Annotation of the Human Genome Assembly [NCBI *hg*17, Build 35, May 2004]

Since the publication of the initial draft of the human genome sequence [1, 2], the quality of the sequence as well as the accuracy of the annotations have improved [3]. This figure provides a snapshot of the current annotation features mapped on the NCBI's *hg*17 (Build 35, May 2004) assembly version of the human genome, summarizing the completion status of both, the sequence and the annotated features. This whole genome map also pinpoints the complexity intrinsic to the vertebrate genomes, which is represented by the variability found at different levels: the nucleotide sequence (exemplified by the G+C content and repeats distribution); the gene density (including transcripts, pseudogenes, and expressed sequence tags—ESTs); the variability of the genic structures (from lengths to their exonic structure and their functional classification); and, finally, the intra-specific variability (represented by the single nucleotide polymorphisms—SNPs).

Each of the 25 human chromosomes (1 to 22, X and Y, plus the mitochondrion circular chromosome), has been drawn separately as a single track on this figure. Each chromosome track is divided into three areas: forward-strand transcripts, sequence analysis, and reverse-strand transcripts (from top to bottom, respectively). The genome sequence is displayed on a nucleotide scale of approximately 600 kbp/cm. The background of the chromosome tracks shows an approximate correspondence of the chromosome cytogenetic map [4]. The centromere is depicted as a blue line crossing the annotation tiers and its position is approximated by the transition from *p* to *q* arms along the chromosome sequences.

Genes are adjacent to the sequence analysis tiers. They are color-coded by the algorithm used to define the transcript structure (see figure key) and are given a minimum length of 1 kbp for display purposes. The structure of transcripts with two or more exons is displayed in one of two expanded transcript tiers at 120 kbp/cm (5x) resolution above or below the genes, for forward- and reverse-strand transcripts, respectively. Untranslated (UTR) and coding (CDS) exons are depicted as grey and black boxes respectively, intronic regions are color-coded for transcripts assigned to 14 Gene Ontology [GO, 5] functional categories. Single-exon transcripts are color-coded by GO classification and are displayed in a tier between the unexpanded transcripts and the pseudogenes tiers. Three different transcript sets were downloaded from the UCSC Genome Browser [6]: *Known* (protein coding genes based on proteins from SWISS-PROT, TrEMBL, and TrEMBL-NEW and their corresponding mRNAs from GenBank), *RefSeq* [7] and *Ensembl* [8] genes. From a merged set of 97796 transcripts (39368 *Known*, 24847 *RefSeq* and 33581 *Ensembl*), a final set of 25201 transcripts (20094, 978 and 4129 respectively) was filtered out to be included in this whole genome map. Gene symbols are shown for the genes on expanded tiers larger than 45 kbp; however, few overlapping labels were removed by hand in order to improve readability. HUGO gene symbols [9] were used when possible; symbols starting with *NM* correspond to *RefSeq* annotated genes, while those starting with *ET* are referring to *Ensembl* transcript identifiers (in this case, *ET* stands for *ENST00000*). Just below genes track, there is a pseudogenes track in which the current pseudogenes annotations were summarized from two different sources: *VEGA* [10] (4451 pseudogenes, only for chromosomes 14, 20 and 22) and the *YALE* Pseudogene Database [11] (7360 pseudogenes). From 11811 pseudogenes, 9693 were filtered out and are drawn as brown ellipses and, for those still retaining introns, the corresponding exons are shown in black.

The next tier shows the projections, depicted as green boxes, along the sequence of the alignments available from the UCSC database for all the human EST sequences to date. The last stranded feature corresponds to a gradient, from white to blue, representing the density of repetitive elements, computed on a running window of length 25 kbp (percent of nucleotides annotated as repeats within that window). The repeats retrieved from the UCSC Genome Browser were produced by RepeatMasker [12] at the -s sensitive setting.

The middle section of the chromosome tracks consists of three sequence analyses: G+C content, CpG Islands and SNP density. G+C content and SNP density, calculated on running windows of 25 and 100 kbp respectively, are depicted as color gradients, ranging from purple/dark blue to red (for minimum to maximum scores respectively). The natural log of the SNP density is used to color-code the SNP density analysis tier. The gradient scales for the repeats density, for the G+C content and for the SNPs, are shown in the figure key. The raw scores for each of those features were normalized to the whole genome quantiles (in steps of 5%). A dark green box indicates the position of CpG islands. Gaps along the sequence are shown in the same track as CpG islands. Three different classes of gaps are illustrated by boxes of different colors: gaps from large blocks of heterochromatin (brown); gaps between clones in the same map contig (orange); and other gaps (red), including telomeric and centromeric gaps, large gaps in the short *p* arm and gaps between map contigs.

Mitochondrial circular chromosome is quite small compared to the rest of human chromosomes. To visualize its gene content, it required a 375x larger scale. This chromosome has small single exon genes, these exons were filled with the corresponding color codes for the gene function (GO), instead of black which is used on all the rest of chromosomes. Sequence analysis tracks contain only EST projections, G+C content and SNPs density.

Each chromosome figure was generated separately with `gff2ps` [13], a genome annotation tool that converts annotation records in General Feature Format [GFF, 14] to a PostScript output.

References

[1] Lander E *et al.* "Initial sequencing and analysis of the human genome." *Nature*, 409(6822):860–921, Feb 2001.

[2] Venter J *et al.* "The sequence of the human genome." *Science*, 291(5507):1304–51, Feb 2001.

