LETTERS

The finished DNA sequence of human chromosome 12

Steven E. Scherer[1], Donna M. Muzny[1], Christian J. Buhay[1], Rui Chen[1], Andrew Cree[1], Yan Ding[1], Shannon Dugan-Rocha[1], Rachel Gill[1], Preethi Gunaratne[1], R. Alan Harris[1], Alicia C. Hawes[1], Judith Hernandez[1], Anne V. Hodgson[1], Jennifer Hume[1], Andrew Jackson[1], Ziad Mohid Khan[1], Christie Kovar-Smith[1], Lora R. Lewis[1], Ryan J. Lozado[1], Michael L. Metzker[1], Aleksandar Milosavljevic[1], George R. Miner[1], Kate T. Montgomery[2], Margaret B. Morgan[1], Lynne V. Nazareth[1], Graham Scott[1], Erica Sodergren[1], Xing-Zhi Song[1], David Steffen[1], Ruth C. Lovering[3], David A. Wheeler[1], Kim C. Worley[1], Yi Yuan[1], Zhengdong Zhang[1], Charles Q. Adams[1], M. Ali Ansari-Lari[1], Mulu Ayele[1], Mary J. Brown[1], Guan Chen[1], Zhijian Chen[1], Kerstin P. Clerc-Blankenburg[1], Clay Davis[1], Oliver Delgado[1], Huyen H. Dinh[1], Heather Draper[1], Manuel L. Gonzalez-Garay[1], Paul Havlak[1], Laronda R. Jackson[1], Leni S. Jacob[1], Susan H. Kelly[1], Li Li[2], Zhangwan Li[1], Jing Liu[1], Wen Liu[1], Jing Lu[1], Manjula Maheshwari[1], Bao-Viet Nguyen[1], Geoffrey O. Okwuonu[1], Shiran Pasternak[1], Lesette M. Perez[1], Farah J. H. Plopper[1], Jireh Santibanez[1], Hua Shen[1], Paul E. Tabor[1], Daniel Verduzco[1], Lenee Waldron[1], Qiaoyan Wang[1], Gabrielle A. Williams[1], JingKun Zhang[1], Jianling Zhou[1], Baylor College of Medicine Human Genome Sequencing Center Sequence Production Team*, David Nelson[1], Raju Kucherlapati[2], George Weinstock[1] & Richard A. Gibbs[1]

Human chromosome 12 contains more than 1,400 coding genes[1] and 487 loci that have been directly implicated in human disease[2]. The q arm of chromosome 12 contains one of the largest blocks of linkage disequilibrium found in the human genome[3]. Here we present the finished sequence of human chromosome 12, which has been finished to high quality and spans approximately 132 megabases, representing ~4.5% of the human genome. Alignment of the human chromosome 12 sequence across vertebrates reveals the origin of individual segments in chicken, and a unique history of rearrangement through rodent and primate lineages. The rate of base substitutions in recent evolutionary history shows an overall slowing in hominids compared with primates and rodents.

Among the human chromosome sequencing projects, the chromosome 12 sequencing effort benefited most from an earlier advanced sequence tagged site (STS) physical map, which contained 5,300 large-insert clones and 3,100 markers with an average resolution of 44 kilobases (kb)[4]. After integration with the whole genome fingerprint map[5], a final tiling path of 1,168 large-insert clones was chosen for sequencing using the clone-by-clone shotgun sequencing strategy. Each clone was finished to community standards (http://www.genome.gov/10001812) yielding 130,683,379 base pairs (bp) of nonoverlapping sequence, independently measured as greater than 99.99% accurate[6]. The features and annotations presented here may be viewed as user-specified tracks within the Genboree genome browser (http://www.genboree.org/Hs.chr12).

The finished sequence contains just five euchromatic gaps, estimated to total 380 kb by fibre fluorescence *in situ* hybridization (FISH) (C. Wagner-McPherson, personal communication) and by comparison to primate and rodent genome assemblies (Supplementary Table 1). The data extend to 16 and 60 kb from the telomere terminus repeats at the p and q arms, respectively (H. Riethman, personal communication; http://www.wistar.upenn.edu/Riethman/)[7]. The pericentromeric sequences on the p and q arms contain approximately 425 and 600 kb of tandem alpha-satellite repeats, respectively. The alpha satellites do not demonstrate the higher-order structure indicative of the 'core centromere' on either arm, but previously established markers[8], which are present in BAC clones flanking the centromere, result in a calculated centromere length of 1.395 megabases (Mb) and an overall chromosome length of 132,449,811 bp.

Starting with an automated analysis of human build 33 using the Ensembl pipeline[9], we manually annotated the chromosome 12 finished sequence using evidence from all publicly available protein and complementary DNA databases, as well as spliced expressed sequence tags (ESTs). We identified a further 282 gene structures beyond the automated output, resulting in a total of 1,435 loci. Using annotation standards developed by the Human Annotation Workshop (http://www.sanger.ac.uk/HGP/havana/hawk.shtml), the loci were categorized as 1,294 'known genes', 12 'novel coding sequences (CDS)', 34 'novel transcripts', 2 'putative genes' and 93 'pseudogenes'.

