Supplementary Note S4

Model-based estimation of the number of genes still duplicated at phylogenetic nodes

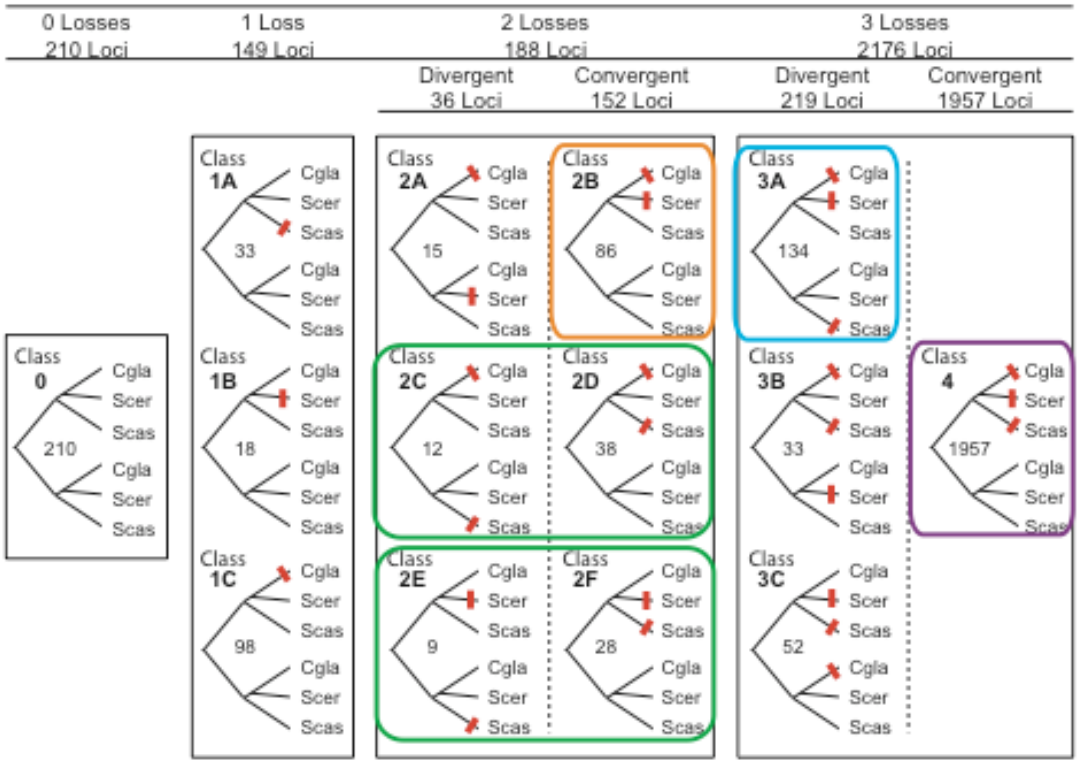


Fig S4.1 Figure 2 from the main text with certain (pairs of) classes highlighted. Green: Double loss classes where the two gene losses must have been independent. Orange: Double loss class where some losses may have occurred on a branch shared by two species (i.e., losses in the common ancestor of *S. cerevisiae* and *C. glabrata*). Blue: Triple loss class where some losses may have occurred on a branch shared by two species. Purple: Triple loss class where some losses may have occurred on a branch shared by two species, and some losses may have occurred on a branch shared by three species.

1. Estimation of the proportion of convergent losses attributable to selection

If *S. castellii* diverged from the lineage leading to *S. cerevisiae* before *C. glabrata* all the loss classes highlighted in green (Fig. S4.1) must have arisen by multiple independent losses. If this is the case, and for all losses the choice of which copy becomes lost is random, we would expect equal frequencies of 2C and 2D and also equal frequencies of 2E and 2F. This is not observed however ($P < .05$ in both cases) suggesting that selection favoured a particular copy. The proportion of ancestrally duplicated loci that are resolved either under selection or neutrally can be estimated from the frequencies of either 2C and 2D, or 2E and 2F (Table S4.1). We use ϕ to denote the proportion of duplicated loci that are resolved neutrally.

Table S4.1 Estimates of the proportion of ancestrally duplicated loci that were resolved neutrally. See the 'Equations' section below for formulae and derivation.

