2013

Network analyses of proteome evolution and diversity

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

NETWORK ANALYSES OF PROTEOME EVOLUTION AND DIVERSITY

by

JASMIN COULOMBE-HUNTINGTON

B.S., McGill University, 2006
M.S., McGill University, 2008

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2013

Approved by

First Reader

_____

Yu Xia, PhD
Assistant Professor of Chemistry

Second Reader

_____

Simon Kasif, PhD
Professor of Biomedical Engineering

Acknowledgements

I have learned a great deal during my time in the Bioinformatics program. I am eternally grateful to my advisor, Yu (Brandon) Xia, for sharing his beautiful view of science with me and opening my eyes to a world of possibilities. I also wish to thank the members of my committee and members of the Xia lab for their very insightful comments and discussion.

I of course thank Xinping Yang for conducting the experiments discussed in Chapter 3, as well as his help in writing, creating the figures and contributing figure panels 3.1A, 3.1B, 3.1D, and 3.4H. I thank Shuli Kang for assembling the isoform sequences in Chapter 3, computing betweenness centrality and contributing figure panels 3.1E, 3.2A and 3.2B. I also thank Marc Vidal and the Center for Cancer Systems Biology for supporting the project which lead to the results presented in Chapter 3.

My work was funded in part by an NSERC scholarship and I have the Bioinformatics program to thank for supporting my participation in scientific conferences around the world. I am also grateful to the entire Bioinformatics program staff, notably David King, for doing an excellent job of keeping me up to date on all the paperwork.

Lastly, I want to thank my wife, Véronik, for her emotional support and unwavering belief in me all that time.

NETWORK ANALYSES OF PROTEOME EVOLUTION AND DIVERSITY

(Order No.)

JASMIN COULOMBE-HUNTINGTON

Boston University, Graduate School of Arts and Sciences
and College of Engineering, 2013

Major Advisor: Yu Xia, Assistant Professor of Chemistry

ABSTRACT

The mapping of biomolecular interactions reveals that the function of most biological components depends on a web of interrelations with other cellular components, stressing the need for a systems-level view of biological processes. This work explores ways in which the integration of evolutionary, functional and network information can lead to a better understanding of eukaryotic biology.

Studying yeast, we show that transcription factor (TF) evolutionary rate follows regulatory network structure, with the strongest correlate being the median evolutionary rate of activated target genes rather than protein abundance or protein-protein interaction degree. We also found that targets of fast-evolving TFs underwent greater evolutionary expression changes and are enriched for other TFs and environment-specific functions as compared to targets of slow evolving TFs. This work highlights the importance of *trans*-regulatory network evolution in species-specific gene expression and network adaptation.

Next, considering gene loss and gain across fungal evolution, we explore the evolutionary life-cycle of eukaryotic genes. By integrating network and functional information, we reveal that network marginalization of genes tends to precede gene loss. We discovered that lost or gained genes are enriched in TFs and have significantly different network properties than universally conserved genes, including a greater number of transcriptional regulators. These results indicate a highly active role of transcriptional network rewiring in gene integration, marginalization and species-specific adaptation.

The final chapter explores how alternative splicing (AS)-driven expansion of human proteome diversity leads to system-level complexity through the AS-mediated rewiring of the protein-protein interaction network. By overlaying network, functional and expression datasets onto the first large-scale isoform-resolution interactome, we found that differentiating between splice variants is essential to capturing the full extent of the network's modularity. We show that AS-mediated rewiring preferentially affects tissue-specific genes and that different rewiring patterns may have distinct functional consequences. Furthermore, we found that most rewiring can be traced to the AS of evolutionarily conserved splicing modules, which promote or inhibit interactions and tend to overlap linear motifs and disrupt known domain-domain interactions.

In summary, by studying the adaptations of proteomes across tissues and through evolution, we uncovered global principles of network organization and condition-dependent regulation.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AS	Alternative splicing
CAI	Codon adaptation index
ChIP	Chromatin immuno-precipitation
HGT	Horizontal gene transfer
PGL	Propensity for gene loss
PPI	Protein-protein interaction
ORF	Open reading frame
TF	Transcription factor

Introduction

Thanks to advances in nucleotide sequencing, we have seen an accelerated pace in the cataloging of biological components such as genes and proteins. Since proteins express their functions within specific biological contexts, including interactions with other cellular components, this simple list of components is insufficient for achieving a full understanding of biological processes and a dynamic system-level view of protein function is required. Fortunately, there are promising avenues which offer hope of understanding this complex web of inter-relations and dynamics. Genome-scale functional assays such as protein-protein, protein-DNA or genetic interaction mapping are beginning to reveal part of the network of inter-relationships relating biological components. Along with considering the network context of proteins, studying the evolutionary and condition-dependent dynamics of proteomes, such as amino acid substitutions, gene gain, gene loss and alternative splicing, also provides a powerful viewpoint into the importance of biological components and their inter-relations. The richness of both biological networks and proteome dynamics as informational resources becomes clear when they are integrated together with other biological data. In this dissertation, through the integration of diverse network, functional, expressional and alternative splicing data, as well as the evolutionary landscapes of proteomes, we reveal principles underlying the organization of the networks and regulatory systems which ultimately determine the role of individual proteins in the cell.

Much of our understanding of biology comes from studying naturally occurring variation within and between organisms. The content and expression levels of proteomes vary significantly between species, as well as between conditions and tissues. By considering both similarities and differences between proteomes in the context of functional and biological network information, we can discover important insights about biological functions and their inter-relations. For example, the selective pressures acting on individual proteins is in part determined by their number of interactions with other proteins (Fraser et al., 2002; Kim et al., 2006), as well as the selective pressures acting on their network neighbors in various types of biological networks (Wang & Lercher, 2011). Other than mutations in protein sequences, many other types of proteomic changes contribute to phenotypic diversity and condition-specific adaptations, including changes in gene expression levels, alternative splicing and gene gain and loss. Placing these changes in a network and functional context can uncover global organizing principles of both proteomes and interactomes.

In order to fully understand phenotypic variations between species, we must understand the genetic determinants of gene expression and how they evolve over time. The evolution of transcriptional regulatory networks has thus far mostly been studied at the level of *cis*-regulatory elements, such as the loss or gain of transcription factor (TF) binding sites. However, *trans*-level variations have been shown to contribute significantly to the expression differences between yeast strains (Tirosh et al., 2009). Therefore, to gain a complete understanding of regulatory network evolution we must also study the

evolutionary role of *trans*-factors, such as transcription factors (TFs). In an effort to better understand the evolution of gene expression through *trans*-regulatory network rewiring as well as the relation between biological networks and evolutionary processes, the first part of this dissertation (Chapter 1) details the systematic analysis of the determinants of TF evolutionary rate in yeast. Specifically, we assess genomic and network-level determinants of TF evolutionary rate in yeast, and how they compare to those of generic proteins, while carefully controlling for differences within the TF protein set, such as expression level. We found that the typical determinants of protein evolutionary rate, such as expression level and protein-protein network interactions have a very different influence on TF evolutionary rate. We found that TF evolutionary rate is most highly correlated to the evolutionary properties of the genes which they regulate and specifically genes which they activate, including median target gene evolutionary rate and the fraction of species-specific target genes. We show that fast evolving TFs tend to regulate other TFs and environment-specific processes and that their targets show larger evolutionary expression changes than targets of other TFs, demonstrating that TF evolutionary rate predicts actual evolutionary expression differences of regulated genes We also show that the positive trend relating TF regulatory in-degree and evolutionary rate is likely related to the species-specificity of the transcriptional regulation modules. Finally, we discuss likely causes for TFs' different evolutionary relationship to the physical interaction network, such as the prevalence of transient interactions in the TF subnetwork or TFs' potential role in adaptive evolution. This work shows that positive and negative regulatory networks follow very different evolutionary rules, and that TF

evolution is best understood at a network- or systems-level. Having shown that TF evolutionary rate is important to the evolution of gene expression levels, this analysis uncovered significant evidence for the modular evolution of transcriptional regulators with their target genes, demonstrating a novel relation between network structure and the selective pressures acting on genes.

Alongside mutations in protein sequences, gene gain and loss constitute some of the most important evolutionary processes creating phenotypic diversity between species, shaping both proteomes and the networks they form. In Chapter 2, we delve further into the evolution of proteomes, systematically applying knowledge of biological networks to the study of gene gain and loss in yeast. While a few network-level analyses have provided insights into gene gain and others into gene loss, an integrative view of the gene evolutionary life-cycle, considering both gene gain and loss in the context of the different types of biological networks available, is still lacking. Here, using orthology mappings across 23 ascomycete fungi genomes, we identify proteins that were gained throughout evolution, those that were lost, as well as proteins which were universally conserved across the tree, defining the “core” genome of the phylum. By overlaying diverse network and functional genomics datasets from the model yeast *S. cerevisiae*, we discover that gain and loss of genes is tightly coupled to the gain and loss of edges in the different networks. We found that an increased propensity for gene loss along a lineage is associated with the progressive network marginalization of genes through network rewiring. By integrating genetic interaction network data with the history of gene gain

and loss we reveal the significantly higher functional independence or poor functional integration of lost and gained genes. We also make the surprising discovery that core genes tend to have fewer regulators than lost or gained genes, demonstrating a link between species-specific adaptation and regulatory complexity. Furthermore, we found that TFs were more likely to have been lost or gained throughout fungal evolution than other genes. Taken together, this systems-level view of the life-cycle of eukaryotic genes reveals the complex interplay between the different levels of cellular organization and species-specific gene importance. This work also highlights the unique and highly active role of regulatory network rewiring, at both the *cis*- and *trans*- regulatory levels, in the processes of gene integration and marginalization. It also exposes some of the global principles of interactome organization and evolution.

Yeast is studied primarily as a model for understanding human cells. However, human cells posses an important additional level of proteome complexity which has no counterpart in yeast. Alternative splicing and alternative transcription vastly expand the diversity of the human proteome and its dynamics across tissues, yet it remains unclear how this diversity contributes to system-level or phenotypic complexity. This problem calls for the application of high-throughput functional assays in order to understand the role of alternative splicing in a system-level context. Targeted studies have demonstrated that AS can turn on or off individual protein-protein interactions (Ellis et al., 2012; Thakar et al., 2012; Wethkamp et al., 2011), establishing a need for an isoform-resolution interactome network to understand the genome-scale influence of AS on the protein-

protein interaction (PPI) network. Chapter 3 details the construction and analysis of the most comprehensive isoform-resolved interactome to date. Through a collaboration with the lab of Dr. Marc Vidal at the Dana-Farber Cancer Institute, we carried out systematic cloning of native splice isoforms from human tissues, followed by genome-scale binary PPI screening against the human ORFeome and pairwise retesting. This isoform-resolved interactome network reveals a deeper layer of modularity in terms of network structure and functional organization than gene-level networks. In addition, this network enables the classification of AS-mediated rewiring patterns which differentially affect network structure and function. In particular, isoforms participating in mutually exclusive interactions define a functionally distinct class of genes, which plays a key role in network rewiring between network modules and tissue types. PPI rewiring events tend to affect tissue-specific proteins and are associated with the alternative inclusion of localized sequence modules, promoting or blocking interactions at comparable frequency. These interaction-regulating modules are evolutionarily conserved, enriched in linear motifs and many disrupt known domain-domain interactions. Taken together, these results demonstrate that AS plays a major role in the organization, function, and cross-tissue dynamics of biomolecular networks.

Chapter 1

Regulatory network structure as a dominant determinant of transcription factor evolutionary rate

1.1 Introduction

The study of regulatory network evolution has so far mostly concentrated on *cis*-regulatory variation, such as the loss or gain of transcription factor (TF) binding sites in the promoter region of a gene. But *trans*-level variations are known to account for a significant amount of the expression variation between yeast strains (Tirosh et al., 2009). TFs are central to decision making in cells, with roles ranging from environmental adaptation in unicellular organisms to controlling cellular differentiation and endocrine response in higher eukaryotes. TFs' unique role might have been exploited by evolution to modulate the activity of entire pathways or re-wiring of the cellular network. An example of pathway activity modulation is the shutdown of the flagellar pathway in non-motile bacterial species through the deletion of the TF activating the pathway (Hershberg & Margalit, 2006). An example of network rewiring through TF protein evolution is how a mutation in Ubx, a Hox protein, led to the loss of a subset of its targets, and is believed to have allowed the transition to a hexapod body plan (Galant & Carroll, 2002; Ronshaugen et al., 2002). Attesting to the usefulness of *trans*-level variation in evolutionary adaptation is the observation that TFs underwent significantly more positive

selection along the human and chimp lineages than other genes (Clark et al., 2003) and significantly more TFs had differential expression between the two species (Clark et al., 2003; Gilad et al., 2006).

The systematic mapping of molecular interactions between pairs of proteins and between proteins and DNA has unveiled a world of complexity not captured by a simple biological parts list. A useful approach to better understand protein functions and relations is to look at these networks through the lens of evolution. Evolutionary rate, typically defined as the ratio of non-synonymous to synonymous substitution rates (K_a/K_s), represents the level of tolerance to mutations of proteins across evolution and can reveal additional information about these networks. For example a strong trend was discovered relating protein-protein interaction (PPI) degree to evolutionary rate (Fraser et al., 2002; Kim et al., 2006), which suggests that physical interactions lead to evolutionary constraints on the protein sequence. Such genome-wide trends however ignore the underlying diversity of the many subnetworks which constitute the global network. Since genes have very distinct functions in the cell and often act together as functional modules, we might expect that the global trends not always hold for all subnetworks. Recent work by Jovelin *et al.* and by Wang *et al.* showed that the TF subnetwork evolved distinctly from the global network between closely related yeast species (Jovelin & Phillips, 2009; Wang et al., 2010). More specifically, it was found that the number of physical interactors or transcriptional regulators correlates much more positively with evolutionary rate than is expected from the genome-wide trend. These previous studies

established that subnetworks can display trends which differ significantly from those of the global network and specifically highlighted TFs as a uniquely evolving gene set.

However, no study to date has looked at how TF evolution may be influenced by other factors that are known to be important in the evolution of generic proteins, such as expression level or the evolutionary rate of network neighbors. In an effort to better understand the unique evolutionary properties of TFs, we conducted a systematic comparison of key determinants of protein evolutionary rate between TFs and generic proteins. The recent increase in the number of fully sequenced species allows us to study short term evolution, before the regulatory network has had much time to rewire. We looked at the coding sequence evolution between *S. cerevisiae* and its closest sequenced relative, *S. paradoxus*. For *S. cerevisiae*, an extensively studied model species, we have access to genome-wide protein-protein and protein-DNA interaction networks, as well as mRNA expression datasets.

Transcriptional networks display extensive evidence of modularity at the functional level. For example, it is well known that metabolic enzymes participating in a common pathway are often regulated by a common TF (Ihmels et al., 2004; Zaslaver et al., 2004). In multicellular organisms, TFs are used to regulate tissue-specific gene expression and execute specific developmental or stimulus response programs. This functional modularity may be detectable at the evolutionary level. To test this hypothesis, we examined how TF evolutionary rate relates to the regulatory network structure, in

particular how the evolutionary properties of target genes influence TF evolution. The average evolutionary properties of proteins regulated by a common TF could serve as a proxy for the amount of selective constraint acting on a transcriptional module. Since the role of a TF is defined through its transcriptional target genes and the way it regulates them, the selective constraint on a TF is expected to be proportional to the selective constraint on the transcriptional module it regulates. This network-centric function of TFs could be at the source of their distinct evolutionary trends. Since the transcriptional network is made up of a combination of activating and repressive regulatory relationships, which could have fundamentally different effects on the evolution of TFs and other genes, we also explored the effects of regulatory sign on TF-target evolutionary relationships.

1.2 Results

1.2.1 The effect of PPI network degree on TF evolution

Protein-protein interactions (PPIs), by imposing additional functions on the structure and interface residues of interacting proteins, often lead to increased selective constraints on these proteins (Kim et al., 2006). This largely explains the empirical observation that proteins with more binding partners tend to evolve at a slower rate (Fraser et al., 2002). The slope of the correlation between network degree and protein evolutionary rate can be interpreted approximately as the average evolutionary pressure on proteins contributed by one network edge, or interaction interface. This effect has been shown to be different within the TF subnetwork than within the global network (Wang et al., 2010). Here, we

re-examined the statistical significance of the previous result using an improved method, described in detail in the Methods section, which avoids specific biases by controlling for the different average degree and evolutionary rate of TFs as compared to generic proteins. We used the ratio of the non-synonymous substitution rate (K_a) over the synonymous substitution rate (K_s), or K_a/K_s, between *S. cerevisiae* and its closest known cousin *S. paradoxus*, as a measure of protein evolutionary rate. As a normalization step, we transformed evolutionary rate and PPI degree into genome-wide ranks for all yeast protein-coding genes. We then calculated the slope for the 174 TFs (list taken from (Wang et al., 2010)) and compared it to a distribution of slopes obtained from random protein samples with the same average degree and evolutionary rate as TFs to within 1% root-mean-square deviation (RMSD). Figure 1.1 shows how genome-wide ranks are preserved when calculating the slope for the TF subset to allow the relative incline to be compared between TFs and generic proteins. We found that the average effect of PPIs on TF K_a/K_s was significantly less pronounced than expected from the sampling procedure, with a p-value of 0.0085. Replacing the evolutionary rate with codon adaptation index (CAI), which allows us to control for the effect of expression, results in a p-value of 0.26, suggesting that expression differences are not driving the different evolutionary rate trend. Repeating the comparison using edges reported in two or more independent experiments, which we term confirmed edges, returns a p-value of 0.0026, confirming that TF evolution is differentially affected by PPI degree as compared to generic proteins. The fact that the number of interaction partners does not influence TF evolutionary rate as strongly as it does for other proteins is potentially explained by the TF subnetwork's

enrichment in transient interactions reusing the same binding interfaces, and depletion of stable complex formations requiring a different binding interface for each partner. Supporting this hypothesis, we used a chi-square test and the Gene Ontology (GO) (Ashburner et al., 2000) term “protein kinase activity” and showed that, compared to other proteins, a significantly greater fraction of TF PPIs involve kinases (2.1-fold enrichment; $p=1.87\times10^{-61}$), which are known to bind transiently. Another contributing factor could be the fact that TFs, which are often bound to the DNA, tend to interact with proteins which are themselves bound to DNA. The greater proximity induced by this tethering reduces the entropy of the unbound state, allowing the protein-protein interaction to be mediated by a relatively weaker binding affinity and thereby relaxing the level of selective constraint imposed by these PPI interfaces. As support for this hypothesis, we showed that a significantly greater fraction of TF PPIs are with DNA binding proteins (2.2-fold enrichment; $p=9.0\times10^{-218}$), using a chi-square test and the GO term “DNA binding”.

1.2.3 The effect of regulatory in-degree on TF evolution

Similarly to PPI degree, but with a much weaker correlation, generic proteins with more regulators (higher in-degree) tend to evolve slower. In contrast, the effect of regulatory in-degree on TFs has been shown to be opposite, with each additional regulator contributing on average towards faster evolution of the TF (Wang et al., 2010). Using our new method and a regulatory network based on a collection of ChIP-chip studies (Teixeira et al., 2006), we confirmed the earlier finding that the slope relating TFs’

regulatory in-degree and evolutionary rate is significantly more positive than expected by chance (p=0.0093, CAI p=0.042, confirmed edges p=0.0041). The opposing trends relating in-degree and K_a/K_s for all proteins and TFs are shown in Figure 1.2A and Figure 1.2B, respectively. To understand why high in-degree TFs tend to evolve at a faster rate, we decided to look at the genes they regulate. Although the median evolutionary rate of target genes is not significantly associated to the in-degree of regulators, we found that TFs' in-degree significantly correlates with the fraction of target genes which are missing an ortholog in the comparison species (ρ=0.20 p=0.016; confirmed edges ρ=0.21 p=0.041), *S. paradoxus*. These results suggest that the regulatory in-degree of TFs is tied to the species specificity of the transcriptional modules they regulate. High in-degree TFs may be more likely to undergo reduced negative selection than low in-degree TFs because the impairment of their regulatory functions is less likely to disrupt core processes. At the same time, high in-degree TFs may be more likely to undergo enhanced positive selection because they tend to regulate more species-specific functions.

1.2.4 The effects of expression level, CAI and PPI network neighbors on TF evolution

Since the trends relating TF evolutionary rate to network degree are significantly distinct from the genome-wide average, we decided to probe whether TF evolutionary rate is differentially affected by other well known correlates, such as mRNA expression level or the evolutionary rate of protein interaction partners. Using our method, we compared the slope of TFs relating K_a/K_s and mRNA expression level from RNA-seq in rich media

(Ingolia et al., 2009) to the slopes produced from random protein sets of the same size, matched for average expression level and evolutionary rate (see Methods). The results show that the trend relating TF K_a/K_s to expression level is too flat to be due to chance (p=0.0025), even accounting for TFs' lower average expression level. TF K_a/K_s is also much less correlated than expected to CAI ($p \leq 0.0001$), a commonly-used surrogate for expression level. This suggests that expression level imposes weaker selective constraints on TFs than on other genes. A similar lack of correlation is also apparent between TF K_a/K_s and the median K_a/K_s of protein-protein interaction (PPI) partners. TF K_a/K_s is too weakly correlated to that of its PPI network neighbors to be the result of chance ($p \leq 0.0001$). This difference is probably related to the greater fraction of TFs involved in transient interactions and thus less likely to co-evolve with their interaction partners. These results demonstrate yet again that TFs are subject to a unique set of evolutionary pressures. Figure 1.2 shows some of the most striking differences in TF evolutionary rate correlations. In addition to other explanations, it is possible that TF evolutionary rate shows weaker correlation to many features because of the dominant influence of other determinants on TF evolution, such as their role in the regulatory network.

1.2.5 The effect of target gene evolutionary rate on TF evolution

To understand the evolutionary behavior of TFs, it is imperative that we study the evolution of their target genes. The function of TFs is inherently expressed through the regulation of their target genes and this network-centric role of TFs might be what distinguishes their evolution from that of other proteins. Using the ChIP-chip based

regulatory network and Spearman's rank correlation coefficient (ρ), we asked whether median target evolutionary rate was predictive of TF evolutionary rate. As shown in Figure 1.3A, we discovered that the evolutionary rate of TFs significantly follows the median rate of its target genes (ρ=0.25, p=0.0033), suggesting that TFs and their target genes constitute co-evolving modules. Figure 1.4A shows that the correlation holds using K_a/K_s values obtained from comparing *S. cerevisiae* to its next closest sequenced cousin, *S. mikatae* (ρ=0.23, p=0.0059). We also confirmed the significance of this effect using the network of confirmed edges (ρ=0.23, p=0.020) and using an alternative regulatory network based entirely on literature curation of small-scale experimental studies (Teixeira et al., 2006) (ρ=0.26, p=0.0018), henceforth referred to as the literature curated network. As shown in Figure 1.5, K_a/K_s itself cannot be used to predict regulatory interactions in general, but it does provide some predictive power in the TF subnetwork (predicting TFs that regulate TFs). Furthermore, we show that targets of the same TF in the network of confirmed edges tend to have closer than expected evolutionary rates (p=0.011) and mRNA expression levels (p=1.13×10^{-4}), using the Wilcoxon rank-sum test (see Methods for details) than targets of different TFs. Although the co-evolution of co-regulated genes is easily explained by their similar expression levels, the co-evolution of TFs and their target genes indicates that TF evolution is directly influenced by their position and role in the regulatory network.

1.2.6 The effect of target gene loss or gain on TF evolution

Evolutionary rate is not the only evolutionary measure of protein importance. We also looked at the fraction of target genes missing an ortholog in the closest yeast species, *S. paradoxus*, indicating the gene was either lost in *S. paradoxus* or gained in *S. cerevisiae*. We discovered that the fraction of target genes missing in *S. paradoxus* is correlated to TF evolutionary rate (ρ=0.22, p=0.0091; Figure 1.3B). The correlation was confirmed using *S. mikatae* as the comparison species (ρ=0.23, p=0.0081), as displayed in Figure 1.4B. The result also holds using the network of confirmed edges (ρ=0.24, p=0.021) and the alternative literature curated network (ρ=0.24, p=0.0042). These results suggest that the evolutionary rate of TFs is tied to the species specificity of the transcriptional modules they regulate. TFs regulating species-specific modules tend to evolve faster, as a result of either relaxed negative selection or enhanced positive selection.

