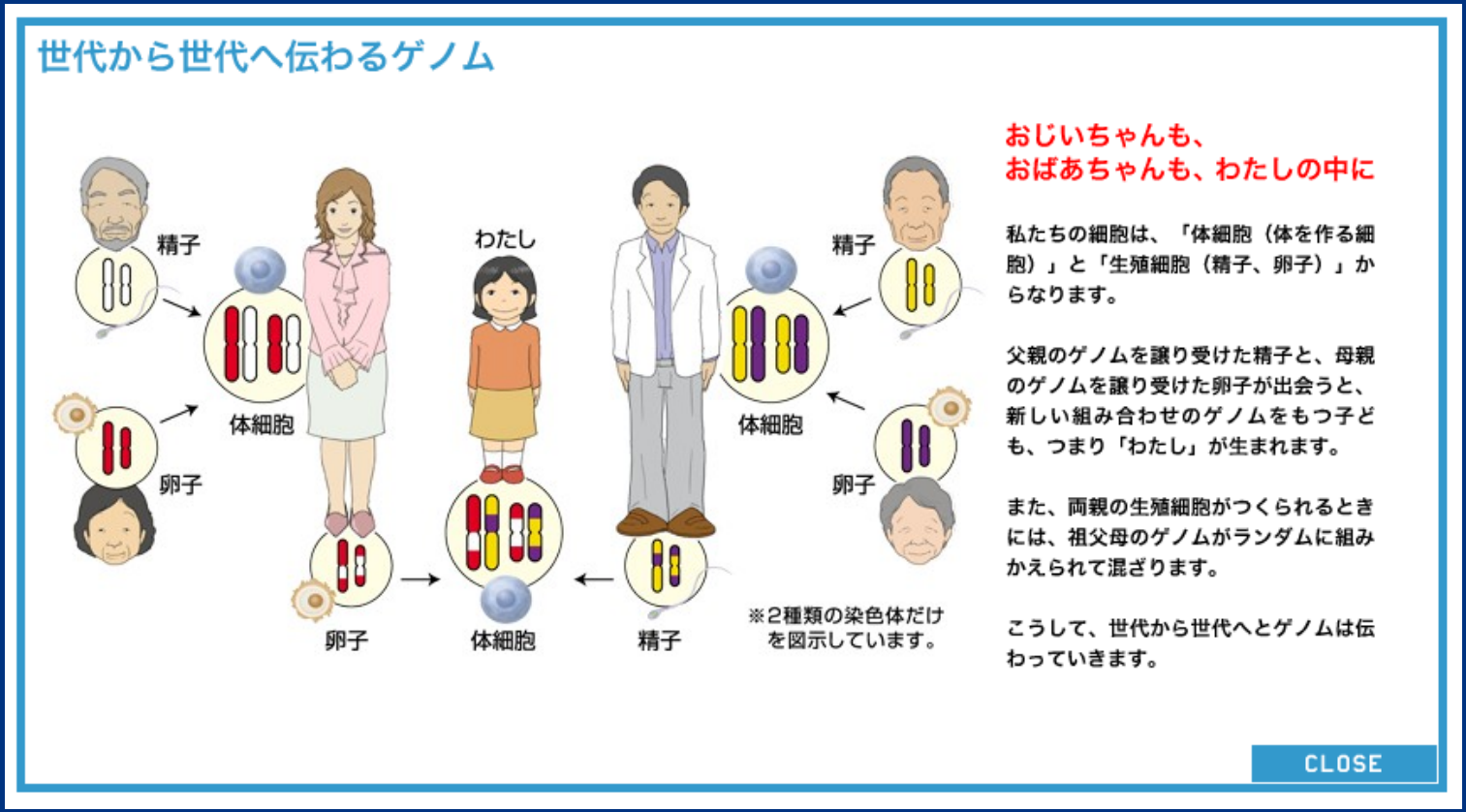
Comparative Genomics, Duplication, and Coevolution of Duplicates

Hideki Innan

Graduate University for Advanced Studies, Japan

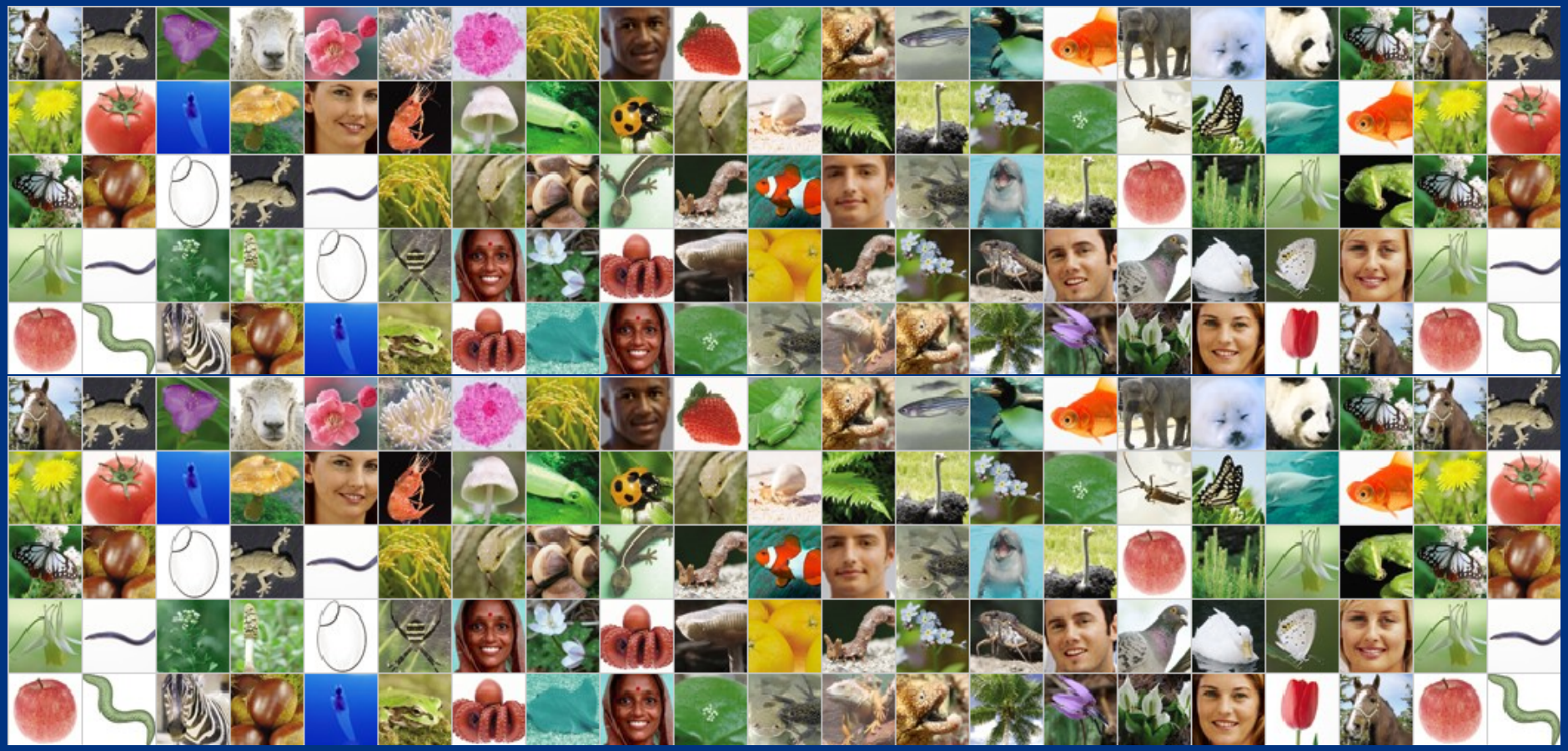
At KITP 2/7/07

Genome as a Genetic Material



from http://www.lif.kyoto-u.ac.jp/genomemap/

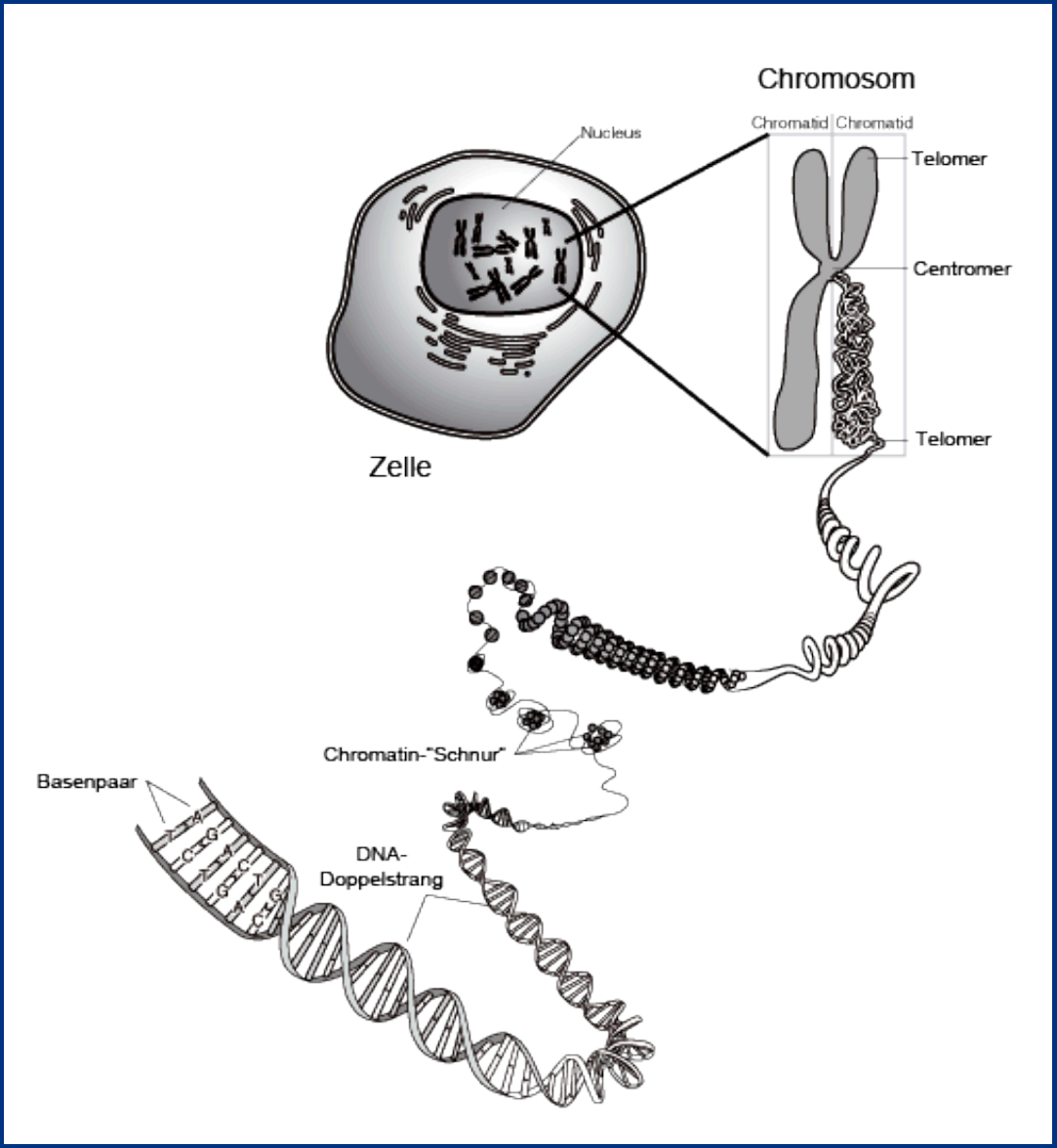
Genomes in Various Species



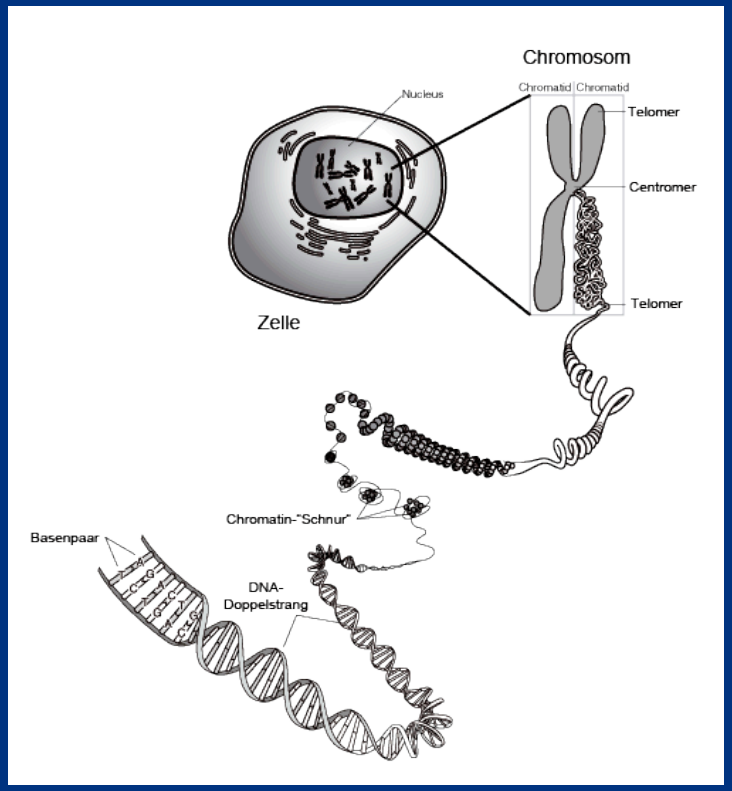
Genome Changes

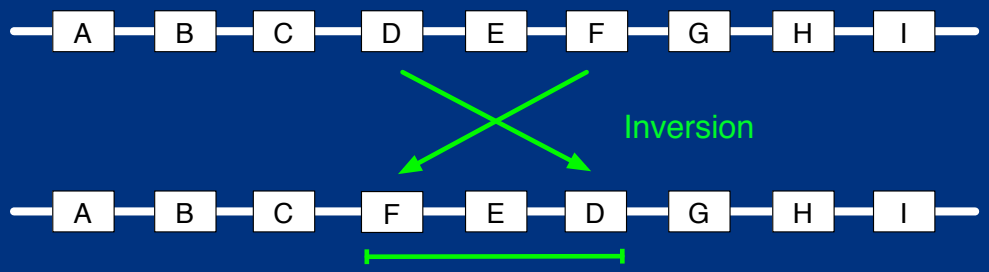
from http://www.tolweb.org/

Genome Changes by Errors

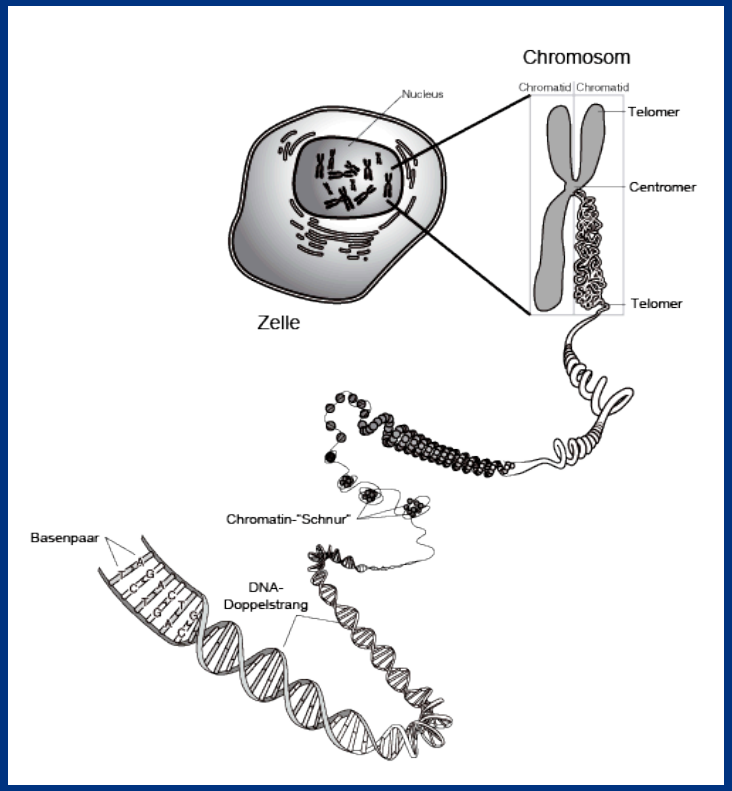


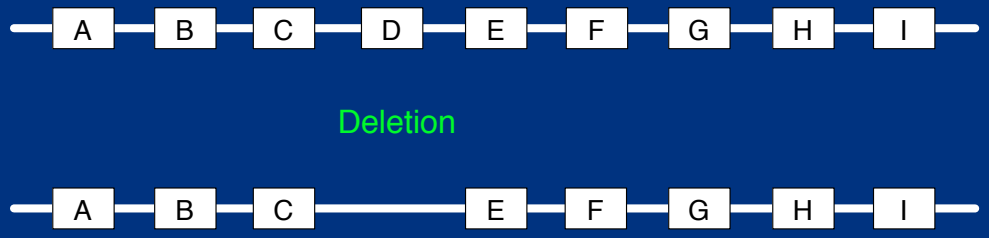
Genome Changes by Errors



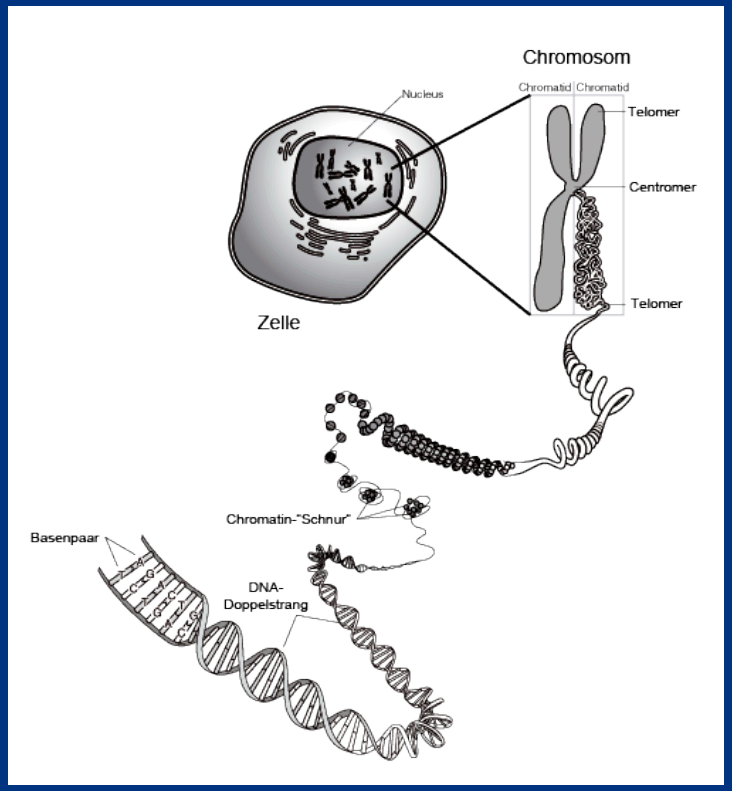


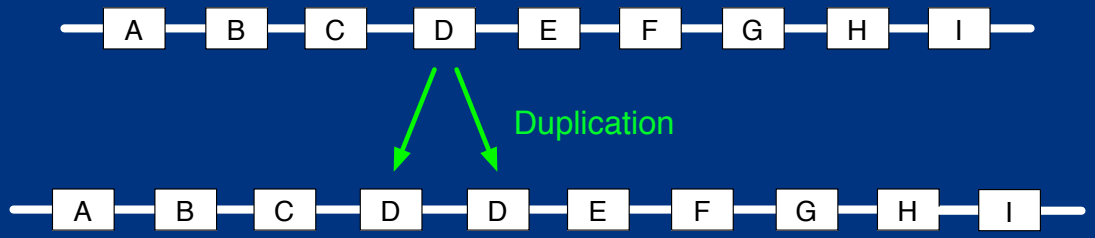
Genome Changes by Errors



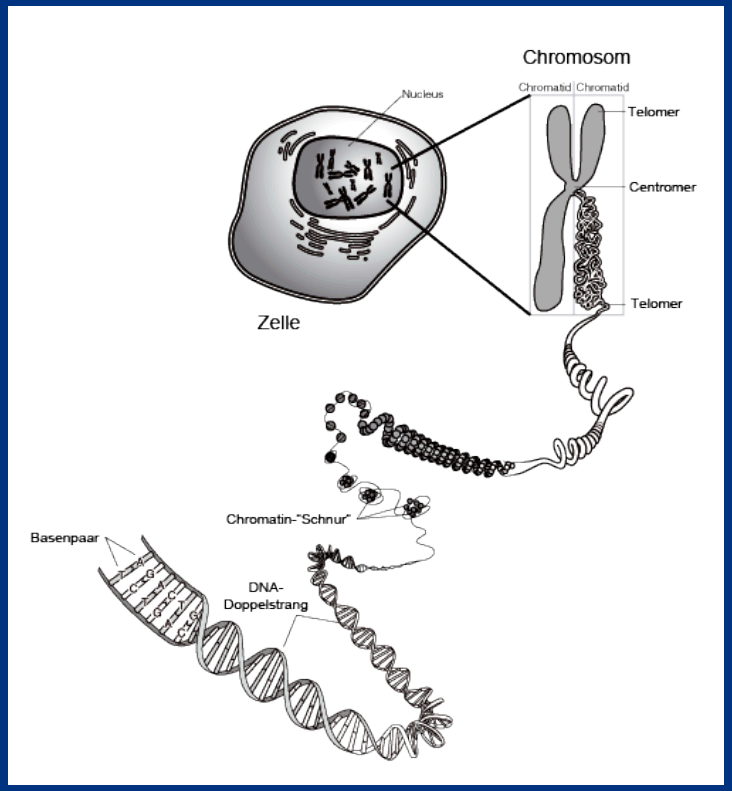


Genome Changes by Errors





Genome Changes by Errors



A-T-G-C-T-G-T-A-T-A-T

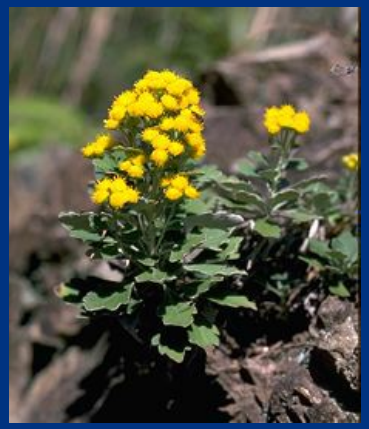
A-C-G-C-T-G-T-G-T-A-T

Genome Duplicates
Autopolyploidy in Dendranthema

D. japonicum
2n=18

D. boreale
2n=36

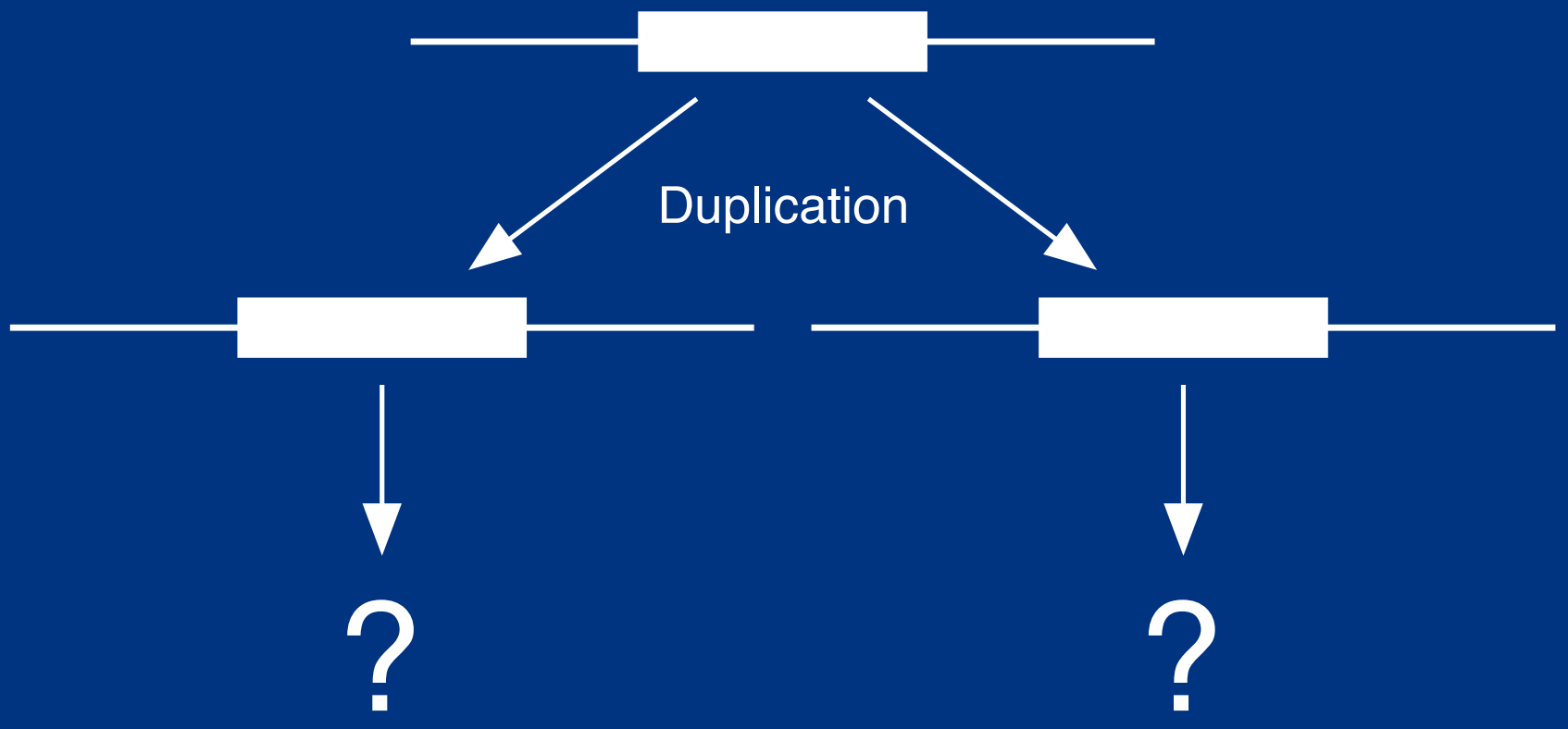
D. japonense
2n=54



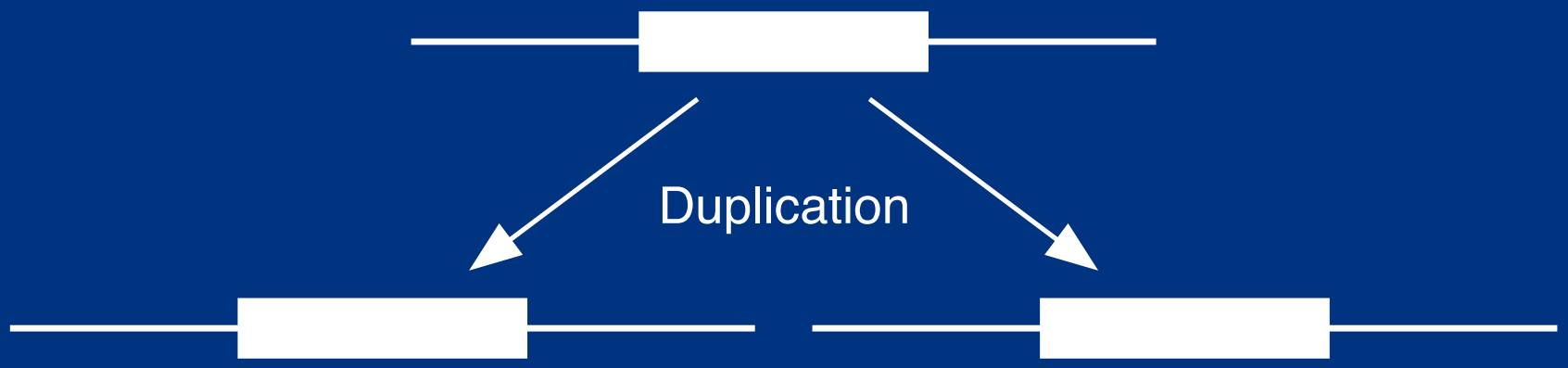
D. shiwogiku
2n=72

D. pacificum
2n=90

What Happens to Duplicated Genes

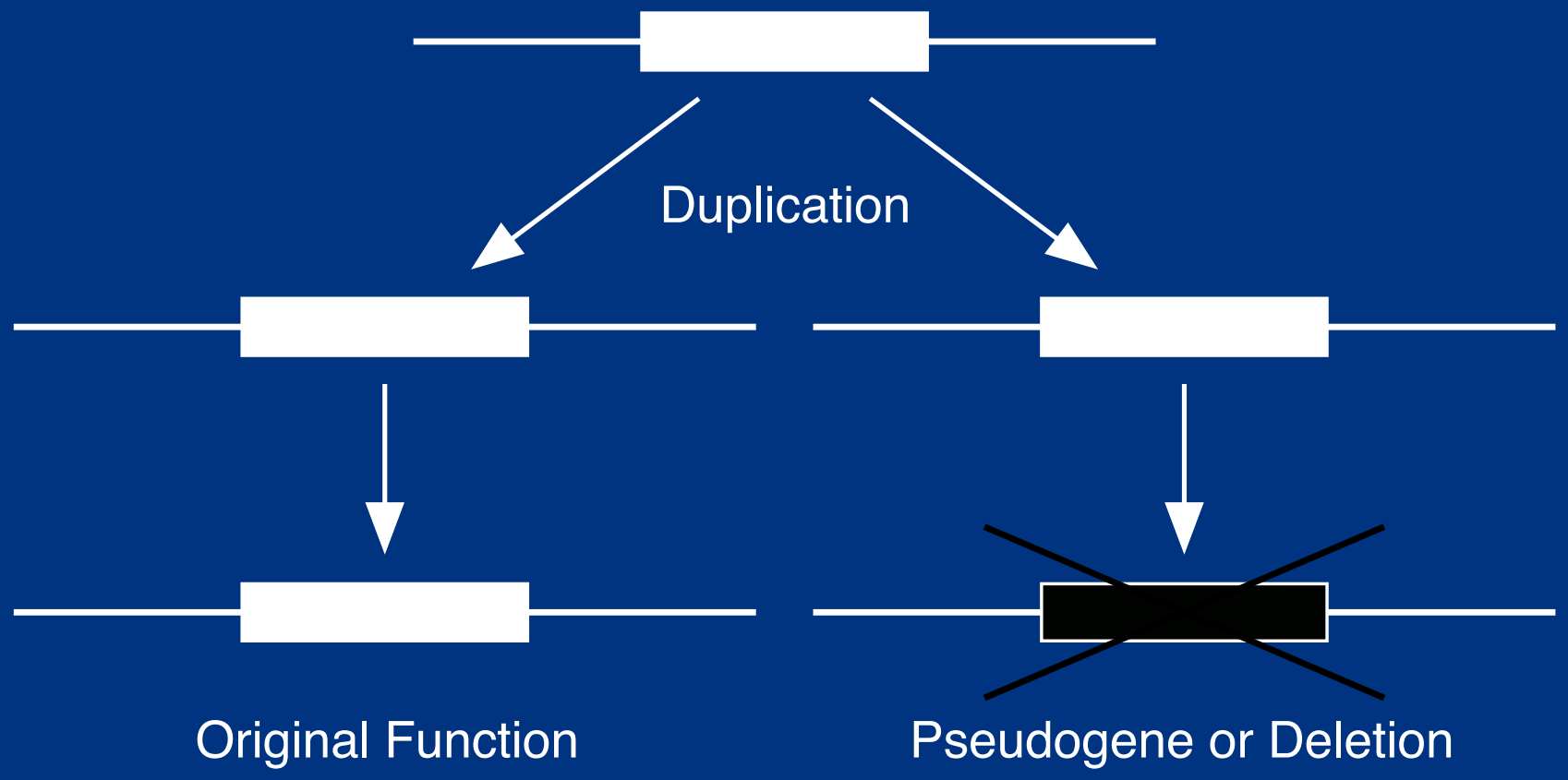


Evolutionary Fate of Duplicated Genes



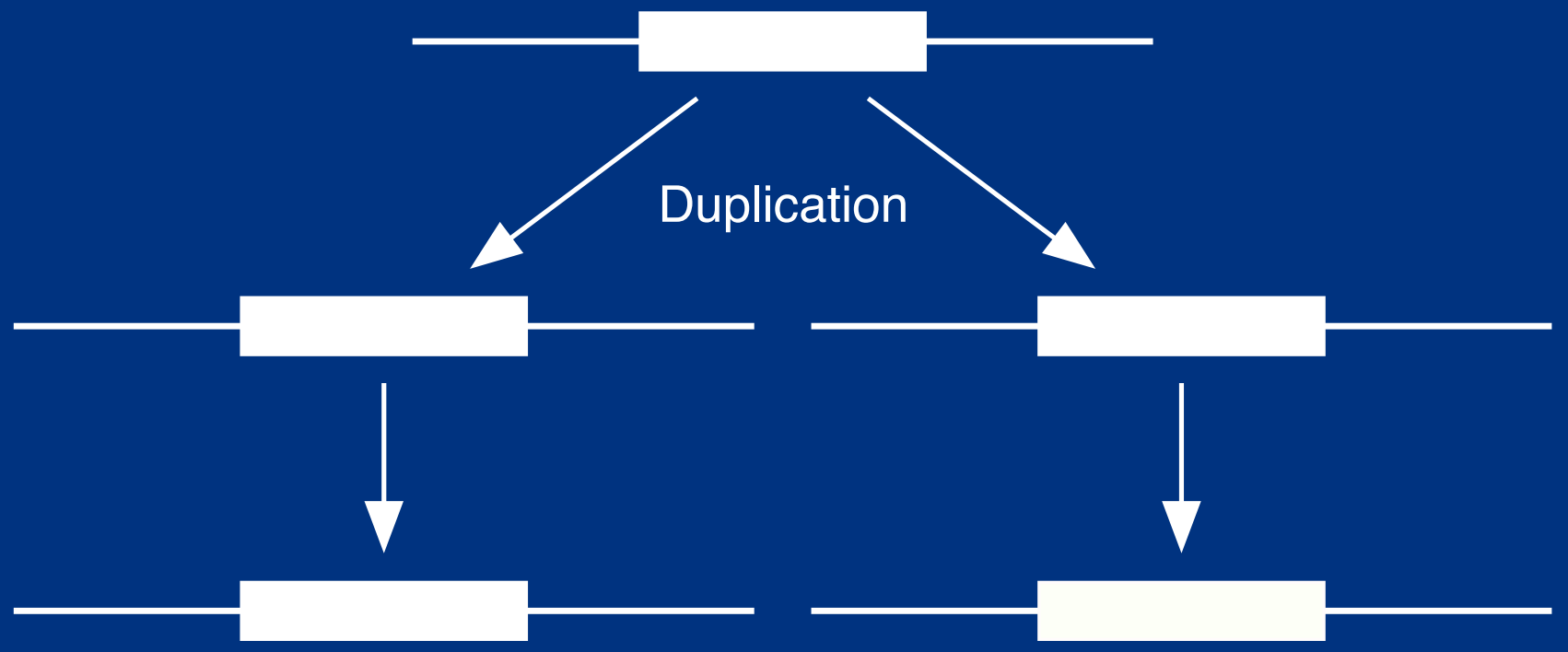
Doubling the product.

