Functional Divergence for Every Paralog

Patricia S. Soria,[†,1] Kriston L. McGary,[†,1] and Antonis Rokas[*,1]
[1]Department of Biological Sciences, Vanderbilt University
[†]These authors contributed equally to this work.
***Corresponding author:** E-mail: antonis.rokas@vanderbilt.edu.
Associate editor: Aoife McLysaght

Abstract

Because genes can be constrained by selection at more than one phenotypic level, the relaxation of constraints following gene duplication allows for functional divergence (FD) along multiple phenotypic axes. Many studies have generated individual measures of FD, but the profile of FD between paralogs across levels of phenotypic space remains largely uncharted. We evaluate paralog pairs that originated via the yeast whole-genome duplication (ohnolog pairs) at three distinct phenotypic levels (properties of proteins, gene expression, and overall organismal growth) using eight complementary measures of FD (protein: evolutionary rates, radical amino acid substitutions, and domain architecture; gene expression: expression differences in a single species and condition, across species in a single condition, and in a single species across conditions; and organismal: genetic interaction profiles and growth profiles in multiple conditions). We find that the majority of ohnolog pairs show FD by multiple phenotypic measures. Within each phenotypic level, measures of FD are strongly correlated but are generally weakly correlated between levels, suggesting that functional constraints exerted on genes from distinct phenotypic levels are largely decoupled. Our results suggest that redundancy is a rare functional fate for retained paralogs and that FD cannot be fully captured by measures at any single phenotypic level.

Key words: **whole-genome duplication, protein sequence evolution, gene expression, genetic interaction, growth assay, *Saccharomyces* sensu stricto, ortholog conjecture.**

Letter

Gene duplication is a major contributor to the evolution of new genes and function (Ohno 1970; Hughes 1994; Wolfe 2001; Zhang 2003). It has long been assumed that functionally redundant duplicate genes cannot be maintained over the long-term because deleterious mutations will not be removed by purifying selection (Nei and Roychoudhury 1973). Early models of gene duplication focused on neofunctionalization (Ohno 1970), but, for some time, the primary theoretical model for understanding the retention of most paralogs has been the duplication–degeneration–complementation model, which predicts that genes are retained in duplicate through partitioning the set of the ancestral gene's functions between the two copies through a selectively neutral process (i.e., subfunctionalization; Force et al. 1999). More recent models have focused on the role of positive selection in gene duplication, due to escape from adaptive conflict (Hittinger and Carroll 2007) or selection of a new weak secondary function that drives gene duplication and subsequent divergence (Bergthorsson et al. 2007; Elde et al. 2012). Irrespectively of the model underlying a particular gene duplication, functional divergence (FD) of duplicates is expected to be the norm for duplicate genes (Force et al. 1999; Bergthorsson et al. 2007; Hittinger and Carroll 2007; Des Marais and Rausher 2008; Elde et al. 2012).

Functional constraints imposed on a gene can come from multiple phenotypic levels; from constraints in protein sequence and structure, to constraints in timing and pattern of gene expression, all the way to constraints on how the gene impacts organismal phenotypes. Higher level (e.g., organismal) phenotypes are obviously dependent on lower level ones (e.g., gene expression or protein sequence); however, the exact nature of this dependency when genes duplicate and diverge is unclear. Several studies have shown FD in gene duplicates using a variety of individual measures, including protein sequence (Gu 2001, 2006; Byrne and Wolfe 2007), domain structure (Grassi et al. 2010; Khaladkar and Hannenhalli 2012), gene expression (Gu et al. 2005; Conant and Wolfe 2006), and genetic interactions (Ihmels et al. 2007; VanderSluis et al. 2010). Although these studies have successfully drawn level-specific profiles of FD between duplicates, a multilevel evaluation of FD measures that not only captures the diversity of phenotypic levels but also describes their interrelationships is yet to be done.

One special category of gene duplicates is ohnolog pairs, that is paralog pairs originating from whole-genome duplication (WGD), which are of specific interest because such pairs originate simultaneously as well as span a wide diversity of functional categories (Wolfe 2001; Davis and Petrov 2005; Guan et al. 2007; Wapinski et al. 2007). The ancient WGD in the ancestor of *Saccharomyces cerevisiae* and related yeasts (Wolfe and Shields 1997; Wong et al. 2002; Kellis et al. 2004; Scannell et al. 2006) provides an excellent model to study the FD of ohnolog pairs across multiple phenotypic levels. Specifically, the availability of superbly annotated genome sequences from several closely related *Saccharomyces* species (Dujon 2010; Scannell et al. 2011) has enabled a remarkably

clear and precise demarcation of nearly all ohnolog pairs retained from the WGD (Byrne and Wolfe 2005). Furthermore, the *Saccharomyces* sensu stricto genus is unparalleled in its breadth of available functional assays that span multiple phenotypic levels and conditions (Hillenmeyer et al. 2008; Nagalakshmi et al. 2008; Costanzo et al. 2010; Caudy et al. 2013; Hittinger 2013).

In this study, we evaluate 499 ohnolog pairs across five *Saccharomyces* sensu stricto species stemming from the yeast WGD at three distinct phenotypic levels (properties of proteins, gene expression, and overall organismal growth) using eight complementary measures of FD (protein: radical substitutions, evolutionary rates, domain architecture; gene expression: expression differences in a single species and condition, across species in a single condition, and in a single species across conditions; and organismal: genetic interaction profiles, and growth profiles in multiple conditions). We find that nearly all ohnolog pairs show FD by at least one measure and, by extension, phenotypic level. For example, 93% of annotated ohnolog pairs show FD by protein sequence-based measures and 90% by growth-based measures. Counter to the perception that gene expression diverges rapidly, only 47% of ohnolog pairs show FD when gene expression is measured across multiple species in a single condition. Measures of FD for protein-level and organismal-level phenotypes show strong internal correlation but weak correlation across levels, suggesting that they are largely decoupled. Most noticeably, measures that capture gene expression FD are not correlated with protein-level FD measures and only moderately correlated with organismal-level FD measures. Our results suggest that redundancy is a very rare functional fate for retained paralogs and that FD cannot be fully circumscribed by measures at any single phenotypic level. We submit that understanding FD following gene duplication and, more broadly, the evolution of functional novelty requires a pluralistic approach.

