

The DNA sequence and analysis of human chromosome 14

Roland Heilig*†, Ralph Eckenberg*†, Jean-Louis Petit*†, Núria Fonknechten*†, Corinne Da Silva*†, Laurence Cattolico*, Michaël Levy*, Valérie Barbe*, Véronique de Berardinis*, Abel Ureta-Vidal*, Eric Pelletier*†, Virginie Vico*, Véronique Anthouard*, Lee Rowen‡, Anup Madan‡, Shizhen Qin‡, Hui Sun§, Hui Du§, Kymberlie Pepin§, François Artiguenave*, Catherine Robert*, Corinne Cruaud*, Thomas Brüls*, Olivier Jaillon*†, Lucie Friedlander*, Gaele Samson*†, Philippe Brottier*, Susan Cure*, Béatrice Séguens*, Franck Anière*, Sylvie Samain*, Hervé Crespeau*, Nissa Abbasi‡, Nathalie Aiach*, Didier Boscus*, Rachel Dickhoff‡, Monica Dors‡, Ivan Dubois*, Cynthia Friedman‡, Michel Gouyvenoux*, Rose James‡, Anuradha Madan‡, Barbara Mairey–Estrada*, Sophie Mangenot*, Nathalie Martins*, Manuela Ménard*, Sophie Oztas*, Amber Ratcliffe‡, Tristan Shaffer‡, Barbara Trask‡, Benoit Vacherie*, Chadia Bellemere*, Caroline Belser*, Marielle Besnard–Gonnet*, Delphine Bartol–Mavel*, Magali Boutard*, Stéphanie Briez–Silla*, Stéphane Combette*, Virginie Dufossé–Laurent*, Carolyne Ferron*, Christophe Lechaplais*, Claudine Louesse*, Delphine Muselet*, Ghislaine Magdelenat*, Emilie Pateau*, Emmanuelle Petit*, Peggy Sirvain–Trukniewicz*, Arnaud Trybou*, Nathalie Vega–Czarny*, Elodie Bataille*, Elodie Bluet*, Isabelle Bordelais*, Maria Dubois*, Corinne Dumont*, Thomas Guérin*, Sébastien Haffray*, Rachid Hammadi*, Jacqueline Muanga*, Virginie Pellouin*, Dominique Robert*, Edith Wunderle*, Gilbert Gauguet*, Alice Roy*, Laurent Sainte–Marthe*, Jean Verdier*, Claude Verdier–Discala*, LaDeana Hillier§, Lucinda Fulton§, John McPherson§, Fumihiko Matsuda||, Richard Wilson§, Claude Scarpelli*, Gábor Gyapay*, Patrick Wincker*, William Saurin*, Francis Quétier*†, Robert Waterston§, Leroy Hood‡ & Jean Weissenbach*†

* Genoscope–Centre National de Séquençage, † UMR–8030, CNRS et Université d'Evry, || Centre National de Génotypage, 91000, Evry, France

‡ Institute for Systems Biology, Seattle, Washington 98103, USA

§ Genome Sequencing Center, Washington University School of Medicine, St Louis, Missouri 63108, USA

Chromosome 14 is one of five acrocentric chromosomes in the human genome. These chromosomes are characterized by a heterochromatic short arm that contains essentially ribosomal RNA genes, and a euchromatic long arm in which most, if not all, of the protein-coding genes are located. The finished sequence of human chromosome 14 comprises 87,410,661 base pairs, representing 100% of its euchromatic portion, in a single continuous segment covering the entire long arm with no gaps. Two loci of crucial importance for the immune system, as well as more than 60 disease genes, have been localized so far on chromosome 14. We identified 1,050 genes and gene fragments, and 393 pseudogenes. On the basis of comparisons with other vertebrate genomes, we estimate that more than 96% of the chromosome 14 genes have been annotated. From an analysis of the CpG island occurrences, we estimate that 70% of these annotated genes are complete at their 5' end.

The draft sequences of the human genome^{1,2} have provided an unprecedented wealth of information on our genome, and have facilitated the identification of genes involved in human diseases. These drafts, however, contain a number of inconsistencies and gaps that have to be resolved to obtain a reliable molecular infrastructure on which we can anchor the entire set of human genes including their transcriptional start and stop signals, exons, splicing variants and regulatory elements, as well as the sequence variations found in various human populations. Nearly complete sequences of chromosomes 22 (ref. 3), 21 (ref. 4) and 20 (ref. 5) have already been achieved in the last three years. As an additional contribution to this goal, we present here the sequence of the euchromatic region of human chromosome 14. This chromosome contains two one-megabase (Mb)-long regions of prime importance for the immune system—the α/δ T-cell receptor (TCR) locus located close to the centromere, and the immunoglobulin heavy chain (IGH) locus adjacent to the telomere—as well as about 60 genes which, when defective, are known to lead to genetic diseases, including spastic paraplegia, Niemann–Pick disease, early onset Alzheimer's disease and a severe form of Usher syndrome.

The sequence quality reaches the internationally adopted standard of 99.99%. We estimate that we cover more than 99.9% of the euchromatic portion of chromosome 14, which, as in the other acrocentric chromosomes, namely 13, 15, 21 and 22, is restricted to its long arm. Particular care has been devoted to evaluating the integrity and accuracy of clone coverage. For its annotation, this sequence has benefited from the availability of the continuously

increasing amount of both genomic and expressed sequence data from humans and other vertebrates.

Sequence assembly and validation

Construction of the tiling path of DNA fragments (clones and polymerase chain reaction (PCR) products) that were used as sequencing material has been described previously⁶. Resolution of two gaps that remained in this sequence-ready map and of a few additional problems that appeared during the validation procedure are detailed in the Supplementary Information.

