Comparative Genomics as a Time Machine: How Relative Gene Dosage and Metabolic Requirements Shaped the Time-dependent Resolution of Yeast Polyploidy

Gavin C. Conant*,[1,2]

[1]Informatics Institute, University of Missouri, Columbia

[2]Division of Animal Sciences, University of Missouri, Columbia

***Corresponding author:** E-mail: conantg@missouri.edu.

Associate editor: Michael Purugganan

Abstract

Using a phylogenetic model of evolution after genome duplication (i.e., polyploidy) and 12 yeast genomes with a shared genome duplication, I show that the loss of duplicate genes after that duplication occurred in three phases. First, losses that occurred immediately after the event were biased toward genes functioning in DNA repair and organellar functions. Then, the main group of duplicate losses appear to have been shaped by a requirement to maintain balance in protein levels: There is a strong statistical association between the number of protein interactions a gene's product is involved in and its propensity to have remained in duplicate. Moreover, when duplicated genes with interactions were lost, it was more common than expected for both members of an interaction pair to have been lost on the same branch of the phylogeny. Finally, in the third phase of the resolution process, overretention of duplicated enzymes carrying high flux and of duplicated genes involved in transcriptional regulation became dominant. I speculate that initial retention of such genes by a requirement to maintain gene dosage set the stage for the later functional changes that then maintained these duplicates for long periods.

Key words: **protein interaction network, metabolic flux, gene dosage, transcriptional regulatory network, polyploidy, *Saccharomyces cerevisiae*.**

Introduction

The evolutionary potential of whole-genome duplication (WGD or polyploidy) has been widely appreciated since Ohno (1970). This potential ranges over many problems in biology, altering our understanding speciation genomics and adaptability and even of basic cell biology (Soltis and Soltis 2012). But most WGD-created duplicate genes do not persist indefinitely, making the essential question regarding polyploidy why some duplicates survive and most others do not. One of the key pieces of data for addressing this question is the fact that those surviving duplicates often have specific functions, including ones such as ribosomal proteins or transcription factors. However, smaller-scale duplications (including tandem duplications) rarely produce surviving duplicate genes from this set of functional categories (Seoighe and Wolfe 1999; Maere et al. 2005; Aury et al. 2006; Wang et al. 2011; Brenchley et al. 2012; D'Hont et al. 2012). To explain this pattern, it has been proposed that, after WGD, selection preserves duplicates that operate together in, for instance, macromolecular complexes so as to maintain the members' relative stoichiometries (Papp et al. 2003; Freeling and Thomas 2006; Freeling 2009; Veitia and Birchler 2010; Birchler and Veitia 2012). Testing this theory, however, is not trivial because selection to maintain balance is only one mechanism of duplicate preservation. Other possible mechanisms include neofunctionalization and subfunctionalization, as well as selection on features such as absolute copy number. Hence, it is very difficult to identify the particular preservation mechanism of individual duplicate gene pairs (Hahn 2009).

These problems are exacerbated by the fact that the work to date has focused on the extant genomes of polyploid species. As a result, both the different ages of the WGD events in these genomes and the different selective environments of these species confound the understanding of duplicate loss. Here, I argue that what is needed is a time-resolved description of the evolution of a polyploidy genome. Fortunately, genomes from multiple yeast species that share an ancient WGD (Goffeau et al. 1996; Wolfe and Shields 1997; Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004; Scannell et al. 2007; Gordon et al. 2011) allow us make exactly these inferences.

The yeast WGD has already enhanced our understanding of the interplay of WGD and functional evolution: There is increasing evidence that *Saccharomyces cerevisiae*'s metabolic preference for aerobic fermentation can be traced to the WGD (Blank et al. 2005; Piškur et al. 2006; Conant and Wolfe 2007; Merico et al. 2007; van Hoek and Hogeweg 2009). More generally, Scannell et al. (2007) used comparative genomics to show that post-WGD duplicate loss was rapid and coincident with speciation, leading to lineages with nearly independent histories of WGD resolution. These nonshared losses may have contributed to reproductive isolation among the respective yeasts (Scannell et al. 2006).

In this work, I take a new approach to the study of the factors driving duplicate preservation and loss after WGD. Rather than focusing on the surviving duplicates (often a small proportion of the genome), I use a phylogenetic model of duplicate resolution (Conant and Wolfe 2008), combined with multiple post-WGD yeast genomes (Gordon et al. 2011), to study the time course of the WGD resolution, seeking to understand not merely which duplicates survive, but the rules governing the order in which the previously duplicated genes were returned to single copy. Using this approach, I find that there were at least three distinct phases of WGD resolution. The very earliest losses involved a small group of genes that may find being duplicated problematic, including genes for DNA repair enzymes and for proteins targeted to the mitochondria. This phase was followed by a large number of losses that followed the predictions of the dosage balance hypothesis. Finally, there is a group of long-lived duplicates that are retained by what appear to be idiosyncratic selection pressures.

