Molecular Evolution of Posttranslational Regulation in Intrinsically Disordered Regions

by

Nghiem (Alex) Nguyen Ba

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Cell and Systems Biology
University of Toronto

Abstract

Molecular Evolution of Posttranslational Regulation in Intrinsically Disordered Regions

Nghiem (Alex) Nguyen Ba

Doctor of Philosophy

Graduate Department of Cell and Systems Biology
University of Toronto

2014

Protein posttranslational regulation is a major facet of protein function, and efforts have been made to systematically characterize the level of control of proteins. For example, systematic determination of protein localization, phosphorylation, and interactions have allowed the examination of the regulatory network of the cell. However, the study of the evolution of this underlying regulatory network requires a higher resolution analysis of the sequences that are responsible for this level of control.

The overarching goal of this thesis was to examine the role of the evolution of protein regulatory sequences as a molecular mechanism driving functional diversity. I developed computational tools and methods for the identification and characterization of these regulatory sequences, as well as experimental approaches to study the evolutionary impact of changes within these sequences.

I first characterized the evolution of phosphorylation sites and used the property that they are strongly conserved relative to their flanking disordered regions as a computational means to

systematically identify motifs in the budding yeast proteome. These results suggest that incorporating evolutionary conservation is sufficient for the prediction of around 30% of the known short linear motifs. Applying these computational approaches to the budding yeast proteome showed that thousands of short linear motifs exist and still remain uncharacterized.

Using a relative rates test, I showed that motifs frequently change selective constraints after gene duplication and showed that these changes can alter protein regulation over evolution. Finally, I designed a high-throughput experimental pipeline to systematically, quantitatively and precisely assess the fitness consequences of rewiring a regulatory network and applied it to test whether a bi-functional protein has sub-functionalized over evolution.

Acknowledgements

Pursuing a doctoral degree has been a thrilling experience and has opened a new world of excitement and challenges that could not be possible without the support and guidance of my advisor Dr. Alan Moses. His ability to convey enthusiasm in scientific pursuit is unparalleled and I am eternally grateful for the encouragements and inspiration throughout my studies. I feel very lucky to have been part of the 'lab start-up' which is now well underway to generating and validating fascinating hypotheses. I also thank my supervisory committee, Dr. Nicholas Provart and Dr. Philip Kim for their providing insights and directions for my projects.

All of the work presented here could not be possible without the stimulating environment provided by the lab and the Department of Cell and Systems Biology. Healthy discussions between present and past lab mates, especially Andy Lai, Louis-Francois Handfield, Annabelle Haudry, Amin Zia, have all been instrumental to my progression in the doctoral program. The Andrews lab at the Donnelly Centre for Cellular Biomolecular Research has provided me with the necessary training and has been a primary source of collaboration. I have to extend my deepest thanks to Dewald van Dyk for taking me through the loops of wet-lab experiments. I also thank all other collaborators of work presented here and elsewhere. My research could also not be possible without the graduate administrator, Ian Buglass, who has provided me with all help required in financial and administrative issues. I would like to thank the TorBUG team, Michelle Brazas in particular, for giving me the opportunity of being both a founding member and part of the committee, which allowed me to center myself on the bioinformatics community of Toronto.

Finally, I would not be in graduate school without the support of my family, mother and sister. My mother has been the one to convince me to approach Dr. Alan Moses and has been incredibly supportive despite all the sacrifices that she had to do to provide me with a life that allowed me to pursue this path. I owe many of my successes to her. Yves, as a great husband to my mother, has also been very supportive and has made me feel less guilty of having made my own life in Toronto. Of course, I would not be the same person without Anastassia, for all the love and time devoted.

Table of contents

List of Tables

List of Figures

List of Abbreviations

APC	Anaphase-promoting complex
CASP	Critical assessment of protein structure prediction
CDK	Cyclin-dependent kinase
cDNA	coding DNA
CFTR	Cystic fibrosis transmembrane conductance regulator
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CKII	Casein-kinase 2
DDC	Duplication-degeneration-complementation
DTT	Dithiothreitol
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
GFP	Green-fluorescent protein
GLEBS	Gle2-binding-sequence
GST	Glutathione S-transferase
HMM	Hidden Markov Model
KL	Kullback-Leibler
LATS	Large tumor suppressor
LR	Likelihood-ratio
LRT	Likelihood-ratio test statistic

MBP	Maltose binding protein
NDR	Nuclear dumbbell forming 2-related
NLS	Nuclear localization signal
NPC	Nuclear pore complex
ORF	Open reading frame
PCR	Polymerase chain reaction
PKA	Protein kinase A
PPV	Positive predictive value
PSSM	Position-specific scoring matrix
PTM	Posttranslational modification
SCF	Skp, Cullin, F-box containing complex
SCP	Single-copy protein
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SGA	Synthetic genetic array
SGD	Saccharomyces genome database
SH2	Sarcoma homology 2
SH3	Sarcoma homology 3
SNP	Single-nucleotide polymorphism
TAP	Tandem affinity purification tag
TPR	Tetratricopeptide

UTR	Untranslated region
WGD	Whole-genome duplication

List of Appendices

Chapter I

General Introduction

I.1 Abstract

Cellular and physiological responses within a single organism are intricately controlled by their underlying genetic regulatory network. Posttranslational control of proteins is mediated by short linear motifs, which are short peptides sequences within disordered regions of proteins. These sequences are recognized by regulator proteins that can alter protein subcellular localization, degradation rates, and even amino acid properties, such as posttranslational modifications. Many proteins contain several short linear motifs and interplay between them can create complex regulatory networks. Over the course of evolution, motif turnover can play a role in rewiring the regulatory network of the cell, leading to different phenotypes that may form the basis of the organismal diversity observed throughout the tree of life. This evolution is facilitated by the position of short linear motifs within fast evolving disordered regions that lack structural constraints.

I.2 Introduction

The central dogma in biology states that the genetic information encodes the proteins that are the major functional components of the cell. This idea has led to exciting research about the nature of the regulation of the expression of this genetic information and to functional characterization of thousands of identified proteins in the kingdoms of life. Despite this important discovery of information flow, there are still several outstanding questions about the nature and the mechanisms of gene regulation, especially regarding the role of regulatory evolution in the observed organismal diversity.

The field of comparative genomics showed remarkable conservation of proteins across the tree of life: up to 31% of the proteins found in the budding yeast *Saccharomyces cerevisiae* have functional homologs in mammals (Botstein et al. 1997). Similarly, chimpanzees and humans show a high degree of genetic similarities (Chimpanzee Sequencing and Analysis Consortium 2005). The challenge of comparative genomics is therefore to decipher how small amount of genetic changes can lead to vastly different and distinct phenotypes. The basis of my research follows the hypothesis that subtle changes in protein enzymatic functions, such as speed of substrate turnover, are unlikely to account for the majority of the genetic basis for the differences in phenotype between individuals and between species. Rather, mutations affecting the regulation of the proteins may be responsible for most of the biological differences between species (Figure I-1; (King and Wilson 1975)).

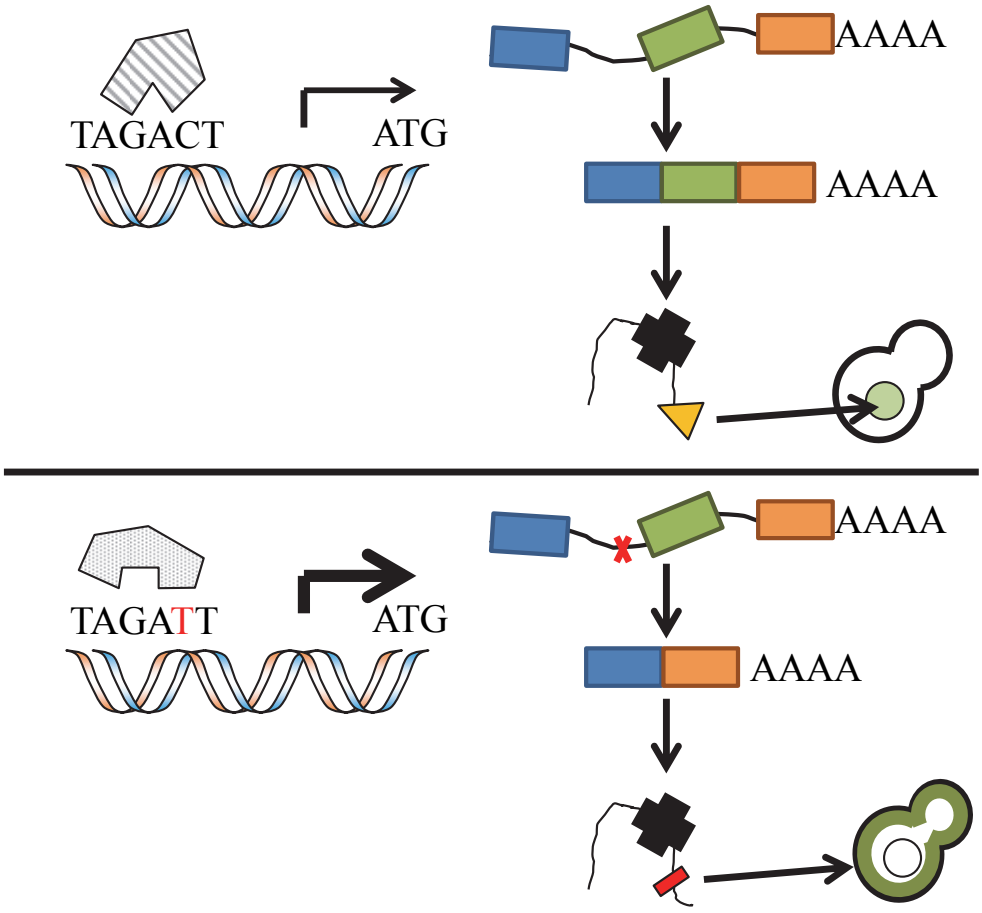


Figure I-1. Evolution by mutations in regulatory sequences. Bottom panel represents several mutations that affect regulation rather than enzymatic functions: transcriptional level (changes in transcription factor binding sites), mRNA splicing level (changes in exons) and changes in protein regulatory sequences. The final outcome can lead to different protein abundance and posttranslational regulation (such as protein localization as shown in the example).

Fueled by technological advances in microarray and ChIP-seq technology, comparative genomics on the differences in gene expression have now shown that transcriptional changes are one of the major contributors of functional differences between species (reviewed in (Wray et al. 2003; Villar et al. 2014)) and it is likely that regulatory divergence affects all facets of protein control. For instance, evolutionary changes in mRNA splicing control can create vastly different arrays and quantities of isoforms across species (see (Barbosa-Morais et al. 2012) for a genome-wide analysis of splicing variants in vertebrates and see (Marshall

et al. 2013) for a well-characterized example of splicing evolution in yeast). Finally, changes in protein regulation once the proteins have been translated could also be contributing to the functional differences of proteins between species.

In the following sections, I will discuss the molecular mechanisms by which proteins can be regulated by posttranslational regulation. Specifically, I will introduce the computational challenges to represent and predict the protein sequences that mediate the posttranslational regulation of the protein. I will then focus this discussion on four molecular mechanisms that cells use to turn proteins on and off as well as related experimental effort to characterize the specific sequences that are responsible for this control. Finally, I will focus on the importance of these regulatory connections in the field of comparative genomics.

I.3 Short linear motifs

I.3.1 Molecular definition of short linear motifs

Protein posttranslational regulation and modifications frequently occur via short linear motifs, which are short 2 to 15 amino acid sequences that are recognized by the modifying enzymes or the binding partners (Gould et al. 2010). These sequences have been labeled as linear as a definition to indicate that the three dimensional nature of the motif is undefined rather than non-existent. Other biological definitions of patterns or motifs that explicitly take into account the three dimensional arrangements of the proteins exist (Doxey et al. 2006; Doxey et al. 2010), however are not discussed further in this thesis.

In contrast with interactions over large surface areas, such as interactions within the nuclear pore complex (Alber et al. 2007), protein interactions through short linear motifs are very dynamic and localized processes (Yaffe et al. 2001). Recognition of short linear motifs by other proteins usually occurs through specialized protein domains, which are typically large, folded and modular structures, such as in the case of SH2 domains (Pawson et al. 2001). A high-resolution molecular view of the short linear motif during these interactions indicates that even short peptide fragments can provide strong specificity of recognition by these domains (Waksman et al. 1992). Although the affinity of the interactions through short linear motifs is typically weaker (nanomolar to low micromolar range (Dinkel et al. 2012)) than

interactions occurring between two large domains (low nanomolar range), this weaker affinity has been thought to promote transient and dynamic interactions within signaling networks (reviewed in (Li 2005) and (Diella et al. 2008)). Therefore, there is an opportunity for multiple interactions to occur within short distances and for these interactions to be highly modular and regulated. This mode of interaction is likely to be the case for hub or scaffold proteins (Cortese et al. 2008). These scaffold proteins can bring multiple different proteins together through multiple short linear motifs that are recognized by specific domains on the cognate proteins.

For these interactions to be possible, a requirement in the flexibility of the local region of the short linear motif must be met. This flexibility in protein backbone is now recognized as an intrinsic state of protein backbone dynamics and is termed 'intrinsically disordered' (discussed below).

I.3.2 Computational representations of short linear motifs

Short linear motifs have often been described as amino acid patterns or consensus sequences in the form of regular expressions (Gould et al. 2010) and several tools can be used to scan protein sequences for the presence of matches (e.g. Scansite (Obenauer et al. 2003)). A major issue with regular expressions is the all-or-nothing character of the representations. This difficulty of representing short linear motifs using patterns is apparent when not all true instances of the short linear motifs match the regular expression (Dinkel et al. 2012). Relaxing the constraints on the expression usually leads to increased spurious matches and fine-tuning these expressions is necessary for appropriate biological conclusions (this procedure is performed manually by two of the most commonly used databases of biological regular expressions: PROSITE (Hulo et al. 2006) and ELM (Dinkel et al. 2012), but users are recommended to also perform this task). Other more popular representations are sequence logos from an underlying PSSM (position-specific scoring matrix), which are more probabilistic interpretations of the characters present at each positions of the short linear motif (Figure I-2; (Schneider and Stephens 1990)). More ambitious representations have also been used (such as anti-motifs (Alexander et al. 2011)) to attempt to model the more complex nature of the recognition of the short linear motif by other proteins. However, not

all short linear motifs obey simple patterns and hidden Markov models have been shown to perform well in some cases (e.g. nuclear localization signals (Nguyen Ba et al. 2009)).

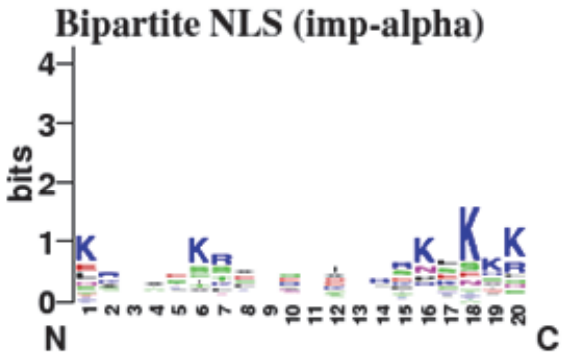


Figure I-2. Sequence logo of the bipartite nuclear localization signal. Characterized bipartite nuclear localization signals recognized by importin alpha were aligned and are represented by their information content using sequence logos.

An important property of short linear motifs is that, because they are short, most matches in the proteome are unlikely to be biologically relevant. Another property of short linear motifs is that they tend to be degenerate (positions with poor information content, see Figure I-2) and therefore, an important computational challenge is the classification of sequence matches of previously characterized short linear motifs.

I.3.2 Computational predictions of short linear motifs

Systematic computational predictions of short linear motifs have been generally divided into two main areas: 1) the classification of instance matches, and 2) the *de novo* prediction of short linear motif consensus sequences.

Classification of instance matches use contextual information and machine learning approaches to differentiate characterized instances to random instances (Blom et al. 1999; Linding et al. 2008). These approaches take advantage of several sequence features or experimental data, such as disorderedness (see Chapter I.8), surface accessibility (Via et al.

2009) and evolutionary conservation (Budovskaya et al. 2005; Chica et al. 2008). However these approaches require a training data set that may not be easily obtained.

De novo prediction of short linear motif consensus sequences uses experimental evidence to gather a list of proteins likely to share the same linear motif. This approach assumes that motifs on proteins have been gained through convergent evolution and that experimental approaches can adequately enrich for at least one short linear motif. Following this, k-mer enrichment (Rigoutsos and Floratos 1998) or other sophisticated statistical approaches (Bailey and Elkan 1994) can be used to predict consensus sequences. Other successful approaches combine features that were discovered to be useful in the classification of instance matches with enrichment of motifs (Neduva and Russell 2006).

I.4 Posttranslation regulation by short linear motifs

I.4.1 Protein-protein interaction

All biologically functional short linear motifs by definition must interact with another protein (or another region of the protein containing them) or to a molecule (such as water). Therefore, precise molecular roles for short linear motifs are usually described by the function of the regulator or by the posttranslational modifications (PTMs) on specific amino acids. Nevertheless, some short linear motifs are strictly defined by their binding preferences to particular protein domains and are typically termed ‘binding motifs’.

Experimental characterization of protein-protein interactions in yeast have taken advantage of the yeast two-hybrid system (Uetz et al. 2000; Ito et al. 2001) to identify binary interactions. This system is able to detect more transient interactions than techniques aimed at identifying complexes, such as co-immunoprecipitation followed by mass spectrometry (Ho et al. 2002). In another study, we found that yeast two-hybrid was well suited to detect kinase-substrate interactions (see Chapter I.4.2 and (Sharifpoor et al. 2011)). Although some kinase-substrate interactions are facilitated by kinase-docking sites, which are short linear motifs that specifically bind to a kinase domain to enhance phosphorylation (Reményi et al. 2006), several substrates of kinases do not contain specific binding motifs, and detection of interactions using the yeast two-hybrid system can be due to the interaction during catalysis,

due to other cellular means (such as protein scaffolds (Cortese et al. 2008)), or due to interactions not mediated by short linear motifs. Nevertheless, because of the sensitivity of yeast two-hybrid in detecting transient interactions, proteins containing domains that specialize in recognizing short linear motifs have often been subjected to proteome-wide yeast two-hybrid screens (e.g. the SH3 domain (Tonikian et al. 2009)). In turn, these screens can allow the discovery of the binding specificity of these domains using computational approaches described in Chapter I.3.2.

Other more direct approaches in deciphering the binding specificities of proteins have taken advantage of high-throughput techniques such as phage display (Tonikian et al. 2009; Teyra et al. 2012). These *in vitro* techniques allow the production of consensus sequences that are recognized by peptide-binding domains. However, they do not provide a list of *in vivo* target that may be required in studying the cellular regulatory networks.

Protein regulation by binding through short linear motifs can facilitate the recognition of regulating proteins (many non-PTM short linear motifs are like this), or alternatively these interactions can be mediated by pseudosubstrates that can inactivate the regulatory protein upon binding (such as in Acm1 and Mad3 pseudosubstrate inactivation of the anaphase promoting complex (Burton and Solomon 2007; Choi et al. 2008)). These pseudosubstrates contain the short linear motifs that are usually recognized by the regulator but these pseudosubstrates contain other sequences that lock the regulator upon binding.

I.4.2 Protein phosphorylation

The discovery of protein kinases in the early 1950's (Burnett and Kennedy 1954; Fischer and Krebs 1955) led to several biochemical studies establishing the molecular mechanisms of phosphate addition to specific amino acids. The role of protein phosphorylation as an activating mechanism was thought to be related to protein folding leading to a final stable conformation. Although still a relevant today, protein phosphorylation as a central mechanism of protein regulation was established shortly after the discovery of protein phosphatases dedicated to removing phosphate groups from proteins (Graves et al. 1960). Therefore, proteins could be turned on or off by the cell after they had been produced. These kinases and phosphatases appeared to have specific targets and, analogous to transcription

factors and their molecular recognition of DNA, recognized specific amino acid peptides on proteins. Several studies have now systematically established *in vitro* the recognition sequence of several kinases (Mok et al. 2010) and the recognition sequences obtained from these studies show remarkable similarities to consensus sequences *in vivo*. Interestingly, several kinases show almost identical consensus sequences. How kinases of similar consensus sequences can have different targets is still an outstanding question, especially considering that signaling complexity through the eukaryotic tree of life partly consists of duplicated kinases (Manning et al. 2008). Several mechanisms to enhance kinase-substrate specificity have been uncovered and are often mediated by short linear motifs that promote binding, perform allosteric regulation of the kinase activity, or colocalizes the substrate with the kinase (reviewed in (Reményi et al. 2006; Ubersax and Ferrell 2007)).

The important processes controlled by phosphorylation propelled the rapidly growing field of phosphoproteomics. Technological advances allowing the rapid determination of kinase recognition site preferences, along with mass spectrometry, allowed the rapid detection of several thousands of phosphorylation sites in eukaryotic proteomes along with the putative kinases responsible for these modifications (Sadowski et al. 2013). A parallel angle to study cell signaling and phosphoproteomics is to systematically catalog all targets of kinases (Sharifpoor et al. 2011). Several assays have been proposed, including protein chips (Ptacek et al. 2005) and mass spectrometry using analog-sensitive kinases (Holt et al. 2009). An important question following the discovery of these thousands of identified sites is whether these sites are biologically relevant. There are currently no easy and reliable way to determine the *in vivo* function of these identified phosphorylation sites, however, efforts have been made to prioritize the interrogation of functional phosphorylation sites (Beltrao et al. 2012).

Notable examples of important kinases are the cyclin-dependent kinases (CDKs), which are proteins first discovered to be necessary for the progression of the cell cycle. Specific temporal activation of CDKs has been found to orchestrate the series of events leading to proper cell division. Interestingly, CDKs usually have more than a single cyclin that can activate them and these cyclins can induce remarkable substrate specificity (Kõivomägi, Valk, Venta, Iofik, Lepiku, Morgan, et al. 2011).

I.4.3 Protein subcellular localization

Protein localization was recognized to be important for their function with the advent of two important technological advances: centrifugation to separate cellular compartments, and microscopy to observe them. But how do the proteins 'know' where to travel within the cell once they have been made? In the 1970s, the signal hypothesis was formulated in the seminal work by Gunter Blobel (Blobel and Dobberstein 1975), which described the presence of genetically encoded localization signals within proteins that were recognized by the cellular machinery and transported to different compartments. Thus, proteins localizing to similar compartments all contained similar localization signals that could be recognized on their primary amino acid sequence. In a similar vein as protein phosphorylation as an activating mechanism, localization toward certain compartments was seen as a rapid way of activating the function of a particular protein (for example translocation of transcription factors involved in stress response (Hao et al. 2013)). Experimentally, protein localization has been one of the simplest regulatory mechanism to assay due to the discovery that green-fluorescent protein (GFP) fused to the protein of interest could be used to visualize the localization of a protein using fluorescence microscopy (Wang and Hazelrigg 1994; Tsien 1998). Indeed, systematic GFP-tagging of the yeast proteome was performed recently (Huh et al. 2003) allowing large-scale collection of yeast protein subcellular localization using high-throughput microscopy.

Among the best studied examples of protein localization as a regulatory mechanism is the nuclear localization signal (Görlich and Mattaj 1996). Active transport of proteins to the nucleus in eukaryotic cells requires the interaction of several import proteins (dubbed importins) with the signal peptide, which then transports the protein possessing the signal peptide to the nucleus via the nuclear pore complex. Classically, the nuclear localization signals have been classified using two patterns: the monopartite pattern, which loosely matches the K[KR]x[KR] consensus sequence, and the bipartite pattern which loosely matches the $KRx_{10\text{-}12}KRxK$ consensus sequence (Lange et al. 2007). However, we and others have shown that this pattern is rarely adequate to predict nuclear localization signals (Nguyen Ba et al. 2009). Finally, nuclear export signals are used to finely tune the temporal nature of protein localization towards the nucleus. These nuclear export signals exhibit

leucine-rich content and are more challenging than nuclear localization signals to predict due to fact that they are slightly structured signals (Xu et al. 2012). Experimentally, using GFP fusions, two major approaches to test nuclear localization/export signals have been employed. First, to test for necessity, site-directed mutagenesis of the putative localization signals can be used to test for disruption of the nuclear import or export. Second, to test for sufficiency, the putative localization signal can be used alone (or in conjunction with a nuclear localization signal, when testing for export signals) to direct 3x-GFP (to control for passive diffusion of 1x-GFP through the nuclear pore complex) to the nucleus or to the cytoplasm. Tandem-GFP (3x-GFP) is used to control for passive diffusion of the monomeric GFP through the nuclear pore complex, which allows but significantly delays GFP passage through the nuclear pore complex (Mohr et al. 2009).

Cellular processes that respond to exogenous signals require the signal transduction from the cytoplasm to the nucleus. The outcome of the several signaling cascades is the rapid translocation of either kinases or transcription factors to the nucleus (Cartwright and Helin 2000; Xu and Massagué 2004), highlighting the importance of protein localization as a regulatory mechanism.

I.4.4 Protein degradation

Cellular programs such as the cell-cycle are composed of several checkpoint sub-programs that ensure complete biogenesis of the necessary molecules before progression to the next sub-program. These sub-programs follow a specific order that must be met for adequate division and this order is governed by the successive degradation and production of proteins between the phases of the cell-cycle. At least two important complexes control the proteins present during the cell-cycle: The Skp, Cullin, F-box containing complex (SCF), which is the major regulator of protein degradation in the G1/S and G2/M transition (Ang and Wade Harper 2005), and the anaphase-promoting complex (APC) which is responsible for the metaphase-anaphase transition and the prevention of G1/S transition. These complexes are E3 ligases that attach ubiquitin and mark proteins for degradation by the proteasome (for a more thorough review on degradation signals see (Ravid and Hochstrasser 2008)).

Ubiquitination is a posttranslational modification of lysine residues that are specifically recognized by the proteasome.

Like phosphorylation and localization, protein degradation is controlled by short peptide sequences in the primary amino acid sequence of the protein. Because the E3 ligases ultimately ubiquitinate the target protein (a posttranslational modification of lysine residues), which flags proteins for degradation by the proteasome, degradation signals fall in the category of "binding sites". The best studied example of these recognition sites is the destruction box (D-box) present in many important cell-cycle proteins (Buschhorn and Peters 2006). This D-box consists of 2 important amino acids (RxxL) that are recognized by the Cdc20 and Cdh1 subunit of the APC. Interestingly, many of these proteins also possess the KEN-box, which is also recognized by the APC subunits (D-boxes and KEN-boxes probably bind the APC at different affinity). Both Cdh1 and Cdc20 are mutually exclusive subunits of the APC, and akin to cyclins, the destruction of several proteins by the same enzyme is clearly controlled by its current regulating subunit. The relative proportion of Cdc20 and Cdh1 can therefore control the relative degradation of D-box and KEN-box containing proteins.

Regulation by the SCF complex is slightly more complicated, requiring priming phosphorylation sites on the target before it can be recognized by the F-box protein. The F-box protein (e.g. Cdc4 in yeast) recognizes specific phosphorylated sequences and transports the substrate to the SCF complex. The short linear motifs controlling these sequences of events are termed 'phosphodegrons' (e.g. Cdc4-mediated phosphodegron), and therefore, the regulation of destruction is also controlled by the cell-cycle kinases (Ang and Wade Harper 2005).

I.5 Regulatory networks

At the heart of these examples of important posttranslational regulatory mechanisms underlies a complex interplay between these regulatory sequences that can be described as a regulatory network. Proteins contain several regulatory signals within their primary sequence that may not act independently and may even absolutely require sequential activation of

signaling components for proper protein function. Because these signals act in a context-specific manner, understanding the full effect of subtle changes on the output of a regulatory network has proved to be a challenging task. Under several assumptions and in the presence of perturbation data, computational modeling has been used with great effect to help deciphering the behavior of the network (Hasty et al. 2001).

An example of a complex regulatory network on a single protein is within the cell-cycle regulator Sic1, which is an intrinsically disordered protein with several regulatory signals that depend on each other for proper function (Nash et al. 2001; Kõivomägi, Valk, Venta, Iofik, Lepiku, Balog, et al. 2011). Importantly, the order and physical distance between the motifs all appear to play a role into proper cell-cycle control. In Sic1, a hydrophobic patch allows docking of Cln2-Cdk1, which primes the initial phosphorylation event. This in turn allows robust docking by Cks1 and enhances the phosphorylation of a second nearby phosphorylation site by Cln2-Cdk1. This second phosphorylation is again recognized by Cks1 to enhance phosphorylation of the first phosphodegron. Once the first phosphodegron is fully phosphorylated, Clb5-Cdk1 interacts with another docking site (RxL) on Sic1, and phosphorylates the second phosphodegron. This phosphodegron in turn is recognized by the F-box component of the SCF complex (see previous section). The requirement on the order of events is dictated by the order of the short linear motifs, which is consistent with the spacing requirements.

I.6 Regulatory divergence and short linear motif evolution

The evolution of posttranslation regulation can be studied on multiple levels. The first high-level view attempts to decipher the regulatory divergence of components between species and is most similar to the large-scale microarray analyses of gene expression across species (see (Qian et al. 2011) for a study on protein-protein interaction evolution). The second level examines more fine-scale details of the evolutionary process using multiple sequence alignments of orthologous proteins once the sequences responsible for the regulation are known (Figure I-3, see (Freschi et al. 2011)).

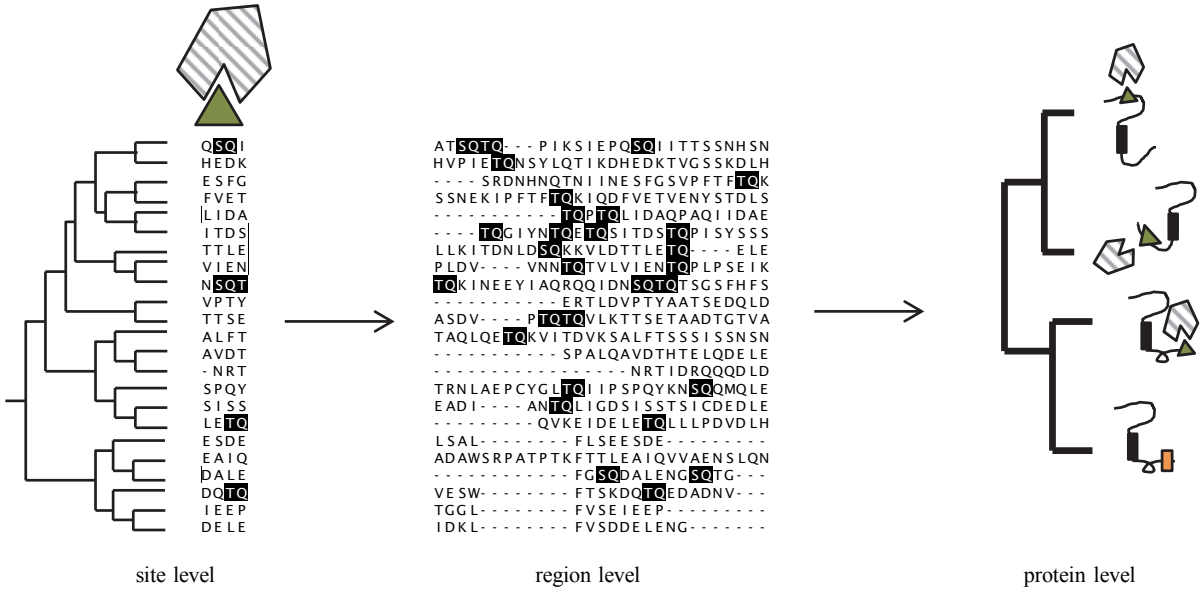


Figure I-3. Different scales to study posttranslational regulatory evolution. Studying the microevolutionary process of regulatory sequence turnover can be performed at the site level, which is useful for modeling the evolution of posttranslational regulatory divergence at the sequence level; however a more global view can provide a better picture of the protein-level changes in regulation.

In general, there is an agreement between the two types of studies in that protein regulation is usually conserved, but the rate at which regulatory changes occur are still under debate. For example, large-scale studies on individual phosphorylation sites have shown that most phosphosites are not conserved, but overall phosphorylation of the protein is maintained (Holt et al. 2009). These seemingly contradicting results can be reconciled by at least two explanations: 1) the majority of regulatory sequences identified from high-throughput experimental studies are non-functional (Landry et al. 2009), or 2) the multiple sites on proteins are under stabilizing selection with some redundancy between each sites (Holt et al. 2009).

Nevertheless, these observations are facilitated by the fact that regulatory sequences can rapidly change over evolution. In contrast to protein-protein interactions through large interfaces, new short linear motifs can be created by few point mutations because of their

short nature. Therefore, even complex regulatory networks can emerge and disappear through a small number of mutations.

A striking example of regulatory divergence over evolution occurs in the Mcm3 protein (Moses, Liku, et al. 2007). In this example, the authors used computational approaches (Moses, Hériché, et al. 2007) to predict proteins likely to be phosphorylated by Cdk1 and found that, when applied to orthologs of the Mcm3 protein, one particular yeast clade gained several phosphorylation sites in a local region. This cluster of phosphorylation sites appeared to be in close proximity of a characterized nuclear localization signal, and the authors hypothesized that the phosphorylation sites regulate the nuclear import of the protein. Consistent with this, cell-cycle regulated nuclear import of the Mcm3 protein could be observed, and this regulated nuclear import was highly correlated with the presence of the predicted phosphorylation sites, indicating that novel regulatory connections in Mcm3 occurred in evolution.

I.7 Gene duplication and regulatory changes

Although many genes between organisms are conserved, there can be an important difference in gene content across species (see (Rubin et al. 2000) for a review on eukaryotic pan-/core-genomes). In many cases, this difference in gene content is facilitated by gene duplication events (see (Cheng et al. 2005) for a comparison of segmental duplications in humans and chimpanzees). Even across human populations, copy number variations have been shown to be extensive (Redon et al. 2006) and can be important drivers of cancer mutations (Zack et al. 2013). Therefore, the evolutionary outcomes of gene duplication events have been of great theoretical interest.

There are at least three models for the fate of duplicate genes (Ohno 1970; Force et al. 1999; Lynch et al. 2001): non-functionalization of the duplicate gene by degenerative mutations, neo-functionalization whereby one copy gains a novel function and sub-functionalization where mutations repartition the functional elements of the ancestral gene. Although sub-functionalization in itself is not expected to provide any fitness advantages (as depicted by the duplication-degeneration-complementation model (Force et al. 1999)) because no new

functions have been created, beneficial mutations can arise by taking advantage of the repartitioning of ancestral gene function to specialize the duplicates (escape from adaptive-conflict (Hittinger and Carroll 2007)). These beneficial mutations can be due to changes in gene expression (such as in expressing different levels of the gene in different tissues) or due to posttranslational regulatory changes (such as in different localization of the protein in different tissues). Therefore, genetic novelty can be the target of mutations that affect protein regulation.

I.8 Intrinsically disordered regions

One of the first evidence for dynamic protein structure comes from X-ray crystallography studies on conformational change of macromolecules (Doucet and Benoit 1987), such as the ones that are derived from a phosphorylation event (Volkman et al. 2001). Despite this, only slight reorganization of the globular nature of proteins was thought to occur and this rigid-body motion of proteins was the dominant view of protein dynamics. Proteins were thought to exist in few states and regulatory mechanisms could switch proteins from one state to another.

However, recently, the discovery of proteins with no single stable tertiary structure revisited an old idea of proteins with constant dynamics (reviewed in (Forman-Kay and Mittag 2013)). Although the term appears to indicate the complete absence of constraints (colloquially described as a spaghetti noodle, see (Uversky 2013)), more precise definitions of intrinsically disordered regions are required for understanding their function. More conservative descriptions of intrinsically disordered simply use the term ‘flexible’, which appears borrowed from the structural biology field to describe changes in rotamers with oscillation in backbone dynamics. Possibly a more accurate description of intrinsic disorder in proteins it that no distinct boundary exists between structured and unstructured regions but that a continuum of movement and backbone dynamics exists (Uversky 2013). Therefore, for many intrinsically disordered proteins there are still some structural constraints on their conformations despite rapid conversions of large-scale changes in overall shape of the disordered region.

Despite current technological advances that have allowed the characterization of intrinsically disordered regions, the *in vivo* existence of intrinsically disordered regions is still controversial (Tompa 2012; Janin and Sternberg 2013). This view stems from the observation that intrinsically disordered proteins were initially described *in vitro* without the proper cellular context. This absence of proper cellular context implied that the binding partners, the effective concentration of the protein within the cell, or the solute concentrations, played important roles in the final structure of those particular disordered regions. Consistent with this, some proteins become folded when purified with their binding partner but are unfolded if purified alone (Daughdrill et al. 1997). This is supported by the existence of chaperones and a general stress response against unfolded proteins (Schröder and Kaufman 2005; Ron and Walter 2007).

However, a key discovery using bioinformatics was that, unlike the prediction of protein tertiary structures, intrinsically disordered regions could be identified by sequence alone (J J Ward et al. 2004; Prilusky et al. 2005; He et al. 2009; Mizianty et al. 2014). It was found that amino acid composition was sufficient for obtaining adequate accuracy in *de novo* discovery of protein disordered regions. The importance of these tools is highlighted by the CASP (Monastyrskyy et al. 2014) benchmarking competition where a category for predicting intrinsically disordered regions exists. In addition, these regions could be differentiated from regions prone to aggregation (Monsellier and Chiti 2007). Arguably, this indicates that disordered regions cannot simply be *in vitro* artifacts that lead to ordered regions inside the cell. Consistent with this, disordered proteins do not tend to interact with chaperones more than globular proteins (Hegyi and Tompa 2008). Using these computational tools, systematic bioinformatics analyses of several proteomes lead to the discovery that eukaryotic proteomes are abundant in intrinsically disordered regions (J J Ward et al. 2004). At least 30% of the human proteins were predicted to have disordered regions. Interestingly, the prokaryotes have very low level of proteins containing disordered regions, and disordered proteins from bacteria have been found to be enriched in secretion effectors during infection of eukaryotic hosts (Marín et al. 2013), suggesting that intrinsic disorder is a general feature of eukaryotic proteomes.

The implication of protein disordered regions in signaling is now well appreciated as many signaling proteins, such as transcription factors (Liu et al. 2006), have been shown to contain disordered regions. Importantly, site directed mutagenesis to directly remove the disordered regions have been shown to severely impair function of several of these proteins (for example, the N-terminus of p53 is disordered (Dawson et al. 2003) and required for p53 activation (Cain et al. 2000)). Because prokaryotes are poor in disordered regions, an attractive hypothesis about the existence of disordered regions revolves around some form of organismal complexity (Schad et al. 2011; Pancsa and Tompa 2012). Although this hypothesis still lacks compelling evidence, signaling complexity has been proposed to be the major reason for the expansion of disordered regions in eukaryotic cells.

Other correlations have been found with disordered proteins. Amongst them are toxicity upon overexpression (Vavouri et al. 2009) and positive correlation with the number of binding partners (Dunker et al. 2005).

I.9 Parallels with transcriptional regulation and evolution

Most of the research on regulatory evolution has been centered on transcriptional evolution and our understanding of molecular evolution affecting transcription can be leveraged to study posttranslational regulatory evolution. For example, transcription factor binding sites are typically found in *cis* near genes they affect and are short degenerate sequences that offer similar computational challenges in their prediction. Although it is thought that the selection constraints on non-coding DNA will be different than the selection constraints on proteins, some interesting properties about the evolution of individual transcription factor binding sites can be mirrored in short linear motifs (Moses et al. 2003; Moses et al. 2006) with analogous roles in regulatory divergence. For example, the probabilistic profile (the sequence logo) of the recognition sequence by transcription factors shows remarkable correlation with the rate of evolution at the individual bases (Moses et al. 2006) whereby more degenerate positions evolve more rapidly. Another interesting pattern in the evolution of transcriptional regulation is that regulatory networks tend to be more conserved than individual sites (Stefflova et al. 2013). Therefore, it is likely that there is some form of stabilizing selection on the clusters of binding sites, but that individual binding sites are free to change location. Finally, because

transcription factor binding sites appear more conserved than the remaining non-coding DNA, comparative genomics can be used to systematically predict regulatory sequences in non-coding DNA (Moses et al. 2004; Siepel et al. 2005).

These micro-evolutionary processes that affect transcription factor binding sites are facilitated by the lack in structural constraints on the non-coding DNA sequence. There exists an interesting analogy with this lack of constraints on non-coding with the lack of constraints on protein disordered regions (Moses and Landry 2010). Intrinsically disordered regions lack stable tertiary conformations and this has led to the hypothesis that selection constraints on amino acids at specific positions may not be necessary. Indeed, several lines of evidence have identified that the disordered state of a particular region can be conserved over evolution without significant residue conservation (Bellay et al. 2011).

At a higher molecular level, DNA chromatin state can alter the accessibility of transcription factors to their recognition sites and it has recently been shown that patterns of methylation can change rapidly over evolution (Gokhman et al. 2014). This is similar to changes in the state of disorder over evolution that has been observed in other studies (Bellay et al. 2011).

I.10 Research objectives and thesis overview

Currently, systematic experimental tools or computational approaches to reliably and systematically identify short linear motifs in proteins are still lacking and this has hampered the study of posttranslational regulatory divergence. To this end, this thesis presents novel computational tools to predict protein regulatory sequences and study their evolution. The specific aims of this thesis are as follows:

1) Understand the evolution of well-characterized short linear motifs.

2) Use this understanding for the *de novo* prediction of short linear motifs using comparative genomics.

3) Characterize the turnover of short linear motifs and the role of turnover in functional divergence over evolution.

4) Present experimental techniques to study the fitness effects of identified regulatory turnover.

In Chapter II, using a set of *in vivo* and functionally characterized phosphorylation sites from different kinases, the effect of purifying selection on the recognition sequences relative to their flanking regions (five amino acids on both sides) is quantified and various properties of the turnover of these phosphorylation sites are characterized (Nguyen Ba and Moses 2010). The low percentage of phosphorylation site turnover and the conservation of the regulatory sequences when compared to their flanking region both form the basis of the computational tool presented in the Chapter III. This tool was used on the whole yeast proteome and identified thousands of putative short linear motifs that are very likely to be functional (Nguyen Ba et al. 2012). In Chapter IV, using this tool and additional statistical methods, I identified short linear motifs that are likely to have relaxed selective constraints after the whole-genome duplication event in budding yeast. Specifically, I found that paralogous genes are much more likely to undergo regulatory changes after gene duplication than single-copy genes (Nguyen Ba et al. submitted). Finally, in Chapter V, I use state-of-the-art genetic techniques to assay the fitness contributions of regulatory network rewiring after gene duplication.

A general discussion of future directions and outstanding questions relating to the evolution of disordered regions and more complex regulatory network is presented in Chapter VI.

Chapter II

Evolution of Characterized Phosphorylation Sites in Budding Yeast

This is an author-produced PDF of an article accepted for publication in Molecular Biology and Evolution following peer review. The version of record:
Evolution of characterized phosphorylation sites in budding yeast
Mol Biol Evol. 2010 Sep;27(9):2027-37. doi:10.1096/molbev/msq090. Epub 2010 Apr 5.
Alex N Nguyen Ba[1,2], Alan M Moses[1,2,¥]
is available online at: http://mbe.oxfordjournals.org/content/27/9/2027.abstract

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, Toronto, Canada
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada

II.1 Abstract

Phosphorylation is one of the most studied and important regulatory mechanisms that modulate protein function in eukaryotic cells. Recently, several studies have investigated the evolution of phosphorylation sites identified by high-throughput methods. These studies have revealed varying degrees of evidence for constraint and plasticity, and therefore, there is currently no consensus as to the evolutionary properties of this important regulatory mechanism. Here, we present a study of high-confidence annotated sites from budding yeast and show that these sites are significantly constrained compared with their flanking region in closely related species. We show that this property does not change in structured or unstructured regions. We investigate the birth, death and compensation rates of the phosphorylation sites and test if sites are more likely to be gained or lost in proteins with greater numbers of sites. Finally, we also show that this evolutionary conservation can yield significant improvement for kinase target predictions when the kinase recognition motif is known, and can be used to infer the recognition motif when a set of targets is known. Our analysis indicates that phosphorylation sites are under selective constraint, consistent with their functional importance. We also find that a small fraction of phosphorylation sites turnover during evolution, which may be an important process underlying the evolution of regulatory networks.

II.2 Introduction

Protein phosphorylation is a ubiquitous posttranslational modification in cells as a means to regulate a variety of cellular processes (Johnson and Hunter 2005). Despite its importance, until recently, few studies had examined the evolution of this regulatory mechanism. Phosphorylation sites are critical functional elements within proteins, and therefore, they are expected to be conserved over evolution. This conservation can be exploited to predict kinase substrate interactions (Budovskaya et al. 2005). However, two recent studies examined the evolution of phosphoregulation in the eukaryotic cell cycle and found evidence for evolutionary changes in the regulatory networks (Jensen et al. 2006; Moses, Liku, et al. 2007). Furthermore, a structural study of phosphorylation sites in mitotic proteins found similar levels of conservation between phosphorylation sites and other similar residues (Jiménez et al. 2007), suggesting no specific constraints on these sites.

With the availability of high-throughput data sets, it has become possible to examine the evolutionary properties of large sets of phosphorylation sites (Macek et al. 2008; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009). Most studies have found evidence for evolutionary conservation of phosphorylated S/T/Y residues compared to unphosphorylated residues (Gnad et al. 2007; Macek et al. 2008; Malik et al. 2008; Landry et al. 2009). In addition, one high-throughput study compared phosphorylation patterns between distantly related yeast species and quantified the rate of evolution of these patterns (Beltrao et al. 2009). Despite providing evidence for constraint, these studies all identified a large number of phosphorylation sites that were not preserved over evolution. These nonconserved sites may contribute to the large difference of patterns of phosphorylation between species (Beltrao et al. 2009). However, it is also important to consider whether many of the sites contained in high-throughput data sets are not critical to protein regulation; for example, some fraction of sites obtained by mass spectrometry may not be functional sites (Lienhard 2008; Landry et al. 2009). These nonfunctional sites are not expected to be preserved over evolution, and therefore may appear as evolutionary changes.

Another important issue is that the alignments used in some of the previous studies include sequences from distantly related species, which creates uncertainty in the analysis because

short degenerate motifs such as phosphorylation sites may not be aligned accurately in distant species comparisons (Balla et al. 2006).

Motivated to address these difficulties, we sought to examine the evolution of a large set of high-confidence phosphorylation sites, where we could obtain high-confidence alignments of orthologous protein sequences. To do so, we assembled 249 characterized phosphorylation sites in budding yeast from the literature, where the likely kinase responsible for phosphorylation is known. By examining alignments of protein sequences from closely related species we can explicitly test evolutionary hypotheses about phosphorylation sites using the ratio of nonsynonymous to synonymous substitutions (Ka/Ks) (Nei and Gojobori 1986). Our results show that the rate of amino acid substitution within the site is lower than the surrounding region and that this property is observed whether the sites appear in structured or unstructured regions of the substrate proteins. As expected, we find that the patterns of substitution in phosphorylation sites are consistent with the specific constraints imposed by the consensus recognition site for the kinase. We also investigate the birth and death and compensation rates of these annotated sites and show that there are evolutionary constraints on the appearance and disappearance of sites in targets of kinases, but only weak constraints on compensation. We also consider the possibility that gain and loss of phosphorylation sites is due to redundancy, but we find no evidence that sites are more likely to be lost or gained in proteins with high number of sites.

Finally, we show that the evolutionary conservation of phosphorylation sites relative to surrounding amino acid sequence can be exploited to improve prediction of kinase substrates or to find the kinase specificity.

II.3 Results

II.3.1 A set of functional phosphorylation sites in *S. cerevisiae* for which the kinase is known

In order to study the evolutionary properties of phosphorylation sites, we searched the literature for experimentally verified phosphorylation sites where the kinase had been identified in low-throughput experiments. Although there is no single experiment that conclusively shows that a specific site is phosphorylated by a specific kinase *in vivo*, we chose to include phosphorylation sites where 1) site-specific mutagenesis on the phosphoacceptor site (usually S/T/Y to A to create nonphosphorylatable mutants) has revealed a functional role for the site or group of sites or 2) low-throughput identification of phosphosites by mass spectrometry had identified sites. The vast majority of the sites included have been confirmed by site-specific mutagenesis. In addition, we required that each site has evidence for the specific kinase responsible, either by *in vitro* experiments showing phosphorylation of the site by that kinase or by *in vivo* experiments showing that the phosphorylation or mutant phenotype depended on a particular kinase. Because kinases usually recognize a short degenerate consensus sequence around the phosphorylated residue (Miller and Blom 2009), knowing the identity of the kinase responsible allows us to accurately define the extent of expected conservation around the phosphorylation site. This contrasts with previous studies on phosphoevolution where phosphorylation sites have been obtained with high-throughput mass spectrometry (Macek et al. 2008; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009) and where the kinase was unknown. In those studies, the evolutionary properties of only the phosphoacceptor sites can be studied.

We focused on seven kinases for which we could define a consensus sequence: CDK (or Cdc28), Mec1, CKII (Cka1/Cka2/Ckb1/Ckb2), Prk1, Ipl1, PKA (or Tpk1/Tpk2/Tpk3) and Pho85 (see Table II-1). We manually aligned all the sites for each kinase and determined the extent of the consensus sequences based on the information content. These consensus sequences are represented as seqlogos (Schneider and Stephens 1990; Crooks et al. 2004) in Figure II-1. We refer to these sites as “annotated” phosphorylation sites, and we believe that

they represent a high-confidence set of *bona fide* phosphorylation sites in budding yeast. A complete table of these sites and references can be found as supplementary table II-S1.

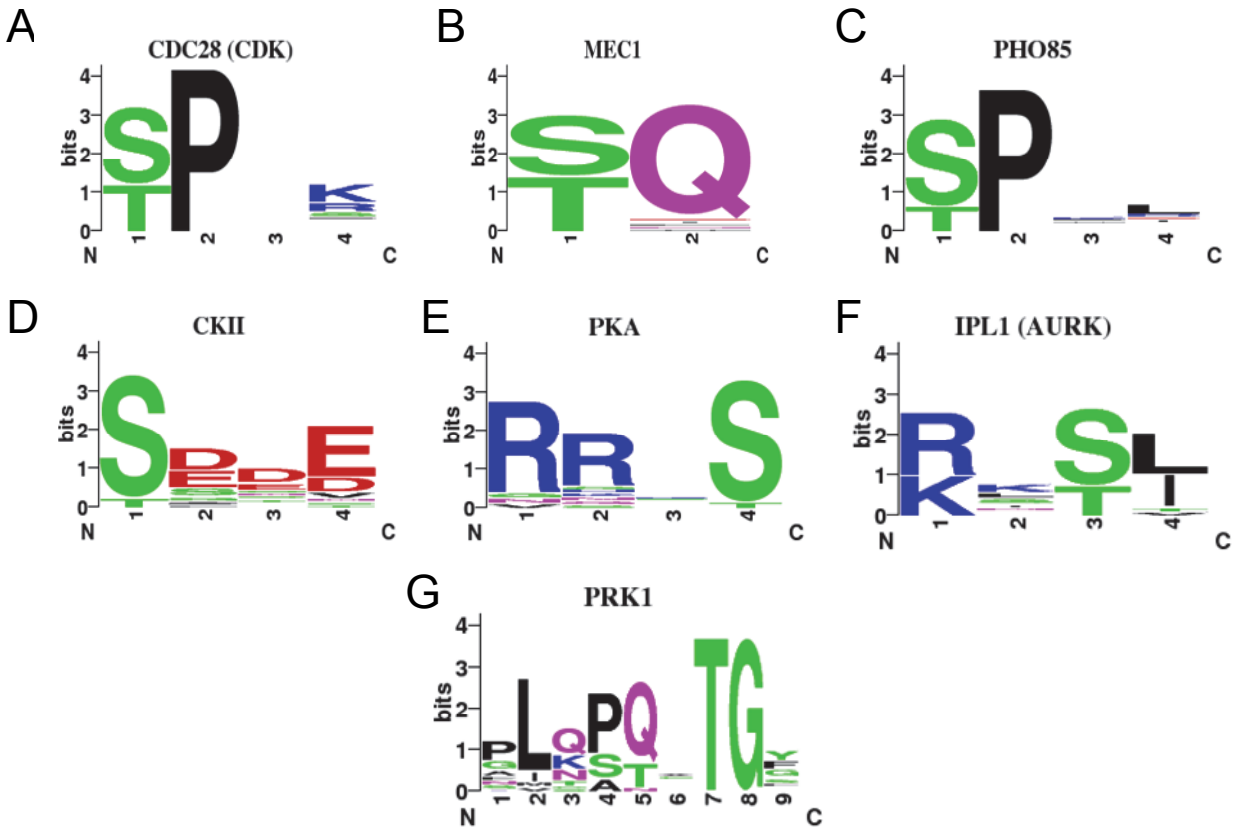


Figure II-1. Sequence logos of aligned annotated sites of seven kinases A-G) Annotated sites from each kinases were aligned along with their flanking regions and boundaries were chosen where the information content was > 1.