The final sequence-ready map consisting of 624 bacterial artificial chromosome (BAC) clones, 33 P1-derived artificial chromosome (PAC) clones and one P1 clone, segments of 8 yeast artificial chromosomes (YACs; subcloned into miniBACs or cosmids), 9 individual cosmids and 6 genomic PCR products spanned the long arm of chromosome 14 with no apparent gaps. A contiguous sequence assembly of this sequence-ready map was established and is presented in Fig. 1. On its centromeric end, the sequence contig begins in a highly conserved segmental duplication. This segmental duplication is highly homologous to a copy located in a similar position at the proximal end of the sequence of chromosome 22^{3,7}. The distal end terminates in a unique sequence, which was mapped 5–8 kilobases (kb) from the telomere on the basis of half-YAC yRM2006 (<http://www.wistar.upenn.edu/Riethman/>) and exonuclease Bal-31 digestion⁸. We therefore conclude that the euchromatic part of chromosome 14q has been entirely sequenced. The mean clone overlap in the tiling path is 29.9 kb (22.5% clone redundancy)

and the mean finished sequence overlap is 4.7 kb (3.6% sequence redundancy).

To identify inconsistencies, we checked continuity, ordering and integrity of the clone and sequence coverage along the whole sequence assembly at several levels of resolution. This validation process is based on two rationales. First, the presence in the sequence of a number of reference points, provided by the marker collections from various maps and, conversely, the absence of sequence-tagged sites (STSs) belonging to other chromosomes. Second, the correlation of the restriction map derived from the sequence assembly with the restriction fragment pattern observed in non-sequenced BACs covering almost the entire chromosome 14 sequence (see Methods).

We first checked for integration of all the markers from two genetic maps and from the HAPPY map⁹ (see below and Supplementary Information). We also compared the marker content of the sequence to a radiation hybrid map⁶ comprising a denser set of 2,350 markers built on the TNG whole-genome radiation hybrid panel¹⁰ and including most markers from this radiation hybrid map. In all instances where markers were apparently missing from the sequence assembly, their location on chromosome 14 could not be confirmed by *de novo* mapping on radiation hybrids or single chromosome hybrids, and most of such markers could be positioned elsewhere in the genome.

Collinearity between the fingerprint contig (FPC)-based physical map¹¹ and the sequence map was also verified (see Supplementary Methods). Finally, we developed a procedure to validate the local clone and sequence assemblies as well as the integrity of the clones selected for sequencing. In this procedure we compared the restriction maps (for four enzymes) deduced from the assembled sequence to the experimental restriction patterns derived from an alternative set of BAC clones that was independent of the sequence-ready map but covered the sequence (see Methods). Ninety-three per cent of the sequence assembly, corresponding to the fraction covered by the alternative set of BACs, could be validated using this procedure (Fig. 1). For regions weakly represented in the two major BAC libraries used and for which we were unable to identify overlapping clones (representing another 5%, see Fig. 1), the validation was limited to the sequenced BAC itself. We ascribed the remaining 2% to experimental failures, escape of some regions from the resolution limits of the restriction pattern analysis, the presence of bacterial transposons in some alternative BACs, or the occurrence of indels (insertion/deletion) or restriction site polymorphisms.

Care was taken to ensure the integrity of the sequence assembly by independent means. However, none of the validation processes used is absolute, particularly in resolving very recent duplications extending beyond the BAC unit, and we cannot formally exclude that a very peculiar genomic situation could escape our validation procedure.

Matching the sequence to chromosome maps

We compared positions of the Généthon microsatellite markers¹² between the sequence of chromosome 14 and the genetic linkage maps from Généthon and deCODE¹³ (Fig. 2). All the Généthon markers could be found on the final sequence assembly. Notably, a region showing an unusually high recombination rate (recombination 'jungle') in female meiosis occurs in a region of 0.7 Mb in the middle third of the genetic maps. The genetic distances of the female meiosis map appear more compact for the deCODE map, especially in the recombination jungle region.

As chromosome 14 is the only human chromosome that was mapped by the HAPPY mapping procedure, it was of particular interest to compare the HAPPY map⁹ to the sequence map (Fig. 3b). In the HAPPY mapping technique, large genomic DNA segments, generated by random breakage, are separated by limiting dilution and analysed for their marker content. Physically close markers tend to segregate together on the same segments. The frequency of

co-segregation provides a measure of distance and information on marker order that can be used to establish maps¹⁴.

The HAPPY map was established using a large set of markers (1,001), of which ten have been reassigned to other chromosomes (dots on ordinate of Fig. 3b). Except for ten other cases that appear as isolated dots, the distances between markers with conflicting order are below 1 Mb, and most frequently below 100 kb. Of note, 63% of the markers were positioned on the HAPPY map in the same order as they occur on the sequence. Such a coincidence obtained with a map showing a mean marker spacing of approximately 85 kb is outstanding and much better than that observed for the radiation hybrid maps. In many instances, orders at the 10-kb range were in agreement, demonstrating the high-resolution power of HAPPY mapping. A quasi-linear correlation between the HAPPY map and sequence distances can be seen for most of the chromosome (Fig. 3b). A more detailed analysis is provided in Supplementary Information.

The HAPPY markers of chromosome 14 were distributed between sequence positions 2,555,234 and 88,773,851, and ranged from no marker in a few intervals larger than 1 Mb, such as between positions 65,037,628 and 67,071,623, to a density of 47 markers per Mb between positions 55.1 and 56.1 Mb. As expected, the distribution was clearly non-uniform, with some marker gaps in the 1-Mb range scattered over the chromosome and which largely reflected the distribution of Alu repeats along the chromosome (Fig. 3a). A more detailed analysis is provided in Supplementary Information.

Sequence variations

Sequence overlaps between consecutive BACs have been analysed for sequence variations, focusing on the large insertion/deletion events (indels) affecting segments greater than 100 base pairs (bp) such as insertions of Alu elements that are stably inherited and provide powerful landmarks for population and human evolution studies. The cumulative size of overlapping segments analysed amounts to approximately 19.7 Mb, of which 12.8 Mb originated from different haplotypes, in which a total of 14 large indels were