1.2.7 The relative contribution of TF evolutionary rate correlates

In addition to separately assessing the genomic and network correlates of TF evolutionary rate, it is important to compare their relative contributions to identify the most dominant determinants of TF evolutionary rate, and whether they differ from those of generic proteins. Figure 1.6 shows the Spearman's rank correlation coefficients (ρ) relating different genomic and network properties to K_a/K_s for TFs and for all proteins. This Figure 1.clearly shows how features like expression, CAI, which is tightly coupled to expression (Sharp & Li, 1987), and PPI degree dominate the evolutionary rate determinant landscape of average proteins. In contrast, median target K_a/K_s dominates the

TF landscape, with other regulatory network properties playing an important role, such as in-degree, median regulator K_a/K_s and the fraction of target genes missing in *S. paradoxus*. This shows that the regulatory network structure is the most important factor determining TF evolutionary rate, suggesting that the function and evolution of TFs is primarily defined at the network level. The dominance of this so far overlooked relationship between TF and target evolution could also potentially explain the eccentricity of other TF evolutionary trends. The observation that TFs have significantly different evolutionary rate determinants was confirmed individually for each variable earlier in the Results section, using sampling of random proteins and rigorous statistical tests as described in the Methods section.

In contrast to random samples, other functionally defined subsets of proteins may also possess a different landscape of evolutionary rate determinants. As case examples, Figure 1.7 shows the same evolutionary rate determinant correlation coefficients for the 240 proteins in the GO term “signal transduction” and for 540 metabolic enzymes taken from the YeastCyc database (Caspi et al., 2008). We see that these functionally defined categories have similar overall evolutionary rate determinant profiles to that of generic proteins in Figure 1.6, with abundance and PPI degree dominating the landscape, suggesting that TFs are unique in this regard among functionally defined protein subsets. The only notable exception is the lack of correlation between signal transduction protein K_a/K_s and its median interactor K_a/K_s, which is consistent with our theory that this effect in TFs may be related to the transience of many interactions in the subnetwork.

1.2.8 Controlling for potential relationships between other TF and target properties

Since target K_a/K_s is apparently the strongest determinant of TF evolutionary rate, it is important to look for potential relationships between key TF and target properties, such as mRNA expression level and PPI degree, to rule out potential confounding effects. Table 1.1 shows the Spearman's rank correlation coefficients relating different TF and target properties. We have repeated each one of these correlations using the network of confirmed edges and using the literature curated network (Teixeira et al., 2006), the results of which are shown in Tables 1.2 and 1.3, respectively. This analysis reveals that median target K_a/K_s remains the strongest predictor of TF K_a/K_s over other important target properties.

Since TFs are often regulated post-translationally, target gene expression has been used in studies to estimate the level of TF activity (Boorsma et al., 2008; Pournara & Wernisch, 2007). Although consistently negative, the correlation between target expression and TF K_a/K_s was only found to be significant using the literature curated network. This result suggests that TF activity as estimated from target gene expression cannot be the only driving force behind the modularity of TF-target evolution. Further studies are needed to investigate the role of TF activity in determining TF evolutionary rate.

1.2.9 TF evolution and target gene function

Having found that TF evolutionary rate is related to the evolutionary rate and species-specificity of target genes, we may expect a similar relation between TF evolutionary rate and target gene function. We looked for enrichments of large GO terms (involving 50 or more target genes) in the targets of the 25% fastest evolving TFs with targets, as compared to targets of other TFs using Fisher's exact test. Table 1.4 shows the enrichments with a p-value below 0.05. Most GO terms that were significantly enriched in targets of fast evolving TFs are indicative of niche-specific functions, such as transporter activity, oxidation-reduction processes, and localization to the extracellular region, plasma membrane or cell periphery, as well as categories likely to show niche-specific expression, like carbohydrate metabolism. Most interestingly, we found that fast evolving TFs were also more likely to regulate other TFs, suggesting that the hierarchical structure of the regulatory network may be useful for adaptive evolution. These results suggest that TF evolution potentially serves as a mechanism for species-specific environmental adaptation through its effect on the expression of multi-gene modules.

1.2.10 TF evolution and the evolution of target gene expression

The role of *trans*-regulatory gene evolution on gene expression is inherently more difficult to study than *cis*-regulatory evolution since the former requires knowledge of the regulatory network structure. To confirm that the evolutionary rate of TFs is related to measurable *trans*-regulatory changes in the gene expression of target genes, we used previously published RNA-seq data from both *S. cerevisiae* and *S. paradoxus* (Busby et

al., 2011). Using the network of confirmed ChIP-chip edges, we found that targets of the top 25% fastest evolving TFs had, on average, larger expression differences between the two species than targets of other TFs, as shown in Figure 1.8 (t-test p=0.00013, see Methods for details). This result confirms that TF evolutionary rate can serve to predict real *trans*-regulatory expression changes of gene modules, which could in turn lead to important phenotypic effects.

1.2.11 The effect of regulatory sign on TF-target co-evolution

Regulatory networks are composed of two inherently distinct edge types, activating (or positive) edges and repressive (or negative) edges, which could potentially play divergent roles on the evolutionary modularity of the network. We used previously published TF knock-out microarray data (Hu et al., 2007) to infer the sign of ChIP-chip based regulatory network edges. Using the microarray fold-changes (see Methods for details), we were able to infer the mode of regulation for 4,010 of the ChIP-chip regulatory edges, 2,628 activating and 1,382 repressive. By overlaying these two datasets, we decomposed the network into positive and negative regulatory subnetworks and studied how the mode of regulation affects TF-target evolutionary relationships. For TFs with 5 or more targets of the same regulatory sign, we found that median K_a/K_s of activated targets significantly follows TF K_a/K_s (ρ=0.26, p=0.0036), while median K_a/K_s of repressed targets shows no significant correlation (ρ=0.068, p=0.46). We also found that TF K_a/K_s predicts the

fraction of activated targets which are missing in the comparison species *S. paradoxus* (ρ=0.29, p=0.0038) but not for repressed targets (ρ=-0.079, p=0.52). Table 1.5 shows the correlation coefficients and associated p-values for activating and repressive networks, where transcriptional edges are inferred either from ChIP-chip or from literature curation of small-scale experimental studies (Teixeira et al., 2006). As shown in Table 1.6, both the significance of the activating edge relations and the lack of a significant trend for repressive edge relations were confirmed using the literature curated network. Figure 1.9 shows how activated and repressed target evolutionary properties have a different effect on TF K_a/K_s. These results demonstrate that TFs evolve in synchrony with the targets they activate but not the targets they repress.

1.3 Discussion

Protein sequence evolutionary rate provides a unique viewpoint into both the importance and the functional relationships between genes and proteins. In this study, we have demonstrated how the function of TFs in the regulatory network is more important in understanding TF evolution than any other property measured. Demonstrating how TF sequence evolution plays an important role in the evolution of gene expression, we have shown that targets of fast evolving TFs are more likely to see their expression change through evolution. We found that TF evolutionary rate is determined by very different rules than that of generic proteins, possessing a unique correlation to expression level, CAI, median evolutionary rate of PPI network neighbors and, as previously reported, to PPI degree and regulatory in-degree. This evidence demonstrates how TFs are subject to their own set of evolutionary pressures.

We have also demonstrated that TF evolutionary rate is strongly related to the evolutionary properties of their target genes, such as evolutionary rate and species-specificity. Remarkably, this network-level influence on TF evolutionary rate trumps even that of gene expression. The fact that TFs and their targets tend to evolve as modules is consistent with similar findings in other types of biological networks. It has previously been reported that neighbors in many types of biological networks tend to evolve at similar rates (Wang & Lercher, 2011), including PPI networks (Fraser et al., 2002), co-expression networks (Carlson et al., 2006), genetic interaction networks (Costanzo et al., 2010) and metabolic networks (Vitkup et al., 2006). What we have demonstrated here is that neighbor genes also tend to evolve at similar rates in the transcriptional network (TFs and their targets) and co-regulation network (genes regulated by a common TF). It is important to note, since TFs have low expression and are often regulated post-translationally, that the regulatory network is the one network among these (including the co-regulation network) for which the co-evolution of protein sequences is the least likely to be explained by similar expression levels.

The lack of correlation with CAI and expression level is especially surprising since it has been thoroughly established that protein abundance is by far the strongest predictor of protein evolutionary rate (Drummond et al., 2006; Fraser et al., 2002; Xia et al., 2009). This is believed to relate to the increased pressure for proper folding and translational accuracy in highly expressed proteins (Drummond et al., 2005). Since TFs' distinct trend

is not explained by their lower average expression level, the difference is likely related to TFs' cellular role. Their expression levels may be subject to more variation across species, as shown across the human-chimp lineage (Gilad et al., 2006). TFs could also be subject to other dominant evolutionary constraints, such as their network-level role.

Looking for unique features of TFs which could explain their distinct evolutionary trends, we found evidence suggesting TFs may play a special role in adaptive evolution. We have shown that targets of fast evolving TFs are more likely to show differential expression between the two yeast species and that targets of these same TFs are also more likely to be involved in environment-specific functions. TFs are themselves more likely to be regulated by fast evolving TFs, suggesting the possibility that adaptive evolution has taken advantage of the hierarchical structure of the regulatory network to achieve desirable phenotypic changes more efficiently.

Another strikingly unique feature of TF evolution is the positive trend relating TF evolutionary rate to regulatory in-degree, while other proteins show a negative trend. Here, we found that this positive trend appears to be a module-level trend, with TF in-degree affecting not only the TFs evolutionary rate but also the species-specificity of genes regulated by those TFs. The fact that TFs were more likely to be regulated by fast evolving TFs than other genes could also help explain this trend, especially considering that TF evolutionary rate is much more sensitive to the evolutionary rates of their regulators than is the case for other proteins, as shown in Figure 1.3. To gain further

insights into the relation between TF in-degree and target function, we calculated the enrichment of GO term memberships comparing targets of high in-degree TFs (≥10 regulators) to targets of low in-degree TFs (≤2 regulators) using Fisher's exact test, considering GO terms with 50 or more targets. As shown in Table 1.6, the GO terms that were significantly enriched for targets of high in-degree TFs are very similar to those of fast evolving TFs, centering around peripheral or niche-specific functions, such as plasma transmembrane transport. These results suggest that TFs regulating niche-specific genes tend to have higher in-degree in part to allow for the integration of environmental signals.

When we decomposed the regulatory network into positive and negative regulatory subnetworks, we found that only positive regulatory relationships predict co-evolution of TFs and their targets. A study by Hershberg *et al.* supports that there are distinct evolutionary pressures on activator and repressor TFs in relation to their role in the transcriptional network. They discovered by comparing different strains of bacteria that activators are more likely than repressors to be lost before all their targets are lost (Hershberg & Margalit, 2006). They suggested that the loss of activator TFs was an "efficient means of shutting down unused pathways". This draws a picture where activator TFs can be used by evolution as on/off switches affecting the activity of multi-gene modules, thereby avoiding the need to silence each gene through individual mutations. The loss of repressors however tends to be avoided regardless of the usefulness of the genes they regulate. Losing a repressor would likely lead to the untimely expression of genes, which will incur an energetic cost and potentially disrupt

homeostasis. Similarly to the loss of a TF, mutations in the protein sequence of a TF are likely to impair the function of that TF. In the case of an activator, this would lead to reduced expression of regulated genes. We therefore expect the more conserved genes modules to be regulated by more conserved TFs, and vice-versa. In the case of a repressor, mutations in its protein sequence would likely lead to the over-expression of target genes, which due to resource expenditure and/or dosage sensitivity can be damaging to the cell independently of the evolutionary importance of target genes. This would explain why the evolutionary rate of repressors is largely independent of the evolutionary properties of target genes. Our results are consistent with Hershberg *et al.*'s earlier findings, but suggest that the loss of activator TFs is an extreme example within the wider spectrum of activator TF protein evolution, which can likely be involved in more subtle and varied modulations of pathway activities than a simple on/off switch. This new perspective on the evolution of *trans*-regulatory gene expression control confirms that positive and negative regulatory subnetworks are subject to very different evolutionary pressures at the regulatory network-level.

This work details the uniqueness of TF evolutionary rate determinants and is the first to establish the modularity of TF-target protein evolution. This new awareness sheds much needed light on the eukaryotic evolution of *trans*-level control of gene expression through TF protein evolution and may help us better understand how subtle differences at the protein level can lead to pathway level variation between species. We also demonstrated that there are fundamental evolutionary differences between positive and negative

regulatory subnetworks. Identifying consistent themes in the ways regulatory networks achieve favorable adaptations can reveal design principles underlying the system's dynamics and evolutionary adaptability. On a wider note, this work has established that for a subset of proteins, systems-level properties can leave evolutionary traces of comparable effect size to physical features such as expression level and PPI degree.

1.4 Methods

1.4.1 Data collection

We used the yeast ChIP-chip data available from the YEASTRACT database (http://yeastract.com) (Teixeira et al., 2006) compiled from multiple studies (Borneman et al., 2007, 2006; Harbison et al., 2004; Horak et al., 2002; Lee et al., 2002; Workman et al., 2006). The literature-curated transcriptional network dataset, which is based on small-scale experimental studies, was also retreived from the YEASTRACT database (Teixeira et al., 2006). We downloaded physical interaction data from the *Saccharomyces* Genome Database (SGD) (Nash et al., 2007), which compiled PPIs from different high-throughput and small-scale studies. Orthology between *S. cerevisiae* and *S. paradoxus* were taken from the Orthogroups database (http://www.broadinstitute.org/regev/orthogroups/) (Wapinski et al., 2007). K_a/K_s values were calculated according to the Yang-Neilsen method (Yang & Nielsen, 2000) using PAML (Yang, 2007). Genes missing an ortholog were assigned a K_a/K_s value higher than the fastest evolving genes with an ortholog. Codon adaptation index (CAI) values were taken from Wang *et al.* (Wang et al., 2010).

TF knock-out microarray expression data (Hu et al., 2007) (accession #: GSE4654) and RNA-seq expression data (Busby et al., 2011; Ingolia et al., 2009) (accessions: GSE13750 and GSE32679) were retrieved from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/).

1.4.2 Calculating and comparing slopes

To allow for the comparison of slopes between TFs and all proteins, without succumbing to the pitfalls associated with the use of highly non-normal distributions, we developed a new normalization procedure. We simply assign ranks to all proteins in the original, genome-wide, distribution. Then we use these ranks to calculate slopes on the different protein sets, rather than re-ranking within the subsets as would Spearman's rank correlation. The problem with re-ranking within the subsets is that the slopes will be normalized to equal the correlation coefficient, which represents the goodness of fit rather than the relative slope. This modified procedure allows us to compare the degree of the slope between the TF subset to the global protein set.

1.4.3 Assigning p-values to subnetwork slopes

We calculated a p-value for the unexpectedness of the TF slope as compared to the slope for generic proteins, using a sampling procedure similar to the approach used in (Wang et al., 2010). We produced a distribution of 10,000 slopes by performing regressions on randomly selected equally sized samples of proteins whose average K_a/K_s and degree (or the relevant pair of variables) in rank space are within 1% root-mean-square deviation

(RMSD) of the TF subset. P-value is calculated as the fraction of slopes generated from random samples whose incline is more extreme than or equal to that of the slope associated with the TF subset. It is essential to control for average rate and degree because having a different distribution in either dimension can systematically bias the slope. As compared to the method applied in (Wang et al., 2010), our new method differs in that we used directly comparable slopes obtained from the genome-wide rank space instead of the correlation coefficient, and in that we controlled for the different average evolutionary rate (and other relevant variables) of TFs as compared to generic proteins. This improved method allows us to draw conclusions in more confidence, having excluded additional potential confounding factors.

1.4.4 Histograms and error bars

For each histogram, we plotted the median value for each bin, which is more robust to outliers than the average, and used bootstrapping with a 100 re-samplings to estimate the standard error of the median. Using the median rather than the mean also produces results which are insensitive to the choice of K_a/K_s assigned to genes which lack an ortholog in the reference species *S. paradoxus*.

1.4.5 Controls

As a control for the K_a/K_s to PPI degree and in-degree trend comparisons, we repeated the calculations, replacing the evolutionary rate of each protein with its codon adaptation index (CAI), which is considered a good proxy for the average expression level (Sharp &

Li, 1987). This way, we can confirm or discard the hypothesis that a surprising slope relating evolutionary rate and degree is explained by a different trend relating the strong correlate, expression, to degree. This approach was used previously in (Wang et al., 2010). To ensure that false positive interactions are not a problem, we also repeated these correlations using only network edges which are supported by two or more independent ChIP-chip experiments, which we termed confirmed edges (CE).

1.4.6 Measuring the relationship between TF and target properties

For every TF with 3 or more targets, based on ChIP-chip edges, we measured the median K_a/K_s, the fraction of targets missing a *S. paradoxus* ortholog as well as the fraction of highly conserved, highly interactive and highly expressed targets (top 20%) and used Spearman's rank correlation to establish the significance of the correlations. We repeated the analysis using literature derived edges and using only confirmed ChIP-chip edges (CE). We used TFs with 2 or more targets for the analysis with confirmed edges, since the resulting network is sparser and edges more reliable. In the case of the fraction of targets missing an ortholog, we still required at least 3 targets because this feature affects a small fraction of genes (~10%). We considered robust the correlations which were found to be significant (p<0.05) in all three networks. For activating and repressive networks, we used TFs with 5 or more targets regulated in the same direction to ensure the correlations are robust to the potential uncertainties in the sign of regulatory edges.

1.4.7 Calculating the spread of K_a/K_s and expression of co-regulated genes

For each TF with 3 or more targets possessing an ortholog in *S. paradoxus*, we calculated the median K_a/K_s and mRNA log read count difference between all pairs of targets and compared the result to the expected difference using the Wilcoxon rank-sum test. The expected median difference was estimated from the average of 100 equally-sized randomized sets of "target" genes, where each gene was chosen with a probability proportional to its in-degree.

1.4.8 Comparing the expression-level differences of genes between two yeast Species

We used previously published RNA-seq data from both *S. cerevisiae* and *S. paradoxus* (Busby et al., 2011) to measure the extent of gene expression change through evolution. We first normalized expression levels by dividing the number of reads mapping to each gene by the number of millions of reads in the sample (reads-per-million), fixing the lowest possible gene expression at 1 read-per-million. This procedure controls for differences in sequencing depth, allowing the levels for each gene to be comparable across the two species. We then measured the $\log_2$ fold expression change between the two species for each gene, using the orthology assignments provided by the expression study. Using logged values makes the fold change distribution closer to a normal distribution. We then used an unpaired t-test to determine the significance of the difference between absolute $\log_2$ fold changes of targets of fast evolving TFs (top 25%) and targets of slow evolving TFs (bottom 75%).

1.4.9 Assigning signs to regulatory edges

To assign a positive or negative sign to regulatory edges, we used previously published TF knock-out microarray data (Hu et al., 2007) which includes 135 TFs with ChIP-chip data. For ChIP-chip derived edges which corresponded to an X score (Hu et al., 2007), a confidence-weighted log ratio, of absolute value greater than 1, we inferred the sign of the edge based on the target gene expression change. The same approach was used for literature-based edges.

Tables

TF properties / Target properties	TF K_a/K_s	TF Expression	TF PPI degree	TF In-degree
Fraction of targets in 20% slowest evolving	**-0.23***	-0.16	**0.20**	0.01
Median target K_a/K_s	**0.25***	**-0.22***	**-0.24***	-0.02
Fraction of targets absent in *S. paradoxus*	**0.22**	-0.06	-0.16	**0.20**
Fraction of targets in 20% most highly expressed	-0.16	**0.21**	**0.29***	0.02
Median target expression	-0.13	0.16	**0.22**	0.01
Fraction of Targets in 20% most interactive	-0.11	**0.29***	**0.22***	-0.13
Median target PPI degree	-0.15	**0.18**	**0.19**	-0.10

Table 1.1: Spearman correlation coefficients relating TF and target properties in the ChIP-chip network. Bold: p-value<0.05. * : p-value<0.01

TF properties / Target properties	TF K_a/K_s	TF Expression	TF PPI degree	TF In-degree
Targets in 20% slowest evolving[1]	-0.19	-0.06	-0.04	-0.07
Median target K_a/K_s [1]	**0.23**	0.06	-0.06	0.04
Targets missing in *S. paradoxus* [2]	**0.24**	-0.07	-0.13	**0.21**
Targets in 20% most highly expressed[1]	-0.16	0.13	**0.23**	0.06
Median target expression	-0.12	0.19	0.20	0.04
Targets in 20% most interactive[1]	0.02	**0.30***	**0.27**	-0.07
Median target PPI degree	-0.14	**0.25**	**0.29***	-0.03

Table 1.2: Spearman correlation coefficients relating TF and target properties in the network of confirmed edges. Bold: p-value<0.05. * : p-value<0.01. 1 : TFs with 2 or more targets. 2 : TFs with 3 or more targets.

