LETTERS

DNA sequence and analysis of human chromosome 18

Chad Nusbaum[1], Michael C. Zody[1], Mark L. Borowsky[1], Michael Kamal[1], Chinnappa D. Kodira[1], Todd D. Taylor[2], Charles A. Whittaker[1]†, Jean L. Chang[1], Christina A. Cuomo[1], Ken Dewar[1]†, Michael G. FitzGerald[1], Xiaoping Yang[1], Amr Abouelleil[1], Nicole R. Allen[1], Scott Anderson[1], Toby Bloom[1], Boris Bugalter[1], Jonathan Butler[1], April Cook[1], David DeCaprio[1], Reinhard Engels[1], Manuel Garber[1], Andreas Gnirke[1], Nabil Hafez[1], Jennifer L. Hall[1], Catherine Hosage Norman[1], Takehiko Itoh[3], David B. Jaffe[1], Yoko Kuroki[2], Jessica Lehoczky[1]†, Annie Lui[1], Pendexter Macdonald[1], Evan Mauceli[1], Tarjei S. Mikkelsen[1], Jerome W. Naylor[1], Robert Nicol[1], Cindy Nguyen[1], Hideki Noguchi[2,4], Sinéad B. O'Leary[1], Bruno Piqani[1], Cherylyn L Smith[1], Jessica A. Talamas[1], Kerri Topham[1], Yasushi Totoki[2], Atsushi Toyoda[2], Hester M. Wain[5], Sarah K. Young[1], Qiandong Zeng[1], Andrew R. Zimmer[1], Asao Fujiyama[2,6], Masahira Hattori[2,7], Bruce W. Birren[1], Yoshiyuki Sakaki[2] & Eric S. Lander[1]

Chromosome 18 appears to have the lowest gene density of any human chromosome and is one of only three chromosomes for which trisomic individuals survive to term[1]. There are also a number of genetic disorders stemming from chromosome 18 trisomy and aneuploidy. Here we report the finished sequence and gene annotation of human chromosome 18, which will allow a better understanding of the normal and disease biology of this chromosome. Despite the low density of protein-coding genes on chromosome 18, we find that the proportion of non-protein-coding sequences evolutionarily conserved among mammals is close to the genome-wide average. Extending this analysis to the entire human genome, we find that the density of conserved non-protein-coding sequences is largely uncorrelated with gene density. This has important implications for the nature and roles of non-protein-coding sequence elements.

The International Human Genome Sequencing Consortium (IHGSC) recently completed a sequence of the human genome and published a report on the finishing of the human genome[2,3]. Now, papers containing detailed reports about each human chromosome are bringing to light aspects of the biomedical and evolutionary implications of this work. Here we describe the completion of a physical map, high-quality finished sequence, and gene catalogue for human chromosome 18, which represents approximately 2.7% of the human genome.

The extremely low density of protein-coding genes on chromosome 18 (Table 1) offers an opportunity to study the conservation of non-protein-coding sequences. It was recently observed that, in addition to protein-coding sequences, ~3% of the human genome shows a degree of evolutionary conservation among mammals that is significantly higher than background[4]. It is unclear whether this sequence consists mostly of regulatory elements related to genes or whether it represents other elements not tightly coupled to genes. These alternatives can be explored by comparing gene-rich and gene-poor chromosomes to see whether the proportion of conserved non-protein-coding sequence tends to scale with gene density or is unrelated to gene density.

The finished sequence of chromosome 18 contains 76,117,153 bases and is interrupted by three euchromatic gaps, one gap at the 18q telomere and one gap containing the centromeric heterochromatin (Fig. 1 and Supplementary Table S2). These gaps are refractory to current cloning and mapping technology. The sizes of the euchromatic gaps were estimated by alignment to the regions of conserved synteny in the mouse genome[4] (see Methods). The size of the telomeric gap was estimated using the size of the telomeric half-YAC (yeast artificial chromosome). The total size of these gaps is estimated at 118 kb. This corresponds to $<0.2\%$ of the euchromatic length of the chromosome, substantially lower than the average across the human genome (cited in ref. 3, also refs 5–7). Of the finished sequence, 79% was generated by the Broad Institute of MIT and Harvard (formerly the Whitehead Institute/MIT Center for Genome Research or WICGR), 20% by the RIKEN Genomic Sciences Center, and the remaining 1% by three other research groups (Supplementary Tables S3, S4). Details of construction of the clone map and sequencing are described in the Supplementary Information.

Several analyses verify that nearly the entire euchromatic region of chromosome 18 is present and accurately represented in the finished sequence. Of the 332 gene sequences in the well-curated RefSeq[8] data set that have been mapped to chromosome 18, all are present and complete in the finished sequence. In addition, the finished sequence shows excellent alignment to genetic and radiation hybrid maps (Supplementary Fig. S1). The genetic map[9] shows perfect alignment, with no discrepancies among 156 sequence-based genetic markers (Supplementary Table S5). The radiation hybrid map[10] shows good agreement, but contains local discrepancies as would be expected from its lower resolution (Supplementary Table S6).