Figure 1 Overview of the euchromatic portion of human chromosome 14 with the centromere on the left and the telomere on the right. The following sections appear in the order from top to bottom: (1) a cytogenetic map positioning the cytogenetic bands on the sequence map; (2) positions of the Généthon microsatellite markers; (3) tiling path of the sequenced genomic clones designated by their GenBank/EMBL/DBJ accession numbers (the length of the double bar beneath is proportional to the length of the clone sequences that were used for the construction of the final finished sequence). Bacterial artificial chromosomes (BACs) from the RPCI-11 library are in black; BACs from the CITD library are in green; genomic polymerase chain reaction (PCR) products are in red; and clones from other origins (see text) are in blue. (4) Experimental sequence validation by the program WATSON (see text) using non-sequenced BACs (green areas) or sequenced clones (yellow areas). Segmental duplications on chromosome 14 with (5) intrachromosomal segmental duplications (blue) and (6) interchromosomal segmental duplications (red)—the thinnest bars represent sequence matches between 1,000 and 8,000 bp. (7) The (G + C) content is expressed in per cent and is represented by a red curve along with a horizontal line representing the average (G + C) percentage (41%) of chromosome 14. The densities of long interspersed nucleotide elements (LINE) and short interspersed nucleotide elements (SINE) expressed by their cumulative length in an interval of 50 kb (%) are represented by an aquamarine and a black line, respectively. Exons are highlighted by filled green vertical bars, the heights of which correspond to cumulative exon length in an interval of 50 kb (%). The scale is in megabases (Mb); '0' corresponds to a position 1,550,000 bp upstream of the sequenced euchromatic start position (see Supplementary Methods). (8) The bottom part of the figure highlights the gene content of the sequence. The annotated gene categories, as described in the text, are represented by coloured bars. The width of the bars is proportional to the gene length. A horizontal strip separates upper- and lower-strand genes, and includes positions of the CpG islands, indicated as vertical bars. (9) The TCR and IGH loci are indicated by boxes in the last two lines.

detected. Among these, eight were Alu insertion polymorphisms (see Supplementary Information) observed in one of two sequenced haplotypes, indicating that Alu insertions account for the most abundant source of large insertion polymorphisms in the human genome. Assuming that chromosome 14 is representative of the human genome, an extrapolation of these events to the entire genome provides a first rough estimate of Alu insertion polymorphisms in the human genome of about 0.7 insertions per million base pairs when matching two haplotypes. Other large indels include a 60-kb segment and are described in Supplementary Information.

The chromosome 14 landscape

The chromosome 14 sequence shows a (G + C) content of 40.86%, which agrees almost exactly with the value (41%) reported for the entire genome. Genes cover a total of 38.069 Mb (43.6% of the total sequence), for a cumulated exon fraction of 2.021 Mb (2.3%) and a potential coding fraction of 1.000 Mb (1.1%). Total repetitive elements cover 40.373 Mb (46.2%), distributed essentially in short interspersed nucleotide elements (SINEs) (11.620 Mb) and large interspersed nucleotide elements (LINEs) (17.327 Mb), which corresponds to 13.3% and 19.8% of the whole sequence, respectively. These characteristics are compared with other finished chromosomes in Table 1. However, the chromosome 14 landscape is far from uniform (see Fig. 1): important fluctuations in (G + C) content occur, between 32.6% and 61.2% (for a window size of 50 kb), with high (G + C) content narrowly correlated with a dense CpG island distribution, a high SINE (versus low LINE) content and a high exon density. At a global level, variations in the (G + C) content occur in a discrete pattern, juxtaposing regions of very different base composition, such as isochore-like domains¹⁵. The gene distribution along the chromosome follows the same discontinuity, with a succession of gene-rich islands tightly superimposed over peaks of SINE elements (and also LINE depressions) and separated by very gene-poor areas; the five longest of them cover 2–6 Mb. Gene organization in gene-dense regions seems to occupy the chromosomal space available in an optimal manner, which is exemplified by a quasi-continuous succession of condensed gene structures often arranged in a forward and reverse alternance so that CpG islands in the region are often shared by genes of inverted polarity. These regions are often followed, with little or

no transition, by long stretches depleted in exons or occupied by very large genes with big introns. For example, the 2.7-Mb-long region between positions 55.2 Mb and 57.9 Mb, which shows a high exon density (5.6%, that is 20.7 genes per Mb), has a global (G + C) composition of 45.2%, with several peaks over 50%, and presents a SINE density of 29.3% versus a LINE density of 10.7%, values that are quite deviant from their respective averages.

A total of 1,768 CpG islands were found along the chromosome sequence, more than 530 of which appear to be related to the 5' end of genes. They are especially concentrated in the subcentromeric and subtelomeric regions (up to one CpG island per 9 kb, compared with the chromosome average of one per 50 kb). These two regions contain an olfactory receptor gene cluster, the α/δ TCR locus (subcentromeric region) and the IGH locus (subtelomeric region), and are rich in potential coding segments and in non-processed pseudogenes. On the contrary, the region between 65.1 Mb and 66.2 Mb, which corresponds to the biggest 'gene desert' (no annotated exon in 1.111 Mb, compared with 43.5 kb for the average intergenic region), shows a low (G + C) content (35.2%) and a weak SINE density (6.6%). This desert also seems to be depleted in CpG islands (one per 75 kb). Three gene deserts larger than 1 Mb and 15 larger than 500 kb cover almost 11.7 Mb (13.4%) of the chromosome 14 landscape.

An analysis of the synteny between human chromosome 14 and its mouse counterpart, at both genomic and gene levels, is reported in the Supplementary Information. The distribution (in position and orientation) of mouse genome segments (essentially from murine chromosomes 12 and 14) that are syntenic with human chromosome 14 are represented in Supplementary Fig. 2.

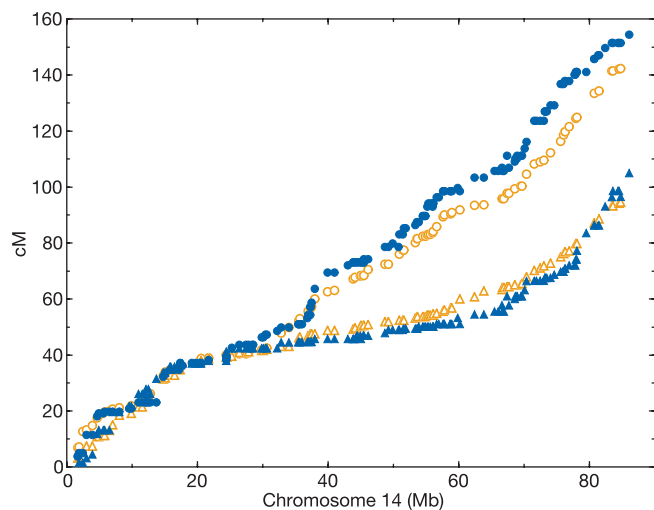


Figure 2 Comparison between genetic linkage maps and the chromosome 14 sequence. For genetic linkage maps, upper curves (circles) represent female recombination maps (filled circles, Généthon map; open circles, deCODE map), and lower curves (triangles) represent male recombination maps (filled triangles, Généthon map; open triangles, deCODE map).

