Introduction to Evolutionary and Functional Genomic Analysis

Cristian I. Castillo-Davis
Post-Doctoral Fellow
Department of Statistics
Harvard University
Cambridge, MA 02138
castillo-davis@stat.harvard.edu

The Computational Systems Bioinformatics Conference Workshop

Biological Synthesis

"Knowing more and more about less and less may mean that relationships are lost and that the grand pattern and great processes of life are overlooked."

- G.G. Simpson, 1944

With the current genomics revolution in biology we are poised to unite the organismic and molecular study of organisms.

Overview

• Introduction and audience background assessment
  • Case study introduction and motivation

Background and Resources

• Molecular biology overview
  • Genomic resources, techniques, and tools
• Functional genomics and expression data

Theory and Methods

• Comparative genomics
• Phylogenetics overview
• Molecular evolution
  • Coding
  • Noncoding
• Data integration - GeneMerge software

Getting Started/Doing it Yourself!

• Word and Excel Methods
• Introduction to Linux
• Introduction Perl
• GeneMerge
Case Study: Of Mice, Men, and Genomes

Based on:

What are the fastest and slowest evolving genes in the human genome?

- Do sex-related genes evolve more quickly?
- If not then what?
- What evolves the slowest?
- Previous studies based on a few genes, mostly in flies
- No rigorous genomic scale studies

Case Study:
What do we need?

- Every gene in the human genome
- Every gene in another related species, for comparison
- Must find the “same” gene in each
- Measure changes in all genes (find fast and slow evolving)
- Relate this information to biological data
  - What do these genes do? (digestion? transport? etc.)
  - Where do they do it? (kidney? brain? etc.)
Molecular Biology Overview

- DNA is a chemical “string” made up of the molecules A, T, C, G
- The sequence is 3 billion letters long (in humans)
- The sequence is broken up into meaningful parts, like a sentence

The genome is the complete list of coded instructions needed to make an organism

Cell Chromosomes DNA

ATGCCCTTTGATGGCTGCTGC ATGATTCAGTAG GTG... Protein-coding sequence regulatory sequences

Proteins are the workers of the cell

Some familiar examples
- collagen – structural component of skin
- lactase – digests milk sugar
Regulatory sequences are the executives of the cell

They tell the proteins
- where to go
- how many of them should go
- and when...

We control gene expression!

ATGCCCTTGTGCTGCTGC ATGATTCCAGTAG GTGG...

"messenger RNA"

string of amino acids

AA, AA, AA

Protein digest milk sugar

The genetic code

- DNA to amino acid code is a triplet code
  - 3 DNA nucleotide bases code for each amino acid (AA)

- The genetic code is redundant (61 triplets but only 20 AAs)
  - Start and Stop "codons" are present

<table>
<thead>
<tr>
<th>Redundancy</th>
<th>Special codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTT = Leucine</td>
<td>ATG = Start</td>
</tr>
<tr>
<td>CTA = Leucine</td>
<td>TAG = Stop</td>
</tr>
<tr>
<td>CTG = Leucine</td>
<td>TGA = Stop</td>
</tr>
<tr>
<td>CTC = Leucine</td>
<td>TAA = Stop</td>
</tr>
</tbody>
</table>

Source: Access Excellence, National Museum of Health
Transcription

mRNA splicing and processing

Translation

The ribosome directs the process of translation and links amino acids to each other.

Transfer RNAs have a unique anticodon and are "charged" with a different amino acid.

The ribosome translates the mRNA into proteins.
A folded amino acid chain is called a protein.

Genomic Methods, Techniques and Resources

Until recently only one or a few genes could be examined at a time. Recent genome sequencing projects have decoded entire DNA sequence (genome) of a few research organisms. Advent of microarrays (DNA chips) allow detection of gene expression level of every gene in the genome at a one time. Analysis of whole genomes in conjunction with microarray and other data allows us to test hypotheses at a general scale not imaginable only a few years ago.
Case Study: What do we need?
- Every gene in the human genome
- Every gene in another related species, for comparison
- Must find the “same” gene in each
- Measure changes in all genes (find fast and slow evolving)
- Relate this information to biological data
  - What do these genes do? (digestion? transport? etc.)
  - Where do they do it? (kidney? brain? etc.)