We assessed the local accuracy of the clone path by aligning paired-end sequences from a human Fosmid library (designated WIBR2,

[1]Broad Institute of MIT and Harvard, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [2]RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [3]Mitsubishi Research Institute Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo 100-8141, Japan. [4]University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-0882, Japan. [5]HUGO Gene Nomenclature Committee, The Galton Laboratory, Department of Biology, University College London, Wolfson House, 4 Stephenson Way, London NW1 2HE, UK. [6]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan. [7]Kitasato Institute for Life Sciences, Kitasato University 1-15-1, Kitasato, Sagamihara, Kanagawa 228-8555, Japan. †Present addresses: MIT Center for Cancer Research, 77 Mass Avenue E18-570, Cambridge, Massachusetts 02139, USA (C.A.W.). McGill University and Genome Quebec Innovation Centre, Montreal, Quebec H3A 1A4, Canada (K.D.). Department of Human Genetics, University of Michigan Medical School, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109, USA (J.L.).

Table 1 | Chromosome 18 gene content

	Gene no.	Gene %	Gene length (bp) *	No. of alternative transcripts	Transcript length (bp)†	No. of exons per transcript‡	Internal exon length (bp)§	Intron length (bp)‖	CpG-5′ association¶
Known genes	243	72	88,523	3.1	3,121	10.7	155 ($n = 2,351$)	9,068 ($n = 2,997$)	73
Novel transcripts	10	6	44,778	1.2	870	4.6	146 ($n = 68$)	7,131 ($n = 101$)	33
Putative genes	11	3	10,427	1.0	560	2.2	97 ($n = 4$)	6,425 ($n = 18$)	36
Novel CDS	49	15	58,294	1.9	987	4.4	145 ($n = 217$)	12,150 ($n = 288$)	0
Gene fragments	13	4	2,095	1.0	2,027	1.0			0
PredictedPlus	11	3	57,060	1.0	1,008	5.6	137 ($n = 40$)	12,576 ($n = 49$)	18
Total	337		75,519						
Pseudogenes	171	51	3,601	1.0	837	1.9	178 ($n = 105$)	2,971 ($n = 159$)	7

* Average chromosomal distance from start of 5′-most exon to 3′-most exon for all transcripts of a gene.
† Average length summed across the footprint of all exons for all transcripts of a gene—total exon space per gene.
‡ Average number of exons in transcripts. Exons common to different transcripts were counted once per transcript.
§ Average length of exons using the footprint of all non-terminal exons for all transcripts of a gene. Unique overlapping exons or contained exons are counted separately, making this an average length of unique exons in a gene.
‖ Average length of unique introns in a gene. In the case of exon skipping, both the shorter and longer overlapping introns were counted towards the average.
¶ Percentage of genes with a transcript having a CpG island (as assessed by FirstEF) within −2 kb and +1 kb of transcription start.

representing 10× physical coverage) to the finished sequence[3]. By identifying discrepancies in the distances between Fosmid ends in the finished sequence and those expected on the basis of insert size constraints, one can detect errors in the clone path[3]. Our analysis revealed a single aberrant region, which was found to result from a bacterial artificial chromosome (BAC) clone containing a 21-kb deletion that was either present in the source genome or occurred in the cloning of the BAC; this clone was replaced with a non-deleted BAC from a different library. Finally, an independent quality assessment exercise commissioned by NHGRI estimated the accuracy of the finished sequence at less than one error per 100,000 bases[11] (J. Schmutz, personal communication).

We produced a manually curated catalogue of genes (see Methods), annotating 337 gene loci and 171 pseudogene loci on chromosome 18. These include all previously known genes on chromosome 18 (Table 1). According to the Hawk2 categorization scheme (http://www.sanger.ac.uk/Info/workshops/hawk2, see Supplementary Information) there are 243 'known' genes, 49 'novel CDS' (coding sequence of a gene), 10 'novel transcripts', 11 'putative genes', 11 'predictedplus genes' and 13 'gene fragments'. All 'novel transcript' genes had expressed-sequence-tag (EST) evidence. For 'putative genes', only a subset of the exons were supported by one or more spliced ESTs. Only a small fraction of all loci, those in the 'novel' and 'putative' categories, were annotated as genes on the basis of spliced EST evidence only. Some 'gene fragment' loci may prove to be pseudogenes.

Using aligned EST evidence, it was possible to extend many of the previously known gene models at their 5′ or 3′ ends (see Supplementary Fig. S2 for an example). Approximately 57% of the RefSeq and mammalian gene collection (MGC) transcripts could be extended. The 5′ end extensions averaged 321 bp, and 3′ end extensions averaged 1,131 bp. In addition, a novel 5′ exon was found for 14% of the RefSeq or MGC transcripts, and a novel 3′ exon was found for 2.2%. The ability to extend the gene models probably reflects expanded databases of transcripts and ESTs. A sampling of the extended gene models was validated in the laboratory (see Supplementary Information).