Evolutionary Fate of Duplicated Genes



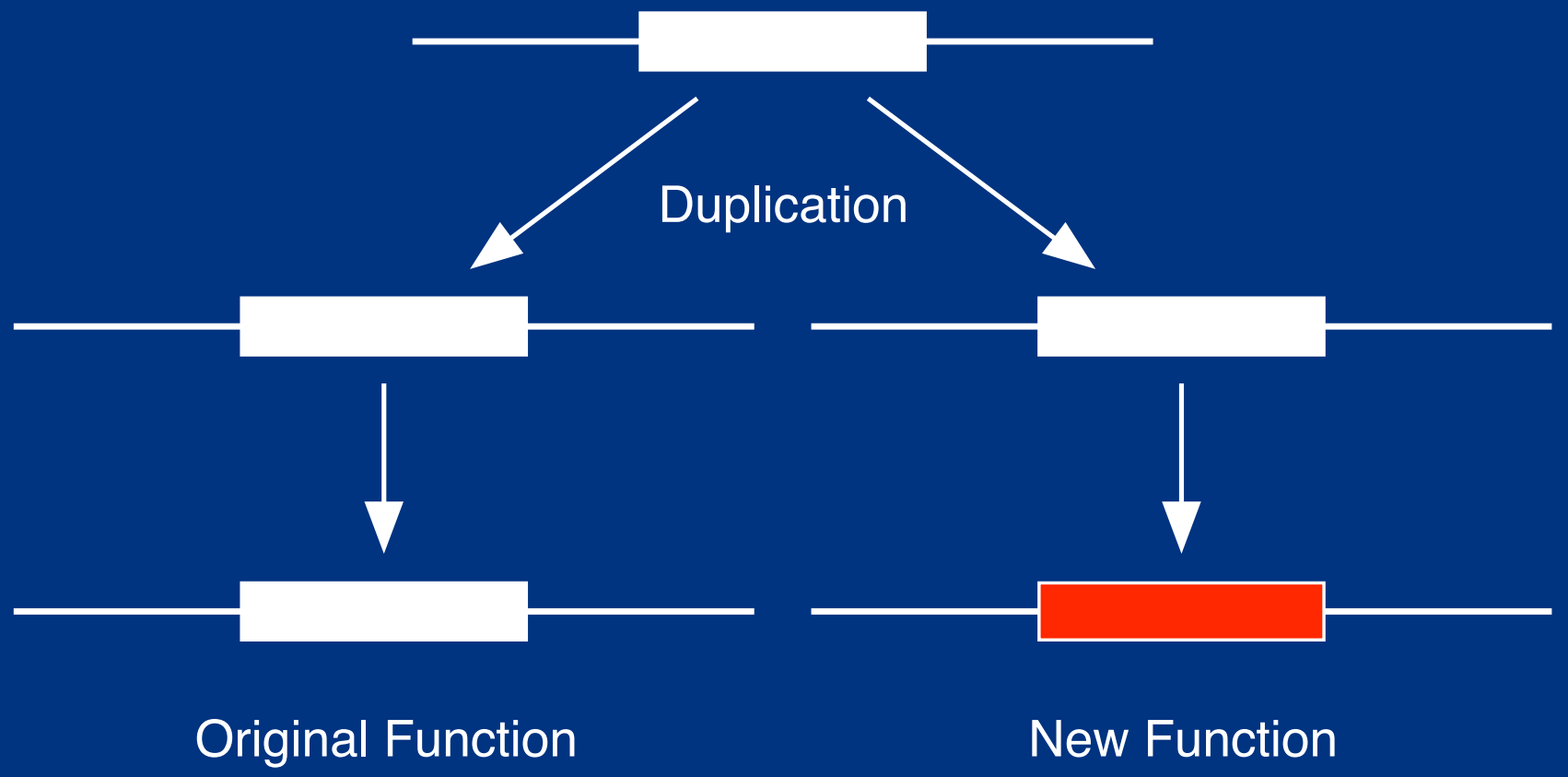
Nonfunctionalization

Evolutionary Fate of Duplicated Genes



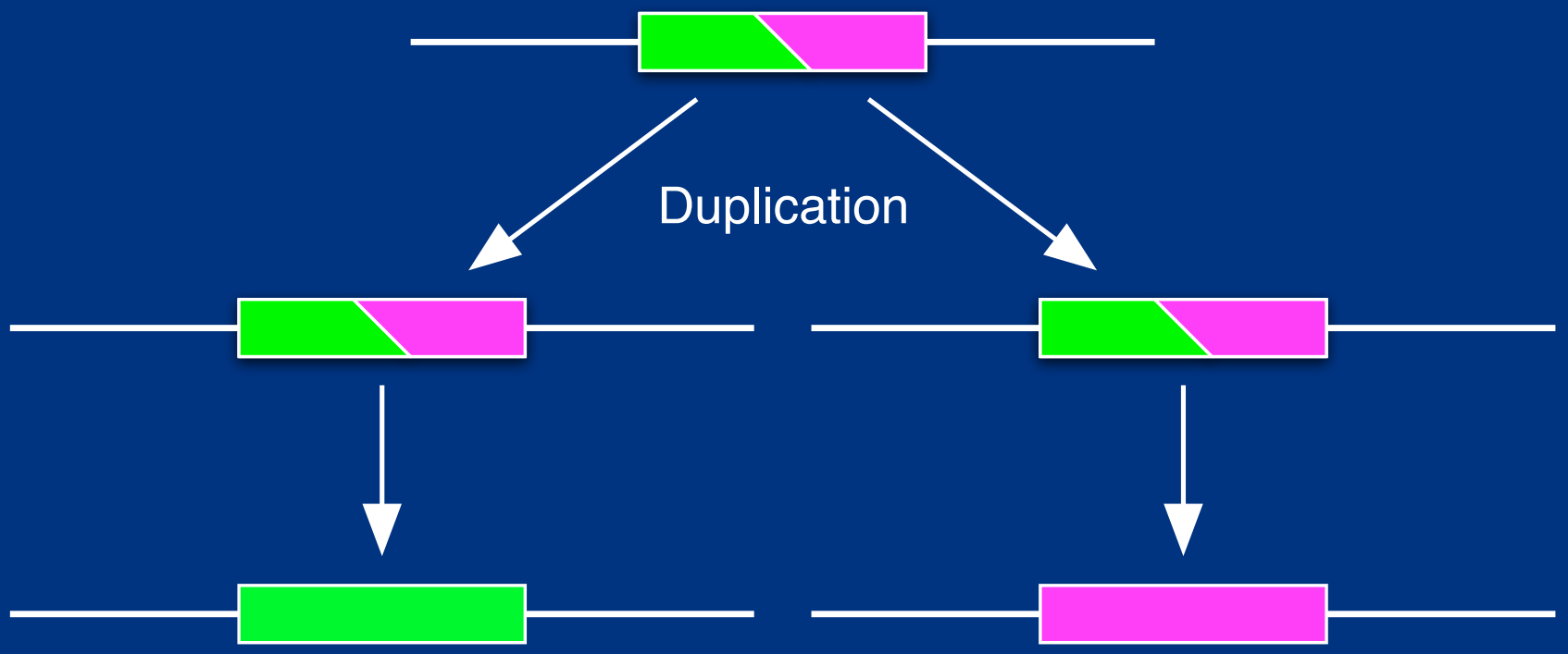
Long-term maintenance of duplicates

Evolutionary Fate of Duplicated Genes



Neofunctionalization

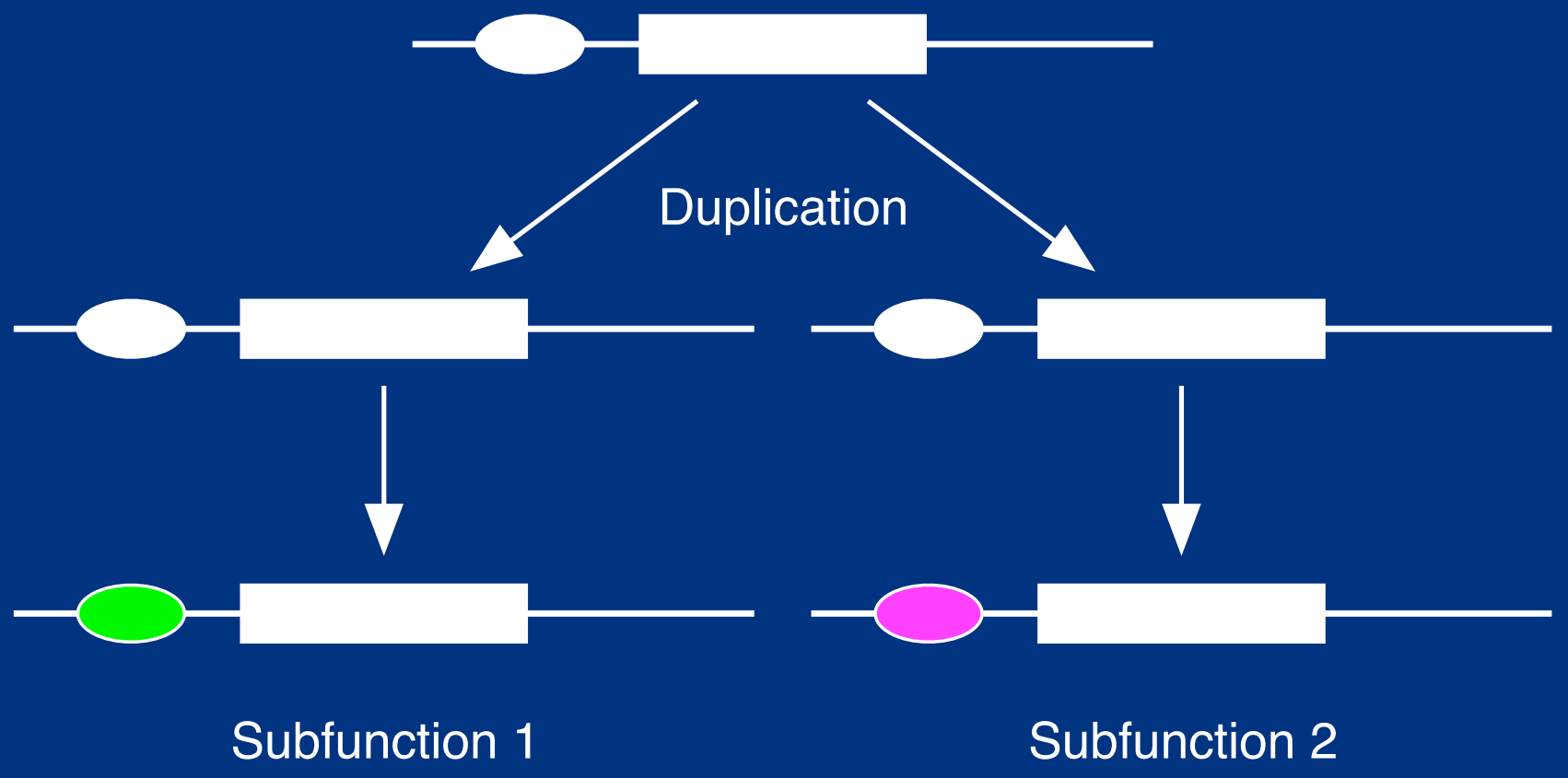
Evolutionary Fate of Duplicated Genes



Partitioning the ancestral functions

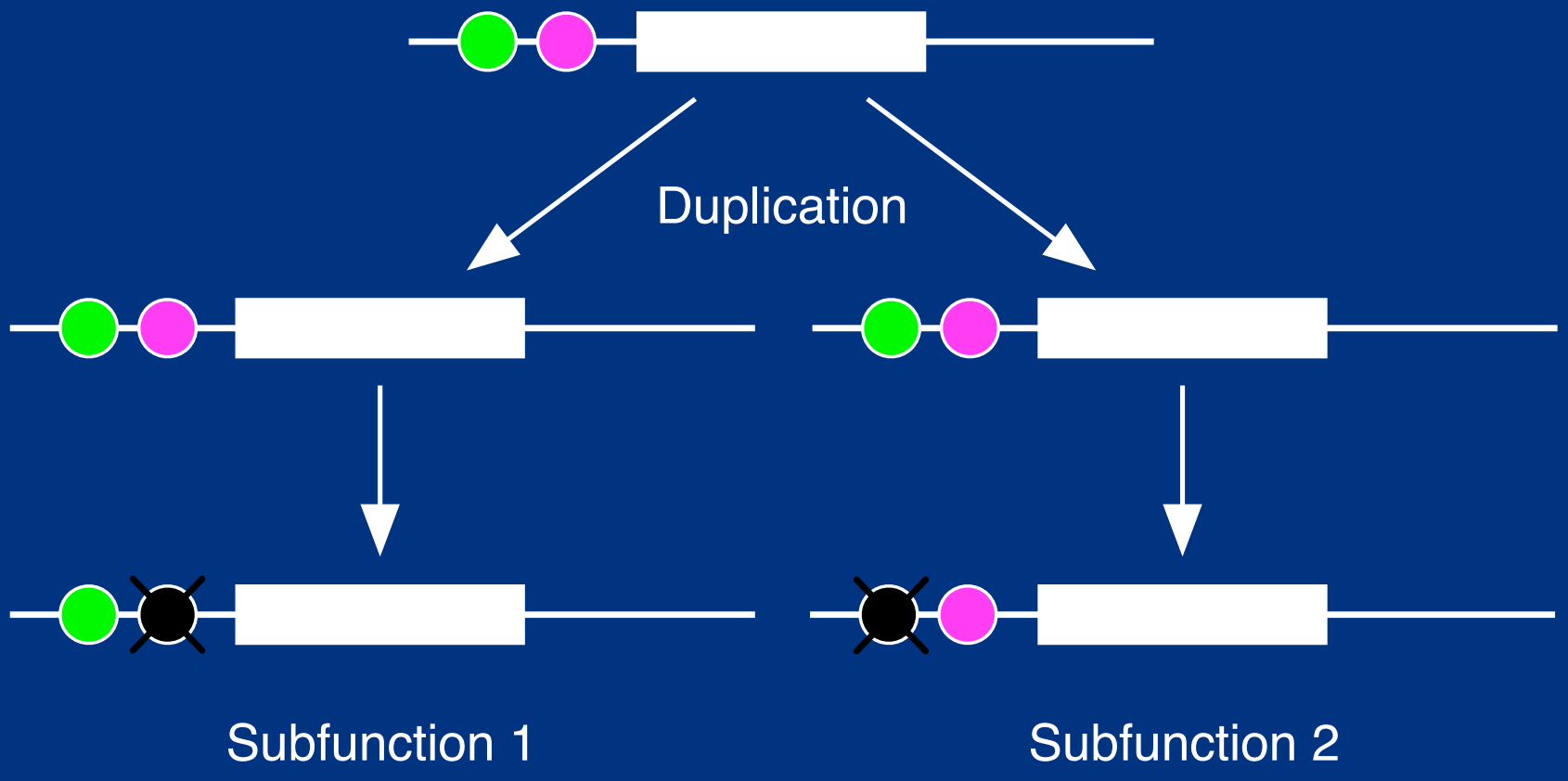
Subfunctionalization

Evolutionary Fate of Duplicated Genes



Subfunctionalization

Evolutionary Fate of Duplicated Genes



Subfunctionalization

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.

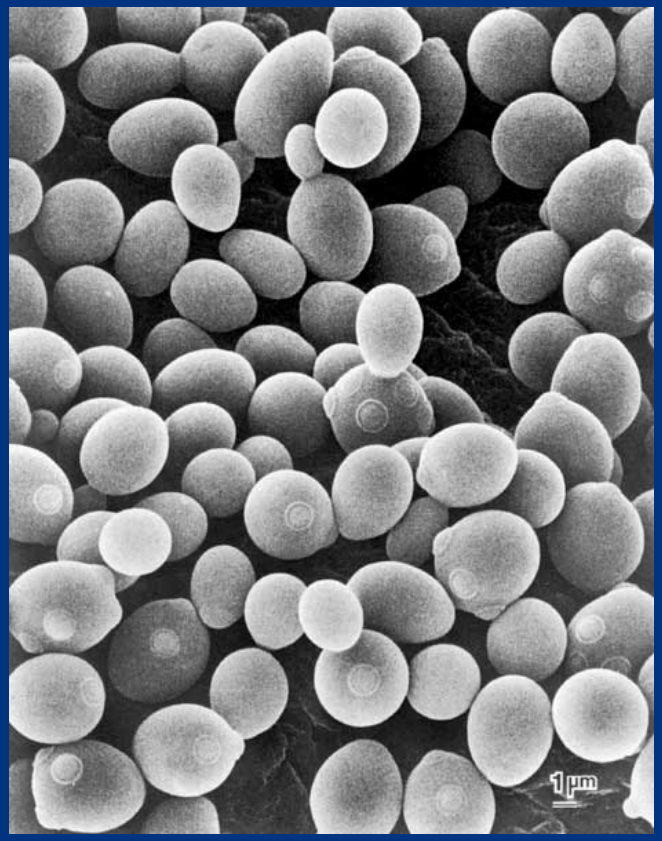
   Good changes are likely accepted.

   How selection works on duplicated genes together with others such as drift, mutation, gene conversion?

Estimating Gene Duplication Rate

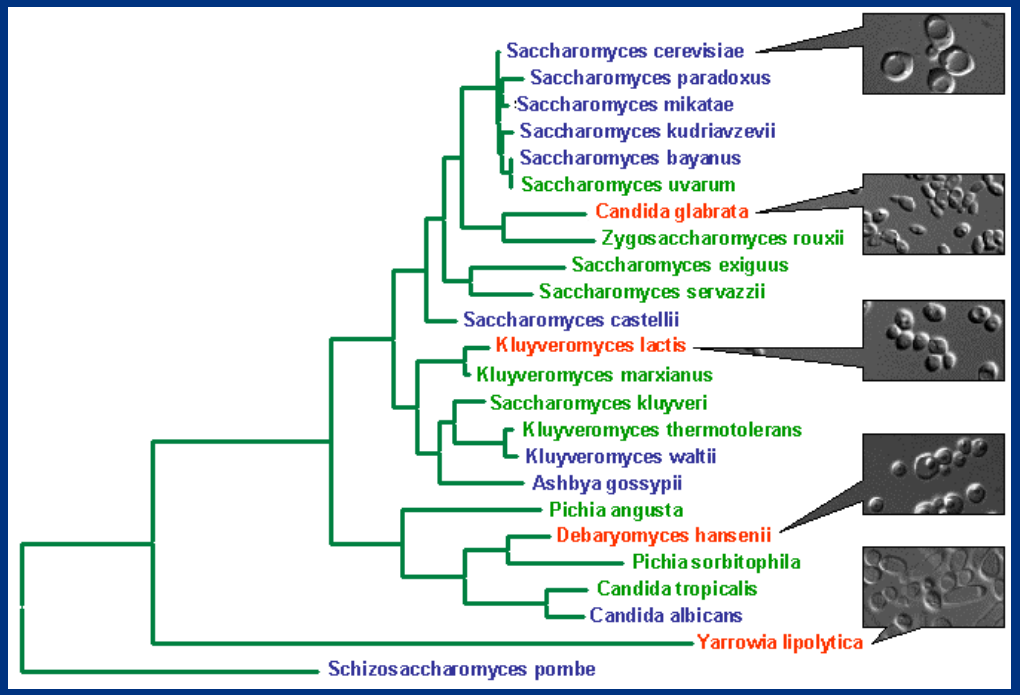
by Comparative Genomics

Model: Yeast (Saccharomyces cerevisiae)



Bread, Beer, Wine...

Yeast Genome



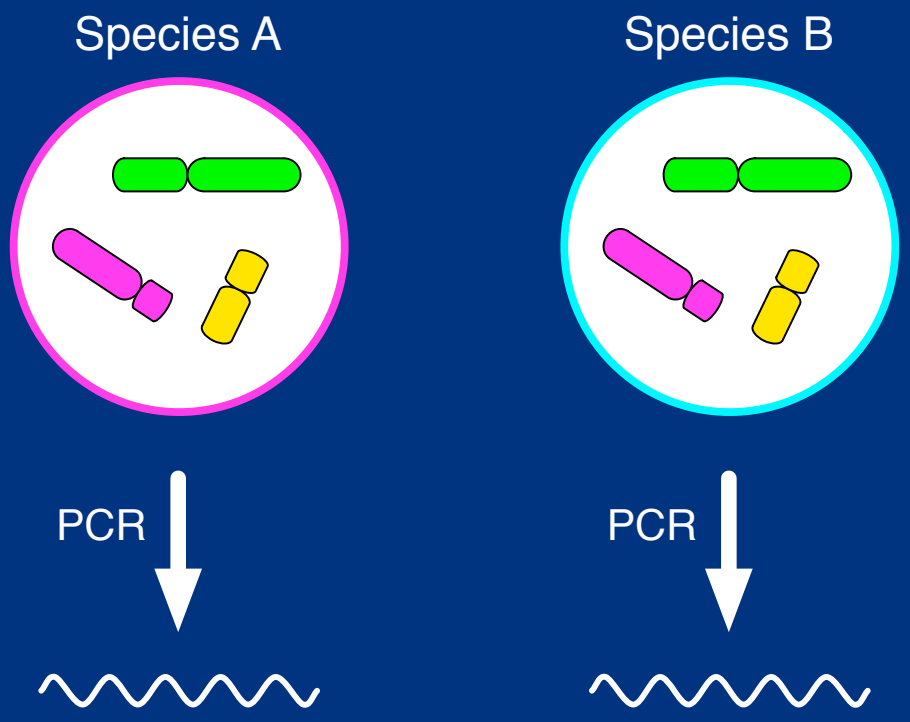
Statistics	
Assembly:	SGD 1, Nov 2005
Genebuild:	SGD, Jun 2006
Database version:	40.1d
Known genes:	6,680
Pseudogenes:	21
snRNA genes:	6
rRNA genes:	14
snoRNA genes:	70
tRNA genes:	299
Genscan gene predictions:	5,077
Genefinder gene predictions:	3,311
Gene exons:	7,026
Gene transcripts:	6,680
Base Pairs:	12,156,590
Golden Path Length:	12,156,590
Most common InterPro domains:	Top 40 Top 500

