Annotation of the Celera Human Genome Assembly

Initial annotations of the Celera compartmentalized shotgun assembly (CSA) of the human genome including transcripts, sequence characteristics, polymorphisms, and molecular markers are presented. Each track of the figure is divided into three areas: forward-strand transcripts, sequence analysis, and reverse-strand transcripts (from top to bottom, respectively). The end of each chromosome tier is depicted as white space as it not yet clear that the CSA includes the telomeres. The genome sequence is displayed on a nucleotide scale of approximately 600 kbp/cm. Molecular genetic markers are shown above the nucleotide scale at the top of each track and are derived from the Marshfield map (http://research.marshfieldclinic.org/genetics/Map_Markers/maps/IndexMapFrames.html). 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 20 kb for display purposes. The structure of transcripts with two or more exons is displayed in one of two expanded transcript tiers at 120 kb/cm resolution above or below the genes for forward- and reverse-strand transcripts, respectively. Exons are depicted as black boxes and intronic regions are color-coded for transcripts assigned to the 14 largest Gene Ontology (GO, http://www.geneontology.org) categories. Single-exon transcripts are color-coded by GO classification and are displayed in a tier between the unexpanded transcripts and the sequence analysis tiers. Transcripts predicted by Celera's annotation algorithm (Otto) that correspond to RefSeq transcripts (http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html) are assigned HUGO gene symbols (http://www.gene.ucl.ac.uk/nomenclature) if the RefSeq transcripts are associated with HUGO symbols by LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink) and if the transcripts are longer than 25 kbp (to prevent overlap of gene symbols). There are three sequence analyses in the middle section of the tracks: G+C content, CpG Islands and SNP density. G+C content is depicted in a nonlinear scale described in the legend. A black box indicates the position of CpG islands. SNPs were identified by comparison of the Celera sequence with a genome assembly available at http://genome.ucsc.edu/. The range of SNP density is depicted above the color gradient in the legend. The natural log of the SNP density is used to color-code the SNP density analysis tier. Gaps within scaffolds are visible as white space in the G+C content tier if the gap is sufficiently large. Gaps between scaffolds are assigned a length of 2 kbp. Scaffold order along the chromosomes was determined by mate-pair information and alignment of scaffold sequence to the GeneMap'99 STS map (http://www.ncbi.nlm.nih.gov/genemap99/) and the Washington University BAC fingerprint map (http://www.genome.wustl.edu/gsc/mapping/). 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 genome sequence, except for acrocentric chromosomes for which the centromere is placed at the beginning of the sequence analysis tiers.

The figure was generated with "gff2ps" (http://www1.imim.es/software/gfftools/GFF2PS.html), a genome annotation tool that converts General Feature Formatted records (http://www.sanger.ac.uk/Software/formats/GFF/) to a PostScript output [J. F. Abril, R. Guigó, *Bioinformatics* 16, 743 (2000)].