We found an average of 10.7 exons per full-length known

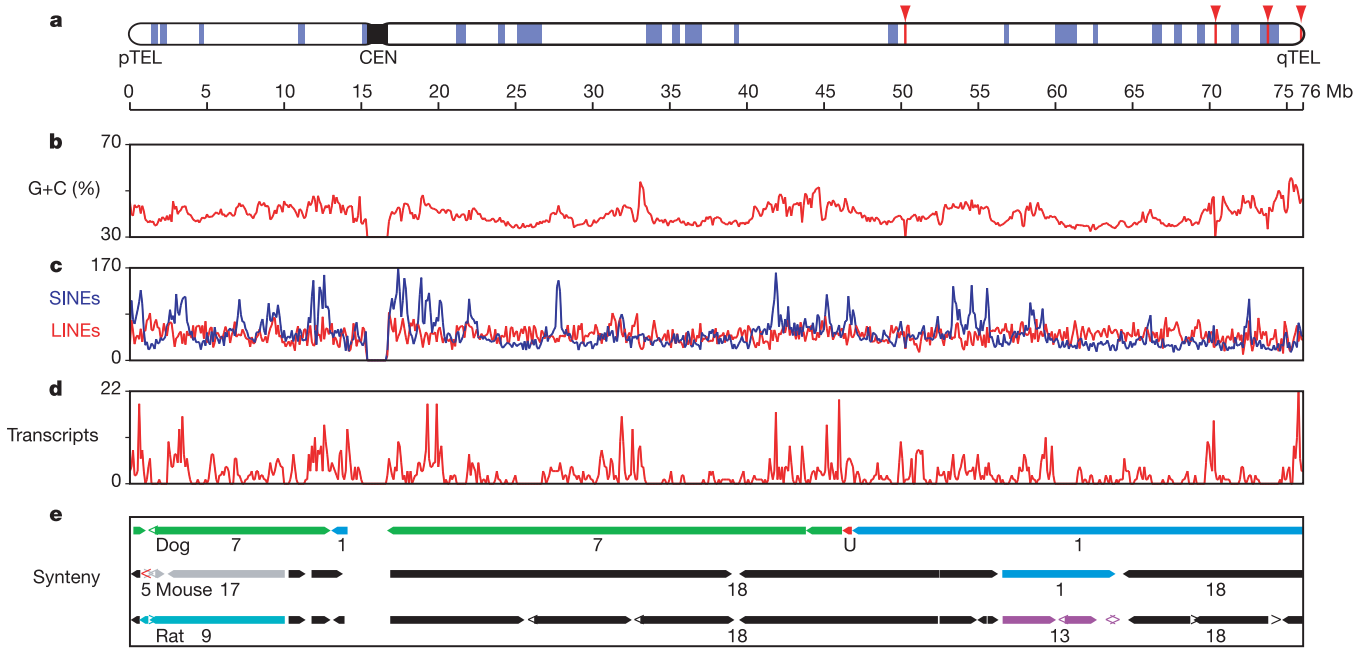

Figure 1 | Overview of human chromosome 18. **a**, Blue shading indicates gene deserts (≥500 kb with no transcript, see Supplementary Table S8). Telomeres (pTEL and qTEL), the centromere (CEN) and euchromatic sequence gaps (red lines) are also indicated. **b**, G+C content in discrete windows of 100 kb. **c**, **d**, Densities of long interspersed nuclear elements (LINEs, red), short interspersed nuclear elements (SINEs, blue) (**c**) and transcripts (**d**) are shown as numbers of these elements in discrete windows of 100 kb. **e**, Blocks of conserved synteny (100-kb resolution) with dog, mouse and rat, determined for this work. Chromosomes are numbered, and are coloured arbitrarily for ease of distinction.

transcript, comparable to recent published reports of human chromosomes. Internal exon lengths average 155 bp, and the average transcript length is 3.1 kb for full-length transcripts of known genes. There is evidence of extensive alternative splicing, with gene loci having an average of 3.1 distinct transcripts and 71% having at least two transcripts. This rate of alternative splicing is comparable to recent reports[5,6].

The longest gene on chromosome 18 is *DCC* (deleted in colorectal carcinoma), spanning 1,190,632 bp. *DCC* also contains the longest intron at 411,177 bp. The longest mature transcript is laminin α3 (*LAMA3*) at 10,585 bp. The longest single exon is found in *TCF4*, being a 3′ exon of 5,700 bp. The gene with the most identified splice forms is *TGIF* (TGFβ-induced factor), which appears to have ten splice forms, of which two are represented by RefSeq transcripts. Of the 171 pseudogenes on chromosome 18, approximately two-thirds are processed (intronless) pseudogenes arising from retroposition, and the remaining one-third are unprocessed. In addition, we identified four transfer RNA genes on the chromosome, listed in Supplementary Table S7. An analysis of gene families revealed that several families have multiple members present on chromosome 18. These include members of the laminin and cadherin families of cell adhesion molecules, and a cluster of ten serpin serine protease inhibitors (see Supplementary Information). Careful analysis of gene models found 59 pairs of overlapping genes on chromosome 18, suggesting that overlapping genes may be 2–4 times more common than previously thought[12,13] (see Supplementary Information).

With an average of 4.4 genes per megabase (Mb), chromosome 18 has the lowest gene density of published human chromosomes (Supplementary Table S1). This gene density cannot be explained by chance fluctuation around a genome-wide mean ($P < 10^{-12}$, see Supplementary Information). The low gene density is reflected both in the low percentage of transcribed sequence (28.5%) and the small fraction of the chromosome included in exons (1.14% in all exons, 1.06% in coding exons). The G+C content (39.8%) is also low, consistent with the known positive correlation between G+C content and gene number[14].