Results and Discussion

FD for Every Ohnolog Pair

Examination of the 499 ohnolog pairs retained in five post-WGD species in the *Saccharomyces* sensu stricto clade (supplementary table S1, Supplementary Material online) reveals high levels of FD (fig. 1*A*; supplementary tables S2–S9, Supplementary Material online, summarized in supplementary table S10, Supplementary Material online). For example, 346/359 (96%) of annotated ohnolog pairs show FD in their genetic interaction profiles (supplementary table S7, Supplementary Material online), and 377/421 (90%) of annotated ohnolog pairs show FD in their growth across conditions (supplementary table S8, Supplementary Material online). Of the 127 ohnolog pairs annotated by all eight measures, all show FD by at least two measures, 125/127 (98%) by at least four measures, and 40/127 (31%) ohnolog pairs show FD by at least seven measures (fig. 1*B* and *C*, supplementary table S9, Supplementary Material online).

Six ohnolog pairs show FD across all eight measures: *IZH1*/*IZH4*, *SBP1*/*RNP1*, *ROM1*/*ROM2*, *HBS1*/*SKI7*, *FLC3*/*FLC1*, and *EMP47*/*EMP46* (fig. 1*C*, supplementary table S11, Supplementary Material online). Notably, one of the six, *HBS1*/*SKI7* (*YKR084C*/*YOR076C*), is the first known case of subfunctionalization by loss of alternative splicing in yeast (Van Hoof 2005; Marshall et al. 2013). Remarkably, the ancestral ortholog of *HBS1*/*SKI7* in the fission yeast *Schizosaccharomyces pombe* underwent independent duplication and subsequent loss of alternative splicing (Marshall et al. 2013). This parallel duplication and divergence suggests that the ancestral *HBS1*/*SKI7* gene faced multiple conflicting functional constraints, possibly due to both species converging on a yeast lifestyle, which is also thought to lead to many parallel gene duplications (Hughes and Friedman 2003).

Sometimes FD affects multiple members of a gene family. Such is the case of the *PCL* family of Pho85-associated cyclins (Measday et al. 1997), which act as regulatory partners with Pho85 to form a cyclin-dependent protein kinase to either begin entry into the mitotic cell cycle (*PCL1*, *PCL2*, *PCL5*, and *PCL9*) or to regulate glycogen metabolism (*PCL6*, *PCL7*, *PCL8*, and *PCL10*) (Measday et al. 1997; Huang et al. 1998; Wang et al. 2001). For example, *PCL6*/*PCL7* (*YER059W*/*YIL050W*) shows FD by seven different measures that span all three phenotypic levels (fig. 2). Two other ohnolog pairs in the *PCL* gene family, namely *PCL2*/*PCL9* (*YDL127W*/*YDL179W*) and *PCL10*/*PCL8* (*YPL219W*/*YGL134W*), also show FD at two or more phenotypic levels. The broad retention of duplicates in this family coupled with their extensive FD across multiple phenotypic levels suggests that these duplications have substantially contributed in tweaking growth rate and cell division in a sugar-rich boom-bust lifestyle (Conant and Wolfe 2007; Van Hoek and Hogeweg 2009).

Interestingly, 16/499 (3%) ohnolog pairs lack evidence for FD (fig. 1*C*, supplementary table S12, Supplementary Material online). Notably, all 16 such pairs have fewer than five measures of FD annotated and seven are documented cases of gene conversion in *S. cerevisiae* (Casola et al. 2012) (supplementary table S13, Supplementary Material online). Of these 16 pairs, 14 consist of members of either the large or small subunit of the ribosome, and the other two are closely associated proteins (*SSB1*/*SSB2* and *TEF1*/*TEF2*). Given this bias toward ribosomal proteins, we decided to further investigate all 47 ribosomal ohnolog pairs (supplementary table S14, Supplementary Material online). We found that 16/47 (34%) of these ohnologs show evidence of gene conversion (Casola et al. 2012). Strong purifying selection pressure to maintain ribosomal function may actively suppress FD between ribosomal paralogs across phenotypic levels. Alternatively, lack of FD could be due to gene conversion, which has the additional effect of preventing the analysis of several of our measures of FD due to the resulting high sequence similarity. For example, only the *RPL35a*/*RPL35b* (*YDL191W*/*YDL136W*) ohnolog pair could be evaluated for Type I sequence divergence (no FD found). Similarly, 44/47 (94%) of ribosomal ohnolog pairs were omitted from the analysis of the single condition gene expression measure data, and only *RPS27b*/*RPS27a* (*YHR021C*/*YKL156W*) demonstrated FD. Finally, 31/47 (66%) of ribosomal ohnolog pairs were omitted from the genetic interaction measure of FD

