The DNA sequence of human chromosome 7

LaDeana W. Hillier*, Robert S. Fulton*, Lucinda A. Fulton*, Tina A. Graves*, Kymberlie H. Pepin*, Caryn Wagner-McPherson*, Dan Layman*, Jason Maas*, Sara Jaeger*, Rebecca Walker*, Kristine Wylie*, Mandeep Sekhon*, Michael C. Becker*, Michelle D. O'Laughlin*, Mark E. Schaller*, Ginger A. Fewell*, Kimberly D. Delehaunty*, Tracie L. Miner*, William E. Nash*, Matt Cordes*, Hui Du*, Hui Sun*, Jennifer Edwards*, Holland Bradshaw-Cordum*, Johar Ali*, Stephanie Andrews*, Amber Isak*, Andrew VanBrunt*, Christine Nguyen*, Feiyu Du*, Betty Lamar*, Laura Courtney*, Joelle Kalicki*, Philip Ozersky*, Lauren Bielicki*, Kelsi Scott*, Andrea Holmes*, Richard Harkins*, Anthony Harris*, Cynthia Madsen Strong*, Shunfang Hou*, Chad Tomlinson*, Sara Dauphin-Kohlberg*, Amy Kozlowicz-Reilly*, Shawn Leonard*, Theresa Rohlfing*, Susan M. Rock*, Aye-Mon Tin-Wollam*, Amanda Abbott*, Patrick Minx*, Rachel Maupin*, Catrina Strowmatt*, Phil Latreille*, Nancy Miller*, Doug Johnson*, Jennifer Murray*, Jeffrey P. Woessner*, Michael C. Wendl*, Shiaw-Pyng Yang*, Brian R. Schultz*, John W. Wallis*, John Spieth*, Tamberlyn A. Bieri*, Joanne O. Nelson*, Nicolas Berkowicz*, Patricia E. Wohldmann*, Lisa L. Cook*, Matthew T. Hickenbotham*, James Eldred*, Donald Williams*, Joseph A. Bedell*, Elaine R. Mardis*, Sandra W. Clifton*, Stephanie L. Chissoe*, Marco A. Marra*††, Christopher Raymond†, Eric Haugen†, Will Gillett†, Yang Zhou†, Rose James†, Karen Phelps†, Shawn Iadanoto†, Kerry Bubb†, Elizabeth Simms†, Ruth Levy†, James Clendenning†, Rajinder Kaul†, W. James Kent‡, Terrence S. Furey‡, Robert A. Baertsch‡, Michael R. Brent§, Evan Keibler§, Paul Flicek§, Peer Bork||, Mikita Suyama||, Jeffrey A. Bailey¶, Matthew E. Portnoy#, David Torrents||, Asif T. Chinwalla*, Warren R. Gish*, Sean R. Eddy☆, John D. McPherson*††, Maynard V. Olson†, Evan E. Eichler¶, Eric D. Green#, Robert H. Waterston*†† & Richard K. Wilson*

* Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA
† University of Washington Genome Center, 2225 Fluke Hall on Mason Road, Campus Box 352 145 Seattle, Washington 98195, USA
‡ Center for Biomolecular Science and Engineering, University of California, 321 BE Santa Cruz, California 95064, USA
§ Department of Computer Science, Washington University, Box 1045, St Louis, Missouri 63130, USA
|| EMBL, Meyerhofstrasse 1, Heidelberg 69117, Germany
¶ Department of Genetics, Center for Computational Genomics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA
# Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 50, Room 5222, South Drive, Bethesda, Maryland 20892, USA
☆ Howard Hughes Medical Institute and Department of Genetics, Washington University School of Medicine, Campus Box 8232, 4566 Scott Ave. St Louis, Missouri 63110, USA

....

Human chromosome 7 has historically received prominent attention in the human genetics community, primarily related to the search for the cystic fibrosis gene and the frequent cytogenetic changes associated with various forms of cancer. Here we present more than 153 million base pairs representing 99.4% of the euchromatic sequence of chromosome 7, the first metacentric chromosome completed so far. The sequence has excellent concordance with previously established physical and genetic maps, and it exhibits an unusual amount of segmentally duplicated sequence (8.2%), with marked differences between the two arms. Our initial analyses have identified 1,150 protein-coding genes, 605 of which have been confirmed by complementary DNA sequences, and an additional 941 pseudogenes. Of genes confirmed by transcript sequences, some are polymorphic for mutations that disrupt the reading frame.

As the reference human genome sequence nears completion, the sequences of individual chromosomes are providing foundational information for genome structure, organization and evolution. Previously, chromosome 7 has been the focal point in the search for the gene associated with cystic fibrosis[1] and the frequent cytogenetic changes associated with some forms of cancer[2]. Here, we describe our analysis of the chromosome 7 sequence, which has exploited orthology with the mouse genome in refining gene predictions and has discovered some unusual structural features that have been implicated in genetic diseases. The unrestricted access of the community to this sequence as it was generated has aided in the discovery of genes on chromosome 7 related to human health and well being, and we anticipate that the full sequence will provide further impetus to these studies.

General features of the chromosome 7 sequence

We generated the sequence of human chromosome 7 using a clone-by-clone shotgun sequencing strategy[3,4] and organized the resulting sequence into 11 contigs (Fig. 1). With the exception of the centromere and one gap near the terminal end of the long arm (pter), the distances between contigs are relatively small (Table 1), with most sized by fluorescence *in situ* hybridization (FISH) of DNA fibres or by comparison to the analogous region in the mouse genome sequence. The DNA in these remaining gaps is repetitive or has proved recalcitrant to isolation in bacterial- or yeast-based cloning systems, including the screening of a series of large-insert genomic libraries that together provided 100-fold coverage of the human genome. On the basis of the size estimates of these gaps, the available sequence represents greater than 99.4% of the total euchromatic sequence.

