Comparative Genomics, Duplication, and Coevolution of Duplicates

Hideki Innan

Graduate University for Advanced Studies, Japan

At KITP 2/7/07
Genome as a Genetic Material

世代から世代へ伝われるゲノム

おじいちゃんも、おばあちゃんも、わたしの中に

私たちの細胞は、「体細胞（体を作る細胞）」と「生殖細胞（精子、卵子）」からなります。

父親のゲノムを譲り受けた精子と、母親のゲノムを譲り受けた卵子が出会うと、新しい組み合わせのゲノムをもつ子ども、つまり「わたし」が生まれます。

また、両親の生殖細胞がつくられるときには、祖父母のゲノムがランダムに組みかえられて混ざります。

こうして、世代から世代へとゲノムは伝わりています。

from http://www.lif.kyoto-u.ac.jp/genomememap/
Genomes in Various Species
Genome Changes

from http://www.tolweb.org/
Genome Changes by Errors
Genome Changes by Errors
Genome Changes by Errors

Deletion
Genome Changes by Errors

A B C D E F G H I

A B C D E F G H I

Duplication
Genome Changes by Errors


A-C-G-C-T-G-T-G-T-A-T
Genome Duplicates
Autopolyploidy in Dendranthema

*D. japonicum*  2n=18

*D. boreale*  2n=36

*D. japonense*  2n=54

*D. shiwogiku*  2n=72

*D. pacificum*  2n=90
What Happens to Duplicated Genes
Evolutionary Fate of Duplicated Genes

Duplication

Doubling the product.
Evolutionary Fate of Duplicated Genes

Duplication

Original Function
Pseudogene or Deletion

Nonfunctionalization
Evolutionary Fate of Duplicated Genes

Duplication

Long-term maintenance of duplicates
Evolutionary Fate of Duplicated Genes

**Duplication**

<table>
<thead>
<tr>
<th>Original Function</th>
<th>New Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neofunctionalization</td>
<td></td>
</tr>
</tbody>
</table>

Diagram showing the evolutionary fate of duplicated genes with duplication leading to original and new functions.
Evolutionary Fate of Duplicated Genes

Duplication

Partitioning the ancestral functions

Subfunctionalization
Evolutionary Fate of Duplicated Genes

Duplication

Subfunction 1

Subfunction 2

Subfunctionalization

Subfunction 1

Subfunction 2
Evolutionary Fate of Duplicated Genes

Duplication

Subfunction 1

Subfunction 2

Subfunctionalization

Subfunction 1

Subfunction 2
Background

1. Genome changes at various levels.
   e.g., Gene duplication

2. Natural selection plays a crucial role.
   Natural selection is one of the key factors to determine the fate.
Questions

1. Genome changes at various levels.
   
   e.g., Gene duplication

   **How often does gene duplication occur?**

2. Natural selection is one of the key factors to determine the fate.
   
   Good changes are likely accepted.

   **How selection works on duplicated genes together with others such as drift, mutation, gene conversion?**
Estimating Gene Duplication Rate by Comparative Genomics
Model: Yeast (*Saccharomyces cerevisiae*)

Bread, Beer, Wine...
The complete genome sequence of the first eukaryote, *Saccharomyces cerevisiae*, appeared in 1996.

Genomic sequences are available for >10 relative species.
Power of Comparative Genomics

Without genomic information

Species A

Species B

Evolutionary Analysis

PCR

PCR
Power of Comparative Genomics

Without genomic information

Species A
Species B

Ortholog

Evolutionary Analysis
Without genomic information

Most Recent common Ancestor (14 MYA)
Human Orangutan

$d=0.028$
Power of Comparative Genomics

Without genomic information

“Molecular Clock”
Power of Comparative Genomics

Without genomic information

Species A

Species B

PCR

Paralog

Evolutionary Analysis

Misleading interpretation!
Power of Comparative Genomics

Without genomic information

Duplication (25 MYA)
Speciation (14 MYA)

Human
Orangutan

d=0.05
Power of Comparative Genomics

With genomic information

Species A

Species B

Ortholog

Evolutionary Analysis

No more confusion!
The Rate of Gene Duplication

Lynch and Conery 2000 Science 290: 1151-1155

8.3 per gene per billion years

Data: *Saccharomyces cerevisiae* genome

Method: Counting young duplicated genes in the genome
The Rate of Gene Duplication

Lynch and Conery 2000 Science 290: 1151-1155

8.3 per gene per billion years

Data: *Saccharomyces cerevisiae* genome

Method: Counting young duplicated genes in the genome

Surprisingly High!
On the order of point mutation rate.
The Rate of Gene Duplication

Lynch and Conery 2000 Science 290: 1151-1155

8.3 per gene per billion years

Data: Saccharomyces cerevisiae genome

Method: Counting young duplicated genes in the genome

What if duplicated genes are cheating their ages?
What is "Young" for Duplicated Gene?

