

Comparative Genomics

Bioinformatics Course 2003

Nathan Hall

Ludwig Institute for Cancer Research



Gene index analysis of the human genome estimates approximately 120,000 genes.

Nat Genet. 2000 Jun;25(2):239-40

Liang F, Holt I, Perlea G, Karamycheva S, Salzberg SL, Quackenbush J.



Human Genome Project

- Sequencing now “*finished*”
- 99 % sequenced
- 400 gaps (technological issues)
- about 25-30,000 genes
- June 2000 draft sequence
 - 90% sequences
 - 150,000 gaps
- **Now comes the understanding . . .**



Understanding the genome

- SNPs
- Disease Genes (100 in 1990, 1,400 now)
- Correlate genomic features with novel traits
- Quantitative Trait Loci (QTL), genetic understanding of diseases and traits
- HapMap, mapping haplotypes (SNP patterns) for help with understanding disease susceptibility
- MicroRNAs
- Regulatory elements
- Alternative splicing



Genome Sequences

- *E. Coli* (1996, 4.6 MB, 3,300 genes)
- *C. elegans* (1998, 97 MB, 19,000 genes)
- *Drosophila* (2000, 180MB, 14,000 genes)
- Human (draft 2000/2001, 2003 finished, 3.3 GB, ~30,000 genes)
- Mouse
- Rat
- Ciona
- Fugu/Tetraodon/Zebrafish

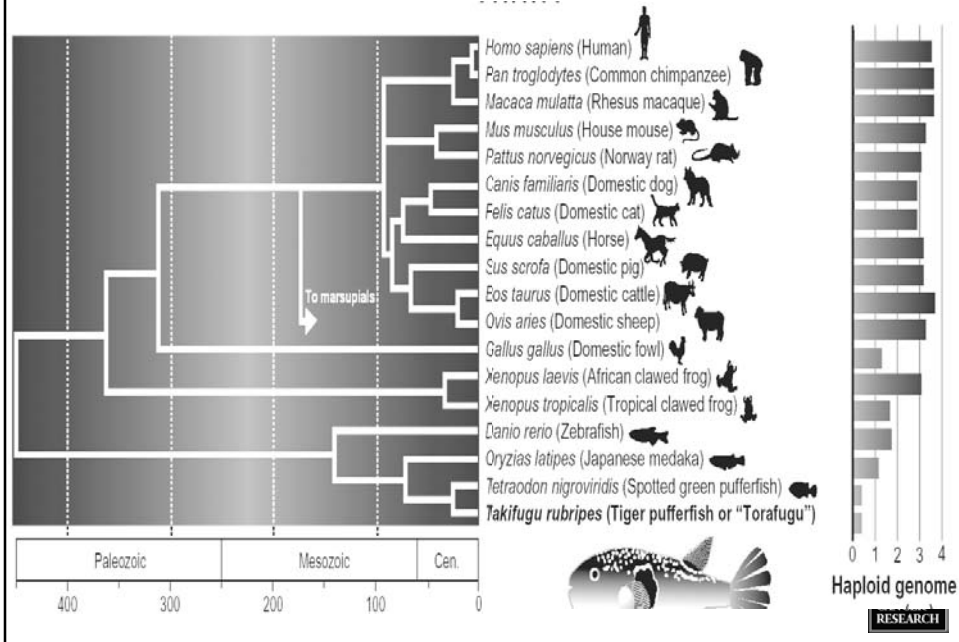


Genome Sequencing List

- Chimpanzee
- Bee
- Chicken
- *Xenopus tropicalis*
- Dog
- Cow
- *Drosophila simulans* and *Drosophila yakuba*
- Pig
- Macaque
- Kargaroo (Tamar Wallaby) ?



Animal Genome Projects



Other Genomes . . .