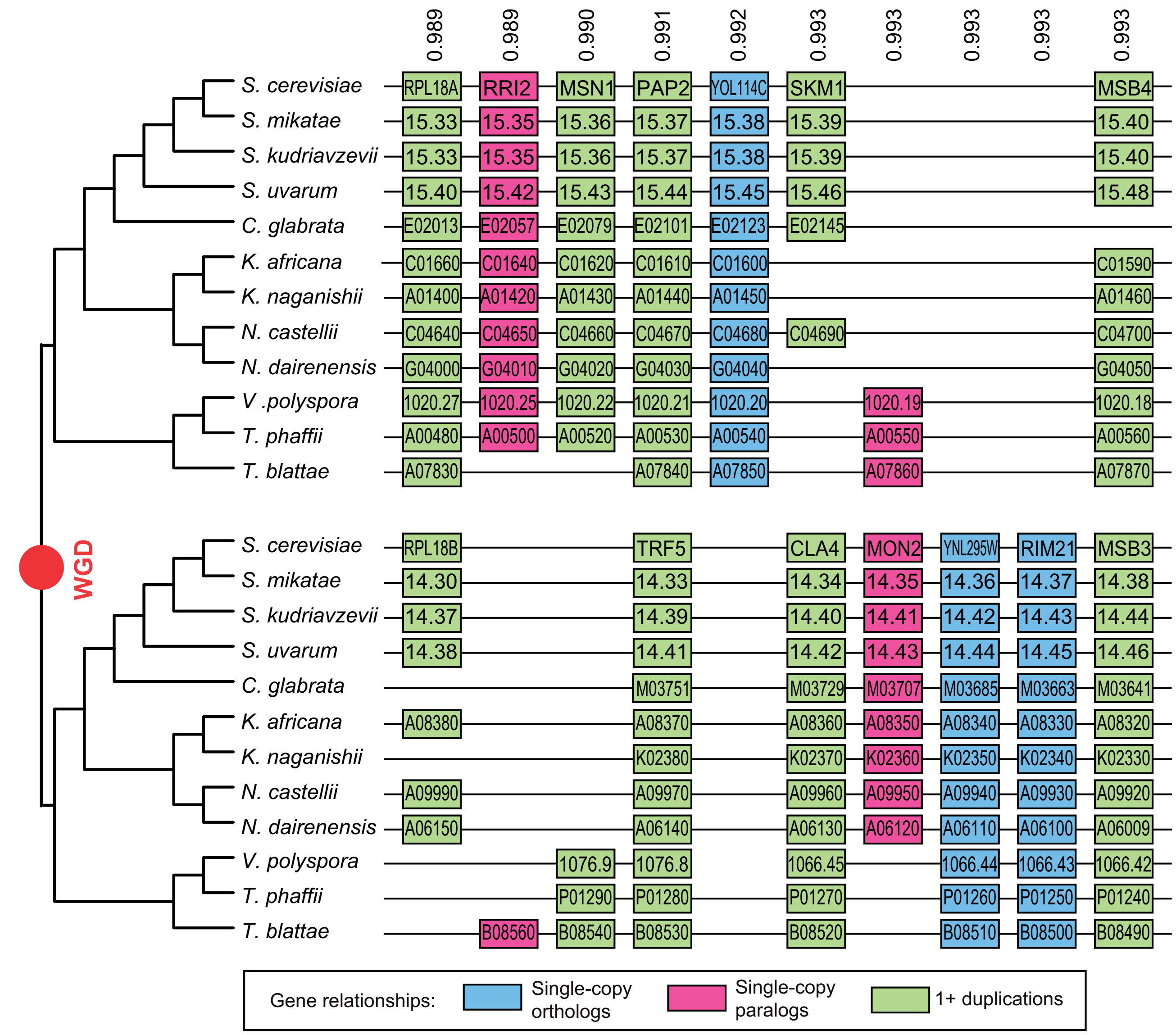
Results

Our probabilistic model of the loss of duplicate genes after WGD (POInT: Polyploidy Orthology Inference Tool; Conant and Wolfe 2008) takes as input an alignment of approximately 4,100 ordered loci (Gordon et al. 2009) in 12 genomes, each duplicated at WGD (fig. 1). Each locus in these data, which were kindly provided by the Yeast Gene Order Browser project (Byrne and Wolfe 2005), can be in one of six states (fig. 2*D*). Transition rates between the states are calculated by solving the system of differential equations implied by the instantaneous substitution rates (Materials and Methods). The calculation also, at every locus, sums over all 2^{12-1} orthology states for these 12 genomes (fig. 1). Because of this exponential scaling in problem size, I used the new Intel Phi coprocessor to make the computation feasible (Jeffers and Reinders 2013). Using POInT, I have estimated, for each ancestral locus duplicated at WGD, the probability that one of the two duplicate genes produced by the WGD was lost on each of the branches of the phylogenetic tree in figure 2*B*. The total number of inferred losses for each branch is given above it in that figure.

POInT returns the predicted history of losses, which is effectively a set of probabilistic reconstructions of the ancestral genomes at each node of figure 2*B* (e.g., fig. 2*C*). Those reconstructions allow specific tests of the dosage balance hypothesis. That hypothesis makes two predictions regarding how duplicates were lost after WGD. First, the losses should have been biased toward genes whose products were involved in fewer interactions with other genes' products. Second, to the extent genes with interactions were lost, the contemporaneous loss of that gene's duplicated interaction partners should have been more frequent. The thinking behind this prediction is that both the fully duplicated and single-copy states are assumed to be balanced and hence evolutionarily stable. There are then at least three ways in which transitions between the two stable states (duplicated and single copy) could occur. If the fitness cost of an imbalance is nontrivial but not large, the losses could occur with dynamics similar to that of the alternating mildly deleterious and mildly advantageous substitutions proposed by Charlesworth and Eyre-Walker (2007), with an initial mildly deleterious loss event being followed by the adaptive fixation of the second loss. A second mechanism might be observed were the cost of imbalance higher: The second null mutation could then need to occur while the deleterious first loss was still segregating in the population. Finally, one might have compensating increases in expression in the retained paralogs that release the dosage constraint and allow neutral losses (Scannell and Wolfe 2008). Under all three scenarios, we might expect to see more shared losses of duplicated interaction partners along a single branch of figure 2*B* than expected.

To test these predictions of the dosage balance hypothesis, I used four types of global interaction network data: Protein–protein interactions, metabolic flux, transcriptional regulatory interactions, and interactions between kinases and their targets (Materials and Methods). Starting with the protein interaction (PPI) network, for each branch of figure 2*B* leading to *S. cerevisiae*, I calculated the mean number of protein interactions for both those genes retained in duplicate along that branch (purple, fig. 2*A*) and those genes either returned to single copy along that branch, or already in single copy (blue, fig. 2*A*). Immediately after WGD, the retained duplicates begin to become enriched for genes whose products possess more protein interactions, although this trend is only statistically significant for the shared branch immediately after the split of *Vanderwaltozyma polysporus*, *Tetrapisispora phaffii*, and *T. blattae* from the other nine species (Branch "B," $P = 0.02$ after a false-discovery rate, or FDR, correction on all branches, supplementary table S1, Supplementary Material online; Benjamini and Hochberg 1995). Later, the enrichment for protein interactions among the surviving duplicates is lost (Branches C-G, $P > 0.05$, FDR correction). One might believe this latter trend is due to subfunctionalization among surviving duplicates, but in fact omitting all of the surviving WGD duplicate pairs from the analysis actually indicates a significant trend in overretention on a later branch as well (Branches B and C; $P = 0.015$, FDR vs. branches D-G; $P > 0.05$, FDR, supplementary table S1, Supplementary Material online). This trend can be seen visually in the "outward-in" pattern of duplicate loss in the extant yeast protein interaction network (e.g., taking the modern *S. cerevisiae* network topology and indicating whether or not each node had a surviving WGD-produced paralog at various times, fig. 2*C*). For that network, proteins at the periphery lost their encoded duplicate genes soon after WGD, whereas duplicate genes encoding more highly interacting proteins survived longer.

