Sequence-function relationships in intrinsically disordered regions through the lens of evolution

by

Taraneh Zarin

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Department of Cell and Systems Biology
University of Toronto

Abstract

Intrinsically disordered regions (IDRs) are regions of proteins that do not autonomously fold into stable secondary or tertiary structures. Though they defy the classical view of proteins as rigidly structured macromolecules, IDRs are widespread in living organisms, and are involved in a diverse array of functions. The majority of IDRs appear to be rapidly evolving at the level of the primary amino acid sequence, which makes it difficult to quantify evolutionary conservation and associate these regions with biological function using standard sequence analysis. The aim of my thesis research has thus been to understand evolutionary constraint and sequence-function relationships in IDRs. Using a functionally characterized IDR in the yeast protein Ste50, I first found that highly diverged amino acid sequences can encode conserved phenotypes in IDRs, showing that sequence divergence does not necessarily imply functional divergence in these regions. Using a phylogenetic comparative framework, I found that the net charge of the Ste50 IDR, rather than the precise amino acids, is a functional molecular feature that is preserved over evolution. I next expanded my evolutionary analysis of IDRs to the yeast proteome, and found that most highly diverged IDRs contain many molecular features that are preserved over evolution. I summarized the evolution of these molecular features with an "evolutionary signature" for each IDR, and found that groups of IDRs with similar evolutionary signatures are enriched for specific biological functions. I also found that IDRs with similar evolutionary signatures can rescue function *in vivo* despite negligible sequence similarity. Finally, I used these evolutionary signatures to train a statistical model, and found that they can be used to classify IDRs for a diverse set of biological functions. I identified the molecular features contributing to these functional predictions, and attributed distinct functions to specific IDRs in proteins with multiple IDRs. Overall, this work shows that there is rich functional information in IDR sequences, and that this information can be revealed through evolutionary analysis.

Acknowledgments

I would first like to acknowledge my advisor, Dr. Alan Moses, who has provided me with an incredible amount of support, mentorship, and countless opportunities during my time as a graduate student. Alan's curiosity, passion for science, commitment to pedagogy, and generosity have continually inspired me. It has been a pleasure to learn from him and work with him on topics that have truly fascinated me over the last several years.

I'm grateful for the guidance and support I have received from those who have served on my supervisory committee: Dr. Alan Davidson, Dr. Sergio Peisajovich, and Dr. Brenda Andrews. Thank you also to Dr. Simon Alberti, with whom I did a visiting fellowship at the lively and inspiring MPI-CBG. Thank you to collaborators and mentors Dr. Julie Forman-Kay and Dr. Christian Landry, and to Dr. Mojca Mattiazzi Usaj, Henry Hong, Thanh Nguyen, Yunchen Gong, and Dr. Helena Friesen for technical support with microscopes, servers, plate readers, and flow cytometers. Thank you to my pre-graduate school research mentors: Dr. Lili-Naz Hazrati, Dr. Cindi Morshead, Dr. Nadia Sachewsky, and Dr. Naomi Visanji.

I'd like to acknowledge past and present members of the Moses lab, who have enriched my graduate school experience as labmates, mentors, mentees, collaborators, and friends: Dr. Alex Nguyen Ba, Gavin Douglas, Bob Strome, Dr. Louis-Francois Handfield, Mitchell Li Cheong Man, Dr. Gelila Tilahun, Ian Hsu, Alex Lu, Dr. Muluye Liku, Nirvana Nursimulu, Caressa Tsai, Dr. Purnima Kompella, Dr. Iva Pritisanac, Shadi Zabad, Tahmid Mehdi, Amy Lu, Xiao Wang, Carolina Weishaar, and Selma Osman. Thanks also to staff, faculty, and students in the Cell and Systems Biology Department for creating a collegial and welcoming atmosphere during research days, retreats, orientations, and informal gatherings.

I am beyond grateful to my family and friends for their (empirically) endless love and support. Thank you to my parents, Hossein and Sharareh, for encouraging me to pursue my interests, my brother Payam for being my role model, and to my uncle Ali, for piquing our interest in scientific research. Thank you to my grandmother Afsar, who embodies persistence, and my late grandparents Hamid, Rouhi, and Ebrahim, who taught me to observe the world around me. Finally, thank you to Alicia, Zoë, Kate, Yasi, Greg, Susie, Anna, Eric, Sam, and Minh for the support, love, adventures, insights, and distractions.

Table of Contents

List of Tables

List of Figures

List of Appendices

1 General Introduction

1.1 Abstract

Intrinsically disordered proteins and protein regions are increasingly appreciated as widespread and important for biological function. Understanding their varied functions has been difficult, particularly due to their dynamic structures, sequences, and evolution compared to ordered, or structured, regions of proteins. Here, I review how intrinsically disordered regions have been defined, the process through which they were discovered, and their unique characteristics. I discuss the application of quantitative trait evolution to molecular features of disordered regions, and how this approach may address some of the challenges associated with understanding the sequence-function relationships in these regions. I then describe the facilitation of functional genomics research using budding yeasts as a model organism. I conclude this chapter by outlining my research objectives and providing an overview of the main findings from this thesis.

1.2 Intrinsically disordered regions

1.2.1 Discovery and definition

The “lock and key” model of enzyme function, posited by Emil Fischer in 1894 (Kunz, 2002), has had a tremendous impact on our understanding of protein function and its relation to protein structure, despite the fact that enzymes were not formally recognized as proteins until the latter part of the 1920s (DeForte and Uversky, 2016; Sumner, 1926). The “lock and key” concept was solidified when the first three-dimensional protein structure was solved by X-ray crystallography a few decades later (Kendrew et al., 1958), perpetuating the view that a stably-folded 3-dimensional protein structure is necessary for biological function (Redfern et al., 2008; Wright and Dyson, 1999). Since that discovery, thousands of protein structures have been solved through X-ray crystallography, and the so-called “structure-function paradigm” has taken hold as the central dogma of structural biology. However, in recent years, the notion that proteins can lack stable secondary or tertiary structures, and that this property could be integral to their functions, has become much more widely appreciated (Forman-Kay and Mittag, 2013). These so-called “intrinsically disordered” proteins and protein regions have been the subject of intensifying research interest since the late 1990s.

Intrinsically disordered regions have been “discovered” many times, mainly since they have always been considered exceptions to the rule, and because no consistent terminology existed for them until a few years ago (Dunker et al., 2013). For example, the prominent disordered protein tau was initially called “natively denatured” (Schweers et al., 1994), and many proteins and protein regions have been denoted as “flexible”, “mobile”, “natively unfolded” (DeForte and Uversky, 2016) and “negative noodles” (Sigler, 1988) in the literature, without any reference to “intrinsic disorder”. This meandering in name space reflects both the treatment of disordered regions as “anomalies” in the face of the structure-function paradigm, and the fact that they occupy a continuum of unstructured states (DeForte and Uversky, 2016).

1.2.2 Sequences, structures, and functions – commonalities and heterogeneity

As more and more intrinsically disordered regions were reported on in the literature, it became clear that they have a biased amino acid composition compared to ordered regions. Hints of this first came about in key studies where rule-based and neural network predictors were trained on amino acid sequences of disordered regions, spurred by the observation that a disordered region in Calcineurin (Kissinger et al., 1995) and several other proteins are devoid of aromatic residues (Romero et al., 1997). These studies were then expanded to the Swiss Protein Database, identifying thousands of likely disordered protein regions, and achieving a prediction accuracy close to that of secondary structure predictors (Romero et al., 1998). Another important study that characterized the sequence bias in intrinsically disordered regions compared several sequence features of 91 protein regions that were experimentally shown to exhibit hallmarks of disorder (e.g. showing characteristics of a random coil, lacking secondary or tertiary structures, or having hydrodynamic radii similar to expanded polypeptide chains) to a set of 275 globular, ordered proteins (Uversky et al., 2000). This study found that the disordered regions could not be distinguished from ordered regions based on their lengths, isoelectric points, and net charge alone, but that they were clearly separated from ordered regions based on a combination of high net charge and low hydrophobicity (Uversky et al., 2000). Further work established that disordered regions are enriched for polar and charged amino acids compared to ordered regions (Romero et al., 2001). Since these early studies, it has become increasingly clear that disordered regions can be predicted based on their amino acid sequences, with over 70 disordered region predictors developed over the last 20 years (Li et al., 2015).

Although intrinsically disordered region sequences are clearly demarcated from ordered regions in a binary classification framework, they also seem to exhibit remarkable sequence heterogeneity within the "disordered" category. For example, many disordered regions are of "low complexity" (Cumberworth et al., 2013; Halfmann, 2016; Mier et al., 2019; Romero et al., 2001; Uversky et al., 2000; Wang et al., 2018; Wootton, 1994), meaning that they are comprised of a reduced alphabet of amino acid residues. However, within this subset of disordered regions, there appear to be different types of low complexity. One example of this is disordered regions that are rich in glutamine and asparagine repeats, and are associated with so-called "prionogenic" proteins (Alberti et al., 2009; Halfmann et al., 2011). Other examples include arginine and glycine repeats, which are found in RNA binding proteins (P. A. Chong et al., 2018). There are also glutamine-rich, proline-rich, or aspartic and glutamic acid-rich disordered regions, which have been associated with transcriptional activation domains (Boija et al., 2018; Dennig et al., 2018; Frietze and Farnham, 2011; Gerber et al., 1994; Klemsz and Maki, 1996). In addition to the diversity of low complexity sequences, there are numerous other sequence features that distinguish different intrinsically disordered regions, including the presence or abundance of post-translational modification sites (Holt et al., 2009; Iakoucheva et al., 2004), as well as different charge properties such as a highly negative (Warren and Shechter, 2017) or positive charge (Garg and Gould, 2016), or a separation of positive and negatively charged blocks of residues (Nott et al., 2015).

Similarly, although intrinsically disordered regions generally do not fold into stable, 3-dimensional structures under physiological conditions, they can be heterogeneous in their lack of structure. Researchers have long recognized that intrinsically disordered regions occupy a "continuum" of unstructured states (Burger et al., 2016; DeForte and Uversky, 2016; Forman-Kay and Mittag, 2013; Uversky, 2002). An example of a protein on the most disordered end of this continuum is Sic1, which remains disordered as it engages in low affinity (but specific) interactions with its partner, Cdc4 (Mittag et al., 2008). In the middle of this continuum, many proteins exhibit folding upon binding (Vacic et al., 2007; Wang et al., 2013), or, in the case of the negative regulator of translation initiation 4E-BP2, folding upon post-translational modification (Bah and Forman-Kay, 2016). A protein region on the more "ordered" end of the continuum is the nuclear co-activator binding domain (NCBD) of the CREB-binding protein (CBP), which acts as a transcriptional co-activator (Kwok et al., 1994). Unlike many disordered

regions that rapidly interconvert between different conformations (e.g. Sic1), NCBD samples different conformations on a relatively long (millisecond) timescale in solution (Kjaergaard et al., 2010). Interestingly, the natural assumption that disordered regions with many weak interacting partners remain disordered, while disordered regions with specific interactions fold upon binding, was recently challenged. Remarkably, the disordered proteins histone H1 and the nuclear chaperone prothymosin-α engage in an ultra-high affinity interaction while remaining completely disordered, possibly because of their extreme opposite charges (Borgia et al., 2018).

Given the range of sequence and structure characteristics exhibited by intrinsically disordered regions, it is perhaps not surprising that they are involved in many different functions in living cells. The variety of functions that intrinsically disordered regions are involved in has been said to rival those of ordered regions, ranging from well-established roles in signaling networks (Holt et al., 2009; Martin-Yken et al., 2016; Nguyen Ba et al., 2012; Tompa et al., 2014), to structural and signaling components of transmembrane proteins (Busch et al., 2015; Kjaergaard and Kragelund, 2017; Meyer et al., 2018; Tusnády et al., 2015) and the nuclear pore (Alber et al., 2007), to their varied roles in transcriptional (Boehning et al., 2018; S. Chong et al., 2018; Minezaki et al., 2006) and translational (Franzmann et al., 2018; Protter et al., 2017) regulation. One classical protein function that has not been attributed to intrinsically disordered regions is catalytic activity, though it is increasingly appreciated that enzymes often have intrinsically disordered regions that play crucial roles in mediating catalysis and substrate binding (Kim et al., 2018; Reed et al., 2015; Szabo et al., 2019).

1.2.3 Evolution of IDRs – evidence for negative and positive selection

One of the key evolutionary studies on intrinsically disordered regions applied the disorder prediction algorithm, ‘DISOPRED2’, to several archaean, bacterial, and eukaryotic genomes (Ward et al., 2004). This study established that predicted intrinsically disordered regions with a minimum length of 30 amino acids are present across these three kingdoms of life, and that long intrinsically disordered regions are particularly widespread in eukaryotic proteomes, with an estimated 33% disorder compared to 4% in bacteria. The results from this study have since been confirmed with other prediction methods (Peng et al., 2013; Xue et al., 2012), and have been accompanied by suggestions that the expansion of intrinsic disorder from prokaryotic to eukaryotic proteomes could be reflective of the increased regulatory capacity that is needed to

contend with multi-compartment cells. Interestingly, there may also be a correlation with disorder and variability or extremity in habitats, as unicellular eukaryotes and viruses display a wide variation in their disorder content (Xue et al., 2012). For example, some virus proteomes are 7% disordered, whereas others are 77% disordered. Intriguingly, a recent study presented evidence in favour of this hypothesis, finding that disordered proteins in so-called "extremophile" tardigrades are specifically expressed and necessary for desiccation tolerance (Boothby et al., 2017).

