Gene Conversion and Codon Usage Bias in the Evolution of Duplicate Genes

Wen-Hsiung Li

Ecology and Evolution

University of Chicago

Why the evolution of many duplicate genes in yeasts has been decelerated?

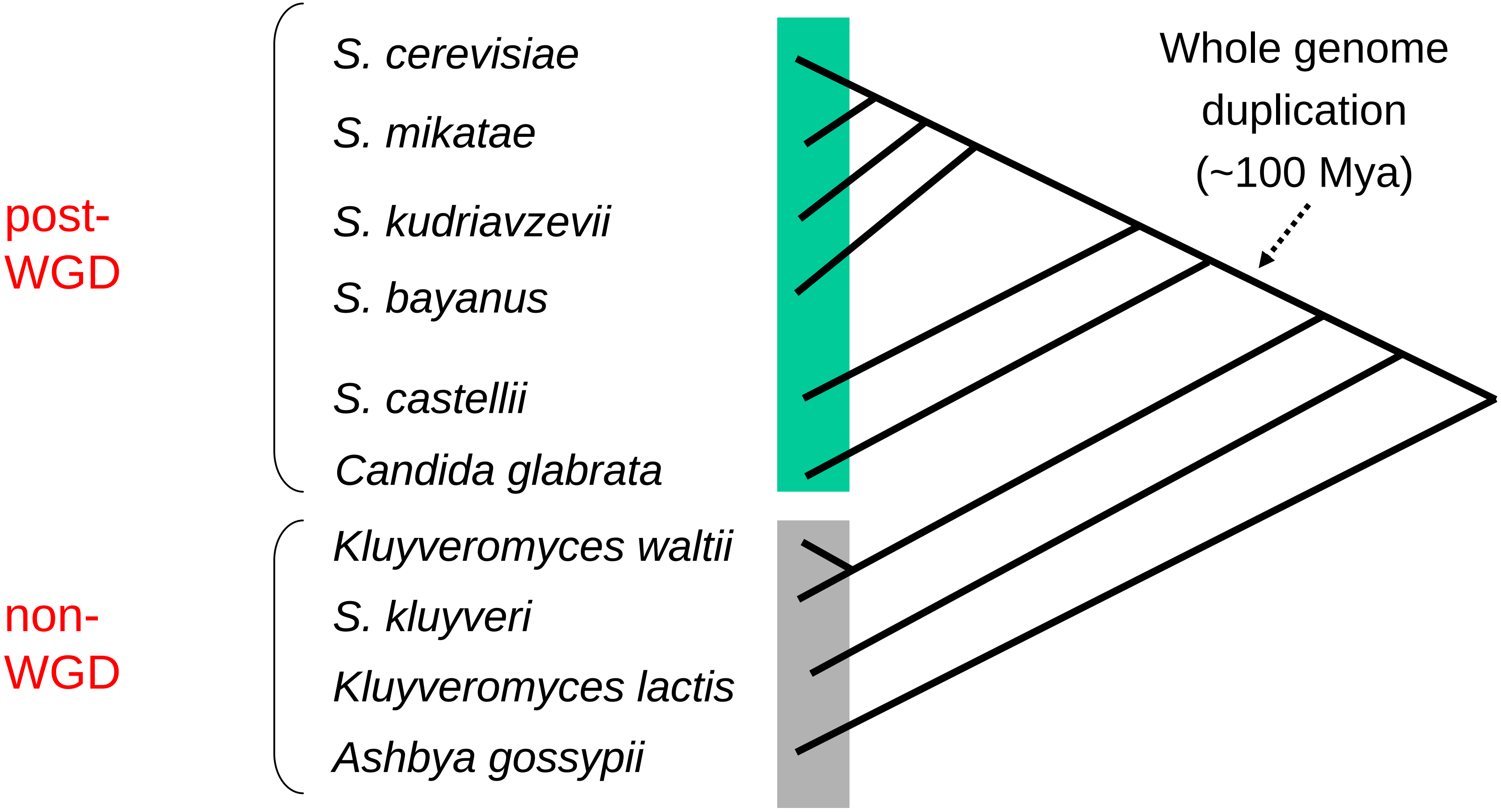
Yeong-Shin Lin, Jake Byrnes, and W.-H. Li

Tetraploid origin of the *Saccharomyces cerevisiae* genome

Wolfe and Shields (1997)

Kellis et al. (2004)

Date of the WGD (whole genome duplication) event



post-WGD

S. cerevisiae

S. mikatae

S. kudriavzevii

S. bayanus

S. castellii

Candida glabrata

non-WGD

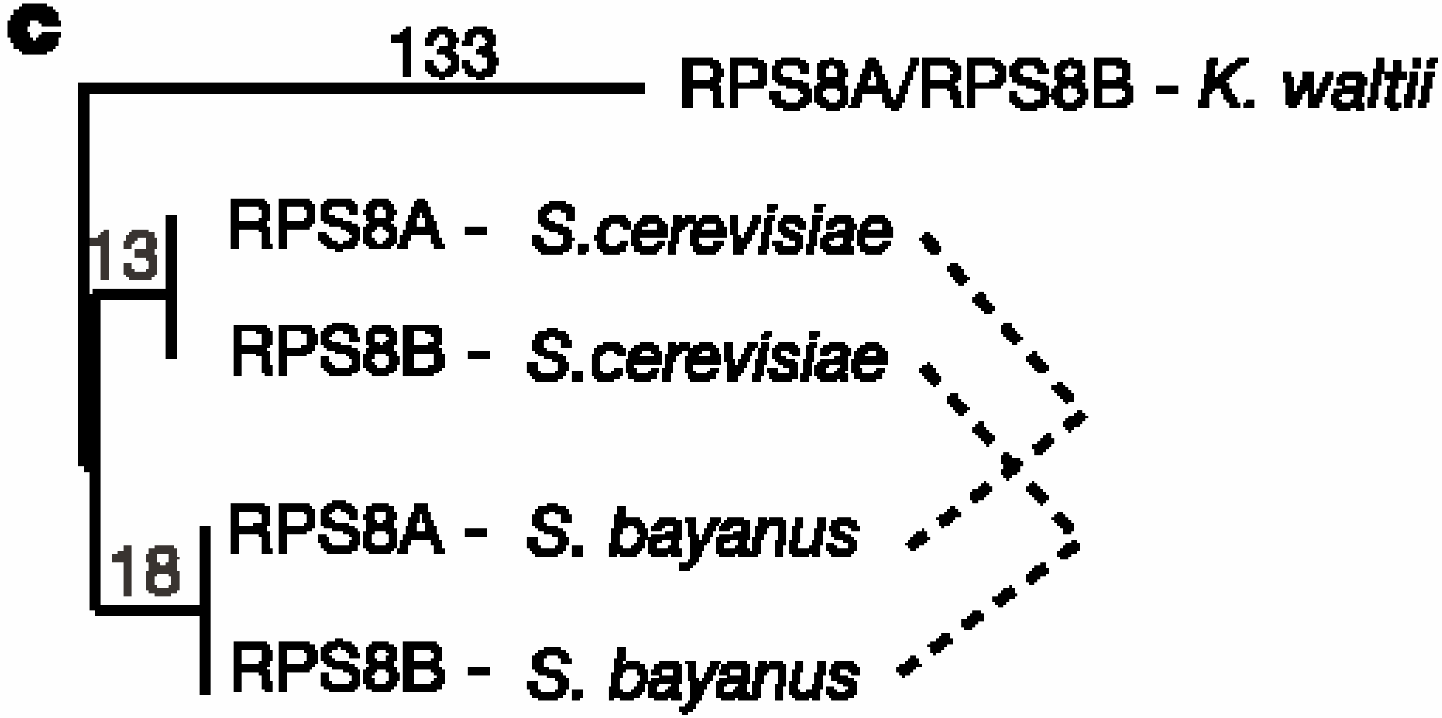
Kluyveromyces waltii

S. kluyveri

Kluyveromyces lactis

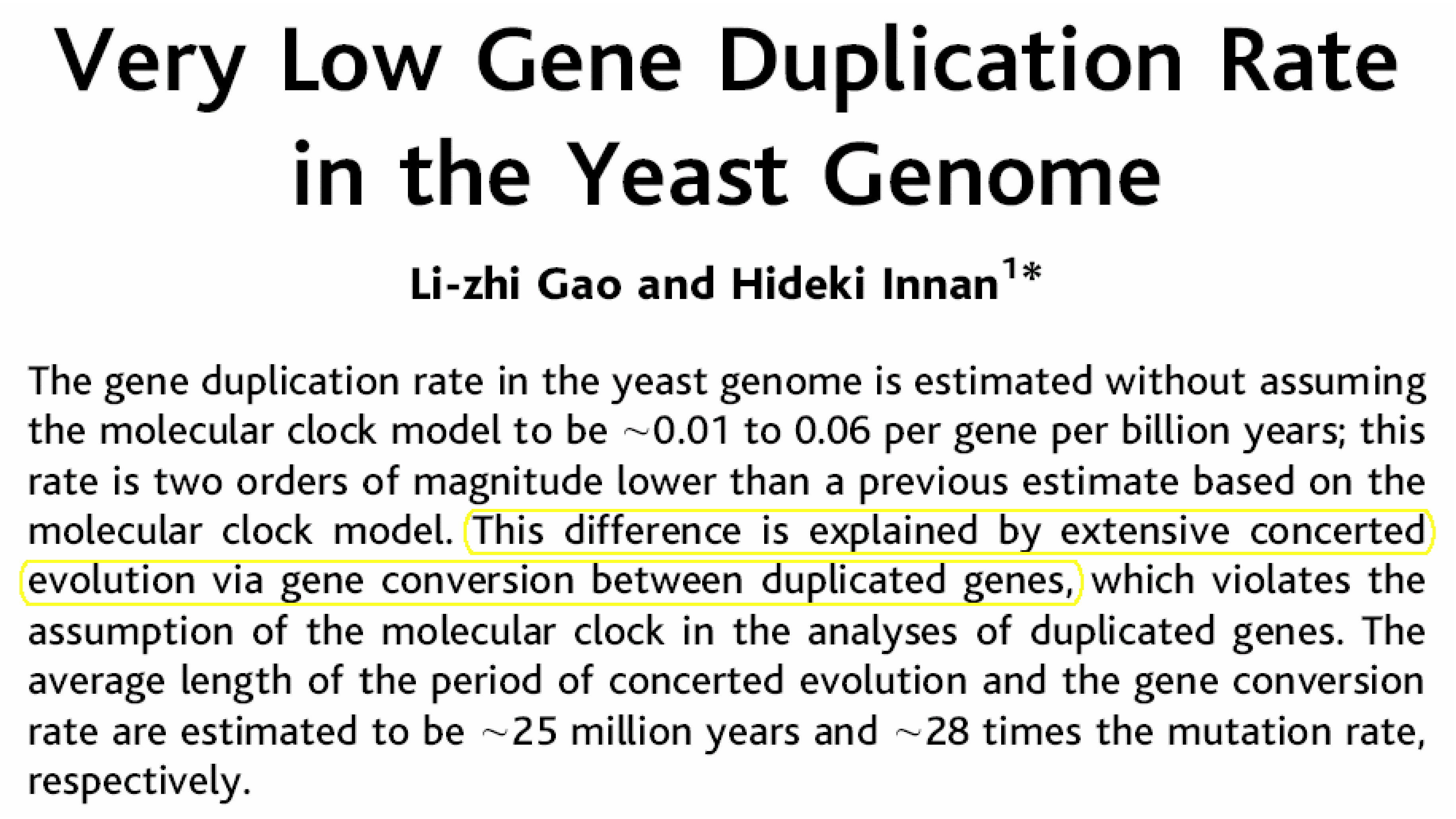
Ashbya gossypii

Whole genome duplication (~100 Mya)



The strong nucleotide similarity between the two duplicate genes at 4-fold degenerate codon positions provides evidence for gene conversion (>90% nucleotide identity versus 41% for all pairs).

60 of the 457 pairs derived from WGD show decelerated protein evolution, defined as instances where one or both copies evolved at least 50% slower than *K. waltii*. In the vast majority of cases (90%), the paralogues both show decelerated evolution and tend to be very similar (98% amino acid identity versus 55% for all pairs), suggesting that they may be subject to periodic gene conversion.