TF properties / Target properties	TF K_a/K_s	TF Expression	TF PPI degree	TF In-degree
Targets in 20% slowest evolving	**-0.22***	-0.08	0.04	**-0.16**
Median target K_a/K_s	**0.26***	0.15	0.00	0.17
Targets missing in *S. paradoxus*	**0.24***	0.16	-0.04	0.16
Targets in 20% most highly expressed	**-0.22***	**0.23***	**0.21**	-0.14
Median target expression	**-0.27***	**0.28***	**0.25***	-0.05
Targets in 20% most interactive	-0.04	**0.23***	**0.28***	-0.02
Median target PPI degree	**-0.26***	**0.21**	**0.31***	-0.06

Table 1.3: Spearman correlation coefficients relating TF and target properties in the network of literature curated edges. Bold: p-value<0.05 **.** *: p-value<0.01

Functional Term	GO ID	# of Genes	Fold enrichment	p-value
fungal-type cell wall	GO:0009277	75	1.56	0.00014
cell wall	GO:0005618	77	1.54	0.00022
external encapsulating structure	GO:0030312	77	1.54	0.00022
cell periphery	GO:0071944	313	1.19	0.0036
extracellular region	GO:0005576	63	1.43	0.0061
plasma membrane	GO:0005886	214	1.22	0.0078
oxidoreductase activity	GO:0016491	172	1.24	0.012
carbohydrate metabolic process	GO:0005975	156	1.25	0.017
transporter activity	GO:0005215	213	1.20	0.018
transmembrane transporter activity	GO:0022857	178	1.21	0.020
substrate-specific transmembrane transporter activity	GO:0022891	162	1.22	0.021
transcription factors, as taken from (Wang et al., 2010)	NA	102	1.28	0.024
substrate-specific transporter activity	GO:0022892	190	1.19	0.026
ion transmembrane transporter activity	GO:0015075	89	1.29	0.031
sequence-specific DNA binding	GO:0043565	123	1.24	0.033
ion transmembrane transport	GO:0034220	93	1.27	0.035
alcohol metabolic process	GO:0006066	102	1.28	0.036
carbohydrate biosynthetic process	GO:0016051	51	1.42	0.036
transmembrane transport	GO:0055085	204	1.17	0.046

Table 1.4: GO terms significantly enriched in targets of 25% fastest evolving TFs as compared to targets of other TFs

		ChIP-chip Network		Literature-curated Network	
Regulatory Sign	Property	Median target K_a/K_s	Fraction of targets lost or gained	Median target K_a/K_s	Fraction of targets lost or gained
Activated	rho (ρ)	0.26	0.29	0.31	0.39
	p-value	0.0036	0.0038	0.00039	0.00029
Repressed	rho (ρ)	0.068	-0.079	0.10	0.065
	p-value	0.46	0.52	0.30	0.61

Table 1.5: Correlations between TF Ka/Ks and evolutionary properties of activated or repressed target genes

GO term	GO term ID	# of genes	Fold enrichment	P-value
Cell periphery	GO:0071944	201	1.32	0.0023
fungal-type cell wall	GO:0009277	54	1.64	0.0042
External encapsulating structure	GO:0030312	56	1.60	0.0052
Cell wall	GO:0005618	56	1.60	0.0052
Oxidoreductase activity	GO:0016491	112	1.42	0.0058
Inorganic cation transmembrane transporter activity	GO:0022890	53	1.61	0.0091
Oxidation reduction process	GO:0055114	147	1.32	0.0096
Plasma membrane	GO:0005886	135	1.34	0.0099
Transporter activity	GO:0005215	139	1.30	0.018
Transmembrane transporter activity	GO:0022857	119	1.32	0.019
Cation transport	GO:0006812	69	1.44	0.022
Ion transport	GO:0006811	82	1.39	0.023
substrate-specific transporter activity	GO:0022892	125	1.30	0.024
Mitochondrial inner membrane	GO:0005743	70	1.46	0.024
substrate-specific transmembrane transporter activity	GO:0022891	107	1.32	0.027
Organelle inner membrane	GO:0019866	71	1.42	0.034
Cation transmembrane transporter activity	GO:0008324	54	1.46	0.035
Ion transmembrane transport	GO:0034220	67	1.40	0.036
Homeostatic process	GO:0042592	74	1.40	0.038
Ion transmembrane transporter activity	GO:0015075	62	1.41	0.042

Table 1.6: GO terms significantly enriched in target genes of TFs with 10 or more regulators as compared to targets of TFs with 2 or less regulators

Figures

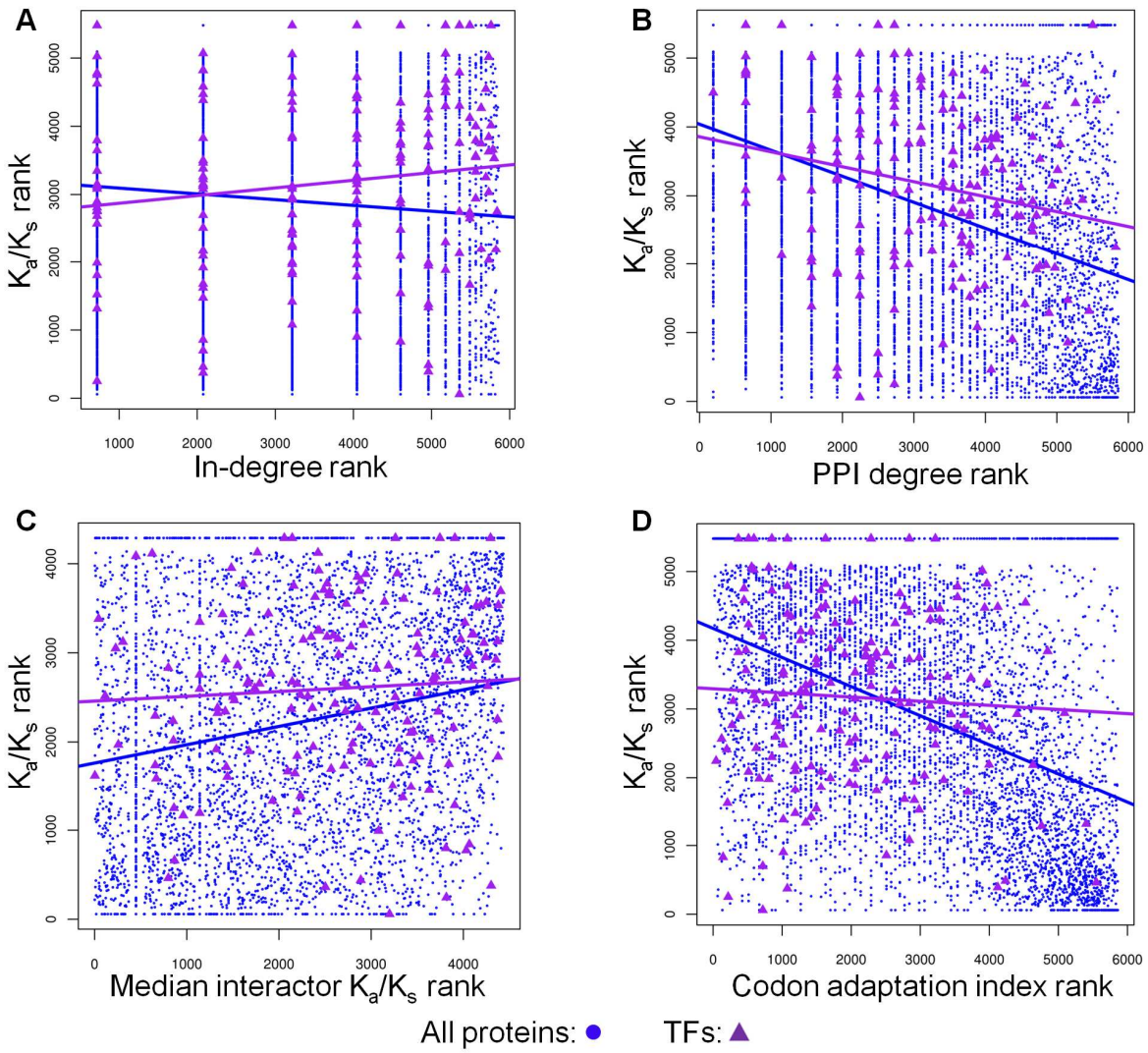

Figure 1.1: Scatter plots for distinct evolutionary trends of TFs compared to generic proteins. Shown are rank-rank plots and trend lines for all proteins (in blue) and TFs (in purple), where K_a/K_s is displayed as a function of regulatory in-degree (A), PPI degree (B), median K_a/K_s of interacting proteins (C), and CAI (D).

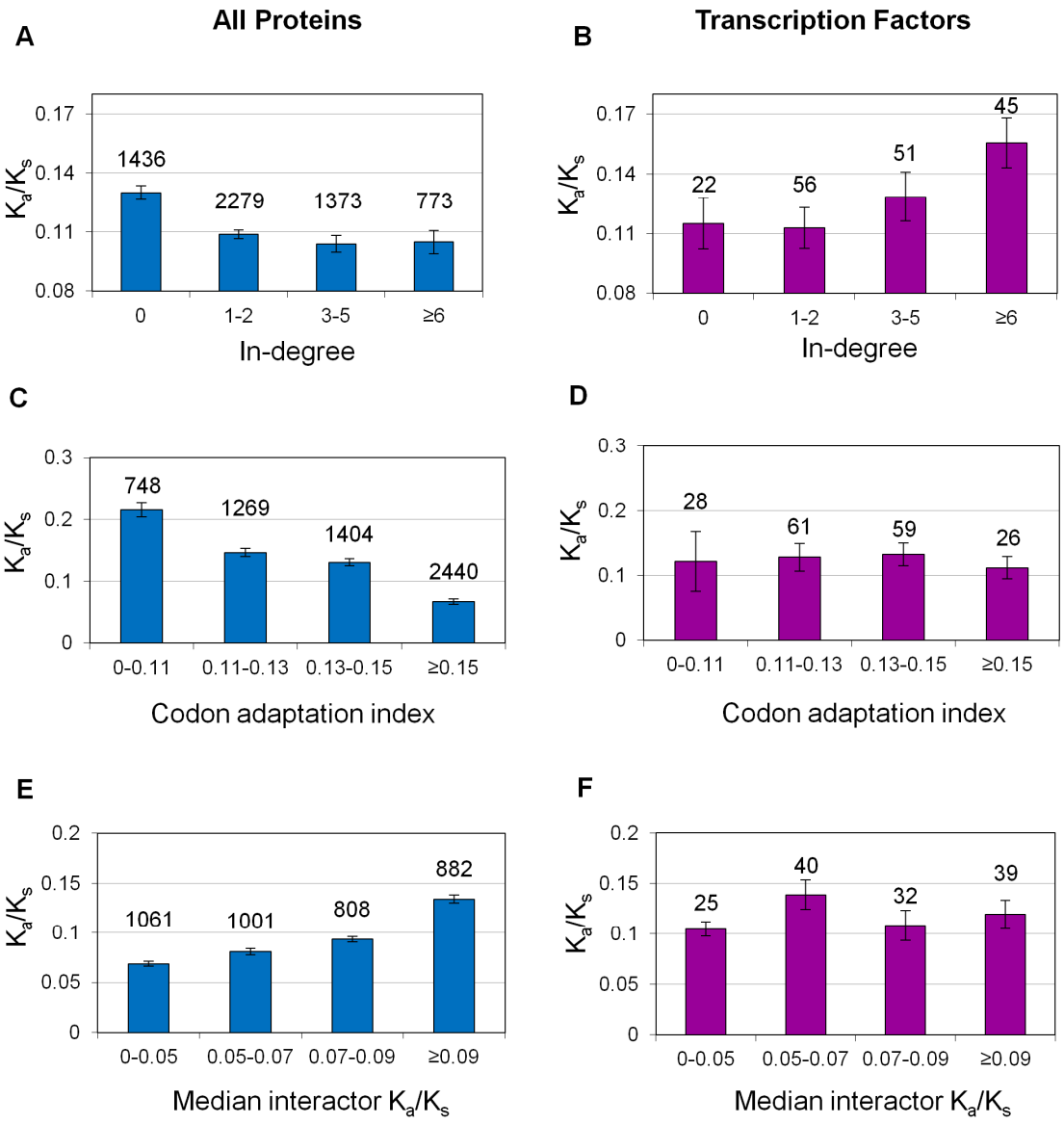

Figure 1.2: Distinct evolutionary trends of TFs. Unlike average proteins, TF K_a/K_s correlates positively with regulatory in-degree and very poorly with CAI and the evolutionary rate of PPI network neighbors. K_a/K_s is displayed as a function of regulatory in-degree (A-B), CAI (C-D) and median K_a/K_s of interacting proteins (E-F) for all proteins (A,C,E) and TFs (B,D,F). Numbers above the bars represent the number of TFs/proteins in the bin.

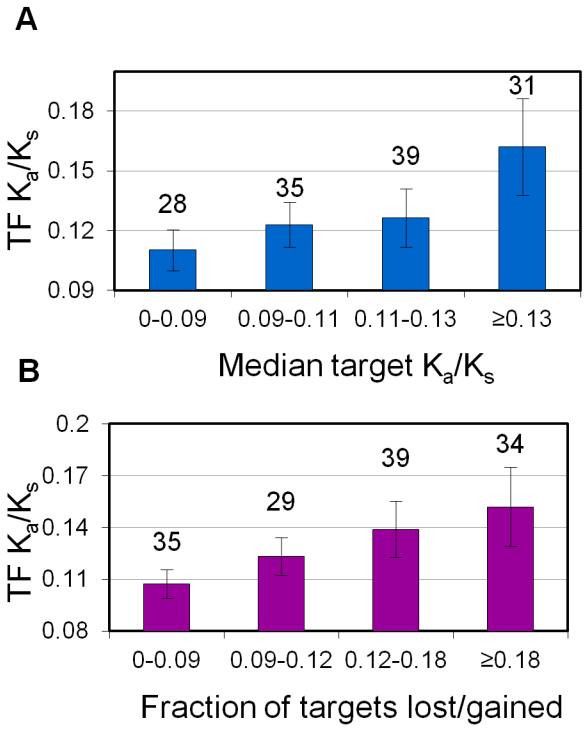

Figure 1.3: TFs and their targets co-evolve as modules. Each data point is based on a TF with 3 or more targets. (A) TF K_a/K_s as a function of the median K_a/K_s of target genes. (B) TF K_a/K_s as a function of the fraction of target genes missing an ortholog in *S. paradoxus* (lost in *S. paradoxus* or gained in *S. cerevisiae*). Numbers above the bars represent the number of TFs in the bin.

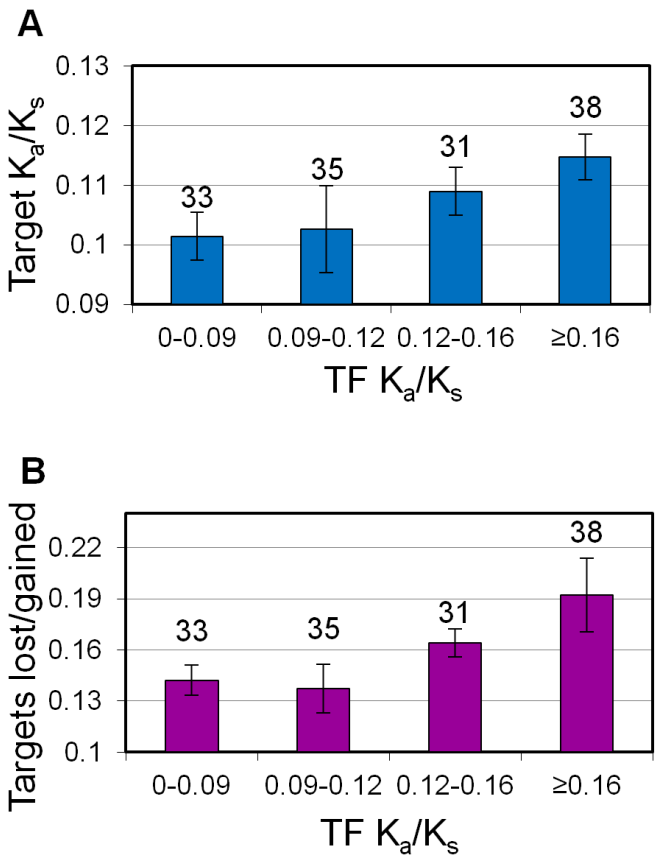

Figure 1.4: TF-target co-evolution between S. cerevisiae and S. mikatae. (A) Median K_a/K_s of target genes as a function of TF K_a/K_s. (B) Fraction of targets missing an ortholog in *S. mikatae* (lost in *S. mikatae* or gained in *S. cerevisiae*) as a function of TF K_a/K_s. Numbers above the bars represent the number of TFs in the bin.

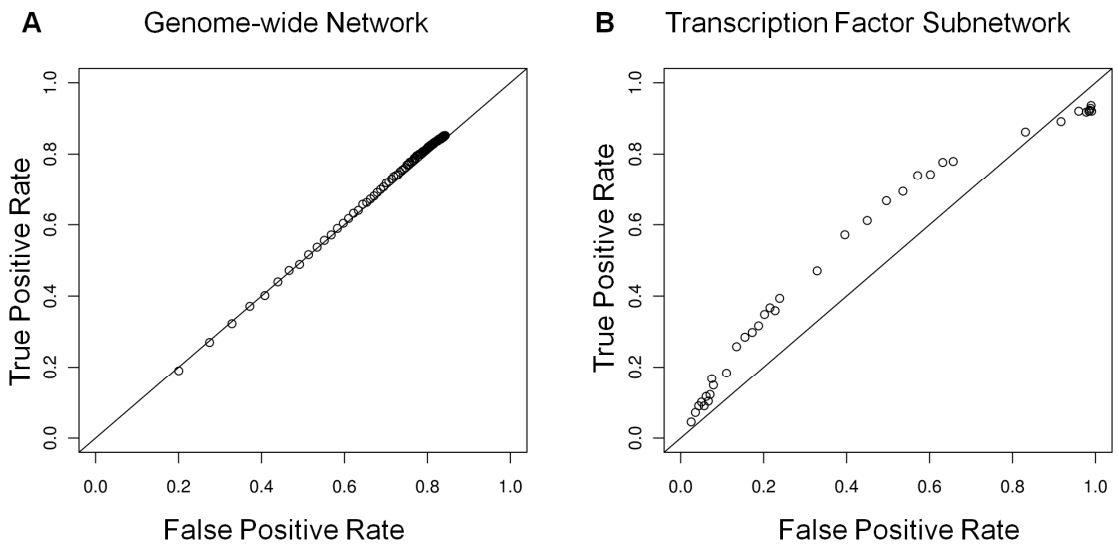

Figure 1.5: Ka/Ks as predictor for transcriptional regulation. Shown are the Receiver Operating Characteristic (ROC) curves over the entire ChIP-chip network (A) and the TF subnetwork (B) of regulatory interaction prediction based on linear regression between TF K_a/K_s and median target K_a/K_s. In each case, TFs were randomly split into a training set, on which regression was performed, and a test set, on which true positive and false positive rates were assessed. The Figure 1.shows that K_a/K_s does not predict regulatory edges in the global network, but it does provide some predictive power when limited to the TF subnetwork (TFs regulating TFs).

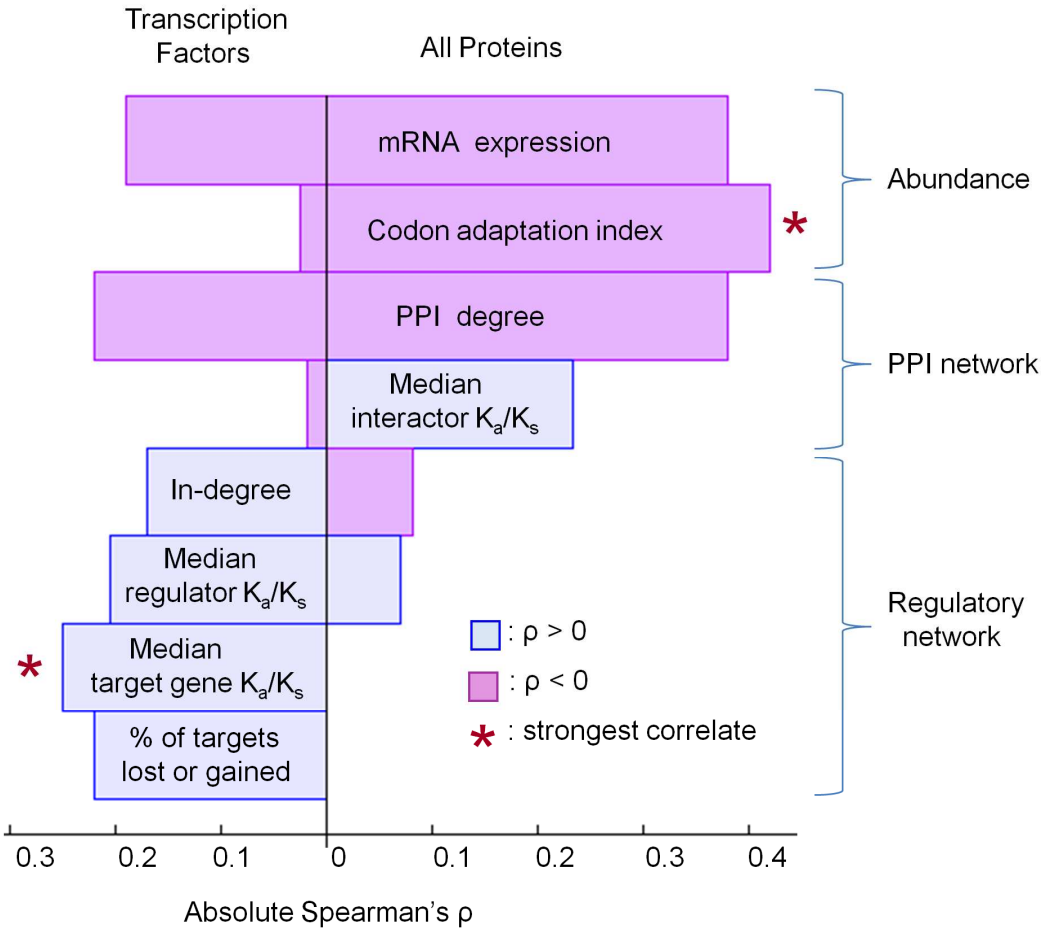

Figure 1.6: Comparison of different genomic and network features influencing TF and protein evolutionary rate. For each determinant, absolute Spearman's rank correlation coefficient (ρ) for TFs is displayed on the left and for all proteins, on the right, with the color of the box representing the direction of the trend. The * indicates the most dominant correlation for each protein set. While CAI is the dominant correlate with K_a/K_s for generic proteins, target gene K_a/K_s is the strongest correlate for TF K_a/K_s.

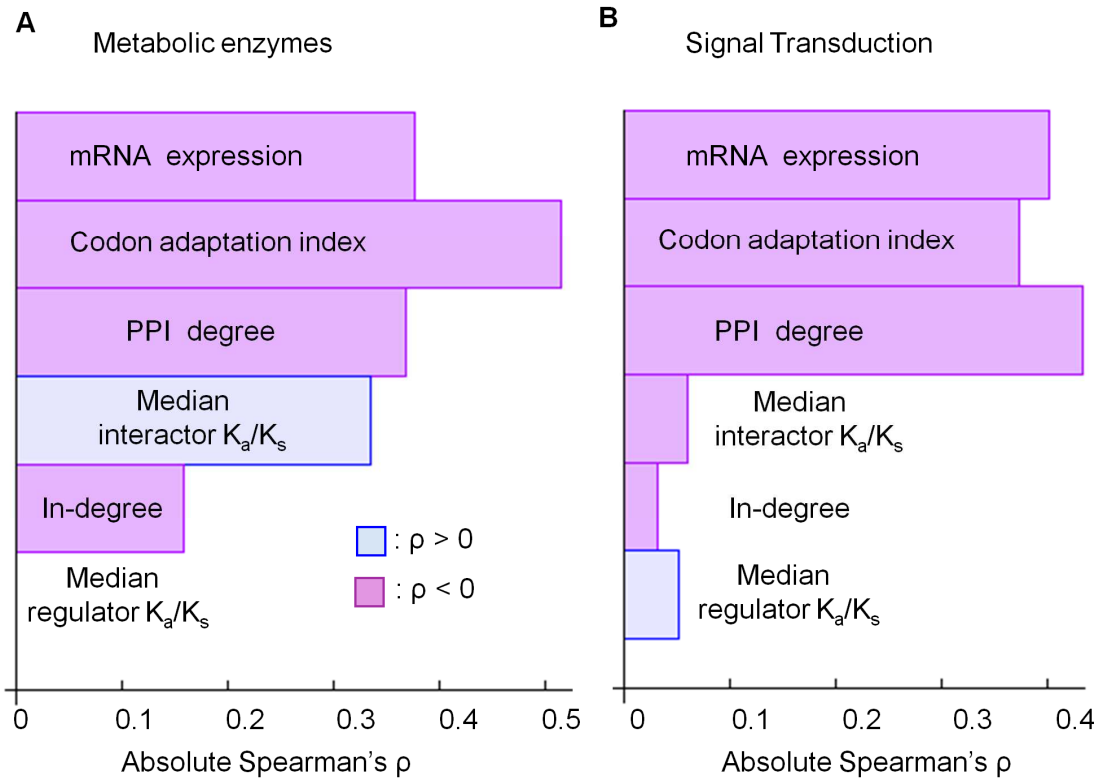

Figure 1.7: Comparison of different genomic and network features influencing evolutionary rate of metabolic enzymes and signal transduction proteins. For each determinant, absolute Spearman's rank correlation coefficient (ρ) is displayed, with the color of the box representing the direction of the trend. (A) Evolutionary rate determinants of 540 metabolic enzymes taken from YeastCyc (Caspi et al., 2008). (B) Evolutionary rate determinants of the 240 proteins in the GO term "signal transduction". This Figure 1.shows that functionally defined protein sets other than TFs have evolutionary rate determinant profiles similar to that of generic proteins.