II.3.2 There is evidence of conservation of phosphorylation sites

We first sought to test for evidence of evolutionary constraint on the annotated phosphorylation sites. To perform our analysis, we aligned orthologous proteins from four closely related species of yeasts (*S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*, see Materials and Methods). We then used maximum parsimony (Durbin et al. 1998) (see Materials and Methods) to calculate the rate of amino acid and synonymous substitution (*Ka* and *Ks*) (Nei and Gojobori 1986) in the phosphorylation sites (termed "site"). Because

phosphorylation sites occur preferentially in unstructured regions of proteins (Gnad et al. 2007; Landry et al. 2009), and phosphoproteins evolve more slowly than other proteins (Gnad et al. 2007), comparing them with a random sample of sites can be misleading. We therefore compared the rates of evolution in the characterized phosphorylation sites with five amino acids on each side (termed "flank"). We use the flanking region to control for structured and unstructured segments of proteins, as well as different rates of protein evolution. To explicitly test for a difference in substitution rate between the sites and flanks, we performed an LRT to compare the hypothesis that the site evolves at a different rate than the flanking region with the hypothesis that a single rate of evolution explains the patterns in both classes (see Materials and Methods). In each case we also performed a nonparametric bootstrap to confirm the significance of our results (see Table II-1).

Table II-1. Summary of the sites included in our analysis. *Ka*/*Ks* was calculated as in the Materials and Methods. The LRT is the likelihood ratio statistic and the P-value is given following a X^2 with a degree of freedom equal to 1 (See Materials and Methods). P-value from bootstrap analysis is given by the number of times the *Ka/Ks* of the flank was observed to be higher than the site in 1000 non-parametric bootstraps (See Methods). Double asterisks show strong significance ($P < 0.01$), and one asterisks shows significance ($P < 0.05$).

Kinase	Number of phosphorylation sites	*Ka*/*Ks* in site	*Ka*/*Ks* in flank	LRT	P-value of LRT	P-value of bootstrap
CDK	114	0.071	0.109	19.62	$9.4 * 10^{-06}$ **	0.005 **
Mec1p	47	0.062	0.173	24.26	$8.4 * 10^{-07}$ **	<0.001 **
Ipl1p	18	0.140	0.157	0.008	0.93	0.37
CKII	17 (27)	0.050	0.099	3.06	0.08	0.003 **
Prk1p	20 (23)	0.034	0.060	4.94	0.03 *	0.128
PKA	14 (18)	0.126	0.057	N/A	N/A	0.986
Pho85p	19 (22)	0.025	0.062	5.64	0.018 *	0.031 *
Total	249 (269)	0.069	0.118	64.36	$1 * 10^{-15}$ **	<0.001 **

[a]In parenthesis is the total number of sites found in the literature. Our analysis excluded overlapping sites.

N/A: Not applicable due to higher Ka/Ks in flank.

Using this method, we found that there is a significant reduction in amino acid substitution rate within the phosphorylation sites as compared with the flanking regions (*Ka*/*Ks* 0.069 vs. 0.118 for sites and flanks respectively, LRT = 64.36, P-value<10^{-14}, Figure II-2a and Table II-1). This indicates that phosphorylation sites evolve under specific evolutionary constraint relative to the regions in which they occur in proteins.

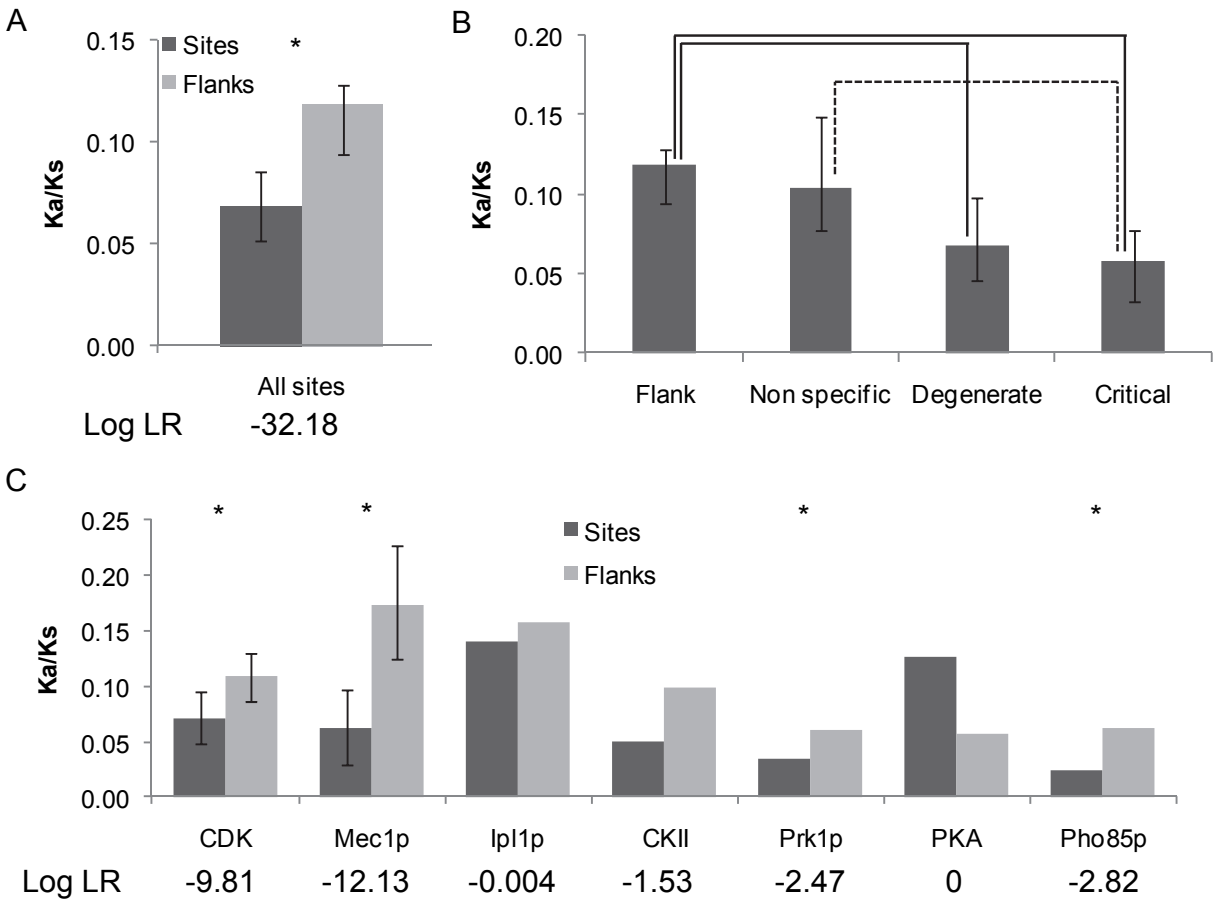


Figure II-2. Difference in *Ka/Ks* ratio between sites and their flanking regions. A) *Ka*/*Ks* ratio of annotated sites and their flanking regions. B) *Ka*/*Ks* ratio of amino acids denoted as critical, degenerate or nonspecific and the flanking region of each site. Lines between each bar shows significant differences under the LRT. C) *Ka*/*Ks* ratio of annotated sites from each kinase. In all graphs, the error bars are obtained from a 95% confidence interval from a nonparametric bootstrap of 1000 replicates. Error bars are only shown for cases with at least 40 phosphorylation sites. Significance from the LRT is shown as an asterisk ($P<0.05$).

Because we defined a phosphorylation site to include more information than the phosphoacceptor, we investigated the *Ka*/*Ks* ratio of amino acids defined as critical (substitution would very likely prevent phosphorylation), degenerate (substitution may lower phosphorylation affinity), or nonspecific (substitution is unlikely to impact phosphorylation). We categorized each position in the recognition motifs based on the information content (see Materials and Methods). We calculated the *Ka* and *Ks* at each position of the

phosphorylation site consensus sequence and binned the sites according to the categorization. As expected, the *Ka*/*Ks* ratio of the amino acids defined as critical (0.057) is lower than degenerate amino acids (0.068), which are lower than nonspecific amino acids (0.104, Figure II-2B).

Performing the analysis on each kinase independently reveals a lower *Ka*/*Ks* ratio in the sites versus the flanking residues for all the studied kinases but PKA (Figure II-2C). The LRT indicates that most of the sites of kinases have a significantly lower rate of substitution than the flanking residues.

It is possible that this lower rate of substitution is due to the difference in frequency of particular amino acids within the consensus sequences, as compared to the flanking regions. To confirm that this could not explain our results, we used as a negative control 514 proteins which were shown not to be targets of CDK by Ubersax et al (Ubersax et al. 2003). We randomly sampled an equivalent number of sequences matching the CDK consensus from these "nontargets" and computed the *Ka*/*Ks* ratio as well as the LRT described above (see Materials and Methods). On average these nontargets showed LRT statistics of 1.8 (s.d=2.23), much less than the LRT=19.62 observed for the annotated sites. This indicates that the differences in amino acid composition between the sites and flanks cannot account for the large LRT statistics that we have observed in the annotated sites. We note that the nontargets include some fraction of false negatives, and therefore this can be regarded as a conservative estimate for the contribution of the difference in residue frequencies. Therefore, at least for CDK sites, the lower rate of substitution within the site compared to the flanking region was not due to amino acid content as our negative and annotated sets show dramatically different results while having similar amino acid content.

We note that most of the sites in our data set appear in unstructured regions of proteins (75% unstructured, 25% structured), and a previous study (Landry et al. 2009) has stressed the importance of studying the context of the phosphorylation site in evolutionary analyses. However, we found that the constraint observed above is similar in both structured and unstructured regions (Figure II-3).

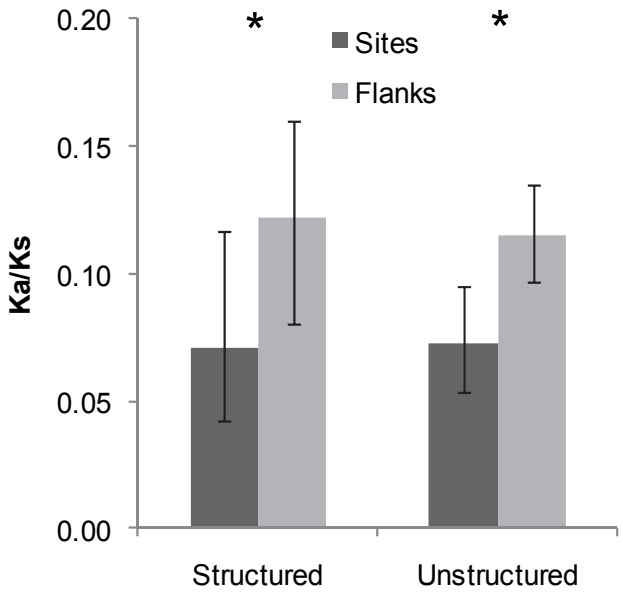


Figure II-3. Difference in *Ka*/*Ks* ratio between sites and their flanking regions categorized by structured and unstructured regions. *Ka*/*Ks* ratio and results of the LRT of annotated sites and their flanking regions for annotated sites separated by structured or unstructured regions. Error bars were obtained from a 95% confidence interval from a nonparametric bootstrap of 1000 replicates. Significance from the LRT is shown as an asterisk ($P<0.05$).

II.3.3 Phosphorylation site turnover

Previous studies have shown that phosphoregulation may change over evolution (Moses, Liku, et al. 2007), and consistent with this, alignments of phosphorylation sites over long evolutionary distances show evidence of change (Holt et al. 2009; Tan et al. 2009). One possible explanation is that the sites may not be required to stay at a particular location in a protein and therefore may shift position over evolution (Moses, Liku, et al. 2007; Holt et al. 2009; Tan et al. 2009), especially in unstructured regions (Brown et al. 2002). Another explanation is that proteins with multiple sites may lose or gain a few sites without changing the regulation of the protein (Moses, Liku, et al. 2007; Serber and Ferrell 2007). In both these cases, functional phosphorylation site turnover does not impact protein function. However, a third possibility is that phosphorylation sites identified in high-throughput experiments may

be nonfunctional, and the evolutionary changes we observe are simply due to the loss of nonfunctional residues.

We decided to test whether we could observe the microevolutionary steps that underlie the changes in functional phosphoregulatory networks. We therefore sought to quantify the rate of turnover of phosphorylation sites within our set of annotated sites. We considered sites that contained substitutions of the critical residues (see Materials and Methods) to be nonconserved. Of the 249 functional sites in our set, we found 22 (8.8% +/-1.8) that were not conserved in the alignments of the closely related species studied here. We confirmed that these nonconserved sites were not due alignment errors or missing data (supplementary table II-S2).

To test for evidence of selection influencing the rates of phosphorylation site turnover, we took advantage of the CDK unbiased "nontarget" set (Ubersax et al. 2003). If selection acts to preserve functional phosphorylation sites, we predict that characterized sites should be lost at a slower rate than similar sequences in the nontargets. On the other hand, if phosphorylation sites were recently added by positive selection, we would expect to see a faster rate of characterized phosphorylation site gain relative to the appearance of matches to the consensus sequence in proteins we know not to be targets. Another hypothesis for the observation of turnover is that constraint at the individual site is superseded by the constraint that total number of sites should be conserved. Therefore, if selection is acting to preserve the total number of sites in a protein, a "death" can be compensated by a "birth" nearby on the same lineage. To test these hypotheses, we compared the rate of birth (Figure II-4A for an example), death (Figure II-4B for an example), and compensation (Figure II-4C) from our annotated CDK sites with the consensus sequences in the set of unbiased negative proteins (nontargets) from Ubersax et al. We chose the set of unbiased nontargets because the birth rate obtained from the total negative set would be biased, as these proteins were chosen to be tested on the basis of the presence of a consensus sequence.

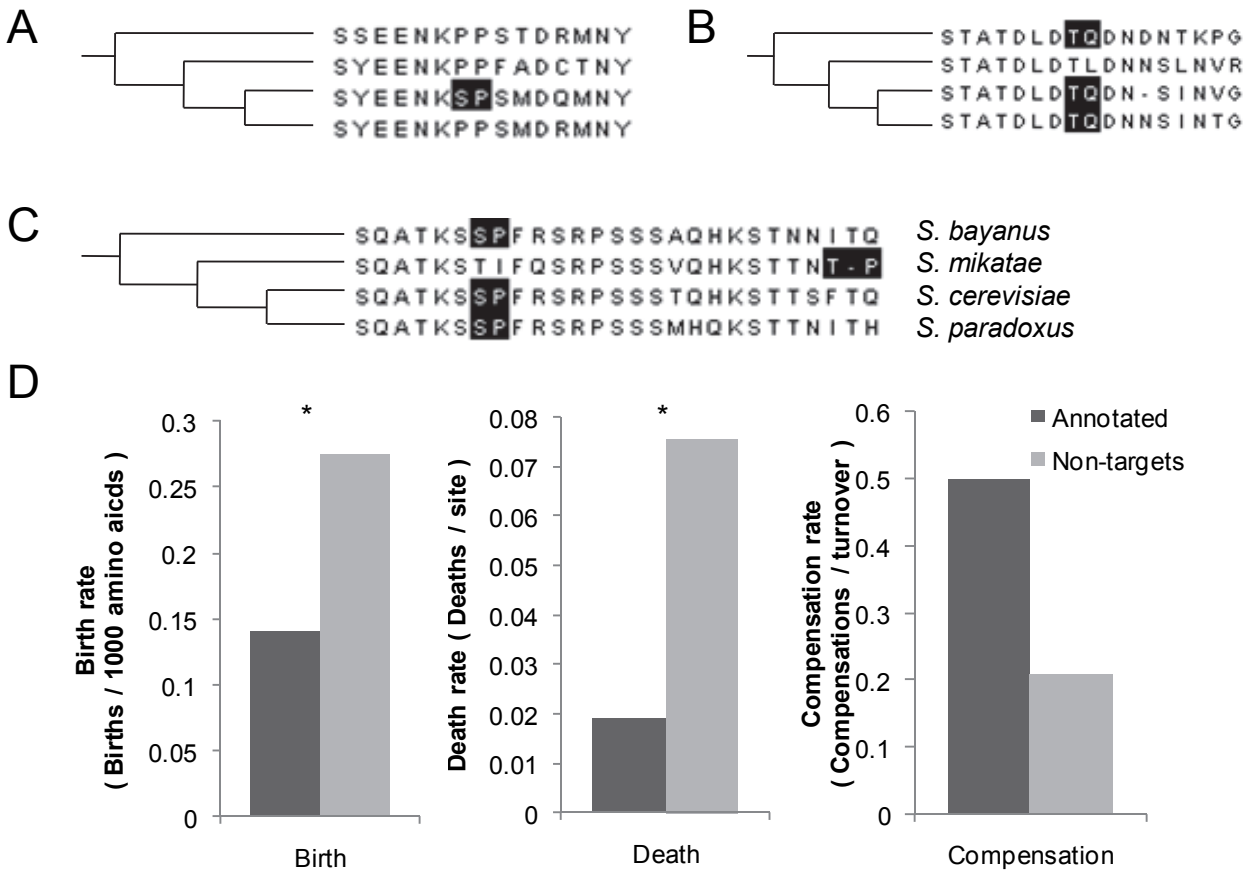


Figure II-4. Birth and death rate comparison. A) Example of a site birth. Site shown is a CDK site in Cnm67. B) Example of a site death. Site shown is a Mec1 site in Mrc1. C) Example of site death compensated by a birth. Site shown is a CDK site in Tgl4. D) Birth, death, and compensation rates of our annotated CDK sites were compared to the consensus sequences appearing in the set of unbiased negative targets by Ubersax et al. Significance from a Fisher's exact test is shown as an asterisk ($P < 0.05$).

We counted birth as sites appearing in the lineage leading to *S. cerevisiae* after the divergence with *S. mikatae* or *S. paradoxus* and deaths where sites disappeared in either *S. mikatae* or *S. paradoxus*, and compensation as a pair of birth and death within the same lineage. Doing so, we found that in our annotated sites, both the birth rate and the death rate were lower than in the set of unbiased nontargets (0.019 vs 0.075 for deaths P=0.04 and 0.14 vs 0.55 P<0.01 for births, Fisher's exact test, Figure II-4D). Although we observed an increased rate of compensation within our annotated set, it was not found to be significant (0.5 compensation/turnover vs 0.2 compensation/turnover, P-value = 0.14, Fisher's exact

test, Figure II-4D). To control for the possibility that this birth and death rate difference is due to the difference in the amount of structured or unstructured regions in both sets, we also calculated the birth and death rates on only unstructured regions or structured regions. Doing so, we found similar results: Both the birth and death rates are lower in the annotated set whether or not we look at structured or unstructured regions (data not shown).

Taken together, this analysis indicates that functional sites are under selective constraint to be preserved and that the *bona fide* targets of the kinase are also less likely to spawn new sites, suggesting selection against spurious matches to the consensus. We propose that in real targets, the appearance of new sites is more likely to disrupt protein function (e.g., inappropriate phosphorylation of a protein domain) than in the nontargets where consensus matches are likely not to be phosphorylated (e.g., because the kinase is never localized close to the substrate).

If selection acts on the number of phosphorylation sites, rather than the specific residues, loss or gain of sites may be permissive in proteins with a high number of phosphorylation sites. We found that the average 8.8% site turnover was seen across all proteins regardless of their site count and found no significance with a simulation of random turnover event (Figure II-5). We tested our statistical power to observe significance in this test and found that, in a simulation where we assumed a site was n times more likely to be lost in a protein with n sites than in a protein with a single site, we did have a large enough sample size to detect this effect.

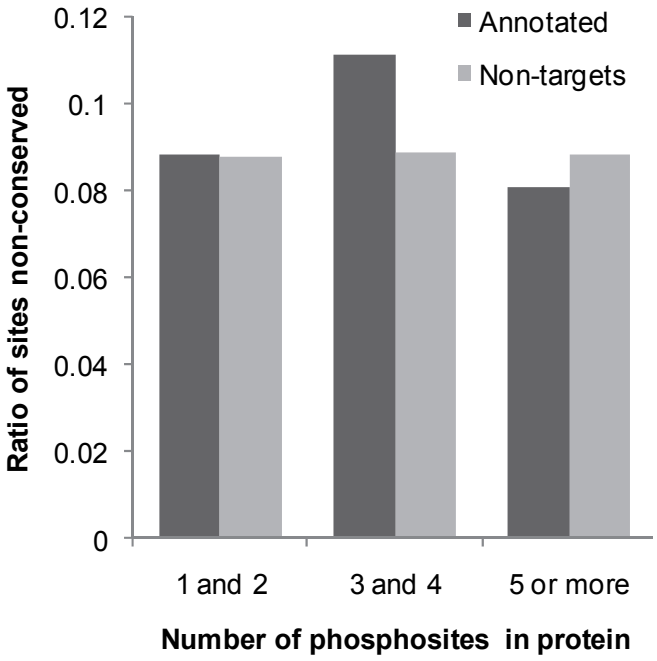


Figure II-5. Turnover in proteins with different number of sites. Percentage of nonconserved site within proteins of different number of sites. Significance of the different distributions was assessed with a X^2 test at a significance level of 5% with degrees of freedom equal to 2.

II.3.4 Conservation of phosphorylation sites can improve kinase target and specificity predictions

Two important challenges in computational biology are predicting kinase substrates based on kinase specificity (Kobe et al. 2005; Turk 2008; Miller and Blom 2009) and predicting kinase specificity given a set of known substrates (Schwartz and Gygi 2005). We observed that experimentally confirmed phosphorylation sites had a lower rate of substitutions than their flanking region. We sought to see if this information could be used to improve kinase target prediction or if it could uncover specific recognition motifs. To perform this analysis, we attempted to predict targets and specificity of CDK and Mec1, the kinases for which we had the most available data.

We first attempted to predict kinase targets. For CDK, our set of positives and negatives were obtained from Ubersax et al.'s set of CDK targets using a CDK-as1 allele (Ubersax et

al. 2003). To incorporate the slower rate of evolution of phosphorylation sites into sequence-based prediction, we applied the LRT described above to the matches to the CDK consensus sequence in each particular protein. Running our LRT on individual proteins, we observe that the likelihood ratio alone is a strong predictor of targets, yielding significant positive predictive value ($P<0.05$, Fisher's Test compared to the consensus sequence alone). Because proteins with more matches to the CDK consensus are more likely to represent *bona fide* targets of this kinase (Moses, Hériché, et al. 2007), we also analyzed targets separately depending on the number of matches to the full consensus.

We find that the improvement in positive predictive value is more pronounced when the number of full consensus matches within the protein is lowest (Figure II-6a). This is likely due to the strong predictive power achieved in the case of large numbers of consensus sites in the absence of evolutionary information. The evidence for constraint on phosphorylation sites improves prediction in the cases where consensus sites alone provide poor predictive power.

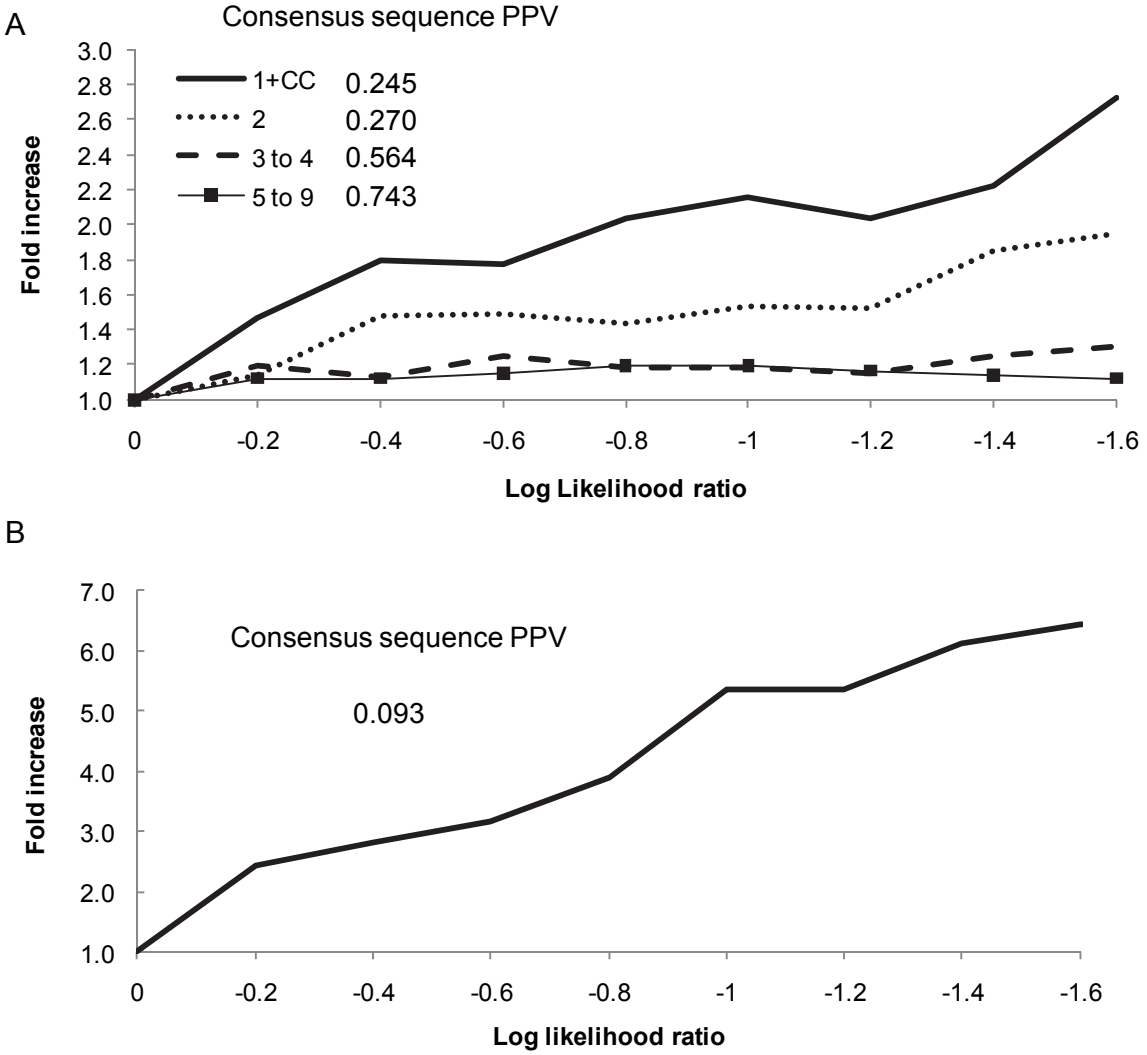


Figure II-6. Improved target prediction using the LRT. A) Ratio of the positive predictive value of the LRT against the consensus sequence alone on the Ubersax et al. set of negative and positive targets of CDK. The number of matches to the full CDK consensus sequence was used to separate the set in multiple categories. B) Ratio of positive predictive value of the LRT against the consensus sequence alone on our set of proteins with annotated phosphorylation site. Positive targets were genes that had annotated Mec1 sites and negative targets were the rest of the proteins with annotated sites having a Mec1 consensus sequence.

We then performed a similar analysis on Mec1 targets. Our set of positives were the proteins with annotated Mec1 sites, and our set of negatives were protein targets of other kinases within our initial data set. Similar to CDK, we observe that the LRT on individual protein is a strong predictor of targets with significant positive predictive value ($P<0.05$, Fisher's Test compared to the consensus sequence alone) (Figure II-6b).

We next tested whether the evolutionary information could be used to search for the kinase recognition motif. For a given set of kinase targets, we identified all k-mers that included a serine or a threonine within unstructured regions and tested their conservation compared with their flanking region using the LRT. For k-mers that are found at least three times in the unstructured regions of the substrates, we found that the LRTs are sufficient to uncover both the CDK recognition motif ([ST]-P, indicated with circles in Figure II-7a) and the Mec1 recognition motif ([ST]-Q, indicated with circles in Figure II-7b). As a negative control, we performed a similar analysis on the non-CDK targets described above, and did not recover the CDK consensus (data not shown).

A

	k-mer	LRT
●	SP	92.8727
●	TP	22.6324
●	PSP	21.7248
●	SPK	19.031
●	SPKK	16.2431
●	TPTK	13.3427
●	TPSK	12.55
●	SPVK	12.2602
	SSSY	11.7418
●	KTPS	11.7328
	SQ	11.4341
●	LSP	11.3636
●	KSPE	11.35
●	GSSP	11.1328
●	SSPVK	10.9455
●	TPRR	10.7002
	SSSSL	10.6253
●	KSP	10.3049
	TKQ	10.127
●	STPTK	9.99264
●	SPLK	9.33111
	SQGS	8.97087
	SSS	8.82163
●	SPV	8.6286
●	TPS	8.58965
●	FTPR	8.51738

B

	k-mer	LRT
●	TQ	13.6924
	SD	9.9984
	SNS	6.95316
●	SQ	6.9488
●	DTQ	6.58132
	DSD	4.93539
	DSE	4.77475
	TL	4.31038

Figure II-7. Kinase specificity prediction using the LRT. A) K-mers with serines or threonines ranked by their likelihood ratio statistics in the unstructured regions of CDK targets. Black circles are k-mers which fit the known CDK consensus sequence. B) K-mers with serines or threonines ranked by their likelihood ratio statistics on Mec1 targets. Black circles are k-mers which fit the known Mec1 consensus sequence.

This indicates that the evolutionary conservation can be complementary to the information about the number of consensus sites when predicting kinase substrates and that it can help in predicting kinase recognition motifs.

II.4 Discussion

Our study differs from previous studies in four main methodologies. First, the phosphorylation sites in our study were known to be functional. Second, we only included closely related species of yeasts. Third, we included important residues other than the phosphosite in the evolutionary rate calculations. Finally, the phosphosites were categorized by their respective kinases in order to verify differences between kinases. Our methodology was chosen to ensure that we obtained reliable alignments and to ensure that our analysis did not include falsely labeled phosphorylation sites. Thus, we have higher confidence in the alignments and the studied sites, but our conclusions are based on less data and substitutions, which are necessary to infer evolutionary properties.

Nevertheless, our analysis of the annotated phosphorylation sites in *S. cerevisiae* yielded several results which have been suggested in other studies (Gnad et al. 2007; Macek et al. 2008; Beltrao et al. 2009; Holt et al. 2009; Landry et al. 2009; Yachie et al. 2009): On average, phosphorylation sites show evidence of functional constraint, but individual sites appear to turnover during evolution. We note that the number of sites in our study is much smaller than many of the previous studies (and represents a small fraction of the total number of phosphorylation sites in the yeast proteome). Furthermore, as we only studied seven kinases, we note that it might not be possible to generalize our study to the whole phosphoproteome. Indeed, results vary between kinases: For example, PKA sites did not show evidence for constraint relative to their flanking sequences. At least in the case of CDK, however, our results seem to be generalizable, as we observe similar results (see Appendix Figure I-1) on putative phosphorylation sites in a large set of CDK targets (Ubersax et al. 2003).

In addition, conservation of sites within very closely related species of yeasts has been observed for sites within targets of CDK (Holt et al. 2009) and the lower death rate in annotated CDK sites that we observed is consistent with the observation that enrichment of sites is also maintained over evolution (Holt et al. 2009).

In another study, it had been shown that phosphorylated residues from high-throughput data in yeast do not appear to be constrained when compared to nonphosphorylated serines/threonines/tyrosines (Landry et al. 2009). Further analysis from that study showed that constraint was significantly observed when comparing phosphorylated residues with known function instead in a human dataset. Consistent with this, we found that constraint could not be observed by comparing the phosphoacceptor with its flanking residues in their yeast high-throughput dataset (see Appendix Figure I-2). Because our sites were all shown to be functional, our analysis confirms the idea that functional sites are more likely to be preserved (Budovskaya et al. 2005; Landry et al. 2009).

The above indicates that evolutionary information can therefore be used to infer functional phosphorylation sites in proteins known to be phosphorylated by a certain kinase. Indeed, some studies in the past have taken advantage of this information (Wang et al. 2005; Koch et al. 2009) for their protein of interest. Evolutionary conservation has also been used systematically to predict novel substrates of PKA (Budovskaya et al. 2005). Although we did not devise a method to predict targets of kinases, we confirmed that, as a proof of concept, simple evolutionary constraint improves predictive power in proteins with small numbers of full consensus sites.

Furthermore, kinase specificity has been predicted in the past using enrichment of linear motifs (Schwartz and Gygi 2005) as proteins phosphorylated by a kinase often share a common recognition motif around the phosphoacceptor. We also have shown that because these motifs tend to be conserved, evolutionary information can help in predicting the kinase recognition motif in the case where the substrates are known.

Although conservation seems to be a general feature of functional phosphorylation sites, by looking at well-characterized sites in closely related species, we quantified the process of evolutionary change in phosphorylation sites: We found that ~9% of phosphorylation sites

were not conserved in the closely related species considered here. Turnover of phosphorylation sites could be consistent either with redundancy of sites in unstructured, multiply phosphorylated regions or with changes in phosphoregulatory networks (Moses, Liku, et al. 2007; Beltrao et al. 2009; Tan et al. 2009).

Consistent with the hypothesis that the total number of sites within a protein is conserved, and the individual sites are free to turnover, we observed examples of compensation within our annotated set, although we could not find statistical evidence for selection on this process. Similarly, we could not find any evolutionary evidence for redundancy of phosphorylation sites in multiply phosphorylated proteins.

On the other hand, it has been proposed that changes in regulatory networks are enabled by the presence of phosphorylation sites in unstructured regions that evolve rapidly (Collins 2009; Holt et al. 2009). However, our data showed similar constraints on phosphorylation sites in structured and unstructured regions, which is in agreement with previous results regarding functional sites (Landry et al. 2009). Another explanation for the prevalence of phosphorylation sites within unstructured regions is that phosphorylation of structured regions is more likely to disrupt function. This is supported by the fact that there are fewer consensus matches in structured regions in the targets than expected (0.1846 in annotated targets vs. 0.2853 per thousand residues in non-targets). However, we also observed a reduction in the birth rate of phosphorylation consensus sites in *bona fide* targets relative to the nontargets, indicating that spurious phosphorylation sites may disrupt function even in the unstructured regions. Understanding the constraints on the organization of regulatory sequences in proteins is an important area for further research.

Nevertheless, our evidence for “turnover” of characterized phosphorylation sites provides the microevolutionary material for the plasticity in regulatory networks that has been observed over longer evolutionary timescales (Jensen et al. 2006; Moses, Liku, et al. 2007; Beltrao et al. 2009). This plasticity in regulatory networks has also been seen in other classes of regulatory elements, namely in transcription factor binding sites. It is interesting to note that the fraction of nonconserved phosphorylation sites found in our study (8.8% over species divergent by 5-20 million years (Kellis et al. 2003)) seems proportional to the fraction of

nonconserved functional transcription factor binding sites when comparing humans to mouse, where 36-40% of sites turnover (Dermitzakis and Clark 2002) during the 65-75 million years separating those species (Mouse Genome Sequencing Consortium et al. 2002)). Our finding that the positions with high information content in phosphorylation sites showed fewer substitutions also mirrors the patterns of evolution seen in transcription factor binding sites (Moses et al. 2003). Therefore, we speculate that conservation with a small rate of turnover is likely to be a general feature of many other classes of regulatory elements.

II.5 Materials and Methods

II.5.1 Alignment of closely related species of yeasts

Genomic sequences from the four species in our study (*Saccharomyces cerevisiae*, *Saccharomyces bayanus*, *Saccharomyces paradoxus* and *Saccharomyces mikatae* (Kellis et al. 2003)) were obtained from the SGD (SGD Project 2011) and translated open reading frames were aligned using t-coffee (Notredame et al. 2000) at default settings. DNA sequences were aligned using the aligned protein sequences by inserting the gaps from the protein sequence alignments into the cDNA sequences. In all, 86% (5045/5884) of the genes in *S. cerevisiae* were aligned successfully. The amino acid sequences of these species are very similar: 73% of the columns in these alignments have no amino acid differences.

II.5.2 Consensus sequences of phosphorylation sites

The phosphoacceptor for each kinase was aligned, and the flanking sequences were added afterwards. We created a seqlogo (Schneider and Stephens 1990; Crooks et al. 2004) for each kinase and set the consensus sequence to start and end where the information content equaled 1.

We defined critical residues as those residues that are likely to be necessary for phosphorylation. These include the phosphoacceptor and residues with information content comparable to the phosphoacceptor. We defined degenerate residues to be residues with lower but observable information content and nonspecific residues as residues with marginal information content.

We defined phosphorylation sites as the consensus sequence match of the respective kinase. This includes the phosphoacceptor as well as critical, degenerate, and nonspecific residues as described above.

II.5.3 *Ka* and *Ks* calculation

To calculate the rate of synonymous (or nonsynonymous) substitution, *Ks* (or *Ka*), we calculated the number of synonymous (or nonsynonymous) substitutions and divided by the number of synonymous (or nonsynonymous) sites. This calculation was done either on individual columns of alignments or on the 'site' and 'flank'. To calculate the number of substitutions, we used the maximum parsimony algorithm (Durbin et al. 1998) with no weighting on the amino acid sequence, and to calculate the number of synonymous or nonsynonymous sites, we used the method presented by Nei and Gojobori (Nei and Gojobori 1986).

Error bars were obtained by nonparametric bootstrapping of 1000 samples with a 95% confidence interval (Nei and Kumar 2000). P-value from bootstrap analysis is obtained by counting the number of times the *Ka*/*Ks* of the flanking region is slower than the site divided by the number of samples. Significance is assessed at P-value < 0.05.

II.5.4 A likelihood ratio test of two rates of substitution against one

We sought to test the hypothesis that the phosphorylation site evolved at a slower rate than its flanking region. To do so, we compared that hypothesis against the null hypothesis that the whole region (site and flank) evolved at a constant rate using a likelihood ratio test. Formally:

$$\log \mathrm{LR} = \log \frac{f(x \mid \lambda)}{f(x_{\mathrm{flanks}} \mid \lambda_{\mathrm{flanks}}) f(x_{\mathrm{site}} \mid \lambda_{\mathrm{site}})}$$

Where x is the observed number of substitutions at a given position, and λ is the rate of evolution, and where $\lambda_{flank} >= \lambda_{site}$. We assumed that substitutions occurred following a Poisson process.

$$f(x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Assuming a single *Ks* rate over the whole region, the likelihood ratio only depends on the amino acid substitution rate, *Ka*. The maximum likelihood estimate of *Ka* is simply the number of nonsynonymous substitutions divided by the number of nonsynonymous sites.

After some algebra, the log likelihood ratio is:

$$\log LR = S_{flank} \log \frac{\lambda}{(\lambda_{flank})} + S_{site} \log \frac{\lambda}{(\lambda_{site})}$$

Where S is the number of nonsynonymous substitutions and λ is the maximum likelihood estimate of the rate of amino acid substitution (*Ka*).

The LRT statistic is given by

$$LRT = -2 \log LR$$

Under the null hypothesis, this statistic follows the X^2 distribution with degrees of freedom equal to 1. Significance is assessed at P-value < 0.05.

II.5.5 Structure

To assess if a site was present in a structured or unstructured region, we used the method presented by Uvarsky et al.(Uversky et al. 2000; Prilusky et al. 2005), first by removing the annotated phosphorylation site from the protein segment, and then using a window of 50 amino acids centered on the region of the phosphorylation site.

II.5.6 Turnover rate calculation

Turnover rate is defined by both death and birth rate. We defined death rate as the number of sites disappearing from the inferred ancestral sequence divided by the number of initial sites in the ancestral sequence, and we defined birth rate as the number of sites appearing along the lineage leading to *S. cerevisiae* after the divergence with *S. mikatae* or *S. paradoxus,* per thousand residues of the ancestral sequence.

Significance of birth and death rate differences were assessed using a two-tailed Fisher's exact test by summing the probability of more extreme possible observations. A P-value < 0.05 was assessed as significant.

Site compensation was defined as a pair of nonconserved phosphorylation site and birth within the same lineage, within a local region of the protein. The distance allowed for the site birth was halfway until the next predicted site within *S. cerevisiae*.

II.5.7 Site and target prediction

For various controls and kinase target prediction, we predicted phosphorylation sites using a profile Hidden Markov Model (HMM) obtained from our initial alignments of sites. Profile HMMs have been used in the past to predict protein domains and model linear states that approximate a consensus sequence (eg. Pfam (Finn et al. 2008)). While this may not be needed for kinases such as Mec1p, which follow a strict consensus sequence, other kinases such as cyclin-dependent kinase (CDK) have "weak" and "strong" consensus matches that offer more leeway in their recognition signal. Because HMMbuild (Eddy 1998) was found to be more reliable for longer sequences than most phosphorylation recognition signal, we built a similar model using a single Dirichlet prior that fitted most with one of the Dirichlet mixture for pseudocount (Sjölander et al. 1996). We then used the posterior algorithm and a threshold that validated most of our annotated sites to predict putative sites.

For CDK, kinase target prediction was assessed using the proteins identified as substrates by Ubersax et al. (Ubersax et al. 2003) as positives and the remaining proteins tested by Ubsersax et al. as negatives (nontargets). We calculated the positive predictive power by

counting the number of positive proteins above a LRT threshold divided by the total number of proteins above the same threshold among the two sets. For Mec1, kinase target prediction was assessed using the proteins with at least one characterized Mec1 phosphorylation site as positives and all proteins with characterized phosphorylation sites for other kinases, but not Mec1, as negatives. Although some of these proteins may indeed be targets of Mec1, we hoped that by using this set as negatives, we would reduce the effect of the bias that is induced by researchers when they choose which proteins to study.

II.6 Acknowledgements and funding information

AMM and ANNB are supported by NSERC. AMM is supported by an infrastructure grant from the CFI. We thank Dr. Nicholas Provart for providing us with a web server to store our data. We also thank Dr. Philip M. Kim for comments on the manuscript.

II.7 Author contribution

ANNB designed and performed the experiments and wrote the paper. AMM designed the experiments and wrote the paper.

II.8 Supplementary Data

Supplementary figures are included in Appendix I, and supplementary tables are included in the supplementary data file.

Chapter III

Proteome-Wide Discovery of Evolutionary Conserved Sequences in Disordered Protein Regions

This is an author-produced PDF of an article accepted for publication in Science Signaling following peer review. The version of record: Proteome-Wide Discovery of Evolutionary Conserved Sequences in Disordered Regions
Sci Signal. 2012 Mar 13;5(215):rs1. doi: 10.1126/scisignal.2002515.
Alex N Nguyen Ba[1,2], Brian J Yeh[3], Dewald van Dyk[4,5], Alan R Davidson[6], Brenda J Andrews[4,5], Eric L Weiss[3], Alan M Moses[1,2,¥]
is available online at: http://stke.sciencemag.org/content/5/215/rs1.abstract

1. Department of Cell & Systems Biology, University of Toronto, Toronto, Canada
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Canada
3. Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, United States of America.
4. The Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada;
5. Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada
6. Department of Biochemistry, University of Toronto, Toronto, Canada

III.1 Abstract

At least 30% of human proteins are thought to contain intrinsically disordered regions, which lack stable structural conformation. Despite lacking enzymatic functions and having few protein domains, disordered regions are functionally important for protein regulation and contain short linear motifs (short peptide sequences involved in protein-protein interactions), but in most disordered regions, the functional amino acid residues remain unknown. We searched for evolutionarily conserved sequences within disordered regions according to the hypothesis that conservation would indicate functional residues. Using a phylogenetic hidden Markov model (phylo-HMM), we made accurate, specific predictions of functional elements in disordered regions even when these elements are only two or three amino acids long. Among the conserved sequences that we identified were previously known and newly identified short linear motifs, and we experimentally verified key examples, including a motif that may mediate interaction between protein kinase Cbk1 and its substrates. We also observed that hub proteins, which interact with many partners in a protein interaction network, are highly enriched in these conserved sequences. Our analysis enabled the systematic identification of the functional residues in disordered regions and suggested that at least 5% of amino acids in disordered regions are important for function.

III.2 Introduction

Intrinsically disordered regions are regions that lack stable secondary or tertiary conformation, and 30% of the human proteins are thought to contain large contiguous disordered regions (J J Ward et al. 2004). These regions are found in many disease-associated proteins, such as the tumor suppressor and transcriptional regulator p53, the DNA repair protein BRCA1, and the chloride channel cystic fibrosis transmembrane conductance regulator (CFTR) (Ostedgaard et al. 2000; Uversky et al. 2008; Wells et al. 2008). Although some of these regions contain recognizable domains or become ordered upon binding (Wright and Dyson 1999; Fong et al. 2009), most of these regions apparently lack enzymatic activity or conserved protein domains that adopt regular structures (Sigalov et al. 2007). Several models have been proposed for their function, including that they are important for (i) protein-protein interactions (Dunker et al. 2001), (ii) protein degradation (Brocca et al. 2009), or (iii) posttranslational modifications that control protein function (Dunker et al. 2002). Indeed, disordered (or unstructured) regions are particularly prevalent in proteins that exhibit many physical interactions (Dunker et al. 2005) and have been associated with the sites of posttranslational modifications (Iakoucheva et al. 2004) [reviewed in (Dyson and Wright 2005)]. Despite the importance of these disordered regions, it is currently difficult to accurately identify which residues within a disordered region might be important.

Many of the proposed functions of disordered regions are mediated by short linear motifs (Gould et al. 2010), which are specific peptides of 2-10 amino acid that physically contact modifying enzymes or binding partners. We tested whether we could systematically identify short linear motifs in disordered regions by using the guiding principle of "comparative genomics" - that critical functional sequences would be preferentially preserved over evolution (Ureta-Vidal et al. 2003; Beltrao and Serrano 2005). One approach to systematically identifying short linear motifs is to combine *in vitro* peptide binding data, protein interaction data, and bioinformatic searches (Linding et al. 2007; Tonikian et al. 2009; Mok et al. 2010). Another approach is to search for matches to a motif pattern derived from sets of co-regulated proteins (Neduva et al. 2005; Lieber et al. 2010). Despite their wide applicability, many of these systematic approaches cannot provide evidence regarding the

functional importance of a particular short peptide *in vivo*. On the other hand, the comparative genomics approach can provide evidence that a particular short sequence is important to the organism. Comparative approaches that use only evolutionary conservation are unbiased in that they do not require information about protein function or whether the short linear motif has been previously associated with a specific function. This is in contrast to other approaches (Neduva et al. 2005; Linding et al. 2007; Tonikian et al. 2009; Lieber et al. 2010; Mok et al. 2010) that take advantage of high-throughput *in vitro* and *in vivo* experimental information.

We applied a comparative genomic approach based on a phylogenetic hidden Markov model (phylo-HMM; (Siepel et al. 2005)) to identify short protein sequences in the proteome of the yeast *Saccharomyces cerevisiae*. The phylo-HMM approach has been used previously to discover conserved elements in DNA (Siepel et al. 2005) by exploiting the pattern of nucleotide substitutions. We modified this phylo-HMM approach to include the pattern of insertion and deletion events, as well as substitutions, within a protein sequence, and with this method, we identified on average 1.44 short sequences per protein that were highly conserved and found within intrinsically disordered regions – these included 30% of previously identified short linear motifs in disordered regions.

When our highly conserved sequences matched known consensuses of short linear motifs, such as the FG motif for interaction with karyopherins, the cyclin-dependent kinase (CDK) consensus phosphorylation site motif, and the KEN box for ubiquitin-mediated protein degradation, we found statistically significant enrichment of proteins known to be regulated by these short linear motifs. We experimentally verified a previously unknown KEN box in the yeast protein Spt21. Furthermore, unsupervised clustering of our conserved sequences on the basis of sequence similarity identified hundreds of motif clusters, many of which were enriched for functional annotations. Of the top clusters we examined, about 60% corresponded to known patterns of short linear motifs, whereas the others represent putative newly identified patterns. We identified one such cluster that was enriched for interacting proteins of the kinase Cbk1, which is a member of the nuclear dumbbell forming 2 (Dbf2)-related (NDR) subfamily of the large tumor suppressor (LATS) family of kinases, and showed that the predicted motif mediated a physical interaction with that kinase. Finally, we

analyzed hub proteins and showed that they contain a higher density of short conserved sequences when compared to the rest of the genome, suggesting that their centrality in protein interaction networks may be facilitated by an overabundance of short linear motifs.

III.3 Results

III.3.1 A phylo-HMM approach can identify short conserved sequences in proteins

Posttranslational regulation of protein activity is often mediated through short linear motifs that are often present within disordered regions (Linding et al. 2003; Iakoucheva et al. 2004). Although these motifs share a common pattern or consensus that is important for their function, they are frequently short and may contain positions that have highly flexible amino acid preference. Thus, pattern matches are expected to occur frequently in random protein sequences, with most matches not corresponding to biologically relevant motifs. It has been suggested that correspondence with biological function can be improved by searching for motifs that are also conserved over evolution (Chica et al. 2008; Nguyen Ba and Moses 2010). For our analysis of the *S. cerevisiae* proteome, we chose related species that have syntenic gene orthologs and are thought to have diverged 100 million to 200 million years ago (Gordon et al. 2009).

We developed a phylo-HMM-based computational framework to systematically detect conserved short linear motifs in unstructured regions in multiple sequence alignments (Figure III-1A). We hypothesize that functionally important short linear motifs will be preferentially conserved such that substitutions and insertions or deletions will occur more frequently adjacent to the motif than within it (Ren et al. 2008). It therefore follows that the amino acids in each multiple sequence alignment column fall into two classes: the conserved class and the background class. The conserved class represents those amino acids with a slow rate of evolution corresponding to the preferentially conserved motif (Figure III-1B; Appendix Figure II-1 “Conserved”, rate = α_c), and the background class represents those with a faster rate of evolution, corresponding to divergent, functionally less important

sequences (Figure III-1B; Appendix Figure II-1 "Background", rate = α_W). We compared the substitution and insertion or deletion rate in each column with the overall rate in a window of surrounding amino acids (Figure III-1B). We then used a statistical approach based on a phylo-HMM to compute the probability (the posterior probability) that each multiple sequence alignment column (Figure III-1B, framed in green) is within the preferentially conserved class. The posterior probability approaches 1 as segments increase in relative conservation or as the number of consecutively conserved residues increases. When the phylo-HMM approach was previously applied to analyze DNA conservation, only substitutions were considered. Because insertion and deletion events are common in disordered regions, we have modified the phylo-HMM approach to include these events as well (Figure III-1B, vertical black bars separating grey highlights, see Methods).

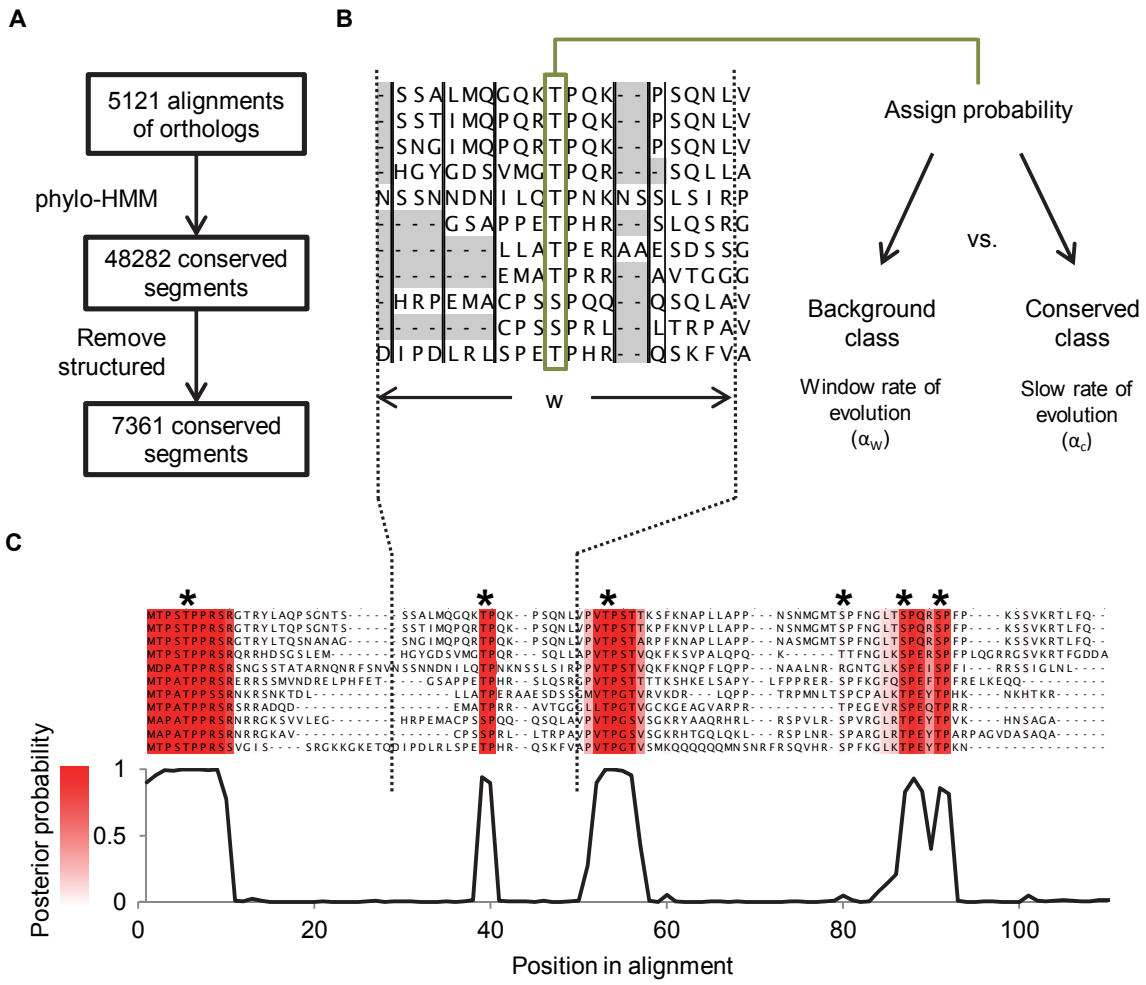


Figure III-1. Schematic of the phylo-HMM approach. A) Flowchart of the computational framework to detect conserved short linear motifs in disordered regions of multiple sequence alignments. B) The rate of evolution for an alignment column (framed in green) is compared to a rate of evolution over a window (w) adjacent to the column. The probability that the column is within the preferentially conserved class is computed. The framework takes advantage of the amino acid substitutions inferred in columns of the alignment and the pattern of insertions and deletions (illustrated as grey highlights) in blocks of the multiple sequence alignment (separated by vertical black lines). C) A posterior probability trace of the region 1 to 110 in the alignment of Sic1 (corresponding to amino acid position 1 to 100 in *S.cerevisiae*). Four strongly conserved segments are detected by the phylo-HMM approach and these overlap with experimentally reported phosphorylation sites in Sic1 (indicated by stars), which are required for Cdc4 binding. The intensity of the red color represents the posterior probability of the conserved state.