Chromosome 18 contains 24 gene deserts (defined as a 500-kb region without a coding gene, Supplementary Table S8), which together comprise 28 Mb or ~38% of the total chromosome length. The sparsest region of the chromosome harbours only three genes over 4.5 Mb. In addition, chromosome 18 also has the longest median length of introns among all chromosomes, reflecting a genome-wide inverse correlation between intron size and gene density (Supplementary Fig. S3).

Despite being gene-poor, chromosome 18 is not enriched in repeat sequences. Transposable element fossils cover 43.5% of the chromosome, which is typical across the genome. Chromosome 18 also has a relatively low proportion of segmental duplication (segmental duplications are defined as having greater than 90% identity and being longer than 1 kb). Segmental duplications constitute ~2.5% (1.92 Mb) of the chromosome, with a greater representation of interchromosomal duplications (2.13%) than intrachromosomal duplications (0.55%). Some sequences are represented in both types of duplication (E. Eichler and X. She, personal communication).

The paucity of genes on chromosome 18 probably explains why it is one of only three autosomes (the others being chromosomes 13 and 21) for which trisomic individuals routinely survive to term[1] (www.trisomy.org, www.ndss.org). Although chromosomes 18 and 21 have roughly the same number of RefSeq genes (332 and 374 genes, respectively), chromosome 18 trisomy (Edwards syndrome) has much more severe health effects than chromosome 21 trisomy

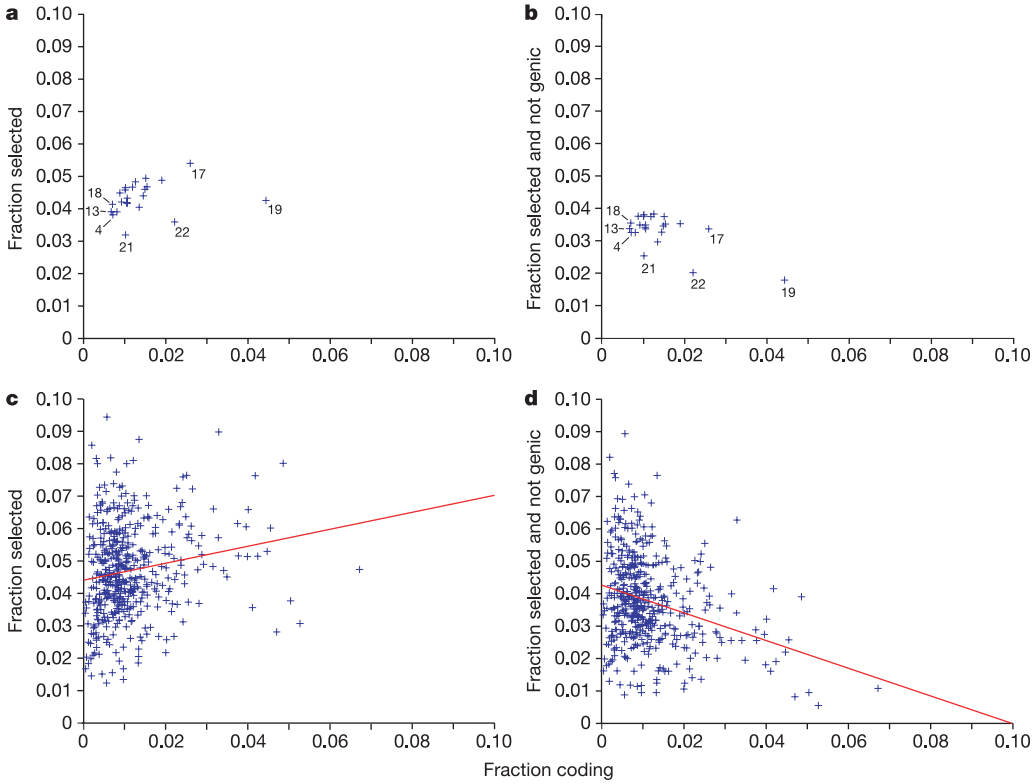

Figure 2 | Scatter plots showing the fraction of syntenic region under selection plotted against the fraction of coding sequence in that region. **a**, By chromosome, the fraction of all sequence under selection versus the coding fraction. **b**, By chromosome, the fraction of all non-protein-coding sequence under selection versus the coding fraction. Numbers refer to specific chromosomes. **c**, The fraction of all sequence under selection within the region versus the coding fraction within the region. **d**, The fraction of all non-protein-coding sequence under selection versus the coding fraction. In **c** and **d**, each point represents a 5-Mb region from a set of non-overlapping 5-Mb regions covering the genome. Lines of regression are shown. We define non-protein-coding sequence as that which is completely disjoint from any predicted mature mRNA product of an annotated protein-coding gene.

(Down syndrome). Edwards syndrome occurs in 1 in 5,000 live births, and ~90% of affected individuals die before one year of age. In constrast, Down syndrome is more common (1 in 800 live births), and affected individuals are frequently able to cope with the numerous health consequences and survive to adulthood. The availability of gene catalogues for these two chromosomes will facilitate work to elucidate how the contributions of specific genes lead to such different clinical outcomes.