Isolation of clones containing the telomeric and pericentromeric regions of the p arm were particularly hampered by the presence of repetitive sequences. A 7p-telomere-containing 'half-YAC'[5] (yeast artificial chromosome) has been identified, but has been too unstable to sequence. However, three cosmid clones estimated to

 157

Table 1 **Contiguous sequence lengths and gap sizes**

Map	Sequence length (Mb)*	Gap size by FISH (kb)	Gap size by mouse (kb)†	Accession (NT_)	Comment‡
Gap	–	–	–	–	~34-kb telomeric gap
1	0.3	–	–	029998.5	–
Gap	–	NA	334	–	–
2	47.4	–	–	007819.11	Extended by 70 kb
Gap	–	50	54	–	–
3	2.1	–	–	030008.5	–
Gap	–	40	45	–	–
4	6.5	–	–	033968.2 034884.1	Merged, ends in α-satellite
Gap	–	20–50	NA	–	–
5	0.8	–	–	023629.10	Centromeric, extended by 300 kb
Gap	–	NA	NA	–	–
6	0.2	–	–	023603.4	Unanchored centromeric (uc) clone, 190 kb uc clone added
Gap	–	NA	NA	–	–
7	12.7	–	–	007758.8 034886.2	Right end extends into WBS region§
Gap	–	NA	NA	–	–
8	64.4	–	–	007933.10	Left end in WBS
Gap	–	NA	87	–	–
9	14.8	–	–	007914.10	–
Gap	–	100	99	–	–
10	0.7	–	–	034885.1	–
Gap	–	80	80	–	–
11	3.9	–	–	007741.10 028233.7	29-kb gap closed

Values were created by FISH and estimation from orthologous mouse sequence. NA, not applicable.
*Status as of 1 July 2002, as reflected in NCBI build 31. These data were used for analyses presented here.
†Estimates of gap sizes based on mouse were obtained from A. Cook, M. Kamal and M. Zody (personal communication).
‡Includes any changes to the sequence since 1 July 2002 from continuing efforts to close all remaining gaps.
§Mouse genome lacks analogous region[9].

be 34 kilobases (kb) from the true telomere were recovered and contain scattered telomere-specific repeats. Such difficulties were not encountered with the 7q telomeres, where the sequence appears to extend into the telomere. On either side of the centromere, we analysed the sequence for higher-order alpha satellite repeats indicative of centromeric boundaries. On the short (q) arm side of the centromere, the sequence clearly contains such boundary features. The boundary was less clear for the p-arm side of the centromere, and the search for it was compounded by the duplication of this region elsewhere on the chromosome.

The quality of the chromosome 7 sequence exceeds the 99.99% accuracy standard established by the International Human Genome Consortium for sequencing the human genome[6]. We further checked the integrity of the sequence and its assembly in two ways. First, as each clone was finished, an *in silico* digest of the sequence was compared to restriction digests of the clone DNA. We also checked the fully assembled sequence by performing *in silico* digests of clone-sized fragments across the chromosome against the underlying fingerprint data used to construct the physical map. In this way, we confirmed more than 99.9% of the testable bands.

Comparison to physical and genetic maps

We evaluated the completeness of the chromosome 7 sequence by looking for its representation of sequence-tagged sites (STSs) from previously constructed physical and genetic maps of the chromosome, specifically a YAC-based STS-content map[7], the Genethon microsatellite-based genetic map[8], and a chromosome-7-specific radiation-hybrid (RH) map[7]. There were only a small number of unassigned STSs (see Supplementary Information), which included those from multi-copy sequences, sequence polymorphisms, regions within the remaining chromosome 7 sequence gaps, and clerical errors that preclude accurate matching of STS names to their true underlying sequence.

We also used these maps to evaluate the assembly of the sequence. The chromosome 7 sequence positions of the identified STSs were plotted relative to their established map positions (Fig. 2). Less than 1% of the identified STSs in the YAC-based map are in serious disagreement (defined as >3 megabases (Mb)) with their sequence position. In only one and four instances of the Genethon genetic and RH maps, respectively, are there discrepancies of more than several megabases between the map and sequence position of the STSs.

Together, these findings reveal an excellent overall concordance between the chromosome 7 sequence and previously constructed physical and genetic maps. The rare discrepancies can be largely accounted for by the inherent lower-resolution nature of the various mapping methods; however, it is also possible that some of the observed differences reflect polymorphisms between the different copies of chromosome 7 used for map construction and sequence generation. Nonetheless, these results, in conjunction with the robustness of the bacterial artificial chromosome (BAC)-based physical map used for sequence generation[3], provide strong support for the established chromosome 7 sequence.

Orthology to mouse

The relationship between the human chromosome 7 sequence and the mouse genome could be readily defined for approximately 92%

Figure 1 Overview of chromosome 7. Headings at each end of the figure indicate the following (from top to bottom): Cytogenetic map; Gaps, positions of gaps within the chromosome 7 sequence; Orthologous mouse, orthologous mouse regions; (G + C) content, (G + C) content (blue) in 20-kb windows (scale = 30–65%); Repeat density, SINES (red) and LINES (blue) in 50-kb windows (scale = 0–70%); Interchromosomal, duplicated interchromosomal regions (yellow); Intrachromosomal, duplicated intrachromosomal regions (red); Genethon markers (green); CpG islands (blue), 200-bp windows on repeat-masked sequence; RNA genes, non-coding RNAs (ncRNA genes in orange; ncRNA predicted pseudogenes in brown); Pseudogenes, 573 processed (purple), 81 non-processed (lime) and 287 unknown (light purple) pseudogenes; Known genes, known (blue) and predicted (red) genes and gene identifiers. ▶

of the chromosome, with 26 identifiable segments sharing the same order of highly conserved sequences in the two species at a resolution of 300 kb (ref. 9) or 46 segments at a resolution of 100 kb (Supplementary Methods). The smallest and largest defined segments are 200 kb and 38 megabases (Mb), respectively, with the latter residing on 7q and containing the cystic fibrosis gene, the ST7 tumour suppressor gene[10], olfactory and taste receptors[11,12], and the T-cell receptor beta gene region[13].

General features

We analysed the chromosome 7 sequence for interspersed repeat content, (G + C) content and the presence of CpG islands. The sequence has an overall repeat content (47%) and distribution of individual repeat classes (for example, short interspersed nucleotide elements (SINEs), long interspersed nucleotide elements (LINEs) and long terminal repeats (LTRs)) that differ only slightly from the whole-genome averages[4] (Supplementary S1). Similarly, its overall (G + C) content (41%) is almost identical to that of the genome as a whole. The regions of highest (G + C) content flank the remaining gaps, consistent with the idea that the (G + C)-rich regions may be difficult to clone. Analysis of the repeat-masked chromosome 7 sequence revealed 1,461 CpG islands. Of the known chromosome-7-derived messenger RNAs (see below), the 5' end of 66% were at or near (5-kb upstream to 1-kb downstream of) a CpG island. For the full gene set presented here, the number with overlapping CpG islands was 51%—the two values bracketing the reported figure of 60% for the genome as a whole[14]. Table 2 provides a comparison of some of the general features of chromosome 7 with the other published chromosomes.