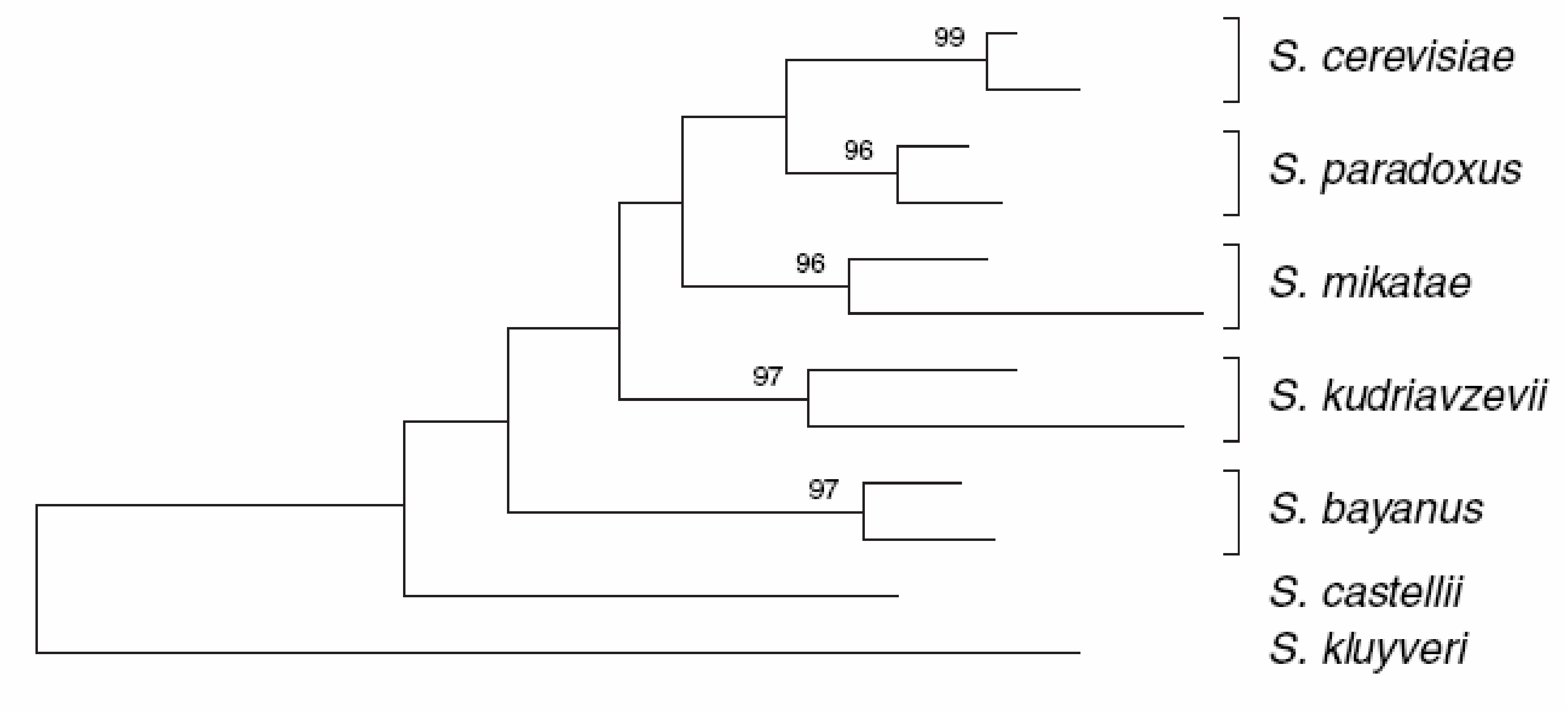
Very Low Gene Duplication Rate in the Yeast Genome

Li-zhi Gao and Hideki Innan[1*]

The gene duplication rate in the yeast genome is estimated without assuming the molecular clock model to be ~0.01 to 0.06 per gene per billion years; this rate is two orders of magnitude lower than a previous estimate based on the molecular clock model. This difference is explained by extensive concerted evolution via gene conversion between duplicated genes, which violates the assumption of the molecular clock in the analyses of duplicated genes. The average length of the period of concerted evolution and the gene conversion rate are estimated to be ~25 million years and ~28 times the mutation rate, respectively.

Gao and Innan (2004) Science

YGL135W and YPL220W



YDL136W and YDL191W

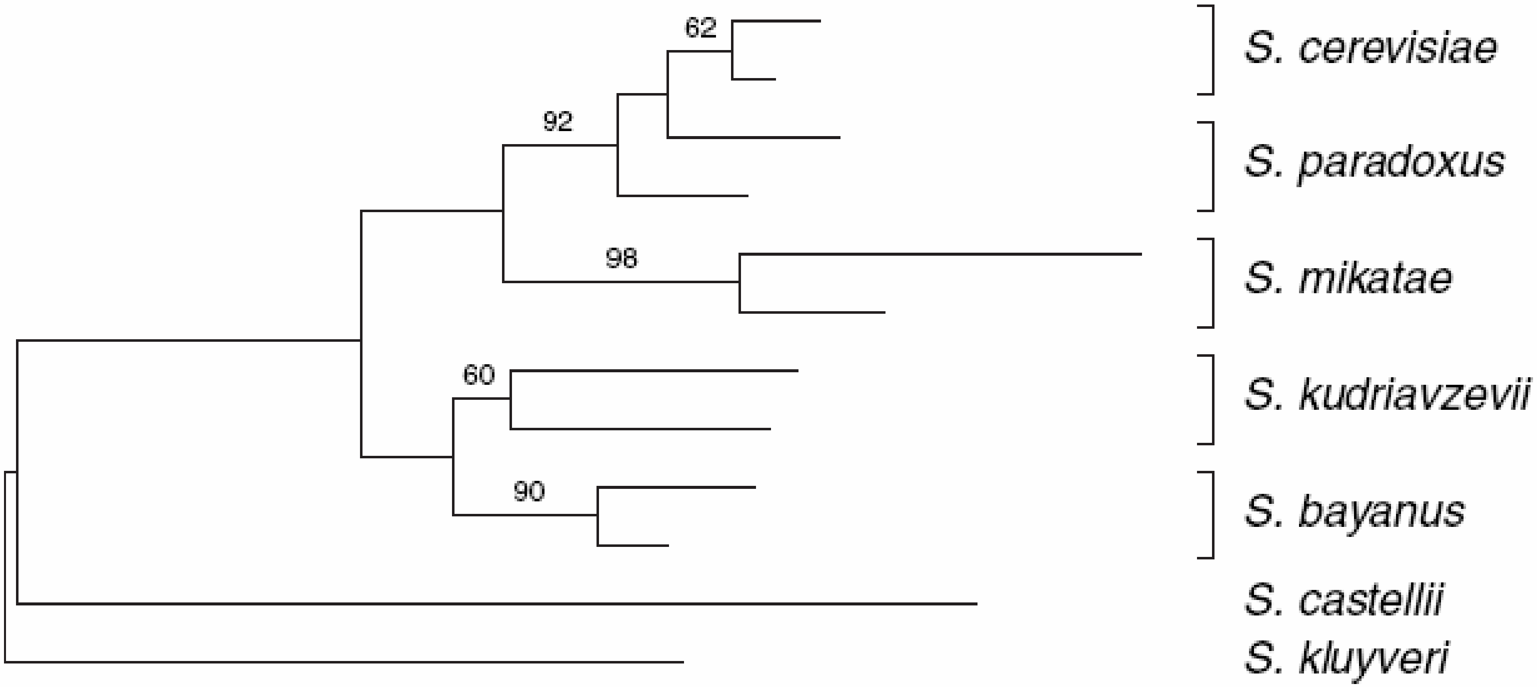


Fig. 2. Evidence for extensive concerted evolution shown in neighbor-joining (NJ) gene trees. (A) Gene tree for the orthologs of YGL135W and YPL220W (the sixth gene pair in Table 1). (**B**) Gene tree for the orthologs of YDL136W and YDL191W (the eighth gene pair in Table 1).

Conclusion:
Gene conversion has occurred frequently between some duplicate genes in yeasts and decelerated the divergence

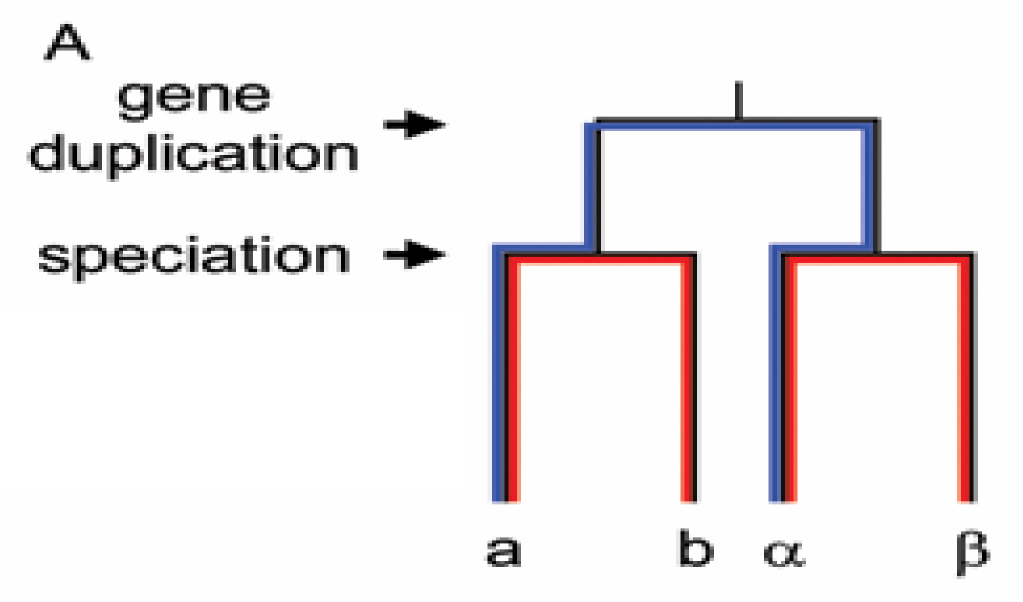
Question: Why gene conversion still continues to occur in these duplicate genes ~100 million years after the WGD?

Codon usage bias

Nonrandom usage of synonymous codons

CAI (Codon adaptation index) (Sharp and Li 1987): A measure of the strength of codon usage bias.

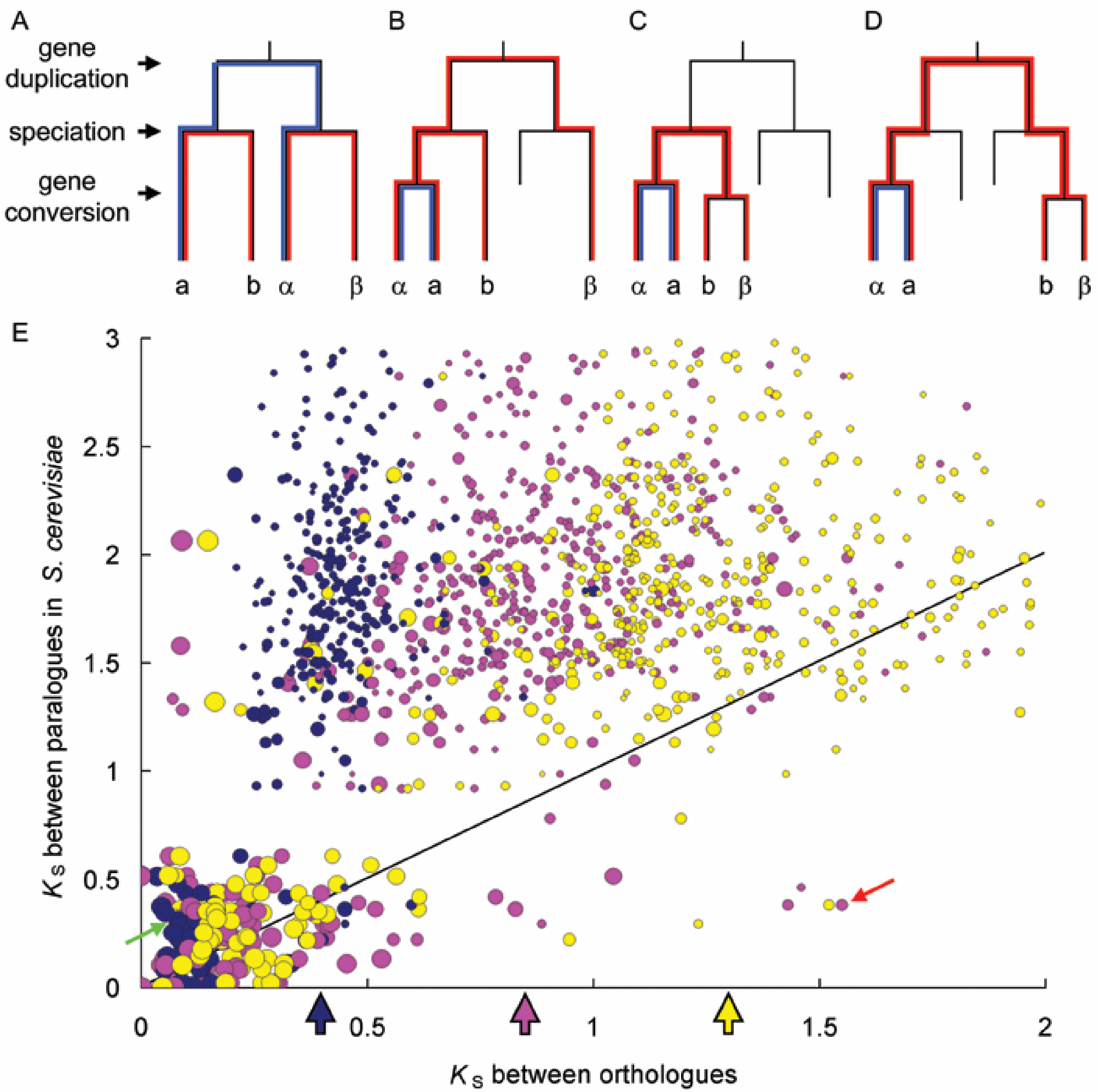
Value between 0 and 1:
The larger the CAI value, the stronger the bias



$$D(\mathrm{a},\mathrm{b}) < D(\mathrm{a},\alpha)$$

- Orthologues in species 1 & 2 2 paralogues in the same species

.