The preservation pattern differs in two regulatory networks. For the bakers' yeast's transcriptional regulatory network (Harbison et al. 2004) and its phosphoregulatory network (Ptacek et al. 2005), there is no association between the number of genes a modern transcription factor (TF) regulates or products a kinase phosphorylates and its propensity to remain duplicated ($P > 0.05$ for all branches in fig. 2*B*). However, a gene's role as a transcription factor or kinase is



	0.989	0.989	0.990	0.991	0.992	0.993	0.993	0.993	0.993	0.993
S. cerevisiae	RPL18A	RRI2	MSN1	PAP2	YOL114C	SKM1				MSB4
S. mikatae	15.33	15.35	15.36	15.37	15.38	15.39				15.40
S. kudriavzevii	15.33	15.35	15.36	15.37	15.38	15.39				15.40
S. uvarum	15.40	15.42	15.43	15.44	15.45	15.46				15.48
C. glabrata	E02013	E02057	E02079	E02101	E02123	E02145				
K. africana	C01660	C01640	C01620	C01610	C01600					C01590
K. naganishii	A01400	A01420	A01430	A01440	A01450					A01460
N. castellii	C04640	C04650	C04660	C04670	C04680	C04690				C04700
N. dairenensis	G04000	G04010	G04020	G04030	G04040					G04050
V .polyspora	1020.27	1020.25	1020.22	1020.21	1020.20		1020.19			1020.18
T. phaffii	A00480	A00500	A00520	A00530	A00540		A00550			A00560
T. blattae	A07830			A07840	A07850		A07860			A07870
S. cerevisiae	RPL18B			TRF5		CLA4	MON2	YNL295W	RIM21	MSB3
S. mikatae	14.30			14.33		14.34	14.35	14.36	14.37	14.38
S. kudriavzevii	14.37			14.39		14.40	14.41	14.42	14.43	14.44
S. uvarum	14.38			14.41		14.42	14.43	14.44	14.45	14.46
C. glabrata				M03751		M03729	M03707	M03685	M03663	M03641
K. africana	A08380			A08370		A08360	A08350	A08340	A08330	A08320
K. naganishii				K02380		K02370	K02360	K02350	K02340	K02330
N. castellii	A09990			A09970		A09960	A09950	A09940	A09930	A09920
N. dairenensis	A06150			A06140		A06130	A06120	A06110	A06100	A06009
V. polyspora			1076.9	1076.8		1066.45		1066.44	1066.43	1066.42
T. phaffii			P01290	P01280		P01270		P01260	P01250	P01240
T. blattae		B08560	B08540	B08530		B08520		B08510	B08500	B08490

FIG. 1. Orthology prediction and estimation of duplicate losses post-WGD from POInT for a small region of the 12 genomes considered. WGD produced two duplicated regions in each taxa (top and bottom panels and phylogenies): I have ordered them using Gordon et al.'s ancestral order (2009). POInT then estimates the orthology ordering of these 2*n* chromosomal regions that gives the highest likelihood. The probability of this orthology arrangement is given above each column. Blue columns indicate that the corresponding single-copy genes in the 12 genomes are inferred to be orthologous, whereas magenta columns indicate genes where some taxa have single-copy genes that are paralogs of their corresponding single-copy genes in another genome. Thus *MON2* from *Saccharomyces cerevisiae* and 1020.19 from *V. polysporus* are paralogs, although *MON2* is an ortholog of gene 14.35 from *S. mikatae*. Green columns indicate the preservation of a WGD-produced duplicate.

very relevant for its survival in duplicate post-WGD: Both classes are overrepresented among surviving duplicates along all branches except A for the transcription factors (fig. 3*D*) or A and S for the kinases (χ^2 test, $P \leq 0.01$; FDR for all comparisons). Moreover, a gene's status as a *target* of one or more transcription factors has a strong influence on retention (fig. 3*A*): The more extant *S. cerevisiae* transcription factors that regulate a gene, the longer that gene is likely to have remained duplicated (fig. 3*A*, e.g., branches B-G in fig. 3*B*, $P \leq 0.017$; FDR). The pattern of figure 3*A* is visually slightly unusual: It might appear surprising that both the duplicate and single-copy genes increase in their number of regulators moving toward the present. The reason is that initially there were many more duplicates of varying number of regulators, with only the least regulated genes having returned to single copy. As the number of duplicates fell, their average number of targets increased, while at the same time an increasing percentage of the genes with more regulators were also returned to single copy. The same visual pattern is seen for a gene's status as the target of a kinase, although it is not statistically significant in most cases (supplementary fig. S1 and table S2, Supplementary Material online, $P \geq 0.046$, FDR).

We previously described a similar pattern of long-term and differential overretention of enzymes catalyzing high flux reactions (e.g., biochemical reactions that have many metabolite turnovers per second in the cell), though without analyzing its significance (Mayfield-Jones et al. 2013). I repeated that analysis with the larger set of post-WGD