We found 4,427 paralogous pairings to genes on chromosome 12, of which 528 were intrachromosomal. Notably, the density of paralogues correlates well with the density of SINEs (short interspersed elements) and breakpoints observed between the human chromosome and its syntenic regions in avian, rodent and canine genomes (see Fig. 1). Excluding pseudogenes, the average gene density on chromosome 12 is 11.0 genes per Mb, which is relatively typical when compared to the gene densities of other chromosomes. A total of just 11 out of 1,264 RefSeq genes[1] are completely or partially absent from genome build 33, while there are only three partially missing genes (NM_001733, *C1R*; NM_018711, *SVOP*; and NM_020993, *BCL7A*) from build 35.

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [2]Harvard Medical School-Partners Healthcare Center for Genetics and Genomics, Boston, Massachusetts 02115, USA. [3]HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK.
*A list of participants and affiliations appears at the end of the paper.

Approximately 58.3% of chromosome 12 genes expressed alternative transcripts, averaging 2.89 transcripts per gene, but ranging as high as 20 (for *UBC*). The majority of these produced altered protein products (2,923 different proteins from among 3,148 alternative transcripts). There were at least 677 partial transcripts, based primarily on spliced EST data, for which we could not identify the complete coding sequence. Therefore, these estimates represent a lower bound on the total alternative splicing activity on the chromosome.

