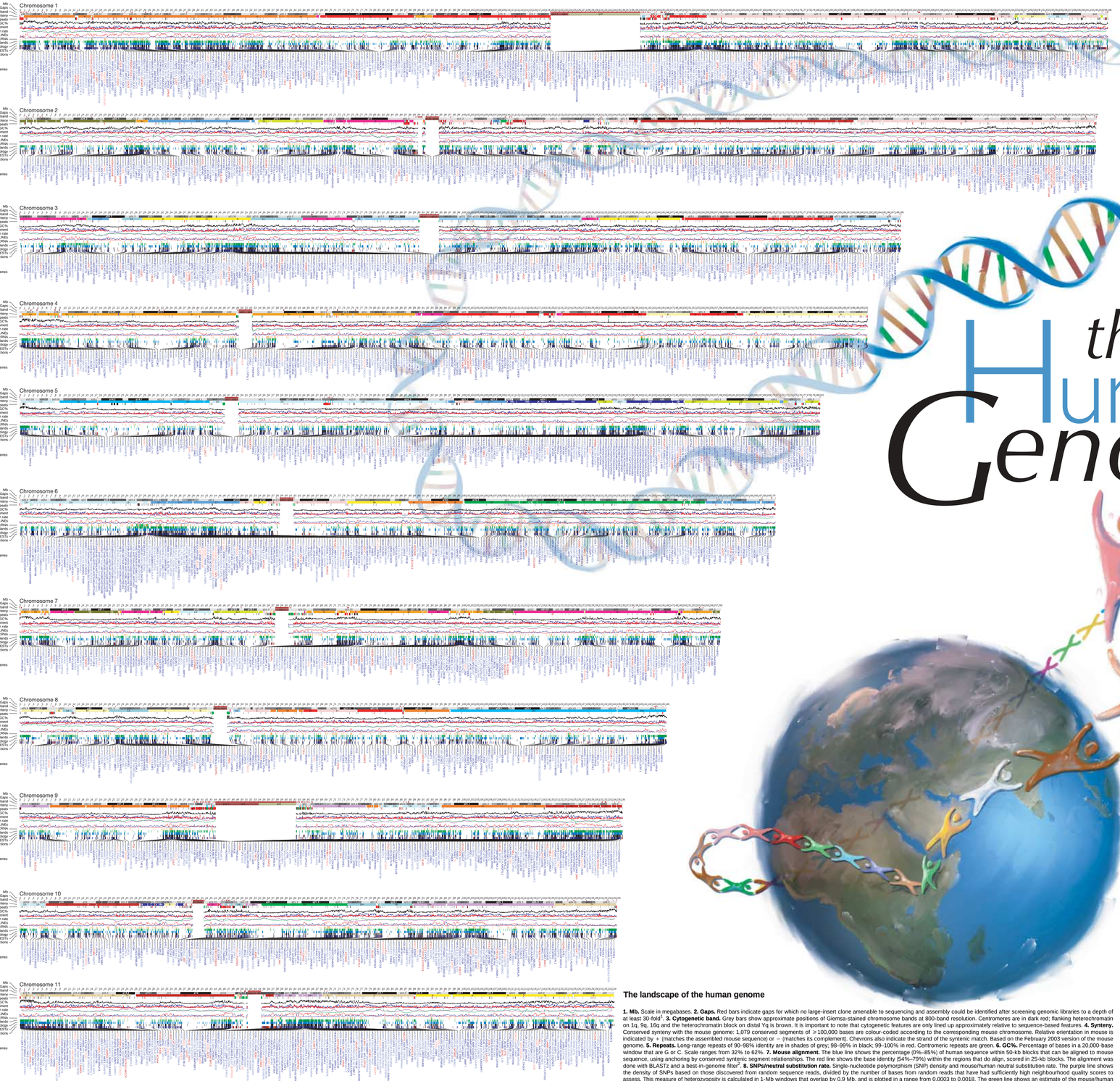
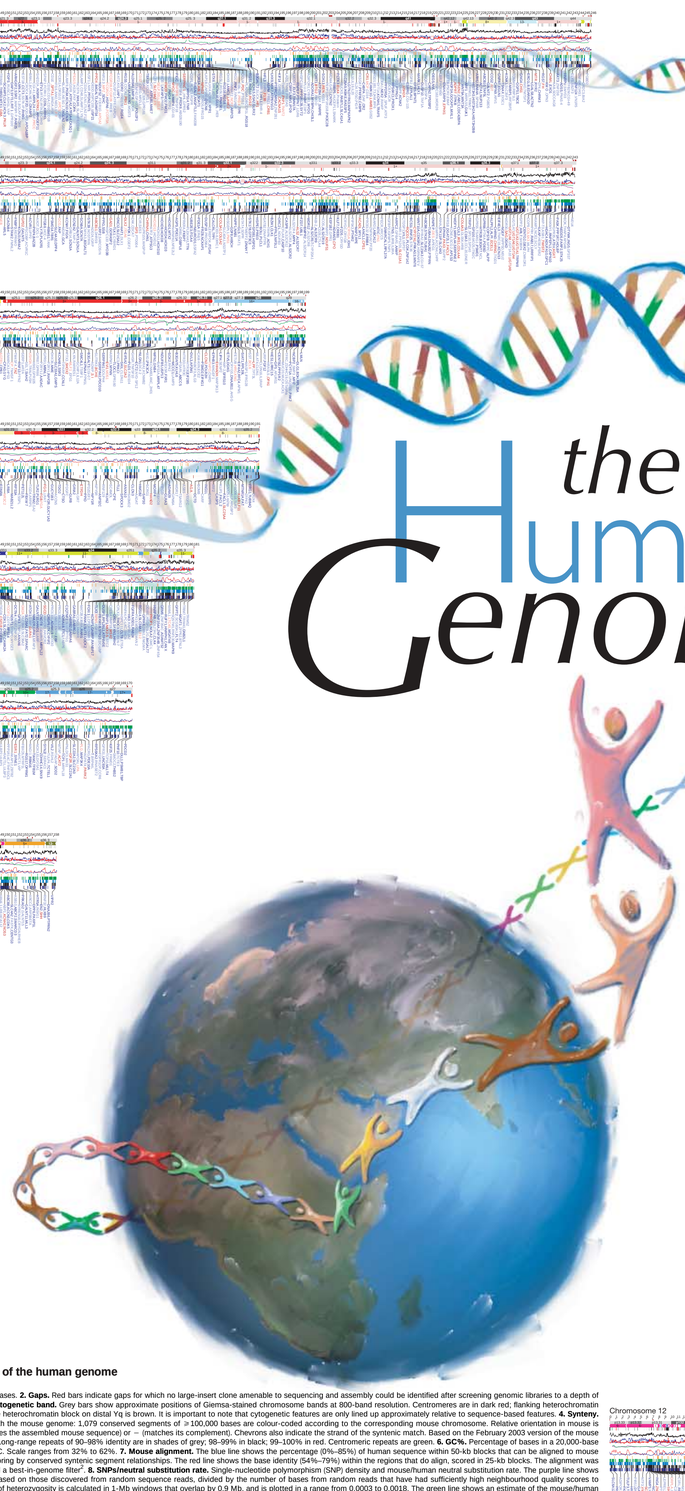
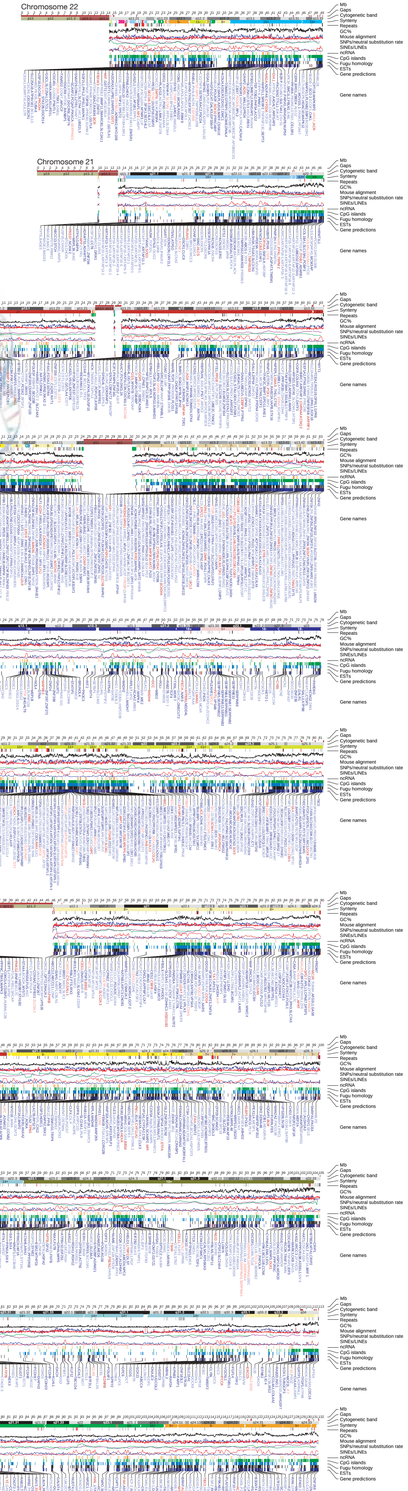
the Human Genome







The landscape of the human genome

1. Mb. Scale in megabases. **2. Gaps.** Red bars indicate gaps for which no large-insert clone amenable to sequencing and assembly could be identified after screening genomic libraries to a depth of at least 30-fold[1]. **3. Cytogenetic band.** Grey bars show approximate positions of Giemsa-stained chromosome bands at 800-band resolution. Centromeres are in dark red; flanking heterochromatin on 1q, 9q, 16q and the heterochromain block on distal Yq is brown. It is important to note that cytogenetic features are only lined up approximately relative to sequence-based features. **4. Synteny.** Conserved synteny with the mouse genome: 1,079 conserved segments of ⩾100,000 bases are colour-coded according to the corresponding mouse chromosome. Relative orientation in mouse is indicated by + (matches the assembled mouse sequence) or − (matches its complement). Chevrons also indicate the strand of the syntenic match. Based on the February 2003 version of the mouse genome. **5. Repeats.** Long-range repeats of 90–98% identity are in shades of grey; 98–99% in black; 99–100% in red. Centromeric repeats are green. **6. GC%.** Percentage of bases in a 20,000-base window that are G or C. Scale ranges from 32% to 62%. **7. Mouse alignment.