Previous analyses of the draft human genome sequence suggested that Alu repeat distribution correlated more strongly with gene content than with (G + C) content[4]. We re-examined this issue by correlating (G + C) content, Alu repeat content and known exonic sequences (see below) across chromosome 7. Using a range (50–800 kb) of non-overlapping windows, we found strong positive correlations between (G + C) content, exonic sequence density and Alu content. However, for all window sizes, Alu elements were more strongly correlated with (G + C) content than with exonic sequence density ($R^2 = 0.66$ compared with $R^2 = 0.41$ for 200-kb windows).
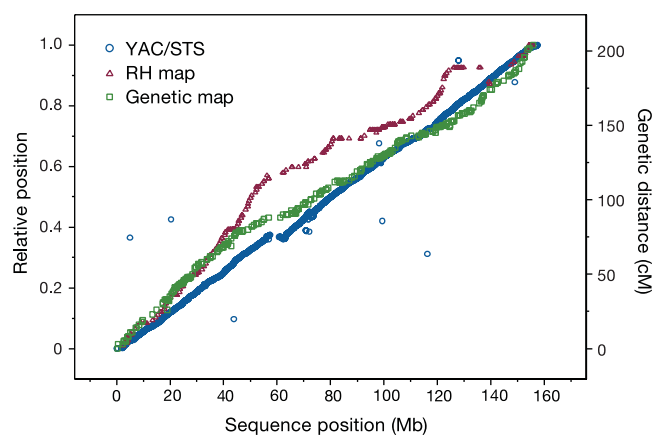
This difference from the overall genomic pattern suggests that the factors influencing Alu distribution are complex.

Known protein-coding genes

The identification of genes within genomic sequence typically uses known mRNA sequences[15], *ab initio* methods[16] and comparative nucleotide or protein sequence data[17,18]. Recognizing the challenges of gene annotation[19] we sought to establish a foundation for the gene catalogue of human chromosome 7 as an intermediate step en route to a full understanding of all the functional elements encoded by the chromosome. First, a set of 1,073 human mRNAs from REFSEQ[19] and the Mammalian Gene Collection[20] were uniquely assigned to the sequence and manually edited, resulting in 605 non-overlapping mRNAs with 45% showing alternative splicing. Only two (<0.3%) of the known genes mapped to chromosome 7 were not identified in the existing sequence.

Detailed examination of these mRNAs aligned against the genome revealed some potential artefacts, even within this experimentally supported set. For example, 23 mRNAs (2.4%) had no similarities to any mouse gene or to any known protein in the database. By contrast, less than 1% of predicted mouse genes have no homologue in the human genome[9]. Although some of these could be true genes, they may also represent untranslated segments of bona fide genes or transcripts of uncertain function. Nonetheless, these were retained in the current set. Additionally, eight other mRNAs had no significant open reading frame and were not included in the final gene set.

We also investigated 61 mRNAs where the matched genomic sequence contained differences from the mRNA that caused in some cases a frameshift and/or truncation of the protein product. To determine the origin of the difference, we re-sequenced the region of interest in a panel of 24 diverse individuals[21], in the starting BAC and in some cases in overlapping BACs. Ten cases could not be uniquely amplified because of surrounding repetitive sequence. In six cases the BAC sequence was found to be in error (representing either a simple sequence error or a mutation at the site during propagation), and the sequence was corrected. For another 35 cases, there was support only for the BAC sequence. Using expressed sequence tag (EST) data, underlying mRNA data and conservation with mouse, these 35 cases were determined to be mRNA error (primarily a deletion or insertion with a second compensatory insertion/deletion downstream). In the remaining ten cases, the site was found to be polymorphic, with support for both BAC and mRNA sequence. One of these polymorphisms is a deletion in zonadhesin (ZAN)[22], a sperm membrane protein that binds in a species-specific manner to the extracellular matrix (zona pellucida) of the egg. This deletion creates a frameshift at position 1,922 of this 2,812 amino acid protein.

Predicted protein-coding genes

To predict additional genes on chromosome 7, we applied three gene prediction programs. One, Genewise[17], uses protein homologies to

Figure 2 Comparisons of mapped positions of STSs and their locations within the chromosome 7 sequence. The position of each identified STS in the chromosome 7 sequence is plotted relative to its position on the chromosome 7 YAC-based map (blue), Genethon genetic map (green), and RH map (red). For the YAC/STS and RH maps, the left *y* axis reflects relative STS positions for those maps. For the Genethon genetic map, the right *y* axis reflects the actual genetic distances in cumulative centimorgans (cM). The apparent break at approximately 57–60 Mb reflects the position of the centromere.

Table 2 Comparison of human genome and published chromosomes

Category*	Chromosome					Human genome
	7	14	20	21	22	
(G + C) content (%)	41.0	40.9	44.1	40.9	47.8	41.0
Repetitive content (%)	45.0	46.2	42.0	40.1	41.9	44.8
Gene coverage (%)	36.5	43.6	42.4	31.0	50.0	27.0
Exon coverage (%)	1.4	2.3	2.4	–	5.0	1.5
Gene density (per Mb)	7.5	10.0	12.2	6.7	16.3	~10.0
Known gene mean size (kb)	61.5	58.7	51.3	57.0	–	27.0
All genes mean size (kb)	45.8	45.7	27.6	36.0	34.1	–
Pseudogenes (% of total genes)	45.0	26.0	18.8	20.8	25.0	–
Sequence length (Mb)	153.8	87.4	59.2	33.5	33.5	2,737.0

*Differences in gene features between chromosomes may in part reflect methodological differences: chromosome 14 (ref. 44), chromosome 20 (ref. 45), chromosome 21 (ref. 46), chromosome 22 (ref. 49), genome (ref. 19).

seed prediction, and the other two, Twinscan[18,23] and FGENESH2[16], use comparative sequence analysis. We used all available protein predictions[24] for Genewise and the mouse genome sequence[9] as informant sequence for Twinscan and FGENESH2. The combined output predicted 90% of all known exons and 98% of the known genes, indicating that for known genes the combined output is reasonably comprehensive with high sensitivity, albeit at the cost of specificity.