The density of genes across chromosome 12 varies quite widely

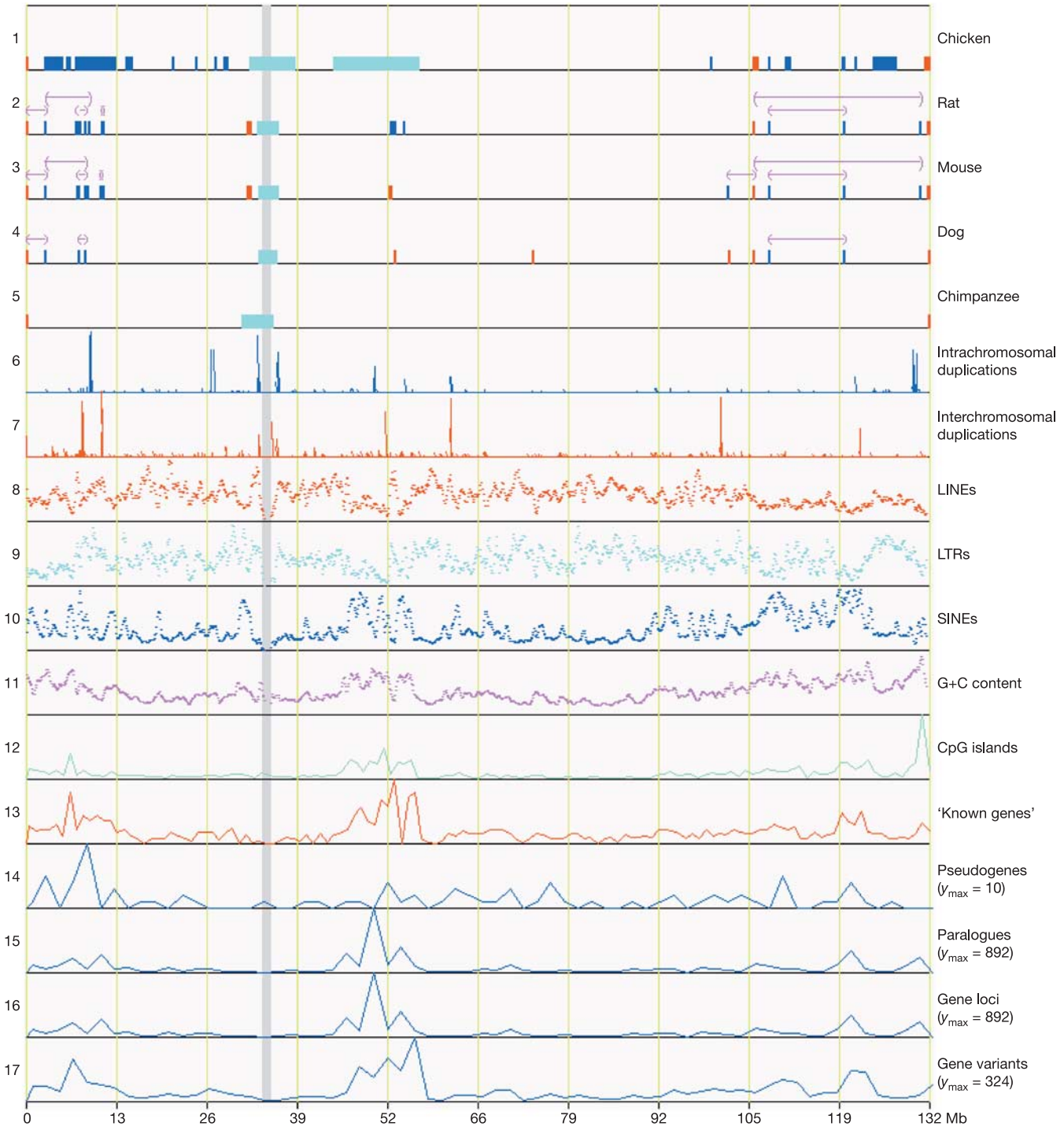

Figure 1 | Correlation of syntenic breakpoints with general chromosome landscape features. Tracks are numbered on the left, and syntenic alignments across human chromosome 12 are shown in the top five tracks: 1, human–chicken; 2, human–rat; 3, human–mouse; 4, human–dog; and 5, human–chimpanzee. The inter- and intrachromosomal breakpoints are represented by red and blue gaps, respectively. Aqua gaps indicate regions without sequence alignment, and the centromere is denoted by the grey bar running through all tracks. Purple brackets portray sequence inversions. The density of recent segmental intra- and interchromosomal duplications from low-copy repeats are shown in tracks 6 and 7, respectively. The incidence of major interspersed (high-copy) repeats is depicted in tracks 8 (LINEs), 9 (LTRs) and 10 (SINEs). The variations in G+C content, and densities of CpG islands, genes and pseudogenes, appear in tracks 11, 12, 13 and 14, respectively, while gene paralogue density, gene density and gene variant density appear in tracks 15, 16 and 17, respectively. Gene density in track 13 is from UCSC 'known genes', and in track 16 is from the nonredundant locus annotations performed in this study. The y_{max} values in tracks 14–17 reflect the maximum *y*-axis values obtained for those tracks.

(see Fig. 1) and there are three large gene clusters: the natural killer cell gene cluster (9 genes) at 12p13.2–12p12.3, the type II keratin gene cluster (14 genes) at 12q13.13, and the homeobox C gene cluster (9 genes) at 12q13. There are also smaller clusters encoding the voltage-gated potassium channels (3 genes) at 12p13, the aquaporin gene cluster (3 genes) at 12q13.1, and a large cluster encoding salivary proline-rich proteins (7 genes) at 12p13.2.

There are 993 segmental duplications (defined as having greater than 90% identity and being >1 kb), which accounts for 2.66% of the chromosome (versus 5.37% for the entire genome), with particular activity in the pericentromeric region of the p arm, and the telomeres. These duplicated regions represent the fraction of the genomic sequence that was most improved in the finishing process, and are a rich resource for the study of gene clusters and large-scale human DNA polymorphism. The chromosome is typical of the remainder of the genome with respect to noncoding RNA (Supplementary Table 2), LINE (long interspersed elements) and SINE distribution, CpG island distribution, and overall G+C content.

A comparison of physical and genetic maps revealed wide variation in recombination activity, with a slightly higher overall recombination rate in females as expected, and an average of 1.3 centimorgans (cM) per Mb (see Supplementary Fig. 1). There are, however, no extensive recombination 'deserts' or 'jungles' (with 'jungles' defined as having a recombination rate of >3 cM per Mb) as previously described[10].

One exceptional region of 12q showed a very large block of linkage disequilibrium spanning 987 kb (compared with a chromosomal average of 26 kb) in all three continental populations assayed. This is one of the largest structures of its kind in the genome, and is associated with evidence for recent positive selection of a pre-expansion *ATXN2* gene CAG repeat allele in Americans of European ancestry[3].

The history of individual human chromosomes can be reconstructed by tracing blocks of conserved synteny across species, and there is sufficient conservation in the chicken genome (>300 million years (Myr) of evolutionary separation from humans) to identify 13 segments representing about 72% of human chromosome 12. Reconstruction of a more recent ancestral mammalian genome shows a double reciprocal rearrangement between two ancient acrocentric chromosomes that were comprised primarily of material syntenic to chromosomes 12 and 22, leading to the current structure (Fig. 2)[11]. The rearrangement occurred in the anthropoid line sometime after the divergence of prosimians. While the chromosome has not undergone any major subsequent rearrangements, pericentric inversions have occurred independently in the chimpanzee and gorilla orthologues[12,13].

Alignment to rodent chromosomes (>80 Myr evolutionary separation) shows that approximately 9 major rearrangements have occurred between human chromosome 12 and the hypothetical common ancestor with rat and mouse. The breakpoints identified