To illustrate this method, we plotted the posterior probability as a function of alignment position (a "probability trace") for the disordered N terminus of Sic1, which contains experimentally verified phosphorylation sites necessary for binding to the E3 ubiquitin ligase adaptor protein Cdc4 (Nash et al. 2001; Mittag et al. 2008). The probability trace showed clear and specific peaks in the N terminus of Sic1 (Figure III-1C) and these peaks corresponded to five of the six known phosphorylation sites (Nash et al. 2001). In sequences lacking known motifs, such as a segment of the transcription factor Swi5, the posterior trace often was flat, despite variation in the local rate of protein evolution (Appendix Figure II-2).

III.3.2 Short conserved sequences predicted by the phylo-HMM contain known motifs

Using the phylo-HMM, we performed a proteome-wide prediction of short conserved sequences in *S. cerevisiae* and identified on average 1.44 short conserved sequences passing our threshold per protein (see Materials and Methods, Figure III-1A, Supplementary data table III-S1). To assess whether these short conserved sequences were biologically relevant, we analyzed a set of 352 literature-curated short linear motifs found in disordered regions (Supplementary data table III-S2, see Materials and Methods for criteria). Although the phylo-HMM predicted short conserved sequences for only ~5% of residues in disordered regions, 104 (30%) of the literature-curated short linear motifs were among the predictions.

We searched our conserved sequences for matches to known patterns of short linear motifs. In an in-vitro kinase assay, of the 695 proteins with at least one Cdc28 phosphorylation site matching the consensus sequence ([ST]Px[RK]), only 185 were phosphorylated by an analog-sensitive mutant of Cdc28, a CDK (Ubersax et al. 2003). (Note that in motif sequences, letters in brackets represent preferred residues for a particular position, and x represents any amino acids.) Thus, simply having a consensus phosphorylation site is not sufficient to predict Cdc28 substrates. Our phylo-HMM identified 40 proteins containing a short conserved sequence that matched the Cdc28 consensus pattern, and 32 of these were positive in the *in vitro* kinase assay (Ubersax et al. 2003), which is a significant enrichment (32/40 vs. 185/695, P-value = $1.4 * 10^{-11}$, Fisher's test, Supplementary data table III-S3). Of the 8 remaining proteins identified by the phylo-HMM, 1 of those (Cdc15) includes consensus sites phosphorylated *in vivo* (Jaspersen and Morgan 2000), and 2 are targets of

kinases that can phosphorylate the canonical Cdc28 consensus sequence - Rim15 phosphorylated by Pho85 (Wanke et al. 2005) and Fus2 phosphorylated by Fus3 (Ydenberg and Rose 2009). Thus, 80% of the proteins identified by the phylo-HMM as containing conserved sequences matching the canonical Cdc28 consensus pattern are likely to be substrates of this kinase or other kinases that recognize the same or similar consensus sequences.

The FG motif pattern (Phe-Gly), which is a canonical motif of common in nuclear pore complex (NPC) proteins and may be important for trafficking of proteins through the nuclear pore (Denning et al. 2003), is found in unstructured regions of these proteins (Denning et al. 2003). Thirteen components of the NPC have been reported to contain FG repeats (Denning et al. 2003), seven of which can be further classified into variants including the FxFG and GLFG motifs (Denning et al. 2003). Using the phylo-HMM, we found 59 proteins in the yeast proteome with at least one conserved FG dipeptide. These included 12 of the previously known FG-containing NPC proteins. Because FG consensus matches are found in 3438 yeast proteins, this is an extremely significant enrichment (12/59 vs. 13/3438, P-value = $7.21 * 10^{-16}$, Fisher's test, supplementary table III-S3). Searches in the yeast proteome for the more specific variants (FxFG and GLFG) yielded six of seven nucleoporins that contain these variant FG motifs.

Of the 59 proteins identified as having a conserved FG dipepetide by phylo-HMM, one of these was Ndc1, which is localized to the nuclear envelope and required for nuclear pore assembly (Chial et al. 1998; Lau et al. 2004), but had not previously been recognized as having an FG motif. The remaining 46 proteins identified by phylo-HMM analysis are not components of the NPC but nevertheless contain a short conserved sequence that matches the minimal FG motif pattern. Because the motif occurred either within known repeat sequences or in proteins that have roles in protein transport and sorting, we believe that the conserved sequences containing an FG dipeptide in the remaining 46 proteins are likely functional. For example, we identified both Sla1 and Pan1, members of the actin cytoskeleton-regulatory complex, as having the FG motif, and the motif in Sla1 is within the functionally important C-terminal repeat region (Warren et al. 2002). Other proteins related to protein transport and sorting that we identified as having the conserved FG dipeptide included Vps15, Ede1, Ent3,

Ent5, Pga2, and Glo3. Thus, rather than being limited to nuclear transport, the FG dipeptide motif may function more broadly in protein transport.

We also identified proteins in the *S. cerevisiae* proteome with a conserved KEN box [a degradation signal that is recognized by the anaphase-promoting complex/cyclosome (APC/C)] (Buschhorn and Peters 2006). The KEN box acts as a binding site for the APC/C and marks target proteins for degradation in different phases of the cell cycle. The phylo-HMM analysis identified only 10 proteins with a conserved KEN sequence. Eight of those contained an experimentally verified KEN degradation signal (King et al. 2007; Choi et al. 2008), were characterized targets of the APC/C (Cohen-Fix et al. 1996; Juang et al. 1997), or were cyclins, including Clb2 which contains a verified KEN sequence (Hendrickson et al. 2001). The two remaining motifs matching the KEN signal are found in Spt21 and Sgd1, neither of which have been associated with the APC or reported to exhibit cell cycle-regulated degradation (supplementary data table III-S3). We noticed that the conserved KEN box in Spt21 was followed by a conserved proline, which is also found conserved following the KEN motif in Clb2 (Figure III-2A) and Mad3 (King et al. 2007), suggesting that the proline may confer additional binding specificity beyond the KEN residues. The presence of a proline after the KEN motif has been reported to mediate more efficient APC/C-mediated degradation of mammalian proteins with KEN boxes (Feine et al. 2007).

To confirm the in silico analyses, we experimentally tested whether the identified KEN sequence in Spt21 served as a degradation signal (Figure III-2A). Spt21 is a protein that promotes transcription of the genes encoding the HTA2 and HTB2 histones, and transcription of the gene encoding Spt21 is cell cycle-regulated (Chang and Winston 2011). We found that the amount of Spt21 coincided with the amount of Clb2, a protein that exhibits changes in abundance during the cell cycle, (Figure III-2B), which indicated that, as at the level of mRNA (Spellman et al. 1998), Spt21 protein abundance varied over the cell cycle.

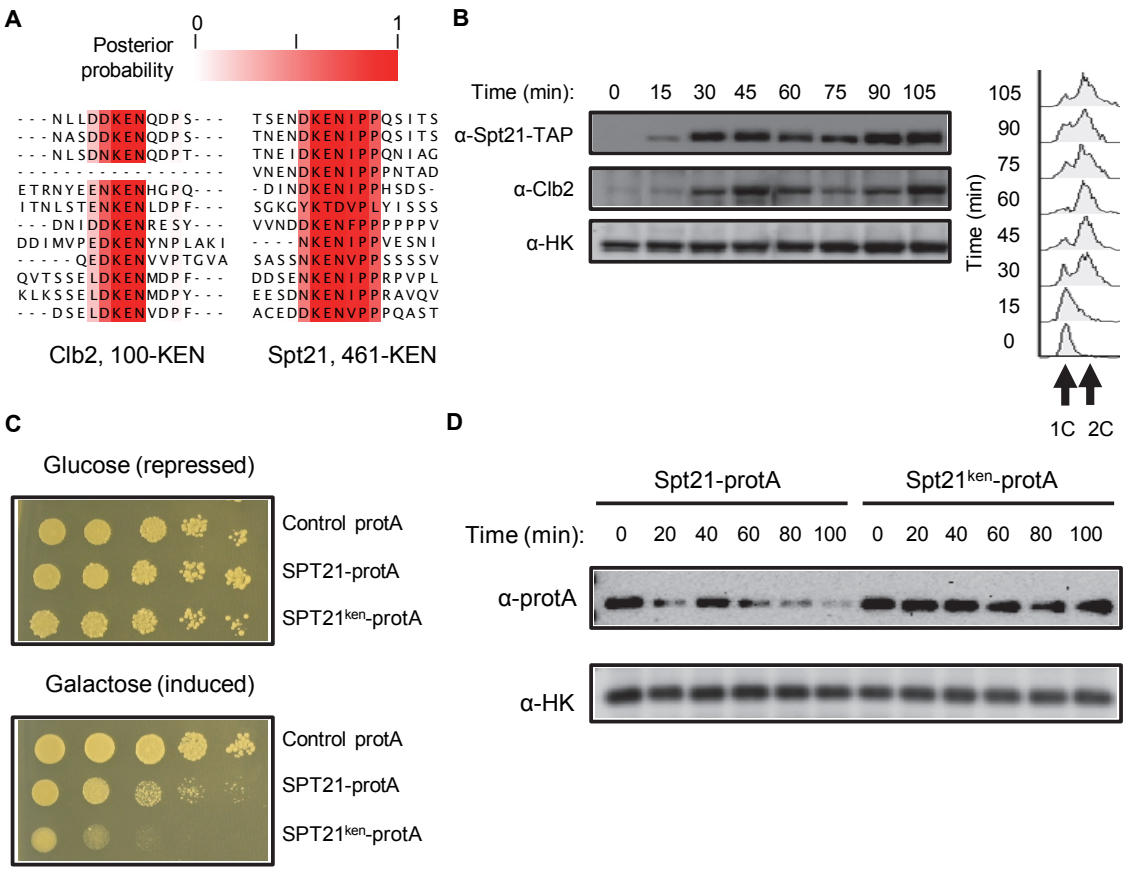


Figure III-2. A KEN box identified by the phylo-HMM approach in Spt21 mediates protein degradation. A) Alignments of the previously characterized Clb2 KEN box degradation signal alongside the predicted KEN box in Spt21. Numbers indicate residue position within the *S. cerevisiae* protein sequence. B) Left panel is a Western blot showing that amount of SPT21-TAP varies within the cell-cycle and, like Clb2, is absent in G_1. Endogenous hexokinase (HK) served as the loading control. Pearson correlation coefficient between the normalized amount of Spt21 and the normalized amount of Clb2 was 0.89 and 0.87 in two independent experiments. Right panel shows FACS analysis as additional validation of cell cycle progression. C) Spotted serial dilutions of strains overexpressing protA-tagged wild-type (*SPT21*) or KEN box mutant (*SPT21*^{ken}) show a stronger fitness defect with overexpression of the KEN box mutant. D) Western blot analysis shows that mutation of the KEN box stabilizes Spt21. Wild-type (*SPT21*) and KEN box mutant (*SPT21*^{ken}) expression was induced in galactose medium for 4 hours. Glucose was added to attenuate protein expression, and protein synthesis was abolished through the addition of cyclohexamide. HK was used as a loading control. Results shown are a representative blot from three independent experiments (two with protA-tag and one with the GST-tag). The time-points after 60 minutes had P-values < 0.05 (*t* test; n=3) when comparing the normalized abundance of the wild-type to the KEN box mutant.

Many proteins have multiple means of regulation, and degradation by the APC/C may act as an additional layer of control, especially because overexpression of Spt21 is deleterious (Sopko et al. 2006). Given this cell cycle regulation and the toxicity of overexpression, we reasoned that if the KEN sequence is a biologically relevant degradation signal, then overexpression of a KEN mutant form of Spt21 would be more toxic than a wild-type form. We mutated the three consecutive KEN amino acids to alanines ($Spt21^{ken}$) and performed serial spot dilution assay to assess growth fitness. Growth was more severely impaired by $Spt21^{ken}$ overexpression than by overexpression of the corresponding Spt21 control (Figure III-2C). To confirm that the KEN box served as a degradation signal, we assayed changes in protein abundance of Spt21 and $Spt21^{ken}$ through the cell cycle by overexpressing the proteins with the GAL promoter followed by shutting off both transcription and translation (see Materials and Methods). The abundance of the KEN mutant form remained high, whereas the abundance of wild-type Spt1 decreased over time (Figure III-2D). These results suggested that the conserved KEN sequence in Spt21 is important for the cell cycle-dependent degradation of this protein.

The evaluation of the KEN box, FG motif, and Cdc28 phosphorylation consensus sites provided evidence that the phylo-HMM approach can predict biologically relevant, short conserved sequences. However, it is possible that the many of the remaining predicted motifs in the yeast proteome were identified by the phylo-HMM because they have not sufficiently diverged, or because alignment errors lead to overestimation of the conservation of residues. To address the possibility of these computational artifacts, we performed extensive simulations of protein evolution (see Chapter III, Materials and Methods), which indicated that such artifacts occurred in alignments of disordered regions at a rate of 1 in 9000 amino acids (fewer than 1 in every 50 proteins examined). Another possible source of error in our classification of disordered regions may be the inclusion of larger protein domains within our disordered regions. However, 63% of the predicted short conserved sequences are within regions of at least 50 disordered amino acids, which are unlikely to be protein domains. Along with the strong enrichment of functional Cdc28 consensus sites, FG motifs, and KEN boxes, this low rate of computational artifacts indicated that short conserved sequences identified by the phylo-HMM likely represent functional elements within unstructured regions.

III.3.3 Known and previously unknown sequence patterns are uncovered by clustering the short conserved segments by sequence similarity

We found that many proteins contained short conserved segments that did not match any known sequence patterns, and, thus, these may represent previously unknown short linear motifs. Conservation in distantly related species would support the biological relevance of these previously unknown motifs and would indicate that these are not computational artifacts and are biologically important.

For example, we found a previously unknown motif in the C terminus of the Dbp6 putative adenosine triphosphate (ATP)-dependent DEAD box RNA helicase (Kressler et al. 1998) that is conserved in plants, yeast, and human (Figure III-3A). If this short conserved sequence is part of biologically relevant, previously unidentified motif pattern, we reasoned that similar short conserved sequences should also be found in other proteins, possibly with shared functions. Dbp6 is required for ribosome biogenesis, and we identified a similar highly conserved short segment in the yeast protein Utp25, which is a DEAD box RNA helicase-like protein also related to ribosome biogenesis (Charette and Baserga 2010). These sequences all match the pattern YxxxLxxL, and the motif is conserved in distant orthologs for these proteins (Figure III-3B); therefore, we speculate that YxxxLxxL may represent an essential short linear motif pattern found in the unstructured regions of proteins involved in ribosome biogenesis.

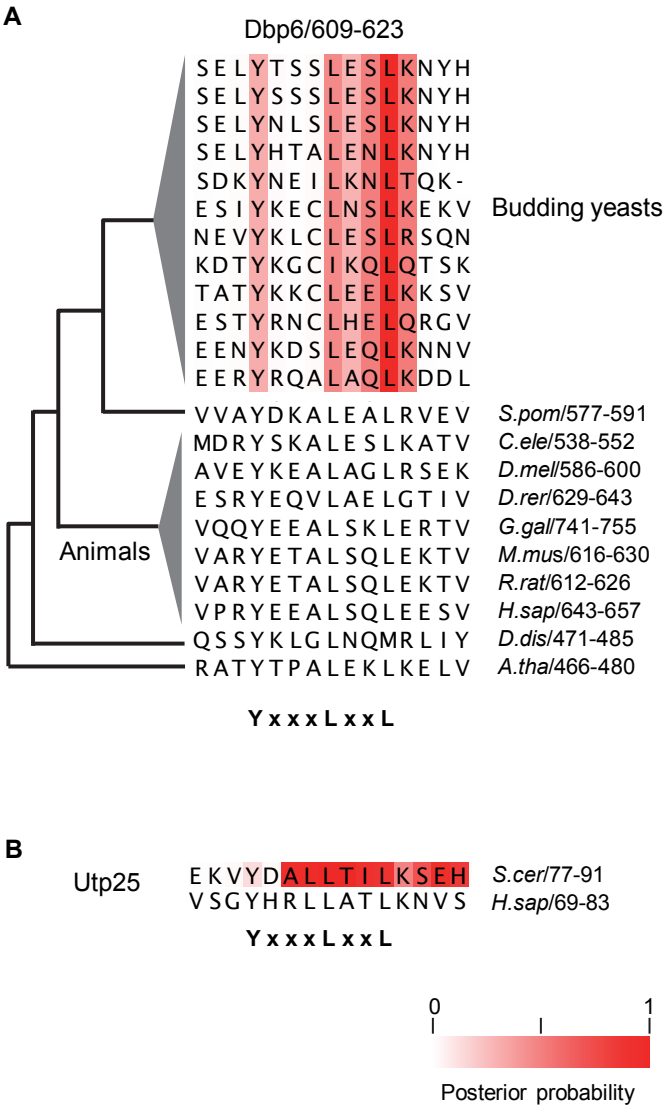


Figure III-3. Predicted motifs are conserved in distant species. A) Alignment region of the predicted YxxxLxxL motif in Dbp6 shows conservation of the motif amongst eukaryotic orthologs. Distant species comparison is shown with a phylogenetic tree. Branch lengths are not to scale. B) Alignment of the predicted YxxxLxxL motif in Utp25 with its human ortholog.

To determine whether other previously unknown patterns were identifiable in our data set, we used an unsupervised graph-clustering algorithm [MCODE (Bader and Hogue 2003)] to group conserved sequences into motif patterns on the basis of their sequence similarity without regard as to which protein contained these motifs (see Materials and Methods). This type of analysis can be visualized by a graph in which conserved sequences are represented

as nodes, edges correspond to sequence distance, and groups of highly interconnected nodes (detected by the graph-clustering algorithm) correspond to motif patterns (Figure III-4A).

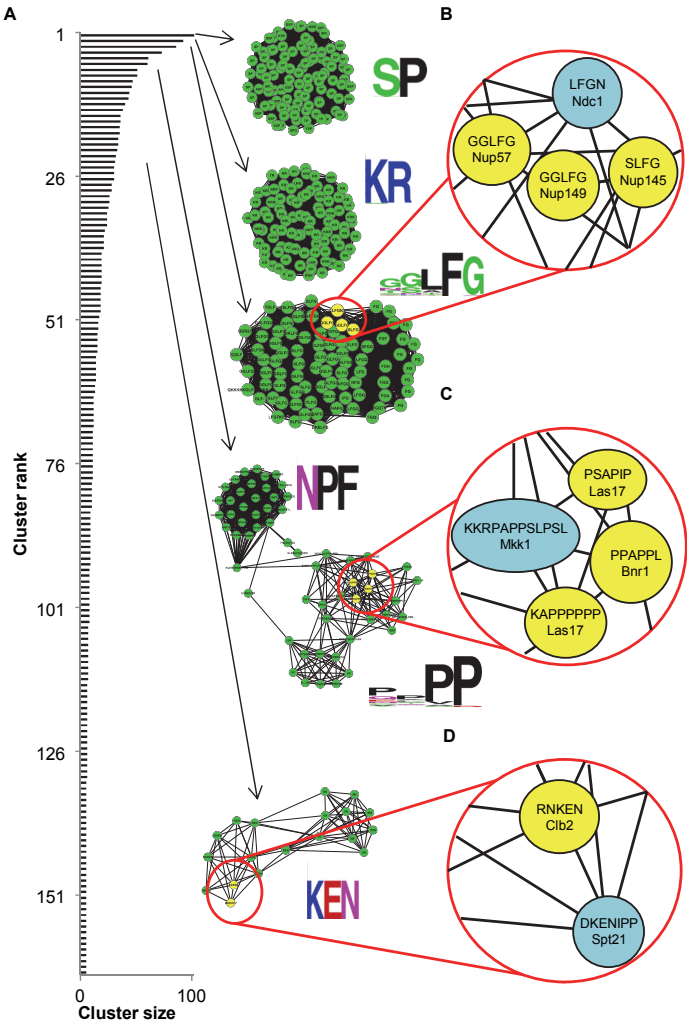


Figure III-4. Known short linear motif patterns are recovered by cluster analysis. A) Distribution of cluster sizes (black bars) and examples of highly interconnected motifs identified in the cluster analysis that represent known sequence patterns (green clusters). Beside each cluster is a graphical representation of the specificity of the motif. B) A close-up representation of the FG motif cluster shows interconnection between known FG motifs in nuclear pore proteins (highlighted in yellow) and a uncharacterized FG motif in Ndc1 (highlighted in blue) . C) A close-up representation of the P-rich motif cluster shows interconnection between a putative SH3-binding peptide (in Mkk1, highlighted in blue) and known SH3-binding peptides (in Las17 and Bnr1, highlighted in yellow). D) A close-up representation of the KEN motif cluster shows connection between the previously uncharacterized KEN motif in Spt21 (highlighted in blue) and an experimentally verified KEN motif in Clb2 (highlighted in yellow). See Supplementary data table III-S4 for a complete list of the proteins identified in each highlighted cluster.

For one set of clustering parameters, this procedure yielded 282 clusters covering 41% of the predicted sequences, with 38 large clusters containing at least 20 short conserved sequences, representing 21% of the predicted sequences, and 45 smaller clusters containing between 10 and 20 conserved sequences, each representing ~9% of the predicted sequences (Figure III-4, Supplementary data table III-S4 and Supplementary data table III-S5). As expected, this uncovered previously described consensus sequences for short linear motifs, such as an SP/TP cluster (proline-directed kinase consensus), a GLFG cluster, and a KEN cluster (Figure III-4). These three motifs corresponded to the patterns described above, and the proteins containing these motifs were enriched in the expected function (see Materials and Methods). For example, the GLFG cluster was enriched in proteins having a nuclear pore subcellular localization (9 nuclear pore localized proteins/16 proteins in cluster vs. 46 nuclear pore localized proteins/5884 proteins in the yeast proteome, P-value = $2.9 * 10^{-15}$, Fisher's test), whereas the proteins in the SP cluster were enriched for cell cycle process (32/88 vs. 520/5884, P-value = $2.4 * 10^{-12}$, Fisher's test). The SP cluster was the largest identified in our analysis (Figure III-4A, Supplementary data table III-S4), likely containing phosphorylation sites for many different proline-directed kinases (including the cell-cycle kinases Cdc28 and Pho85), which suggested that the most frequently observed conserved short sequences in disordered regions in yeast are consensus phosphorylation sites.

Other clusters matching known consensus sequences included the NPF cluster, a motif found in EH domain interacting proteins (de Beer et al. 2000), which was enriched in endocytosis-related proteins (7/20 vs. 59/5884, P-value = $1.09 * 10^{-9}$, Fisher's test), a KR cluster, which is a signature of nuclear localization signals (Lange et al. 2007; Nguyen Ba et al. 2009) and was enriched in proteins identified in the nuclear compartment (70/88 vs. 2077/5884, P-value = $4.3 * 10^{-17}$, Fisher's test), and a cluster of proline-rich sequences that resemble binding sites for peptide-binding domains, such as SH3 (Src homology 3) and WW (Macias et al. 2002). This cluster contained known SH3-binding proteins, such as Las17 (Rodal et al. 2003), and predicted the presence of an uncharacterized proline-rich binding site in the mitogen-activated protein kinase kinase Mkk1.

We repeated the cluster analysis with different parameter settings (see Materials and Methods, Supplementary data table III-S4 and Appendix table II-5) and searched for clusters

representing motif patterns that to our knowledge are uncharacterized, but had strong enrichment in functional annotations (Figure III-5A-C). With this analysis, we identified an NPY cluster, which may be related to the NPF motif and was enriched in vesicle and nuclear membrane proteins and enriched in proteins associated with protein transport process (7/12 vs. 419/5884, P-value = $5.64 * 10^{-6}$, Fisher's test). We also identified an FxDSF[RK]R motif, which was present in many amino acid permeases (6/8 vs. 36/5884, P-value = $2.5 * 10^{-12}$, Fisher's test), and those permeases that contained this motif also had a C-terminal palmitoylation motif, FWC (Roth et al. 2006). Finally, we identified a [YF][KQ]FP motif (also referred to as FxFP), which was found in Cbk1-interacting proteins (Ho et al. 2002; Breitkreutz et al. 2010) (4/6 vs. 27/5884, P-value = $9.4 * 10^{-9}$, Fisher's test).

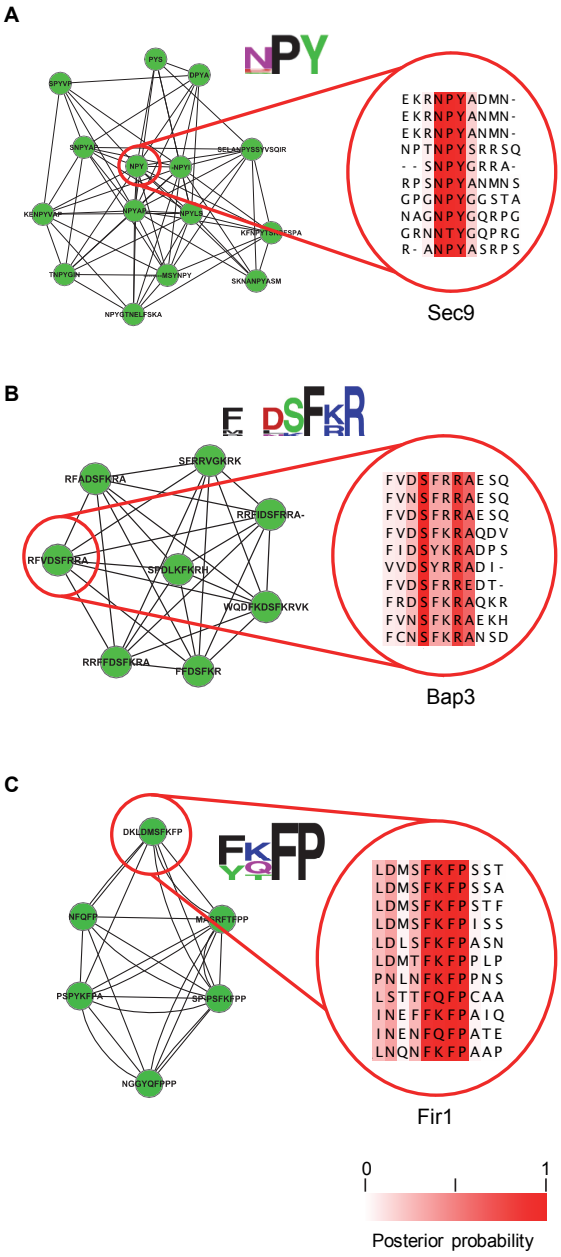


Figure III-5. Previously unknown short linear motif patterns are predicted by cluster analysis. This figure shows representative examples of highly interconnected motifs that represent uncharacterized sequence patterns. A) The NPY motif cluster consists of proteins enriched in vesicle and nuclear membrane proteins related to protein transport. We show an example aligned segment from Sec9 at position 231 to 234. B) The FxDSF[KR]R motif cluster consists of proteins enriched in amino acid permease function. We show an example aligned segment from Bap3 at position 56 to 61. C) The FxFP motif cluster consists of proteins enriched in Cbk1 kinase targets. We show an example aligned segment from Fir1 at position 416 to 419. See Supplementary data table III-S4 for a complete list of the proteins identified in each cluster.

Of these uncharacterized putative consensus sequences, we focused on the [YF][KQ]FP motif (Figure III-5C). This cluster was enriched for proteins that interact with the kinase Cbk1 (Ho et al. 2002; Breitkreutz et al. 2010) and contained two known Cbk1 substrates, Ssd1 and Ace2 (Mazanka et al. 2008; Jansen et al. 2009), Table 1. The [YF][KQ]FP pattern is not similar to the known Cbk1 phosphorylation site consensus (Mazanka et al. 2008) but is similar to the reported kinase docking motif (FxFP) for the extracellular signal-regulated kinases (ERKs) in mammals (Jacobs et al. 1999). This docking motif facilitates kinase-substrate recognition by specific binding of the substrate to a docking site on the kinase domain that is distinct from the catalytic site (Reményi et al. 2006). Therefore, we hypothesized that this motif was important for the physical interaction of the kinase with its substrates. To test this, we fused fragments containing the conserved sequences to maltose-binding protein (MBP) and assayed binding to Cbk1 in a pull-down assay (see Materials and Methods). We detected reproducible binding with five of six tested peptides (Figure III-6), indicating that that the peptide fragments containing the newly identified [YF][KQ]FP motif interacted with Cbk1.

Table III-1. Members of the FxFP cluster. Unsupervised clustering of the conserved sequences revealed a cluster enriched for Cbk1 interactors and contained two known Cbk1 kinase targets (underlined).

ORF	Gene Name	Start	Stop	Sequence
YLR131C	ACE2	280	288	NGGYQFPPP
YNL042W	BOP3	152	159	PSPYKFPA
YER075C	PTP3	371	375	NFQFP
YDR293C	SSD1	231	239	SPPSFKFPP
YER032W	FIR1	410	419	DKLDMSFKFP
YIL129C	TAO3	1	9	MASRFTFPP

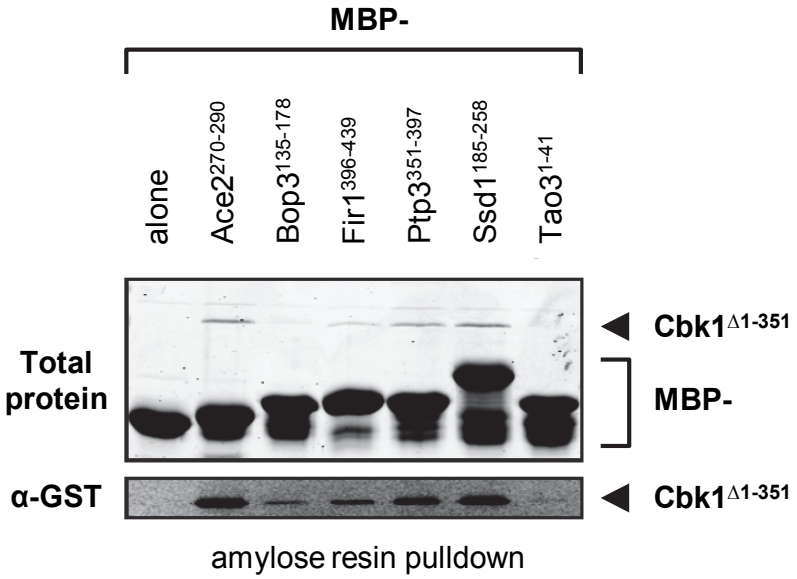


Figure III-6. [YF][KQ]FP peptides interact with the Cbk1 kinase domain. Fragments from proteins identified in the [YF][KQ]FP cluster were expressed as MBP fusions and immobilized on amylose resin. The beads were assayed for binding to GST-tagged Cbk1 ($Cbk1^{\Delta 1-351}$) in a pulldown assay. Binding was detected by Western blot for all six protein fragments tested, while MBP alone could not pull down Cbk1 (lower panel). Shown is a representative blot from three independent experiments. See Appendix Figure II-6 for a shorter exposure of the blot that shows the loading control.

III.3.4 Protein hubs show higher motif density

One hypothesis for the existence of unstructured regions is that they serve as regulatory hubs where multiple regulatory motifs can act in a concerted way to finely regulate function and interaction (Iakoucheva et al. 2004; Dunker et al. 2005). This model is consistent with the idea that unstructured regions can undergo multiple different transient structural configurations to accommodate the multiple regulatory sequences (Wright and Dyson 1999). Proteome-wide analyses of protein-protein interactions (Uetz et al. 2000; Ito et al. 2001) have revealed a small number of "hub" proteins that interact with many partners (Jeong et al. 2001). Because protein-protein interactions are often mediated by short linear motifs, we analyzed the short conserved sequences in a high-confidence set of hub proteins (Bertin et al. 2007).

Using our definition of unstructured regions, we found, consistent with previous studies (Dunker et al. 2005; Haynes et al. 2006), that hub proteins had significantly more large segments (>=30 amino acids) of disordered amino acids (13% increase, P-value = 0.0009, Poisson distribution, Figure III-7A). Thus, relative to the entire proteome, hub proteins should contain more predicted short linear motifs per protein because they have more disordered regions. Indeed, hub proteins contained significantly more predicted short conserved sequences per protein (46% increase, P-value = $2.8 * 10^{-12}$, Poisson distribution). However, the increase in short conserved sequences was not due only to the fact that hub proteins contained more large segments of disordered amino acids: We found that hub proteins had a significantly higher density of short conserved sequences per amino acid (29% increase in disordered regions of >=30, P-value = $1.83 * 10^{-12}$, Poisson distribution, Figure III-7B), indicating that these short conserved sequences may mediate their high degree of connectivity. Thus, the centrality of hub proteins to interaction networks may, in part, be due to their high prevalence of short linear motifs.

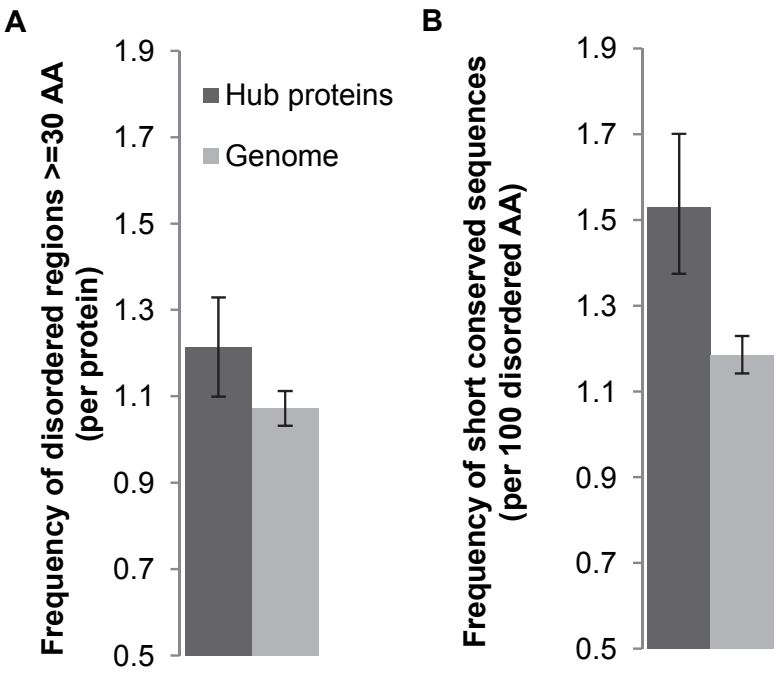


Figure III-7. Hub proteins are enriched in short conserved sequences. A) The frequency of long regions of disordered amino acids (>=30) for hub proteins is higher than for the rest of the genome. B) The number of predicted conserved sequences per amino acid that are present in long regions of disordered amino acids (>=30) is higher for hubs than the genome. Error bars represent the 95% confidence interval obtained by non-parametric bootstrapping.

III.4 Discussion

Although unstructured regions are ubiquitous in eukaryotic proteomes, it is difficult to identify the critical functional residues within them. For example, despite detailed characterization of Utp25 (Charette and Baserga 2010), using the phylo-HMM approach we identified a short sequence in disordered regions of this protein. This sequence was conserved in all eukaryotes but had not been previously characterized. Systematic application of the phylo-HMM approach to the yeast proteome identified on average 1.44 short conserved sequences per proteins, totalling about 5% of the unstructured amino acids. Proteins containing known sequences showed strong functional enrichment, suggesting that the conserved sequences are involved in specific biological functions. Because the false-positive rate was 1 in 9000 unstructured amino acids, we only expect about <1% or 70 of the thousands of identified short sequences to be false positives, resulting from computational

artifacts. However, in many individual cases, we are confident that the identified sequences are important, because the motif is conserved across divergent species, representing a long evolutionary period. For example, the previously uncharacterized KEN motif in Spt21 is conserved in its orthologs within the *Candida* clade; and the Cbk1-interacting motifs in Ssd1 are conserved even further within the Ascomycetes (Appendix Figure II-3). Although many short linear motifs are well conserved, other functional sequence segments may be species-specific, or they may not have been captured by our analysis (for example, the phylo-HMM approach that we used does not detect motifs embedded in large conserved protein fragments because these large regions are excluded from the analysis). Consequently, the short linear motifs predicted in this study only provide a lower bound of the number and frequency of these motifs in unstructured regions. Because 30% of the known characterized short linear motifs in disordered regions in our data set were predicted by the phylo-HMM, and because our phylo-HMM identified short conserved sequences totalling about 5% of the unstructured amino acids, we estimate that short linear motifs correspond to roughly 17% of the unstructured amino acids in yeast proteins.

Our approach for identifying short linear motifs is different than other computational methods designed for this goal (Davey et al. 2010). Two other bioinformatic approaches involve either the classification of matches to a known consensus (Obenauer et al. 2003) or the prediction of a consensus given known co-regulation (Bailey and Elkan 1994; Davey et al. 2006; Neduva and Russell 2006; Edwards et al. 2007; Lieber et al. 2010), both of which rely on previously obtained experimental data. Other structure-based methods, such as ANCHOR (Mészáros et al. 2009), identify disordered regions that have the propensity to become ordered upon binding. Our phylo-HMM approach requires only the underlying evolutionary relationship between genes and that regulatory function is preserved in most of the species considered. Therefore, our study is complementary to previous methods and opens the framework of phylogenetic footprinting (Tagle et al. 1988; Cliften et al. 2003) (a method to identify functional elements in noncoding DNA by exploiting evolutionary conservation) to protein sequences. Because this analysis requires only sequence information from orthologous proteins, it can be applied in many clades for which these data are now available (Kent et al. 2002; Flicek et al. 2011). However, the success of the phylo-HMM approach is directly related to the choice of species and their evolutionary distance.

Computational artifacts increase at short evolutionary distances (Appendix Figure II-4D), whereas biologically relevant motifs may no longer be conserved at the same position at very long evolutionary distances and, therefore, will not be detected (Appendix Figure II-5). In general, the performance of the phylo-HMM approach can be assessed by simulations of molecular evolution where conserved motifs have been inserted and by analysis of previously characterized short linear motifs. Another important issue concerning the performance of the phylo-HMM approach is that the posterior probability output depends on both the length of the conserved segment and on its relative conservation compared with the background evolutionary rate. Therefore, the predictions with the highest posterior probability tend to be longer regions (more than five amino acids), which we speculate may be high-specificity biomolecular binding sites. Equally important short linear motifs are very short (about two amino acids) and will tend to have lower posterior probabilities.

Because our analysis is independent of functional data, it led to the discovery of important elements from the sequence data without attaching any specific function to the results. Although we could propose functions for some previously unknown motif patterns through enrichment analysis for biological processes, in other cases we also observed clusters that matched known sequence patterns but were not present in proteins enriched in the expected function. For example, we identified the well-characterized acidic dileucine ([DE]xxxL[LI]) motif (Supplementary data table III-S5) found in transmembrane proteins of endosomes and lysosomes in metazoans or in yeast vacuolar proteins (Bonifacino and Traub 2003) in one of our clusters. However, the proteins forming this cluster were not significantly enriched for any particular compartment even though it includes the experimentally verified acidic dileucine motif from Vam3, a vacuolar t-SNARE (Darsow et al. 1998). We speculate that the conserved motifs in this cluster likely serve other functions. Even when functional enrichment of a cluster can be found, the function of the motif cannot always be ascertained. Despite these potential difficulties in assigning functional relationships, our unbiased methods (prediction of conserved sequences and the clustering analysis) were successful in discovering an interaction motif for the NDR/LATS kinase member Cbk1. We speculate that some sequence patterns, such as the newly identified motif for Cbk1 interaction, are associated with only one function, whereas others such as the acidic dileucine motif and the FG dipeptide are involved in multiple processes.

Our analysis suggested that intrinsically disordered regions contain large numbers of functional sequences that are involved in protein regulation and interaction, and this may partly explain the prevalence of disordered regions. Consistent with the hypothesis that the functional sequences may contribute to protein interactions, we observed a higher density of predicted short linear motifs in hub proteins, which is consistent with previous reports that the disordered regions in hub proteins are particularly important for their interactions *(11, 66)*. The observation that the increase in conserved sequence density (29%) (Figure III-7b) is greater than the increase in disordered segments (13%) (Figure III-7A) suggests that the conserved sequences identified by the phylo-HMM approach are more indicative of hub functions than the presence of disordered regions alone. We found no differences in the types of conserved motifs in disordered regions of hub proteins when compared to the rest of the genome, indicating that there are no specific "hub motifs", nor any differences in conserved sequence density between "date" vs "party" hubs (Bertin et al. 2007) (date hubs 1.535; party hubs 1.520, per 100 amino acids in disordered regions >=30 amino acids). Instead, these highly connected proteins simply have more functional sequences within their disordered regions than do proteins that are not hubs. Given the importance of protein regulation and interaction to cellular physiology (Gould et al. 2010) and an increasing appreciation of its importance in evolution (Jensen et al. 2006; Moses and Landry 2010), disordered regions seem poised to play a critical role in these contexts.

III.5 Materials and Methods

III.5.1 Alignment of related species of yeasts

Protein sequences from 13 related species of yeasts [*S. cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces mikatae*, *Saccharomyces bayanus*, *Candida glabrata*, *Saccharomyces castellii* (now renamed to *Naumovia castellii*), *Kluyveromyces polysporus* (now renamed to *Vanderwaltozyma polyspora*), *Zygosaccharomyces rouxii*, *Kluyveromyces lactis*, *Ashbya gossypii*, *Kluyveromyces waltii* (now renamed *Lachancea waltii*), *Kluyveromyces thermotolerans* (now renamed *Lachancea thermotolerans*) and *Saccharomyces kluyveri (*now renamed *Lachancea kluyveri)*] were obtained from the SGD (SGD Project 2011) and the Yeast Genome Order Browser (Gordon et al. 2009). These

species were chosen because of the high quality of the sequence information and of the annotation associated with each gene or protein. Orthologous genes were aligned using MAFFT (Katoh et al. 2002) at automatic settings. Branch lengths for the species tree (Gordon et al. 2009) were obtained by 10 replicates of 50 random concatenations of orthologous genes and analyzed using PAML v3.15 (Yang 2007). Analysis showed that the expected substitution per site for these alignments was 3.189. We aligned 5121 proteins from *S. cerevisiae* to at least one of the related species.

Conservation of motifs over more distantly related orthologs was performed with sequences from the Fungal Orthogroup Repository (Wapinski et al. 2007) and the Princeton Protein Orthology Database (Heinicke et al. 2007) or by using BLASTP (Altschul et al. 1997) on the uniref90 database (Suzek et al. 2007). Other species analyzed were *Candida lusitaniae, Debaryomyces hansenii, Candida guillermondii, Candida tropicalis, Candida albicans, Candida parapsilosis, Lodderomyces elongisporus, Pichia stipitis, Yarrowia lipolytica, Uncinocarpus reesii, Aspergillus niger, Penicillium chrysogenum, Sclerotinia sclerotiorum, Schizosaccharomyces pombe, Caenorhabditis elegans, Drosophila melanogaster, Danio rerio, Gallus gallus, Mus musculus, Rattus rattus, Homo sapiens, Dictyostelium discoideum* and *Arabidopsis thaliana*.

III.5.2 Creation of a two-state phylogenetic hidden Markov model

Our two-state phylo-HMM (Appendix Figure II-1) has a rate parameter associated with each state: one for the background (α_w = background rate of evolution) and one for the conserved segment (α_C = conserved rate). The local rate of evolution is the maximum likelihood estimate within a window (w = 21), which was obtained by gradient ascent. The conserved rate of evolution was set to be the smaller of (i) one third of the local rate of evolution or (ii) the maximum likelihood rate estimate at that column. These rates are then used to obtain the likelihood of the data under specific models of protein evolution, and the phylo-HMM then outputs a posterior probability of the conserved state at a particular column.

We used Felsenstein's algorithm (Felsenstein 1981; Durbin et al. 1998) to calculate the likelihood of the data [P(data|tree)] with an empirical amino acid substitution matrix obtained from the four closest related species of yeasts: *S.cerevisiae*, *S. paradoxus*, *S. mikatae* and *S.*

bayanus. The tree used was the species tree described above (Appendix Figure II-7), where the branch lengths were scaled by the rate of evolution for each HMM state. The likelihood of the substitution process is therefore:

$$L_s = \prod_{i=1}^{m} P(x_i^{\bullet} \mid tree, \alpha)$$

Where m is the number of alignment columns and the tree indicates the phylogenetic relationship between species (Appendix Figure II-7). α is the rate of evolution, which scales the branch lengths, and $x^{\bullet}$ represents the amino acid sequences in the alignment.

One of the assumptions of the traditional probabilistic approaches to protein evolution (such as the phylo-HMM) is that every amino acid column in an alignment can be treated as independent (Durbin et al. 1998). Because insertions or deletions do not follow this assumption (they can span multiple residues), most current phylogenetic models only account for residue substitutions (Durbin et al. 1998; Rivas and Eddy 2008). Gaps are usually ignored in phylogenetic analyses. Because short linear motifs occur often in unstructured regions that tend to create gapped regions in alignments, ignoring gaps would be a considerable problem in our analysis. Probabilistic models accounting for gaps have been proposed (Rivas and Eddy 2008), but their complexity and incompleteness have motivated us to create another simpler model. In our protein evolution model, blocks of gaps (illustrated as vertical black lines in Figure III-1B) are treated as insertion or deletion events. We consider a gap process that operates on one block at a time, contrasting with the substitution process that operates on one column of an amino acid alignment at a time. The two processes are considered independently and combined at the end. Having assigned each insertion or deletion as a contiguous block, we can obtain the likelihood of the gap process:

$$L_G = \prod_{j=1}^{b} P(y_j^{\bullet} \mid tree, \alpha, k_j) = \prod_{j=1}^{b} \frac{\mathrm{k_j}^{-1.5}}{\sum_{\mathrm{n}=1}^{\infty} \mathrm{n}^{-1.5}} P(y_j^{\bullet} \mid tree, \alpha)$$

Where b is the number of blocks and k the length of each block, which follows an empirically derived power law distribution (Cartwright 2006). In this likelihood, the substitution matrix consists of only two characters (gap or amino acids) and is calculated similarly as before (Appendix Figure II-1). Therefore $y_j^{\bullet}$ represents the pattern of gap characters and amino acid characters of the j^{th} block in the alignment.

The total likelihood of an alignment can then be written as:

$$L = L_S L_G$$

The HMM requires a likelihood for each alignment column. Therefore, we distributed uniformly the gap likelihood of each block to its alignment columns (x_i is within y_j). The likelihood of a single alignment column is therefore:

$$L_i = P(x_i^{\bullet} \mid tree, \alpha) P(y_j^{\bullet} \mid tree, \alpha, k)^{1/k}$$

To find regions in alignments that are conserved, we then computed the posterior probability of the conserved state with the likelihood of single columns and the forward and backward algorithm (Durbin et al. 1998). Because the insertion or deletion lengths do not depend on the evolutionary rate, the likelihoods given by the empirically derived power law distribution get canceled in the calculation of the posterior probability. This means that, for our method, the appearance and disappearance of insertions and deletions over the phylogenetic tree modeled as "blocks" are the sole contributor of insertion and deletion likelihoods in the final posterior probability. Transition frequencies between states were obtained with the expectation-maximization procedure described by Baum-Welch (Durbin et al. 1998).

We used multiple heuristics on the posterior probability to find peaks corresponding to short conserved residues. First, the analysis ignored the first three residues of the alignment, because the conserved methionine is usually aligned by MAFFT. Second, peaks were found by initially finding regions above a threshold of 0.2. These peaks were later pruned if the maximal posterior threshold within the region was lower than 0.6, or if they did not fit the desired size (2 to 20 amino acids). Although peaks longer than 20 amino acids were rare

because we calculated the local rate of evolution using a window of size 21, we excluded these signals because we did not consider them representative of typical short linear motifs.

We visualized the alignments with Jalview (Waterhouse et al. 2009) and red color intensity represented the posterior probability with full color intensity indicating a posterior probability of 1.

III.5.3 Defining unstructured regions

To find functional segments in unstructured regions of proteins, we used several filters to select regions of interest (unstructured regions) and to remove regions that may be conserved due to chance or as a property of the alignment program. We used DISOPRED2 (Jonathan J Ward et al. 2004) to remove structured regions from proteins, as well as pFilt (Jones et al. 1994) for coiled-coils. Large repetitive regions were removed using the SEG algorithm (Wootton and Federhen 1993). If long domains were interspersed with short highly degenerate sequences, these were not captured by any of the above filters, so we also removed regions of high conservation that were longer than 20 amino acids. Overall, of the total length of yeast proteins with orthologs, 24% of the amino acids passed all our filters.

III.5.4 Analysis of literature-curated short linear motifs

To estimate the effectiveness of our approach in identifying previously known short linear motifs, we identified 526 characterized short linear motifs in budding yeast by performing literature searches for known posttranslational regulatory proteins and detailed reading of the primary literature and determined how many of these were correctly identified by the phylo-HMM. The modifications were mostly phosphorylation sites but also included degradation signals, localization signals, interaction motifs, and SUMOylation sites (supplementary data table III-S2). Of these, 352 were found in regions that passed our filters for classification as disordered and of these 352 (346 that did not overlap with another motif), we considered 123 (119 that did not overlap with another motif) conserved, such that they could be identified (by consensus sequences within a window of six amino acids or by eye for localization signals) in at least 90% of the orthologous proteins.

Our phylo-HMM approach predicted 104 (or 30% of the 346 motifs that were classified as disordered) of the motifs present in disordered regions. However, because the underlying assumption of the phylo-HMM is that the motifs are fully conserved, we do not expect this method to find a large portion of the regulatory elements that may diverge over long evolutionary distances. Consistent with this, the phylo-HMM predicts 75 (or 63% of the 119 motifs that were classified as disordered and conserved) of the conserved motifs (Supplementary data table III-S2, Appendix Figure II-5).

III.5.5 Simulations of protein evolution

To address the issue of computational artifacts from misalignment in distant species and to low sequence divergence, we performed simulations of protein evolution. In our simulations, an ancestral protein is randomly generated and evolved through point mutations, insertions, and deletions according to the desired phylogenetic tree. Proteins contained three regions (see Appendix Figure II-4A for an example): the first region (on average 75 amino acids) and third region (on average 87.5 amino acids) evolved at a "background" rate (the average rate of yeast proteins) or at 70% or 130% of this rate. The first region contained a single simulated short sequence (two to nine amino acids) that evolved at a slow rate that we varied between 2.5% and 100% of the background rate. The second region (on average 75 amino acids) evolved slowly to simulate a conserved protein domain. Because unstructured regions often include gaps from insertions and deletions, we modeled the simulations such that the evolved proteins also evolved insertions or deletions of various sizes (k) following an empirically derived power-law distribution with z = 1.5 (Cartwright 2006) in:

$$P(k \mid z) = \frac{k^{-z}}{\sum_{n=1}^{\infty} n^{-z}}$$

We aligned the simulated protein sequences with MAFFT (Katoh et al. 2002) (Appendix Figure II-4A). We performed 100 simulations per data point.