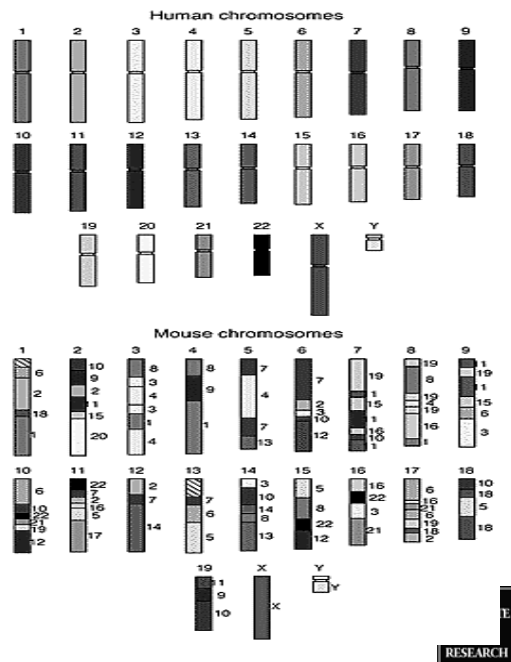
- 242 bacterial/ 18 archaeal/ 42 eukaryotic at NCBI
- Multiple Fungal projects, already a number of yeast genomes sequenced
- Mustard Weed *Arabidopsis thaliana* (125 Mb ~26,000 genes, 11% repeat, human 50%)
- 2 Rice genomes *Oryza sativa indica* & *japonica* (430 Mb, ~30,000 genes)
- 64% rice genes homologous to *Arabidopsis*



Why Compare Genomes?

- Powerful for gene prediction
- Helps understand evolution (species & chromosomal)
- Reconstruct gene family evolution
- Reconstruct phylogeny
- ~2% coding, ~98% non-coding
- ~70% intergenic, ~30% intronic
- **Non-coding sequences conserved between species are reliable guides to regulatory elements**

Human-Mouse Chromosome Map



- **Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences.**

Gene. 2002 Jul 10;294(1-2):35-44

- **Gilligan P, Brenner S, Venkatesh B.**



Chimpanzee Genome

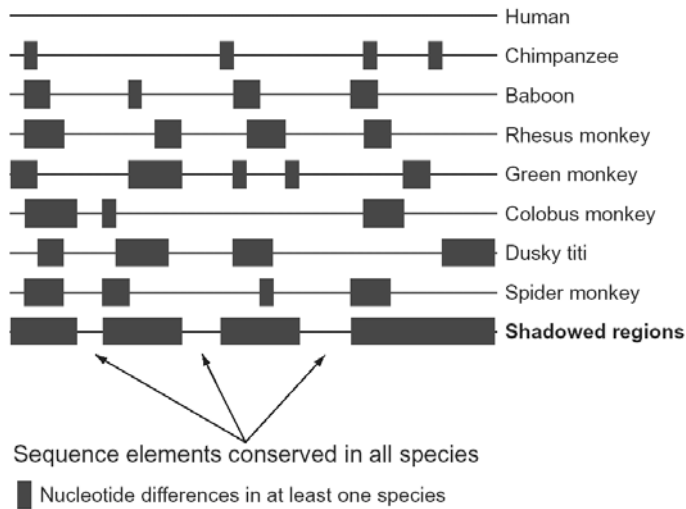
Almost human...



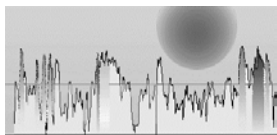
Sequencing the chimpanzee has emerged as a top genomic priority. David Cyranoski asks the chimp's champions what they hope to gain from studying the genome of our closest living relative.