The presence of intrinsically disordered regions across living organisms raises many questions about whether and how they persist or expand in genomes through evolutionary time. Many of these questions have begun to be addressed through population genomics and comparative genomics studies. Based on pairwise genetic distances between homologs, one of the first comparative genomics studies on disordered regions found that experimentally verified disordered regions in 19/26 protein families are more rapidly evolving than ordered regions (Brown et al., 2002). This study corroborated some of the previous anecdotal evidence about intrinsically disordered regions being highly diverged compared to ordered regions in individual proteins. Further studies found similar results, with intrinsically disordered regions showing a much higher tolerance for insertions and deletions (InDels) than ordered regions (Khan et al., 2015; Light et al., 2013; Tóth-Petróczy and Tawfik, 2013) and overall relaxed purifying, or negative selection (Khan et al., 2015). On the molecular level, this implies that there is less selective pressure to keep nonsynonymous mutations out of IDRs. An important caveat with studies that measure distances between genes using standard models is that intrinsically disordered regions are likely not evolving under the same substitution models as ordered regions. A study that addressed this point directly used substitution matrices specific for either ordered or disordered regions, and found that intrinsically disordered regions are still more likely to tolerate evolutionary changes compared to ordered regions when these differences are accounted for (Brown et al., 2010). Overall, evolutionary analysis has so far revealed that the presence of intrinsically disordered regions is conserved across orthologs, but that there can be considerable divergence in the specific amino acids that make up these orthologous disordered regions (Bellay et al., 2011; Chen et al., 2006a, 2006b; Colak et al., 2013). However, despite the rapid change in amino acid residues in intrinsically disordered regions, there is evidence of conservation for amino acid composition (Moesa et al., 2012) and length (Schlessinger et al., 2011).

Along with the evidence for negative selection on intrinsically disordered regions, there is also evidence for positive selection in these regions, complementing the view that they may be sources of evolutionary novelties or could have high "evolvability". On the molecular level, positive selection implies that there is a selective advantage to acquiring nonsynonymous mutations. An early study combined population sequence data from *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* strains, and found a much higher ratio of positively selected codons in intrinsically disordered regions as compared to structured regions in these species (Nilsson et al., 2011). Recently, another study used a powerful comparative approach to analyze mammalian sequences, and found that despite the presence of negative selection, human intrinsically disordered regions have a much higher rate of adaptive substitutions than their ordered counterparts (Afanasyeva et al., 2018).

1.2.4 Stabilizing selection on molecular features of intrinsically disordered regions

An interesting hypothesis about intrinsically disordered regions of proteins is that they resemble non-coding regions of DNA in their rapid evolution and regulatory functions, and therefore may be bound by similar constraints (Landry et al., 2014; Moses and Landry, 2010). Specifically, this idea has been applied to phosphorylation sites in intrinsically disordered regions (Landry et al., 2014; Moses and Landry, 2010), as they display a high rate of turnover and can appear in clusters, where their density, but not their specific positions, can be conserved (Beltrao et al., 2009; Holt et al., 2009; A. C. W. Lai et al., 2012; Moses et al., 2007b). Clusters of phosphorylation sites that undergo turnover through evolution are reminiscent of clusters of transcription factor binding sites that display turnover, and have been shown to be evolving under stabilizing selection (Ludwig et al., 2000). Within this model, specific phosphorylation sites can be under weak functional constraint, but contribute to an aggregate property, or phenotype, that is under selection. Thus, they could be gained or lost between lineages (Figure 1-1), leaving signatures of rapid evolution. The function that they contribute to, however, could be conserved, as there would be a loss of fitness associated with the loss or diminishment of such a function, perhaps beyond a certain threshold.

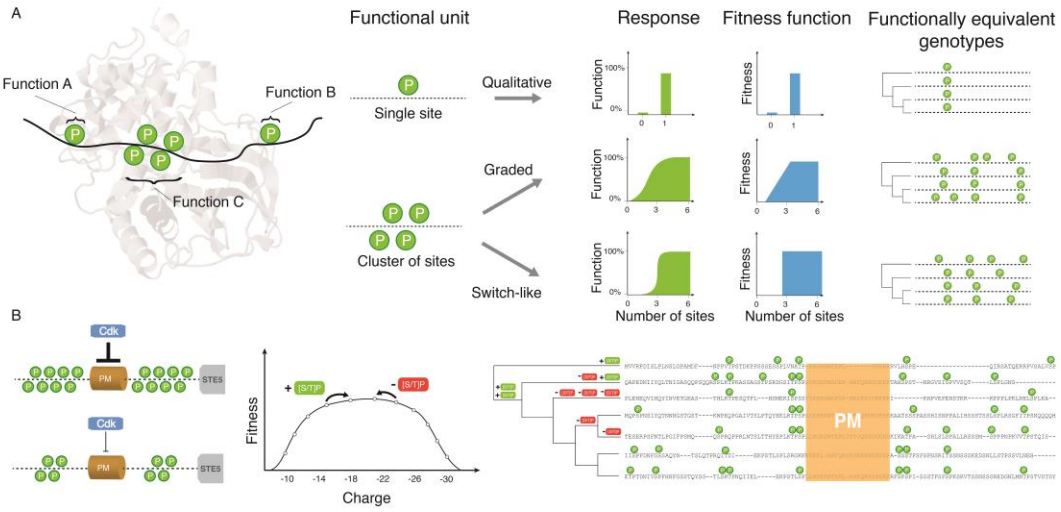


Figure 1-1. The relationship between site phosphorylation, localization and protein functions determines how much conservation is expected among species under purifying or stabilizing selection. (A) Toy examples of phosphorylation sites (indicated as "P"s) and cluster of sites and how they may affect protein functions individually or collectively. Phosphorylation sites regulate three putative functions A, B, C. The aggregate function of phosphorylation sites affects the fitness function of the protein and thus determines how many possible equivalent genotypes may give rise to equivalent functions or fitness. Only few possible examples are shown to illustrate the complex relationships expected and their impact on the evolution of phosphorylation profiles and many more are possible. (B) Shows a possible fitness landscape for CDK inhibition of Ste5. Ste5 inhibition is proportional to the charge (twice the number of phosphorylated residues) in the disordered region surrounding the PM domain (Strickfaden et al. 2007). Evolutionary changes that create CDK consensus sites ([ST]-P) will increase the strength of the inhibition, while changes that destroy consensus sites will reduce the strength of inhibition. The stabilizing selection model suggests that as long as the total strength of inhibition is within an acceptable range, the exact number and location of phosphorylation sites will drift nearly neutrally. A sequence alignment of the disordered regions surrounding the PM domain of Ste5 from *S. cerevisiae* and related yeasts is shown on the right. During evolution consensus sites are gained and lost (+ [ST]-P or − [ST]-P) on the phylogenetic tree, leading to a large diversity in number and location of phosphorylation sites in this region. Figure from (Landry et al., 2014).

Besides clusters of phosphorylation sites, intrinsically disordered regions contain aggregate molecular features within their amino acid sequences such as net charge (Mao et al., 2010; Strickfaden et al., 2007), length (Schlessinger et al., 2011), or physicochemical properties (Ravarani et al., 2018; Warren and Shechter, 2017) that appear to be important for their functions. Thus, it is interesting to posit that these aggregate properties could contribute to quantitative traits under stabilizing selection, where a phenotypic "optimum" could shape the variation of their aggregate molecular features, as has been found for gene expression (Bedford and Hartl, 2009; Charlesworth, 2013; Ludwig et al., 2000). Indeed, there are now several examples in intrinsically disordered regions where there is experimental evidence for stabilizing selection acting on features such as net charge rather than specific amino acids (Daughdrill et al., 2007; Lemas et al., 2016; Zarin et al., 2017).

1.3 Functional genomics using yeast as a model organism (with modified text from (Zarin and Moses, 2014))

1.3.1 Comparative genomics of yeasts

Enabled by comparative genomics, yeasts have increasingly developed into a powerful model system for molecular evolution. The beginning of 'comparative genomics' was a turning point for molecular evolution. Two types of genome sequences were most often compared: those of 'closely' related species whose entire genomes could be aligned at the level of the nucleic acids (Waterston et al., 2002), and genomes of more 'distantly' related species that showed interesting variation in lifestyle and physiology, but were close enough that most genes had clear orthologues (Aparicio et al., 2002). Because budding yeast was the first eukaryote to have its genome completely sequenced (Goffeau et al., 1996), it was naturally at the forefront of the comparative genomics work (Cliften et al., 2003; Dujon et al., 2004; Kellis et al., 2003). The hemiascomycetous yeast species whose genomes are now available span an evolutionary distance similar to that of the chordates (Dujon, 2006), making these genomes a model system for animal evolution.

Comparative genomics vastly expanded the scope of molecular evolution. The availability of evolutionary measurements for thousands of genes could be used to evaluate evolutionary hypotheses in general, using statistical analysis (Wolfe and Li, 2003) as opposed to analysing single genes anecdotally, as had often been done before. Comparative genomics also enabled

functional genomics in multiple species, leading to a further expansion of the questions that could be tackled: evolution of genome-wide expression patterns, protein interactions and post-translational modifications. Once available, the comparative sequence and functional data could be applied in several areas of interest in molecular evolution. For example, one of the fortuitous discoveries made soon after the completion of the yeast genome was the identification of a whole-genome duplication event (Wolfe and Shields, 1997). The comparative data for this set of gene duplicates ('ohnologues') continues to allow unprecedented large-scale studies of gene duplication and divergence, facilitating studies of classical topics in molecular evolution (Kimura and Ohta, 1974).

Despite the excitement about regulatory evolution, what was known about the evolution of non-coding DNA prior to comparative genomics was largely anecdotal (reviewed in (Wray, 2007)). Analysis of 'closely' related yeast genomes showed that functional non-coding DNA was conserved, but that over longer evolutionary distances there was little conservation of regulatory sequences at the DNA level. Gene expression and other genomic data for multiple species have supported the idea that gene regulation has changed considerably among yeast species (reviewed in (Weirauch and Hughes, 2010); (Wohlbach et al., 2009)).

Recently, proteomics experiments have begun to characterize the evolution of signalling networks, also referred to as regulatory evolution at the post-translational level (reviewed in (Beltrao et al., 2013)). Like transcriptional regulatory sequences, post-translational regulatory sites are apparently largely conserved between 'closely' related species of yeast (Holt et al., 2009; Nguyen Ba et al., 2012; Nguyen Ba and Moses, 2010). At further evolutionary distances, some modifications and interactions show high levels of divergence (Beltrao et al., 2009; Holt et al., 2009; Sun et al., 2012), while other protein–protein interactions evolve much more slowly (Qian and Zhang, 2009). Further research will be needed to determine the major patterns of protein-regulatory evolution, but it is clear that regulatory evolution at levels other than transcription is an emerging area (Moses and Landry, 2010), with yeast a leading model system.

1.3.2 Testing hypotheses about regulatory evolution in laboratory experiments with *Saccharomyces cerevisiae*

It has become increasingly possible to reconstruct evolutionary history at the molecular level and to infer the corresponding changes in cellular function and physiology. This so-called 'functional

synthesis' (Dean and Thornton, 2007) holds great appeal to evolutionary biologists, who have been historically limited to correlative experiments and statistical inferences. Yeast is an ideal organism for mechanistic evolutionary experiments, for several reasons. First, the largely tree-like evolution of yeast genes (Rokas et al., 2003) and bioinformatic resources (such as the Yeast Gene Order Browser (Byrne and Wolfe, 2005)) allow molecular history to be reconstructed accurately. For example, reconstructed ancestral maltases from a large gene family show evidence for multiple mechanisms of diversification, including natural selection on key residues that control substrate specificity (Voordeckers et al., 2012).

Perhaps more importantly, yeast evolution is experimentally accessible. For example, cryogenic preservation of intermediate genotypes creates a living record of the dynamic evolutionary process (Buckling et al., 2009). In addition, the short generation time of yeast enables techniques for systematic, quantitative measurements of fitness (Breslow et al., 2008) and genotype frequencies (Gresham et al., 2008; Parts et al., 2011; Taylor and Raes, 2004), allowing systematic investigation of evolutionary properties that have been discussed extensively in abstract, but have been hard to measure.

Thus, it is now possible in yeast to: (a) infer specific molecular changes during evolution; (b) test functional impacts on protein function and cellular traits; and (c) measure whether those changes lead to fitness advantages (at least in the environments that are possible to simulate in the laboratory). This implies the prospect of discovering (after centuries of speculation) how evolution actually happened (Dean and Thornton, 2007). One of the first and most compelling studies to use direct fitness measurements of re-engineered evolutionary changes was a study of the gene duplication of GAL1 and GAL3 within the classical GAL regulatory network (Hittinger and Carroll, 2007), which showed direct evidence for fitness increases after gene duplication, consistent with 'escape from adaptive conflict' (Hughes, 2005). Experimental fitness measurements were also used to provide direct evidence that gene expression differences in an endocytosis complex (implicated through statistical analysis) in the transition to pathogenicity conferred a growth advantage at high temperatures (Fraser HB et al., 2012). Most recently, ancestral reconstructions were used to identify specific amino acid changes in the paralogous transcription factors Mcm1 and Arg80 that led to subdivision of the ancestral gene function, but also (fascinatingly) to avoid interfering with each other through spurious vestigial interactions (Baker et al., 2013).

The mechanistic perspective of evolution that is now possible in yeast is still only beginning to take hold. However, the early studies in this area have already demonstrated that it will be possible to directly address fundamental questions about regulatory evolution using these approaches.

1.4 Research objectives and thesis overview

The majority of disordered regions of proteins have thus far remained uncharacterized, partly due to their fundamental differences with ordered regions of proteins, and a current sparsity of methods that take these differences into account. The goal of my thesis has been to understand, through functional genomics experiments in budding yeast, as well as computational methods and evolutionary analysis, the sequence-to-function relationships in disordered regions. Specifically, the aims of this thesis are:

1. To understand whether highly diverged, orthologous disordered regions are functionally divergent, or if disordered regions with highly diverged primary amino acid sequences can perform the same functions in a model protein in budding yeast
2. To explore the extent to which molecular features of highly diverged disordered regions are preserved through evolution proteome-wide, and assess their association with protein function
3. To use the evolution of molecular features in disordered regions to predict function of specific proteins and disordered regions

In chapter 2, I find that despite the high divergence in primary amino acid sequences of orthologous disordered regions, they can perform similar functions *in vivo* and confer similar fitness. Through evolutionary analysis, I find that rather than preserving the precise amino acids, natural selection is preserving a molecular feature (specifically the net charge) of the disordered region in question (Zarin et al., 2017). In chapter 3, I apply an evolutionary analysis to the budding yeast proteome, and find that the preservation of molecular features in disordered regions is a general phenomenon. I find that many disordered regions share sets of molecular features that are under selection, and that these "signatures" of evolution are associated with specific biological functions (Zarin et al., submitted). In chapter 4, I use these evolutionary signatures of disordered regions in a machine-learning framework to predict functions and phenotypes of specific proteins and disordered regions. In chapter 5, I provide a general discussion about insights that have been made and questions that remain in relation to evolutionary analysis, functional annotation, and sequence-function relationships of intrinsically disordered regions.