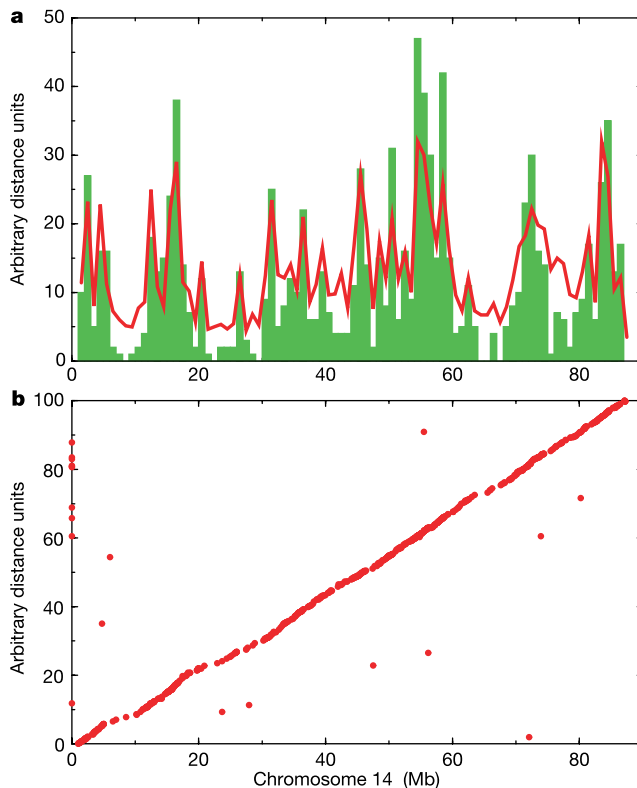


Figure 3 Comparison between the HAPPY map and the sequence of human chromosome 14. **a**, Distribution of the HAPPY map markers (histogram) and Alu elements (curve) per Mb along chromosome 14. **b**, Representation of the HAPPY map on the chromosome 14 sequence. Each dot represents the position of a marker on both the HAPPY map and the sequence assembly. Dots on the ordinate correspond to markers not found in the chromosome 14 sequence, but are found instead on other human chromosomes.

Table 1 Comparison of chromosome features

Genome property			Chromosome		Whole genome
	14	20	21	22	
(G + C) content (%)	40.9	44.1	40.9	47.8	41.0
Repetitive elements (%)	46.2	42.0	40.1	41.9	44.8
Gene coverage (%)	43.6	42.4	–	39.0	27.0
Exon coverage (%)	2.3	2.4	–	3.0	–
Gene density (genes per Mb)	10.0	12.2	6.7	16.3	9.3–10.8
Known genes mean size (kb)	58.7	51.3	57.0	–	27.0
All genes mean size (kb)	45.7	34.7	39.0	–	–
Pseudogenes (%)	26.0	18.8	20.8	19.7	–
Alternative splicing (%)	54.0	35	–	59	–
Putative genes (%)	25.0	16	(*)	(*)	–

Statistics on gene categories of sequenced chromosomes or whole-genome draft sequence¹. (*) Indicates that a different definition of putative genes was used for chromosomes 21 and 22.

Segmental duplications

An ancient duplication involving 70% of chromosome 14 and a portion of chromosome 2 has been reported². This event is, however, only visible at the protein level and predates the mouse–human separation. In our study we focused on more recent events that are detectable at the DNA level.

Using genome-wide sequence alignments (see Supplementary Methods), we found that 1.6% of chromosome 14 consists of interchromosomal segmental duplications contained in fragments of 1 kb or more that show at least 90% sequence identity. A comparable value, based on a different comparison procedure, was reported earlier¹⁶, and confirms that chromosome 14 has the lowest content of interchromosome segmental duplications in the human genome. These duplications are scattered along chromosome 14, with an increased density in the proximal third and some areas of high concentration in the subcentromeric and subtelomeric regions (Fig. 1), which contains members of the TCR and immunoglobulin superfamilies. The largest duplicated segment is shared with chromosome 22. As already shown¹⁷, these paralogous segments are adjacent to the centromeres of the two chromosomes and show a sequence identity above 98%.

Very divergent values for internal segmental duplications, ranging from 0.6%¹ to 4%¹⁶ have been reported previously. It should be noted, however, that an undetermined fraction of sequence redundancy characterized the early genome assemblies, which increases the predicted values of intrachromosomal segmental duplications. In a similar analysis that excluded repetitive DNA, we found that internal duplications account for 1.1% of chromosome 14 and were

clustered in four segments (Fig. 1). The largest includes an 800-kb region adjacent to the centromere, which is also part of the segmental duplication shared with chromosome 22.

RNA genes and mitochondrial DNA insertions

A total of 14 of the 20 transfer RNA genes identified on chromosome 14 using the tRNAscan-SE program are clustered in a segment of 75 kb (positions 2.947 to 3.021 Mb), which appears as the most dense cluster of transfer RNA genes in the human genome¹. Several small non-coding RNAs were localized along the chromosome sequence, including a complex tandem organization in an imprinted domain in 14q32. No complete ribosomal RNA-coding gene was found on the chromosome 14 long arm. Finally, 11 mitochondrial DNA insertions were characterized (see Supplementary Information for details).