We assessed alignment and prediction accuracy with simulations performed with different background rates of evolution. We plotted the accuracy of the alignment (Appendix Figure

II-4B, fraction of simulated motifs with correct motif alignment), sensitivity of the phylo-HMM (Appendix Figure II-4C, fraction of simulated motifs that were predicted by the phylo-HMM), and rate of computational artifacts (Appendix Figure II-4D, number of predictions that do not correspond to a simulated motif per 100 unstructured amino acids)

To estimate the rate at which the phylo-HMM identified motifs that were truly conserved, we compared the number of simulated motifs that were correctly aligned with the number of simulated motifs identified by the phylo-HMM. We found that 95% of the simulated artificial motifs were correctly aligned even when the surrounding region had minimal sequence similarity (at motif rate 10% of the background rate, Appendix Figure II-4B). At the same motif evolution rate, 93% of the simulated motifs were correctly identified and the fraction of simulated motifs that were predicted by the phylo-HMM was dependent on the relative rate of evolution of the motif to the background (Appendix Figure II-4C). The difference of the simulation results (93% correct predictions) with results from literature-curated conserved motifs (63% correct predictions) is likely due to an oversimplification of the evolution of disordered regions in our simulations. We also addressed the prediction of computational artifacts using the simulations. Because we know the location of the true motifs in the simulations, any other motifs identified by the phylo-HMM are false predictions. For yeast proteins evolving at the background rate, the phylo-HMM predicted about 1 false conserved motif every 9000 rapidly evolving amino acids; however this was dependent on the background rate of evolution (Appendix Figure II-4D).

To calculate the proportion of unstructured regions that contain short functional sequences, we first estimated the fraction of our predicted conserved sequences that are computational artifacts (1/9000 times 636409 unstructured amino acids divided by 7361 predicted motifs = ~0.95%). To estimate the fraction of unstructured amino acids that are biologically important, we divided the number of amino acids in predicted conserved sequences by the total number of unstructured amino acids in the yeast proteome (33626 divided by 636904 = 5.3%) and multiplied by 99.05% (100% - 0.95%) to take into account the expected number of predicted computational artifacts, which yields our estimate of 5.2%.

III.5.6 Motif clustering, alignment, and enrichment

We performed an all-by-all pairwise comparison and alignment of each sequence alignment to another using the Smith-Waterman algorithm (Smith and Waterman 1981; Durbin et al. 1998). In the initial distance metric, we divided the alignment score by the square root of the length of the alignment and corrected for the initial length of the sequence. This was done so that poor but long alignments would not score as well as short but strong alignments. We also performed another clustering with the same distance metric, but first the sequences were extended by five amino acids on each side, and the positions with an information content lower than 1 (positions with high sequence diversity) were eliminated from the beginning and end of the extended sequences. To easily identify subclusters, we tried clustering by finding the top 10 "partners" of each sequence, removing hits between paralogs and within the same gene. For this final cluster, we first extended the sequences by five amino acids; however, we did not divide by the square root of the length of the alignment. Alignments that passed a threshold (as described in Supplementary data table III-S5) were then plotted as an interaction network using Cytoscape (Shannon et al. 2003), and we used the MCODE's (Bader and Hogue 2003) k-core clustering algorithm to form similarity clusters. MCODE often links multiple clusters by a single node and therefore forms "subclusters". We either analyzed the whole clusters or these subclusters by creating sequence logos and functional enrichment as described below (see Supplementary data table III-S5 for the top 20 predictions of each clustering analysis with annotation and more details).

Enrichment in protein function or interaction was performed with data from the MIPS functional catalog using FunSpec (Robinson et al. 2002), , and with data from the Gene Ontology (GO) Slim Mapper at the SGD (SGD Project 2011) for GO. Statistical significance was assessed at a P-value < 0.05.

Motif patterns are represented as sequence logos (Schneider and Stephens 1990; Crooks et al. 2004), which were obtained with a heuristic multiple alignment of the *S. cerevisiae* representative of each motif.

III.5.7 Strains, plasmids, and primers

We used an endogenously tagged *SPT21* strain from the TAP-fusion library (Ghaemmaghami et al. 2003) to assess Spt21 stability throughout the cell cycle. *SPT21*

overexpression plasmids were obtained from the MORF (Gelperin et al. 2005) and GAL-ORF-GST (Zhu et al. 2001) libraries. Mutagenesis was performed with the QuikChange™ Site-Directed Mutagenesis System developed by Strategene (La Jolla, CA). For the KEN box, all three codons were mutated to the alanine coding GCT with oligos SPT21kenbox1-FP 5'-GATATCTTTAACTAGTGAAAATGATGCTGCTGCTATTCCACCCCAAAGCATAA CTAGTA and SPT21kenbox1-RP 5'-TACTAGTTATGCTTTGGGGTGGAATAGCAGCAGCATCATTTTCACTAGTTAAAGA TATC. The desired mutations were confirmed by sequence analysis. BY4741 or isogenic derivatives were used for all of our experiments.

III.5.8 Cell-cycle induction of SPT21

Yeast cells expressing Spt21-TAP from its endogenous promoter were grown to early log phase in YEPD (1% yeast extract, 2% bactopeptone, 2% glucose) and then arrested in G_1 with α-factor. After 2 hours (>95% cells arrested), the cells were washed twice with fresh media and samples were taken every 15 minutes. Both fluorescence-activated cell sorting (FACS) analysis and the amount of Clb2 protein were used to follow cell cycle progression. Hexokinase was used as the loading control for the Western blot. Spt21 or Clb2 abundance was quantified, and significance was assessed by Pearson correlation coefficient.

III.5.9 Pulse-chase assay

Cells carrying galactose-inducible overexpression plasmids were grown in synthetic dextrose medium lacking uracil overnight. Spt21 and $Spt21^{ken}$ expression was induced by culturing cells in galactose-containing medium (2% concentration) for 4 hours. Glucose was subsequently added to a final concentration of 2% to attenuate protein expression, and protein synthesis was abolished through the addition of cyclohexamide (100 mg/ml final). Cells were collected at 20-min time intervals. To ensure reproducibility, we performed pulse-chase experiments on both the glutathione *S*-transferase (GST)-tagged and the protA-tagged version of the Spt21 and corresponding KEN box mutant. Protein abundance was quantified and analyzed for significant changes in abundance by *t* test.

III.5.10 Protein extracts and Western blotting

Protein extracts were prepared by trichloroacetic acid and separated by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) on 8% polyacrylamide gels. Western blotting was performed with anti-protA antibody (peroxidase anti-peroxidase soluble complex Sigma, P1291) for detection of SPT21-protA. Clb2 and hexokinase detection was performed using α-Clb2 [Santa Cruz Biotechnology, Clb2 (y-180) sc-9071], and α-hexokinase (yeast) (Rockland Immunochemicals, Inc., 200-4359), respectively. For data requiring quantification, we quantified the amount of protein with images of the Western blots obtained from the VersaDoc MP System (Bio-Rad Laboratories, Inc.). Mean band intensities of the relevant proteins were normalized to the mean intensity of the hexokinase band with ImageJ (Abramoff et al. 2004).

III.5.11 *In vitro* pull-down assays

A GST-tagged Cbk1 fragment containing the kinase domain and the C-terminal extension (~76 kD) was expressed in *Escherichia coli* Rosetta(DE3)pLysS; purified on Ni-NTA resin (Qiagen) and glutathione-Sepharose (GE Biosciences), and dialyzed into 20 mM Tris, 150 mM NaCl, 2 mM dithiothreitol (DTT) (pH 8.0). Purified Cbk1 was flash-frozen in liquid nitrogen and stored at -80°C. Fragments containing putative interaction motifs were expressed as MBP fusions in BL21(DE3)RIL. Cell lysates containing the interaction motif constructs were incubated with amylose resin (New England Biolabs) on a rotator at 4°C for 15 min, and the beads were washed with phosphate-buffered saline [137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4 (pH 7.3)] + 2 mM DTT. Washed amylose beads (approximately 50 μg of MBP fusions) were incubated with 1 μM purified Cbk1 (~3.8μg) for 15 min at 4°C (total volume 50 μL) and then washed with TBST [50 mM Tris, 150 mM NaCl, 0.1% Tween 20 (pH 7.5)] and resuspended in SDS-PAGE loading buffer. A third of the reactions were loaded on SDS-PAGE gels, which were directly stained by GelCode Blue (Pierce) or transferred to nitrocellulose for Western blotting. Cbk1 was detected with a GST primary antibody (Santa Cruz Biotechnology), followed by an IRDye800 anti-mouse (Rockland) secondary antibodies. Blots were visualized using a Li-Cor Odyssey system.

III.6 Acknowledgements and funding sources

We thank Philip Kim, Nicholas Provart, John Parkinson, and members of the Moses' lab for discussions. We also thank an anonymous reviewer for suggestions on hub proteins analysis. ANNB is supported by an Ontario Graduate Scholarship and a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada. BJY was a Damon Runyon Fellow supported by the Damon Runyon Cancer Research Foundation (DRG-1976–08). Research in ELW's laboratory is supported by an NIH-NIGMS grant (#GM-084223). ARD is supported by Canadian Institutes of Health Research (Grant number MOP-13609). AMM is supported by a Natural Sciences and Engineering Research Council of Canada Discovery grant. This research was supported by an infrastructure grants from the Canadian Foundation for Innovation to AMM and BJA. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

III.7 Author contributions

ANNB designed and performed the computational and SPT21 related experiments, identified characterized short linear motifs and wrote the paper. BJY designed and performed the pull-down assays for the Cbk1 interaction motif and edited the paper. DvD designed and performed the SPT21 related experiments and edited the paper. ARD designed the computational experiment and edited the paper. ELW designed the Cbk1 interaction motif experiments and edited the paper. AMM designed the computational and SPT21 related experiments and wrote the paper.

III.8 Data and materials availability

All strains and source code used for this study are available upon request. A browser, and the datasets produced are available on our website: http://www.moseslab.csb.utoronto.ca/phylo_HMM/

III.9 Supplementary Data

Supplementary text and figures can be found in the Appendix II. Supplementary tables and source codes are included in the supplementary data file.

Chapter IV

Detecting Functional Divergence After Gene Duplication Through Evolutionary Changes in Posttranslational Regulatory Sequences Using a Non-central Correction to the Likelihood-ratio Test

This work has been submitted as: Detecting functional divergence after gene duplication through evolutionary changes in posttranslational regulatory sequences using a non-central correction to the likelihood-ratio test

Submitted to PLoS Comp. Biol.

Alex N Nguyen Ba[1,2], Bob Strome[1], Jun Jie Hua[1], Jonathan Desmond[1], Isabelle Gagnon-Arsenault[3], Eric L Weiss[4], Christian R Landry[3], Alan M Moses[1,2]

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, M5S 3B2, Toronto, Canada

2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada

3. Département de Biologie, IBIS and PROTEO, Pavillon Charles-Eugene-Marchand, 1030 Avenue de la Medecine, Laval University, Québec City, QC G1V 0A6, Canada.

4. Department of Molecular Biosciences, Northwestern University, Evenston, Illinois, United States of America

IV.1 Abstract

Gene duplication is an important evolutionary mechanism that can result in functional divergence in paralogs due to neo-functionalization or sub-functionalization. Consistent with functional divergence after gene duplication, recent studies have shown accelerated evolution in retained paralogs. However, little is known in general about the impact of this accelerated evolution on the molecular functions of retained paralogs. For example, do new functions typically involve changes in enzymatic activities, or changes in protein regulation? Here we study the evolution of posttranslational regulation by examining the evolution of important regulatory sequences (short linear motifs) in retained duplicates created by the whole-genome duplication in budding yeast. To do so, we identified short linear motifs whose constraint has relaxed after gene duplication with a likelihood-ratio test that can account for heterogeneity in the evolutionary process by using a non-central chi-squared null distribution. We find that short linear motifs are more likely to show changes in evolutionary constraints in retained duplicates compared to single-copy genes. We examine changes in constraints on known regulatory sequences and show that for the Rck1/Rck2, Fkh1/Fkh2, Ace2/Swi5 paralogs, they are associated with changes in posttranslational regulation. Finally, we experimentally confirm our prediction that for the Ace2/Swi5 paralogs, Cbk1 regulation was lost along the lineage leading to *SWI5* after gene duplication. Our analysis suggests that changes in posttranslational regulation mediated by short regulatory motifs systematically contribute to functional divergence after gene duplication.

IV.2 Introduction

Gene duplication is thought to be one of the major sources of evolutionary innovation (reviewed in (Conant and Wolfe 2008)). Several molecular mechanisms of functional change have been proposed: 1) changes at the transcriptional level can alter the expression of the paralogous copy (Force et al. 1999; Levine and Tjian 2003; Hittinger and Carroll 2007; Gagnon-Arsenault et al. 2013) , 2) changes at the enzymatic level can alter the activity or specificity of the protein (Conant and Wolfe 2008; Voordeckers et al. 2012), 3) changes at the posttranslational level can modify the regulation or localization of the protein (Marques et al. 2008; Amoutzias et al. 2010), and 4) changes within the splicing sites can change the isoforms produced at each loci (Su et al. 2006; Marshall et al. 2013). Studies on genome-wide mRNA expression patterns have established that transcriptional changes are one of the major contributors of functional differences within duplicated genes (Gu et al. 2004; Huminiecki and Wolfe 2004; Gu et al. 2005). However, the relative roles of functional divergence by gene regulation and changes within the amino acid coding sequence of the proteins are still unclear (Li et al. 2005).

Coding sequences of paralogous genes show increased evolutionary rates after duplication (Byrne and Wolfe 2007; Scannell and Wolfe 2008), consistent with the hypothesis that changes within the amino acid coding sequences are also important contributors to functional divergence. However, because some functional features in proteins comprise a small number of amino acids, statistical studies comparing evolutionary rates of whole proteins do not provide mechanistic explanations for changes in function (Dean and Thornton 2007). For example, many proteins contain short linear motifs such as phosphorylation sites, localization signals and interaction motifs, and these motifs are only 2-15 amino acids long (Gould et al. 2010). For instance, the cell-cycle regulator Sic1 is a disordered protein with several phosphorylation and binding sites that comprise less than 20% of the protein (Kõivomägi, Valk, Venta, Iofik, Lepiku, Balog, et al. 2011). Computational identification of short linear motifs is an important challenge, often relying on experimental data (Neduva et al. 2005; Lieber et al. 2010). However, because they are short and are frequently found in fast evolving disordered regions, it is still difficult to accurately and systematically identify

them; indeed, most short linear motifs in disordered regions probably remain uncharacterized (Nguyen Ba et al. 2012). Therefore, analyses on whole proteins may underestimate the level of functional divergence after gene duplication because changes in constraints in short linear motifs may lead to regulatory changes and therefore functional divergence (Amoutzias et al. 2010). Recently, several studies have investigated specific types of posttranslational regulatory changes (Beltrao et al. 2009; Lim and Pawson 2010; Sun et al. 2012) (reviewed in (Beltrao et al. 2013)), such as differences in patterns of phosphorylation between paralogs (Freschi et al. 2011) or differences in localization in paralogous proteins (Marques et al. 2008), and have shown that regulatory changes can also contribute to functional divergence. However, these regulatory changes can also be attributed in part to *trans*-regulatory changes (changes in proteins that control posttranslational regulation). Identification of changes in the protein regulatory sequences would allow us to determine *cis*-regulatory divergence (changes within duplicated proteins), and provide amino acid level mechanistic explanations for protein regulatory changes after duplication (Moses and Landry 2010).

Formally, functional divergence in amino acid sequences after gene duplication has been divided into two types of evolution (Gu 1999). The first (type I) describes so-called "changes in constraint" where the rate of evolution in a site or region changes after duplication, and remains different in one of the paralogous clades. The second (type II) describes a burst of rapid evolution immediately after gene duplication, and then a restoration of similar levels of constraint in the two paralogous lineages. Several statistical methodologies have been developed to identify sites or regions in proteins that fall into these classes (Huang and Golding 2012; Gu et al. 2013). These approaches have largely focused on identifying sites in globular regions of proteins for which large numbers of homologues can be accurately aligned (Abhiman and Sonnhammer 2005). These approaches often use likelihood-ratio tests based on advanced probabilistic models of phylogeny and amino acid substitution to compare the rates of evolution in individual sites (Knudsen and Miyamoto 2001) or groups of sites (Huang and Golding 2012; Gu et al. 2013) to the rest of the protein. For example, previous applications of these methods have identified possible positions in the globular domain of carbonic anhydrase III that are responsible for posttranslational addition of glutathione (Knudsen et al. 2003). In principle, these methods could be applied to identify changes in short linear motifs within disordered regions that contribute to posttranslational

regulatory change. However, because real protein evolution can be more complicated than even the most sophisticated models (Pond et al. 2005) and real protein alignments include non-biological sources of heterogeneity (such as alignment errors and missing data), the likelihood-ratio test can falsely identify type I functional divergence (Gu et al. 2013). One strategy to tackle these issues is to estimate the rejection rate of the likelihood-ratio test using empirical data, for example using permutation tests (Lanfear 2011). However, the distribution of the likelihood-ratio test statistic must be obtained through permutations performed for every protein and therefore may be too laborious for genome-wide studies.

We set out to study the change in selective constraints in short linear motifs within disordered regions after the whole-genome duplication (WGD) in budding yeast by asking whether the rates of evolution of these segments significantly differed after the whole-genome duplication event. We first developed a statistical method to correct the p-value distributions of likelihood-ratio tests in the presence of a heterogeneous background evolution process that violates the models assumed by the test, but where the background can be "fit" to some extent by the alternative hypothesis. We show how this approach can be applied to predicted short linear motifs and to protein sequences simulated under various complex evolutionary processes to eliminate false rejections. We then show that the turnover of predicted motifs within retained paralogs is faster than in genes whose paralogs were lost after duplication (which we refer to as single-copy genes or proteins) and that, for these putative short linear motifs, correlated loss of selective constraints appear to be common, consistent with changes in function specific to one of the two paralogs.

Finally, we identify examples of experimentally verified motifs present in one paralog that are unlikely to be present in the other copy, and verify our prediction for one of these examples (Ace2 and Swi5). Our results show that a view of molecular evolution with amino acid resolving power can allow us to propose specific hypotheses about the functional divergences between paralogs.

IV.3 Results

IV.3.1 Detection of type I functional divergence after gene duplication using a non-central chi-squared null distribution for likelihood-ratio tests

We have previously shown that short linear motifs can be predicted based on their conservation relative to their surrounding regions (Nguyen Ba et al. 2012). We sought to detect regulatory divergence in proteins by looking for statistical signals of lineage-specific changes in rate of evolution in predicted short linear motifs in multiple sequence alignments, which would indicate changes in function. Likelihood-ratio tests have previously been used to detect differences in rate of evolution of full-length yeast proteins after the whole-genome duplication (Byrne and Wolfe 2007). We sought to perform essentially the same test to identify short linear motifs whose rate of evolution changed significantly after gene duplication. To do so, we first predicted short linear motifs within proteins of species that have diverged prior to the yeast whole-genome duplication (see Methods). Because this procedure does not involve the post-WGD clade (which therefore allows us to detect changes in constraints), we then mapped the location of the predicted short linear motifs to the genes post-duplication (Figure IV-1). Using a likelihood-ratio test (Yoder and Yang 2000), we tested whether two rates of evolution (one for the post-duplication clade and one for the remainder of the phylogenetic tree) explain the data significantly better than one single rate of evolution common to the whole tree (see Methods). This test is performed once for genes that reverted to single-copy, and twice in retained duplicates (one for each post-WGD protein).

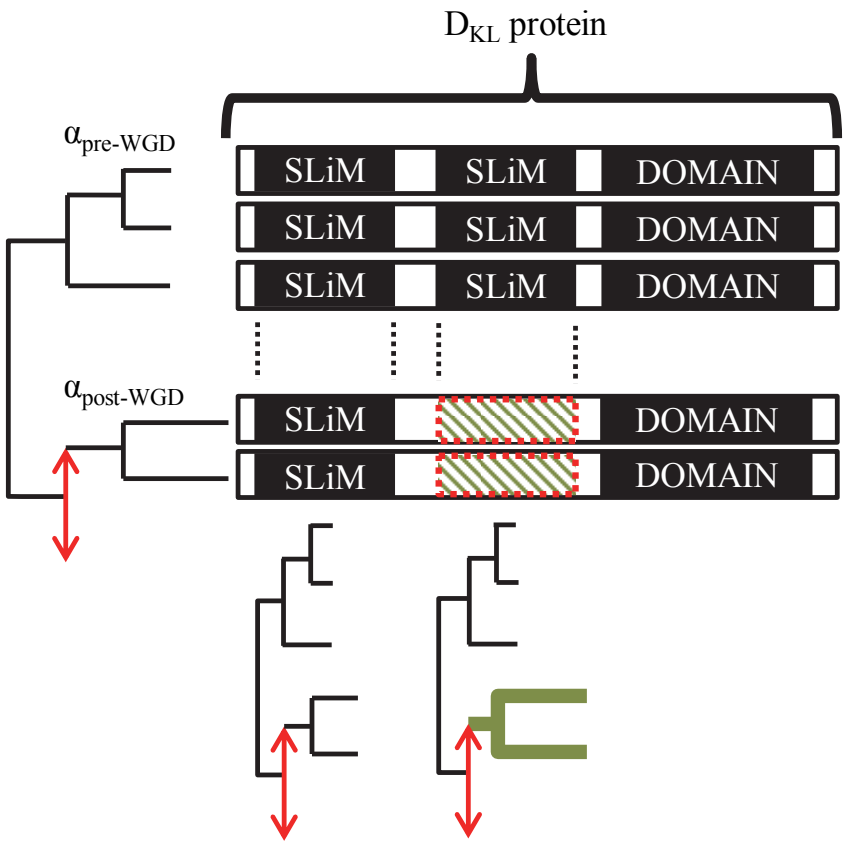


Figure IV-1. Schematic of the motif-specific likelihood-ratio test. Rates of evolution are computed for each motif before ($\alpha_{pre\text{-}WGD}$) and after (α_{WGD}) gene duplication and compared with the rates that were observed for the whole protein (see Methods). Red double arrow illustrates the duplication event. Bolded clades are clades with significant changes in constraints. Striped patterned boxes indicate short linear motifs with significantly different change in rate of evolution. D_{KL} indicates the expected deviation of the likelihood-ratio test from the whole protein.

Previous efforts to identify changes in evolutionary rate have showed that the likelihood-ratio test statistic often deviates from the expected chi-squared null distribution even when there is truly no change in rate of evolution (Fletcher and Yang 2010; Lanfear 2011). Indeed, when we performed simulations of molecular evolution with no changes in rate of evolution specific to the short linear motifs (Figure IV-2, see Methods), but included realistic aspects of the evolutionary process (such as rate heterogeneity, insertions and deletions, etc.), we found that the likelihood ratio test falsely identified increased rates of evolution after gene duplication (Figure IV-3, black circles, Appendix III).

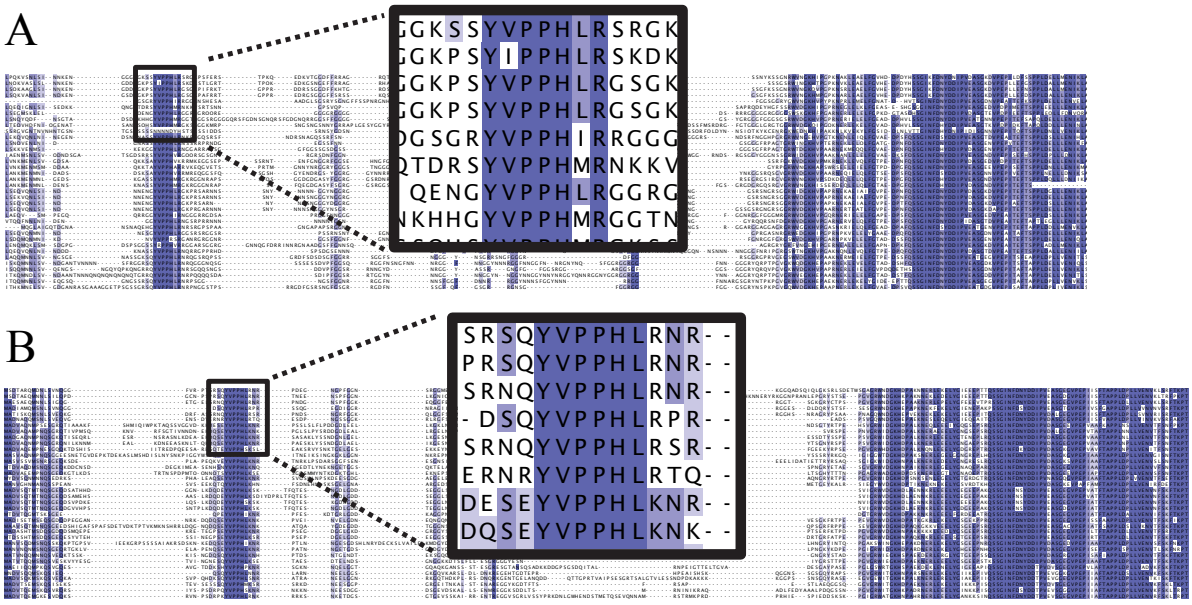


Figure IV-2. Simulation of protein evolution. A) Alignment of the N-terminus of the Dbp1/Ded1 homologs illustrates the rate heterogeneity amongst columns and highlights the short length of a putative motif (black rectangle zoom). Blue shade represents the percentage identity. B) Alignment of the N-terminus of a simulated set of proteins based on Dbp1/Ded1 using our 'realistic' simulation of evolution (see Methods).

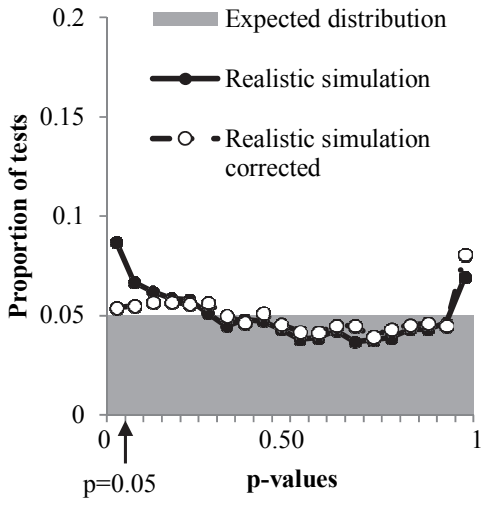


Figure IV-3. Likelihood-ratio test on short linear motifs after gene duplication. Protein sequences were evolved as in Figure IV-2B). Grey shaded area indicates the expected proportion of tests. Circles indicate the distribution of p-values obtained from the likelihood-ratio test described in Figure IV-1) when the test statistic is assumed to be chi-squared distributed (black circles) or non-central chi-squared distributed (white circles, "corrected").

We hypothesized that the increased rate of false rejections was because to the additional evolutionary rate parameter in the alternative hypothesis (that is supposed to capture the change in selective constraints) can also model some of the background heterogeneity in evolutionary rate (due to alignment errors, non-stationary and non-homogeneous evolution, etc.).

Under assumptions that 1) the majority of the tests performed are truly null, and that 2) the deviation of the real data from the models assumed by the test is consistent over the columns of the multiple sequence alignment, the distribution of the likelihood-ratio test follows a non-central chi-squared distribution with a data-dependent non-central parameter (see Methods). This non-central parameter (the expected increase in the test statistic from 'fitting' some of the heterogeneous background process using the likelihood ratio test) is the product of the

Kullback-Leibler (KL) divergence D_{KL}, which is the "fit" or the expected log-likelihood ratio of the alternative hypothesis over the null hypothesis given the data (see Methods) and the number of data points used to compute the likelihood-ratio test. Larger KL divergence means larger deviation of the background distribution from the null model assumed by the test. To use this in practice, we first estimate a non-central parameter using sequence data generated by a background heterogeneous evolution process and then use the non-central chi-squared distribution to obtain p-values for our test (see Methods). Extensive simulations on large proteins with non-stationary and non-homogeneous evolution, including alignment errors, showed that this approach works as expected and yields uniform p-values (see Appendix III).

We applied this approach to our 'realistic' simulation (Figure IV-2 for an example protein) by calculating a KL divergence parameter for each protein (see Methods) and obtained p-values for each likelihood-ratio test (for each short linear motif) in that protein. This procedure reduced the false-rejection rate (Figure IV-3, white circles) and p-values were nearly uniform.

IV.3.2 Frequent post-duplication changes in constraints in motifs

Having confirmed that our approach to detect type I functional divergence could be applied on short linear motifs, we then analyzed our set of protein alignments. After correction for multiple testing, we identified 159 short linear motifs with significantly different rates of evolution after gene duplication at a false discovery rate of 5% (see Methods, Supplementary data table IV-S1). This corresponds to 1.2% of the motifs identified in single-copy genes (Figure IV-4A) and 9.8% of the identified motifs in retained duplicates (Figure IV-4B). Changes in constraints are approximately 4.5 times more frequent in retained duplicates versus single-copy proteins (5.26% vs 1.15% of LRTs, p-value $< 10^{-20}$, Fisher's exact test).

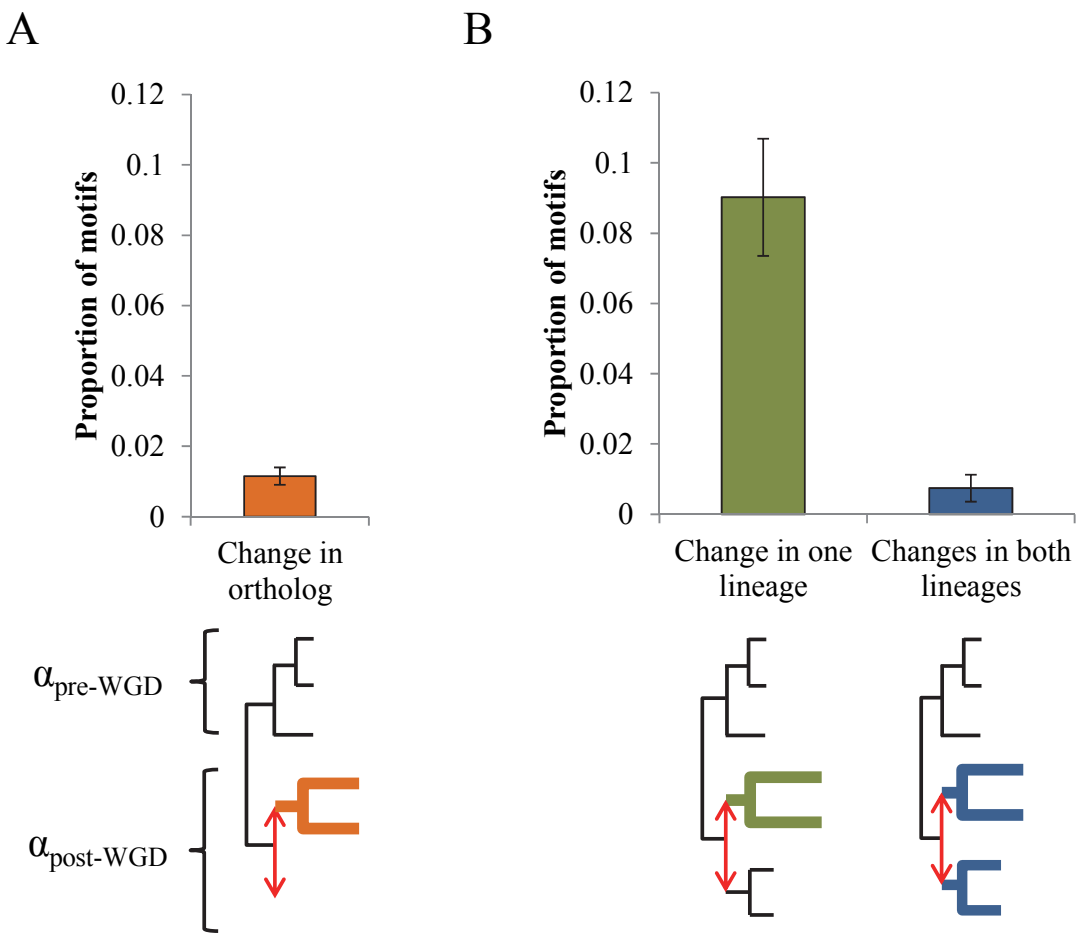


Figure IV-4. Regulatory turnover after gene duplication. A-B) The proportion of motifs with changes in constraints at a 5% false-discovery rate is significantly larger than in genes with retained duplicates (B) than in single-copy genes (A). Error bars represent the 95% confidence interval of the estimated proportion (binomial distribution). Bolded clades are clades with significant changes in constraints. α is the rate of evolution.

Our previous 'realistic' simulation had no intended site-specific changes in constraints. Despite this, our pipeline (including the non-central correction) identified 0.059% and 0.55% of the motifs in simulated single-copy genes and retained paralogs respectively to have significantly different rates of evolution after false-discovery rate correction. Using these values as our estimate of false positives due to possible computational artifacts (such as misalignments) or due to incorrect non-central parameter estimation for the null distribution of the likelihood-ratio test statistic, we expect that 5 motifs in duplicates and 3 motifs in

single-copy genes are artifacts. Therefore, although the false positive rate due to artifacts in retained duplicates is higher than in single-copy genes, the increased proportion of motifs identified with changes in constraints in duplicates cannot be explained by these computational artifacts.

As another negative control, we also looked at whether the flanking regions of the putative short linear motifs (five amino acids on each side of the motifs) showed changes in constraints after gene duplication. After correction for multiple testing, only two flanking regions were identified as having significantly different rates of evolution after gene duplication. Given that these identified changes in constraints on the flanking regions are consistent with our false positive rate, this result indicates that the type I functional divergence we identify in predicted short linear motifs is specific to the motifs and not due to some local change in constraint.

Most of the motifs with changes in constraints in duplicates only occurred in one of the two copies (85/92 motifs retained in duplicates), consistent with the idea of sub-/neo-functionalization after gene duplication through posttranslational regulatory changes (Amoutzias et al. 2010) (Figure IV-4B).

IV.3.3 Lineage bias in post-duplication changes in constraints

One hypothesis as to the fate of paralogous proteins is the duplication-degeneration-complementation (DDC) model (Force et al. 1999) which explains the preservation of paralogous proteins by the neutral generation of sub-functionalized copies of proteins. Under this hypothesis, one might therefore expect that both paralogous proteins would show signs of relaxed evolution, but that specific functional regions of each protein showing relaxation in selective constraints would be complementary, such that they partition the functional regions in the ancestral protein. We sought to test whether signs of the DDC model could be detected at the posttranslational regulatory level and found 20 paralog pairs where more than one short sequence was detected as having different rate of evolution after gene duplication (see Methods). Of these, seven showed reciprocal changes in constraints on their motifs, which is consistent with degeneration and complementarity at the posttranslational regulatory level as predicted by the DDC model.

Despite some evidence for complementarity, the majority of paralogs (13/20) with more than a single change in constraints appeared to have a lineage bias in their posttranslational regulatory changes. We tested this using the set of 20 paralog pairs described above and asked whether the motifs were more likely to have correlated evolution than expected by chance. To do so, we randomly permutated the changes in constraints across paralogous pairs to establish the null expectation of random assortment and counted the lineage differences in changes in constraints (see Methods). This analysis revealed a lineage bias in changes in constraints for regulatory sequences (p-value = 0.01106, one-tailed non-parametric permutation test, Figure IV-5). Therefore, proteins that change function after duplication may typically change multiple short linear motifs in concert, consistent with the idea that multiple regulatory mechanisms often work together to control protein function. For example, multisite phosphorylation from individual or multiple kinases can form intricate regulatory modules on single proteins (reviewed in (Cohen 2000)) and these clusters of phosphorylation sites have been found to be frequently conserved through evolution (Holt et al. 2009; Lai et al. 2012) and have been shown to turnover (Moses, Liku, et al. 2007).

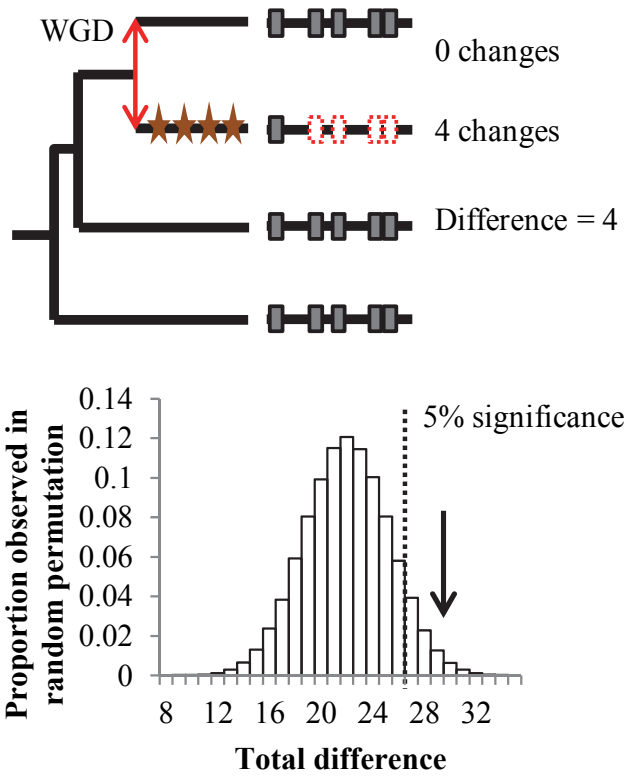


Figure IV-5. Correlated evolution of short linear motifs. Top panel shows the procedure to obtain the number of lineage specific changes in constraints in a single protein. Red double arrow illustrates the duplication event. Stars represent significant changes in constraints along the lineage. Significant changes in constraints detected on short linear motifs are shown in dotted red boxes. Bottom panel shows the distribution of the total cumulated number of lineage specific changes in constraints from a non-parametric permutation test. Arrow shows the observed total difference for all 20 paralog pairs.

IV.3.4 Amino acid level resolving power allows detection of additional changes after gene duplication

The increase in resolving power obtained by analysing short linear motifs allowed us to determine whether specific regions within the paralogous proteins differed in their selective constraints. We wanted to test if this amino acid level analysis could also allow us to detect signatures of functional divergence even when the rate of evolution of the whole protein after duplication did not appear to be different than the pre-WGD clade.

Using similar methodologies as previous studies (Byrne and Wolfe 2007), we found that 57% of the paralog pairs showed no evidence of significant increase in rate of evolution of the whole protein in either of the two lineages. This value is slightly higher than that obtained previously (44% (Byrne and Wolfe 2007)), which we attribute to either a different gene set or methodology, or to the non-central correction that we applied. Nevertheless, we then searched within these proteins for motifs with significant changes in constraints. Doing so, we identified 37 motifs in 28 paralogous pairs, and 46 motifs in 43 single-copy proteins. This indicates that an analysis of evolutionary rate differences using higher resolving power of functional sequences within proteins can identify additional sources of functional divergences than analyses at the whole protein level.

IV.3.5 Post-duplication changes in constraints are associated with changes in regulation

If changes in posttranslational regulation are important for functional divergence after gene duplication, we expect the changes in constraints in short linear motifs that we detected to point to functional differences between paralogous proteins. A previous study investigated changes in localization after gene duplication by taking advantage of the systematic green fluorescently-tagged protein collection in budding yeast (Huh et al. 2003; Marques et al. 2008) and categorized paralog pairs as having different or similar subcellular localization. We sought to test if motifs present in paralog pairs with different subcellular localizations were more likely to turnover after gene duplication. Motifs with changes in constraints were more than twice as likely to appear in proteins with detected changes in localization (26/209 motifs with changes in constraints in proteins with different localization vs 12/197 in proteins with similar localization, p-value = 0.032, permutation test), providing evidence that proteins with changes in localization are more likely to have evolved differences in short linear motifs.

We also tested if the changes in constraints we predicted corresponded to interpretable differences in posttranslational regulation by analyzing experimentally characterized motifs (same set as in (Nguyen Ba et al. 2012)) identified in the paralogous protein that overlapped with segments predicted to have a change in constraint.

Of these, the paralog pair Rck1/Rck2 contained two predicted motifs that were found to have significant changes in constraints in the Rck1 protein. Interestingly, both motifs are involved in Hog1 signaling (Bilsland-Marchesan et al. 2000; Teige et al. 2001). Consistent with our predictions, Rck2 is known to be regulated by Hog1, while Rck1 is thought not to be regulated by Hog1 (Teige et al. 2001). However, while our algorithm identified that the motif required for Hog1 binding in Rck2 was evolving more rapidly in Rck1, it is clear that Rck1 preserved some of the critical residues required for binding to Hog1, yet its binding activity to Hog1 has been shown to be poor (Teige et al. 2001). This suggests that: 1) the protein ancestral to Rck1/Rck2 is likely to also be regulated by Hog1, and 2) that Rck1 is likely to be regulated in a different manner, having lost or changed critical regulatory sequences after the duplication event (Figure IV-6).

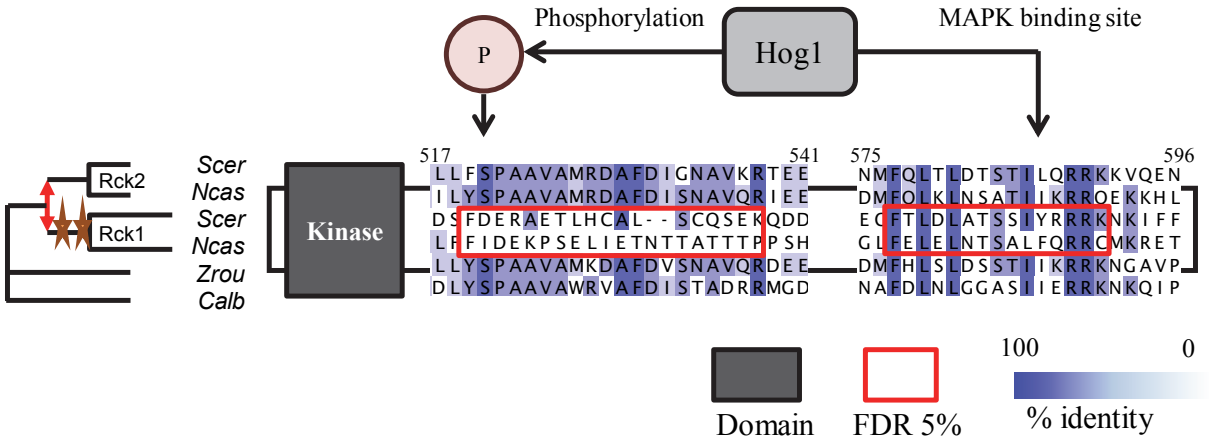


Figure IV-6. Known regulatory motifs with changes in constraints in Rck2/Rck1. Alignment of the short linear motifs with known function (indicated with arrows) and significant changes in constraints (red boxes) after gene duplication from representative species. The Rck2 protein is known to bind and be phosphorylated by Hog1 kinase at two motifs that have significant changes in constraints after gene duplication. Numbers indicate residue position within the *S. cerevisiae* Rck2 protein. The two identified motifs occur at aa519-538 and aa577-591 for Rck2, and changed constraints within the aligned region aa439-456 and aa492-506 in Rck1. These overlap with the known phosphorylation site in Rck2 (aa520) and the MAP kinase binding site (aa492-506) in Rck2. Both Rck2 and Rck1 retain kinase function.

Another clear example where experimentally characterized regulation of one paralog appears to have been lost in the other following gene duplication is in the Fkh2/Fkh1 paralogous pair of transcription factors. While both proteins play a role in cell-cycle progression, they are known to have non-redundant functions (Hollenhorst et al. 2000). For example, Fkh2, but not Fkh1, associates with Mcm1 (Hollenhorst et al. 2001). Another important function of the Fkh2 protein that is absent in Fkh1 is its ability to recruit the transcriptional co-activator Ndd1. This interaction is mediated by at least two adjacent Cdk1 phosphorylation sites (Pic-Taylor et al. 2004), one of which is found to have significant changes in constraints in the Fkh1 lineage. The other phosphorylation site is not predicted by our motif prediction algorithm but is also likely to have changed constraints. We speculate that the ancestral protein to Fkh1/Fkh2 may also have bound Ndd1 in a Cdk1-dependent manner, but Fkh1's regulation appears to have changed, possibly to accommodate new functional roles (Figure IV-7).

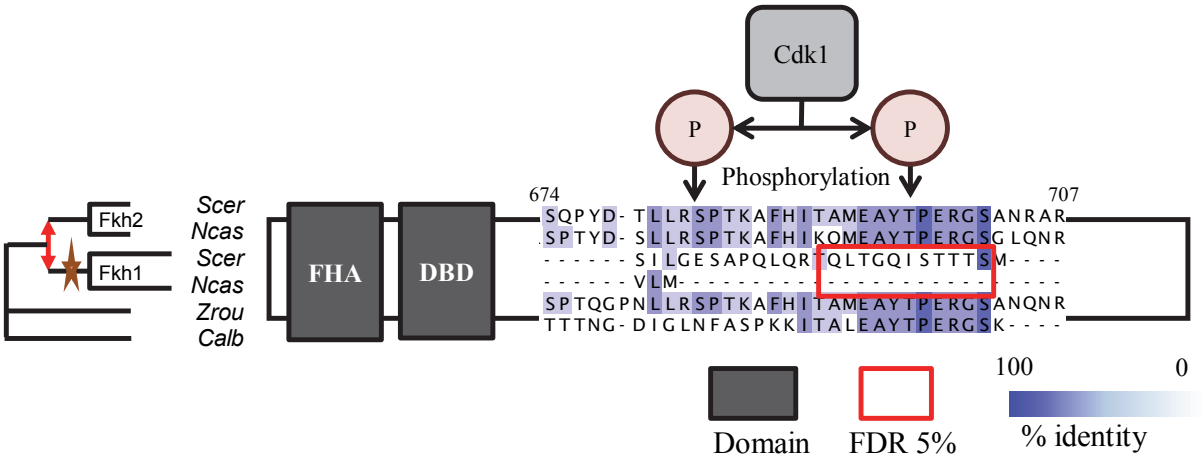


Figure IV-7. Known regulatory motifs with changes in constraints in Fkh2/Fkh1. Alignment of the short linear motifs with known function (indicated with arrows) and significant changes in constraints (red boxes) after gene duplication from representative species. The Fkh2 protein is known to be phosphorylated by Cdk1 at two phosphorylation sites on a region shown to have significant changes in constraints after gene duplication. Numbers indicate residue position within the *S. cerevisiae* Fkh2 protein. The identified motif occurs at region aa692-702 in Fkh2 and has changed constraint in the aligned region aa459-469 in Fkh1. One of the known phosphorylation site in Fkh2 occurs within this region at aa697. Fkh2 and Fkh1 retain their forkhead-associated domain (FHA) and DNA binding domain (DBD).

A third example could be found in the Ace2/Swi5 paralog pair, important cell-cycle regulated proteins known to localize differently in budding yeast (Sbia et al. 2008). These two proteins have been extensively characterized, with several major posttranslational regulatory sequences identified (Mazanka et al. 2008). Two of these have significant p-values in our analysis, suggesting that changes in constraints occurred within the Swi5 lineage. One of these is the Cbk1-regulated nuclear export signal, known to give Ace2 its daughter-cell specific nuclear localization (Mazanka et al. 2008), and another other is a putative Cbk1-binding motif (Nguyen Ba et al. 2012) (Figure IV-8). In Ace2, Cbk1 phosphorylation prevents nuclear export and Cbk1 is only active in daughter cells (Mazanka et al. 2008). Therefore, we hypothesize that the ancestral protein to the Ace2/Swi5 paralog pair was also regulated by Cbk1 to provide daughter-cell specific nuclear localization, but

that loss of these important signals allowed Swi5 to localize to both mother and daughter cells' nuclei.

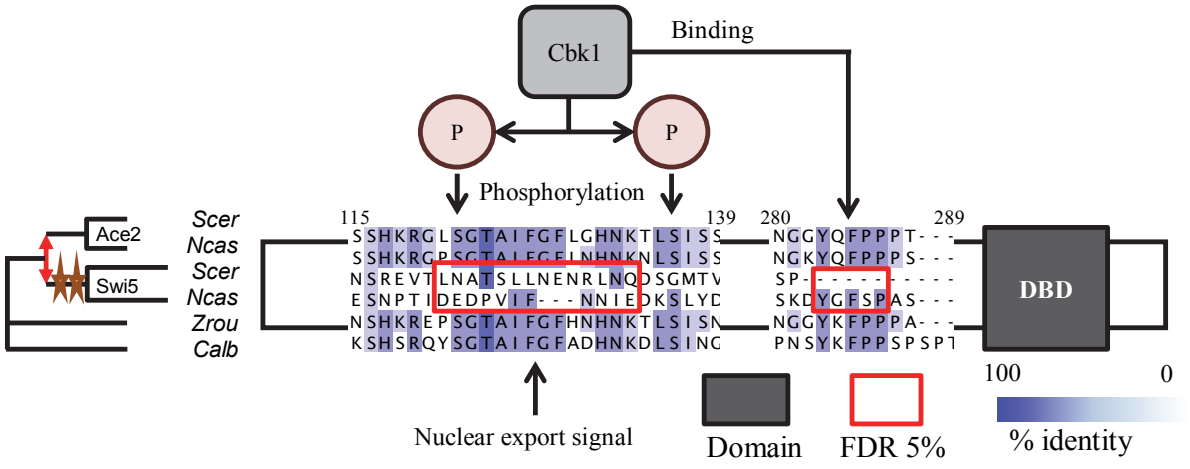


Figure IV-8. Known regulatory motifs with changes in constraints in Ace2/Swi5. Alignment of the short linear motifs with known function (indicated with arrows) and significant changes in constraints (red boxes) after gene duplication from representative species. The Ace2 protein is known to bind and be phosphorylated by Cbk1 kinase at two motifs that have significant changes in constraints after gene duplication. Numbers indicate residue position within the *S. cerevisiae* Ace2 protein. The two identified motifs occur at aa121-134 and aa283-287 in Ace2, and changed constraints within the aligned region aa115-128 and aa247-248 (it is a gap) in Swi5. These overlap with the known phosphorylation site in Ace2 (aa122) and the Cbk1 binding site (aa283-286) in Ace2. Both Ace2 and Swi5 retain their DNA binding domain (DBD). Stars represent significant changes in constraints along the lineage. Red double arrow illustrates the duplication event. aa: amino acid position. Scer: *S. cerevisiae*, Ncas: *N. castellii*, Zrou: *Z. rouxii*, Calb: *C. albicans*.

IV.3.6 Pre-WGD Ace2 localizes asymmetrically

To confirm our sequence-based predictions about evolutionary divergence, we focused on the Swi5/Ace2 paralog pair. Because the ancestral protein likely contained critical regulatory motifs, we hypothesized that it was also regulated by Cbk1, and localized asymmetrically in the daughter cell (Figure IV-8). To test this, we cloned and replaced the *S. cerevisiae* endogenous *SWI5* gene with GFP-tagged Swi5/Ace2 homologs from species before and after the whole-genome duplication and quantitatively assayed their localization pattern using fluorescence microscopy (Figure IV-9A, see Methods).

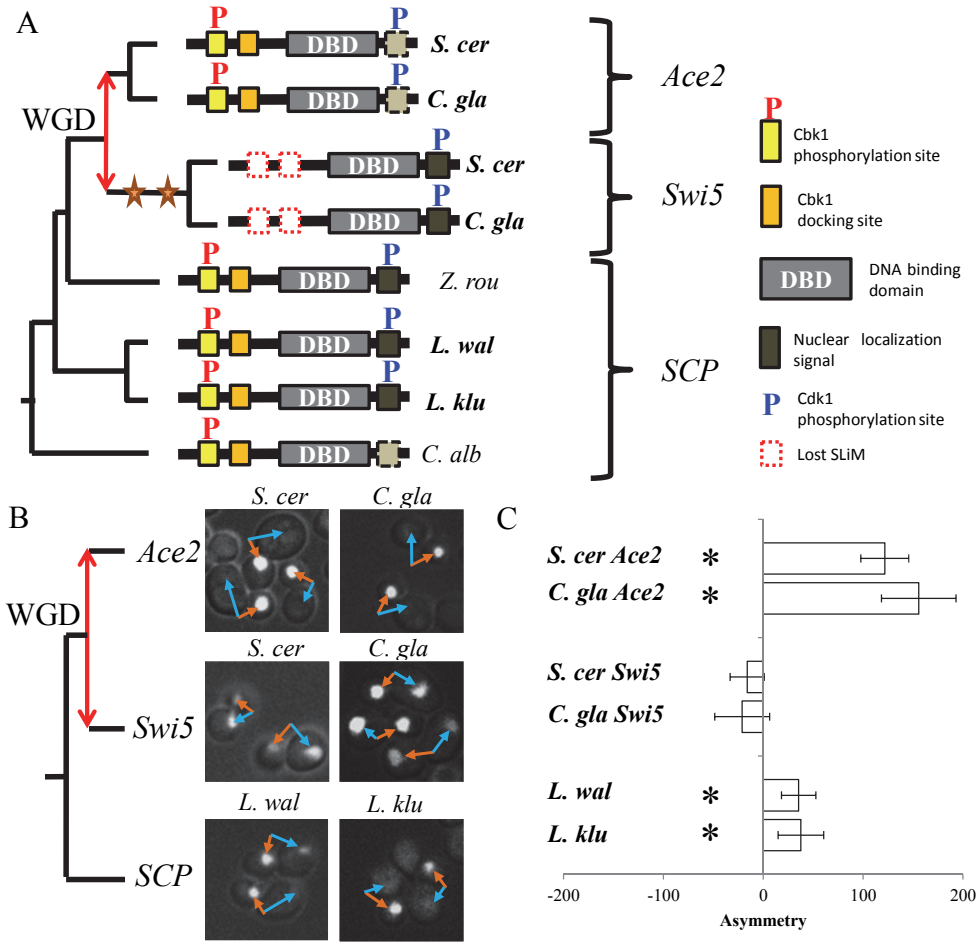


Figure IV-9. Posttranslational change in regulation after gene duplication in Swi5 and Ace2. A) Schematic of the gene tree relating the Ace2/Swi5 paralog pair with diagram of protein features found in proteins from different yeast species. Bolded species name indicate cloned genes assayed for localization in *S. cerevisiae*. The nuclear localization signal characterized in Swi5 is putatively altered and may not be functionally homologous in Candida and Ace2, but this difference was not predicted in our analysis (see Discussion and Appendix Figure III-3). B) Green-fluorescent protein tagged genes cloned from the labeled species were assayed for their localization in unsynchronized *S. cerevisiae* cells. Two representatives of each pre-/post-WGD genes were assayed. Orange and blue arrows indicate representative bud and mother nucleus pairs. C) The fluorescence intensity of the nucleus in cells expressing the labeled proteins was quantified, and mean difference of the intensity (bud-mother) is used as the measure of asymmetry (unfilled bars). Error bars show 95% confidence interval of the mean. Stars indicate 5% statistical significance. Red double arrow illustrates the duplication event. Scer: *S. cerevisiae*, Cgla: *C. glabrata*, Zrou: *Z. rouxii*, Lwal: *L. waltii*, Lklu: *L. kluyveri*, Calb: *C. albicans*.