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

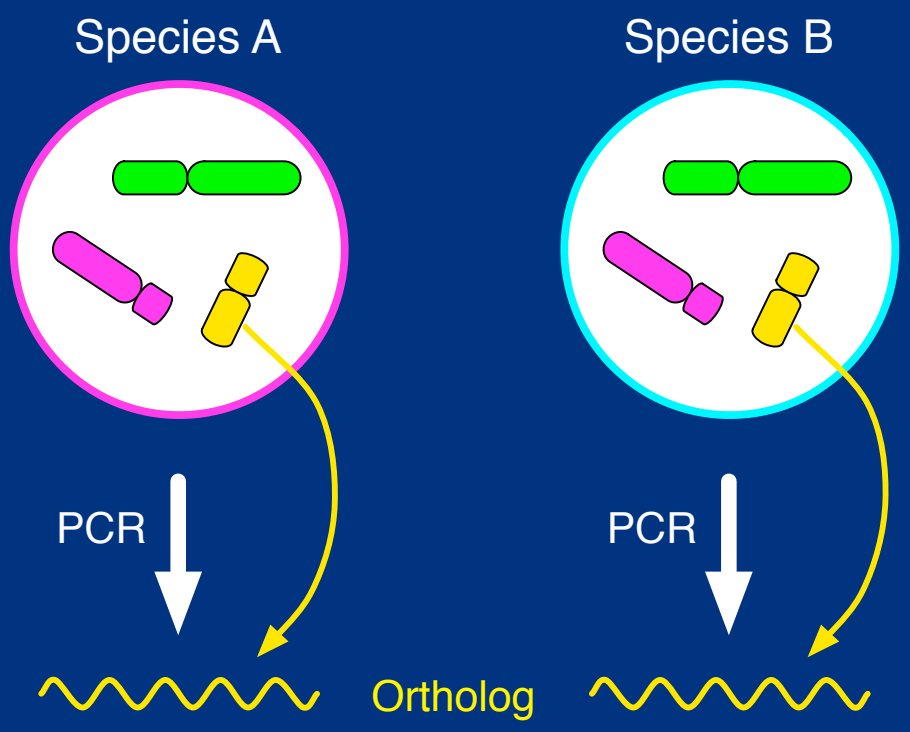
Without genomic information



Evolutionary Analysis

Power of Comparative Genomics

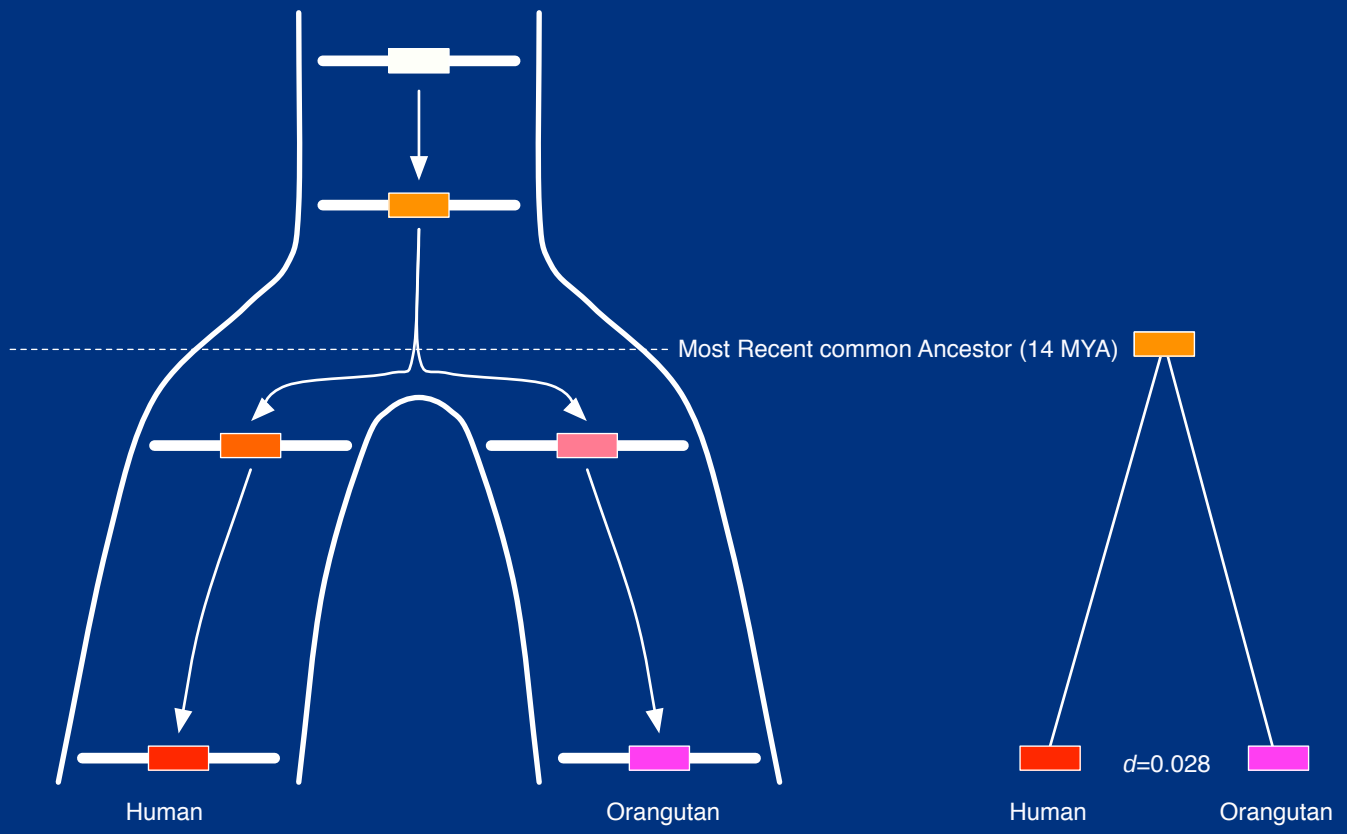
Without genomic information



Evolutionary Analysis

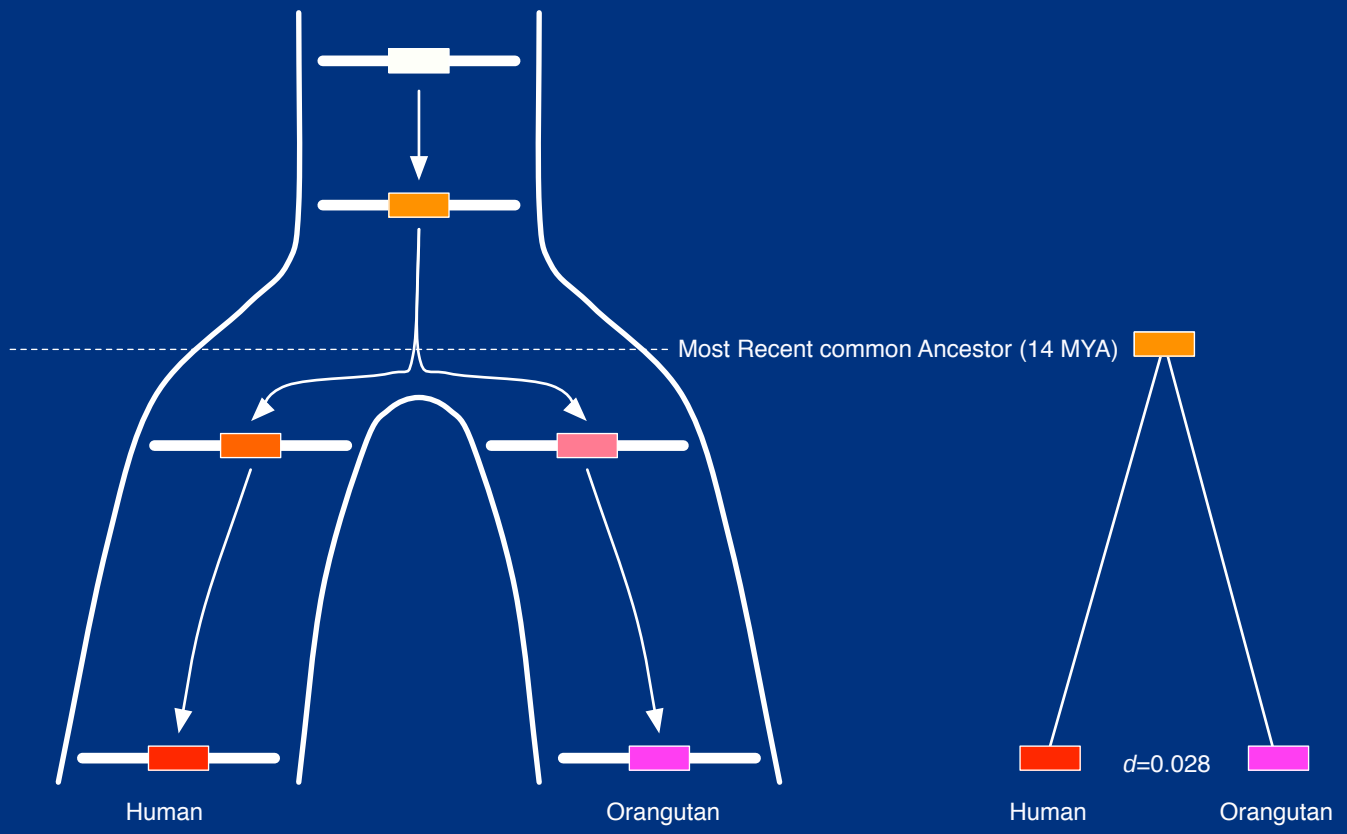
Power of Comparative Genomics

Without genomic information



Power of Comparative Genomics

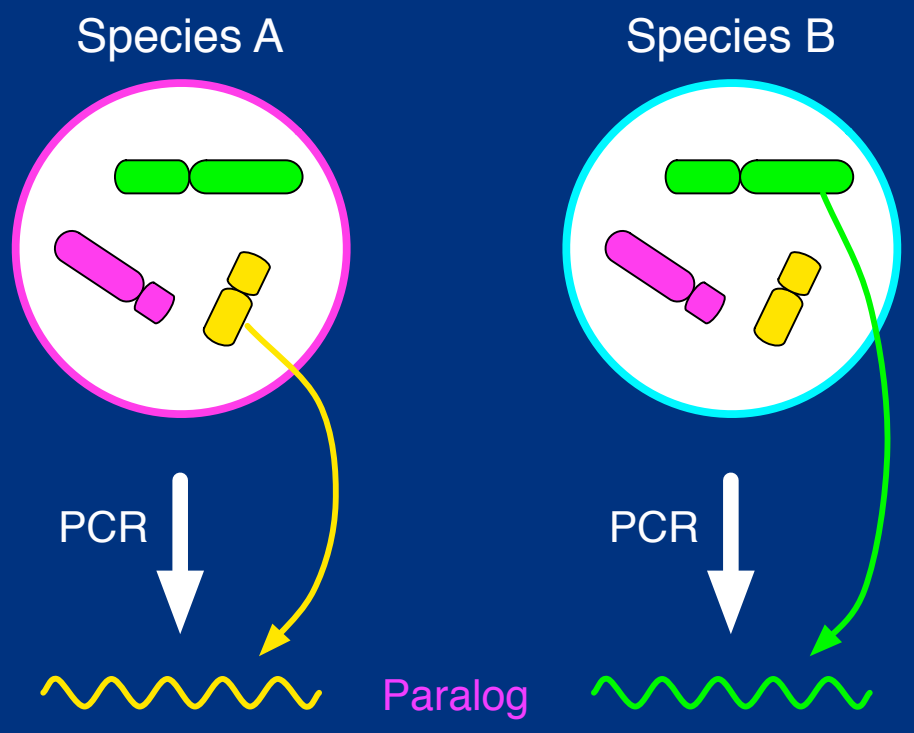
Without genomic information



"Molecular Clock"

Power of Comparative Genomics

Without genomic information

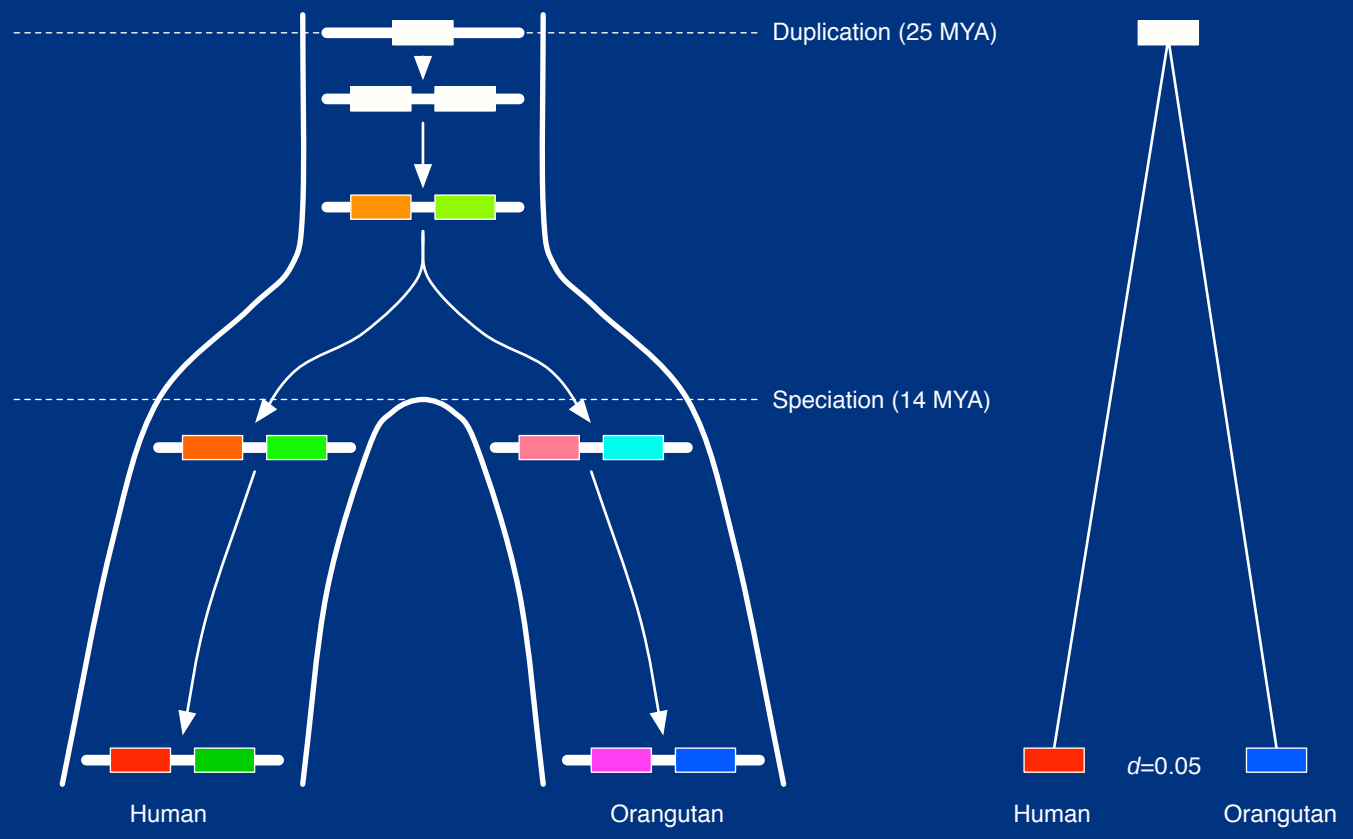


Evolutionary Analysis

Misleading interpretation!

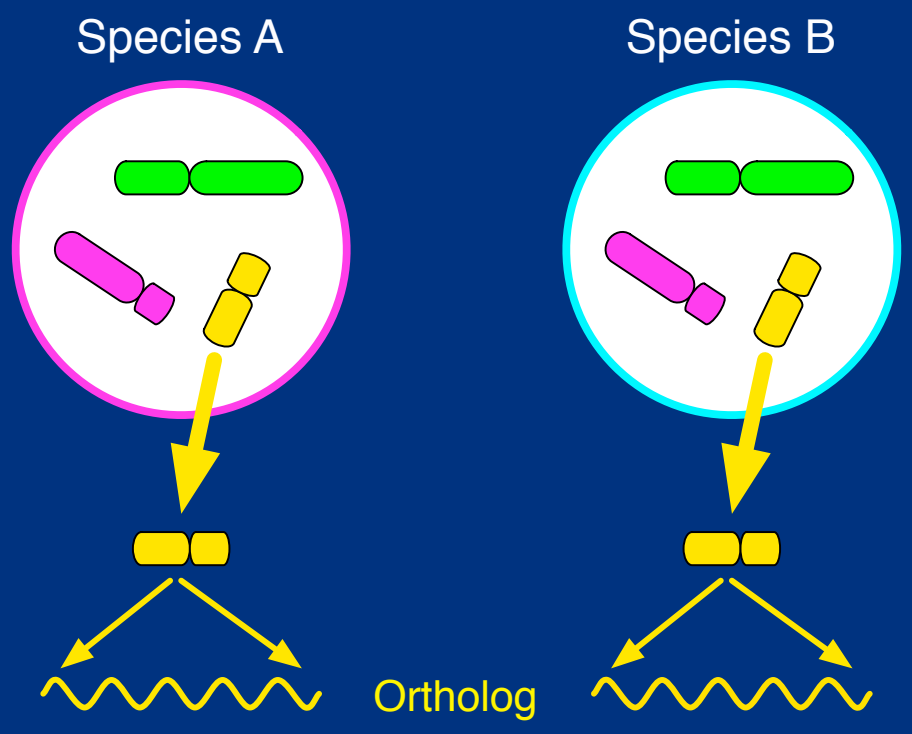
Power of Comparative Genomics

Without genomic information



Power of Comparative Genomics

With genomic information



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

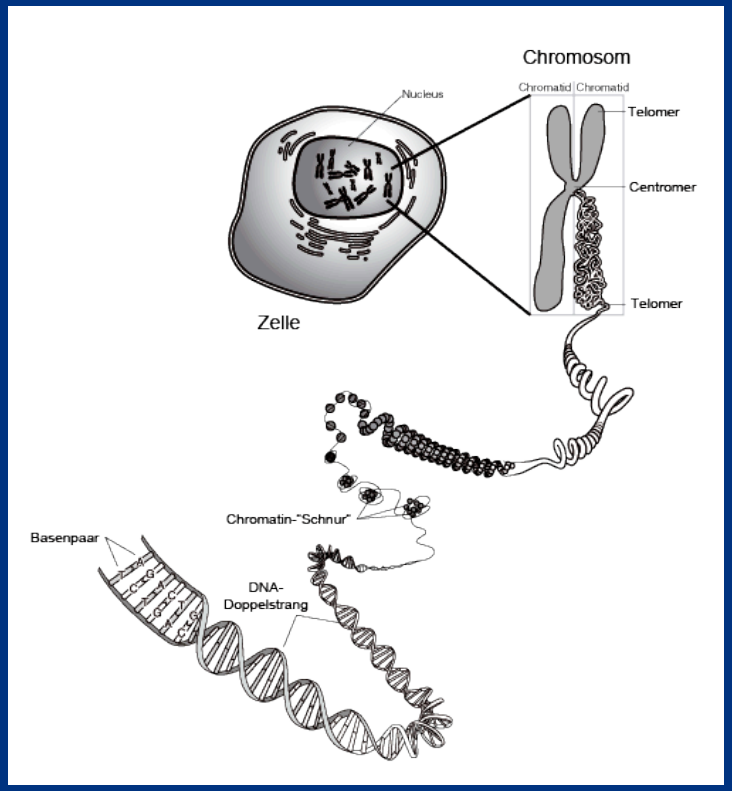
Lynch and Conery 2000 Science 290: 1151-1155

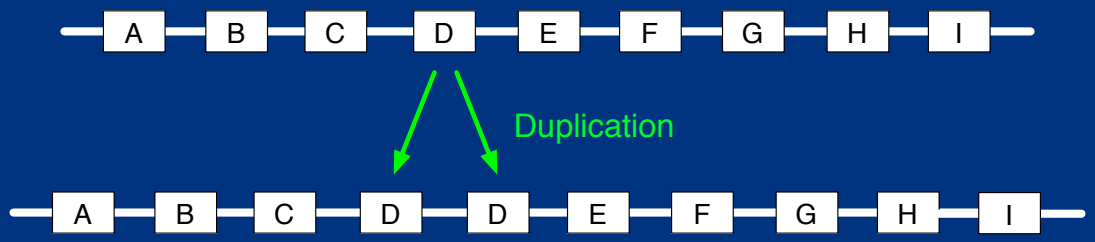
8.3 per gene per billion years

Data: *Saccharomyces cerevisiae* genome

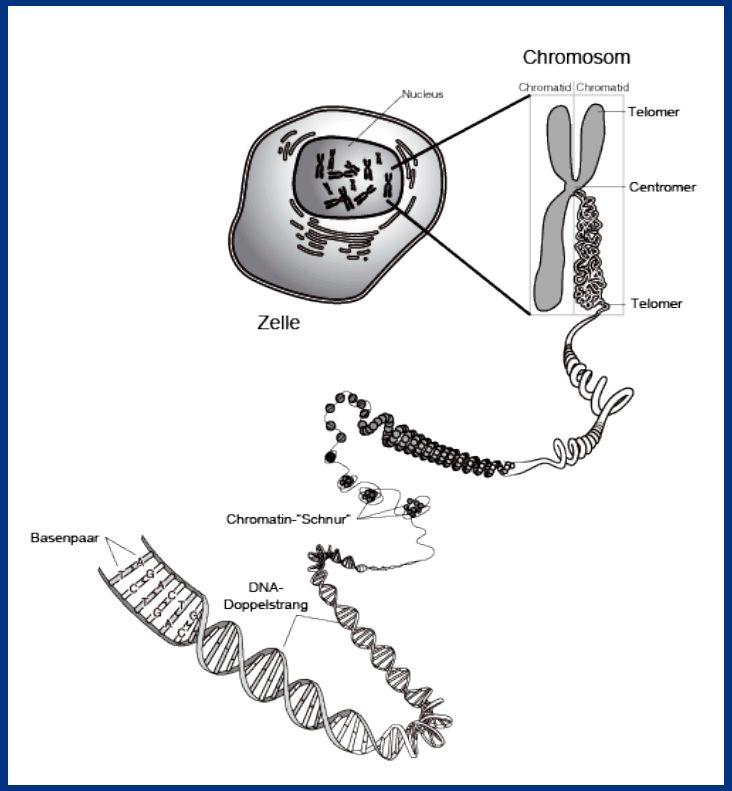
Method: Counting "looking young" duplicated genes

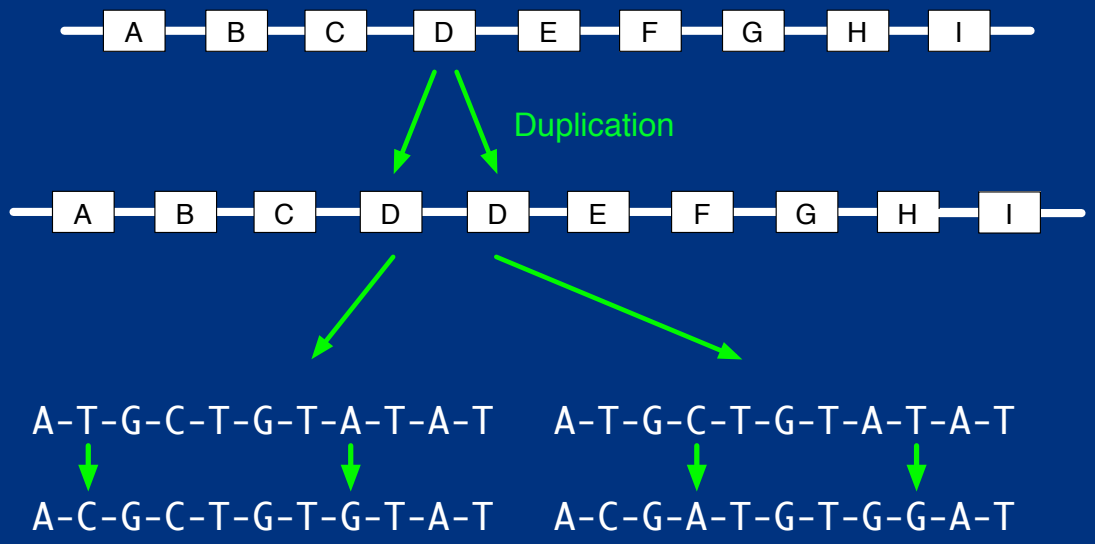
Nucleotide Evolution in Duplicated Genes





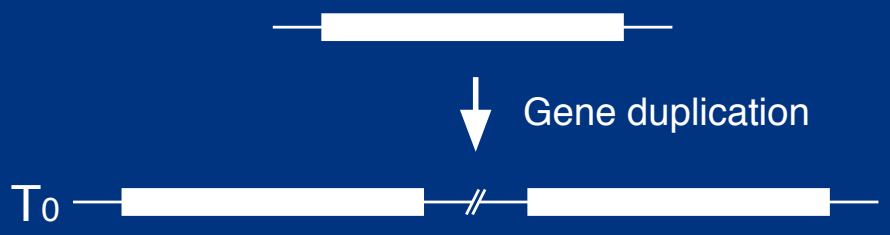
Nucleotide Evolution in Duplicated Genes

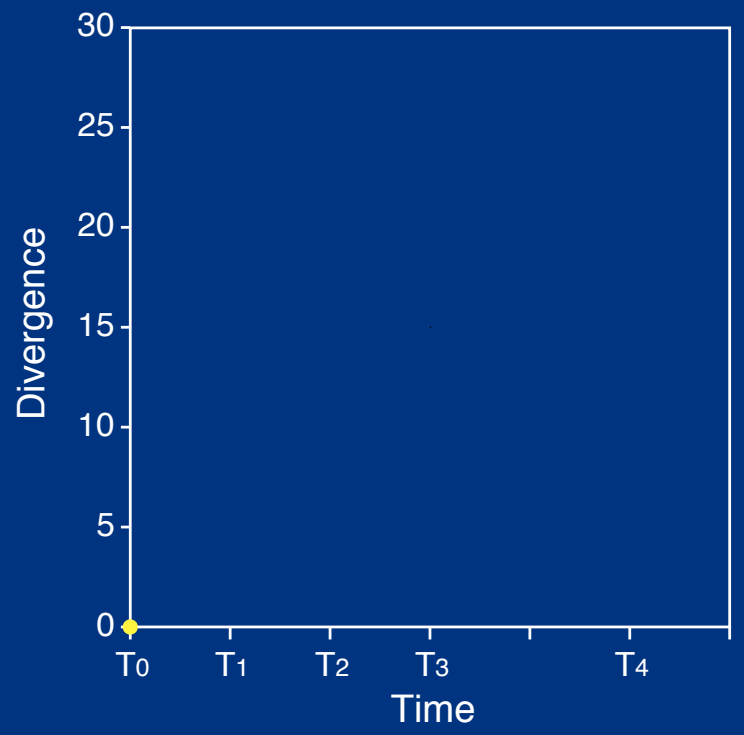




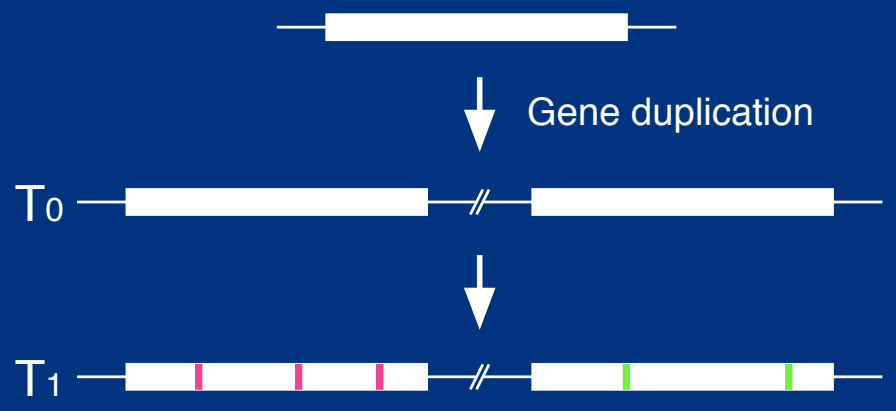
How does a molecular clock work for the divergence between duplicates?