“Young”: Genes duplicated recently
“Old”: Genes duplicated a long time ago

“Look Young”: Duplicated genes with low divergence
“Look Old”: Duplicated genes with high divergence

according to Molecular Clock
What is ”Young” for Duplicated Gene?

“Young”: Genes duplicated recently  
“Old”: Genes duplicated a long time ago  

“Look Young”: Duplicated genes with low divergence  
“Look Old”: Duplicated genes with high divergence  

according to Molecular Clock

Question

“Look Young” = ”Young”?  

Does a molecular clock hold for duplicated gene?
The Rate of Gene Duplication

Lynch and Conery 2000 Science 290: 1151-1155

8.3 per gene per billion years

Data: *Saccharomyces cerevisiae* genome

Method: Counting “looking young” duplicated genes
Nucleotide Evolution in Duplicated Genes
How does a molecular clock work for the divergence between duplicates?
Independent Evolution

Gene duplication

T₀ ————

Divergence

Time

30
25
20
15
10
5
0

T₀ T₁ T₂ T₃ T₄

Divergence

0
5
10
15
20
25
30

T₀ T₁ T₂ T₃ T₄

Time
Independent Evolution

Gene duplication

T₀

Divergence

Time

30
25
20
15
10
5
0
T₀ T₁ T₂ T₃ T₄

T₁
Independent Evolution

Gene duplication

$T_0$ → $T_1$ → $T_2$

![Graph showing divergence over time](image-url)
Independent Evolution

Gene duplication

T0

T1

T2

T3

Time

Divergence

T0  T1  T2  T3  T4

0  5  10  15  20  25  30
Independent Evolution

Gene duplication

\[ \text{Time} \]

\[ T_0, T_1, T_2, T_3, T_4 \]

\[ \text{Divergence} \]

Graph showing divergence over time:

- \( T_0 \)
- \( T_1 \)
- \( T_2 \)
- \( T_3 \)
- \( T_4 \)
Independent Evolution

Gene duplication

T₀
T₁
T₂
T₃
T₄
T∞

Divergence

Time

0 5 10 15 20 25 30

T₀ T₁ T₂ T₃ T₄
Independent Evolution

Gene duplication

Old Pair

Young Pair

Molecular Clock

Old Pair
Concerted Evolution

Gene duplication

T₀

Divergence

30
25
20
15
10
5
0

Time

T₀ T₁ T₁' T₂ T₃

Divergence

0
5
10
15
20
25
30
Concerted Evolution

Gene duplication

T0

T1

Divergence

Time

T0 T1 T1' T2 T3

Divergence

0 5 10 15 20 25 30
Concerted Evolution

Gene duplication

Gene conversion

T0

T1

T1'

Time

Divergence

T0

T1

T1'

T2

T3
Concerted Evolution

Gene duplication

T₀

Gene conversion

T₁

T₁'

T₂

Divergence

Time

T₀ T₁ T₁' T₂
Concerted Evolution

Gene duplication

Gene conversion

T3
T2
T1'
T1
T0

Divergence

Time

T0 T1 T1' T2 T3

Graph showing divergence over time with points at T0, T1, T1', T2, and T3.
Concerted Evolution

Gene duplication

Gene conversion

Divergence

Time

T0

T1

T1'

T2

T3

T∞
Concerted Evolution

Gene duplication

T₀

Gene conversion

T₁

T₁'

Molecular Clock

T₂

T₃

T∞

Old Pair Still Look Young

Young Pair Look Young
Model-Free Estimation of the Gene Duplication Rate in Yeast

- **S. cerevisiae**
- **S. paradoxus**
- **S. mikatae**
- **S. kudriavzevii**
- **S. bayanus**
- **S. castellii**
- **S. kluyveri**

Complete Genome Sequences:
- Gofieau et al. (1996)
  - Science 274: 546-567
- Shotgun Sequences:
  - Cliften et al. (2003)
    - Science 301: 71-76
  - Kellis et al. (2003)
    - Nature 423: 241-254
Model-Free Estimation of the Gene Duplication Rate in Yeast

