ARTICLES

The DNA sequence and biological annotation of human chromosome 1

S. G. Gregory[1,2], K. F. Barlow[1], K. E. McLay[1], R. Kaul[3], D. Swarbreck[1], A. Dunham[1], C. E. Scott[1], K. L. Howe[1], K. Woodfine[4], C. C. A. Spencer[5], M. C. Jones[1], C. Gillson[1], S. Searle[1], Y. Zhou[3], F. Kokocinski[1], L. McDonald[1], R. Evans[1], K. Phillips[1], A. Atkinson[1], R. Cooper[1], C. Jones[1], R. E. Hall[1], T. D. Andrews[1], C. Lloyd[1], R. Ainscough[1], J. P. Almeida[1], K. D. Ambrose[1], F. Anderson[1], R. W. Andrew[1], R. I. S. Ashwell[1], K. Aubin[1], A. K. Babbage[1], C. L. Bagguley[1], J. Bailey[1], H. Beasley[1], G. Bethel[1], C. P. Bird[1], S. Bray-Allen[1], J. Y. Brown[1], A. J. Brown[1], D. Buckley[3], J. Burton[1], J. Bye[1], C. Carder[1], J. C. Chapman[1], S. Y. Clark[1], G. Clarke[1], C. Clee[1], V. Cobley[1], R. E. Collier[1], N. Corby[1], G. J. Coville[1], J. Davies[1], R. Deadman[1], M. Dunn[1], M. Earthrowl[1], A. G. Ellington[1], H. Errington[1], A. Frankish[1], J. Frankland[1], L. French[1], P. Garner[1], J. Garnett[1], L. Gay[1], M. R. J. Ghori[1], R. Gibson[1], L. M. Gilby[1], W. Gillett[3], R. J. Glithero[1], D. V. Grafham[1], C. Griffiths[1], S. Griffiths-Jones[1], R. Grocock[1], S. Hammond[1], E. S. I. Harrison[1], E. Hart[1], E. Haugen[3], P. D. Heath[1], S. Holmes[1], K. Holt[1], P. J. Howden[1], A. R. Hunt[1], S. E. Hunt[1], G. Hunter[1], J. Isherwood[1], R. James[3], C. Johnson[1], D. Johnson[1], A. Joy[1], M. Kay[1], J. K. Kershaw[1], M. Kibukawa[3], A. M. Kimberley[1], A. King[1], A. J. Knights[1], H. Lad[1], G. Laird[1], S. Lawlor[1], D. A. Leongamornlert[1], D. M. Lloyd[1], J. Loveland[1], J. Lovell[1], M. J. Lush[6], R. Lyne[1], S. Martin[1], M. Mashreghi-Mohammadi[1], L. Matthews[1], N. S. W. Matthews[1], S. McLaren[1], S. Milne[1], S. Mistry[1], M. J. F. Moore[1], T. Nickerson[1], C. N. O'Dell[1], K. Oliver[1], A. Palmeiri[3], S. A. Palmer[1], A. Parker[1], D. Patel[1], A. V. Pearce[1], A. I. Peck[1], S. Pelan[1], K. Phelps[3], B. J. Phillimore[1], R. Plumb[1], J. Rajan[1], C. Raymond[3], G. Rouse[3], C. Saenphimmachak[3], H. K. Sehra[1], E. Sheridan[1], R. Shownkeen[1], S. Sims[1], C. D. Skuce[1], M. Smith[1], C. Steward[1], S. Subramanian[3], N. Sycamore[1], A. Tracey[1], A. Tromans[1], Z. Van Helmond[1], M. Wall[1], J. M. Wallis[1], S. White[1], S. L. Whitehead[1], J. E. Wilkinson[1], D. L. Willey[1], H. Williams[1], L. Wilming[1], P. W. Wray[1], Z. Wu[3], A. Coulson[1], M. Vaudin[1], J. E. Sulston[1], R. Durbin[1], T. Hubbard[1], R. Wooster[1], I. Dunham[1], N. P. Carter[1], G. McVean[4], M. T. Ross[1], J. Harrow[1], M. V. Olson[3], S. Beck[1], J. Rogers[1] & D. R. Bentley[1,7]

The reference sequence for each human chromosome provides the framework for understanding genome function, variation and evolution. Here we report the finished sequence and biological annotation of human chromosome 1. Chromosome 1 is gene-dense, with 3,141 genes and 991 pseudogenes, and many coding sequences overlap. Rearrangements and mutations of chromosome 1 are prevalent in cancer and many other diseases. Patterns of sequence variation reveal signals of recent selection in specific genes that may contribute to human fitness, and also in regions where no function is evident. Fine-scale recombination occurs in hotspots of varying intensity along the sequence, and is enriched near genes. These and other studies of human biology and disease encoded within chromosome 1 are made possible with the highly accurate annotated sequence, as part of the completed set of chromosome sequences that comprise the reference human genome.

The sequence of each human chromosome underpins an extremely broad range of biological, genetic and medical studies. Sequence annotation—the process of gathering all of the available information and relating it to the sequence assembly—is essential to develop our understanding of the information stored in human DNA. Initially, there was a strong focus on annotating genes that allowed us to define the genetic information that determines biochemical function and to characterize the functional consequence of genetic aberrations. More recently, we have undertaken systematic identification and annotation of single nucleotide polymorphisms (SNPs) on genomic sequence. This has enabled us to measure the genetic diversity of the genome in geographically distinct population groups, to estimate recombination at a new high-level of resolution, and to identify signals of selection that may reveal new functions encoded in the genome. In parallel, reagents provided by chromosome mapping and sequencing have provided the basis for acquiring additional experimental data: for example, on gene expression and replication timing. These data sets may be used to elucidate the mechanisms that are

[1]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. [2]The Duke University Center for Human Genetics, Durham, North Carolina 27708, USA. [3]Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine, Seattle, Washington 98195, USA. [4]King's College London, Department of Medical and Molecular Genetics, Guy's Tower, London SE1 9RT, UK. [5]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [6]HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, London NW1 2HE, UK. [7]Solexa Ltd, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK.

used by the cell to regulate the use of chromosomal sequences—at the level of transcription, epigenetic modification or gross chromosomal behaviour—during replication and cell division.