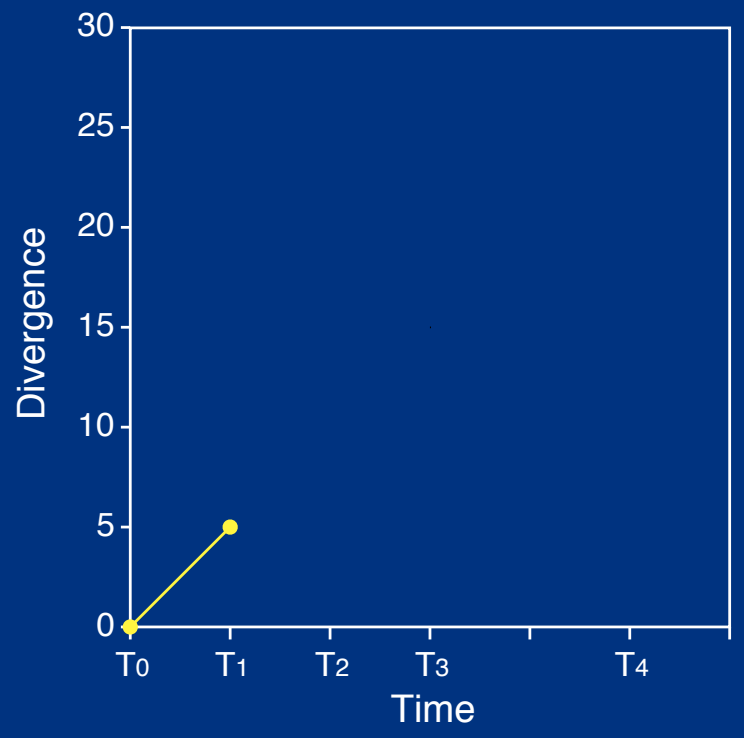
Independent Evolution



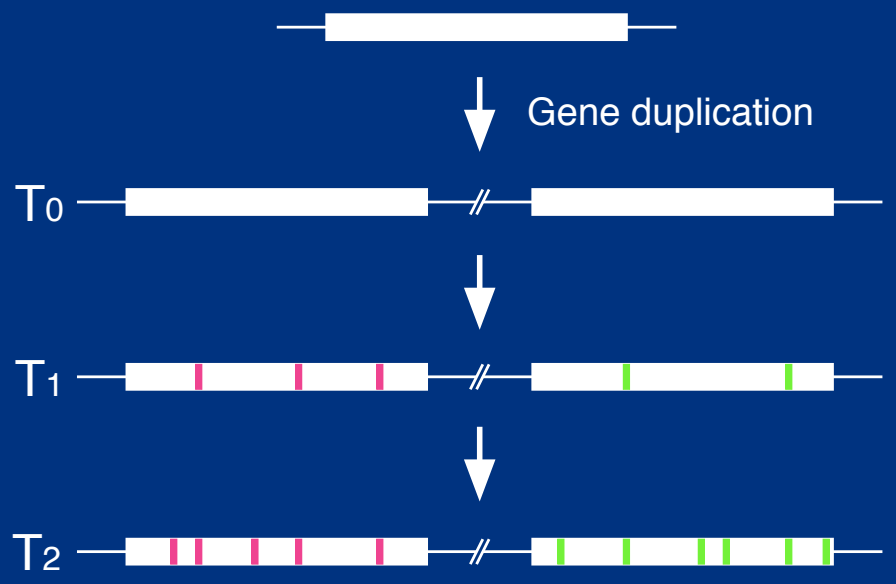


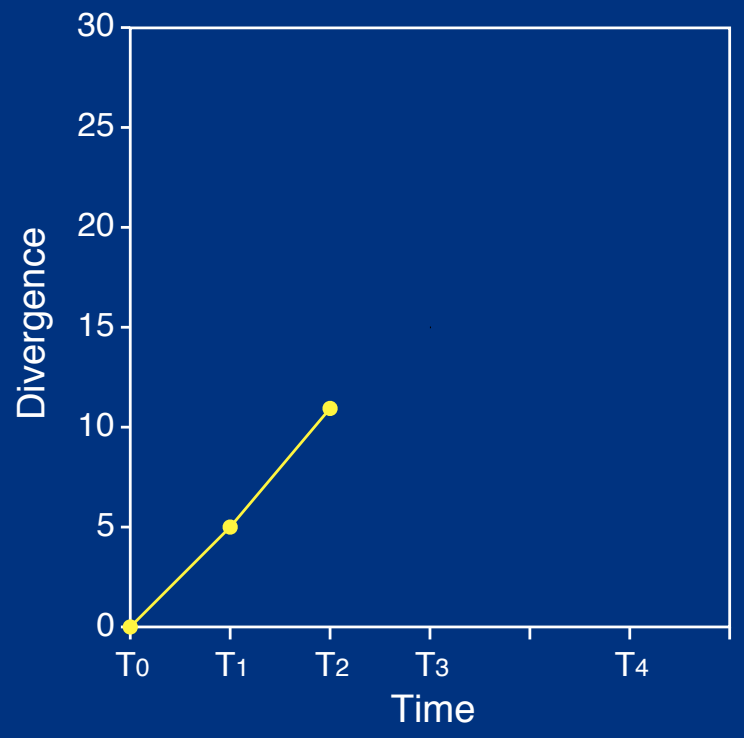
Independent Evolution



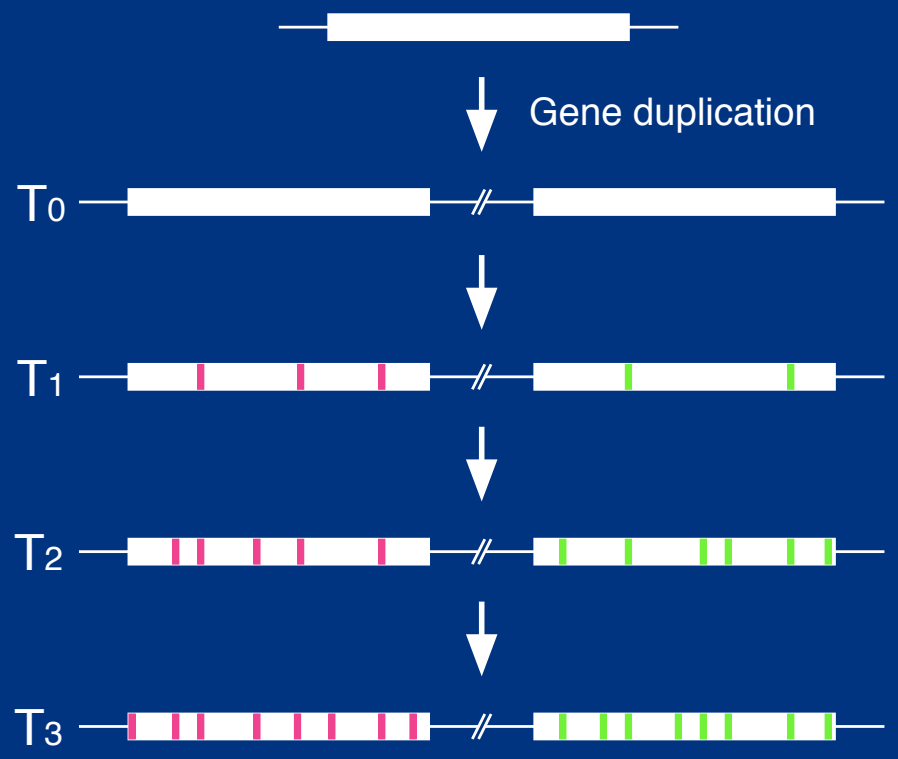


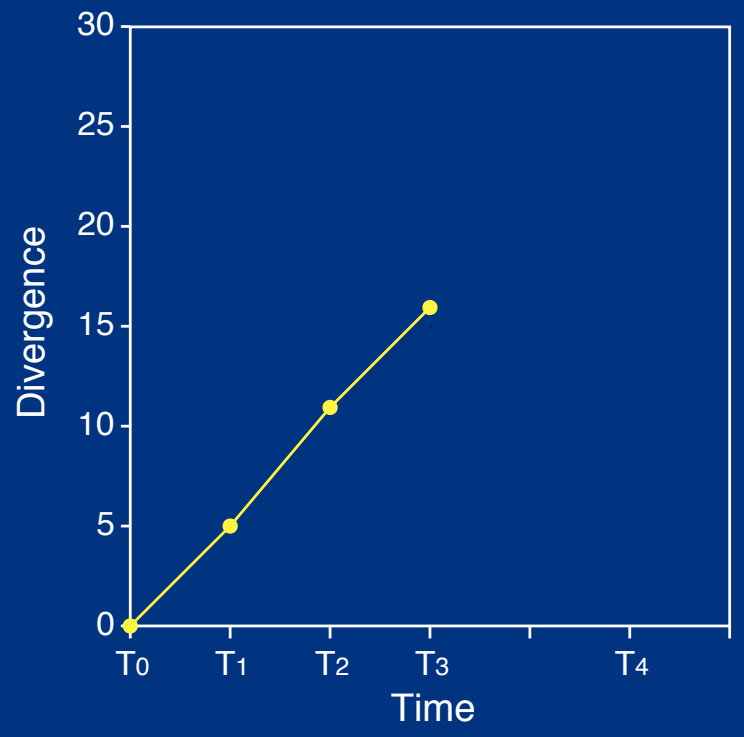
Independent Evolution



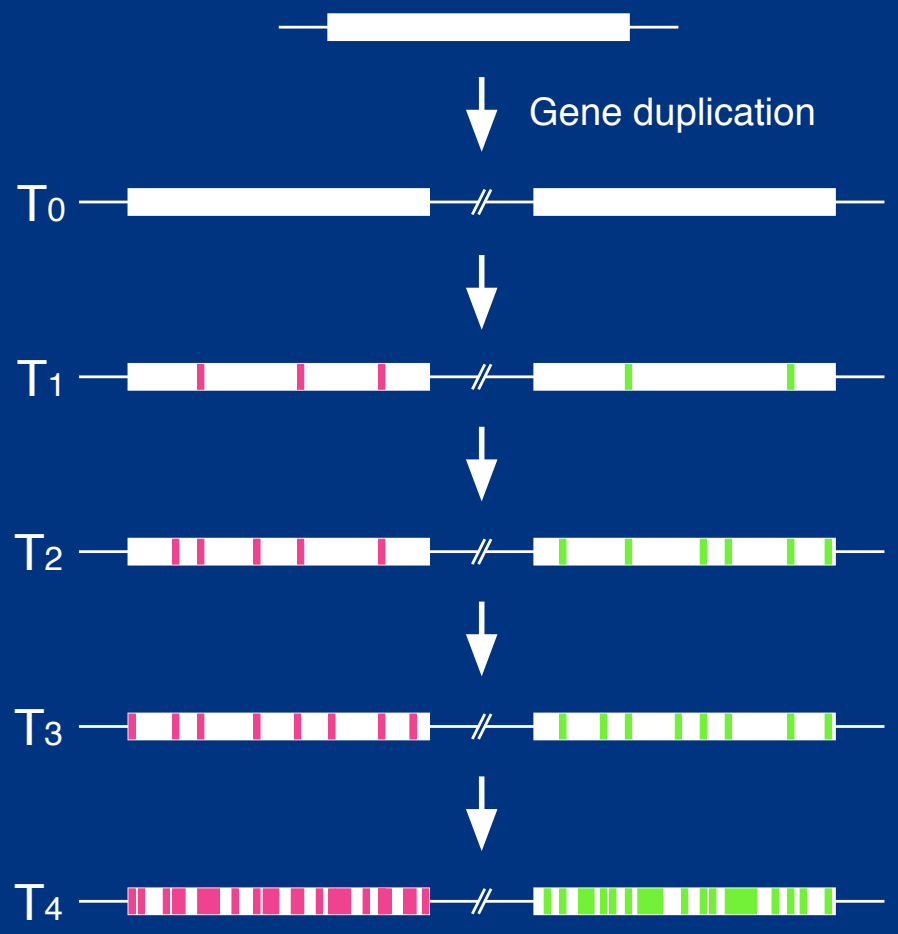


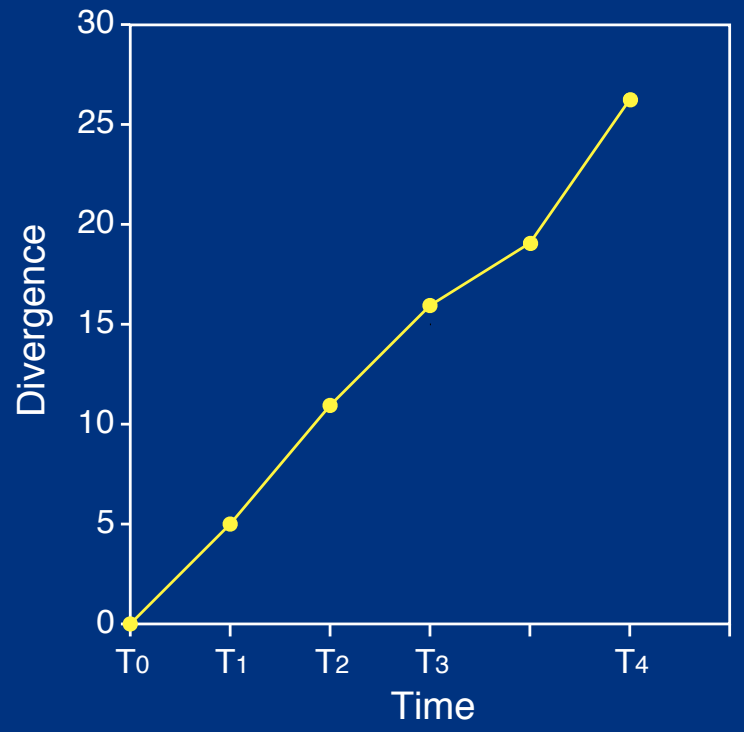
Independent Evolution



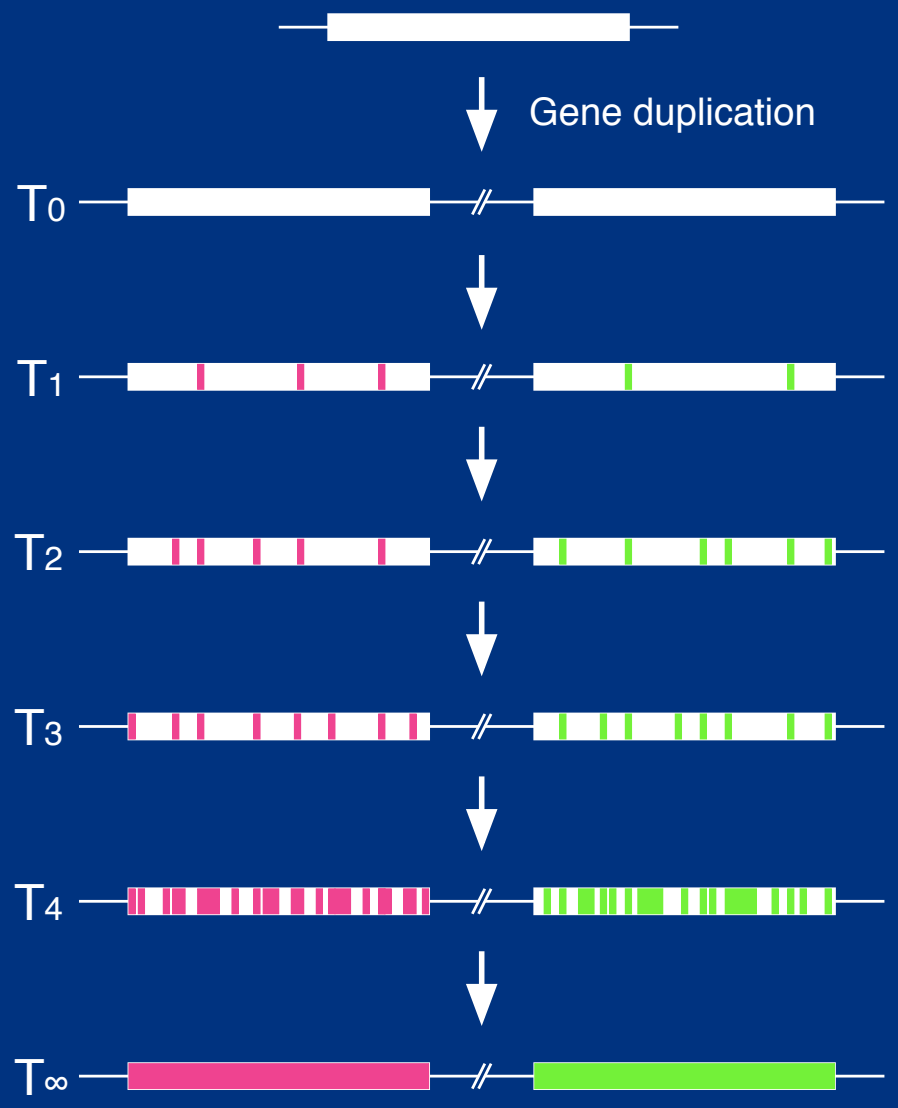


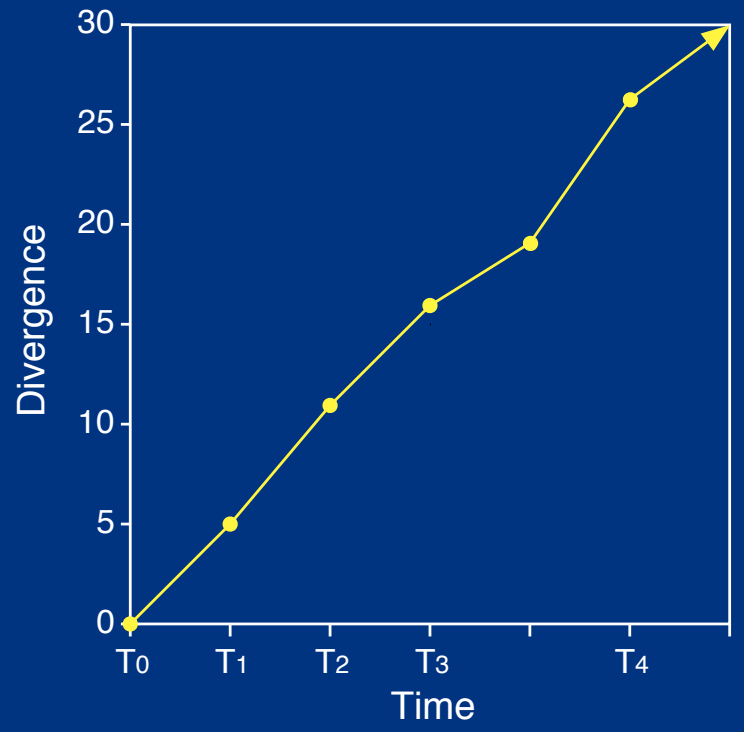
Independent Evolution



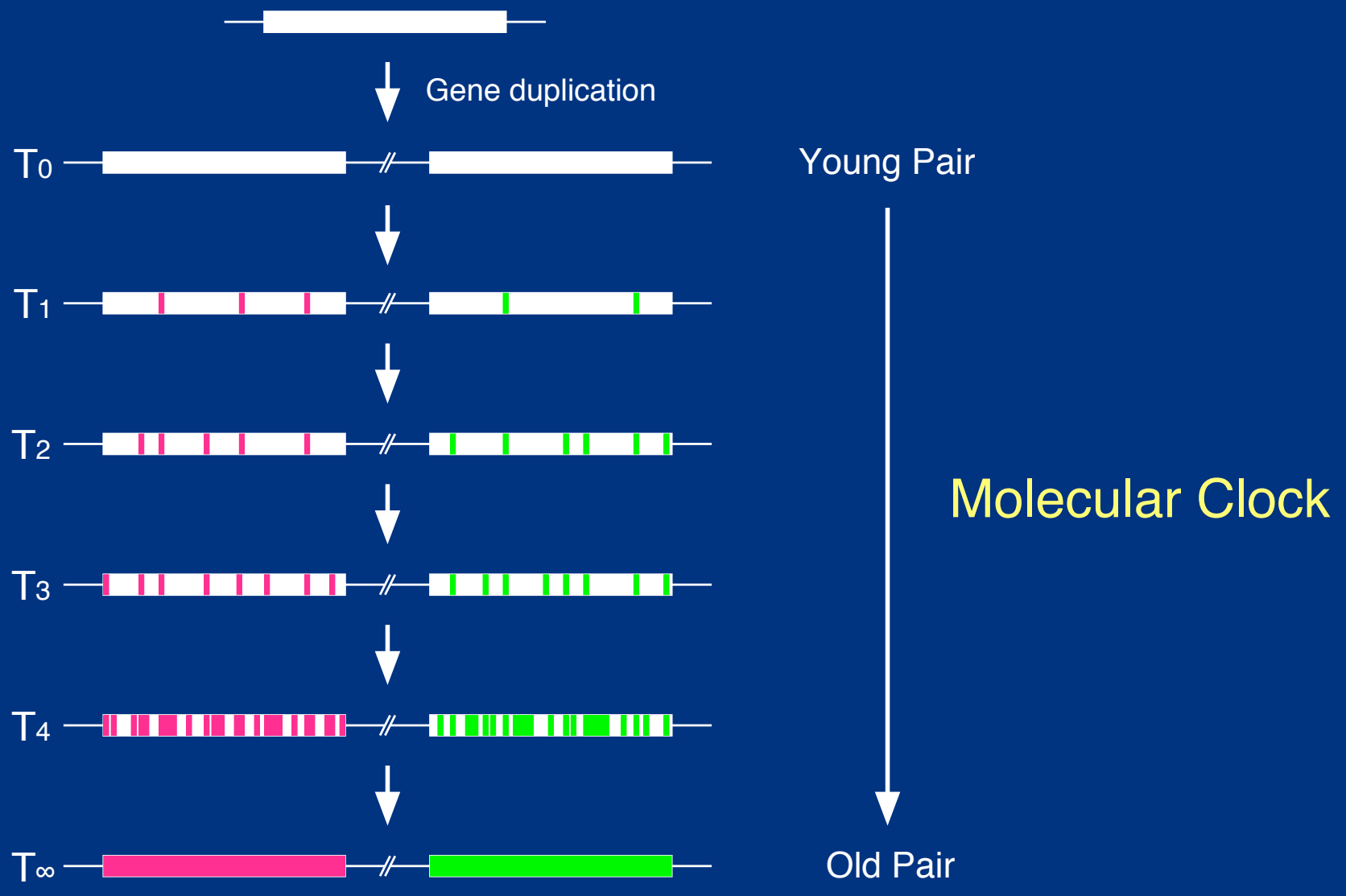


Independent Evolution

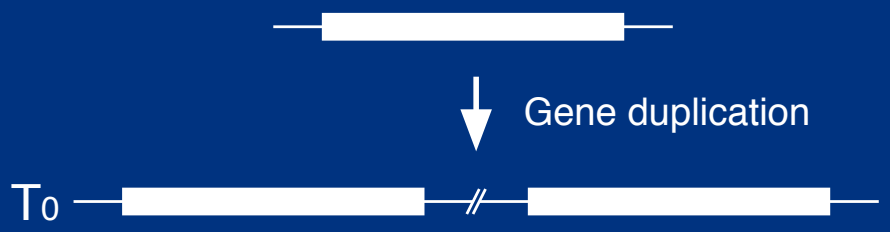


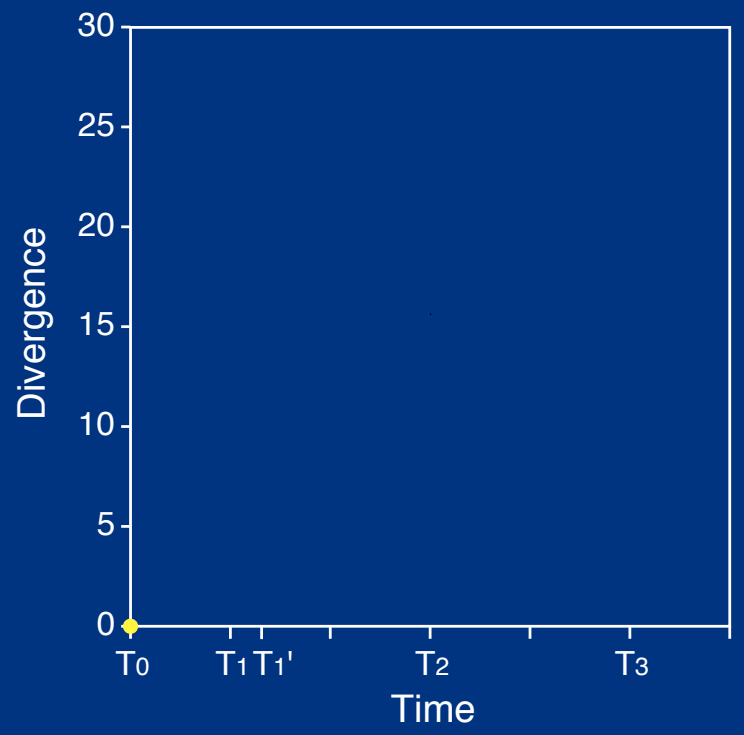


Independent Evolution

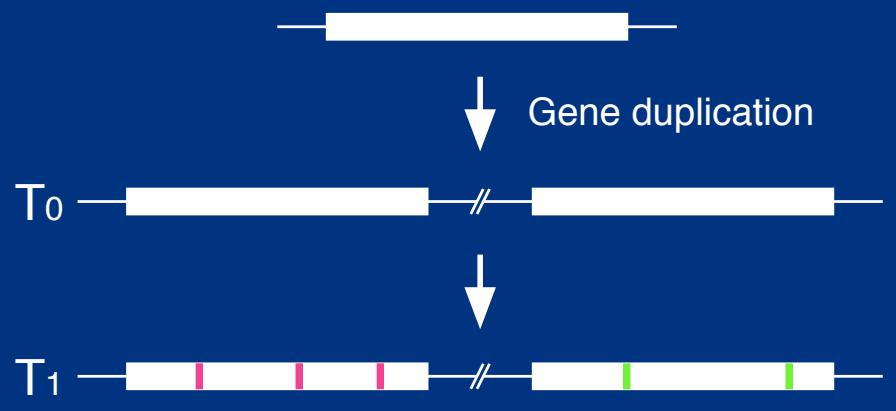


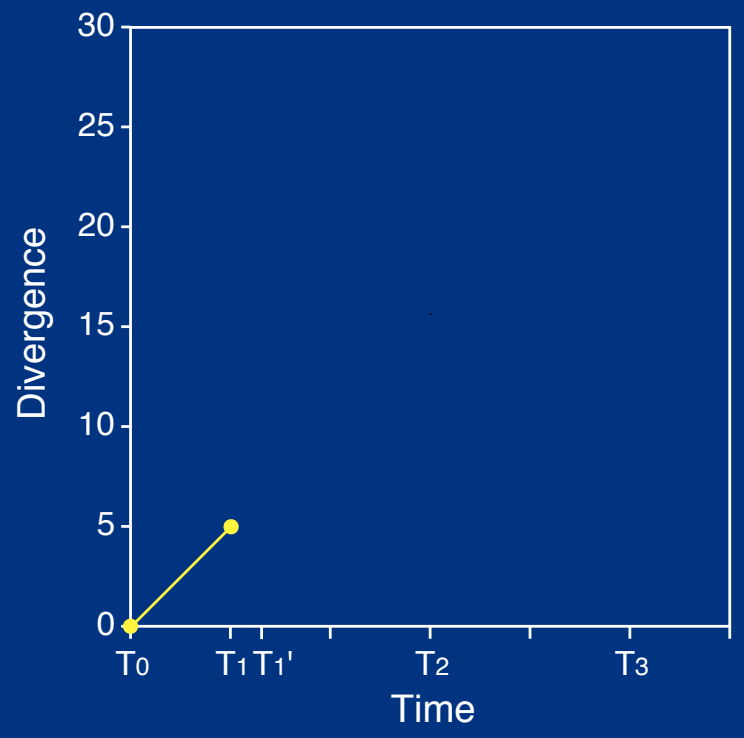
Concerted Evolution



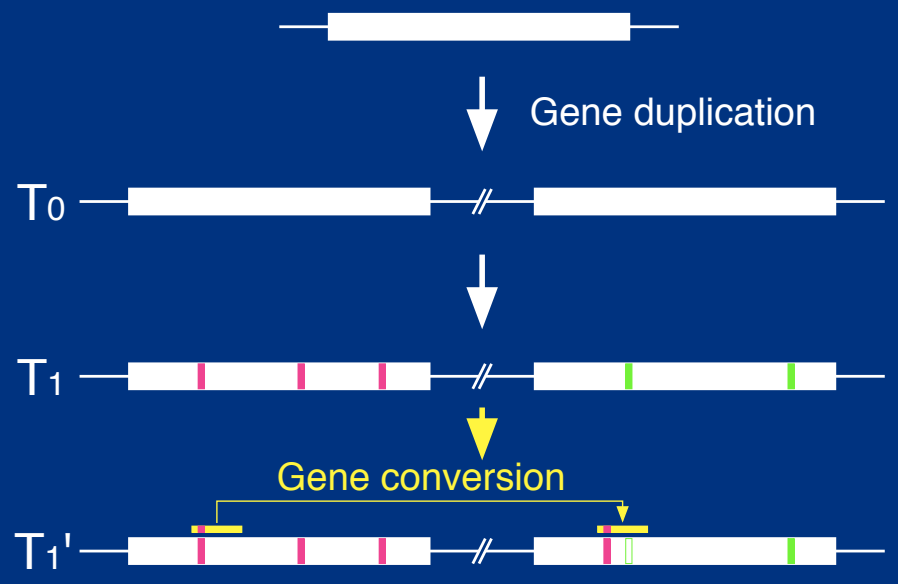


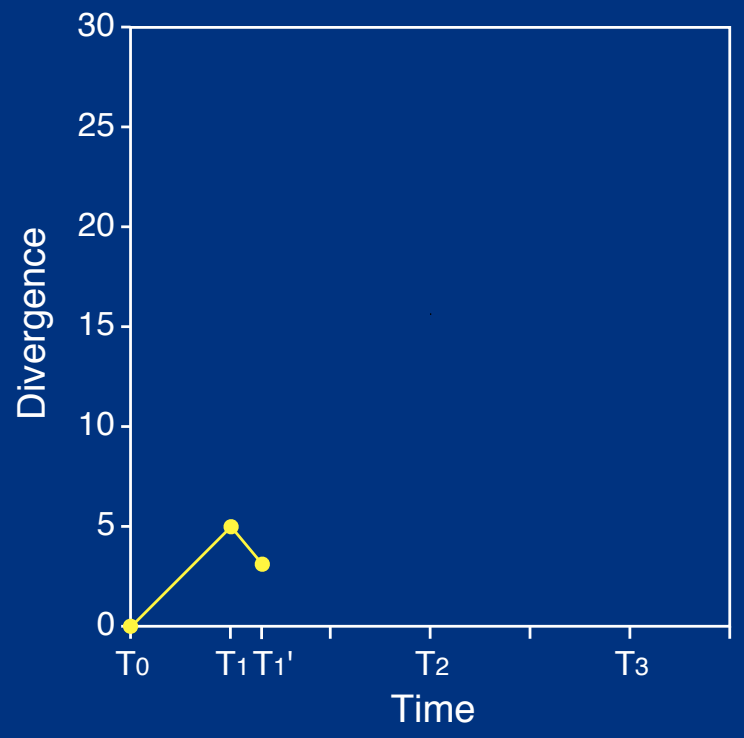
Concerted Evolution



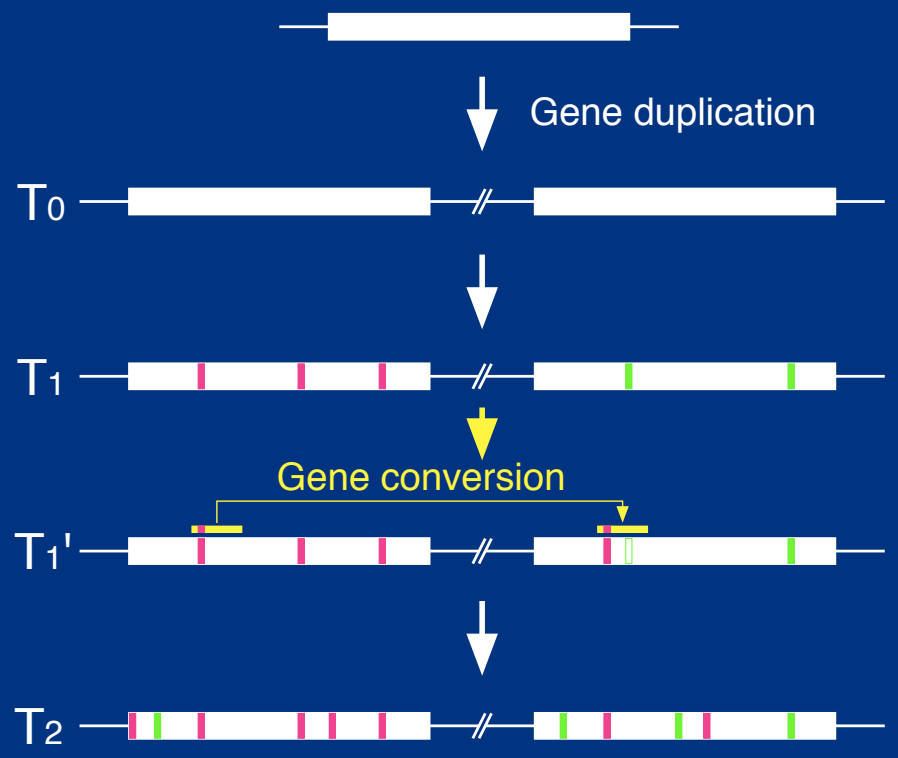


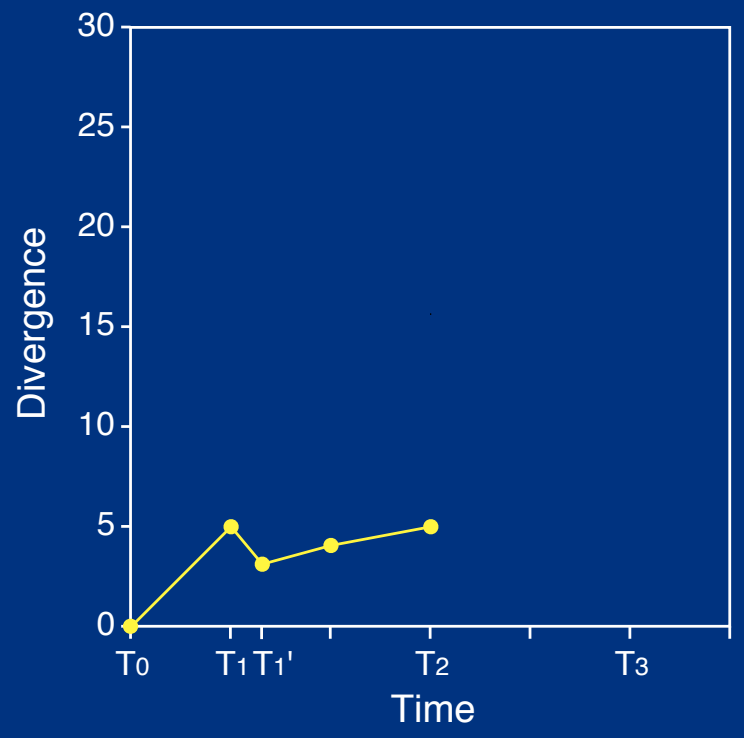