** The blue line shows the percentage (0%–85%) of human sequence within 50-kb blocks that can be aligned to mouse sequence, using anchoring by conserved syntenic segment relationships. The red line shows the base identity (54%–79%) within the regions that do align, scored in 25-kb blocks. The alignment was done with BLASTZ and a best-in-genome filter[2]. **8. SNPs/neutral substitution rate.** Single-nucleotide polymorphism (SNP) density and mouse/human neutral substitution rate. The purple line shows the density of SNPs based on those discovered from random sequence reads, divided by the number of bases from random reads that have had sufficiently high neighbourhood quality scores to assess. This measure of heterozygosity is calculated in 1-Mb windows that overlap by 0.9 Mb, and is plotted in a range from 0.0003 to 0.0018. The green line shows an estimate of the mouse/human neutral substitution rate, calculated using the REV model from aligned human, mouse and rat sites in transposon relics that predate the primate–rodent divergence, and depicted in 1-Mb windows. Units are substitutions per site; the scale ranges from 0.2 to 1.0. **9. SINEs/LINEs.** The most common dispersed repetitive elements in the human genome. Short interspersed repeated sequences (SINEs) are in red; long interspersed repeated sequences (LINEs) are in blue; both are calculated in 100-kb windows. **10. ncRNA.** RNA genes that do not code for protein — for example, transfer RNAs, small nucleolar RNAs and ribosomal RNAs — are in green. RNA pseudogenes are in orange. **11. CpG islands.** Each green tick represents a segment of 200 bases or more with CpG dinucleotide levels significantly higher than in the genome as a whole, and GC percentage of at least 50%. **12. Fugu homology.** Blue ticks show positions of translated BLAT alignments[3] between the *Takifugu rubripes* genome[4] and the human genome. **13. ESTs.** Black ticks show the location of expressed sequence tags (ESTs) with at least one intron, aligned against genomic DNA. **14. Gene predictions.** The start points of genes from RefSeq are indicated in blue; the starts of additional gene predictions from Ensembl are in dark red[5]. **15. Gene names.** Genes are named in blue, using HUGO Gene Nomenclature Committee abbreviations. Red indicates a known disease gene, bold indicates genes with a *Caenorhabditis elegans* or *Drosophila melanogaster* ortholog[6].

[1]*Nature* (this issue). [2]Schwartz, S. *et al. Genome Res.* **13,** 103–107 (2003). [3]Kent, W. J. *Genome Res.* **12,** 656–664 (2002). [4]Aparicio, S. *et al. Science* **297,** 1301–1310 (2002). [5]Birney, E. *et al. Nucleic Acids Res.* **32,** D468–D470 (2004). [6]Enright, A. J. *et al. Nucleic Acids Res.* **30,** 1575–1584 (2002).

Data: Ewan Birney, Jim Kent, Krishna M. Roskin, Jim Mullikin, Daryl Thomas, Robert Baertsch. Design: Darryl Leja.