ARTICLES

Human chromosome 11 DNA sequence and analysis including novel gene identification

Todd D. Taylor[1], Hideki Noguchi[1]†, Yasushi Totoki[1], Atsushi Toyoda[1], Yoko Kuroki[1], Ken Dewar[2]†, Christine Lloyd[3], Takehiko Itoh[4], Tadayuki Takeda[1], Dae-Won Kim[5], Xinwei She[6], Karen F. Barlow[3], Toby Bloom[2], Elspeth Bruford[7], Jean L. Chang[2], Christina A. Cuomo[2], Evan Eichler[6], Michael G. FitzGerald[2], David B. Jaffe[2], Kurt LaButti[2], Robert Nicol[2], Hong-Seog Park[5], Christopher Seaman[2], Carrie Sougnez[2], Xiaoping Yang[2], Andrew R. Zimmer[2], Michael C. Zody[2], Bruce W. Birren[2], Chad Nusbaum[2], Asao Fujiyama[1,8], Masahira Hattori[1,9], Jane Rogers[3], Eric S. Lander[2] & Yoshiyuki Sakaki[1]

Chromosome 11, although average in size, is one of the most gene- and disease-rich chromosomes in the human genome. Initial gene annotation indicates an average gene density of 11.6 genes per megabase, including 1,524 protein-coding genes, some of which were identified using novel methods, and 765 pseudogenes. One-quarter of the protein-coding genes shows overlap with other genes. Of the 856 olfactory receptor genes in the human genome, more than 40% are located in 28 single- and multi-gene clusters along this chromosome. Out of the 171 disorders currently attributed to the chromosome, 86 remain for which the underlying molecular basis is not yet known, including several mendelian traits, cancer and susceptibility loci. The high-quality data presented here—nearly 134.5 million base pairs representing 99.8% coverage of the euchromatic sequence—provide scientists with a solid foundation for understanding the genetic basis of these disorders and other biological phenomena.

Human chromosome 11 (HSA11), which represents approximately 4.4% of the human genome[1,2], has had a significant role in the history of molecular genetics, beginning long before its complete sequencing was undertaken. The haemoglobin beta gene, encoding one of the best-studied proteins, was one of the first genes mapped to the human genome (11p15.5) and was the first protein to have its crystal structure solved[3]. It is also the cause of sickle cell anaemia, the first human genetic disease for which a molecular basis was demonstrated[4]. Three megabases (Mb) distal lies the insulin gene, encoding the first fully-sequenced protein[5], and the intensely studied imprinting region responsible for Beckwith–Wiedemann syndrome[6]. The physical map, high-quality finished sequence and gene catalogue presented here are but the latest landmark in an effort to understand the unique characteristics and functions of this chromosome.

The clone map and finished sequence

Chromosome 11 was sequenced using a clone-by-clone shotgun sequencing approach. The sequence is in eight finished contigs (Supplementary Tables S1–S3), the largest being 49.6 Mb, with seven gaps remaining, including one at 11p-tel (~50 kilobases (kb)), one heterochromatic gap (207 kb) near 11p-cen and five small internal clone gaps (totalling ~64.5 kb). Where possible, all of the gaps were size-estimated by fibre-fluorescence *in situ* hybridization (FISH) analysis. On 11q, we reached both the telomeric repeats and the centromeric alpha satellite repeats, and higher-order repeat structure was observed in clone AC126345 at 11p-cen. To ensure production of the most reliable data, sequence quality control checks were performed both internally (Supplementary Table S4) and externally[7]. In total, we finished 131,130,853 base pairs (bp) and estimate the total size of the chromosome, including the gaps and centromere, to be approximately 134.5 Mb (May 2004, NCBI build 35). The coverage of the euchromatic portion of the chromosome is an estimated 99.8%. Of the finished sequence, 60% was generated by RIKEN Genomic Sciences Center, 36% by the Broad Institute of MIT and Harvard, 3% by the Wellcome Trust Sanger Institute, and 1% by the Washington University School of Medicine Genome Sequencing Center.