Chapter 2
Selection maintains signaling function of a highly diverged intrinsically disordered region

This is an author-produced PDF of an article accepted for publication in Proceedings of the National Academy of Sciences of the United States of America following peer review. The version of record:

Selection maintains signaling function of a highly diverged intrinsically disordered region

Proc Natl Acad Sci U S A. 2017 Feb 21;114(8):E1450-E1459. doi: 10.1073/pnas.1614787114. Epub 2017 Feb 6.

Zarin, T.[1], Tsai, C.[2], Nguyen Ba, A.N.[3], Moses, A.M.[1,2]

is available online at: https://www.pnas.org/content/114/8/E1450.long

1. Department of Cell and Systems Biology, University of Toronto, 25 Harbord St., Toronto, ON, Canada, M5S 3G5
2. Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks St., Toronto, ON, Canada, M5S 3B2
3. FAS Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, MA, USA, 02138

2 Selection maintains signaling function of a highly diverged intrinsically disordered region

2.1 Abstract

Intrinsically disordered regions (IDRs) are characterized by their lack of stable secondary or tertiary structure, and comprise a large part of the eukaryotic proteome. Although these regions play a variety of signaling and regulatory roles, they appear to be rapidly evolving at the primary sequence level. In order to understand the functional implications of this rapid evolution, we focused on a highly diverged IDR in *Saccharomyces cerevisiae* that is involved in regulating multiple conserved MAP Kinase pathways. We hypothesized that under stabilizing selection, the functional output of orthologous IDRs could be maintained, such that diverse genotypes could lead to similar function and fitness. Consistent with the stabilizing selection hypothesis, we find that diverged, orthologous IDRs can mostly recapitulate wildtype function and fitness in *S. cerevisiae*. We also find that the electrostatic charge of the IDR is correlated with signaling output, and using phylogenetic comparative methods, find evidence for selection maintaining this quantitative molecular trait despite underlying genotypic divergence.

2.2 Significance statement

Intrinsically disordered regions (IDRs) are widespread, have diverse functions, and are involved in human disease. Because standard sequence analysis methods identify little sequence homology in IDRs, it’s not currently understood whether (or how) the functions of these protein regions are preserved over evolution. Here we show that orthologous IDRs can preserve regulatory functions despite near-complete sequence divergence. This suggests that natural selection maintains aggregate molecular properties in IDRs, which we propose to be quantitative traits. Consistent with this, we find signatures of stabilizing selection on the electrostatic properties of IDRs. Thus, in analogy to the rapid evolution of non-coding DNA in eukaryotic enhancers, divergence in primary amino acid sequence does not imply functional divergence in IDRs.

2.3 Introduction

Current predictions suggest that close to 40% of all proteins in eukaryotic organisms are either entirely disordered, or contain sizeable regions that are disordered, meaning they do not

autonomously fold into defined secondary or tertiary structures (Peng et al., 2013; Ward et al., 2004). These intrinsically disordered regions (IDRs) are thought to have important implications for protein function (Liu et al., 2009; Vavouri et al., 2009), and are known to play regulatory roles, often through short linear motifs (SLiMs) that control protein-protein interactions, localization, degradation, and post-translational modifications (Forman-Kay and Mittag, 2013; Tompa et al., 2014). While proteome-wide studies have provided in silico evidence for conservation of length (Schlessinger et al., 2011) and composition (Moesa et al., 2012) in some IDRs, reports of increased rates of insertions and deletions (de la Chaux et al., 2007; Khan et al., 2015; Light et al., 2013; Nido et al., 2012; Tóth-Petróczy and Tawfik, 2013) and amino acid substitutions (Brown et al., 2002) in IDRs are indicative of their rapid evolution compared to ordered regions. In addition, while some SLiMs are indeed conserved in IDRs (Beltrao and Serrano, 2005; Davey et al., 2012; Nguyen Ba et al., 2012), others appear in clusters where precise position and number are not conserved (Beltrao et al., 2012; Holt et al., 2009; Moses et al., 2007b). Although it is reasonable to assume that conservation of sequence in IDRs is indicative of functional conservation of SLiMs, it is more difficult to interpret the functional consequences of IDRs that are highly diverged at the sequence level: these may represent either non-functional sequences evolving in the absence of constraint, or weakly constrained functional elements that are gained or lost in a compensatory manner (undergoing evolutionary turnover [as described in (Ludwig et al., 2000; Moses et al., 2007b)]), such that they are not conserved at the amino acid sequence level.

Like IDRs, non-coding DNA often shows relatively rapid evolution and weak constraints at the sequence level (Bergman and Kreitman, 2001). Interestingly, IDRs show other parallels with non-coding DNA (Beltrao et al., 2013; Moses et al., 2007b; Moses and Landry, 2010). For example, non-conserved clusters of phosphorylation sites in IDRs are reminiscent of non-conserved transcription factor binding sites in enhancers. Although these enhancers and the binding sites within them are not conserved, they can lead to the same expression patterns (Ludwig et al., 1998). Preservation of expression patterns despite underlying sequence divergence in these regions is thought to result from stabilizing selection on quantitative phenotypes (Ludwig et al., 2000). Stabilizing selection could allow for quantitative phenotypes to be maintained within an optimal range while allowing tolerance of mutations or insertions and deletions, as these individually exert weak functional and selective effects (Charlesworth, 2013;

Ludwig et al., 2000). Although it is likely that some of these highly diverged IDRs, like non-coding regions, are either non-functional or sites of lineage-specific evolution (Wray, 2007), at least a portion of these IDRs may be performing quantitative functions that are under stabilizing selection (Landry et al., 2014).

In this study, we investigate whether the observed molecular divergence in IDRs implies functional divergence, or whether the diversity in these regions could accumulate while functional output is preserved under stabilizing selection. Under stabilizing selection, we expect that diverged, orthologous IDRs have similar functional outputs, and confer similar fitness. To test this, we take advantage of a model IDR that plays roles in multiple signaling pathways in *Saccharomyces cerevisiae*. We show that orthologous disordered regions can recapitulate wildtype morphology and quantitative regulatory function. This represents, to our knowledge, the first *in vivo* evidence that disordered signaling protein regions that are highly divergent at the primary sequence level can perform similar functions and confer similar fitness. We also find that the basal net charge of the IDRs is correlated with the signaling output, and, by applying phylogenetic comparative methods to the basal net charge in these IDRs, find evidence for selection on this quantitative molecular trait.

2.4 Results

2.4.1 An intrinsically disordered region in the adaptor protein Ste50 that is involved in multiple signaling pathways is highly diverged at the primary amino acid sequence level

We chose to focus our study on an IDR in the adaptor protein Ste50 that is involved in several highly studied MAP Kinase (MAPK) signaling pathways in *S. cerevisiae* (Figure 2-1a). We chose this IDR in part because it is situated between two highly conserved protein domains: the Sterile Alpha Motif (SAM) and the Ras-association (RA) domain (Jansen et al., 2001; Tatebayashi et al., 2006; Truckses et al., 2006) (Figure 2-1b). We can therefore confidently identify the orthologous protein sequence in other hemiascomycete species, even though the primary amino acid sequence has diverged rapidly (Figure 2-1b). We find that the Ste50 IDR shows only 27.76% (s.d.=11.94) average pairwise percent identity (Figure 2-1c), which is similar to scrambled IDR sequences (see methods), which show 24.40% pairwise percent identity (mean of 100 simulations), and the 24.26% (s.d.=12.27) pairwise percent identity that we get from

aligning randomly chosen non-homologous disordered regions of the same length as the Ste50 IDR (see methods). The divergence of the Ste50 IDR also appears to saturate with divergence time (Appendix Figure 1-1). This rapid divergence is not due to overall divergence of the Ste50 protein, as the adjacent structured domains show strong conservation at the primary amino acid level (SAM: 43.02% [s.d.=9.92] and RA domain: 83.61% [s.d.=5.28] pairwise percent identity).

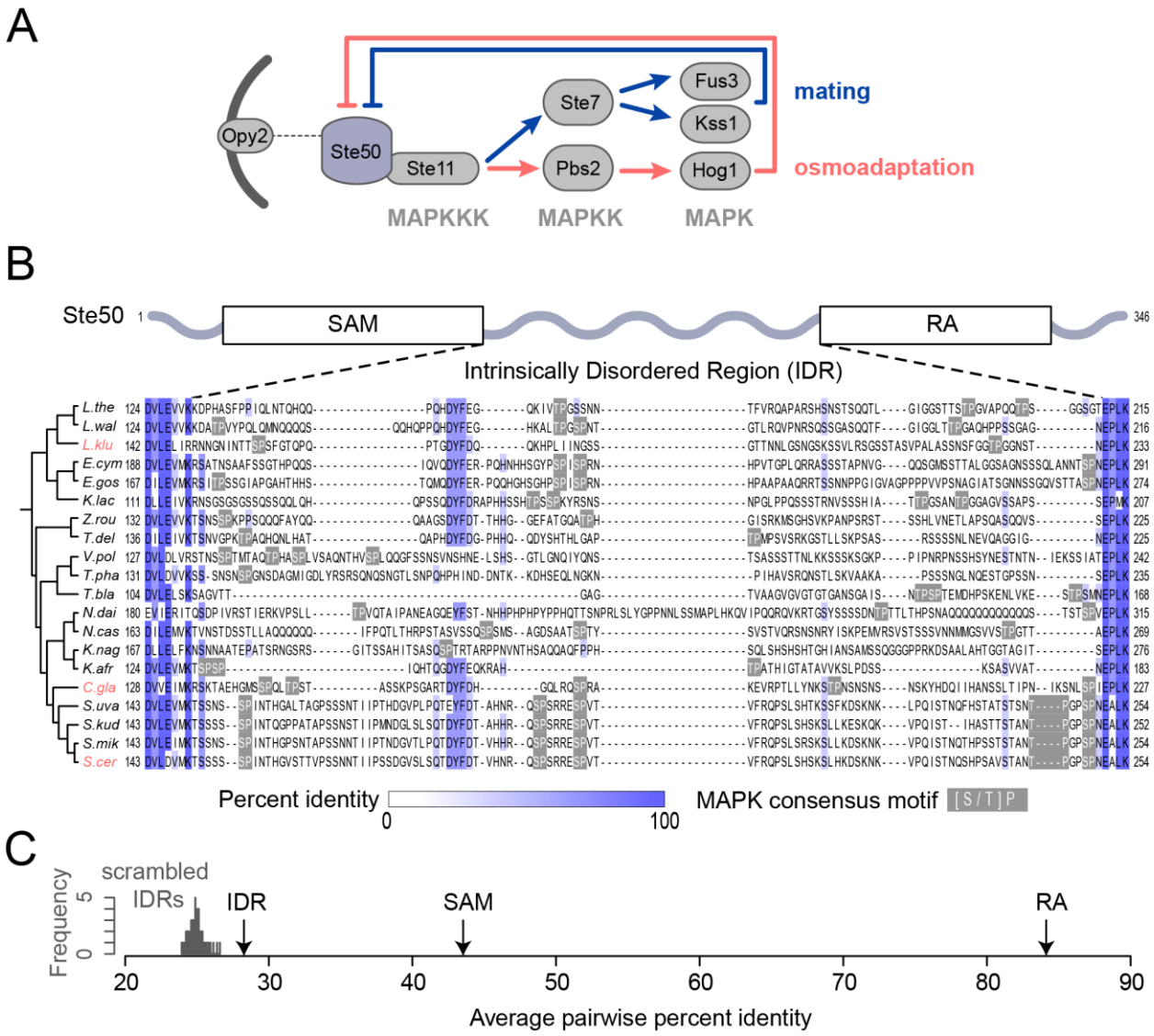


Figure 2-1. a) The adaptor protein Ste50 is phosphorylated by multiple MAPKs, which results in dissociation of the adaptor and associated proteins from membrane-bound Opy2, and subsequent negative regulation of downstream MAPKs. Not all pathway components are shown in the schematic. b) Alignment of the Ste50 IDR for hemiascomycetes (displayed using Jalview (Waterhouse et al., 2009)). Percentage identity is shown in blue, MAPK phosphorylation

consensus motifs ([S/T]P) are boxed in grey. Species names of IDRs that were used for downstream functional and fitness experiments are highlighted in red. C) Average pairwise percent identity of the real Ste50 IDR alignment (IDR), compared to a distribution of IDRs with randomly scrambled residues (scrambled IDRs), the Ste50 Sterile Alpha Motif (SAM), and the Ste50 Ras Association (RA) domain. Y-axis shows the frequency of scrambled IDRs.

The Ste50 IDR also represents a good candidate for evolutionary analysis because it contains a cluster of MAPK consensus phosphorylation sites (S or T, followed by a P) which contribute to signal modulation of MAPK pathways (English et al., 2015; Hao et al., 2008; Yamamoto et al., 2010). Evolutionary turnover within clusters of phosphorylation sites in disordered regions is thought to be widespread (Beltrao et al., 2009; Freschi et al., 2011; Holt et al., 2009; Landry et al., 2014; Nguyen Ba and Moses, 2010), and the alignment of Ste50 shows that MAPK consensus sites differ in position, spacing, and number, consistent with evolutionary turnover of these sites within the IDR (Figure 2-1b).