1. Identify duplicated genes (denoted by X and Y) in the baker’s yeast (*Saccharomyces cerevisiae*) genome.

2. Find evidence for the presence of the orthologs of X and Y.

- **S. cerevisiae** genome
  - A
  - X
  - B
  - C
  - Y
  - D

- **Target genome**
  - A
  - X
  - B
  - C
  - Y
  - D

BLASTN

AXB/CYD/0
Both present
Estimating the Minimum Age, $T_m$

$K_s$ $T_6$ $T_5$ $T_4$ $T_3$ $T_2$ $T_1$ $T_0$

$S. paradoxus$

$S. cerevisiae$ AXB/CYD/0

$S. paradoxus$

$S. mikatae$

$S. kudriavzevii$

$S. bayanus$

$S. castellii$

$S. kluyveri$
Estimating the Minimum Age, $T_m$

- $S. paradoxus$
- $S. cerevisiae$
- $S. mikatae$
- $S. kudriavzevii$
- $S. bayanus$
- $S. castellii$
- $S. kluyveri$

AXB/CYD/0: Both present
Estimating the Minimum Age, $T_m$

- $T_0$
- $T_1$
- $T_2$
- $T_3$
- $T_4$
- $T_5$
- $T_6$

- $K_s$
- 0.25
- 0.5
- 1

- S. cerevisiae
- AXB/CYD/0
- Both present

- S. paradoxus

- S. mikatae
- Both present

- S. kudriavzevii

- S. bayanus

- S. castellii

- S. kluyveri
Estimating the Minimum Age, $T_m$

$T_6 \quad T_5 \quad T_4 \quad T_3 \quad T_2 \quad T_1 \quad T_0$

$K_s \quad 0.5 \quad 0.25 \quad 0$

- **S. cerevisiae**
- **S. paradoxus**
- **S. mikatae**
- **S. kudriavzevii**
- **S. bayanus**
- **S. castellii**
- **S. kluyveri**