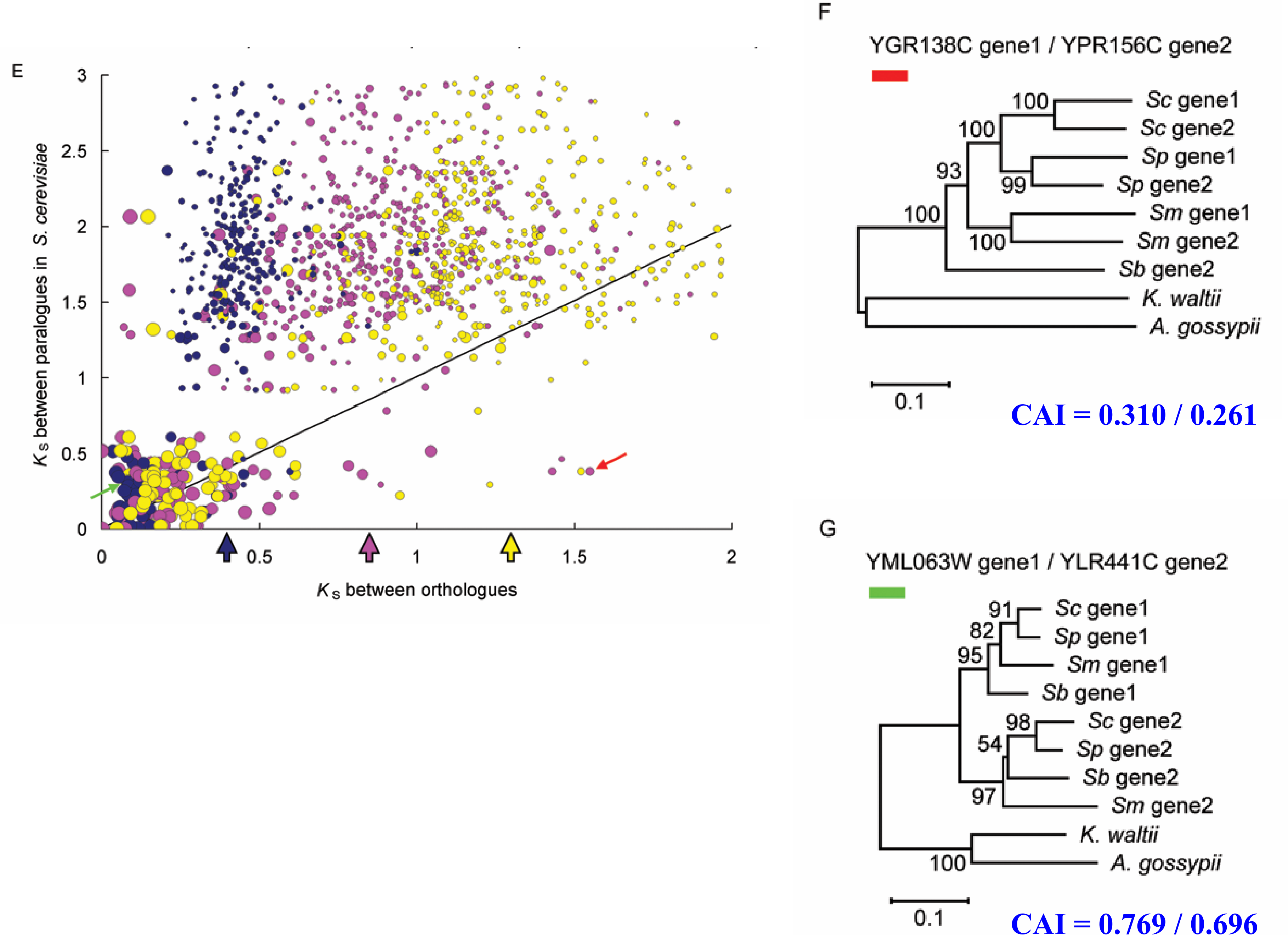
A: *D*(a,b) < *D*(a,α)
B, C, and D:
D(a,b) > *D*(a,α)

S. cerevisiae **vs.** *S. paradoxus*

S. cerevisiae **vs.** *S. mikatae*

S. cerevisiae vs. *S. bayanus*

Circle sizes indicate the CAI values of the genes in *S. cerevisiae*.


E
K_S between paralogues in *S. cerevisiae*
3
2.5
2
1.5
1
0.5
0
0
0.5
1
1.5
2
K_S between orthologues
F
YGR138C gene1 / YPR156C gene2
100
100
93
99
100
100
Sc gene1
Sc gene2
Sp gene1
Sp gene2
Sm gene1
Sm gene2
Sb gene2
K. waltii
A. gossypii
0.1
CAI = 0.310 / 0.261
G
YML063W gene1 / YLR441C gene2
91
82
95
98
54
97
100
Sc gene1
Sp gene1
Sm gene1
Sb gene1
Sc gene2
Sp gene2
Sb gene2
Sm gene2
K. waltii
A. gossypii
0.1
CAI = 0.769 / 0.696

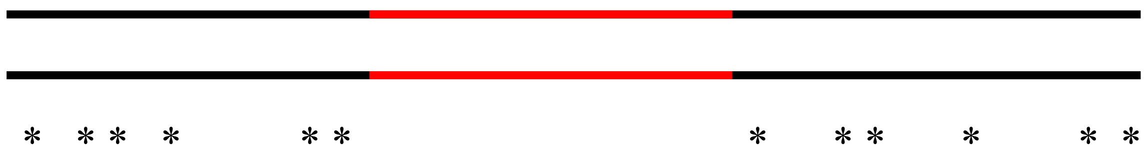
Gene pairs	CAI values	Gene pairs	CAI values
With completely distorted tree topology		With no topology distortion	
YHL033C / YLL045C	0.842 / 0.849	YER074W / YIL069C	0.816 / 0.756
YGL135W / YPL220W	0.832 / 0.821	YJL190C / YLR367W	0.812 / 0.523
YBR031W / YDR012W	0.803 / 0.812	YMR230W / YOR293W	0.802 / 0.840
YBR009C / YNL030W	0.734 / 0.627	YDR450W / YML026C	0.775 / 0.733
YHR203C / YJR145C	0.709 / 0.695	YDR418W / YEL054C	0.766 / 0.605
YMR186W / YPL240C	0.581 / 0.518	YER056C-A / YIL052C	0.763 / 0.781
YDL131W / YDL182W	0.329 / 0.321	YDL061C / YLR388W	0.760 / 0.653
YDR312W / YHR066W	0.160 / 0.189	YDR447C / YML024W	0.757 / 0.810
With partially distorted tree topology		YNL302C / YOL121C	0.757 / 0.794
YBR181C / YPL090C	0.846 / 0.837	YOR234C / YPL143W	0.730 / 0.747
YEL034W / YJR047C	0.814 / 0.704	YGR118W / YPR132W	0.726 / 0.789
YBR189W / YPL081W	0.810 / 0.507	YER131W / YGL189C	0.711 / 0.781
YCR031C / YJL191W	0.805 / 0.590	YDR500C / YLR185W	0.711 / 0.700
YHR141C / YNL162W	0.795 / 0.769	YLR441C / YML063W	0.696 / 0.769
YLR029C / YMR121C	0.783 / 0.436	YJL136C / YKR057W	0.693 / 0.596
YFR031C-A / YIL018W	0.773 / 0.764	YBR191W / YPL079W	0.691 / 0.733
YGL147C / YNL067W	0.771 / 0.778	YJL177W / YKL180W	0.680 / 0.809
YDL083C / YMR143W	0.764 / 0.677	YGR034W / YLR344W	0.677 / 0.631
YDL136W / YDL191W	0.759 / 0.798	YGR214W / YLR048W	0.668 / 0.733
YGL031C / YGR148C	0.759 / 0.756	YMR242C / YOR312C	0.665 / 0.697
YBL072C / YER102W	0.747 / 0.718	YLR448W / YML073C	0.627 / 0.672
YDL075W / YLR406C	0.737 / 0.630	YMR194W / YPL249C-A	0.620 / 0.800
YBR048W / YDR025W	0.733 / 0.705	YIL133C / YNL069C	0.611 / 0.723
YGR085C / YPR102C	0.727 / 0.781	YNL096C / YOR096W	0.597 / 0.747
YGR027C / YLR333C	0.716 / 0.612	YJR094W-A / YPR043W	0.571 / 0.872
YBL027W / YBR084C-A	0.708 / 0.686	YLR264W / YOR167C	0.561 / 0.528
YNL301C / YOL120C	0.680 / 0.812		
YHL001W / YKL006W	0.680 / 0.684		
YDL082W / YMR142C	0.652 / 0.742		
YLR287C-A / YOR182C	0.642 / 0.748		
YBL087C / YER117W	0.624 / 0.648		
YBL002W / YDR224C	0.563 / 0.658		

Gao and Innan identified 57 gene pairs for which gene conversion likely occurred at the divergence time between *S. cerevisiae* and *S. bayanus*

Detection of Gene Conversion

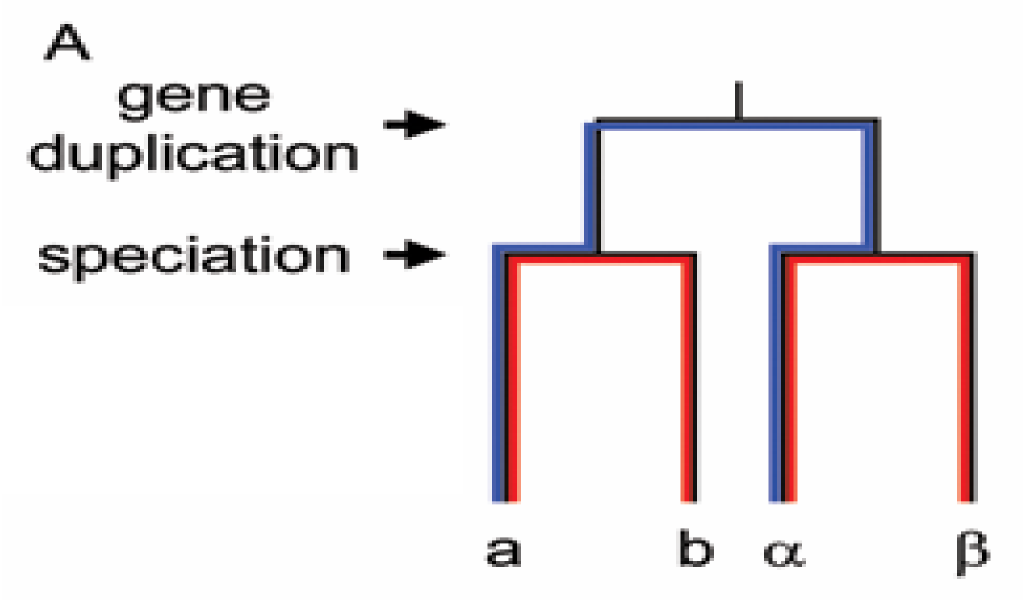
Possible gene conversion

Duplicate genes



Synonymous substitutions

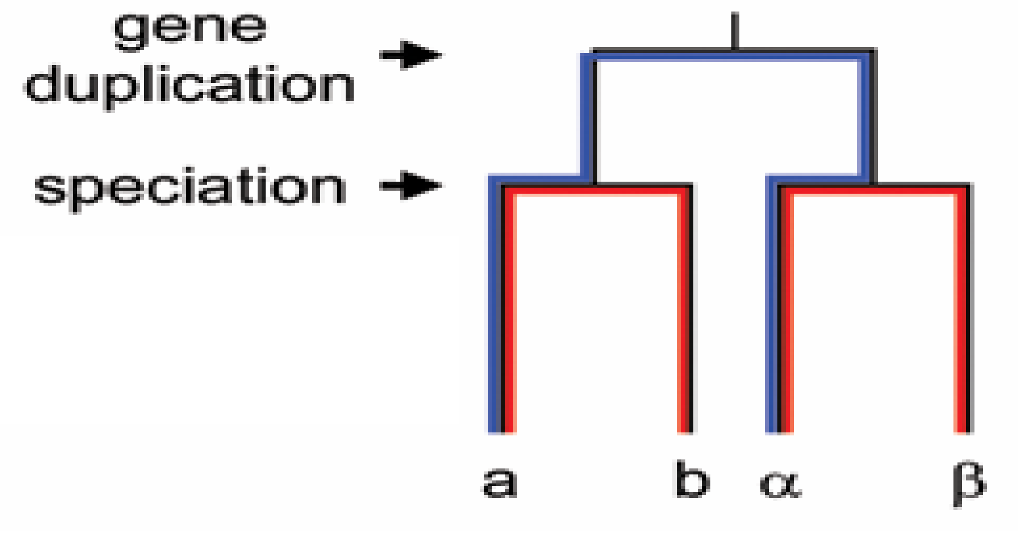
Sawyer (1989) MBE



$$D(\mathrm{a},\mathrm{b}) < D(\mathrm{a},\alpha)$$

- Orthologues in species 1 & 2 2 paralogues in the same species

.