To reduce the number of false positives and pseudogenes in the collection, we demanded that the predicted genes have a highly significant match in the mouse gene set in the orthologous region of mouse where possible, and in turn that the matching mouse gene have among its best matches the original chromosome 7 predicted gene ('reciprocal match'). Furthermore, single-exon genes were removed from the collection if they had matches to multi-exon genes in either the human or mouse genomes. Redundancy between the three sets among themselves and with known genes was eliminated, accepting in order known genes, FGENESH2, Twinscan and Genewise predictions, with gene models with a reciprocal match taking precedence. Predictions showing signs of non-functionality (truncation or absence of introns) and those that produced L1/reverse transcriptase were also removed from the set. This yielded 545 predicted genes, bringing the total number of protein-coding genes on chromosome 7 to 1,150.

We next examined ESTs to look for genes that the above process may have missed. Of 41,399 spliced ESTs that had their best match to chromosome 7, 93% at least partially overlapped existing exons, and an additional 1% lay near or within existing genes and suggested alternative splice forms (all were represented in the redundant gene prediction set). The remainder lacked significant open reading frames, and none satisfied the reciprocal match criteria used for the gene predictions above (indeed only 5% had any match to mouse predicted genes). Although these unplaced spliced ESTs could represent protein-coding genes, there is currently little corroborative evidence that they do. More likely they represent other transcription products, including non-coding RNA genes or untranslated fragments of protein-coding genes.

By several criteria, the predicted gene set is robust. As expected given the methods used to establish it, 94% have a reciprocal best match (that is, the predicted gene has as its best match a mouse gene, which in turn has as its best match the starting human gene) with the mouse genome (known genes show a 92% best match). The remainder have a reciprocal match. High percentages of the predictions are supported by similarities to non-mammalian vertebrate genomes and by EST matches (Table 3a), but as expected the rates for each of these are not as high as seen for known genes. The known genes are enriched in highly expressed genes, which are more likely to be represented in EST sets and are more likely to be conserved across evolution. The predicted genes also compare favourably to known genes in coding exon number and in total coding sequence (Table 3b). The fact that both exon number and total coding sequence is smaller for the predicted genes suggests that either some terminal exons have been missed or that there are fragmented genes in the set that reduce average values. Finally, the pseudogene analysis carried out below shows that the set is remarkably free of likely pseudogenes, and that the set is not missing many genes (<60) with similarities to known proteins.

Pseudogenes

To identify pseudogenes on chromosome 7, we adapted an approach used for analysing the *Anopheles*[25] and mouse[9] genome sequences. This involved identifying sequence with significant similarity to known proteins in regions that reside between known or predicted genes, and then evaluating the ratio of non-synonymous to synonymous coding changes (the K_A/K_S ratio) for each potential coding sequence. Although not absolute, this ratio is an indicator of selective constraints associated with particular DNA regions, and can be used to assess differences between genes that evolve under purifying selection and pseudogenes that evolve in a neutral fashion[26]. Of the 941 such regions identified, nearly all (97 ± 3%) seem to evolve under neutrality and therefore are considered to contain pseudogenes. In contrast, only 5 ± 3% of the predicted and known genes have a K_A/K_S ratio consistent with neutral evolution. As the search did not attempt to identify highly diverged copies and may have merged two or more pseudogenes in the same interval, this probably corresponds to a lower limit for the number of pseudogenes. Indeed, there may be more pseudogenes than true genes on chromosome 7. As with the mouse genome sequence[9], a significant fraction (33%) of the identified pseudogenes contain neither stop codons nor frameshifts. Virtually all pseudogenes (94%) could be aligned to another region in the human genome with higher sequence identity than to any region in the mouse genome, suggesting that they originated after the human–mouse divergence.

Pseudogenes are generally formed by two independent mechanisms of duplication: retrotransposition (giving rise to processed pseudogenes) and segmental duplication (often leading to non-processed pseudogenes). In an attempt to classify the pseudogenes on chromosome 7, we exploited the available mouse sequence. First, in contrast to non-processed pseudogenes, processed pseudogenes integrate throughout the genome and are unlikely to have sequence similarity in orthologous mouse regions. Furthermore, pseudogenes that arose before the divergence of mouse and human are probably so diverged as to be below the thresholds of detection used here. For 654 pseudogenes within regions of chromosome 7 with identified mouse orthology, 573 (88%) appear to be processed, and 81 (12%) appear to be segmentally duplicated pseudogenes. Processed pseudogenes are broadly distributed along chromosome 7, with a slight tendency to cluster near the telomeres, whereas non-processed pseudogenes are concentrated in gene-rich regions (see Fig. 1). A minority of the former (100) has inserted into the introns of unrelated functional genes. The direction of these processed pseudogenes relative to the host intron varies almost randomly: 41 (41%) are integrated in the same strand as the host gene, whereas 59 (59%) are in the opposite strand.

Non-coding RNAs

We identified non-coding RNA (ncRNA) genes in the chromosome 7 sequence as described for the draft human genome sequence[19], followed by a refinement step using matches to mouse orthologous regions. *De novo* computational gene-finding methods for most non-coding RNA genes are not yet sufficiently robust for automated genome annotation[27]. Annotation was restricted to transfer RNA

Table 3 **Coverage and characteristics of exons and genes on chromosome 7**

a Coverage of predicted exons and genes by various data sets

	Predicted (%)		Known (%)		Ratio	
	Exons	Genes	Exons	Genes	Exons	Genes
Human ESTs	42	59	85	94	0.49	0.63
Non-mammalian*	20	38	36	52	0.56	0.73
Total	47	66	89	95	0.53	0.69
Pfam	–	45	–	69	–	0.65
Interpro	–	60	–	78	–	0.77

b Characteristics of predicted versus known genes

	Predicted	Known	Ratio
Exons per gene	6.7	9.5	0.71
Coding bases per gene (bp)	1,231	1,457	0.84
Genic bases per gene (bp)	28,474	61,439	0.44

*Translated non-mammalian vertebrate genomic sequence: *Fugu rubripes, Tetraodon nigroviridis, Gallus gallus*. Ratio is that of predicted to known.

genes, for which there is a robust identification programme, tRNAscan-SE[28], and to strong primary sequence similarities to known mammalian ncRNAs, including recently discovered human microRNAs[29].