Chromosome 1 is the largest of the human chromosomes, containing approximately 8% of all human genetic information. Because of its size, we can expect it to be more representative of the human genome than some other chromosomes with respect to genomic landscape and genetic properties. It is medically important: over 350 human diseases are associated with disruptions in the sequence of this chromosome—including cancers, neurological and developmental disorders, and mendelian conditions—for which many of the corresponding genes are unknown. There are also important biological implications of the size of chromosome 1: it is approximately six times longer than the smallest human chromosomes (21, 22 and Y), which raises the question of how all human genetic information is replicated in a coordinated manner before each cell division. This study reports the finished sequence of human chromosome 1, and provides a detailed annotation of the landscape, gene index and sequence variations of the chromosome. Our annotation also brings together information from a wide range of additional genetic and biological studies to describe features such as profiles of recombination, signals of natural selection and replication timing, and their relation to each other along the chromosome sequence. In turn, we show that this level of annotation reveals clues to the location of functionally important sequences that are currently unknown and merit targeted investigation.

Genomic sequence and landscape

We determined the sequence of a set of 2,220 minimally overlapping clones representing the euchromatic portion of chromosome 1 (Supplementary Table S1). The sequence comprises 223,875,858 base pairs (bp) at >99.99% accuracy[1] (Supplementary Table S2); 120,405,438 bp lie in 14 contigs on the short arm (1p) and 103,470,420 bp lie in 13 contigs on the long arm (1q). The sequence reaches telomeric repetitive motifs $(TTAGGG)_n$ on both chromosome arms and pericentromeric alpha-satellite sequence at the proximal end of the short arm (1pcen). There are 18 megabases (Mb) of heterochromatin on 1q adjacent to the centromere that has not been sequenced.

Twenty-six gaps remain after exhaustive screening of bacterial and yeast-derived clone libraries with a combined coverage of 90 genomic equivalents (Supplementary Table S3). Eight gaps are clustered in 1p36 and eight in 1q21.1 (Fig. 1). These regions are GC-rich and contain low-copy repeats, which we believe contribute to the absence of clones in these regions. Seventeen gaps, measured using fluorescent *in situ* hybridization (FISH) of flanking clones to chromosomal DNA, cover a total of 0.8 Mb (data not shown). By aligning the human contigs to the genome sequences of mouse, rat and chimpanzee, we estimated that the remaining nine gaps total 0.53 Mb (Supplementary Table S2). Therefore, the euchromatic fraction of chromosome 1 is 225.2 Mb, and 99.4% is available as finished sequence.

We assessed sequence integrity and completeness by three separate measures. First, all except one of the 2,580 RefSeq genes assigned to chromosome 1 (release number 7; http://www.ncbi.nlm.nih.gov/RefSeq/)[2] are present in the sequence. The missing gene, *RAB7B*, maps to 1q32 in the GB4 radiation hybrid map (http://www.ncbi.nlm.nih.gov/genemap/)[3] and should lie in gap 23 or 24. Two genes, *IPP* and *PHACTR4,* were only partially represented in the sequence reported here, but have since been completely sequenced (see http://www.sanger.ac.uk/HGP/sequence/). Second, we compared the order of 467 chromosome 1 markers in the finished sequence and in the deCODE genetic map[4] and found no discrepancies. Third, we aligned 32,984 pairs of fosmid clone end sequences to unique positions in the finished sequence and found eleven discordances. Three were sequence misassemblies caused by low-copy repeats, which have been corrected. The remainder are either deletions in the finished sequence or naturally occurring length polymorphisms. For example, the *GSTM1* gene is absent from 50% of individuals. This gene is present in the reference sequence, but was absent from the fosmid clones mapped to the region.

The G + C, repeat and CpG island content of the chromosome (41%, 48% and 8.9 islands per Mb, respectively) match the genome average[5] (Supplementary Tables S4 and S5). Areas of high G + C content (Fig. 1c), gene density (Fig. 1d), light Giemsa-staining (Fig. 1a), and SINE (short interspersed element) and LINE (long interspersed element) repeat density (Supplementary Fig. S1a) all correlate[5]. Chromosome 1 has an overall gene density of 14.2 genes per Mb—almost twice the genome average (7.8 genes per Mb) and is, therefore, one of the most gene-dense chromosomes. The 2 Mb light Giemsa-staining region of 1p36.33, adjacent to the 1p telomere, exemplifies a section of extreme sequence content on the chromosome (58.4% G + C, 98 predicted CpG islands, and 104 genes).

Gene index

We curated all available complementary DNA (cDNA) and protein information that provided evidence for gene features, and annotated a total of 3,141 structures (Supplementary Table S6). These are contained in the Vertebrate Genome Annotation (VEGA) database (http://vega.sanger.ac.uk/Homo_sapiens/index.html)[6]. The gene index includes 1,669 known genes, 332 novel coding sequences, 720 novel transcripts, and 420 putative transcripts (defined as described previously[7]), which cover 49.5% of the sequence. We found that 1,189 genes (39%) share overlaps on opposite strands and 655 loci (21%) share overlapping coding regions on the same strand. We also identified 991 pseudogenes, of which 840 are processed, and determined that CpG islands associated with 56% of the known genes and 40% of the novel coding sequences (see Supplementary Methods for details).