D_i = # of nucleotide differences between the two nucleotides at site i in paralogous gene 1 and gene 2 (a & α) in species 1

B_{ji} = # of nucleotide differences between gene j (j = 1, 2) in species 1(a or α) and its orthologue in species 2 (b or β).

Let $B_i = (B_{1i} + B_{2i}) / 2$. **Note:** $\boldsymbol{D_i \geq B_i}$

The null hypothesis: no gene conversion

$$D_i - B_i \geq 0$$

Dynamic programming is used to select the segment from site *m* to *n* that maximizes

$$\sum_{i=m}^{n} (B_i - D_i)$$

This segment has N sites, $N = n - m + 1$

Let $D = \sum_{i=m}^{n} D_i$. and $B = \sum_{i=m}^{n} B_i$

If $N \geq 20$, the probability to observe $D \leq$ B for a segment of N sites is calculated, assuming $D = B$. This is a stringent criterion because D ≥ B, as shown above.

The estimated probability is

$$P(B,D,N) = \sum_{k=0}^{D} [N!/k!(N-k)!](B/N)^{k}(1-B/N)^{(N-k)}$$

A segment thus identified might have undergone a gene conversion.

However, many possible segments of *N* sites can be selected from the entire gene sequence, so we need to take this factor into consideration.

For each segment with $P < 0.01$, we construct an empirical distribution of B for a segment of length N using 10,000 bootstrap samples from $\{B_1, B_2, ..., B_L\}$, where L = alignment length for the gene under consideration.

Then, we determine the significance of D by counting the proportion of samples for which $D < B$.

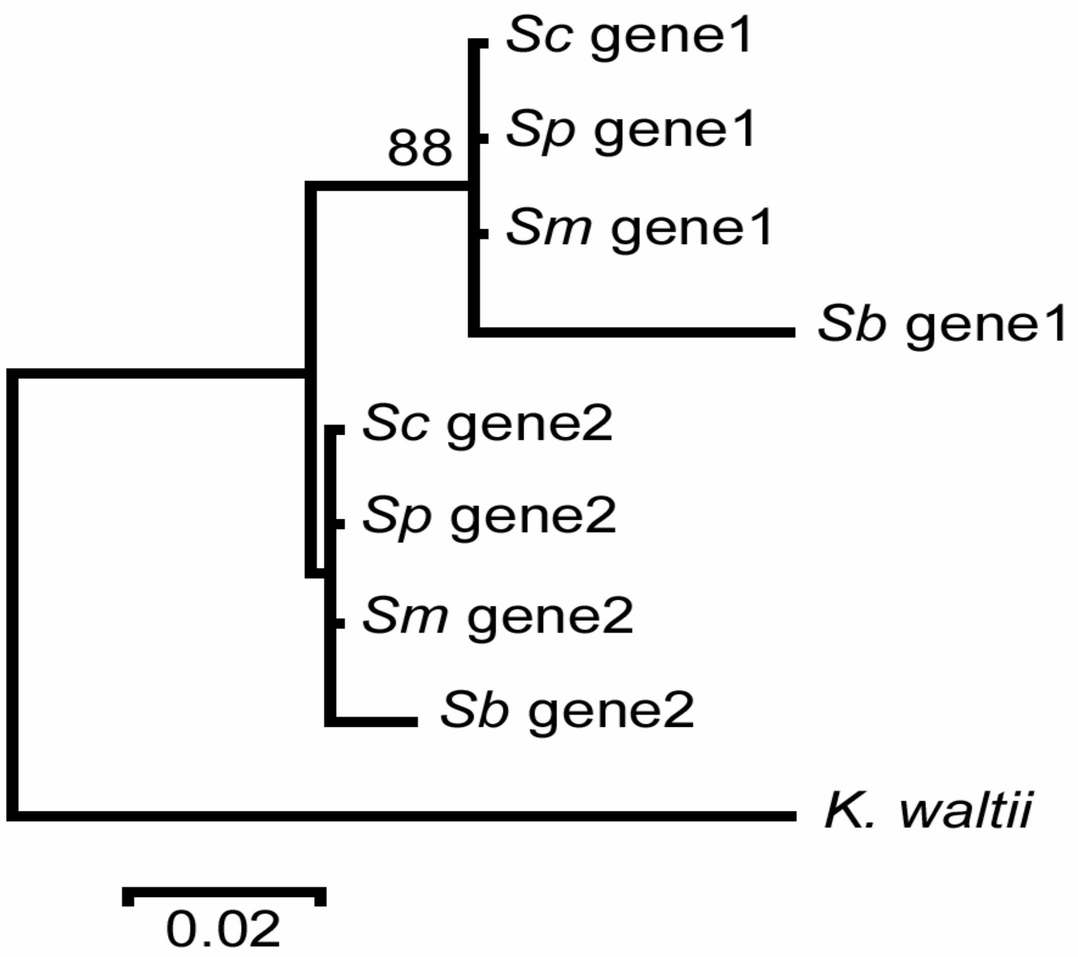
Segments with a binomial probability *P* < 0.01 and with an empirical probability < 0.01 are considered candidate gene conversions.

Number of gene pairs with detected gene conversion events / total

	CAI ≥ 0.7	0.7 > CAI ≥ 0.5	CAI < 0.5	*P* value
S. cerevisiae	**12 / 21**	**4 / 31**	**4 / 238**	**$< 10^{-8}$**
S. paradoxus	**9 / 18**	**3 / 28**	**3 / 215**	**$< 10^{-8}$**
S. mikatae	**8 / 20**	**4 / 29**	**3 / 161**	**$< 10^{-8}$**
S. bayanus	**15 / 21**	**11 / 31**	**6 / 246**	**$< 10^{-8}$**

Only detected conversion events longer than 20 bp were reported.

Codon usage bias increases the rate of gene conversion by reducing the rate of sequence divergence. In the absence of strong codon usage bias, synonymous divergence between duplicate genes increases with time, and the chance of gene conversion is concomitantly reduced.

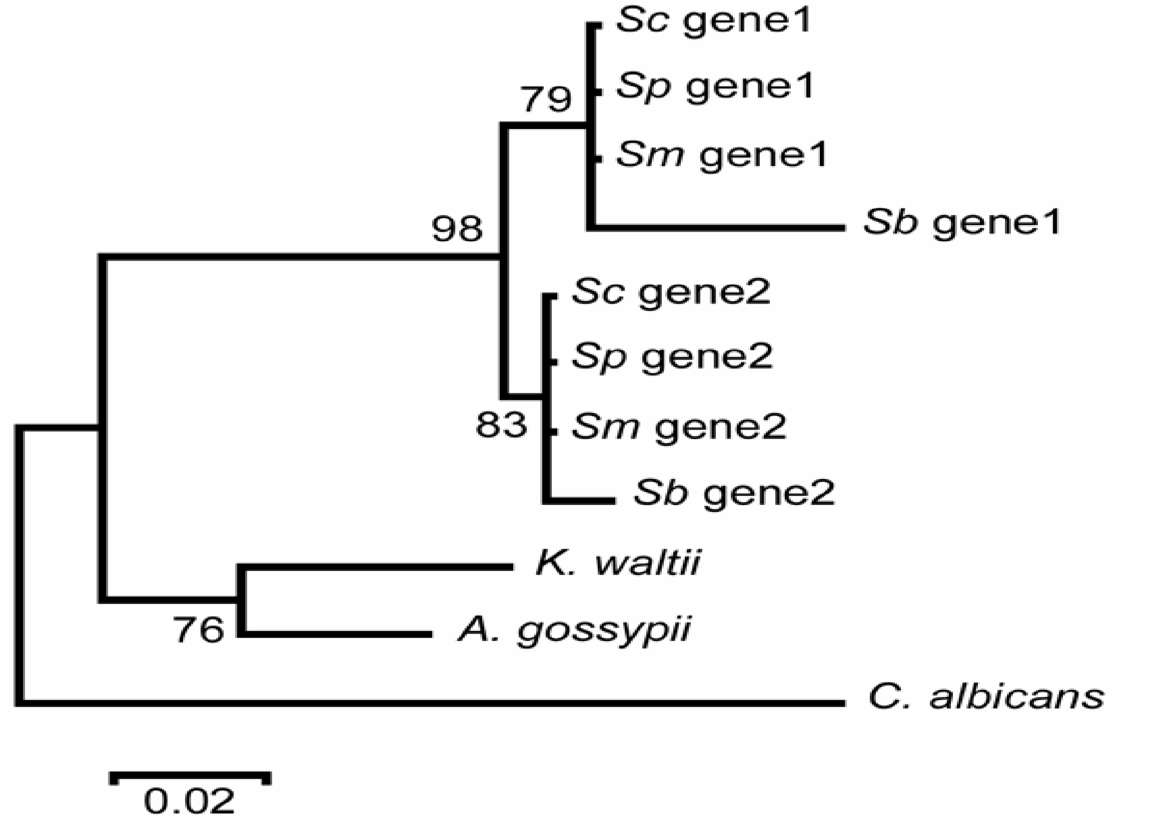


Protein sequences

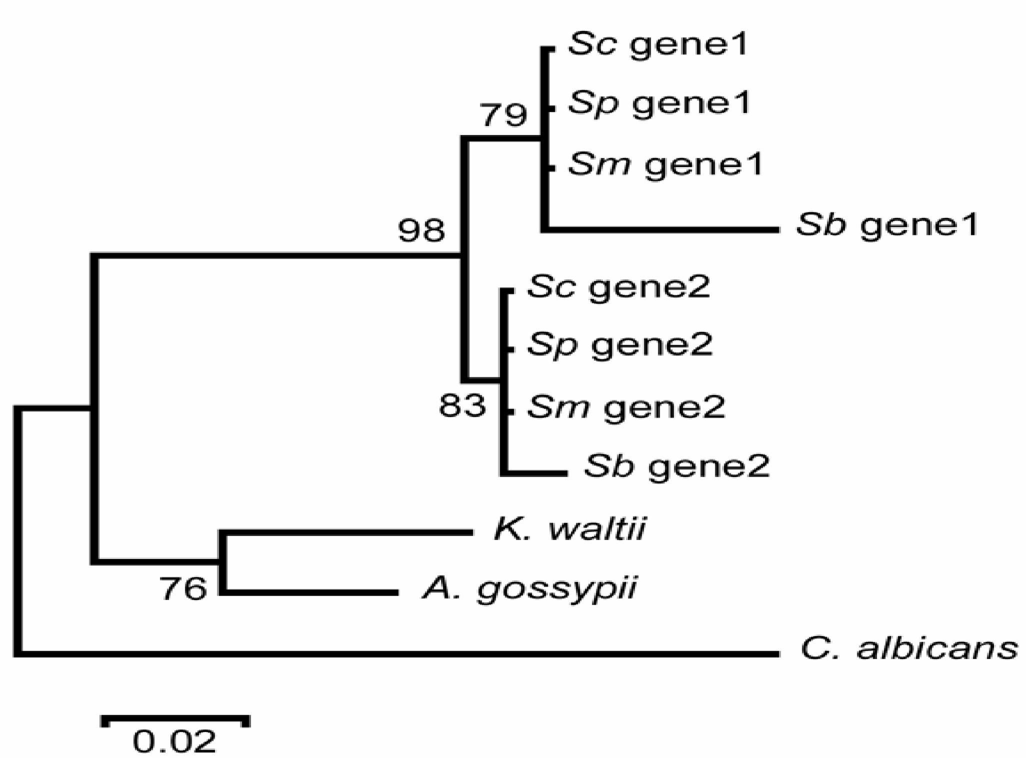
Decelerated evolution compared to *K. waltii*

Kellis et al.(2004)

Neighbor-joining tree of the whole genome duplicated ORFs of *S. cerevisiae* (*Sc*), and their orthologues in *S. paradoxus* (*Sp*), *S. mikatae* (*Sm*), and *S. bayanus* (*Sb*), and outgroups *K. waltii* for YER131W (gene1) / YGL189C (gene2) (cytoplasmic small ribosomal subunits; CAI = 0.711 / 0.781). The tree was constructed using protein Poisson distances.