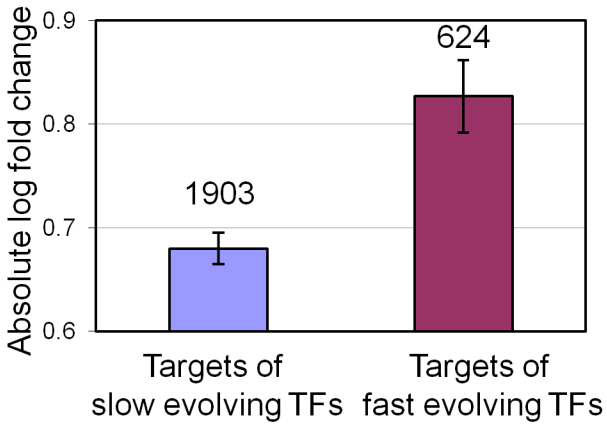

Figure 1.8: Targets of fast evolving TFs have larger expression changes through evolution. The targets of the 25% fastest evolving TFs, on the right, have on average larger absolute fold changes in expression between *S. cerevisiae* and *S. paradoxus* than targets of other TFs, on the left, as determined by RNA-seq. Numbers above the bars represent the number of TFs in the bin.

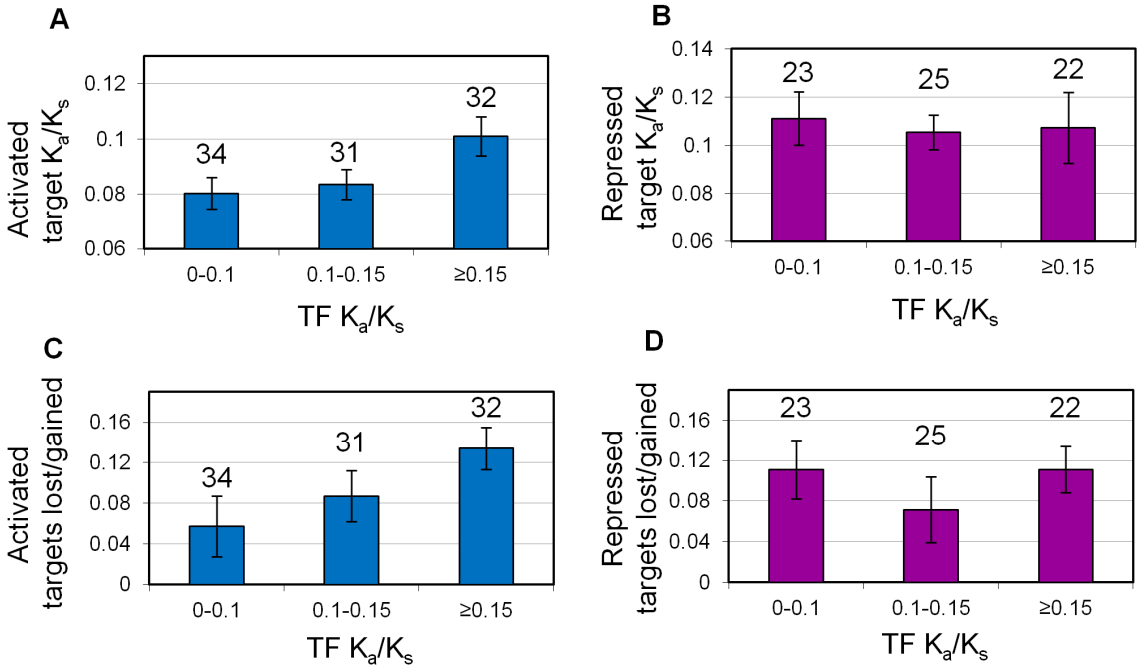

Figure 1.9: TFs co-evolve with activated targets, but not with repressed targets. Edge signs are inferred from TF knock-out expression data. Each data point is based on a TF with 5 or more targets regulated in the same direction. (A) Median K_a/K_s of activated target genes as a function of TF K_a/K_s. (B) Median K_a/K_s of repressed target genes as a function of TF K_a/K_s. (C) Fraction of activated targets missing an ortholog in *S. paradoxus* as a function of TF K_a/K_s. (D) Fraction of repressed targets missing an ortholog in *S. paradoxus* as a function of TF K_a/K_s. Numbers above the bars represent the number of TFs in the bin.

Chapter 2
Network analysis reveals complex regulation of lost and gained genes

2.1 Introduction

Gene gain and loss are very important components of evolution and interspecies differences. For example, a dozen distant eukaryotes have been shown to share as little as 9% of their combined gene families (Ptitsyn & Moroz, 2012). Proteomes are constantly evolving and the dynamics of gene gain and loss processes shape the networks of interactions that determine the behavior of higher-level systems. While protein sequence evolution provides an informative evolutionary landscape over the length of a single protein, gene gain and loss, as binary components of protein evolution, necessitate a genomic-level view and are ideally studied over many species.

While some network-level analyses have provided insights into the process of gene gain and others, gene loss, none has considered these two types of evolutionary events in conjunction, as part of the entire gene evolutionary life-cycle. The set of genes which are universally conserved across the phylogenetic tree has been termed the “core” genome of the lineage (Harris et al., 2003; Lefébure & Stanhope, 2007). Studies of gene loss comparing distant Eukaryotes have shown that lost genes differ significantly from core

genes in many ways. They have fewer protein-protein interaction (PPI) partners, lower mRNA expression, lower sequence conservation and their deletion is less likely to produce a lethal phenotype, known as gene essentiality (Krylov et al., 2003). Studies on horizontally transferred genes and *de novo* gene birth in Eukaryotes and prokaryotes have shown similar features for gained genes, with the most recently gained genes harboring the most extreme values (Carvunis et al., 2012; Lercher & Pál, 2008). No study however has considered gene loss in the context of the regulatory network, nor have lost or gained genes been placed in the context of the genetic interaction network, which defines the map of functional inter-dependencies within the proteome.

In this study, we classified *S. cerevisiae* genes according to the presence of their orthologs throughout the Ascomycota phylogenic tree and overlaid diverse network and functional datasets to attempt to understand how lost and gained genes differ from evolutionary core genes. We identify genes which have been lost along the tree as well as genes which were likely gained along the *S. cerevisiae* lineage. Studying only extant species, we cannot follow the same genes through their different life stages. We instead study different genes, classified according to their position along the average evolutionary trajectory of genes from gene birth to gene loss. We thus reconstruct the life-cycle of genes from a snapshot of genes as they appear today. Many proteins will not make it through all stages and may for example be lost after having been recently gained. But the proteins that survive this pruning process may eventually grow to become part of the stable evolutionary core genome of a phylogenetic clade.

Gene duplication, including whole-genome duplication, is one of the most common mechanism for gene gain in eukaryotes (Blomme et al., 2006; Ohno, 1970). However, there is an important functional distinction between gene duplication, which merely increases the number of genes in a family, and horizontal gene transfer or *de novo* gene birth, which introduce an entirely new gene family into a genome. For this reason, we considered duplication events separately from other gene gain events, and mostly concentrated on loss and gain events which changed whether any ortholog was present or not in a lineage, systematically excluding gene duplication events as well as the loss of duplicated genes. There exists some previous work providing a network perspective on gene duplication, such as Hughes and Friedman 2005(Hughes & Friedman, 2005) and Jian *et al.* 2011 (Jiang et al., 2011).

While some studies have classified gained genes according to their age (Carvunis et al., 2012; Lercher & Pál, 2008), no study has considered a time-resolved view of gene loss. Just as gained genes tend to integrate in the PPI and regulatory networks over time(Lercher & Pál, 2008), we may expect genes to be marginalized via network rewiring before they are eventually lost. To test this hypothesis, we classified gene loss events according to their distance to the model species, and asked whether phylogenetically closer gene loss events were associated with more peripheral genes in the different types of networks.

2.2 Results

2.2.1 Identifying gene loss and gain

Using gene orthology assignments across 23 ascomycete fungi genomes from the Orthogroups database (Wapinski et al., 2007), we classified genes into lost, gained or "core" genes, based on their representation across the tree, as detailed in the Methods section. We identified 2,012 *S. cerevisiae* protein-coding genes universally conserved across all 23 species. Starting from the roughly 350 million year old divergence of the *N. crassa* lineage (Fitzpatrick et al., 2006), we identified 6,348 gene loss events, implicating orthologs of 2,445 unique proteins in *S. cerevisiae* (see Methods for details). To identify gene gain events, we identified cases where a gene's orthologs were absent from all species which branch off from the *S. cerevisae* lineage before a specific branch point. In order to minimize the misclassification of parallel loss events as gain events, we only considered gains represented in at least 75% of the species in the affected lineage. We further filtered cases with significant homology to any older gene (see Methods), indicating the gain is potentially a missed duplication event. We thus identified 652 *S. cerevisiae* proteins that we could confidently assign as arising from a gain event other than through the duplication of existing genes. Figure 1 shows the number of loss and gain events by branch in the context of the tree. Based on relative branch lengths (see Methods), there is a 3.9 fold reduction in the rate of gene gain and a 3.3 fold increase in the rate of gene loss along the *S. cerevisiae* lineage after its divergence from *K. Waltii*, following the whole-genome duplication event (Kellis et al., 2004). This suggests that

duplicated genes may compete with other gained genes for the same functions and that the whole-genome duplication event caused a greater number of functions to be assumed by duplicate genes.

2.2.2 Reconstructing the gene life-cycle

As depicted in Figure 2, gained and conserved genes can be classified by their inferred phylogenetic age and positioned along an average evolutionary path, from species-specific new genes to universally conserved “core” genes. Analogously to the birth of new genes and their subsequent integration, the deletion or pseudogenisation of genes may be preceded by a phase of network and functional marginalization. We thus distinguished genes which were lost solely on distant branches from genes lost on proximal branches, based on a distance cut-off shown in Figure 1.

Most genes specific to S. cerevisiae are expected to be lost before becoming shared across two or more species, what we refer to as gene “pruning”. Assuming a constant rate of gene gain equal to that observed in *S. cerevisiae*, we expect 4116 new genes to have been gained after the divergence from *K. waltii* and before the divergence from *S. paradoxus*, based on relative branch lengths (see Methods). Since the number of gained genes we actually observe (i.e., conserved in *S. cerevisiae*) is only 63, we infer that approximately 98.5% of species-specific gained genes are pruned early on, although this estimate is likely to be higher than the general trend because of the influence of the whole-genome duplication event. Most *S. cerevisiae*-specific genes are likely created *de*

novo, 90% possessing no homology to any sequence within or outside yeast (see Methods), and the high rate of gene pruning we infer is consistent with recent findings on *de novo* gene birth in yeast (Carvunis et al., 2012).

2.2.3 Protein-protein interaction degree of lost and gained genes

By overlaying different types of *S. cerevisiae* networks onto the reconstructed gene content evolutionary history, we can explore how the connectivity of core genes differs from that of lost and gained genes and how the distance of the loss or gain events from the model species influences this property. We exclude genes specific to *S. cerevisiae* from these comparisons as they have no evolutionary evidence of encoding genuine functional proteins. As shown in Figure 3A, we found that the average PPI degree of core proteins is significantly higher than both lost (Wilcoxon test $p<2.2x10^{-16}$) and gained (Wilcoxon test $p<2.2x10^{-16}$) proteins, confirming earlier reports (Carvunis et al., 2012; Krylov et al., 2003; Lercher & Pál, 2008). The differences between gained proteins of different ages and core proteins is partially explained by the preferential loss of low-degree genes and partially by network rewiring, including the loss and gain of other proteins in the network.

2.2.4 Genetic interaction degree of lost and gained genes

Genetic interactions are a very different concept than that of physical interactions, representing the functional integration of genes into the organism. Genetic interactions

are epistatic interaction detected through the non-linearity of the phenotypic effects of double-mutations, such as synthetic lethality, when two viable mutations combine to create a non-viable double-mutant. These interactions indicate that a pair of genes is functionally linked. This functional linkage can for the most part be considered a type of conditional essentiality and its degree indicates how independent a gene is functionally from the rest of the genome. Excluding essential genes (Winzeler, 1999), which cannot be tested for genetic interactions, we found that core genes have many more genetic interaction partners on average than lost or gained genes. In the case of gained genes which likely possess completely novel functions which have not had time to integrate into the system, it is expected that their function be relatively independent from the rest of the genome. The lower degree of lost genes indicates that gene loss preferentially targets genes with relatively independent functions, as compared to the tightly nit functions of the core genome.

2.2.5 Regulatory in-degree of lost and gained genes

As we have shown in a recent study in yeast, the regulatory network plays a unique role in species-specific adaptation, with fast evolving TFs preferentially regulating species-specific or fast evolving genes (Coulombe-Huntington & Xia, 2012). The regulatory network, which is known to rewire more rapidly than most other biological networks (Shou et al., 2011), may thus play a relatively more active role in the integration of new genes. Based on a collection of ChIP-chip studies (Teixeira et al., 2006) , providing a relatively unbiased measure of regulatory in-degree, we found to our surprise that

universally conserved genes had significantly fewer regulators than both lost (Wilcoxon rank sum test $p=4.3x10^{-10}$) and gained genes ($p=5.8x10^{-6}$), excluding genes specific to *S. cerevisiae*. The significance of the results was confirmed using only regulatory interactions reported in two or more studies (gained genes: $p=1.5x10^{-3}$, lost genes: $p=1.1x10^{-7}$) and also using a completely independent network based entirely on literature curation of small scale studies (gained genes: $p=5.0x10^{-4}$, lost genes $p<2.2x10^{-16}$). The relative centrality of lost and gained genes in the regulatory network contrasts sharply with the trend observed for other networks. Given regulatory in-degree's negative correlation with evolutionary rate(Wang et al., 2010; Xia et al., 2009) (Spearman's $\rho=-0.082$, $p=1.7x10^{-7}$) and positive correlation with PPI degree ($\rho=0.048$, $p=0.00018$) and mRNA expression ($\rho=0.10$, $p=4.7x10^{-15}$), this result cannot be explained by typical covariates and is all the more surprising, meaning it is likely an inherent property of species-specific gene regulation. It suggests a highly active role of TFs in regulating species-specific gene expression, to the extent that the regulatory complexity of species-specific genes tends to surpass that of the core genome. A study in mammals has previously shown that conditionally expressed genes have more conserved promoters than constitutively expressed genes(Lee et al., 2005), suggesting they possess more complex regulatory programs. While the core genome is largely composed of constitutively expressed house-keeping genes, most species-specific genes likely participate in niche-specific functions which are likely to be condition-specific and thus may require more complex regulation. The rapid adaptation of the transcriptional regulatory program of new genes may allow the cell to tightly regulate their abundance,

minimizing energetic costs and potentially unfavorable interactions, as they more slowly become integrated into the other types of networks.

Since gene pruning is a form of gene loss and gene loss preferentially targets genes with high regulatory in-degree, the age-dependent increase in the regulatory in-degree of gained genes is very unlikely to be explained by gene pruning. This means that gained genes must have recruited a large number of new regulators through *cis*- or *trans*-regulatory rewiring during their integration phase.

2.2.6 Gain and loss of transcription factors

Given the highly active role of the transcriptional network in species-specific gene regulation, we decided to explore the role of gene gain and loss in *trans*-regulatory network evolution. As shown in Table 1, we found that TFs are highly enriched in all types of species-specific genes, including duplicated, lost and gained genes, excluding those gained in *S. cerevisiae* (see Methods). This result demonstrates the central role of *trans*-regulatory network evolution in species-specific adaptation and suggests a mechanism for the *trans*-regulatory network integration of newly gained genes.

It is noteworthy that 27 of the 28 gained TFs were identified as potential horizontal gene transfers (see Methods). The relative under-representation of potential *de novo* TFs is potentially due to the reliance on the presence of known DNA-binding domains in the

identification of TFs or because the specific function of TFs may not easily be achieved by *de novo* gene creation mechanisms.

2.2.7 Function of lost and gained genes

Having shown TFs are enriched in lost and gained gene sets, we may expect lost and gained genes to be enriched for other specific functions. To identify such functional enrichments, we compared the fraction of gained or lost genes annotated for specific Gene Ontology (GO) terms of more than 200 genes to that of universally conserved genes. To avoid potential biases caused by the whole-genome duplication event or the relatively poor annotation of new genes, we concentrated on gene loss and gain events which occurred outside of the whole-genome duplication-affected lineage. Tables 2 and 3, show, with highly overlapping GO terms (>90%) removed, that both lost and gained genes are significantly enriched for terms involving the physical periphery of the cell, transcriptional regulation and sexual reproduction. In other words, proteins found in these regions or involved in these processes are thus more likely to be species-specific than other proteins. These functions provide good examples of processes which likely require complex regulatory programs. In contrast, metabolic enzymes from YeastCyc (Caspi et al., 2008) are found to be depleted in gained (Fisher's exact test $p=3.3\text{x}10^{-12}$) and in lost genes ($p=5.0\text{x}10^{-8}$), which is consistent with the existence of an evolutionary core metabolism and the likely difficult network integration of gained enzymes given the extremely slow rewiring rates of metabolic networks (Shou et al., 2011).

2.2.8 Network marginalization as a lineage-specific predictor of gene loss

As proposed in an earlier study, the propensity for gene loss can be modeled as an, unchanging, inherent property of a gene (Krylov et al., 2003). This model captures the spectrum of evolutionary gene dispensability, from peripheral niche-specific genes to core, universally conserved genes. However, underlying the this view is the implicit assumption that network structure is either static throughout evolution, or has no influence on the propensity for gene loss of individual genes. Here, we investigate the possibility that gene loss propensity could be modeled more accurately as a branch-specific property. Considering that networks tend to rewire over time (Shou et al., 2011), we expect that the propensity for gene loss of a gene should change depending on its network and genomic context in each organism. To assess this possibility, we distinguished between genes lost only in distant species (distant loss), which may have preserved or gained indispensability in the *S. cerevisiae* lineage, from genes lost in closely related species (proximal loss), which we consider to be at highest risk of gene loss in the model species. While lost genes are known to display signatures of peripherality (Krylov et al., 2003), considering the temporal dimension in the study of gene loss allows us to look for an evolutionary progression of these signatures. As shown in Figure 3 and Figure 4, genes lost in close species have stronger network and genomic signatures of marginalization than genes lost in distant species. Specifically, we found that genes lost in close species have significantly lower PPI interaction degree (Wilcoxon test p=0.00037), lower genetic interaction degree (Wilcoxon test p=1.9x10^{-5}), lower mRNA expression (Wilcoxon test p=2.2x10^{-9}) and higher regulatory in-degree (Wilcoxon

test p=8.4x10^{-5}), considering the union of ChIP-chip and literature-derived edges, than genes lost solely on distant branches. These results are based only on genes which possess an ortholog in the outer-most branch of the tree, in order to control for the potential confounding effects of gene age. Furthermore, we controlled for differences in the average propensity for gene loss over the entire tree, in order to disassociate the lineage-specific component of the propensity for gene loss from the lineage independent component considered in a previous study (Krylov et al., 2003) (see Methods). Evolutionary rate and gene essentiality show a similar trend supporting gene marginalization after applying the controls but the differences fall below statistical significance (t-test p=0.35 and Fisher's exact p=0.16, respectfully). These results demonstrate that the propensity for gene loss is not solely an inherent, fixed property of genes but that it has a significant branch-specific component, being dependant on the network and genomic context of a gene in a given species. These results can be explained by a phase of progressive network and functional marginalization of genes preceding gene loss, similarly to the integration phase which follows gene gain. Inversely, ancient genes which are likely to be lost in one lineage may integrate further and become part of the core genome in a different lineage. Both of these processes can equivalently account for the observed differences between distantly and proximally lost genes and both are likely to be exist given that networks have been shown to gain and lose edges over time (Shou et al., 2011).

2.2.9 Mechanisms of gene gain

Genes can be gained through very different mechanisms, and the mechanism of gain could have an influence on many of the properties measured in this study. To assess this, we classified gained genes as having been gained through either horizontal gene transfer (HGT) or *de novo* gain events, excluding genes gained in *S. cerevisiae* (see Methods). We also looked at duplicated genes in order to compare their properties with other gained genes, and further classified these according to whether they were gained through a single duplication event or the whole-genome duplication event (see Methods). We found that duplicated genes have much higher PPI degree and slower sequence evolution than HGTs, and the differences are even greater with respect to *de novo* genes. Regulatory in-degree however is roughly uniform across the different types of gained genes, supporting the hypothesis that the relatively fast rewiring allows for the relatively quick integration of new genes into the regulatory network. Furthermore, *de novo* genes tend not to be essential, have low genetic interaction degree and low mRNA expression, as compared to all other types of gained genes. The significantly higher proportion of *de novo* genes in the first two evolutionary life stages of genes, before the divergence from *K. waltii*, could thus contribute to the relatively strong signatures of peripherality of these gene sets as compared to older gained genes.

2.2.10 Gene gain by duplication

In multicellular eukaryotes, duplication accounts for the vast majority of gene gain events (Ohno, 1970). In contrast to horizontal gene transfers or *de novo* gene birth, duplicated

genes might inherit functions and network edges from their pre-integrated parent gene. As such, we expect their integration to be very different from that of genes gained by other mechanisms. In order to study the integration of new duplicate genes, we separated duplicate genes into three categories based on their relative age: before, during or after the divergence of *K. waltii*, corresponding to the branch where the whole-genome duplication (WGD) occurred. This way, we can attempt to disentangle the effects of gene integration from those of the WGD. As we show in Figure 2.5, although more subtle than the integration of genes gained by other mechanisms, we find that duplicate genes also appear to require time to fully integrate into the different networks. Specifically, genes duplicated before the WGD (pre-WGD) appear to be more integrated than genes duplicated after the WGD (post-WGD), as evidenced by their significantly higher PPI degree (Figure 2.5A, Wilcoxon test $p=2.2\text{x}10^{-10}$), genetic interaction degree (Figure 2.5B, Wilcoxon test $p<2.2\text{x}10^{-16}$), lower indegree (Figure 2.5C, Wilcoxon test $p=0.0016$), considering the union of ChIP-chip and literature-derived edges, and higher mRNA expression (Figure 2.5D, Wilcoxon test level $p<2.2\text{x}10^{-16}$). As shown in Figures 2.5E and 2.5F, essentiality and evolutionary rate show similar trends supporting the scenario of time-dependant integration but the differences between duplicate genes of different age groups fail to reach significance.

2.2.11 Gene integration and evolutionary rewiring rates

Although it is difficult to isolate the relative contribution of network rewiring in the integration of new genes, we can ask whether the agreement between the relative rates of

network integration and experimentally measured rates of network rewiring extends to other networks, such as the genetic interaction network or phosphorylation network. To assess this question, we used the ratio of the average network degree of anciently gained genes to that of universally conserved genes as a measure of the relative rate of interaction gain. This measure orders the different types of interactions, from fast to slow rate of gain, in the following order: transcriptional regulatory interactions, kinase interactions, genetic interactions, and PPIs, where kinase interaction degree was measured as the number of PPI partners annotated with GO term “protein kinases” (Ashburner et al., 2000). This ordering follows exactly the order established by experimental measures of evolutionary network rewiring rates (Shou et al., 2011), which is unlikely the result of chance, given 24 possible orderings (p=0.042). This suggests that rewiring rate is the dominant force determining the rate of network integration.

2.3 Discussion

In this study, we have demonstrated that lost and gained genes are very distinct from universally conserved genes in terms of various network and genomic properties. We have also shown that these properties are influenced by the phylogenic distance of the loss or gain event from the model species, shedding light on the complex processes of gene pruning, integration and marginalization. This classification scheme allows us to reconstruct the complete evolutionary life-cycle of genes, through the various

evolutionary stages leading from gene birth to gene loss. For the first time, we have established that lost and gained genes tend to have lower genetic interaction degree, demonstrating that species-specific genes tend to have more independent, or less integrated, functions than core genes.

We have also found that the regulatory network plays a unique role in the evolutionary integration and marginalization of genes. For one, the regulatory network is the only network for which species-specific genes tend to have a higher degree than core genes, demonstrating a strong association between regulatory complexity and species-specific adaptation through gene gain or loss. Secondly, TFs are highly enriched in lost, gained or duplicated genes, as compared to the evolutionary core, highlighting the important role of *trans*- as well as *cis*- regulatory network rewiring in species-specific adaptation.