The chromosome landscape

Figure 1 shows an overview of the chromosome 11 landscape. HSA11 is very gene rich and there are many clustered gene families located on the chromosome. According to a recent survey of the Ensembl genome browser[8], HSA11 contains the fourth highest number of genes in the human genome, after human chromosomes 1, 2 (ref. 9) and 19 (ref. 10), respectively. These data show 10.6 protein-coding genes per Mb on HSA11, as compared to the genome-wide average of 7.3. In fact, manual annotation of the chromosome identifies a slightly higher gene density of 11.6 genes per Mb, with genes spaced

[1]RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [2]Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]Mitsubishi Research Institute, Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan. [5]Korea Research Institute of Bioscience & Biotechnology, 52 Oun-dong, Yusong-gu, Daejeon 305-333, South Korea. [6]University of Washington, Genome Sciences, HSB K336B, Box 357730, Seattle, Washington 98195, USA. [7]HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. [8]National Institute of Informatics, Hitotsubashi 2-1-2, Chiyoda-ku, Tokyo 101-8430, Japan. [9]Kitasato Institute for Life Sciences, Kitasato University 1-15-1, Kitasato, Sagamihara, Kanagawa 228-8555, Japan. †Present addresses: University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-0882, Japan (H.N.); McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada (K.D.).

© 2006 Nature Publishing Group

an average of 86 kb apart. Both the repeat density (47.98%; Supplementary Table S5 and Supplementary Information) and G+C content (41.57%) are close to genome-wide averages. Table 1 lists various features of the chromosome.

The finished sequence of HSA11 shows strong concordance with existing physical and genetic maps. All sequence-tagged sites from the Généthon microsatellite-based genetic map[11], the deCODE map[12] and the Marshfield genetic maps[13] are present in the HSA11 sequence (Supplemental Methods). We compared recombination rates in the deCODE female, male and sex-averaged meiotic maps (which average 1.53, 0.85 and 1.19 cM per Mb, respectively) with the physical distance as determined from the sequence assembly (Supplementary Fig. S1). Recombination statistics for HSA11 are similar to other human chromosomes, showing a relatively linear relationship between recombination rate and physical distance.

Gene catalogue

We annotated a total of 2,347 gene loci consisting of 1,524 potentially active protein-coding genes, 765 pseudogenes and 58 RNAs (Supplementary Table S6 and Supplementary Methods). The 1,524 protein-coding genes comprise 1,195 known genes (including 166 olfactory receptor genes), 104 novel coding sequences (CDSs), 221 novel transcripts and four putative genes. Some of these genes were identified by our *ab initio* gene prediction program DIGIT[14], as described below. The 765 pseudogenes include at least three unprocessed pseudogenes and 203 olfactory receptor pseudogenes. In total, we annotated 230 previously unknown genes (that is, no RefSeq or Ensembl location, Supplementary Methods) consisting of 48 novel CDSs, 178 novel transcripts and four putative genes. These novel genes are scattered throughout the chromosome, with many located in potential disease candidate regions.

There are 296 single-exon genes, of which 168 belong to the olfactory receptor gene family. The remaining 1,228 multi-exon genes (80.53%) have an average of 9.39 exons per gene. In addition to the olfactory receptor gene clusters described later, we identified 142 genes in 37 clusters that belong to gene families with at least two members on HSA11 (Supplementary Table S7).

Co-transcribed or read-through genes do not appear to be a very common phenomenon in the human genome, but this could be due to the current lack of uniform genome-wide gene annotation. We found 12 cases on HSA11 (Supplementary Table S8), which are each supported by just one messenger RNA (and in a few cases by expressed sequence tags (ESTs)). Besides these examples, we found only a few other examples on chromosomes 17 and 22 (ref. 15) (Supplementary Methods). Of these, only two were found that probably result in a protein fusion product, *TRIM6-TRIM34* and *BSCL2-HNRPUL2*. Whether or not these read-through transcripts should be considered as alternative transcripts or separate genes, with functions different from the two genes they connect, remains to be investigated. Because the supporting evidence for such read-through transcripts is usually minimal, additional experiments should first be carried out to determine whether or not they are real, or just represent cellular mistakes or artefacts.

For the protein-coding genes, we attempted to identify all possible splice variants using currently available mRNA data (and, in a few cases, EST information). We found that 805 (52.8%) of the genes have at least two or more variants, consisting of 738 known genes, 36 novel CDSs, 30 novel transcripts and one putative gene. The genes with at least two variants have an average of 3.73 variants per locus. The *CTNND1* gene showed the largest number of variants with 28. In total, we identified 3,723 variants for the 1,524 expressed genes. Of these nearly 4,000 splice variants, there are many instances where the transcripts splice correctly but do not have definitive or long (>100 amino acids) open reading frames and may be examples of incompletely spliced RNAs, incorrectly spliced RNAs or non-coding RNAs.