	Class 2 Divergent losses	Class 2 Convergent losses	ϕ
2C:2D	12	38	.480
2E:2F	9	28	.486

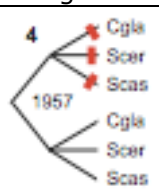
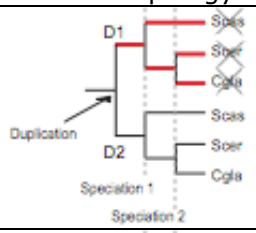
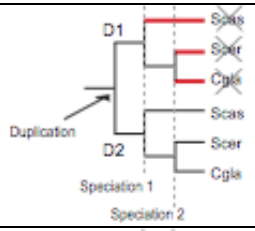
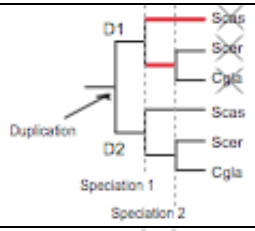
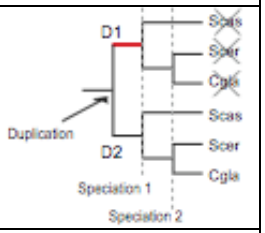
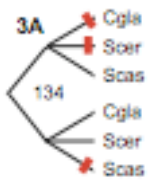
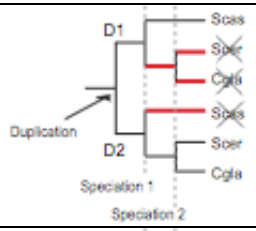
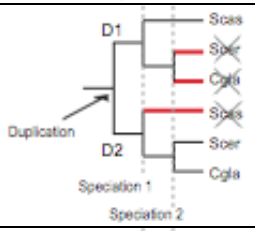
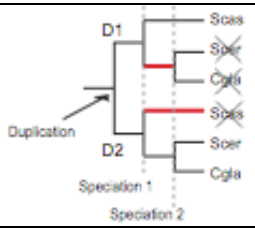
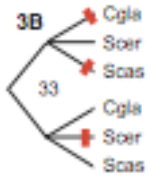
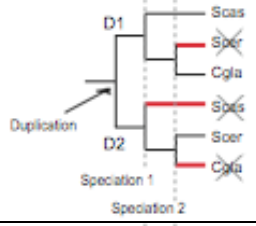
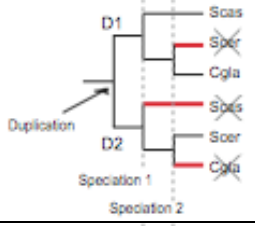
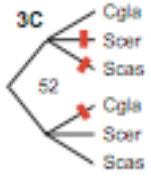
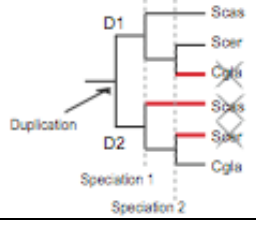
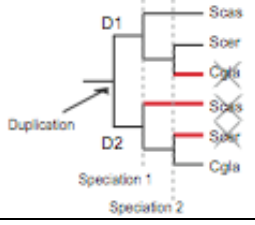
2. Estimation of the number of apparent double losses that occurred on a shared branch

Some of the losses in Class 2B (orange in Fig. S4.1) may be attributable to single losses on the shared branch leading to *S. cerevisiae* and *C. glabrata*. We can estimate the number of these by subtracting the number of convergent losses that we expect to find in Class 2B if all losses are independent, from the observed total of Classes 2A and 2B. From equation XI (in 'Equations', below) we therefore expect that of the 86 losses in Class 2B, 38.9 occurred on the shared branch leading to *S. cerevisiae* and *C. glabrata* and the remaining 47.1 were due to convergent losses after the speciation. This is calculated using equation XI as $SB_2 = 38.9 = (86+15) - 2*15/\phi$, where ϕ is estimated to be 0.483 by comparing Class 2C to 2D, and 2E to 2F (Table S4.1).

3. Estimation of the number of apparent triple losses that occurred on a shared branch either before the first speciation or before the second speciation

The process for partitioning apparent triple losses into those that occurred immediately after WGD (Speciation 0), after the first speciation (Speciation 1) or after the second speciation (Speciation 2) is identical to that just described for double losses. It is outlined in Figure S4.2 below.

Figure S4.2 Assigning convergent losses from triple loss classes (3A, 3B, 3C and 4) to time periods delimited by speciation events.