We identified evolutionarily conserved regions (ECRs) by

Figure 1 | The genomic landscape of human chromosome 1. **a**, Chromosome 1 ideogram according to Francke[50], showing the differential Giemsa staining pattern. **b**, Sequence scale in intervals of 1 Mb. Note that the correlation between cytogenetic band positions and physical distance is imprecise, owing to varying levels of condensation of different Giemsa bands. **c**, G + C content (on a scale of 30–70%) of 100-kb sequence windows. **d**, Gene density (the number of genes, excluding pseudogenes, per Mb) in 1-Mb sequence windows. **e**, Replication timing ratio (S/G1) at tile-path resolution (horizontal red line denotes the midpoint of replication). **f**, Log of the probability of gene expression at tile-path resolution (horizontal red line denotes the midpoint of log(expression)). **g**, Positions of copy number polymorphisms (CNPs; 1.4 kb–1 Mb in size). **h**, Positions of selected gene families and genes lying within regions of copy number polymorphism (CNP). **i**, Positions of other selected genes of interest on chromosome 1. **j**, Population differences in SNP allele frequency between the CEU, YRI and JPT + CHB HapMap analysis panels[24] (see the text for population definitions). The differentiation measure (0–240, on the *y* axis) is the log-likelihood ratio test statistic. Triangles indicate the most highly differentiated SNPs (see the text for details), and are colour-coded as follows: black, intergenic; purple, intronic; turquoise, untranslated region; red, non-synonymous coding variant. **k**, Haplotype diversity measured for each of the three HapMap panels separately. The blue traces measure the average SNP heterozygosity in windows of 21 SNPs. The horizontal red lines indicate extended haplotypes (defined as the most extreme 1% in terms of length) associated with derived (that is, recent) mutations. The heights of the haplotype bars are arbitrary. **l**, Recombination profile of chromosome 1. The histogram shows recombination rate (cM per Mb) in 100-kb windows. Bars are coloured according to the number of recombination hotspots. Dark red shading indicates the highest level of recombination (5 hotspots per Mb). Dark blue represents regions of least recombination (1 hotspot per Mb). Vertical grey lines in **c**–**f** and **j**–**l** represent gaps in the euchromatic sequence of the chromosome. The grey bar located between approximately 121 Mb and 141 Mb shows the position of the centromere and the long-arm heterochromatic block.

alignment of the chromosome 1 sequence to the genome sequences of mouse, rat, zebrafish and two pufferfish species (*Tetraodon nigroviridis* and *Takifugu* (*Fugu*) *rubripes*). We found that 10,669 of the 10,971 ECRs conserved in all six genomes overlap with annotated exons, suggesting that exon annotation is at least 97.2% complete (see Supplementary Table S7). The remaining 302 ECRs may represent additional exons without supporting evidence, or highly conserved regulatory or structural elements.

We predicted 459 non-coding RNA (ncRNA) genes or pseudogenes using the Rfam database of structural RNA alignments (http://www.sanger.ac.uk/Software/Rfam/) (Supplementary Table S8). We also identified 22 microRNAs (miRNAs), a class of ncRNAs with a mature length of approximately 22 nucleotides that regulate gene expression by post-transcriptional control, through BLAST analysis of the 1,345 miRNA entries within the miRNA Registry (http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml) (Supplementary Table S8).

Sequence duplications

Duplication gives rise to new genetic material that can subsequently diverge and specify novel functions. We analysed the sequence for all repeats ≥10 kilobases (kb) in length with ≥90% identity, and observed 3.49% intra- and 1.64% inter-chromosomal duplication (Supplementary Fig. S2). A 5-Mb region of 1q21.1 has a complex pattern of intrachromosomal duplication (Fig. 2a). Previously, a bacterial artificial chromosome (BAC) clone derived from 1q21.1 showed FISH signals at 1p36.13 and 1p12, and a broad band of hybridization in 1q21.1 (ref. 8). We found one sequence element at eleven locations in 1q21.1, three locations in 1p36.13, and one location in 1p12. Each copy includes a tandem 1.5-kb repeat array of varying size (≤75 kb) that results in exon duplication within different members of the neuroblastoma breakpoint family (NBPF)[9], so called because one gene (*NBPF1*) was shown to be disrupted by translocation in a neuroblastoma patient. The inter- and intra-genic duplications may foster illegitimate recombination leading to NBPF gene variation[9].