Concerted Evolution



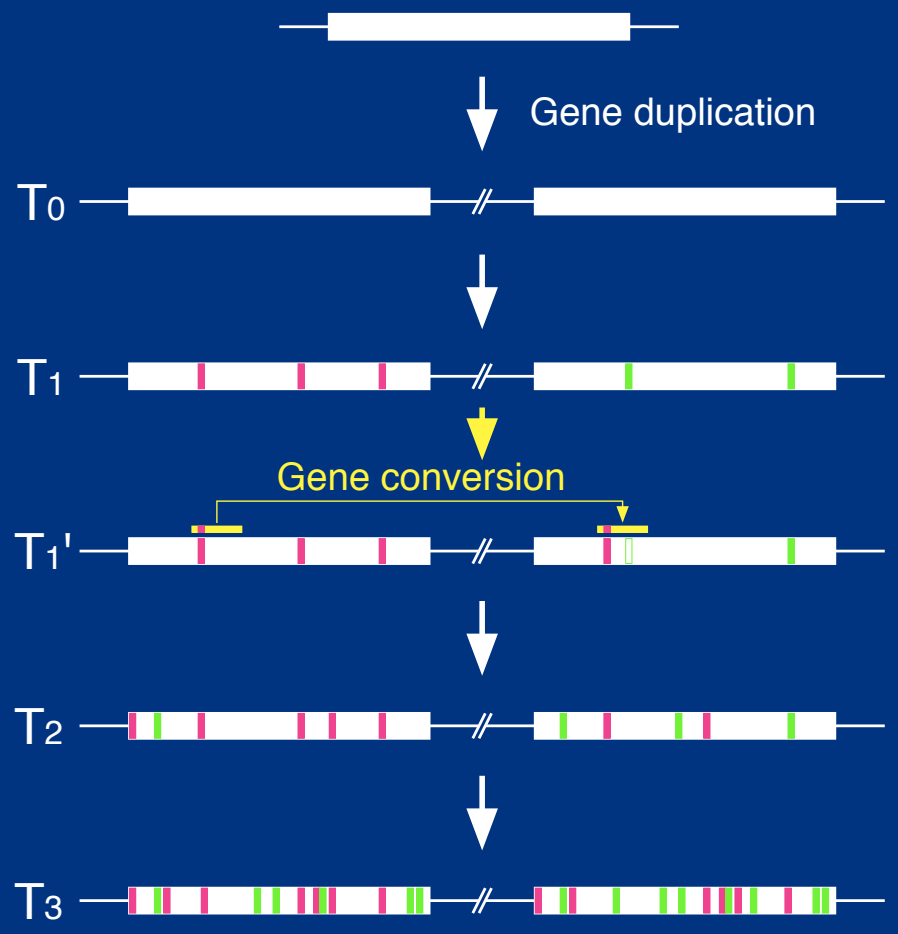


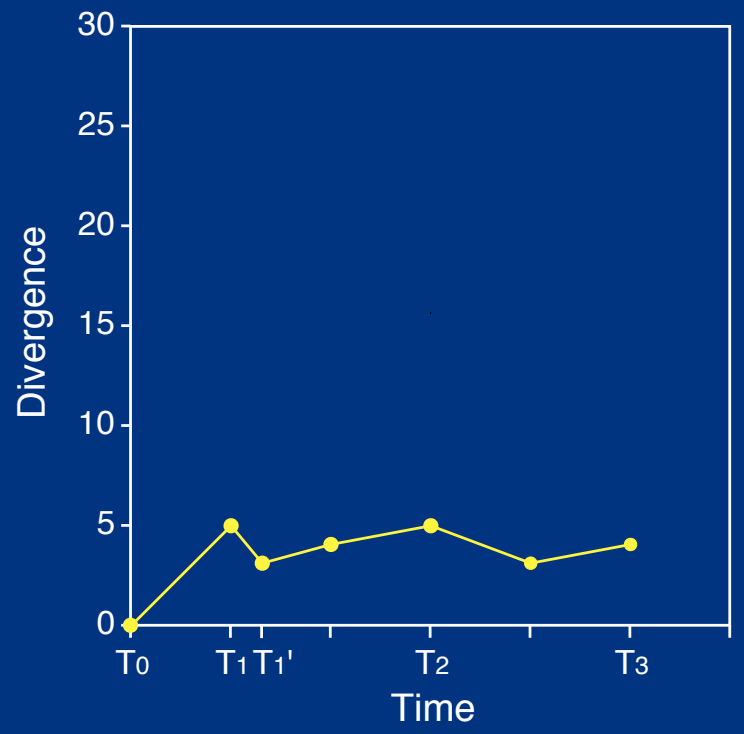
Concerted Evolution



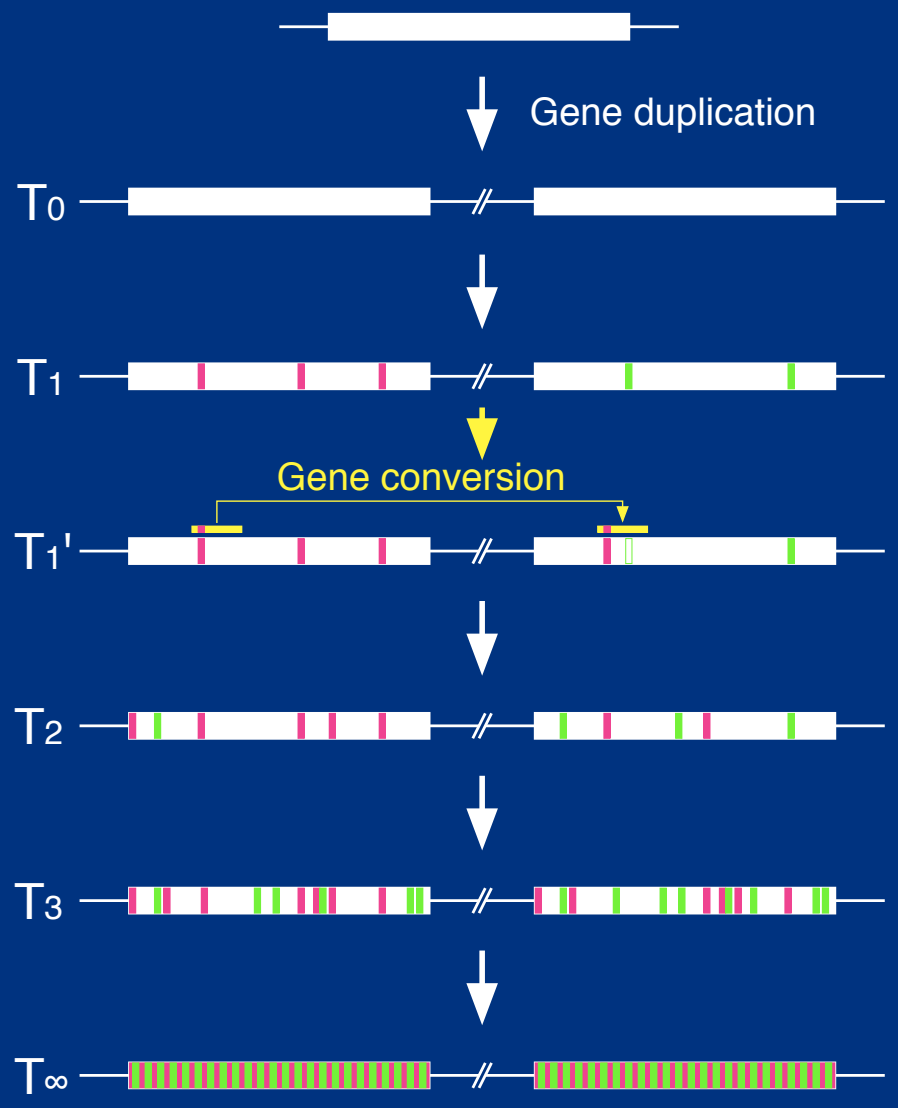


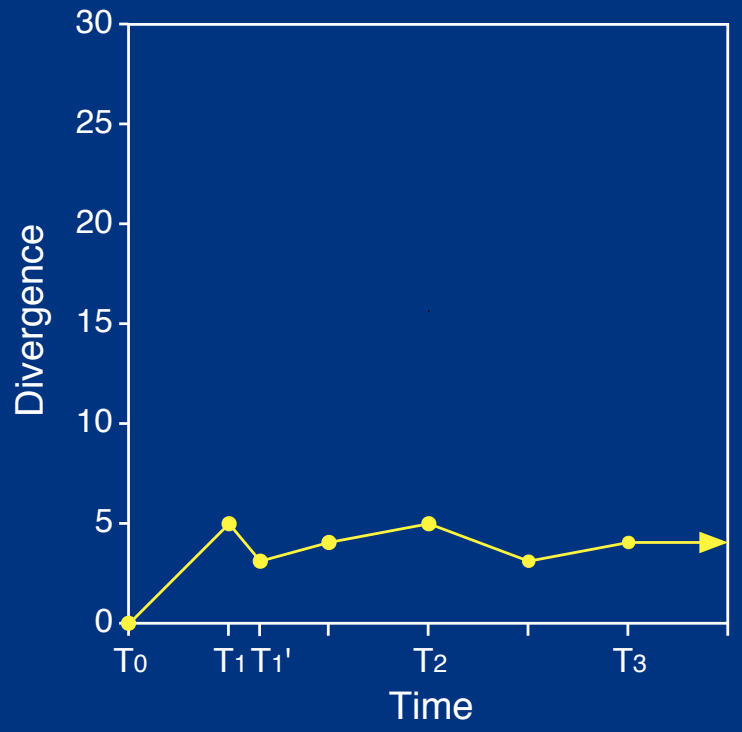
Concerted Evolution



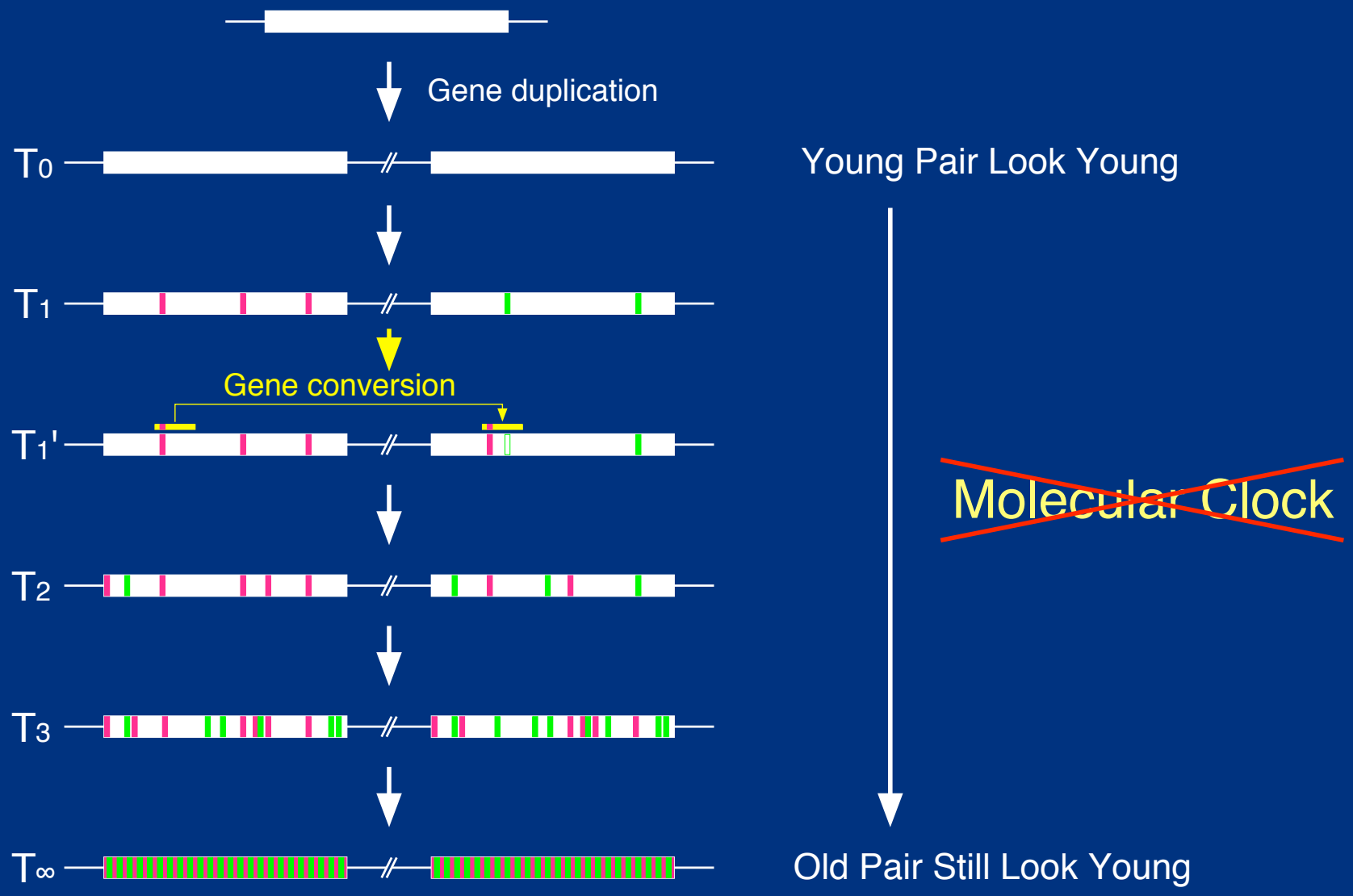


Concerted Evolution

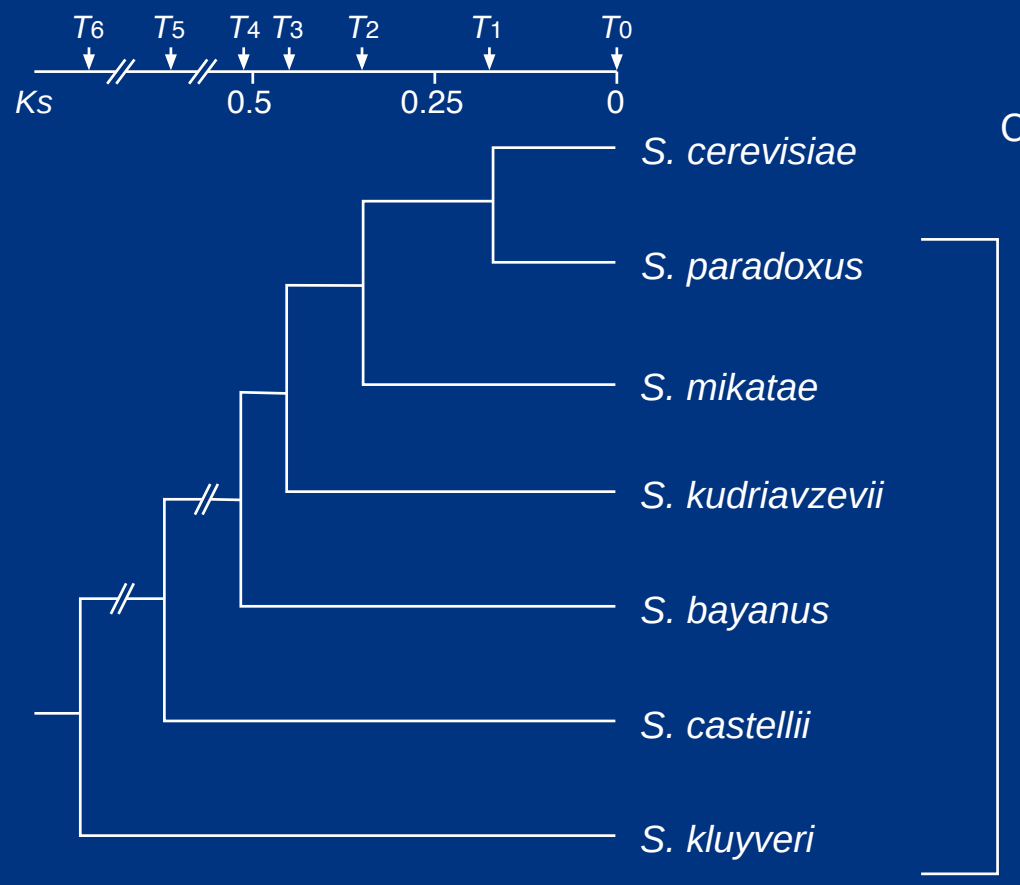




Concerted Evolution



Model-Free Estimation of the Gene Duplication Rate in Yeast

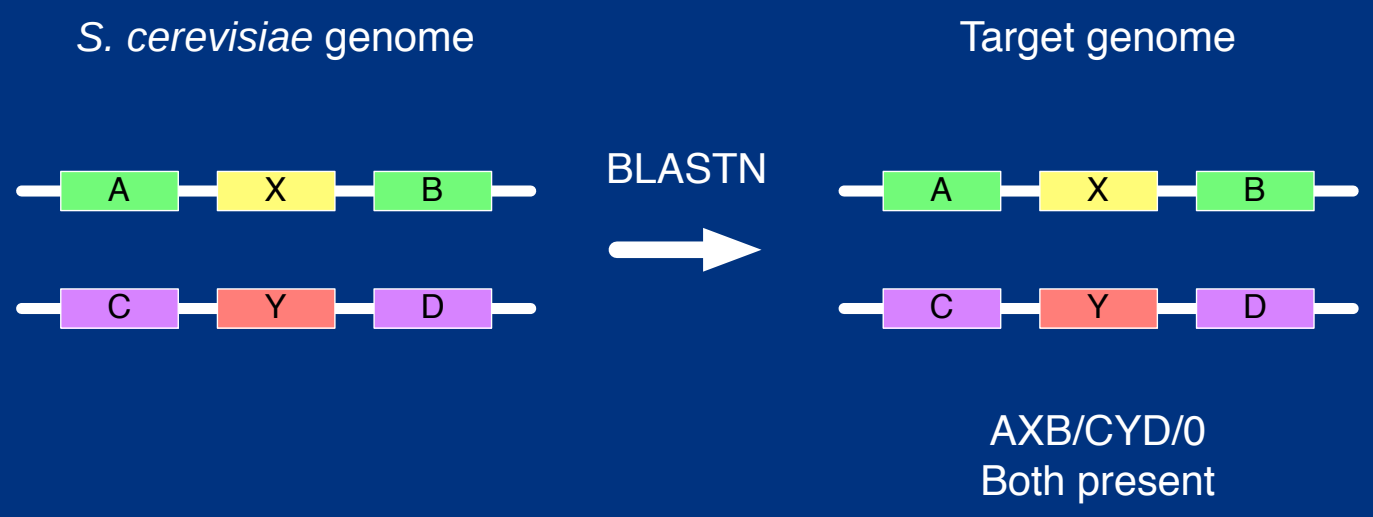


Complete Geneome Sequnece
Gofieau et al. (1996)
Science 274: 546-567

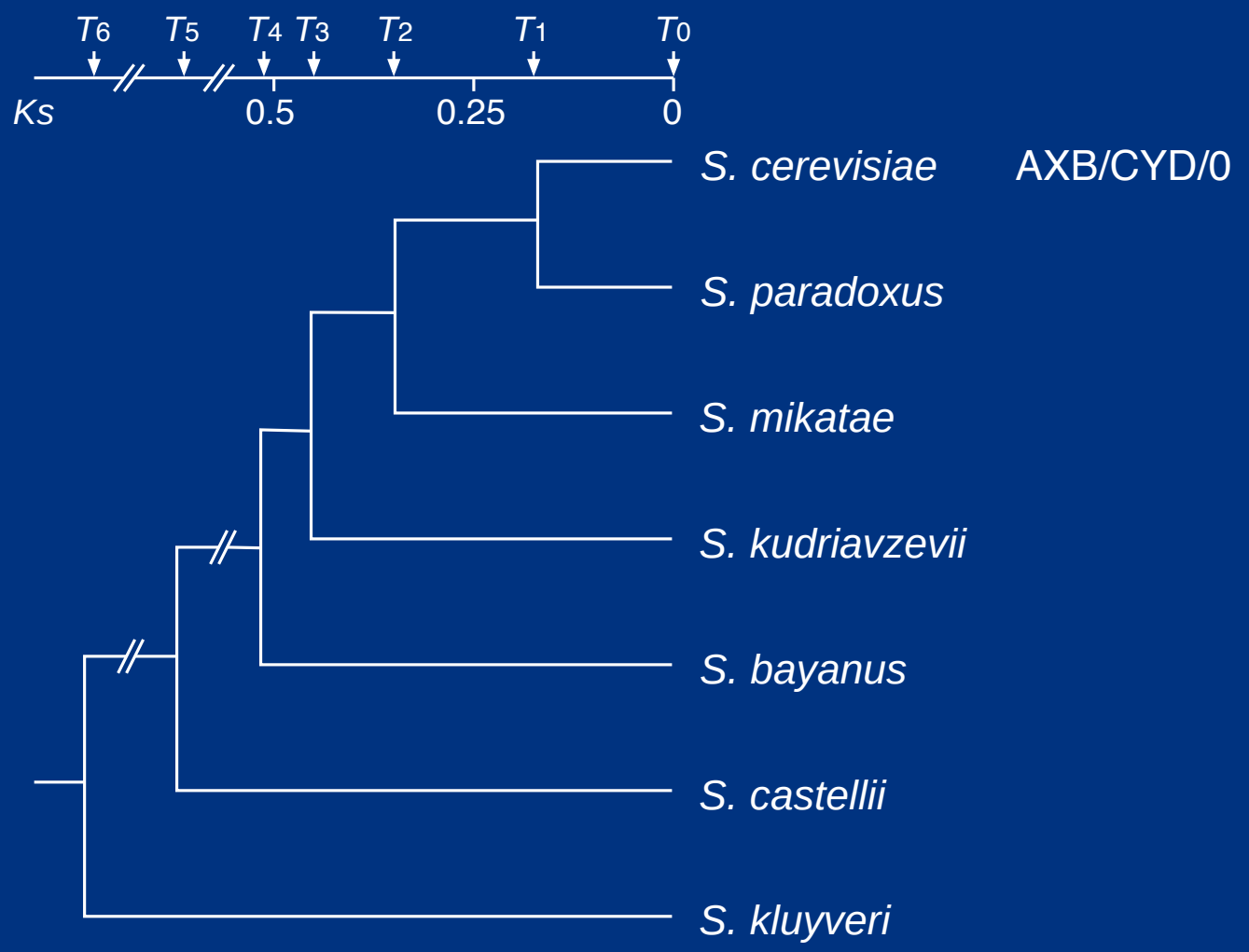
Shotgun Sequneces
Cliften et al. (2003)
Science 301: 71-76
Kellis et al. (2003)
Nature 423: 241-254

Model-Free Estimation of the Gene Duplication Rate in Yeast

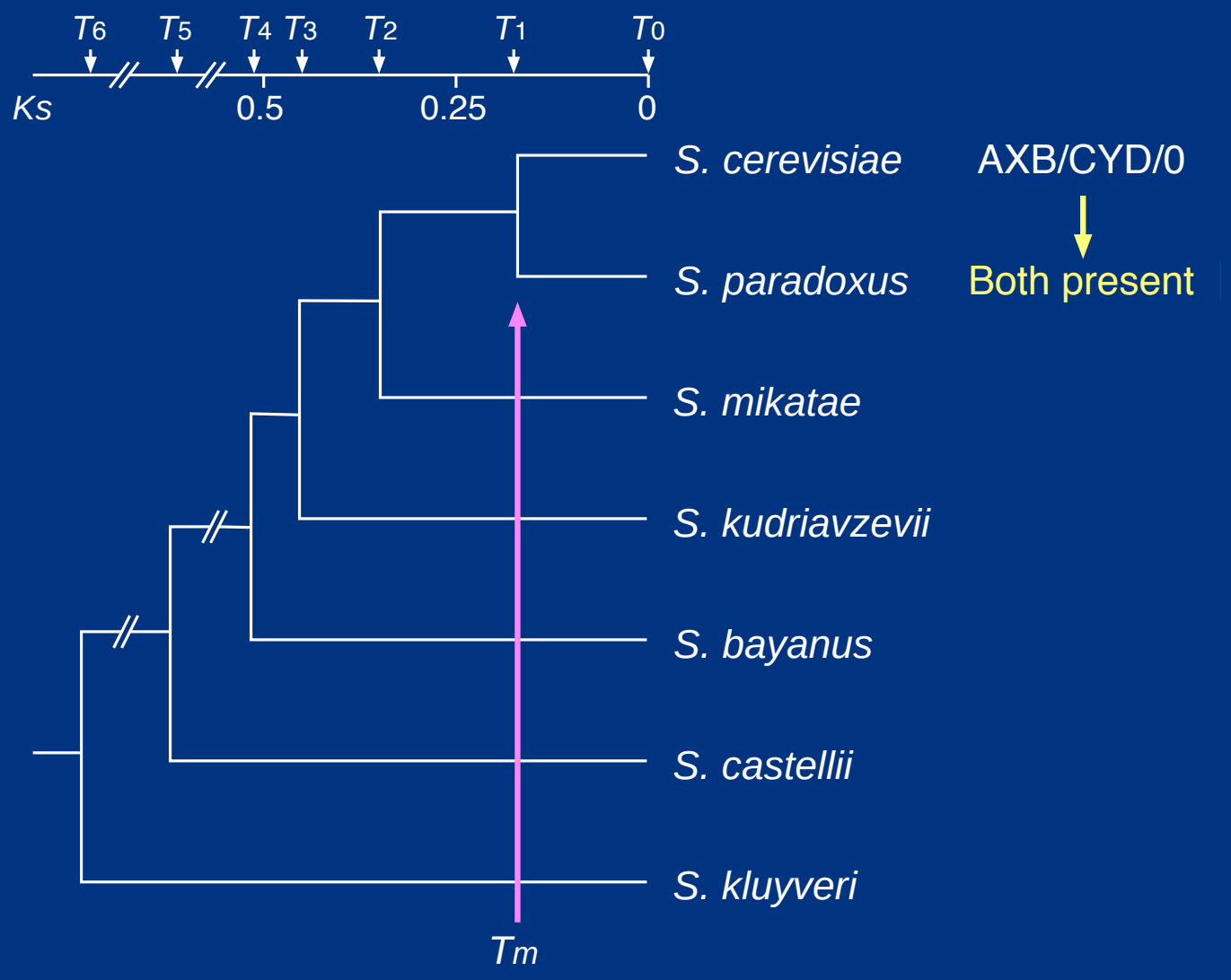
1. Identifiy dupicated genes (denoted by X and Y) in the baker's yeast (*Saccharomyces cerevisiae*) genome
2. Find evidence for the presence of the orthologs of X and Y