Gene index of chromosome 14

Annotation of chromosome 14 results in 1,443 gene models that were classified into eight categories (see Table 2 and Methods): (1) 506 ‘known genes’, identical to human complementary DNA or protein sequences identified by a LocusID in the LocusLink database (<http://www.ncbi.nlm.nih.gov/LocusLink>); (2) 110 ‘novel genes’, having a coding sequence (CDS) and which are identical or homologous to spliced expressed sequence tags (ESTs) (vertebrates) or identical (but without a LocusID) or homologous to known cDNAs (vertebrates) and/or proteins (all species); (3) 11 ‘novel transcripts’, similar to novel genes, but for which a unique CDS could not be defined unambiguously; (4) 212 ‘putative genes’, homologous to cDNAs or spliced ESTs (vertebrates) but devoid of a significant CDS; (5) 11 ‘predicted genes’, based on *ab initio* predictions and for which at least one exon is supported by biological data (unspliced ESTs, protein sequence similarities with mouse or *Tetraodon* genomes or expression data from Rosetta); (6) 296 ‘pseudogenes’, sequences homologous to cDNAs or proteins (in at least 50% of the resource length) with a disrupted CDS and for which an active gene could generally be found at another locus. Finally, two specific categories were created for the annotation of the elements belonging to the TCR α/δ locus (positions 3.959 to 5.073 Mb) and the IGH locus (positions 87.685 to 88.954 Mb). These loci have an intrinsic germline organization correlated with inherent biological mechanisms essential for the generation of active genes (namely recombination events between ordered sets of segments), such that a classical exon/intron structure cannot be

Table 2 Distribution of genes among the different categories in human chromosome 14 annotations

Categories of genes	No. of genes	Gene length (nt)*	Transcript length (nt)*	No. of exons (exons per gene)	Exon length (nt)*	No. of alternative transcripts	CDS length (nt)*	CpG (–2,000 to 1,000 bp)
Known genes	506	58,748 (1,691,847)	3,080.3 (21,776)	5,206 (10.3)	299.4 (10,334)	287 (57%)	1,788 (20,658)	379 (75%)
Novel genes	110	35,818 (302,130)	1,901.3 (6,139)	765 (7.0)	273.4 (4,607)	51 (46%)	1,079 (5,616)	69 (63%)
Novel transcripts	11	13,316 (51,010)	2,326.4 (4,691)	32 (2.9)	799.7 (4,691)	1 (9%)	487 (954)	8 (73%)
Predicted genes	11	34,334 (170,900)	1,503.4 (4,227)	68 (6.2)	243.2 (2,770)	0	1,237 (2,217)	6 (55%)
Putative genes	212	22,005 (388,767)	732.0 (3,296)	600 (2.8)	258.6 (3,296)	23 (11%)	–	69 (33%)
All genes	850	45,713	2,311.9	6,671 (7.8)	294.6	362 (43%)	–	531 (62%)
Gene segments	200	581 (10,131)	280.4 (2,330)	–	–	6 (3%)	–	11 (5%)
Pseudogene segments	97	506 (5,462)	356.4 (2,870)	–	–	–	–	7 (7%)
Pseudogenes	296	1,431 (40,592)	1,244.5 (8,514)	–	–	–	–	–

Definitions of the categories are described in the text. CpG islands found in an interval spanning from 2,000 bp upstream to 1,000 bp downstream from the most 5' end of the annotated gene were scored. nt, nucleotides. CDS, coding sequence.

* Average gene-length values are shown, with longest gene-length values indicated in parentheses.

assigned for these genes. We therefore defined these structural elements as: (7) 200 'gene segments'; and (8) 97 'pseudogene segments' (when inactivated), and analysed these and the corresponding genomic regions independently. The gene index of human chromosome 14 is presented in Supplementary Table 4 and will be regularly updated at <http://www.genoscope.cns.fr/chr14/>. Gene names will be regularly reviewed by the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) (<http://www.gene.ucl.ac.uk/nomenclature/>).

Excluding pseudogenes and 'segments', chromosome 14 presents a mean gene density of 10 genes per Mb (850 genes per 85 Mb), which is close to the mean value estimated for the whole human genome (9.3–10.8 genes per Mb¹). Table 1 gives a comparison of the gene density between the four finished chromosomes. Pseudogenes accounted for 26% of the total, which is slightly higher than for the other finished chromosomes (Table 1). The mean gene length is 45.7 kb on the basis of the first five categories above, with significant variations between them (see Table 2 and an interchromosomal comparison in Table 1). Six known genes exceed 500 kb, whereas 104 genes (82 known, 13 novel, 8 putative and one predicted) are over 100 kb in length. The largest gene, the neurexin 3 (*NRXN3*) gene¹⁸, extends over 1,691,847 bp and the largest messenger RNA is 21,776 bp long and consists of 115 exons that encode SYNE-2, a protein of 6,649 amino acids expressed in synaptic nuclei at the neuromuscular junction¹⁹.

Among the 506 known genes, three CDSs use internal TGA triplet(s) to encode selenocysteine, namely the glutathione peroxidase 2 (*GPX2*) gene²⁰ and the genes for deiodinase iodothyronine types II and III (*DIO2* and *DIO3*)^{21,22}. Seventeen known genes have no CDS longer than 100 amino acids. These genes might encode a short peptide product. Alternatively, as hypothesized for the maternally expressed genes *MEG3* and *MEG8* from the imprinted domain in 14q32 (ref. 23), which is syntenic to the ovine callipyge locus, some genes that have a clear mosaic structure but do not present a significant CDS might act by means of a non-coding mRNA-like element²⁴ rather than a protein product.

Notably, about a quarter of annotated genes (212 out of 850) were classified as putative, showing an exon/intron structure but lacking a significant CDS, according to our criteria. They correspond in part to incomplete transcription units, annotated from very partial resources and for which experimental determination of the complete structure will be needed. As discussed above, some of them might also correspond to genes encoding short peptides or act by means of a non-coding mRNA-like transcript.

We found 362 (43% of 850) annotated genes that have one or more alternative transcript. The prevalence of alternative splice events in each gene category is given in Table 2. Excluding the putative and predicted genes, which are usually incomplete, alternative splicing occurs in 339 genes (54% of 627). This value is comparable to that reported for chromosome 22 (59%)¹ and is noticeably higher than that reported for chromosome 20 (35%)⁵ (Table 1). Indeed, owing to our conservative approach in annotation of alternative transcripts, in which we did not discriminate for the presence or absence of a CDS, we cannot exclude the possibility that a fraction of rare splicing events might be due to splicing machinery background²⁵.