Phylogenetic Shadowing



LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

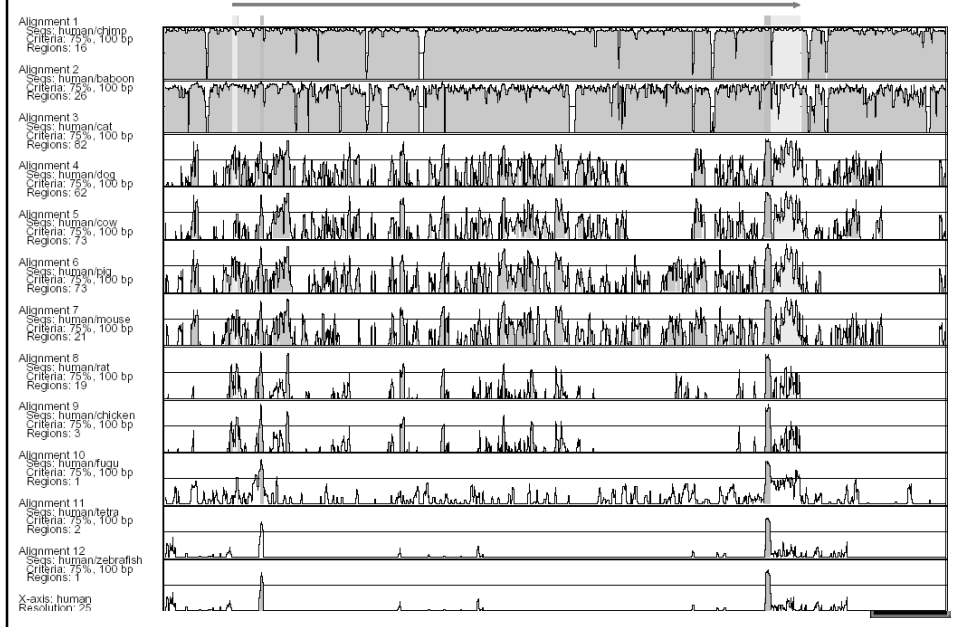


VISTA VISUALIZATION TOOLS FOR ALIGNMENTS

- VISTA is an integrated system for global sequence alignment and visualization, designed for comparative genomic analysis
- www-gsd.lbl.gov/vista
- Uses AVID global DNA alignments (Lior Pachter)

LUDWIG
INSTITUTE
FOR
CANCER
RESEARCH

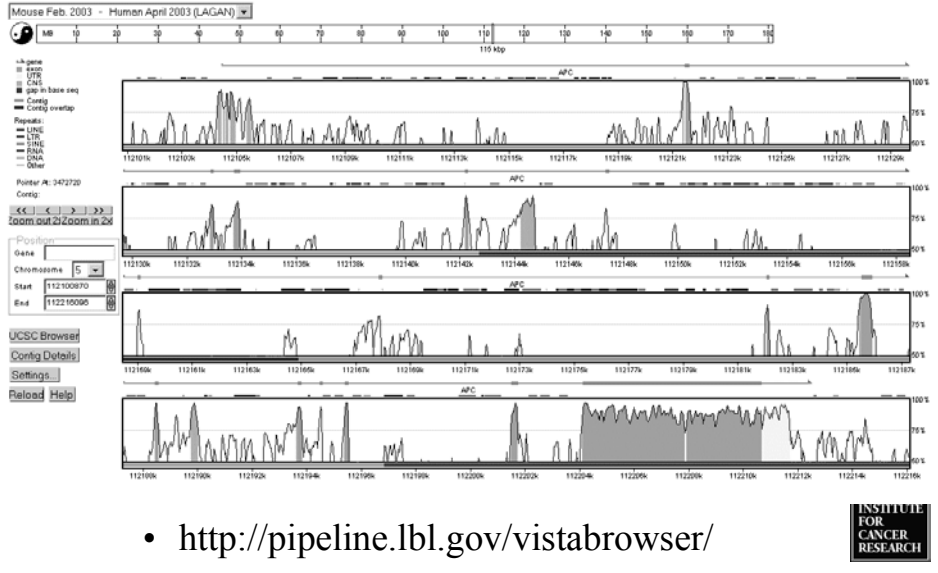
VISTA example



VISTA analysis

- Degree of conservation is variable between genes
- Highly conserved genes require comparisons with more distant organisms
- Less conserved genes are best compared with more closely related species

VISTA genome browser



rVISTA

Transcription Factor Binding Site Analysis

- use TRANSFAC to find possible TFBS
- Utilises TFBS clustering and interspecies conservation
- 95% eliminated by Hs/Mm VISTA conservation
- 88% of known sites located

<http://dcode.berkeley.edu/ecrBrowser/>
precomputed genome analysis for rVISTA

<http://dcode.berkeley.edu/rvista/>
perform your own rVISTA analysis