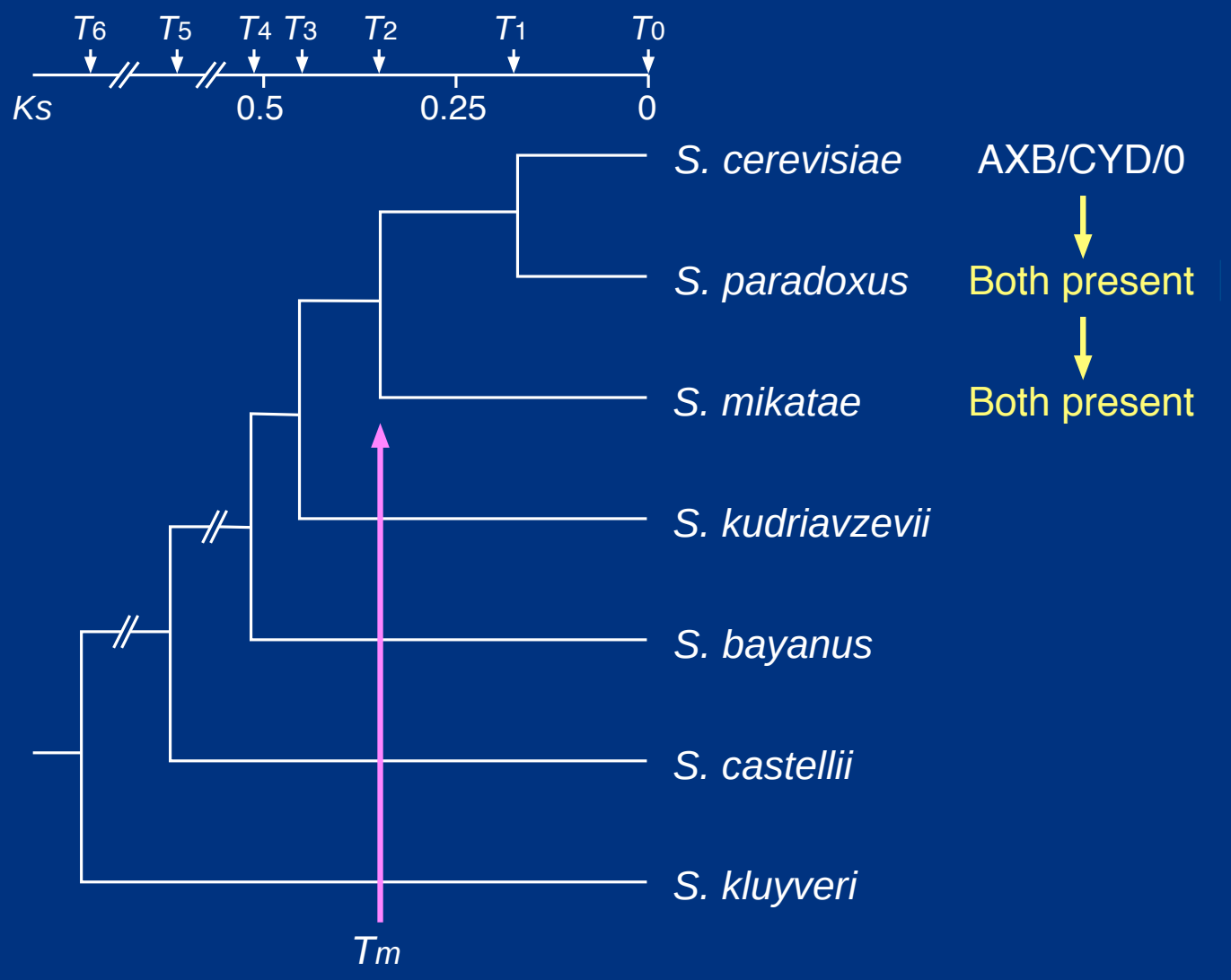
Estimating the Minimum Age, T_m



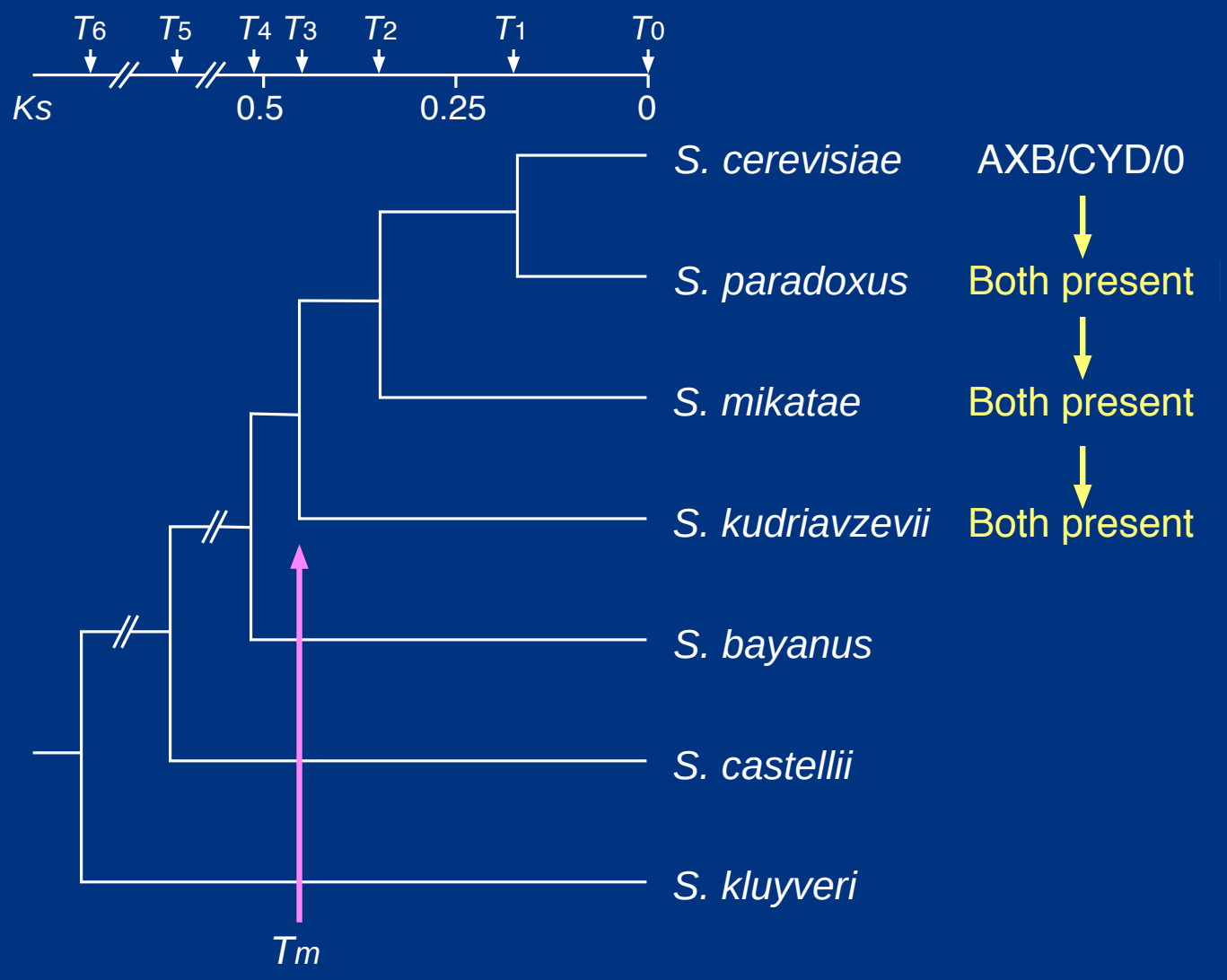
Estimating the Minimum Age, T_m



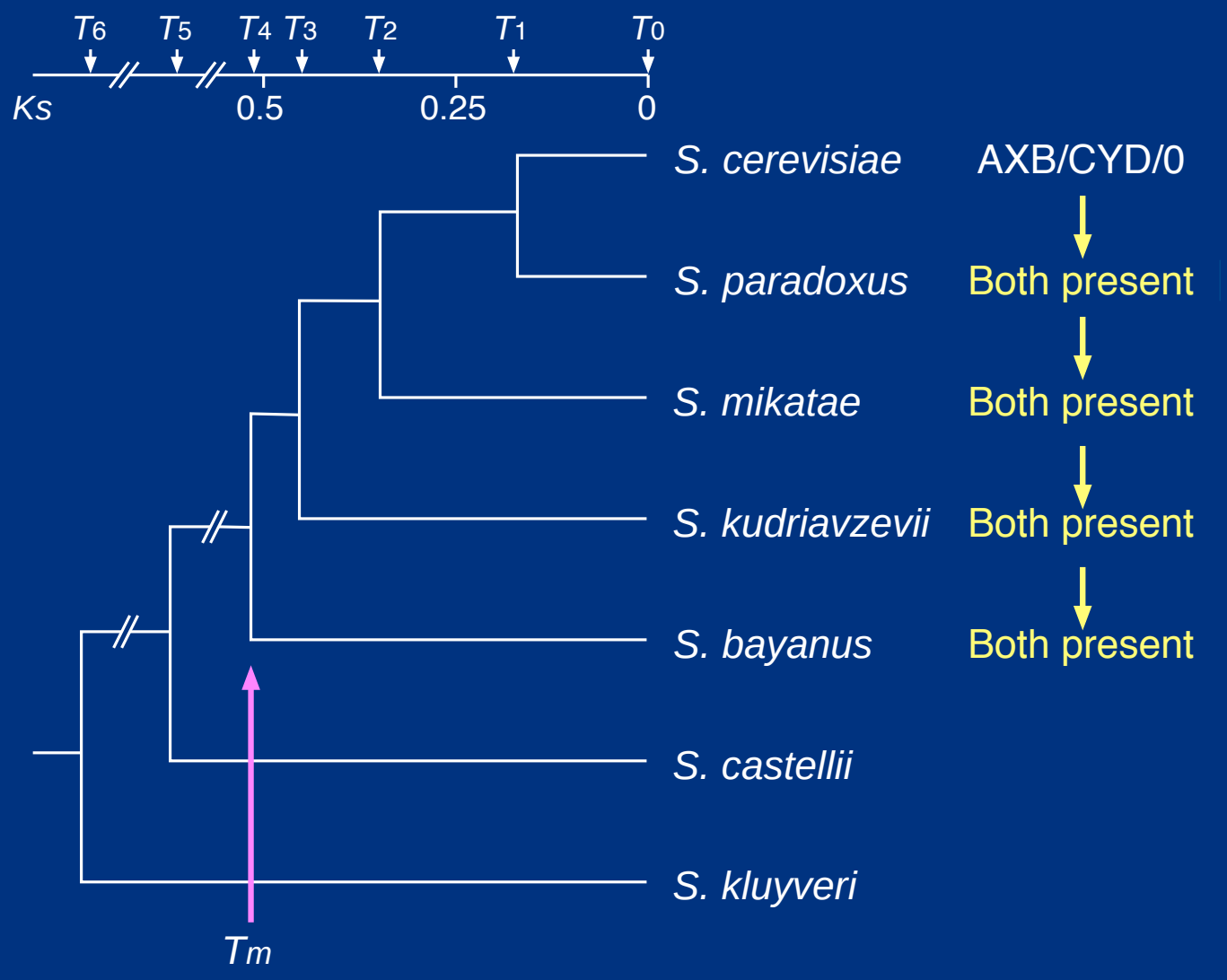
Estimating the Minimum Age, T_m



Estimating the Minimum Age, T_m



Estimating the Minimum Age, T_m



Minimum Ages of Duplicated Genes in the Baker's Yeast Genome

	X	Y	Ks	*S. paradoxus*	*S. mikatae*	*S. kudriavzevii*	*S. bayanus*	*S. castellii*	*S. kluyveri*	T_m
1*	YHR053C	YHR055C	0	AXB	AXB	AX-	AXB	—	—	0
2	YCL066W	YCR040W	0	AXB/-YD/2	—/CYD/2	XB/CYD×2/1	AXB/-YD/1	A-B/—/0	—/—/0	4
3	YNL019C	YNL033W	0	AXB/-YD/1	—/-YD/0	—/CYD/0	AX-/CYD/0	—/—/0	—/—/0	1
4	YAR064W	YHR213W-B	0	—/-YD/3	—/—/0	—/—/1	—/—/0	—/—/0	—/—/0	0
5	YFL061W	YNL335W	0.0053	—/-YD/0	—/—/1	—/—/0	—/—/0	—/—/0	—/—/0	0
6	YGL135W	YPL220W	0.0209	AXB/CYD/0	AXB/CYD/0	-XB/-YD/0	AXB/CY-/0	A-B/CYD/0	AX-/—/0	4
7	YOR390W	YPL279C	0.0209	AXB/—/0	A-B/—/1	-XB/—/1	—/—/1	—/—/0	—/—/0	0
8	YDL136W	YDL191W	0.0265	AXB/CYD/0	-XB/CY-/0	AXB/CY-/0	AXB/CYD/0	A-B/CYD/0	—/CYD/0	4
9	YBR181C	YPL090C	0.0298	—/CYD/1	AX-/CYD/1	AXB/CYD/0	-XB/—/1	—/CYD,-YD/0	—/—/1	3
10	YNL018C	YNL034W	0.0329	AXB/CY-/1	—/CY-,-YD/0	—/CYD/0	—/CYD/2	—/—/0	—/—/0	1
11	YBR031W	YDR012W	0.0354	AXB/CYD/0	AXB/CY-,-YD/0	AXD/CYB/0	AXD/CYB/0	AXD/—/0	AX-/—/0	4
12	YHR141C	YNL162W	0.0677	AXB/CYD/0	-XB/CYD/0	AXB/CYD/0	AXB/CYD/0	AXB/-YD/0	—/—/1	5
13	YHR203C	YJR145C	0.0699	AXB/CYD/0	AXB/CYD/0	AXB/—/2	AXB/CY-/0	—/—/2	—/—/1	4
14	YER074W	YIL069C	0.0699	AXB/CYD/0	AXB/CYD/0	AXB/CYD/0	AXB/CYD/0	—/CY-/1	—/—/1	4
15	YBL072C	YER102W	0.0947	AXB/CYD/0	AX-/CYD/0	AX-/CYD/0	AX-/—/1	—/CY-/1	—/—/1	3
16	YBR009C	YNL030W	0.1339	AXB/CYD/0	AXB/CYD/0	—/—/2	-XB/CYD/0	-XB/CYD×2/0	—/CY-/1	5
17	YHL001W	YKL006W	0.1343	AXB/CYD/0	AX-/-YD/1	AXB/CYD/0	AXB/CY-/0	-XB/CYD/0	—/—/1	5
18	YIL018W	YFR031C-A	0.1412	AXB/CYD/0	AX-/CYD/0	AXB/CYD/0	AXB/CY-/0	-XB/—/1	—/CY-/0	4
19	YGR085C	YPR102C	0.1445	AXB/CYD/0	AX-,-XB/CYD/0	AXB/CYD/0	AXB/CYD/0	—/—/1	—/—/1	4
20	YHL033C	YLL045C	0.1457	AXB/CYD/0	AXB/CY-/0	AXB/CYD/0	AXB/CYD/0	—/—/1	—/—/0	4

* 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

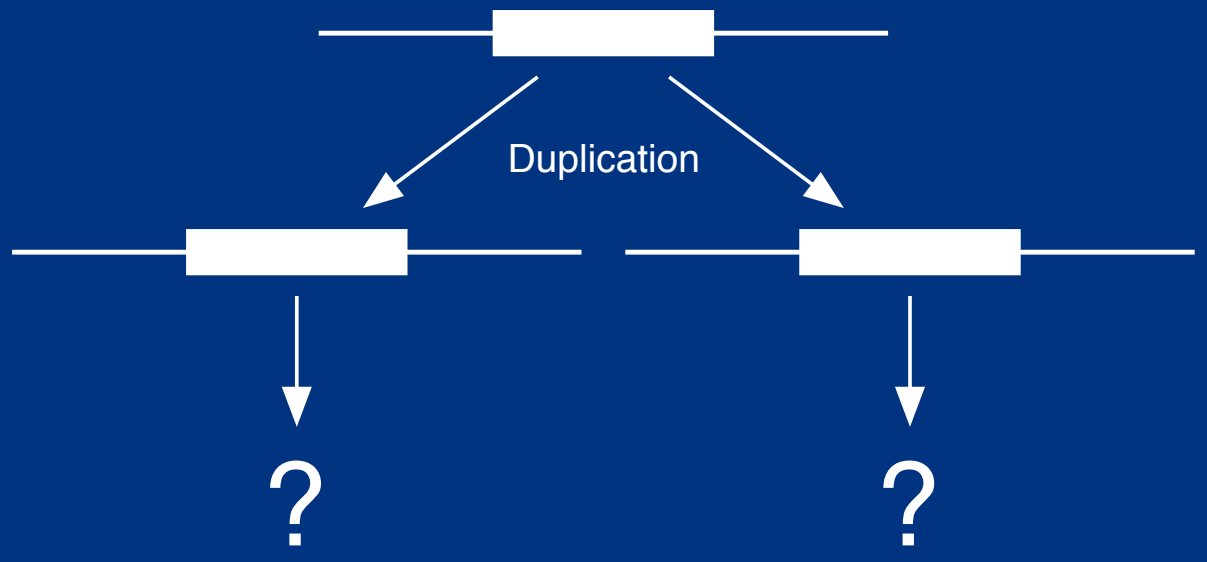
   Analysis with a single genome do not work well...

2. Genome-wide demonstration of concerted evolution in duplicated genes

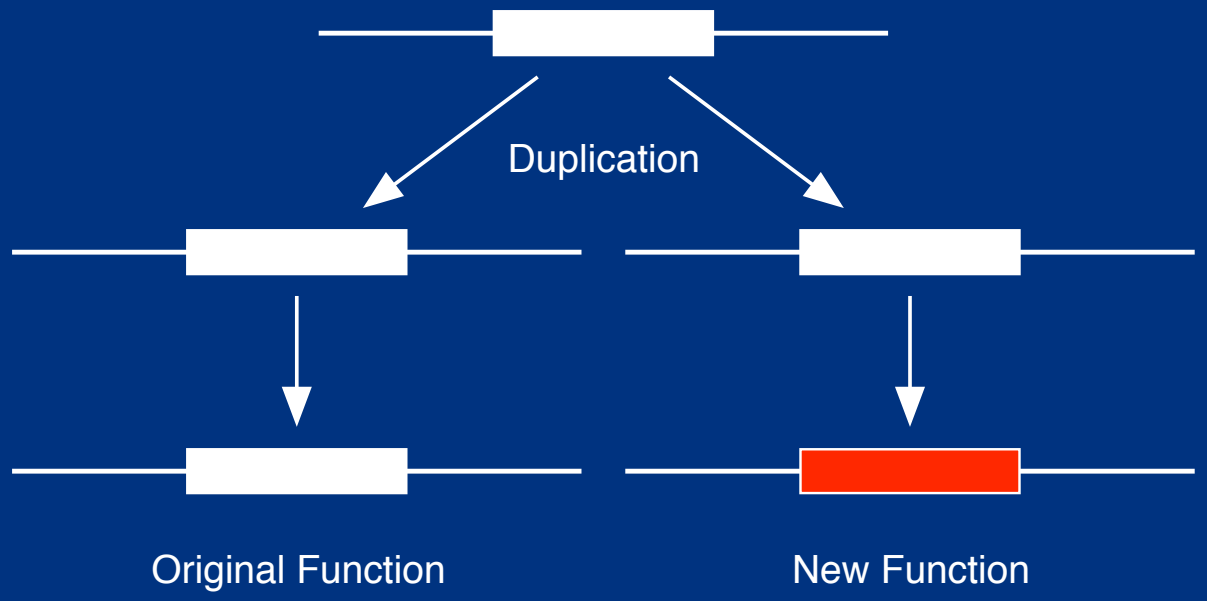
   Molecular clock does not hold for duplicated genes.

Selection and Evolutionary Fate of Duplicated Genes under the Pressure of Gene Conversion

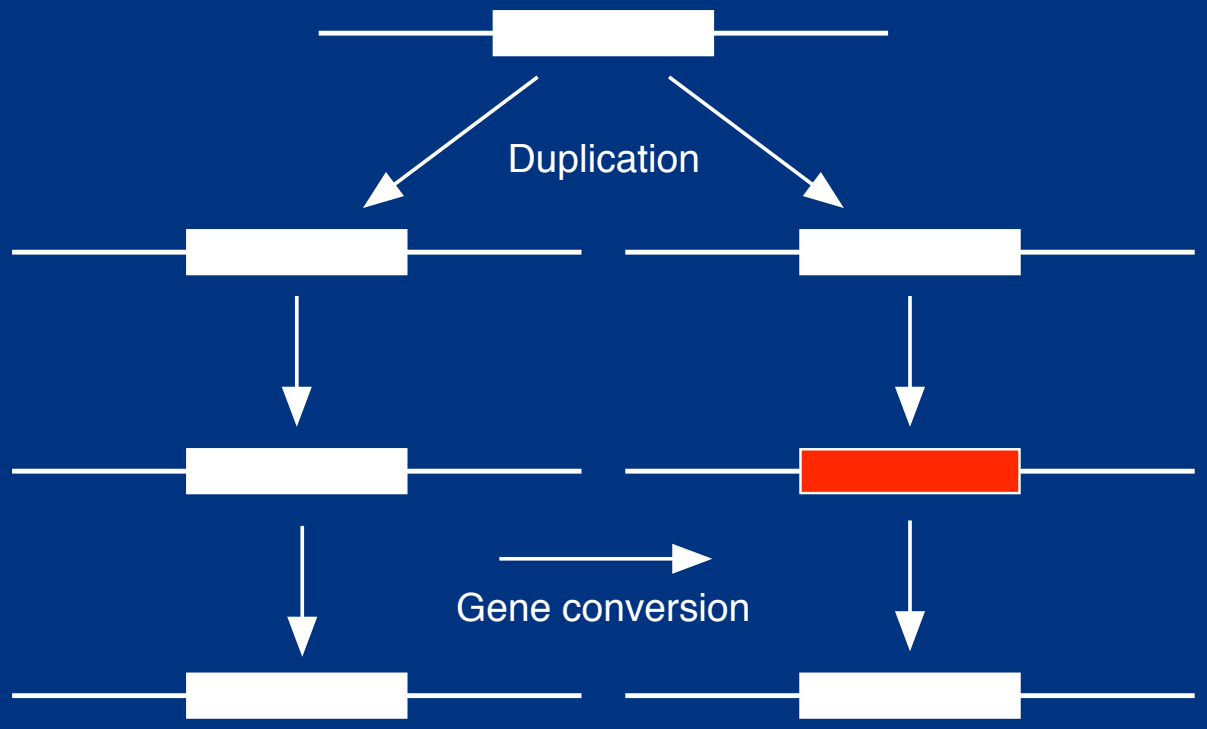
Evolutionary Fate of Duplicated Genes



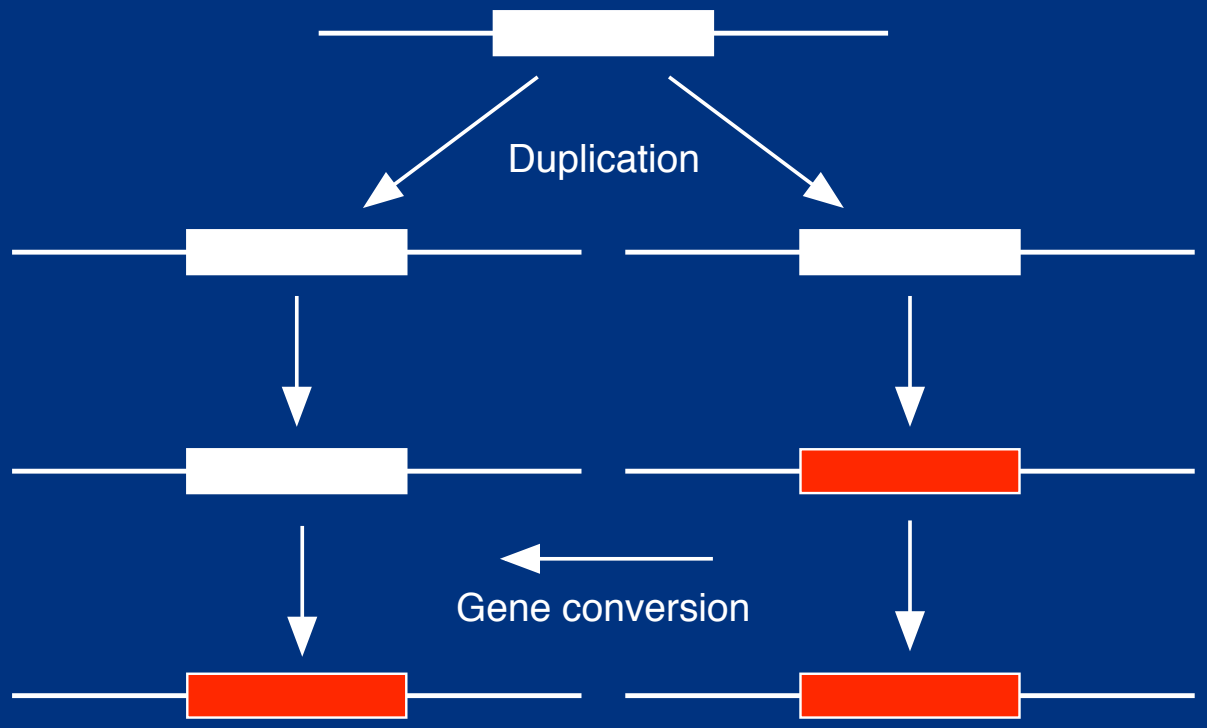
Evolutionary Fate of Duplicated Genes



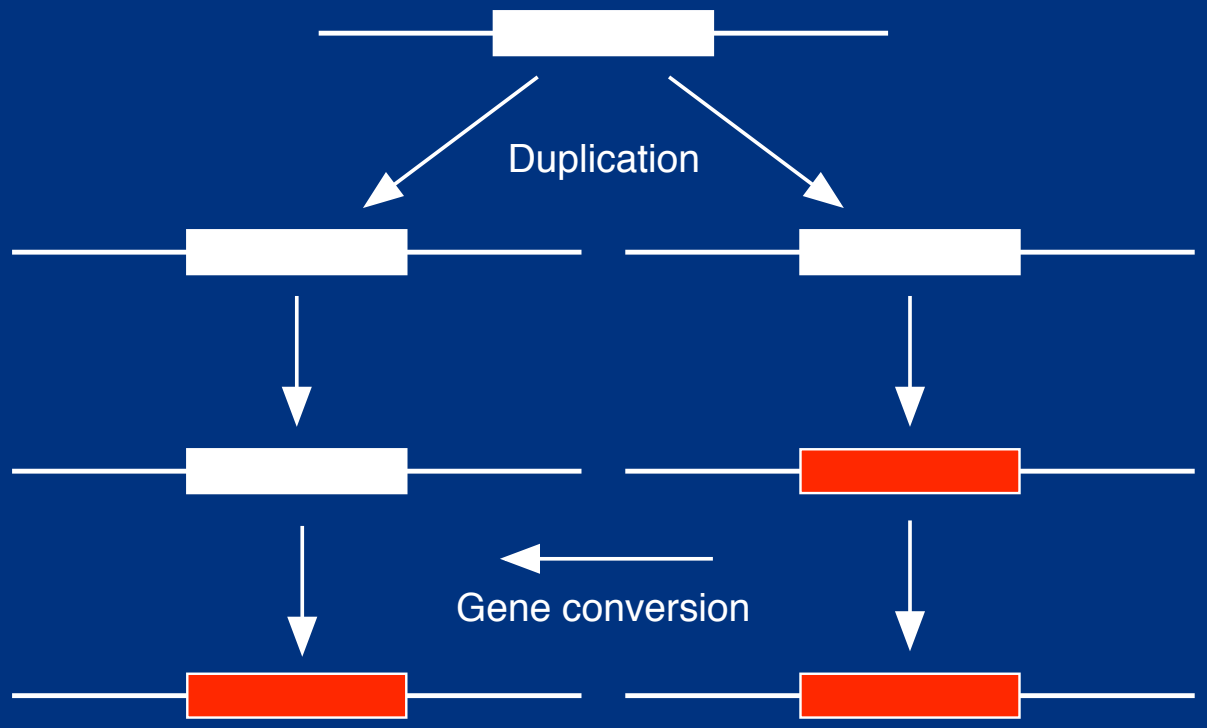
Evolutionary Fate of Duplicated Genes



Evolutionary Fate of Duplicated Genes

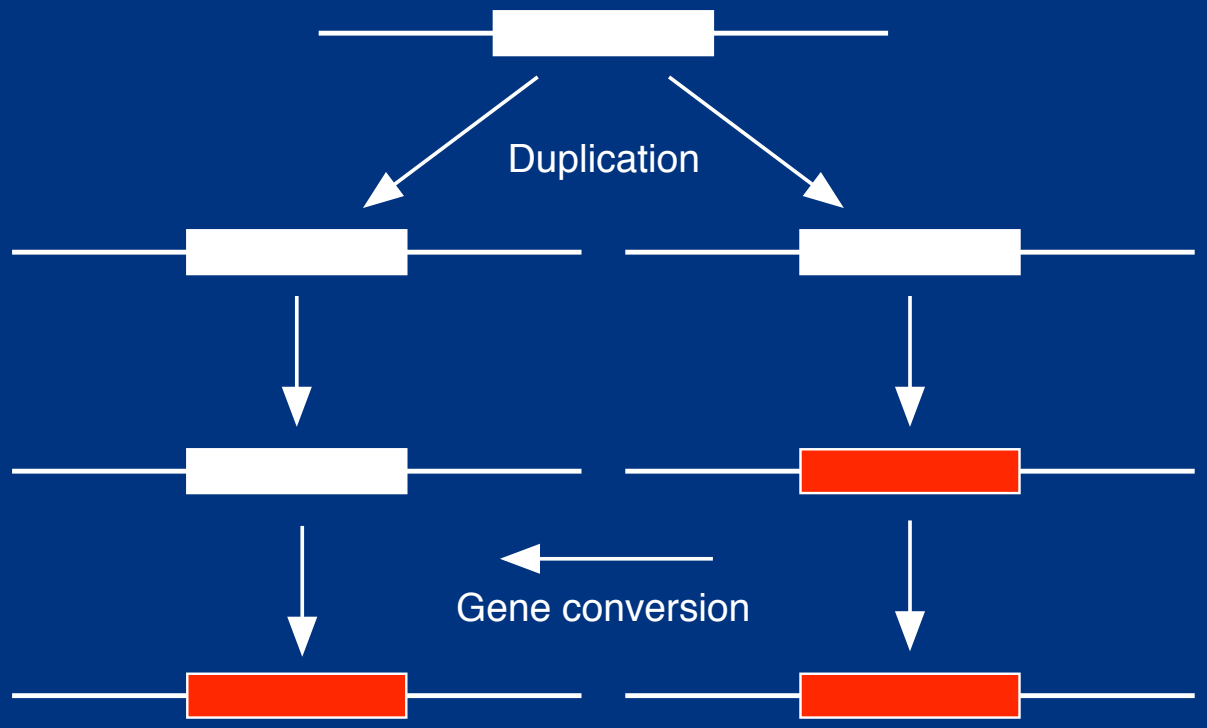


Evolutionary Fate of Duplicated Genes



Under what condition does neofunctionalization occur?

Evolutionary Fate of Duplicated Genes



Under what condition does neofunctionalization occur?

$\Rightarrow$ Look at nucleotide polyorphism!

Single Nucleotide Polymorphisms (SNPs)

in Duplicated Genes

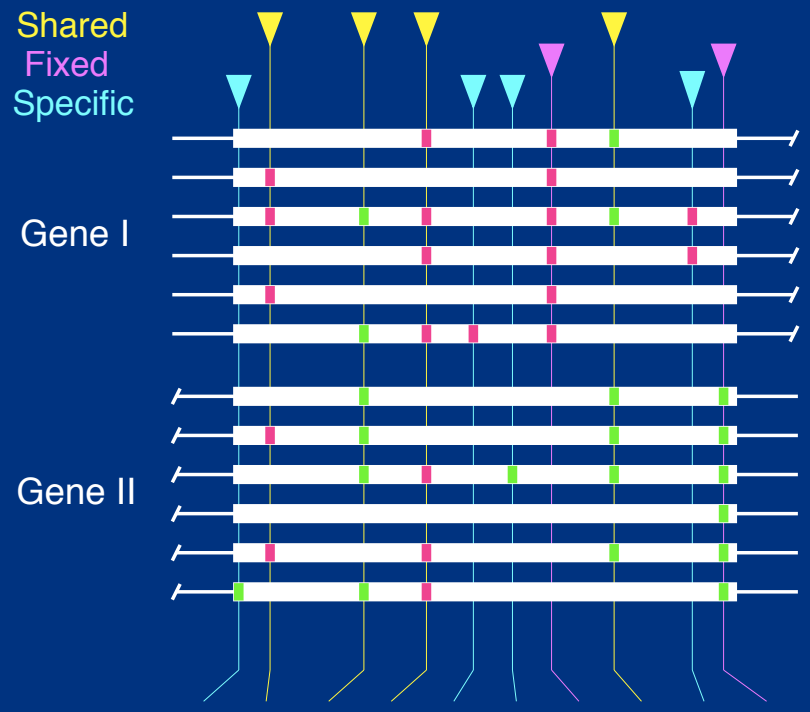
Typical Pattern of Polymorphism in Duplicated Genes

Proximal & Distal Amy in D. melanogaster

```
                                                  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                  1 1 2 2 2 3 3 4 4 5 5 5 6 7 7 7 8 8 8 9 8 9 9 9 9 0 0 0 1 0 1 1 2 2 3 2 3 3 4 4 4 4 4
                4 5 4 5 4 1 3 7 6 6 6 1 2 8 6 6 0 5 5 7 3 4 7 2 7 5 4 9 9 0 1 1 4 9 9 8 0 3 5 8 9 7 1 0 3 2 4
                8 1 8 0 1 3 7 6 2 1 8 3 5 5 7 0 8 0 9 7 2 6 3 1 9 4 5 9 0 8 4 1 6 5 2 8 8 3 3 1 3 1 6 1 1 6 9

D. simulans     C C C G G C G G A G C G C C C C G G C C G C C C T G C C C G C C C C G G C G C C G C G C T T G

Proximal Amy    * * A A A * * A * * * * * T T * * * * * * * T * T * * * * * * T * A A A * * * T * * * A *
                * * A A A * * A * * * * * T * * * * * * * * T * T * * * * * * T * A A A * * * T * * * A T
                * * A A A * * A * * * * T T * * * * * * * * T * T * * * * * * T * A A A * * * T * * * A *
                * * A A A * * A * * * * T T * * * * * * * * * * * * * * * C * T * * * * * * * T * * * * A *
                * * A A A * * A * * * * T T * * * * * * * * * * * * * * * C * T * * * * * * * T * * * * A *
                * * A A A * * A G * G * * T T T * * A * * * * * * A T A * * * T * * A * * T * * A T * A *
                * * A A A * * A G * G * * T T T * * A * A T * T * * A T A * * * T * * A * * T A * A T * A *
                * * A A A * * A G * G * * T T T * * A * A T * * * * T A * * * T * * A * * T * * A T * A *
                * * A * * G A * * A G * * T T * * * A * A T * * * * A T A * * * T * * A * * T * * A T * A *

Distal Amy      * * A * * * A * * * * * * * T C * * A A T * * * * * * C * T * * * * * * * T * * * * * A *
                * * * * * * A * * * * * * T T * * * * * T * * * T * * * * A * T * A A A * T * * * * * * *
                * * A * * G A * G * G * * * T * * * A A T * * * * * T A C * T * * A * * * * * * * * * A *
                T * * * * * A * * * C * * * * C T A A * * A * C * A * * * * * * A A A * * * * T A * A A *
                * * A * * * A * * * * * * * * T * * * A A T * * * * * * C * T * * * * * * * T * * * * A *
                * * A * * G A * G * G * * * T * * * A T * * * * * T A C * T * * * * * * * * * * * * * A *
                * * A * * G A * G * G * * * T * * * A A T * * * * * T A C * T * * * * * * * * * * A * * A *
                * T A * * * A * * * * * * T * * * * * * * * T * T * * * * * * * * A A A T * * A * * * * A *
                * * A * * G A * * A G * * T * T * * * A A T * * * * * T A C * T * * * * * * * * * * * * A *
```

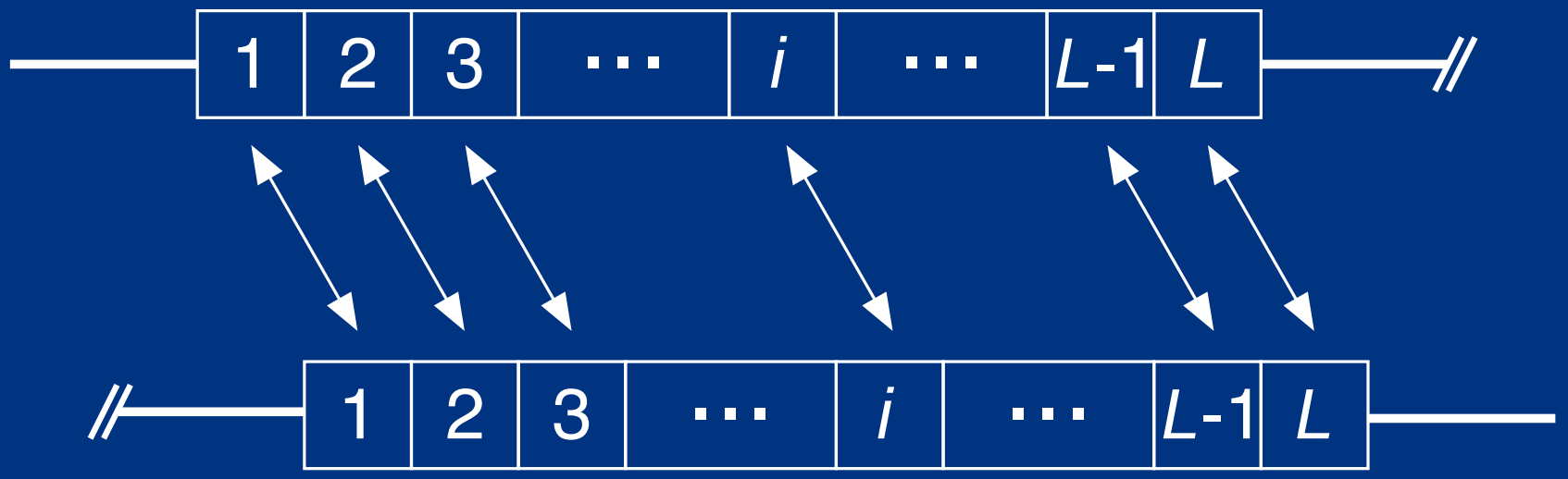
Data: Inomata et al. 1995 Genetics 141: 237-244

Effect of Gene Conversion on Polymorphism



h_{w1}	0	.6	.53	.53	.33	0	0	.53	.53	0	$K_{w1} = 3.05$
h_{w2}	.33	.53	.53	.6	0	.33	0	.53	0	0	$K_{w2} = 2.85$
h_b	.2	.57	.6	.5	0	0	1	.6	.33	1	$K_b = 4.80$
D	0	.2	.13	0	0	0	0	.13	0	0	$D_{sum} = 0.43$