The number of distinct transcripts averages 2.5 per gene for the

'known' category; however, several genes exhibit extensive alternative splicing potential (see Supplementary Information). The most notable is the large neurexin 3 gene, that contains a functional internal alternative promoter. At five sites of the long form, which are reduced to two sites in the shorter form, large numbers of possible splicing variants were postulated, resulting in a theoretical set of 1,728 combinatorial possibilities for alternative transcripts¹⁸. Because the exact function of neurexin 3 is unknown, the reasons for the large gene size and extensive alternative splicing remain a mystery. Another interesting example is given by the imprinted genes *MEG8* and *MEG3*, for which five and eight alternative transcripts have been described respectively, although none of them seems to correspond to a CDS²³.

From an analysis of 10,164 splice junctions, we observed that minor types of splice sites (different from the canonical GT-AG pattern) account for less than 1.47% (see Supplementary Information for details), with a relative prevalence of GC-AG (87 cases) and AT-AC (5 cases).

The 5' end of annotated genes

To evaluate the fraction of complete 5' ends in annotated genes after human curation, we investigated the presence of CpG islands in the 5' regions of genes (-2 kb, +1 kb), for each category and determined its correlation with the annotation of unambiguous 'start' codons. Our *in silico* definition of an initiator methionine was conservative, on the basis of the presence of upstream in-phase 'stop' codon(s) (see Methods). Using this procedure, unambiguous start codons were annotated for 358 known genes (73%), of which 266 (74%) have their 5' end in the vicinity of a CpG island. Considering previous estimations that 60–67% of genes are associated with a CpG island at their 5' end^{5,26}, we assumed that most of the 92 known genes (26%) with an unambiguous start codon but not linked to a CpG island at their annotated 5' end are nevertheless complete, or nearly so. Similarly, 77% of the 130 known genes for which the program did not find an unambiguous initiator Met are associated with a CpG island in their annotated 5' end. We therefore considered that most (if not all) genes from the known category and, to a lesser extent from the novel categories are essentially complete at their annotated 5' end, because of the close proximity of a CpG island in 75% and 64% of the cases, respectively. Assuming correct identification of the 5' end of genes associated with CpG islands, an initiator methionine could be defined for 100 additional known and 33 novel genes. The 5' end of the putative and predicted genes however, is currently too hypothetical to be associated significantly with CpG islands. On the basis of these data, we estimate that about 90% of the annotated genes (excluding putative genes) or about 70% (including putative ones) should be complete.

Contribution of comparative genomics resources

The process of annotation includes comparisons derived from partially sequenced genomes of *Mus musculus* ('BLAT-mouse'; see Methods) and *Tetraodon nigroviridis* (Exofish²⁷). No gene could be identified on the basis of a comparison with the deduced proteome of *Caenorhabditis elegans* and *Drosophila melanogaster* alone. On the other hand, mouse and fish genome sequence comparisons contributed significantly to the refinement of gene annotation on

Table 3 Specificity and sensitivity of the non-expressed resources in the annotation of human chromosome 14

	Exofish	BLAT-mouse	GENSCAN exons	FGENESH exons	Exofish + BLAT-mouse	Exofish + BLAT-mouse + GENSCAN exons + FGENESH exons
Annotation features	4,396	18,583	10,192	7,901	3,862	2,609
Specificity*	96%	31%	52%	61%	96%	99%
Sensitivity†	42%	70%	68%	63%	74%	83%

* The fraction of the annotation features of the various resources or a combination of resources that were included in the exons of the final gene annotation is indicated.

† The total number of annotated exons for all gene categories is 7,305 exons. The fraction of exons identified with the various resources or with a combination of resources is indicated.

human chromosome 14 (5' or 3' extensions, fusions of gene structures, putative alternative exons) and provided support for annotation as predicted genes for 11 *ab initio* determinations (an example of the contribution of comparative genomics data is given in Supplementary Information). Although the precise structure of such determinations needs to be experimentally confirmed, this annotation already constitutes a working model.

We tried to evaluate the contribution of non-expressed resources to the detection and refining of gene structures retrospectively (see Supplementary Tables 1 and 2). Table 3 shows the values of specificity and sensitivity for Exofish, BLAT-mouse, GENSCAN²⁸ and FGENESH²⁹, in several combinations. Although Exofish might be insufficient to identify all of the exons despite its high specificity, particularly when they belong to rapidly evolving genes, its sensitivity attains 82% for the detection of the RefSeq transcripts and 84% for the known genes. This last sensitivity value reaches 96% when both genomic resources are combined (see Supplementary Table 2), providing an estimation of the completeness of the gene identification.

Missing genes

Models resulting from the gene prediction algorithms (GENSCAN and FGENESH) and supported, even partially, by non-expressed resources only (Exofish and/or BLAT-mouse) were examined and, in 11 occurrences, annotated as predicted genes. The remaining *ab initio*-only models include 116 (out of 586) GENSCAN-only, 70 (out of 157) FGENESH-only and 58 (out of 990) combined (that is overlapping in at least one exon) models. Although we cannot exclude the possibility that some of these could pin-point a missing gene (especially for the combined type), we estimate that they essentially represent a background that is intrinsic to the prediction algorithms. Indeed, the mean probability for a GENSCAN-predicted exon is 0.80 inside annotated exons versus 0.40 outside. Such *ab initio*-only models were excluded from the present annotation.

On the other hand, a fraction of single ecores (evolutionarily conserved regions) (and, to a lesser extent owing to their lower specificity, of BLAT-mouse comparisons) that match outside annotated genes might correspond to exons of missing genes, particularly when conserved in all three species. To purge genomic DNA contamination of EST libraries we excluded unspliced ESTs devoid of CDS from annotation, as well. Such products could, however, correspond to some 5' or, most likely, 3' untranslated portions of actual transcripts. We are considering experimental testing of some of these products that are questionable individually, but for which the number of matching ESTs suggests biological relevance.

Finally, some genes that undergo rapid evolution, that are specific for minor cell types, or that show a very sharp pattern of expression or display an uncommon base composition might have escaped all current detection or prediction strategies.