2.4.2 Diverged orthologous IDRs recapitulate multiple signaling functions in *S. cerevisiae*

Phosphorylation of the MAPK consensus sites in the Ste50 IDR results in attenuation of signaling by dissociation of the signaling complex from the membrane (Yamamoto et al., 2010). Phospho-proteomic studies also indicate phosphorylation of a subset of these sites in standard growth conditions (Albuquerque et al., 2008; Bodenmiller et al., 2010; Gnad et al., 2009; Holt et al., 2009; Kanshin et al., 2015; Smolka et al., 2007; Soulard et al., 2010), which we refer to as basal phosphorylation. To test the function of this region in *S. cerevisiae*, we therefore made an unphosphorylatable mutant, where each consensus site was mutated to alanine (referred to as 5A mutant) (see methods). Previous studies have shown that this mutant is defective in Hog1 signaling dynamics and displays increased basal expression of FUS1, presumably because of overactive effector kinases Fus3 and Kss1 (Hao et al., 2008; Yamamoto et al., 2010). In order to determine whether or not diverged sequences are divergent in function, we swapped orthologous IDR sequences from two yeast species (*C. glabrata* and *L. kluyveri*) into *S. cerevisiae* (see methods) and quantified the function of these chimaeric Ste50s compared to the wildtype and 5A mutant (Figure 2-2a).

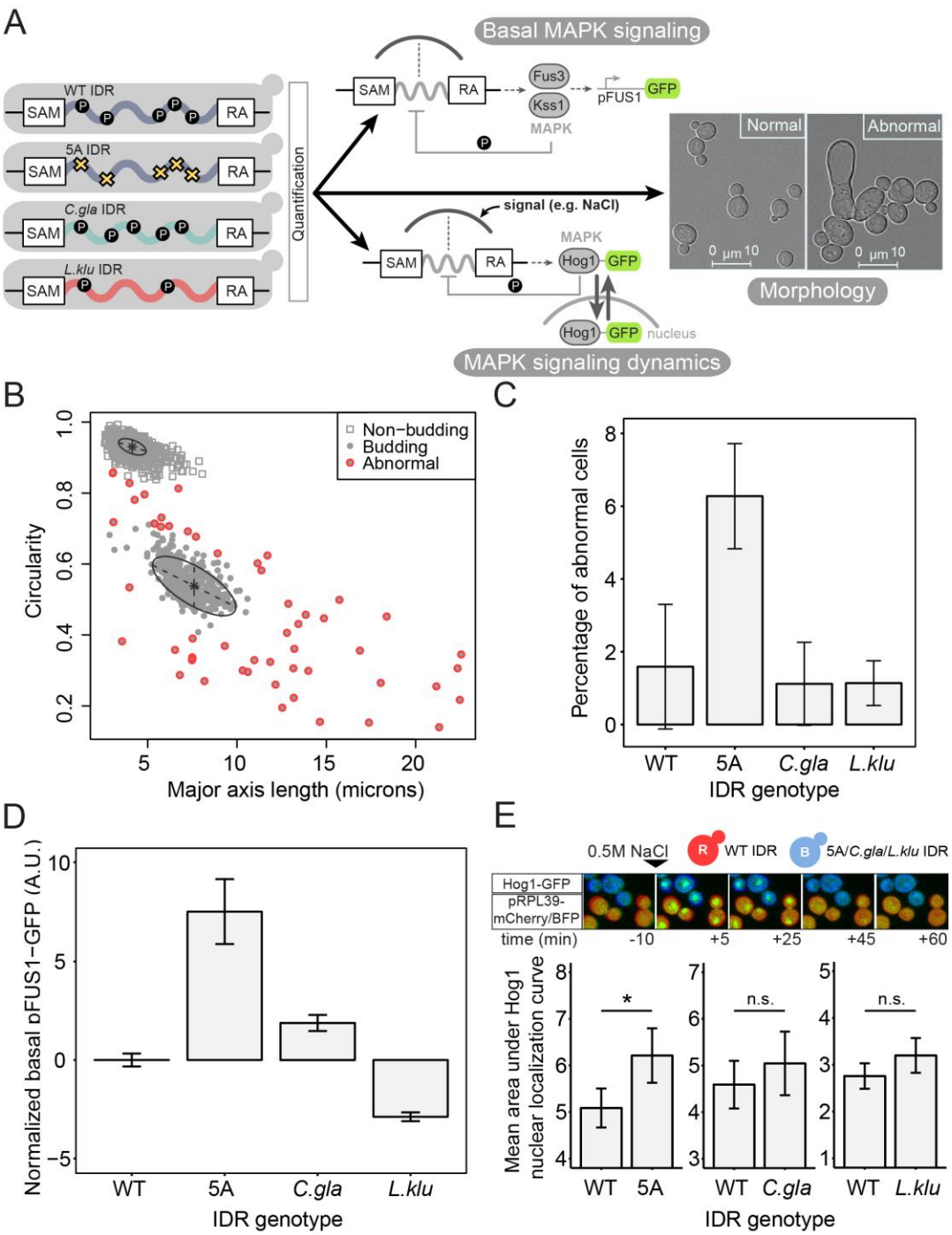


Figure 2-2. Diverged orthologous IDRs recapitulate *S. cer* IDR functions compared to the 5A mutant a) Diverged IDRs were swapped with the *S. cer* IDR and 3 different functional outputs were quantified: morphology, basal MAPK (Fus3) signaling, and MAPK (Hog1) signaling

dynamics b) Cell morphology clusters along two dimensions. Each point represents one cell for which major axis length and circularity features were extracted. Figure shows example plot from one biological replicate, where cells have been classified as non-budding, budding, and abnormal (see methods for details). c) Average percentage of cells with abnormal morphology for each IDR genotype. Error bars represent 1.96 s.e. between 3 biological replicates (average of ~400 cells per replicate). d) Diverged IDRs mostly recapitulate wildtype basal pFUS1-GFP levels. Error bars represent 1.96 s.e. between 6-12 biological replicates (50 000 cells per replicate) for each strain. e) Top: representative images of time-lapse movies capturing Hog1-GFP localization in co-cultured wildtype and experimental strains (constitutively expressing mCherry and mTagBFP2, respectively). Bottom: Diverged IDRs recapitulate wildtype Hog1 signaling dynamics. Error bars represent 1.96 s.e. Asterisk represents statistical significance ($P<0.01$, Student's t-test, N=15-35 cells).

Interestingly, we noticed that the 5A mutant displays abnormal morphology in a small subset of cells (Figure 2-2a zoomed-in micrograph, Appendix Figure 1-2 wide field of view), which, to our knowledge, was previously unreported. We therefore first tested whether the chimaeric Ste50 proteins could rescue these abnormal morphologies. We quantified morphology using the length of the major axis (a measure to capture the elongated shape of the abnormal cells), as well as circularity (a measure to capture the irregular, non-circular shape of the abnormal cells) (see methods for details). Along these dimensions, the vast majority of cells fall into two clear clusters based on their shape: non-budding cells, which are highly circular and have a small major axis length, and budding cells, which are less circular and have a higher major axis length (Figure 2-2b). We defined the cells that fell outside of these clusters as "abnormal" cells, and quantified the fraction of abnormal cells for each genotype (Figure 2-2c). We found 6.3% (s.d.=1.3) abnormal cells in the 5A mutant population, compared to less than 2% (s.d.$\leq$1.5) abnormal cells for the wildtype strain and the orthologous, diverged IDRs (Figure 2-2c). We therefore conclude that the diverged IDRs quantitatively recapitulate wildtype morphology.

We then sought to quantify the basal activity of the Fus3 and Kss1 MAPKs, as the IDR is known to be involved in negative regulation of these kinases (Hao et al., 2008; Yamamoto et al., 2010). We quantified basal MAPK signaling by using a genomically-integrated GFP reporter driven by

the FUS1 promoter (pFUS1), a transcriptional target of Ste12, the effector of Fus3 and Kss1 signaling (Elion et al., 1993; Hagen et al., 1991). As expected, we found that the 5A mutant had significantly higher levels of basal pFUS1-GFP expression compared to the wildtype in flow cytometry analysis (Figure 2-2d) (see methods). This is consistent with the IDR being important for negative regulation of basal Fus3 and Kss1 signaling, as suggested by previous studies (Hao et al., 2008; Yamamoto et al., 2010). In contrast, we found that the diverged, orthologous IDRs mostly recapitulated wildtype basal pFUS1 expression.

The Ste50 IDR has also been shown to modulate the dynamics of Hog1 activity following activation by osmotic stress. Previous studies have shown that Hog1 is active for a longer amount of time when the 5 phosphorylation sites in the Ste50 IDR are mutated to alanine – this is thought to happen because of relaxed negative feedback on the HOG (High Osmolarity Glycerol) pathway (Hao et al., 2008; Yamamoto et al., 2010). Based on previous work (Ferrigno et al., 1998), we used Hog1-GFP nuclear localization as a proxy for Hog1 activity. To eliminate experimental day-to-day variation in the length of Hog1 activity following stimulation, we devised an assay through which we could directly compare Hog1 signaling for different IDR genotypes in an identical environment (Figure 2-2e, top). To do so, we constitutively expressed different fluorescent proteins in wildtype and experimental (i.e. 5A or orthologous IDR) strains to differentially label IDR genotypes in each experiment. We were thus able to co-culture strains, and, following addition of stimulus, captured Hog1 nuclear localization for single cells with different IDRs in the same field of view through time-lapse imaging (see methods for details). As expected, Hog1 in the 5A mutant displayed a significantly slower return to baseline activity compared to the wildtype, as evidenced by a longer duration and magnitude of Hog1 nuclear localization (Figure 2-2e). However, the diverged orthologous IDRs recapitulated the wildtype signaling dynamics, showing no significant deviation from wildtype in the duration and magnitude of Hog1 localization.

2.4.3 Diverged orthologous IDRs rescue fitness in *S. cerevisiae*

Having established that the diverged IDRs from other species could perform the known signaling functions of the *S. cerevisiae* IDR, we tested whether they were able to support wild-type growth and reproduction. We therefore quantified the fitness of the genotypes carrying diverged orthologous IDRs. For this we used a quantitative competitive growth assay, where we directly

competed the wildtype strain against all experimental strains (Figure 2-3a; see methods for details). We did this by labeling the wildtype with one fluorescent protein (ymCherry or mTagBFP2) and the experimental strains with a different fluorescent protein (ymCherry or mTagBFP2) and measuring growth of serially diluted, co-cultured cells over time. We found that while the 5A mutant displayed a significant fitness defect compared to the wildtype strain (mean selection coefficient of -0.038, s.e.= 0.005), the diverged IDRs displayed a much lower fitness defect compared to the wildtype strain (mean selection coefficient of -0.014, s.e.=0.004 for *C.glabrata* and -0.012, s.e.=0.002 for *L.kluyveri*) (Figure 2-3b). This is consistent with these IDRs recapitulating not only the function of the *S. cerevisiae* IDR *in vivo*, but also recapitulating most of the fitness of the wildtype IDR (see Discussion).

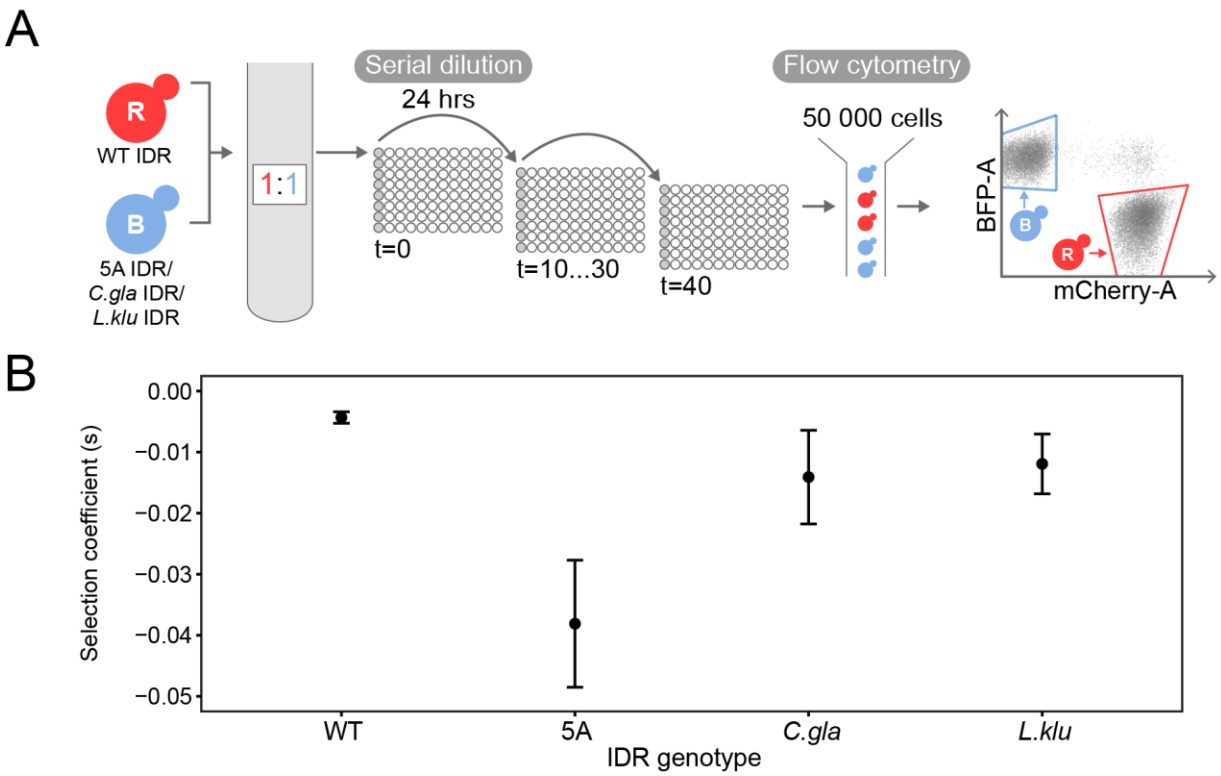


Figure 2-3. Diverged orthologous IDRs rescue fitness of wildtype *S. cer* IDR compared to 5A mutant. a) High-throughput quantitative competition assay captures growth rate of co-cultured cells over time b) Relative selection coefficients of 5A mutant and orthologous IDRs versus wildtype. Error bars represent 1.96 s.e. N=2 for wildtype, N=4 for 5A, *C.gla* and *L.klu* IDRs.