Twenty-three tRNA genes were identified in the chromosome 7 sequence, including a cluster of 20 tRNAs with 18 tRNACys-GCA genes in one 400-kb interval. Eleven other ncRNA genes were found: four microRNAs, two U6 small nuclear RNA (snRNA) genes, and five genes for the four cytoplasmic Y RNAs (hY1, hY4, hY5 and two near-identical copies of hY3); all of the human Y RNA genes were already known to be located on chromosome 7 (ref. 30). Additionally, 302 putative pseudogenes were detected by BLAST similarity to other human ncRNA genes; this included 65 apparent U6 snRNA pseudogenes, 43 hY3 RNA pseudogenes and 37 non-Alu SRP-RNA pseudogenes.

Protein index

Using the above-generated gene set, we derived an index of predicted protein sequences. In turn, this was compared to the Interpro database[31] using Interproscan[32], which predicts protein families, domain and repeat families, and sequence motifs. The Interpro results were used to assign Gene Ontology (GO) codes[33]: 51% of proteins were in the category of molecular function, 31% in biological process and 18% in cellular components. Of the 68% of the proteins that had an Interpro classification, 56% were multi-domain. The two most prevalent families are also two of the most prevalent in the human genome: the immunoglobulins and zinc fingers.

The general homeobox domain proteins are heavily represented on chromosome 7, accounting for 21 of the 211 currently annotated records in the Ensembl gene set[34] of the human genome. In addition, the chromosome 7 gene set contains one-third of each of the more

specific homeobox proteins in the genome: antennapedia homeo-box proteins, engrailed-type homeobox proteins and homeo-domain protein CUT domains. The chromosome 7 HOX region, one of the four homeobox clusters in the human genome, contains ten two-exon genes (with one alternative form). This 90.7-kb region has a (G + C) content of 52% and is completely devoid of known interspersed repeats.

Williams syndrome critical region

Williams–Beuren syndrome (WBS) is associated with large (typically about 1.6 Mb) deletions[35] within 7q11.23. A series of large (>140 kb) duplicated segments (duplicons) span over 2 Mb in this region and are shown as coloured boxes in Fig. 3. The commonly deleted region is flanked by a duplicon that contains the p47-phox gene (or pseudogene), which has served as a useful marker for differentiating among the different duplicated segments. A third copy of the duplicon resides distal to the other pair within the commonly deleted region, in inverted orientation. At present there is a gap in the physical map and the chromosome sequence between the second and third copies of this duplicon.

The WBS region presented the single greatest challenge to the mapping and sequencing of the euchromatic regions of chromosome 7. Indeed, even after considerable effort, there remains some uncertainty about the location and orientation of some sequences. The duplicated segments approach the size of BACs, and the frequency of sequence differences among duplicons approaches (or is less than) the human polymorphism rate. To add to this complexity, the long-range organization of the region, perhaps including the number of duplicons, differs among individuals.

In attempting to establish a representative sequence of the WBS region, we used clones from a single BAC library to reduce the allelic complexity to just two variants. We also deliberately sequenced
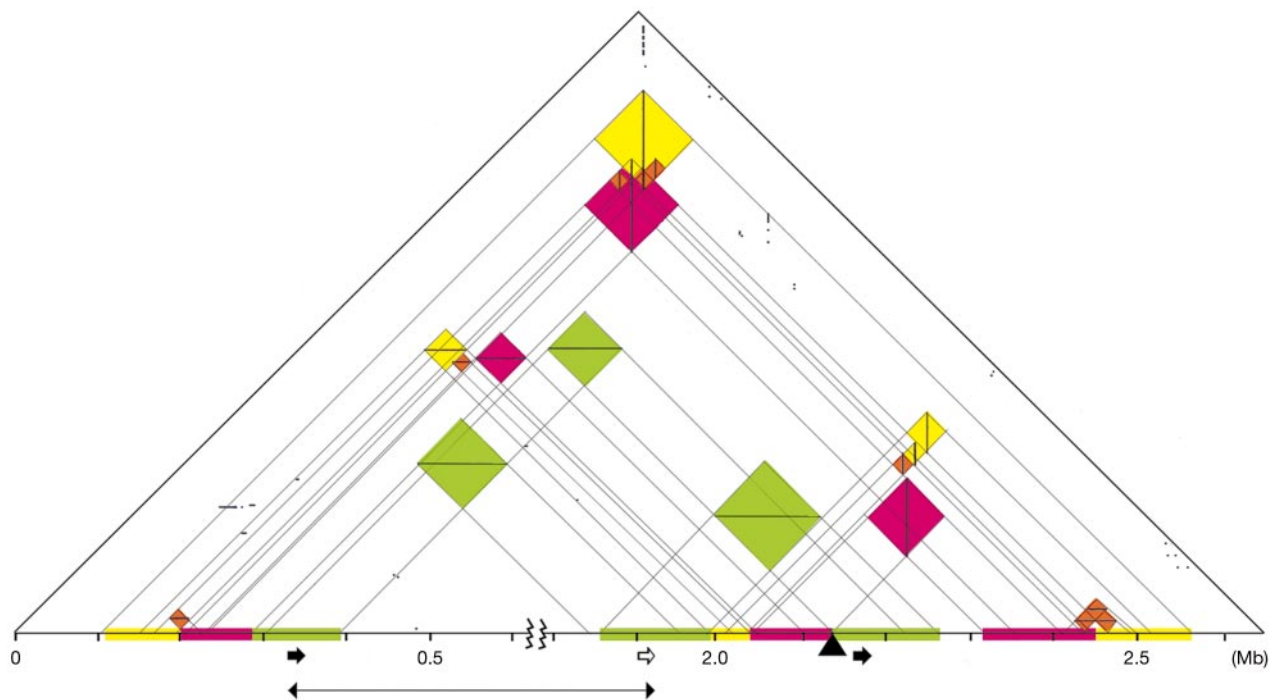

Figure 3 Repetitive content surrounding the commonly deleted region of the Williams–Beuren Syndrome area of 7q11.23 illustrated using self_dot_plot (H. Skaletsky and S. Rozen, personal communication). Each dot represents an identical match of 200 bp. Direct repeats are shown by horizontal lines, inverted repeats as vertical lines and palindromes as vertical lines that nearly intersect the baseline. Large sequence

duplications are indicated by coloured boxes. As indicated by the line break, a central 1.15 Mb of non-duplicated sequence has been omitted. The start of the region shown corresponds to base 70,945,566. Arrows mark the positions of the two phox genes (closed) and the phox pseudogene (open).

BACs with extensive overlaps to establish linkage of variant sites and used polymerase chain reaction (PCR) analysis of clones and a human population panel to distinguish polymorphic sites from differences between duplicated segments. This process eventually yielded the two sequence contigs shown in Fig. 3.