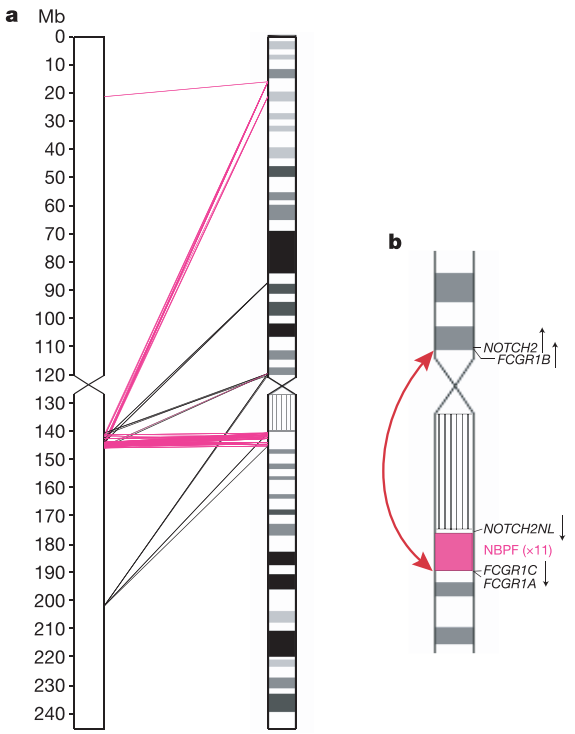

Figure 2 | Segmental duplications involving 1q21.1. a, Intrachromosomal segmental duplications within 1q21.1, and those between 1q21.1 and other parts of chromosome 1. The pink lines indicate duplications that incorporate members of the neuroblastoma breakpoint family (NBPF) genes. Three NBPF genes at 1p13 (ref. 9) are not detected in the segmental duplication analysis but are present in the sequence. **b**, Enlarged section of chromosome 1 encompassing the proximal short and long arm (1p13.3–1q23.1). The hatched box is the heterochromatic block on 1q. The positions of the *NOTCH2* and *NOTCH2NL* genes, and the *FCGR1A*, *B* and *C* genes, are shown, and the arrows indicate the direction of their transcription. Eleven members of the NBPF lie in the region of the pink box. The red arrow indicates a suggested pericentric inversion that occurred in the human lineage following duplication of *NOTCH2* and *FCGR1*.

The NBPF repeat region in human 1q21.1 is flanked by the *NOTCH2NL* gene proximally and the *FCGR1C* and *FCGR1A* genes distally (Fig. 2b). Their homologues—the *NOTCH2* and *FCGR1B* genes—lie together on the other side of the centromere at 1p12, consistent with the occurrence of a pericentric inversion after divergence of the human and chimpanzee lineages[10]. *NOTCH2NL* is a truncated copy of *NOTCH2* spanning the 5′ end of the gene as far as 8 kb into intron 4. It encodes a functional protein that interacts with neutrophil elastase and has been implicated in hereditary neutropenia[11]. The NOTCH2NL protein contains several of the epidermal growth factor (EGF) repeats found in NOTCH2, plus a novel 25-amino-acid carboxy terminus that is required for the interaction with neutrophil elastase[11]. Our analysis shows that this C-terminal region is derived from the splicing of exon 4 to a region in intron 4.

Paralogous gene pairs result from segmental duplications that diverged by accumulation of mutations at either locus. We found 56 clusters of genes that duplicated after the human–murine divergence (see Supplementary Table S9). The ratio of non-synonymous to synonymous nucleotide substitution rates (K_a/K_s) provides a measure of the rate of divergence in each gene pair. We observed the highest K_a/K_s ratio (1.8) between *SPRR2A* and *SPRR2F* (compared with the chromosome average of 0.4). The SPRR genes encode small proline-rich proteins that are primary constituents of the cornified cell envelope—a cross-linked protein scaffold that protects the body from the environment. Many other proteins in this envelope are encoded by genes that cluster with the SPRR genes in 1q21.3, and together constitute the epidermal differentiation complex (EDC). The EDC was recently noted as the most rapidly diverging gene cluster in a comparison between human and chimpanzee[12].

Copy number polymorphisms (CNPs) of up to several hundred kilobases occur within phenotypically normal individuals[13–16]. We positioned CNPs from two CNP databases (http://paralogy.gs.washington.edu/structuralvariation and http://projects.tcag.ca/variation/) in the chromosome 1 sequence (Fig. 1g and Supplementary Table S10) and localized a subset of gene families within these regions of duplication (Fig. 1h). Genes within CNP regions show structural polymorphism that may lead to disease susceptibility. For example, a *GSTM1* polymorphism may confer an increased cancer risk[17,18], and polymorphisms of *FCGR3* have recently been shown to predispose humans to glomerulonephritis[19].

Sequence variation

We mapped 800,653 SNPs from the public databases (dbSNP; release 121) to unique positions in the chromosome 1 sequence. We found 7,917 (1.26%) in protein coding regions, of which 4,471 are non-synonymous and are therefore putative functional variants. We also identified 90 SNPs that introduce premature stop codons in the annotated coding sequence. These mutations would truncate the proteins encoded by 88 genes, 15 of which are associated with genetic diseases and include *COL11A1* in Marshall and Stickler type II syndromes (Online Mendelian Inheritance in Man (OMIM) entry number: 120280), *FY/DARC* in malarial susceptibility (OMIM: 110700), and *UROD* in porphyria cutanea tarda (OMIM: 176100) (see Fig. 1i and Supplementary Table S11). Many of the SNPs on chromosome 1 have been used to determine patterns of genetic variation, providing important new information about molecular

and evolutionary processes that can be annotated along the chromosome sequence (discussed below).

Chromosome recombination

We compared physical and genetic distances between markers from the deCODE genetic map[4] (Supplementary Fig. S3), and observed high rates of recombination near the telomeres. The lowest rates are near the p-arm centromere (male: 0.04 centimorgans (cM) per Mb; female: 0.77 cM per Mb) and the q-arm heterochromatin (male: 0.04 cM per Mb; female: 0.31 cM per Mb). The sex-averaged recombination rate across the chromosome is 1.13 cM per Mb, equalling the genome average. Recombination is higher in females than males (1.43 versus 0.82 cM per Mb), except at 1ptel and 1qtel (Supplementary Fig. S3).