2.4.4 Basal net charge of diverged sequences is correlated with functional output

Despite the sequence divergence of this IDR in orthologous yeast proteins, the IDRs we tested were able to mostly recapitulate function and fitness in *S. cerevisiae*. This led us to ask if there are certain features in the sequence that are contributing to function, and are therefore likely to be under selection. Although we know that the 5 MAPK consensus phosphorylation sites are important for function in *S. cerevisiae* (Hao et al., 2008; Yamamoto et al., 2010), the *L.klu* IDR only has 2 consensus sites (Figure 2-1b), and has a similar functional output (basal FUS1 signaling and morphology) to *S. cerevisiae* (Figure 2-2). Further, the *C.gla* and *L.klu* IDRs conferred almost identical fitness in the *S. cerevisiae* background despite the former having 5 consensus phosphorylation sites and the latter having 2 (Figure 2-3b). Taken together, these results suggest that the number of MAPK consensus phosphorylation sites alone does not explain the functional output of the Ste50 IDR. However, the multiply-phosphorylated Ste50 IDR's interactions with membrane-bound Opy2 (Yamamoto et al., 2010) are reminiscent of the Ste5 disordered signaling region in *S. cerevisiae,* whose multiple MAPK consensus phosphorylation sites are thought to electrostatically modulate its interactions with the membrane (Strickfaden et al., 2007). Net charge is also thought to be a general functional property of intrinsically disordered regions (Forman-Kay and Mittag, 2013; Uversky, 2013), and has been shown to modulate conformational and binding properties of other intrinsically disordered proteins (Mao et al., 2010; Mittag et al., 2010; Müller-Späth et al., 2010). We therefore speculated that the salient sequence feature influencing the functional output of the Ste50 IDR could be its net charge.

Since our simplest quantitative measure of functional output is the basal expression of pFUS1, we correlated this with the basal net charge for each of the IDRs that we tested in our previous experiments (Figure 2-4; blue points). We calculated the basal net charge of each IDR by considering its net charge (sum of positive and negatively charged residues) including basal phosphorylation at up to two SP sites, as mass-spectrometry studies have found that up to two serines are phosphorylated in this IDR under basal conditions in *S. cerevisiae* (as reported in (Albuquerque et al., 2008; Bodenmiller et al., 2010; Gnad et al., 2009; Holt et al., 2009; Kanshin et al., 2015; Smolka et al., 2007; Soulard et al., 2010)). Therefore, if the IDR has two or more SP sites, we assume that two of these serines are phosphorylated under basal conditions, and add a

charge of -4 (-2 for each phosphorylation site) to the net charge of the IDR (see methods for details). To test the hypothesis that two SP sites are phosphorylated under basal conditions and contribute to net charge, we constructed an *S.cerevisiae* IDR where three out of five [S/T]P MAPK consensus phosphorylation sites were mutated to alanine, but two of the [S/T]P MAPK consensus phosphorylation sites were mutated to double glutamic acids (EE), as phospho-charge mimics (see Appendix Figure 1-3a). By our calculation, this IDR has the same basal net charge as the basally phosphorylated wildtype *S.cerevisiae* IDR. We find that this IDR (which we refer to as "WT-charge") has wildtype-like pFUS1 expression levels (Appendix Figure 1-3b) and wildtype-like morphology (Appendix Figure 1-3c), supporting our assertion that basally phosphorylated sites in the Ste50 IDR contribute to net charge, which is associated with wildtype function.

In order to further test the correlation between basal net charge and functional output in the form of pFUS1 expression, we engineered a series of IDRs broadly falling into the following categories: point mutations in the *S. cerevisiae* IDR, more examples of orthologous IDRs, and chimaeric IDRs (Appendix Figure 1-4). We also tested 16 other sequence features that could potentially impact the functional output of these IDRs (Appendix Figure 1-5), but only found a strong and significant positive correlation ($R^2=0.69$, Bonferroni corrected $P=0.001$) between the basal net charge of these sequences and their functional output (Figure 2-4; black points). The positive relationship between charge and signalling is similar to previous evidence from Ste5 suggesting that an increase in negatively charged residues weakens the interaction of the disordered region with the membrane, thus decreasing signal. Taken together, these data suggest that the amino acid composition of these sequences can modulate the functional output of the IDRs via the basal net charge.

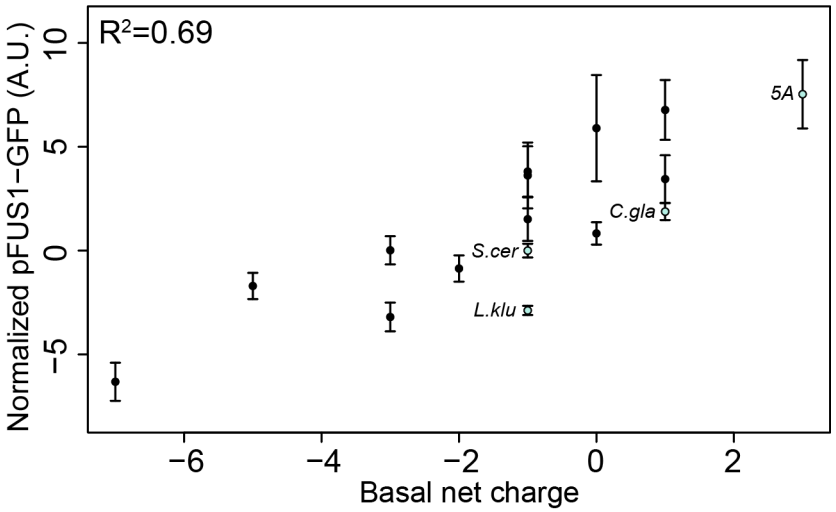


Figure 2-4. Basal net charge and MAPK reporter pFUS1-GFP expression are positively correlated. Each point represents a different IDR genotype (with blue corresponding to previously shown orthologous IDRs, the wildtype *S. cer* IDR, and the 5A IDR mutant, and black corresponding to engineered IDRs with varying phosphorylatable residues, charge, and length). Error bars represent 1.96 s.e.

2.4.5 Selection maintains functional output despite divergence at the primary sequence level

We next wanted to understand whether selection is preserving the function of these IDRs, despite the apparent divergence at the level of the primary amino acid sequence. Because the basal net charge of the IDRs is strongly correlated with their functional output (Figure 2-4), we considered this to be a quantitative trait that selection could act on. Stabilizing selection is expected to decrease trait variance by removing extreme phenotypes from the population (Bedford and Hartl, 2009; Hansen, 1997; Lande, 1976). We therefore used a phylogenetic comparative approach to test for reduced trait variance, which indicates selective constraint to preserve basal net charge across species (Figure 2-5a). To do so, we applied a Brownian motion (BM) model (Felsenstein, 1973) (see methods) to estimate the evolutionary variance of basal net charge and compared it to a null expectation of disordered region evolution without selection on basal net charge.

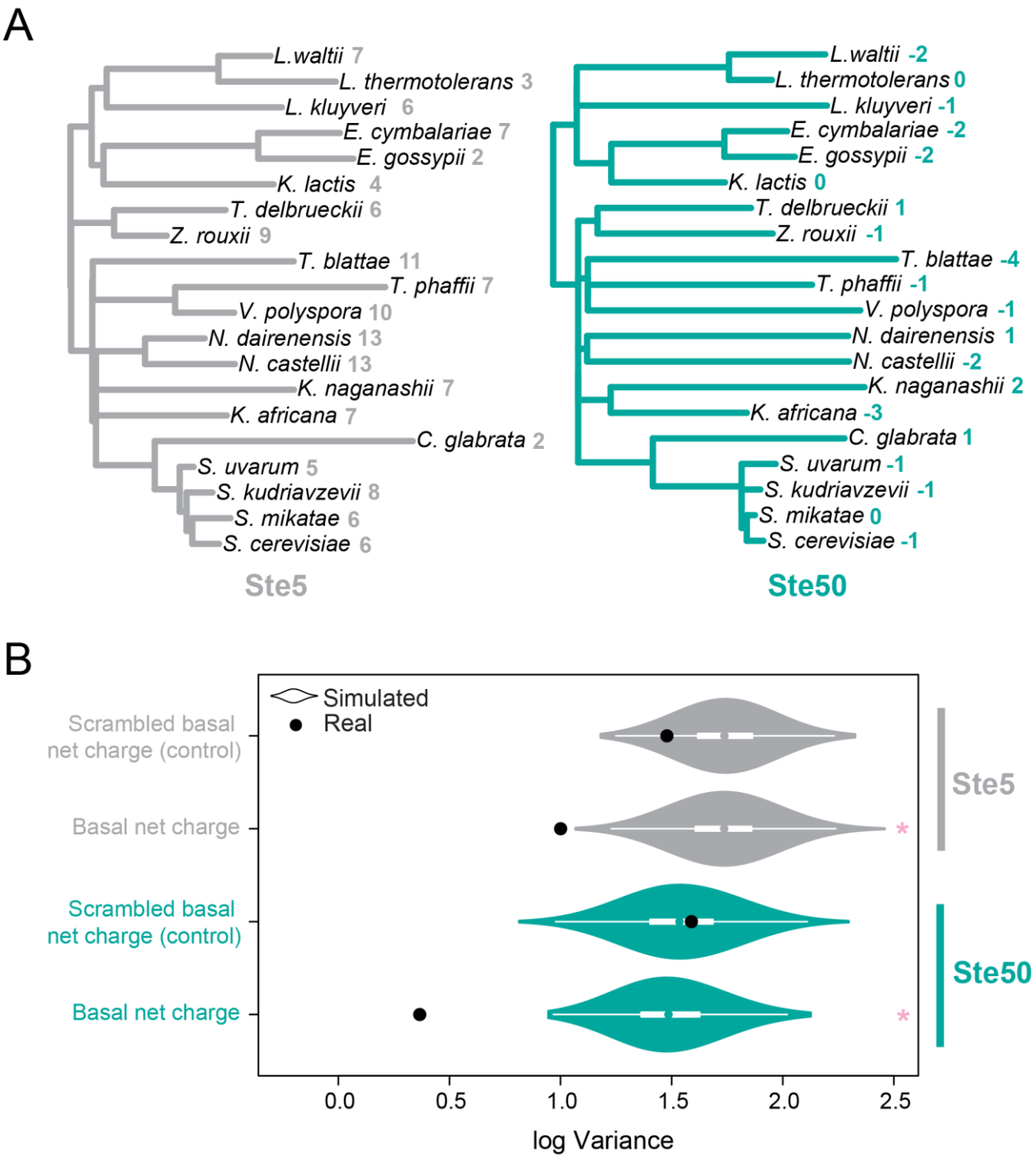


Figure 2-5. Stabilizing selection constrains the evolution of basal net charge in Ste50 a) Phylogenetic trees inferred from Ste5 (left) and Ste50 (right) IDRs with constrained resolved species topology. Quantitative trait value (basal net charge) for each species is indicated on tree tips b) Log evolutionary variance compared between real proteins (black dots) and 1000 simulated proteins (violin plots) for Ste5 (top) and Ste50. White boxes show interquartile range and median. Basal net charge was calculated as the sum of positively and negatively charged

residues accounting for basal phosphorylation of two ‘SP’ motifs, and with basal phosphorylation of two scrambled ‘PSX’ motifs (‘Scrambled basal net charge (control)’). Asterisks indicate statistical significance between the real and simulated proteins (<5% of the distribution – 50 proteins).

To obtain an expectation for the evolution of basal net charge in the absence of selection on basal net charge, we simulated molecular evolution of the Ste50 IDR. To do so, we used a simulator that includes disordered region-specific substitution patterns as well as position-specific local evolutionary rates, such that short linear motifs in disordered regions are preserved in the simulations through purifying selection (Nguyen Ba et al., 2014, 2012) (see methods). In using this simulation as a null expectation, we do include selection that can be inferred from the multiple sequence alignments, but do not include additional selection on basal net charge. Thus our null expectation includes selection, and deviations from it imply additional selection that is not apparent in sequence alignments. (See discussion)

We then compared the variance in the basal net charge inferred using the BM model on these simulated sequences to that inferred for the real Ste50 IDR alignment. We found that the variance of the real Ste50 IDR sequences was lower than all simulated sequences (Figure 2-5b, turquoise plots). Lower variance in basal net charge for Ste50 implies evolutionary constraint on basal net charge, consistent with stabilizing selection.

To confirm that the findings were not a result of unrealistic assumptions in our simulations, as a negative control, we also performed the analysis with positive and negative charges reassigned to four different residues (asparagine, glycine, threonine, alanine) than those known to be charged under physiological pH conditions (glutamic acid, aspartic acid, lysine, arginine), and assuming basal phosphorylation of two ‘scrambled’ phosphorylation sites (‘PSX’ motifs, where X is any amino acid other than proline). We found that the evolutionary variance in these negative controls (scrambled charged residues and phosphorylation motifs) was not different from the null expectation (Figure 2-5b).

We conducted the same analysis on Ste5, the previously mentioned signalling protein known to rely on net charge for functional output (Strickfaden et al., 2007). If selection is acting to

preserve basal net charge in the Ste5 disordered region, we also expect reduced evolutionary variance relative to simulations. We find similar results for Ste5 as for Ste50 (Figure 2-5b, grey plots), consistent with selection preserving its function, and suggesting that the phylogenetic comparative approach may be a general method to detect selection on basal net charge in diverged disordered regions.

2.5 Discussion

To date, experimental studies of protein evolution have focused on structured classes of proteins such as enzymes, where point mutations in the primary amino acid sequence are consistently coupled with functional divergence (Soskine and Tawfik, 2010). However, the functional consequences of evolutionary divergence in intrinsically disordered regions of proteins have remained largely unexplored, save for two *in vitro* studies (Daughdrill et al., 2007; Lemas et al., 2016).

In this study, we show that highly diverged, orthologous IDRs can perform similar signaling functions and confer similar fitness to a wildtype IDR in Saccharomyces cerevisiae. To do so, we took advantage of several quantitative signaling assays, including a dynamic fluorescence–microscopy experiment that allows comparison of different genotypes in co-culture. This allows for quantitative comparisons of signaling dynamics by controlling for imaging and culture conditions.

Using these quantitative assays, we found that the orthologous IDRs did not precisely recapitulate wildtype signaling and fitness in *S. cerevisiae*. Although we chose an IDR that is involved in conserved signaling pathways, there could be co-evolution of the orthologous IDRs with other proteins in their native signaling pathways or with the rest of the orthologous protein itself. Thus, inserting the orthologous IDRs in a *S. cerevisiae* context could be slightly detrimental to their function. This is an important caveat for experimental studies where protein regions are expressed outside their native context.