Segmental duplications

Segmental duplications are large low-copy repeats that arise as a consequence of duplication of genomic DNA and may range up to hundreds of kilobases in length. We performed a detailed analysis of duplicated sequence ($\geq$90% sequence identity and $\geq$1 kb in length; see Methods), comparing the chromosome 7 sequence against a recent assembly of the human genome. We identified a total of 3,215 pairwise alignments that met these criteria of length and sequence identity (Fig. 4a; see also Supplementary S3 and S4), making chromosome 7 one of the most duplicated human chromosomes. Overall, 8.2% (12,588 kb) of the sequence shares sequence homology to more than one location in the genome (Fig. 4a; see also

Supplementary S4). The enrichment is predominantly due to an increase in intrachromosomal duplications (7.0% of the sequence) rather than interchromosomal duplications (2.2%) with 0.5% (729,982 bp) sequence overlap between the two types. The spatial distribution of the interchromosomal and intrachromosomal duplications is clustered (Fig. 4a; see also Supplementary S3 and S4). As expected, large blocks of interchromosomal duplication locate preferentially within the pericentromeric and subtelomeric regions[36,37]. A marked asymmetry, however, was observed between the short and long arms of chromosome 7. The short arm of chromosome 7 has large blocks of recent interchromosomal duplications within both the pericentromeric (600 kb) and subtelomeric (about 150 kb) regions. The p arm subtelomere contains the most recent interchromosomal duplications (>99%). The low degree of sequence divergence suggests that such regions may have duplicated and/or undergone gene conversion since the separation of the human and chimpanzee lineages from a common ancestor.

In contrast to the short arm, the q arm pericentromeric region shows a much smaller block of duplication (<200 kb) followed by a large (about 600 kb) tract of monomeric alpha-satellite repeat sequence. These duplications are highly divergent (<93%). Although possibly a much more extensive domain of pericentromeric duplication remains to be sequenced closer to the centromere, the current configuration and the low degree of sequence similarity among these interchromosomal duplications suggest that the 7q pericentromeric region has been relatively quiescent over the last 25 million years of evolution. The subtelomeric region of chromosome 7q is even more striking in this regard. There is virtually no evidence of segmental duplication within a megabase of the telomere. The molecular mechanisms that underlie the difference in subtelomeric and pericentromeric duplication architecture between the two arms of chromosome 7 are unknown, but the asymmetry may represent a general property of metacentric chromosomes as has been suggested previously[36,38].

Comparison of the features of inter- and intrachromosomal duplications reveals some intriguing differences. In general, the intrachromosomal alignments tend to be larger than interchromosomal alignments (8 out of 10 pairwise alignments in excess of 100 kb were intrachromosomal duplications; Supplementary S5). This trend may, in part, be due to the fragmented nature of the draft sequence for other unfinished chromosomes. Furthermore, intrachromosomal duplications show an abundance of highly similar duplications (>98% identity), whereas most (57.1%) of the interchromosomally duplicated bases cluster between 93% and 96% identity (mode of 94.5–95.0%) (Fig. 4b; see also Supplementary S5). This interchromosomal mode is significantly different from the
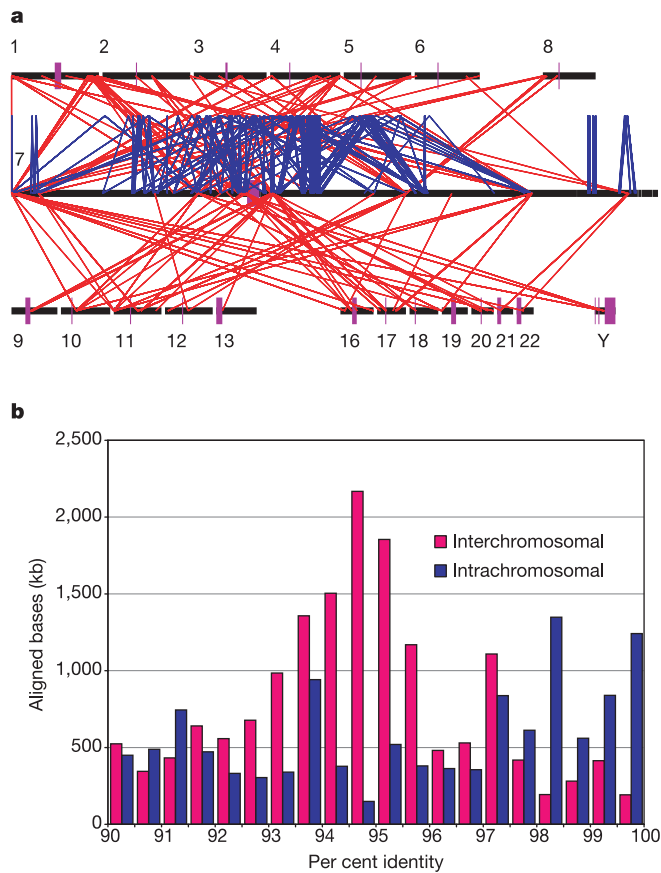

Figure 4 Recent segmental duplications on chromosome 7. **a**, Large (>10 kb) highly similar (>95%) intrachromosomal (blue) and interchromosomal (red) segmental duplications are shown for chromosome 7. Chromosome 7 is magnified in scale relative to the other chromosomes. The intrachromosomal duplications shown may be sites of unequal homologous recombination resulting in large-scale rearrangements such as those that cause WBS. The centromere is coloured purple. For more details regarding these regions see Supplementary S3–S6. **b**, Sequence similarity of segmental duplications. For all pairwise alignments, the total number of aligned bases was calculated and binned based on per cent sequence identity. Sequence identity distributions for interchromosomally (red) and intrachromosomally (blue) duplicated bases are shown. Assuming a clock-like rate for nuclear DNA substitution, per cent similarity is linear with evolutionary time providing a surrogate for the age of the duplication or gene conversion event.

