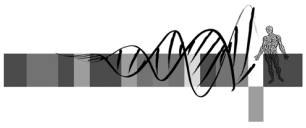


# Sequence Analysis

III:

Genomics and Genome Browsers

George Bell, Ph.D.
WIBR Bioinformatics and Research Computing

## Genomics and Genome Browsers

- Introduction to genomics
- Genomics with genome browsers
- Conservation and evolution
- Introduction to comparative genomics
- Genome-wide data analysis




## Genomics: some big questions

- What is a gene?
  - one definition: a region of DNA that encodes functional RNA or protein.
- What is the sequence of the genome? SNPs?
- Where are all of the genes?
- What are the proteins they encode? What do they do?
- Where's the regulatory sequence? What does it do?
- How can one integrate all of this information?




## The human genome

## The human genome

- Last assembly: July 2003

  3.2 billion bases, mostly complete
  Ensembl annotation: 23,531 genes; 31,609 transcripts
  Heterochromatin (light staining) is not sequenced

| | |
|---|---|
| Mean GC content: | 41% |
| Repetitive DNA: | 50% |
| Coding sequence: | 1.5% |
| Under selection: | 5% |

- Reference genome sequence comprises one strand of each chromosome.




## Identifying genes

- Optimal protocol: Collect all RNA from all cell types in all conditions, sequence it and map it to the genome.
- Practical protocols:
  - predict genes de novo
  - cluster ESTs
  - sequence full-length clones
  - search with known genes in another species
  - a combination of those techniques above
- Still problems with pseudogenes

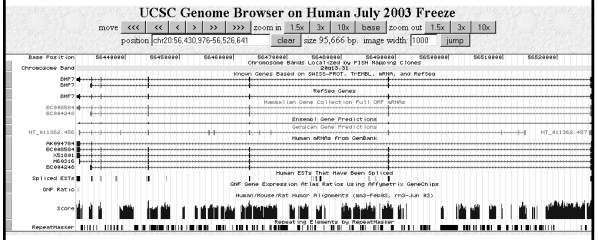
# How many genes and transcripts?

- Gene-centric databases (one entry per gene)
  - Ensembl (Hs=23,531; Mm=26,762)
  - LocusLink (37,497; 69,612) incl. other "stuff"
- Human-curated full-length cDNA resources (one entry per transcript)
  - RefSeq (21,150; 17,017)
  - Mammalian Gene Collection (11,196; 10,216)
- EST-centric clusters (one entry per cluster)
  - UniGene (118,517; 82,482)
  - TIGR Gene Indices (201,258; 145,559)




# Genome Browsers

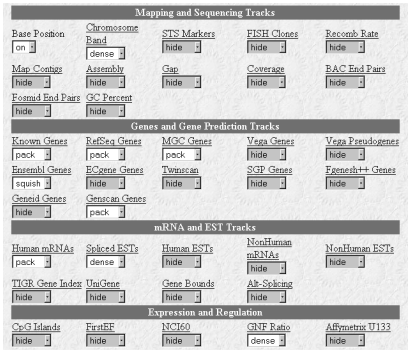
Examples: UCSC, Ensembl, NCBI, WIBR







# Genome Browser tracks







# Genome Browser data

- Potential to show any data that can be mapped to a genome.
- Visual examination can be more powerful than any automated analysis tool.
- Positive strand of reference chromosome is shown.
- Conventions: gene "start" < "end"
- Coordinates change with each assembly.
- Sequence is often soft- or hard-masked for repetitive DNA.



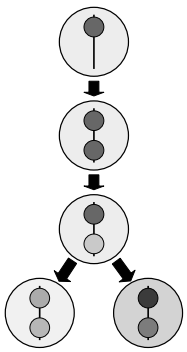

# Conservation and evolution

- Functional regions of a genome can be difficult to find in a large, repetitive sequence.
- During evolution, pressure for selection leads to greater conservation of some regions of a genome.
- Searching for regions of purifying selection is hoped to lead to elements of functional significance.