We found evidence that the electrostatic charge of the Ste50 IDR is correlated with signaling output of the mating pathway. Although previous studies had identified the phosphorylation sites in this region as being important for signaling (Hao et al., 2008; Yamamoto et al., 2010), we found no correlation between just the number of MAPK consensus sites in the IDR and the

functional output we tested. We speculate that the phosphorylation sites contribute to the net charge of the region, and allow the cell to modulate the charge of the region in response to signals. This is consistent with the model for the evolution of phosphorylation sites as a mechanism for modulation of charge (Pearlman et al., 2011). The importance of the basal net charge of the Ste50 IDR in signaling function is also consistent with recent studies suggesting that 'cryptic' electrostatic properties encoded in amino acid sequences of IDRs are important for their function (Das et al., 2016; Pak et al., 2016). We speculate that the charge of the Ste50 IDR affects interactions with the cell membrane as has been demonstrated for Ste5 (Strickfaden et al., 2007), but understanding the precise biophysical and biochemical properties of the Ste50 IDR that translate charge into signaling output is an important area for further study.

Lastly, by treating the charge as a quantitative trait (Lande, 1976), we were able to apply phylogenetic comparative methods (Beaulieu et al., 2012; Felsenstein, 1973; Hansen, 1997) to the disordered protein sequences and found evidence that these electrostatic properties are likely under stabilizing selection. Because disordered regions show little conservation at the sequence level, functional prediction methods based on amino acid sequence similarity have limited power. We believe that phylogenetic comparative methods represent a new approach to detect functional features within disordered regions. Selection on quantitative traits is often inferred using the Ornstein-Uhlenbeck (OU) model, a stochastic model that includes the tendency of a trait to evolve towards an adaptive optimum (Hansen, 1997). However, due to the limitations of the OU model (Cooper et al., 2016), we used the simpler approach of testing for reductions in trait variance to infer selection (Bedford and Hartl, 2009).

To test for selection, we compared real to simulated protein sequences. Evidence for selection may be represented by any disparity between these and real protein sequences. It is important to note that our simulations do include selection to preserve short conserved motifs (through position specific rates) as well as selection to retain the amino acid frequencies of disordered regions (through disordered region specific substitution models (Nguyen Ba et al., 2014, 2012)). Therefore, when we find evidence for selection, it is specifically evidence for conservation beyond what can be expected based on our models of disordered sequence evolution alone. Thus, we believe the reduction in variance observed in real proteins relative to simulated proteins is sufficient to conclude that the evolution of net charge within disordered regions is selectively

constrained. We propose that this is an example of a quantitative trait under stabilizing selection, in which the molecular phenotype of net charge is maintained within an optimal range.

Our results suggest the following picture of disordered region evolution: rapid evolution within IDRs introduces many mutations of individually small fitness effects, creating slight perturbations in net charge that fall within the nearly neutral range. These mutations contribute to a significant amount of protein sequence divergence. However, stabilizing selection will remove mutations that perturb quantitative traits such as net charge beyond an acceptable range, leading to reduced evolutionary variance. This reflects a form of mutation-selection balance, and offers an explanation for the existence of highly divergent genotypes within disordered regions, despite functional constraints. Although mutational input in intrinsically disordered regions is sufficient to generate abundant variation between species, our results are evidence of stabilizing selection constraining a molecular phenotype in spite of this variation.

2.6 Materials and methods

2.6.1 Ste50 alignment and quantification of divergence

The multiple sequence alignment for the Ste50 protein and its orthologs (MUSCLE (Edgar, 2004)) as well as their illustrated phylogenetic relationship (Figure 2-1b) were acquired from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005) and visualized using Jalview (Waterhouse et al., 2009). Boundaries for the Ste50 IDR (A.A. 151-251) were acquired using disorder predictions from DISOPRED3 (Jones and Cozzetto, 2015). IDR boundaries for the Ste50 orthologs were determined via the multiple sequence alignment, using the boundaries from the S. cer Ste50 IDR. All pairwise-percentage identities were calculated using Jalview (Waterhouse et al., 2009), which calculates the pairwise percent identity as the number of identical residues divided by the number of aligned residues for each pairwise re-alignment. For the set of randomly scrambled IDRs, the amino acids in the IDRs from Ste50 and each of its orthologs were randomly scrambled 100 times (leaving the remainder of each protein unscrambled), and the average percent identity of each pairwise alignment was calculated (distribution of these averages is plotted in Figure 2-1c). For the comparison of pairwise percent identity to random sequences, we calculated the pairwise percent identity of 19 (same number as

the YGOB orthologs) random IDRs in the yeast proteome that had the same length as the Ste50 IDR (YBL081W, YBR033W, YBR081C, YDR282C, YDR527W, YIL105C, YJL090C, YLL027W, YLR399C, YML045W, YMR266W, YMR277W, YNL047C, YNL288W, YOR153W, YOR316C, YPL053C, YPL270W, YPR115W).

We also calculated dN/dS ratios for the IDR, SAM domain, and RA domain (Appendix Figure 1-6). To do this, we first used PAL2NAL (Suyama et al., 2006) to obtain a codon alignment based on the protein alignment and DNA sequences of Ste50 for the Saccharomyces sensu stricto species available from YGOB (*Saccharomyces cerevisiae, mikatae, kudriavzevii,* and *uvarum*). We then used the PAML CODEML package (Yang, 2007) on the respective alignments and the corresponding species topology tree, and estimated the dN/dS ratio across the respective trees using the M0 model and the F1x4 codon frequency model.

2.6.2 Strain construction and growth conditions

All strains (

Appendix Table 1-1) were constructed in the S. cerevisiae BY4741 ssk22Δ0::kanMX4 ssk2Δ0 background (ssk2 and ssk22 were knocked out to disable the partially redundant SLN1 branch of the HOG pathway, as Ste50 is only active in the SHO1 branch (Maeda et al., 1995)). All genomic transformations were confirmed by Sanger sequencing. Mutagenized IDRs, chimaeric IDRs, and reporters were constructed using Gibson cloning (Gibson et al., 2009) and standard site-directed PCR mutagenesis. IDRs from orthologous proteins were amplified from purified genomic DNA of *C.glabrata, L.kluyveri, Z.rouxii, L.waltii, L.thermotolerans,* and *K.lactis* (see Appendix Figure 1-4 for IDR a.a. boundaries for each species). All transformations were performed using the standard lithium acetate procedure (Schiestl and Gietz, 1989). Genomic integration of IDR transformants was done using the seamless Delitto Perfetto in vivo site-directed mutagenesis method (Storici et al., 2001) at the endogenous Ste50 IDR locus. Genomic integration of the pFUS1-GFP reporter was done at the HO locus using a selectable marker (URA3). Genomic integration of pRPL39-ymCherry and pRPL39-mTagBFP2 was done at the pCAN1 locus using a selectable marker (LEU2). Hog1 was tagged with yemGFP at the C-terminus using Delitto Perfetto.

All experiments were performed on log-phase cells grown at 30°C in rich (YEP) or synthetic complete (SC) media lacking appropriate nutrients to maintain selection of markers, unless otherwise stated. 2% glucose was used as the carbon source for all strains. Where necessary, Geneticin (G418) or 5-Fluoroorotic acid (5-FOA) (Boeke et al., 1987) were used for selection or counter-selection, respectively.

2.6.3 Confocal microscopy and image analysis

All images were acquired on a TCS-SP8 confocal microscope (Leica).

For the morphology experiment, cells were imaged in brightfield on standard, uncoated glass slides. For quantification of morphology, single cells in micrographs were segmented using the thresholding function in ImageJ (Schneider et al., 2012) applied to brightfield images (see Appendix Figure 1-2 for example images) slightly below the focal plane. The features of each segmented cell, particularly the length of the major axis and circularity, were quantified in ImageJ, and Gaussian mixture modeling (using the “mclust” package in R (Fraley et al., 2012)) was employed to recognize budding cell (long major axis, lower circularity) and non-budding cell (shorter major axis, high circularity) clusters for each replicate experiment, which included 4 micrographs for each of the 4 genotypes. In each replicate (16 images), we automatically identified 271-532 cells of each of the 4 genotypes.

Abnormal cells (and mis-segmented objects) were exclusively assigned to the budding cell cluster by the Gaussian mixture model due to their elongated shape. To identify these abnormal cells, we quantified the Mahalanobis distance of each cell in the budding cell cluster to the centre of that cluster (identified independently in each replicate, which includes all 4 genotypes). The 10% most distant cells to the centre of the budding cell cluster were classified as being abnormally-shaped (for each replicate). We divided the number of abnormally shaped cells by the total number cells of that genotype, and reported the average over the 3 replicates in Figure 2-2c.

To control for possible variation in the fraction of budded cells for each genotype (for example due to cell cycle effects of the mutations, or other types of variation) that could lead to a bias to identify abnormal cells, we also computed the percentage of budded cells classified as abnormal

for each genotype and found the same results as reported above: the 5A strain has a significantly higher fraction of abnormal cells than the WT or orthologous IDR strains.

For the dynamic Hog1 signaling assay, co-cultured Hog1-GFP tagged wildtype and experimental strains (expressing constitutive ymCherry and mTagBFP2, respectively [see Strain construction, above]) were imaged simultaneously on glass dishes coated with 0.1 mg/mL concanavalin A (conA) as a binding agent (in order to allow for continuous imaging of the same cells in media over time) (as described in (Pemberton, 2014)). Briefly, glass dishes were spotted with conA solution for 15 minutes, after which point the conA was aspirated and the spot was washed with sterile water. Once the conA spot was dry, the cells were incubated on the conA spot for 10 minutes, excess cells were washed off, and the dish was filled with synthetic complete media lacking histidine and leucine (the same media the cells were cultured in). Hog1-GFP was visualized in the cells at baseline levels every 5 minutes for 10 minutes, after which NaCl (dissolved in media) was added to the dish on the microscope stage to a final concentration of 0.5 M, serving as the stimulus for the HOG pathway. After addition of NaCl, the same cells were imaged every 5 minutes for 60 minutes. Eight evenly-spaced z-slices covering approximately 6 microns in the z-plane were imaged, and the maximum projections of these z-stacks were used for downstream analysis. After visualization of Hog1-GFP using the 488 nm laser, the 558 nm and 405 nm lasers were switched on to identify the genotype of the cells (wildtype or experimental IDR, based on which fluorescent tag [mCherry or mTagBFP2] was being constitutively expressed). We sequentially switched on the lasers in this way to prevent the cells from exposure to blue (UV) light during the experiment.

Automated segmentation and quantification of Hog1-GFP time-lapse microscopy images was done using previously-described methods (Handfield et al., 2013). Images were manually filtered to remove out-of-focus or mis-segmented cells as well as buds lacking nuclei. Normalized spatial spread (Handfield et al., 2013) of Hog1-GFP fluorescence was used as a measure of nuclear localization. We plotted this measure over time for each cell, and reported the average area under the curve in Figure 2-2e. All comparisons are made between co-cultured cells that were imaged on the same dish and in the same field of view.

2.6.4 Quantification of basal FUS1 expression

Flow cytometry was performed on a MACSQuant VYB (Miltenyi Biotec Inc.). GFP expression of the integrated pFUS1-GFP reporter (see Strain construction, above) was quantified for 50 000 cells per biological replicate. All GFP intensity values were normalized to the mean wildtype GFP intensity value of the day the experiment was run. Normalized mean GFP intensity values are reported in Figure 2-2d.

2.6.5 Quantitative fitness assay

The quantitative fitness assay was adapted from (Breslow et al., 2008). Briefly, individual strains were grown for 48 hours at 30°C in 5mL of cultures on a rolling wheel. To start the competitive fitness experiment, equal proportions of wildtype and experimental strains (constitutively expressing ymCherry or ymTagBFP2) were mixed in deep 96-well blocks (100ul of a single ymCherry expressing strain and 100ul of a single ymTagBFP2 expressing strain into 600ul distilled water) at a final 1024-fold dilution. The cells were then serially diluted 1024-fold every 24 hours. With an estimated 2x108 yeast cells per mL at saturation, the population size (Ne) is approximately 3.44x105.

Each genotype was labeled with both fluorescent proteins, and there were four biological replicates of each competition (two with each colour combination). Therefore, we controlled for potential competitive advantage of expressing one fluorescent protein over the other by pooling equal replicates of each colour combination (e.g. two biological replicates of blue wildtype vs. red experimental strain plus two biological replicates of red wildtype vs. blue experimental strain). Using the MACSQuant VYB (Miltenyi Biotec Inc.) flow cytometer, 50 000 cells per competition were counted at the 20th and the 40th generation. We analysed the data using Flowing software (by Perttu Terho, freely available at flowingsoftware.com) to identify the two differently-coloured populations of cells. Gates for each population were drawn manually to exclude cells fluorescing in both red and blue channels (dead cells), and were kept consistent throughout the experiment. We then calculated the relative selection coefficient (s) as the increase in logarithmic ratio of the wildtype (WT) and experimental (EXP) cells every generation (Chao and Cox, 1983; Hegreness et al., 2006; Hietpas et al., 2011), as follows:

$$\frac{\ln\frac{EXP_t}{WT_t} - \ln\frac{EXP_0}{WT_0}}{t} = \ln(1+s)$$

Where t indicates the number of generations, and s is the selection coefficient. We report s in Figure 2-3.