Four other syndromes are caused by gross abnormalities in chromosome 18, including three partial monosomies caused by deletion of part of the p or q arms (18p-, 18q- and ring18) and tetrasomy of the p arm (www.chromosome18.org). The gene catalogue presented here should facilitate identification of the critical genes associated with each syndrome.

At least 45 loci on chromosome 18 have been implicated in genetic disorders[15] (Supplementary Table S9). The list includes at least four disorders for which the responsible gene and molecular mechanism of disease have been characterized (Supplementary Table S9). For two such diseases (methemoglobinaemia and erythropoietic protoporphyria), we found evidence for novel alternative splice forms that would result in coding sequence alterations (not shown).

Comparative gene analysis revealed one locus that may represent a newly evolved gene in the primate lineage, although its function is unknown. Among the annotated multi-exon genes contained in blocks of conserved synteny among mammals, only one lacks exonic conservation with rodents and dog: *C18orf2*, a predicted RefSeq gene. Within this block of conserved synteny there is a primate-specific ~100-kb inversion in the region (present in both human and chimpanzee). One of the endpoints of this inversion lies in the middle of the coding region of the gene, with the result that the region is not contiguous in either dog or rodent genomes. Partial sequencing of this gene in apes suggests that it is conserved at least as far back as orangutan (see Supplementary Information).

We compared chromosome 18 to its homologue chimpanzee chromosome 18 (ref. 16). The average sequence divergence is 1.25%, which is close to the genome-wide average. On a larger scale, the karyotype of human chromosome 18 differs from its homologues in the great apes by a human-specific pericentric inversion with an associated human-specific inverted duplication of 19 kb (refs 17, 18). As a consequence, human 18p corresponds to the proximal region of chimpanzee 18q. As large-scale chromosomal rearrangements can facilitate speciation[19,20], it is possible that this inversion had had a role in hominid evolution.

Finally, we sought to explore the still-mysterious nature of conserved non-protein-coding sequences. Recent comparison of the human and mouse genomes[4] led to the surprising discovery that ~5% of the human genome shows evolutionary conservation higher than the background rate (defined as the rate seen in ancestral repeat elements, which are presumed to be non-functional). Similar results have been seen in comparisons between the human and rat genomes[21]. As only 1–2% of the human genome encodes protein-coding exons, this indicates that the majority of human sequence under purifying selection is non-protein-coding. In principle, these non-protein-coding sequences could be (1) associated with protein-coding genes, such as those that directly or indirectly regulate the expression of protein-coding genes, or (2) independent of protein-coding genes, such as those that play a structural role in chromosome architecture or those that encode RNA genes.

We calculated the overall proportion of bases on each chromosome that are under purifying selection, and allocated this proportion as either protein-coding or non-protein-coding (see Methods). The computational analysis closely followed that used in recent mammalian comparisons[4,22] (see Methods). We compared the proportion of total sequence under selection (Fig. 2a) and non-protein-coding sequence under selection (Fig. 2b) to the proportion of coding sequence for each human chromosome. Chromosome 18 contains a low overall proportion of sequence under selection, but this is almost entirely explained by its low coding density, as there is no deficit in non-protein-coding sequence under selection. Approximately 4.2% of the bases on chromosome 18 appear to be under purifying selection, consisting of 0.6% in exons of protein-coding genes and 3.6% in non-protein-coding elements. The proportion of non-protein-coding sequence under selection is typical for human chromosomes. (Note that chromosomes 19 and 22 are outliers in this analysis; the many local gene family expansions make it difficult to assign orthology.)

As chromosomes vary widely in size, we repeated the analysis for 5-Mb windows across the human genome (Fig. 2c, d). Although there is more scatter in the data, the overall conclusion is very similar. Notably, the average proportion of non-protein-coding selected sequence in a window is ~3.8%, and is slightly negatively correlated ($R^2 = 0.08$) with the proportion of coding sequence in the window.

Our analysis shows that the density of conserved non-protein-coding sequences is largely independent of the density of protein-coding genes. It is interesting to note that examination of non-coding aligned sequences between human and chicken[23] showed a negative correlation with coding content, and a study of highly conserved non-coding sequences in intergenic regions of human chromosome 21 did not identify tight coupling to the starts and ends of genes[24,25].

What is the nature of the non-protein-coding elements? First, the elements might encode transcripts that are not translated into proteins, such as small RNA genes or large regulatory RNAs[26]. Second, they might serve a structural role, with a constant density of such elements required to maintain chromosome structure independent of gene density. Such structural elements could be evolutionarily essential for maintenance of a region, but might be dispensable if the entire region were to be deleted; this might explain the recent observation in mouse that a 1-Mb deletion in a gene desert containing highly conserved elements has no discernable phenotypic effect[27]. Third, the elements may be largely related to the regulation of protein-coding genes, but their distribution may be inversely correlated with gene density[28,29]. It is possible that genes in gene-poor regions tend to have more elaborate regulatory controls, and this could partially explain the relative sparsity of genes in such regions. In any case, it is clear that the finished sequence of the human genome will reveal many features of biological function and provide a firm foundation for future systematic analyses.