We explored whether there was any correlation between the presence of a CpG island and the number of variant transcripts for a gene (Supplementary Table S9). Interestingly, we found a significant correlation ($\chi^2 = 224.29$, $P < 0.0001$, 6 degrees of freedom). Out of the 894 genes with CpG islands, 650 (70%) have two or more variants. By contrast, of the 626 genes with no CpG islands, only 154 (24.6%) have two or more variants.

Olfactory receptor genes

Olfactory receptor genes comprise the largest multi-gene family in metazoans. All human chromosomes except HSA20 (ref. 16) and HSAY (ref. 17) contain olfactory receptor genes, but HSA11 is by far the richest. In human there are 856 olfactory receptor genes, 369 (43%) of which are located on HSA11 (ref. 18). These are mostly single-exon genes, with an average length of about 1 kb. Of the 369 loci on HSA11, 166 (45%) are protein-coding and the other 203 (55%) are pseudogenes; this is close to the genome-wide average (47% versus 53%). All but 10 of the olfactory receptor genes on HSA11 lie within 18 clusters, separated by at least 100 kb (Figs 1 and 2; see also Supplementary Table S10). The largest cluster contains 97 genes over a range of 1.5 Mb. The average distance between genes within a cluster is about 17 kb. The olfactory receptor genes on HSA11 are classified into 13 different families (having >40% protein identity), containing from as few as one to as many as 81 members. The olfactory receptor regions on HSA11 generally are rich in L1 repeats, poor in Alu repeats, CpG islands (Supplementary Table S11 and Supplementary Information) and predicted transcription starts (based on the Eponine program[19]), and have a G+C content of 40% or lower. Functional olfactory receptor genes are evenly distributed within the clusters.

Olfactory receptor genes are roughly classified into two classes: I and II. Class I olfactory receptor genes are known as fish-like olfactory receptors and are believed to be receptors for water-soluble ligands. They have expanded in mammalian lineages (Fig. 2) and many belong to one large cluster in mammalian genomes. In the human genome, all of the class I olfactory receptor genes are found in three closely spaced clusters on the subtelomeric region of 11p, from

Figure 1 | Chromosome 11 landscape. The following sections appear in order from top to bottom. (1) Cytogenetic banding pattern. (2) Non-olfactory receptor (OR) expressed genes. Genes are colour coded according to category type (known genes, red; novel CDS, green; novel transcript, blue; putative, orange) and are shown according to their orientation on the chromosome (upper, forward orientation; lower, reverse orientation: note that this applies to sections 2–6, which means that these sections appear twice (with sections 7–9 sandwiched between them), once for forward and once for reverse orientation). (3) Non-olfactory receptor clustered gene families. (4) Non-olfactory receptor pseudogenes. (5) Miscellaneous RNAs including tRNAs, microRNAs and snoRNAs. (6) Eponine TSS predicted transcription start sites. (7) CpG islands. (8) Exofish evolutionarily conserved regions (ECRs). (9) Gene deserts larger than 650 kb. (10) Olfactory receptor genes. The olfactory receptor genes are coloured coded (connecting line, see key on left) according to family and whether a coding sequence (blue symbol name) or pseudogene (grey symbol name). Upper, forward orientation on chromosome; lower, reverse orientation on chromosome. (11) G+C content. This was calculated using a sliding window of 100 kb. (12) Short interspersed element (SINE)/Alu (blue) and long interspersed element (LINE)/L1 (red) repeat densities. This was also calculated using a sliding window of 100 kb. (13) Interspersed repeat elements as reported by RepeatMasker (SINEs, LINEs, long terminal repeat (LTR) elements, DNA elements, unclassified repeats, simple repeats, low-complexity regions, RNAs and satellite repeats). Also included in this region are tandem repeats as predicted by the Tandem Repeat Finder program. (14) Interchromosomal duplications (red) and intrachromosomal duplications (blue). (15) Gene disorders (mapped only). Forty-one disorders that have been mapped to HSA11 are shown as coloured horizontal bars, which represent the possible genetic location of the disorder according to OMIM (Online Mendelian Inheritance in Man). Clone gaps (including the centromere) are indicated by the vertical grey blocks. ►