Conclusion and medical implications

The chromosome 14 long-arm sequence was generated essentially by a sequence tag connector strategy (STC)⁶. This sequence covers, in a single 87,410,661-bp-long contig, the entire euchromatic portion of the chromosome, from the subcentromeric region to the telomere. We made a special effort to resolve cloning and sequencing gaps and to validate the whole assembly combining various mapping resources and experimental procedures. Comparisons to genomic or expressed sequence data from human and other organisms, combined with *in silico* gene predictions, allowed us to establish a gene index for human chromosome 14. This annotation, which represents the present state of knowledge, is available on our web site (<http://www.genoscope.cns.fr/chr14/>) and will be regularly updated, including results of experimental validation, especially those for the numerous putative genes. When extended to the whole genome, such a catalogue should be considered as an invaluable

informational basis for the resolution of the complex network of functions and interactions between genes and gene products.

Functional studies can also greatly benefit from mapping of biological activities and disease gene phenotypes that have accumulated during the past decades. We have therefore examined the OMIM entries localized to chromosome 14 for which no sequence data is yet available. This analysis (which is described in Supplementary Information) enabled us to clarify some mapping issues and to identify some alternative strategies for functional studies.

Two large gene clusters of immunological importance, namely the α/δ TCR and the immunoglobulin heavy chain, are located on chromosome 14. The complete structure and organization of these gene loci form the basis of our knowledge of the germline origins of TCR and immunoglobulin repertoires as well as the molecular pathology of lymphocyte malignancies that often involve chromosomal translocations within these gene loci. The availability of the complete sequence of chromosome 14 confirms that the α/δ TCR complex has already been exhaustively characterized. However, the sequence of the IGH constant region gene locus, now complete, will shed new light on the dissection of the molecular mechanisms of isotype switching and affinity maturation of immunoglobulin heavy-chain molecules.

Several disease gene identifications have benefited from the availability of draft or finished sequences from chromosome 14 BACs. These include autosomal dominant spastic paraplegia (*SPG3A*)³⁰, molybdenum cofactor deficiency³¹, oligodontia³², type I Leber congenital amaurosis³³, microphthalmia³⁴ and type C2 Niemann–Pick disease (*NPC2*)³⁵. Some 25–30 loci are not yet linked to sequences and represent future targets for positional cloning³⁶. Most correspond to mendelian disease phenotypes segregating in families. These diseases are quite diverse and include one locus for the very severe Usher syndrome (*USH1A*) and several other hearing and vision impairments. The identification of all these genes should greatly benefit from the complete sequence and thorough analyses that are now available. □

Methods

Sequence integrity assessment

The program WATSON (E. Pelletier, personal communication) constructs the chromosome sequence assembly from individual BAC sequences and deduces the restriction map from the master-sequence obtained; inventories all of the overlapping BACs from BAC-end sequence resources (see Supplementary Information) and proposes an alternative tiling path excluding clones selected for sequencing; reconstructs the clone sequence (including the specific vector in the correct orientation) and deduces a virtual restriction pattern for four selected enzymes; compares (in the 0.5–50 kb range) this virtual pattern to the experimental restriction pattern performed from BACs of this alternative tiling path; and scores the concordant results on a 0–4 scale (for non-redundant segments) or 0–8 (for redundant segments). Taking into account experimental biases, segments scoring 2 or more were considered as valid, in terms of assembly and of clone integrity. Discordant results were reanalysed after selection of other alternative overlapping BACs. This allowed resolution of experimental artefacts (owing to error in overlapping clone selection, non-integrity of overlapping BACs, namely partial deletions, bacterial insertion elements, and so on) and also pin-pointed the actual non-validated segments in the master sequence.

Annotation procedure for protein-coding genes

We developed a two-step annotation process. The first step consisted of *in silico* comparisons performed between the repeat-masked genomic sequence and expressed sequence resources (RNAs, ESTs and proteins) from human and other vertebrates, leading to the definition of preliminary gene models. Then, the structure of each gene model (including small exons and repeat-containing fractions) was refined by processing the SIM4 (ref. 37) alignment between the relevant transcribed sequences and the corresponding unmasked genomic segment (extended by 5 kb at both ends). In addition, comparative sequence analysis was performed against the genome (or proteome) of completely or partially sequenced organisms: *M. musculus* (BLAT³⁸ alignment coordinates extracted from the University of California, Santa Cruz genome annotation database and mapped on our chromosome 14 assembly); *T. nigroviridis* (Exofish²⁷); *D. melanogaster* (http://www.fruitfly.org/sequence/sequence_db/); and *C. elegans* (http://www.sanger.ac.uk/Projects/C_elegans/wormpep). Together with this homology analysis, *ab initio* gene predictions were performed using two algorithms: FGENESH²⁸ and GENSCAN²⁹. In addition, experimental data based on microarray expression measurements performed by Rosetta³⁹, which support GENSCAN predictions, were used. Finally, CpG islands were detected along the chromosome 14 sequence using GRAIL

(<http://compbio.ornl.gov/grailexp/>). In a second step, the preliminary gene structures defined automatically were submitted to manual curation, to incorporate additional data leading to the extension, fusion or splitting of gene models and to characterize alternative splicing events. Transcripts were summarized for each gene by their most complete representative (one for each form, in case of alternative splicing) and a 'proposed model' of that gene was constructed that included the additional data. Human expertise was also required to exclude from the final annotation suspicious data as unsplined ESTs and cDNAs or very partial matches, showing nonsignificant CDSs, to reduce the background owing to experimental artefacts (that is, genomic contaminants in cDNA libraries).

CDS determination

Exon sequences of a gene model were concatenated and start and stop triplets were identified. Each putative coding region extending from a start triplet (or the beginning of the first exon), to a stop triplet (or the end of the last exon) was selected if its length was at least 300 bp. The CDSs contained, in the same phase, in another CDS were eliminated and the proximal methionine was identified. If no stop codon was found in phase between this proximal methionine and the 5' extremity of the concatenated virtual cDNA, this extremity was considered as a provisional CDS starting point. In a last step, three kinds of scores were calculated on the remaining CDSs: (1) a score equal to the ratio of the length of individual CDSs versus the length of the longer one; (2) a score relative to the 5' position (the CDS starting at the 5' extremity of the gene model had a score equal to 1 and dropped to 0 when the CDS started at a 10,000-bp distance); (3) a score relative to coding exons (a CDS overlapping all the exons of a gene model received a score equal to 1). The presence of 1, 2, 3 and 4 non-coding exons in 5' ends (and of 1, 2 and 3 non-coding exons in 3' ends) decreased the score by 0.025, 0.075, 0.2 and 0.375 (and 0.075, 0.2 and 0.375) respectively. The mean of these three scores was the final score for a CDS. The CDSs with a final score greater than 0.6 were retained, and ultimately CDSs with final scores below 90% of the best CDS final score were rejected.

