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

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

### The age of comparative genomics

<table>
<thead>
<tr>
<th>14 yeasts</th>
<th>12 flies</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>chimp</td>
</tr>
<tr>
<td>opossum</td>
<td>armadillo</td>
</tr>
<tr>
<td>bat</td>
<td>dolphin</td>
</tr>
<tr>
<td>llama</td>
<td>Tree shrew</td>
</tr>
</tbody>
</table>
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?

Part I: Using genomics to study evolution

Genomics

Evolution

Part II: Using evolution to study genomes

<table>
<thead>
<tr>
<th>Part I</th>
<th>Part II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Yeast</td>
<td>1. Genes</td>
</tr>
<tr>
<td>2. Alignment</td>
<td>2. Regulation</td>
</tr>
<tr>
<td>4. Duplication</td>
<td>4. Human</td>
</tr>
</tbody>
</table>
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
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)**

Extend concept of best-bidirectional hits

**Implementation: Iterative refinement**

Iterative refinement with increasing relative threshold
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

S. cerevisiae

S. paradoxus

S. mikatae

S. bayanus

S. cerevisiae Chromosome VI 250-300kbp

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

YBR184W
13% aa identity. Ka/Ks 7-fold higher. 120bp deletion Spore-specific. Cell wall. 3 conserved domains. 100% aa id. 100% nucleotide id. Mat Alpha2 counterpart. Role in mRNA complementation? Mating type switching?

Dynamic view of a changing genome

Mutation rates by functional classification
Protein domain evolution

- DNA-binding domain
- Dimerization domain
- Transcription activation domain

Rapid protein change

- **Protein domain creation**
  - Q/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

Evolutionary shortcuts apparent in recent evolution
TFP1 intein is a recent event

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

Different species, few novel genes
**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

**Specific mechanisms mediate rearrangements**

- 10 translocations
  - 8 across Ty elements
  - 2 across nearly identical genes
- 19 inversions
  - All flanked by tRNA genes

Evolutionary features
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

Transposons selectively kept in specific loci
Chromosomal Evolution

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

Ability to ask different set of questions
Gene correspondence

S. cerevisiae chromosomes

K. waltii scaffolds

Gene correspondence

S. cerevisiae chromosomes

K. waltii scaffolds
Sister regions show gene interleaving

Gene interleaving is evidence of complete duplication

Few genes remain in 2 copies

Duplicate mapping tiles K. waltii

S. cer.
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

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 → 4-way synteny
Genome duplication evidence in a single species

Evidence of duplication using a single genome?

- 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

- **Supporting evidence**
  - "There was a whole-genome duplication." Wolfe, Nature '97
  - "There was no whole-genome duplication." Dujon, FEBS 2000
  - "At least some chrom dup. occurred independently" Langkjær, JMB, 2000
  - "Dynamic equilibrium of duplications and loss" Llorente, FEBS, 2000
  - "Recent evidence supports single event" Wong, PNAS '02
  - "Continuous block duplications and deletions" Dujon, Yeast 2003
  - "Telomere-mediated duplication events" Coissac, Mol Bio Evo 1997
  - "Multiple closely spaced events" Friedman, Genome Res, 2003
  - "Spontaneous duplication of large chromosomal segments" Koszul, EMBO '04

**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

Post-duplication evolution
Whole-genome duplication resolved

Number of genes

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

Fate of duplicated genes

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

Evidence of accelerated protein divergence?
Scenarios for rapid gene evolution

One copy faster

- Scer - copy1
- Scer - copy2
- Kwal

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

Both copies faster

- Scer - copy1
- Scer - copy2
- Kwal

Emerging gene functions after duplication

- Origin of replication → silencing
  - 4-fold acceleration
  - Scer - Sir3 (silencing)
  - Scer - Orc1 (origin of replication)
  - Kwal - Orc1

- Translation initiation → anti-viral defense
  - 3-fold acceleration
  - Scer - Ski7 (anti-viral defense)
  - Scer - Hbs1 (translation initiation)
  - Kwal - Hbs1

Asymmetric divergence → recognize ancestral / derived
### Distinct functional properties

<table>
<thead>
<tr>
<th>Gene deletion</th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td></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>Gene deletion</th>
<th>Ancestral function</th>
<th>Derived function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lethal (20%)</td>
<td>Never lethal</td>
</tr>
<tr>
<td>Expression</td>
<td>Abundant</td>
<td>Specific (stress, starvation)</td>
</tr>
<tr>
<td>Localization</td>
<td>General</td>
<td>Specific (mitochondrion, spores)</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 4fold 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. waltii 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

Genome duplication in a vertebrate
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?

WGD → Rate 1 → S. cerevisiae copy 1

Rate 2 → S. cerevisiae copy 2

K. waltii

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**

Ancestral network motifs

Scenario 1

Scenario 2

Lost Duplicate

Duplication

Gain

Modern network motif

Duplicated gene

Interaction partners

Modern network

Pre-WGD

Network motif
Mechanisms of network motif emergence

- 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

Lost Interactions
Kept Interactions
Gained Interactions

Creation Probability
\( p \cdot (1-q) \cdot q \)

Transition Probability
\([(1-P_{\text{plus}}) \cdot (1-P_{\text{minus}})^3 - P_{\text{minus}}^2]\)

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

Study emergence of network motifs

Abundance of network motifs predicted by duplication
2. High frequency of ohnolog pair interaction

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

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

Ancestral self-interaction or gain of ohnolog interaction

3. Abundance of global properties and network hubs

Duplication + asymmetric divergence model
Traditional preferential attachment model

Model matches local and global network properties