2.6.6 Ste50 IDR sequence feature calculations

We calculated a series of different features for the wildtype Ste50 IDR as well as each IDR that we engineered, and regressed the mean pFUS1-GFP levels as a quantitative functional output on these values (correlation shown in Figure 2-4). We calculated length, proportion of TP sites, SP sites, or TP/SP sites, number of TP sites, SP sites, or TP/SP sites, net charge of TP sites, SP sites, TP/SP sites, and net charge plus varying levels of basal phosphorylation on TP sites, SP sites, or TP/SP sites for each IDR. For net charge, we added positively charged residues (lysine, arginine) to negatively charged residues (glutamic acid, aspartic acid) for each IDR, unless otherwise indicated. For net charge with basal phosphorylation ("basal net charge"), we calculated net charge with the above-mentioned method, but added a charge of -2 for each phosphorylation site that could potentially be phosphorylated in the IDR. For example, if an IDR had 3 phosphorylation sites and a net charge of +2, and we considered the net charge with basal phosphorylation of 2 SP/TP sites, we calculated a value of 2 + 2*-2 = -2. All trait calculations were made using base functions in R, except the proportion of TP, SP or TP/SP sites, which was calculated using the "protr" package in R (Xiao et al., 2015), the Henderson-Hasselbalch net charge and hydrophobicity calculations, which were done using the "Peptides" package in (Osorio et al., 2015), and the polarity calculation, which was done using the "alakazam" package in R (Gupta et al., 2015).

2.6.7 Test for selection on IDR sequence features/quantitative traits

To estimate evolutionary time for the phylogenetic comparative method, we assumed that evolutionary distance could serve as a proxy for evolutionary time (following (Bedford and Hartl, 2009)). Multiple sequence alignments for Ste5 and its orthologs were obtained from the Yeast Gene Order Browser (YGOB) (Byrne and Wolfe, 2005). All evolutionary analyses were performed on only the longest disordered region within Ste5; the boundaries of this region across

all orthologs were determined with DISOPRED3 (Jones and Cozzetto, 2015) predictions for S. cerevisiae, as with Ste50 (described above). Evolutionary distances for both Ste50 and Ste5 disordered regions were estimated across the YGOB species' phylogeny (Byrne and Wolfe, 2005) using PAML (Yang, 2007) under the WAG model, with an initial kappa of 2, initial omega of .4, and clean data set to 0.

To obtain the expectation of quantitative trait evolution in the absence of selection on the quantitative trait, we simulated a set of 1000 IDRs, following methods and using software from (Nguyen Ba et al., 2014). Briefly, we used a phylogenetic hidden Markov model to infer 1) the location of conserved functional SLiMs, 2) a column (per site) rate of evolution, and 3) a local (window of 31 residues) rate of evolution. The simulated disordered regions were generated using the S. cerevisiae disordered region as the root sequence, the constraints inferred from the phylogenetic hidden Markov model, as well as an amino acid substitution model that accounts for the exchangeability of amino acid pairs specific to disordered regions (Nguyen Ba et al., 2014).

We applied a Brownian motion (BM) model to both real and simulated sequences. BM is a model that can be used to describe the evolution of quantitative traits (Felsenstein, 1973). This model is given by the equation: $dX(t) = \boldsymbol{\sigma} dB(t)$, where $dX(t)$ represents the change in a trait value (X) over time (t), $\boldsymbol{\sigma}$ represents the intensity of random fluctuations, and B(t) is drawn at random from a normal distribution with a mean of 0 and a variance of $\boldsymbol{\sigma}2$ (Butler and King, 2004). We applied this model using the "GEIGER" package in R (Harmon et al., 2008). Basal net charge was calculated (see Ste50 IDR sequence feature calculations, above) assuming basal phosphorylation of up to two 'SP' motifs (each phosphorylation event decreases the net charge by 2). As a negative control, we defined another quantitative trait, i.e., "scrambled" charge, with positive and negative charges reassigned to four different residues (asparagine, glycine, threonine, alanine) than those known to be charged under physiological pH conditions, and assuming basal phosphorylation of up to two "scrambled" phosphorylation sites ('PSX' motifs, where X is any amino acid other than proline).

Estimation of evolutionary variance with Brownian motion assumes mutations have approximately symmetrically distributed effects on quantitative traits with mean equal to zero. We therefore tested the average effect of a random mutation on the basal net charge trait

(Appendix Figure 1-7). We did this by using evolver in the PAML package (Yang, 2007). We simulated nucleotide evolution using the Ste50 IDR nucleotide sequence as the root sequence, under the HKY85 model with parameters estimated from the Ste50 IDR alignment of sensu stricto species: kappa of 3.36, and base frequencies of 0.20370, 0.31145, 0.31481, and 0.17003 for T, C, A, and G, respectively. We ran the simulation 2000 times, and calculated the difference in basal net charge from the initial root sequence for the 1472 Sequences that only had 1 mutation.

2.7 Acknowledgements

Thanks to Henry Hong and Drs. Mojca Mattiazzi Usaj, Yihan Lin and Michael Elowitz for technical advice and/or assistance with time-lapse microscopy, Dr. Louis-François Handfield and Mitchell Li Cheong Man for advice and assistance with scripts for time-lapse image analysis and phylogenetic comparative analysis, respectively, Drs. Julie Forman-Kay, Sergio Peisajovich, Alan Davidson, and Muluye Liku and other members of the Moses lab for helpful discussions, and the Peisajovich lab for use of the MACSQuant VYB flow cytometer. This work was funded by a Natural Sciences and Engineering Research Council (NSERC) doctoral Canada Graduate Scholarship (T.Z.), Ontario Graduate Scholarship (T.Z.), an NSERC Discovery Grant (A.M.M.), Canadian Institutes of Health Research Grant no. PJT-148532 (A.M.M.), and infrastructure grants from the Canada Foundation for Innovation (A.M.M.).

2.8 Author contributions

T.Z., A.N.N.B., and A.M.M. designed research; T.Z. and C.N.T. performed research; A.N.N.B. contributed new reagents/analytic tools; T.Z. and C.N.T. analyzed data; and T.Z., C.N.T., and A.M.M. wrote the paper.

2.9 Supplementary materials

Supplementary materials are available in Appendix 1.

Chapter 3
Proteome-wide signatures of function in highly diverged intrinsically disordered regions

This work has been submitted as: Proteome-wide signatures of function in highly diverged intrinsically disordered regions.

Submitted to eLife.

Taraneh Zarin[1], Bob Strome[1], Alex N Nguyen Ba[2], Simon Alberti[3,4], Julie D Forman-Kay[5,6], Alan M Moses[1,7,8]

1. Department of Cell and Systems Biology, University of Toronto, Toronto, Canada
2. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA 02138
3. Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
4. Technische Universität Dresden, Center for Molecular and Cellular Bioengineering, Biotechnology Center, Dresden, Germany
5. Program in Molecular Medicine, Hospital for Sick Children, Toronto, Canada
6. Department of Biochemistry, University of Toronto, Toronto, Canada
7. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada
8. Department of Computer Science, University of Toronto, Toronto, Canada

3 Proteome-wide signatures of function in highly diverged intrinsically disordered regions

3.1 Abstract

Intrinsically disordered regions make up a large part of the proteome, but the sequence-to-function relationship in these regions is poorly understood, in part because the primary amino acid sequences of these regions are poorly conserved in alignments. Here we use an evolutionary approach to detect molecular features that are preserved in the amino acid sequences of orthologous intrinsically disordered regions. We find that most disordered regions contain multiple molecular features that are preserved, and we define these as "evolutionary signatures" of disordered regions. We demonstrate that intrinsically disordered regions with similar evolutionary signatures can rescue function in vivo, and that groups of intrinsically disordered regions with similar evolutionary signatures are strongly enriched for functional annotations and phenotypes. We propose that evolutionary signatures can be used to predict function for many disordered regions from their amino acid sequences.

3.2 Introduction

Intrinsically disordered protein regions are associated with a large array of functions (reviewed in (Forman-Kay and Mittag, 2013)), including cell signaling (Iakoucheva et al., 2004; Tompa, 2014; Wright and Dyson, 2014), mediation of protein-protein interactions (Borgia et al., 2018; Tang et al., 2012; Tompa et al., 2015), and the formation of membraneless organelles through phase separation (Banani et al., 2017; Franzmann et al., 2018; Nott et al., 2015; Patel et al., 2015; Riback et al., 2017). These regions are widespread in eukaryotic proteomes (Peng et al., 2013; Ward et al., 2004), but do not fold into stable secondary or tertiary structures, and do not typically perform enzymatic functions (Uversky, 2011). Although intrinsically disordered regions can readily be identified based on their primary amino acid sequence (Dosztányi et al., 2005; Uversky, 2002), it remains a challenge to associate these regions with specific biological and biochemical functions based on their amino acid sequences, limiting systematic functional analysis. In stark contrast, for folded regions, protein function can often be predicted with high specificity based on the presence of conserved protein domains (El-Gebali et al., 2018) or enzymatic active sites (Ondrechen et al., 2001). Analogous methods to assign function to

intrinsically disordered regions based on evolutionary conservation (or other sequence properties) are of continuing research interest (reviewed in (Van Der Lee et al., 2014)).

We and others (Davey et al., 2012; Nguyen Ba et al., 2012) have shown that short segments of evolutionary conservation in otherwise rapidly evolving disordered regions point to key functional residues, often important for posttranslational modifications, or other transient protein interactions (Tompa et al., 2014). However, these conserved segments make up a small fraction of disordered regions (5%), and the vast majority of disordered amino acids show little evidence for evolutionary constraint in alignments of primary amino acid sequences (Colak et al., 2013). It is currently unclear how intrinsically disordered regions persist at high frequency in the proteome, given these apparently low levels of evolutionary constraint.

One hypothesis for the preponderance of disordered regions despite high amino acid sequence divergence, is that the "molecular features" of disordered regions that are important for function (such as length (Schlessinger et al., 2011), complexity (Alberti et al., 2009; Halfmann, 2016; Kato et al., 2012; Molliex et al., 2015), amino acid composition (Moesa et al., 2012), and net charge (Mao et al., 2010; Strickfaden et al., 2007; Zarin et al., 2017)) do not lead to detectable similarity in primary amino acid sequence alignments. Indeed, recently, evidence that such molecular features can be under evolutionary constraint has been reported for some proteins (Daughdrill et al., 2007; Lemas et al., 2016; Zarin et al., 2017). For example, we showed that signaling function of a disordered region in the *Saccharomyces cerevisiae* protein Ste50 appears to depend on its net charge, and we found evidence that this molecular feature is under evolutionary constraint, despite no evidence for homology of the primary amino acid sequence in alignments (Zarin et al., 2017).

Here we sought to test whether evolutionary preservation of molecular features is a general property of highly diverged intrinsically disordered protein regions. To do so, we obtained a set of 82 sequence features reported in the literature to be important for disordered region function (Appendix Table 2-1). We computed these for *S.cerevisiae* intrinsically disordered regions and their orthologs, and compared them to simulations of molecular evolution where conserved segments (if any) are retained, but where there is no selection to retain molecular features (Nguyen Ba et al., 2014, 2012). Deviations from the simulations indicate that the highly diverged

intrinsically disordered regions are preserving molecular features during evolution through natural selection (Zarin et al., 2017).

We find that many intrinsically disordered regions show evidence for selection on multiple molecular features, which we refer to as an “evolutionary signature”. Remarkably, we show that intrinsically disordered regions with similar evolutionary signatures appear to rescue function, while regions with very different signatures cannot, strongly supporting the idea that the preserved molecular features are important for disordered region function. By clustering intrinsically disordered regions based on these evolutionary signatures, we obtain (to our knowledge) the first global view of the functional landscape of these enigmatic protein regions. We recover patterns of molecular features known to be associated with intrinsically disordered region functions such as subcellular organization and targeting signals. We also identify new patterns of molecular features not previously associated with functions of disordered regions such as DNA repair and ribosome biogenesis. Finally, we show that similarity of evolutionary signatures can generate hypotheses about the function of completely disordered proteins. Taken together, our results indicate that evolutionary constraint on molecular features in disordered regions is so widespread that sequence-based prediction of their functions should be possible based on molecular features.

3.3 Results

3.3.1 Proteome-wide evolutionary analysis reveals evolutionarily constrained sequence features are widespread in highly diverged intrinsically disordered regions

We identified more than 5000 intrinsically disordered regions (IDRs) in the *S.cerevisiae* proteome and quantified their evolutionary divergence (see Methods). As expected, we found that the IDRs evolve more rapidly than the regions that were not identified as disordered (Appendix Figure 2-1). We also confirmed that the vast majority of these IDRs are distinct from Pfam domains (Appendix Figure 2-2). These results are consistent with previous reports (Brown et al., 2010; Colak et al., 2013; de la Chaux et al., 2007; Khan et al., 2015; Light et al., 2013; Tóth-Petróczy and Tawfik, 2013) that the primary amino acid sequence alignments of IDRs show high levels of divergence and it is not possible to annotate IDR functions using standard homology-based approaches.

To test for selection on molecular features in these IDRs, we applied a method that we recently used to show evidence of selection on an IDR in the *S.cerevisiae* Ste50 protein (Zarin et al., 2017). We obtained 82 molecular features that have been reported or hypothesized to be important for IDR function (Appendix Table 2-1) and tested whether these molecular features are under selection in the *S. cerevisiae* IDRs (see Methods for details). Briefly, we compare the distribution of a given molecular feature in a set of orthologous IDRs to a null expectation, which is formed by simulating the evolution of each IDR. When the mean or variance of the molecular feature across the orthologous IDRs deviates from the distribution of means or variances in our null expectation, we predict that this feature is under selection, and thus could be important for the function of the IDR in question. For example, in the Ste50 IDR, as reported previously (Zarin et al., 2017), we found that the variance of the net charge with phosphorylation of the IDR falls outside of our null expectation, while the mean falls within our null expectation (Figure 3-1A).

We applied this analysis to 5149 IDRs (see Methods) and computed the percentage of IDRs where the evolution of each molecular feature fell beyond our null expectation (empirical $p<0.01$, Figure 3-1B). We find that charge properties such as net charge and acidic residue content are most likely to deviate from our null expectation (more than 50% of IDRs) (Figure 3-1B). This is in contrast to non-conserved motif density, which deviates from our null expectation in 21.6% of IDRs at most (for CDK phosphorylation consensus sites). Other molecular features that frequently deviate from our null expectation are sequence complexity (43.0%), asparagine residue content (43.3%), and physicochemical features such as isoelectric point (53.9%). We also found that the mean of each molecular feature deviates from our null expectation more often than the variance (Figure 3-1B). These results suggest that there are many more molecular features that are under selection in IDRs than is currently appreciated (Daughdrill et al., 2007; Lemas et al., 2016; Zarin et al., 2017).