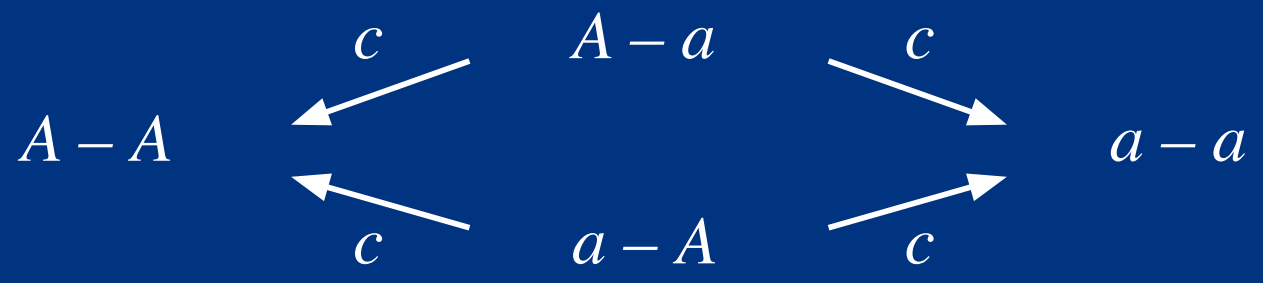
Modeling Duplicated Genes



$L \times$ two-locus (site) models

Two-Locus Model

1. Two duplicated loci (sites), I and II, with two neutral alleles (A and a) at mutation-drift equilibrium
2. Four haplotypes, $A-A$, $A-a$, $a-A$ and $a-a$
3. Symmetrical mutation rate, μ, 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: frequency of $A - A$
x_2: frequency of $A - a$
x_3: frequency of $a - A$
x_4: frequency of $a - a$

$$
\begin{aligned}
x_1' &= (1-2\mu)x_1 + (\mu + c)(x_2 + x_3) - rD \\
x_2' &= (1-2\mu - c)x_2 + \mu(x_1 + x_4) + rD \\
x_3' &= (1-2\mu - c)x_3 + \mu(x_1 + x_4) + rD \\
x_4' &= (1-2\mu)x_4 + (\mu + c)(x_2 + x_3) - rD
\end{aligned}
$$

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

$$
\begin{aligned}
L(g) = & \\
& \frac{x_1(1-x_1)}{4N}\frac{\partial g}{\partial^2 x_1^2} + \frac{x_2(1-x_2)}{4N}\frac{\partial g}{\partial^2 x_2^2} + \frac{x_3(1-x_3)}{4N}\frac{\partial g}{\partial^2 x_3^2} \\
& + \frac{x_1 x_2}{2N}\frac{\partial g}{\partial x_1 \partial x_2} + \frac{x_1 x_3}{2N}\frac{\partial g}{\partial x_1 \partial x_3} + \frac{x_2 x_3}{2N}\frac{\partial g}{\partial x_2 \partial x_3} \\
& + [-2\mu x_1 + (\mu + c)(x_2 + x_3) - rD]\frac{\partial g}{\partial x_1} \\
& + [-2(\mu + c)x_2 + \mu(1 - x_2 - x_3) + rD]\frac{\partial g}{\partial x_2} \\
& + [-2(\mu + c)x_3 + \mu(1 - x_2 - x_3) + rD]\frac{\partial g}{\partial x_3}
\end{aligned}
$$

Transformation of the three variables

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

$$\begin{aligned} p &= x_1 + x_2 \\ q &= x_1 + x_3 \\ D &= x_1 x_4 - x_2 x_3 \end{aligned}$$

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

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

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

Differential Operator

$$L'(g) =$$

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

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

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

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

$$+[Cp(1-p) - Cq(1-q) + (2 + 4\theta + 2C + R)D]\frac{\partial g}{\partial D}$$

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

$$\alpha = 2\theta + C,\ \beta = 2 + 2\alpha + R$$
$$\lambda = 4C^2 + 4\beta[2\theta C + 2\alpha(1+\theta)]$$
$$\omega = 8C^2 + 4\beta[\alpha(1+\alpha) - c^2]$$

Expectations of h_w, h_b and D

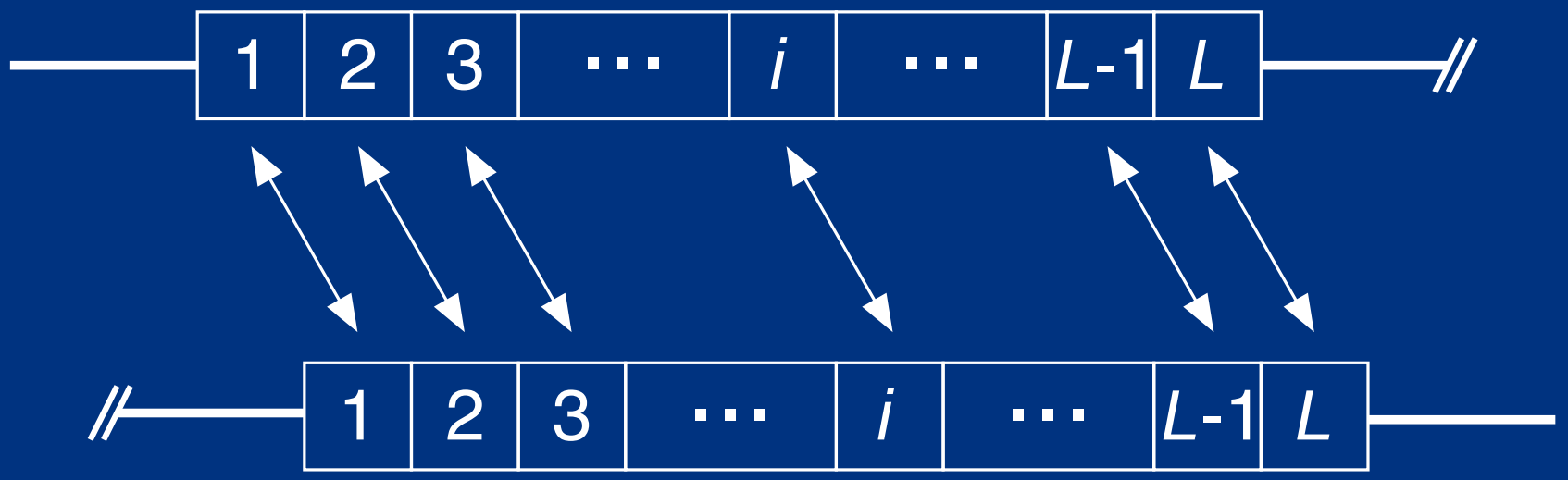
$$E(h_w) = 1 - 2\frac{\lambda}{\omega}$$

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

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

Innan 2002 Genetics 161: 865-872

Two-Locus Model → Finite-Site Model



$$E(K_w) = L \times E(h_w)$$

$$E(K_b) = L \times E(h_b)$$

$$E(D_{sum}) = L \times E(D)$$

Finite-Site Model → Infinite-Site Model

$$E(K_w) = \lim_{L \to \infty} L \times E(h_w) = \frac{2\Theta(2C + R + 2)}{4C + R + 2}$$

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

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

where

$\Theta = L\theta$: Mutation rate per gene

Innan 2003 Genetics 163: 803-810

Estimating Θ, C, and R

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

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

$$\hat{R} = \frac{K_w^2 + 4D_{sum}^2 - 4K_b D_{sum}}{2(K_b - K_w)D_{sum}}$$

Proximal & Distal Amy in D. melanogaster

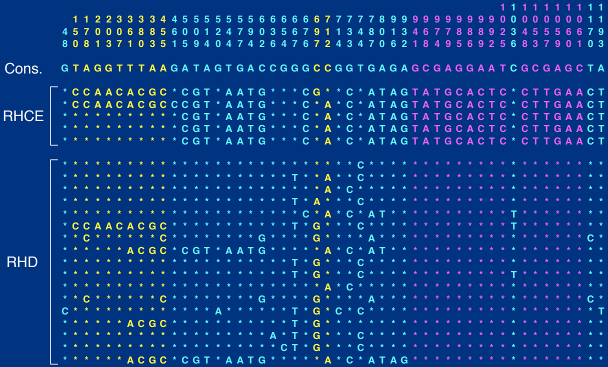
```
                                                  1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                1 1 2 2 2 3 3 4 4 5 5 5 6 7 7 7 8 8 8 9 8 9 9 9 9 0 0 0 1 0 1 1 2 2 3 2 3 3 4 4 4 4 4
            4 5 4 5 4 1 3 7 6 6 6 1 2 8 6 6 0 5 5 7 3 4 7 2 7 5 4 9 9 0 1 1 4 9 9 8 0 3 5 8 9 7 1 0 3 2 4
            8 1 8 0 1 3 7 6 2 1 8 3 5 5 7 0 8 0 9 7 2 6 3 1 9 4 5 9 0 8 4 1 6 5 2 8 8 3 3 1 3 1 6 1 1 6 9

D. simulans C C C G G C G G A G C G C C C C G G C C G C C C T G C C C G C C C C G G C G C C G C G C T T G

            * * A A A * * A * * * * * T T * * * * * * * T * T * * * * * * T * A A A * * * T * * * A *
            * * A A A * * A * * * * * T * * * * * * * * T * T * * * * * * T * A A A * * * T * * * A T
            * * A A A * * A * * * T T * * * * * * * * T * T * * * * * * T * A A A * * * T * * * A *
            * * A A A * * A * * * T T * * * * * * * * * * * * * * C * T * * * * * * * T * * * * * A *
Proximal Amy * * A A A * * A * * * * T T * * * * * * * * * * * * * C * T * * * * * * * T * * * * * A *
            * * A A A * * A G * G * * T T T * * A * * * * * * A T A * * * T * * A * * T * * A T * A *
            * * A A A * * A G * G * * T T T * * A * A T * T * * A T A * * * T * * A * * T A * A T * A *
            * * A A A * * A G * G * * T T T * * A * A T * * * * T A * * * T * * A * * T * * A T * A *
            * * A * * G A * * A G * * T T * * * A * A T * * * * A T A * * * T * * A * * T * * A T * A *

            * * A * * * A * * * * * * * T C * * A A T * * * * * * C * T * * * * * * * T * * * * * A *
            * * * * * * A * * * * * * T T * * * * * T * * * T * * * * A * T * A A A * T * * * * * * *
            * * A * * G A * G * G * * * * T * * * A A T * * * * * T A C * T * * A * * * * * * * * * A *
            T * * * * * * A * * * C * * * * C T A A * * A * C * A * * * * * * * A A A * * * T A * A A *
Distal Amy  * * A * * * A * * * * * * * * T * * * A A T * * * * * * * C * T * * * * * * * T * * * * * A *
            * * A * * G A * G * G * * * * T * * * * A T * * * * * T A C * T * * * * * * * * * * * * A *
            * * A * * G A * G * G * * * * T * * * A A T * * * * * T A C * T * * * * * * * * * * A * * A *
            * T A * * * A * * * * * * T * * * * * * * * T * T * * * * * * * * A A A T * * A * * * * A *
            * * A * * G A * * A G * * T * T * * * A A T * * * * * T A C * T * * * * * * * * * * * * A *
```

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

RHCE & RHD in Human: Anything Strange?

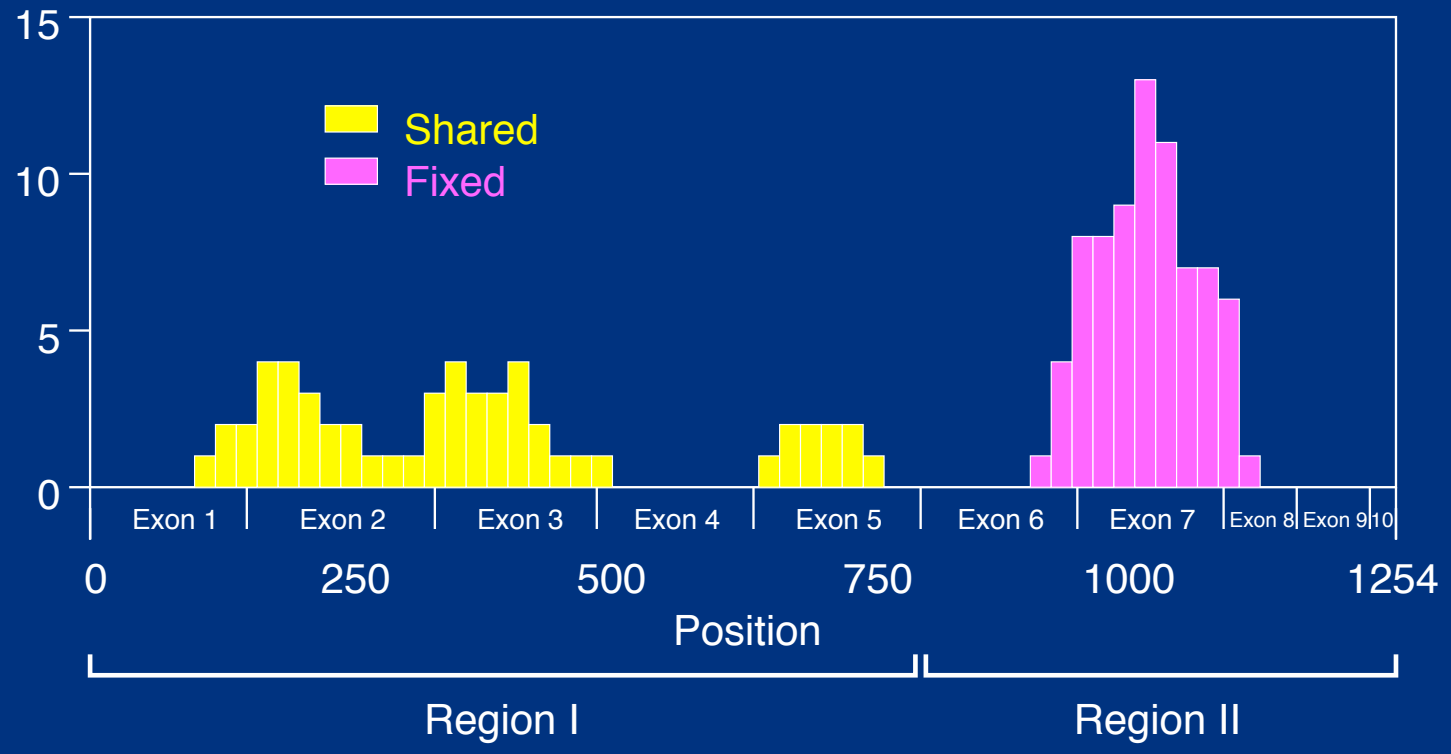


```
                                                                                        1 1 1 1 1 1 1 1 1 1
          1 1 2 2 3 3 3 3 4 4 5 5 5 5 5 5 5 6 6 6 6 6 6 7 7 7 7 7 8 9 9 9 9 9 9 9 9 9 9 0 0 0 0 0 0 0 0 1 1
        4 5 7 0 0 0 6 8 8 5 6 0 0 1 2 4 7 9 0 3 5 6 7 9 1 1 3 4 8 0 1 3 4 6 7 7 8 8 8 9 2 3 4 5 5 5 6 6 7 9
        8 0 8 1 3 7 1 0 3 5 1 5 9 4 0 4 7 4 2 6 4 7 6 7 2 4 3 4 7 0 6 2 1 8 4 9 5 6 9 2 5 6 8 3 7 9 0 1 0 3

Cons.   G T A G G T T T A A G A T A G T G A C C G G G C C G G T G A G A G C G A G G A A T C G C G A G C T A

        * C C A A C A C G C * C G T * A A T G * * * C G * * C * A T A G T A T G C A C T C * C T T G A A C T
        * C C A A C A C G C C C G T * A A T G * * * C * A * C * A T A G T A T G C A C T C * C T T G A A C T
RHCE    * * * * * * * * * * * C G T * A A T G * * * C * A * C * A T A G T A T G C A C T C * C T T G A A C T
        * * * * * * * * * * * C G T * A A T G * * * C * A * C * A T A G T A T G C A C T C * C T T G A A C T
        * * * * * * * * * * * C G T * A A T G * * * C * A * C * A T A G T A T G C A C T C * C T T G A A C T

        * * * * * * * * * * * * * * * * * * * * * * * * * * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * T * * A * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * * * * A * C * * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * T * A * * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * * C * A * C * A T * * * * * * * * * * * T * * * * * * * *
        * C C A A C A C G C * * * * * * * * * * * T * G * * * C * * * * * * * * * * * * * T * * * * * * * *
        * * C * * * * * * C * * * * * * * * G * * * * G * * * * A * * * * * * * * * * * * * * * * * * * C *
        * * * * * * A C G C * C G T * A A T G * * * * * A * C * A T * * * * * * * * * * * * * * * * * * * *
RHD     * * * * * * * * * * * * * * * * * * * * * T * G * * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * T * G * * * C * * * * * * * * * * * * * T * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * * * * * A * C * * * * * * * * * * * * * * * * * * * * * * *
        * * C * * * * * * C * * * * * * * * G * * * * G * * * * A * * * * * * * * * * * * * * * * * * * C *
        C * * * * * * * * * * * * * A * * * * * * T * G * C * C * * * * * * * * * * * * * * * * * * * * * T
        * * * * * * A C G C * * * * * * * * * * * T * G * * * * * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * A * T * G * * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * * * * * * * * * * * * * * * C T * G * * * C * * * * * * * * * * * * * * * * * * * * * *
        * * * * * * A C G C * C G T * A A T G * * * * * A * C * A T A G * * * * * * * * * * * * * * * * * *
```

Innan 2003 PNAS 100: 8793-8798

RHCE & RHD in Human: Anything Strange?

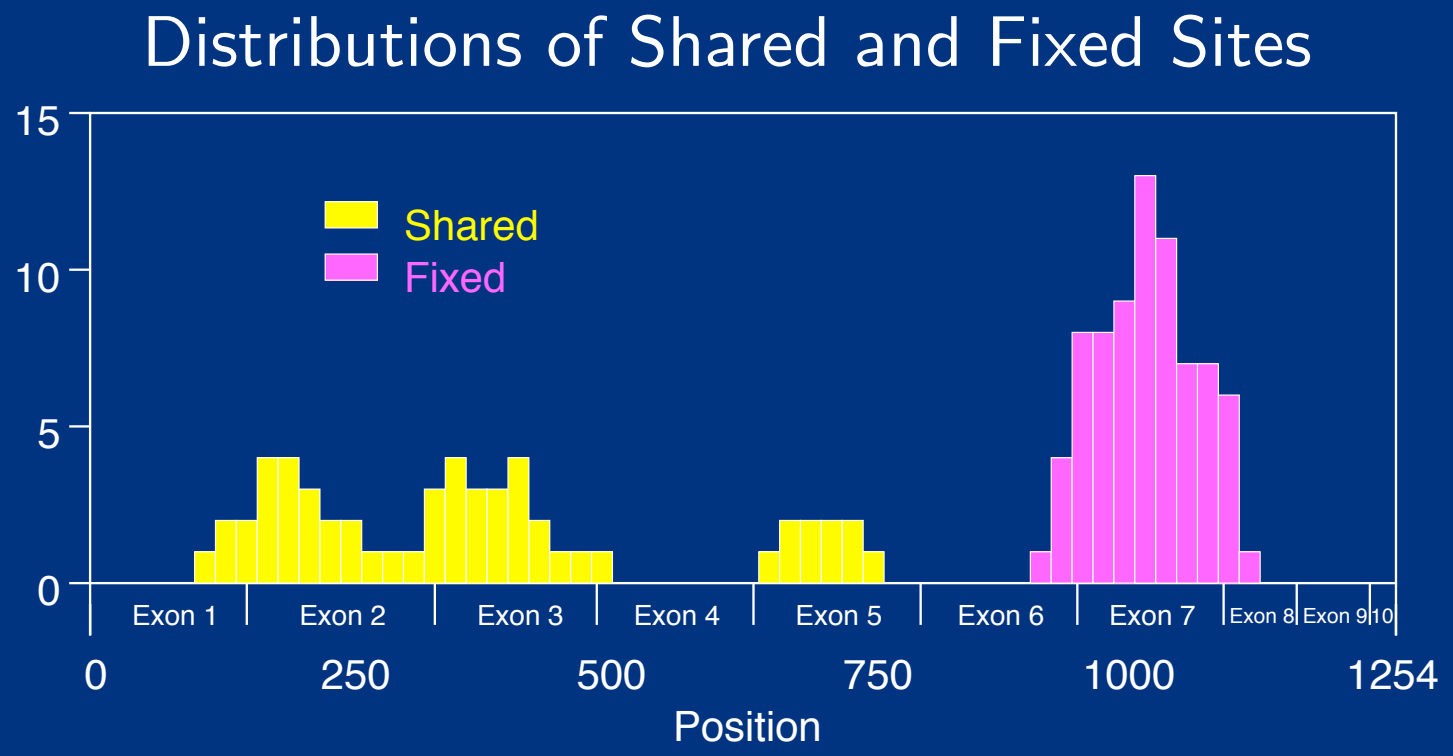
Distributions of Shared and Fixed Sites



	Shared	Fixed
Region I (801 bp)	11	0
Region II (453 bp)	0	15

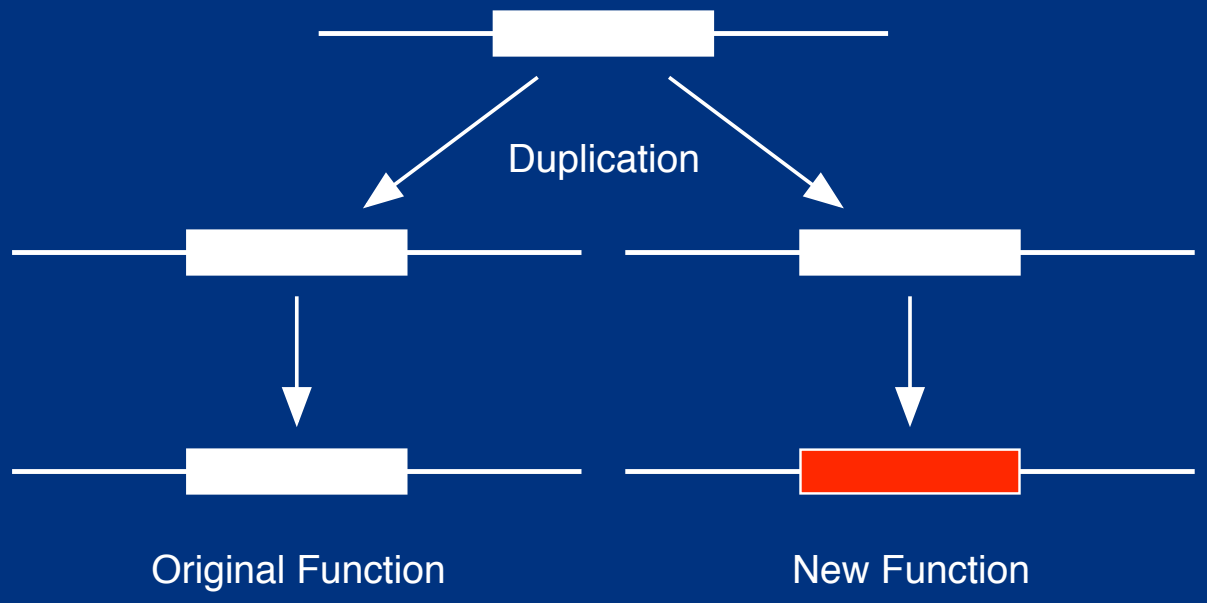
$P < 10^{-6}$

RHCE & RHD in Human: Anything Strange?

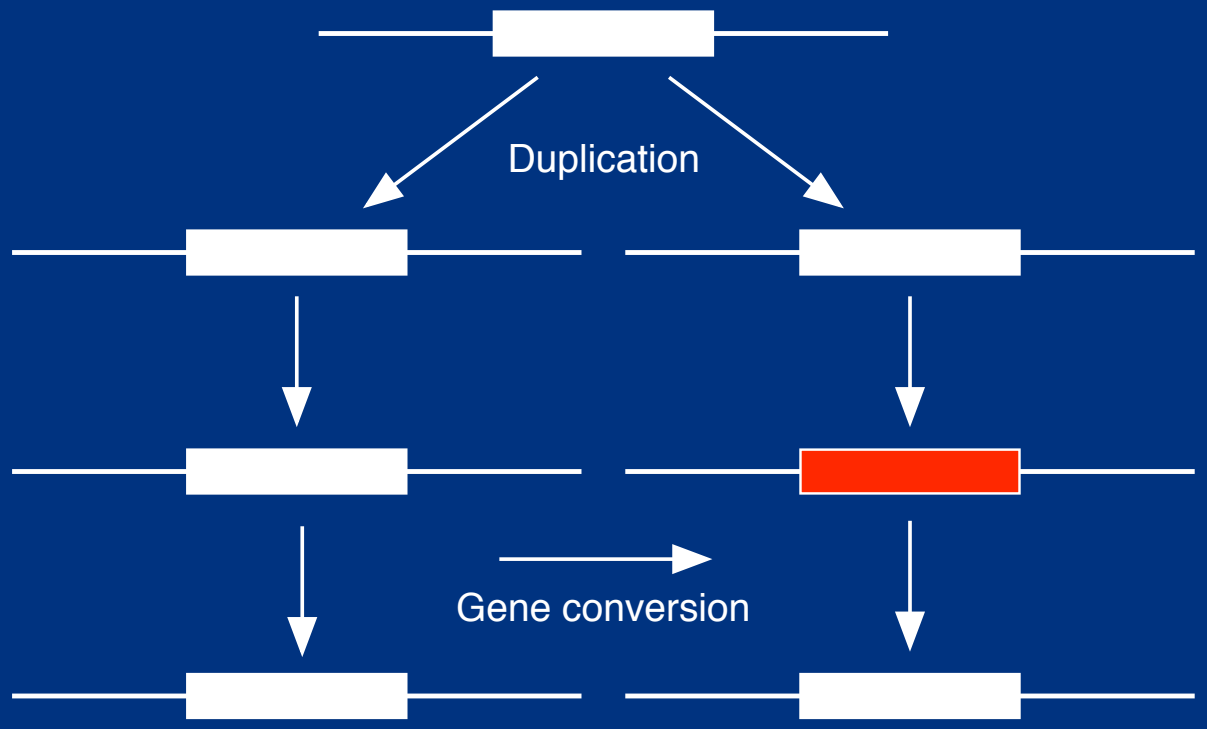


Does Selection Against Gene Conversion in Exon 7 Explain the Data?

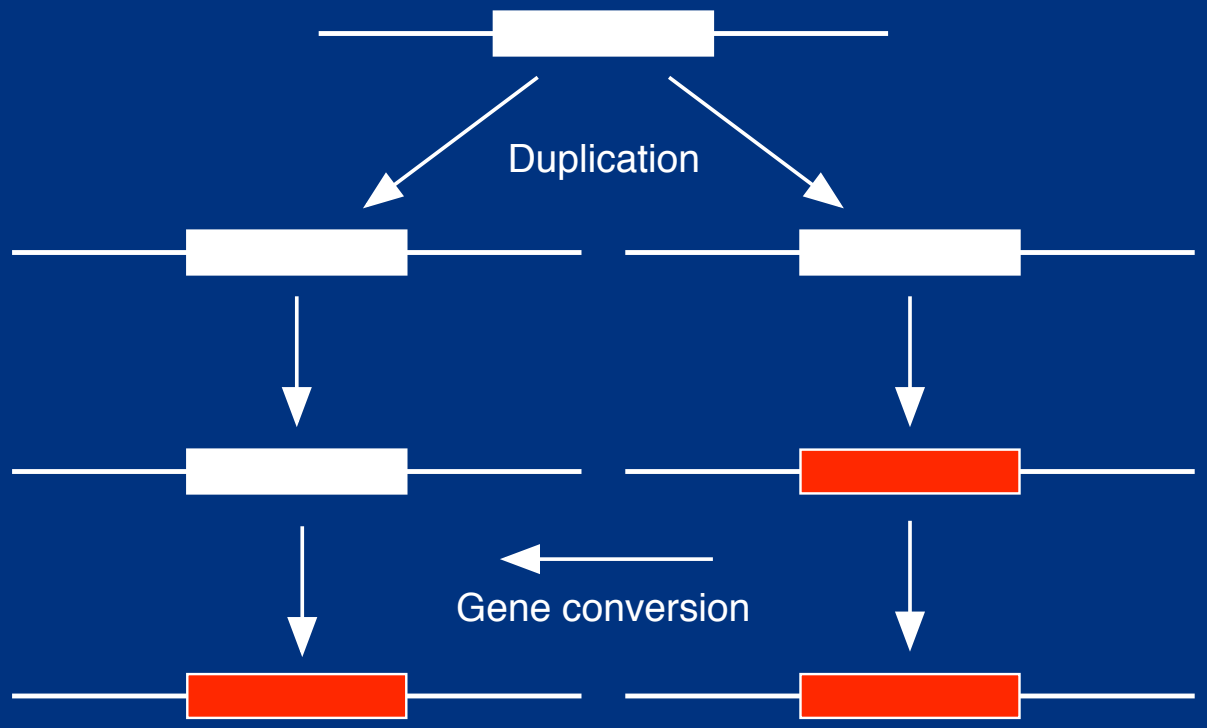
Selection Against Gene Conversion: Selection for Neofunctionalization



Selection Against Gene Conversion: Selection for Neofunctionalization

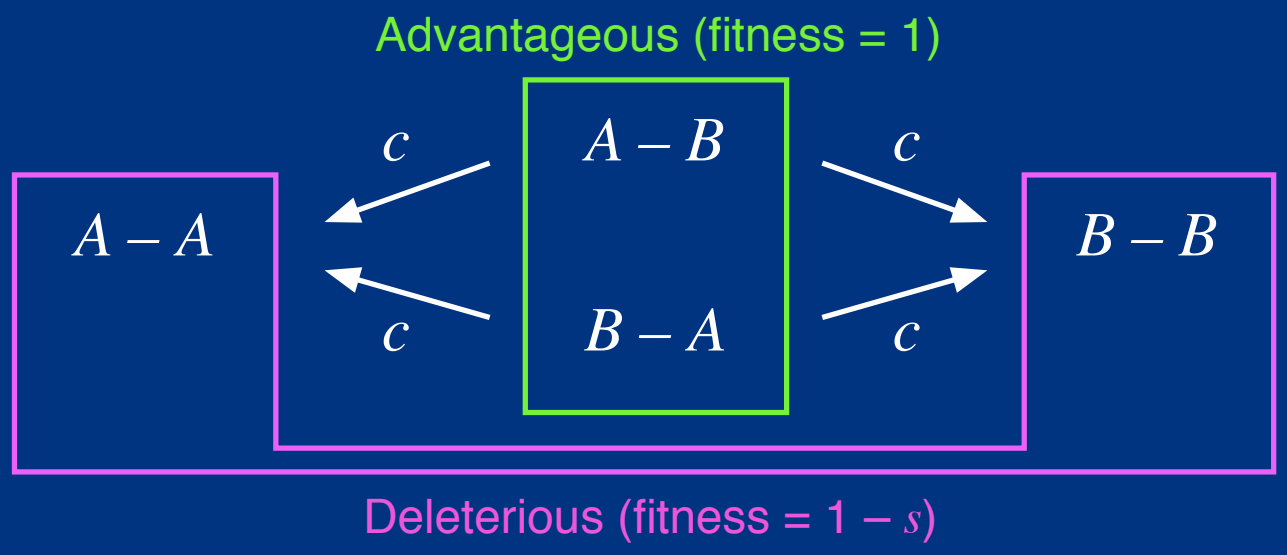


Selection Against Gene Conversion: Selection for Neofunctionalization



Modeling the Battle
Selection vs. 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: frequency of $A - A$
x_2: frequency of $A - B$
x_3: frequency of $B - A$
x_4: frequency of $B - B$

$$
\begin{aligned}
x_1' &= (1-2\mu)x_1 + (\mu + c)(x_2 + x_3) - rD - sx_1(x_2 + x_3) \\
x_2' &= (1-2\mu - c)x_2 + \mu(x_1 + x_4) + rD + sx_2(x_1 + x_4) \\
x_3' &= (1-2\mu - c)x_3 + \mu(x_1 + x_4) + rD + sx_3(x_1 + x_4) \\
x_4' &= (1-2\mu)x_4 + (\mu + c)(x_2 + x_3) - rD - sx_4(x_2 + x_3)
\end{aligned}
$$