Class	Fig 2		Forced Topology		3 Losses		2 Losses		1 Loss
4		=		=		+		+	
3A		=		=		+			
3B		=		=					
3C		=		=					
Neutral Estimator					Average(3B, 3C) * 4		(3A - Average(3B, 3C)) * 2		
=> Neutral loci					170		183		
Selected Estimator					Neutral*(1 - ϕ) / ϕ		Neutral*(1 - ϕ) / ϕ		
=> Selected loci					182		196		
Total loci	2176		2176		352		379		1445
Timing (After)					Speciation 2		Speciation 1		Speciation 0

Assumptions

1) We assume that selection on copy number (whether due to dosage, neofunctionalization or subfunctionalization) and selective differences between duplicates are independent. We ignore the former.

2) Selective differences between duplicate pairs we treat as either negligible (duplicates are functionally indistinguishable; $\Delta s_{duplicates} = 0$), in which case alternative copies may be retained in different lineages, or absolute (one of the duplicates is 'superior' to the other in all lineages; $\Delta s_{duplicates} = 1$), in which case a particular copy may be lost repeatedly in independent lineages. In the former case we consider duplicates to be resolved neutrally (N in Table S4.3 below) and in the latter case to be resolved under the influence of selection (S in Table S4.3 below).

3) We assume that ϕ, the fraction of duplicate pairs for which $\Delta s_{duplicates} = 0$, is a constant.

4) We classify the pattern of loss at loci where two or more losses have occurred as convergent if all single-copy lineages have retained the same syntenic copy. If alternative copies have been retained in different lineages the pattern of loss is considered to be divergent.

Duplicate Resolution

Under the assumptions above, the total number of loci in each loss class (defined by number of losses: 0-3) is fixed, but the frequencies of subclasses may be distorted due to preferential retention of one or other copy ($\Delta s_{duplicates} = 1$). This will be observed as an excess of convergent losses over divergent losses: Compare Classes 2A and 2B, 2C and 2D, or 2E and 2F (Table S4.2).

Table S4.2. Gene loss classes, their component classes, and paired divergent/convergent subclasses.

Gene Loss Class	Total	Component classes	Divergent/Convergent Pairs
0 (no losses)	210	0	n/a
1 (one loss)	149	1A, 1B, 1C	n/a
2 (double losses)	188	2A, 2B, 2C, 2D, 2E, 2F	(2A, 2B), (2C, 2D), (2E, 2F)
3 (triple losses)	2176	3A, 3B, 3C, 4	(3A+3B+3C, 4)

Also, under the assumptions above, different paralogs may not be selectively favored in different lineages. All incidences of divergent resolution must therefore be due to neutral loss of alternative copies and SD in Table S4.3 must always be 0.

Table S4.3. Duplicate resolution and pattern of loss.

	Resolution	
Pattern of loss	Neutral (N)	Selection (S)
Convergent (C)	NC	SC
Divergent (D)	ND	SD (=0)

Note: Convergent and divergent losses are observed. Neutral resolution and resolution under selection must be inferred.

If no losses occurred on shared branches, then (where subscripts denote loss class and *n* refers to any class):

$NC_n+ND_n+SC_n+SD_n = Total_n$	*from model (assumptions 1,2,4)*	Eqn 0
$(NC_n+ND_n)/Total_n = \phi$	*by definition (assumption 3)*	Eqn 1
$D_n = ND_n$	*since $SD_n = 0$ for all n*	Eqn 2
$ND_2/NC_2 = 1$	*see Figure S3.3*	Eqn 3
$ND_3/NC_3 = 3$	*see Figure S3.4*	Eqn 4

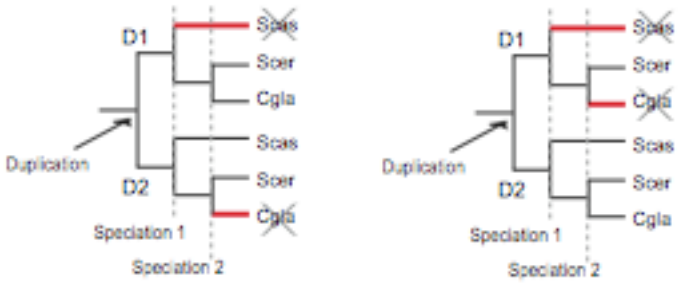


Figure S4.3. Classes 2C and 2D. These outcomes and the other pairs of convergent/divergent losses in Table S4.2 are equally likely if two random losses occur (assuming that both copies of a gene may not be lost and that there are no shared branches).

