Lecture 5  
Whole-genome comparative genomics (part II - Evolution)

6.047 / 6.878 / HST.507 : CompBio: Genomes, Networks, Evolution

The age of comparative genomics

- human
- chimp
- mouse
- rat
- dog
- 14 yeasts
- 12 flies
- opossum
- armadillo
- rabbit
- cow
- hyrax
- elephant
- bat
- dolphin
- lemur
- bushbaby
- pika
- hedgehog
- tenrec
- lama
- Tree shrew
- pangolin
- 32 mammals

Questions for phylogenetic analyses

- Panda
  - Bear or raccoon?
- Out of Africa
  - mitochondrial evolution story?
- Human evolution
  - Did we ever meet Neanderthal?
- Primate evolution
  - Are we chimp-like or gorilla-like?
- Vertebrate evolution
  - How did complex body plans arise?
- Recent evolution
  - What genes are under selection?

Our goals for today

- Mechanisms of genome evolution
  - Resolving region correspondence
  - Rapid and slow evolution
- Whole genome duplication
  - Evolutionary signature of WGD
  - Emergence of new functions

Determining corresponding regions

Part I: Using genomics to study evolution

Part II: Using evolution to study genomes

Part I
1. Yeast
2. Alignment
3. Evolution
4. Duplication

Part II
1. Genes
2. Regulation
3. Grammar
4. Human
Goal: Find corresponding regions

• Aligning four genomes: methodology
  – Anchor genomic segments by the genes they contain
  – Resolve the correspondence of genes for each pair of species
  – Construct nucleotide-level alignment

• What makes it hard?
  – Not all regions have one-to-one correspondence
  – Gene divergence, duplication and loss.
  – Genome rearrangements.

• What information is available
  – Amino-acid similarity of every gene pair across genomes
  – Locations of genes in each genome

Framework: graph of gene correspondence

• Weighted bipartite graph
  – Graph represents gene correspondence
  – Nodes: genes (w/ coordinates)
  – Edges: sequence similarity (w/ weights)

• Two types of evolutionary relationships
  – Orthologs (1-to-1 matches)
  – Paralogs (1-to-many / many-to-many)

• Method
  – Eliminate spurious edges (simplify graph)
  – Select edges based on available information
    • Blocks of conserved gene order
    • Protein sequence similarity

Definition: Best Unambiguous Subgroups (BUS)

<table>
<thead>
<tr>
<th>Species</th>
<th>Species</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>80</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>60</td>
<td>80</td>
</tr>
</tbody>
</table>

Extend concept of best-bidirectional hits

Conservation of gene order (synteny)

Preferentially select edges in synteny blocks

Applying method to genome (6000 x 6000)

We are comparing orthologous regions
Conservation of local gene order and spacing

Genome alignment: Contributions

• BUS algorithm (Best Unambiguous Subgroups)
  – Resolve correspondence of genes and regions
  – Uses complete bipartite graph connectivity
  – Integrates protein similarity and gene order information

• Correctly resolved gene correspondence
  – More than 90% of genes have 1-to-1 correspondence
  – Identified regions and protein families of rapid change

Our goals for today

• Mechanisms of genome evolution
  – Resolving region correspondence
  – Rapid and slow evolution

• Whole genome duplication
  – Evolutionary signature of WGD
  – Emergence of new functions

Specific regions of rapid evolution

• HXT, FLO, COS, PAU, YRF families

• 80% of ambiguities in 5% of the genome

• 31 of 32 telomeres in ambiguity clusters

Identify regions of rapid change

Protein family expansions in chromosome ends
**Fast and slow evolving genes**

- **Percent amino acid identity**

YBR184W

13% aa identity

100

MatA2

Spore

- **Dynamic view of a changing genome**

- **Role in meiotic recombination - mating type switching?**

**Mutation rates by functional classification**

- **Ribosomal proteins**

- **Mitochondrial ribosomal proteins**

**Protein domain evolution**

- **DNA-binding domain**

- **Dimerization domain**

- **Gal4**

**Rapid protein change**

- **Protein domain creation**
  - G/N stretches
  - Protein-protein interaction

- **Compensatory frame-shifts**
  - Explore new reading frames
  - RNA editing signals

- **Stop-codon variation**
  - Gain enables rapid change
  - Loss explores new diversity
  - Read-through is regulated

- **Intein gain**
  - Recent, present in *S. cerevisiae* only

**Differences in gene content**

- **8-10 genes unique to each genome**
  - Metabolism, regulation/silencing, stress

