirano, M. Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. *Science* **283,** 689–692 (1999).
15. Peyrard, M. *et al.* The human LARGE gene from 22q12.3-q13.1 is a new, distinct member of the glycosyltransferase gene family. *Proc. natl Acad. Sci. USA* **96,** 598–603 (1999).
16. Kao, H. T. *et al.* A third member of the synapsin gene family. *Proc. Natl Acad. Sci. USA* **95,** 4667–4672 (1998).
17. Mittman, S., Guo, J., Emerick, M. C. & Agnew, W. S. Structure and alternative splicing of the gene encoding alpha1I, a human brain T calcium channel alpha1 subunit. *Neurosci. Lett.* **269,** 121–124 (1999).
18. Seroussi, E. *et al.* TOM1 genes map to human chromosome 22q13.1 and mouse chromosome 8C1 and encode proteins similar to the endosomal proteins HGS and STAM. *Genomics* **57,** 380–388 (1999).
19. Seroussi, E. *et al.* Duplications on human chromosome 22 reveal a novel ret finger protein-like gene family with sense and endogenous antisense transcripts. *Genome Res.* **9,** 803–814 (1999).
20. Ning, Y., Rosenberg, M., Biesecker, L. G. & Ledbetter, D. H. Isolation of the human chromosome 22q telomere and its application to detection of cryptic chromosomal abnormalities. *Hum. Genet.* **97,** 765–769 (1996).
21. Edelmann, L., Pandita, R. K. & Morrow, B. E. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am. J. Hum. Genet.* **64,** 1076–1086 (1999).
22. McDermid, H. E. *et al.* Long-range mapping and construction of a YAC contig within the cat eye syndrome critical region. *Genome Res.* **6,** 1149–1159 (1996).
23. Johnson, A. *et al.* A 1.5-Mb contig within the cat eye syndrome critical region at human chromosome 22q11.2. *Genomics* **57,** 306–309 (1999).
24. Kawasaki, K. *et al.* One-megabase sequence analysis of the human immunoglobulin lambda gene locus. *Genome Res.* **7,** 250–261 (1997).
25. 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).
26. Mewes, H. W. *et al.* Overview of the yeast genome. *Nature* **387,** 7–65 (1997).
27. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* **277,** 1453–1474 (1997).
28. The C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282,** 2012–2018 (1998).
29. Solovyev, V. & Salamov, A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Ismb* **5,** 294–302 (1997).
30. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
31. Cross, S. H. & Bird, A. P. CpG islands and genes. *Curr. Opin. Genet. Dev.* **5,** 309–314 (1995).
32. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
33. Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.* **27,** 260–262 (1999).
34. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acis Res.* **27,** 215–219 (1999).
35. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* **27,** 49–54 (1999).
36. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
37. Holmquist, G. P. Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51,** 17–37 (1992).
38. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228**, 953–958 (1985).
39. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* **401,** 921–923 (1999).
40. Bernardi, G. The isochore organization of the human genome. *Annu. Rev. Genet.* **23,** 637–661 (1989).
41. Collins, J. E., Mungall, A. J., Badcock, K. L., Fay, J. M. & Dunham, I. The organization of the gamma-glutamyl transferase genes and other low copy repeats in human chromosome 22q11. *Genome Res.* **7,** 522–531 (1997).
42. Edelman, L. *et al.* A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum. Mol. Genet.* **8,** 1157–1167 (1999).
43. Eppig, J. T. & Nadeau, J. H. Comparative maps: the mammalian jigsaw puzzle. *Curr. Opin. Genet. Dev.* **5,** 709–716 (1995).
44. Bucan, M. *et al.* Comparative mapping of 9 human chromosome 22q loci in the laboratory mouse. *Hum. Mol. Genet.* **2,** 1245–1252 (1993).
45. Carver, E. A. & Stubbs, L. Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Res.* **7,** 1123–1137 (1997).
46. Puech, A. *et al.* Comparative mapping of the human 22q11 chromosomal region and the orthologous region in mice reveals complex changes in gene organization. *Proc. Natl Acad. Sci. USA* **94,** 14608–14613 (1997).
47. Bentley, D. R. Genomic sequence information should be released immediately and freely in the public domain. *Science* **274,** 533–534 (1996).
48. Guyer, M. Statement on the rapid release of genomic DNA sequence. *Genome Res.* **8,** 413 (1998).
49. Dunham, I., Dewar, K., Kim, U.-J. & Ross, M. T. in *Genome Analysis: A Laboratory Manual Series, Volume 3: Cloning Systems* (eds Birren, B. *et al.*) 1–86 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1999).