© 2006 Nature Publishing Group

Table 1 | Chromosome 11 sequence features

Chromosome property	Value
Chromosome length (bp)	134,452,3843
Finished sequence length (bp)	131,130,853
Total gap length including 3 Mb centromere (bp)	3,321,531
Euchromatic region coverage (%)	99.76
Protein-coding gene loci	1,524
Known	1,195
Novel CDS	104
Novel transcripts	221
Putative	4
Pseudogenes	765
tRNAs (all)	19
tRNA pseudogenes	2
MicroRNAs	12
snoRNAs	27
Gene coverage (bp)	61,581,901 (46.96%)
Exon coverage (bp)	3,543,101 (2.70%)
Known gene mean size (bp)	45,007
CpG islands	1,369
Genes with 5′ CpG islands	806
Gene deserts (>650 kb)	19
Gene desert coverage (bp)	23,378,900 (17.83%)

snoRNAs, small nucleolar RNAs.

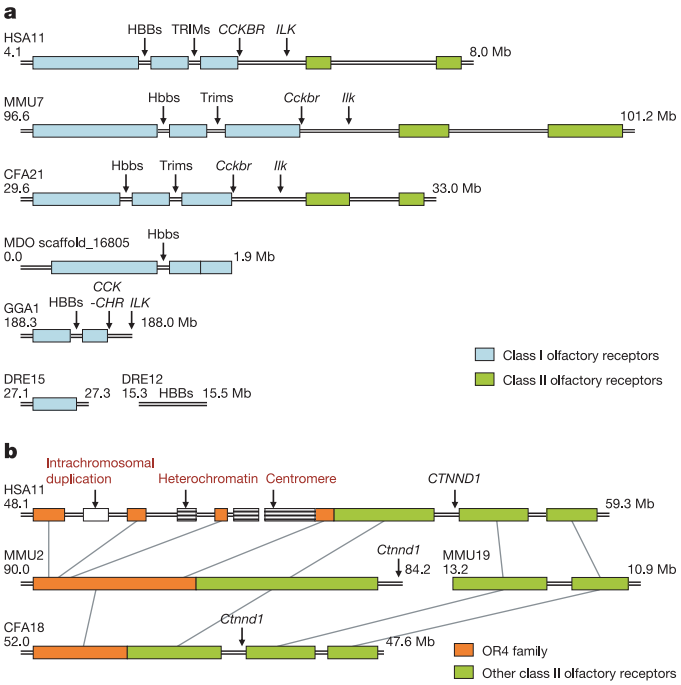

Figure 2 | Conservation and expansion of olfactory receptor clusters. **a**, Structure of the class I olfactory receptor genes and neighbouring class II genes (HSA11 4.1–8.0 Mb). The basic structure of the class I cluster is highly conserved. The beta-globin gene cluster was inserted after the divergence of fish and other vertebrates. The TRIM cluster was inserted in the placental mammal lineage. **b**, The olfactory receptor 4 (OR4) superfamily cluster (HSA11 48.1–55.2 Mb) was divided into four pieces by insertion of the centromere, a heterochromatic region and an intrachromosomal duplication. Overall, the olfactory receptor gene clusters in mouse were expanded. HSA11, human chromosome 11; MMU2, MMU7 and MMU19, mouse chromosomes 2, 7 and 19; CFA18 and CFA21, dog chromosomes 18 and 21; MDO, opossum; GGA1, chicken chromosome 1; DRE12 and DRE15, zebrafish chromosomes 12 and 15.

4.1 to 6.2 Mb. Approximately 50% (54 out of 103) of these genes are intact in human, which is close to the genome-wide average for all olfactory receptors. The class I region is interrupted by a few genes including the beta-globin gene cluster and a TRIM gene cluster.

The most significant class II cluster is located around the centromere of HSA11. The corresponding clusters for the mouse[20], rat[21] and dog[22] genomes are also the largest ones, and are comprised of many families of class II olfactory receptor genes. Notably, the human cluster has a significantly different structure from that of other mammals: the human cluster is divided by insertion of the centromere, a heterochromatic region and an intrachromosomal duplication. Despite the structural changes, and although some members of the cluster are on different chromosomes in rodents, analysis of conserved order of orthologous sequences suggests that they belonged to one large cluster in their last common ancestral genome.