[3] International Human Genome Sequencing Consortium, IHGSC. "Finishing the euchromatic sequence of the human genome." *Nature*, 431(7011):931–45, Oct 2004.

[4] Furey T and Haussler D. "Integration of the cytogenetic map with the draft human genome sequence". *Hum Mol Genet*, 12(9):1037–1044, May 2003.

[5] Ashburner M *et al.* "Gene Ontology: tool for the unification of biology". *Nature Genetics*, 25(1):25–9, May 2000.
URL `http://www.geneontology.org`

[6] Karolchik D *et al.* "The UCSC Genome Browser Database." *Nucleic Acids Res*, 31(1):51–4, Jan 2003.
URL `http://genome.ucsc.edu/`

[7] Pruitt K, Tatusova T and Maglott D. "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Res*, 33(DB Issue):D501–4, Jan 2005.
URL `http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html`

[8] Curwen V *et al.* "The Ensembl automatic gene annotation system." *Genome Res*, 14(5):942–50, May 2004.
URL `http://www.ensembl.org/`

[9] Wain H *et al.* "Genew: the Human Gene Nomenclature Database, 2004 updates". *Nucleic Acids Res*, 32(DB Issue):D255–257, Jan 2004.
URL `http://www.gene.ucl.ac.uk/nomenclature/`

[10] Ashurst J *et al.* "The VErtebrate Genome Annotation (VEGA) database." *Nucleic Acids Res*, 33(DB Issue):D459–65, Jan 2005.
URL `http://vega.sanger.ac.uk/`

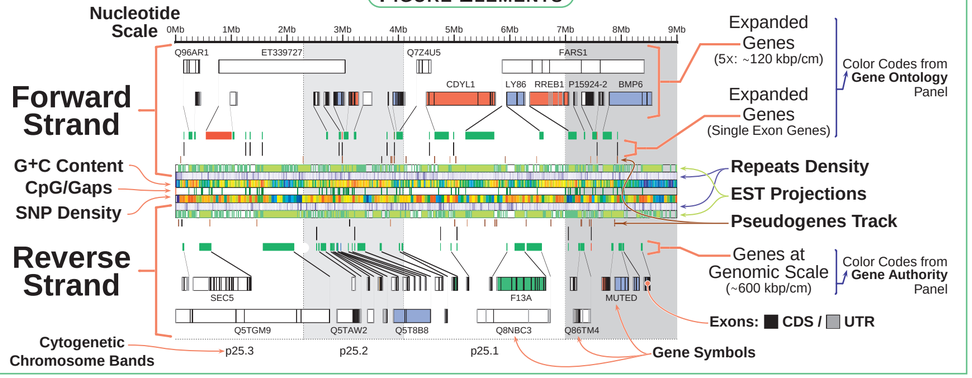
[11] Zhang Z and Gerstein M. "Large-scale analysis of pseudogenes in the human genome." *Curr Op in Genet and Devel*, 14(4):328–335, Aug 2004.
URL `http://www.pseudogene.org/`

[12] Smit A, Hubley R and Green P. "`RepeatMasker`". *Institute for Systems Biology*, *unpublished*, 1996–2006.
URL `http://www.repeatmasker.org/`

[13] Abril J and Guigó R. "`gff2ps`: visualizing genomic annotations." *Bioinformatics*, 16(8):743–4, Aug 2000.
URL `http://genome.imim.es/software/gfftools/GFF2PS.html`

[14] R. Durbin and D. Haussler, with updates by L. Stein, S. Lewis, A. Krogh and others. "GFF Protocol Specification". *Sanger Center*, *unpublished*, 1997–2006.
URL `http://www.sanger.ac.uk/Software/formats/GFF/`

Figure Elements



Gene Authority

Known 20094 [79.74%]	**RefSeq** 978 [3.88%]	**Ensembl** 4129 [16.38%]

GO Functional Categories (Expanded Genes)

Antioxidant Activity 36 [0.14 %]	**Binding** 1800 [7.15 %]
Catalytic Activity 1915 [7.60 %]	**Chaperone Regulator** 3 [0.01 %]
Enzyme Regulator 511 [2.03 %]	**Motor Activity** 66 [0.26 %]
Nutrient Reservoir 1 [0.004%]	**Obsolete Molecular Function** 60 [0.24 %]
Signal Transducer 228 [0.91 %]	**Structural Molecule** 208 [0.83 %]
Transcription Regulator 467 [1.85 %]	**Translation Regulator** 32 [0.13 %]
Transporter 3287 [13.05 %]	**Unknown Molecular Function** 16574 [65.80 %]

Gap Classes

- **Heterochromatin**
- **Clone**
- **Other**

Other Features

- **VEGA Pseudogenes**
- **YALE Pseudogenes**
- **EST Projections**
- **CpG Islands**

Color Gradient Features

Repeats density (% per 25 kbp)

5 10 12 14 16 18 20 22 24 26 28 30 32 35 40 60

G+C content (% per 25 kbp)

20 30 35 36 37 38 39 40 41 42 43 44 45 46 48 50 60 70

5 50 75 100 125 150 175 200 300 1000 4000

SNP coverage (counts per 100 kbp, logarithmic scale)

Chromosomes Cytogenetic Map (Background Bands)

Giemsa[+] 100%	**Giemsa[+] 75%**	**Giemsa[+] 50%**	**Giemsa[+] 25%**
Giemsa[-]	**Stalk**	**Centromere Mark**	**Centromeric and Heterochromatic**