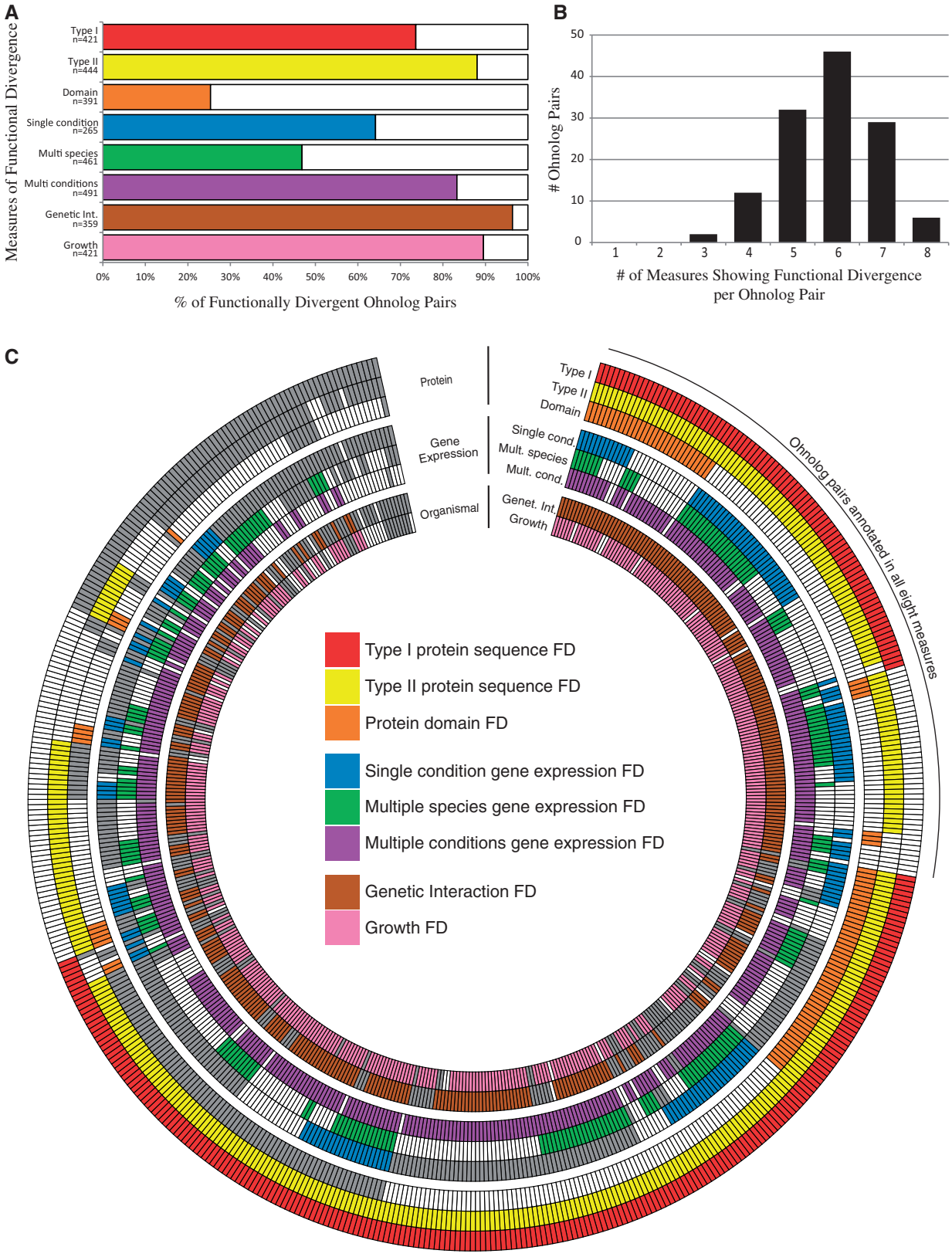

Fig. 1. FD for every gene duplicate that originated from an ancient WGD in the *Saccharomyces* yeasts. (*A*) Percent of functionally divergent (FD) and nonfunctionally divergent ohnolog pairs for each of the eight measures used in this study. *N* values correspond to the number of ohnolog pairs without data missing (out of 499) at each measure. Filled bars represent the percentage of FD ohnolog pairs, and empty bars represent the percentage with no evidence of FD. (*B*) The distribution of fully annotated ohnolog pairs ($N = 127$ ohnolog pairs) that show FD for one or more measures. (*C*) The presence or absence of FD for every ohnolog pair across eight measures at three phenotypic levels. Each box in each circular track corresponds to one of the 499 ohnolog pairs; and concentric data tracks correspond to the eight distinct measures of FD used in this study. Empty boxes represent the absence of FD and light gray boxes represent missing data; all other filled boxes represent the presence of FD.