AXB/CYD/0

Both present

Both present

Both present
Estimating the Minimum Age, $T_m$

$T_6$ $T_5$ $T_4$ $T_3$ $T_2$ $T_1$ $T_0$

$K_s$ 0.5 0.25 0

S. cerevisiae
S. paradoxus
S. mikatae
S. kudriavzevii
S. bayanus
S. castellii
S. kluyveri

AXB/CYD/0
Both present
Both present
Both present
Both present

$T_m$
## Minimum Ages of Duplicated Genes in the Baker’s Yeast Genome

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Ks</th>
<th>S. paradoxus</th>
<th>S. mikatae</th>
<th>S. kudriavzevi</th>
<th>S. bayanus</th>
<th>S. castellii</th>
<th>S. kluveyri</th>
<th>Tm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>YHR053C</td>
<td>YHR055C</td>
<td>0</td>
<td>AXB</td>
<td>AXB</td>
<td>AX-</td>
<td>AXB</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>2</td>
<td>YCL066W</td>
<td>YCR040W</td>
<td>0</td>
<td>AXB/-YD/2</td>
<td>—/-CYD/2</td>
<td>XB/CYD×2/1</td>
<td>AXB/-YD/1</td>
<td>A-B/*-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>3</td>
<td>YNL019C</td>
<td>YNL033W</td>
<td>0</td>
<td>AXB/-YD/1</td>
<td>—/-YD/0</td>
<td>—/-CYD/0</td>
<td>AX-/-CYD/0</td>
<td>—/-/-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>4</td>
<td>YAR064W</td>
<td>YHR213W-B</td>
<td>0</td>
<td>—/-YD/3</td>
<td>—/-/-/0</td>
<td>—/-/-/1</td>
<td>—/-/-/0</td>
<td>—/-/-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>5</td>
<td>YFL061W</td>
<td>YNL335W</td>
<td>0.0053</td>
<td>—/-YD/0</td>
<td>—/-/-/1</td>
<td>—/-/-/0</td>
<td>—/-/-/-/0</td>
<td>—/-/-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>6</td>
<td>YGL135W</td>
<td>YPL220W</td>
<td>0.0209</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>-XB/-YD/0</td>
<td>AXB/CYD/0</td>
<td>A-B/CYD/0</td>
<td>AX-/*-/-/0</td>
</tr>
<tr>
<td>7</td>
<td>YOR390W</td>
<td>YPL279C</td>
<td>0.0209</td>
<td>AXB/—/0</td>
<td>A-B/**/1</td>
<td>-XB/*-/-/1</td>
<td>—/-/-/0</td>
<td>—/-/-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>8</td>
<td>YDL136W</td>
<td>YDL191W</td>
<td>0.0265</td>
<td>AXB/CYD/0</td>
<td>-XB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>A-B/CYD/0</td>
<td>—/-CYD/0</td>
</tr>
<tr>
<td>9</td>
<td>YBR181C</td>
<td>YPL090C</td>
<td>0.0298</td>
<td>—/-CYD/1</td>
<td>AX-/-CYD/1</td>
<td>AXB/CYD/0</td>
<td>-XB/*-/-/1</td>
<td>—/-CYD,-YD/0</td>
<td>—/-/-/1</td>
</tr>
<tr>
<td>10</td>
<td>YNL018C</td>
<td>YNL034W</td>
<td>0.0329</td>
<td>AXB/CYD/0</td>
<td>—/-CYD,-YD/0</td>
<td>—/-CYD/0</td>
<td>—/-CYD/2</td>
<td>—/-/-/-/0</td>
<td>—</td>
</tr>
<tr>
<td>11</td>
<td>YBR031W</td>
<td>YDR012W</td>
<td>0.0354</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD,-YD/0</td>
<td>AXD/CYD/0</td>
<td>AXD/CYD/0</td>
<td>AXD/*-/-/0</td>
<td>AX-/*-/-/0</td>
</tr>
<tr>
<td>12</td>
<td>YHR141C</td>
<td>YNL162W</td>
<td>0.0677</td>
<td>AXB/CYD/0</td>
<td>-XB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/-YD/0</td>
<td>—/-/-/1</td>
</tr>
<tr>
<td>13</td>
<td>YHR203C</td>
<td>YJR145C</td>
<td>0.0699</td>
<td>AXB/CYD/0</td>
<td>AXB/*-/-/2</td>
<td>AXB/CYD/0</td>
<td>—/-/-/-/2</td>
<td>—/-/-/-/-/1</td>
<td>—/-/-/-/1</td>
</tr>
<tr>
<td>14</td>
<td>YER074W</td>
<td>YIL069C</td>
<td>0.0699</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>—/-CY-/-/1</td>
<td>—/-/-/1</td>
</tr>
<tr>
<td>15</td>
<td>YBL072C</td>
<td>YER102W</td>
<td>0.0947</td>
<td>AXB/CYD/0</td>
<td>AX-/-CYD/0</td>
<td>AX-/*-/-/1</td>
<td>—/-CY-/-/1</td>
<td>—/-/-/-/1</td>
<td>—/-/-/-/1</td>
</tr>
<tr>
<td>16</td>
<td>YBR099C</td>
<td>YNL030W</td>
<td>0.1339</td>
<td>AXB/CYD/0</td>
<td>—/-/-/-/2</td>
<td>-XB/CYD/0</td>
<td>-XB/CYD×2/0</td>
<td>—/-CY-/-/1</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>YHL001W</td>
<td>YKL006W</td>
<td>0.1343</td>
<td>AXB/CYD/0</td>
<td>AX-/-YD/1</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/*-/-/1</td>
<td>—/-/-/1</td>
</tr>
<tr>
<td>18</td>
<td>YIL018W</td>
<td>YFR031C-A</td>
<td>0.1412</td>
<td>AXB/CYD/0</td>
<td>AX-/-CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>-XB/*-/-/1</td>
<td>—/-CY-/-/0</td>
</tr>
<tr>
<td>19</td>
<td>YGR085C</td>
<td>YPR102C</td>
<td>0.1445</td>
<td>AXB/CYD/0</td>
<td>AX-,-XB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>—/-/-/-/1</td>
<td>—/-/-/-/1</td>
</tr>
<tr>
<td>20</td>
<td>YHL033C</td>
<td>YLL045C</td>
<td>0.1457</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>AXB/CYD/0</td>
<td>—/-/-/-/1</td>
<td>—/-/-/-/0</td>
</tr>
</tbody>
</table>

* Tandem duplicated genes for which the gene order of X, Y and markers is given by AXYB in *S. cerevisiae.*
Model-Free Estimation of the Gene Duplication Rate in Yeast

1-5 gene pairs with $T_m = T_0$

$\Rightarrow$ duplication rate $= 0.01$-0.06 per billion years

About 1/100 of molecular clock-based estimate by Lynch and Conery

So many old genes look as if they are young.

Gao and Innan 2004 Science 306: 1367-1370
Can we do that?
Summary 1

1. Low estimate of gene duplication rate by comparative genomics
   
   Analysis with a single genome do not work well...

2. Genome-wide demonstration of concerted evolution in duplicated genes
   
   Molecular clock does not hold for duplicated genes.
Selection and Evolutionary Fate of Duplicated Genes under the Pressure of Gene Conversion
Evolutionary Fate of Duplicated Genes

Duplication

[Diagram of gene duplication and its evolutionary fate]
Evolutionary Fate of Duplicated Genes

Duplication

Original Function

New Function
Evolutionary Fate of Duplicated Genes

Diagram showing the evolutionary fate of duplicated genes through duplication and gene conversion processes.
Evolutionary Fate of Duplicated Genes

Duplication

Gene conversion
Evolutionary Fate of Duplicated Genes

Under what condition does neofunctionalization occur?
Evolutionary Fate of Duplicated Genes

Under what condition does neofunctionalization occur?