Identification of weakly expressed novel genes

As mentioned above, we applied DIGIT[14] (Methods), an *ab initio* gene-finder, to HSA11. The program predicted 65 novel protein-coding gene loci, with an average open-reading length of 366 amino acids, the genomic regions of which did not overlap at the time with human mRNAs from GenBank. We found experimental support for 34 (52%) of these predictions, based on reverse transcription polymerase chain reaction (RT–PCR) experiments that show evidence of the predicted splice junctions (including four full CDSs) (Supplementary Tables S12 and S13). Most (26 out of 34) of these genes appear to be expressed at an extremely low level (detectable only by nested PCR), which might explain why they had not been previously detected by any high-throughput EST or full-length cDNA sequencing strategy. Of the 34 genes with experimental support (Supplementary Table S14), 12 were identified only through this method because they have no orthologous sequence in any currently available genome (six also have no related human sequence, whereas six have related human sequence). Eight of the 34 genes may simply be extensions of nearby human genes. The remaining 14 genes either have orthologous sequence in other species or are highly similar to known human genes. Many of the 34 genes are predicted by the InterProScan[23] program to contain a functional domain (Supplementary Table S15). Further experimental evidence to support the expression of these genes and to identify their full-length structures is necessary. In order to obtain a more complete catalogue of all protein-coding genes in the human genome, this type of analysis should ideally be extended to include all chromosomes, especially as some genes were only identified by this '*ab initio* plus experimental verification' approach.

Conclusions

This work describes just a few of the interesting features of human chromosome 11. Notably, the chromosome is very rich in genes overall and disease genes in general (Supplementary Fig. S2). It contains many clustered gene families, the most significant being 369 members of the olfactory receptor gene family. Many medically important loci are associated with chromosome 11 for which the genetic cause of the disorder has yet to be elucidated (Supplementary Table S16). This includes various cancers, susceptibility genes and loci implicated in behavioural and psychiatric disease variation.

Some findings that stand out in our analysis include a significant correlation between the presence of CpG islands and the number of splice variants, a large number of overlapping genes (Supplementary Information) and genes sharing CpG islands, and genes that were only initially identified through *ab initio* methods. Although these phenomena may not necessarily be specific to chromosome 11, they do emphasize the need for further uniform analyses and annotation across the entire human genome. With the availability of the high-quality human genomic sequence as presented here, scientists have a solid foundation for identifying and understanding all of the genes and functional elements it holds.

METHODS

Construction of the chromosome 11 large insert libraries. We prepared chromosome-specific bacterial artificial chromosome (BAC; CMB9) and fosmid

© 2006 Nature Publishing Group

(CMF9) libraries by using flow-sorted chromosomal DNA derived from human chromosomes 9–12 (these chromosomes cannot be separated by flow cytometry due to their similar size). For construction of the CMB9 library, sorted DNA derived from cultured lymphoblastoid cells was partially digested by *Sac*I and the fragments were ligated into the pKS145 vector. Transformation was carried out by electroporation into *Escherichia coli* DH10B. The CMF9 library was prepared according to previously described methods[24]. We screened these two libraries, the RPCI-11 whole-genome BAC library, and a few other BAC and P1-derived artificial chromosome (PAC) whole-genome libraries (Supplementary Table S2). The chromosome-specific libraries proved especially useful during the gap-filling stage of the project and for identifying clones near the complex centromeric and telomeric regions.

Clone path construction. Initial seed clones were selected by using the restriction enzyme digest fingerprint data of WUGSC and MIT for HSA11p, and by screening the RPCI-11 BAC library with evenly spaced markers taken from a highly-integrated STS map of the whole human genome for HSA11q. These approaches allowed us to construct quickly a tiling path across most of the chromosome. The remaining gaps were filled by walking from clone end sequences and by re-screening of the clone libraries. The chromosomal locations for some clones in the minimum tiling path were confirmed by FISH analysis, and the lengths of the clone gaps were estimated by fibre-FISH according to previous methods[25]. The procedures for large-insert clone sequencing are described in Supplementary Methods.