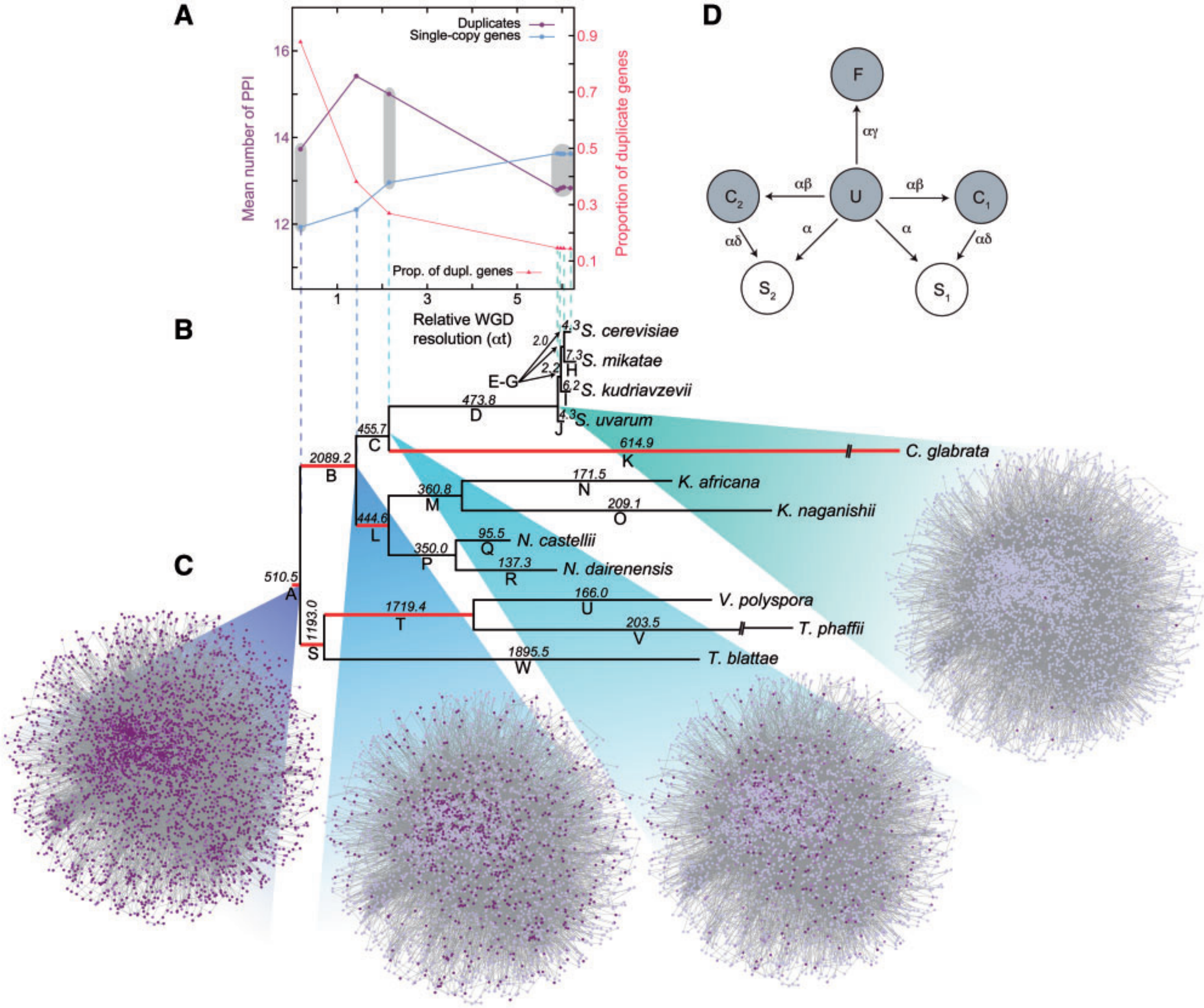


FIG. 2. Resolution of the yeast WGD was driven by relative dosage constraints in its early phases. In (panel *A*), I show the mean number of protein interactions (left axis) for the duplicated (purple) and single-copy (blue) genes at six time points after the WGD (Materials and Methods). Gray boxes around pairs of points indicate no statistical difference between the mean number of PPIs for duplicates and single-copy genes after FDR correction (Results). On the *x* axis is the relative proportion of duplicate losses at that point (branch lengths from the POInT model or αt, panel *B*). On the right axis is the number of duplicates produced by WGD surviving to that point (pink). (Panel *B*) gives the relative degree of duplicate resolution for each of these six points (all on the lineage leading to *Saccharomyces cerevisiae*) and shows the phylogenetic comparisons used to make the inferences. Branch names are given below each branch: Above the branches are POInT's estimates of the number of duplicate loci returned to single copy along that branch. Branches shown in red indicate a statistical excess of shared losses of both members of a PPI pair (Materials and Methods). Because branch lengths are proportional to the probability of a loss along each branch, more recent branches have fewer total losses because most genes were already single copy at that point. (Panel *C*) shows my reconstruction of the copy-number condition of the extant yeast protein interaction network at four of the six points in (*A*) (i.e., the network topology is fixed to the extant *S. cerevisiae* network, with nodes colored based on their inferred copy number). For clarity, I have omitted all surviving duplicates from the WGD. Again, shortly after WGD, most modern single-copy genes were duplicated (purple): Over time, losses started at the periphery of the network and moved inwards, producing single-copy genes (blue). Finally, (panel *D*) gives a diagrammatic description of the likelihood model of post-WGD gene losses. All genes start in state **U**: They may then become fixed (**F**) at rate $\alpha\gamma$, lost ($\mathbf{S_1}$ and $\mathbf{S_2}$) at rate α or move to a converging state ($\mathbf{C_1}$ and C$_2$) at rate $\alpha\beta$, after which only one type of loss is permitted (rate $\alpha\delta$, see Materials and Methods).

genomes considered here. There is a trend toward increasing retention of genes encoding high-flux enzymes through time. However, this trend is not universal: Only in the *Saccharomyces* clade is the enrichment significant (blue oval in fig. 3C, $P = 0.026$, FDR).

The second prediction of the dosage balance hypothesis is that, when interacting genes are lost post-WGD, it should be more likely that both members of an interaction pair are lost contemporaneously (e.g., on the same branch of the phylogeny). I took the extant yeast protein interaction network and removed all surviving WGD duplicate pairs. For the remaining single-copy genes, I took each interacting pair and asked, for each branch, the probability that both members of that interaction pair were lost along that same branch. I then summed this probability over the entire network. To assess whether there were more such shared losses than would

