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

- Last assembly: May 2004 ("NCBI 35")
  - 3.0 billion bases, mostly complete
  - Ensembl annotation: 24,194 genes; 35,838 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=24,194; Mm=28,069)
  - LocusLink (32,688; 67,653) incl. other "stuff"
- Human-curated full-length cDNA resources (one entry per transcript)
  - RefSeq (23,534; 30,462)
  - Mammalian Gene Collection (17,747; 14,639)
- EST-centric clusters (one entry per cluster)
  - UniGene (52,888; 45,719)
  - TIGR Gene Indices (227,631; 161,499)

Genome Browsers

Examples: UCSC, Ensembl, NCBI, WIBR

Genome Browser tracks

Other groups:
- Expression and Regulation
- Comparative Genomics
- ENCODE Tracks
- Variation and Repeats

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.
- Paralogs: homologs (in the same species) that arose from a duplication event
- Orthologs: homologs (in different species) that arose from a speciation 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
   \[
   \begin{array}{c|c|c|c}
   & \text{AA substitution rate} & \text{Neutral substitution rate} & \text{Non-synonymous substitution rate} \\
   \hline
   \text{Synonymous substitution rate} & & & \\
   \end{array}
   \]
   For human-mouse orthologs, median K_a/K_s = 0.12
   \Rightarrow 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

Conserved synteny

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


Exercises

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