METHODS

Generation of the gene catalogue. We started by aligning all available human RefSeq, MGC and GenBank messenger RNA sequences, as well as GenPept sequences from several species, to the finished sequence. Gene models were inspected manually to ensure accurate transcriptional start and stop sites, and to correct splice sites. Non-canonical splice sites were used only if supported by sufficient complementary DNA-based evidence. Partial transcripts (those containing a partial open reading frame (ORF) or overlapping non-coding exons of sibling transcripts) were annotated in cases for which there was firm evidence of their existence. Gene symbols for biologically characterized loci were assigned by the HUGO Gene Nomenclature Committee. See Supplementary Table S10 for a complete list of gene symbols. Our annotations are available from the Vertebrate Genome Annotation database (VEGA, http://vega.sanger.ac.uk/Homo_sapiens).

Comparative analysis: creation of synteny maps. We performed full genomic alignments of repeat masked sequence from mouse[4] (builds 31 and 33), rat[21] and dog (CanFam 1.0; K. Lindblad-Toh, personal communication) with the human genome sequence using the PatternHunter program[30]. We did this for human build 34 with the Broad finished chromosomes (8, 15, 17, 18) inserted, and also for human build 35 (mouse build 31 was used against human build 34, and mouse build 33 against human build 35). From these alignments we identified collinear clusters of conserved microsynteny, which were then used to form larger syntenic segments in a hierarchical fashion. Syntenic maps and their underlying syntenic anchors serve as the basis for identification of conserved elements.

Comparative analysis: identification of conserved elements. Starting with large-scale syntenic blocks defined by the human–mouse and human–dog syntenic maps, we generated pair-wise alignments within these syntenic blocks using the PatternHunter program[30]. We then scanned 50-bp windows with 5-bp

offset and calculated the fraction of aligning bases that were matches (discarding windows with fewer than 20 aligning bases). These percentage conservation values were locally normalized to the average conservation in the surrounding 5 Mb to generate *Z*-scores measuring divergence from the local average (0) for every window. We examined the joint empirical distribution of mouse and dog *Z*-scores for windows contained within ancestral repeat sequence (undergoing neutral evolution and believed to predate the mouse–human split) and windows overlapping coding exons (Supplementary Fig. S4a). Coding sequence is defined as all bases that are annotated as coding in any transcript. All analysis presented uses Ensembl[31] genes on human build 35; analysis with both Ensembl and Broad annotations on build 34 yields substantially similar results (Supplementary Information).

We combined dog and mouse *Z*-scores to generate a 'composite' *Z*-score (see Supplementary Information). We estimated the distribution of composite *Z*-scores for selected sequence by decomposing the global distribution of *Z*-scores into two components: a 'neutral distribution' centred at zero and corresponding to the conservation scores for ancestral repeat sequences, and a 'selected distribution' consisting of the residual after subtraction of the neutral distribution (Supplementary Fig. S4b). Taking into account the relative fractions of the aligning windows in each distribution, we were able to assign a probability that a window at a given score is under purifying selection.

We then divided the genome into non-overlapping 5-Mb windows. Within each such window, we counted the number of syntenic bases, the number of syntenic 50-bp windows, and the number of 50-bp windows under selection. The fraction of coding sequence (the explanatory variable in all regressions) was taken as the number of syntenic bases annotated as coding divided by the number of syntenic bases. The fraction under selection was calculated as the sum of all selection probabilities for all windows divided by the number of syntenic windows. If windows of only a certain class were considered, the probabilities were calculated only for windows in that class. We note that, on average, windows contained within coding exons scored only slightly higher than 0.67 probability of selection, owing to the large prior probability of neutrality. Thus, the slopes of all regressions are <1. For all analyses, we discarded any 5-Mb window with less than 4 Mb of syntenically assigned sequence (retaining $>85\%$ of all windows of non-zero euchromatic length). Similar results are obtained if the discarded windows are included, but the variance is higher.

Annotation. RefSeq (release 1), mammalian gene collection (MGC, 3 February 2003), dbEST and GenBank (29 December 2002) mRNAs were aligned to the genomic assembly using BLAT[32]. GenPept protein sequences (3 February 2003) were aligned using BLASTX[33] and GeneWise[34]. All gene models were created manually using these aligned sequences as evidence, following HAWK2 (www.sanger.ac.uk/Info/workshops/hawk2) transcript type conventions. Gene models derived from aligned mRNA evidence were extended when possible using spliced EST evidence at the 5′ end and spliced and unspliced EST evidence in the 3′ untranslated region (UTR). Evidence was given relative priority as follows (high–low): RefSeq/MGC, GeneWise, other mRNAs, spliced ESTs and unspliced ESTs. We found CpG islands within 2-kb upstream and 1-kb downstream of the 5′ end of 73% of known category loci, which is somewhat higher than previous reports (in the range of 61–66%; cited in ref. 3, also refs 5–7).

Received 21 January; accepted 27 June 2005.