⇒ Look at nucleotide polymorphism!
Single Nucleotide Polymorphisms (SNPs) in Duplicated Genes
Typical Pattern of Polymorphism in Duplicated Genes

Proximal & Distal Amy in D. melanogaster

Data: Inomata et al. 1995 Genetics 141: 237-244
### Effect of Gene Conversion on Polymorphism

<table>
<thead>
<tr>
<th></th>
<th>hw1</th>
<th>hw2</th>
<th>hb</th>
<th>D</th>
<th>Kw1</th>
<th>Kw2</th>
<th>Kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shared</td>
<td>0</td>
<td>.6</td>
<td>.53</td>
<td>.53</td>
<td>0</td>
<td>.33</td>
<td>0</td>
</tr>
<tr>
<td>Fixed</td>
<td>.33</td>
<td>.53</td>
<td>.53</td>
<td>.6</td>
<td>0</td>
<td>.33</td>
<td>0</td>
</tr>
<tr>
<td>Specific</td>
<td>.2</td>
<td>.57</td>
<td>.6</td>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ D_{\text{sum}} = 0.43 \]
Modeling Duplicated Genes

$L \times$ two-locus (site) models
Two-Locus Model

1. Two duplicated loci (sites), I and II, with two neutral alleles \((A \text{ and } a)\) at mutation-drift equilibrium

2. Four haplotypes, \(A - A, A - a, a - A, \text{ and } a - a\)

3. Symmetrical mutation rate, \(\mu\), between \(A\) and \(a\) at each loci \((\theta = 4N\mu)\)

4. Recombination rate, \(r\) \((R = 4Nr)\)

5. Gene conversion rate, \(c\) \((C = 4Nc)\)
Recursions for Haplotype Frequencies

\[ x_1: \text{frequency of } A - A \]
\[ x_2: \text{frequency of } A - a \]
\[ x_3: \text{frequency of } a - A \]
\[ x_4: \text{frequency of } a - a \]

\[ x'_1 = (1 - 2\mu)x_1 + (\mu + c)(x_2 + x_3) - rD \]
\[ x'_2 = (1 - 2\mu - c)x_2 + \mu(x_1 + x_4) + rD \]
\[ x'_3 = (1 - 2\mu - c)x_3 + \mu(x_1 + x_4) + rD \]
\[ x'_4 = (1 - 2\mu)x_4 + (\mu + c)(x_2 + x_3) - rD \]
Backward Diffusion Equation

At equilibrium, \( g = g(x_1, x_2, x_3) \) satisfies

\[
E[L(g)] = 0,
\]

where

\( L(g) \): differential operator (next slide)
Diffusion Operator

\[ L(g) = \]
\[ \frac{x_1 (1 - x_1)}{4N} \frac{\partial g}{\partial^2 x_1^2} + \frac{x_2 (1 - x_2)}{4N} \frac{\partial g}{\partial^2 x_2^2} + \frac{x_3 (1 - x_3)}{4N} \frac{\partial g}{\partial^2 x_3^2} \]
\[ + \frac{x_1 x_2}{2N} \frac{\partial g}{\partial x_1 \partial x_2} + \frac{x_1 x_3}{2N} \frac{\partial g}{\partial x_1 \partial x_3} + \frac{x_2 x_3}{2N} \frac{\partial g}{\partial x_2 \partial x_3} \]
\[ + \left[ -2\mu x_1 + (\mu + c)(x_2 + x_3) - rD \right] \frac{\partial g}{\partial x_1} \]
\[ + \left[ -2(\mu + c)x_2 + \mu(1 - x_2 - x_3) + rD \right] \frac{\partial g}{\partial x_2} \]
\[ + \left[ -2(\mu + c)x_3 + \mu(1 - x_2 - x_3) + rD \right] \frac{\partial g}{\partial x_3} \]
Transformation of the three variables

$$(x_1, x_2, x_3) \rightarrow (p, q, D)$$

\[
p = x_1 + x_2 \\
q = x_1 + x_3 \\
D = x_1x_4 - x_2x_3
\]