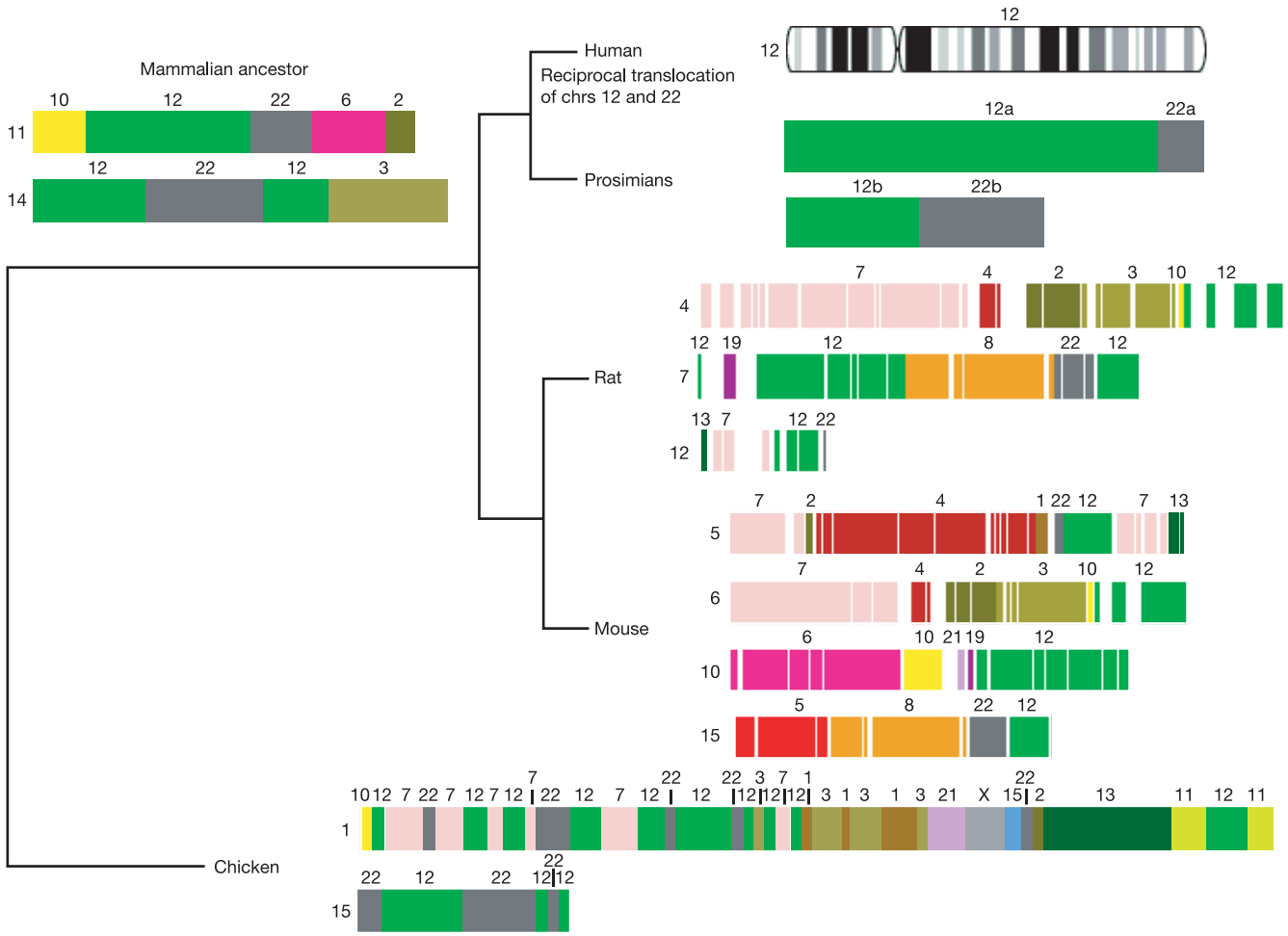

Figure 2 | Human chromosome 12 evolution. Numbers above the chromosome indicate the human orthologous chromosome, and numbers to the left of the chromosome refer to the chromosome of the compared organism. Regions orthologous to human chromosome 12 are shown in green. The rodent comparisons to human were performed using Pash[28], and the ancestral mammal and chicken are adapted from GRIMM (Genome Rearrangements In Man and Mouse)-Synteny computations presented in ref. 30. On the basis of cytogenetic evidence, a double reciprocal translocation of human chromosomes 12 and 22 occurred after the divergence of prosimian primates, and is present in all anthropoid primates.

Table 1 | Human- and chimpanzee-specific base-pair substitution rates

	Intergenic	Intronic	UTR	CDS	Substitutions per year* (non-CDS)
Human	0.0060	0.0056	0.0043	0.0024	9.70×10^{-10}
Chimpanzee	0.0074	0.0069	0.0069	0.0034	1.19×10^{-9}

Base-pair substitution rates were calculated per site within the genomic regions indicated.
*Substitutions per year was calculated assuming a human–chimpanzee divergence at 6 Myr ago.

between the human and rodent genomes are summarized in Fig. 1. An unexpected feature is the pattern of increased activity at the ends of the chromosome, with a substantial portion of the proximal q arm relatively unaffected by gross rearrangement across mammals. The detailed examination of breakpoints between the available genome sequences does not suggest any obvious sequence features that correlate with the evolutionary change, and the pattern of breaks otherwise generally conforms to that predicted from the known evolutionary relationships. There is an observed increase in the number of intra- and interchromosomal duplications in the regions where breaks occur, but a full analysis awaits a comprehensive description of duplication in the nonhuman species, which are all currently at draft coverage.

There are innumerable 'small-scale' differences between regions of the rodent and human chromosomes with conserved synteny. The data presented here confirm earlier studies[14] showing that the rate of change of single base differences and small insertions was greater in the rodent than in hominid lineages, while the deletion rates were slightly lower. Several studies have attempted to correlate the frequency and distribution of these small-scale changes with other genome features. For example, the relative increase in insertions correlates with the active expansion of the overall size of human chromosomes, and a genome-wide burst of SINE insertions (mainly AluJ and AluS repeats) in the human lineage. Furthermore, the density of SINEs shows a modest positive correlation with the microinsertion rate ($r = 0.66$), a strong negative correlation with the substitution rate ($r = -0.73$), but no correlation with microdeletions. The SINE distribution correlations may be due to local G+C content, but the relationships are complicated. For example, G+C density shows modest-to-strong positive correlation with both SINEs and insertions. However, SINEs are known to prefer to insert in locally G+C-rich DNA, and this alone may govern the connection between G+C content and the other events. The human chromosomes have a lower G+C content than in rodents, leading to a generally lower rate of G or C gain substitutions (compared with A/T gains) (Supplementary Table 3) exacerbating the complexity of the effect of G+C SINE insertions[14,15].