For a more detailed profile, we used data generated by a recent survey in which the genotypes of 60,000 chromosome 1 SNPs were determined in 269 individuals (as part of the HapMap project[20]). Recombination rates along the chromosome were estimated using coalescent methodologies[21], whereby a high level of association between nearby SNP alleles indicates a low level of historic recombination, and *vice versa*. We observed a highly non-random distribution of recombination, with 80% of all recombination occurring in 15% of the sequence, in agreement with previous studies[21]. Peaks of recombination are greater towards the telomeres (Fig. 1l). The majority of recombination occurs in hotspots (discrete segments of <2 kb)[21] of very variable density, but with a trend of higher densities towards the telomeres (vertical red-shaded bars, Fig. 1l). In some areas (for example, at 107.5 Mb and 156.5 Mb), a high overall recombination rate is due to a high density of discrete hotspots, whereas elsewhere (for example, at 11.5 Mb and 151 Mb) it is due to a few extremely active hotspots. Elevated recombination is positively correlated with gene density and G + C content. However, further analysis at a finer scale[22] reveals that the density of recombination hotspots actually peaks near (within 50 kb), but outside, genes and is suppressed within coding regions. This can be accounted for if double-strand breaks in recombination are accompanied by mutagenesis and, therefore, are sometimes deleterious compared to recombination in non-essential flanking DNA. We also identified an interesting relationship between recombination rate (Fig. 1l), expression level (Fig. 1f) and G + C content (Fig. 1c). Contrary to our expectations, we found that higher rates of recombination and G + C content were associated with genes of lower expression (Supplementary Fig. S5).

Natural selection

Natural selection causes altered patterns of genetic variation in populations. Marked differences in the frequency of SNP alleles in one population group relative to another indicate that variants have been selected in one geographically restricted population compared with another. The selected variant is linked with alleles at nearby loci, and evidence for selection may be observed through the existence of extended haplotypes. This effect is influenced by local recombination rate. Therefore, patterns of variation provide a powerful new form of annotation to target searches for sequences that may be important for human fitness. We outline several analyses below.

First, we plotted along chromosome 1 a profile of population-specific differences in SNP allele frequency between populations of Western and Northern European (CEU, Centre d'Etude du Polymorphisme Humain collection (CEPH)/Utah residents from Western and Northern Europe), West African (YRI, Yoruba from Ibadan) or East Asian (JPT + CHB, Japanese from Tokyo + Han Chinese from Beijing) ancestry using information from the HapMap project (Fig. 1j; population samples are described elsewhere[23]). The highest peaks identified 68 SNPs (triangles above peaks in Fig. 1j) that provided evidence for geographically restricted selection (see Fig. 1j and Supplementary Table S12). These SNPs included three non-synonymous variants. The best known of these—the *FY**A mutation in the Duffy gene [OMIM: 110700], which protects against *Plasmodium vivax* malaria—was used as a baseline (log-likelihood ratio test statistic of 150), so that all other 67 SNPs show the same or greater degree of allelic differentiation compared to *FY**A. The two other non-synonymous variants found were in the *ACOT11* gene (a cold-induced thioesterase expressed in adipose tissue and involved in obesity in mice[24]) and *OR9H1P* (an olfactory receptor that may be a pseudogene). Higher-resolution haplotype analysis of *ACOT11* (Fig. 3) suggests near-fixation for the ancestral haplotype within the YRI population (that is, the ancestral type at all three SNPs—Pro-Asp-Met), whereas the JPT + CHB populations have much greater diversity. This observation may be attributable to strong purifying selection in Africa, relaxed constraints in Asia, and a recent selective sweep in Europe. Most notable among the other SNPs were those clustered in or near *NOTCH2*, *ATP1A1* and *SLC35F3*, and single non-coding SNPs included examples in the cholinergic receptor gene *CHRM3*, a gene encoding a helix–loop–helix transcription factor, and several olfactory receptor genes. We also observed marked allelic

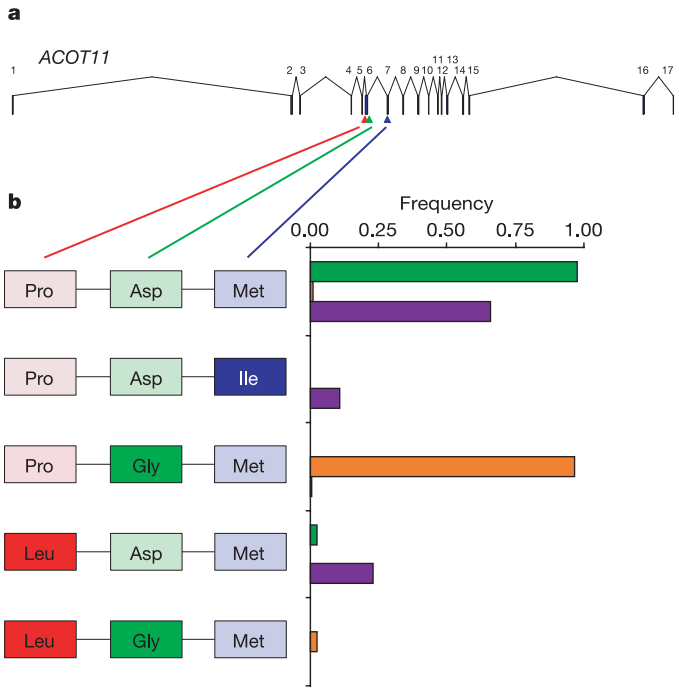

Figure 3 | Population differences in the frequency of non-synonymous haplotypes within the *ACOT11* gene—a cold-induced thioesterase expressed in brown adipose tissue and associated with obesity in mice. **a**, Structure of the *ACOT11* gene, with exons indicated by vertical bars numbered 1–17. Three non-synonymous SNPs, rs2304306, rs1702003 and rs2304305 (red, green and blue arrowheads, respectively), which represent the mutations Pro165Leu, Asp202Gly and Met212Ile, respectively (where the direction of the mutation has been inferred by comparison with the chimpanzee sequences), have been typed across the four HapMap populations. **b**, Different non-synonymous haplotypes in the *ACOT11* gene are shown on the left. Light or dark coloured shading of the boxes containing the amino acids indicate the ancestral or derived mutations at the three sites, respectively. The frequency of each haplotype in each population is shown on the right, colour-coded by population: YRI, green; JPT + CHB, purple; CEU, orange (see the text for population definitions). Note that the population of European origin almost exclusively carries a haplotype with the Asp202Gly mutation—a mutation that is nearly absent from the other populations. In contrast to the situation for most genes, the two Asian populations show the greatest diversity (here grouped together as they have very similar haplotype frequencies), with the African population almost exclusively carrying the ancestral haplotype. These patterns indicate strongly varying selection pressures across the three populations; one possible interpretation being the presence of strong purifying selection in Africa, reduced selection pressures in Asia, and a recent selective sweep in Europe.

differentiation of SNPs associated either with protein coding genes of unknown function, or regions containing no annotated features.

Second, we identified the longest of the extended haplotypes that are associated with the derived (that is, recent) allele for each SNP along the chromosome (horizontal red lines in Fig. 1k)[25]. These features suggest the occurrence of partial selective sweeps around new beneficial mutations. This analysis revealed several strong candidates for recent adaptive evolution, including some with signals in all populations (for example, at position 92 Mb, which coincides with a high-differentiation SNP in a novel gene (RP11-163M2.4), and at position 35 Mb, which contains the claspin (*CLSPN*) and neurochondrin (*NCDN*) genes). We also found regions with signals in two of the three populations—for example, the extended haplotypes at 50 Mb and 170 Mb in the European (CEU; Fig. 1k) and the Asian (JPT + CHB; Fig. 1k) populations, but not in the African (YRI; Fig. 1k) population. In some cases, these regions are also accompanied by a marked drop in diversity (blue traces in Fig. 1k), which arises because the selected haplotype predominates in the region (see regions 17 Mb, 182 Mb and 221 Mb, where sharp troughs all coincide with high-differentiation SNPs). The correlation of low recombination with the length of the extended haplotype is also evident from co-alignment of these tracks with the recombination rate profile (Fig. 1l). For example, the multiple extended haplotypes at 50 Mb and 170 Mb coincide with the lowest estimated recombination rates on the chromosome.

Replication timing

Human DNA replicates in a distinct temporal manner during the S phase of the cell cycle. It is initiated at replication origins of unknown sequence specificity. Replication timing may be influenced by a chromosome's position within the nucleus[26,27], local transcriptional activity[28], or base composition[29–33], DNA methylation or histone modification[34,35]. Previously, we surveyed replication timing in lymphoblastoid cells by comparative genomic hybridization of S- or G1-phase nuclear DNA to a microarray of BAC clones at 1-Mb separation across the genome[33,36]. Chromosome 1 showed the greatest variability in replication timing of any chromosome, indicating that chromosome 1 takes the longest time to replicate. We have performed a higher-resolution study using a microarray of 1,961 overlapping BAC clones (a tile path representing the entire chromosome 1 sequence[37]). The results confirm the substantial variation in replication timing along the chromosome and suggest some correlations between replication timing and features of the chromosome 1 landscape (see Fig. 1). For example, the distal 45 Mb of the chromosome, which has a high gene-density and G + C content, replicates early in S phase whereas the section of 1p from 55–107 Mb, which contains relatively few genes and a low G + C content, is mostly late in replicating. Linear regression analysis for selected sequence features showed a modest correlation between replication timing and G + C content or SINE content of 0.45 and 0.51, respectively (Supplementary Table S13a). We also tested the relationship between replication timing (Fig. 1e) and gene expression. We obtained expression data points for chromosome 1 genes in 647 of the 1,931 arrayed BAC clones (Fig. 1f) and observed a strong correlation between the probability of gene expression and early replication ($r^2 = 0.83$; Supplementary Fig. S4 and Supplementary Table S13b). This correlation is significantly higher than that obtained from the low-resolution whole-genome study[33], and agrees with the results obtained for equivalent high-resolution analyses on chromosomes 22q and 6 (ref. 36). No significant correlation was observed between replication timing and the level at which a gene is transcribed (Supplementary Table S13c). Replication timing therefore correlates with transcriptional activity, and not necessarily the abundance of specific transcripts in a particular region of the chromosome1, in agreement with previous reports in *Drosophila*[38] and other genome wide studies of the human[33,36].

Medical significance

So far, 356 mendelian conditions have been localized to chromosome 1 (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). These include Parkinson's disease (OMIM: 168600), Alzheimer's disease (OMIM: 104300), several types of Charcot–Marie–Tooth disease (OMIM: 118200, 609260, 118210, 607736, 607791, 607677 and 605253), Gaucher disease (the most common lipid-storage disorder; OMIM: 230800), Stargardt disease 1 (the most common form of inherited juvenile macular degeneration; OMIM: 248200) and the Duffy blood group *FY* gene (malaria susceptibility; OMIM: 110700). The chromosome 1 sequence has contributed to finding the genes involved in 35 mendelian disorders (Supplementary Table S14). Genes for a further 53 diseases remain unknown.

Alterations of chromosome 1—in particular, loss of 1p or gains of 1q—are among the most common chromosomal abnormalities in human cancer. Terminal and interstitial deletions of chromosome 1p occur in as many as 1/5,000 to 1/10,000 live births, and are believed to contribute to mental retardation syndromes. Microarrays consisting of large insert bacterial clones from the chromosome 1 map have been used to characterize genomic rearrangements associated with these diseases and identify candidate genes. This approach has been adopted to elucidate the genes associated with neoplasias such as sarcoma[39], meningioma[40] and pheochromocytoma[41], in addition to developmental phenotypes such as the 1p36 deletion syndrome[42].

Concluding remarks

At the turn of this century, the finished sequence of the first human chromosomes[43,44], and then the draft genome sequence[5], provided us with a first view of the landscape of the human genome and a partial annotation of genes. Since then, continued efforts to determine the finished sequence of each chromosome has provided a near-complete reference, accurate base-pair scale on which to place all other genetic information. The explosion of parallel investigations using this resource to characterize features of genomic biology, such as sequence variation, disease-causing mutations, recombination, and replication, provide a new level of information with which to annotate chromosome sequences. Our description of chromosome 1 illustrates the importance of continuing the effort to characterize the complexity of information that is stored in finished sequence. Bringing this information together in one place, and providing convenient views of the data and full access to it, is essential to enable us to increase our understanding of genome biology.

METHODS

Mapping, sequencing and sequence analysis. The hierarchical strategy used for the construction of the sequence-ready physical map of chromosome 1, before clone-by-clone sequencing, as well as the tools of the gene annotation pipeline, are as previously described[6,45]. Manual annotation of gene structures followed the guidelines agreed in the human annotation workshop (HAWK; http://www.sanger.ac.uk/HGP/havana/hawk.shtml), and the approved HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/) gene symbols where possible. Known genes are defined as genes with an RNA entry in RefSeq; novel coding sequences are gene structures with experimental evidence and with an open reading frame (ORF); novel transcripts are gene structures with supporting evidence but no obvious ORF; and, putative transcripts are supported gene structures based on alternative species but have no ORF (ref. 7). Protein translations were analysed with InterProScan (http://www.ebi.ac.uk/InterProScan/), which was run via the Ensembl protein annotation pipeline to obtain Pfam, Prosite, Prints and Profiles domain matches.

Methods used for shotgun sequencing, finishing strategies, comparative analysis and identification of ECRs, sequence annotation predictions and SNP identification are as previously described[46], and are also available in Supplementary Methods.

Alignments for inter- and intra-chromosomal duplications were performed with WU-BLASTN (http://blast.wustl.edu) using the current sequence assembly of chromosome 1 and National Center for Biotechnology Information (NCBI) build 35 for the rest of the genome. All sequences were repeat-masked with RepeatMasker (http://repeatmasker.genome.washington.edu) and low-quality

alignments (E-value $>10^{-30}$; sequence identity <90%; length <80 bp) were discarded. For intrachromosomal duplications, self-matches were discarded. For interchromosomal duplications, the sequence was split into 400-kb segments. Adjacent matches in the same orientation were joined together as described[47]. Only blocks of 10 kb or greater were retained.

SNPs in sequence overlaps were identified using a modification of the SSAHA software[48]. The chromosome 1 SNPs (dbSNP; release 121) were mapped to the sequence assembly of this chromosome (from this study), first with SSAHA and then with Cross-Match.

Replication timing. Replication timing and correlated expression data were generated as previously described[33]. Briefly, cells from a human male cultured lymphoblastoid cell line with a normal (46, XY) karyotype (C0202-JAT; European Collection of Cell Cultures (ECCAC) number 94060845) were stained with Hoechst 33258 and flow-sorted into S- and G1-phase fractions. DNA fractions were labelled with dCTP-Cy5 using a random hexamer labelling kit and spin-column-purified before hybridization to a chromosome 1 array.

The overlapping tile-path array of chromosome 1 was constructed by selecting clones from the minimum tiling path of the chromosome[42] with array hybridization being carried out as described[37]. The arrays were scanned and images quantified using 'Spot' software[49]. Raw-fluorescence ratios were normalized by dividing each ratio by the mean ratio of all clones and scaled by a value representing the median DNA content of the S-phase fraction calculated from the cell-cycle histogram. The median DNA content of the S-phase fraction was calculated from the mean replication timing ratio reported for chromosome 1 on the 1-Mb-resolution array.

Total RNA was extracted from lymphoblastoid cells and first-strand, second-strand cDNA synthesis was performed on 10 μg of total RNA using 100 pmol of a high-performance liquid chromatography (HPLC)-purified T7-$(T)_{24}$ primer. Amplified, biotinylated complementary RNA was then produced with an *in vitro* transcription labelling reaction. Samples with a yield greater than 40 μg of cRNA were subsequently hybridized to Affymetrix U133A oligonucleotide arrays. Hybridization was performed at 45 °C for 16 h. Arrays were washed and stained with streptavidin–phycoerythrin. Signal amplification was performed using a biotinylated anti-streptavidin antibody following the recommended Affymetrix protocol for high-density chips. Scans were carried out on a GeneArray scanner. The fluorescence intensities of scanned arrays were analysed with Affymetrix GeneChip software. The Affymetrix Microarray Suite 5.0 was used for the quantification of gene expression levels. Global scaling was applied to the data to adjust the average recorded intensity to a target intensity of 100. Quantification data was exported from Affymetrix Microarray Suite 5.0 into Microsoft Office Excel for further analysis. Presence or absence of gene expression was determined by a 'present' call in any of the oligonucleotides representing a gene, as determined by Affymetrix Microarray Suite 5.0.

Methods for the determination of genome parameters associated with replication timing analysis are available in Supplementary Methods, and the analysis of segmentation data was carried out as previously described[33].

Received 24 December 2005; accepted 13 March 2006.