Upon visual inspection, consistent with our predictions, both single-copy genes localized in an Ace2-like pattern with clear daughter specific localization (Figure IV-9B). To quantitatively compare the localization asymmetry of the retained duplicates and the single-copy proteins, we manually quantified the nuclear fluorescence (see Methods) and computed the difference between fluorescence intensity in bud and mother cells, and used this as measure of asymmetry. While we could not reject the null hypothesis of symmetry in bud and mother cell localization for Swi5, the single-copy proteins and Ace2 showed statistically significant asymmetry, consistent with our visual inspections (Figure IV-9C, p-value < 0.05). The most parsimonious explanation for these results is that the ancestral protein also showed asymmetrical nuclear localization.

Interestingly, we noted that the quantitative measure of asymmetry for the single-copy proteins was not as extreme as the post-duplicate Ace2 (Figure IV-9C). We also observed several cells with clear mother cell GFP localization just as observed for Swi5 (data not shown). Because cells in our assay were not synchronized, and the localization of Ace2 and Swi5 are cell-cycle dependent, we suggests that the localization of the single-copy proteins may actually represent a mixture of the Ace2 and Swi5 localization patterns, and may be more consistent with sub-functionalization of the ancestral function, as opposed to the simple lineage specific losses predicted based on sequence analysis alone (see Discussion).

To confirm our prediction that the changes in regulation were not specific to the *S. cerevisiae* lineage and occurred during the period of rapid diversification immediately following the whole-genome duplication, we also examined the corresponding genes from *C. glabrata* (a budding yeast species that diverged from *S. cerevisiae* after the whole genome duplication) and found similar patterns of localization to *S. cerevisiae*. This supports our prediction that the change in localization in the two paralogs most likely occurred shortly after the gene duplication event (Figure IV-9B,C) and rules out the possibility that the changes we observe are simply due to a problem with expressing foreign proteins in *S. cerevisiae*. Although we cannot rule out more complicated artifacts due to the expression of heterologous proteins, because we observe consistent localization in two proteins that diverged before and two proteins that diverged after the gene duplication, we consider such artifacts unlikely.

Taken together, these experiments support our prediction that the asymmetric localization pattern of Ace2 was present in the single-copy ancestral protein, and this asymmetry was lost after the gene duplication in Swi5 due to losses of specific posttranslational regulatory sequences.

IV.4 Discussion

In this study, we have analyzed the evolution of short linear motifs in protein disordered regions after gene duplication and found that regulatory change is likely to contribute to functional divergence in paralogous genes. An important outstanding question in this analysis is whether the functional changes we identify are adaptive. Previous studies have shown adaptation due to specific changes in posttranslational regulation (Rosso et al. 2008), however general molecular mechanisms for these adaptive posttranslational regulatory change are still under study. The resolution of adaptive conflicts has been suggested as a model for adaptation of paralogous copies of multifunctional genes after duplication (Hittinger and Carroll 2007) and differential patterns of posttranslational regulation could be an example of resolved 'multifunctionality'. For example, in our analysis of the Ace2 and Swi5 paralogous pair, we observed that the asymmetry of the single-copy proteins was reduced when compared to the post-duplicate Ace2 (Figure IV-9C). Although we cannot rule out that these single-copy proteins have other mechanisms within these species that confer daughter specific localization (as we use a heterologous system to test for their localization), we believe that this observation may instead be due to a Swi5-specific motif. Indeed, the characterized nuclear localization signal (NLS) of Swi5 (Moll et al. 1991) was not predicted in our analysis, most likely due to its proximity to the DNA-binding domain, or to the weak conservation of the residues associated with the NLS in the *Candida* species. This NLS of 20 amino acids spans 50 alignment columns within our alignment, and upon close inspection appears to show that the single-copy protein contains high sequence similarity to the Swi5 NLS and that the Ace2 protein and proteins from *Candida* have a more dissimilar one, suggesting that they might not be functionally homologous (Appendix Figure III-3). This hypothesis is consistent with the predominantly Ace2-like localization pattern of the orthologous protein in the *Candida* clade (Kelly et al. 2004). We speculate that this NLS is responsible for the Swi5-like pattern of localization in both Swi5 and the single-copy protein.

Given that Swi5 is known to enter the nucleus slightly before Ace2 and becomes degraded before Ace2 exits the daughter-cell nucleus (Sbia et al. 2008; Di Talia et al. 2009), the observed pattern for the single-copy protein is consistent with first localizing to both mother and bud nucleus as Swi5, and subsequent nuclear export from the mother cell as Ace2. We therefore propose that the differential localization pattern of the Ace2/Swi5 paralogs is a repartitioning of localization of the ancestral protein due to sub-functionalization of the short linear motifs present in the ancestral protein.

In this study, we have identified several putative motifs that have changed constraints within proteins after the whole-genome duplication in budding yeasts. Our methodology to identify changes in evolutionary rate in very small motifs relies on a correction to the distribution of the likelihood-ratio test statistic to control for possible 'protein level' background heterogeneous evolution that can be encountered. These 'protein level' effects, such as changes in protein expression levels (Guan et al. 2007) and divergence due to changes in essentiality or gene function (van Hoof 2005; DeLuna et al. 2008), have been shown to be major issues in evaluating correlated changes in evolutionary rates between interacting proteins (Agrafioti et al. 2005; Swapna et al. 2012). These effects are likely to be encountered in our set of paralogous proteins. Therefore, we ensured that the identification of divergent short linear motifs is unlikely to be caused by these "protein level" effects by correcting the null distribution of the likelihood-ratio test to take account of the whole protein's deviation to the null model assumed by the test. Other methodologies have been previously proposed to empirically obtain the distribution of the likelihood-ratio test statistic (Lanfear 2011). Our approach is similar; however we only estimate one parameter (the non-central parameter) because in our case it sufficiently describes the null distribution. Both approaches (empirically-derived null distribution and estimation of the non-central parameter) have the caveat that they rely on having several data points (in our case alignment columns) that are assumed to be null distributed. An additional constraint of our approach is that it requires that the null distributed data evolves under a shared and constant background heterogeneous evolutionary process to obtain the KL divergence. Therefore, it cannot accurately produce an adequate null distribution under cases where recombination has occurred in a gene, for example. Nevertheless, this approach can be simpler and faster than the permutation tests when performed on genome-wide data where we expect a small

proportion of tests to reveal functional divergence. We believe that the non-central chi-squared null-distribution can be applied to other important tests in molecular evolution where genome-scale data are available and where the assumptions of the chi-squared distribution of the likelihood-ratio test statistic are violated; however this is still under study.

Our study on short linear motifs reveals that posttranslational regulatory evolution is widespread after gene duplication. However, an important limitation of our study is that it cannot identify novel regulatory sequences that have appeared along any lineage or that occur within structured regions, in part due to the way motifs are predicted. Additional genomic sequences such as population data or from additional post-WGD species may allow further analyses of functional changes in the budding yeast after gene-duplication. These types of analyses are likely to uncover even more functional variations between paralogous proteins than were suggested by protein-wide and motif-wide analyses.

Nevertheless, our results are consistent with several results suggested by other studies (Amoutzias et al. 2010): posttranslational regulatory change may underlie an important number of observed functional differences between paralogous proteins. This appears analogous to the models of functional divergence after gene duplication due to transcriptional regulatory change (Force et al. 1999). These parallels between transcriptional and post-translational regulatory evolution (Moses and Landry 2010) suggest that transcription factor binding sites in non-coding DNA are analogous to short linear motifs in proteins. In the former, the rapid transcriptional regulatory evolution is facilitated by the rapid evolution and lack of strong constraints on non-coding DNA. In the case of post-translational regulatory evolution, because short linear motifs are typically found in protein disordered regions which evolve rapidly due to lack of structural constraints, changes in motifs in disordered regions may be a general means to facilitate functional evolution (Neduva and Russell 2005).

IV.5 Materials and Methods

IV.5.1 Alignment of related species of yeasts

We based the orthology assignment on the data from the Fungal Orthogroups Repository (Wapinski et al. 2007) because it contained both sequences from Candida species and

budding yeasts. Protein sequences and orthology assignment from six Candida yeast species [*Candida tropicalis*, *Candida albicans*, *Candida parapsilosis*, *Candida lusitaniae*, *Candida guilliermondii*, *Debaryomyces hansenii*] were obtained from the Fungal Orthogroups Repository. When several protein sequences form the Fungal Orthogroups Repository were mapped to a single budding yeast orthology group, only the most similar sequence as assessed by blast scores was chosen. The six Candida genes, along with the *Saccharomyces cerevisiae* gene, were supplemented with protein sequences and orthology assignment from 19 additional related budding yeast species [*Saccharomyces mikatae*, *Saccharomyces bayanus var. uvarum*, *Saccharomyces kudriavzevii*, *Candida glabrata*, *Kazachstania Africana*, *Kazachstania naganishii*, *Naumovozyma castellii*, *Naumovozyma dairenensis*, *Tetrapisispora blattae*, *Tetrapisispora phaffii*, *Vanderwaltozyma polyspora*, *Zygosaccharomyces rouxii*, *Torulaspora delbrueckii*, *Kluyveroymces lactis*, *Eremothecium gossypii*, *Eremothecium cymbalariae*, *Lachancea kluyveri*, *Lachancea thermotolerans*, *Lachancea waltii*] that were obtained from the Yeast Gene Order Browser (Byrne and Wolfe 2005). 452 alignments of retained duplicates and 3566 alignments of single-copy proteins were used in our analysis.

Protein sequences were then aligned using MAFFT v6.864b with the --auto flag at default settings (Katoh et al. 2002).

IV.5.2 Conserved segment prediction

We sought to predict small functional regions that could be labeled as short linear motifs. Because we were interested in functional segments that could be identified before the whole-genome duplication (Kellis et al. 2004), we first removed from the multiple sequence alignment the sets of proteins from species that had undergone the whole-genome duplication and predicted short linear motifs within the remaining species (which we refer to as the 'pre-WGD clade'). To identify short linear motifs, we used a phylogenetic hidden Markov model (phylo-HMM) (Nguyen Ba et al. 2012). Briefly, this method identifies highly conserved short amino acid sequences within disordered regions of proteins. The unstructured regions are predicted by DISOPRED2 (Jonathan J Ward et al. 2004), filtered for coiled coils using pFilt (Jones et al. 1994) and for repetitive regions using the SEG algorithm (Wootton and

Federhen 1993). We also use the phylo-HMM to filter out large conserved regions as we consider them likely to be structural regions. In a previous study, the phylo-HMM approach identified 104 of 352 known motifs with a false positive rate of 1 in 9000 amino acids (Nguyen Ba et al. 2012).

In addition to the heuristics used in (Nguyen Ba et al. 2012), we now also assume that a scaling factor of rates of evolution within the conserved state is sampled from a discretized Gamma distribution with eight categories (Yang 1994) with a fixed alpha and beta parameter of 0.6, which was chosen as a heuristic that allowed predictions of large conserved regions (>35aa) interspersed by a few fast evolving columns. We now obtain the rates of evolution through a Newton-Raphson procedure, and used a window size of 31 alignment columns for the calculation of the background rate.

Because the phylo-HMM tends to classify single insertion/deletion events as slow evolving regions, motifs are trimmed on either ends to remove regions that are over 50% gaps or is filtered out if the prediction itself contains over 50% gaps.

Flanking regions of the predicted conserved segments consisted of five alignment columns on each side.

IV.5.3 Likelihood-ratio test of multiple rates of evolution

We sought to systematically identify short linear motifs that evolve at a different rate after the whole-genome duplication. To do so, each predicted motif from the pre-WGD clade was mapped back into the complete alignment.

Each predicted motif was then analyzed using the PAML package (Yang 2007) by a likelihood-ratio test that compares the null hypothesis (H_0) that motifs before and after the whole-genome duplication are evolving at the same rate, to a model (H_1) with two distinct rates (Yoder and Yang 2000) (PAML program: AAML, clock=2, cleandata=0, fix_omega=0, ncatG=8). Likelihood-ratio tests have been previously used to study the evolution of the yeast paralogs generated in the WGD (Byrne and Wolfe 2007). Our test differs from this previous application of the likelihood-ratio test, because we compared the evolutionary rate

on each paralogous clade (post-WGD_1 and post-WGD_2) to the evolutionary rate on the lineages that diverged before the whole-genome duplication (pre-WGD) one at a time. Formally, the likelihood-ratio test (LRT) is:

$$LRT = 2\log LR = 2\ \log\frac{P(data|\widehat{H_1})}{Q(data|\widehat{H_0})} = 2\ \log\frac{P(data|1=\alpha_{pre-WGD},\hat{\alpha}_{post-WGD})}{P(data|1=\alpha_{pre-WGD}=\alpha_{post-WGD})}$$

where the data corresponds to the motif segment within the multiple sequence alignment, and α_{clade} represents the rate for corresponding clades. In this model (Yoder and Yang 2000), α is a scaling factor by which the estimated branch lengths are multiplied, and one of the rates always defaults to 1. Therefore, under the null hypothesis H_0, the single rate is equal to 1, while the alternative hypothesis H_1 allows one of the two rates to be different than 1 and it is estimated by maximum likelihood. Because these models are nested, under the null hypothesis H_0, the distribution of the likelihood-ratio test statistic (LRT) follows the chi-squared distribution with degrees of freedom equal to 1 (Wilks 1938) (see the next Methods section for the correction to the chi-squared distribution performed when assumptions of the test are violated). Although it is in principle possible using this test to find short linear motifs that evolve either slower or faster than the proteins in which they are found, because short linear motifs are predicted on the basis of their conservation in the pre-WGD clade, we only expect to identify motifs with faster rates of evolution after the whole-genome duplication.

We estimated the false discovery rate using a slight modification of the procedure described in (Storey and Tibshirani 2003) to obtain a threshold for significant p-values. We modified this approach because when applying the LRT described above to our alignments of the yeast proteome, we observed a large number of tests resulting in LRTs of exactly zero (thus having a p-value of 1, e.g. Figure IV-3), many of which correspond to motifs where no information can be inferred about their rate of evolution. For example, in our real data, for 284/498 of these LRTs of exactly zero, we observed no amino acid differences in the multiple alignments and therefore have no power to estimate a change in evolutionary rate. Because we observed that p-values between 0.6 and 0.95 appeared uniform as expected for the distribution of truly null p-values, we used this range only to estimate the false discovery rate (FDR). We counted 1836 p-values between 0.6 and 0.95 out of a total of 7709 tests. If we

assume that all these p-values correspond to truly null hypotheses, then we can estimate the proportion of null tests (π_0) by 1836/(7709*(0.95-0.6)) = 0.6804. The FDR at p-value threshold *t* is therefore estimated as:

$$FDR(t) = \frac{\pi_0 7709t}{\#\{p_i \leq t\}}$$

We considered p-values as significant where this FDR is lower than 0.05.

IV.5.4 Correction for data heterogeneity due to violations of model assumptions about protein evolution

Increased evolutionary rate after gene duplication is frequently observed in entire proteins (Scannell and Wolfe 2008). We reasoned that short linear motifs within these proteins may also show the same changes in protein-level selective constraints. Furthermore, because mutations may not be homogeneous over the phylogeny (e.g., due to lineage specific changes in GC content), proteins might show biases in their substitution process that are not accounted for by the models assumed in the LRT. Because we were interested in short linear motif evolution, we wished to test for *additional* changes in motifs using the heterogeneity of protein evolution as the "background". In this case, we can still compute the LRT statistic, but the test statistic no longer follows the standard chi-squared null distribution because the heterogeneity in rates and patterns of protein evolution can be 'fit' using the additional parameter in the alternative hypothesis. This biases the test to reject the null hypothesis and leads to detection of false positives. A permutation test has been proposed for this case (Lanfear 2011) however, in our case, this test must be performed for each individual predicted motif, and these permutation tests may lack power for genome-wide analyses. We therefore devised another strategy by which we can approximate the distribution of the LRT statistic under a heterogeneous background process in protein evolution.

We assume that evolution of each alignment column is independent and is possibly evolving under a heterogeneous background process after the whole genome duplication event. This heterogeneity that affects the whole protein could be due, for example, to changes in expression level, lineage-specific changes in GC content or alignment errors. The likelihood

of the data generated under this scenario can be computed under the alternative hypothesis H_1 where there has been a change in constraints P(data|H_1), or under the 'null hypothesis' where evolutionary rate has remained constant, Q(data|H_0). We note that H_1 can capture only some of the true heterogeneity in the data using the additional rate parameter, and the null model H_0 captures even less. If θ is a parameter space and β the possible values of those parameters, then there may exist sets of values β* in the parameter space of the alternative hypothesis θ_{H1} that captures some of this heterogeneity and that cannot be captured by the values β_0 in the parameter space of the null hypothesis (θ_{H0}). Although this heterogeneous background process does not produce data following a generative process with parameters and values β*, we only seek the extra 'fit' obtained from the parameter space θ_{H1} that cannot be captured by the parameter space θ_{H0}.

This fit can be summarized by the expectation of the log-likelihood-ratio of the two models, where the expectation is taken using the probabilities P, which is the Kullback-Leibler (KL) divergence D_{KL}(P||Q). This measures the additional amount of deviation of the possibly heterogeneous background captured by the alternative hypothesis relative to the null hypothesis.

$$\int \log \frac{P\left(data \middle| \theta_{H_1} = \beta^*\right)}{Q\left(data \middle| \theta_{H_0} = \beta_0\right)} P\left(data \middle| \theta_{H_1} = \beta^*\right) = D_{KL}(P||Q)$$

In practice, we cannot necessarily parameterize the heterogeneity in the background evolutionary process, for example if it is due to alignment errors (i.e. it is difficult to estimate β* or how data is generated from this heterogeneous process). Nevertheless, the distribution of the likelihood-ratio test statistic (LRT) when we test the alternative hypothesis H_1 vs H_0 (by maximizing the 'fit'), is related to the KL divergence as follows. Given that the data used to compute the LRT are truly drawn from P, the distribution of the likelihood-ratio test statistic converges to a data-dependent *non-central* chi-squared distribution, $\chi^2(k,\lambda)$, parametrized by the "non-centrality parameter" λ and the degrees of freedom k. The non-centrality parameter is given by $\lambda = 2\, L\, D_{KL}(P||Q)$, where L is the number of data points used in the LRT (van der Hoeven 2005). To estimate D_{KL}(P||Q), we note that the mean of the LRT

when data is drawn from P must be equal to the mean of the non-central chi-squared, which is given by $k + \lambda$. Therefore,

$$E[LRT] = \frac{2}{L}\sum_{i=1}^{L} \log\frac{P(X_i|\widehat{H_1})}{Q(X_i|\widehat{H_0})} = k + 2LD_{KL}(P||Q)$$

where X_i is the data at an alignment column i, k is 1 in our case and L in our case is the number of alignment columns.

Under the assumption of independence between alignment columns, D_{KL} can be estimated from the whole alignment using a single likelihood-ratio test, which we believe is reliable since L is the number of alignment columns in the whole protein and is typically large, and we assume that the background process operates uniformly over the alignment columns. Therefore, we let $E[LRT] = LRT_{protein}$ and use:

$$D_{KL}(P||Q) \approx \frac{LRT_{protein} - k}{2L}$$

We note that because the motif is small in comparison to the whole protein (which we use to estimate P), its contribution to the calculation of D_{KL} is small and unlikely to affect the results.

While the expectation of the likelihood-ratio test statistic (E[LRT]) is always greater or equal to the degrees of freedom k, the obtained likelihood-ratio test statistic for a single protein $LRT_{protein}$ may be smaller than k, especially when D_{KL} is small. In these cases, we assume that D_{KL} is equal to the parameter estimated for proteome-wide (species) evolution (see below).

We note that P = Q implies $\beta^* = \beta_0$, which indicates that the data has no source of background heterogeneity that is better captured by the alternative hypothesis than by the null hypothesis. In that case, D_{KL} is zero and this approach simplifies to the standard chi-squared distribution. Further, although it is possible to formulate a likelihood-ratio test with estimated β^* as the values of the parameters of the null hypothesis (akin to modeling more complex evolutionary processes in the test), there are several advantages of modeling the

extra 'fit' instead. First, it is a single value, and second, it is directional (such that rejection of the null hypothesis occurs when values of the parameters are farther from β_0 than from β^*).

This estimate of the non-centrality parameter gives us a new null distribution for the LRT statistic for the predicted motifs in each protein. Since these motifs are short segments chosen from the entire alignment, we can compute the probability of having observed an LRT statistic as extreme (or more) in a short segment, given the length of the motif and the null distribution estimate for that protein. Therefore, the p-value for each motif, *m*, is given by the non-central chi-squared with 1 degree of freedom and non-centrality λ_m.

$$\lambda_m = 2\, L_m\, D_{KL} = L_m \frac{LRT_{protein} - 1}{L}$$

where L_m is the length of the short linear motif. A closed-form solution exists, which we used, for the cumulative distribution of the non-central chi-squared with one degree of freedom:

$$P(LRT \leq LRT_m | \lambda_m) = \frac{\mathrm{erf}\left(\frac{\sqrt{LRT_m} - \sqrt{\lambda_m}}{\sqrt{2}}\right) - \mathrm{erf}\left(\frac{-\sqrt{LRT_m} - \sqrt{\lambda_m}}{\sqrt{2}}\right)}{2}$$

Where erf is the error function, LRT_m is the LRT statistic computed (by PAML) for the motif *m*, and λ_m is as above. In more general cases (i.e. $k > 1$), this computation can be performed using several algorithms (see e.g. (Benton and Krishnamoorthy 2003)).

We also noticed that the species used in our study appeared to evolve in a manner that differed from the single rate of evolution null hypothesis (H_0), even for single-copy proteins. To correct for this additional source of heterogeneity, we estimated another D_{KL} parameter using the whole proteome to rule out any effect on the short linear motifs that could be explained simply by species-level evolution. This D_{KL} parameter was estimated to be 0.014552523. We therefore obtained two D_{KL} parameters for each motif, and because we wanted to correct for rate differences which could be explained by genome-wide deviation or the individual protein's deviation, we chose the larger parameter while computing the p-values. This chooses the larger p-value, for which we believe no additional multiple-testing

correction needs to be performed (in that we believe we are still performing only one test per motif) and allows us to perform a likelihood-ratio test using the standard tools for molecular clock hypothesis testing. Importantly, this global correction means our p-values are always more conservative than the significance values obtained using the standard central chi-squared distribution.

IV.5.5 Simulation of protein evolution

To simulate more 'realistic' protein evolution (Figure IV-2), we use a similar simulation program as in (Nguyen Ba et al. 2012). We evolve sequences to closely mirror our protein alignments by using every protein in our analysis as a template for a simulated protein. First, AAML is used on every protein alignment to obtain protein-specific branch lengths for the phylogenetic tree (we use the species tree for all proteins). The root sequence is one of the sequences of the alignment (we chose the protein sequence of median length), and a site-specific rate of evolution for each amino acid is inferred by the phylogenetic hidden Markov model, which we use as a scaling factor to evolve the root according to the branch lengths obtained by AAML. Indels are generated as in (Nguyen Ba et al. 2012) but site specific rates are propagated to indels, such that insertions have the same rate of evolution as the amino acid positions that created it. To ensure that the sequences were as realistic as possible, we also use two amino acid substitution models: one for ordered regions, and one for disordered regions. These two models differ by their equilibrium, or stationary frequencies, of the 20 amino acids, which is estimated based on DISOPRED2 predictions on the *S. cerevisiae* proteome. The exchangeabilities of amino acid pairs was estimated as a whole on closely related species as in (Nguyen Ba et al. 2012). Because the rate matrix is a product of the stationary frequencies with the exchangeabilities of amino acids (Whelan and Goldman 2001), the substitution matrix for disordered and ordered regions will tend to create amino acids found in disordered and ordered regions, respectively. These stationary frequencies of amino acids are also used in the production of insertions.

We assigned ordered or disordered regions in the root sequence, and propagated them across the phylogenetic tree. Finally, to ensure that some motifs can be predicted, we do not allow indels within regions that have been predicted as motifs in the ancestor. Therefore, no site

specific changes in constraints are ever simulated but our simulated proteins are evolved according to estimated phylogenetic trees with two different substitution processes (and therefore two different stationary frequencies of amino acids), and with indels. After alignment by MAFFT, the full pipeline used to predict short linear motifs and calculate the likelihood-ratio test is then used on the full set of simulated proteins. In principle, none of the motifs are intended to have lineage-specific changes in constraints. However, in practice, computational artifacts may occur during the simulation (such as misalignments, deletions of motifs within a clade, mispredictions of short linear motifs) and these can cause signatures of type I functional divergence. Deletions causing a motif to be removed in one of the lineage are computational artifacts of the simulation because they are unintended; however they also would represent genuine changes in constraints on the motif. However, misalignments and mispredictions of short linear motifs are actual computational artifacts that can also occur within our data. Using this set of simulated proteins, it is therefore possible to conservatively assess how many of the predicted changes in constraint can be explained by these computational artifacts or by incorrect non-central parameter estimation for the null distribution of the likelihood-ratio test statistic.

IV.5.6 Test of correlated evolution

We define correlated evolution to be a tendency for changes in constraints on several functional sequences to occur within only one of the two paralogous proteins. Our test for correlated evolution cumulated the number of conserved segments with changes in constraints within each of the paralogs and asked whether the changes occurred more in a particular direction than expected by chance. Under the null hypothesis, the expected difference in the number of motifs changing in one direction minus the other on one protein should be zero. The sum of all the differences is used as the final test statistic, for which a p-value was obtained by a non-parametric permutation test.

To correct for the possibility that the phylo-HMM mistakenly separated a functional fragment as two motifs due to rapid evolution between the regions, we counted multiple motifs that were close to each other (within 35aa) and that had accelerated evolution on the same lineage as a single motif for the purpose of this test.

IV.5.7 Strains and plasmids

BY4741 or isogenic derivatives were used for all of our experiments. Single-copy genes were PCR amplified from purified genomic DNA (Fermentas, #K0512) of *L. kluyveri* (NRRL Y-12651) and *L. waltii* (UCD 72-13), and Ace2/Swi5 orthologs were PCR amplified from purified DNA from *C. glabrata* (CBS 138). Allele replacement for single-copy genes was performed using a modification of the method as in (Li et al. 2011) with single-copy genes replacing the *SWI5* gene. Briefly, the 3'UTR of *ACE2* or *SWI5* was cloned into pFA6a (Wach et al. 1994) (PCR primers P7/P8 and P9/P10 respectively), after which genes of interests were cloned upstream (Ace2(L.klu) – PCR primers P34/P35; and Ace2(L.wal) – PCR primers P36/P37). Two PCR fragments were then transformed to target the *SWI5* locus (PCR primers CaURA3MX: P29/P44, Δ*swi5*::*ACE2(L.klu)* P32/P31, Δ*swi5*::*ACE2(L.wal)* P46/P31). The selection marker used for our experiments was the CaURA3MX cassette, which allowed subsequent marker removal using 5-FOA (Boeke et al. 1987). Once the marker was removed, all strains contained precise gene replacement of *SWI5* and these were then tagged with monomeric yeast-enhanced GFP using the same method as in (Huh et al. 2003) with either the CaURA3MX or KanMX4 resistance marker instead of the HIS3MX4 (Ace2(S.cer) [YBS31] – PCR primers P5/P41, Swi5(S.cer) [YBS32] – PCR primers P2/P44, Ace2(L.klu) [YBS14] – PCR primers P12/P44 and Ace2(L.wal) [YBS13] – PCR primers P11/P44). For the *C. glabrata* orthologous genes, gene tagging and allele replacement was performed in a single step by transforming two fragments: one containing the gene with homology to the S288C genome on the 5' end and to the GFP cassette on the 3' end, and one containing the GFP with homology to the S288C genome on the 3' end (Δ*ace2*::*ACE2(C.gla)* [YBS17] – PCR primers P38/P40 and P40/P41, and Δ*swi5*::*SWI5(C.gla)* [YBS18] – PCR primers P42/P43 and P40/P41). All strains were verified using genomic PCR and sequencing of the homologous recombination junctions. Aberrant cell morphology was observed for Δ*swi5*::*SWI5(C.gla)* (the cells appeared larger than normal and this phenotype did not appear to be due to the Δ*swi5* deletion based on comparison to Δ*swi5* deletion mutants (data not shown)).

All primer sequences and strains used for these experiments are included in Supplementary data table IV-S2.

Strains were then imaged by growing the cells to log-phase in minimal defined media with appropriate auxotrophic requirements and imaged with a standard 491nm blue laser on a Leica spinning-disc confocal microscope.

IV.5.8 Localization analysis

We wished to test that the localization of Ace2/Swi5 homologous proteins differed by quantifying the intensity of the green fluorescent protein with respect to bud or mother nuclei. We chose to quantify solely the nuclear intensity as these proteins are transcription factors known to shuttle to the nucleus during the cell cycle, and show distinct patterns of nuclear localization (Sbia et al. 2008). To obtain normalized fluorescence intensity, images were analyzed by manually quantifying the cell and nuclear median green fluorescence. Cell size in pixel count was also quantified in this manner and was used to identify the daughter cells. The difference in fluorescence intensity between the bud and mother cell was used as the index of asymmetry. Cells where the median fluorescence intensity observed was over 240 were discarded as they were potentially too saturated to obtain reliable measures. Statistical significance was calculated using a Z-test.

To determine statistical significance when testing for association between changes in constraints and localization differences as determined by (Marques et al. 2008), we asked whether the observed fold increase in rate of motif changes in constraints was higher than random permutations of the ‘different’ and ‘similar’ labels of localization.

IV.6 Acknowledgments and funding information

We are grateful for the *L. kluyveri*, *L. waltii* and *C. glabrata* strains that were generously donated by Dr. Marc-André Lachance. We thank Brenda Andrews for access to the microscope, as well as lab members of the Andrews lab for useful help during the project. We also thank Moses lab members, Dr. Nicholas Provart and Dr. Philip Kim for useful feedback during the course of this study. ANNB is funded by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC). JJH was funded by an undergraduate student research award from NSERC. IG and CL are funded by a Canadian Institute of Health Research grant (grant GMX-191597). AMM is supported by a

NSERC Discovery grant and Canadian Institutes of Health research (grant MOP-119579). This research was supported by infrastructure grants from the Canadian Foundation for Innovation to AMM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

IV.7 Software availability

The updated phylo-HMM and simulation programs can be found at www.moseslab.csb.utoronto.ca/phylo_HMM and in the supplementary data files.

IV.8 Author contributions

ANNB designed and performed the computational and Ace2/Swi5 related experiments, and wrote the paper. BS, JJH, JD performed the Ace2/Swi5 related experiments. IG and CRL performed experiments that were not ultimately included in the manuscript, and provided discussion and comments for the manuscript. ELW provided comments and useful discussions throughout the project. AMM designed the computational and Ace2/Swi5 related experiments, and wrote the paper.

IV.9 Supplementary material

Supplementary text and figures can be found in Appendix III and supplementary tables are included in the supplementary data.

Chapter V

Experimental Evidence for Non-adaptive Increases in Complexity in an Ancient Eukaryotic Regulatory Network

This work has not been previously published.

Alex N Nguyen Ba[1,2], Sergio Peisajovich[1], Alan M Moses[1,2,¥]

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, Toronto, Canada
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada

V.1 Abstract

Cellular processes are frequently conserved across multiple domains of life, yet the individual components, and the circuits regulating them, often change over evolution. These circuits can increase in complexity over evolution though gene duplication and divergence of regulatory genes, and there are two models for how this happens: 1) adaptive changes after gene duplication, such as resolution of adaptive conflicts, and 2) non-adaptive processes such as duplication, degeneration and complementation. Both of these models predict complementary changes in the retained duplicates, but they can be distinguished by direct fitness measurements in organisms with short generation times. Previously, it has been demonstrated that repeated sub-functionalization of an essential protein in the spindle checkpoint pathway has occurred multiple times over the eukaryotic tree of life. Here, we use comparative genomics approaches to guide the reconstruction of the ancestral spindle checkpoint network and perform high-throughput and systematic quantitative fitness measurements of evolutionary intermediates during gene network rewiring after gene duplication. We show that, at the resolution of our assay, no fitness advantage of the rewired network can be detected when compared to the ancestral network. We also find stepwise mutational paths from the simpler to more complex network with no fitness defects. Our results indicate that, even in cases of parallel evolution which has been taken as strong evidence for natural selection, increase in network complexity after gene duplication may be explained by neutral processes.

V.2 Introduction

The molecular details of well-conserved cellular processes often differ surprisingly between distant organisms. For example, kinetochore assembly in eukaryotes is divided into two types: monocentric attachment and holocentric attachment (Dernburg 2001) and it is currently unclear whether one is more advantageous than the other. As with the kinetochore assembly process, the spindle-checkpoint pathway is divided into two distinct molecular mechanisms (Suijkerbuijk et al. 2012). In most laboratory studied organisms, the spindle checkpoint pathway consists of the paralogous Bub1 and Mad3 proteins. However, other organisms have a single protein that presumably can perform both Bub1 and Mad3 function. Gene duplication and sub-/neo-functionalization are processes that, not only increase genomic complexity, but are thought to be the major sources of genetic novelty in organisms (Conant and Wolfe 2008). Particularly interesting are cases of repeated, parallel evolution of genes that has been taken to be strong evidence for natural selection (several examples reviewed in (Gompel and Prud'homme 2009)). Most striking is that this single protein that performs both Bub1 and Mad3 function in the spindle checkpoint pathway has been observed to duplicate at least nine times throughout the tree of life, always leading to Bub1 and Mad3 functional homologs. This increase in complexity has therefore been thought to be adaptive (Murray 2012; Suijkerbuijk et al. 2012). However, theoretical work suggests that a causal link between increased genetic complexity and adaptation may not be as prevalent as commonly assumed (Lynch 2007a; Lynch 2007b). For example, if degeneration and complementation of the ancestral bi-functional protein must always lead to Bub1 and Mad3 functional homology, then our repeated observation of Bub1 and Mad3 might be due to a rate of degeneration that is higher than the rate of reconstitution.

Whether sub-functionalization of a bi-functional protein is adaptive or neutral cannot be easily distinguished by sequence analysis alone. However, precise quantitative fitness measurements in organisms with short generation time, such as yeasts, can be used to address questions about these two models of sequence evolution (adaptive vs neutral). Using comparative approaches that delineate functional regions of proteins, we found that this gene duplication leads to sequence signatures that indicate repartitioning of the ancestral function

in the extant paralogs. To test whether the increase in genome complexity is adaptive, we employ high-throughput quantitative fitness measurements and we systematically dissect the evolutionary trajectory of an ancestral bi-functional gene. To our surprise, we could not detect any increase in fitness for the stepwise increase in complexity (duplication, degeneration and complementation) from a simpler ancestral spindle checkpoint pathway.

Our results provide evidence for neutral evolution of a simpler to a more complex regulatory network, even in the case of parallel evolution.

V.3 Results

V.3.1 Subfunctionalization in the spindle checkpoint network

Preservation of duplicate genes as explained by the duplication-degeneration-complementation model is a neutral process by which repartitioning of the functions of the ancestral protein occurs within the two extant duplicates (Force et al. 1999). Although sub-functionalization is a neutral process, it can lead to adaptation in the case of adaptive conflicts: mutations that are precluded from occurring in the 'ancestral' gene but are adaptive when the functions of the ancestral gene are separated, such as in the case of mutations that 'specialize' multifunctional proteins (Hittinger and Carroll 2007). Even in the cases where mutations leading to neo-functionalization have occurred, we would still expect at least that one of the duplicate genes retain protein sequences that are responsible for the functional roles of the ancestral protein, whether or not they are repartitioned.

It has previously been shown that Bub1 and Mad3 sub-functionalization had occurred several times over the eukaryotic tree of life, leading to similar sequence profiles (Figure V-1A,(Suijkerbuijk et al. 2012)). To obtain an amino acid resolution of the evolution of the paralogous genes, we performed sequence analysis of the duplication that occurred in the whole-genome duplication of budding yeasts (see Methods). This analysis revealed several protein regions, in addition to the KEN boxes and kinase domain identified in previous studies (Suijkerbuijk et al. 2012))., that have repartitioned from the ancestral sequence to the paralogs (Figure V-1B).

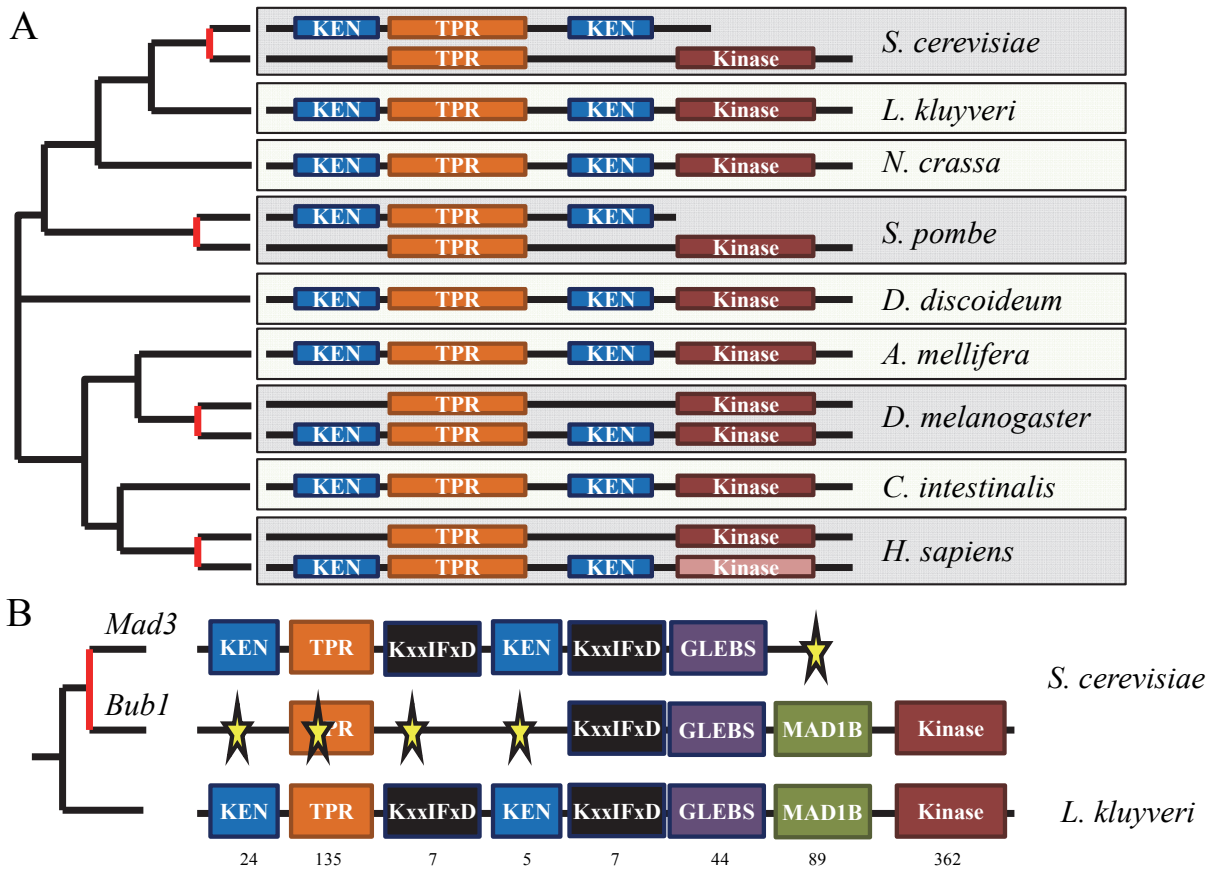


Figure V-1. Sequence analysis of the Mad3/Bub1 paralog. A) Independent duplication events in the bi-functional gene lead to similar domain arrangements. Red line on the phylogenetic tree indicates an inferred duplication event (species highlighted in dark grey). Several species retain the bi-functional gene (species highlighted in light grey). Schematics are not to scale. B) Amino-acid resolution analysis reveals several regions with changes in constraints in both Bub1 and Mad3 proteins. KEN: KEN-box, TPR: Tetratricopeptide-domain, KxxIFxD: Unknown motif, GLEBS: Gle2-binding-sequence-domain (binds Bub3), MAD1B: Mad1-binding region, Kinase: Kinase-domain. Numbers indicate the amino acid size of each segments of *L. kluyveri*. The whole protein is 982 amino acids.

This more detailed view of the changes in constraints allowed us to propose why some of the changes were correlated (Figure V-2A for an example). For example, according to 3D structural information of Mad3 (Figure V-2B), specific residues in the TPR domain (shown in yellow in Figure V-2B) appear to contact the KEN box (shown in red in Figure V-2B) and may stabilize the interaction of Mad3 with Cdc20 (Figure V-2C). This indicates that the degeneration of residues in the TPR domain or the KEN box will disrupt the same function

(binding to Cdc20) and that loss of selection constraints on that function will lead to degeneration of both. Indeed, the same correlated changes in constraints as exemplified is also observed in the mammalian Bub1 and Mad3, which occurred through independent gene duplication event (results not shown). We speculate that the second pair of motif losses (the first KxxIFxD motif and the second KEN box) is also structurally related: the KxxIFxD sequence may orient the KEN box to the WD40 repeat containing Cdc20 (Tian et al. 2012). This is consistent with the conservation of the second KxxIFxD motif next to the GLEBS domain as it may help orient the GLEBS domain to the WD40 repeat containing Bub3 (Larsen and Harrison 2004). Therefore, the KxxIFxD motif may be a general WD40 repeat binding motif or it may be used to kink the disordered region in a specific orientation for the adjacent short linear motif.

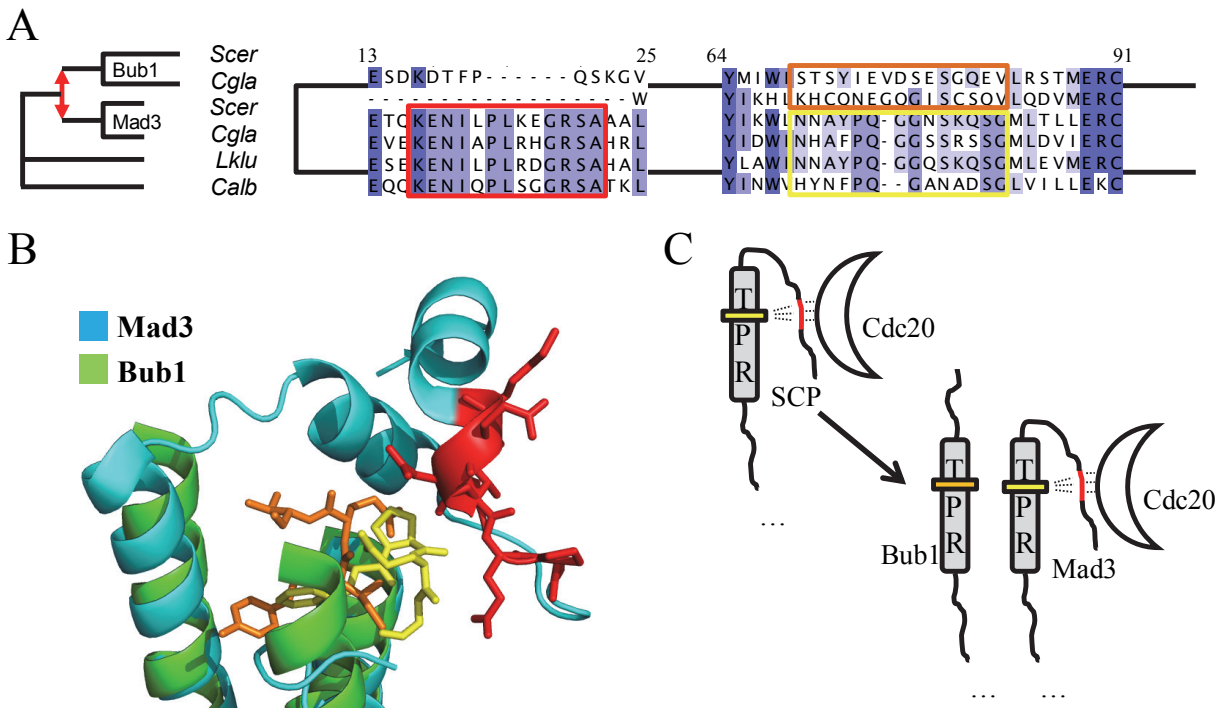


Figure V-2. Amino acid resolving power of changes in constraints are biologically relevant. A) Alignment of an example change in constraint in the Bub1 lineage. Numbers indicate residue position within the *S. cerevisiae* Bub1 protein. B) Structural alignment of the region shown in A). The TPR domain of Bub1 (green) and Mad3 (cyan) are shown, along with the N-terminal tail of Mad3. Regions highlighted in red, orange and yellow are the same as shown in A). C) Schematic representation of the binding interaction of Mad3 with Cdc20 can help elucidate why the regions shown in A) have correlated evolution.

Consistent with important functions for the domains identified in the ancestral protein, there were no conserved regions that had changes in constraints in both paralogs. However, the repartitioning of function suggests that the duplication lead to sub-functionalization. We therefore sought to experimentally verify that sub-functionalization had occurred in the paralog pair. As a proxy for the 'ancestral gene', we used the gene from *Lachancea kluyveri* (which we refer to as the "single-copy protein"), which diverged prior to the whole-genome duplication event, and integrated the gene within *S. cerevisiae*. We first assessed the localization of the ancestral protein because it was known that Bub1 and Mad3 localize to

different subcellular compartment: Bub1 is localized to the kinetochores in a Mps1 dependent manner during specific phases of the cell-cycle (London et al. 2012) and Mad3 is constitutively localized in the nucleus (Hardwick et al. 2000). If Bub1 and Mad3 were products of a sub-functionalization event, we would also expect the *L. kluyveri* protein to localize to both subcellular compartments. Consistent with this, we observed that the *L. kluyveri* protein localized constitutively to the nucleus, with distinct re-localization of a subset of proteins to the kinetochore during the cell-cycle (Figure V-3).

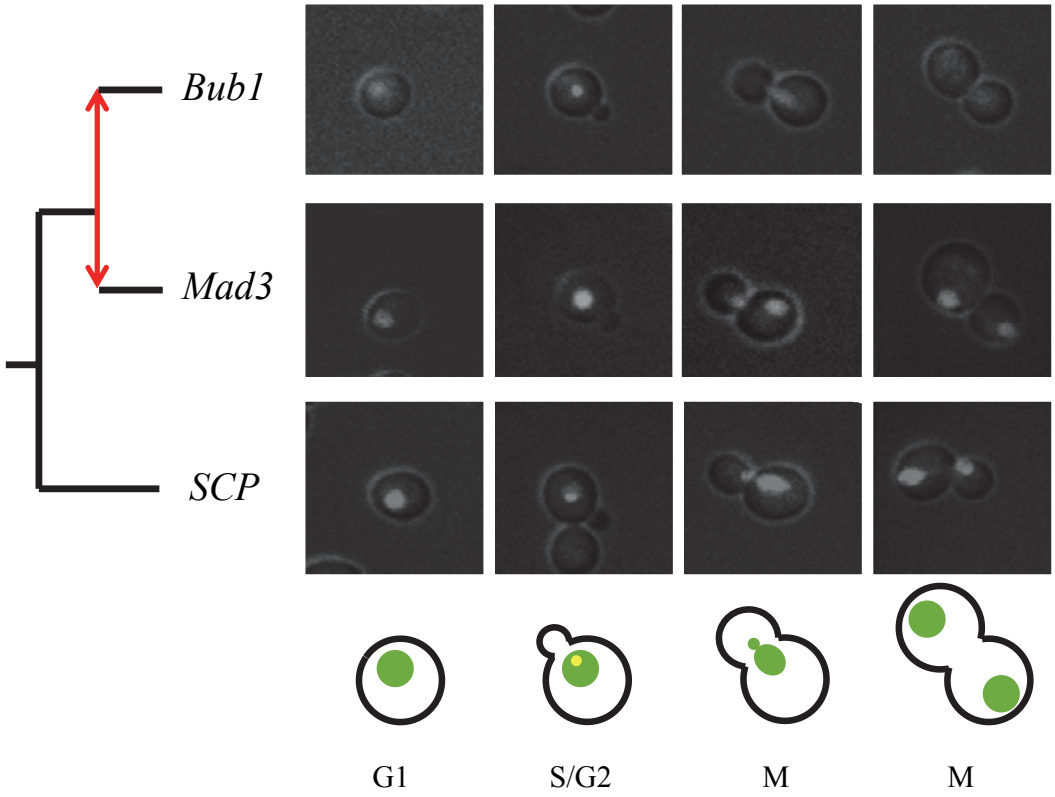


Figure V-3. Localization of the Bub1/Mad3 paralogs. Green-fluorescent protein tagged genes cloned from *L. kluyveri* (SCP) localizes as a mixture of Bub1 and Mad3 protein from *S. cerevisiae*. Bub1 is clearly visible as a puncta in S/G2 phase of the cell cycle, presumably when it localizes to the kinetochores, while Mad3 is seen as nuclear throughout the cell-cycle. The single-copy protein is also seen as nuclear throughout the cell-cycle, however is also seen as a puncta in S/G2.

We next sought to functionally verify if the single-copy protein could rescue the cell-cycle functions of Bub1 and Mad3. Cells lacking Bub1 or Mad3 are highly sensitive to benomyl, a microtubule destabilizing drug (Fernius and Hardwick 2007), because cells cannot detect that their chromosomes are not attached to the mitotic spindles. If the single-copy protein can perform the functions of both Bub1 and Mad3, we expect that the single-copy protein would rescue the fitness defect of cells lacking Bub1 or Mad3. If the phenotype is not fully rescued, it may indicate neo-functionalization and adaptation in the Bub1 or Mad3 protein or it may be due to an artifact of expressing a heterologous gene. We performed spot dilution assays

and our results indicate that cells lacking Bub1 and Mad3 can grow adequately in the presence of benomyl if rescued by the single-copy protein with similar growth characteristics to wild-type *S. cerevisiae* cells (Figure V-4). We could not simply test the function of the spindle checkpoint in other species as we observed that other yeasts other than *S. cerevisiae* were highly resistant to benomyl, irrespective of whether they had duplicated the Bub1/Mad3 gene (Appendix Figure IV-3).

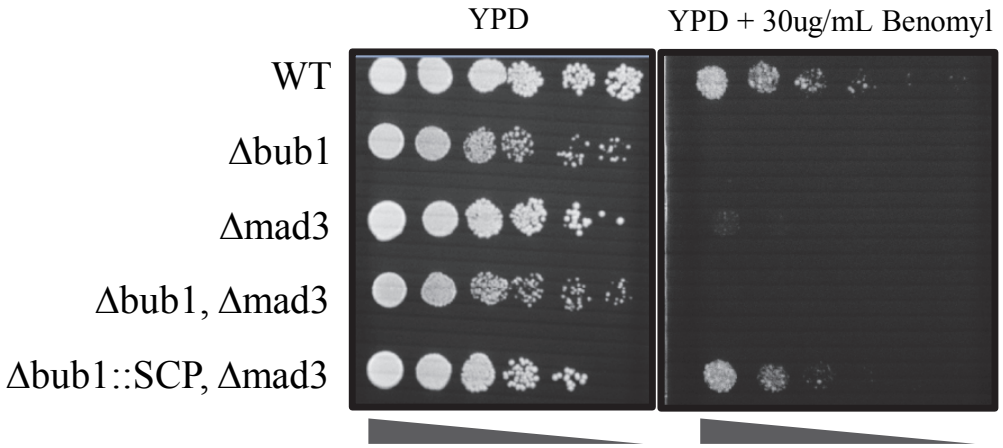


Figure V-4. Spot dilution assays showing phenotype rescue of the spindle checkpoint mutants. Yeast cells expressing the single-copy protein (SCP) at the BUB1 locus can grow on YPD plates with 30ug/mL benomyl even in the absence of Mad3, as assessed by a 10-fold spot dilution assay. YPD plate was imaged after 2 days. YPD plate containing benomyl was imaged after 3 days.