Table 4 **Duplicated features for non-redundant transcripts**

Non-redundant source	Duplicated features				All features	
	Exons (%)		Transcripts (%)		Exons	Transcripts
mRNA-based transcripts	533	(8.7%)	121	(20.6%)	6,122	588
Gene	508	(8.6%)	117	(20.4%)	5,926	573
Pseudogene	25	(12.8%)	4	(26.7%)	196	15
Predicted transcripts	694	(22.8%)	136	(36.2%)	3,045	376
Gene	353	(15.3%)	64	(26.0%)	2,314	246
Pseudogene	341	(46.6%)	72	(55.4%)	731	130
Spliced EST clusters	202	(19.3%)	91	(24.6%)	1,049	370
Total	1,429	(14.0%)	348	(26.1%)	10,216	1,334

The non-redundant transcript set was screened for confirmatory transcriptional support by best genomic placement of ESTs. mRNAs required one supporting EST, and EST clusters and predicted transcripts required two ESTs and evidence of splicing (two or more exons). Thus, EST clusters of a single EST were removed. Additionally, each gene feature was binned as duplicated if at least 50 bp overlapped a duplicated region. Thus, exons less than 50 bp were lost from this analysis. Overall, 14.0% of all exons were binned as duplicated. Similar results were obtained requiring the entire exon to be encompassed by duplication (13.8% of all exons duplicated). Coding potential was only assessed as part of the analysis of mRNA-based and predicted transcripts. For these, the pseudogene fraction of duplicated exons (366/1,227 = 30%) was greater than for unique exons (927/7,940 = 11%).

average computed for the draft genome as well as other published chromosomes (96.0–98.0%)[39]. Using sequence divergence as an indicator of evolutionary age, the data suggest that chromosome 7 has been the target of more ancient interchromosomal duplication and/or gene conversion events. In contrast, more recent chromosome-specific duplication/gene conversion events[40] have occurred since the separation of the human and African ape lineages.

Segmental duplications are known sites of both pathological and evolutionary instability[41,42]. To identify regions on chromosome 7 that may be associated with genomic disorders, we searched for regions (between 50 kb and 10 Mb apart) that were flanked by large (≥10 kb), highly homologous (≥95%) segmental duplications (Supplementary S6). In addition to the WBS region, we detected ten regions of chromosome 7 (Supplementary S7 and S8), nine of which contained genes and would be considered candidates for genomic disorders. In total, these 11 regions (corresponding to 103 pairwise alignments) of chromosome 7 implicate almost a quarter (40.7 Mb) of the chromosome as being susceptible to duplication-mediated rearrangement. It will be important to investigate these regions experimentally for large-scale variation and association with disease.

Segmental duplications have long been noted for their potential role in the evolution of new genes[43]. To examine the transcriptional and coding potential of duplicated regions, we analysed a hierarchical, non-overlapping set of known genes, predicted genes and remaining spliced EST clusters (see above). For each group, we categorized every exon as unique or duplicated on the basis of its overlap with duplicated sequence (Table 4). Almost 14% (1,244 out of 9,890) of all exons are duplicated, and most of these lie within intrachromosomal duplicated sequence as opposed to interchromosomal duplicons. Our analysis shows that the relative number of transcribed exons is significantly greater for duplicated DNA when compared with non-duplicated DNA on chromosome 7. These results support a previous observation[39] that recently duplicated regions are rich in genes/transcripts. It should be noted, however, that many transcripts within the duplicated sequence have poor translational potential (44% compared with 20% of exons in pseudogenes for duplicated and unique regions), based on analysis of open reading frames (see pseudogene analysis above). Indeed, our analyses suggest that genes within duplicated regions show relaxed selective constraint when compared with genes encoded within unique portions of chromosome 7. Duplicate regions of chromosome 7 are, therefore, enriched for a particular class of pseudogene, which may be transcribed and may possess intron–exon structure but is unlikely to be translated. Most of the pseudogenes probably represent dying transcripts, which may on rare occasions lead to the formation of new genes.

Conclusions

The sequence of human chromosome 7 described here, and that of several other human chromosomes[44–47], represent landmark steps in the Human Genome Project. As chromosome sequences advance from their initial 'draft' status to a high-accuracy comprehensive stage, the molecular landscape becomes clearer, and the ability to perform detailed analyses becomes more robust. For our studies of chromosome 7, the combination of a high-quality, nearly complete sequence and a draft sequence of the mouse genome allowed us to perform rigorous gene analyses that included an improved ability to distinguish pseudogenes from bona fide genes. In addition to generating a gene index for the chromosome, our data provide evidence for a small subset of proteins that contain a polymorphism leading to a truncated protein in the human lineage. A chromosome-wide view of segmental duplications revealed that, compared with other chromosomes analysed so far, chromosome 7 exhibits a much higher rate of intrachromosomal duplication. Furthermore, there seems to be evolutionary asymmetry between the long and short arms. Taken together, these findings illustrate the dynamic

nature of a mammalian chromosome. Such dynamic behaviour also has adverse consequences, as revealed by the sequence features of the WBS region, where large duplicons of remarkably high sequence similarity mediate disease-causing deletions. Finally, the sequence we report here for chromosome 7 has directly facilitated the identification of a number of genes associated with human disease (for example, refs 10, 48). But these examples, although highly gratifying, simply represent the beginning of efforts to capitalize on the knowledge provided by finished genomic sequence for better understanding the genetic bases for human health and disease.   □

Methods

Tiling-path verification

To evaluate clone overlaps where the rate of difference between overlapping clone sequences was higher than 1 in 1,000 bases, a PCR product encompassing the differences was sequenced from each BAC in the overlap region and from each of a panel of 24 ethnically diverse genomic DNA samples[21]. If the 24 samples showed allelic variation, the overlap was judged to be correct, but if the 24 samples yielded persistent heterozygosity, the sequence was judged to be derived from a repeated sequence, with sequence differences between the copies.

Assaying mRNA/genomic discrepancies

To investigate discrepancies between mRNA and genomic sequence, PCR products were generated, re-sequenced and the polymorphic bases examined in the DNA from 24 individuals, the original BAC, and in some cases other BACs.

Pseudogene detection

After masking all predicted and known genes and common repeats, we performed homology searches by comparing the DNA sequence of chromosome 7 with a non-redundant protein database (see Supplementary Information). All regions matching non-viral and non-transposon known proteins (E value <0.001) were further processed by merging those likely to be parts of the same gene or pseudogene. This step, although likely to cause the loss of some real pseudogenes, is essential to eliminate possible fragmented predictions, which could otherwise lead to an overestimation. We refined the prediction for the resulting DNA fragments by comparing them with the closest protein sequence using Genewise[17]. We finally confirmed the integrity of the predictions by removing elements without significant matches (E values <0.001) in a second round of BLASTX against NRDB.