To better understand more recent small-scale changes leading to the current human sequence, we compared chromosome 12 to the available orthologous regions of the chimpanzee and rhesus macaque genomes. Although neither nonhuman assembly represents a 'deep draft', high sequence identity enabled alignment of about 87% of the human chromosome to both nonhuman primates. The availability of two additional primate genomes allowed reconstruction of the ancestral hominoid genome from the three-way alignment. Human- and chimpanzee-specific evolutionary changes, listed in Table 1, were defined by differences in each species compared with the ancestral hominoid.

Previous comparisons of substitution rates per year in neutral sites between human, rat and mouse[14] revealed a lower rate in the lineage from the human–rodent ancestor to human (1.73×10^{-9}) compared with the rate from the human–rodent ancestor to rat (4.95×10^{-9}) and mouse (4.84×10^{-9}). Other findings indicate hominoid substitution rates are lower than the rates in Old World monkeys and in other eutherian mammals[16–18]. Our analysis reveals further slow down in the branch from the hominoid ancestor to human (9.70×10^{-10}) compared with chimpanzee (1.19×10^{-9}). (See note added in proof.)

The pattern of insertion–deletions (indels) in human chromosome 12 relative to the hominoid ancestor is summarized in Table 2. Microsatellite expansion events, predominantly A/T mononucleotide and CA dinucleotide expansions, comprise 19% of insertion events in the 1–100-bp range. Trinucleotide expansions are more common in coding (1.62 per million base pairs (Mbp)) than noncoding regions (0.727 per Mbp). Comparisons of insertions in the 101–8,000-bp range to the "Retroposed Genes" track of the University of California at Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/; ref. 19) revealed 12 human-specific pseudogenes. Comparisons to known segmental duplications[20] revealed nine human-specific duplications. Human-specific retrotransposon insertions in the 101–8,000-bp size range are classified by type in Table 3. Consistent with previous results[21], SINE insertion rates in the human lineage are twofold higher than in the chimpanzee lineage. All retrotransposon insertions were intergenic or intronic, except for an AluY insertion in the 3′ untranslated region (UTR) of *RAB21*, a member of the *RAS* oncogene family. The overall ratio of inserted to deleted base pairs was 1.67, lengthening the human chromosome by 2.69% relative to the ancestral hominoid and consistent with previously detected evolutionary increases in the size of the human genome[22].

Chromosome 12 is rich in disease-associated loci, and a total of 487 disease genes have been assigned to this chromosome[2], accounting for 5.2% of all disease genes. In Supplementary Tables 4 and 5, we present the most cited genes on chromosome 12 together with the medically relevant genes sorted by function, respectively.

With regard to disease, chromosome 12 is best known for its links with cancer. Three genes mapping to the chromosome have been associated with cancer-related chromosome translocations. The most prominent of these is the *ETV6* (*TEL1*) gene at 12p13,

Table 2 | Human-specific insertion and deletion rates

	Intergenic	Intronic	UTR	CDS
1–100-bp indels				
Insertion events	0.3636	0.3938	0.2748	0.0316
Inserted bases	1.9768	2.0958	1.2722	0.2662
Microsatellite expansion events	0.0669	0.0779	0.0344	0.0024
Microsatellite expansion bases	0.4483	0.4512	0.0022	0.0002
Deletion events	0.3659	0.3672	0.3092	0.0267
Deleted bases	2.5017	2.4870	1.5683	0.2007
101–8,000-bp indels				
Insertion events	0.0131	0.0173	0.0059	0.0049
Inserted bases	9.0789	12.9904	9.3402	2.1110
Retrotransposon insertion events	0.0061	0.0075	0.0012	0.0000
Retrotransposon insertion bases	4.8921	5.4302	0.3897	0.0000
Deletion events	0.0235	0.0206	0.0154	0.0000
Deleted bases	5.6285	4.7809	5.1019	0.0000

Indels were calculated per kilobase of sequence within the genomic regions indicated.