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
2. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29,** 137–140 (2001).
3. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
4. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
6. Ashurst, J. L. *et al.* The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33,** D459–D465 (2005).
7. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
8. Weise, A., Starke, H., Mrasek, K., Claussen, U. & Liehr, T. New insights into the evolution of chromosome 1. *Cytogenet. Genome Res.* **108,** 217–222 (2005).
9. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & Van Roy, F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22,** 2265–2274 (2005).
10. Maresco, D. L. *et al.* Localization of *FCGR1* encoding Fcγ receptor class I in primates: molecular evidence for two pericentric inversions during the evolution of human chromosome 1. *Cytogenet. Cell Genet.* **82,** 71–74 (1998).
11. Duan, Z. *et al.* A novel Notch protein, N2N, targeted by neutrophil elastase and implicated in hereditary neutropenia. *Mol. Cell. Biol.* **24,** 58–70 (2004).
12. Chimp Sequencing Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
13. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
14. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36,** 949–951 (2004).
15. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77,** 78–88 (2005).
16. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37,** 727–732 (2005).
17. Strange, R. C. *et al.* The human glutathione *S*-transferases: a case-control study of the incidence of the GST1 0 phenotype in patients with adenocarcinoma. *Carcinogenesis* **12,** 25–28 (1991).
18. van Poppel, G., de Vogel, N., van Balderen, P. J. & Kok, F. J. Increased cytogenetic damage in smokers deficient in glutathione *S*-transferase isozyme μ. *Carcinogenesis* **13,** 303–305 (1992).
19. Aitman, T. J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439,** 851–855 (2006).
20. The International HapMap Consortium, The International HapMap Project. *Nature* **426,** 789–796 (2003).
21. McVean, G. A. T. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584 (2004).
22. Myers, S., Bottolo, L., Freeman, C., McVean, G. & Donnelly, P. Genetics: A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310,** 321–324 (2005).
23. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
24. Adams, S. H. *et al.* BFIT, a unique acyl-CoA thioesterase induced in thermogenic brown adipose tissue: cloning, organization of the human gene and assessment of a potential link to obesity. *Biochem. J.* **360,** 135–142 (2001).
25. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
26. Ferreira, J., Paolella, G., Ramos, C. & Lamond, A. I. Spatial organization of large-scale chromatin domains in the nucleus: a magnified view of single chromosome territories. *J. Cell Biol.* **139,** 1597–1610 (1997).
27. Schermelleh, L., Solovei, I., Zink, D. & Cremer, T. Two-color fluorescence labeling of early and mid-to-late replicating chromatin in living cells. *Chromosome Res.* **9,** 77–80 (2001).
28. Holmquist, G. P. Role of replication time in the control of tissue-specific gene expression. *Am. J. Hum. Genet.* **40,** 151–173 (1987).
29. Hassan, A. B. & Cook, P. R. Does transcription by RNA polymerase play a direct role in the initiation of replication? *J. Cell Sci.* **107,** 1381–1387 (1994).
30. Hassan, A. B., Jackson, D. A., Cook, P. R., Errington, R. J. & White, N. S. Replication and transcription sites are colocalized in human cells. *J. Cell Sci.* **107,** 425–434 (1994).
31. Gilbert, D. M. Making sense of eukaryotic DNA replication origins. *Science* **294,** 96–100 (2001).
32. Gilbert, D. M. Replication timing and metazoan evolution. *Nature Genet.* **32,** 336–337 (2002).
33. Woodfine, K. *et al.* Replication timing of the human genome. *Hum. Mol. Genet.* **13,** 191–202 (2004).
34. Cimbora, D. M. *et al.* Long-distance control of origin choice and replication timing in the human β-globin locus are independent of the locus control region. *Mol. Cell. Biol.* **20,** 5581–5591 (2000).
35. Schübeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18,** 1263–1271 (2004).
36. Woodfine, K. *et al.* Replication timing of human chromosome 6. *Cell Cycle* **4,** 172–176 (2005).
37. Fiegler, H. *et al.* DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosom. Cancer* **36,** 361–374 (2003).
38. Schübeler, D. *et al.* Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nature Genet.* **32,** 438–442 (2002).
39. Kresse, S. H. *et al.* Mapping and characterization of the amplicon near *APOA2* in 1q23 in human sarcomas by FISH and array CGH. *Mol. Cancer* **4,** 39 (2005).
40. Buckley, P. G. *et al.* Comprehensive DNA copy number profiling of meningioma using a chromosome 1 tiling path microarray identifies novel candidate tumor suppressor loci. *Cancer Res.* **65,** 2653–2661 (2005).
41. Jarbo, C. *et al.* Detailed assessment of chromosome 22 aberrations in sporadic pheochromocytoma using array-CGH. *Int. J. Cancer* **118,** 1159–1164 (2005).
42. Redon, R. *et al.* Tiling path resolution mapping of constitutional 1p36 deletions by array-CGH: contiguous gene deletion or "deletion with positional effect" syndrome? *J. Med. Genet.* **42,** 166–171 (2005).
43. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
44. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
45. 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).

46. Mungall, A. J. *et al.* The DNA sequence and analysis of human chromosome 6. *Nature* **425,** 805–811 (2003).
47. Cheung, J. *et al.* Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.* **4,** R25 (2003).
48. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).
49. Jain, A. N. *et al.* Fully automatic quantification of microarray image data. *Genome Res.* **12,** 325–332 (2002).
50. Francke, U. Digitized and differentially shaded human chromosome ideograms for genomic applications. *Cytogenet. Cell Genet.* **65,** 206–218 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank the numerous collaborators who have contributed experimental data to the construction of the physical map and assembly of the finished sequence, the EMBL and Ensembl database teams at the European Bioinformatics Institute, S. Povey, E. A. Bruford, T. A. Eyre, V. K. Khodiyar, R. C. Lovering, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright at the HUGO Gene Nomenclature Committee for assignment of official gene symbols, and T. Furey for data mining from the UCSC database. Work at the Sanger Institute was funded by the Wellcome Trust, work at the University of Washington was funded by the NIH, and work at HUGO was funded by the NIH and the MRC.

Author Information The updated human chromosome 1 sequence can be accessed through GenBank accession number NC_000001. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to S.G. (sgregory@chg.duhs.duke.edu).