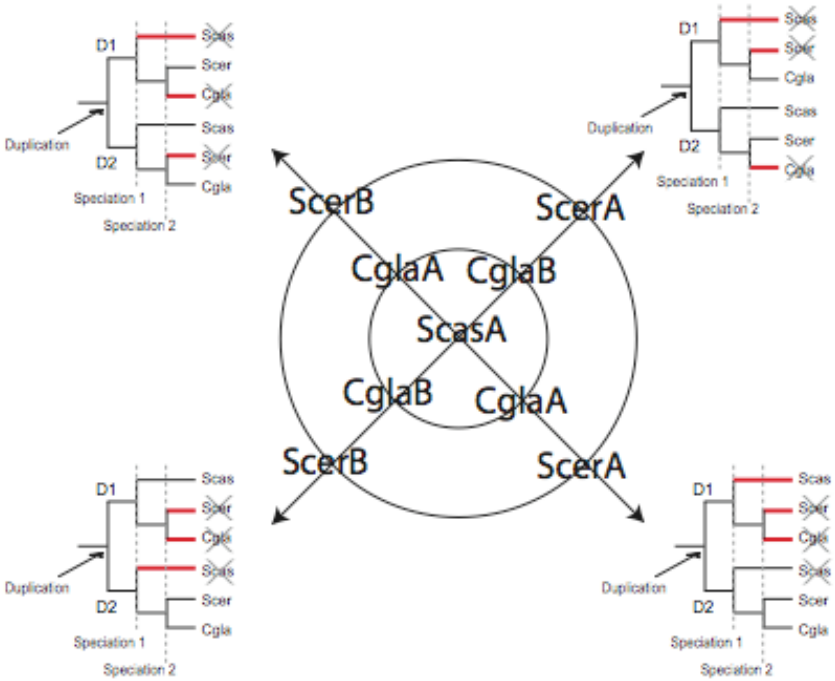


Figure S4.4. Four outcomes are equally likely if three random losses occur (assuming that both copies of a gene may not be lost and that there are no shared branches). These correspond to Classes 3A, 3B, 3C and 4.

Equations

1. ϕ from pairs of convergent/divergent double loss loci (e.g., 2E and 2F)

$(NC_n+ND_n)/(NC_n+ND_n+SC_n+SD_n) = \phi$	*from Eqns 0,1*	
$(NC_2+ND_2)/(NC_2+ND_2+SC_2+SD_2) = \phi$	*for class 2 loci*	*I*
$D_2 = ND_2 = NC_2$	*from Eqn 2 and Eqn 3*	*II*
$2*D_2 / (SC_2 + 2*D_2) = \phi$	*from I and II*	*III*
$SC_2 = C_2 - D_2$	*from $C_2 = SC_2 + NC_2$ and Eqn II*	*IV*
$\phi = 2*D_2 / (C_2 + D_2)$	*from III and IV* *ϕ in terms of observed classes*	*V*

2. Selected convergent losses from ϕ and the number of neutral divergent losses

(a) Double Loss Loci

$D_2 = ND_2 = NC_2$	*from Eqn 2 and Eqn 3*	*VI*
$SC_2 = 2*ND_2*(1 - \phi) / \phi$	*from I and II*	*VII*

(b) Triple Loss Loci

$D_3 = ND_3 = 3*NC_3$	*from Eqn 2 and Eqn 4*	*VIII*
$SC_3 = (4/3)*ND_3*(1 - \phi) / \phi$	*from I and VIII*	*IX*

3. Shared branch (SB) losses for double loss loci, assuming *S. castellii* to be the outgroup

$NC_2+ND_2+SC_2+SD_2 + SB_2 = Total_{2A+2B}$	*Eqn 0 modified*	*X*
$SC_2 = 2*ND_2*(1 - \phi) / \phi$	*from VII*	
$SB_2 =$ $Total_{2A+2B}$ –	*from X*	
D_2 –	*from II*	
D_2 –	*from II*	
$[2*D_2*(1 - \phi) / \phi]$	*from VII and II*	
$SB_2 = Total_{2A+2B} - 2*D_2 / \phi$	*SB_2 in terms of observed classes*	*XI*