1. Hernandez, D. & Fisher, E. M. Mouse autosomal trisomy: two's company, three's a crowd. *Trends Genet.* **15,** 241–247 (1999).
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
5. Grimwood, J. *et al.* The DNA sequence and biology of human chromosome 19. *Nature* **428,** 529–535 (2004).
6. Deloukas, P. *et al.* The DNA sequence and comparative analysis of human chromosome 10. *Nature* **429,** 375–381 (2004).
7. Martin, J. *et al.* The sequence and analysis of duplication-rich human chromosome 16. *Nature* **432,** 988–994 (2004).
8. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33,** D501–D504 (2005).
9. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31,** 241–247 (2002).
10. Schuler, G. D. *et al.* A gene map of the human genome. *Science* **274,** 540–546 (1996).
11. Schmutz, J. *et al.* Quality assessment of the human genome sequence. *Nature* **429,** 365–368 (2004).
12. Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nature Biotechnol.* **21,** 379–386 (2003).
13. Veeramachaneni, V., Makalowski, W., Galdzicki, M., Sood, R. & Makalowska, I. Mammalian overlapping genes: the comparative perspective. *Genome Res.* **14,** 280–286 (2004).
14. Mouchiroud, D. *et al.* The distribution of genes in the human genome. *Gene* **100,** 181–187 (1991).
15. Rebhan, M. *et al.* GeneCards: encyclopedia for genes, proteins and diseases. (Weizmann Institute of Science, Bioinformatics Unit and Genome Center, Rehovot, Israel) ⟨http://bioinformatics.weizmann.ac.il/cards⟩ (1997).
16. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
17. Dennehey, B. K., Gutches, D. G., McConkey, E. H. & Krauter, K. S. Inversion, duplication, and changes in gene context are associated with human chromosome 18 evolution. *Genomics* **83,** 493–501 (2004).
18. Goidts, V., Szamalek, J. M., Hameister, H. & Kehrer-Sawatzki, H. Segmental duplication associated with the human-specific inversion of chromosome 18: a further example of the impact of segmental duplications on karyotype and genome evolution in primates. *Hum. Genet.* **115,** 116–122 (2004).
19. King, M. *Species Evolution* 72–91 (Cambridge Univ. Press, Cambridge, 1993).
20. Delneri, D. *et al.* Engineering evolution to study speciation in yeasts. *Nature* **422,** 68–72 (2003).
21. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
22. Chiaromonte, F. *et al.* The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harb. Symp. Quant. Biol.* **68,** 245–254 (2003).
23. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).
24. Dermitzakis, E. T. *et al.* Comparison of human chromosome 21 conserved nongenic sequences (CNGs) with the mouse and dog genomes shows that their selective constraint is independent of their genic environment. *Genome Res.* **14,** 852–859 (2004).
25. Dermitzakis, E. T., Reymond, A. & Antonarakis, S. Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nature Rev. Genet.* **6,** 151–157 (2005).
26. Rastegar, M. *et al.* Sequential histone modifications at Hoxd4 regulatory regions distinguish anterior from posterior embryonic compartments. *Mol. Cell. Biol.* **24,** 8090–8103 (2004).
27. Nobrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V. & Rubin, E. M. Megabase deletions of gene deserts result in viable mice. *Nature* **431,** 988–993 (2004).
28. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145 (2005).
29. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning human gene deserts for long-range enhancers. *Science* **302,** 413 (2003).
30. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18,** 440–445 (2002).
31. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).
32. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).
33. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
34. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).

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

Acknowledgements Special thanks are due to L. Gaffney for help with the manuscript, figures and tables, and to K. Lance for help with the manuscript. We are grateful to E. Eichler and X. She for sharing data on segmental duplications, T. Furey for help with lists of genetic markers and placement of RefSeqs, and K. Lindblad-Toh for sharing data from the dog genome project. In addition, we thank Ming Li and Bin Ma (BioInformatics Solutions Inc.) for providing PatternHunter and advice about how to choose appropriate parameters. We also acknowledge the HUGO Gene Nomenclature Committee (S. Povey, E. A. Bruford, V. K. Khodiyar, R. C. Lovering, M. J. Lush, T. P. Sneddon, C. C. Talbot Jr and M. W. Wright) for assigning official gene symbols. We are grateful to all the members, present and past, of the Broad (and Whitehead) sequencing platform for the consistent high quality of their data.

Author Information Accession numbers for all clones contributing to the finished sequence of human chromosome 18 can be found in Supplementary Table S3. The updated human chromosome 18 sequence can be accessed through GenBank accession number NC_000018. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.N. (chad@broad.mit.edu).

CORRIGENDUM

doi:10.1038/nature04362

A network-based analysis of systemic inflammation in humans

Steve E. Calvano, Wenzhong Xiao, Daniel R. Richards, Ramon M. Felciano, Henry V. Baker, Raymond J. Cho, Richard O. Chen, Bernard H. Brownstein, J. Perren Cobb, S. Kevin Tschoeke, Carol Miller-Graziano, Lyle L. Moldawer, Michael N. Mindrinos, Ronald W. Davis, Ronald G. Tompkins, Stephen F. Lowry & the Inflammation and Host Response to Injury Large Scale Collaborative Research Program†

Nature **437**, 1032–1037 (2005) doi:10.1038/nature03985

In this Letter, the affiliations of authors participating in the Inflammation and Host Response to Injury Large Scale Collaborative Research Program are incorrectly listed. The renumbered and amended footnote listing is given here.