At equilibrium, $g = g(p, q, D)$ satisfies

$$E[L'(g)] = 0$$

Ref: Ohta and Kimura 1969 Genetics 63: 229-238
Differential Operator

\[ L'(g) = \]
\[ p(1 - p) \frac{\partial^2 g}{\partial p^2} + q(1 - q) \frac{\partial^2 g}{\partial q^2} \]
\[ + [pq(1 - p)(1 - q) + D(1 - 2p)(1 - 2q)] \frac{\partial^2 g}{\partial x_3^2} \]
\[ + 2D \frac{\partial g}{\partial p} + 2D(1 - 2p) \frac{\partial g}{\partial p} \frac{\partial}{\partial D} + 2D(1 - 2q) \frac{\partial g}{\partial q} \frac{\partial}{\partial D} \]
\[ + [\theta(1 - 2p) - C(p - q)] \frac{\partial g}{\partial p} + [\theta(1 - 2q) - C(p - q)] \frac{\partial g}{\partial q} \]
\[ + [Cp(1 - p) - Cq(1 - q) + (2 + 4\theta + 2C + R)D] \frac{\partial g}{\partial D} \]
Four Equations of $E(p^2)$, $E(pq)$ and $E(D)$

$g = p^2 :$

$$1 + \theta - 2(1 + 2\theta + C)E(p^2) + 2CE(pq) = 0$$

$g = pq :$

$$2E(D) + 2CE(p^2) + (4\theta + 2C)E(pq) + \theta = 0$$

$g = D :$

$$C - 2CE(p^2) - (4\theta + 2C + R)E(D) = 0$$
Exact Solutions for $E(p^2)$, $E(pq)$ and $E(D)$

$$E(p^2) = \frac{\lambda}{\omega}$$

$$E(pq) = -\frac{1 + \theta}{2C} + \frac{(1 + \alpha)\lambda}{C\omega}$$

$$E(D) = \frac{C}{\beta} \left( 1 - \frac{2\lambda}{\omega} \right)$$

where

$$\alpha = 2\theta + C, \quad \beta = 2 + 2\alpha + R$$

$$\lambda = 4C^2 + 4\beta[2\theta C + 2\alpha(1 + \theta)]$$

$$\omega = 8C^2 + 4\beta[\alpha(1 + \alpha) - c^2]$$
Expectations of $h_w$, $h_b$ and $D$

\[
E(h_w) = 1 - 2\frac{\lambda}{\omega}
\]

\[
E(h_b) = 1 + \frac{1 + \theta}{C} - \frac{2(1 + \alpha)\lambda}{C\omega}
\]

\[
E(D) = \frac{C}{\beta} \left( 1 - \frac{2\lambda}{\omega} \right)
\]

Innan 2002 Genetics 161: 865-872
Two-Locus Model → Finite-Site Model

\[ E(K_w) = L \times E(h_w) \]
\[ E(K_b) = L \times E(h_b) \]
\[ E(D_{sum}) = L \times E(D) \]
\( E(K_w) = \lim_{L \to \infty} L \times E(h_w) = \frac{2\Theta(2C + R + 2)}{4C + R + 2} \)

\( E(K_b) = \lim_{L \to \infty} L \times E(h_b) = \frac{\Theta(4C^2 + 4C + 2CR + R + 2)}{C(4C + R + 2)} \)

\( E(D_{\text{sum}}) = \lim_{L \to \infty} L \times E(D) = \frac{2\Theta C}{4C + R + 2} \)

where

\( \Theta = L\theta: \text{Mutation rate per gene} \)

Innan 2003 Genetics 163: 803-810
Estimating $\Theta$, $C$, and $R$

\[
\hat{\Theta} = \frac{K_w + 2D_{sum}}{2} \quad \text{or} \quad \hat{\theta} = \frac{K_w + 2D_{sum}}{2L}
\]

\[
\hat{C} = \frac{K_w - 2D_{sum}}{2(K_b - K_w)}
\]

\[
\hat{R} = \frac{K_w^2 + 4D_{sum}^2 - 4K_bD_{sum}}{2(K_b - K_w)D_{sum}}
\]
Proximal & Distal Amy in D. melanogaster

\[
\begin{array}{c}
\text{D. simulans} \\
\text{Proximal Amy} \\
\text{Distal Amy}
\end{array}
\]

\[
\begin{array}{cccccc}
CCC & G & G & A & G & C \ \text{CC} \\
\text{Proximal Amy} \\
\text{Distal Amy}
\end{array}
\]

Observed:  \( K_w = 45.89, \ K_b = 68.16, \ D_{sum} = 0.67 \)

Estimated:  \( \theta = 0.0172, \ C = 1.03, \ R = 66.6 \)

Data: Inomata et al. 1995 Genetics 141: 237-244
Distributions of Shared and Fixed Sites