Table 3 | Human-specific insertions of retrotransposons

	Insertion events		Inserted bases per Mb
	Total count	Per Mb	
SINEs	419	3.72	1083.07
AluY	338	3.00	358.90
AluYa5	116	1.03	316.44
AluYb8	79	0.70	220.21
LINEs	228	2.02	3010.20
L1	225	2.00	3002.83
L1PA2	52	0.46	1494.21
L1HS	13	0.12	312.74
LTRs	51	0.45	328.00
SVAs*	50	0.44	628.13

*SVAs are hominoid-specific composite retrotransposons comprised of SINE-R, VNTR (variable number of tandem repeats) and Alu sequences.

encoding an ETS-like transcription factor, which has a central role in haematopoietic malignancies including acute lymphoblastic leukaemia (ALL), acute myeloblastic leukaemia (AML) and myelodysplastic syndromes (MDS). Furthermore, 5% of children with ALL have 12p13–p12 deletions[23]. The other two genes on this chromosome involved in oncogenic gene fusions are *DDIT3* (or *CHOP*; 12q13.1), often rearranged in myxoid liposarcoma, and *HMGA2* (12q15), which has been found fused to various genes in lipoma, salivary adenoma, uterine leiomyoma and multiple lipomatosis. In addition, several genes mapped to this chromosome have been directly or indirectly associated with cancer, including *BCL7A* (12q24.13) in B-cell non-Hodgkin lymphoma, *P2RX7* (12q24) in susceptibility to chronic lymphatic leukaemia (CLL), *YEATS4* (12q13–q15) in glioma, *CDK4* (12q14) in melanoma, *ACVR1B* (12q13) in pancreatic cancer, and *KRAS* (12p12.1) in colorectal adenoma.

There are 16 genes associated with movement disorders mapped to chromosome 12, as well as two diabetes-related genes, genes associated with four separate mood disorders, and four genes involved in heart disease. Of the 28 members of the keratin gene family mapped to this chromosome, 12 have been linked to skin and hair disorders. Chromosome 12 also harbours the human *CD4* locus, which encodes the main receptor for the human immunodeficiency virus, at 12p3.1. This locus is within one of the most dense gene clusters in the genome, and has been the focus of many functional and population genetic studies[24].

A primary goal in generating a high quality, finished reference sequence for chromosome 12 is to provide the research community with the resources to accelerate the search for additional disease-causing genes. As an example, Li and co-workers[25] were recently able to demonstrate an association between sequence variants of the *GAPDH* gene on chromosome 12 and late-onset Alzheimer's disease (LOAD).

The finished sequence of chromosome 12 reported here marks the beginning of more extensive studies aimed at understanding our evolutionary history, together with the variation in genome structure and sequence that defines us as individuals.

Note added in proof: Recent data[31] showed that of the nine human chromosomes assayed, chromosome 12 demonstrated the greatest slow down in single nucleotide substitution rate.

METHODS

Mapping and sequencing. Gap closure, sequencing and finishing strategies, and clone integrity assays, are described in Supplementary Methods. BAC clone overlaps were verified by BLAST (http://www.ncbi.nlm.nih.gov/BLAST/) in-house using a locally installed copy of the software, and polymorphic regions within overlaps were confirmed by polymerase chain reaction (PCR) using a bi-gender, multi-ethnic pool of genomic DNA isolated from eight individuals (J. Belmont, personal communication). Coverage, integrity and clone order were analysed using available genetic and radiation hybrid map markers (see Supplementary Methods and Supplementary Fig. 2). Additionally, we aligned the unique paired-end fosmid reads to the finished chromosome sequence, and concluded that the limited set of clustered size discrepancies reflected probable polymorphisms between clone libraries. The unique paired fosmid end-sequence (Broad Institute) analysis was performed using sequences downloaded from the UCSC Genome Browser that were checked for both paired end-sequence orientation and resulting insert size.

Annotation. Manual curation identified each known gene, novel gene and novel transcript locus, defined as a set of one or more transcripts that share at least one exon of coding sequence (in frame) and supported by full-length and partial human or vertebrate cDNA sequences having a best-in-genome BLAST or blat (http://genome.ucsc.edu/cgi-bin/hgBlat) hit (≥98% identity). All analyses were performed in-house using locally installed copies of the software. The cDNA/RefSeq sequences were compared to the genomic sequence to place exons. All intron/exon splice sites for all predicted exons were examined for canonical splice motifs. Coding regions were examined for a best-fit open reading frame (ORF). The 5′ and 3′ UTRs were annotated and extended using available EST and cDNA evidence, and poly(A) sites and poly(A) signals were annotated on each gene where identified. Alternative splice variants were identified from cDNA, EST and protein evidence. The translation product for each coding sequence was verified using Swissprot. Pseudogenes were defined as sequences with no direct evidence for expression but that match with a high score to a spliced messenger RNA or spliced EST from elsewhere in the genome. This is a more stringent definition than has been applied by others in broad genomic screens of pseudogenes, and results in a fivefold-lower count across chromosome 12 than previously reported[26]. For paralogue analysis, protein sequences corresponding to the 'known genes' track of the UCSC Genome Browser[27] were compared in an all-against-all BLAST search. Two loci were defined as paralogues if there was a match of any of their transcript variants with the following criteria: expect value cutoff of 10^{-10} or less, the lengths of the matching transcripts were within 20% of each other, and the match length extended over 70% of the average length of the two sequences. The complete set of annotations has been submitted to the Vega database (http://vega.sanger.ac.uk/Homo_Sapiens/).