- **Changes in gene dosage**
  - 10-20 tandem duplications (1-2 genes)
  - 2 segment duplications (5-6 genes)

- **Protein family expansions**
  - 211 genes (3%) with ambiguous correspondence
  - Paralog duplication and/or loss

**TFP1 intein is a recent event**

|===============================| 99 UNCLASSIFIED PROTEINS (2399 ORFs) |
|=============================| 05 PROTEIN SYNTHESIS (359 ORFs)     |

**05.01 ribosome biogenesis (215 ORFs)***

| Ribosomal proteins |

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mitochondrial ribosomal proteins</td>
</tr>
</tbody>
</table>

**Least conserved**

**Most conserved**

**Evolutionary shortcuts apparent in recent evolution**
Gene loss / Gene conversion

- Observe positions of paralogs in sensu stricto to identify recently lost duplicates
  - Two copies in S. bayanus, one copy in S. cerevisiae. Recently lost in S. cerevisiae lineage
  - One copy in each genome, different chromosomes. Recently lost independently in both genomes
- Observe rates of change for both paralogs

Transposon locations are conserved

- Transposons are active
  - Full-length Ty elements are recent
  - Typically appear in only one genome
- Transposon locations are conserved
  - Recent insertions reuse old loci
  - LTR remnants found in other genomes

Evolutionary advantage of transposons?

- Studied 8 strains resulting from experimental evolution
  - 3 strains duplicate Chr4R, containing HXT genes
  - 3 strains display extensive overlapping Chr5R deletions
  - 3 strains reuse same breakpoint on Chr14R
- Population advantage in transposon location
  - Ability to mediate reversible rearrangements

Our goals for today

- Mechanisms of genome evolution
  - Resolving region correspondence
  - Rapid and slow evolution
- Whole genome duplication
  - Evolutionary signature of WGD
  - Emergence of new functions
Genome Duplication

Further back in evolutionary time

- S. cerevisiae
- S. paradoxus
- S. mikatae
- S. bayanus
- K. waltii

Ability to ask different set of questions

Gene correspondence

Sister regions show gene interleaving

Duplicate mapping tiles K. waltii

Few genes remain in 2 copies

Gene interleaving is evidence of complete duplication
Duplicate mapping of centromeres

Recognize sister regions solely based on gene order

Duplicate mapping tiles S. cerevisiae

145 blocks cover 88% of genome

Conclusion: Whole Genome Duplication has happened

Whole Genome Duplications are everywhere!

Yeast Duplication
- Most genes 1-to-1 mapping
- Gene interleaving evidence of duplication
- Complete tiling of the genome

Vertebrate Duplication in Fish
- Fish: Gene order not conserved, only chromosomes
- Mammals: Gene order conserved, not chromosomes

Two rounds of WGD in base of vertebrate lineage
- Build clusters of related genes (use Ciona as outgroup)
- Count duplications by reconciliation
- Find regions of duplicate overlap \( \rightarrow \) 4-way synteny

Evidence of duplication using a single genome?

Genome duplication evidence in a single species

Genomic evidence
- Conserved order of paralogous genes
- Same transcriptional orientation

However
- Interspersed with single-copy genes

Interpretation: Genome duplication followed by gene loss
Whole genome duplication is controversial

- Insufficient evidence
  - Only 50% of genome in duplicate regions
  - Only 8% of genes present in two copies
  - Extensive redundancy outside duplicate regions

- Evidence against WGD
  - Divergence-based dating show multiple times
  - Other species have similar level of redundancy

- Alternative evolutionary scenario proposed
  - Independent segmental duplications
  - Also consistent with the evidence

Conclusion: Whole Genome Duplication has happened

Our goals for today

- Mechanisms of genome evolution
  - Resolving region correspondence
  - Rapid and slow evolution

- Whole genome duplication
  - Evolutionary signature of WGD
  - Emergence of new functions

- Phylogenomics
  - Distinguishing orthologs and paralogs
  - Machine learning approach to phylogeny

Whole-genome duplication resolved

- 457 genes kept in two copies, result of selection
  - Involved in sugar metabolism and fermentation

Fate of duplicated genes

Number of genes

<table>
<thead>
<tr>
<th>Time</th>
<th>5,000</th>
<th>10,000</th>
<th>~500 gained</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 Myrs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Today</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evidence of accelerated protein divergence?
**Scenarios for rapid gene evolution**