Neighbor-joining tree of the whole genome duplicated ORFs of *S. cerevisiae* (*Sc*), and their orthologues in *S. paradoxus* (*Sp*), *S. mikatae* (*Sm*), and *S. bayanus* (*Sb*), and outgroups *K. waltii*, *A. gossypii* and *C. albicans* for YER131W (gene1) / YGL189C (gene2) (cytoplasmic small ribosomal subunits; CAI = 0.711 / 0.781). The tree was constructed using protein Poisson distances.



In the period immediately following the WGD event the duplicate proteins had apparently evolved rather rapidly, likely due to relaxed functional constraints following WGD or the emergence of anaerobic growth.

During this period gene conversion might have played a key role in maintaining the sequence similarity between the two paralogues.

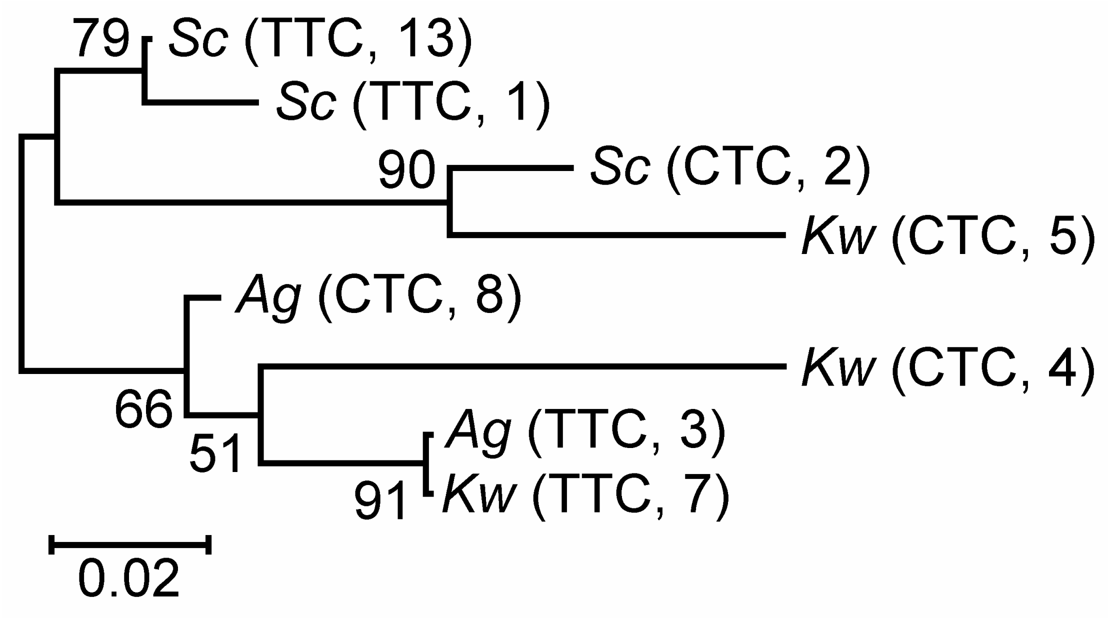
However, the rate of evolution had evidently become very slow prior to the radiation of the four *Saccharomyces* species and this largely explains why the sequence divergence is small between not only paralogues but also orthologues.

Genome GC content	52%	44%	38% ~ 40%			
	A. gossypii	*K. waltii*	*S. cerevisiae*	*S. paradoxus*	*S. mikatae*	*S. bayanus*
a. a. - codon	Relative frequencies of codon usage					
Asp - GAU	0.14 / 0.44	0.16 / 0.55	0.43 / 0.67	0.45 / 0.66	0.46 / 0.68	0.39 / 0.61
Asp - GAC	**0.86** / 0.56	**0.84** / 0.45	**0.57** / 0.33	**0.55** / 0.34	**0.54** / 0.32	**0.61** / 0.39
Cys - UGU	**0.43** / 0.35	**0.61** / 0.48	**0.89** / 0.61	**0.91** / 0.60	**0.92** / 0.63	**0.92** / 0.58
Cys - UGC	0.57 / 0.65	0.39 / 0.52	0.11 / 0.39	0.09 / 0.40	0.08 / 0.37	0.08 / 0.42
Gln - CAA	0.21 / 0.30	0.56 / 0.54	**0.98** / 0.67	**0.97** / 0.66	**0.97** / 0.66	**0.99** / 0.65
Gln - CAG	**0.79** / 0.70	0.44 / 0.46	0.02 / 0.33	0.03 / 0.34	0.03 / 0.34	0.01 / 0.35
Glu - GAA	0.13 / 0.38	0.31 / 0.54	**0.95** / 0.69	**0.95** / 0.68	**0.93** / 0.68	**0.96** / 0.67
Glu - GAG	**0.87** / 0.62	**0.69** / 0.46	0.05 / 0.31	0.05 / 0.32	0.07 / 0.32	0.04 / 0.33
His - CAU	0.08 / 0.44	0.09 / 0.54	0.28 / 0.66	0.30 / 0.65	0.33 / 0.67	0.27 / 0.62
His - CAC	**0.92** / 0.56	**0.91** / 0.46	**0.72** / 0.34	**0.70** / 0.35	**0.67** / 0.33	**0.73** / 0.38

Genes in *K. waltii* and *A. gossypii* have a stronger preference for G and C at third codon positions than genes in the four *Saccharomyces* species. Therefore, the *K. waltii* - *A. gossypii* lineage and the *Saccharomyces* lineage might have evolved under different selection pressures and rapidly diverged in highly expressed genes.

	A. gossypii	*K. waltii*	*S. cerevisiae*	*S. paradoxus*	*S. mikatae*	*S. bayanus*
a. a. - codon	Relative frequencies of codon usage					
Asp - GAU	0.14 / 0.44	0.16 / 0.55	0.43 / 0.67	0.45 / 0.66	0.46 / 0.68	0.39 / 0.61
Asp - GAC	**0.86** / 0.56	**0.84** / 0.45	**0.57** / 0.33	**0.55** / 0.34	**0.54** / 0.32	**0.61** / 0.39
Cys - UGU	**0.43** / 0.35	**0.61** / 0.48	**0.89** / 0.61	**0.91** / 0.60	**0.92** / 0.63	**0.92** / 0.58
Cys - UGC	0.57 / 0.65	0.39 / 0.52	0.11 / 0.39	0.09 / 0.40	0.08 / 0.37	0.08 / 0.42
Gln - CAA	0.21 / 0.30	0.56 / 0.54	**0.98** / 0.67	**0.97** / 0.66	**0.97** / 0.66	**0.99** / 0.65
Gln - CAG	**0.79** / 0.70	0.44 / 0.46	0.02 / 0.33	0.03 / 0.34	0.03 / 0.34	0.01 / 0.35
Glu - GAA	0.13 / 0.38	0.31 / 0.54	**0.95** / 0.69	**0.95** / 0.68	**0.93** / 0.68	**0.96** / 0.67
Glu - GAG	**0.87** / 0.62	**0.69** / 0.46	0.05 / 0.31	0.05 / 0.32	0.07 / 0.32	0.04 / 0.33
His - CAU	0.08 / 0.44	0.09 / 0.54	0.28 / 0.66	0.30 / 0.65	0.33 / 0.67	0.27 / 0.62
His - CAC	**0.92** / 0.56	**0.91** / 0.46	**0.72** / 0.34	**0.70** / 0.35	**0.67** / 0.33	**0.73** / 0.38
a. a. - Anticodon	Numbers of tDNA genes					
Asp - ATC	0	0	0	0	0	0
Asp - GTC	10	10	16	19	16	16
Cys - ACA	0	0	0	0	0	0
Cys - GCA	3	4	4	3	4	4
Gln - TTG	4	6	9	9	9	9
Gln - CTG	4	4	1	1	1	1
Glu - TTC	3	7	14	15	14	14
Glu - CTC	8	9	2	2	2	2
His - ATG	0	0	0	0	0	0
His - GTG	5	4	7	7	7	7

The neighbor-joining tree of tDNA-Glu genes among four yeast species (*Sc*, *S. cerevisiae*; *Kw*, *K. waltii*; *Ag*, *A. gossypii*). The triplet and number in the parenthesis indicate, respectively, the tDNA anticodon and the gene copy number in corresponding genome. The numbers at branch nodes are bootstrap values.



This phylogeny suggests the switch between anticodons occurred at least twice in tDNA-Glu evolutionary history for these yeast species.

Conclusion

Our analysis suggests that codon usage bias and protein functional conservation might have been more important than gene conversion for the decelerated evolution of WGD duplicate genes in yeasts.

Remarks

Gene conversion occurs only occasionally, whereas codon usage constraint and functional constraint of proteins are constant forces that slow down sequence evolution.

Remarks

The rate of gene conversion decreases as sequence divergence increases. For this reason, gene conversion may not be an effective means for long-term maintenance of sequence similarity between duplicate genes in the absence of codon usage constraint or functional constraint. In contrast, both codon usage constraint and protein functional constraint can slow down sequence evolution in the absence of gene conversion.

Detecting Gene Conversions with a HMM (Hidden Markov Model)

Jake Byrnes and Wen-Hsiung Li

Divergence of two duplicate genes

```
  acctgatggggactatgccact
  ||||||||||||||||||||||
  acctgatggggactatgccact

  **********************


Time = 0
Substitutions = 0
```

Divergence of two duplicate genes

```
acctgatggggactatgacact
||||||||||||||||||||||
accagatggggtctatgccact

*** ******* ***** ****
```

Time = 5

Substitutions = 3

Divergence of two duplicate genes

acc**t**ga**t**gggg**a**ctatg**ag**act

||||||||||||||||||||||

acc**a**ga**c**gggg**t**ctatg**cc**act

*** ** **** ***** ***

Time = 10

Substitutions = 5

Divergence of two duplicate genes

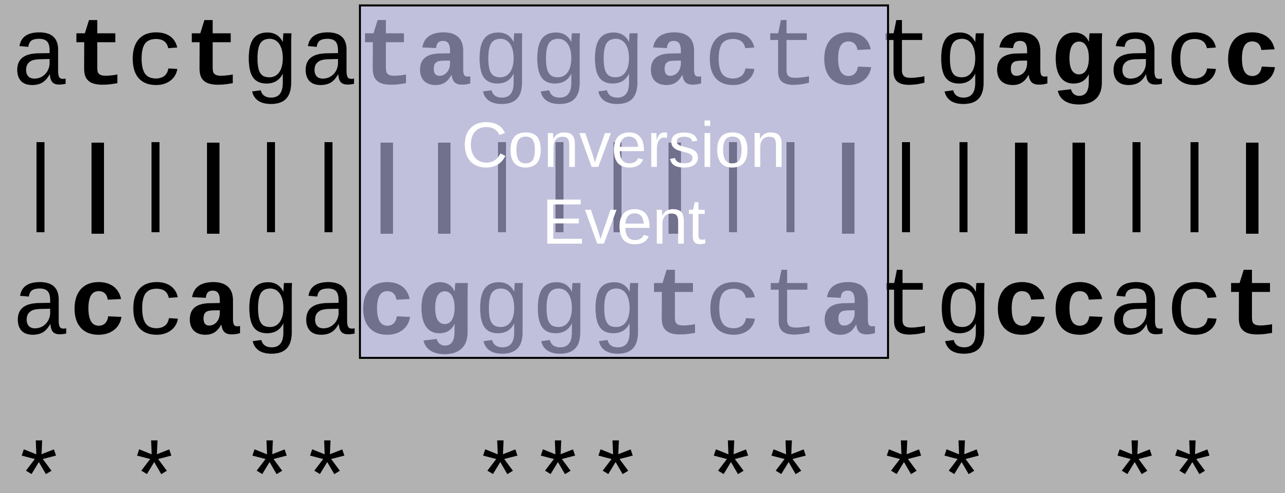
```
atctgatagggactctgagacc
||||||||||||||||||||||
accagacggggtctatgccact

* * **  *** ** **  **
```

Time = 20

Substitutions = 9

Gene Conversion



```
atctgatagggactctgagacc
||||||||||||||||||||||
accagacggggtctatgccact

* * **  *** ** **  **
```

Time = 25

Substitutions = 9

After Gene Conversion

```
atctgatagggactctgagacc
||||||||||||||||||||||
accagatagggactctgccact

* * *************  **
```

Time = 26

Substitutions = 5

HMMs

- A hidden Markov model (HMM) is a statistical model that takes a linear sequence of data as input.
- ***Hidden Property:***