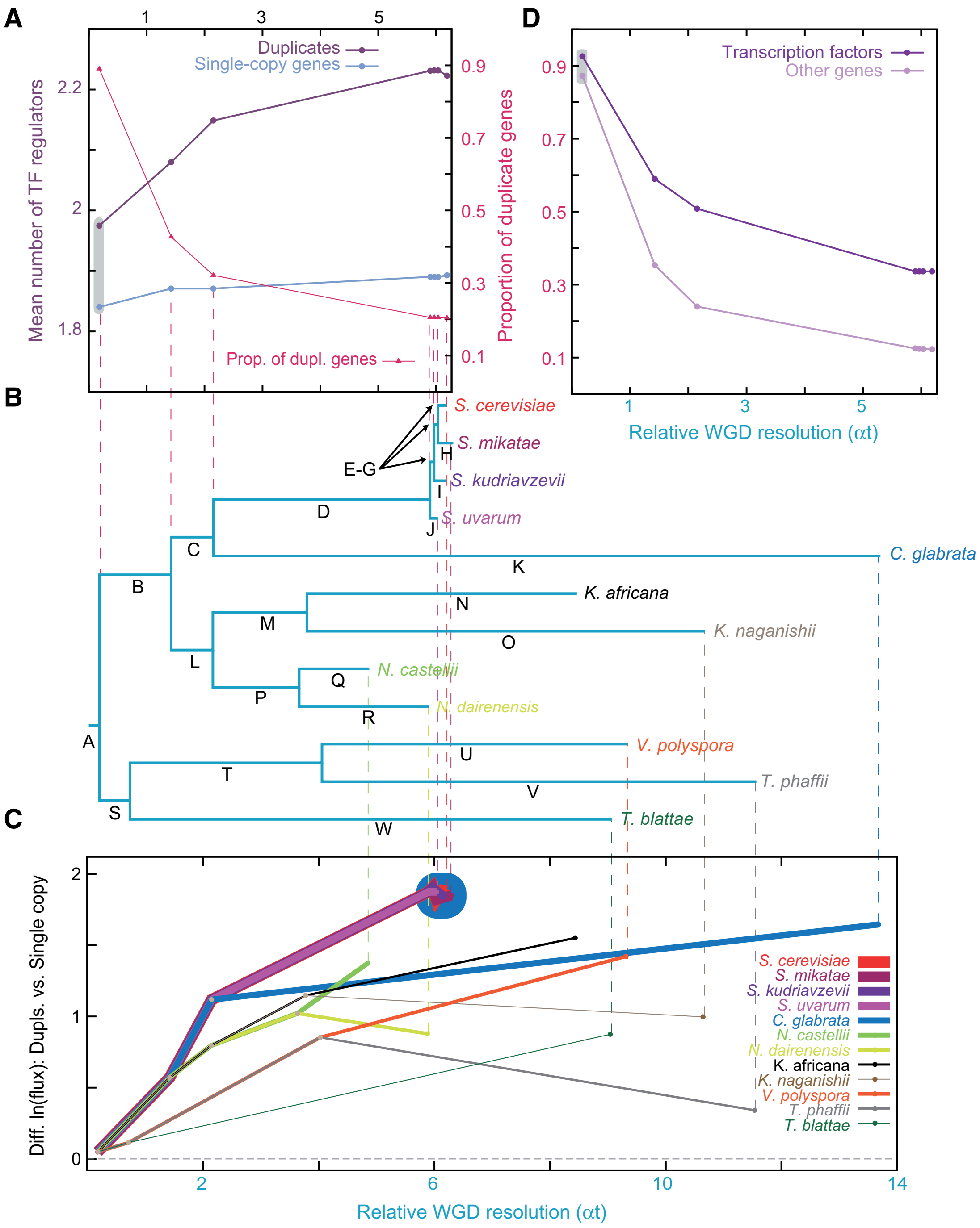


Fig. 3. Later phases of the yeast WGD saw duplicate gene preservation due to several forces. In (panel *A*), the mean number of transcription factors regulating the duplicated (purple) and single-copy (blue) genes at each of six time points are shown (left axis). Note that these regulatory counts are based on the extant *Saccharomyces cerevisiae* network and hence not adjusted for the instantaneous redundancy created by WGD. Gray boxes indicate nonsignificant comparisons ($P > 0.05$, FDR). The *x* axis and right axis are as for figure 2*A*. (Panel *B*) is as for figure 2. In (panel *C*) is the association between maximal flux for each enzyme (estimated with flux-balance analysis) in the *S. cerevisiae* metabolic network and its propensity to remain duplicated in each taxa (Materials and Methods). On the *y* axis is the mean difference in the log-transformed flux between the duplicated and single-copy genes. Thus, a value of 0 (dashed line) indicates no difference, values greater than zero indicate that duplicated genes on average carry higher flux. The blue oval indicates the cases where there is a significant difference between duplicated and single-copy genes ($P = 0.026$, FDR; Results). The *x* axis is as in (*A*). (Panel *D*) confirms that transcription factors (dark purple) themselves were overretained in duplicate after the WGD throughout the *S. cerevisiae* lineage relative to the remainder of the genome (light purple).

expected by chance, I randomized the network 1,000 times and recomputed this same shared loss statistic using the original branch loss patterns but randomized networks (Materials and Methods). For six of the earliest branches in the phylogeny, I observe more cases of shared loss than can be explained by chance (red branches in fig. 2*B*, $P \leq 0.015$; FDR). No such excess of shared losses is observed in regulatory networks ($P > 0.05$; FDR). Probably the fact that transcription factors interact with the DNA coding for a gene rather than with the gene product gives rise to this difference: I have previously argued that a significant degree of transcriptional redundancy survives among the duplicate TFs created by the WGD (Conant 2010).

Figures 2 and 3 suggest that in the very earliest stages of the WGD resolution, many of the trends just described were either not yet established, or at a minimum, are not statistically detectable. Part of the reason is that relatively few losses had occurred by this point: The model predicts that only roughly 511 genes had returned to single copy by the split between *S. cerevisiae* and *V. polysporus* (fig. 2*B*—the fractional values indicate statistical uncertainty in when on the phylogeny each gene was lost). I thus selected the 903 genes from the model with a probability > 0.4 of having been lost along branch A (Materials and Methods). Two of the eight most common GO terms among this group were significantly overrepresented compared with other genes: "mitochondrion organization" and "DNA repair" (Materials and Methods; $P \leq 0.045$, χ^2 tests with FDR correction).

Clearly, several factors are predictive of duplicate loss. However, these factors themselves might be related. For instance, transcriptional factors often have multiple protein interactions: Is it the function of transcriptional regulation or the number of interactions that drives retention? For each branch in figure 2*B*, I asked what was the R^2 of the association between the ratio of the probability of returning to single copy ($P^{\mathrm{DUPL} \rightarrow \mathrm{SING}}$) and the probability of having remained duplicated (P^{DUPL}) along that branch and its 1) protein interaction degree, 2) metabolic flux, and 3) number of transcriptional regulators. These R^2 values are generally not high, but given the complex distributions of these variables, that fact is not unexpected (supplementary table S3, Supplementary Material online). What is more striking is that the relative order of the ranked R^2 values changes over time: On early branches, PPI degree is the best predictor of loss probability, while for most extant branches metabolic flux dominates.

Discussion

Analyses of the temporal pattern of duplicate gene losses identify at least three phases of WGD resolution. There are multiple lines of evidence for this conclusion: Figures 2 and 3 strongly suggest that the dosage-based trends were not fully established by the first speciation event (neither overretention of highly interacting genes nor of transcription factors is statistically distinguishable by this point). Similarly, the overretention of highly interacting genes and the shared losses of interacting partners (fig. 2*B*) were clearly transient and are not observable in the modern *S. cerevisiae* genome. Likewise, selection on enzymatic flux does not become statistically significant until after dosage-based selection on interaction degree is no longer detectable. Finally, our correlation analyses of the various factors' association with loss rates also support differential timing in these patterns.