50. Asakawa, S. *et al.* Human BAC library: construction and rapid screening. *Gene* **191**, 69–79 (1997).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank S. Povey, J. White and H. Wain for the help with gene nomenclature, and M. Adams for making available the sequence trace files of U62317. We thank M. Elharam, H. Jia, L. Lane, R. Morales-Diaz, F. Najar, P. Pham, R. Rahhal, M. Rao, Y. Tilahun, R. Wayt, H. Wright, E. Nakato, J. L. Schmeits, K. Schooler, J. Wang, M. Asahina, M. Takahashi, H. Harigai, Y. G. Xie, F. Y. Han, S. Swahn, B. Funke, R. K. Pandita, C. Chieffo, D. Michaud and all members of the Sanger Centre past and present for their assistance. This work was supported by grants from the Wellcome Trust, the NIH National Human Genome Research Institute to B.A.R. and to B.S.E., the NSF to B.A.R., the University of Oklahoma, the Fund for the Human Genome Sequencing Project of Japan Science and Technology Corporation, the Fund for the 'Research for the Future' Program from the Japan Society for the Promotion of Science, the UK Medical Research Council, the Medical Research Council of Canada to H.E.M., the Swedish Cancer Foundation, the Swedish Medical Research Council, and a Senior/Established Investigator Award from the Swedish Cancer Foundation to J.P.D.

Correspondence and requests for materials should be addressed to I.D. (e-mail: id1@sanger.ac.uk).

Contributors: I. Dunham[1], A. R. Hunt[1], J. E. Collins[1], R. Bruskiewich[1], D. M. Beare[1], M. Clamp[1], L. J. Smink[1], R. Ainscough[1], J. P. Almeida[1], A. Babbage[1], C. Bagguley[1], J. Bailey[1], K. Barlow[1], K. N. Bates[1], O. Beasley[1], C. P. Bird[1], S. Blakey[1], A. M. Bridgeman[1]*, D. Buck[1]*, J. Burgess[1]*, W. D. Burrill[1], J. Burton[1], C. Carder[1], N. P. Carter[1], Y. Chen[1], G. Clark[1], S. M. Clegg[1], V. Cobley[1], C. G. Cole[1], R. E. Collier[1], R. E. Connor[1]*, D. Conroy[1]*, N. Corby[1], G. J. Coville[1], A. V. Cox[1], J. Davis[1]*, E. Dawson[1], P. D. Dhami[1], C. Dockree[1], S. J. Dodsworth[1]*, R. M. Durbin[1], A. Ellington[1], K. L. Evans[1], J. M. Fey[1], K. Fleming[1], L. French[1], A. A. Garner[1], J. G. R. Gilbert[1], M. E. Goward, D. Grafham[1], M. N. Griffiths[1], C. Hall[1]*, R. Hall[1], G. Hall-Tamlyn[1], R. W. Heathcott[1]*, S. Ho[1]*, S. Holmes[1], S. E. Hunt[1], M. C. Jones[1], J. Kershaw[1], A. Kimberley[1], A. King[1], G. K. Laird[1], C. F. Langford[1], M. A. Leversha[1], C. Lloyd[1], D. M. Lloyd[1], I. D. Martyn[1], M. Mashreghi-Mohammadi[1], L. Matthews[1], O. T. McCann[1], J. McClay[1], S. McLaren[1], A. A. McMurray[1], S. A. Milne[1], B. J. Mortimore[1], C. N. Odell[1], R. Pavitt[1], A. V. Pearce[1], D. Pearson[1], B. J. Phillimore[1], S. H. Phillips[1], R. W. Plumb[1], H. Ramsay[1], Y. Ramsey[1], L. Rogers[1], M. T. Ross[1], C. E. Scott[1], H. K. Sehra[1], C. D. Skuce[1], S. Smalley[1], M. L. Smith[1], C. Soderlund[1], L. Spragon[1], C. A. Steward[1], J. E. Sulston[1], R. M. Swann[1], M. Vaudin[1]*, M. Wall[1], J. M. Wallis[1], M. N. Whiteley[1]*, D. Willey[1], L. Williams[1], S. Williams[1], H. Williamson[1]*, T. E. Wilmer[1], L. Wilming[1], C. L. Wright[1], T. Hubbard[1], D. R. Bentley[1], S. Beck[1], J. Rogers[1], N.Shimizu[2], S. Minoshima[2], K. Kawasaki[2], T. Sasaki[2], S. Asakawa[2], J. Kudoh[2], A. Shintani[2], K. Shibuya[2], Y. Yoshizaki[2], N. Aoki[2], S. Mitsuyama[2], B. A. Roe[3], F. Chen[3], L. Chu[3], J. Crabtree[3], S. Deschamps[3], A. Do[3], T. Do[3], A. Dorman[3], F. Fang[3], Y. Fu[1], P. Hu[3], A. Hua[3], S. Kenton[3], H. Lai[3], H. I. Lao[3], J. Lewis[3], S. Lewis[3], S.