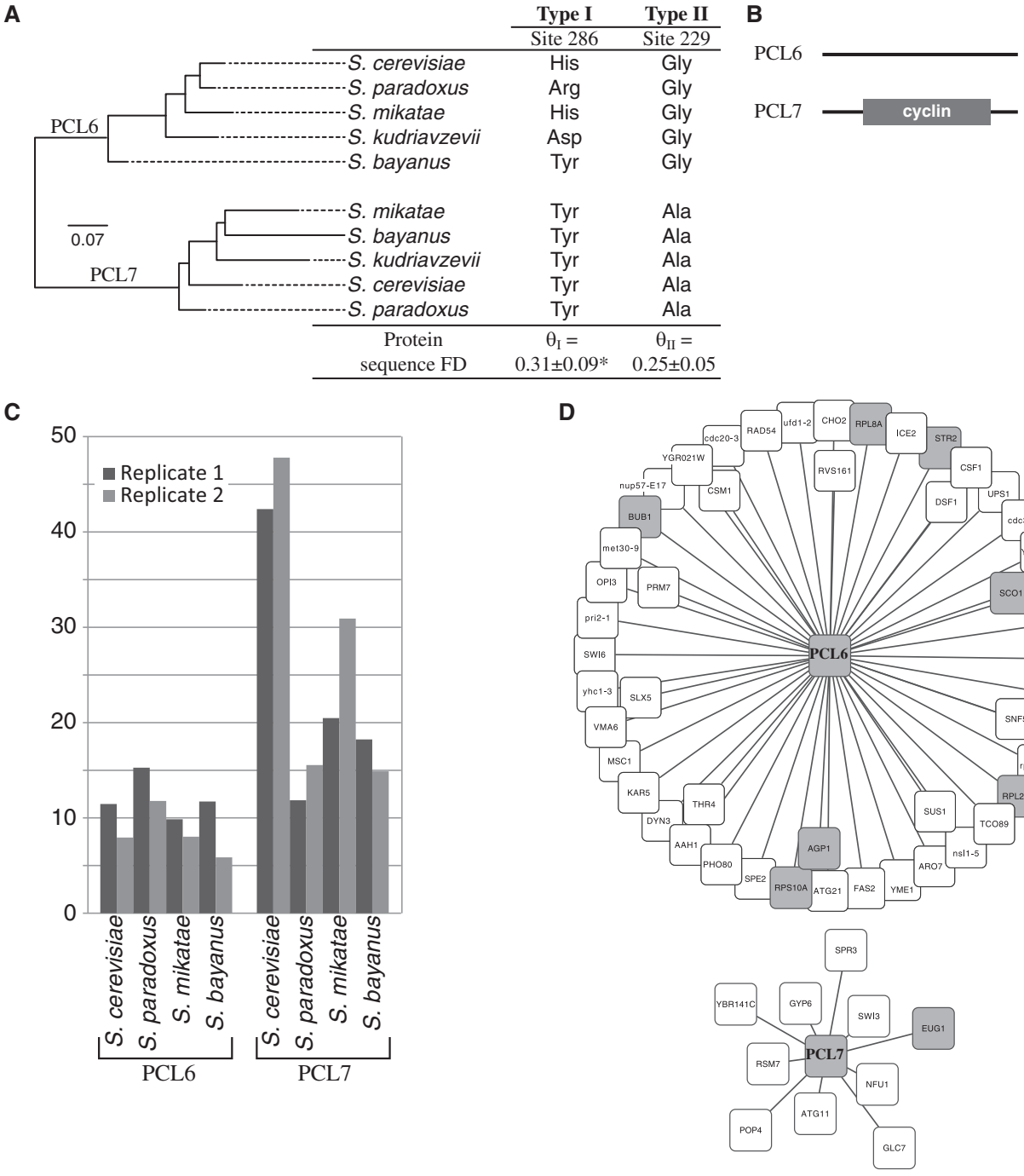

Fig. 2. The *PCL6*/*PCL7* ohnolog pair that is involved in the regulation of glycogen biosynthesis shows FD across seven measures at three phenotypic levels. (*A*) Maximum likelihood gene tree of *PCL6*/*PCL7* orthologs in *Saccharomyces* sensu stricto clade. θ_I and θ_{II} values and their respective standard errors measure the degree of FD between the two members of the ohnolog pair. $\theta \pm 2SE > 0$ indicates significant FD ($P < 0.05$, H_0: $\theta = 0$, for no FD). The asterisk for θ_I indicates significance for Type I FD using the likelihood ratio test (χ^2 critical value = 3.84, $P > 0.05$, $df = 1$). (*B*) Domain assignments for the *PCL6*/*PCL7* ohnolog pair from SUPERFAMILY, a database of HMM models for assigning SCOP domains. *PCL6* has no domain assignments, whereas *PCL7* contains a cyclin domain (scop 47955). (*C*) Gene expression FD of the *PCL6*/*PCL7* ohnolog pair, as measured by two replicates of whole-transcriptome sequencing in four species (Busby et al. 2011). The pair is considered functionally divergent for single condition, single species gene expression because

(continued)

likely due to the essential nature of many ribosomal genes, even though 11 pairs showed evidence of FD by this measure. Despite these limits, 33/47 ribosomal ohnolog pairs show FD by at least one measure; the remaining pairs have either undergone gene conversion (6/47) or are poorly annotated by our chosen measures (8/47; supplementary table S14, Supplementary Material online).

FD Is Strongly Correlated within but Not between Phenotypic Levels

Examination of FD values across the 499 ohnolog pairs showed that FD measures are significantly correlated within each phenotypic level (table 1; average correlation within levels, $r = 0.24$). For example, at the level of protein sequence, Type I (evolutionary rate-based divergence) and Type II (amino acid biochemical property-based divergence) (Gu 1999) measures of FD are strongly correlated ($r = 0.48$) and significantly overlapping ($P = 4.7 \times 10^{-14}$). Similarly, the two organismal measures of FD, genetic interactions, and environmental growth show good correlation ($r = 0.3$) and significant overlap ($P = 2.8 \times 10^{-5}$). In contrast, measures of FD across phenotypic levels are typically considerably less correlated and lack significant overlap (average correlation between levels, $r = 0.06$). Interestingly, the single condition-based measure of gene expression FD shows no correlation with Type I ($r = 0$) or Type II FD ($r = 0.04$), and negative correlation with protein domain FD ($r = -0.18$) with the other measures of gene expression FD showing a similar trend, suggesting that the functional constraints relieved by WGD typically involve either protein function or gene expression, but not both. The strongest correlation found between levels with significant overlap is for the comparison of multicondition gene expression FD (expression level measure) and growth FD (organismal level measure) ($r = 0.27$; $P = 4.1 \times 10^{-11}$), which highlights the value of data from multiple conditions for understanding FD.