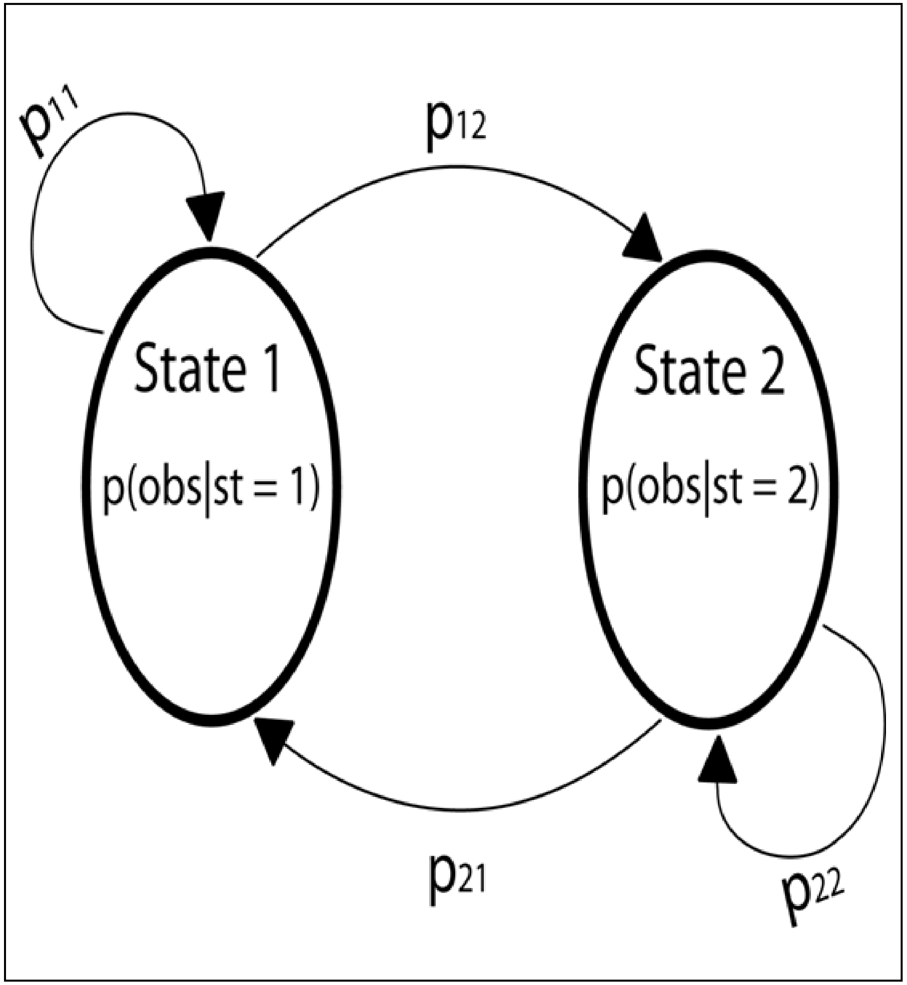
  Observations are dependent on the hidden state at each position.

$$P(o_l) = P(o_l|z_l)$$

- ***Markovian Property:***

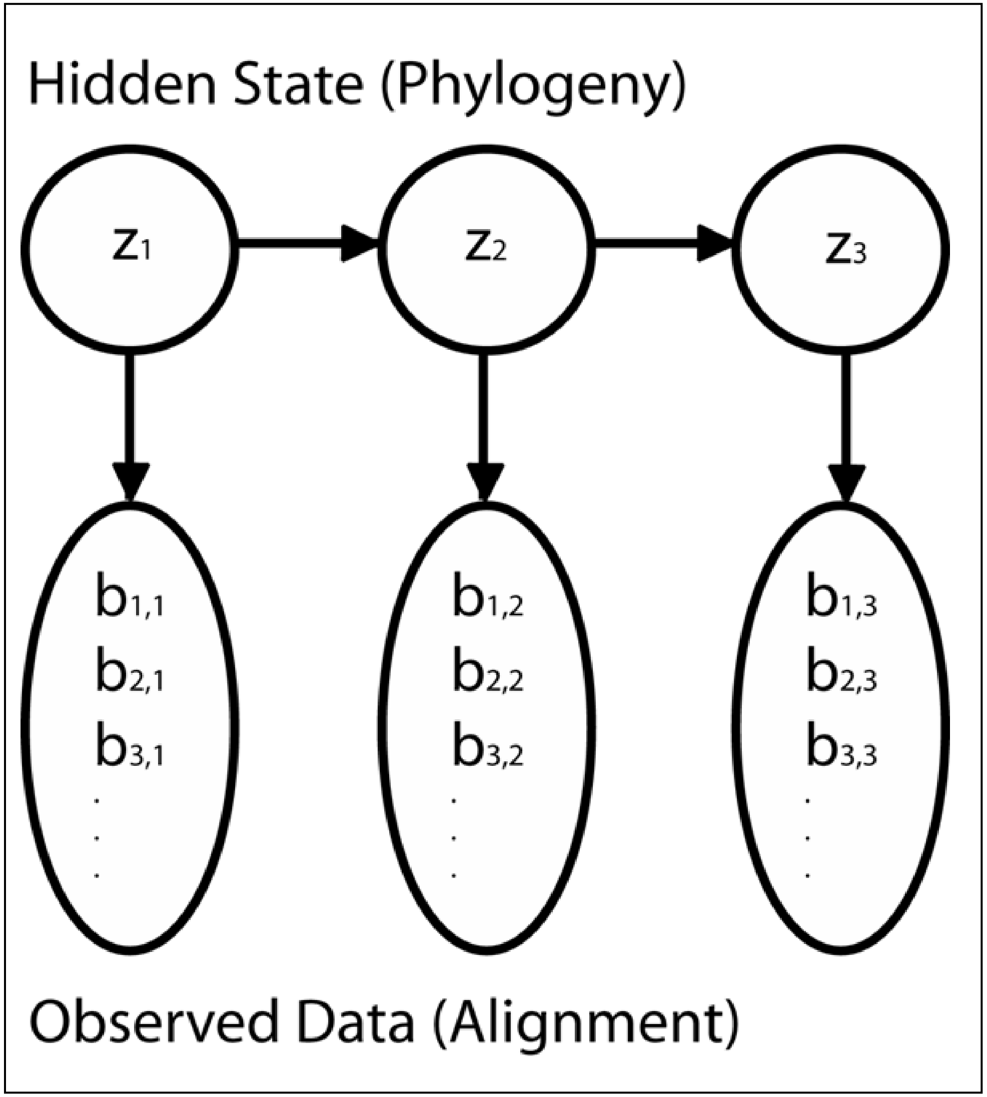
Current state dependent on the state at the previous position.

$$P(z_l = i|z_1, \ldots, z_{l-1}) = P(z_l = i|z_{l-1})$$



Phylo-HMMs and Conversion

- We use a “Phylo-HMM” to find converted regions
- Each hidden state is a phylogeny with different branch lengths
- Each observation is a column of the multiple alignment


Hidden State (Phylogeny)
z_1
z_2
z_3
$b_{1,1}$ $b_{2,1}$ $b_{3,1}$ . . .
$b_{1,2}$ $b_{2,2}$ $b_{3,2}$ . . .
$b_{1,3}$ $b_{2,3}$ $b_{3,3}$ . . .
Observed Data (Alignment)

Three calculations for HMMs

1. Given the model parameters, calculate the probability of the observed data sequence $\vec{O}$ given the model parameters $\vec{\lambda}$

$$P(\vec{O}|\vec{\lambda})$$

2. Calculate the most likely state sequence $\vec{Z}$

$$\underset{\vec{Z}}{\operatorname{argmax}}\, P(\vec{Z}|\vec{O},\vec{\lambda})$$

3. Find the parameters that maximize

$$\underset{\vec{\lambda}}{\operatorname{argmax}}\, P(\vec{\lambda}|\vec{O})$$

Forward and Backward Probabilities

- The forward probability is,

$$f_i(l) = P(o_1, \ldots, o_l, z_l = i)$$

- The forward recursive algorithm allows us to calculate $P(\vec{O}|\vec{\lambda})$

- The backward probability is,

$$b_i(l) = P(o_{(l+1)}, \ldots, o_L | z_l = i)$$

- The backward recursive algorithm also allows us to calculate $P(\vec{O}|\vec{\lambda})$

Forward/Backward Sequence Decoding

- We can calculate the probability of each state for each position as

$$P(z_l = i|\vec{O}) = \frac{f_i(l)b_i(l)}{P(\vec{O}|\vec{\lambda})}$$

- We decode the sequence of hidden states for each position as

$$\hat{z}_l = \underset{k}{\operatorname{argmax}}\, P(z_l = k|\vec{O})$$

Likelihood

- The likelihood equation for an alignment $\mathbf{B}$ given the matrix of branch lengths Φ and the matrix of transition probabilities $\mathbf{M}$

$$P(\mathbf{B}|\Phi,\mathbf{M}) = \sum_{allpossible\vec{Z}} m_{0,z_1} \prod_{l=1}^{L} P(\mathbf{B}[,l]|\Phi[z_l,])m_{z_l,z_{l+1}}$$

where

$$P(\mathbf{B}[,l]|\Phi[z_l,]) = \sum_{b\in\{a,c,g,t\}} p_{b,b_{1,l}}(\Phi[z_l,1])p_{b,b_{2,l}}(\Phi[z_l,2])p_{b,b_{3,l}}(\Phi[z_l,3])$$

and where

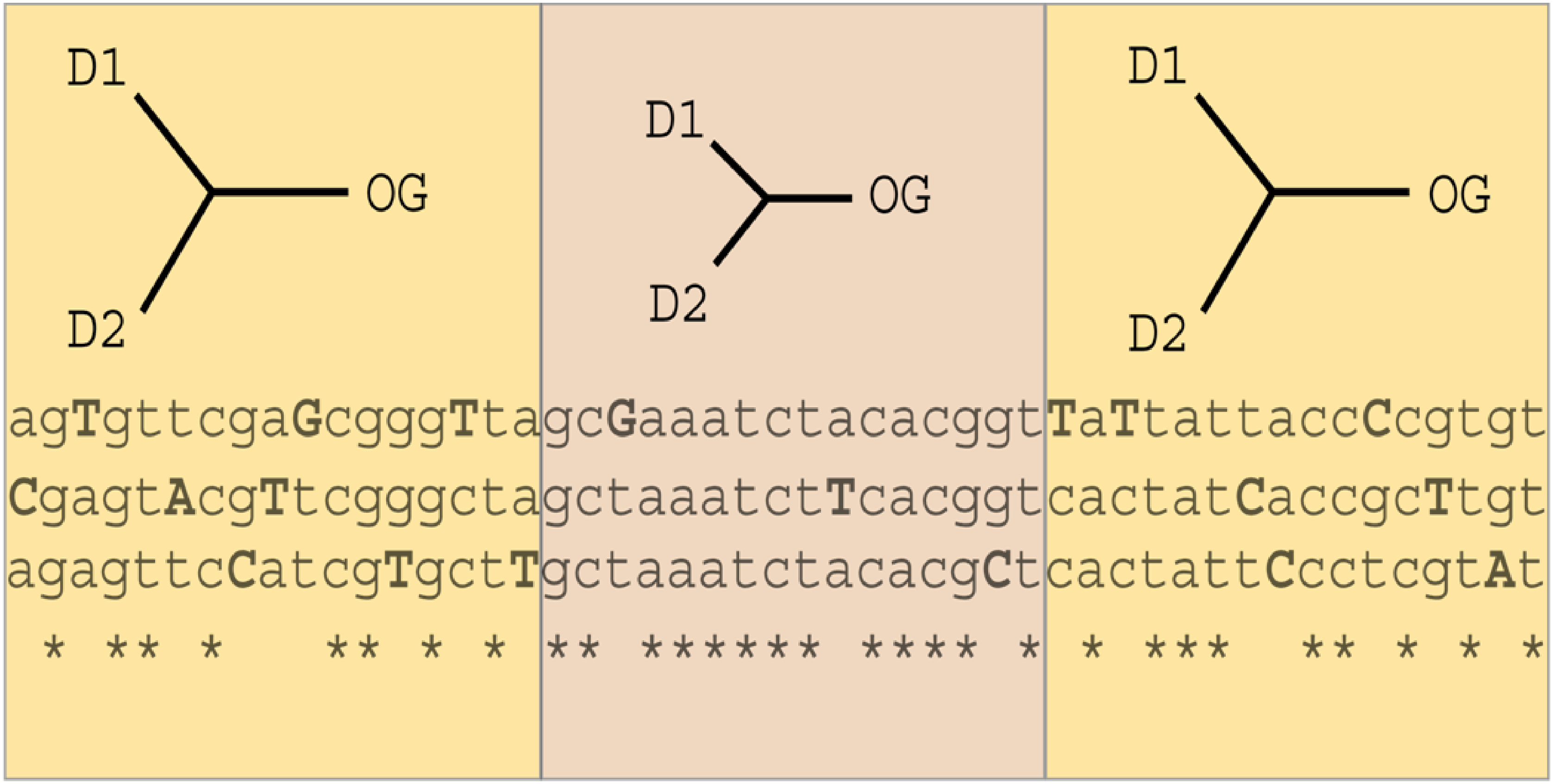
$$p_{b,b_{k,l}}(\Phi[z_l,k]) = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-4\alpha t} & \text{if } b = b_{k,l} \\ \frac{1}{4}-\frac{1}{4}e^{-4\alpha t} & \text{if } b \neq b_{k,l} \end{cases}$$

Conversion vs. Protein Constraint

- An outgroup sequence allows us to distinguish conversion from selective constraint on a functional protein domain.

- A conversion will only shorten the branches to the duplicates, while constraint should act to shorten all branches of the phylogeny

Phylo-HMM Constraint



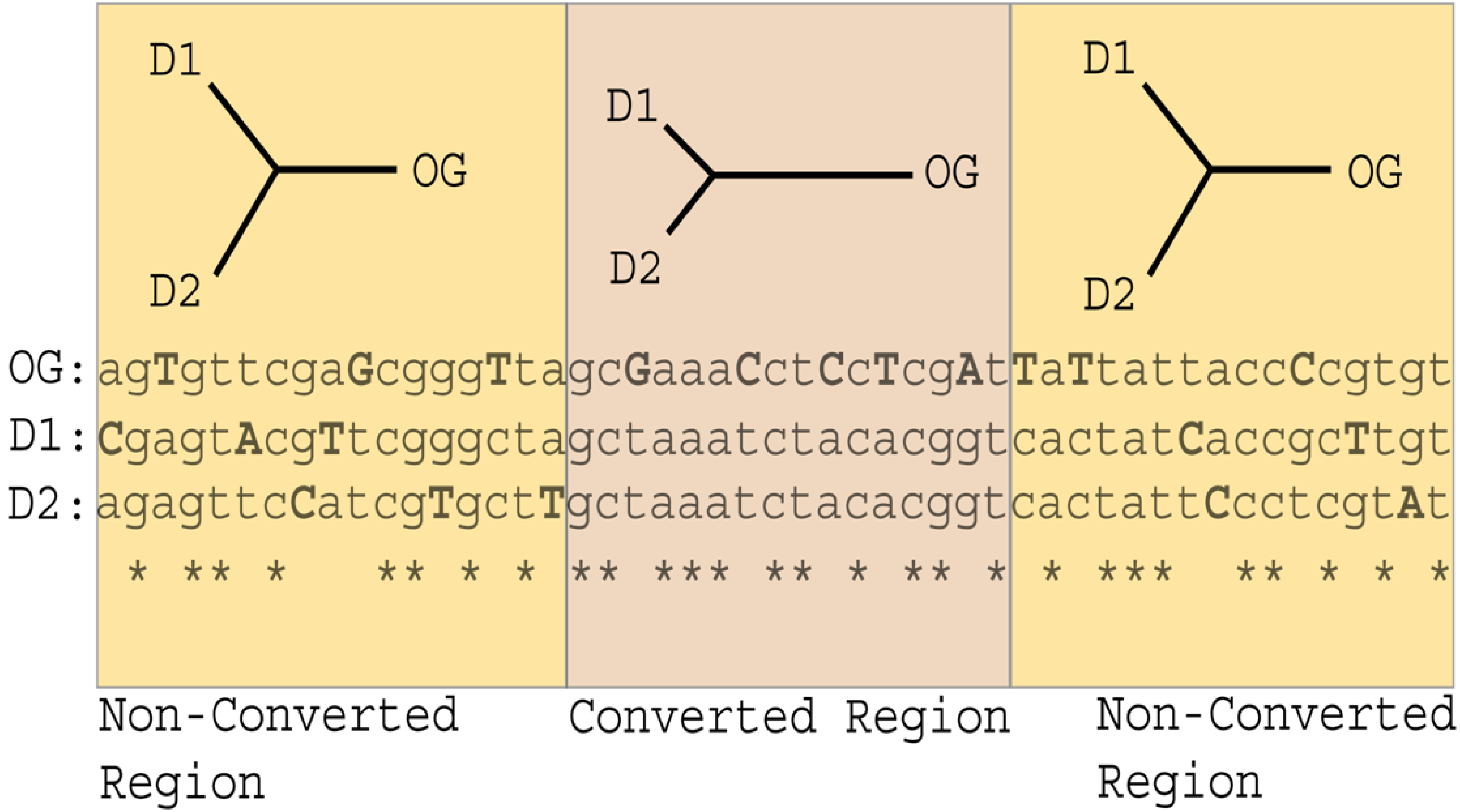
```
OG: agTgttcgaGcgggTtagcGaaatctacacggtTaTtattaccCcgtgt
D1: CgagtAcgTtcgggctagctaaatctTcacggtcactatCaccgcTtgt
D2: agagttcCatcgTgctTgctaaatctacacgCtcactattCcctcgtAt
     * ** *   ** * * ** ****** **** * * ***  ** * * *
```

Non-Constrained Region

Constrained Region

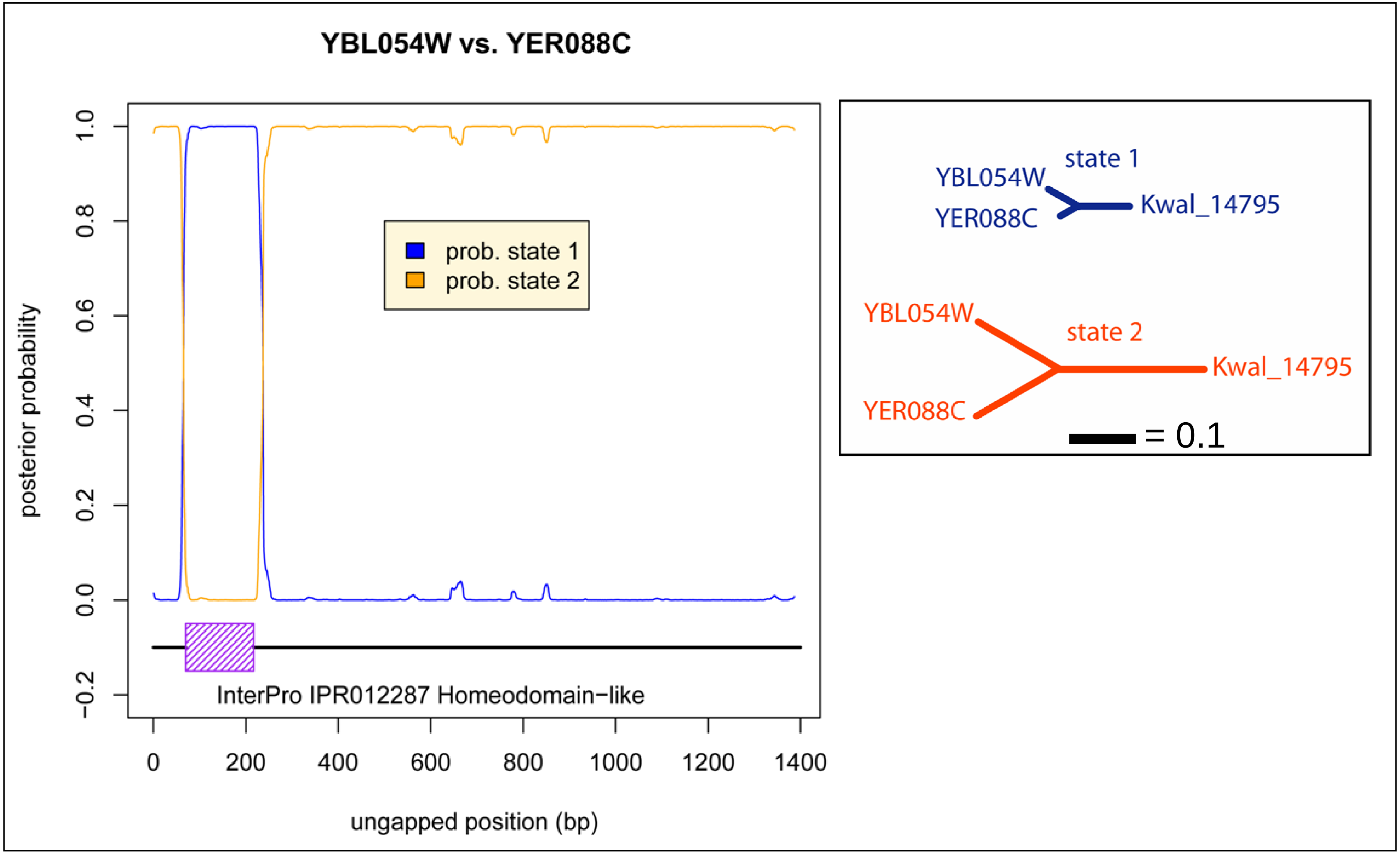
Non-Constrained Region

Phylo-HMM Conversion

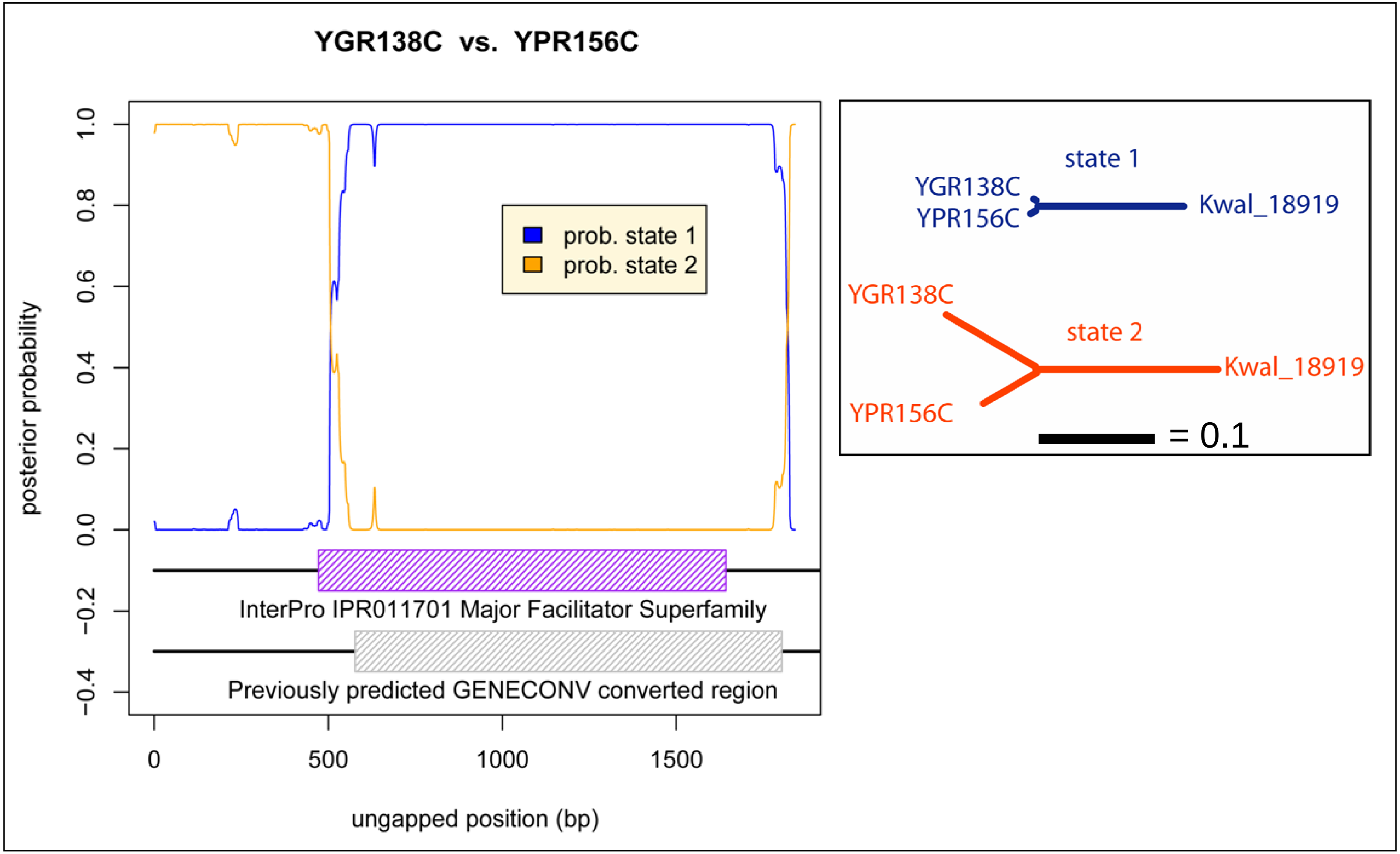


```
OG: agTgttcgaGcgggTtagcGaaaCctCcTcgAtTaTtattaccCcgtgt
D1: CgagtAcgTtcgggctagctaaatctacacggtcactatCaccgcTtgt
D2: agagttcCatcgTgctTgctaaatctacacggtcactattCcctcgtAt
     * ** *   ** * * ** *** ** * ** * * ***  ** * * *
```