Prediction of novel human genes. For exhaustive and efficient rare gene prediction we used DIGIT[14], an *ab initio* gene-finder that finds genes by combining gene predictions from multiple *ab initio* gene-finders such as FGENESH[26], GENSCAN[27] and HMMgene[28]. The reason we used DIGIT is that *ab initio* gene-finders, which do not use sequence similarity, have the potential for exhaustive rare gene prediction. The most remarkable feature of *ab initio* gene-finders is their high sensitivity, especially at the nucleotide level. Conversely, *ab initio* gene-finders also predict many false-positive genes. DIGIT successfully discards many false-positive exons predicted by the individual gene-finders and yields remarkable improvements in specificity without lowering sensitivity as compared with the best accuracies achieved by any single gene-finder. For experimental verification of the candidate genes, RT–PCR was performed using primer sets designed from the predicted exon sequences with a single-strand cDNA library prepared from various human tissues. If, in the first round of RT–PCR, a product could not be detected, a second round of PCR using nested PCR primer sets with the diluted RT–PCR products was conducted. When a PCR product was amplified, sequence analysis was used to confirm that the cDNA fragment was located at the predicted genomic location.

Received 17 October 2005; accepted 7 February 2006.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
3. Bragg, W. L. & Perutz, M. F. The structure of haemoglobin. *Proc. R. Soc. Lond. A* **213,** 425–435 (1952).
4. Ingram, V. M. A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature* **178,** 792–794 (1956).
5. Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.* **49,** 463–481 (1951).
6. Wiedemann, H.-R. Complexe malformatif familial avec hernie ombilicale et macroglossie – un 'syndrome nouveau'? *J. Genet. Hum.* **13,** 223–232 (1964).
7. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
8. Hubbard, T. *et al.* Ensembl 2005. *Nucleic Acids Res.* **33,** D447–D453 (2005).
9. Hillier, L. W. *et al.* Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature* **434,** 724–731 (2005).
10. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
11. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380,** 152–154 (1996).
12. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
13. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
14. Yada, T., Takagi, T., Totoki, Y., Sakaki, Y. & Takaeda, Y. DIGIT: a novel gene finding program by combining gene-finders. *Pac. Symp. Biocomput.* **2003,** 375–387 (2003).
15. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
16. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).
17. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).
18. Olender, T., Feldmesser, E., Atarot, T., Eisenstein, M. & Lancet, D. The olfactory receptor universe – from whole genome analysis to structure and evolution. *Genet. Mol. Res.* **3,** 545–553 (2004).
19. Down, T. A. & Hubbard, T. J. P. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12,** 458–461 (2002).
20. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
21. Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
22. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
23. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33,** W116–W120 (2005).
24. Park, H.-S. *et al.* Newly identified repeat sequences, derived from human chromosome 21qter, are also localized in the subtelomeric region of particular chromosomes and 2q13, and are conserved in the chimpanzee genome. *FEBS Lett.* **475,** 167–169 (2000).
25. Suto, Y., Tokunaga, K., Watanabe, Y. & Hirai, M. Visual demonstration of the organization of the human complement C4 and 21-hydroxylase genes by high-resolution fluorescence *in situ* hybridization. *Genomics* **33,** 321–324 (1996).
26. Salamov, A. & Solovyev, V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10,** 516–522 (2000).
27. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
28. Krogh, A. Two methods for improving performance of a HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5,** 179–186 (1997).

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

Acknowledgements Thanks to the staff, past and present, at RIKEN Genomic Sciences Center and the Broad Institute. We also acknowledge Y. Arai and M. Ohki (mapping), M. Hirai, Y. Suto and Y. Kanoh (fibre-FISH analysis technical support), C. Kawagoe and T. Katayama (computational data management), R. Baertsch and J. Mudge (annotation), V. Heyningen (historical insights), K. Linblad-Toh (preliminary assembly of the *Monodelphis domestica* genome), and the HUGO Gene Nomenclature Committee: S. Povey (chair), T. A. Eyre, V. K. Khodiyar, R. C. Lovering, K. M. B. Sneddon, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright (assignment of official gene symbols). The zebrafish sequence data (assembly Zv4) were produced by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/Projects/D_rerio/wgs.shtml). The authors also acknowledge the Ministry of Education, Culture, Sports, Science and Technology (Japan), the National Human Genome Research Institute (USA) and the Wellcome Trust Sanger Institute (UK) for funding this work.

Author Information The entire sequence for the chromosome is deposited in DDBJ/EMBL/GenBank under accession numbers NT_035113, NT_009237, NT_035158, NT_033903, NT_078088, NT_033927, NT_008984 and NT_033899. Accession numbers for individual clones and genes identified in this study can be found in Supplementary Information. 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 T.D.T. (taylor@gsc.riken.jp) or Y.S. (sakaki@gsc.riken.jp).

© 2006 Nature Publishing Group