In principle, it may be possible to quantify fitness effects from plate dilution assays; however we found that this assay was unreliable in assessing small differences (we estimate that differences higher than 10% are required if plating exactly the same amount of cells at each dilution). At the limits of the resolution of this plate assay, and taken together with the localization data, these results suggest that the ancestral protein can rescue Bub1 and Mad3 function in *S. cerevisiae* and that, consistent with the DDC model, the sub-functionalization of Bub1 and Mad3 may confer no fitness advantage to *S. cerevisiae*.

V.3.2 A precise and rapid quantitative fitness assay for gene network rewiring

However, to further support that the gene duplication confers no fitness advantage, we also expect that the step-wise degeneration of the ancestral gene would be neutral. Alternatively it is possible that the spot dilution assay did not have enough resolution to detect the fitness advantage of the gene duplication. To address this, we designed a method to more precisely quantify the relative selection coefficients (s_n, see Methods) of intermediates during gene network rewiring after gene duplication to assess whether a neutral (or adaptive) path could exist between the ancestral network and the current extant network. We first show how the strains are made and then show an analysis describing the resolution limit of our fitness assay.

Because we do not know the order of mutations, we created all strains with the genetic make-up of the evolutionary intermediates. The evolutionary paths assayed occur on multiple loci and we therefore took advantage of the SGA cloning strategy (Tong et al. 2001) to combine alleles at multiple loci into single strains (see Methods). Briefly, libraries of mutations were cloned into different starting strains (Figure V-5A). Query strains carrying the SGA markers and different fluorophores (shown to have no fitness effects (Figure V-5A)) were crossed to the library in an ordered array and selected such that the final products of several rounds of mating were otherwise genetically identical haploid spores carrying different marked alleles (Figure V-5B). Diploids can also be generated by a slight extension of the procedure (see Appendix IV).

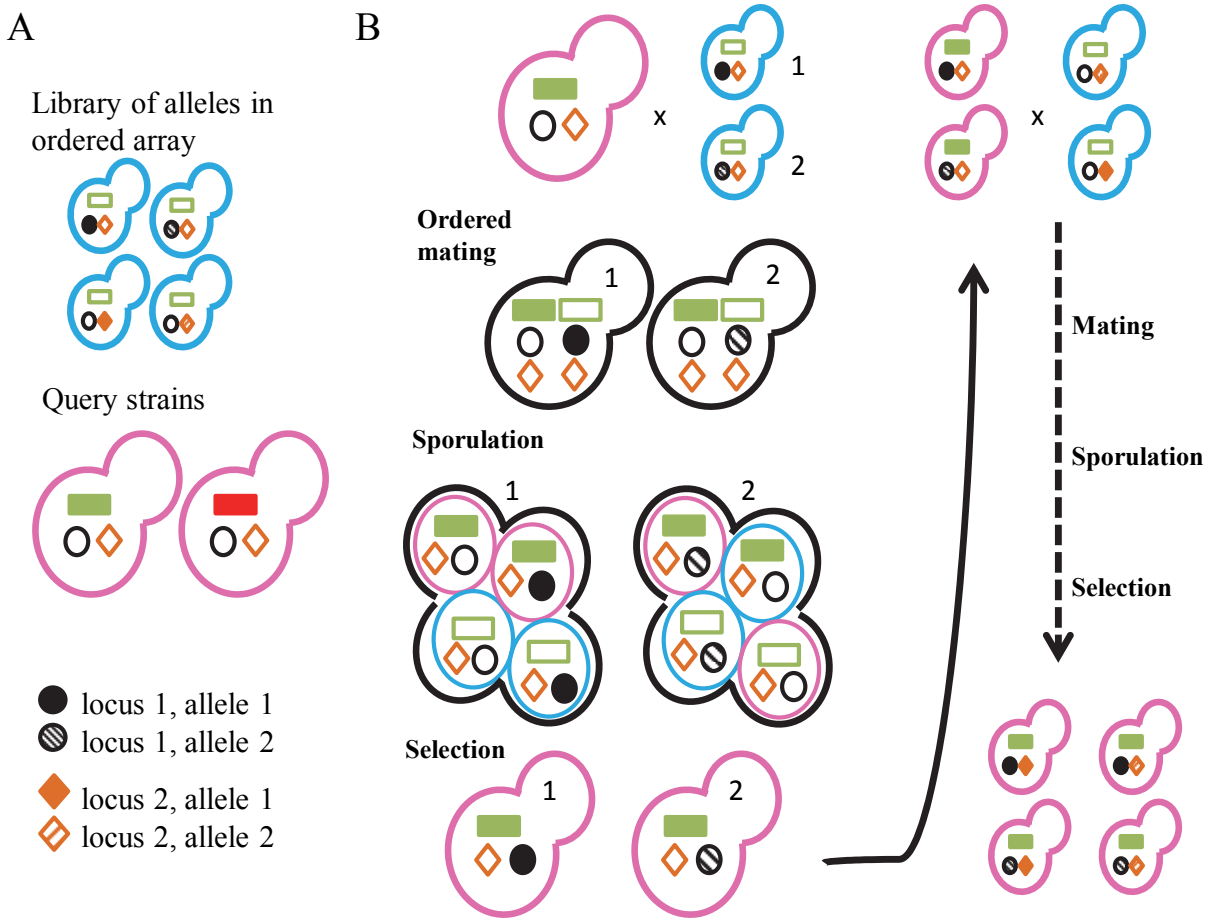


Figure V-5. Synthetic genetic array (SGA) design. A) A library of MATa strains (blue) all contain a single mutation. MATα SGA query strains (pink) contain a fluorophore (red or green rectangle) and SGA marker. Strains carrying the library of alleles have marked mutations. B) The SGA process allows the recombination of alleles from one strain to a query strain using a series of simple replica-pinning procedure on selective media. In contrast with previously described SGA designs (Tong et al. 2001), the product of a single SGA process is the same genotype as the query strain except that it now contains the marked allele taken from the ordered array of libraries. The process is reiterated for every locus with marked alleles.

Having described how the strains are cloned, we now turn to our assay for quantitative fitness measurements. Briefly, our assay is a competitive fitness assay that takes advantage of flow-cytometry to provide relative counts of fluorescently labeled cells within a growing

population (Breslow et al. 2008) and this assay can be performed in high-throughput (see Methods). Because of the sensitive nature of the competitive fitness assay, we first performed an analysis of the resolution of the assay. This was found to be important because BY4741 and BY4742 cells differ from each other at several loci with at least 20 non-synonymous single-nucleotide polymorphisms (SNPs) (SGD Project 2011). Therefore, haploids created by SGA may have the same marked alleles but different background genetic make-up and are therefore mosaics of the BY4741 and BY4742 starting strains. Presumably, the most important SNP found in the whole-genome sequences of BY4741 and BY4742 is a SNP causing an early stop-codon in the *WHI2* gene of BY4742. *WHI2* is a stress response gene that becomes frequently mutated in laboratory *S. cerevisiae* strains (Lang et al. 2013), with a large portion (>25%) of strains in the yeast-deletion collection having independent mutations causing loss of Whi2 function (Teng et al. 2013). Loss of Whi2 causes several phenotypes where strains grow more rapidly under starvation conditions, but this causes aberrant mitochondrial function (tubular to large punctae morphology) and ultimately promotes cell-death upon prolonged growth after diauxic shift (Leadsham et al. 2009). This phenotype is not seen on plates (Cheng et al. 2008) where yeast cells have more space to grow, but can be problematic under conditions of growth that are required by the high-throughput assay (96-well plates) where oxygen is more limiting and the media quickly accumulates fermentation by-products that require mitochondrial function for growth (Goldring et al. 1970). We first confirmed that, contrary to the whole-genome sequence of BY4742, none of our query strains had a *WHI2* mutation. We also discovered that the *mad3* deletion strain from the yeast deletion collection (Giaever et al. 2002) contained an undocumented Y104X mutation in *WHI2* (causing s_n = -0.08 in our hands) and subsequently remade the strain. Although we cannot be certain that new mutations would not arise again, or that there are no additional mutations causing large effects on phenotypes, we nevertheless assessed the resolution of our assay by selecting eight spores carrying the green fluorescent protein with the wild-type Bub1 allele marked with CaURA3, and eight spores carrying the red fluorescent protein with the wild-type Bub1 allele marked with CaURA3. These cells were competed to form 64 competitions. We reasoned that if any additional SNP in the BY4741/BY4742 background had detectable fitness effect (but masked due to epistasis within their respective backgrounds), that they would be uncovered within these 16 spores.

We compared the relative proportions of cells expressing the green fluorescent protein and red fluorescent protein at the 20^{th} generation to the 40^{th} generation and we calculated the selection coefficient for each competition. Because our strains are supposedly genetically identical except for the non-shared SNPs, we expect an average selection coefficient of zero and the standard deviation obtained from this test can be used to estimate the resolution of our assay (deviations due to growth conditions or to the SNPs). Doing this, we observed a mean estimated selection coefficient of 0.00069, and a standard deviation of 0.0017 (Figure V-6). This indicates that the resolution of our assay is in the order of s_n = 0.0033 and we believe this represents the difference in growth rate that we can detect. To account for other possible variations that may occur during the course of the study (changes in media, etc) we therefore chose to report as deleterious/beneficial any differences in fitness where both replicates of a competition exceeded a selection coefficient with an absolute value of 0.005 or greater while remaining consistent with all other competitions. See Appendix IV for another analysis on the reproducibility in selection coefficient measurements.

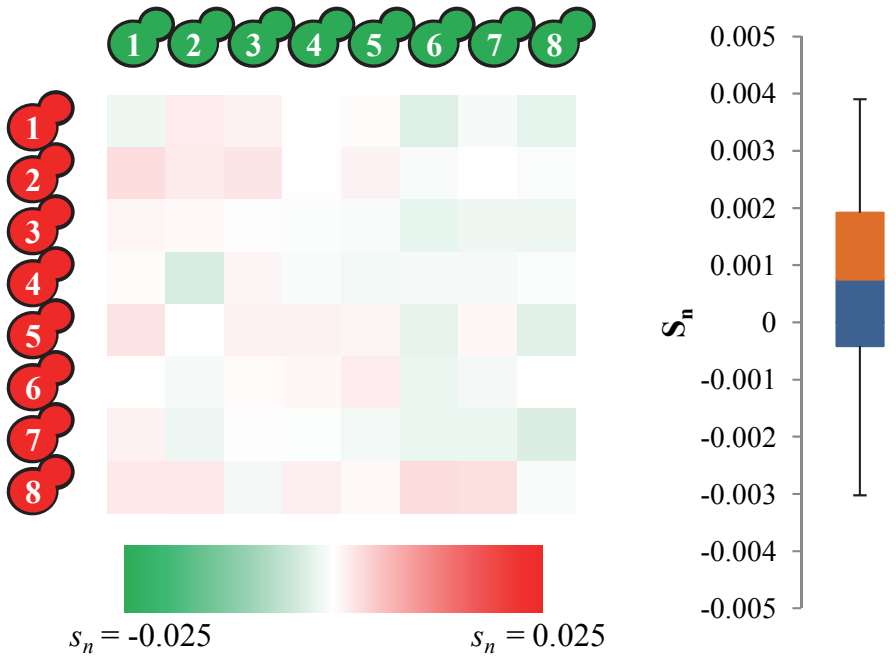


Figure V-6. High-throughput fitness assay on genetically identical strains. Spores from the SGA process carrying identical marked alleles form a mosaic of BY4741/BY4742 parental background and are assayed using our high-throughput fitness assay. Results show low calculated fitness effect (background colors of the grid) with median of s=0.00069 and a standard deviation of 0.0017. A quantile plot of the calculated fitness effects is shown on the right.

V.3.3 Subfunctionalization of the Bub1/Mad3 ancestral protein is neutral

We then performed fitness measurements on the strains created by our high-throughput procedure. Each screen competes 8 genotypes against the others (64 wells), with an additional 16 wells used as contamination control. We use the diagonal of the competitions as additional negative control as these should be competing genetically identical strains constructed independently with different fluorophores.

To assay possible paths to sub-functionalization, we first divided all of our strains into two categories (‘functional genotypes’ and ‘non-functional genotypes’ corresponding to whether or not the degenerations are complementary) and confirmed that paths leading to genotypes lacking important sequence elements would be deleterious (Figure V-7).

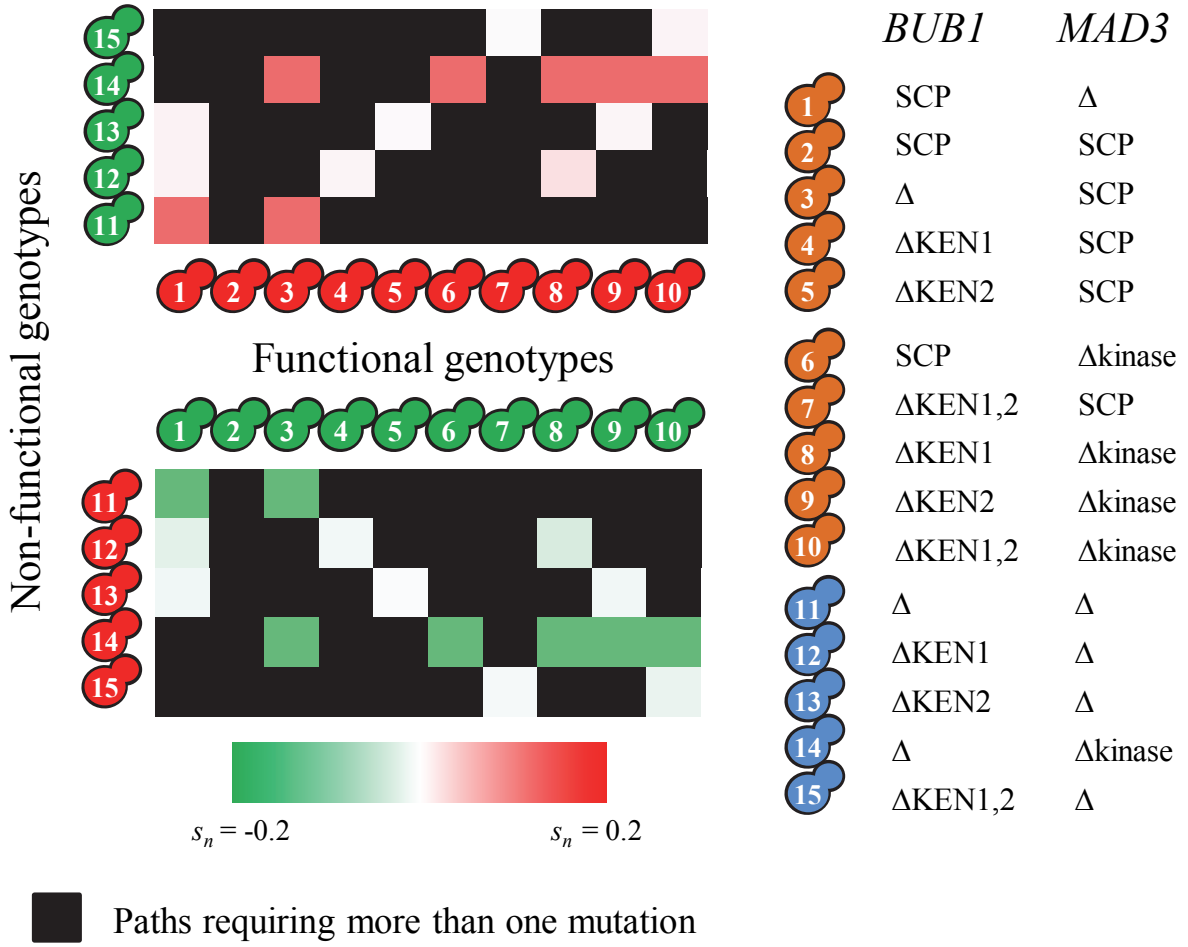


Figure V-7. Degeneration without complementation is deleterious. Functional genotypes (1-10) are genotypes where the degenerations are complemented in the duplicates. Non-functional genotypes are non-complementary degenerations. When the compared genotypes are separated by a single degeneration of a functional element, a colour scale is used to indicate the measured normalized selection coefficient of the competition. Black squares indicate compared genotypes that would require more than a single degeneration of a functional element and are excluded from the analysis of putative evolutionary paths. Figure shows the results of reciprocal experiments where the fluorescent proteins are swapped.

We then focused on the genotypes where the degenerations are complementary. Although in principle it is possible for a non-functional genotype to revert, we have not considered them in the next described analysis for simplicity. Given a current genotype, we considered that an evolutionary path towards a future genotype was possible if it involved a single molecular event (such as a deletion of a whole functional element). The results of our analysis are displayed in Figure V-8 where all possible current genotype have multiple paths and are less likely to occur when the steps are deleterious. Our results show that several possible evolutionary intermediates have slight defects when compared with the extant network and we also observed clearly that the single-copy protein could not fully rescue the spindle-checkpoint defects when placed at the *MAD3* locus, but that the protein could still perform some function (compare Figure V-8, genotype 3 vs 14, and Figure V-8 genotype 1 vs 3, see Discussion). Interestingly, we were able to find at least one neutral path consisting of three degenerations (Figure V-8, blue arrow).

Although we have not tested all the possible mutations that may have occurred during the yeast evolutionary history, the fact that we can find at least one neutral evolutionary path strongly supports the DDC hypothesis that, at the resolution of our assay, the increased complexity of the spindle checkpoint pathway can be explained by simple degenerative mutations.

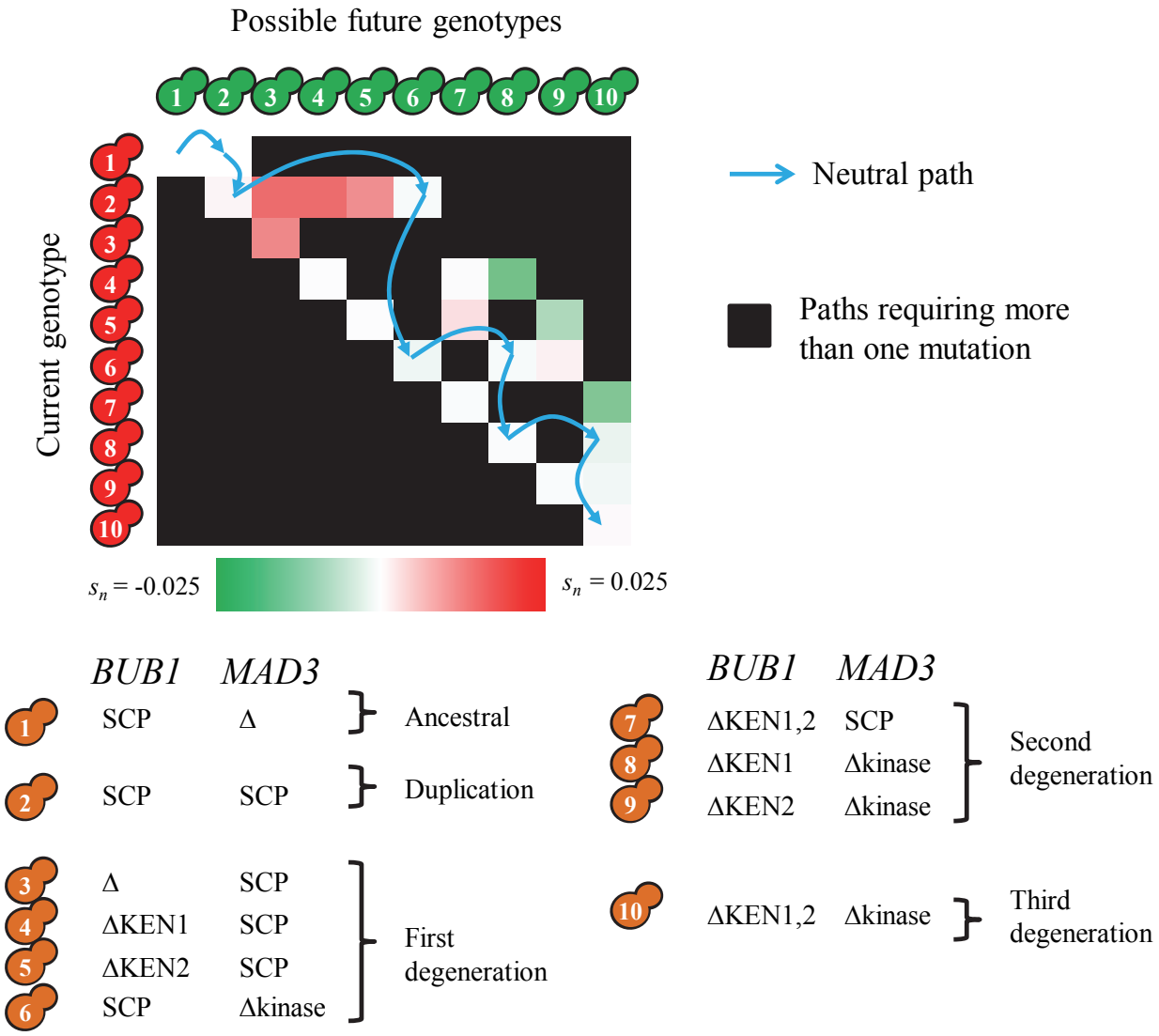


Figure V-8. Evolutionary paths during network rewiring. Schematic of the intermediates we have assayed in our experiment if we assume changes in the *MAD3* promoter occurred first following the gene duplication event. When the compared genotypes are separated by a single mutational event (such as a degeneration of a functional element), the accessible evolutionary paths are shown horizontally and a colour scale is used to indicate the measured normalized selection coefficient such that a redder square indicates a deleterious step. Black squares indicate compared genotypes that would require more than a single mutational event and are excluded from the analysis of putative evolutionary paths. The schematic shows a putative neutral path through the fitness surface from the ancestral genotype (1) to the extant genotype (10).

V.4 Discussion

Neutral processes, such as described by the DDC model (Force et al. 1999), have been shown to be important in increasing genomic complexity (see (Fernández and Lynch 2011) for a study on non-adaptive increase in interactome complexity). At the limit of the resolution of our assay (discussed below), we find no evidence that the Bub1/Mad3 sub-functionalization in budding yeast provides a fitness advantage over a single-copy protein in the spindle checkpoint pathway.

Consistent with the DDC model, none of the evolutionary intermediates that we consider functional through the comparative genomics analysis showed a fitness defect when driven by the *BUB1* promoter. However, the resolution of our assay meant that we could only detect fitness effects in the range of $s_n = 0.005$. Because of the population size of budding yeasts in nature (Lynch and Conery 2003) and our estimated effective population size during the experiment (see Methods), selection could be efficient even on undetected differences in fitness. Therefore, it is possible that the sub-functionalization to Bub1/Mad3 is truly adaptive. However, even if we assume a very small beneficial effect that was not detected, because we performed our assay in the presence of high concentration of benomyl which had been used to characterize all the components of the spindle checkpoint pathway (Straight and Murray 1997), it is likely that under ideal conditions the adaptive effect of this sub-functionalization would be even smaller. It is not possible to know the environmental context, nor the genetic context, of the ancestral yeast. However, the spindle checkpoint pathway is activated every cell division to ensure proper chromosomal attachment prior to anaphase so we believe that any significant large effect would have been captured even in the laboratory environment.

Although in our study all functional evolutionary paths lead to the same genotype through the same number of degenerations, the probability that a path is taken is dependent on the rate of mutation, the selection coefficient of the intermediates and the effective population size (Weissman et al. 2009). We here discuss only the scenario of large population size (as discussed above), because small population sizes would allow all genotypes, including the sub-functionalization, to be effectively neutral. When population size is large and mutation

rate is low, then crossing a fitness valley (as measured by our quantitative fitness assay) requires the population to fix each intermediate genotypes (this is the deleterious sequential fixation regime discussed in (Weissman et al. 2009)). In this large population size and low mutation rate regime, paths through deeper fitness valleys are essentially never taken because selection is very effective. In this regime, the relative frequencies of the effectively neutral states are entirely dependent on the mutational rate between these states (Force et al. 1999). Because the mutational rate to degenerate is likely to be higher than the mutational rate to re-create a functional element, the evolutionary path of Bub1/Mad3 homologs after sub-functionalization could be nearly deterministic. Therefore, the relative probability of observing Bub1/Mad3 functional homologs is equal to the rate of sub-functionalization divided by the rate of non-functionalization. We propose that the rate of sub-functionalization can be high due to the very small number and mechanistically simple degenerations: we showed here that sub-functionalization and degeneration of the Bub1/Mad3 protein can occur within only two mutations following the gene duplication (generation of stop codon to remove the kinase function in Mad3, and a shift in start position to remove the first KEN box in Bub1), both of which have been observed frequently over evolution in other genes (Kellis et al. 2004; Main et al. 2013). If this sub-functionalization is truly neutral, we propose that the repeated observation of the Bub1/Mad3 functional homologs may be due to the fact that no other possible outcome of the duplication can be easily observed (fitness valleys or reversion to single-copy protein) and that further degeneration would be unfit. We note that we have identified at least one phylogenetic clade where the gene duplication reverted back to the single-copy functional homolog in the clade leading to *Vanderwaltozyma polysporus* (Appendix Figure IV-4). It is estimated that the ancestor to the lineages leading to *V. polysporus* and *S. cerevisiae* had already non-functionalized ~20% of the duplicates (Scannell et al. 2007) suggesting that both sub-functionalization of Bub1/Mad3 and non-functionalization were not rapid. In the absence of data confirming how often the gene has duplicated and reverted back, it is difficult to determine if the sub-functionalization truly occurred more than expected under neutral evolution.

Unlike other similar studies (Finnigan et al. 2012; Baker et al. 2013), we have not performed ancestral gene resurrection via gene synthesis, but instead we have chosen a gene from

another species which we believe is representative of the ancestral allele. Our approach has several advantages. First, it is more likely that the gene is functional in at least one genetic environment. Second, the reconstruction of the ancestral gene might not be accurate in proteins with disordered regions. Finally, the gene has evolved for the same period of time as the duplicate genes, providing a direct test of whether adaptation from the ancestral allele is due to the resolution of an adaptive conflict.

Transcriptional evolution has been shown to be an important aspect of functional divergence after gene duplication (Gu et al. 2004; Huminiecki and Wolfe 2004; Gu et al. 2005). An important caveat of our study is that we have not explicitly tested the evolution at the promoter, due to the difficulty in finding the functional elements of the promoter region (Siepel et al. 2005). However, we have evidence that the Mad3 promoter and the Bub1 promoter are not functionally equivalent as the ancestral protein does not fully rescue the spindle-checkpoint defects when placed at the *MAD3* locus. Although there is clear evidence that the promoters of Mad3 and Bub1 have diverged, we do not yet understand the effect of these changes. That differences also exist at the transcriptional level is interesting but not within the scope of our study because we aimed to address whether the repeated repartitioning of functional elements that were observed at the protein level (Suijkerbuijk et al. 2012) were adaptive. Our study shows otherwise and that there is no evidence for the acquisition of novel functions or mutations to escape an 'adaptive conflict' for the spindle checkpoint pathway in mitotic cells. Our data is consistent with the DDC model and our study suggests that parallel evolution through degenerative processes do not have to be rare or adaptive.

Interestingly, although the core spindle checkpoint pathway is conserved in all eukaryotic life, several other differences exist between them (Vleugel et al. 2012). These differences include non-conserved proteins important for spindle checkpoint function (such as p31) or different copy number of paralogous proteins (such as Cdc20 in human). Our study provides experimental techniques to test the step-wise effects of evolutionary changes that have been detected through comparative genomics on multiple loci (such as the ones in the spindle checkpoint). We anticipate that these sensitive quantitative measurements will be useful in the computational modeling of sequence and protein evolution within the context of a

complete regulatory network (Hasty et al. 2001), as has been performed on other important regulatory networks (Hao et al. 2007; Josephides and Moses 2011).

V.5 Materials and Methods

V.5.1 Yeast strains and culturing

All strains were derived from either BY4741 or BY4742 using standard yeast genetic techniques or synthetic gene arrays (see next method subsection). The 'ancestral' gene was PCR amplified from purified genomic DNA (Fermentas, #K0512) of *L. kluyveri* (NRRL Y-12651)

Strain construction for the library of alleles was performed using the same method as previously described in Chapter IV and in (Li et al. 2011). SGA query strains were created by transferring the Ste3pr_LEU2 marker from Y8205 into the *CAN1* locus of BY4742. Fluorophores with the Ste2pr_LkHIS3 were cloned into the pAN200a plasmid (based on pFA6a (Wach et al. 1994)) using standard cloning techniques and transformed into the *CAN1pr* locus using delitto perfetto (Storici et al. 2001).

Cells are grown according to slight modifications to the protocols outlined in (Tong and Boone 2006). Specifically, the protocol was modified to allow selection of uracil auxotrophy instead of nourseothricin antibiotics resistance and lysine auxotrophic MATα haploid cells are selected after each round of mating. Additional modifications will be outlined below.

Benomyl (10mg/mL DMSO stock) is used at outlined concentrations and added to boiling-hot media until completely dissolved. 5-fluoroanthranilic acid (5-FAA, 70mg/mL DMSO stock) was used to select against tryptophan biosynthesis prototrophs (Toyn et al. 2000) and plates were poured at 0.7g/L 5-FAA final concentration (supplemented with all amino acids, including 72ug/mL tryptophan). This concentration was found to allow adequate growth of Δtrp1 cells while preventing growth of TRP1 and TRP1/Δtrp1 cells. However, concentrations above 0.8g/L were toxic even to Δtrp1 cells. 5-fluoroorotic acid (5-FOA, 100mg/mL DMSO stock) was used to select against uracil biosynthesis prototrophs (Boeke et al. 1987) and plates were poured at 1g/L 5-FOA final concentration (supplemented with all

amino acids, including 72ug/mL uracil). Geneticin (G418) was used at 200ug/mL to select for geneticin resistance.

V.5.2 Synthetic genetic arrays

Synthetic genetic arrays (SGA) were performed with several modifications to the procedure used in (Tong and Boone 2006). We modified our query strains to have the following cassette integrated at the *CAN1* locus: RPL39pr_fluorophore_Ste2pr_LkHIS3_Ste3pr_LEU2. Fluorophores used for our study were yeast-enhanced monomeric green fluorescent protein (ymEGFP) and yeast-mCherry (ymCherry). For the general construction of our strains, a query strain is first crossed with all the desired alleles at a particular locus. Diploid selection is performed by selecting for complementary auxotrophies. Overnight diploid cells from plate patches are then scraped into liquid sporulation media (1% Potassium acetate, 0.005% Zinc acetate) and supplemented with amino acid requirements for diploid strains at 25% of the normal usage and incubated on a roller wheel for three days at room temperature. Usually, about 30% sporulation is observed and 5ul of the mixture is spread or spotted on selection plates that select for MATα and other selection markers.

Interestingly, we found that modifications to the germination and outgrowth procedure were necessary in our hands to obtain colonies after the SGA procedure. In our hands, S/MSG media was adequate for germination of Ste3pr_LEU2 MATα cells. However, S/MSG media did not allow robust growth for Ste2pr_SkHIS3 and any form of growth for Ste2pr_LkLYS2 following germination (this media is recommended in the random-spore analysis but not in the standard SGA protocol).We speculate that either the nitrogen source is too low for adequate germination in the presence of histidine or lysine selection, or that the Ste2 promoter behaves with different dynamics and molecular control than the Ste3 promoter (except for the mating type exclusivity). Further, we found that adding lysine to the media greatly enhances the initial outgrowth of spores for all strains that were constructed using our query strains (even the ones that did not express a fluorophore). Spore outgrowth was normal for standard SGA query strains, indicating that heterozygous *lys2* deletion strains or homozygous *LYP1* affected the outgrowth of our strains (*LYS2*/Δ*lys2 LYP1*/*LYP1* compared with *LYS2*/*LYS2 LYP1*/Δ*lyp1*). Therefore, to select for lysine prototrophs, the colonies are

replicated to media lacking lysine after the initial growth. To select for lysine auxotrophy, replica plating was used to isolate colonies that did not grow on media lacking lysine, however alpha-aminoadipate could be used instead (Chattoo et al. 1979). We have since created a BY4741 strain with Δ*lys2*, which bypasses the need to select for lysine auxotrophs.

A schematic of constructs that were used for assaying fitness is included in Appendix Figure IV-5.

V.5.3 Genomic sequences and comparative analyses

Protein sequences used for the comparative analyses were from the Yeast Gene Order Browser (Byrne and Wolfe 2005). These proteins were then aligned with MAFFT (Katoh et al. 2002). Genomic sequences for BY4741 and BY4742 were obtained from the Saccharomyces Genome Database (SGD Project 2011).

To perform our comparative analyses, protein sequences were analyzed using methods described in Chapter IV and by visual inspection. Briefly, regions under purifying selection in the clade that diverged prior to the whole-genome duplication in yeast are analyzed for relaxation of these constraints after the whole-genome duplication. Partitioning of these losses in selection constraints in the two paralogs is an indication of sub-functionalization.

X-ray crystallography files (3ESL: Bub1 (Bolanos-Garcia et al. 2009) and 4AEZ: Mad3 (Chao et al. 2012)) were obtained from the Protein Data Bank (Berman et al. 2000) and analyzed using PyMOL (Schrödinger 2010).

V.5.4 Quantitative fitness assay

Quantitative fitness assays were performed using the MACSQuant VYB (Miltenyi Biotec Inc.). Briefly, strains are grown for 48 hours in 5mL of cultures on a rolling wheel. The competitive fitness experiment is setup by mixing relatively equal proportion of green cells and red cells in 96-well blocks (100ul of a single ymCherry expressing strain and 100ul of a single ymEGFP expressing strain into 600ul distilled water) at a 4-fold dilution. 20ul of this mixture is transferred into 100ul water that is distributed on a 96-well plate and analyzed by flow cytometry. To continue growth of the competition mixture, 20ul of the competition is

also diluted into 300ul distilled water to form a 16 fold dilution, then 20ul of this dilution is diluted into 300ul of defined media supplemented with amino acids and 100ug/mL ampicillin to form a final 16-fold dilution. The cells are therefore diluted 1024-fold every 24 hours. Given a conservative estimate of $2*10^8$ yeast cells per mL at saturation, we estimate an effective population size (Ne) of approximately $3.44 * 10^5$.

In total, 50 000 cells are counted for each competition experiment and well-defined proportions of green and red cells are gated to remove doublets and relative cell counts was obtained from each competition (Lang et al. 2009; Frenkel et al. 2014). We define the relative selection coefficient (s) as the increase in logarithmic ratio of red fluorescent cells (R) and green fluorescent cells (G) every generation as follows:

$$\frac{R_1}{G_1} = \frac{R_0}{G_0}(1+s)$$

$$\frac{R_t}{G_t} = \frac{R_0}{G_0}(1+s)^t$$

Where t indicates the number of generation, and s is the selection coefficient. Clearly, if s is positive, then the proportion of red cells to green cells increases at the next generation. Transforming to log-space:

$$\frac{\ln\frac{R_t}{G_t} - \ln\frac{R_0}{G_0}}{t} = \ln(1+s)$$

The relative selection coefficient s is not symmetric (s of 0.1 is equal to s of -0.090909 if we do the inverse calculation G/R) so we cannot directly compare our control experiments. To correct for this, we normalized the selection coefficient (s_n) by taking the inverse calculation when $\ln(1+s) < 0$, and multiplied by -1 to indicate a negative effect. The absolute value of the normalized selection coefficient is therefore the selective effect of a beneficial allele as in (Kimura 1962). For the purpose of this study, we only report the normalized selection coefficients s_n.

The upper limit of detection occurs when we expect fewer than 1 out of 50000 cells of the worst genotype at the 40th generation, which occurs at approximately s_n = -0.3. In practice, some genotypes can no longer be detected at the 20th generation, and we simply report these as s_n < -0.3. In some other cases, the number of red or green cells is fewer than 50, which leads to highly inaccurate estimates of the selection coefficient and we also report these as s_n < -0.3. When the total number of cells counted was fewer than 50, we reported s_n = 0 to mean equally lethal.

In the absence of all pairwise measurements, it is also possible to measure the selection coefficients against a reference strain (typically the wild-type strain). This can be done by simply using the following relationship:

$$s_{a\_b} = \frac{1 + s_{a\_c}}{1 + s_{b\_c}} - 1$$

Where s_{x_y} is the measured (not normalized) relative selection coefficient of x over y.

V.5.5 Localization analysis

Cells were grown in low-fluorescence media with appropriate auxotrophic requirements to log-phase and imaged using a standard 491nm blue laser on a Leica spinning-disc confocal microscope.

V.6 Acknowledgments and funding information

We are grateful for the *L. kluyveri* strain that were generously donated by Dr. Marc-André Lachance. We thank Brenda Andrews for access to the microscope, as well as lab members of the Andrews lab for useful help during the project. We especially thank the Peisajovich lab for helping with the flow cytometer. We also thank Moses lab members, Dr. Nicholas Provart and Dr. Philip Kim for useful feedback during the course of this study. We also thank Dr. Andrew Murray for suggesting the competitive fitness assay and for other comments during the course of the study. ANNB is funded by a postgraduate scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC). AMM is supported by a NSERC Discovery grant and Canadian Institutes of Health research (grant

MOP-119579). This research was supported by infrastructure grants from the Canadian Foundation for Innovation to AMM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

V.7 Author contributions

ANNB designed and performed the experiments and wrote the paper. AMM designed the experiments.

V.8 Supplementary materials

Supplementary text and figures are included in Appendix IV.

Chapter VI

Discussions and future directions

VI.1 Summary

Proteins contain short linear motifs and these short linear motifs are important for protein posttranslational regulation. In the past, these short linear motifs have been characterized using laborious experiments that were frequently performed using model organisms and therefore questions about their evolution could not be simply answered. For example, it was difficult to know how prevalent short linear motifs were and whether regulatory evolution could lead to functional divergence through evolution. With the advent of large-scale genomics data and sequencing, computational approaches to identify functional elements in the DNA sequence successfully uncovered sequences responsible for transcriptional evolution and allowed the exploration of the properties of transcriptional changes between organisms. My thesis work aimed to address similar questions regarding the evolution of short linear motifs.

By studying the evolution of protein regulation from manually curated and characterized phosphorylation sites, I showed that phosphorylation sites were highly conserved relative to their flanking regions, but that turnover could still be detected across closely related species of yeast. Motivated by these findings, I hypothesized that regulatory sequences such as localization signals and degradation signals would also show similar evolutionary properties. I therefore set out to design state-of-the-art computational and comparative approaches to identify short linear motifs in proteins and identified thousands of putative short linear motifs in the budding yeast proteome – many of which still remain uncharacterized. Using statistical models of protein evolution, I identified motifs that are likely to have changed constraints throughout the budding yeast phylogenetic tree. Finally, I developed experimental techniques to study the fitness contribution of regulatory changes across a regulatory network.

VI.2 Discussions and future directions

VI.2.1 Large scale quantification of birth, death, and compensation rate of short linear motifs

The quantification of turnover in Chapter II and Chapter IV showed that changes in protein regulation are very likely to contribute to functional divergence over evolution. For example, we could identify up to 9% of phosphorylation sites as not being conserved over 5-20 million years of evolution, which showed a turnover rate that is similar to the turnover rate of transcription factor binding sites. Furthermore, our study on very well conserved short linear motifs estimated that (after correcting for tests with no power) approximately 28% of the conserved short linear motifs have different rates of evolution before and after gene duplication, which occurred around 100 million years ago (Wolfe and Shields 1997). This suggests that turnover of regulatory sequences may frequently occur through evolution. However we observed in Chapter II that about half of the changes in regulatory sequences were accompanied by signatures of compensation, which was almost more than expected by chance. Therefore, it is possible that natural selection on the regulation of a protein occurs through stabilizing selection of the motifs in disordered regions of the protein. This is supported by the presence of conserved clusters of phosphorylation sites without conservation of any individual sites (Holt et al. 2009; Lai et al. 2012; Freschi et al. 2014). Nevertheless, in Chapter IV we also observed clear instances of regulatory turnover in duplicate proteins. Whether compensation is more frequently observed in single-copy proteins vs. duplicate proteins is an interesting future direction to reconcile changes in regulatory turnover of proteins as opposed to regulatory turnover of motifs. Observing a lower compensation rate in duplicate proteins would support the idea that regulatory evolution of duplicate proteins can be a direct mechanism for neo-/sub-functionalization. On the other hand, observing the same compensation rate in duplicate proteins would indicate that regulatory evolution affects all genes equally but that accelerated rate of evolution (or longer evolutionary time) in duplicate proteins is a more important contributor to the observed differences in regulation.

To obtain an estimate of the birth rate of short linear motifs, additional genomic sequences may be used to perform similar analysis by predicting motifs on the post-WGD clades. Although current genomic sequences do not cover enough phylogenetic distance in the post-WGD clade to accurately predict strongly conserved motifs, the rate at which new genomes are sequenced (at least 50 fungal genomes per year (Grigoriev et al. 2014)) suggests that this approach to study birth of motifs in an unbiased way will soon be possible.

VI.2.2 Proposing functions for putative short linear motifs

Although work presented in Chapter III strongly suggests that the motifs we identified using comparative genomics are likely to be functional, no function could be proposed for most of the motifs. Therefore, thousands of short linear motifs remain uncharacterized. Interestingly, some motifs consist of patterns that appear more than once in the yeast genome. Indeed, this lead to the discovery of a novel pattern which we suggest to be the Cbk1 kinase docking motif in yeast (Reményi A et al, *submitted*). Our study identified other patterns that are present in multiple proteins with similar cellular function and even more patterns may be identified using supervised approaches.

We also identified a motif pattern that was found to be present in around half of the yeast amino acid permeases and that motif was strongly associated with a palmitoylation signal. Interestingly, in another study, both the putative pattern and the palmitoylation signal have been mutated in *GAP1* (one of the amino acid permeases in budding yeast) and the authors were unable to identify any defect in Gap1 function related to these specific mutations (Merhi et al. 2011). This is not surprising considering the very complex regulation of the multifunctional Gap1 permease (Van Zeebroeck et al. 2014). Therefore, the role of this novel pattern may not easily be elucidated. However, the pattern can be detected over a wide phylogenetic distance, which suggests that it is likely to be important. The pattern, as is the terminal cysteine used in palmitoylation, can also be found in several *Schizosacchromyces* and *Candida*. In addition, the novel pattern can also be detected in basidiomycetes (but not in mammals), indicating that the pattern is broadly conserved. The presence of this pattern provides an interesting avenue for targeted drug design. For example, the fungal lysine biosynthesis via the alpha-aminoadipate pathway has been under extensive study for antifungal targets because the pathway is unique to fungi (Xu et al. 2006). However, some studies have shown that the lysine concentration in the blood is sufficient for survival of *Candida albicans* and that auxotrophy does not reduce virulence (Kur et al. 2010). There have been promising results in targeting the lysine biosynthesis pathway in *Aspergillus* lung infection models (Liebmann et al. 2004; Schöbel et al. 2010), but free lysine rescued pathogenicity even in these cases. Encouragingly, at least in budding yeast, lysine auxotrophy cannot be rescued by exogenous lysine when lysine permeases are depleted.

Although the lysine permease in budding yeast does not contain this pattern, permease specificity has been observed to change over evolution. For example, the pattern is present in the *Schizosaccharomyces*' lysine permease (*CAT1,* presumably not orthologous to the *Saccharomyces* lysine permease) and it may be present in the lysine permease of important fungal pathogens. Understanding the function of the novel pattern, which is not present in mammals, could therefore allow the use of antifungal blends targeting both import and synthesis of lysine.

The other pattern we identified appeared to be specific to proteins related to protein transport, and is also broadly conserved to *Candida*. It can also be identified in some genes from *Schizosaccharomyces* but the pattern of conservation is not as clear as in *Candida* species. This pattern, the NPY pattern, is biochemically similar to the NPF pattern which interacts with EH-domains and is known to be important in protein transport (de Beer et al. 2000). It is therefore likely that the NPY pattern interacts with a different EH-domain with subtle changes in amino acid specificities (for example, after a gene duplication of an EH-domain containing protein). If this is the case, then it may be interesting to use the NPY pattern to study posttranslational evolution mediated by evolution in *trans* (discussed below in the next section).

Nevertheless, site-directed mutagenesis of a single motif may not always yield a detectable phenotype without prior knowledge of what that phenotype may be. To remedy this, systematic deletion of all desired motifs can be rapidly performed using methodologies that were used in Chapter V, and fitness assays can be performed on 96-well plates containing different environment per wells, as has been used to uncover the phenotypic diversity of wild yeast strains (Liti et al. 2009). These high-throughput strategies are likely to help uncovering the function of putative motifs that match the same pattern (such as the NPY pattern).

VI.2.3 Evolution of novel short linear motif patterns

Novel patterns of short linear motifs appear due to evolution of the regulatory protein (evolution in *trans*). Evolution in *trans* is thought to be rare because mutations that affect the affinity of the regulatory protein to all regulatory sites are likely to be deleterious (and slightly equivalent to mutating all regulatory sites at the same time). However, novel motif

patterns are partly spared from this effect at the time of their introduction and can therefore produce different patterns in different species over the course of evolution (Figure VI-1). Even well-established regulatory networks can undergo evolution of regulatory proteins: evolution in *trans* has been observed on the Matα1 transcription factor (Baker et al. 2011), even in the absence of a duplication event. This evolution of the Matα1 transcription factor along the *Candida* lineage is particularly interesting because it controls the same set of genes in the hemiascomycetes, which is possible because the DNA recognition sites co-evolved with the transcription factor. Similar evolution in *trans* is also likely to occur in protein regulators such as protein kinases.

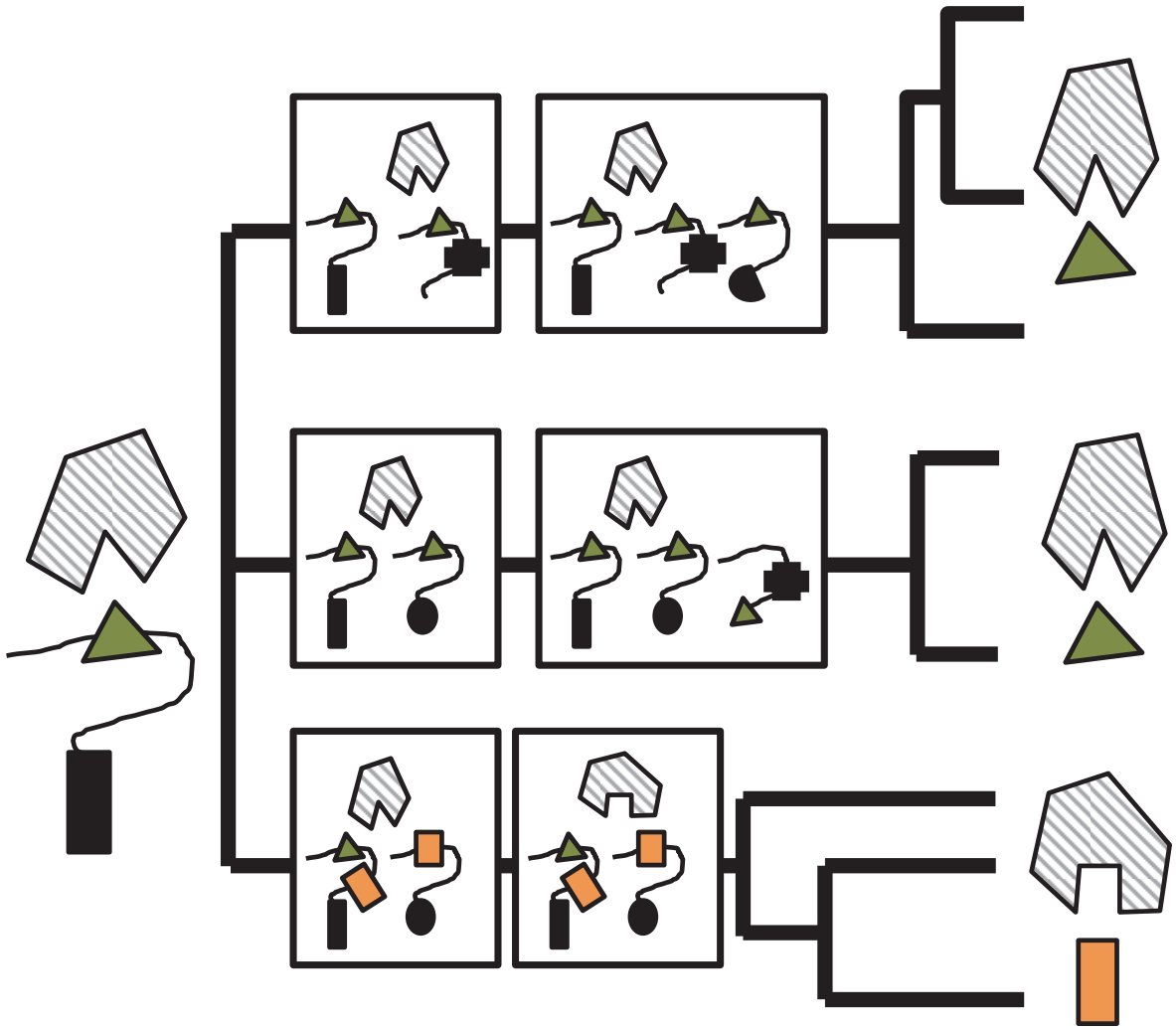


Figure VI-1. Model for acquisition of motif patterns. Evolution of motifs in *cis* can produce different sets of proteins containing the motif over the phylogenetic tree (Top two clades). Evolution in *trans* early in the phylogenetic tree can allow the acquisition of a novel pattern (Bottom clade). Striped shape indicates the recognition domain of a regulator. Black shapes indicate distinct domains on different proteins. Colored shapes represent motifs that can be recognized by the striped shape. Overlaid box show possible molecular events.

Studying the evolution of motif patterns instead of individual regulatory sequences requires at least two phylogenetically distinct clades that both contain several sequenced species with high amino acid diversity in their disordered regions. However, these clades must also be closely related, such that they would have a large degree of shared orthologs. Based on these criteria, species of yeasts that include the ones studied in Chapter IV can be compared to the *Candida* and related species. Unfortunately, current genomic sequences from the *Candida* clade are still underrepresented but efforts have been made to extend the comparative analyses that was performed on *S. cerevisiae* to the *Candida* clade (Maguire et al. 2013).

Given the pathogenicity of Candida, as well as documented changes in the genus' regulatory networks (e.g. (Baker et al. 2012)), these types of analyses may provide a more global view of how molecular evolution of short linear motifs impact the regulatory networks of the cell.

VI.2.4 Short linear motifs in higher eukaryotes

Most of the research performed in this thesis took advantage of budding yeast as a model organism for the understanding of posttranslational regulatory divergence. Given the important differences in population structures and effectiveness of natural selection between unicellular and multicellular eukaryotic organisms, it is difficult to assess whether the methods used in this thesis and whether the findings will be directly applicable to another lineage such as the mammals.

Specifically, we have attempted to use the phylogenetic hidden Markov model approach on sequence data obtained from Ensembl (Flicek et al. 2011). Preliminary analysis on the mammalian clade showed that the gene predictions were severely problematic with mispredictions of exons causing poor quantification of evolutionary rates in protein regions. Other problems surrounding the quality of the ortholog assignments also reduced the precision of the algorithm. Interestingly, the evolutionary distance required for the phylo-HMM approach is responsible for many of these issues because the annotation quality tends to be worst on species more distantly related to humans. Consistent with this, encouraging preliminary results were obtained when predicting short linear motifs in the insect lineages, which are thought to have better gene prediction models due to their simpler genome

structure while maintaining a higher nucleotide diversity than the mammals (Leffler et al. 2012).

Another important issue with applying our phylo-HMM approach is that several parameters must be chosen in order to obtain meaningful predictions. Of these, the size of the local region (the window size) is one of the most crucial parameter that must absolutely be tweaked based on the quality and evolutionary time scale of the data. This is because a higher evolutionary distance tends to create more indels within the alignment, causing the estimation of local rate of evolution to be based on indels rather than substitutions, which are more informative than indels for our model. It may be possible to use Bayesian statistical approaches to remove the effects of choosing an inappropriate window size although this is still under study. Therefore, current limitations in using our phylo-HMM approach to other clades appear to be due to technical reasons.

Nevertheless, it is likely that important short linear motifs would also be under purifying selection in more complex eukaryotes and that regulatory turnover is responsible for a portion of the organismal diversity observed throughout the multicellular eukaryotes.

VI.2.5 Regulatory sequences in ordered regions

Posttranslational regulatory sequences in disordered regions of proteins were a major focus of this thesis. However, our set of manually curated phosphorylation sites in Chapter II showed that 25% of the phosphorylation sites occurred in structured regions. This is not surprising because phosphorylation sites on kinases are known to control their catalytic activity (Johnson et al. 1996). There are most likely other motifs in ordered regions that could play regulatory roles

Our current approach to predict short linear motifs is unlikely to be adequate for predictions inside ordered regions. However, as we refine the rules of disordered regions that allow accurate recognition of motifs by other proteins, we may uncover short segments in structured regions that have similar properties and it may be possible to adapt the phylo-HMM approach for the prediction of motifs inside structured regions.