Landscape features. For CpG analysis, a CpG island has been defined as an expanse of >200 nucleotides in which the G+C content is greater than 50% and the ratio of observed CG dinucleotides to expected in the segment is >0.6. We developed databases of known microRNAs, small nuclear (sn) RNAs and small nucleolar (sno) RNAs, and used tRNAscan-SE v.1.23 (http://www.genetics.wustl.edu/eddy/tRNAscan-SE; installed locally) and BLAST to search for non-coding RNA sequences. We identified recent intra- and interchromosomal segmental duplications by BLAST searching the repeat-masked chromosome sequence against itself and the rest of the human genome. The duplication densities were calculated by averaging the duplications of each base over nonoverlapping 100-kb windows after filtering low identity matches (<90%). The densities of SINEs, LINEs and long terminal repeats (LTRs) were calculated from repeat-masked data using 100-kb windows. The G+C density was calculated by counting the G+C content over nonoverlapping 100-kb windows. The densities of CpG islands (UCSC), genes (Baylor College of Medicine Human Genome Sequencing Center annotations), and pseudogenes (as defined above) were counted and displayed using 1-Mb windows. Markers were placed on the genomic sequence using a combination of locally installed e-PCR (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/) and BLAST software packages.

Comparative analysis. The multiple alignments of human, chimp (panTro1), dog (canFam1), mouse (mm5), rat (rn3), chicken (galGal2), zebrafish (danRer1) and fugu (fr1) were downloaded from the UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/; May 2004/hg17-Build35). The pairwise synteny blocks between human and other species were parsed out with Synteny-Parser (X. Song & G. Weinstock, unpublished perl script), which was tuned to include all visible chromosome rearrangements in the dot plot. Rhesus scaffolds from the Mmul_0.1 preliminary assembly were mapped to human using Pash[28]. Rhesus scaffolds that mapped to human chromosome 12 by both Pash and the human–rhesus Alignment Net (UCSC) were aligned with orthologous human regions and chimpanzee regions from the Human–Chimpanzee Reciprocal-Best Chain alignments (UCSC) using MLAGAN[29].

Received 17 December; accepted 31 December 2005.

1. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
2. Online Mendelian Inheritance in Man (OMIM). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, Maryland) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, Maryland) ⟨http://www.ncbi.nlm.nih.gov/omim/⟩ (2000).
3. Yu, F. *et al.* Positive selection of a pre-expansion CAG repeat of the human *SCA2* gene. *PLoS Genet.* **1,** e41 (2005).
4. Montgomery, K. T. *et al.* A high-resolution map of human chromosome 12. *Nature* **409,** 945–946 (2001).
5. McPherson, J. D. *et al.* A physical map of the human genome. *Nature* **409,** 934–941 (2001).
6. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
7. Knight, S. J. *et al.* An optimized set of human telomere clones for studying telomere integrity and architecture. *Am. J. Hum. Genet.* **67,** 320–332 (2000).
8. Vermeesch, J. R. *et al.* A physical map of the chromosome 12 centromere. *Cytogenet. Genome Res.* **103,** 63–73 (2003).
9. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14,** 942–950 (2004).
10. Yu, A. *et al.* Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
11. Wienberg, J. Fluorescence *in situ* hybridization to chromosomes as a tool to understand human and primate genome evolution. *Cytogenet. Genome Res.* **108,** 139–160 (2005).
12. Nickerson, E. & Nelson, D. L. Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50,** 368–372 (1998).
13. Kehrer-Sawatzki, H., Sandig, C. A., Goidts, V. & Hameister, H. Breakpoint analysis of the pericentric inversion between chimpanzee chromosome 10 and

the homologous chromosome 12 in humans. *Cytogenet. Genome Res.* **108,** 91–97 (2005).
14. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
15. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
16. Li, W. H. & Tanimura, M. The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326,** 93–96 (1987).
17. Steiper, M. E., Young, N. M. & Sukarna, T. Y. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. *Proc. Natl Acad. Sci. USA* **101,** 17021–17026 (2004).
18. Yi, S., Ellsworth, D. L. & Li, W. H. Slow molecular clocks in Old World monkeys, apes, and humans. *Mol. Biol. Evol.* **19,** 2191–2198 (2002).
19. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **100,** 11484–11489 (2003).
20. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. & Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).
21. Hedges, D. J. *et al.* Differential *Alu* mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res.* **14,** 1068–1075 (2004).
22. Liu, G. *et al.* Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res.* **13,** 358–368 (2003).
23. Stegmaier, K. *et al.* Frequent loss of heterozygosity at the *TEL* gene locus in acute lymphoblastic leukemia of childhood. *Blood* **86,** 38–44 (1995).
24. Ansari-Lari, M. A. *et al.* A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human chromosome 12p13. *Genome Res.* **6,** 314–326 (1996).
25. Li, Y. *et al.* Association of late-onset Alzheimer's disease with genetic variation in multiple members of the *GAPD* gene family. *Proc. Natl Acad. Sci. USA* **101,** 15688–15693 (2004).
26. Zhang, Z., Harrison, P. M., Liu, Y. & Gerstein, M. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13,** 2541–2558 (2003).
27. Hsu, F. *et al.* The UCSC Proteome Browser. *Nucleic Acids Res.* **33** (suppl. 1), D454–D458 (2005).
28. Kalafus, K. J., Jackson, A. R. & Milosavljevic, A. Pash: Efficient genome-scale sequence anchoring by Positional Hashing. *Genome Res.* **14,** 672–678 (2004).
29. Brudno, M. *et al.* LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721–731 (2003).
30. Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A. & Tesler, G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res.* **15,** 98–110 (2005).
31. Elango, N., Thomas, J. W., NISC Comparative Sequencing Program & Soojin, V. Y. Variable molecular clocks in hominoids. *Proc. Natl Acad. Sci. USA* **103,** 1370–1375 (2006).