<table>
<thead>
<tr>
<th>Region</th>
<th>Shared</th>
<th>Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region I (801 bp)</td>
<td>11</td>
<td>0</td>
</tr>
</tbody>
</table>
| Region II (453 bp)| 0     | 15    | $P < 10^{-6}$
Does Selection Against Gene Conversion in Exon 7 Explain the Data?
Selection Against Gene Conversion: Selection for Neofunctionalization

Duplication

Original Function

New Function
Selection Against Gene Conversion: Selection for Neofunctionalization
Selection Against Gene Conversion: Selection for Neofunctionalization

- Duplication
- Gene conversion
A and $B$ have different functions so that haplotypes with two different alleles are favored.

Population parameters, $\theta = 4N\mu$, $R = 4Nr$, and $C = 4Nc$.
Recursions for Haplotype Frequencies

\[ x_1: \text{ frequency of } A - A \]
\[ x_2: \text{ frequency of } A - B \]
\[ x_3: \text{ frequency of } B - A \]
\[ x_4: \text{ frequency of } B - B \]

\[ x'_1 = (1 - 2\mu)x_1 + (\mu + c)(x_2 + x_3) - rD - sx_1(x_2 + x_3) \]
\[ x'_2 = (1 - 2\mu - c)x_2 + \mu(x_1 + x_4) + rD + sx_2(x_1 + x_4) \]
\[ x'_3 = (1 - 2\mu - c)x_3 + \mu(x_1 + x_4) + rD + sx_3(x_1 + x_4) \]
\[ x'_4 = (1 - 2\mu)x_4 + (\mu + c)(x_2 + x_3) - rD - sx_4(x_2 + x_3) \]
Backward Diffusion Equation

At equilibrium, \( g = g(x_1, x_2, x_3) \) satisfies

\[
E[L(g)] = 0,
\]

where

\( L(g) \): differential operator (next slide)
Differential Operator

\[ L'(g) = \]
\[
\frac{x_1(1-x_1)}{4N} \frac{\partial^2 g}{\partial x_1^2} + \frac{x_2(1-x_2)}{4N} \frac{\partial^2 g}{\partial x_2^2} + \frac{x_3(1-x_3)}{4N} \frac{\partial^2 g}{\partial x_3^2} \\
+ \frac{x_1 x_2}{2N} \frac{\partial g}{\partial x_1 \partial x_2} + \frac{x_1 x_3}{2N} \frac{\partial g}{\partial x_1 \partial x_3} + \frac{x_2 x_3}{2N} \frac{\partial g}{\partial x_2 \partial x_3} \\
+ \left[ -2\mu x_1 + (\mu + c)(x_2 + x_3) - rD - sx_1(x_2 + x_3) \right] \frac{\partial g}{\partial x_1} \\
+ \left[ -2(\mu + c)x_2 + \mu(x_1 + x_4) + rD + sx_2(x_1 + x_4) \right] \frac{\partial g}{\partial x_2} \\
+ \left[ -2(\mu + c)x_3 + \mu(x_1 + x_4) + rD + sx_3(x_1 + x_4) \right] \frac{\partial g}{\partial x_3}
\]
Transformation of the three variables

$$(x_1, x_2, x_3) \rightarrow (p, q, D)$$

\[
\begin{align*}
p &= x_1 + x_2 \\
q &= x_1 + x_3 \\
D &= x_1 x_4 - x_2 x_3
\end{align*}
\]

At equilibrium, $g = g(p, q, D)$ satisfies

$$E[L'(g)] = 0$$
Differential Operator

\[ L'(g) = \]

\[ p(1 - p) \frac{\partial^2 g}{\partial p^2} + q(1 - q) \frac{\partial^2 g}{\partial q^2} \]

\[ + [pq(1 - p)(1 - q) + D(1 - 2p)(1 - 2q)] \frac{\partial^2 g}{\partial x_3^2} \]

\[ + 2D \frac{\partial g}{\partial p \partial q} + 2D(1 - 2p) \frac{\partial g}{\partial p \partial D} + 2D(1 - 2q) \frac{\partial g}{\partial q \partial D} \]

\[ + \ldots \text{continue} \]
\[ + [ \theta (1 - 2p) - C (p - q) \\
\quad + 4N s p (1 - p - 2q + 2pq) - 4N s D (1 - 2p)] \frac{\partial g}{\partial p} \]

\[ + [ \theta (1 - 2q) - C (p - q) \\
\quad + 4N s q (1 - 2p - q + 2pq) - 4N s D (1 - 2q)] \frac{\partial g}{\partial q} \]

\[ + [C p (1 - p) - C q (1 - q) + (2 + 4 \theta + 2C + R) D \\
\quad - 8N s p q (1 - p - q + pq) + 8N s D^2] \frac{\partial g}{\partial D} \]
Three Equations of $E(p^2)$, $E(p^3)$ and $E(p^4)$

\[ g = p : \]
\[ 1 - 6E(p^2) + E(p^3) = 0 \]

\[ g = pq : \]
\[ -\theta - C + 4Ns + 4(\theta + C - 4Ns)E(p^2) - 8NsE(p^3) + 16NsE(p^4) = 0 \]

\[ g = D : \]
\[ C - 4Ns - 2(C - 8Ns)E(p^2) - 8NsE(p^4) = 0 \]

Assumption: Strong selection (i.e., $p + q \approx 1$)
Solutions

\[ E(p^2) = \frac{\theta - C + 2Ns}{4(\theta + Ns)}, \quad E(p^3) = \frac{\theta - 3C + 4Ns}{8(\theta + Ns)} \]

\[ E(p^4) = \frac{C(\theta - 8Ns) + 8(Ns)^2 + C^2}{16(\theta + Ns)} \]

Then,

\[ E(h_w) = 1 - 2E(p^2) = \frac{\theta + C}{2(\theta + Ns)} \]

\[ E(h_b) = 1 - 2E(pq) = \frac{\theta - C + 2Ns}{2(\theta + Ns)} \]

Innan 2003 PNAS 100: 8793-8798
Interpretation of Theoretical Results

If selection is sufficiently strong, the target site of selection could be a fixed (or nearly fixed) site.

Additional mutations likely fix in the surrounding region of the target site by hitchhiking effect.

Signature of Selection

A peak of the divergence between duplicated genes in regions of functional importance.
1. All 15 fixed sites are around exon 7, which encodes amino acids that characterize the difference between RHCE and RHD antigens.

2. Most fixed sites are non-synonymous (13 non-synonymous vs. 2 synonymous)
Hypothetical History of the Rh genes

Gene duplication

Position in the Rh gene

Divergence (%)
Hypothetical History of the Rh genes

Gene duplication

T0 → T1

T0 T1 T2 T3

Divergence (%)

T0
T1

Time

Position in the Rh gene

Divergence (%)
Hypothetical History of the Rh genes

Gene duplication

T₀ → T₁ → T₂

T₀

T₁

T₂

Divergence (%)

15 10 5 0

Time

T₀ T₁ T₂

Divergence (%)

1000 750 500 250

Position in the Rh gene

0 250 500 750 1000 1254

T₀ T₁ T₂
Hypothetical History of the Rh genes

Gene duplication

T0

T1

T2

T3

Selection

Divergence (%) vs. Position in the Rh gene

Divergence (%) vs. Time
Hypothetical History of the Rh genes

Gene duplication

T0 → T1 → T2 → T3 → T4

Selection

Neofunctionalization?

Exon 7

Divergence (%)

T0 T1 T2 T3 T4

Time

Position in the Rh gene

0 250 500 750 1000 1254
Similar Pattern in Human Opsin Genes

Two amino acid differences in exon 5 contribute to the changes between the red and green opsins

1. Simulated region ($L$ bp) assigned to a (0,1) interval

2. Mutation rate $\mu$ per region per generation

3. Gene conversion rate $c$ per site per generation
   - Gene conversion is initiated at rate $g$
   - Tract length follows a Geometric distribution with a mean tract length of $1/Q$ ($1/q = L/Q$ bp)

Teshima and Innan 2004 Genetics 166: 1553-1560
How Concerted Evolution is Terminated?

1. Accumulation of a number of point mutations
   Gene conversion is terminated once
   the divergence between the duplicates (\(d\)) hits a
   threshold value, \(d_t\).
   important parameters: \(c/\mu, d_t, Q\)

2. Drastic changes of DNA sequences
   Mutations such as transposons and large in/dels
   automatically terminates gene conversion.
   important parameters: \(\mu T\)

3. Selection for neofunctionalization
   Gene conversion is deleterious.
   important parameters: \(s/c\)
Summary

1. Demonstrating how comparative genomics works
2. Estimate of gene duplication rate
3. Genome-wide demonstration of concerted evolution in yeast
4. Development of basic theory to analyze DNA polymorphism data in duplicated genes
5. Selection for neofunctionalization works against gene conversion
Acknowledgement

Graduate Univ. for Advanced Studies
Kosuke Teshima (currently at Sokendai)
Ryuichi Sugino (currently at Sokendai)

Univ. Texas at Houston
Lizhi Gao (currently at UT at Austin)