These results teach us not only about the evolutionary processes surrounding gene gain and loss but also about the organization of biological networks themselves. Every stage along the gene evolutionary life-cycle is associated with different network properties, introducing some degree of predictability to the overall network structure and helping to explain important topological features of biological networks, including their scale-free nature (Jeong et al., 2000). The processes and gene gain, loss, integration and marginalization each exert significant influence on network structure, stressing the importance of considering the evolutionary context of genes when trying to make sense of networks in any species.

By considering the branch specificity of gene loss events, we put forward evidence that ancient genes can be marginalized or integrated further via network rewiring, as their propensity for gene loss evolves along a specific branch. This demonstrates that a gene's propensity for gene loss is dependant on its network context and that network structure evolves significantly over time. The evolution of gene content is thus intertwined with the evolution of network structure and through network rewiring, individual genes can migrate along a continuum between highly species-specific roles and core roles.

In multicellular eukaryotes, duplication accounts for almost all gene gain events (Ohno, 1970). In contrast to horizontal gene transfers or *de novo* gene birth, duplicate genes might inherit functions and network edges from their pre-integrated parent gene. As such, we expect their integration to be very different from that of genes gained by other mechanisms. As we have shown, the increased number of duplicate genes created by the whole-genome duplication following *S. cerevisiae*'s divergence from *K. waltii*, has had a significant effect on the subsequent rates of gene loss and gene gain by other mechanisms, suggesting newly duplicated genes compete with both new and old genes to fulfill biological functions from a limited pool of naturally selected functions. It would therefore be interesting to consider gene duplication and the network integration of duplicate genes alongside other forms of gene gain in order to fully understand the contribution of gene gain to network structure and organization.

Proteomes and networks are constantly evolving and are therefore best understood in an evolutionary context. This work shows how the evolutionary dynamics of nodes and edges in biological networks are strongly correlated. Only from this multi-dimensional systems-level perspective can the processes of gene integration and marginalization be understood, and in turn help to explain the organization and evolutionary dynamics of biological networks.

2.4 Methods

2.4.1 Data collection

We downloaded the orthology mappings provided by the Orthogroups database (Wapinski et al., 2007). PPI and genetic interaction network data were retrieved from the *Saccharomyces* Genome Database (Nash et al., 2007) and regulatory network data from YEASTRACT (Teixeira et al., 2006). mRNA expression information was downloaded from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) (Accession: GSE13750) and based on RNA-seq performed on yeast grown in rich media (Ingolia et al., 2009).

2.4.2 Identifying gene loss and gain events

We used the orthology mappings provided by the Orthogroups (Wapinski et al., 2007) database covering 23 fungal species, as well as the phylogenic tree from the same source.

Aiming to study the features of lost and gained genes in *S. cerevisiae*, we only considered genes which are present in *S. cerevisiae*. We therefore only identified loss events which happened on branches leading away from *S. cerevisiae* and gain events on branches ancestral to *S. cerevisiae*. Species belonging to the two outer-most branches were used as the outgroup for the identification of gene loss and gain events, allowing newly gained genes to be distinguished from older genes with sparse representation (parallel loss events). In order to identify gene loss events, we simply identified proteins which were present in a common ancestor and missing in a descendant species. Identifying gain events posed the additional challenge of filtering false positives caused by parallel loss events. Assuming that a gene can not be gained more than once independently, a confidently assigned gain event should involve a gene which is present in most of the species in a given lineage and absent in all other lineages. Gains were thus defined according to the following three conditions: (1) the gene is found in a single lineage which contains *S. cerevisiae*, (2) it is present in the first branch to diverge from *S. cerevisiae* in the affected lineage and (3) in at least 75% of the species in the affected lineage.

2.4.3 Identifying duplicated genes from the orthology map

Duplicated genes are genes for which an ortholog in another species maps to two or more genes in *S. cerevisiae*. As expected, a large number of duplication events were observed at the *S. cerevisiae-K. waltii* split (data not shown), where a whole-genome duplication in the *S. cerevisiae* lineage was shown to have occurred (Kellis et al., 2004). For each pair

or family of duplicates, we identify the parent gene as the copy with the highest level of sequence homology to the ortholog in the closest species not affected by the duplication event. Duplicated genes not identified as parents were considered duplicates for the purpose of identifying the ratio of transcription factors in genes gained by duplication.

2.4.4 Identifying potential duplications missed in orthology map

Gene duplication events do not lead to an increase in the number of gene families and were therefore discarded from the set of gene gains used in this study. While the Orthogroups data structure clearly distinguishes dupliucations from other gain events. We opted to further filter out any potential duplication events that may have been misclassified as gains by Orthogroups. We used BLAST (Altschul, 1997) with default settings to compare all against all *S. cerevisiae* proteins. We then considered as potential duplication events cases where a gained protein bares significant homology ($e<1x10^{-4}$) to an older gene. Out of 812 genes initially identified as gain events, 141 showed evidence of duplication and were thus discarded from the analysis, most of them specific to *S. cerevisiae*.

2.4.5 Identifying potential horizontal gene transfers

The two well known mechanisms of gene gain which can introduce new gene families are horizontal gene transfers (HGT), where a gene is transferred from a different species, and *de novo* gene birth, where a gene is created from mutations to non-coding DNA. Given the large number of sequenced genomes available, it should be possible to identify

potential horizontal gene transfers based on homology. Assuming that gene duplication can be ruled out, the identification of sequence homology between a newly gained gene and any gene in any species outside of the gain-affected lineage would indicate the gene was gained through HGT. We used BLAST to compare gained proteins to all complete genomes in the BLAST database and looked for significant homology ($e<1x10^{-4}$) with any species outside of the fungal lineage. Non-duplicated genes with such homology were flagged as potential HGTs and those with no homology, as potential *de novo* genes.

2.4.6 Measuring transcription factor enrichment

We used the list of *S. cerevisiae* TFs compiled in Wang et al. (Wang et al., 2010) and used Fisher's exact test to compare the ratios of TFs in lost, gained and duplicated genes to the ratio of TFs in core genes.

2.4.7 Calculating evolutionary rate

In order to estimate the level of selective constraint on individual proteins, or evolutionary rate, we compared *S. cerevisiae* to its second closest relative, *S. mikatae*. We measured the rate of non-synonymous substitutions over the rate of synonymous substitutions (K_a/K_s) for all *S. cerevisiae* proteins not lost in *S. mikatae*. K_a/K was calculated according to the Yang-Neilsen method (Yang & Nielsen, 2000) using PAML (Yang, 2007).

2.4.8 Controlling for lineage-independent propensity for gene loss

We defined the overall propensity for gene loss (PGL) as the number of independent loss events divided by the total branch length where a loss could have occurred (see estimating relative branch lengths). Genes lost exclusively in distant lineages tend to have a lower average PGL than genes lost in close species. To control for this difference, we discarded the proximally lost genes with the highest PGL values, one at a time, until their average PGL fell below that of distantly lost genes.

2.4.9 Estimating relative branch lengths

In order to estimate relative branch lengths along the tree, we selected 3 slowly evolving, universally conserved proteins (UBA1, URA2 and EFT2), calculated the rate of missense substitutions (K_a) between all pairs of species with PAML 4 (Yang, 2007) and used the median K_a as the distance between two species. Then, we calculated the branch lengths in a stepwise manner, starting from the closest pairs of organisms/phyla and progressing upwards along the tree, until the all branch lengths are inferred.

Tables

	Core	Lost	Gained	Duplicated
Number of TFs	16	97	28	45
Percent of TFs	0.8	4.0	5.8	4.3
P-value*	-	$2.1x10^{-11}$	$2.5x10^{-5}$	$1.7x10^{-7}$

Table 2.1: Transcription factor enrichment in lost and gained genes. *:based on Fisher's exact test (see Methods)

GO term	GO term ID	# of genes	Fold enrichment	P-value*
cell wall organization	GO:0071555	30	3.18	0.00082
sequence-specific DNA binding	GO:0043565	30	2.49	0.0065
positive regulation of gene expression	GO:0010628	32	2.13	0.016
positive regulation of metabolic process	GO:0009893	44	1.84	0.016
regulation of transcription from RNA polymerase II promoter	GO:0006357	44	1.80	0.016
positive regulation of transcription, DNA-dependent	GO:0045893	31	2.06	0.022
reproductive process	GO:0022414	45	1.74	0.025
cell periphery	GO:0071944	51	1.64	0.027
regulation of RNA metabolic process	GO:0051252	76	1.42	0.035
positive regulation of RNA metabolic process	GO:0051254	32	1.90	0.038
regulation of RNA biosynthetic process	GO:2001141	74	1.42	0.043
regulation of transcription, DNA-dependent	GO:0006355	74	1.42	0.043

Table 2.2: GO terms significantly enriched in gained genes as compared to core genes. *: based on Fisher's exact test

GO term	GO term ID	# of genes	Fold enrichment	P-value*
transmembrane transport	GO:0055085	110	2.11	$2.5x10^{-6}$
cell periphery	GO:0071944	138	1.86	$4.3x10^{-6}$
plasma membrane	GO:0005886	92	2.21	$9.7x10^{-6}$
intrinsic to membrane	GO:0031224	260	1.35	0.00035
sequence-specific DNA binding	GO:0043565	60	2.09	0.0010
reproductive process	GO:0022414	103	1.67	0.0011
zinc ion binding	GO:0008270	86	1.61	0.0065
cell wall organization	GO:0071555	39	1.74	0.037

Table 2.3: GO terms significantly enriched in lost genes as compared to core genes.
*: based on Fisher's exact test

Figures

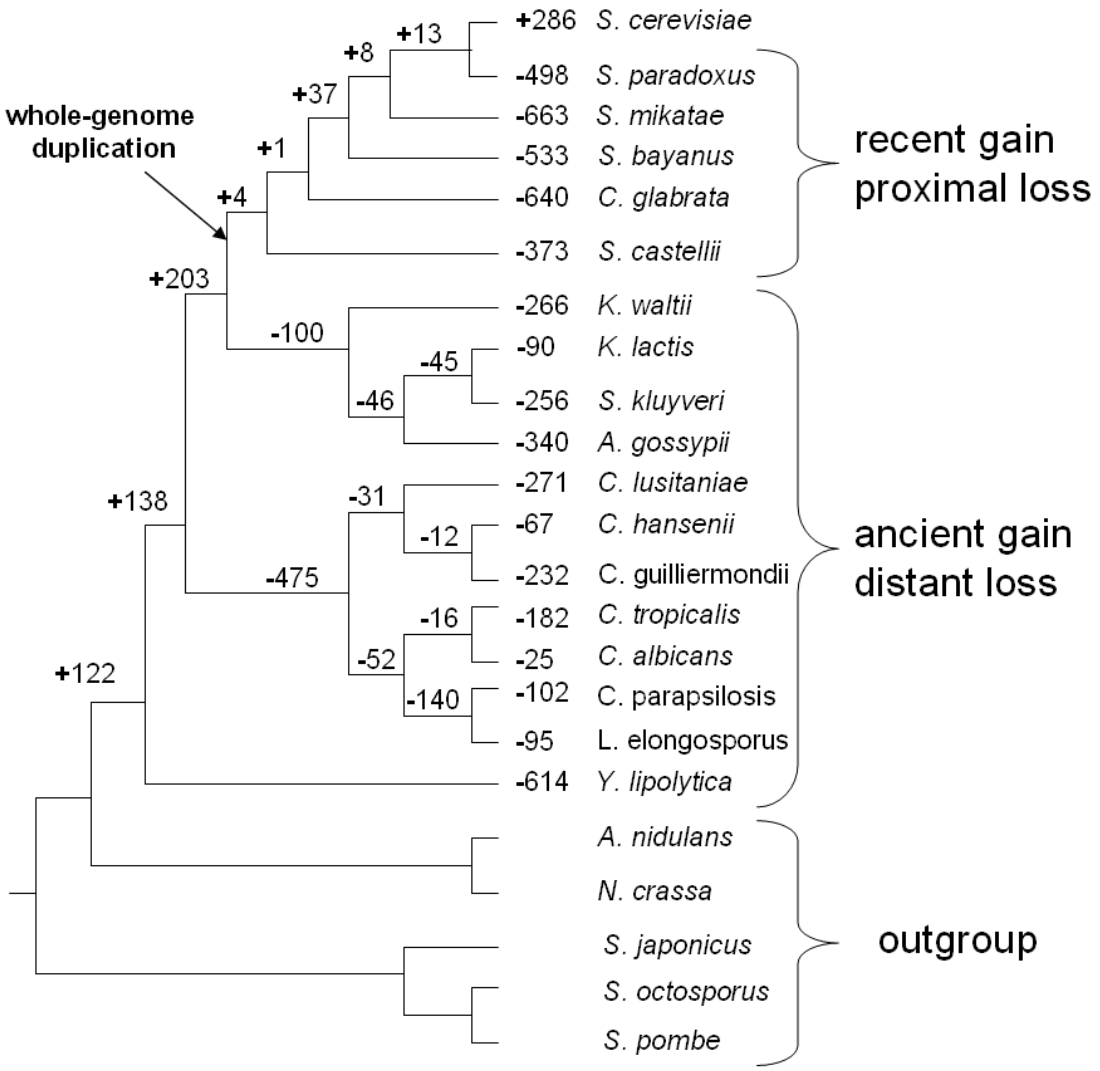

Figure 2.1: Inferred gene loss and gain events displayed along the yeast phylogenetic tree. The "+" sign denotes gains and the "-" sign, losses. Recent gains and proximal losses were defined as those having occurred after the split with *K. waltii*, with the exception of genes gained in *S. cerevisiae*.

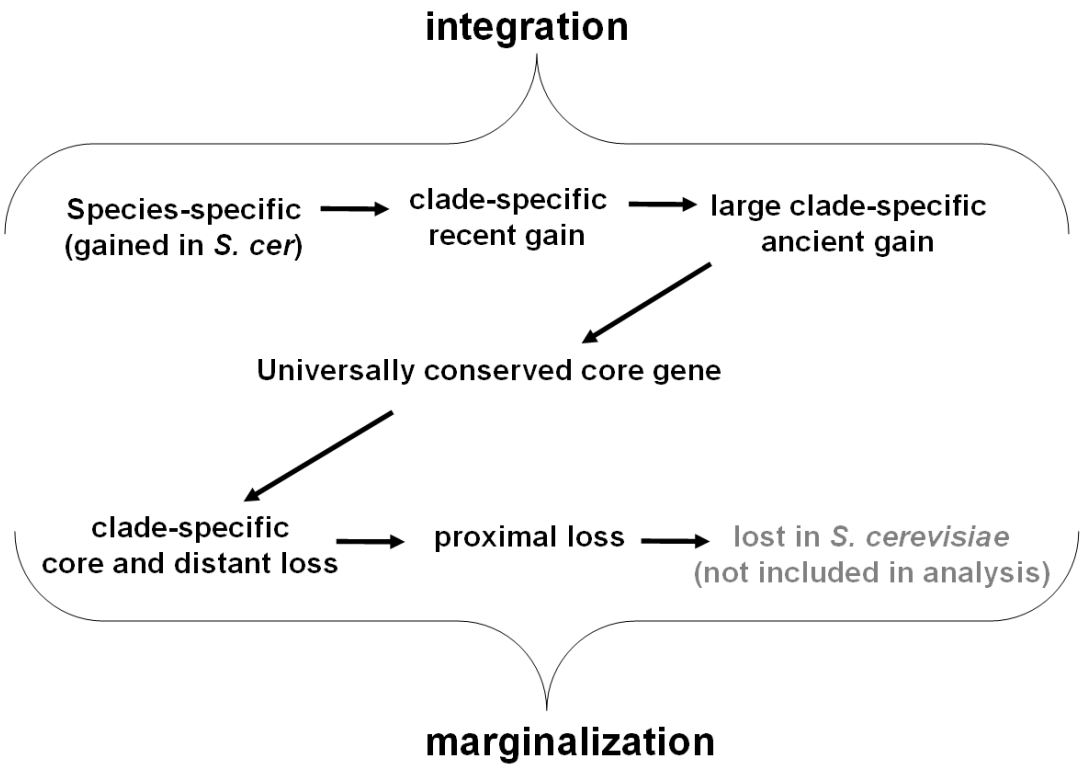

Figure 2.2: The life-cycle of genes. Depiction of the complete life-cycle of genes, from gene gain to gene loss, and how we inferred the life-cycle stage of different genes based on the representation of their orthologs across the tree.

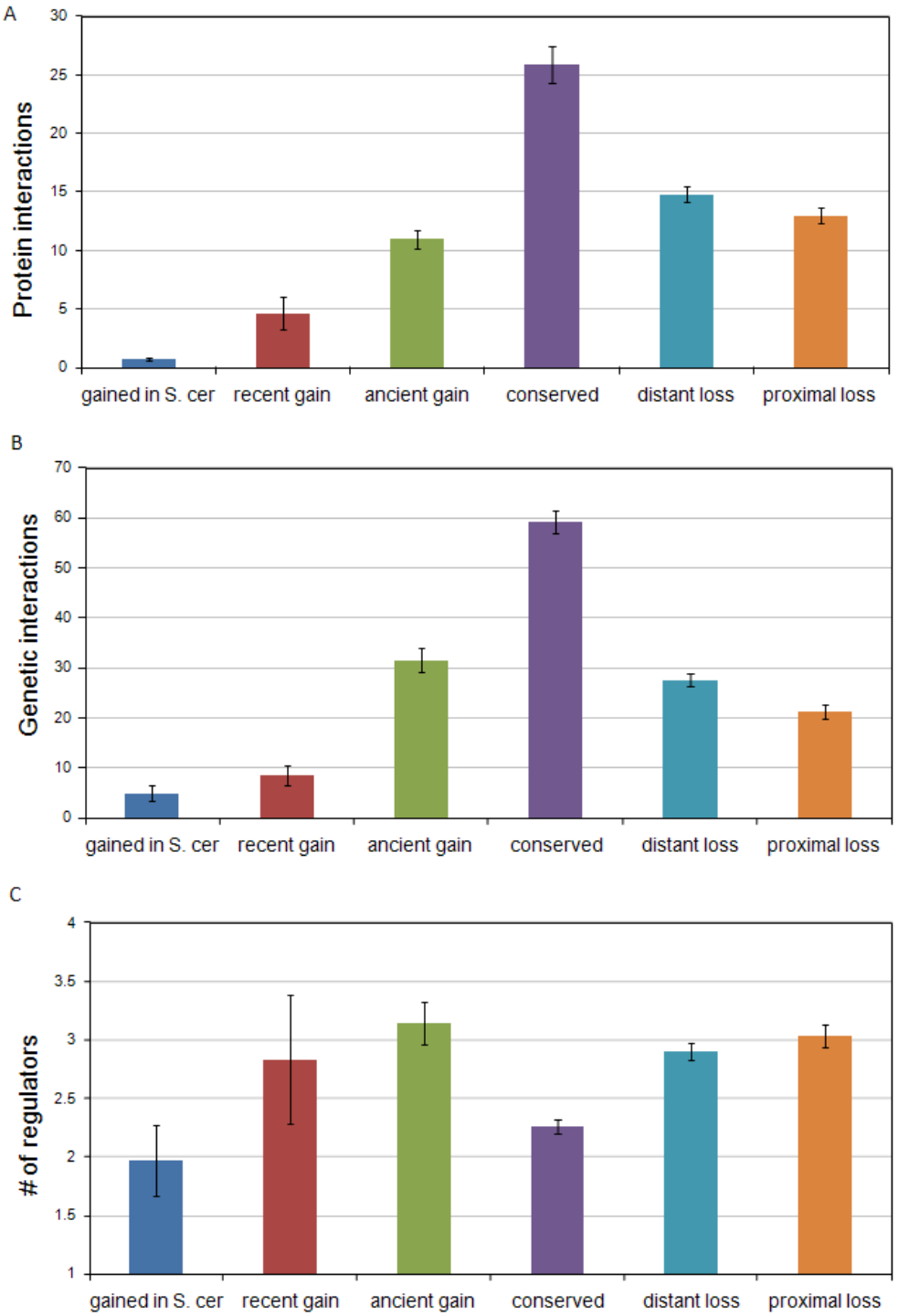

Figure 2.3: Network properties of lost and gained genes, including (A) PPI degree, (B) genetic interaction degree and (C) regulatory in-degree. Conserved genes are those which are universally conserved across all 23 species. Loss and gain categories are defined based on a threshold distance from *S. cerevisiae*, as shown in Figure 2.1.

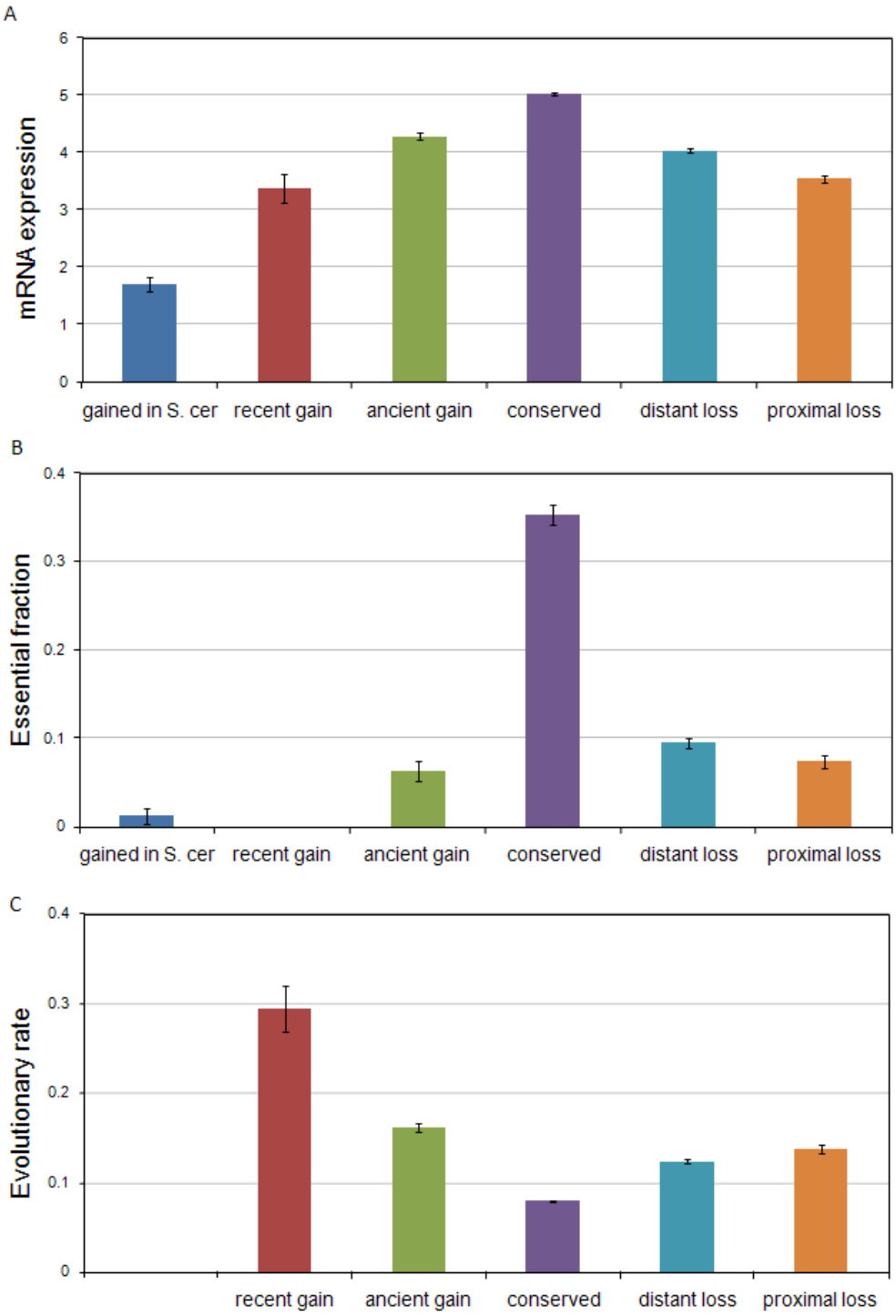

Figure 2.4: Genomic properties of lost and gained genes, including (A) log of mRNA read count in rich media RNA-seq (Ingolia et al., 2009), (B) the fraction of genes which are essential (Winzeler, 1999) and (C) protein evolutionary rate between *S. cerevisiae* and *S. mikatae* (see Methods)