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

Acknowledgements This work was supported by NIH grants to R.K. and to R.A.G. We acknowledge and thank the genome sequencing community for generating the data sets used in our comparative analysis. We also acknowledge the following members of the HUGO Gene Nomenclature Committee: S. Povey (chair), E. A. Bruford, V. K. Khodiyar, M. J. Lush, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright.

Author Information The chromosome 12 sequence has been deposited into GenBank under the accession number NC_000012. 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.S. (sscherer@bcm.tmc.edu).

Baylor College of Medicine Human Genome Sequencing Center Sequence Production Team Carlana C. Allen[1], Anita G. Amin[1], Vivian Anyalebechi[1], Michael Bailey[1], Joseph A. Barbaria[1], Kesha E. Bimage[1], Nathaniel P. Bryant[1], Paula E. Burch[1], Carrie E. Burkett[1], Kevin L. Burrell[1], Eliana Calderon[1], Veronica Cardenas[1], Kelvin Carter[1], Kristal Casias[1], Iracema Cavazos[1], Sandra R. Cavazos[1], Heather Ceasar[1], Joseph Chacko[1], Sheryl N. Chan[1], Dean Chavez[1], Constantine Christopoulos[1], Joseph Chu[1], Raynard Cockrell[1], Caroline D. Cox[1], Michelle Dang[1], Stephanie R. Dathorne[1], Robert David[1], Candi Mon'Et Davis[1], Latarsha Davy-Carroll[1], Denise R. Deshazo[1], Jeremy E. Donlin[1], Lisa D'Souza[1], Kristy A. Eaves[1], Amy Egan[1], Alexandra J. Emery-Cohen[1], Michael Escotto[1], Nicole Flagg[1], Lisa D. Forbes[1], Abdul M. Gabisi[1], Melissa Garza[1], Cerissa Hamilton[1], Nicholas Henderson[1], Omar Hernandez[1], Sandra Hines[1], Marilyn E. Hogues[1], Mei Huang[1], DeVincent G. Idlebird[1], Rudy Johnson[1], Angela Jolivet[1], Sally Jones[1], Ryan Kagan[1], Laquisha M. King[1], Belita Leal[1], Heather Lebow[1], Sandra Lee[1], Jaclyn M. LeVan[1], Lakeshia C. Lewis[1], Pamela London[1], Lorna M. Lorensuhewa[1], Hermela Loulseged[1], Demetria A. Lovett[1], Alice Lucier[1], Raymond L. Lucier[1], Jie Ma[1], Renita C. Madu[1], Patricia Mapua[1], Ashley D. Martindale[1], Evangelina Martinez[1], Elizabeth Massey[1], Samantha Mawhiney[1], Michael G. Meador[1], Sylvia Mendez[1], Christian Mercado[1], Iracema C. Mercado[1], Christina E. Merritt[1], Zachary L. Miner[1], Emmanuel Minja[1], Teresa Mitchell[1], Farida Mohabbat[1], Khatera Mohabbat[1], Baize Montgomery[1], Niki Moore[1], Sidney Morris[1], Mala Munidasa[1], Robin N. Ngo[1], Ngoc B. Nguyen[1], Elizabeth Nickerson[1], Ogechi O. Nwaokelemeh[1], Stanley Nwokenkwo[1], Melissa Obregon[1], Maryann Oguh[1], Njideka Oragunye[1], Rodolfo J. Oviedo[1], Bridgette J. Parish[1], David N. Parker[1], Julia Parrish[1], Kenya L. Parks[1], Heidie A. Paul[1], Brett A. Payton[1], Agapito Perez[1], William Perrin[1], Adam Pickens[1], Eltrick L. Primus[1], Ling-Ling Pu[1], Maria Puazo[1], Miyo M. Quiles[1], Juana B. Quiroz[1], Dina Rabata[1], Kacy Reeves[1], San Juana Ruiz[1], Hongmei Shao[1], Ida Sisson[1], Titilola Sonaike[1], Richard P. Sorelle[1], Angelica E. Sutton[1], Amanda F. Svatek[1], Leah Anne Svetz[1], Kavitha S. Tamerisa[1], Tineace R. Taylor[1], Brian Teague[1], Nicole Thomas[1], Rachel D. Thorn[1], Zulma Y. Trejos[1], Brenda K. Trevino[1], Ogechi N. Ukegbu[1], Jeremy B. Urban[1], Lydia I. Vasquez[1], Virginia A. Vera[1], Donna M. Villasana[1], Ling Wang[1], Stephanie Ward-Moore[1], James T. Warren[1], Xuehong Wei[1], Flower White[1], Angela L. Williamson[1], Regina Wleczyk[1], Hailey S. Wooden[1], Steven H. Wooden[1], Jennifer Yen[1], Lillienne Yoon[1], Vivienne Yoon[1] & Sara E. Zorrilla[1]

Affiliation for participants: [1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA.