Published in final edited form as:

J Proteome Res. 2012 January 1; 11(1): 261–268. doi:10.1021/pr201065k.

Phosphorylation of Yeast Transcription Factors Correlates with the Evolution of Novel Sequence and Function

Mark Kaganovich and Michael Snyder*

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, United States

Abstract

Gene duplication is a significant source of novel genes and the dynamics of gene duplicate retention vs loss are poorly understood, particularly in terms of the functional and regulatory specialization of their gene products. We compiled a comprehensive data set of *S. cerevisiae* phosphosites to study the role of phosphorylation in yeast paralog divergence. We found that proteins coded by duplicated genes created in the Whole Genome Duplication (WGD) event and in a period prior to the WGD are significantly more phosphorylated than other duplicates or singletons. Though the amino acid sequence of each paralog of a given pair tends to diverge fairly similarly from their common ortholog in a related species, the phosphorylated amino acids tend to diverge in sequence from the ortholog at different rates. We observed that transcription factors (TFs) are disproportionately present among the set of duplicate genes and among the set of proteins that are phosphorylated. Interestingly, TFs that occur on higher levels of the transcription network hierarchy (i.e., tend to regulate other TFs) tend to be more phosphorylated than lower-level TFs. We found that TF paralog divergence in expression, binding, and sequence correlates with the abundance of phosphosites. Overall, these studies have important implications for understanding divergence of gene function and regulation in eukaryotes.

© 2011 American Chemical Society

*Corresponding Author mpsnyder@stanford.edu..

Supporting Information

Supplementary Table 1. Phosphorylation data sources and the number of unique phosphosites reported by each study. Supplementary Table 2. Data from Ptacek et al.¹¹ on the activity of 82 unique protein kinases was used to calculate the correlation between the number of phosphosites on a protein and the number of kinases targeting the protein. The Pearson correlation for all proteins and only for TFs is shown in the above table, along with associated p-values. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Keywords

phosphorylation; gene duplication; paralogs; transcription factors

INTRODUCTION

Gene duplication is widely thought to be the main evolutionary mechanism for the development of biological functional complexity.^{1–3} Genomic analysis of many species spanning hundreds of millions of years of evolutionary time has provided well-supported hypotheses of gene duplications. In contrast, little is known about the functional consequences of most duplication events. In particular, it remains largely unknown why some duplicates are retained while others are not. There is a wealth of theoretical work on the subject, reviewed by Innan and Kondrashov.³ The models attempt to explain how gene copies that arise from a duplication event become fixed in a population and then preserved and maintained as paralogs, assuming the cost-benefit model of a fitness landscape central to natural selection. In essence, the selection model suggests that paralogs are retained if at least one develops a new function, or if higher dosage or the presence of a back-up copy confers a selective advantage.³ Often, either one of the genes develops an entirely new function (neofunctionalization) or the two genes split the functional burden of the original copy (subfunctionalization). Expression analysis suggests that paralog divergence can be attributed in many cases more to the divergence of expression regulatory elements rather than the divergence of coding sequence itself.^{4–6} Understanding gene duplicate retention and functional specialization (or lack thereof) is central to understanding evolution and gene function. The scenario of genes closely related in coding sequence, yet of separate and distinct use to the cell, is relevant to much ongoing research: phenotypic divergence among species, cell differentiation in multicellular organisms, and in cancer where the consequences of genomic instability are often gene disruptions or multiplications. In this study, we attempt to better understand yeast paralog divergence.

Recent advances in proteomics have allowed for the proteome-wide detection and mapping of post-translational protein phosphorylation.⁷ As an important mechanism of protein function regulation, we asked whether phosphorylation plays a role in paralog retention and divergence. Previous work by Amoutzias et al. found that yeast Whole Genome Duplication (WGD) paralogs are more phosphorylated than other proteins.⁸ This is consistent with the hypothesis that WGD paralogs were retained due to novel regulation disproportionately to other paralogs that originated in smaller-scale duplications (SSD).^{4,5,8} We used the phylogenetic analysis of SSD events by Wapinski et al. (Figure 1A) to calculate the number of phosphorylation events on the *S. cerevisiae* proteins descendant from each of the duplications.⁶ We found that an event prior to the WGD also contains a high level of phosphorylation. We compared *S. cerevisiae* phospho-sites to their orthologous positions on *K. waltii*. *K. waltii* is a distant relative of *S. cerevisiae* that formed prior to the WGD.⁹ Thus, most *S. cerevisiae* paralog pairs that originated in the WGD and post-WGD duplications have the same ortholog in *K. waltii*.^{6,9} We observed that phosphorylated amino acids diverge differently between two paralogs when each paralog is aligned to their common *K. waltii* orthologs, whereas nonphosphorylated amino acids tend to have similar divergence

rates. We further investigated the relationship of phosphorylation with Transcription Factor (TF) paralogs and found that TF duplicates tend to be highly phosphorylated and the number of phosphosites among the pair is correlated to the functional divergence between the TFs.

RESULTS

Phosphorylation of Genes that Originated in Duplication Events

Recent high throughput proteome Mass Spectrometry (MS) studies in *S. cerevisiae* have resulted in data on thousands of phosphosites in the yeast genome.¹⁰ We compiled data from seven studies for a total of approximately 10000 serine, threonine, or tyrosine phosphosites on over 2000 yeast proteins (Suppl Table 1, Supporting Information, and Methods). The number of phosphosites per protein correlated weakly though significantly with the number of kinases targeting the protein as detected by kinase protein arrays in Ptacek et al. (Suppl Table 2, Supporting Information, and Methods).¹¹ To study the role of phosphorylation on the evolution of proteins from gene duplicates we used the phylogenetic classification of the history of gene duplication events in yeast compiled by Wapinski et al.⁶ A summary of the duplication events and the yeast species descendent from the resulting evolutionary divergence is presented in Figure 1A. Four-hundred thirty-seven paralog pairs are said to have originated and subsequently retained in the Whole Genome Duplication event (WGD) and 346 other pairs originated in Smaller Scale Duplication events (SSD).^{4,6,9} The orthologs and paralog gene groups were defined by Wapinski et al. using gene sequence similarity combined with the yeast phylogenetic tree to estimate gene ancestry.^{6,12}

We calculated the number of phosphorylation sites on the *S. cerevisiae* proteins retained and descendant from each duplication event (Figure 1A) and found that the WGD paralogs and paralogs from SSD duplication events prior to WGD tend to be enriched in phosphosites as compared to post-WGD proteins. Amoutzias et al. have already observed that WGD gene proteins tend to be phosphorylated at higher rates than average yeast proteins including SSD-generated paralogs.⁸ However, we find that several pre-WGD SSD events also have higher levels of phosphorylation than singleton, nonduplicated genes. This may suggest that phosphorylation was a more significant mechanism for paralog functional differentiation for duplicates created and retained prior and during the WGD than for more recent duplicates.

We further investigated the potential role of phosphorylated amino acids in paralog divergence. As illustrated in Figure 2a we compared the sequence divergence of amino acids in a paralog pair when aligned to their common orthologs in *K. waltii* (Figure 2A). For each paralog pair, the divergence rate of amino acids in paralog 1 (a_1) from the ortholog in *K. waltii* was compared to the divergence rate of paralog 2 from the ortholog (a_2). We refer to the difference between a_1 and a_2 as the difference of divergence rates for each paralog pair. In other words, if $|a_1 - a_2|$ is large then the gene copies have a highly different rate of divergence from the ortholog. We performed a similar calculation for the phosphosites, $|p_1 - p_2|$ (Figure 2A). The calculations are further described in the Methods section. An example alignment is illustrated in Figure 2B. The results (Figure 2B) show that phosphorylated amino acids tend to diverge more asymmetrically than nonphosphorylated amino acids. This supports the hypothesis that phosphorylation plays a role in paralog divergence.

Transcription Factor Phosphorylation

TFs are central to gene regulation and have been previously shown to be often phosphorylated. As such, we investigated TF paralog functional divergence. TFs tend to be enriched in the set of genes retained in yeast duplication events as compared to the global proportion of TFs in the genome (approximately 5% of the proteome are TFs whereas about 10% of duplicated genes are TFs) (Figure 3A). This points to the possibility that as yeast species diverged through evolutionary innovation the nascent genes generated from duplication events that had TF properties were more likely be retained than other genes, perhaps because of their contribution to the cell regulatory network.

TF Phosphorylation Is Correlated with the Transcription Regulation Network Level

We found that TFs are more phosphorylated than non-TF genes, regardless of duplication category (Figure 3B). To further investigate the role of phosphorylation in TF function we explored the relationship between the number of phosphosites on a protein and its level in the transcriptional regulatory network hierarchy as described by Yu and Gerstein (2006).¹³ Yu and Gerstein categorized the 287 *S. cerevisiae* TFs as belonging to one of four levels in a hierarchical network model of yeast transcriptional regulation.¹³ Level 1 TFs are those that directly regulate the transcription of a target (non-TF) gene. Level 2 TFs are those that regulate the transcription of Level 1 TFs, Level 3 regulate Level 2, and Level 4 regulate Level 3. Most TFs are Level 1 (145 out of 287), 104 are Level 2, 30 are Level 3, and eight are Level 4.¹³ In Figure 4 we plot the mean number of phosphosites per protein for each category except for Level 4 because the small size of the Level 4 category makes sample statistics comparisons insignificant (Figure 4). Next to each mean we indicate the p-values of a *t* test where the null hypothesis is that the distribution of phosphosites for the given category is the same as for Level 1 (Figure 4). Surprisingly, we observed that higher level TFs tend to be significantly more phosphorylated than Level 1 TFs. This means that TFs that regulate other TFs are more regulated by phosphorylation. Yu and Gerstein suggest that higher-level TFs respond to environmental signals by controlling whole transcriptional modules at once rather than just individual genes. Our results may suggest that phosphorylation is a mechanism by which these environmental signals propagate through the gene regulatory network.

TF Paralog Divergence Correlates with Phosphosite Abundance

We next investigated whether the amount of phosphorylation on a TF paralog pair relates to the functional divergence between the two proteins. Figure 2 suggests that phosphosites play a disproportionately higher role (compared to nonphosphorylated amino acids) in paralog divergence. Figure 4 provides evidence for the importance of phosphorylation in TF functional on a global level. Thus, we hypothesized that the number of phosphosites detected on a TF paralog pair is directly related to the amount by which the pair diverges functionally, as measured by expression divergence and DNA binding.

We tested this hypothesis by comparing the coexpression modules of each TF to its paralog (when available). We used microarray data of *S. cerevisiae* gene expression, under 1327 conditions, compiled by McCord et al.¹⁴ We calculated the pairwise correlation of every gene expression profile (a 1×1327 vector) with the expression profiles of the other genes.

As a result, every gene had a coexpression profile (a 1×5800 vector). The Euclidean distance between this vector for gene X and the vector for gene Y we took to be an estimate of coexpression module divergence between X and Y. This process is described further in the Methods section.

The distance between the coexpression profiles of TF paralogs was compared to the number of phosphorylation events per protein. Interestingly, we saw a significant correlation of 0.3 between the phosphosite number and coexpression distance (Figure 5A, left panel). The further apart a TF paralog pair is in terms of its coexpression patterns, the more phosphosites there are among the paralogs. The same measurement for non-TF proteins does not yield a correlation (Figure 5A, left panel). We then calculated the correlation of phosphosite number with sequence divergence between paralogs (Figure 5A, middle panel) and sequence divergence with coexpression distance (Figure 5A, right panel). We found a greater correlation of phosphorylation amount and sequence divergence for TF paralogs than for all paralogs as a whole. The same was true for the correlation of sequence divergence with coexpression distance. These results suggest sequence divergence and expression divergence are more directly correlated among TF paralogs than other paralogs and that one or both of these measurements of genetic divergence correlates with the degree to which the paralog pairs are phosphorylated.

We sought to further investigate TF paralog pair divergence by comparing the binding profile of a TF with the binding profile of its paralog. Harbison et al. used chromatin immunoprecipitation combined with array hybridization (ChIP-chip) to evaluate the binding of yeast TFs to intergenic regions.¹⁵ Zhu et al. measured the *in vitro* binding profiles of proteins to DNA oligos to hypothesize binding motifs, or specific sequence patterns that drive the highest affinity binding, for each protein tested.¹⁶ Integrating their approach with the Harbison et al. data, Zhu et al. provide data on 89 TF binding profiles across the genomic intergenic regions measured by these two separate methods.¹⁶ Of the 89 available TF binding profiles we selected those TFs that resulted from gene duplication. We calculated the correlation of binding profiles between the paralogs and measured the relationship between this correlation and the number of phosphosites among the pair (Figure 5B). Although the number of pairs we were able to identify among the TF binding data was relatively small (15 pairs) we found a trend suggesting that the number of phosphosites on the paralog pairs potentially correlates with the degree to which a paralog TF pair's binding profiles correlate among themselves (Figure 5B, left panel). This supports the earlier observation that TF functional divergence is related to abundance of phosphorylated amino acids on the TF paralog pair. We also measured the relationship of phosphorylation events and sequence divergence (Figure 5B, middle panel). TF paralog sequence divergence correlates with phosphosite number, as in Figure 5A, middle panel. Sequence divergence itself anticorrelates with TF paralog binding correlation (Figure 5B, right panel), corroborating the result of Figure 5A, right panel that sequence divergence and TF paralog functional divergence are correlated with each other and with the number of phosphorylation sites on the paralogs.

DISCUSSION

Phosphorylation is a reversible post-translational modification important in cellular signal processing.¹⁷ It is presently thought that approximately one-third of proteins in eukaryotic organisms are phosphorylated, though our knowledge of phosphorylation sites in the proteome is far from complete.^{11,18} Advances in mass spectrometry have allowed for the global study of the *in vivo* phosphoproteome.⁷ We compiled the available *S. cerevisiae* phosphosite data covering ~10000 sites on ~2000 proteins to analyze phosphorylations' effect on paralog divergence in yeast. Our results suggest a previously unexplored role for phosphorylation in protein evolution.

Earlier work on the comparative genomics of yeast by Landry et al. and others suggests that phosphosites evolve slightly slower than average amino acids when adjusted for their prevalence in “unstructured”, flexible protein regions that tend to exhibit low sequence-specific conservation.¹⁰ This assigns a degree of functional significance to amino acids that are phosphorylated. However, this aggregate measure of phospho-site conservation in a yeast species multiple alignment is complicated by the potential presence of many different categories of phosphosites. A phosphorylation event on a protein may be functionally conserved in a position-specific manner whereas for other phosphorylation events the exact position on the protein may be of little consequence. In contrast, we probed the functional significance of phosphosites by assessing their asymmetric divergence among paralogs. Paralogs provide a system where sequence comparison can be triangulated with the ortholog in common, thus accentuating any potentially meaningful signal of divergence between the three entities (ortholog, paralog 1, paralog 2). Our observation that phosphosites tend to diverge more asymmetrically than other amino acids among paralog pairs when compared to the ortholog (Figure 2) suggests that they may more strongly contribute to, or result from, the functional divergence of retained gene duplicates than other coding sequence.

As a major signal processing mechanism in the cell, phosphorylation has been widely reported to affect TF behavior.¹⁹ In global proteomic studies the large representation of TFs among phosphorylated proteins has been reported *in vitro* in yeast by our lab¹¹ and *in vivo* in HeLa cells.¹⁸ We find this enrichment to be the case in our *S. cerevisiae* *in vivo* data set among both singleton and paralog phosphoproteins (Figure 3). Interestingly, the number of phosphosites per TF was related to the TF level in the transcriptional network hierarchy (Figure 4). Yu and Gerstein observe that higher-level TFs tend to have more interaction partners and are more influential in terms of the number of genes they affect.¹³ In the context of this network hierarchy model, the observation that higher-level TFs tend to be more phosphorylated on average is a striking illustration of the role of phosphorylation in TF signal propagation on a global scale. These results may suggest a model where the phosphoregulatory arsenal is concentrated on TFs that affect the most change. The lower-level TFs are also highly phosphorylated, but there, there may be less of a need for as complex a combinatorial regulatory scheme provided by the presence of many phosphosites as on the higher-level TFs.

In the context of paralogs, TF phosphorylation has been, to our knowledge, previously unexplored. Our observation that the coexpression profile distance (Figure 5B) and binding

profile correlation (Figure 5C) is correlated with phosphorylation abundance suggests a direct role for phosphorylation in TF paralog functional divergence.

There could be the possibility that those paralog TFs that do diverge happen to be phosphorylated, without there being a causative relationship. This, however, is similar in effect to the converse. Whether the evolution of phosphosites leads to functional divergence or divergence leads to phosphorylation, phosphorylation seems to be a manifestation of both functional and sequence differentiation among many yeast TF paralogs.

One may hypothesize that some phosphosites are more inducible in response to stimuli than others.¹⁸ Further work is needed to understand which types of phosphosites modulate which signals, and how these signal transduction mechanisms are involved in paralog divergence and TF modules.

Other studies have observed a role or consequence for phosphorylation in evolution by comparing phosphorylation across species.^{10,20} Beltrao et al. found that kinase–substrate genetic interactions evolve slower than TF–gene interactions, but are more divergent than average.²⁰ We provide evidence of phosphosite effect on, or correlation with, a measure of functional and sequence divergence for TFs. This may be an example of functional innovation in eukaryotic proteins. Much work still remains to be done in deciphering the actual functional consequences of post-translational modification in greater detail.

METHODS

Phosphorylation Data

S. cerevisiae phosphorylation data was obtained from the following seven studies: Ficarro et al., Gruhler et al., Reinders et al., Chi et al., Li et al., Smolka et al., and Albuquerque et al.^{21–27} The data sets were generated under several different conditions as summarized by Supplementary Table 1 (Supporting Information). The peptides detected by Mass Spectrometry were mapped to their respective ORFs. As a result, over 10000 nonredundant positions in the proteome on over 2000 ORFs composed the *S. cerevisiae* phosphosite data set used for this study. Kinase array assay data was downloaded from Ptacek et al.¹¹ We calculated the number of kinases that showed any phosphorylation activity for each protein in the yeast proteome. We then calculated the Pearson correlation between the number of phosphosites per protein and the number of kinases per protein.

Paralog and Orthologs Identification and Comparison

The whole genome duplication (WGD) occurred ~100 million years ago and smaller scale duplication events occurred before and after the WGD.^{6,28} Those gene duplicates that were retained in the duplication events we refer to as paralogs that originated as those events. Kellis et al. and Wapinski et al. identified 437 gene pairs that were retained in the WGD and a roughly equal number of paralogs retained in SSD events.^{6,9} We use those categories in this work.

Kellis et al. and others identified *S. cerevisiae* orthologs in *K. waltii* and other species through phylogenetic analysis and synteny alignment.⁹ To assess the rate of conservation of

phosphorylated *S. cerevisiae* amino acids as compared to their corresponding orthologous positions in *K. waltii*, we aligned *S. cerevisiae* genes to their *K. waltii* orthologs using the MAFFT alignment program.²⁹ Similarly, to assess paralog sequence conservation we aligned *S. cerevisiae* paralog pairs using MAFFT. Asymmetric sequence divergence, as described in Figure 2, was assessed by aligning *S. cerevisiae* paralog 1 with the common *K. waltii* ortholog and paralog 2 with the ortholog. To accurately estimate the rates of nonphosphorylated amino acid divergence we performed sequence comparison on P random nonphosphorylated amino acids per protein where P is the number of phosphosites on the given protein. This simulation was run 103 times and the mean was used in the final calculation. This way, we minimized potential biases such as if highly phosphorylated proteins are shorter in length or were much more asymmetrically divergent.

TF Expression and Binding Comparison

The gene expression profiles for the 5885 *S. cerevisiae* profiles for the 5885 *S. cerevisiae* ORFs examined in this work were obtained from the McCord et al.¹⁴ study in which a data set of 1327 expression conditions measured by microarray was compiled from previously published studies. This data set was converted into a 5885×5885 pairwise correlation matrix. Thus, each row *i* of this matrix represents the correlation of gene *i* with all the other genes in terms of the 1327 condition expression data. The distance between gene *i* and gene *j* was calculated as the Euclidean distance between rows *i* and *j*.

TF binding data from Harbison et al. was combined by Zhu et al. with *in vitro* data measuring on TF affinities to DNA arrays as a function of the sequence of the oligos.^{15,16,30} As a result, the Zhu et al. study provided estimates of TF binding affinities for 89 TFs based on this integrated approach. The correlation of TF *i* and TF *j* binding tendencies was calculated by correlating rows *i* and *j* of the 89×5885 matrix of TF binding affinities to the genomic intergenic regions.

Statistical Analysis

All statistical analysis performed was done using the R statistical programming language. All data parsing, manipulation, and exploration was done using the Python programming language version 2.6.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Peter Philippsen, Jen Gallagher, and Doug Phanstiel for comments and helpful discussion.

REFERENCES

1. Wolfe KH, Li WH. Molecular evolution meets the genomics revolution. *Nat. Genet.* 2003; 33(Suppl):255–65. [PubMed: 12610535]
2. Sankoff D, Nadeau JH. Chromosome rearrangements in evolution: from gene order to genome sequence and back. *Proc. Natl. Acad. Sci. U.S.A.* 2003; 100(20):11188–9. [PubMed: 14506293]

3. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 2010; 11(2):97–108. [PubMed: 20051986]
4. Guan Y, Dunham MJ, Troyanskaya OG. Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics*. 2007; 175(2):933–43. [PubMed: 17151249]
5. Tirosh I, Barkai N. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol.* 2007; 8(4):R50. [PubMed: 17411427]
6. Wapinski I, Pfeffer A, Friedman N, Regev A. Natural history and evolutionary principles of gene duplication in fungi. *Nature*. 2007; 449(7158):54–61. [PubMed: 17805289]
7. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003; 422(6928):198–207. [PubMed: 12634793]
8. Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y. Posttranslational regulation impacts the fate of duplicated genes. *Proc. Natl. Acad. Sci. U.S.A.* 2010; 107(7):2967–71. [PubMed: 20080574]
9. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*. 2004; 428(6983):617–24. [PubMed: 15004568]
10. Landry CR, Levy ED, Michnick SW. Weak functional constraints on phosphoproteomes. *Trends Genet.* 2009; 25(5):193–7. [PubMed: 19349092]
11. Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R, McCartney RR, Schmidt MC, Rachidi N, Lee SJ, Mah AS, Meng L, Stark MJ, Stern DF, De Virgilio C, Tyers M, Andrews B, Gerstein M, Schweitzer B, Predki PF, Snyder M. Global analysis of protein phosphorylation in yeast. *Nature*. 2005; 438(7068):679–84. [PubMed: 16319894]
12. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007; 23(13):i549–58. [PubMed: 17646342]
13. Yu H, Gerstein M. Genomic analysis of the hierarchical structure of regulatory networks. *Proc. Natl. Acad. Sci. U.S.A.* 2006; 103(40):14724–31. [PubMed: 17003135]
14. McCord RP, Berger MF, Philippakis AA, Bulyk ML. Inferring condition-specific transcription factor function from DNA binding and gene expression data. *Mol. Syst. Biol.* 2007; 3:100. [PubMed: 17437025]
15. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004; 431(7004):99–104. [PubMed: 15343339]
16. Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* 2009; 19(4):556–66. [PubMed: 19158363]
17. Chen WG, White FM. Proteomic analysis of cellular signaling. *Expert Rev. Proteomics*. 2004; 1(3):343–54. [PubMed: 15966830]
18. Olsen JV, Blagoev B, Gnäd F, Macek B, Kumar C, Mortensen P, Mann M. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*. 2006; 127(3):635–48. [PubMed: 17081983]
19. Edmunds JW, Mahadevan LC. Cell signaling. Protein kinases seek close encounters with active genes. *Science*. 2006; 313(5786):449–51. [PubMed: 16873633]
20. Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, Burlingame AL, Krogan NJ. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol.* 2009; 7(6):e1000134. [PubMed: 19547744]
21. Ficarro SB, McClelland ML, Stukenberg PT, Burke DJ, Ross MM, Shabanowitz J, Hunt DF, White FM. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.* 2002; 20(3):301–5. [PubMed: 11875433]
22. Gruhler A, Olsen JV, Mohammed S, Mortensen P, Faergeman NJ, Mann M, Jensen ON. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics*. 2005; 4(3):310–27. [PubMed: 15665377]
23. Reinders J, Wagner K, Zahedi RP, Stojanovski D, Eyrich B, van der Laan M, Rehling P, Sickmann A, Pfanner N, Meisinger C. Profiling phosphoproteins of yeast mitochondria reveals a role of