So what are these three phases? First, there was a group of early gene losses. Although these were few in number, they included genes involved in mitochondrial organization. The analysis of GO terms indicates that the early losses are distinct from later losses: Certainly the terms mitochondrion organization and DNA repair are both strongly underrepresented among the surviving WGD duplicates in *S. cerevisiae* ($P < 10^{-10}$, χ^2 test). It has previously been suggested (Edger and Pires 2009) that genes whose products are targeted to the organelles experience different dosage constraints post-WGD than other genes, as organellar genomes are not altered by WGD. It is therefore gratifying to detect this early trend toward a return to "normal" organellar dosage. Our results are also consistent with those of De Smet et al. (2013), who found that genes involved in organellar processes and DNA damage repair were single copy in many plant genomes, despite the genomes in question having undergone many nested WGD events.

In the second phase of resolution, maintenance of relative dosage was clearly key, as indicated by both the overretention of genes for interacting proteins and the subsequent shared losses of these interactors. However, selection to maintain dosage becomes less evident as time progresses (fig. 2*A* and *B*). Instead, as we previously argued for plants (Bekaert et al. 2011), the species eventually entered a final resolution phase where a smaller group of retained duplicates began to alter the organisms' phenotypes. Of course, it may be that the maintenance of duplicates that initially occurred due to reasons of dosage then allowed for subsequent functional changes (Hittinger and Carroll 2007; Komili et al. 2007; Kim et al. 2009; Evangelisti and Conant 2010; Fusco et al. 2010), including further evidence for the importance of WGD in reshaping metabolism in bakers' yeast (Blank et al. 2005; Conant and Wolfe 2007; Merico et al. 2007; van Hoek and Hogeweg 2009). One might also speculate that genes such as transcription factors, while retained initially due to relative dosage, are particularly prone to subfunctionalization, accounting for their longer survival (Birchler and Veitia 2012). On the other hand, there are clear cases of post-WGD neofunctionalization as well (Conant et al. 2014).

Of course, the notion of "phases" is an oversimplification: Dosage constraints will not suddenly cease to act at a certain point in time, nor will all the duplicate pairs have responded to the same selection pressures or be expected to have the same fate should a subsequent WGD occur (Conant et al. 2014). Instead, I prefer to think of the three phases as describing what one might think of as the main force acting at different points in time after WGD. Similarly, my analyses only consider the set of approximately 4,500 genes identified in syntenic positions in all 12 yeast genomes (Gordon et al. 2009)—it is at least possible that the faster-evolving genes that do not show sufficient synteny and sequence conservation to be placed in this framework might show different patterns. One final caveat is that it would be preferable to actually reconstruct the ancestral regulatory and protein interaction networks for these species rather than using the extant networks from *S. cerevisiae*. Unfortunately, we completely lack the sort of large-scale comparative data necessary to do so: The state of the art in this area is the comparison of the interaction patterns of a handful of proteins across species (Qian et al. 2011). However, here I have considered the aggregate network properties of each gene—a protein's interaction degree or a transcription factor's number of targets will be much more slowly evolving than is any particular interaction.

The complex temporal dynamics of post-WGD evolution help illuminate some of the mysteries and controversies surrounding polyploidy evolution. For instance, there is some danger in naively comparing WGD events of different ages, because they may have very different retention patterns (e.g., an overabundance of genes with high interaction degree early on vs. no such overabundance later). It may be similarly dangerous to make claims about whether particular features are important in WGD resolution when considering only extant genomes: The role of protein interactions in retaining duplicates would not be detected in the extant *S. cerevisiae* genome (Zhu et al. 2013).

More generally, if we accept the power of polyploidy for innovation (Fawcett et al. 2009), it is surprising that it may be followed by a reduction in speciation rates (Mayrose et al. 2011). These potentially conflicting facts have reconciled by the suggestion of a post-WGD speciation lag (Schranz et al. 2012): One source of this lag is the dosage-balance phase of WGD resolution. It is possible that it is only after dosage has been stabilized that the power of WGD becomes manifest (Arrigo and Barker 2012). At that point there will be a set of duplicates with functions especially suited as substrates of innovation: They are more likely to be essential (Conant and Wolfe 2008), have many epistatic constraints that can be resolved by subfunctionalization (Scannell and Wolfe 2008) and are involved in critical regulatory and metabolic functions (fig. 3). We should not dismiss the innovative potential of WGD by focusing on its immediate effects—the power of these events may well be felt many millions of years after the initial doubling (Van de Peer et al. 2009; Soltis and Soltis 2012).

Materials and Methods

Modeling Gene Loss Post-WGD

Our probabilistic model takes as input the chromosomal gene orders from 12 extant genomes that all share a WGD (Goffeau et al. 1996; Cliften et al. 2003; Dujon et al. 2004; Scannell et al. 2007; Gordon et al. 2011). POInT then orders these genes relative to each other using an inferred ancestral order, presumed to reflect the gene order immediately before the WGD. In yeast we are fortunate to have a high-quality estimate of this gene order, provided by Gordon et al. (2009).

The model uses six states (fig. 2*D*) to describe the process of post-WGD duplicate loss. **U** represents undifferentiated duplicated genes that are free to be lost, state **F** represents fixed duplicate genes and states $\mathbf{S_1}$ and $\mathbf{S_2}$ are the two single-copy states. Finally, states $\mathbf{C_1}$ and $\mathbf{C_2}$ are "partisan" states where even though a duplicate pair is retained at that locus, the two duplicates have differentiated, such that for $\mathbf{C_1}$ only copy 2 can be lost and hence only state $\mathbf{S_1}$ is reachable and vice versa. Under this model, we can generate a set of instantaneous transition rates among the six states:

$$\begin{aligned} R(\mathbf{U} \rightarrow \mathbf{S_1}) &= \alpha \\ R(\mathbf{U} \rightarrow \mathbf{F}) &= \alpha \cdot \gamma \\ R(\mathbf{U} \rightarrow \mathbf{C_1}) &= \alpha \cdot \beta \\ R(\mathbf{C_1} \rightarrow \mathbf{S_1}) &= \alpha \cdot \delta \\ R(\mathbf{C_1} \rightarrow \mathbf{F}) &= 0 \end{aligned} \qquad (1)$$

We then use standard techniques to compute the model's time-dependent substitution probabilities (Lewis 2001).