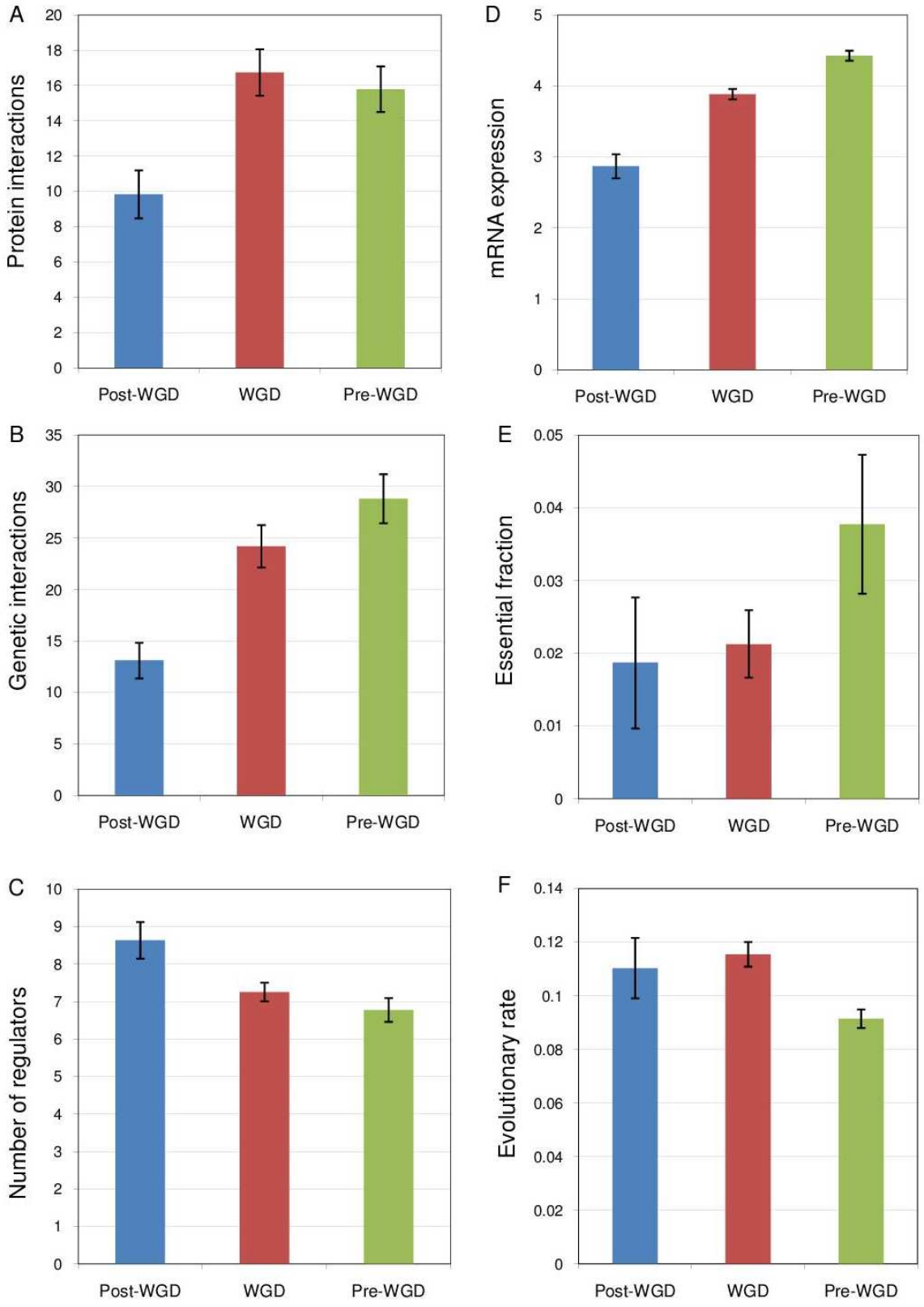
A
Protein interactions
20
18
16
14
12
10
8
6
4
2
0
Post-WGD
WGD
Pre-WGD
D
mRNA expression
5
4
3
2
1
0
Post-WGD
WGD
Pre-WGD
B
Genetic interactions
35
30
25
20
15
10
5
0
Post-WGD
WGD
Pre-WGD
E
Essential fraction
0.05
0.04
0.03
0.02
0.01
0
Post-WGD
WGD
Pre-WGD
C
Number of regulators
10
9
8
7
6
5
4
3
2
1
0
Post-WGD
WGD
Pre-WGD
F
Evolutionary rate
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Post-WGD
WGD
Pre-WGD

Figure 2.5: Properties of genes gained by duplication (see methods), and including (A) PPI degree, (B) genetic interaction degree, (C) regulatory in-degree, considering the union of ChIP-chip and literature-derived edges (Teixeira et al., 2006) (D) log of mRNA read count in rich media (E) the fraction of genes which are essential and (C) protein evolutionary rate between *S. cerevisiae* and *S. mikatae* (see Methods). Duplication events are separated into three age groups: before the whole-genome-duplication (pre-WGD), during (WGD) or after (post-WGD).

Chapter 3

Alternative Splicing and Interactome Complexity

3.1 Introduction

Alternative splicing (AS) is known to vastly increase proteome diversity in mammals and especially in primates (Barbosa-Morais et al., 2012), yet the mechanisms by which this proteome diversity contributes to phenotypic complexity remains largely unknown. Transcriptome sequencing on different tissues in human and mouse shows that almost all multi-exon genes exhibit alternative splicing (Pan et al., 2008; Wang et al., 2008) and AS events have been shown to play crucial roles in development, phosphorylation-network rewiring and cell type differentiation (Gabut et al., 2011a; Kalsotra & Cooper, 2011; Merkin et al., 2012) among other functions. However, a global view of the influence of AS on biomolecular networks is still lacking. Just as high-throughput methods were essential to studying AS at the transcriptome level (Modrek & Lee, 2002), high-throughput methods are also desperately needed to study its effects at the systems level.

Recent computational studies based on transcriptome profiling found that tissue-specific exons are enriched in conserved disorder residues and in linear binding motifs (Buljan et al., 2012; Dinkel et al., 2012), highlighting PPI rewiring as a potentially important function of tissue-specific splicing. The differential protein binding affinities of isoforms

could potentially allow for specialization of tissue-specific PPI networks without increasing the number of gene loci in the genome. A recent study, experimentally deleting tissue-specific exons in 43 genes and systematically testing its effects on known interactions, demonstrated that AS can turn on or off PPIs (Ellis et al., 2012). These studies, along with a growing body of anecdotal examples (Thakar et al., 2012; Wethkamp et al., 2011), all point to the possibility that PPI network rewiring could be a primary function of AS. Since the majority of human genes undergo AS, it is plausible that a large fraction of PPIs may be influenced by AS. This creates an urgent need for a new interactome systematically mapping the interactions of naturally occurring isoforms in order to achieve a systems-level view of AS-mediated PPI rewiring. As illustrated in Figure 3.1A, such a network will also allow us to observe and study the network's precise structure, which has so far been blurred in traditional interactome studies by the reliance on the implicit assumption that isoforms of a gene behave identically in the network.

In order to study the genome-wide influence of AS-mediated PPI rewiring, we first selected 1,518 genes as a representative sample of the human protein-coding genome, where roughly half are known disease-associated proteins and half were selected randomly (see Methods for details). Starting from these genes, we carried out systematic cloning and sequencing of native mRNA transcripts in 5 tissues using a high-throughput method we have recently developed (Yang et al., 2011), allowing us to study both novel and known isoforms. We then performed genome-scale yeast-two-hybrid interaction screening (Braun et al., 2009; Li et al., 2004; Rual et al., 2005; Yu et al., 2008) against

>15,000 open reading frame (ORF) clones from ~13,000 genes (ORFeome v5.1). This is the first interactome of systematically cloned native isoforms, as well as the first glimpse into the genome-wide interaction profiles of protein isoforms. This unprecedented dataset allows us to assess at the global level the scale and general principles of the influence of AS on the interactome network.

3.2 Results

3.2.1 Mapping isoform interactome network

For 1,518 genes, mRNA transcripts from 5 human tissues were systematically amplified, gateway cloned and submitted to deep sequencing (Figure 3.1B). We obtained 2,130 alternative spliced isoforms from 1270 genes. Approximately half of these genes are disease-associated genes from the OMIM database (Hamosh et al., 2005) (see methods). 71% of the full length transcripts are present in the AceView (Thierry-Mieg & Thierry-Mieg, 2006) or Gencode (Harrow et al., 2012) databases, along with 95% of exon-exon junctions and 97% of splice-site positions (Figure 3.1C). This suggests that the isoforms are largely the product of naturally occurring splicing in the cells. Nevertheless, many transcripts contained premature stop codons (~30%), and were thus discarded from further sequence-level analyses.

We obtained at least two splice isoforms for 447 genes, resulting in a total of 1,091 newly cloned splice isoforms and 175 reference ORFs from our ORFeome collection. These

splice isoforms were screened for interactions using yeast-two-hybrid against ~13,000 genes in human ORFeome v5.1 (Figure 3.1B). For each interaction detected during screening, all isoforms of the gene were then tested for the interaction in four independent pairwise tests, followed by Sanger sequencing to confirm their identities. We found a total of 1,440 isoform interactions between 345 isoform baits from 212 genes and 477 preys from human ORFeome (Figure 3.1D), collapsing into 967 gene-level interactions.

In contrast to traditional PPI networks where each node represents a gene locus (Braun et al., 2009; Li et al., 2004; Rual et al., 2005; Yu et al., 2008), the isoform-resolved network can reveal differences in the interaction profiles of isoforms derived from the same locus. As a way of quantifying the amount of PPI rewiring between two isoforms, we defined the rewiring score as the fraction of total interactions that are not shared between two isoform interaction profiles. As shown in Figure 3.1E, comparing isoform interaction profiles reveals a large amount of rewiring, with 88.8% of isoform pairs having at least one rewired interaction (rewiring score>0). This result strongly supports that PPI rewiring is a major function of AS in coding regions.

3.2.2 Enhanced network modularity at isoform resolution

Just as proteins from different genes play distinct roles in the network, protein isoforms of the same gene may similarly assume different roles and locate in different network neighborhoods. To assess the degree of this network-level functional diversification

between isoforms, we measured for each isoform pair the average probability of interaction between the binding partners of the two isoforms (cross-AS clustering coefficient). As shown in Figure 3.2A, we found that isoform pairs with a high rewiring score tend to have binding partners which are less likely to interact with each other (Spearman's rank correlation $p<2.2\text{x}10^{-16}$). Connecting isoform pairs of the same gene and treating these connections as PPI network edges (AS-edges), we found that AS-edges between highly rewired isoforms tend to bare larger betweenness centrality (cross-AS betweenness) than other AS-edges (Figure 3.2B, Spearman's rank correlation $p=7.5\text{x}10^{-5}$). Both of these results indicate that highly rewired isoform pairs from the same gene tend to locate in different network neighborhoods, similarly to proteins from different genes.

As illustrated in Figure 3.1A, due to isoforms participating in different interactions, isoform-level resolution may reveal a considerably different network structure than is visible in gene-level interactomes. This can have important implications for our understanding of the network organization and complexity underlying cellular and disease processes. To assess this we overlaid the isoform network with different types of functional genomics information and asked whether signatures of functional relatedness could be influenced by proteins interacting with the same or with different isoforms. As shown in Figures 3.2C and 3.2D, for those interactions rewired by AS, pairs of proteins each interacting with different isoforms of the same gene tend to be more distant in an independent PPI network (CCSB, 2012) (t-test $p<2.2\text{x}10^{-16}$), and less likely to share

specific functional annotations (Gene Ontology (Harris et al., 2004) [GO] categories with <25 genes) (t-test p=7.7x10^{-5}) than proteins binding to the same isoform(s) (see Figure 3.2 legend for details). These results demonstrate how AS co-rewires functionally similar interactors and differentially rewires functionally distinct interactors. In addition, we found that proteins binding the same isoforms are significantly more co-expressed across 16 human tissues (Illumina, 2011) than proteins binding to different isoforms (Figure 3.2D, t-test p=0.023, see Methods for details). This shows that the transcriptional and splicing regulatory machineries work in synchrony to rewire the PPI network across different tissues. We also found that proteins binding the same isoforms are more likely to be associated with the same disease (Safran et al., 2010), or interact with proteins associated with the same disease (t-test p=1.7x10^{-15}). This enhanced modularity of disease associated genes suggests that disease genes as well as potential drug targets can be predicted with higher accuracy and disease pathways better understood when the splicing sensitivity of interactions is considered. Together, these results demonstrate that splicing regulated PPI rewiring is a non-random, well-regulated process which affects very specific interactions, and thus contributes in a major way to the overall organization and function of the system. As compared to gene-level interactome networks, the consideration of the isoform specificity of interactions results in a more modular and coherent network from which we can more accurately delineate the pathways and functional modules which form the basis of our understanding of cellular and disease processes.

3.2.3 AS-mediated rewiring types, network pleiotropy, and tissue-specificity

While most AS is associated with either loss or gain of specific interactions, some AS is associated with the simultaneous loss of some interactions and gain of other interactions. The first type of AS-mediated rewiring creates isoforms with a subset of the interactions of other isoforms ("subset On/Off rewiring") or no interactions at all ("On/Off rewiring"), which could be considered subfunctional isoforms from a network perspective. The second type of AS-mediated rewiring creates isoforms with mutually exclusive interactions ("change-over rewiring") and thus locating to different network neighborhoods, an AS-regulated form of interaction pleiotropy. While the first type of rewiring (On/Off rewiring) is analogous to On/Off switches in electrical circuits, the second type of rewiring (change-over rewiring) is analogous to change-over switches. This isoform interactome reveals different patterns of rewiring between isoforms (Fig. 3A), enabling the classification of genes according to these rewiring patterns (Fig. 3B). By locating in different network modules, isoforms of change-over rewiring genes can potentially participate in different cellular or disease processes. As an example (Fig. 3C), different isoforms of the gene CD99L2 interact exclusively with proteins from different disease subnetworks, revealing a mechanism by which AS-mediated rewiring can explain the genetic pleiotropy of certain genes. Change-over rewiring genes have interactors which are more distant in an independent network (The Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute, 2012) (Wilcoxon test $P = 0.0022$, Fig. 3D) and show higher betweenness centrality (Wilcoxon test $P = 3.0 \times 10^{-6}$, Fig. 3E) than other rewiring or non-rewiring genes.

AS occurs prevalently in multi-cellular organisms and much of AS is tissue-specific (Castle et al., 2008; Pan et al., 2008; Wang et al., 2008). Furthermore, AS events have been demonstrated to play key roles in cell-type differentiation and tissue specialization (Bland et al., 2010; Gabut et al., 2011b; Huot et al., 2012; Shapiro et al., 2011). To assess whether AS-mediated interaction rewiring is associated with tissue-specific functions, we looked at the expression profiles of binding partners whose interactions are either rewired or non-rewired by AS across a publicly available 16-tissue RNA-seq dataset (Illumina, 2011). We found that binding partners affected by AS-mediated rewiring are expressed in a more tissue-specific manner than non-rewired binding partners (t-test $P = 0.024$), as measured by the range of their expression levels across the 16 tissues, despite having similar expression levels on average (t-test $P > 0.05$). This indicates that AS-mediated PPI network rewiring tends to affect tissue-specific functions. This result is consistent with the recent studies that showed tissue-specific exons are enriched in conserved linear motifs (Buljan et al., 2012; Ellis et al., 2012), and provides the first genome-scale empirical evidence that AS-mediated network rewiring preferentially regulates tissue-specific functions. However, after breaking down rewiring events by type (Fig. 3F), we found that binding partners whose interactions are affected by change-over rewiring are expressed in a more tissue-specific manner than other rewired or non-rewired interaction partners. Since most rewired interactors are affected by multiple types of rewiring, we further examined the tissue-specificity of rewired interactors that are not affected by change-over rewiring (Fig. 3G). We found that change-over rewired binding partners have significantly more tissue-specific expression profiles than other rewired (t-test

p=0.018) or non-rewired (*P* = 0.0035) binding partners (Fig. 3G), while having comparable average expression levels (t-test P > 0.05). This suggests that change-over rewiring alone may account for all of the observed tissue-specificity of AS-mediated PPI rewiring.

These significant differences between change-over rewiring and other rewiring types suggest that there are two topologically and functionally distinct classes of rewiring. While On/Off rewiring and subset On/Off rewiring may serve to regulate the activity of a protein by shutting off some or all interactions, change-over rewiring creates isoforms with distinct interactions in the network, leading to functional diversification. In contrast to On/Off and subset On/Off rewiring genes, which are in many ways similar to non-rewiring genes, change-over rewiring genes play a unique role in rewiring the interactome between tissues and across network modules. This new dimension of interactome network dynamics thus reveals key differences between genes which would have remained hidden from a gene-level protein interactome perspective.

3.2.4 Sequence modules and mechanisms of AS-mediated network rewiring

Comparing the sequences of isoforms enables the identification of the sequence-level determinants of PPI rewiring, which may reveal interaction sites and precise rewiring mechanisms. For each rewired interaction, we used a sliding window approach to define, when possible, alternatively spliced (AS) regions which are unique to, and universally shared by, either all interacting isoforms or all non-interacting isoforms (see Methods for details). These regions presumably mediate PPI rewiring events. For rewiring genes with

only 2 isoforms, it is trivial to identify a region associated with rewiring. In genes with 3 or more isoforms, we observed that about 56% of rewiring events can be explained by a single sequence module promoting or blocking a set of interactions, which is significantly higher than the expected (26%) based on random shuffling of the isoforms participating in each gene-level interaction (Figure 3.4A, Fisher's exact test p=0.00040, see Methods for details). This observation suggests that the alternative inclusion of a single protein region is sufficient to promote or block interactions in most cases. As shown in Figure 3.4B, we found that AS regions associated with rewiring have a significantly lower DNA mismatch rate between human and mouse than the average over the entire coding region (paired t-test p=0.00012), or other AS regions normalized by the average for each gene (unpaired t-test p=8.7e-5, see Methods). The high sequence-level constraint on these interaction-regulating sequence modules suggests that these regions constitute specific functional units, similar to protein domains. At the same time, the rewiring modules are complementary to the concept of domains in that they are derived entirely from interaction rewiring and isoform sequences rather than from sequence homology. We identified potential interaction promoting and blocking regions in a roughly 2-to-1 ratio, where each mechanism rewires interactions globally at comparable frequency (Figure 3.4C). There are a smaller number of cases which can be explained by either mechanism. In addition, some cases cannot be explained by either mechanism ("complex" in Figure 3.4C), which includes rewiring events potentially mediated by unique exon-exon junctions or combinations of multiple AS regions.

To evaluate the "sequence module" model of AS-mediated PPI rewiring, we asked whether rewiring events could be predicted from the inclusion or exclusion of rewiring-associated regions. We separated our dataset into two mutually exclusive sets: the test set consisting of isoforms with no interactions (zero-degree), and the training set consisting of isoforms with one or more interactions. After defining rewiring regions using only the interacting set of isoforms (training set), we found we could correctly predict the rewiring of 260 out of 293 (89%) interactions lost in zero-degree isoforms (test set) based solely on the inclusion or exclusion of rewiring-associated AS regions. This result reinforces the model whereby interactions are regulated by the AS of localized sequence modules and suggests most of the interactions lost in zero-degree isoforms are due to the AS of these sequence modules rather than the sensitivity of the interaction detection method.

PPIs are generally mediated either via domain-domain interactions or short linear motifs which interact with linear motif binding domains (LMBDs) (Dinkel et al., 2012; Neduva & Russell, 2006), such as the SH3 domain binding to the PxxP motif. Previous studies have shown that tissue-specific exons tend to contain linear motifs (Buljan et al., 2012; Ellis et al., 2012), but it remains to be shown whether AS of these linear motifs actually causes PPI rewiring and whether this is a dominant mechanism. To answer these questions we first asked whether rewired binding partners contain more LMBDs than expected. As shown in Figure 3.4D, we found a ~3-fold enrichment of LMBDs in rewired partners as compared to non-rewired partners (Fisher's exact test $p=1.8\text{x}10^{-11}$). This enrichment suggests that as many as 41% of AS-mediated rewiring events may

involve known LMBDs. We then identified linear motif matches (Dinkel et al., 2012; Neduva & Russell, 2006) in AS regions and found that they appear at greater frequency in AS regions associated with PPI rewiring than other AS regions (Figure 3.4E, Wilcoxon test p=0.0037), suggesting that the alternative inclusion of linear binding motifs constitutes a common mechanism for AS-mediated PPI rewiring. Figure 3.4F shows an example of PPI rewiring likely mediated through the AS of a known linear motif.

Many domain-domain interactions have been documented (Finn et al., 2005; Raghavachari et al., 2008; Stein et al., 2009) and the alternative inclusion of these domains may explain some of our observed rewiring events. As shown in Figure 3.4G, we found that in 47 out of 53 cases (89%) where a rewiring-associated AS region overlaps a domain known to interact with a domain in a rewired binding partner, we could correctly predict which isoforms participate in the interaction (Fisher's exact test $p=6.3\text{x}10^{-5}$). This attests to the high quality of detected rewiring events and demonstrates that AS disrupting known domain-domain interactions is an important mechanism for mediating PPI rewiring alongside AS of linear binding motifs.

This analysis has revealed that AS-mediated rewiring is most often traceable to single localized AS sequence modules, which represent conserved functional units, and that many rewiring events can be explained by AS disrupting linear motifs or known domain-domain interactions. Figure 3.4H illustrates how both promoting and blocking rewiring

mechanisms can be explained by the alternative inclusion of localized sequence modules overlapping domains or linear motifs.

3.3 Discussion

The systematic cloning of native splice isoforms and genome-scale mapping of isoform interactions enabled us to capture the different types of AS-mediated PPI rewiring, providing much needed insight into its global influence on the interactome network. With this first comprehensive isoform-resolved interactome, we have thoroughly established that PPI network rewiring is a major function of AS. We have shown that AS not only increases proteome diversity but also network complexity. Compared to traditional interactomes, where each node represents a locus, the isoform interactome more accurately captures the precise structure of the network, as demonstrated by its enhanced organizational and functional coherence. In addition, we found that AS-mediated PPI rewiring preferentially affects tissue-specific functions, demonstrating the large-scale importance of AS to the functional specialization of tissues. By uncovering a previously hidden dimension of network structural dynamics, we discovered that rewiring patterns can be classified into different types, with distinct topological and functional consequences. In particular, we identified change-over rewiring as an important modulator of tissue-specific function and network organization, and suggested that AS-mediated interaction pleiotropy may serve as a driving force for genetic and disease pleiotropy. We found that most interaction rewiring events are mediated by conserved,

localized sequence modules and tend to be traceable to the AS of known interaction elements, such as linear motifs and protein interaction domains. In summary, we have shown that AS plays a crucial role in network organization, function and cross-tissue dynamics, demonstrating the importance of a splicing-sensitive global view of biomolecular networks to our understanding of disease and systems biology in multicellular organisms.

3.4 Methods

3.4.1 Binding partner co-expression

Using all 75 base pair runs from the Illumina Body Map 2.0 16-tissue RNA-seq dataset (Illumina, 2011) and the Bowtie alignment tool (Langmead & Salzberg, 2012) with default settings, we mapped reads to all ORFeome v5.1 (Rual et al., 2004) clone sequences and calculated the $\log_2$ read count for each gene for each tissue. We then normalized expression values for each gene to that of the upper-quartile most highly expressed gene for each tissue, as suggested in (Bullard et al., 2010), and performed Pearson correlation on all pairs of binding partners.

3.4.2 Defining alternatively spliced (AS) regions and rewiring-associated AS regions

Sliding a window of 10 residues over each isoform of a gene, we asked for each position, to which other isoforms the window matches perfectly. The AS region is then defined as

the widest merged window which maps to the same subset of isoforms as all 10 residue windows within. Regions which map to all isoforms of a gene are considered constitutive regions. Regions which map to the entire set of isoforms either participating in a rewired interaction or not participating, and map only to these isoforms, are considered rewiring-associated AS regions.

3.4.3 Measuring tissue-specific splicing

Using all 75 base pair runs from the Illumina Body Map 2.0 16-tissue RNA-seq dataset (Illumina, 2011) and the Bowtie alignment tool (Langmead & Salzberg, 2012), we mapped reads to all isoform clone sequences. Applying the same logic as for the "sliding window" described in the last section, reads were classified according to the subset of isoforms to which they mapped. Groups of reads mapping to all isoforms were considered to map to constitutive regions and were thus used to estimate transcriptional expression changes. Groups of reads mapping to only a subset of isoforms were considered as mapping to AS regions and were thus used to estimate splicing-level changes across tissues. The $\log_2$ upper-quartile gene expression-normalized read count was used as the expression measure for each read group in each tissue. The level of tissue-specific splicing for each gene was defined as the average range of expression normalized for each gene by the range of expression of constitutive regions, to control for transcriptional tissue-specificity.

3.4.4 Conservation of AS regions

AS regions and rewiring-associated AS regions were defined as described earlier. After mapping isoform clone sequences onto the human genome as described earlier, we used the MULTIZ human-mouse pairwise whole-genome alignment (Blanchette et al., 2004) to identify which bases differ between the two genomes. First, we performed a paired t-test, comparing the DNA mismatch rate in rewiring-associated AS regions to the average for the entire coding region of the gene. This shows that rewiring-associated AS regions are more conserved than the rest of the coding sequence. Then, for each AS region, we divided the region's mismatch rate by the average rate for the gene, allowing us to compare the conservation of rewiring-associated AS regions to other AS regions while controlling for gene-level variation in mutation rate and selective constraint. This shows that rewiring-associated AS regions are more conserved than AS regions not associated with rewiring.

Figures

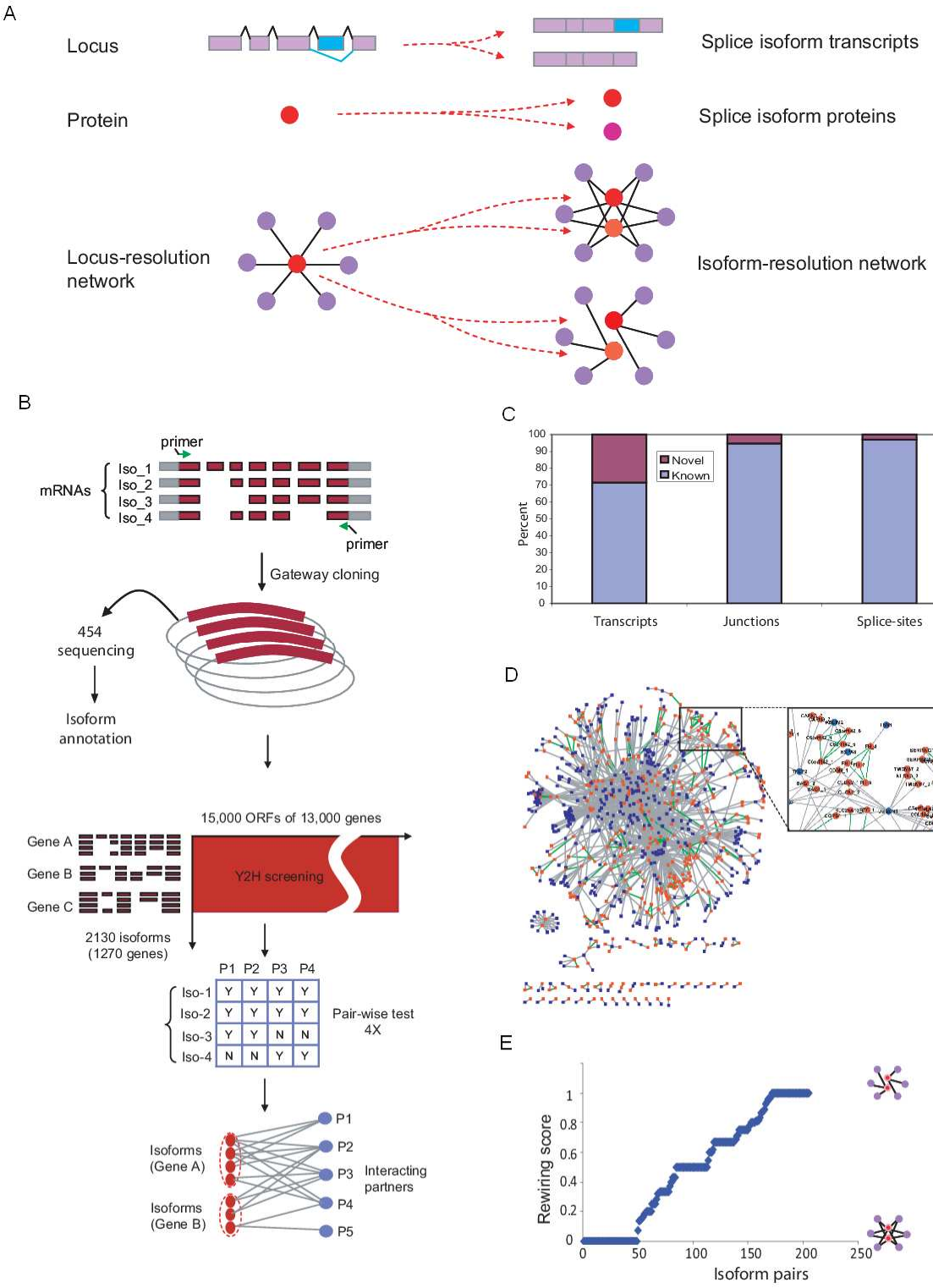
A
Locus
Splice isoform transcripts
Protein
Splice isoform proteins
Locus-resolution network
Isoform-resolution network
B
primer
mRNAs
Iso_1
Iso_2
Iso_3
Iso_4
primer
Gateway cloning
454 sequencing
Isoform annotation
15,000 ORFs of 13,000 genes
Gene A
Gene B
Gene C
Y2H screening
2130 isoforms (1270 genes)
P1 P2 P3 P4
Iso-1 Y Y Y Y
Iso-2 Y Y Y Y
Iso-3 Y Y N N
Iso-4 N N Y Y
Pair-wise test 4X
Isoforms (Gene A)
Isoforms (Gene B)
P1
P2
P3
P4
P5
Interacting partners
C
Percent
Novel
Known
Transcripts
Junctions
Splice-sites
D
E
Rewiring score
Isoform pairs

Figure 3.1: Experimental procedure and overview of the isoform interactome. A) Illustration of the concept of isoform resolution. B) Outline of experimental pipeline. C) Percent of transcripts, junctions and splice-sites found in the AceView (Thierry-Mieg & Thierry-Mieg, 2006) or Gencode (Harrow et al., 2012) databases. D) Isoform interactome network. Red nodes: splice isoforms; blue nodes: interacting partners from ORFeome; grey edges: protein-protein interaction; green edges: connecting two isoforms belonging to the same gene. E) Distribution of rewiring scores between two isoforms, defined as the fraction of total interactions that are not shared between two isoform interaction profiles.

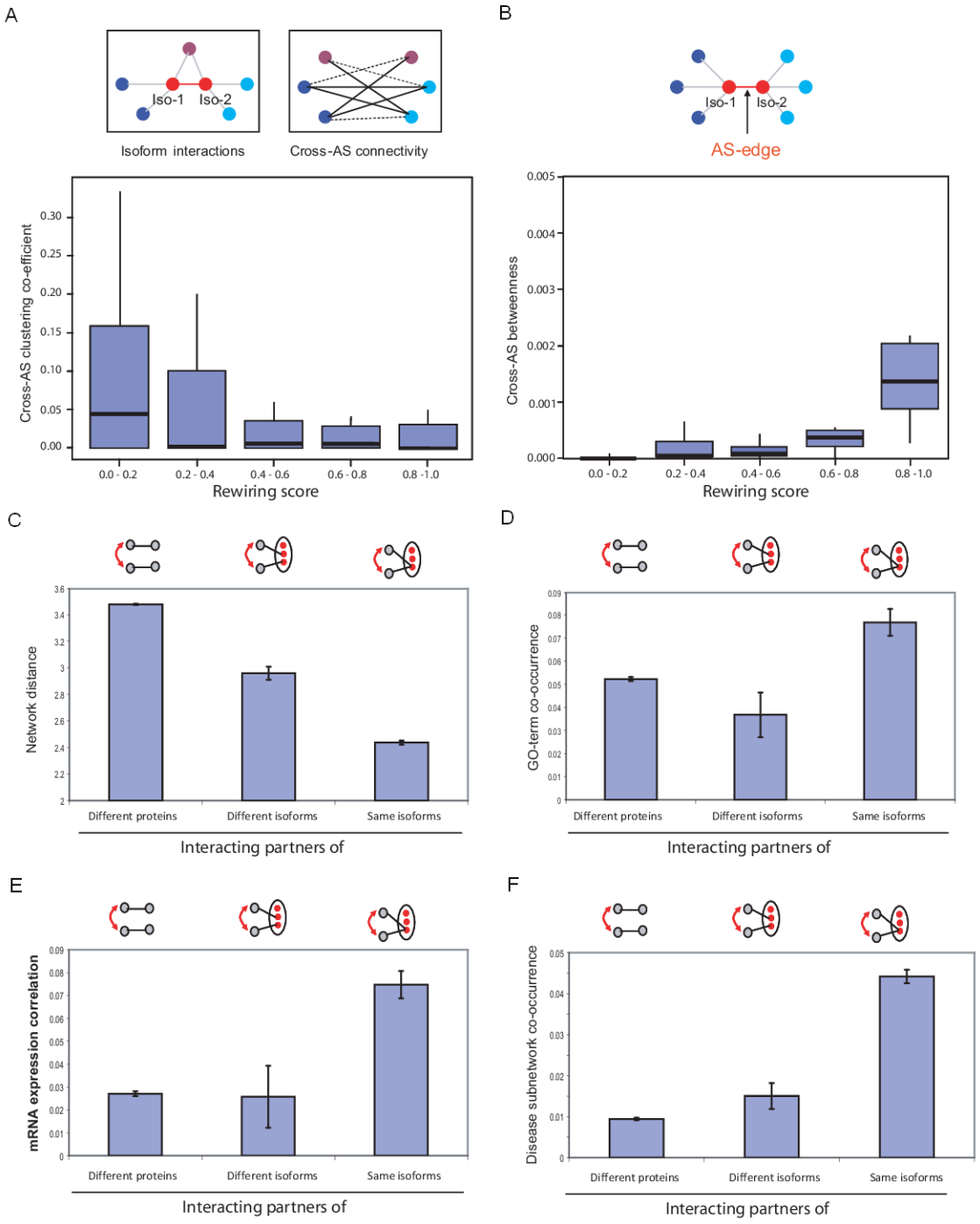

Figure 3.2: Network modularity at the isoform resolution. A) Cross-AS clustering coefficient as a function of rewiring score. Cross-AS clustering coefficient is defined as

the probability that the interactors of different isoforms interact with each other based on an independent human interactome dataset (CCSB,2012). This demonstrates that interactors of highly rewired isoform pairs are less likely to interact with each other. B) Cross-AS betweenness as a function of rewiring score. Cross-AS betweenness is defined as the betweenness centrality in the human interactome of the AS "edge" when treated as another PPI edge. Betweenness centrality of a node/edge measures the fraction of shortest paths in the network that pass through it. This result shows that highly rewired isoform pairs tend to locate to different modules in the network. C) Mean shortest path distance in the human interactome between pairs of proteins interacting with the same subset of isoforms ("same isoforms"), one or more different isoforms ("different isoforms"), or never interacting with proteins of the same gene ("different proteins"). D) Mean Jaccard index of protein pair co-occurrence in GO (Harris et al., 2004) categories of less than 25 genes. The Jaccard index is defined as the number of shared occurrences over the union. E) Mean Pearson correlation coefficient of upper-quartile-normalized $\log_2$ RNA-seq read counts from 16 human tissues (Illumina, 2011) (see Methods). F) Mean Jaccard index of disease subnetwork co-occurrence of protein pairs, with disease subnetworks defined for each disease as the set of disease associated genes from GeneCards (Safran et al., 2010) and their first neighbors in the human interactome (CCSB,2012).

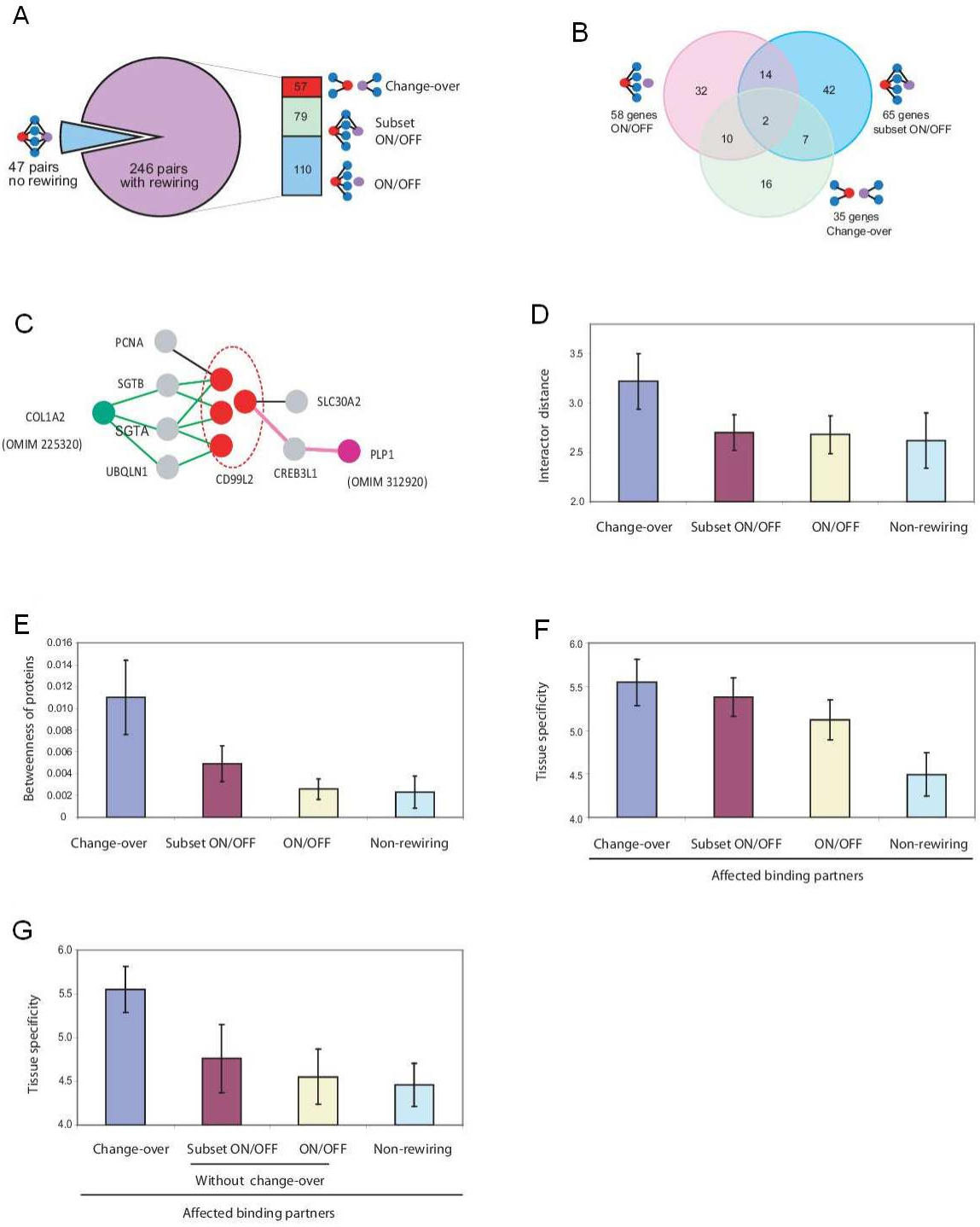

Figure 3.3: Rewiring types, network pleiotropy, and tissue specificity. A) Isoform pairs from interacting genes, classified as either non-rewiring (blue) or rewiring (purple),

where at least one interaction is rewired. Rewired isoform pairs are further subclassified into On/Off rewiring, where one isoform loses all interactions, subset On/Off rewiring, where one isoform loses part of the interactions, and change-over rewiring, where each isoform has at least one exclusive interaction partner. B) The numbers of genes involved in the 3 different types of rewiring. C) The isoform interaction profiles of the CD99L2 gene as an example illustrating how change-over rewiring can explain the genetic pleiotropy of certain genes through different isoforms participating in different disease subnetworks exclusively from each other. D) Mean shortest path distance in the human interactome (CCSB, 2012) of protein pairs interacting with the same rewiring or non-rewiring gene, with rewiring genes classified by the type of rewiring in which they participate. E) Betweenness centrality of proteins in the human interactome (CCSB, 2012), as classified by their type of rewiring. F) Range of $\log_2$ RNA-seq read counts from 16 human tissues (Illumina, 2011) of interactors rewired by the different rewiring types, and non-rewired interactors. G) Same as last panel but with change-over rewired interactors removed from the other categories.

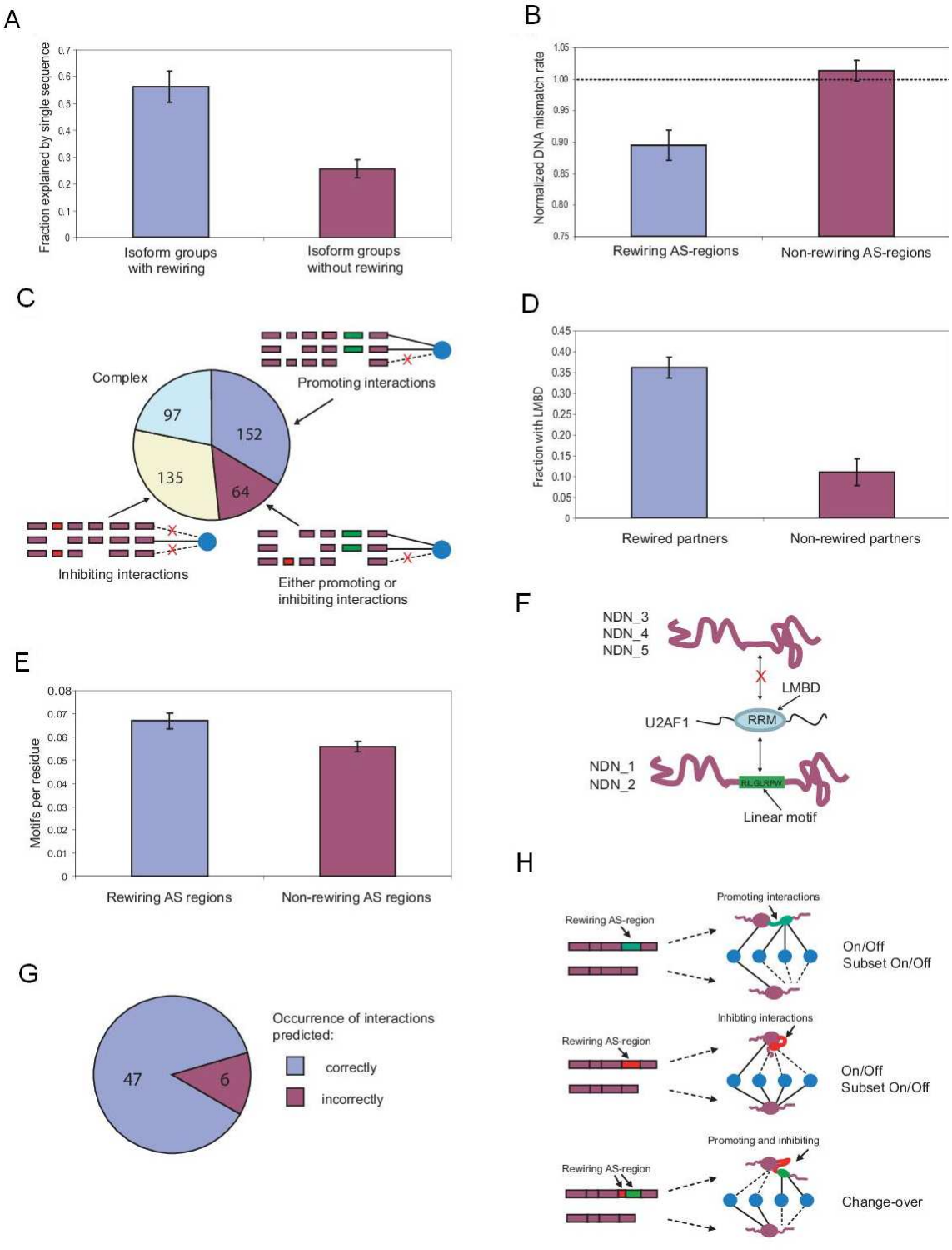
A
Fraction explained by single sequence
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Isoform groups with rewiring
Isoform groups without rewiring
B
Normalized DNA mismatch rate
1.05
1.00
0.95
0.90
0.85
0.80
0.75
Rewiring AS-regions
Non-rewiring AS-regions
C
Complex
97
152
135
64
Promoting interactions
Inhibiting interactions
Either promoting or inhibiting interactions
D
Fraction with LMBD
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0
Rewired partners
Non-rewired partners
E
Motifs per residue
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Rewiring AS regions
Non-rewiring AS regions
F
NDN_3
NDN_4
NDN_5
LMBD
U2AF1
RRM
NDN_1
NDN_2
RILGLRPW
Linear motif
G
47
6
Occurrence of interactions predicted:
correctly
incorrectly
H
Rewiring AS-region
Promoting interactions
On/Off
Subset On/Off
Rewiring AS-region
Inhibting interactions
On/Off
Subset On/Off
Rewiring AS-region
Promoting and inhibiting
Change-over

Figure 3.4: Rewiring-associated AS regions and molecular mechanisms. A) Interacting and non-interacting isoform groups as defined by a rewiring interaction are more likely to be distinguished by a localized sequence than other combinations of isoforms. B) AS regions associated with PPI rewiring have a significantly lower DNA mismatch rate between human and mouse than the coding region average for the same gene (dashed line at 1) or other AS regions in the same or different genes. C) Rewiring events classified according to the type of regions associated with the rewiring, either promoting, blocking, both (either promoting or blocking) or neither ("complex"). D) For interactors with at least one assigned Pfam (Bateman et al., 2004) domain (e-value<0.01), the fraction of rewired and non-rewired interactors which contain a linear motif binding domain (LMBD) from the ELM database (Dinkel et al., 2012) or from DILIMOT (Neduva & Russell, 2006). E) ELM database linear motif matches (≥7 residues long) per residue for AS regions associated with or not associated with rewiring. F) As an example of rewiring driven by AS of a linear motif, only the protein isoforms of the NDN gene which contain the ELM database RRM_1 binding motif interact with the RRM_1-containing U2AF1 protein. G) In cases where a rewiring-associated AS region disrupts a Pfam (Bateman et al., 2004) domain (e-value<0.01) which is known to bind to a domain in a rewired target protein according to iPfam (Finn et al., 2005), 3DID (Stein et al., 2009) or Domine (Raghavachari et al., 2008), the interacting isoform(s) contain the intact version of the domain (blue) significantly more often than cases where the non-interacting isoform(s) contain the intact domain (purple). H) Potential PPI rewiring

mechanisms through AS of localized sequence modules. Blue regions: AS regions causing interaction rewiring.

Conclusions

In this work, we have shown how the integration of network, functional, expressional and evolutionary data can reveal organizing principles of proteomes and interactomes. We have found that the structure of biological networks is shaped by evolutionary and tissue-specific adaptations of proteomes and that these proteome dynamics are in turn influenced by network structure. An integrative systems-level view, considering networks and proteome dynamics in conjunction, is therefore essential to fully understanding either biological networks or the dynamics of proteomes.