- phosphorylation in assembly of the ATP synthase. *Mol. Cell. Proteomics*. 2007; 6(11):1896–906. [PubMed: 17761666]
24. Chi A, Huttenhower C, Geer LY, Coon JJ, Syka JE, Bai DL, Shabanowitz J, Burke DJ, Troyanskaya OG, Hunt DF. Analysis of phosphorylation sites on proteins from *Saccharomyces cerevisiae* by electron transfer dissociation (ETD) mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* 2007; 104(7):2193–8. [PubMed: 17287358]
25. Li X, Gerber SA, Rudner AD, Beausoleil SA, Haas W, Villén J, Elias JE, Gygi SP. Large-scale phosphorylation analysis of alpha-factor-arrested *Saccharomyces cerevisiae*. *J. Proteome Res.* 2007; 6(3):1190–7. [PubMed: 17330950]
26. Smolka MB, Albuquerque CP, Chen SH, Zhou H. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc. Natl. Acad. Sci. U.S.A.* 2007; 104(25):10364–9. [PubMed: 17563356]
27. Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell. Proteomics*. 2008; 7(7):1389–96. [PubMed: 18407956]
28. Byrne KP, Wolfe KH. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 2005; 15(10):1456–61. [PubMed: 16169922]
29. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 2002; 30(14):3059–66. [PubMed: 12136088]
30. Bulyk ML, Gentale E, Lockhart DJ, Church GM. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.* 1999; 17(6):573–7. [PubMed: 10385322]

Figure 1.

(A) Phylogenetic tree showing the predicted evolutionary relationship among major yeast species. Alphabetical letters (A–I) near diverging branches indicate small-scale duplication (SSD) events that are predicted to have occurred during the species divergence. Both SSD and WGD events and the resulting retained genes are as predicted by Wapinski et al.⁶ (B) Phosphosites are enriched in WGD and I category duplicates as compared to singleton genes. The number of phosphosites per gene for each duplication event (A–) and WGD was compared to the distribution of phosphosites on singleton genes. The negative log of the resulting p-values of a Wilcoxon signed-rank test is graphed for each category. We indicate the $p = 0.05$ level with a vertical line. WGD and I category duplicates are phosphorylated significantly above the singleton rate.

Figure 2.