Examples: YBL054W – YER088C



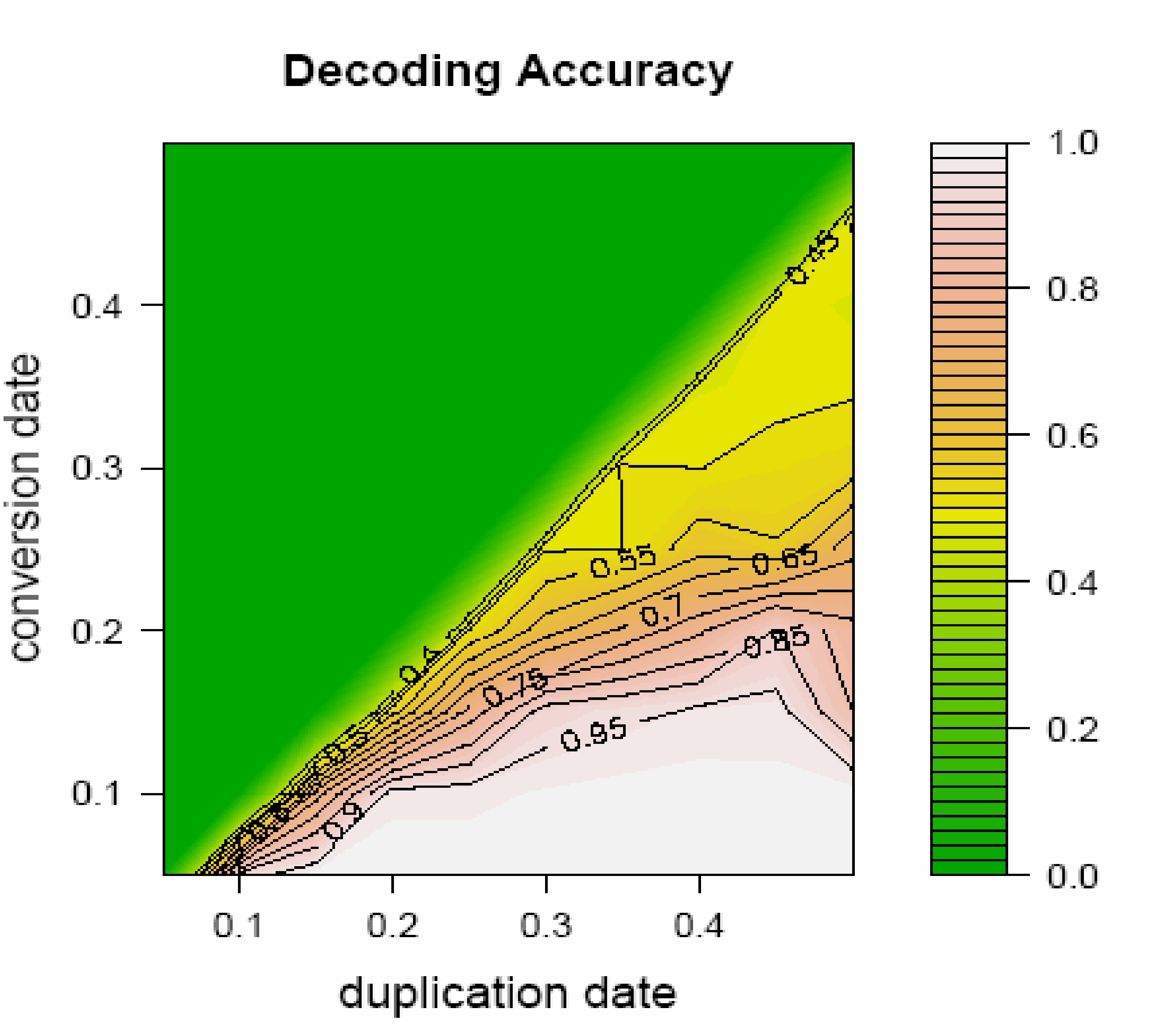
Examples: YGR138C – YPR156C



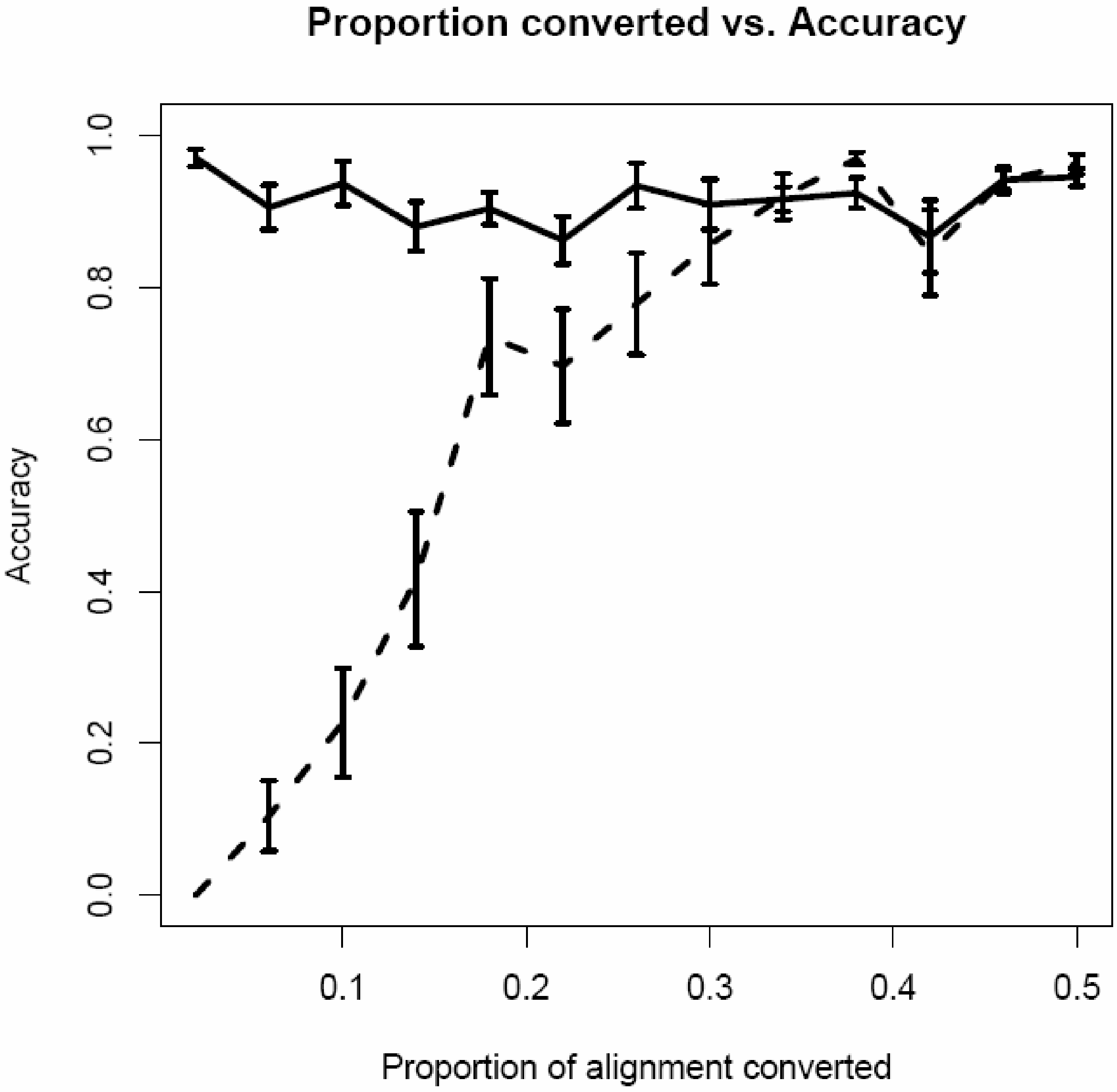
Sensitivity testing

- We tested the identifiability of gene conversions as a function of the duplication date $(\alpha t_{duplication})$ and the conversion date $(\alpha t_{conversion})$
- We used MCMC to get parameter estimates
- We then used these estimates to decode the sequence
- Finally, we determined the accuracy as the proportion of accurately called bases.

Sensitivity Example 1



Sensitivity Example 2



Converted region accuracy

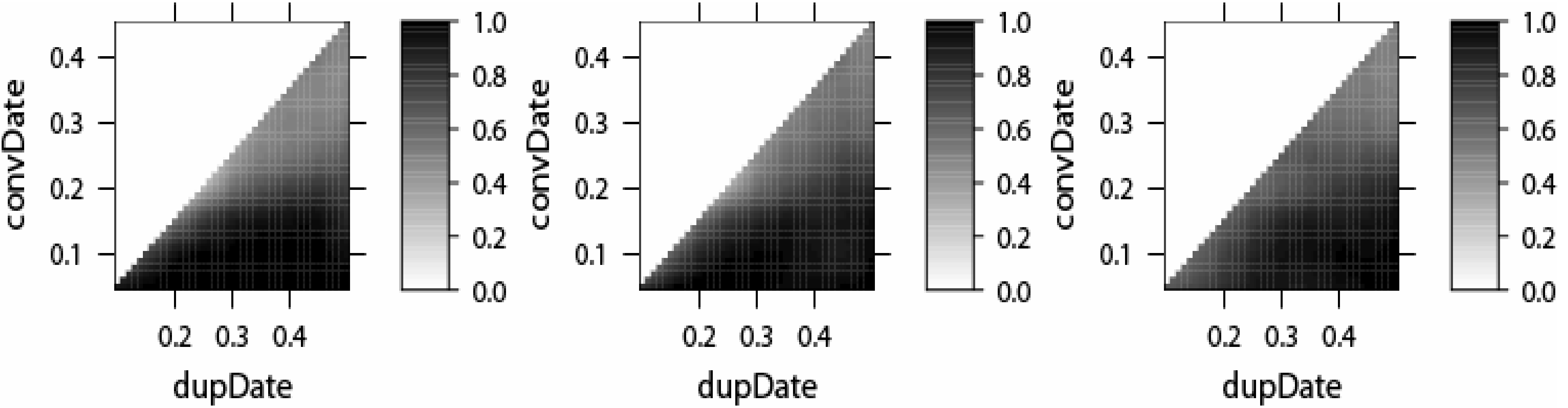
- - - - - - - -

Non-converted region accuracy

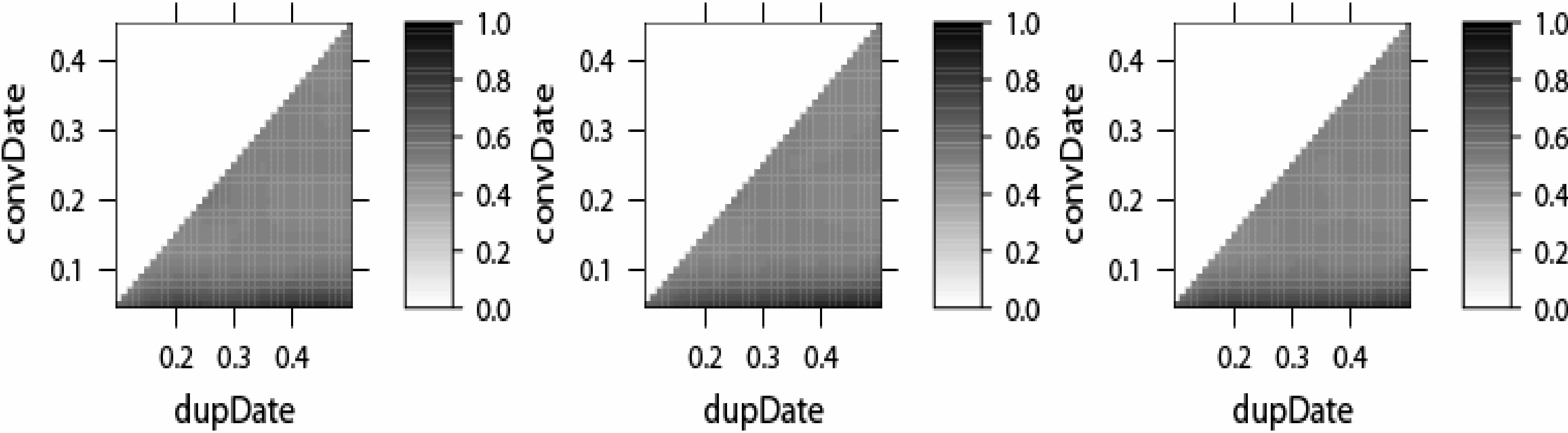
————————

Sensitivity Comparison

New method



Sawyer Method



Remarks

- Our HMM method has a higher detection power than Sawyer’s method.

- Our HMM method can define the boundaries more accurately.

Conclusion

HMM is a powerful approach to detecting gene conversion events.

Thanks!