The transcription factor *SWI5*/*ACE2* (*YDR146C*/*YLR131C*) ohnolog pair nicely illustrates the lack of coupling between protein sequence-based and gene expression-based measures of FD. *SWI5* and *ACE2* regulate expression of several target genes involved in mating type switching, exit from mitosis, and cell wall function (Doolin et al. 2001). Systematic molecular dissection of the functions of these two proteins has shown that their FD is due to changes in the domains present in each of the two proteins (McBride 1999; Sbia et al. 2008). In contrast, the two ohnologs are expressed at similar levels and do not show FD in gene expression when measured in a single condition in one or multiple species but do show expression FD across multiple conditions (supplementary tables S4–S6, Supplementary Material online).

Studying FD Requires a Pluralistic Approach

When selection impacts multiple phenotypic levels, sometimes in opposite directions, evaluating FD from a single-phenotypic-level perspective is at best partial and at worst misleading. For example, measures of molecular sequence divergence show that rodents have accumulated many more changes than primates (Ohta 1995); however, measures of gene expression divergence in rodents and primates show the opposite pattern (Brawand et al. 2011). Prior investigations into the extent of FD have called for pervasive redundancy among duplicates based on the classical genetics measure of phenotypic buffering against null mutations (Dean et al. 2008; DeLuna et al. 2008; Kafri et al. 2008; Musso et al. 2008; Kafri et al. 2009) or based on similarity in functional annotation (Wapinski et al. 2007). If true redundancy between duplicates is a theoretically rare, if not impossible, fate then these findings may not reflect true evolutionary redundancy (Brookfield 1992), but rather the lack of precision stemming from sole reliance on a single measure at a single phenotypic level.

Of greatest potential concern is that perhaps the most commonly used quantitative measure of FD, which is gene expression in a single condition (in one or multiple species), is only very weakly correlated with FD measures at both the protein level and organismal level (table 1). However, our results suggest that it is possible to improve the assessment of gene expression-based FD by measuring it across multiple conditions. Doing so results in much higher and significant correlation between gene expression FD and the two organismal level measures of FD, although the collection of gene expression and growth data across multiple species and conditions in the same set of experiments would likely be of even greater value.

Differential patterns of FD across phenotypic levels might help explain aspects of the recent controversy surrounding the ortholog conjecture, which postulates that orthology is more predictive of protein function than paralogy (Eisen 1998). A recent examination of the ortholog conjecture based on gene expression provided evidence that questioned its validity (Nehrt et al. 2011). Although our results are not aimed at testing the ortholog conjecture, the significant nonoverlap between ohnolog pairs that exhibit protein sequence-based FD with those that exhibit gene expression-based FD suggests that any testing of the ortholog

Fig. 2. Continued
the ratios of RPKM values of the two members are more than 2-fold different in both *Saccharomyces cerevisiae* replicates (4-fold change in replicate 1 and 6-fold change in replicate 2). The *PCL6*/*PCL7* pair also demonstrates single-condition, multispecies expression FD because *PCL6* is expressed at a lower level than *PCL7* across replicates in all four species (Wilcoxon paired signed rank test, $P = 0.02$). Data from the *PCL6*/*PCL7* pair are missing from the single species, multicondition gene expression experiments. (*D*) Genetic interaction networks for *PCL6* and *PCL7* show no overlap in interaction partners, which is indicative of FD. Other ohnologs in the data set are highlighted in gray. In each case, only one member of the ohnolog pair is present in the *PCL6*/*PCL7* interaction networks. (*E*) Heat map depicting log ratio fitness defect scores of deletion mutants relative to wild-type across 418 environmental and chemical stresses (Hillenmeyer et al. 2008). The negative Pearson correlation ($r = -0.36$) between the environmental growth profiles for *PCL6* and *PCL7* indicates that the two members are in FD for this measure.

Table 1. FD Measures Show Weak or Negative Correlations between Phenotypic Levels (Protein, Expression, and Organismal) but Strong Correlations within Phenotypic Levels[a].

		Protein			Expression			Organismal	
		Type I	Type II	Domain	Single Condition	Multispecies	Multicondition	Genetic Interaction	Growth
Protein	Type I	—	0.48*	0.16*	0.00	0.01	−0.05*	0.17	0.12
	Type II	4.7E-14*	—	0.05	0.04*	−0.02	0.08	0.08	0.14*
	Domain	9.3E-4*	0.09	—	−0.18*	0.01	−0.05	0.00	0.02
Expression	Single condition	0.11	0.02*	0.05*	—	0.40*	0.20	0.19	0.11
	Multispecies	0.27	0.50	0.02*	2.9E-12*	—	0.08	0.07	−0.05
	Multicondition	0.02*	0.12	0.37	0.05	0.04*	—	0.24*	0.27*
Organismal	Genetic interaction	0.14	0.61	0.46	0.75	0.16	8.7E-3*	—	0.30*
	Growth	0.57	9.7E-5*	0.49	0.23	0.12	4.1E-11*	2.8E-5*	—

[a]Upper diagonal values correspond to Pearson correlations (r) between FD measures. Dark gray boxes represent strong positive correlations ($r \geq 0.1$), white boxes represent weak and no correlations, and the light gray box represents a strong negative correlation ($r \leq -0.1$). Lower diagonal values correspond to the probabilities of significantly overlapping FD measures (dark gray boxes) and probabilities of significantly nonoverlapping FD measures (light gray boxes). Values marked with an asterisk and in underline are significant by the cumulative hypergeometric probability ($P < 0.05$).

conjecture from a single-phenotypic-level perspective is likely to be idiosyncratic.

Understanding the diverse fates of gene duplicates is a substantial research agenda due to gene duplication's role as a major substrate for organismal adaptation and the evolution of functional novelty. To mention just a few examples, gene duplication is thought to have contributed to radiations in plants (Lewis 1979; Masterson 1994; Otto and Whitton 2000), speciations in fungi (Scannell et al. 2006), and to the evolution of developmental innovations in vertebrates (Holland et al. 1994). Our results, as well as several individual examples of gene duplications that have been dissected in exquisite detail (Hittinger and Carroll 2007; Des Marais and Rausher 2008; Marshall et al. 2013), indicate that understanding FD following gene duplication and, more broadly, the evolution of functional novelty requires a pluralistic approach.