The resulting likelihood model is uniquely appropriate to understanding post-WGD evolution for two reasons. First, it makes the inference of the loss pattern part of the same computation as inferences regarding the orthology of the various single-copy genes in genomes sharing the WGD. Figure 1 illustrates this problem: Pink columns are single-copy genes that were differentially resolved: For example, the *V. polysporus* homolog of *S. cerevisiae* gene *MON2* (gene 1020.19) is actually a paralog created by the WGD and not an ortholog from the more recent common ancestor of *S. cerevisiae* and *V. polysporus*. We previously found that only approximately 56% of shared, single-copy, genes in *S. cerevisiae* and *V. polysporus* are orthologs (Scannell et al. 2007). To address this issue, the model considers the set of 2^{n-1} possible orthology relationships between the 2*n* different loci (e.g., two duplicated loci in each of *n* genomes). We sum over these 2^{n-1} combinations for each locus, accounting for the potential uncertainty in whether a pair of single-copy genes in two different genomes are paralogs or orthologs. As this computation scales as 2^{n-1}, I used OpenMP (Dagum and Menon 1998) and the new Intel Phi coprocessor (Jeffers and Reinders 2013) to parallelize and accelerate the computation. Running times were between 3 and 13 days on one of these devices.

Depending on the state of the 12 genomes, different orthology assignments will be more or less probable: For a shared single-copy gene, a single common loss will be more likely than multiple independent losses. However, loci are not treated as independent: Using the hidden-Markov approach of Felsenstein and Churchill (1996), the likelihood of site *i*+1 having orthology state *j* given that site *i* has that orthology assignment is (1-θ), where θ is a small constant estimated from data ($\theta = 0.002$ for this analysis). In cases where there is a break in gene order (e.g., a new contig or chromosome in an extant genome with respect to two adjacent genes in the ancestral order), $\theta = 0.5$.

The second advantage of POInT is that it places the losses of duplicated genes onto a phylogeny, allowing the inference of when each loss event occurred. This placement would not be possible without the orthology assignments, because one can easily mistake the loss of alternative copies of a duplicate pair in different lineages for a common loss. Using a likelihood framework is also important, because homoplasy (the independent loss of the same duplicated gene in two lineages) is common, making parsimony-based inferences problematic.

Inferring Ancestral Duplication Status

Using the model, it is straight forward to infer probabilistic estimates of the state of each duplicated locus at each internal node of the tree: Effectively these values are simply the conditional probabilities that form part of the tree-transversal algorithm used to compute the tree likelihood (Felsenstein 1981). There are three cases that are relevant for these analyses: The probability that a locus remained duplicated from the start of the branch to the current node (P^{DUPL}, combining states **U**, **F** and $\mathbf{C_1}$ and $\mathbf{C_2}$ in fig. 2*D*), the probability that it transitioned from duplicated to single-copy state ($P^{\mathrm{DUPL}\rightarrow\mathrm{SING}}$, e.g., the gene entered state $\mathbf{S_1}$ or $\mathbf{S_2}$ from state **U**, $\mathbf{C_1}$, or $\mathbf{C_2}$ along that branch) and the probability that it started the branch in single-copy and remained as such (P^{SING}, e.g., the gene was in state $\mathbf{S_1}$ or $\mathbf{S_2}$ at the start of the branch). These three probabilities necessarily sum to 1.0 for a given locus and branch.

Network Data and Randomization

To assess the role of network position in shaping the patterns of gene loss, I employed four different biological networks:

1) A protein interaction network from release 3.1.80 of BioGRID (Stark et al. 2011),
2) Estimates of metabolic flux computed as previously described (Pérez-Bercoff et al. 2011) for the metabolic network of Duarte et al. (2004),
3) The transcriptional regulatory network of Harbison et al. (2004) and
4) The phosphoregulatory network of Ptacek et al. (2005).

For each network, one can compute the various network statistics (see Results), using the three probabilities above. For instance, if a particular gene has PPI degree 2 and has $P^{\mathrm{DUPL}} = 0.5$ and $P^{\mathrm{DUPL}\rightarrow\mathrm{SING}} = 0.5$, I added degree 2 to each total with probability 0.5. As this description implies, I have only considered each protein/transcription factor/kinase's modern behavior: I did not, for instance, correct for the potential 4-fold redundancy in transcription factor binding induced by the WGD. Instead, in the case of extant duplicates from the WGD, I have taken the mean of the duplicates' interaction degree or number of regulators. Although it would be possible instead to correct for redundancy immediately after the WGD, computing the decay rate of that redundancy would be challenging. Using the modern networks at least allows me to distinguish between the duplicate survival patterns for different network positions.

In order to assess if there were more shared losses than expected of two members of a PPI pair, a transcription factor and its target or a kinase and its target on a single branch, I used our previously described network randomization approach. This approach exactly preserves the interaction degree of all network nodes, whereas randomizing their interaction partners (Pérez-Bercoff et al. 2011). This fact means that the overretention patterns seen in figure 2*A* will not bias this analysis. I then compared the number of shared losses (shared cases of $P^{\mathrm{DUPL}\rightarrow\mathrm{SING}}$ for both nodes) for the real genomes to that seen with the randomized networks.

Early WGD Losses and GO Term Analyses

I sought the set of genes lost along branch A in figure 2*B*. No duplicates were inferred to have been lost along branch A with greater than 47% confidence. The reason for this low probability is that the model allows the early transition of an ancestral locus to a converging state ($\mathbf{C_1}$ or $\mathbf{C_2}$ in fig. 2*D*). At that point, any later loss in any lineage will remove the same copy. It is thus difficult to distinguish an early loss from an early entry into this converging state. One might think that this problem could be overcome by removing the requirement for these converging states. However, doing so significantly reduces the quality of the model fit to the data ($\beta = 0$ and $\delta = 0$ in fig. 2*D*, $P < 10^{-50}$, likelihood ratio test with 2 degrees of freedom). I thus used the set of genes with a probability of loss along branch A of > 0.4 (Results).

I compared the frequency of the eight GO SLIM terms from Saccharomyces Genome Database (Cherry et al. 1998) that were common to more than 5% of these 903 genes to the frequency of those same terms among the remaining 3,668 GO-annotated genes not predicted to have been lost along branch A using a chi-square test with one degree of freedom.

Correlation of Loss Probabilities and Network Statistics

To compute the statistical association of the probability that a gene was lost along a branch and various network statistics, for every duplicated locus I computed the association between that statistic and the ratio $P^{\mathrm{DUPL}\rightarrow\mathrm{SING}}/P^{\mathrm{DUPL}}$. This ratio only indicates the probability that a duplicate was lost relative to the probability that it remained duplicated. In many cases, both of these quantities will be small because the most probable state for that locus along that branch is having already returned to single copy (P^{SING}). To account for this issue, I used the weighted regression package in R (R Development Core Team 2008), weighing each locus by the probability that it was not in single copy (1.0-P^{SING}).