VI.2.6 The effect of selection on disordered regions

Results from my thesis showed that disordered regions have short linear motifs that are likely to be important for function. However, in Chapter III I showed that short linear motifs were estimated to only populate 17% of the disordered regions. A compelling outstanding question remains as to whether the remaining disordered regions are also under purifying selection (not just junk protein) and if so how much of it. A parallel question aims to answer what are the functions that natural selection is preserving in disordered regions. Evidence that disordered regions must be maintained exists: disordered regions typically do not conform to the amino acid residues of proteins found to aggregate (Monsellier and Chiti 2007) and the state of disorder is frequently maintained over evolution (Bellay et al. 2011). Similarly, in Chapter IV, the simulation of protein evolution showed that a single substitution matrix could not be used to adequately model both ordered and disordered regions. It is therefore likely that evolution can maintain a state of disorderedness through mutations that favor amino acids found in these regions. The increased rate of evolution is likely to be due to several more accepted mutations and more accepted indels within these regions while maintaining the required level of disorder. A more thorough analysis of these regions could be performed using population data or closely related species where the amino acid residues can be aligned accurately.

Experimentally, studying the evolution of disordered regions is an exciting avenue to understand the function of disordered regions. Although disordered regions for specific proteins have been studied, these previous studies did not address how these regions can evolve. One way to address whether disordered regions are only selected to remain disordered is by using simulation of protein evolution to simulate neutral evolution using the disordered amino acid substitution matrix or using the ordered substitution matrix. These simulated regions could then be introduced into the yeast genome along with the conserved short linear motifs and assayed for fitness. Alternatively, it would be interesting to know whether different disordered regions from non-orthologous (or orthologous) proteins can serve the same ‘function’. A possible experimental design to directly assay this would be to swap disordered regions from different proteins and assay for fitness consequences using similar techniques as outlined in Chapter V. This would allow us to disentangle some of the correlations with disordered regions that have been found. For example, as described in Chapter I, proteins containing disordered regions are more likely to be toxic upon

overexpression (Vavouri et al. 2009). However, because proteins containing disordered regions are more likely to be signaling proteins, it is difficult to untangle the effect of overexpressing the domain with the effect of overexpressing the disordered region. This can be partly solved by changing the length of the disordered region using other disordered regions from orthologs and assaying for toxicity, although care should be taken to ensure that the functions of the disordered regions have been preserved.

VI.2.7 Competitive fitness assay using pooled yeast strains

In Chapter V, I used the standard competitive fitness assay using fluorophores to differentiate genotypes. Although this assay was performed in high-throughput, the assay cannot be scaled easily because the number of competition grows exponentially for each new strain. One quick solution for this is to employ high-throughput liquid handling to significantly ease the process. In practice, it may also be possible to only compete against a reference strain at the expense of sensitivity in measurements of selection coefficient.

One way to partly resolve the scalability problem is to employ more fluorophores, which allows more competitions to be performed in a single well. Indeed, we have generated query strains that express several other colours that can be spectrally differentiated using the flow-cytometer or fluorescent microscopy with minimal spectral compensation (Figure IV-2).

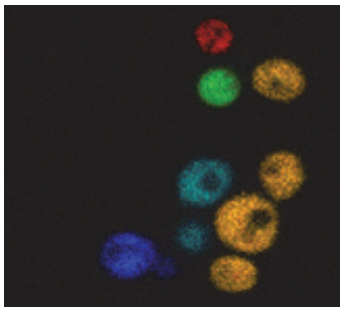


Figure VI-2. Possible fitness assay using multiple colours. Yeast query strains carrying spectrally differentiable fluorophores are available and can be used to multiplex the fitness assay.

However, the additions of these fluorophores complicate the strain-making process and it is still unclear whether these fluorophores can be used effectively. An alternative to using fluorophores is to perform deep sequencing of en-masse competitions (Hietpas et al. 2011). However this approach is limited to studying a single allele that can be covered by sequencing reads, which is not applicable to our study that assayed multiple loci. To remedy this, another approach which is still under study is the use of site-specific recombinases (Lee and Saito 1998) flanking barcodes for each allele that can be combined into a single region of the yeast genome and finally sequenced (unpublished work by the Roth lab). A more recent approach has been to barcode yeast strains containing multiple alleles rather than marking individual alleles (unpublished work by the Sherlock lab). However, these approaches suffer from low resolution ($s = 0.05$), which is of an order of magnitude lower than what we can detect using the fluorophore system.

An alternative approach to using one fluorophore per genotype is to use one fluorophore per allele. This requires more fluorescent reporters to be used and currently limits our choice of alleles. However, this can be partly solved by bypassing the use of florescent proteins. Instead, cell-surface display of epitopes can be used (Kondo and Ueda 2004) and recognized by chemical fluorophore-conjugated antibodies, which have better spectral properties than fluorescent proteins. For example, quantum dot bioconjugates can be engineered to cover a broad range of the visible light spectra from a single emission wavelength (Medintz et al. 2005). Using this technique, seventeen colours have been used to track populations of phenotypically distinct cells in the peripheral blood of humans (Perfetto et al. 2004). An even higher level of multiplexing has be achieved with mass cytometry (Bendall et al. 2011). If these systems could be applied to yeast, then we would be able to follow several different alleles inside a single tube because the cytometer provides the measurements for each cellular event.

Finally, these new technologies may even allow the creation of different strains from pools of yeasts rather than precisely forming ordered arrays using SGA (Tong et al. 2001) by sorting the cells with the desired allele combinations. This may also lead to other exciting areas of research where evolution of sexual populations can be studied, allowing the marked alleles to be followed in mixing populations.

Appendix I

Supplementary data for Chapter II

This is an author-produced PDF of an article accepted for publication in Molecular Biology and Evolution following peer review. The version of record:

Evolution of characterized phosphorylation sites in budding yeast

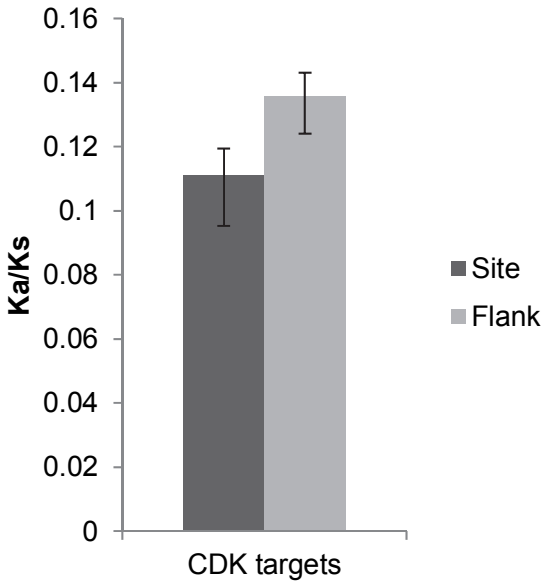
Mol Biol Evol. 2010 Sep;27(9):2027-37. doi:10.1096/molbev/msq090. Epub 2010 Apr 5.

Alex N Nguyen Ba[1,2], Alan M Moses[1,2,¥]

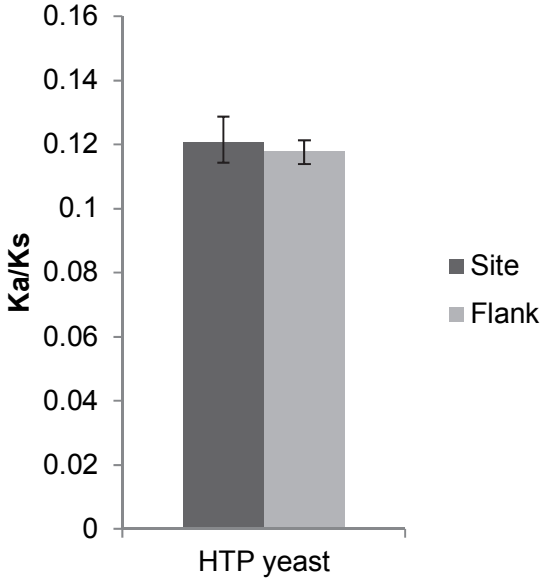
is available online at: http://mbe.oxfordjournals.org/content/27/9/2027.abstract

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, Toronto, Canada

2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada



Appendix Figure I-1. Ka/Ks of CDK phosphorylation sites in yeast *in vitro* CDK targets.



Appendix Figure I-2. Ka/Ks of phosphorylation sites from yeast high-throughput studies.

Appendix II

Supplementary data for Chapter III

This is an author-produced PDF of an article accepted for publication in Science Signaling following peer review. The version of record: Proteome-Wide Discovery of Evolutionary Conserved Sequences in Disordered Regions

Sci Signal. 2012 Mar 13;5(215):rs1. doi: 10.1126/scisignal.2002515.

Alex N Nguyen Ba[1,2], Brian J Yeh[3], Dewald van Dyk[4,5], Alan R Davidson[6], Brenda J Andrews[4,5], Eric L Weiss[3], Alan M Moses[1,2,¥]

is available online at: http://stke.sciencemag.org/content/5/215/rs1.abstract

1. Department of Cell & Systems Biology, University of Toronto, Toronto, Canada

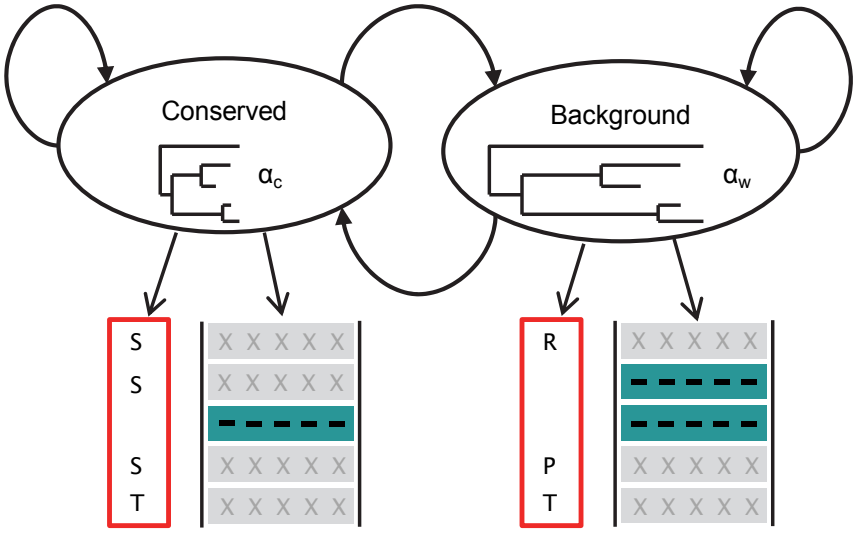
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Canada

3. Department of Molecular Biosciences, Northwestern University, Evanston, Illinois, United States of America.

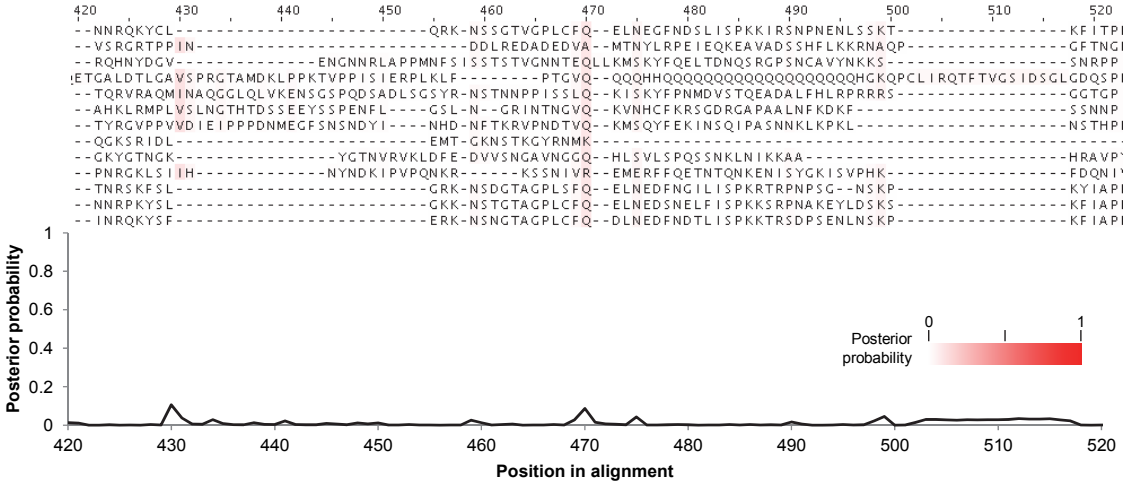
4. The Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada;

5. Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

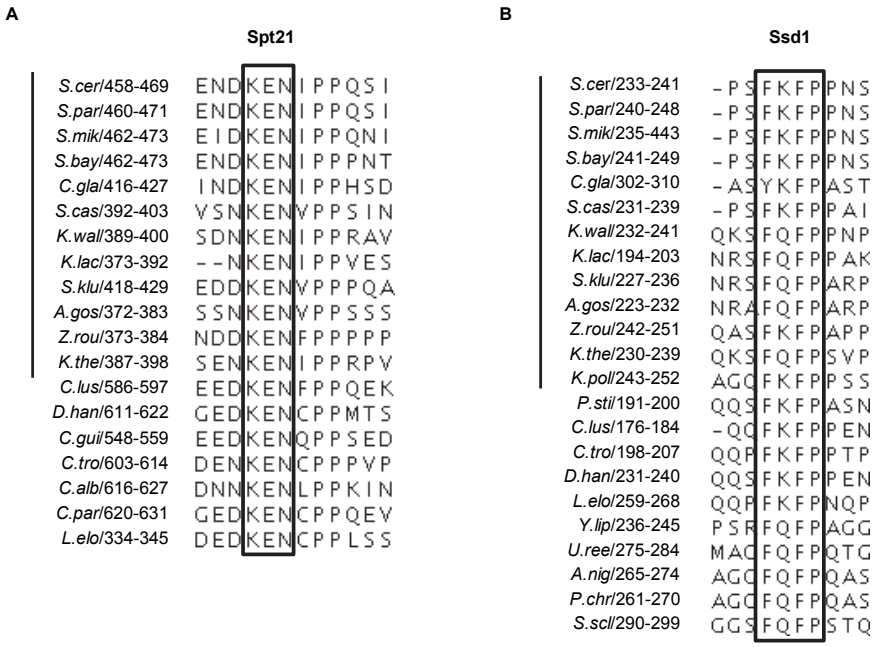
6. Department of Biochemistry, University of Toronto, Toronto, Canada



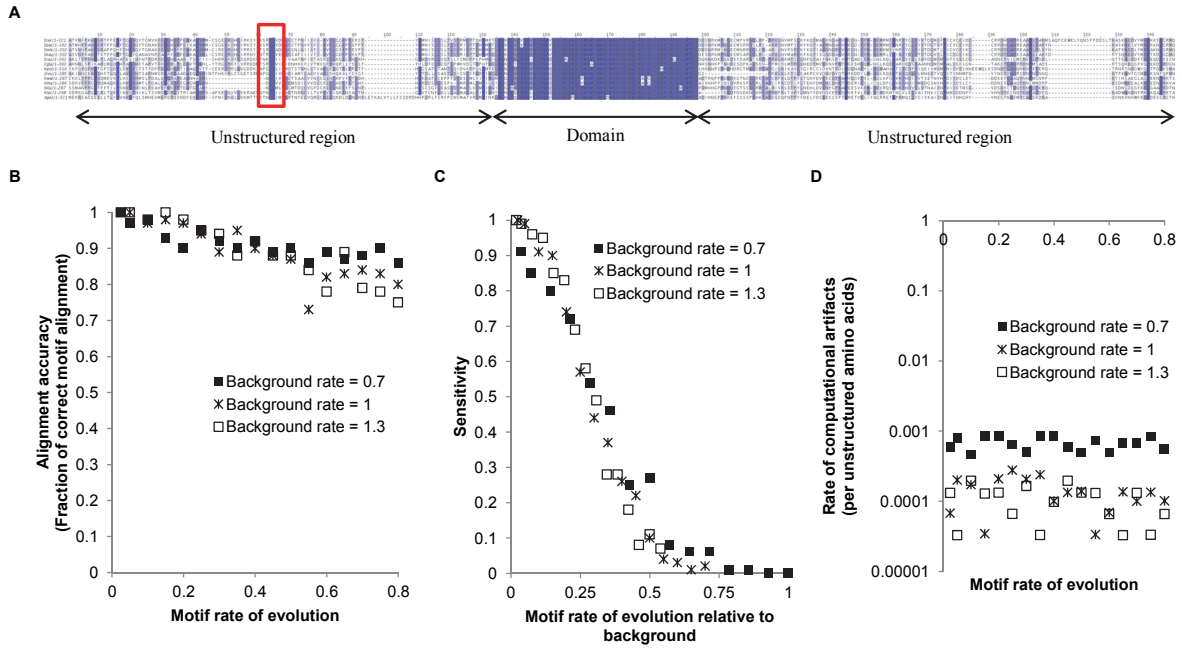
Appendix Figure II-3. Schematic of the phylo-HMM approach. In the phylo-HMM framework, a column of the sequence alignment is assumed to belong to either a conserved state ("Conserved") or a background state ("Background") and probabilities of observing alignment columns in each state depend on a phylogenetic model of protein evolution with a rate parameter α.



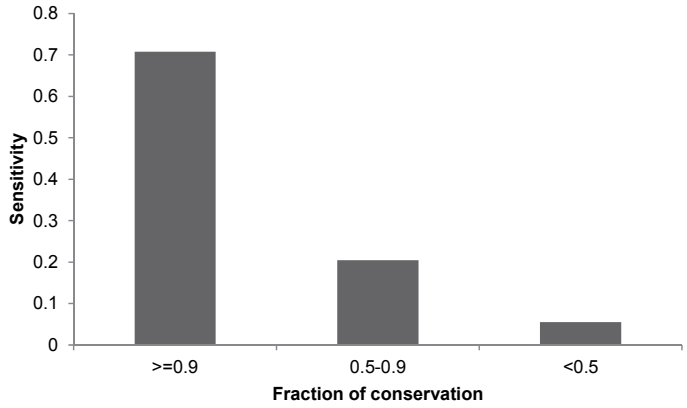
Appendix Figure II-4. Regions without conserved segments are not detected by the phylo-HMM approach. Posterior trace of the region 420-520 in the alignment of Swi5. No locally conserved segments were identified. The region shown corresponds to position 266-322 in *S.cerevisiae*. Red color intensity represents the posterior probability of the conserved state.



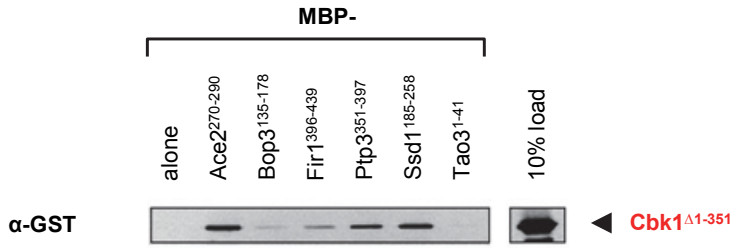
Appendix Figure II-5. Newly identified KEN box in Spt21 and Cbk1 interaction motif in Ssd1 are conserved in farther yeast species. A) Alignment of the KEN box (in rectangle) in Spt21 with distantly related yeast species. We note that the KEN motif in the Candida clade does not align with species used in our study when performing a whole gene alignment. However, the motif is well conserved. B) Alignment of one of the Cbk1 interaction motif (in rectangle) in Ssd1 with distantly related yeast species. In both panels, species used for the phylo-HMM analysis are labelled with a vertical black bar.



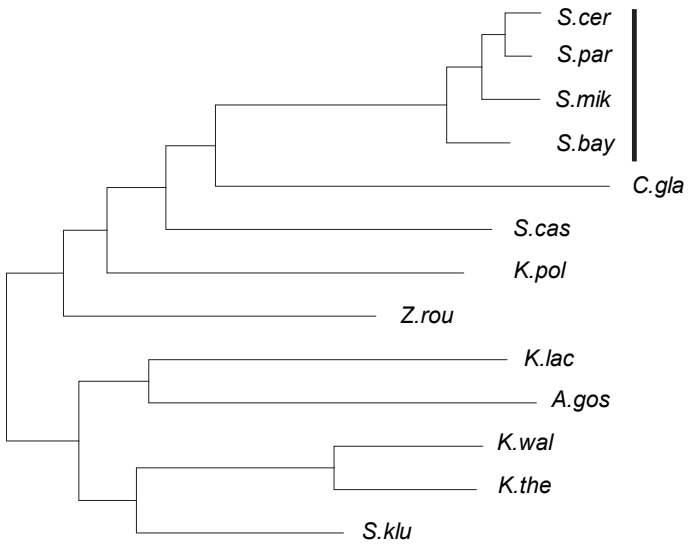
Appendix Figure II-6. Simulation of protein evolution. A) Simulations of proteins evolution were performed by randomly generating a protein sequence containing a motif in an unstructured region. An alignment of a typical simulated protein is shown with the motif, properly aligned, boxed in red. B) Alignment accuracy of the motif using MAFFT depends on the background rate of evolution and on the rate of motif evolution. C) The sensitivity of the phylo-HMM on simulated data shows strong dependence on the relative rate of evolution of the motif to the rate of evolution of the background. D) The rate of computational artifacts of the phylo-HMM on simulated data is dependent on the background rate of evolution. Each point represents the results from 100 simulated proteins.



Appendix Figure II-7. Performance of the phylo-HMM approach on literature-curated short linear motifs. On characterized short linear motifs, the phylo-HMM performs best on conserved regulatory sequences. Data shown here only includes motifs with consensus sequences (no localization signals). Regulatory motif were binned on the basis of the fraction of species in which the consensus sequence could be found (Fraction of conservation) and sensitivity of the phylo-HMM was calculated for each bin (See Chapter III, Methods).



Appendix Figure II-8. Binding of [YF]xFP peptides to Cbk1. Fragments from proteins identified in the [YF]xFP cluster (Figure III-5C) were expressed as maltose-binding protein (MBP) fusions and immobilized on amylose resin. The beads were assayed for binding to GST-tagged Cbk1 (Cbk1Δ1-351) in a pulldown assay. At the exposure time shown in Figure III-6, the lane showing 10% of the GST-tagged Cbk1 input (on a different nitrocellulose membrane but imaged at the same time and incubated with the same conditions) is overexposed. A reduced exposure of the assay is shown in this figure for comparison.



Appendix Figure II-9. Phylogenetic tree of species used for this study. Protein sequences from related yeast species were used to predict short conserved sequences by our phylo-HMM approach. The vertical black line indicates the four closest species of yeast that were used to obtain the amino acid substitution model. The branch lengths were estimated from random concatenations of proteins (See Methods) and the Newick format tree representation is the following: (((((((Scer: 0.0317431, Spar: 0.022837): 0.0200533, Smik: 0.0499671): 0.0302187, Sbay: 0.0545286): 0.1984531, Cgla: 0.3382801): 0.042494, Scas: 0.2796456): 0.0506147, Kpol: 0.3061641): 0.0374508, Zrou: 0.2684242): 0.0489862, ((Klac: 0.3075333, Agos: 0.3326945): 0.0600818, ((Kwal: 0.1277869, Kthe: 0.1223815): 0.1697185, Sklu: 0.1779484): 0.0492655): 0.0622827);

Appendix Table II-5. Annotation of top 20 clusters of predicted short conserved sequences using three different clustering parameters. Three tables show annotations of the top 20 clusters of each three cluster analyses that used three different metrics to assign sequence similarity between motifs (see Supplementary data table III-S4). Each set of annotation includes a sequence logo representing the pattern formed by the motifs, whether this pattern is known and notes regarding functional enrichments. The first table describes the results using an all-by-all pairwise distance measure (described in Methods) between the predicted segments. The second table describes the results after first extending either side of the motifs by five residues before applying the all-by-all distance measure. Using the same extended motifs, the third table describes the results considering only the top 10 most similar motifs as similar.

Patterns obtained from clustering of pairwise sequence distance between conserved sequences.

Cluster number	Known	Motif	Notes
1	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
2	Yes		Enriched in nuclear proteins Probably related to nuclear localization signals
3	Yes		Enriched in nucleoporins GLFG-repeat motif
4	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
5	Yes		Proline-rich motif
6	Yes		Basophilic-kinase phosphorylation sites

7	No		
7	No		
8	No		
9	Yes		Probably related to nuclear localization signals
10	Yes		Enriched in endocytosis genes EH-interacting motif
10	Yes		Proline-rich motif
11	Yes		Similar to Ime2 phosphorylation sites
12	Yes		Probably related to nuclear localization signals
13	No		
14	No		
15	Yes		Probably related to nuclear localization signals
16	No		
17	Yes		Proline-rich motif
17	Yes		Probably related to proline-directed phosphorylation sites
18	Yes		Glutamine-repeat

19	No	
19	No	
20	No	

Motif profiles were aligned using the average pairwise substitution score derived from the empirical substitution matrix described in Methods. Pairs of profiles (m_i and m_j) were then clustered using the Smith-Waterman score ($S[m_i,m_j]$) and divided by the square root of the aligned region length ($l_{i,j}$). Edges were pruned if the score exceeded a threshold of $\min(7.7,S[m_i,m_i]/\mathrm{sqrt}(l_{i,i}))$.

Patterns obtained from clustering of pairwise sequence distance between conserved sequences. Each predicted conserved sequences were first extended on both side and trimmed (See Methods).

Cluster number	Known	Motif	Notes
1	Yes		Enriched in nucleoporins GLFG-repeat motif
2	Yes		Proline-rich motif
3	Yes		Enriched in nuclear localization Probably related to nuclear localization signal
4	Yes		Enriched in endocytosis genes EH-interacting motif
5	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
6	Yes		Enriched in nucleoporins FxFG-repeat motif
7	Yes		Enriched in cell-cycle proteins Probably proline-directed phosphorylation sites
7	No		
8	No		Enriched in vesicle and nuclear membrane proteins, enriched in protein transport process. Probably related to NPF motif

9	No		
10	Yes		Probably basophilic-kinase phosphorylation sites
10	Yes		Probably related to nuclear localization signal
11	Yes		Probably related to nuclear localization signals
12	Yes		Probably related to nuclear localization signals
12	Yes		Probably related to nuclear localization signals
12	Yes		Basophilic-kinase phosphorylation sites
12	Yes		Probably proline-directed phosphorylation sites
13	Yes		Probably acidophilic-kinase phosphorylation sites
13	No		
14	Yes		Enriched in ER localization ER-localization signal
15	Yes		KEN-box APC/C degradation signal
16	Yes		Proline-rich motif

16	Yes		Probably proline-directed phosphorylation sites
17	No		Enriched in amino acid permeases
17	No		
17	No		
18	No		
19	Yes		Probably proline-directed phosphorylation sites
20	Yes		Enriched in mitochondrial localization Mitochondrial targeting signal

Motif profiles were aligned and scored as above.

Patterns obtained from clustering of conserved sequences and their top ten most similar conserved sequences, without allowing matches to paralogs or to the same protein. Each predicted conserved sequences were first extended on both side and trimmed (See Methods).

Cluster number	Known	Motif	Notes
1	Yes		Resembles motif in vacuolar proteins in yeast
2	Yes		Proline-rich motif
3	Yes		Enriched in ER localization ER-localization signal
4	Yes		Proline-rich motif of class II (PxxPx+)
5	Yes		Enriched in mitochondrial localization Mitochondrial targeting signal
6	Yes		Probably related to nuclear localization signal
6	Yes		PCNA-interacting motif
6	No		
7	No		
8	Yes		Probably basophilic-kinase phosphorylation sites
9	No		Enriched in Cbk1 interactors
10	Yes		Cbk1 phosphorylation motif

11	No		
12	Yes		Disulfide isomerase motif
13	Yes		Probably related to nuclear localization signal
14	Yes		eIF4e binding site
15	Yes		EH-interacting motif
16	No		N-terminal motif
17	No	No particular motif	
18	No		
19	No		
19	Yes		Disulfide isomerase motif
19	No		
20	Yes		FxFG-repeat motif

Motif profiles were aligned using the average pairwise substitution score derived from the empirical substitution matrix described in Methods. These profiles were then clustered using the Smith-Waterman score ($S[m_i,m_j]$). The top ten most similar profiles were taken.

Appendix III
Supplementary data for Chapter IV

This work has been submitted as: Detecting functional divergence after gene duplication through evolutionary changes in posttranslational regulatory sequences using a non-central correction to the likelihood-ratio test

Submitted to PLoS Comp. Biol.

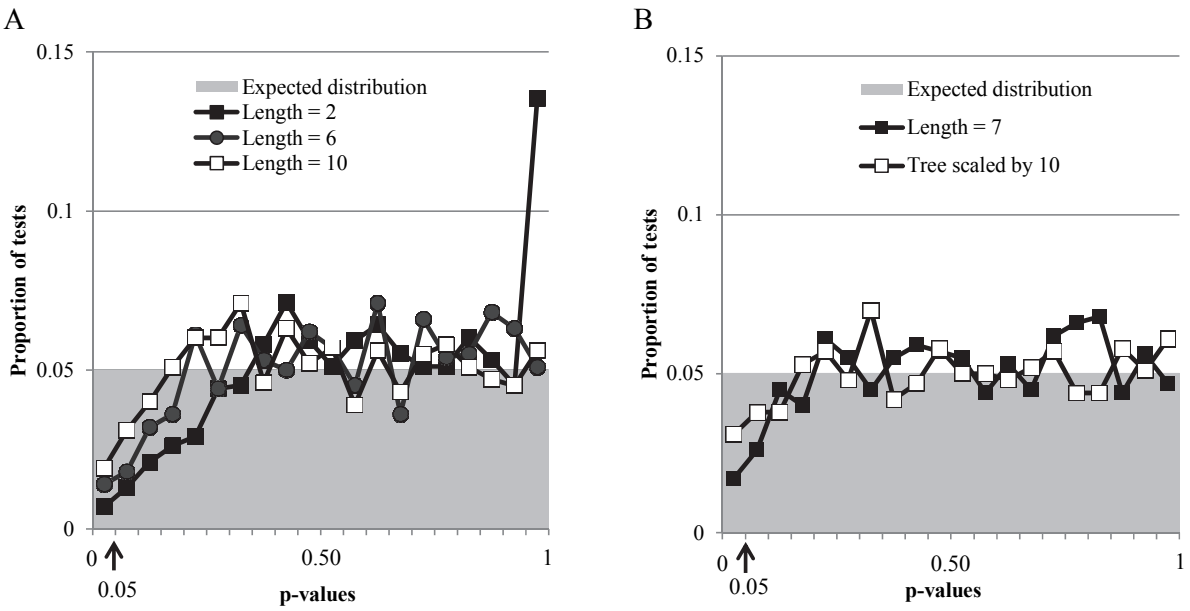
Alex N Nguyen Ba[1,2], Bob Strome[1], Jun Jie Hua[1], Jonathan Desmond[1], Isabelle Gagnon-Arsenault[3], Eric L Weiss[4], Christian R Landry[3], Alan M Moses[1,2]

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, M5S 3B2, Toronto, Canada
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada
3. Département de Biologie, IBIS and PROTEO, Pavillon Charles-Eugene-Marchand, 1030 Avenue de la Medecine, Laval University, Québec City, QC G1V 0A6, Canada.
4. Department of Molecular Biosciences, Northwestern University, Evenston, Illinois, United States of America

Supplementary text

Behavior of the likelihood-ratio test on short sequences under the null models

We were first concerned by the fact that motifs identified in our study are short, which may prevent the use of the chi-squared approximation for the distribution of the likelihood-ratio test statistics (Wilks 1938). To address this, we evolved sequences of different lengths according to the null model assumed by the test (see Methods) and assessed whether the likelihood-ratio test statistics were chi-squared distributed. If the likelihood-ratio test statistics were chi-squared distributed, then we would expect uniformly distributed p-values. In our simulation of short sequence evolution, we observed that the p-values obtained were not uniform and that there were much fewer rejections of the null hypothesis than expected (Appendix Figure III-1A). This indicates that this test has low power for very short sequences and that assuming chi-square distribution for the likelihood-ratio test statistic will be conservative. Increasing the branch lengths to allow for more substitutions improved the uniformity slightly, but we never observed higher rates of rejection of the null hypothesis than expected (Appendix Figure III-1B). In phylogenetics studies involving proteome-wide analyses, increased acceptance of the null hypothesis has been suggested as an acceptable compromise to decrease the levels of false positives that may arise due to invalid model assumptions (Zhang et al. 2005; Yang and dos Reis 2011). See the main text for the behavior of the test on more 'realistic' simulations.



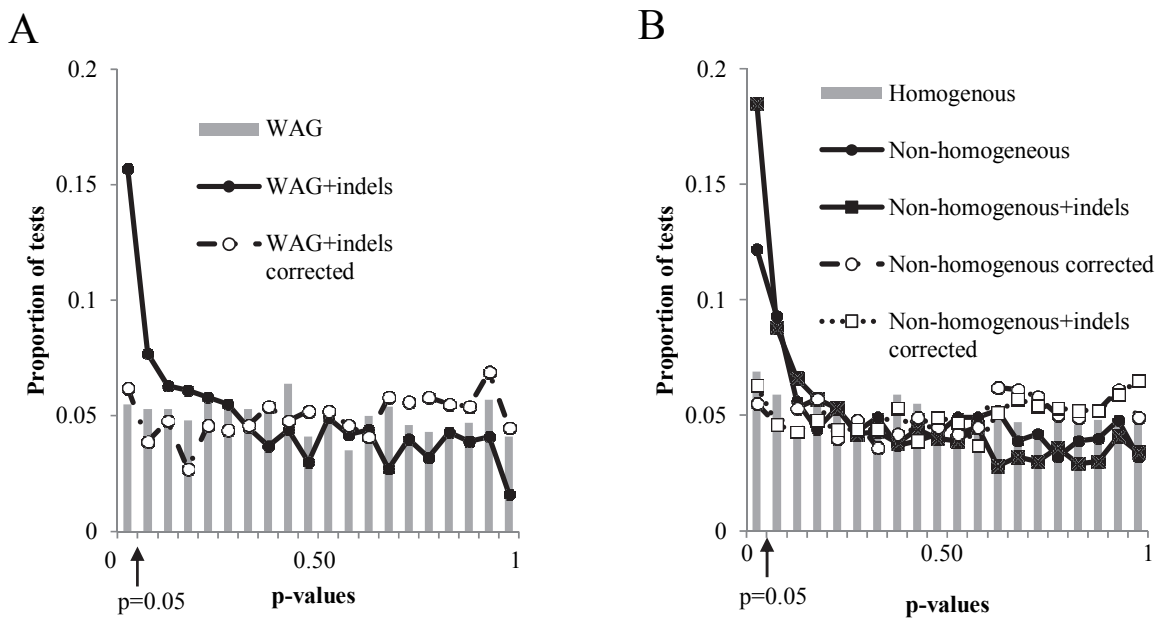
Appendix Figure III-1. The chi-squared approximation of the distribution of likelihood-ratio test on short sequences is conservative. A) Short amino acid sequences of various lengths were evolved under the WAG model with the same phylogenetic tree that follows a global clock (corresponding to the null model assumed by the test). Grey bars show the expected distribution of p-values if the chi-squared approximation is correct. Data points are the obtained distribution of p-values. B) Short linear motifs of length 7 were evolved using the same procedure as in A) but the phylogenetic tree was scaled to allow for more substitutions per sites, showing that more substitutions do not lead to more false rejections than expected for short sequences.

Simulation of protein evolution under various evolutionary models

To illustrate the use of this non-central correction to eliminate false rejections of the null hypothesis when the background evolutionary process deviates from the model assumed by the test, we performed extensive simulation of sequence evolution where there were truly no changes in constraints after gene duplication, but where the simulation violated the model assumed by the test. In each case, we used a likelihood-ratio test to test whether the data is better explained by two rates of evolution as opposed to a single rate of evolution (using AAML (Yoder and Yang 2000), see Methods). Under the null hypothesis, p-values show a

uniform distribution and we use this as a measure for deviations of our test statistic under various simulations. For example, if the p-values are uniform in a simulation with no true positives, then we infer that the null distribution used is correct.

We first simulated proteins evolving under the WAG model (Whelan and Goldman 2001) according to a single rate of evolution, which is the model assumed by the test, and tested for changes in constraints (see Methods). As expected, the distribution of the likelihood-ratio test statistic followed a chi-squared distribution, and the p-values are uniformly distributed (Appendix Figure III-2A, grey columns). Next, because short linear motifs are found in rapidly evolving disordered regions that contain indels, we simulated proteins according to the same model but also allowed indels (using INDELIBLE (Fletcher and Yang 2009), see Methods for indel parameters). We then aligned the proteins (MAFFT (Katoh et al. 2002), see Methods), and repeated the likelihood-ratio test for changes in constraints. The inclusion of indels in the simulation led to an increased rejection rate of the null hypothesis (Appendix Figure III-2A, black circles), presumably because alignment errors lead to analysis of non-homologous residues and the extra rate parameter in the alternative hypothesis can 'fit' some of this heterogeneity. However, after performing the non-central correction to the chi-squared distribution, the distribution of p-values was uniform, indicating that the indel and alignment process was adequately captured by the non-central parameter (Appendix Figure III-2A, white circles). The KL divergence for this particular set of parameters was 0.000837.



Appendix Figure III-2. P-value distribution of the likelihood-ratio test obtained from chi-squared and non-central chi-squared on simulated data. A) Amino acid sequences were evolved under the WAG model with or without indels. Grey bars show the distribution of p-values obtained from the likelihood-ratio test when the data are generated according to the model assumed by the test. Circles indicate the distribution of p-values when indels are also included and data is aligned, and the test statistic is assumed to be chi-squared distributed (black circles) or non-central chi-squared distributed (white circles, "corrected"). B) Protein coding DNA sequences were evolved. Grey bars show the distribution of p-values when sequences are evolved under a homogenous and stationary codon frequency model assumed by the test. Circles indicate the distribution of p-values when the model is non-homogenous and the test statistic is assumed to be chi-squared distributed (black circles) or non-central chi-squared distributed (white circles, "corrected"). Squares indicate the distribution of p-values when the indels are also included and the test statistic is assumed to be chi-squared distributed (black squares) or non-central chi-squared distributed (white squares, "corrected").

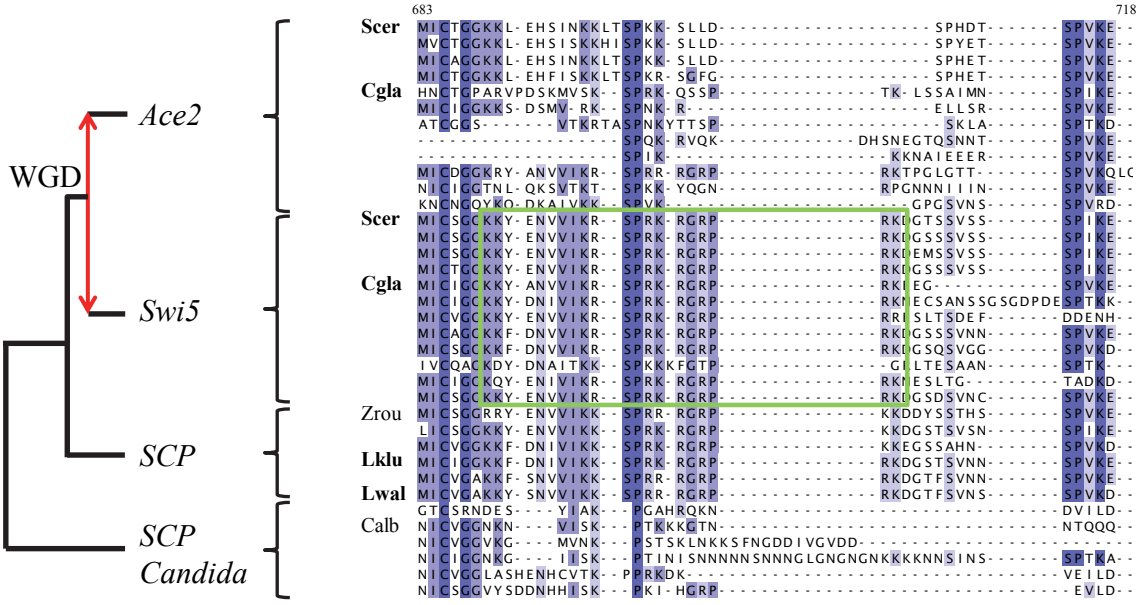
To test whether the non-central correction could account for heterogeneity in the substitution process, we next performed codon-based simulations where we could vary the stationary codon frequencies. We simulated proteins according to a codon model (with Ka/Ks is equal to 1) using the codon frequency table from *Thermus aquaticus* (which is GC-biased), and

found that the likelihood-ratio test statistic followed a chi-squared distribution (Appendix Figure III-2B, grey columns) despite the GC-biased codon model likely leading to amino acid frequencies different than those assumed by the WAG model in the test. Next, we simulated a similar set of proteins, except that one of the tested clades was evolving under the *S. cerevisiae* codon frequency table (which is AT-biased). Because the substitution model changes on the phylogeny, this set of proteins corresponds to proteins evolving under a non-homogenous and non-stationary process. As expected, we observed increased false rejections of the null hypothesis (Appendix Figure III-2B, black circles) because the alternative hypothesis can account for some of this heterogeneity using the additional rate parameter. However, after correction by the non-central parameter, the p-values were now uniformly distributed (Appendix Figure III-2B, white circles). The KL divergence for this set of parameters was 0.001031.

To illustrate how all these deviations of the evolutionary models can be compounded in the analyses, we also evolved proteins under the same non-homogenous, non-stationary processes, but now also included indels (see Methods). Doing this, we observed a much higher false rejection rate of the null hypothesis (Appendix Figure III-1B, black squares); however, even while several factors compounded to deviate from the models assumed in the test, it was still possible to capture the deviation to the null hypothesis using a single non-central parameter (Appendix Figure III-1B, white squares). Only a very small increase in the KL divergence for the combined deviations to the model was observed (0.001059 vs 0.001031 for the heterogeneous substitution model with no indels); however, we observed a much higher false rejection rate of the null-hypothesis because the likelihood-ratio test is performed on more columns (more data points in the test) due to the indels.

Taken together, these results indicate that a non-central chi-squared distribution can be used to capture some of the evolutionary complexities that can be encountered when detecting functional divergence after gene duplication.

Supplementary Figure



Appendix Figure III-3. Alignment of the characterized nuclear localization signal of Swi5 and other species. Green box indicates the characterized nuclear localization signal in the Swi5 lineage. Coordinates of Ace2 are indicated in the alignment. Bolded species indicate species that we tested for their localization in Figure IV-6.

Methods

To simulate short linear motif evolution of different lengths (Appendix Figure III-1), we estimated a phylogenetic tree from Cdc20 (a protein conserved in all eukaryotes) using

AAML and under the global clock model. We used this tree with the program INDELIBLE (Fletcher and Yang 2009) to evolve short sequences of different lengths under the WAG model. These sets of simulated short linear motifs therefore correspond to the null model of the likelihood-ratio test.

To simulate protein sequences under various models of evolution for the purpose of testing the non-central chi-squared correction to the likelihood-ratio test (Appendix Figure III-1), we used the program INDELIBLE (Fletcher and Yang 2009). The same phylogenetic relationship was always used: (((a,b),(c,d)),((e,f),(gh))) with equal branch lengths of 0.8 and a root length of 300 (codons or amino acids). We arbitrarily set the (e,f) clade to be the post-WGD clade and performed likelihood-ratio tests on proteins simulated by the program. Sequences were evolved with the following parameters:

1) WAG model,

2) WAG model with power law distributed indels, with a=1.7 for inserts, and a=1.8 for deletions, with an indel rate of 0.1,

3) Homogeneous codon models with codon stationary frequency equals to the *Thermus aquaticus* codon frequency with kappa=2, omega=1,

4) Non-homogeneous codon models, and codon stationary frequency equals to the *S. cerevisiae* codon frequency for all the tree, except for the (e,f) clade which was set to the *Thermus aquaticus* codon frequency,

5) same as 4) except with indels similar to test 2) except with indel rate of 0.05.

The likelihood-ratio test was then performed on the resulting proteins (or aligned first with MAFFT if indels were simulated).

Appendix IV
Supplementary data for Chapter V

This work has not been previously published.

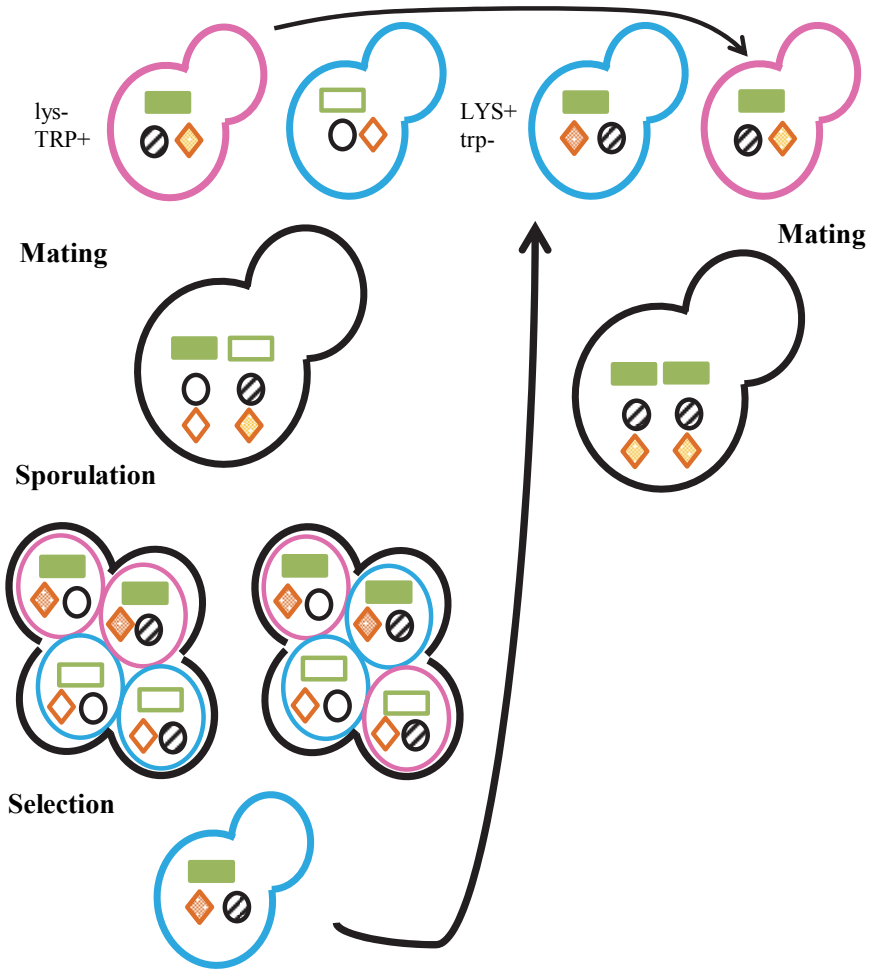
Alex N Nguyen Ba[1,2], Sergio Peisajovich[1], Alan M Moses[1,2,¥]

1. Department of Cell & Systems Biology, University of Toronto, 25 Willcocks Street, M5S 3B2, Toronto, Canada
2. Centre for the Analysis of Genome Evolution and Function, University of Toronto, 25 Willcocks Street, Toronto, Canada

Supplementary text

High-throughput construction of homozygous and heterozygous diploids

Diploids containing all combinations of the marked alleles can also be created (Appendix Figure IV-I). A previous approach to efficiently construct diploid strains used a plasmid that contained a haploid-specific promoter driving transcription of a counter-selectable marker and controlled expression of the HO endonuclease to alter the mating type of cells within a colony (Furukawa et al. 2011). In practice, this approach requires the transformation of a plasmid and requires strains that can only survive with URA3 complementation. Importantly, this approach does not allow the creation of heterozygous diploids which we believe are important when studying the evolution of beneficial mutations. To address this, we designed another strategy to create diploid strains, taking advantage of the previously applied SGA methodology. Briefly, SGA output strains (MATα,Δlys2, TRP1, yfg::marker) are mated with a Δtrp1 strain (MATa,LYS2,Δtrp1) and diploids are selected on media lacking tryptophan and lysine. The cells are then sporulated, and a haploid of the opposing mating type is selected along with lysine prototrophy and tryptophan auxotrophy (MATa,LYS2,Δtrp1, yfg::marker). We use lysine and tryptophan auxotrophy because the markers can be conveniently selected and counter-selected (see Chapter V, Methods). This creates a library of strains with identical marker as the original set of strains but of opposite mating type and with complementary auxotrophies.

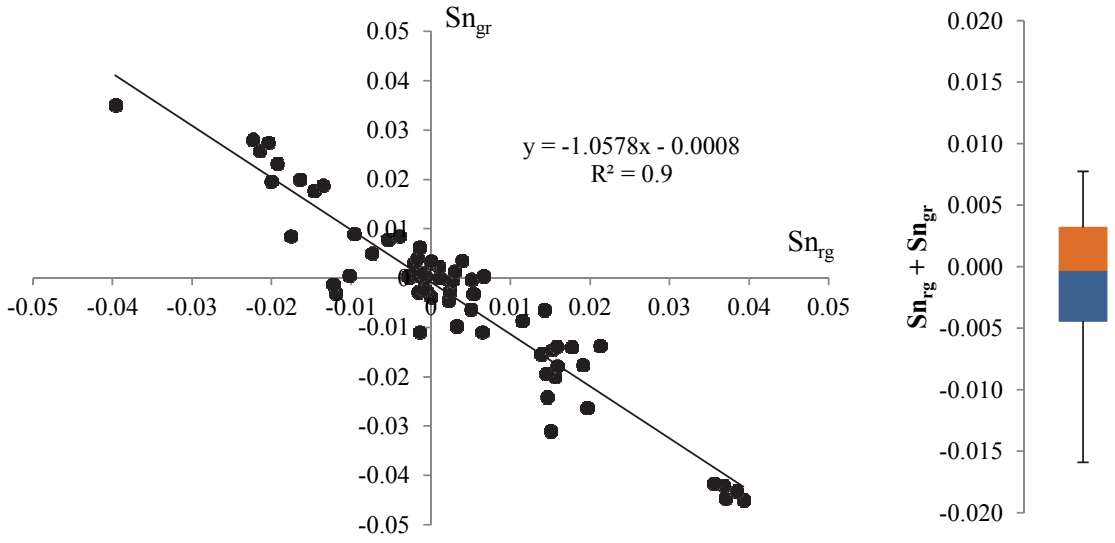


Appendix Figure IV-1. SGA methodology to create diploids. The SGA methodology can be used to create homozygous or heterozygous diploids from starting SGA output strains. The SGA output strain (pink: MATα) is converted to the opposite mating type (blue: MATa) along with different auxotrophic markers. The converted strain can be effectively mated back to the original ordered array to create a diploid strain.