Materials and Methods

Data Set Generation

A total of 547 *S. cerevisiae* ohnolog pairs are identified through synteny in YGOB (Byrne and Wolfe 2005; Scannell et al. 2011), of which 499 ohnolog pairs have both sets of orthologs present in four other post-WGD species in the sensu stricto clade (*S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, and *S. bayanus*). The amino acid sequences for these 499 ohnolog pairs were retrieved from the high-quality annotations of 5,261 orthologous groups among five species of the sensu stricto clade (Scannell et al. 2011). Because a prior study (Scannell et al. 2006) classified the set of ohnolog pairs that exhibited no gene losses in *S. cerevisiae*, *S. castellii*, and *Candida glabrata* as class 0 loci, we adopted this classification and named these 499 duplicate pairs as class 0 ohnologs, labeling each pair with a Class0_ID in supplementary tables S1–S15, Supplementary Material online. Supplementary table S15, Supplementary Material online, provides SGD annotation for all ohnolog pairs.

Inferring FD from Protein Sequence Data

Type I and Type II FD

The amino acid sequences of each of the 499 ohnolog pairs were aligned with MAFFT (version 6.847, default settings [Katoh et al. 2002]). In turn, these alignments were used as input into DIVERGE software version 2.0 (Gu 1999, 2001; Gaucher et al. 2002; Gu and Vander Velden 2002) for measuring FD between the two members of each pair. DIVERGE2.0 estimates two types of FD from amino acid sequence data, Type I FD (evolutionary rate) and Type II FD (amino acid properties), which are measured by the coefficients of FD θ_I and θ_{II}, respectively. For Type I FD, we used a two-step significance test for rejecting the null hypothesis of no FD ($\theta = 0$): two times the standard error of θ and a likelihood ratio test (critical value = 3.84, $df = 1$, $P < 0.05$) (Gu 2001; Gu and Vander Velden 2002). For Type II, pairs with θ values greater than 0 after subtracting two times the standard error were annotated as having undergone FD ($P < 0.05$, H_0: $\theta = 0$ or no FD). Cases in which the groups of orthologs that define each of the two members of the ohnolog pair were not reciprocally monophyletic were recorded as missing data in this analysis. All values using these measures are reported in supplementary table S2, Supplementary Material online.

Protein Domain Architecture FD

We used the SUPERFAMILY hidden Markov models available in the SUPERFAMILY database (ver 1.75, last accessed April 3, 2013 [Gough et al. 2001; Wilson et al. 2009]) to assign protein domain annotations for each protein. Protein sequences were submitted for SCOP domain annotations to identify differences in protein domain architecture between members of each ohnolog pair. SCOP unique identifiers (sunid) for every ohnolog are available in supplementary table S3, Supplementary Material online. HMM library classifications with E values less than 0.0001 indicated significant hits for superfamily domain assignments.

The criteria for inferring FD were adopted from those outlined by Grassi et al. (2010) for homology classification based on domain architecture similarity. Specifically, we annotated members of ohnolog pairs with identical domain architecture (class A) as having no FD in protein domain architecture. Members of ohnolog pairs whose domain architectures differed only in domain copy number (class B) and ohnologs pairs whose domain architectures differed in the presence/absence of one or more domains (class C) were considered to show FD in protein domain architecture. Members of

ohnolog pairs in which one ohnolog had no significant hits and the other had significant hits for domain assignments were considered class C pairs and therefore to show FD. FD classifications were done based on the 499 *S. cerevisiae* ohnolog pairs, which to a great extent coincided with the FD classifications done based on the other four species. All values using this measure are reported in supplementary table S3, Supplementary Material online.

Inferring FD from Gene Expression

Inferring FD from Gene Expression Measurements of S. cerevisiae in a Single Condition

To measure FD in gene expression between members of each ohnolog pair, we used the *S. cerevisiae* transcriptome data from Busby et al. (2011). Gene annotations and unique reads were downloaded from Gene Expression Omnibus (GSE32679) and used to calculate RPKM values (number of mapped reads per kilobase of exon per million mapped reads) from uniquely mapped reads. The RPKM value for a transcript was reported as missing data if the ratio between the number of all mapped reads for that transcript and the number of uniquely mapped reads for that transcript exceeded 1.3. The mapped reads/uniquely mapped reads ratio was part of a conservative strategy to minimize the effect of read mapping uncertainty between highly similar ohnolog sequences. Because RPKM is a relative and normalized measure of gene expression and both members of an ohnolog pair are included in the mapped transcriptome, we considered an ohnolog pair to show FD when there was a more than 2-fold change of RPKM values between the two members. To accurately quantify significant fold changes, we omitted ohnolog pairs for which one or both RPKM values of their members was less than 5 and the difference between their RPKMs was less than 10. Ohnolog pairs were annotated as showing FD only when the expression of two members of the pair had the same behavior in both replicates of the RNA-Seq data generated by Busby et al. (2011). When the results from one biological replicate disagreed with those of the other, the ohnolog pair in question was considered as missing data. All values using this measure are reported in supplementary table S4, Supplementary Material online.

Inferring FD from Gene Expression Measurements of Multiple Species in a Single Condition

To measure FD in gene expression across multiple species, we used the transcriptome data that contain two biological replicate measurements of *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus* grown in a single condition generated by Busby et al. (2011). RPKM values for each ohnolog pair were calculated from uniquely aligned reads. We annotated only those ohnolog pairs with RPKM values in at least two species (261/499 ohnolog pairs), with RPKMs for each pair in each species treated as matched pairs in a Wilcoxon signed ranks test. Expression values and reported *P* values for the Wilcoxon Signed Ranks test (R Core Team 2013) are reported in supplementary table S5, Supplementary Material online. We used a $P < 0.05$ cut off to annotate ohnolog pairs as showing FD.

Inferring FD from Gene Expression Measurements of S. cerevisiae in Multiple Conditions

To measure FD in gene expression across multiple conditions, we used the list of overexpressed and underexpressed *S. cerevisiae* genes across 18 growth conditions provided by Waern and Snyder (2013). We converted these results into a numerical matrix (1 = overexpressed relative to expression in YPAD media, 0 = same expression as YPAD, − 1 = underexpressed relative to expression in YPAD) and calculated the Pearson correlation for each ohnolog pair. Based on the distribution of correlation values of 500 randomly chosen pairs, we chose an $r < 0.63$ cut off (97.5th percentile of genomic background) as the criterion for significant FD. Pearson correlation values for ohnolog pairs are reported in supplementary table S6, Supplementary Material online.

Inferring FD from Organismal Phenotypes

Inferring FD from Genetic Interaction Profiles

Epistasis, that is the presence of genetic interactions, occurs when one gene influences the effect of another gene's impact on phenotype. We classified genetic interactions as organismal phenotypes because they are quantified by comparing the organismal growth rate of a strain with two gene knockouts to the organismal growth rates of the single gene knockout strains. Pearson correlation coefficients for 359 of the 499 ohnolog pairs were extracted from the yeast genetic interaction data set generated by Costanzo et al. (2010) (http://drygin.ccbr.utoronto.ca, last accessed February 5, 2014), with missing correlations annotated as missing data. Based on the distribution of correlations between all 8 million random pairs from the genomic background, we annotated a given ohnolog pair as showing FD only when their correlation coefficient was $r < 0.2$ (approximately the 95th percentile of background correlation). All values using this measure are reported in supplementary table S7, Supplementary Material online.

Inferring FD from Environmental Growth Profiles

We obtained growth fitness assay data for whole-genome *S. cerevisiae* homozygous deletion strains from the study of Hillenmeyer et al. (2008), which is available as a web supplement on the Yeast Fitness Database (http://fitdb.stanford.edu/, last accessed February 5, 2014). Log-ratio fitness defect values across over 418 stress conditions were extracted for each member of an ohnolog pair, giving an environmental growth profile. Custom Python and R scripts were used to calculate Pearson correlations between environmental growth profiles of 499 randomly selected pairs of deletion strains and based on the distribution, we chose an $r < 0.4$ cut off (97.5th percentile of background) as the criteria for significant FD. Only ohnolog pairs with negative correlation or positive correlation smaller than 0.4 were annotated as showing FD. Pairs in which one or both of the ohnologs were missing from the data were omitted. All values using this measure are reported in supplementary table S8, Supplementary Material online.

Evaluating FD Relationships between Phenotypic Levels

To calculate Pearson correlations between measures of FD, we used the following quantitative values from each measure: For protein sequence-based FD, we used θ_I and θ_{II} for Type I and Type II FD, respectively, whereas numeric values were assigned to classes of domain divergence ($A = 0$, $B = 1$, $C = 2$) for protein domain architecture-based FD; for gene expression FD based on a single condition from a single species, we used the average of the $\log_2$ of the ratio of RPKMs between ohnologs; for gene expression FD based on a single condition from multiple species, we used the P values of the Wilcoxon signed ranks tests; for gene expression FD based on multiple conditions from a single species, we used the Pearson correlation coefficient between paralogs across the various conditions as calculated above; and for both organismal phenotypes, that is for genetic interaction and growth profile measures, the r value of correlation between members of ohnolog pairs was used. In each case, the available quantitative values were used to calculate Pearson correlation with maximum precision, rather than using just the FD/no-FD binary values. To calculate the sign of the correlation correctly, numeric measures of FD that decrease with increasing FD were multiplied by −1. In calculating the correlation between measures of FD, ohnolog pairs with missing values in either measure of FD were omitted.

In addition to calculating correlation, we also calculated the significance of the similarity of the sets of ohnologs annotated as FD by each measure. To do so, we used the cumulative hypergeometric probability of the overlap (or nonoverlap) between the sets for each pair of FD measures (reported in the lower diagonal of table 1). Ohnolog pairs with missing data were omitted from this calculation. The more significant of the two probabilities (overlap or nonoverlap) is reported in table 1. Asterisks throughout table 1 refer to this probability (dark gray, significance of overlap; light gray, significance of nonoverlap).

Supplementary Material

Supplementary tables S1–S15 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University. This work was supported by National Science Foundation grant DEB-0844968 to A.R.

References

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A.* 104:17004–17009.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478: 343–348.

Brookfield J. 1992. Can genes be truly redundant? *Curr Biol.* 2:553–554.

Busby MA, Gray JM, Costa AM, Stewart C, Stromberg MP, Barnett D, Chuang JH, Springer M, Marth GT. 2011. Expression divergence measured by transcriptome sequencing of four yeast species. *BMC Genomics* 12:635.

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

Byrne KP, Wolfe KH. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* 175:1341–1350.

Casola C, Conant GC, Hahn MW. 2012. Very low rate of gene conversion in the yeast genome. *Mol Biol Evol.* 29:3817–3826.

Caudy AA, Guan Y, Jia Y, Hansen C, DeSevo C, Hayes AP, Agee J, Alvarez-Dominguez JR, Arellano H, Barrett D, et al. 2013. A new system for comparative functional genomics of *Saccharomyces* yeasts. *Genetics* 195:275–287.

Conant GC, Wolfe KH. 2006. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol.* 4:e109.