-P. Lin[3], P. Loh[3], E. Malaj[3], T. Nguyen[3], H. Pan[3], S. Phan[3], S. Qi[3], Y. Qian[3], L. Ray[3], Q. Ren[3], S. Shaull[3], D. Sloan[3], L. Song[3], Q. Wang[3], Y. Wang[3], Z. Wang[3], J. White[3], D. Willingham[3], H. Wu[3], Z. Yao[3], M. Zhan[3], G. Zhang[3], S. Chissoe[4], J. Murray[4], N. Miller[4], P. Minx[4], R. Fulton[4], D. Johnson[4], G. Bemis[4], D. Bentley[4], H. Bradshaw[4], S. Bourne[4], M. Cordes[4], Z. Du[4], L. Fulton[4], D. Goela[4], T. Graves[4], J. Hawkins[4], K. Hinds[4], K. Kemp[4], P. Latreille[4], D. Layman[4], P. Ozersky[4], T. Rohlfing[4], P. Scheet[4], C. Walker[4], A. Wamsley[4], P. Wohldmann[4], K. Pepin[4], J. Nelson[4], I. Korf[4], J. A. Bedell[4], L. Hillier[4], E. Mardis[4], R. Waterston[4], R. Wilson[4], B. S. Emanuel[5], T. Shaikh[5], H. Kurahashi[5], S. Saitta[5], M. L. Budarf[6], H. E. McDermid[6], A. Johnson[6], A. C. C. Wong[6], B. E. Morrow[7], L. Edelmann[7], U. J. Kim[8], H. Shizuya[8], M. I. Simon[8], J. P. Dumanski[9], M. Peyrard[9], D. Kedra[9], E. Seroussi[9], I. Fransson[9], I. Tapia[9], C. E. Bruder[9], K. P. O'Brien[9].

Addresses: 1, The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK; 2, Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan; 3, Department of Chemistry and Biochemistry, The University of Oklahoma, 620 Parrington Oval, Room 311, Norman, Oklahoma 73019, USA; 4, Genome Sequencing Center, Washington University School of Medicine, 4444 Forest Park Blvd, St. Louis, Missouri 63108, USA 5, Division of Human Genetics and Molecular Biology, The Children's Hospital of Philadelphia and the Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA; 6, Department of Biological Sciences, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada; 7, Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA; 8, California Institute of Technology, Division of Biology, Pasadena, California 91125, USA; 9, Department of Molecular Medicine, Clinical Genetics Unit, Karolinska Hospital, CMM bldg. L8:00, 17176 Stockholm, Sweden

** Present addresses: Division of Virology, Department of Pathology, Tennis Court Road, Cambridge CB12 1QP, UK (A. M. Bridgeman); Lark Technologies, Radwinter Road, Saffron Walden Essex, CB11 3HY, UK (D. Buck, R. E. Connor, S. Ho); Australian Genome Research Facility, Gehrmann Laboratories, University of Queensland, St Lucia QLD 4072, Australia (J. Burgess, J. Davis); Incyte Europe Ltd, 214 Cambridge Science Park, Cambridge CB4 0WA, UK (D. Conroy); Tepnel Life Sciences, Innovation Centre, Scotscroft Building, Wilmslow Road, Didsbury, Manchester M20 8RY, UK (S. J. Dodsworth); Department of Brassica & Oilseeds Research, Cambridge Laboratory, John Innes Centre, Norwich, Norfolk, UK (C. Hall); Cancer Genetics Lab, Department of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand (R. W. Heathcott); Dept. of Zoology, University of Cambridge, Downing Street, CB2 3EJ, UK (M. N. Whiteley); Monsanto, Genomic Team, Mail Zone N3SA, 800 North Lindbergh Boulevard, St. Louis, Missouri 63167, USA (M. Vaudin); Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK (H. Williamson)*