Each candidate pseudogene and all predicted and known genes were subjected to a K_A/K_S analysis. We first inferred the ancestral sequence of each of the target sequences (A) using a protein-based DNA multiple alignment of A and its two closest matches in NRDB between 50% and 95% identical. We next estimated the number of synonymous and non-synonymous substitutions occurring in sequence A by comparing it with its ancestral sequence using the YN00 program. In addition to the program's pre-set, we excluded those K_A/K_S ratios that were based on excessively low (<20) or high ($K_S > 1$) numbers of substitutions, resulting in reliable calculations for approximately 50% of analysed sequences.

The fraction of neutrally evolving sequences included in the pseudogene and gene sets was calculated comparing their K_A/K_S distributions with benchmark distributions for functional and pseudogenic elements. These distributions were obtained from the K_A/K_S analysis of 2,000 functional human genes randomly selected from a 50% non-redundant RefSeq[19] (reviewed) collection, and 1,730 processed pseudogenes with open reading frame truncations identified from a homology search through the whole human genome as performed for chromosome 7 (manuscript in preparation). Each of the K_A/K_S distributions associated to the pseudogene and gene sets of chromosome 7 was compared with benchmark distributions using a 'least squares fitting' to obtain estimates of neutrally evolving fractions of sequences. The error rate associated with this estimate is <3% according to cross-validation analysis with the benchmark sequences.

We compared each identified pseudogene at the protein level with all mouse genes (Ensembl[34]) located in the corresponding orthologous regions. Positive match was considered when the associated E-value was $<10^{-8}$. To avoid mis-annotations owing to the possible absence of decisive sequences in the mouse gene set, we also compared each translated pseudogene with the whole mouse orthologous region using tBLASTn with the same E-value cutoff as above.

1. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245,** 1066–1073 (1989).

2. Liang, H. *et al.* Molecular anatomy of chromosome 7q deletions in myeloid neoplasms: evidence for multiple critical loci. *Proc. Natl Acad. Sci. USA* **95,** 3781–3785 (1998).

3. Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).

4. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

5. Riethman, H. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409,** 948–951 (2001).

6. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from the Human Genome Project. *Genome Res.* **9,** 1–4 (1999).

7. Bouffard, G. G. *et al.* A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**, 673–692 (1997).

8. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152–154 (1996).

9. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

10. Zenklusen, J. C., Conti, C. J. & Green, E. D. Mutational and functional analyses reveal that ST7 is a highly conserved tumor-suppressor gene on human chromosome 7q31. *Nature Genet.* **27**, 392–398 (2001).

11. Nadezhdin, E. V. *et al.* Identification of paralogous HERV-K LTRs on human chromosomes 3, 4, 7 and 11 in regions containing clusters of olfactory receptor genes. *Mol. Genet. Genom.* **265**, 820–825 (2001).

12. Chandrashekar, J. *et al.* T2Rs function as bitter taste receptors. *Cell* **100**, 703–711 (2000).

13. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272**, 1755–1762 (1996).

14. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).

15. Wheelan, S. J., Church, D. M. & Ostell, J. M. Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.* **11**, 1952–1957 (2001).

16. Solovyev, V. V. *Statistical Approaches in Eukaryotic Gene Prediction* (ed. Balding, D.) (Wiley, New York, 2001).

17. Birney, E. & Durbin, R. Using GeneWise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547–548 (2000).

18. Korf, I., Flicek, P., Duan, D. & Brent, M. R. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, 140–148 (2001).

19. Pruitt, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137–140 (2001).

20. Mammalian Gene Collection Program Team Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA* **99**, 16899–16903 (2002).

21. The International Human SNP Working Group A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).

22. Hardy, D. M. & Garbers, D. L. A sperm membrane protein that binds in a species-specific manner to the egg extracellular matrix is homologous to von Willebrand factor. *J. Biol. Chem.* **270**, 26025–26028 (1995).

23. Flicek, P., Keibler, E., Hu, P., Korf, I. & Brent, M. R. Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Res.* **13**, 46–54 (2003).

24. Benson, D. A. *et al.* GenBankM. *Nucleic Acids Res.* **30**, 17–20 (2002).

25. Zdobnov, E. M. *et al.* Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**, 149–159 (2002).

26. Ohta, T. & Ina, Y. Variation in synonymous substitution rates among mammalian genes and the correlation between synonymous and non-synonymous divergences. *Mol. Evol.* **41**, 717–720 (1995).

27. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinform.* **2**, 18 (2002).

28. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

29. Mourelatos, Z. *et al.* miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728 (2002).

30. Maraia, R. J., Sasaki-Tozawa, N., Driscoll, C. T., Green, E. D. & Darlington, G. J. The human Y4 small cytoplasmic RNA gene is controlled by upstream elements and resides on chromosome 7 with all other hY scRNA genes. *Nucleic Acids Res.* **22**, 3045–3052 (1994).

31. Mulder, N. J. *et al.* InterPro: an integrated documentation resource for protein families, domains and functional sites. *Brief Bioinform.* **3**, 225–235 (2002).

32. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

33. The Gene Ontology Consortium Creating the Gene Ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).

34. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).

35. Korenberg, J. R. *et al.* Genome structure and cognitive map of Williams syndrome. *J. Cogn. Neurosci.* **12**, 89–107 (2000).

36. Horvath, J. E., Bailey, J. A., Locke, D. P. & Eichler, E. E. Lessons from the human genome: transitions between euchromatin and heterochromatin. *Hum. Mol. Genet.* **10**, 2215–2223 (2001).

37. Mefford, H. C. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* **3**, 91–102 (2002).

38. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9**, 1048–1058 (1999).

39. 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).

40. DeSilva, U., Massa, H., Trask, B. J. & Green, E. D. Comparative mapping of the region of human chromosome 7 deleted in Williams syndrome. *Genome Res.* **9**, 428–436 (1999).

41. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).

42. Stankiewicz, P. & Lupski, J. R. Genome architecture, rearrangements and genomic disorders. *Trends Genet.* **18**, 74–82 (2002).

43. Ohno, S., Wolf, U. & Atkin, N. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).

44. Heilig, R. *et al.* The DNA sequence and analysis of human chromosome 14. *Nature* **421**, 601–607 (2003).

45. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414**, 865–871 (2001).

46. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–318 (2000).

47. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).

48. Everett, L. A. *et al.* Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17**, 411–422 (1997).

49. Collins, J. E. *et al.* Reevaluating human gene annotation: A second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36 (2003).