Supplementary Material

Supplementary figure S1 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The author would like to thank K. Byrne, J. Gordon and K. Wolfe for assistance with YGOB and access to genomic data, without which this project would have been impossible. The author also appreciates helpful discussions with J. Birchler, P. Edger, M. Freeling, and J. C. Pires and assistance with developing parallel versions of these codes and access to computational resources from M. Becchi. This work was supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri.

References

Arrigo N, Barker MS. 2012. Rarely successful polyploids and their legacy in plant genomes. *Curr Opin Plant Biol.* 15:140–146.

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178.

Bekaert M, Edger PP, Pires JC, Conant GC. 2011. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative followed by absolute dosage constraints. *Plant Cell* 23: 1719–1728.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat.* 57:289–300.

Birchler JA, Veitia RA. 2012. Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. *Proc Natl Acad Sci U S A.* 109:14746–14753.

Blank LM, Lehmbeck F, Sauer U. 2005. Metabolic-flux and network analysis of fourteen hemiascomycetous yeasts. *FEMS Yeast Res.* 5: 545–558.

Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–710.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15:1456–1461.

Charlesworth J, Eyre-Walker A. 2007. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. *Proc Natl Acad Sci U S A.* 104:16992–16997.

Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, et al. 1998. SGD: *Saccharomyces* genome database. *Nucleic Acids Res.* 26:73–80.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA. 2003. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71–76.

Conant GC. 2010. Rapid reorganization of the transcriptional regulatory network after genome duplication in yeast. *Proc Biol Sci.* 277: 869–876.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr Opin Plant Biol.* 19:91–98.

Conant GC, Wolfe KH. 2007. Increased glycolytic flux as an outcome of whole-genome duplication in yeast. *Mol Syst Biol.* 3:129.

Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* 179:1681–1692.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488:213–217.

Dagum L, Menon R. 1998. OpenMP: an industry standard API for shared-memory programming. *IEEE Comput Sci Eng.* 5:46–55.

De Smet R, Adams KL, Vandepoele K, Van Montagu MC, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc Natl Acad Sci U S A.* 110:2898–2903.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, et al. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 304:304–307.

Duarte NC, Herrgård MJ, Palsson BØ. 2004. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* 14:1298–1309.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuveglise C, Talla E, et al. 2004. Genome evolution in yeasts. *Nature* 430:35–44.

Edger PP, Pires JC. 2009. Gene and genome duplications: the impact of dosage sensitivity on the fate of nuclear genes. *Chromosome Res.* 17: 699–717.

Evangelisti AM, Conant GC. 2010. Nonrandom survival of gene conversions among yeast ribosomal proteins duplicated through genome doubling. *Genome Biol Evol.* 2:826–834.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A.* 106:5737–5742.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 17:368–376.

Felsenstein J, Churchill GA. 1996. A hidden markov model approach to variation among sites in rate of evolution. *Mol Biol Evol.* 13: 93–104.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol.* 60:433–453.

Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* 16:805–814.

Fusco D, Grassi L, Bassetti B, Caselle M, Cosentino Lagomarsino M. 2010. Ordered structure of the transcription network inherited from the yeast whole-genome duplication. *BMC Syst Biol.* 4:77.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, et al. 1996. Life with 6000 genes. *Science* 274:546,563–567.

Gordon JL, Armisen D, Proux-Wera E, OhEigeartaigh SS, Byrne KP, Wolfe KH. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proc Natl Acad Sci U S A.* 108: 20024–20029.

Gordon JL, Byrne KP, Wolfe KH. 2009. Additions, losses and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet.* 5: e1000485.

Hahn MW. 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. *J Hered.* 100:605–617.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449:677–681.

Jeffers J, Reinders J. 2013. Intel Xeon Phi coprocessor high performance programming. Waltham, MA: Morgan Kaufmann.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–624.

Kim T-Y, Ha CW, Huh W-K. 2009. Differential subcellular localization of ribosomal protein L7 paralogs in *Saccharomyces cerevisiae*. *Mol Cells.* 27:539–546.

Komili S, Farny NG, Roth FP, Silver PA. 2007. Functional specificity among ribosomal proteins regulates gene expression. *Cell* 131: 557–571.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst Biol.* 50:913–925.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102:5454–5459.

Mayfield-Jones D, Washburn JD, Arias T, Edger PP, Pires JC, Conant GC. 2013. Watching the grin fade: tracing the effects of polyploidy on different evolutionary time scales. *Semin Cell Dev Biol.* 24: 320–331.

Mayrose I, Zhan SH, Rothfels CJ, Magnuson-Ford K, Barker MS, Rieseberg LH, Otto SP. 2011. Recently formed polyploid plants diversify at lower rates. *Science* 333:1257.

Merico A, Sulo P, Piškur J, Compagno C. 2007. Fermentative lifestyle in yeasts belonging to the *Saccharomyces* complex. *FEBS J.* 274:976–989.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.

Papp B, Pal C, Hurst LD. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.

Pérez-Bercoff.Å, McLysaght A, Conant GC. 2011. Patterns of indirect protein interactions suggest a spatial organization to metabolism. *Mol Biosyst.* 7:3056–3064.

Piškur J, Rozpedowska E, Polakova S, Merico A, Compagno C. 2006. How did *Saccharomyces* evolve to become a good brewer? *Trends Genet.* 22:183–186.

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, et al. 2005. Global analysis of protein phosphorylation in yeast. *Nature* 438:679–684.

Qian W, He X, Chan E, Xu H, Zhang J. 2011. Measuring the evolutionary rate of protein-protein interaction. *Proc Natl Acad Sci U S A.* 108: 8725–8730.

R Development Core Team. 2008. R: a language and environment for statistical computing: R Foundation for Statistical Computing. Available from: http://cran.case.edu/web/packages/dplR/vignettes/timeseries-dplR.pdf.

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440:341–345.

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci U S A.* 104:8397–8402.

Scannell DR, Wolfe KH. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.* 18:137–147.

Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr Opin Plant Biol.* 15:147–153.

Seoighe C, Wolfe KH. 1999. Yeast genome evolution in the post-genome era. *Curr Opin Microbiol.* 2:548–554.

Soltis PS, Soltis DE. 2012. Polyploidy and genome evolution. Heidelberg: Springer. p. 415.

Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, et al. 2011. The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.* 39:D698–D704.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nat Rev Genet.* 10:725–732.

van Hoek MJ, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol.* 26:2441–2453.

Veitia RA, Birchler JA. 2010. Dominance and gene dosage balance in health and disease: why levels matter!. *J Pathol.* 220:174–185.

Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet.* 43:1035–1039.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Zhu Y, Lin Z, Nakhleh L. 2013. Evolution after whole-genome duplication: a network perspective. *G3* 3:2049–2057.