Where to obtain whole genome sequences
- There are several major databases that are “official” repositories of genetic data
  - mRNA sequences
  - protein sequences
  - predicted coding sequences
  - genomic sequence (whole chromosome sequences)
- NCBI
- Swiss-Prot/TrEMBL
- UCSC Genome Browser
- Organism-specific databases (WormBase, FlyBase, etc.)

Many journals require sequence submission to a public database... thus
- many versions of same gene in GenBank
- many errors
- Billions of sequences from every imaginable species!
- good place to search if you want to search “the world” of sequences
- not highly curated, but there are parts that are: RefSeq
Genome sequences can be downloaded via FTP.


<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023-03-14</td>
<td>10:00:00</td>
<td>100MB</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2023-03-14</td>
<td>10:00:00</td>
<td>200MB</td>
</tr>
</tbody>
</table>
Case Study:
What do we need?

• Every gene in the human genome
• Every gene in another related species, for comparison
• Must find the “same” gene in each
• Measure changes in all genes (find fast and slow evolving)
• Relate this information to biological data
  What do these genes do? (digestion? transport? etc.)
  Where do they do it? (kidney? brain? etc.)
Organism-specific and General Genomic Databases

- Most every model organism has its own online database
- Rich source of genomic information
  - curated gene sequences, functions and literature resources
  - depository for expression and other functional genomic data
  - some limited tools for genomic analysis
- Movement toward standardization
- General genomic databases are often organized around a topic
  - gene function
  - transcription factor sites
  - ligand binding / chemistry

<table>
<thead>
<tr>
<th>General Genomic Databases</th>
<th>Organism-specific Genome Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>FlyBase</td>
<td>Saccharomyces Genome Database (SGD)</td>
</tr>
<tr>
<td>SGD (yeast)</td>
<td>- repository of information on brewer’s yeast</td>
</tr>
<tr>
<td>TRANSFAC</td>
<td>- A search for gene name FC72 brings up this page</td>
</tr>
<tr>
<td>SubtiList</td>
<td>Strains</td>
</tr>
<tr>
<td>PlasmoDB</td>
<td>- standard gene name</td>
</tr>
<tr>
<td>MGI (mouse)</td>
<td>- aliases</td>
</tr>
</tbody>
</table>

Example 1: Saccharomyces Genome Database (SGD)
- repository of information on brewer’s yeast
- A search for gene name FC72 brings up this page

Strains
- standard gene name
- aliases
- English description
- known functions
- phenotype
- chromosome location
- genome coordinates
- other links

Similarity (image of the glio.)
Organism-specific Genome Databases

Example #2:
Flybase
- repository of information on the research fruit fly Drosophila melanogaster
- A search for gene name lace brings up this page
Shows
- standard gene name
- synonyms
- mutant phenotypes
- expression information
- cytogenetic map location
- genetic interactions
- text summary on the gene

Example of a General Genomic Database #1
Gene Ontology Consortium
- repository of information on gene function with controlled vocabulary and IDs for gene function for multiple species
- Main page contains link to the controlled vocabularies ("ontologies") and search area for GO Terms
Shows
- current ontologies
- current annotations
- links to ontology browsers
- software utilizing GO
- FTP archive

Example of a General Genomic Database #2:
Kyoto Encyclopedia of Genes and Genomes (Pathway data and diagrams)
Organism-specific and General Genomic Databases

- Almost every online database is useless for performing genomic analyses with built in “query tools”
- Fortunately, almost all databases make their information available as so-called “flat files” (text files) which can be downloaded
- If not, database curators will sometimes provide files if asked!
- Flat files can be downloaded over the web and parsed using
  - Word processors like Word
  - Spreadsheet programs like Excel
  - Unix/Linux shell utilities like awk and grep
  - Custom scripts and programs using PERL, PYTHON, or C
- The computational manipulation, organization, and analysis of large volumes of data is the heart of modern genomics

FlyBase
WormBase
PlasmoDB
SGD (yeast)
SubtiList
TRANSFAC
MGI (mouse)

DNA Microarrays

Allow one to measure the level of gene expression of thousands of genes at a time

1. Gene sequence is laid down in specific spot on a chip
2. mRNA is extracted from organism
3. mRNA labeled with a fluorescent dye
4. mRNA is washed over chip
5. mRNA sticks to complementary DNA spot and glows
6. Optical reading of spot intensity measures mRNA abundance

DNA Microarrays

cDNA arrays
1. DNA is obtained from cDNA clones, full-length or partial
2. Each gene is spotted on glass slides with tiny pins (arrayer)
3. Since spot size is variable can’t use spot intensity as indication of sample mRNA abundance
4. So… hybridize two mRNA samples and compare ratios (log₂[Cy5/Cy3])

Oligonucleotide arrays (Affymetrix)
1. DNA is built up one base at a time with great precision
2. Multiple 30mers per gene (probes and “probesets”)
3. Mismatch probes used to infer absolute mRNA abundance (500.3 ppm)
Expressed sequence tags (ESTs) and gene expression

- Extract mRNA from tissue
- Use poly-T primers to reverse transcribe (mRNA ⇆ DNA)
- PCR amplification
- DNA sequence fragments
- mRNA in highest abundance gets sequenced the most! → database

Protein-Protein Interaction Data

- What proteins interact with each other in the cell?
- Difficult to measure in vivo so…
  - Yeast-2-Hybrid method
    - Proteins from another species put into yeast "bait and prey" reporter construct
  - Large-scale experiments in Drosophila, C. elegans, others

High-throughput yeast-2-hybrid efforts catalog a significant fraction of (potential) protein-protein interactions and provide us with an "interactome" map → database

Comparative Genomics
Comparative Genomics — common ancestry + time

Case Study:
What do we need?

- Every gene in the human genome
- Every gene in another related species, for comparison
- Must find the “same” gene in each
- Measure changes in all genes (find fast and slow evolving)
- Relate this information to biological data
  - What do these genes do? (digestion? transport? etc.)
  - Where do they do it? (kidney? brain? etc.)

Orthology and Paralogy
How to Establish Orthology?

Since duplications can occur within a genome before or after speciation (species split) we need to make sure that we are comparing the "same" gene in each species.

There may have been gene loss in one species or incomplete genome sequencing...

"orthologous genes"
"paralogous genes"

Orthology – 3rd cervical vertebrae in human
3rd cervical vertebrae in chimp

How to Find Paralogs?

Duplicate genes can be found within a genome by comparing all genes within a genome to each other!

Genes that have "significant" similarity to each other are paralogs
"duplicates" "gene families"

Paralogy – 1st cervical vertebrae in human
3rd cervical vertebrae in human
But in practice, how does one compare sequences?

Sequence Alignment

*a measurement of "sameness"

- There are a variety of methods that can be used to align protein and DNA sequences... including "by eye!"
- Algorithms align sequences automatically using rules about how matches, mismatches, and gaps are "rewarded" or "penalized" ⇒ different alignments

<table>
<thead>
<tr>
<th>Orig. Seq</th>
<th>gaps penalized more</th>
<th>mismatch penalized more</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGCTTG</td>
<td>ATGCTTG-</td>
<td>ATGC-TTG</td>
</tr>
<tr>
<td>ATGTTTTG</td>
<td>ATGTTTTG</td>
<td>ATGTTTTG</td>
</tr>
</tbody>
</table>

- Example scoring scheme: match +4
  mismatch -4
  gap opening -10
  gap extension -1
- There is no one optimum scoring scheme... context/biology dependent!
**Sequence Alignment**

- Two main kinds of sequence alignment
  - "Global alignment" forces sequences to match from beginning to end
  - "Local alignment" finds best matching segment between two sequences

```
<table>
<thead>
<tr>
<th>Global</th>
<th>Local</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Dynamic programming - finds the best possible alignment... but is CPU intensive:

Needleman-Wunsch

Smith-Waterman

**Clustal**

- Global alignment algorithm and software for pairwise and *multiple alignment* of DNA and protein sequences
- Based on finding the "closest" sequence, then next closest, etc.
- Builds a dendrogram or "tree" of the sequences first
- Efficient and reasonably accurate
- Command-line and graphical interface versions
- What one might use to align sequences before evolutionary analysis (for example human and chimp)

**BLAST**

- A program that allows efficient searches of gene databases based on DNA or protein sequence similarity (alignment)
- Sequences that are similar to a given query sequence can be retrieved by "blasting"
- "Hits" against the database are returned in order of best to worst matches (LOCAL ALIGNMENT matches)
- "E-value" statistics are assigned to the probability of a sequence match based on database size
  - Longer and more similar matches are assigned a low E-value
- Not a good alignment algorithm per se (heuristic not optimal)
- Web-based and command-line programs available
Case Study:

What do we need?

• Every gene in the human genome
• Every gene in another related species, for comparison
• Must find the “same” gene in each
• Measure changes in all genes (find fast and slow evolving)

• Relate this information to biological data
  – What do these genes do? (digestion? transport? etc.)
  – Where do they do it? (kidney? brain? etc.)
Molecular Evolution

multiple hits

Since there are only 4 possible nucleotide bases at a given site, it is possible that a site could mutate many times but have the same nucleotide when observed. Thus we must correct for the possibility of "multiple hits".

Jukes Cantor (1969) single parameter model:
The rate of substitution from one nucleotide to any other is equal to $\alpha$

\[
p_{A(0)} = 1 \\
p_{A(1)} = 1 - 3\alpha \\
p_{A(2)} = (1 - 3\alpha)p_{A(1)} + 3\alpha[1 - p_{A(1)}]
\]

\[K = -\frac{3}{4}\ln\left(1 - \frac{4p}{3}\right), \text{where } p = \text{observed # nucleotide differences between two sequences}
\]

Example: $p = 18/163 = 0.110 \Rightarrow K = 0.119$ nucleotide substitutions per site

Molecular Evolution

transitions and transversions

Some types of mutations are more common than others...

- $C \leftrightarrow T$ and $G \leftrightarrow A$ are transitions (tr)
- rest are transversions (tv)
- $\kappa = tr/tv$

Transitions are more common than transversions, 2:1 or higher in most organisms (mutation bias)
Molecular Evolution

transitions and transversions

Some types of mutations are more common than others...

\[ \begin{array}{c}
C & \xrightarrow{\alpha} & T \\
\beta & \xrightarrow{\beta} & \beta \\
G & \xrightarrow{\alpha} & A \\
\end{array} \]

- \( \alpha = \) transitions
- \( \beta = \) transversions

Kimura 2-parameter model (Kimura 1980) takes this into account when estimating the number of nucleotide substitutions per site.

Many other complex models with greater numbers of parameters...

Molecular Evolution - Protein change

Nonsynonymous vs. Synonymous Substitution

The genetic code is redundant...

Example:
- CTT = Leucine
- CTA = Leucine
- CTG = Leucine
- CTC = Leucine

- nonsynonymous substitution (replacement)
- synonymous substitution (silent)

Molecular Evolution - Protein change

Nonsynonymous vs. Synonymous Substitution Rate

Given the genetic code, there are not an equal number of possible synonymous and nonsynonymous "targets."
- all second position substitutions result in an AA change
- some first position and third position substitutions are silent
- multiple hits and "paths" to change complicate things further

Nei and Gojobori (1986) develop a method to estimate the number of nonsynonymous and synonymous substitutions per site using Jukes-Cantor correction etc. ...

- \( d_\alpha = \) nonsynonymous substitution rate (\( K_A \))
- \( d_s = \) synonymous substitution rate (\( K_S \))
Molecular Evolution - Protein change

Codon Bias

- Codons that match the most abundant tRNAs in a cell are "preferred" in highly expressed genes since they make translation faster, more efficient, or less error prone.
- Selection on translation of proteins and therefore 3rd codon usage is biased.

Thus highly expressed genes show less synonymous change!

<table>
<thead>
<tr>
<th>codon</th>
<th>most common codon for leucine</th>
<th>in highly expressed genes is</th>
<th>tRNA copies in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTT</td>
<td>CTA = Leucine</td>
<td>CTT = Leucine</td>
<td>CTT = Leucine</td>
</tr>
<tr>
<td>CTA</td>
<td>CTA = Leucine</td>
<td></td>
<td>CTA = Leucine</td>
</tr>
<tr>
<td>CTG</td>
<td>CTG = Leucine</td>
<td></td>
<td>CTG = Leucine</td>
</tr>
<tr>
<td>CTC</td>
<td>CTC = Leucine</td>
<td></td>
<td>CTC = Leucine</td>
</tr>
</tbody>
</table>

Spearman Rank Correlation

\[ r_s = -0.30, \ p < 10^{-117} \]

Codon bias is significantly correlated with mRNA expression genome-wide in *C. elegans*.


- Given all these possibilities, what we'd like is to incorporate all parameters in a complete model to infer the evolution of protein coding DNA sequences.
- Goldman and Yang (1996)
- Markov process model of change between all 60 codons represented as a transition matrix dependent on the following parameters:

\[ dN \] - nonsynonymous substitution rate
\[ dS \] - synonymous substitution rate
\[ \kappa \] - transition/transversion ratio
\[ \pi_{c1}, \pi_{t1}, \pi_{a1}, \pi_{g1} \] - 1st position codon nucleotide frequency
\[ \pi_{c2}, \pi_{t2}, \pi_{a2}, \pi_{g2} \] - 2nd position nucleotide frequency
\[ \pi_{c3}, \pi_{t3}, \pi_{a3}, \pi_{g3} \] - 3rd position nucleotide frequency
Molecular Evolution - Protein change
Codon Based Models of Molecular Evolution

• Ziheng Yang's Phylogenetic Analysis by Maximum Likelihood (PAML) package implements the codon-based model discussed (Yang 2000).
  - uses maximum likelihood procedure to estimate parameters
  - command-line program (Windows, Linux)
  - most sophisticated/accurate measurement of DNA evolution
  - many options and possibility for user control
  - simulate DNA evolution
  - tests for positive selection

<table>
<thead>
<tr>
<th>g</th>
<th>f</th>
<th>f0</th>
<th>f1</th>
<th>f2</th>
<th>k1</th>
<th>k2</th>
<th>k3</th>
<th>k4</th>
<th>k5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.37539</td>
<td>0.00882</td>
<td>0.04963</td>
<td>0.00854</td>
<td>0.72535</td>
<td>0.13818</td>
<td>0.70774</td>
<td>0.11850</td>
<td>43.64</td>
<td>94.36</td>
</tr>
</tbody>
</table>

Molecular Evolution - Noncoding sequence evolution

• most proteins are present and conserved across species
• 99.7% amino-acid identity between human and chimpanzee
• changes in gene expression associated with profound phenotypic changes
• regulatory changes may play a major role in the evolution of phenotypic diversity

understanding cis-regulatory evolution is crucial!
Molecular Evolution - Noncoding sequence evolution

- motifs are rare and/or unknown (Ex: only 12bp in 500 bp matters!)
- abundant neutral variation hides important changes
∴ functional changes are difficult to characterize

TF
aatttctgcatctactgttatttctgcatctact
aatttctgcatctactgttatttctgcatctact

Measuring functionally important regulatory sequence divergence

* Simple alignment methods often fail to detect conservation of important regulatory regions.

* Proper functioning of regulatory regions can be flexible with respect to the order, spacing, and even orientation of motifs (Ptashne & Gann 1997; Ludwig 2000)

* Why not look for “motifs” without respect to
  - global alignment
  - strict identity
  - order
  - orientation?

Shared motif method

1. Find the best sub-alignment between noncoding sequences using derivation of Smith-Waterman algorithm w/o respect to order, spacing, orientation
2. Find next best region using Smith-Waterman
3. Iterate until reach a minimum size
4. Calculation of fraction shared “motifs” —> $d_{SM} = 1 - \text{fraction of shared motifs}$

Shared motif divergence ($d_{SM}$) measures functional regulatory evolution by comparative sequence analysis alone.

62 of 79 experimentally verified Human/Mouse motifs (78%) were found within conserved regions identified by the SMM.

- Arabidopsis/Brassica
- C. elegans/C. briggsae
- D. mel/D. pseudo


**Case Study Question**

What are the fastest and slowest evolving genes in the human genome? Statistical overrepresentation?

**Our Approach**

- Genome-wide data, thousands of genes
- Rate of amino-acid replacement
- Rate of insertion/deletion
- Functional genomic data
Case Study Methods

Sequence
- *H. sapiens* GenBank Release 30
- *M. musculus* GenBank CDS

Alignment
- ClustalW
- AA alignment
- back-translation to nucleotides

Molecular Evolutionary Analysis
- Maximum likelihood codon-based model (PAML: Yang 2000)
- mouse-human pairwise

Functional Genomics Analysis
- Gene Ontology annotations
- EST tissue libraries for expression specificity

![Figure 1: dN/dS, n = 8000](image1.png)
8000 orthologs  fast/slow evolving genes

- top 10%
- lower 10%
- 0% indel

proportion of particular functional classes among 8000 orthologs
GeneMerge is a versatile genomics program that can be used to analyze a wide range of functional genomic data. It returns functional/categorical information for all types of genomic data and performs statistical tests for over-representation. For more information, visit http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge/

Castillo-Davis & Hartl 2003, Bioinformatics 19(7):891-8920
dN/dS, n = 8000

- Top 10% fastest-evolving gene names
- All examined genes
- Function/Category
### Functional enrichment of top 10% fastest evolving genes

#### Lowest dN (lower 10%)

<table>
<thead>
<tr>
<th>GO Term</th>
<th>PopSize</th>
<th>N-Freq</th>
<th>P-Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006446</td>
<td>9/659</td>
<td>0.00260</td>
<td>0.0024060</td>
<td>translational regulation, initiation</td>
</tr>
<tr>
<td>GO:0006350</td>
<td>9/659</td>
<td>0.00305</td>
<td>0.0125789</td>
<td>transcription</td>
</tr>
<tr>
<td>GO:0006913</td>
<td>12/659</td>
<td>0.00382</td>
<td>0.0005430</td>
<td>nucleocytoplasmic transport</td>
</tr>
<tr>
<td>GO:0007264</td>
<td>28/659</td>
<td>0.01314</td>
<td>2.97E-006</td>
<td>small GTPase mediated signal transduction</td>
</tr>
<tr>
<td>GO:0008152</td>
<td>254/658</td>
<td>0.31511</td>
<td>0.0118111</td>
<td>metabolism</td>
</tr>
<tr>
<td>GO:0006511</td>
<td>22/659</td>
<td>0.01023</td>
<td>0.0001019</td>
<td>ubiquitin-dependent protein catabolism</td>
</tr>
<tr>
<td>GO:0006440</td>
<td>1/660</td>
<td>1.98E-010</td>
<td>ribosome</td>
<td></td>
</tr>
<tr>
<td>GO:0006412</td>
<td>48/659</td>
<td>1.19E-012</td>
<td>protein biosynthesis</td>
<td></td>
</tr>
<tr>
<td>GO:0006886</td>
<td>43/659</td>
<td>2.62E-010</td>
<td>intracellular protein transport</td>
<td></td>
</tr>
<tr>
<td>GO:0005665</td>
<td>5/660</td>
<td>0.0157550</td>
<td>DNA-directed RNA polymerase II subunit</td>
<td></td>
</tr>
<tr>
<td>GO:0005794</td>
<td>20/660</td>
<td>0.0229683</td>
<td>Golgi apparatus</td>
<td></td>
</tr>
<tr>
<td>GO:0005718</td>
<td>5/660</td>
<td>7.95E-002</td>
<td>nucleosome</td>
<td></td>
</tr>
<tr>
<td>GO:0008151</td>
<td>349/659</td>
<td>0.0010793</td>
<td>cell growth and/or maintenance</td>
<td></td>
</tr>
<tr>
<td>GO:0005839</td>
<td>7/660</td>
<td>0.0088780</td>
<td>20S core proteasome</td>
<td></td>
</tr>
<tr>
<td>GO:0005884</td>
<td>8/660</td>
<td>3.33E-005</td>
<td>actin filament</td>
<td></td>
</tr>
<tr>
<td>GO:0005681</td>
<td>9/660</td>
<td>1.48E-005</td>
<td>spliceosome</td>
<td></td>
</tr>
<tr>
<td>GO:0005634</td>
<td>121/660</td>
<td>1.41E-005</td>
<td>nucleus</td>
<td></td>
</tr>
<tr>
<td>GO:0005840</td>
<td>24/660</td>
<td>3.90E-010</td>
<td>ribosome</td>
<td></td>
</tr>
</tbody>
</table>

#### GO Term Pop-frc S-frc P-Value Description

- GO:0005840: 24/660, 3.90E-010, ribosome
- GO:0005634: 121/660, 1.41E-005, nucleus
- GO:0005681: 9/660, 1.48E-005, spliceosome
- GO:0005884: 8/660, 3.33E-005, actin filament
- GO:0005839: 7/660, 0.0088780, 20S core proteasome
- GO:0005665: 5/660, 0.0157550, DNA-directed RNA polymerase II subunit
- GO:0005794: 20/660, 0.0229683, Golgi apparatus
- GO:0005718: 5/660, 7.95E-002, nucleosome
- GO:0008151: 349/659, 0.0010793, cell growth and/or maintenance
- GO:0006412: 48/659, 1.19E-012, protein biosynthesis
- GO:0006886: 43/659, 2.62E-010, intracellular protein transport
- GO:0005665: 5/660, 0.0157550, DNA-directed RNA polymerase II subunit
- GO:0005839: 7/660, 0.0088780, 20S core proteasome
- GO:0005840: 24/660, 3.90E-010, ribosome
- GO:0005634: 121/660, 1.41E-005, nucleus
- GO:0005681: 9/660, 1.48E-005, spliceosome
- GO:0005884: 8/660, 3.33E-005, actin filament
- GO:0005839: 7/660, 0.0088780, 20S core proteasome
- GO:0005665: 5/660, 0.0157550, DNA-directed RNA polymerase II subunit
- GO:0005794: 20/660, 0.0229683, Golgi apparatus
- GO:0005718: 5/660, 7.95E-002, nucleosome
- GO:0008151: 349/659, 0.0010793, cell growth and/or maintenance
- GO:0006412: 48/659, 1.19E-012, protein biosynthesis
- GO:0006886: 43/659, 2.62E-010, intracellular protein transport
- GO:0005665: 5/660, 0.0157550, DNA-directed RNA polymerase II subunit
- GO:0008151: 349/659, 0.0010793, cell growth and/or maintenance

*Figures depict histograms and tables showing functional enrichment of the top 10% fastest evolving genes.*
### Genes with 0% insertion/deletions

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Pop frac</th>
<th>Study frac</th>
<th>P-Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:003735</td>
<td>0.01146</td>
<td>49/2319</td>
<td>5.68E-005</td>
<td>structural constituent of ribosome</td>
</tr>
<tr>
<td>GO:0005843</td>
<td>0.00214</td>
<td>13/2319</td>
<td>0.0017939</td>
<td>cytosolic small ribosomal subunit (sensu Eu k)</td>
</tr>
<tr>
<td>GO:0005842</td>
<td>0.00260</td>
<td>14/2319</td>
<td>0.0138429</td>
<td>cytosolic large ribosomal subunit (sensu Eu k)</td>
</tr>
<tr>
<td>GO:0005794</td>
<td>0.01314</td>
<td>56/2319</td>
<td>2.34E-006</td>
<td>Golgi apparatus</td>
</tr>
<tr>
<td>GO:0005739</td>
<td>0.02887</td>
<td>98/2319</td>
<td>0.0003915</td>
<td>mitochondrion</td>
</tr>
<tr>
<td>GO:0005747</td>
<td>0.00183</td>
<td>11/2319</td>
<td>0.0387217</td>
<td>actin cytoskeleton reorganization</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>0.02138</td>
<td>89/2319</td>
<td>5.17E-009</td>
<td>protein biosynthesis</td>
</tr>
<tr>
<td>GO:0006886</td>
<td>0.02016</td>
<td>85/2319</td>
<td>5.61E-009</td>
<td>intracellular protein traffic</td>
</tr>
<tr>
<td>GO:0006446</td>
<td>0.01314</td>
<td>19/2319</td>
<td>0.0017939</td>
<td>cell growth and/or maintenance</td>
</tr>
<tr>
<td>GO:0005794</td>
<td>0.01314</td>
<td>56/2319</td>
<td>2.34E-006</td>
<td>structural constituent of ribosome</td>
</tr>
<tr>
<td>GO:0008151</td>
<td>0.44539</td>
<td>1188/2319</td>
<td>3.53E-013</td>
<td>cell growth and/or maintenance</td>
</tr>
<tr>
<td>GO:0006397</td>
<td>0.00183</td>
<td>11/2319</td>
<td>0.0387217</td>
<td>ATP biosynthesis</td>
</tr>
<tr>
<td>GO:0007204</td>
<td>0.01314</td>
<td>19/2319</td>
<td>2.24E-007</td>
<td>small GTPase mediated signal transduction</td>
</tr>
</tbody>
</table>

### Mining EST data for tissue expression specificity information

- **Since annotations are not complete or necessarily unbiased, explore gene function using other functional genomic data...**
- **Over 500,000 ESTs deposited in GenBank derived from various mouse tissue "libraries."**

**How to use it?**
- Organize EST sequences by tissue