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

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. *Science* 327:425–431.

Davis JC, Petrov DA. 2005. Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet.* 21:548–551.

Dean EJ, Davis JC, Davis RW, Petrov DA. 2008. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.* 4: e1000113.

DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, Kishony R. 2008. Exposing the fitness contribution of duplicated genes. *Nat Genet.* 40:676–681.

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454: 762–765.

Doolin MT, Johnson AL, Johnston LH, Butler G. 2001. Overlapping and distinct roles of the duplicated yeast transcription factors Ace2p and Swi5p. *Mol Microbiol.* 40:422–432.

Dujon B. 2010. Yeast evolutionary genomics. *Nat Rev Genet.* 11:512–524.

Eisen JA. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8: 163–167.

Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, Geballe AP, Malik HS. 2012. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell* 150: 831–841.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci.* 27:315–321.

Gough J, Karplus K, Hughey R, Chothia C. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol.* 313: 903–919.

Grassi L, Fusco D, Sellerio A, Corà D, Bassetti B, Caselle M, Lagomarsino MC. 2010. Identity and divergence of protein domain architectures after the yeast whole-genome duplication event. *Mol Biosyst.* 6: 2305–2315.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol.* 16:1664–1674.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol.* 18:453–464.

Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol.* 23:1937–1945.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* 18:500–501.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A.* 102:707–712.
Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* 175: 933–943.
Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, et al. 2008. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* 320: 362–365.
Hittinger CT. 2013. *Saccharomyces* diversity and evolution: a budding model genus. *Trends Genet.* 29:309–317.
Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* 449:677–681.
Holland PW, Garcia-Fernàndez J, Williams NA, Sidow A. 1994. Gene duplications and the origins of vertebrate development. *Dev Suppl.* 125–133.
Huang D, Moffat J, Wilson WA, Moore L, Cheng C, Roach PJ, Andrews B. 1998. Cyclin partners determine Pho85 protein kinase substrate specificity in vitro and in vivo: control of glycogen biosynthesis by Pcl8 and Pcl10. *Mol Cell Biol.* 18:3289–3299.
Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci.* 256:119–124.
Hughes AL, Friedman R. 2003. Parallel evolution by gene duplication in the genomes of two unicellular fungi. *Genome Res.* 13:794–799.
Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS. 2007. Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Mol Syst Biol.* 3:86.
Kafri R, Dahan O, Levy J, Pilpel Y. 2008. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A.* 105:1243–1248.
Kafri R, Springer M, Pilpel Y. 2009. Genetic redundancy: new tricks for old genes. *Cell* 136:389–392.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.
Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428:617–624.
Khaladkar M, Hannenhalli S. 2012. Functional divergence of gene duplicates—a domain-centric view. *BMC Evol Biol.* 12:126.
Lewis WH. 1979. Polyploidy in species populations. *Basic Life Sci.* 13: 103–144.
Marshall AN, Montealegre MC, Jiménez-López C, Lorenz MC, van Hoof A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. *PLoS Genet.* 9:e1003376.
Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science* 264:421–424.
McBride HJ. 1999. Distinct regions of the Swi5 and Ace2 transcription factors are required for specific gene activation. *J Biol Chem.* 274: 21029–21036.
Measday V, Moore L, Retnakaran R, Lee J, Donoviel M, Neiman A, Andrews B. 1997. A family of cyclin-like proteins that interact with the Pho85 cyclin-dependent kinase. *Mol Cell Biol.* 17: 1212–1223.
Musso G, Costanzo M, Huangfu M, Musso G, Costanzo M, Huangfu M, Smith AM, Paw J, San Luis BJ, Boone C, et al. 2008. The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18:1092–1099.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.
Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol.* 7:e1002073.
Nei M, Roychoudhury A. 1973. Probability of fixation of nonfunctional genes at duplicate loci. *Am Nat.* 107:362–372.
Ohno S. 1970. Evolution by gene duplication. New York: Springer.
Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J Mol Evol.* 40:56–63.
Otto SP, Whitton J. 2000. Polyploid incidence and evolution. *Annu Rev Genet.* 34:401–437.
R Core Team. 2013. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. [cited 2013 Dec 18]. Available from: http://www.R-project.org/.
Sbia M, Parnell EJ, Yu Y, Olsen AE, Kretschmann KL, Voth WP, Stillman DJ. 2008. Regulation of the yeast Ace2 transcription factor during the cell cycle. *J Biol Chem.* 283:11135–11145.
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, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, Rine J, Johnston M, Hittinger CT. 2011. The awesome power of yeast evolutionary genetics: new genome sequences and strain resources for the *Saccharomyces* sensu stricto genus. *G3* 1:11–25.
Van Hoek MJA, Hogeweg P. 2009. Metabolic adaptation after whole genome duplication. *Mol Biol Evol.* 26:2441–2453.
Van Hoof A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics* 171:1455–1461.
VanderSluis B, Bellay J, Musso G, Costanzo M, Papp B, Vizeacoumar FJ, Baryshnikova A, Andrews B, Boone C, Myers CL, et al. 2010. Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol.* 6:429.
Waern K, Snyder M. 2013. Extensive transcript diversity and novel upstream open reading frame regulation in yeast. *G3* 3:343–352.
Wang Z, Wilson WA, Fujino MA, Roach PJ. 2001. The yeast cyclins Pc16p and Pc17p are involved in the control of glycogen storage by the cyclin-dependent protein kinase Pho85p. *FEBS Lett.* 506:277–280.
Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449: 54–61.
Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. 2009. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.* 37:D380–D386.
Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet.* 2:333–341.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.
Wong S, Butler G, Wolfe KH. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc Natl Acad Sci U S A.* 99: 9272–9277.
Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol Evol.* 18:292–298.