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

$$L'(g) =$$

$$\frac{x_1(1-x_1)}{4N}\frac{\partial^2 g}{\partial x_1^2} + \frac{x_2(1-x_2)}{4N}\frac{\partial^2 g}{\partial x_2^2} + \frac{x_3(1-x_3)}{4N}\frac{\partial^2 g}{\partial x_3^2}$$

$$+\frac{x_1 x_2}{2N}\frac{\partial g}{\partial x_1 \partial x_2} + \frac{x_1 x_3}{2N}\frac{\partial g}{\partial x_1 \partial x_3} + \frac{x_2 x_3}{2N}\frac{\partial g}{\partial x_2 \partial x_3}$$

$$+[-2\mu x_1 + (\mu + c)(x_2 + x_3) - rD - s x_1(x_2 + x_3)]\frac{\partial g}{\partial x_1}$$

$$+[-2(\mu + c)x_2 + \mu(x_1 + x_4) + rD + s x_2(x_1 + x_4)]\frac{\partial g}{\partial x_2}$$

$$+[-2(\mu + c)x_3 + \mu(x_1 + x_4) + rD + s x_3(x_1 + x_4)]\frac{\partial g}{\partial x_3}$$

Transformation of the three variables

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

$$\begin{aligned} p &= x_1 + x_2 \\ q &= x_1 + x_3 \\ D &= x_1 x_4 - x_2 x_3 \end{aligned}$$

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

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

Differential Operator

$$L'(g) =$$

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

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

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

$$+ \ldots \text{continue}$$

Differential Operator

$$+[\theta(1-2p) - C(p-q)$$

$$+4Nsp(1-p-2q+2pq) - 4NsD(1-2p)]\frac{\partial g}{\partial p}$$

$$+[\theta(1-2q) - C(p-q)$$

$$+4Nsq(1-2p-q+2pq) - 4NsD(1-2q)]\frac{\partial g}{\partial q}$$

$$+[Cp(1-p) - Cq(1-q) + (2+4\theta+2C+R)D$$

$$-8Nspq(1-p-q+pq) + 8NsD^2]\frac{\partial g}{\partial D}$$

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

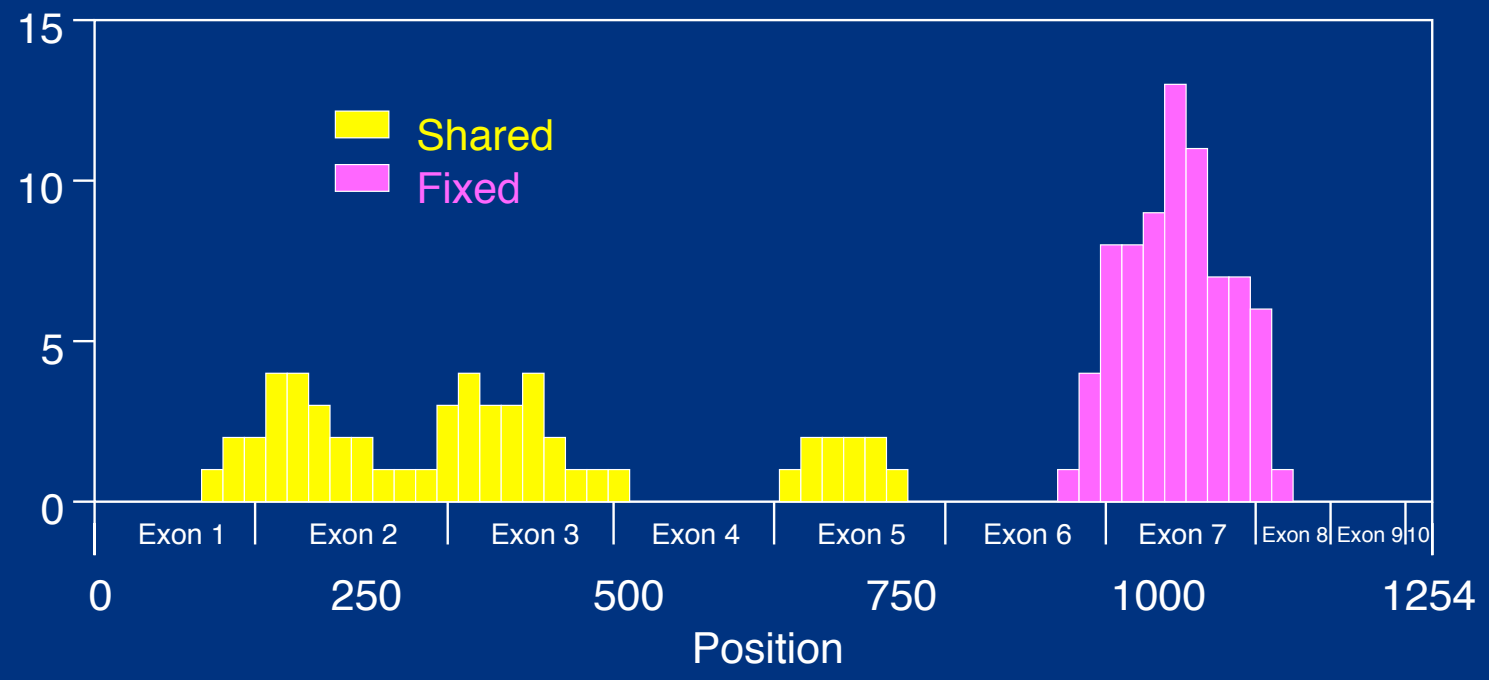
If selection is sufficiently strong, the target site of selection could be a fixed (or nearly fixed) site

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

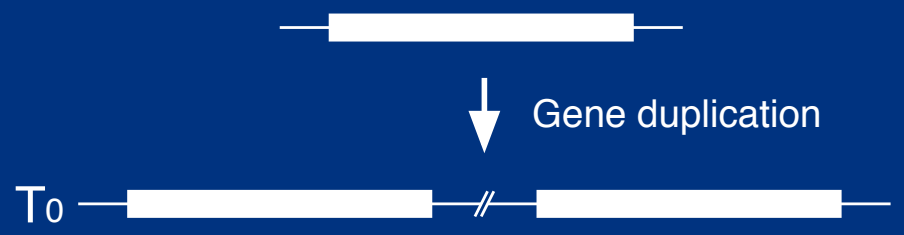
Selection Around Exon 7 in the RH Genes?

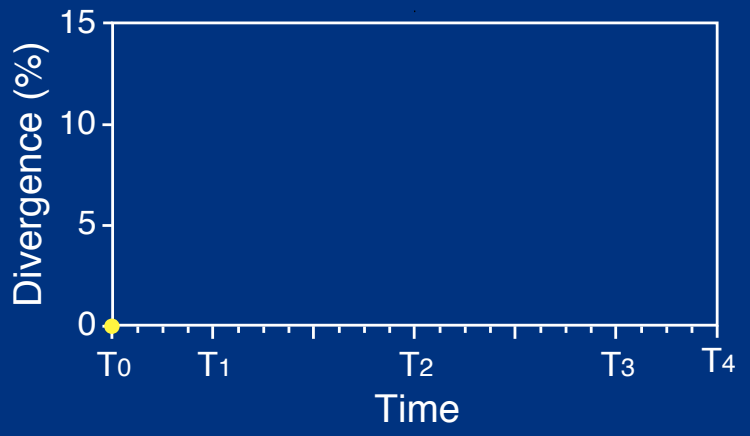


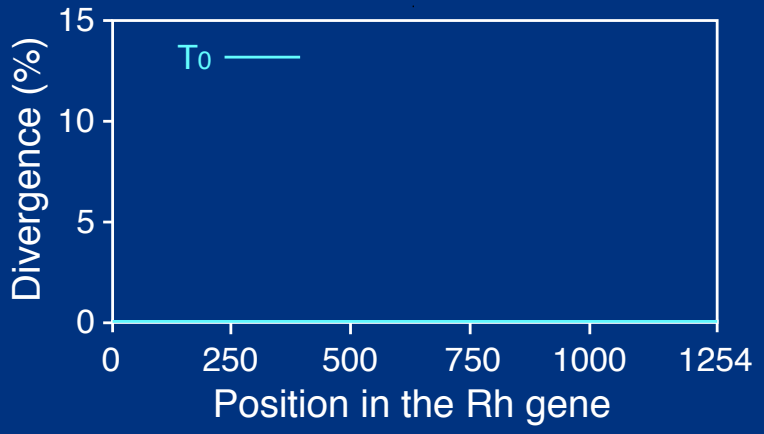
1. All 15 fixed sites are around exon 7, which encodes amino acids that characterize the difference between RHCE and RHD antigens.

2. Most fixed sites are non-synonymous (13 non-synonymous vs. 2 synonymous)

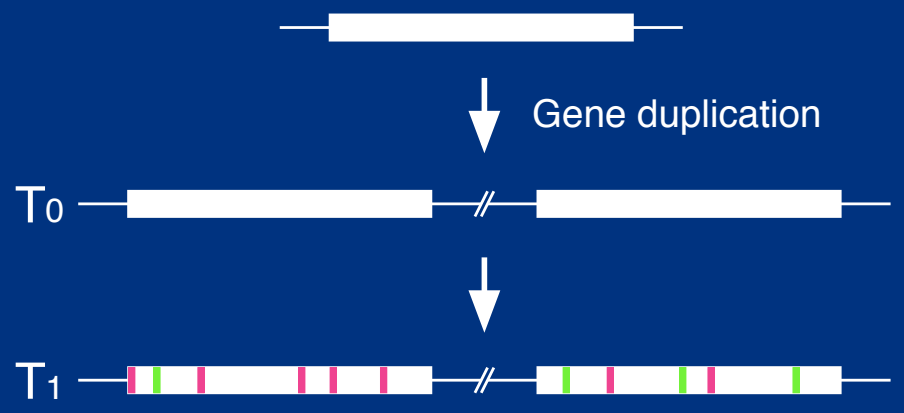
Hypothetical History of the Rh genes

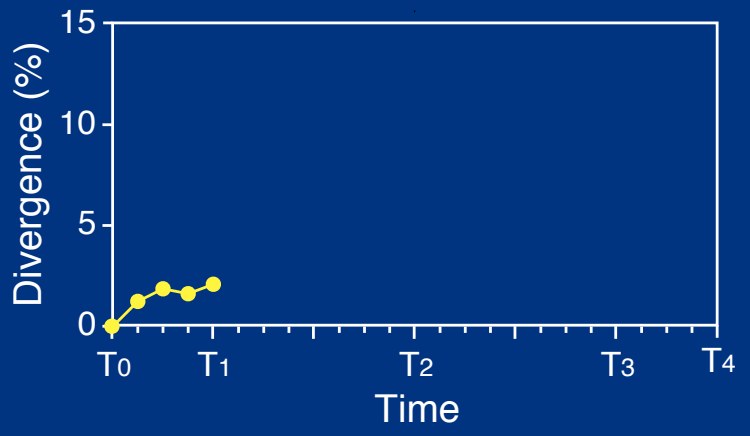


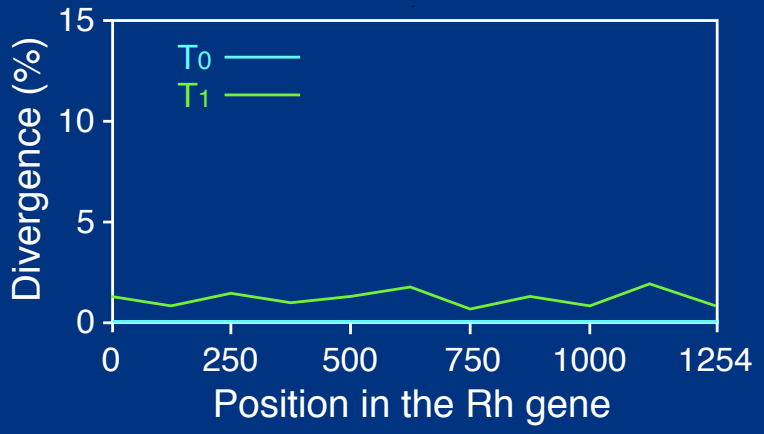




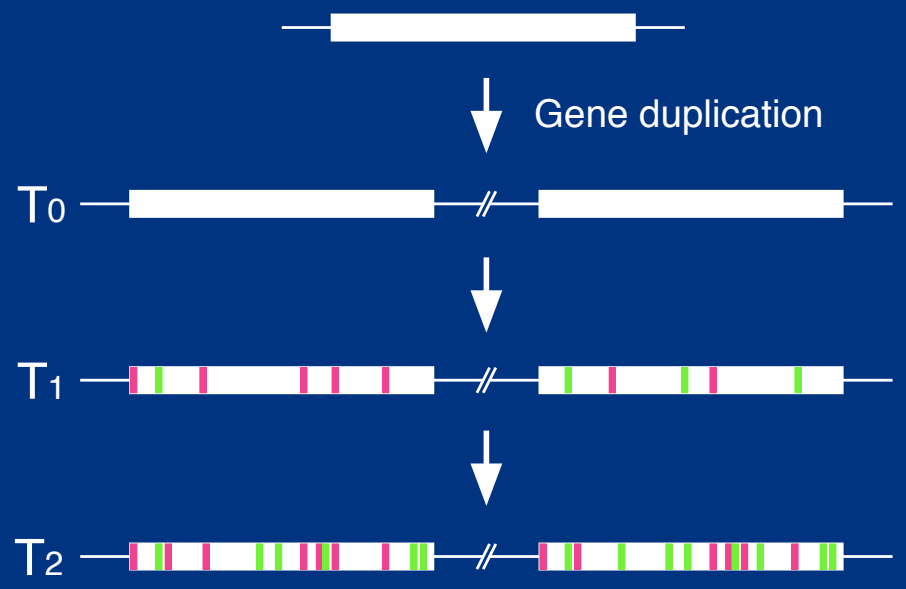
Hypothetical History of the Rh genes

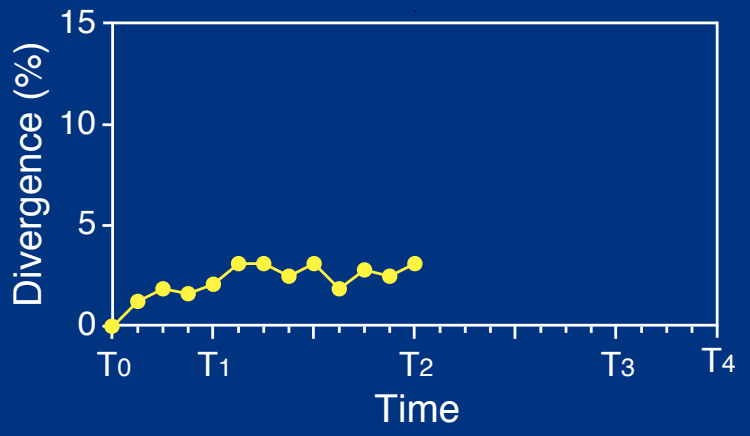


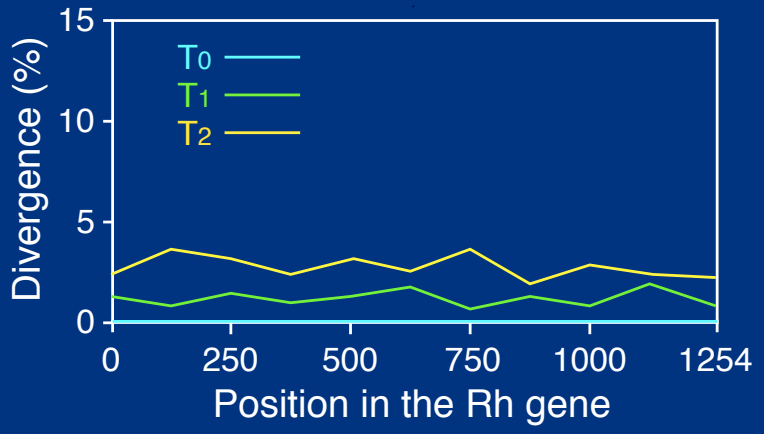




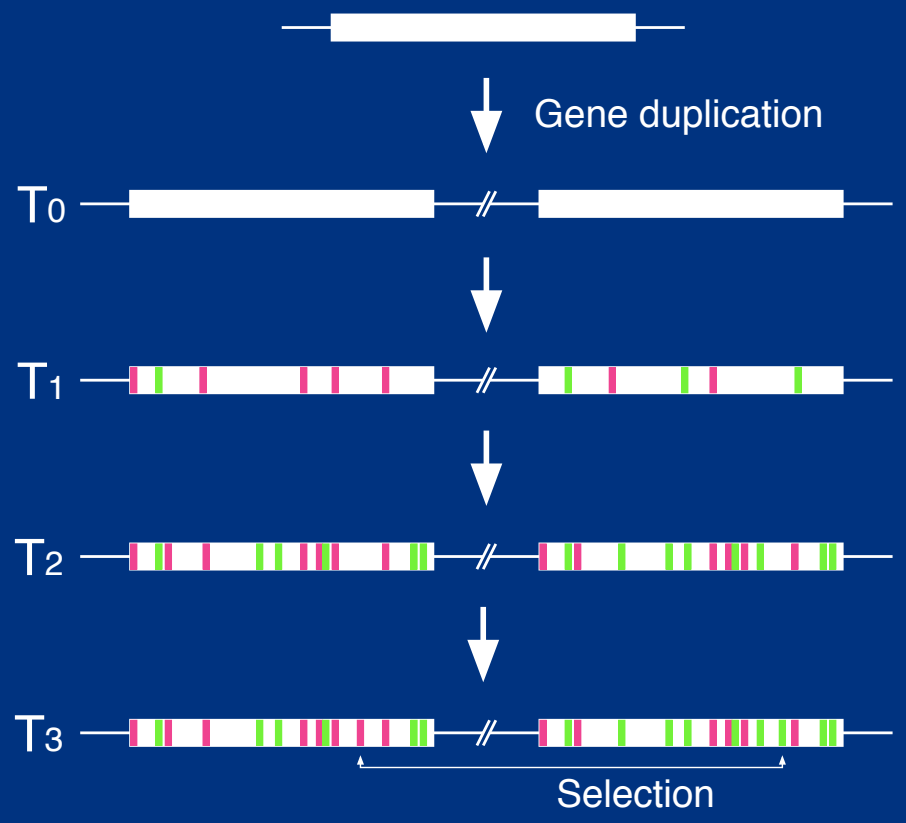
Hypothetical History of the Rh genes

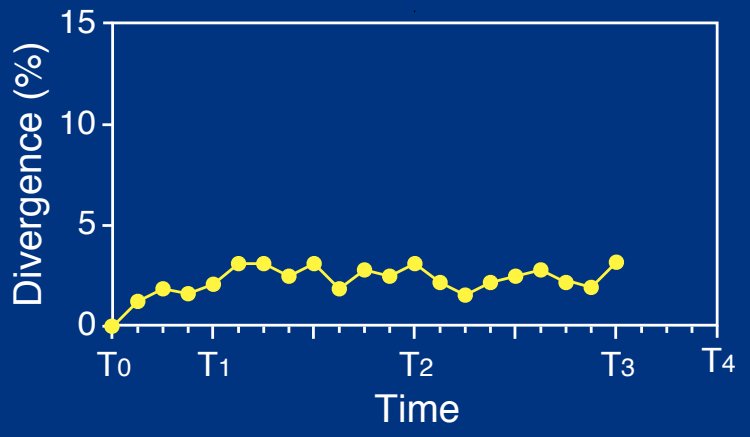


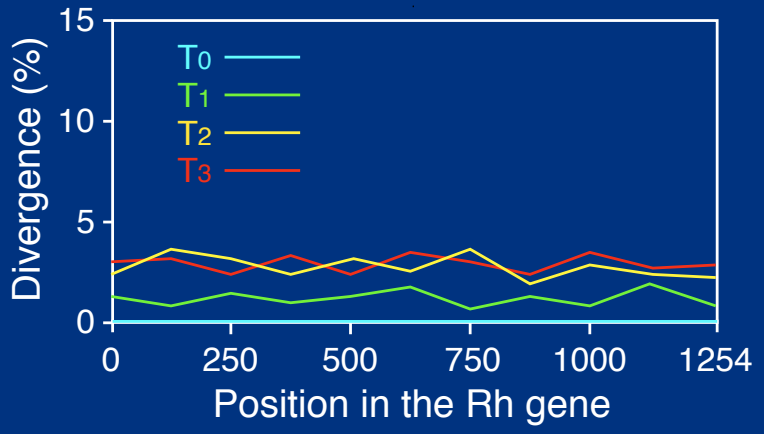




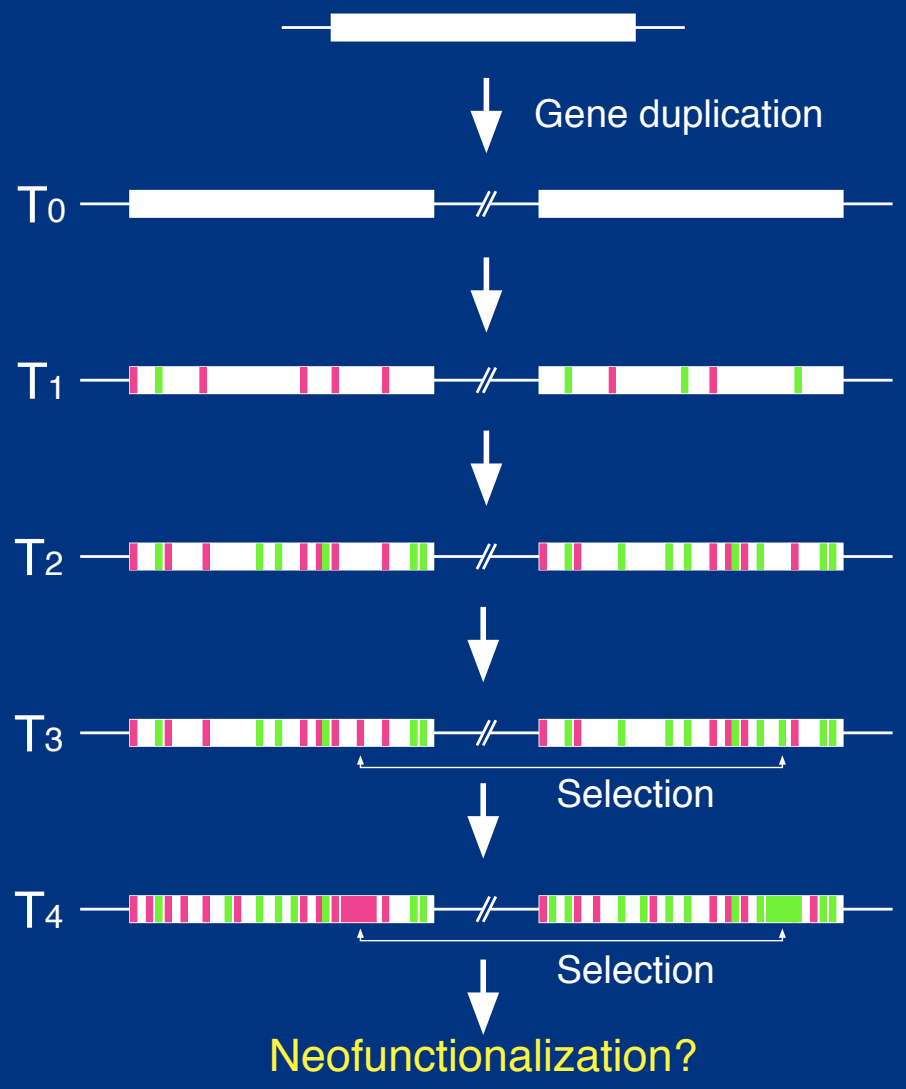
Hypothetical History of the Rh genes

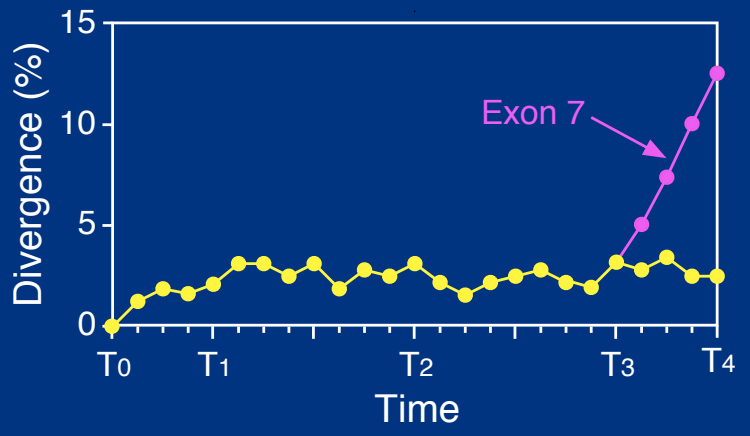


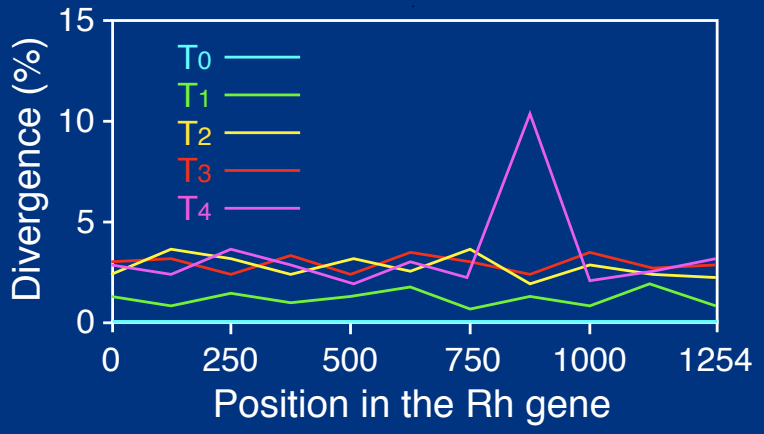




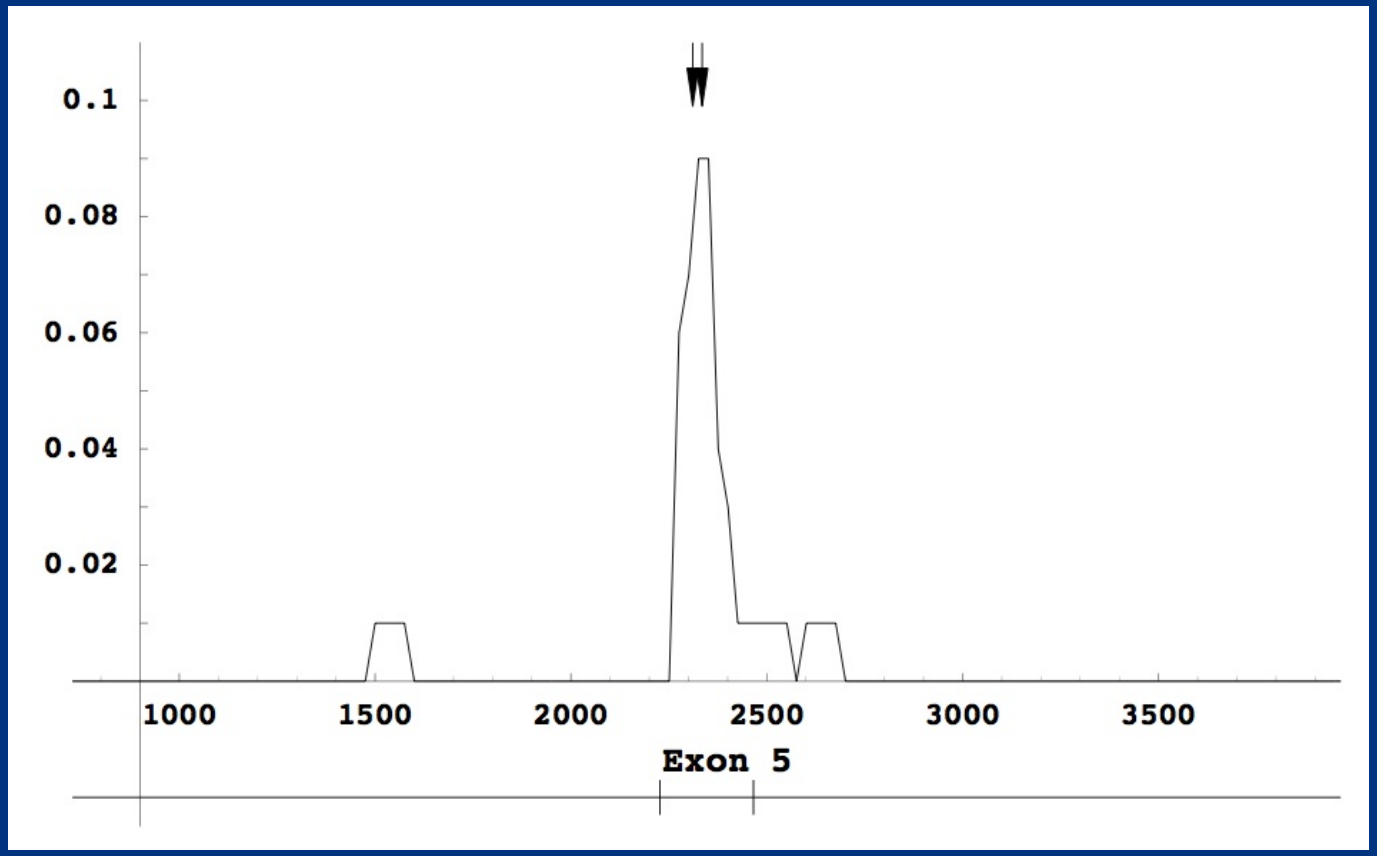
Hypothetical History of the Rh genes







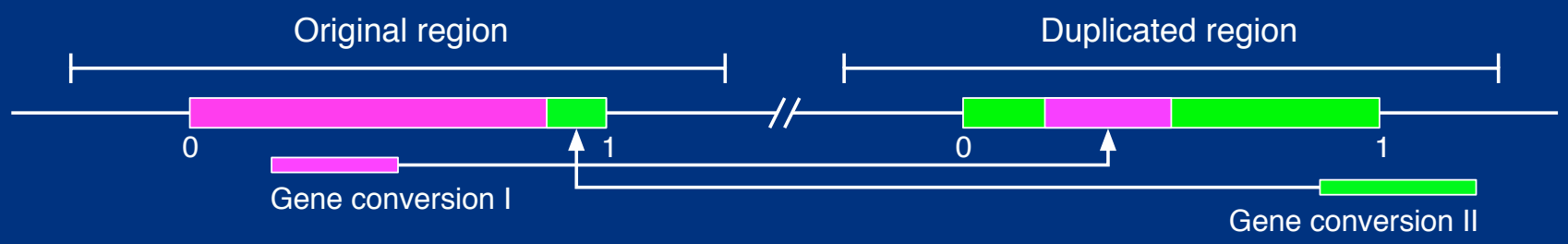
Similar Pattern in Human Opsin Genes



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

Data: Zhao, Hewett-Emmett, and Li 1997 J. Mol. Evol. 46: 494-496.

Simulating Divergence



1. Simulated region (L bp) assigned to a (0,1) interval
2. Mutation rate μ 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.

   imporatnt parameters: c/μ, d_t, Q

2. Drastic changes of DNA sequences

   Mutations such as transposons and large in/dels automatically terminates gene conversion.

   imporatnt parameters: μ_T

3. Selection for neofunctionalization

   Gene conversion is deleterious.

   imporatnt 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)