In Chapter 1, we have shown that transcription factor protein sequence evolution is an important component of gene expression variation between species and that selective constraints acting on transcription factors are best understood in the context of their position in the regulatory network, as they evolve at a rate proportional to that of their activated target genes. While such patterns of modular evolution have been observed in other types of biological networks (Wang & Lercher, 2011), this result contrasts sharply with trends observed for generic proteins, for which the dominant determinants of evolutionary rate are protein abundance and their position in the protein-protein interaction network. Having further shown that adaptive evolution specifically targets high-level transcription factors, those which regulate other transcription factors, this work has uncovered some of the global strategies of *trans*-regulatory network evolution.

Further exploring the relation between biological networks and proteome evolutionary dynamics, Chapter 2 showed that lost and gained genes across yeasts have more regulators, fewer epistatic interactions, and involved proportionally more transcription factors than universally conserved “core” genes. This shows that the lost and gained genes have functions which are more peripheral and more complex regulation than core genes. Furthermore, by considering a time-resolved view of gene loss, we established that network marginalization of genes through network rewiring tends to precede gene loss. This demonstrates that networks rewire over time and that these changes influence the selection against gene loss acting on individual genes.

The discoveries presented in Chapters 1 and 2 open the door to intriguing new questions. One such question is whether these results hold in other organisms, such as prokaryotes or higher eukaryotes. Yeast is used as a system to study eukaryotic biology primarily because of its relative simplicity as compared to multi-cellular eukaryotes. However, the complexity inherent to the biology of higher eukaryotes needs to be understood as well as the more basic processes in order to understand many diseases processes or complex phenotypic traits. Alternative splicing, the network-level consequences of which were explored in the last chapter, is only one of the many layers of complexity associated with human biology. Transcriptional regulation is also significantly more complex in humans as compared to yeast. In Chapters 1 and 2, we considered a comprehensive list of yeast TFs, comprising 174 proteins, and found only 16 TFs to be universally conserved across yeasts. In contrast, humans have at least 1,400 TFs (Vaquerizas et al., 2009) and

mammals have highly conserved tissue-specific expression programs (Merkin et al., 2012), suggesting the existence of a relatively large set of core TFs universally shared across mammals. Porting the analyses conducted in yeast to the study of mammalian evolution could help us understand the organization of the human regulatory network. Organizing principles such as evolutionary modularity within the transcriptional network or functional distinctions between core and species-specific components could help provide a framework for tackling the immense complexity of these systems. We have shown in yeast that *trans*-regulatory network evolution, through mutations in transcription factor protein sequences as well as through the loss and gain of transcription factors, plays a key role in species-specific evolutionary adaptation. Evidence exists of a similar evolutionary strategy used during human evolution. Transcription factors have been shown to be enriched among positively selected genes along the human-chimp lineage (Clark et al., 2003) as well as in genes with large expression changes (Gilad et al., 2006), suggesting that *trans*-regulatory network evolution constitutes a key component of human-specific evolutionary adaptation. A global network perspective on this *trans*-regulatory network evolution could allow us to observe broader patterns and adaptive strategies during human evolution, such as the potential targeting of high-level regulators.

Inter-species differences, by revealing the selective forces acting on individual elements such as nucleotides or proteins, can teach us about the function and organization of biological systems and components. Similarly, intra-species variation can reveal newly gained or lost selective forces on cellular components. Intra-species genetic variation

tends to reflect inter-species variation (Castle, 2011). Therefore, principles which guide evolutionary processes may also guide population-level variation. The findings presented in Chapters 1 and 2 may thus help to understand population-level variation both at the DNA level as well as at the systems or phenotypic level. For example, having shown the importance of *trans*-regulatory network evolution on gene expression phenotypes as well as its relatively fast evolution through gene loss and gain as compared to other functions, it is possible that *trans*-regulatory network variation may explain much of the phenotypic variation between individuals as well as genetic predispositions to disease.

The fact that humans have highly similar protein sequences to their closest cousins while displaying striking phenotypic and behavioral differences has long puzzled evolutionary biologists. Gene gain and loss may play a key role in human evolution and thus help to explain the uniqueness of the human species. Gene gain and loss have been shown to be accelerated along the primate lineage (Hahn et al., 2007), potentially accounting for much of the observed phenotypic variation. Since gene gain and loss are highly related to network structure, as shown in Chapter 2, a network-level analysis of gene gain and loss in primates may help to better understand the selective forces driving these processes and provide novel insights into human-specific evolutionary adaptation.

Proteomes are dynamic at many different levels. As shown in Chapters 1 and 2, their content, sequences and expression evolves significantly over millions of years. Another dimension of proteome dynamics, which is virtually absent in yeast, is alternative

splicing, which in multi-cellular eukaryotes, allows for different versions of the proteome to be expressed in different tissues, conditions and developmental stages. Similarly to evolutionary adaptations of proteomes, as we have shown in Chater 3, these dynamics can also be studied in the context of networks in order to better understand their functions at the local and system-wide level. It was shown that a splicing-sensitive human interactome is essential to capturing the full extent of the network's organizational coherence. These results create an urgent need for the mapping of new genome-wide isoform-resolved interactomes from which we will be able to better understand individual protein functions and roles in disease processes. Given the important influence of alternative splicing on the protein-protein interaction network, we must also asses its role in other realms, such as nucleotide binding or the regulation of other protein functions. This work has established the power of high-throughput functional assays for understanding the system-level influence of alternative splicing. The role of alternative splicing on other networks and functions may similarly be revealed through the application of high-throughput assays, such as nucleotide binding assays.

Together, this work solidifies the notion that biological networks and data integration constitute powerful tools for the study of biology and evolution. We have shown the importance of systematically exploiting the knowledge gained from networks to better understand the dynamics of proteomes through evolution and across tissues, and thus exposed core organizational principles underlying biological and evolutionary processes.

References

Altschul, S (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402.

Ashburner, M, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, AP Davis, K Dolinski, SS Dwight, JT Eppig, MA Harris, DP Hill, L Issel-Tarver, A Kasarskis, S Lewis, JC Matese, JE Richardson, M Ringwald, GM Rubin, and G Sherlock (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25: 25-29.

Barbosa-Morais, NL, M Irimia, Q Pan, HY Xiong, S Gueroussov, LJ Lee, V Slobodeniuc, C Kutter, S Watt, R Colak, T Kim, CM Misquitta-Ali, MD Wilson, PM Kim, DT Odom, BJ Frey, and BJ Blencowe (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587-93.

Bateman, A, L Coin, R Durbin, RD Finn, V Hollich, S Griffiths-Jones, A Khanna, M Marshall, S Moxon, ELL Sonnhammer, DJ Studholme, C Yeats, and SR Eddy (2004) The Pfam protein families database. *Nucleic Acids Research* 32: D138-41.

Blanchette, M, WJ Kent, C Riemer, L Elnitski, AFA Smit, KM Roskin, R Baertsch, K Rosenbloom, H Clawson, ED Green, D Haussler, and W Miller (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14: 708-15.

Bland, CS, ET Wang, A Vu, MP David, JC Castle, JM Johnson, CB Burge, and TA Cooper (2010) Global regulation of alternative splicing during myogenic differentiation. *Nucleic Acids Research* 38: 7651-7664.

Blomme, T, K Vandepoele, S De Bodt, C Simillion, S Maere, and Y Van de Peer (2006) The gain and loss of genes during 600 million years of vertebrate evolution. *Genome biology* 7: R43.

Boorsma, A, X-J Lu, A Zakrzewska, FM Klis, and HJ Bussemaker (2008) Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS ONE*. Public Library of Science 3: e3112.

Borneman, AR, TA Gianoulis, ZD Zhang, H Yu, J Rozowsky, MR Seringhaus, LY Wang, M Gerstein, and M Snyder (2007) Divergence of transcription factor binding sites across related yeast species. *Science* 317: 815-819.

Borneman, AR, JA Leigh-Bell, H Yu, P Bertone, M Gerstein, and M Snyder (2006) Target hub proteins serve as master regulators of development in yeast. *Genes & Development* 20: 435-448.

Braun, P, M Tasan, M Dreze, M Barrios-Rodiles, I Lemmens, H Yu, JM Sahalie, RR Murray, L Roncari, AS de Smet, K Venkatesan, JF Rual, J Vandenhaute, ME Cusick, T Pawson, DE Hill, J Tavernier, JL Wrana, FP Roth, and M Vidal (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nature methods* 6: 91-97.

Buljan, M, G Chalancon, S Eustermann, GP Wagner, M Fuxreiter, A Bateman, and MM Babu (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Molecular Cell* 46: 871-883.

Bullard, JH, E Purdom, KD Hansen, and S Dudoit (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11: 94.

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

Carlson, MR, B Zhang, Z Fang, PS Mischel, S Horvath, and SF Nelson (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* 7: 40.

Carvunis, A-R, T Rolland, I Wapinski, MA Calderwood, MA Yildirim, N Simonis, B Charloteaux, CA Hidalgo, J Barbette, B Santhanam, GA Brar, JS Weissman, A Regev, N Thierry-Mieg, ME Cusick, and M Vidal (2012) Proto-genes and de novo gene birth. *Nature* 487: 370-4.

Caspi, R, H Foerster, CA Fulcher, P Kaipa, M Krummenacker, M Latendresse, S Paley, SY Rhee, AG Shearer, C Tissier, TC Walk, P Zhang, and PD Karp (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research* 36: D623-31.

Castle, JC (2011) SNPs occur in regions with less genomic sequence conservation. *PloS one*. Public Library of Science 6: e20660.

Castle, JC, C Zhang, JK Shah, AV Kulkarni, A Kalsotra, TA Cooper, and JM Johnson (2008) Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics* 40: 1416-1425.

Clark, AG, S Glanowski, R Nielsen, PD Thomas, A Kejariwal, MA Todd, DM Tanenbaum, D Civello, F Lu, B Murphy, S Ferriera, G Wang, X Zheng, TJ White, JJ Sninsky, MD Adams, and M Cargill (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960-1963.

Costanzo, M, A Baryshnikova, J Bellay, Y Kim, ED Spear, CS Sevier, H Ding, JL Koh, K Toufighi, S Mostafavi, J Prinz, RP St Onge, B VanderSluis, T Makhnevych, FJ Vizeacoumar, S Alizadeh, S Bahr, RL Brost, Y Chen, M Cokol, et al. (2010) The genetic landscape of a cell. *Science* 327: 425-431.

Coulombe-Huntington, J, and Y Xia (2012) Regulatory network structure as a dominant determinant of transcription factor evolutionary rate. *PLoS Computational Biology* 8: e1002734.

Dinkel, H, S Michael, RJ Weatheritt, NE Davey, K Van Roey, B Altenberg, G Toedt, B Uyar, M Seiler, A Budd, L Jodicke, MA Dammert, C Schroeter, M Hammer, T Schmidt, P Jehl, C McGuigan, M Dymecka, C Chica, K Luck, et al. (2012) ELM--the database of eukaryotic linear motifs. *Nucleic Acids Research* 40: D242-51.

Drummond, DA, JD Bloom, C Adami, CO Wilke, and FH Arnold (2005) Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14338-14343.

Drummond, DA, A Raval, and CO Wilke (2006) A single determinant dominates the rate of yeast protein evolution. *Molecular Biology and Evolution* 23: 327-337.

Ellis, JD, M Barrios-Rodiles, R Colak, M Irimia, T Kim, JA Calarco, X Wang, Q Pan, D O'Hanlon, PM Kim, JL Wrana, and BJ Blencowe (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular Cell* 46: 884-892.

Finn, RD, M Marshall, and A Bateman (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* 21: 410-2.

Fitzpatrick, DA, ME Logue, JE Stajich, and G Butler (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology* 6: 99.

Fraser, HB, AE Hirsh, LM Steinmetz, C Scharfe, and MW Feldman (2002) Evolutionary rate in the protein interaction network. *Science* 296: 750-752.

Gabut, M, P Samavarchi-Tehrani, X Wang, V Slobodeniuc, D O'Hanlon, HK Sung, M Alvarez, S Talukder, Q Pan, EO Mazzoni, S Nedelec, H Wichterle, K Woltjen, TR Hughes, PW Zandstra, A Nagy, JL Wrana, and BJ Blencowe (2011a) An alternative

splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* 147: 132-146.

Gabut, M, P Samavarchi-Tehrani, X Wang, V Slobodeniuc, D O'Hanlon, HK Sung, M Alvarez, S Talukder, Q Pan, EO Mazzoni, S Nedelec, H Wichterle, K Woltjen, TR Hughes, PW Zandstra, A Nagy, JL Wrana, and BJ Blencowe (2011b) An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming. *Cell* 147: 132-146.

Galant, R, and SB Carroll (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* 415: 910-913.

Gilad, Y, A Oshlack, GK Smyth, TP Speed, and KP White (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. *Nature* 440: 242-245.

Hahn, MW, JP Demuth, and S-G Han (2007) Accelerated rate of gene gain and loss in primates. *Genetics* 177: 1941-9.

Hamosh, A, AF Scott, JS Amberger, CA Bocchini, and VA McKusick (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* 33: D514-7.

Harbison, CT, DB Gordon, TI Lee, NJ Rinaldi, KD Macisaac, TW Danford, NM Hannett, JB Tagne, DB Reynolds, J Yoo, EG Jennings, J Zeitlinger, DK Pokholok, M Kellis, PA Rolfe, KT Takusagawa, ES Lander, DK Gifford, E Fraenkel, and RA Young (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431: 99-104.

Harris, JK, ST Kelley, GB Spiegelman, and NR Pace (2003) The genetic core of the universal ancestor. *Genome Research* 13: 407-12.

Harris, MA, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, GM Rubin, JA Blake, C Bult, M Dolan, H Drabkin, JT Eppig, DP Hill, L Ni, M Ringwald, et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* 32: D258-61.

Harrow, J, A Frankish, JM Gonzalez, E Tapanari, M Diekhans, F Kokocinski, BL Aken, D Barrell, A Zadissa, S Searle, I Barnes, A Bignell, V Boychenko, T Hunt, M Kay, G Mukherjee, J Rajan, G Despacio-Reyes, G Saunders, C Steward, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* 22: 1760-1774.

Hershberg, R, and H Margalit (2006) Co-evolution of transcription factors and their targets depends on mode of regulation. *Genome Biology* 7: R62.

Horak, CE, NM Luscombe, J Qian, P Bertone, S Piccirrillo, M Gerstein, and M Snyder (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes & Development* 16: 3017-3033.

Hu, Z, PJ Killion, and VR Iyer (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nature Genetics* 39: 683-687.

Hughes, AL, and R Friedman (2005) Gene duplication and the properties of biological networks. *Journal of molecular evolution* 61: 758-64.

Huot, ME, G Vogel, A Zabarauskas, CT Ngo, J Coulombe-Huntington, J Majewski, and S Richard (2012) The Sam68 STAR RNA-binding protein regulates mTOR alternative splicing during adipogenesis. *Molecular Cell* 46: 187-199.

Ihmels, J, R Levy, and N Barkai (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nature Biotechnology* 22: 86-92.

Illumina (2011) Human Body Map 2.0 Project. (GEO accession: GSE30611).

Ingolia, NT, S Ghaemmaghami, JR Newman, and JS Weissman (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324: 218-223.

Jeong, H, B Tombor, R Albert, ZN Oltvai, and AL Barabasi (2000) The large-scale organization of metabolic networks. *Nature* 407: 651-654.

Jiang, H, L Xu, and Z Gu (2011) Growth of novel epistatic interactions by gene duplication. *Genome biology and evolution* 3: 295-301.

Jovelin, R, and PC Phillips (2009) Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biology* 10: R35.

Kalsotra, A, and TA Cooper (2011) Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics* 12: 715-729.

Kellis, M, BW Birren, and ES Lander (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428: 617-24.

Kim, PM, LJ Lu, Y Xia, and MB Gerstein (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science* 314: 1938-1941.

Krylov, DM, YI Wolf, IB Rogozin, and EV Koonin (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research* 13: 2229-35.

Langmead, B, and SL Salzberg (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 9: 357-9.

Lee, S, I Kohane, and S Kasif (2005) Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics* 6: 168.

Lee, TI, NJ Rinaldi, F Robert, DT Odom, Z Bar-Joseph, GK Gerber, NM Hannett, CT Harbison, CM Thompson, I Simon, J Zeitlinger, EG Jennings, HL Murray, DB Gordon, B Ren, JJ Wyrick, JB Tagne, TL Volkert, E Fraenkel, DK Gifford, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* 298: 799-804.

Lefébure, T, and MJ Stanhope (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biology* 8: R71.

Lercher, MJ, and C Pál (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Molecular Biology and Evolution* 25: 559-67.

Li, S, CM Armstrong, N Bertin, H Ge, S Milstein, M Boxem, PO Vidalain, JD Han, A Chesneau, T Hao, DS Goldberg, N Li, M Martinez, JF Rual, P Lamesch, L Xu, M Tewari, SL Wong, LV Zhang, GF Berriz, et al. (2004) A map of the interactome network of the metazoan C. elegans. *Science* 303: 540-543.

Merkin, J, C Russell, P Chen, and CB Burge (2012) Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* 338: 1593-9.

Modrek, B, and C Lee (2002) A genomic view of alternative splicing. *Nature Genetics* 30: 13-19.

Nash, R, S Weng, B Hitz, R Balakrishnan, KR Christie, MC Costanzo, SS Dwight, SR Engel, DG Fisk, JE Hirschman, EL Hong, MS Livstone, R Oughtred, J Park, M Skrzypek, CL Theesfeld, G Binkley, Q Dong, C Lane, S Miyasato, et al. (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Research* 35: D468-71.

Neduva, V, and RB Russell (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Research* 34: W350-5.

Ohno, S (1970) *Evolution by gene duplication.* Springer-Verlag, New-York.

Pan, Q, O Shai, LJ Lee, BJ Frey, and BJ Blencowe (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40: 1413-1415.

Pournara, I, and L Wernisch (2007) Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics* 8: 61.

Ptitsyn, A, and LL Moroz (2012) Computational workflow for analysis of gain and loss of genes in distantly related genomes. *BMC Bioinformatics* 13 Suppl 1: S5.

Raghavachari, B, A Tasneem, TM Przytycka, and R Jothi (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Research* 36: D656-61.

Ronshaugen, M, N McGinnis, and W McGinnis (2002) Hox protein mutation and macroevolution of the insect body plan. *Nature* 415: 914-917.

Rual, JF, T Hirozane-Kishikawa, T Hao, N Bertin, S Li, A Dricot, N Li, J Rosenberg, P Lamesch, PO Vidalain, TR Clingingsmith, JL Hartley, D Esposito, D Cheo, T Moore, B Simmons, R Sequerra, S Bosak, L Doucette-Stamm, C Le Peuch, et al. (2004) Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Research* 14: 2128-2135.

Rual, JF, K Venkatesan, T Hao, T Hirozane-Kishikawa, A Dricot, N Li, GF Berriz, FD Gibbons, M Dreze, N Ayivi-Guedehoussou, N Klitgord, C Simon, M Boxem, S Milstein, J Rosenberg, DS Goldberg, LV Zhang, SL Wong, G Franklin, S Li, et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437: 1173-1178.

Safran, M, I Dalah, J Alexander, N Rosen, T Iny Stein, M Shmoish, N Nativ, I Bahir, T Doniger, H Krug, A Sirota-Madi, T Olender, Y Golan, G Stelzer, A Harel, and D Lancet (2010) GeneCards Version 3: the human gene integrator. *Database* 2010: baq020.

Shapiro, IM, AW Cheng, NC Flytzanis, M Balsamo, JS Condeelis, MH Oktay, CB Burge, and FB Gertler (2011) An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. *PLoS Genetics* 7: e1002218.

Sharp, PM, and WH Li (1987) The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research* 15: 1281-1295.

Shou, C, N Bhardwaj, HYK Lam, K-K Yan, PM Kim, M Snyder, and MB Gerstein (2011) Measuring the Evolutionary Rewiring of Biological Networks. *PLoS Computational Biology*. Public Library of Science 7: e1001050.

Stein, A, A Panjkovich, and P Aloy (2009) 3did Update: domain-domain and peptide-mediated interactions of known 3D structure. *Nucleic Acids Research* 37: D300-4.

Teixeira, MC, P Monteiro, P Jain, S Tenreiro, AR Fernandes, NP Mira, M Alenquer, AT Freitas, AL Oliveira, and I Sa-Correia (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Research* 34: D446-51.

Thakar, K, I Votteler, D Kelkar, T Shidore, S Gupta, S Kelm, and F Dietz (2012) Interaction of HRP-2 isoforms with HDGF: chromatin binding of a specific heteromer. *The FEBS Journal* 279: 737-751.

The Center for Cancer Systems Biology (CCSB) at the Dana-Farber Cancer Institute (2012) The Human Interactome Project 2012 (prepublication).

Thierry-Mieg, D, and J Thierry-Mieg (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biology* 7 Suppl 1: S12 1-14.

Tirosh, I, S Reikhav, AA Levy, and N Barkai (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324: 659-662.

Vaquerizas, JM, SK Kummerfeld, SA Teichmann, and NM Luscombe (2009) A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*. Nature Publishing Group 10: 252-63.

Vitkup, D, P Kharchenko, and A Wagner (2006) Influence of metabolic network structure and function on enzyme evolution. *Genome Biology* 7: R39.

Wang, ET, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470-476.

Wang, GZ, and MJ Lercher (2011) The effects of network neighbours on protein evolution. *PLoS ONE* 6: e18288.

Wang, Y, EA Franzosa, XS Zhang, and Y Xia (2010) Protein evolution in yeast transcription factor subnetworks. *Nucleic Acids Research* 38: 5959-5969.

Wapinski, I, A Pfeffer, N Friedman, and A Regev (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. Oxford Univ Press 23:

Wethkamp, N, H Hanenberg, S Funke, CV Suschek, W Wetzel, S Heikaus, E Grinstein, U Ramp, R Engers, HE Gabbert, and C Mahotka (2011) Daxx-beta and Daxx-gamma,

two novel splice variants of the transcriptional co-repressor Daxx. *Journal of Biological Chemistry* 286: 19576-19588.

Winzeler, EA (1999) Functional Characterization of the S. cerevisiae Genome by Gene Deletion and Parallel Analysis. *Science* 285: 901-906.

Workman, CT, HC Mak, S McCuine, JB Tagne, M Agarwal, O Ozier, TJ Begley, LD Samson, and T Ideker (2006) A systems approach to mapping DNA damage response pathways. *Science* 312: 1054-1059.

Xia, Y, EA Franzosa, and MB Gerstein (2009) Integrated assessment of genomic correlates of protein evolutionary rate. *PLoS Computational Biology* 5: e1000413.

Yang, X, JS Boehm, K Salehi-Ashtiani, T Hao, Y Shen, R Lubonja, SR Thomas, O Alkan, T Bhimdi, TM Green, CM Johannessen, SJ Silver, C Nguyen, RR Murray, H Hieronymus, D Balcha, C Fan, C Lin, L Ghamsari, M Vidal, et al. (2011) A public genome-scale lentiviral expression library of human ORFs. *Nature methods* 8: 659-661.

Yang, Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591.

Yang, Z, and R Nielsen (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* 17: 32-43.

Yu, H, P Braun, MA Yildirim, I Lemmens, K Venkatesan, J Sahalie, T Hirozane-Kishikawa, F Gebreab, N Li, N Simonis, T Hao, JF Rual, A Dricot, A Vazquez, RR Murray, C Simon, L Tardivo, S Tam, N Svrzikapa, C Fan, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science* 322: 104-110.

Zaslaver, A, AE Mayo, R Rosenberg, P Bashkin, H Sberro, M Tsalyuk, MG Surette, and U Alon (2004) Just-in-time transcription program in metabolic pathways. *Nature Genetics* 36: 486-91.

Curriculum Vitae