(A) Schematic of yeast protein sequence comparison measuring asymmetric sequence evolution of *S. cerevisiae* paralogs. We compared the amino acid sequences of *S. cerevisiae* paralogs with their orthologs in *K. waltii*, a descendent of yeast that did not undergo WGD. We then performed the same comparison for amino acids that are phosphorylated. p refers to the phosphosite conservation rate, a to other amino acid conservation (see Methods). We investigated the difference between $|a_1 - a_2|$ and $|p_1 - p_2|$ as illustrated in the figure. As depicted by the length of the arrows, in some cases one paralog diverges more rapidly than the other from their common ortholog in either amino acid sequence as a whole or only phosphosite sequence. (B) Example alignment comparison. The ASK10 and GCA1 paralogs of *S. cerevisiae* were aligned to their common *K. waltii* orthologs Kwal55.20547; a sample

stretch of the alignment is depicted. The average amino acid sequence diverged fairly symmetrically in both paralogs so in this case $|a_1 - a_2|$ is small. There are 40 phosphosites between the two paralogs. Those sites are more asymmetric in their divergence: the ASK10 phosphosite conservation rate is 0.22 and the GCA1 phosphosite conservation rate is 0.35. (C) Phosphosites sequence evolution is different among paralogs, whereas nonphosphorylated amino acids evolve similarly. The distributions of the differences between the sequence conservation of the paralogs are plotted for nonphosphorylated amino acids (black) and phosphosites (green). The means of the paralog differences in amino acids conservation and phosphosite amino acid conservation are presented in the table along with the p -value of a Wilcoxon signed-rank test comparing the two distributions.

Figure 3.

(A) TFs are enriched among genes retained in WGD and several duplication events prior to WGD. We compared the fraction of TFs among singleton genes to the fraction of TFs for each duplication category. The negative log p -values of the Fisher exact test comparing the fractions are plotted. (B) TFs are more phosphorylated than the average *S. cerevisiae* gene. The number of phosphosites per gene is graphed for all genes (red) and TFs (blue) in the indicated duplication category.

Figure 4.

TF phosphorylation is related to TF regulatory hierarchy. Depicted are the *S. cerevisiae* TF regulatory network hierarchy levels as defined by Yu and Gerstein.¹³ Level 1 TFs are those that directly regulate non-TF genes whereas higher-level TFs are those that tend to regulate the expression of other TFs. The corresponding average number of phosphosites per gene is plotted next to each TF level category. The top level only contains four genes so comparisons of phosphosite distributions of that TF level were not statistically significant and thus not included. The *p*-values of a *t* test comparison of level 1 phosphosites per gene distribution to levels 2 and 3 are indicated to the right of the graph. The differences in the means suggest that level 2 and 3 TFs are more phosphorylated on average than level 1 TFs. The *t* test *p*-values do not reject this hypothesis.

Figure 5.

(A) Distance between gene coexpression vectors correlates with the number of phosphosites for TF paralogs but not for non-TF paralogs. We plotted the distance between the gene coexpression vectors described against the total number of phosphosites for a given paralog pair for pairs where both are TFs (blue) and pairs where neither are TFs (red) (left plot). The correlations between the coexpression distances and the number of phosphosites are presented in the leftmost table along with p -values (t test). The middle panel shows the comparison between phosphorylation number and sequence divergence between the paralogs, and the right panel is the comparison of sequence divergence vs coexpression distance. (B) TF binding divergence correlates with phosphosite number. We calculated the correlation among TF paralog pairs of their binding profiles. We investigated those paralog pairs where both paralogs belonged to the 89 TFs whose intergenic region binding profiles were calculated by Zhu et al.¹⁶ These correlations were plotted against the total number of phosphosites on the TF pair. The more phosphorylated a TF paralog pair, the lower the correlation of binding profiles among the pair (left panel). Similarly, the number of

phosphosites correlates with sequence divergence (middle panel) and sequence divergence anticorrelates with the TF binding correlation metric (right panel).