# Homology

- Genes are *homologous* if they arose from the same ancestor.
- Orthologs: homologs (in different species)
  that arose from a speciation event
- Paralogs: homologs (in the same species)
  that arose from a duplication event



# Quantifying evolution of coding regions

1. Percentage of AA identity or similarity
   For human-mouse orthologs, median identity = 79%
2. The $K_a/K_s$ ratio

$$\frac{\text{AA substitution rate}}{\text{Neutral substitution rate}} = \frac{\text{Non-synonymous substitution rate}}{\text{Synonymous substitution rate}}$$

   For human-mouse orthologs, median $K_a/K_s = 0.12$
   => 88% of AA-changing mutations are deleterious

- Domain-containing regions have evolved less.
- Pseudogenes have a $K_a/K_s$ ratio close to 1.

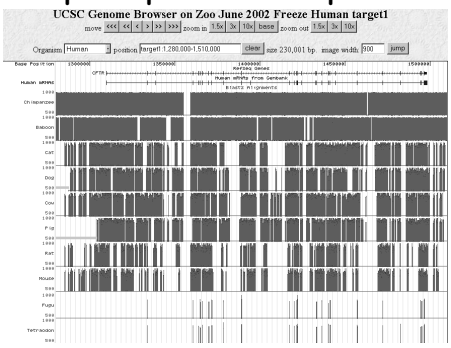



# Comparative genomics

- Conservation between genomes is a very effective way to identify genes and regulatory regions.
- Comparison of multiple genomes can identify functional elements without any previous understanding of their function.
- With increasing conservation of a region of interest, comparisons between more distant species becomes more informative.
- Comparison of two species is rarely as effective as that of multiple species.



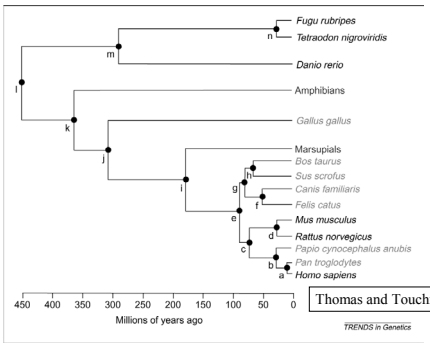

# Multiple-species comparisons







# Vertebrate sequencing projects




Thomas and Touchman, 2002




# Conserved synteny

Nature 420(6915):520-62, 2002

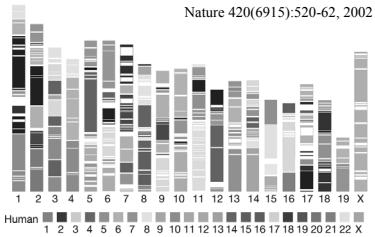



**Figure 3** Segments and blocks >300 kb in size with conserved synteny in human are superimposed on the mouse genome. Each colour corresponds to a particular human chromosome. The 342 segments are separated from each other by thin, white lines within the 217 blocks of consistent colour.




# Finding orthologous genes

- Traditional method 1: reciprocal best BLASTP hits in all vs. all searches
- Traditional method 2: synteny maps
- Current methods: sequence analysis and conserved synteny
- Resources:
  Ensembl, NCBI, genome browsers
- Complicated by paralogous genes

# What do all the genes do?

Q: How can every molecular function and biological process be systematically organized?

A: The Gene Ontology Consortium

- The three GO ontologies:
  - Molecular function
  - Biological Process
  - Cellular Component


- Components of the ontologies are like hierarchies except that a "child" can have more than one "parent".
- Evidence for annotation varies.




# Genome-wide data analysis

- Ensembl and UCSC genome downloads
- NCBI flat file downloads
- EnsMart for genome-wide queries on the web
- Ensembl and WIBR LocusLink for SQL queries
- Analyzing sequence vs. annotations
- Transitivity of sequences and annotations?
- Check with BaRC about data on their servers




# Summary

- Introduction to genomics
- Genomics with genome browsers
- Conservation and evolution
- Introduction to comparative genomics
- Genome-wide data analysis




# Selected references


- Initial sequencing and analysis of the human genome. Nature. 409:860-921, 2001.
- Initial sequencing and comparative analysis of the mouse genome. Nature. 420:520-62, 2002.
- A User's Guide to the Human Genome II. Nature Genetics. 35 Suppl 1:4, 2003. ("web special")





# Exercises

- Browsing for genomic information
- Extracting annotated genomic sequence
- Gene-finding with comparative mammalian genomics
- Gene and genome analysis through annotation
- Command-line applications