†**Inflammation and Host Response to Injury Large Scale Collaborative Research Program** Paul E. Bankey[1], Timothy R. Billiar[2], David G. Camp[3], George Casella[4], Irshad H. Chaudry[5], Mashkoor A. Choudhry[5], Charles Cooper[6], Asit De[1], Constance Elson[7], Bradley Freeman[8], Richard L. Gamelli[9], Celeste Campbell-Finnerty[10], Nicole S. Gibran[11], Douglas L. Hayden[7], Brian G. Harbrecht[2], David N. Herndon[10], Jureta W. Horton[12], William J. Hubbard[5], John L. Hunt[13], Jeffrey Johnson[14], Matthew B. Klein[15], James A. Lederer[16], Tanya Logvinenko[7], Ronald V. Maier[11], John A. Mannick[16], Philip H. Mason[6], Bruce A. McKinley[17], Joseph P. Minei[12], Ernest E. Moore[14], Frederick A. Moore[17], Avery B. Nathens[11], Grant E. O'Keefe[11], Laurence G. Rahme[18], Daniel G. Remick[19], David A. Schoenfeld[7], Martin G. Schwacha[5], Michael B. Shapiro[20], Geoffrey M. Silver[9], Richard D. Smith[3], John D. Storey[21], Mehmet Toner[22], H. Shaw Warren[23] & Michael A. West[20]

Affiliations for participants: [1]Department of Surgery, University of Rochester School of Medicine, Rochester, New York 14642, USA. [2]Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA. [3]Pacific Northwest National Laboratory, Richland, Washington 99352, USA. [4]Department of Statistics, University of Florida, Gainesville, Florida 32611, USA. [5]Department of Surgery, University of Alabama School of Medicine, Birmingham, Alabama 35294, USA. [6]Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Cambridge, Massachusetts 02139, USA. [7]Department of Biostatistics, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [8]Department of Surgery, Washington University School of Medicine, St. Louis, Missouri 63110, USA. [9]Department of Surgery, Loyola University Stritch School of Medicine, Maywood, Illinois 60153, USA. [10]Department of Surgery, University of Texas Medical Branch, Shriners Burns Hospital, Galveston, Texas 77550, USA. [11]Department of Surgery, University of Washington Harborview Medical Center, Seattle, Washington 98104, USA. [12]Department of Surgery, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [13]Division of Trauma, Burns, and Critical Care, University of Texas Southwestern Medical Center, Dallas, Texas 75390, USA. [14]Department of Surgery, University of Colorado Denver Health Medical Center, Denver, Colorado 80204, USA. [15]Burn Center and Division of Plastic Surgery, University of Washington Harborview Medical Center, Seattle, Washington 98104, USA. [16]Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. [17]Department of Surgery, University of Texas Houston Health Science Center, Houston Medical School, Houston, Texas 77030, USA. [18]Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [19]Department of Medical Science, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA. [20]Department of Surgery, Northwestern University Medical School, Chicago, Illinois 60611, USA. [21]Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA. [22]Center for Engineering in Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [23]Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02129, USA.

CORRIGENDUM

doi:10.1038/nature04363

DNA sequence and analysis of human chromosome 18

Chad Nusbaum, Michael C. Zody, Mark L. Borowsky, Michael Kamal, Chinnappa D. Kodira, Todd D. Taylor, Charles A. Whittaker, Jean L. Chang, Christina A. Cuomo, Ken Dewar, Michael G. FitzGerald, Xiaoping Yang, Amr Abouelleil, Nicole R. Allen, Scott Anderson, Toby Bloom, Boris Bugalter, Jonathan Butler, April Cook, David DeCaprio, Reinhard Engels, Manuel Garber, Andreas Gnirke, Nabil Hafez, Jennifer L. Hall, Catherine Hosage Norman, Takehiko Itoh, David B. Jaffe, Yoko Kuroki, Jessica Lehoczky, Annie Lui, Pendexter Macdonald, Evan Mauceli, Tarjei S. Mikkelsen, Jerome W. Naylor, Robert Nicol, Cindy Nguyen, Hideki Noguchi, Sinéad B. O'Leary, Keith O'Neill, Bruno Piqani, Cherylyn L. Smith, Jessica A. Talamas, Kerri Topham, Yasushi Totoki, Atsushi Toyoda, Hester M. Wain, Sarah K. Young, Qiandong Zeng, Andrew R. Zimmer, Asao Fujiyama, Masahira Hattori, Bruce W. Birren, Yoshiyuki Sakaki & Eric S. Lander

Nature **437**, 551–555 (2005) doi:10.1038/nature03983

The name of Keith O'Neill was accidentally omitted from the published author list. He is at the first affiliation in the address list.

ERRATUM

doi:10.1038/nature04361

Astronomical pacing of methane release in the Early Jurassic period

David B. Kemp, Angela L. Coe, Anthony S. Cohen & Lorenz Schwark

Nature **437**, 396–399 (2005)

In the labelling of Fig. 1 of this Letter, the spelling of '*D. semicelatum*' was accidentally reversed to read '*D. mutalecimes*'. It appears correctly in the text.