Received 20 August; accepted 3 December 2002; doi:10.1038/nature01348.
Published online 1 January 2003.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Dunham, I., Shimizu, N., Roe, B. A. & Chissov, S. The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
4. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
5. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).
6. Brüls, T. *et al.* A physical map of human chromosome 14. *Nature* **409**, 947–948 (2001).
7. Chen, C. & Birshstein, B. K. Virtually identical enhancers containing a segment of homology to murine 3' IgH-E(hs1,2) lie downstream of human IgCα1 and Cα2 genes. *J. Immunol.* **159**, 1310–1318 (1997).
8. Wintle, R. F., Nygaard, T. G., Herbrick, J. A., Kvaloy, K. & Cox, D. W. Genetic polymorphism and recombination in the subtelomeric region of chromosome 14q. *Genomics* **40**, 409–414 (1997).
9. Dear, P. H., Bankier, A. T. & Piper, M. B. A high-resolution metric HAPPY map of human chromosome 14. *Genomics* **48**, 232–241 (1998).
10. Olivier, M. *et al.* A high-resolution radiation hybrid map of the human genome draft sequence. *Science* **291**, 1298–1302 (2001).
11. The International Human Mapping Consortium. A physical map of the human genome. *Nature* **409**, 934–941 (2001).
12. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).
13. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
14. Dear, P. H. & Cook, P. R. Happy mapping: linkage mapping using a physical analogue of meiosis. *Nucleic Acids Res.* **21**, 13–20 (1993).
15. Clay, O. & Bernardi, G. Compositional heterogeneity within and among isochores in mammalian genomes. II. Some general comments. *Gene* **276**, 25–31 (2001).
16. 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).

17. Bailey, J. A. *et al.* Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. *Am. J. Hum. Genet.* **70**, 83–100 (2002).
18. Rowen, L. *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* **79**, 587–597 (2002).
19. Zhen, Y. Y., Libotte, T., Munck, M., Noegel, A. A. & Korenbaum, E. NUANCE, a giant protein connecting the nucleus and actin cytoskeleton. *J. Cell Sci.* **115**, 3207–3222 (2002).
20. Chu, F. F. *et al.* Polymorphism and chromosomal localization of the GI-form of human glutathione peroxidase (GPX2) on 14q24.1 by *in situ* hybridization. *Genomics* **32**, 272–276 (1996).
21. Araki, O. *et al.* Assignment of type II iodothyronine deiodinase gene (DIO2) to human chromosome band 14q24.2 → q24.3 by *in situ* hybridization. *Cytogenet. Cell Genet.* **84**, 73–74 (1999).
22. Hernandez, A., Park, J. P., Lyon, G. J., Mohandas, T. K. & St Germain, D. L. Localization of the type 3 iodothyronine deiodinase (DIO3) gene to human chromosome 14q32 and mouse chromosome 12F1. *Genomics* **53**, 119–121 (1998).
23. Charlier, C. *et al.* Human-ovine comparative sequencing of a 250-kb imprinted domain encompassing the callipyge (cplg) locus and identification of six imprinted transcripts: DLK1, DAT, GTL2, PEG11, antiPEG11, and MEG8. *Genome Res.* **11**, 850–862 (2001).
24. Erdmann, V. A., Szymanski, M., Hochberg, A., de Groot, N. & Barciszewski, J. Collection of mRNA-like non-coding RNAs. *Nucleic Acids Res.* **27**, 192–195 (1999).
25. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nature Genet.* **30**, 13–19 (2002).
26. Antequera, F. & Bird, C. CpG islands as genomic footprints of promoters that are associated with replication origins. *Curr. Biol.* **9**, R661–R667 (1999).
27. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
28. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
29. Salamov, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.* **10**, 516–522 (2000).
30. Zhao, X. *et al.* Mutations in a newly identified GTPase gene cause autosomal dominant hereditary spastic paraplegia. *Nature Genet.* **29**, 326–331 (2001).
31. Reiss, J. *et al.* A mutation in the gene for the neurotransmitter receptor-clustering protein gephyrin causes a novel form of molybdenum cofactor deficiency. *Am. J. Hum. Genet.* **68**, 208–213 (2001).
32. Stockton, D. W., Das, P., Goldenberg, M., D'Souza, R. N. & Patel, P. I. Mutation of PAX9 is associated with oligodontia. *Nature Genet.* **24**, 18–19 (2000).
33. Dryja, T. P. *et al.* Null RPGRIP1 alleles in patients with Leber congenital amaurosis. *Am. J. Hum. Genet.* **68**, 1295–1298 (2001).
34. Ferda Percin, E. *et al.* Human microphthalmia associated with mutations in the retinal homeobox gene CHX10. *Nature Genet.* **25**, 397–401 (2000).
35. Naureckiene, S. *et al.* Identification of HE1 as the second gene of Niemann-Pick C disease. *Science* **290**, 2298–2301 (2000).
36. Kamnasaran, D. & Cox, D. W. Current status of human chromosome 14. *J. Med. Genet.* **39**, 81–90 (2002).
37. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8**, 967–974 (1998).
38. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
39. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank H. Riethman, M. Meugnier, C. Sarlande, B. Baude and D. Le Paslier for their respective contributions. We also thank HUGO Gene Nomenclature Committee (H. Wain, R. Lavering, E. Bruford, M. Lush, M. Wright and S. Povey) for determining the chromosome 14 gene symbols.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H. (e-mail: heilig@genoscope.cns.fr). The entire chromosome 14q sequence is deposited in EMBL under accession number AL954800.