- **One copy faster**
  - Scer - copy1
  - Kwaf
  - Ohno, 1970

- **Both copies faster**
  - Scer - copy1
  - Scer - copy2
  - Kwaf
  - Lynch, 2000

20% of duplicated genes show acceleration
95% of cases: Only one copy faster

**Emerging gene functions after duplication**

- **Origin of replication → silencing**
  - 4-fold acceleration
    - Scer - Sir3 (silencing)
    - Kwaf - Orc1

- **Translation initiation → anti-viral defense**
  - 3-fold acceleration
    - Scer - Hbs1 (translation initiation)
    - Kwaf - Hbs1

Asymmetric divergence → recognize ancestral / derived

**Distinct functional properties**

<table>
<thead>
<tr>
<th></th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene deletion</td>
<td>Lethal (20%)</td>
<td>Never lethal</td>
</tr>
</tbody>
</table>

Gain new function and lose ancestral function

**Distinct functional properties**

<table>
<thead>
<tr>
<th></th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene deletion</td>
<td>Lethal (20%)</td>
<td>Never lethal</td>
</tr>
<tr>
<td>Expression</td>
<td>Abundant</td>
<td>Specific</td>
</tr>
<tr>
<td></td>
<td>(stress, starvation)</td>
<td></td>
</tr>
<tr>
<td>Localization</td>
<td>General</td>
<td>Specific</td>
</tr>
<tr>
<td></td>
<td>(mitochondrion, spores)</td>
<td></td>
</tr>
</tbody>
</table>

Gain new function and lose ancestral function

**Decelerated evolution**

- 60 gene pairs (13% of 457 pairs)
  - 98% protein identity (all pairs: 55%)
  - 90% identity in 4-fold degenerate sites (all pairs: 41%)

- **Not recent duplication**
  - Gene order argues ancestral WGD pairs

Gene conversion?

**Evidence of gene conversion**

- Tree root reveals time of duplication
  - No acceleration in the K. walti branch
  - The two genes have recently replaced each other
- Branching order reveals gene conversion
  - Paralogs are closer to each other than to their ortholog
  - Both S. cerevisiae and S. bayanus show gene conversion

Periodic gene conversion
Evolutionary genomics in yeast

- Genome ancestry resolved
  - Whole-genome duplication
  - Massive gene loss

- Emergence of new functions
  - Asymmetric acceleration
  - Ancestral and derived functions
  - Repository for buffering mutations

Gene duplication and network growth

Functions of duplicated genes

- As a group
  - Biased towards environment adaptation
  - Sugar metabolism, fermentation, regulation

- Individual pairs
  - Are new gene functions gained by WGD?
  - How are new gene functions emerging?

Evidence of accelerated protein divergence?

Asymmetry also found in network connectivity

Interaction loss more likely than gain.
One protein maintains ancestral function?

Study network in context of duplication

Network evolution by duplication

Scenario 1
- Ancestral network motif

Scenario 2
- Modern network motif
Mechanisms of network motif emergence

- Duplication
- Divergence
- Creation Probability: $p \cdot (1-q)$
- Transition Probability: $[(1-P_{\text{plus}}) \cdot (1-P_{\text{minus}})]^2$ $P_{\text{minus}}$

- Pre-Duplication Probabilities:
  - $p$ = probability of interaction
  - $q$ = probability of self-interaction

- Post-Duplication Probabilities:
  - $P_{\text{plus}}$ = probability of adding an interaction
  - $P_{\text{minus}}$ = probability of eliminating an interaction

Emergence of post-duplication network motifs

- All have either 4 or 0 edges across the pairs (4-across or 0-across)

Modeling network evolution

- Parameters:
  - Fraction Duplicated vs Spontaneous Generation
  - Fraction Edges Deleted
  - Number of Edges for Spontaneous Genes

- 90% of timesteps: duplication
  - Pick a gene at random
  - Duplicate with all its connections
  - Delete on average 35% of new connections

- 10% of timesteps: creation
  - "Create" a new gene
  - Randomly connect it to the existing network with 0 – 20 connections

Abundance of network motifs predicted by duplication

Lessons Learned

1. Asymmetry in network connectivity
   - Interaction loss more likely than gain.
   - One protein maintains ancestral network function?

2. High frequency of ohnolog pair interaction
   1. Abundance of ancestral self-interactions
   2. Gain of ohnolog interaction by proximity due to common interactions

3. Abundance of global properties and network hubs
   - Duplication + asymmetric divergence model
   - Traditional preferential attachment model

Model matches local and global network properties