- BLAST every gene in mouse genome against each tissue’s ESTs
- Count significant hits to each organ
- Decide if a given gene is heart-specific, kidney-specific, thymus/spleen co-expressed, etc. based on a clustering algorithm of some sort
Mining EST data for tissue expression specificity information

Table of significant* BLAST hit results

<table>
<thead>
<tr>
<th>Gene</th>
<th>Heart</th>
<th>Lung</th>
<th>Kidney</th>
<th>Thymus</th>
<th>Spleen</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_00001</td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NM_00002</td>
<td>1</td>
<td>16</td>
<td>15</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>NM_00003</td>
<td>2</td>
<td>14</td>
<td>9</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>NM_00004</td>
<td>2</td>
<td>14</td>
<td>9</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>NM_00005</td>
<td>1</td>
<td>14</td>
<td>9</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>NM_00006</td>
<td>0</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>NM_00007</td>
<td>0</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

* >100 bp & >90% & E-value < 10^-20

• How do we decide what is “tissue-specific?”
• If we cluster based on just the “profile” we ignore the magnitude information and need to pre-specify # of clusters (Self Organizing Map)
• If we cluster based on just the magnitude we ignore the profile information (Hierarchical clustering)
• SOTA - Self-Organizing Tree Algorithm (Herrero et al. 2000)
  one approach to combine best of both methods

Tissue specific expression based on EST data and SOTA Analysis

mean dN
(95% CI by
nonparametric
bootstrap,
1,000 replicates)
Case Study Conclusions

- Slowest evolving genes in mammalian genome are involved in core molecular metabolism (either by % indel or AA subs)
- Fastest evolving genes in mammalian genome are not those under sexual selection but are immune related, cell surface receptors/ligand signal transduction
- EST expression data show fastest evolving proteins are co-expressed in the thymus and spleen, corroborating the annotation data


Phylogenetic Trees

Historical Perspective

Criteria
- Subjective similarity (500 B.C.)
- Evolutionary systematics (1900s)
- Willi Hennig and the cladistic method (1966)
- Principle of Parsimony and Occam's Razor

- Wedding party! Who's related to whom?
- Raw similarity vs. special similarity
  - synapomorphy = special homology

Phylogenetic Trees
Phylogenetic Trees

<table>
<thead>
<tr>
<th>species</th>
<th>chlorophyll</th>
<th>vascular system</th>
<th>seeds</th>
<th>flowers</th>
<th>ray and disc flowers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

a "character matrix"

Most parsimonious explanation of the data

chlorophyll     vascular system     seeds     flowers     ray and disc flowers
minimize hypothesized independent evolutionary events...

Most parsimonious explanation of the data

chlorophyll     vascular system     seeds     flowers     ray and disc flowers
the evolution of seeds, their loss, and re-evolution is not most parsimonious
Some Problems with Parsimony
"long-branch attraction"

When branch lengths (mutation or time passed) are large, many multiple hits occur and the two related sequences are essentially randomized with respect to each other. Thus the probability of two sites being the same is 0.25!

If faster mutation along two lineages the long-branches "attract"

Maximum Likelihood
Models of Phylogenetics

Using an explicit model of sequence evolution, branch-length (mutation rate) and various tree topologies are simultaneously explored to yield the topology and combination of parameters that maximize the likelihood of the data.

Computationally intensive (>20 taxa), parameter space large, heuristic methods

Getting Started/Doing It Yourself!

- Word and Excel Techniques
- Introduction to Linux
- Introduction Perl
- Using GeneMerge

Further handouts and a CD with software will be provided...