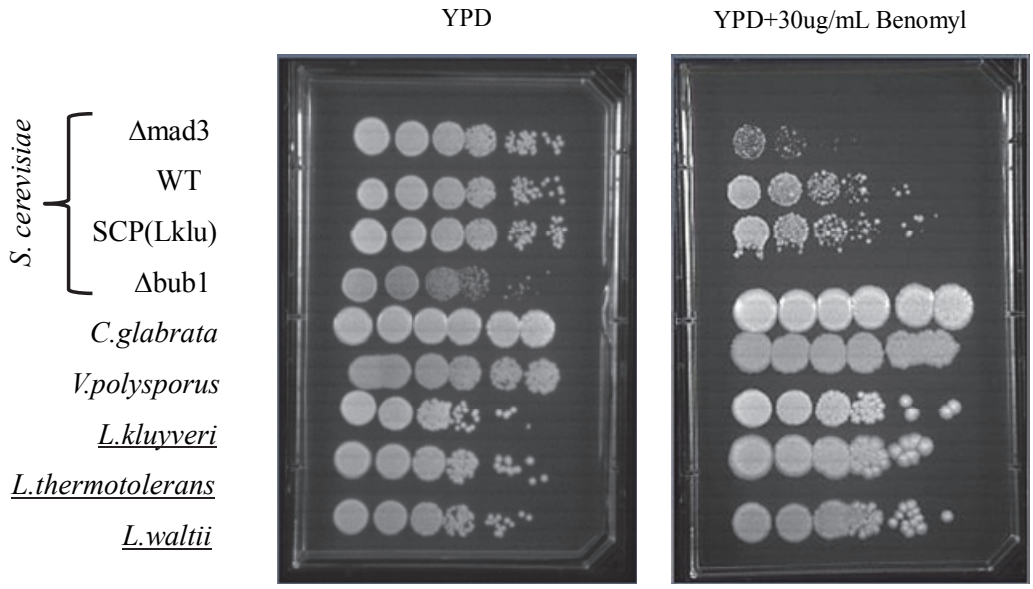
Reproducibility in selection coefficient measurements

Our high-throughput strain construction and fitness measurements allows several controls to be included that can assess the reproducibility of the selection coefficient measurements. Briefly, strains are constructed twice independently, marked with red and green fluorescent protein. Therefore, when performing high-throughput fitness measurements, competition between genotypes is performed twice on the 96-well plate (swapping the fluorescent

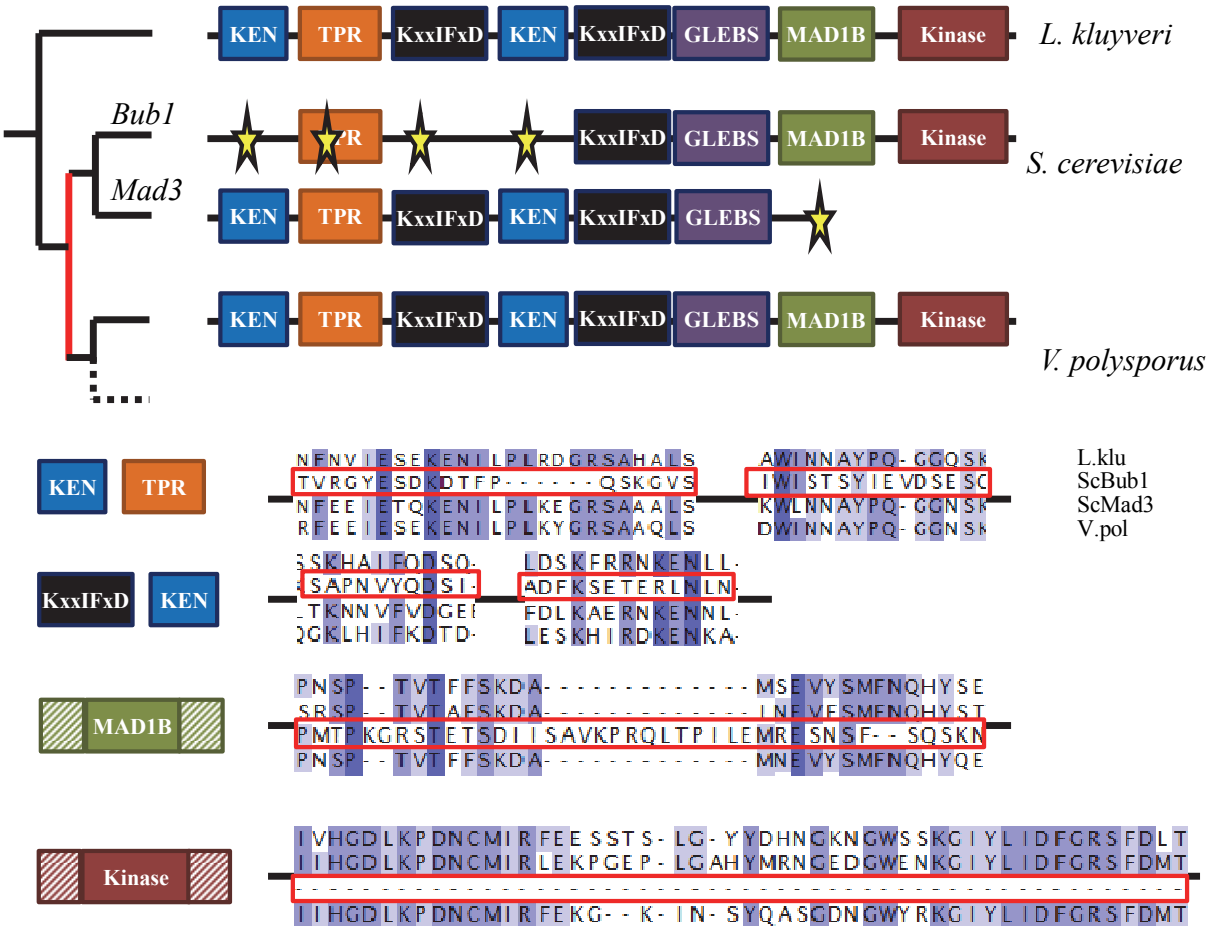
proteins between competitions). If our measurements are reproducible, we would expect very small variations between the two replicates. Indeed, on all the fitness measurements produced for this study, we observed an R^2 of 0.9983 with a slope of -1.0014 between the two replicates. We were concerned that this correlation was driven by very sick genotypes (where we arbitrarily set the relative selection coefficient to 0.3) and assessed the correlation of our measurements without these. This resulted in a slightly lower correlation (R^2 of 0.9) with a slope of -1.058 (Appendix Figure IV-2). To further quantify the variation between replicates, we also plotted the sum of the reciprocal measurements. Under ideal conditions, we would expect this sum to be zero. As expected for highly reproducible measurements, we observed that the median of this sum was -0.00042. However, we did observe cases where the measured fitness effects varied by a large amount (0.016 as the strongest deviation in measurements). These deviations occurred when strains were phenotypically sick and similar effects have been observed in previous use of the competitive fitness assay (see Breslow et al, 2008).



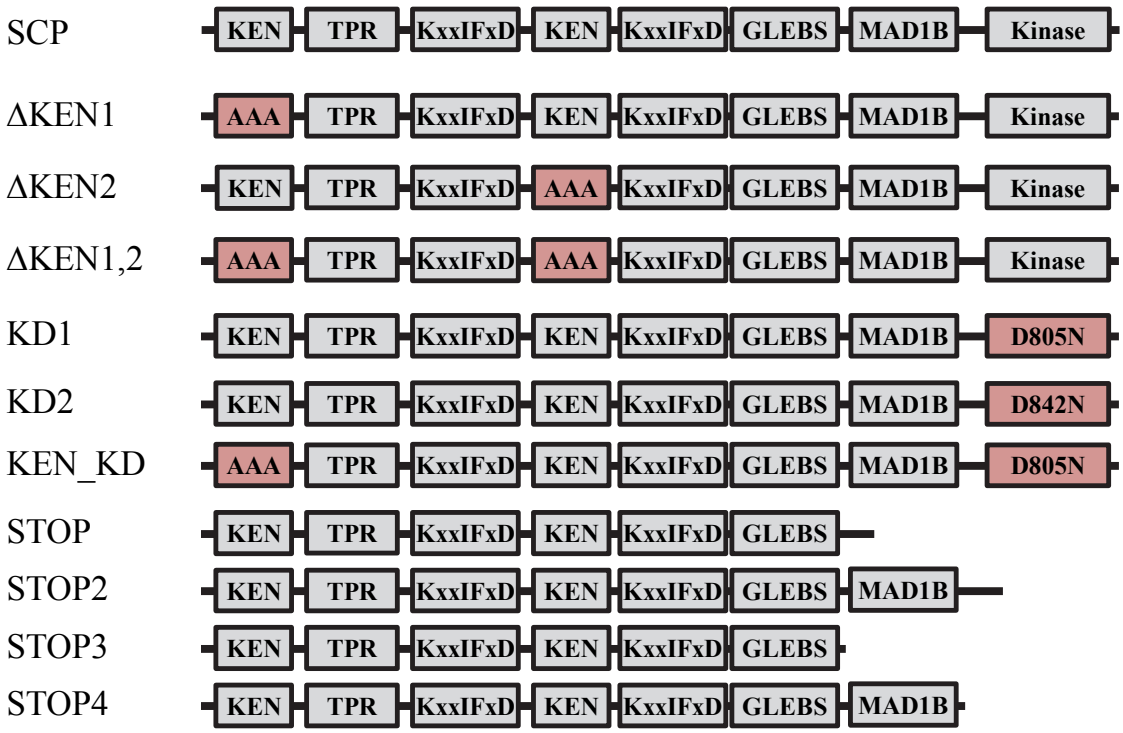
Appendix Figure IV-2. Variations in selection coefficient measurements between replicates.
Competitions between genotypes with swapped fluorescent proteins produce highly reproducible fitness measurements. Sn_{gr} and Sn_{rg} are the measured selection coefficients for reciprocal experiments. The sum of the two measurements is expected to be zero for reciprocal measurements, and the observed distribution of the sum is represented in a quantile plot.



Appendix Figure IV-3. Spot dilution assay of yeast species on benomyl. *S.cerevisiae* is highly sensitive to benomyl, and the sensitivity can be rescued using the single-copy protein from *L. kluyveri*. Species underlined have not undergone the whole-genome duplication, but still grow well on benomyl containing plates.



Appendix Figure IV-4. Sequence analysis of *V. polysporus* shows reversion to single-copy protein.
Sequence alignment of the regions of interests indicates that *V. polysporus* reverted to a single-copy protein. Shown is a schematic of the phylogenetic gene tree of the studied proteins. Branch lengths are not to scale. The red line indicates the duplication event. Stars and boxes are as in Figure V-1B.



Appendix Figure IV-5. Schematic of constructs. Various mutations of the single-copy proteins were assayed for fitness using spot dilution assays (all), or our high-throughput fitness assay (first four constructs). In red are the indicated mutations that abolish function of the functional element. *KD*: Kinase-dead.

References

Abhiman S, Sonnhammer ELL. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins 60:758–768.

Abramoff MD, Magalhaes PJ, Ram SJ. 2004. Image Processing with ImageJ. Biophotonics Int. 11:36–42.

Agrafioti I, Swire J, Abbott J, Huntley D, Butcher S, Stumpf MPH. 2005. Comparative analysis of the Saccharomyces cerevisiae and Caenorhabditis elegans protein interaction networks. BMC Evol. Biol. 5:23.

Alber F, Dokudovskaya S, Veenhoff LM, et al. 2007. The molecular architecture of the nuclear pore complex. Nature 450:695–701.

Alexander J, Lim D, Joughin BA, et al. 2011. Spatial exclusivity combined with positive and negative selection of phosphorylation motifs is the basis for context-dependent mitotic signaling. Sci. Signal. 4:ra42.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, Van de Peer Y. 2010. Posttranslational regulation impacts the fate of duplicated genes. Proc. Natl. Acad. Sci. U. S. A. 107:2967–2971.

Ang XL, Wade Harper J. 2005. SCF-mediated protein degradation and cell cycle control. Oncogene 24:2860–2870.

Bader GD, Hogue CWV. 2003. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4:2.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol. 2:28–36.

Baker CR, Booth LN, Sorrells TR, Johnson AD. 2012. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. Cell 151:80–95.

Baker CR, Hanson-Smith V, Johnson AD. 2013. Following gene duplication, paralog interference constrains transcriptional circuit evolution. Science 342:104–108.

Baker CR, Tuch BB, Johnson AD. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. Proc. Natl. Acad. Sci. U. S. A. 108:7493–7498.

Balla S, Thapar V, Verma S, et al. 2006. Minimotif Miner: a tool for investigating protein function. Nat Meth 3:175–177.

Barbosa-Morais NL, Irimia M, Pan Q, et al. 2012. The evolutionary landscape of alternative splicing in vertebrate species. Science 338:1587–1593.

De Beer T, Hoofnagle AN, Enmon JL, Bowers RC, Yamabhai M, Kay BK, Overduin M. 2000. Molecular mechanism of NPF recognition by EH domains. Nat. Struct. Biol. 7:1018–1022.

Bellay J, Han S, Michaut M, et al. 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 12:R14.

Beltrao P, Albanèse V, Kenner LR, et al. 2012. Systematic functional prioritization of protein posttranslational modifications. Cell 150:413–425.

Beltrao P, Bork P, Krogan NJ, van Noort V. 2013. Evolution and functional cross-talk of protein post-translational modifications. Mol. Syst. Biol. 9:714.

Beltrao P, Serrano L. 2005. Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. PLoS Comput. Biol. 1:e26.

Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim WA, Shokat KM, Burlingame AL, Krogan NJ. 2009. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. PLoS Biol. 7:e1000134.

Bendall SC, Simonds EF, Qiu P, et al. 2011. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332:687–696.

Benton D, Krishnamoorthy K. 2003. Computing discrete mixtures of continuous distributions: noncentral chisquare, noncentral t and the distribution of the square of the sample multiple correlation coefficient. Comput. Stat. Data Anal. 43:249–267.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235–242.

Bertin N, Simonis N, Dupuy D, Cusick ME, Han J-DJ, Fraser HB, Roth FP, Vidal M. 2007. Confirmation of organized modularity in the yeast interactome. PLoS Biol. 5:e153.

Bilsland-Marchesan E, Ariño J, Saito H, Sunnerhagen P, Posas F. 2000. Rck2 kinase is a substrate for the osmotic stress-activated mitogen-activated protein kinase Hog1. Mol. Cell. Biol. 20:3887–3895.

Blobel G, Dobberstein B. 1975. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. J. Cell Biol. 67:835–851.

Blom N, Gammeltoft S, Brunak S. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. 294:1351–1362.

Boeke JD, Trueheart J, Natsoulis G, Fink GR. 1987. 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. Methods Enzymol. 154:164–175.

Bolanos-Garcia VM, Kiyomitsu T, D'Arcy S, et al. 2009. The crystal structure of the N-terminal region of BUB1 provides insight into the mechanism of BUB1 recruitment to kinetochores. Struct. Lond. Engl. 1993 17:105–116.

Bonifacino JS, Traub LM. 2003. Signals for sorting of transmembrane proteins to endosomes and lysosomes. Annu. Rev. Biochem. 72:395–447.

Botstein D, Chervitz SA, Cherry JM. 1997. Yeast as a model organism. Science 277:1259–1260.

Breitkreutz A, Choi H, Sharom JR, et al. 2010. A global protein kinase and phosphatase interaction network in yeast. Science 328:1043–1046.

Breslow DK, Cameron DM, Collins SR, et al. 2008. A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. Nat. Methods 5:711–718.

Brocca S, Samalíková M, Uversky VN, Lotti M, Vanoni M, Alberghina L, Grandori R. 2009. Order propensity of an intrinsically disordered protein, the cyclin-dependent-kinase inhibitor Sic1. Proteins 76:731–746.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J. Mol. Evol. 55:104–110.

Budovskaya YV, Stephan JS, Deminoff SJ, Herman PK. 2005. An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. Proc. Natl. Acad. Sci. U. S. A. 102:13933–13938.

Burnett G, Kennedy EP. 1954. The enzymatic phosphorylation of proteins. J. Biol. Chem. 211:969–980.

Burton JL, Solomon MJ. 2007. Mad3p, a pseudosubstrate inhibitor of APCCdc20 in the spindle assembly checkpoint. Genes Dev. 21:655–667.

Buschhorn BA, Peters J-M. 2006. How APC/C orders destruction. Nat. Cell Biol. 8:209–211.

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

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

Cain C, Miller S, Ahn J, Prives C. 2000. The N terminus of p53 regulates its dissociation from DNA. J. Biol. Chem. 275:39944–39953.

Cartwright P, Helin K. 2000. Nucleocytoplasmic shuttling of transcription factors. Cell. Mol. Life Sci. CMLS 57:1193–1206.

Cartwright RA. 2006. Logarithmic gap costs decrease alignment accuracy. BMC Bioinformatics 7:527.

Chang JS, Winston F. 2011. Spt10 and Spt21 Are Required for Transcriptional Silencing in Saccharomyces cerevisiae. Eukaryot. Cell 10:118–129.

Chao WCH, Kulkarni K, Zhang Z, Kong EH, Barford D. 2012. Structure of the mitotic checkpoint complex. Nature 484:208–213.

Charette JM, Baserga SJ. 2010. The DEAD-box RNA helicase-like Utp25 is an SSU processome component. RNA N. Y. N 16:2156–2169.

Chattoo BB, Sherman F, Azubalis DA, Fjellstedt TA, Mehnert D, Ogur M. 1979. Selection of lys2 Mutants of the Yeast SACCHAROMYCES CEREVISIAE by the Utilization of alpha-AMINOADIPATE. Genetics 93:51–65.

Cheng W-C, Teng X, Park HK, Tucker CM, Dunham MJ, Hardwick JM. 2008. Fis1 deficiency selects for compensatory mutations responsible for cell death and growth control defects. Cell Death Differ. 15:1838–1846.

Cheng Z, Ventura M, She X, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88–93.

Chial HJ, Rout MP, Giddings TH, Winey M. 1998. Saccharomyces cerevisiae Ndc1p is a shared component of nuclear pore complexes and spindle pole bodies. J. Cell Biol. 143:1789–1800.

Chica C, Labarga A, Gould CM, López R, Gibson TJ. 2008. A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC Bioinformatics 9:229.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87.

Choi E, Dial JM, Jeong D-E, Hall MC. 2008. Unique D box and KEN box sequences limit ubiquitination of Acm1 and promote pseudosubstrate inhibition of the anaphase-promoting complex. J. Biol. Chem. 283:23701–23710.

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

Cohen P. 2000. The regulation of protein function by multisite phosphorylation--a 25 year update. Trends Biochem. Sci. 25:596–601.

Cohen-Fix O, Peters JM, Kirschner MW, Koshland D. 1996. Anaphase initiation in Saccharomyces cerevisiae is controlled by the APC-dependent degradation of the anaphase inhibitor Pds1p. Genes Dev. 10:3081–3093.

Collins MO. 2009. Cell biology. Evolving cell signals. Science 325:1635–1636.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9:938–950.

Cortese MS, Uversky VN, Dunker AK. 2008. Intrinsic disorder in scaffold proteins: getting more from less. Prog. Biophys. Mol. Biol. 98:85–106.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Darsow T, Burd CG, Emr SD. 1998. Acidic di-leucine motif essential for AP-3-dependent sorting and restriction of the functional specificity of the Vam3p vacuolar t-SNARE. J. Cell Biol. 142:913–922.

Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. 1997. The C-terminal half of the anti-sigma factor, FlgM, becomes structured when bound to its target, sigma 28. Nat. Struct. Biol. 4:285–291.

Davey NE, Edwards RJ, Shields DC. 2010. Computational identification and analysis of protein short linear motifs. Front. Biosci. J. Virtual Libr. 15:801–825.

Davey NE, Shields DC, Edwards RJ. 2006. SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res. 34:3546–3554.

Dawson R, Müller L, Dehner A, Klein C, Kessler H, Buchner J. 2003. The N-terminal domain of p53 is natively unfolded. J. Mol. Biol. 332:1131–1141.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat. Rev. Genet. 8:675–688.

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

Denning DP, Patel SS, Uversky V, Fink AL, Rexach M. 2003. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. Proc. Natl. Acad. Sci. U. S. A. 100:2450–2455.

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. 19:1114–1121.

Dernburg AF. 2001. Here, there, and everywhere: kinetochore function on holocentric chromosomes. J. Cell Biol. 153:F33–38.

Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ. 2008. Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front. Biosci. J. Virtual Libr. 13:6580–6603.

Dinkel H, Michael S, Weatheritt RJ, et al. 2012. ELM--the database of eukaryotic linear motifs. Nucleic Acids Res. 40:D242–251.

Doucet J, Benoit JP. 1987. Molecular dynamics studied by analysis of the X-ray diffuse scattering from lysozyme crystals. Nature 325:643–646.

Doxey AC, Cheng Z, Moffatt BA, McConkey BJ. 2010. Structural motif screening reveals a novel, conserved carbohydrate-binding surface in the pathogenesis-related protein PR-5d. BMC Struct. Biol. 10:23.

Doxey AC, Yaish MW, Griffith M, McConkey BJ. 2006. Ordered surface carbons distinguish antifreeze proteins and their ice-binding regions. Nat. Biotechnol. 24:852–855.

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. Biochemistry (Mosc.) 41:6573–6582.

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 272:5129–5148.

Dunker AK, Lawson JD, Brown CJ, et al. 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19:26–59.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Twelfth. Cambridge: Cambridge University Press

Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. 6:197–208.

Eddy SR. 1998. Profile hidden Markov models. Bioinforma. Oxf. Engl. 14:755–763.

Edwards RJ, Davey NE, Shields DC. 2007. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PloS One 2:e967.

Feine O, Zur A, Mahbubani H, Brandeis M. 2007. Human Kid is degraded by the APC/C(Cdh1) but not by the APC/C(Cdc20). Cell Cycle Georget. Tex 6:2516–2523.

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

Fernández A, Lynch M. 2011. Non-adaptive origins of interactome complexity. Nature 474:502–505.

Fernius J, Hardwick KG. 2007. Bub1 kinase targets Sgo1 to ensure efficient chromosome biorientation in budding yeast mitosis. PLoS Genet. 3:e213.

Finn RD, Tate J, Mistry J, et al. 2008. The Pfam protein families database. Nucleic Acids Res. 36:D281–288.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. Nature 481:360–364.

Fischer EH, Krebs EG. 1955. Conversion of phosphorylase b to phosphorylase a in muscle extracts. J. Biol. Chem. 216:121–132.

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol. Biol. Evol. 27:2257–2267.

Flicek P, Amode MR, Barrell D, et al. 2011. Ensembl 2011. Nucleic Acids Res. 39:D800–806.

Fong JH, Shoemaker BA, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV, Panchenko AR. 2009. Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. PLoS Comput. Biol. 5:e1000316.

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

Forman-Kay JD, Mittag T. 2013. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Struct. Lond. Engl. 1993 21:1492–1499.

Frenkel EM, Good BH, Desai MM. 2014. The fates of mutant lineages and the distribution of fitness effects of beneficial mutations in laboratory budding yeast populations. Genetics 196:1217–1226.

Freschi L, Courcelles M, Thibault P, Michnick SW, Landry CR. 2011. Phosphorylation network rewiring by gene duplication. Mol. Syst. Biol. 7:504.

Freschi L, Osseni M, Landry CR. 2014. Functional divergence and evolutionary turnover in mammalian phosphoproteomes. PLoS Genet. 10:e1004062.

Furukawa K, Furukawa T, Hohmann S. 2011. Efficient construction of homozygous diploid strains identifies genes required for the hyper-filamentous phenotype in Saccharomyces cerevisiae. PloS One 6:e26584.

Gagnon-Arsenault I, Marois Blanchet F-C, Rochette S, Diss G, Dubé AK, Landry CR. 2013. Transcriptional divergence plays a role in the rewiring of protein interaction networks after gene duplication. J. Proteomics 81:112–125.

Gelperin DM, White MA, Wilkinson ML, et al. 2005. Biochemical and genetic analysis of the yeast proteome with a movable ORF collection. Genes Dev. 19:2816–2826.

Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. Nature 425:737–741.

Giaever G, Chu AM, Ni L, et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. Nature 418:387–391.

Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M. 2007. PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol. 8:R250.

Gokhman D, Lavi E, Prüfer K, Fraga MF, Riancho JA, Kelso J, Pääbo S, Meshorer E, Carmel L. 2014. Reconstructing the DNA Methylation Maps of the Neandertal and the Denisovan. Science [Internet]. Available from: http://www.sciencemag.org/content/early/2014/04/16/science.1250368.abstract

Goldring ES, Grossman LI, Krupnick D, Cryer DR, Marmur J. 1970. The petite mutation in yeast. Loss of mitochondrial deoxyribonucleic acid during induction of petites with ethidium bromide. J. Mol. Biol. 52:323–335.

Gompel N, Prud'homme B. 2009. The causes of repeated genetic evolution. Dev. Biol. 332:36–47.

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

Görlich D, Mattaj IW. 1996. Nucleocytoplasmic transport. Science 271:1513–1518.

Gould CM, Diella F, Via A, et al. 2010. ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. 38:D167–180.

Graves DJ, Fischer EH, Krebs EG. 1960. Specificity studies on muscle phosphorylase phosphatase. J. Biol. Chem. 235:805–809.

Grigoriev IV, Nikitin R, Haridas S, et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 42:D699–704.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc. Natl. Acad. Sci. U. S. A. 102:707–712.

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. 2013. An update of DIVERGE software for functional divergence analysis of protein family. Mol. Biol. Evol. 30:1713–1719.

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

Gu Z, Rifkin SA, White KP, Li W-H. 2004. Duplicate genes increase gene expression diversity within and between species. Nat. Genet. 36:577–579.

Guan Y, Dunham MJ, Troyanskaya OG. 2007. Functional analysis of gene duplications in Saccharomyces cerevisiae. Genetics 175:933–943.

Hao N, Behar M, Parnell SC, Torres MP, Borchers CH, Elston TC, Dohlman HG. 2007. A systems-biology analysis of feedback inhibition in the Sho1 osmotic-stress-response pathway. Curr. Biol. CB 17:659–667.

Hao N, Budnik BA, Gunawardena J, O'Shea EK. 2013. Tunable signal processing through modular control of transcription factor translocation. Science 339:460–464.

Hardwick KG, Johnston RC, Smith DL, Murray AW. 2000. MAD3 encodes a novel component of the spindle checkpoint which interacts with Bub3p, Cdc20p, and Mad2p. J. Cell Biol. 148:871–882.

Hasty J, McMillen D, Isaacs F, Collins JJ. 2001. Computational studies of gene regulatory networks: in numero molecular biology. Nat. Rev. Genet. 2:268–279.

Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput. Biol. 2:e100.

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. 2009. Predicting intrinsic disorder in proteins: an overview. Cell Res. 19:929–949.

Hegyi H, Tompa P. 2008. Intrinsically disordered proteins display no preference for chaperone binding in vivo. PLoS Comput. Biol. 4:e1000017.

Heinicke S, Livstone MS, Lu C, Oughtred R, Kang F, Angiuoli SV, White O, Botstein D, Dolinski K. 2007. The Princeton Protein Orthology Database (P-POD): a comparative genomics analysis tool for biologists. PloS One 2:e766.

Hendrickson C, Meyn MA, Morabito L, Holloway SL. 2001. The KEN box regulates Clb2 proteolysis in G1 and at the metaphase-to-anaphase transition. Curr. Biol. CB 11:1781–1787.

Hietpas RT, Jensen JD, Bolon DNA. 2011. Experimental illumination of a fitness landscape. Proc. Natl. Acad. Sci. U. S. A. 108:7896–7901.

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

Ho Y, Gruhler A, Heilbut A, et al. 2002. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183.

Van der Hoeven N. 2005. The probability to select the correct model using likelihood-ratio based criteria in choosing between two nested models of which the more extended one is true. J. Stat. Plan. Inference 135:477–486.

Hollenhorst PC, Bose ME, Mielke MR, Müller U, Fox CA. 2000. Forkhead genes in transcriptional silencing, cell morphology and the cell cycle. Overlapping and distinct functions for FKH1 and FKH2 in Saccharomyces cerevisiae. Genetics 154:1533–1548.

Hollenhorst PC, Pietz G, Fox CA. 2001. Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. Genes Dev. 15:2445–2456.

Holt LJ, Tuch BB, Villén J, Johnson AD, Gygi SP, Morgan DO. 2009. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. Science 325:1682–1686.

Van Hoof A. 2005. Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. Genetics 171:1455–1461.

Huang Y-F, Golding GB. 2012. Inferring sequence regions under functional divergence in duplicate genes. Bioinforma. Oxf. Engl. 28:176–183.

Huh W-K, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. 2003. Global analysis of protein localization in budding yeast. Nature 425:686–691.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJA. 2006. The PROSITE database. Nucleic Acids Res. 34:D227–230.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. 14:1870–1879.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 32:1037–1049.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U. S. A. 98:4569–4574.

Jacobs D, Glossip D, Xing H, Muslin AJ, Kornfeld K. 1999. Multiple docking sites on substrate proteins form a modular system that mediates recognition by ERK MAP kinase. Genes Dev. 13:163–175.

Janin J, Sternberg MJE. 2013. Protein flexibility, not disorder, is intrinsic to molecular recognition. F1000 Biol. Rep. 5:2.

Jansen JM, Wanless AG, Seidel CW, Weiss EL. 2009. Cbk1 regulation of the RNA-binding protein Ssd1 integrates cell fate with translational control. Curr. Biol. CB 19:2114–2120.

Jaspersen SL, Morgan DO. 2000. Cdc14 activates cdc15 to promote mitotic exit in budding yeast. Curr. Biol. CB 10:615–618.

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P. 2006. Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature 443:594–597.

Jeong H, Mason SP, Barabási AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jiménez JL, Hegemann B, Hutchins JRA, Peters J-M, Durbin R. 2007. A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. Genome Biol. 8:R90.

Johnson LN, Noble ME, Owen DJ. 1996. Active and inactive protein kinases: structural basis for regulation. Cell 85:149–158.

Johnson SA, Hunter T. 2005. Kinomics: methods for deciphering the kinome. Nat Meth 2:17–25.

Jones DT, Taylor WR, Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry (Mosc.) 33:3038–3049.

Josephides C, Moses AM. 2011. Modeling the evolution of a classic genetic switch. BMC Syst. Biol. 5:24.

Juang YL, Huang J, Peters JM, McLaughlin ME, Tai CY, Pellman D. 1997. APC-mediated proteolysis of Ase1 and the morphogenesis of the mitotic spindle. Science 275:1311–1314.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

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

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–254.

Kelly MT, MacCallum DM, Clancy SD, Odds FC, Brown AJP, Butler G. 2004. The Candida albicans CaACE2 gene affects morphogenesis, adherence and virulence. Mol. Microbiol. 53:969–983.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.

Kimura M. 1962. On the probability of fixation of mutant genes in a population. Genetics 47:713–719.

King EMJ, van der Sar SJA, Hardwick KG. 2007. Mad3 KEN boxes mediate both Cdc20 and Mad3 turnover, and are critical for the spindle checkpoint. PloS One 2:e342.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. Science 188:107–116.

Knudsen B, Miyamoto MM, Laipis PJ, Silverman DN. 2003. Using evolutionary rates to investigate protein functional divergence and conservation. A case study of the carbonic anhydrases. Genetics 164:1261–1269.

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. U. S. A. 98:14512–14517.

Kobe B, Kampmann T, Forwood JK, Listwan P, Brinkworth RI. 2005. Substrate specificity of protein kinases and computational prediction of substrates. Biochim. Biophys. Acta 1754:200–209.

Koch R, Ledermann R, Urwyler O, Heller M, Suter B. 2009. Systematic functional analysis of Bicaudal-D serine phosphorylation and intragenic suppression of a female sterile allele of BicD. PloS One 4:e4552.

Kõivomägi M, Valk E, Venta R, Iofik A, Lepiku M, Balog ERM, Rubin SM, Morgan DO, Loog M. 2011. Cascades of multisite phosphorylation control Sic1 destruction at the onset of S phase. Nature 480:128–131.

Kõivomägi M, Valk E, Venta R, Iofik A, Lepiku M, Morgan DO, Loog M. 2011. Dynamics of Cdk1 substrate specificity during the cell cycle. Mol. Cell 42:610–623.

Kondo A, Ueda M. 2004. Yeast cell-surface display--applications of molecular display. Appl. Microbiol. Biotechnol. 64:28–40.

Kressler D, de la Cruz J, Rojo M, Linder P. 1998. Dbp6p is an essential putative ATP-dependent RNA helicase required for 60S-ribosomal-subunit assembly in Saccharomyces cerevisiae. Mol. Cell. Biol. 18:1855–1865.

Kur K, Gabriel I, Morschhäuser J, Barchiesi F, Spreghini E, Milewski S. 2010. Disruption of homocitrate synthase genes in Candida albicans affects growth but not virulence. Mycopathologia 170:397–402.

Lai ACW, Nguyen Ba AN, Moses AM. 2012. Predicting kinase substrates using conservation of local motif density. Bioinforma. Oxf. Engl. 28:962–969.

Landry CR, Levy ED, Michnick SW. 2009. Weak functional constraints on phosphoproteomes. Trends Genet. TIG 25:193–197.

Lanfear R. 2011. The local-clock permutation test: a simple test to compare rates of molecular evolution on phylogenetic trees. Evol. Int. J. Org. Evol. 65:606–611.

Lang GI, Murray AW, Botstein D. 2009. The cost of gene expression underlies a fitness trade-off in yeast. Proc. Natl. Acad. Sci. U. S. A. 106:5755–5760.

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM. 2013. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500:571–574.

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. 2007. Classical nuclear localization signals: definition, function, and interaction with importin alpha. J. Biol. Chem. 282:5101–5105.

Larsen NA, Harrison SC. 2004. Crystal structure of the spindle assembly checkpoint protein Bub3. J. Mol. Biol. 344:885–892.

Lau CK, Giddings TH, Winey M. 2004. A novel allele of Saccharomyces cerevisiae NDC1 reveals a potential role for the spindle pole body component Ndc1p in nuclear pore assembly. Eukaryot. Cell 3:447–458.

Leadsham JE, Miller K, Ayscough KR, Colombo S, Martegani E, Sudbery P, Gourlay CW. 2009. Whi2p links nutritional sensing to actin-dependent Ras-cAMP-PKA regulation and apoptosis in yeast. J. Cell Sci. 122:706–715.

Lee G, Saito I. 1998. Role of nucleotide sequences of loxP spacer region in Cre-mediated recombination. Gene 216:55–65.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Ségurel L, Venkat A, Andolfatto P, Przeworski M. 2012. Revisiting an old riddle: what determines genetic diversity levels within species? PLoS Biol. 10:e1001388.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature 424:147–151.

Li SS-C. 2005. Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. Biochem. J. 390:641–653.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. Trends Genet. TIG 21:602–607.

Li Z, Vizeacoumar FJ, Bahr S, et al. 2011. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. Nat. Biotechnol. 29:361–367.

Lieber DS, Elemento O, Tavazoie S. 2010. Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. PloS One 5:e14444.

Liebmann B, Mühleisen TW, Müller M, Hecht M, Weidner G, Braun A, Brock M, Brakhage AA. 2004. Deletion of the Aspergillus fumigatus lysine biosynthesis gene lysF encoding homoaconitase leads to attenuated virulence in a low-dose mouse infection model of invasive aspergillosis. Arch. Microbiol. 181:378–383.

Lienhard GE. 2008. Non-functional phosphorylations? Trends Biochem. Sci. 33:351–352.

Lim WA, Pawson T. 2010. Phosphotyrosine signaling: evolving a new cellular communication system. Cell 142:661–667.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. Struct. Lond. Engl. 1993 11:1453–1459.

Linding R, Jensen LJ, Ostheimer GJ, et al. 2007. Systematic discovery of in vivo phosphorylation networks. Cell 129:1415–1426.

Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, Yaffe MB, Pawson T. 2008. NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res. 36:D695–699.

Liti G, Carter DM, Moses AM, et al. 2009. Population genomics of domestic and wild yeasts. Nature 458:337–341.

Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. 2006. Intrinsic disorder in transcription factors. Biochemistry (Mosc.) 45:6873–6888.

London N, Ceto S, Ranish JA, Biggins S. 2012. Phosphoregulation of Spc105 by Mps1 and PP1 regulates Bub1 localization to kinetochores. Curr. Biol. CB 22:900–906.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401–1404.

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics 159:1789–1804.

Lynch M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. U. S. A. 104 Suppl 1:8597–8604.

Lynch M. 2007b. The evolution of genetic networks by non-adaptive processes. Nat. Rev. Genet. 8:803–813.

Macek B, Gnad F, Soufi B, Kumar C, Olsen JV, Mijakovic I, Mann M. 2008. Phosphoproteome analysis of E. coli reveals evolutionary conservation of bacterial Ser/Thr/Tyr phosphorylation. Mol. Cell. Proteomics MCP 7:299–307.

Macias MJ, Wiesner S, Sudol M. 2002. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. 513:30–37.

Maguire SL, ÓhÉigeartaigh SS, Byrne KP, Schröder MS, O'Gaora P, Wolfe KH, Butler G. 2013. Comparative genome analysis and gene finding in Candida species using CGOB. Mol. Biol. Evol. 30:1281–1291.

Main BJ, Smith AD, Jang H, Nuzhdin SV. 2013. Transcription start site evolution in Drosophila. Mol. Biol. Evol. 30:1966–1974.

Malik R, Nigg EA, Körner R. 2008. Comparative conservation analysis of the human mitotic phosphoproteome. Bioinforma. Oxf. Engl. 24:1426–1432.

Manning G, Young SL, Miller WT, Zhai Y. 2008. The protist, Monosiga brevicollis, has a tyrosine kinase signaling network more elaborate and diverse than found in any known metazoan. Proc. Natl. Acad. Sci. U. S. A. 105:9674–9679.

Marín M, Uversky VN, Ott T. 2013. Intrinsic disorder in pathogen effectors: protein flexibility as an evolutionary hallmark in a molecular arms race. Plant Cell 25:3153–3157.

Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol. 9:R54.

Marshall AN, Montealegre MC, Jiménez-López C, Lorenz MC, van Hoof A. 2013. Alternative splicing and subfunctionalization generates functional diversity in fungal proteomes. PLoS Genet. 9:e1003376.

Mazanka E, Alexander J, Yeh BJ, Charoenpong P, Lowery DM, Yaffe M, Weiss EL. 2008. The NDR/LATS family kinase Cbk1 directly controls transcriptional asymmetry. PLoS Biol. 6:e203.

Medintz IL, Uyeda HT, Goldman ER, Mattoussi H. 2005. Quantum dot bioconjugates for imaging, labelling and sensing. Nat. Mater. 4:435–446.

Merhi A, Gérard N, Lauwers E, Prévost M, André B. 2011. Systematic mutational analysis of the intracellular regions of yeast Gap1 permease. PloS One 6:e18457.

Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5:e1000376.

Miller ML, Blom N. 2009. Kinase-specific prediction of protein phosphorylation sites. Methods Mol. Biol. Clifton NJ 527:299–310, x.

Mittag T, Orlicky S, Choy W-Y, Tang X, Lin H, Sicheri F, Kay LE, Tyers M, Forman-Kay JD. 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. Proc. Natl. Acad. Sci. U. S. A. 105:17772–17777.

Mizianty MJ, Uversky V, Kurgan L. 2014. Prediction of intrinsic disorder in proteins using MFDp2. Methods Mol. Biol. Clifton NJ 1137:147–162.

Mohr D, Frey S, Fischer T, Güttler T, Görlich D. 2009. Characterisation of the passive permeability barrier of nuclear pore complexes. EMBO J. 28:2541–2553.

Mok J, Kim PM, Lam HYK, et al. 2010. Deciphering protein kinase specificity through large-scale analysis of yeast phosphorylation site motifs. Sci. Signal. 3:ra12.

Moll T, Tebb G, Surana U, Robitsch H, Nasmyth K. 1991. The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SWI5. Cell 66:743–758.

Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. 2014. Assessment of protein disorder region predictions in CASP10. Proteins 82 Suppl 2:127–137.

Monsellier E, Chiti F. 2007. Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep. 8:737–742.

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB. 2003. Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evol. Biol. 3:19.

Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 5:R98.

Moses AM, Hériché J-K, Durbin R. 2007. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. Genome Biol. 8:R23.

Moses AM, Landry CR. 2010. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? Trends Genet. TIG 26:462–467.

Moses AM, Liku ME, Li JJ, Durbin R. 2007. Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. Proc. Natl. Acad. Sci. U. S. A. 104:17713–17718.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li X-Y, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput. Biol. 2:e130.

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.

Murray AW. 2012. Don't make me mad, Bub! Dev. Cell 22:1123–1125.

Nash P, Tang X, Orlicky S, Chen Q, Gertler FB, Mendenhall MD, Sicheri F, Pawson T, Tyers M. 2001. Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication. Nature 414:514–521.

Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB. 2005. Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol. 3:e405.

Neduva V, Russell RB. 2005. Linear motifs: evolutionary interaction switches. FEBS Lett. 579:3342–3345.

Neduva V, Russell RB. 2006. DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res. 34:W350–355.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. New York, USA: Oxford University Press

Nguyen Ba AN, Moses AM. 2010. Evolution of characterized phosphorylation sites in budding yeast. Mol. Biol. Evol. 27:2027–2037.

Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. 2009. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. BMC Bioinformatics 10:202.

Nguyen Ba AN, Yeh BJ, van Dyk D, Davidson AR, Andrews BJ, Weiss EL, Moses AM. 2012. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci. Signal. 5:rs1.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Obenauer JC, Cantley LC, Yaffe MB. 2003. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res. 31:3635–3641.

Ohno S. 1970. Evolution by Gene Duplication. London, UK: Allen and Unwin

Ostedgaard LS, Baldursson O, Vermeer DW, Welsh MJ, Robertson AD. 2000. A functional R domain from cystic fibrosis transmembrane conductance regulator is predominantly unstructured in solution. Proc. Natl. Acad. Sci. U. S. A. 97:5657–5662.

Pancsa R, Tompa P. 2012. Structural disorder in eukaryotes. PloS One 7:e34687.

Pawson T, Gish GD, Nash P. 2001. SH2 domains, interaction modules and cellular wiring. Trends Cell Biol. 11:504–511.

Perfetto SP, Chattopadhyay PK, Roederer M. 2004. Seventeen-colour flow cytometry: unravelling the immune system. Nat. Rev. Immunol. 4:648–655.

Pic-Taylor A, Darieva Z, Morgan BA, Sharrocks AD. 2004. Regulation of cell cycle-specific gene expression through cyclin-dependent kinase-mediated phosphorylation of the forkhead transcription factor Fkh2p. Mol. Cell. Biol. 24:10036–10046.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinforma. Oxf. Engl. 21:676–679.

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinforma. Oxf. Engl. 21:3435–3438.

Ptacek J, Devgan G, Michaud G, et al. 2005. Global analysis of protein phosphorylation in yeast. Nature 438:679–684.

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

Ravid T, Hochstrasser M. 2008. Diversity of degradation signals in the ubiquitin-proteasome system. Nat. Rev. Mol. Cell Biol. 9:679–690.

Redon R, Ishikawa S, Fitch KR, et al. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Reményi A, Good MC, Lim WA. 2006. Docking interactions in protein kinase and phosphatase networks. Curr. Opin. Struct. Biol. 16:676–685.

Ren S, Uversky VN, Chen Z, Dunker AK, Obradovic Z. 2008. Short Linear Motifs recognized by SH2, SH3 and Ser/Thr Kinase domains are conserved in disordered protein regions. BMC Genomics 9 Suppl 2:S26.

Rigoutsos I, Floratos A. 1998. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinforma. Oxf. Engl. 14:55–67.

Rivas E, Eddy SR. 2008. Probabilistic phylogenetic inference with insertions and deletions. PLoS Comput. Biol. 4:e1000172.

Robinson MD, Grigull J, Mohammad N, Hughes TR. 2002. FunSpec: a web-based cluster interpreter for yeast. BMC Bioinformatics 3:35.

Rodal AA, Manning AL, Goode BL, Drubin DG. 2003. Negative regulation of yeast WASp by two SH3 domain-containing proteins. Curr. Biol. CB 13:1000–1008.

Ron D, Walter P. 2007. Signal integration in the endoplasmic reticulum unfolded protein response. Nat. Rev. Mol. Cell Biol. 8:519–529.

Rosso L, Marques AC, Weier M, Lambert N, Lambot M-A, Vanderhaeghen P, Kaessmann H. 2008. Birth and rapid subcellular adaptation of a hominoid-specific CDC14 protein. PLoS Biol. 6:e140.

Roth AF, Wan J, Bailey AO, Sun B, Kuchar JA, Green WN, Phinney BS, Yates JR, Davis NG. 2006. Global Analysis of Protein Palmitoylation in Yeast. Cell 125:1003–1013.

Rubin GM, Yandell MD, Wortman JR, et al. 2000. Comparative genomics of the eukaryotes. Science 287:2204–2215.

Sadowski I, Breitkreutz B-J, Stark C, et al. 2013. The PhosphoGRID Saccharomyces cerevisiae protein phosphorylation site database: version 2.0 update. Database J. Biol. Databases Curation 2013:bat026.

Sbia M, Parnell EJ, Yu Y, Olsen AE, Kretschmann KL, Voth WP, Stillman DJ. 2008. Regulation of the yeast Ace2 transcription factor during the cell cycle. J. Biol. Chem. 283:11135–11145.

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

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

Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 12:R120.

Schneider TD, Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18:6097–6100.

Schöbel F, Jacobsen ID, Brock M. 2010. Evaluation of lysine biosynthesis as an antifungal drug target: biochemical characterization of Aspergillus fumigatus homocitrate synthase and virulence studies. Eukaryot. Cell 9:878–893.

Schröder M, Kaufman RJ. 2005. The mammalian unfolded protein response. Annu. Rev. Biochem. 74:739–789.

Schrödinger L. 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

Schwartz D, Gygi SP. 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23:1391–1398.

Serber Z, Ferrell JE Jr. 2007. Tuning bulk electrostatics to regulate protein function. Cell 128:441–444.

SGD Project. 2011. Saccharomyces Genome Database. Available from: http://www.yeastgenome.org/

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13:2498–2504.

Sharifpoor S, Nguyen Ba AN, Youn J-Y, et al. 2011. A quantitative literature-curated gold standard for kinase-substrate pairs. Genome Biol. 12:R39.

Siepel A, Bejerano G, Pedersen JS, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15:1034–1050.

Sigalov AB, Zhuravleva AV, Orekhov VY. 2007. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. Biochimie 89:419–421.

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput. Appl. Biosci. CABIOS 12:327–345.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197.

Sopko R, Huang D, Preston N, et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. Mol. Cell 21:319–330.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9:3273–3297.

Stefflova K, Thybert D, Wilson MD, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell 154:530–540.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U. S. A. 100:9440–9445.

Storici F, Lewis LK, Resnick MA. 2001. In vivo site-directed mutagenesis using oligonucleotides. Nat. Biotechnol. 19:773–776.

Straight AF, Murray AW. 1997. The spindle assembly checkpoint in budding yeast. Methods Enzymol. 283:425–440.

Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. Genome Res. 16:182–189.

Suijkerbuijk SJE, van Dam TJP, Karagöz GE, et al. 2012. The vertebrate mitotic checkpoint protein BUBR1 is an unusual pseudokinase. Dev. Cell 22:1321–1329.

Sun MGF, Sikora M, Costanzo M, Boone C, Kim PM. 2012. Network evolution: rewiring and signatures of conservation in signaling. PLoS Comput. Biol. 8:e1002411.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinforma. Oxf. Engl. 23:1282–1288.

Swapna LS, Srinivasan N, Robertson DL, Lovell SC. 2012. The origins of the evolutionary signal used to predict protein-protein interactions. BMC Evol. Biol. 12:238.

Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. 1988. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. J. Mol. Biol. 203:439–455.

Di Talia S, Wang H, Skotheim JM, Rosebrock AP, Futcher B, Cross FR. 2009. Daughter-specific transcription factors regulate cell size control in budding yeast. PLoS Biol. 7:e1000221.

Tan CSH, Bodenmiller B, Pasculescu A, et al. 2009. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Sci. Signal. 2:ra39.

Teige M, Scheikl E, Reiser V, Ruis H, Ammerer G. 2001. Rck2, a member of the calmodulin-protein kinase family, links protein synthesis to high osmolarity MAP kinase signaling in budding yeast. Proc. Natl. Acad. Sci. U. S. A. 98:5625–5630.

Teng X, Dayhoff-Brannigan M, Cheng W-C, et al. 2013. Genome-wide consequences of deleting any single gene. Mol. Cell 52:485–494.

Teyra J, Sidhu SS, Kim PM. 2012. Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS Lett. 586:2631–2637.

Tian W, Li B, Warrington R, Tomchick DR, Yu H, Luo X. 2012. Structural analysis of human Cdc20 supports multisite degron recognition by APC/C. Proc. Natl. Acad. Sci. U. S. A. 109:18419–18424.

Tompa P. 2012. Intrinsically disordered proteins: a 10-year recap. Trends Biochem. Sci. 37:509–516.

Tong AH, Evangelista M, Parsons AB, et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science 294:2364–2368.

Tong AHY, Boone C. 2006. Synthetic Genetic Array Analysis in Saccharomyces cerevisiae. In: Xiao W, editor. Yeast Protocol. Methods in Molecular Biology. Humana Press. p. 171–191. Available from: http://link.springer.com.myaccess.library.utoronto.ca/protocol/10.1385/1-59259-958-3%3A171

Tonikian R, Xin X, Toret CP, et al. 2009. Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. PLoS Biol. 7:e1000218.

Toyn JH, Gunyuzlu PL, White WH, Thompson LA, Hollis GF. 2000. A counterselection for the tryptophan pathway in yeast: 5-fluoroanthranilic acid resistance. Yeast Chichester Engl. 16:553–560.

Tsien RY. 1998. The green fluorescent protein. Annu. Rev. Biochem. 67:509–544.

Turk BE. 2008. Understanding and exploiting substrate recognition by protein kinases. Curr. Opin. Chem. Biol. 12:4–10.

Ubersax JA, Ferrell JE Jr. 2007. Mechanisms of specificity in protein phosphorylation. Nat. Rev. Mol. Cell Biol. 8:530–541.

Ubersax JA, Woodbury EL, Quang PN, Paraz M, Blethrow JD, Shah K, Shokat KM, Morgan DO. 2003. Targets of the cyclin-dependent kinase Cdk1. Nature 425:859–864.

Uetz P, Giot L, Cagney G, et al. 2000. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403:623–627.

Ureta-Vidal A, Ettwiller L, Birney E. 2003. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4:251–262.

Uversky VN, Gillespie JR, Fink AL. 2000. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins 41:415–427.

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu. Rev. Biophys. 37:215–246.

Uversky VN. 2013. A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci. Publ. Protein Soc. 22:693–724.

Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 138:198–208.

Via A, Gould CM, Gemünd C, Gibson TJ, Helmer-Citterich M. 2009. A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinformatics 10:351.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. Nat. Rev. Genet. 15:221–233.

Vleugel M, Hoogendoorn E, Snel B, Kops GJPL. 2012. Evolution and function of the mitotic checkpoint. Dev. Cell 23:239–250.

Volkman BF, Lipson D, Wemmer DE, Kern D. 2001. Two-state allosteric behavior in a single-domain signaling protein. Science 291:2429–2433.

Voordeckers K, Brown CA, Vanneste K, van der Zande E, Voet A, Maere S, Verstrepen KJ. 2012. Reconstruction of ancestral metabolic enzymes reveals molecular mechanisms underlying evolutionary innovation through gene duplication. PLoS Biol. 10:e1001446.

Wach A, Brachat A, Pöhlmann R, Philippsen P. 1994. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. Yeast Chichester Engl. 10:1793–1808.

Waksman G, Kominos D, Robertson SC, et al. 1992. Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. Nature 358:646–653.

Wang S, Hazelrigg T. 1994. Implications for bcd mRNA localization from spatial distribution of exu protein in Drosophila oogenesis. Nature 369:400–403.

Wang X, Goshe MB, Soderblom EJ, et al. 2005. Identification and functional analysis of in vivo phosphorylation sites of the Arabidopsis BRASSINOSTEROID-INSENSITIVE1 receptor kinase. Plant Cell 17:1685–1703.

Wanke V, Pedruzzi I, Cameroni E, Dubouloz F, De Virgilio C. 2005. Regulation of G0 entry by the Pho80-Pho85 cyclin-CDK complex. EMBO J. 24:4271–4278.

Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449:54–61.

Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. Bioinforma. Oxf. Engl. 20:2138–2139.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337:635–645.

Warren DT, Andrews PD, Gourlay CW, Ayscough KR. 2002. Sla1p couples the yeast endocytic machinery to proteins regulating actin dynamics. J. Cell Sci. 115:1703–1715.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinforma. Oxf. Engl. 25:1189–1191.

Weissman DB, Desai MM, Fisher DS, Feldman MW. 2009. The rate at which asexual populations cross fitness valleys. Theor. Popul. Biol. 75:286–300.

Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Fersht AR. 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc. Natl. Acad. Sci. U. S. A. 105:5762–5767.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18:691–699.

Wilks SS. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Ann. Math. Stat. 9:60–62.

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

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17:149 – 163.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. 20:1377–1419.

Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293:321–331.

Xu D, Farmer A, Collett G, Grishin NV, Chook YM. 2012. Sequence and structural analyses of nuclear export signals in the NESdb database. Mol. Biol. Cell 23:3677–3693.

Xu H, Andi B, Qian J, West AH, Cook PF. 2006. The alpha-aminoadipate pathway for lysine biosynthesis in fungi. Cell Biochem. Biophys. 46:43–64.

Xu L, Massagué J. 2004. Nucleocytoplasmic shuttling of signal transducers. Nat. Rev. Mol. Cell Biol. 5:209–219.

Yachie N, Saito R, Sugahara J, Tomita M, Ishihama Y. 2009. In silico analysis of phosphoproteome data suggests a rich-get-richer process of phosphosite accumulation over evolution. Mol. Cell. Proteomics MCP 8:1061–1071.

Yaffe MB, Leparc GG, Lai J, Obata T, Volinia S, Cantley LC. 2001. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat. Biotechnol. 19:348–353.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Ydenberg CA, Rose MD. 2009. Antagonistic regulation of Fus2p nuclear localization by pheromone signaling and the cell cycle. J. Cell Biol. 184:409–422.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

Zack TI, Schumacher SE, Carter SL, et al. 2013. Pan-cancer patterns of somatic copy number alteration. Nat. Genet. 45:1134–1140.

Van Zeebroeck G, Rubio-Texeira M, Schothorst J, Thevelein JM. 2014. Specific analogs uncouple transport, signaling, oligo-ubiquitination and endocytosis in the yeast Gap1 amino acid transceptor. Mol. Microbiol.

Zhu H, Bilgin M, Bangham R, et al. 2001. Global analysis of protein activities using proteome chips. Science 293:2101–2105.