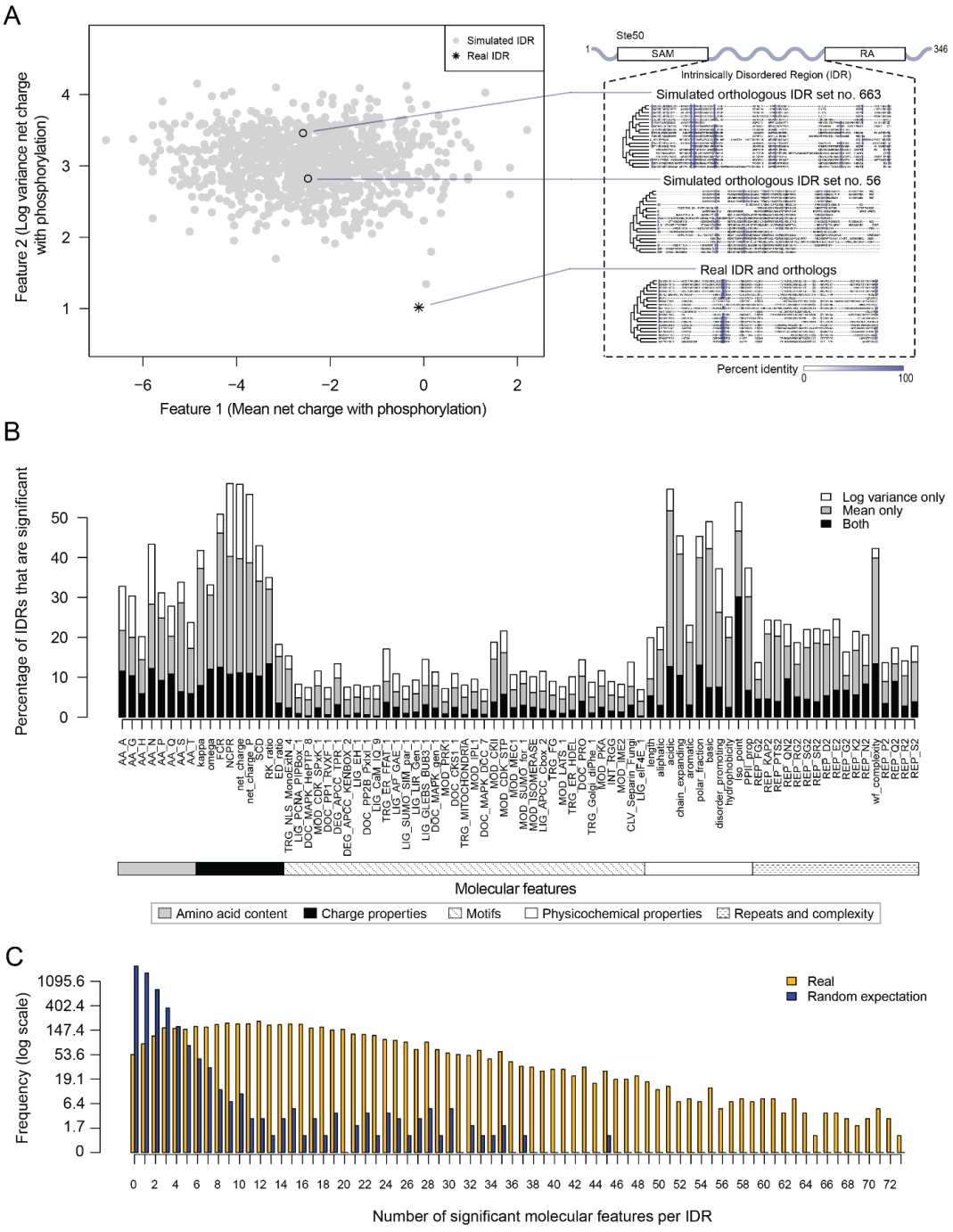


Figure 3-1. Proteome-wide evolutionary analysis reveals evolutionarily constrained sequence features are widespread in highly diverged intrinsically disordered regions. A) Left: Mean versus log variance of the "net charge with phosphorylation" molecular feature for the real Ste50 IDR (a.a. 152-250) ortholog set and simulated Ste50 orthologous IDR sets (N=1000). Right: Example simulated Ste50 orthologous IDR sets (no. 663 and no. 56 out of 1000) and the real Ste50 IDR and its orthologs, coloured according to percent identity in the primary amino acid sequence. B)

Percentage of IDRs that are significantly deviating from simulations in mean, log variance, or both mean and log variance of each molecular feature. C) Frequency [1+log(frequency)] of number of significant molecular features per IDR for the real IDRs (yellow) versus the random expectation (blue) obtained from a set of simulated IDRs.

Next, we quantified the number of molecular features that are significant per IDR, assigning significance to a molecular feature if either the mean, variance, or both mean and variance of the molecular feature deviated from our null expectation (empirical $p<0.01$, Figure 3-1C). Surprisingly, many IDRs have many significant molecular features, with a median of 15 significant molecular features per IDR (compared to 1 significant feature expected by chance; see Methods). Although many of our features are correlated (see Discussion), these results suggest that the deviation from our expectations of molecular feature evolution is not due to a few outlier IDRs, but rather that most IDRs tend to have multiple molecular features that are under selection.

3.3.2 Intrinsically disordered regions with similar molecular features can perform similar functions despite negligible similarity of primary amino acid sequences

The analysis above indicates that highly diverged IDRs typically contain multiple molecular features that are under selection. To summarize the set of preserved molecular features in each IDR, we computed Z-scores comparing either the observed mean or variance of each molecular feature in the orthologous IDRs to our simulations (see Methods). We call these summaries of evolution of molecular features (vectors of Z-scores) “evolutionary signatures”. If the features are important for function, IDRs with similar evolutionary signatures are predicted to perform (or at least be capable of performing) similar molecular functions. To test this hypothesis, we replaced the endogenous Ste50 IDR with several IDRs from functionally unrelated proteins: Pex5, a peroxisomal signal receptor (Erdmann and Blobel, 1996), Stp4, a predicted transcription factor (Abdel-Sater et al., 2004), and Rad26, a DNA-dependent ATPase involved in Transcription Coupled Repair (Gregory and Sweder, 2001; Guzder et al., 1996) (Figure 3-2A). Ste50 is an adaptor protein in the High Osmolarity Glycerol (HOG) and mating pathways (Hao et al., 2008; Jansen et al., 2001; Tatebayashi et al., 2007; Truckses et al., 2006; Yamamoto et al.,

2010) whose IDR is important for basal mating pathway activity (as measured by expression of a reporter driven by the Fus1 promoter) (Hao et al., 2008; Zarin et al., 2017). The IDRs that we used to replace the Ste50 IDR all have negligible similarity when their primary amino acid sequences are aligned, but vary in the similarity of their evolutionary signatures (to the Ste50 IDR, Figure 3-2A). We found that the basal mating reporter expression in each strain corresponded to how similar the evolutionary signature of the replacing IDR was to that of the Ste50 IDR (all mutants significantly different from wildtype and each other, Wilcoxon test $p<0.05$, Figure 3-2B). To further assay mating pathway activity, we exposed the wildtype and chimaeric strains with IDRs from Pex5, Stp4 and Rad26 to mating pheromone. We found that the two chimaeric strains that were more similar in their evolutionary signatures to the wildtype (Pex5 and Stp4) began the process of "shmooing", or responding to pheromone, whereas the strain that had the IDR with the most different evolutionary signature (Rad 26) could not shmoo (Figure 3-2C; full micrographs in Appendix Figure 2-3). That the evolutionary signature of molecular features of IDRs can be used to predict which IDRs can rescue signaling function suggests that these signatures may be associated with IDR function.

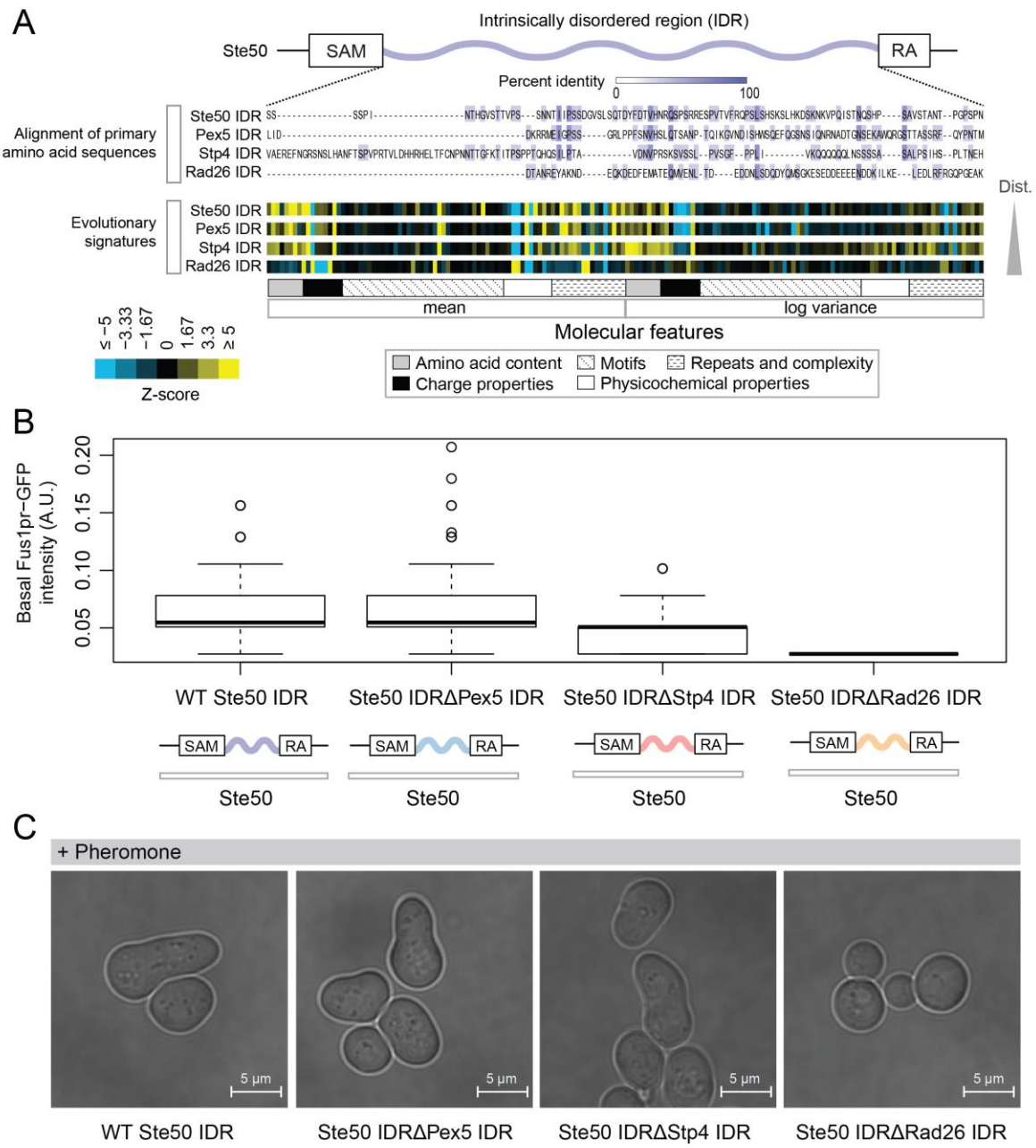


Figure 3-2. Intrinsically disordered regions with similar evolutionary signatures can rescue wildtype phenotypes, while those with different evolutionary signatures cannot. A) Multiple sequence alignment of Ste50 IDR (a.a. 152-250), Pex5 IDR (a.a. 77-161), Stp4 (a.a. 144-256), and Rad26 IDR (a.a. 163-239) shows negligible similarity when their primary amino acid sequences are aligned, while evolutionary signatures show that the Pex5 and Stp4 IDRs are more similar to the Ste50 IDR than the Rad26 IDR. IDRs are presented in order of increasing Euclidian distance between their evolutionary signatures. The Ste50 IDR is located between the

Sterile Alpha Motif (SAM) and Ras Association (RA) domains in the Ste50 protein. B) Boxplots show distribution of values corresponding to basal Fus1pr-GFP activity in an *S.cerevisiae* strain with the wildtype Ste50 IDR compared to strains with the Pex5, Stp4, or Rad26 IDR swapped to replace the Ste50 IDR in the genome. Boxplot boxes represent the 25^{th}-75^{th} percentile of the data, the black line represents the median, and whiskers represent 1.5*the interquartile range. Outliers fall outside the 1.5*interquartile range, and are represented by unfilled circles. Distribution of GFP activity is based on quantification of GFP intensity in single cells pooled from 4 colonies (which we define as biological replicates) for each strain; sample sizes for each distribution are as follows: WT n=588 cells, Pex5 IDR n=196 cells, Stp4 IDR n=228 cells, Rad26 IDR n=271 cells. C) Brightfield micrographs showing each strain from part B following exposure to pheromone. Shmooing cells are those which have elongated cell shape, i.e. mating projections.

3.3.3 Proteome-wide view of evolutionary signatures in disordered regions reveals association with function

To test the association of function with evolutionary signatures in highly diverged IDRs, we clustered and visualized the evolutionary signatures for 4646 IDRs in the proteome (see Methods) (Figure 3-3). Remarkably, the evolutionary signatures reveal a global view of disordered region function. The IDRs fall into at least 23 clusters based on similarity of their evolutionary signatures (groups A through W, Figure 3-3) that are significantly associated with specific biological functions (enriched for Gene Ontology (GO) term, phenotype, and/or literature annotations, False Discovery Rate [FDR]=5%, Benjamini-Hochberg corrected) (Table 3-1; full table of enrichments in supplementary data; clustered IDRs and evolutionary signatures in supplementary data). Given that this level of specificity of biological information has not been previously associated with sequence properties of highly diverged IDRs, we performed a series of controls, ensuring that our clusters are not based on homology between IDRs, that our annotation enrichment results are not due to a mis-specification of the null hypothesis, and to confirm that these annotation enrichment results cannot be obtained simply based on amino acid frequencies of IDRs (Appendix Table 2-2; see Methods).

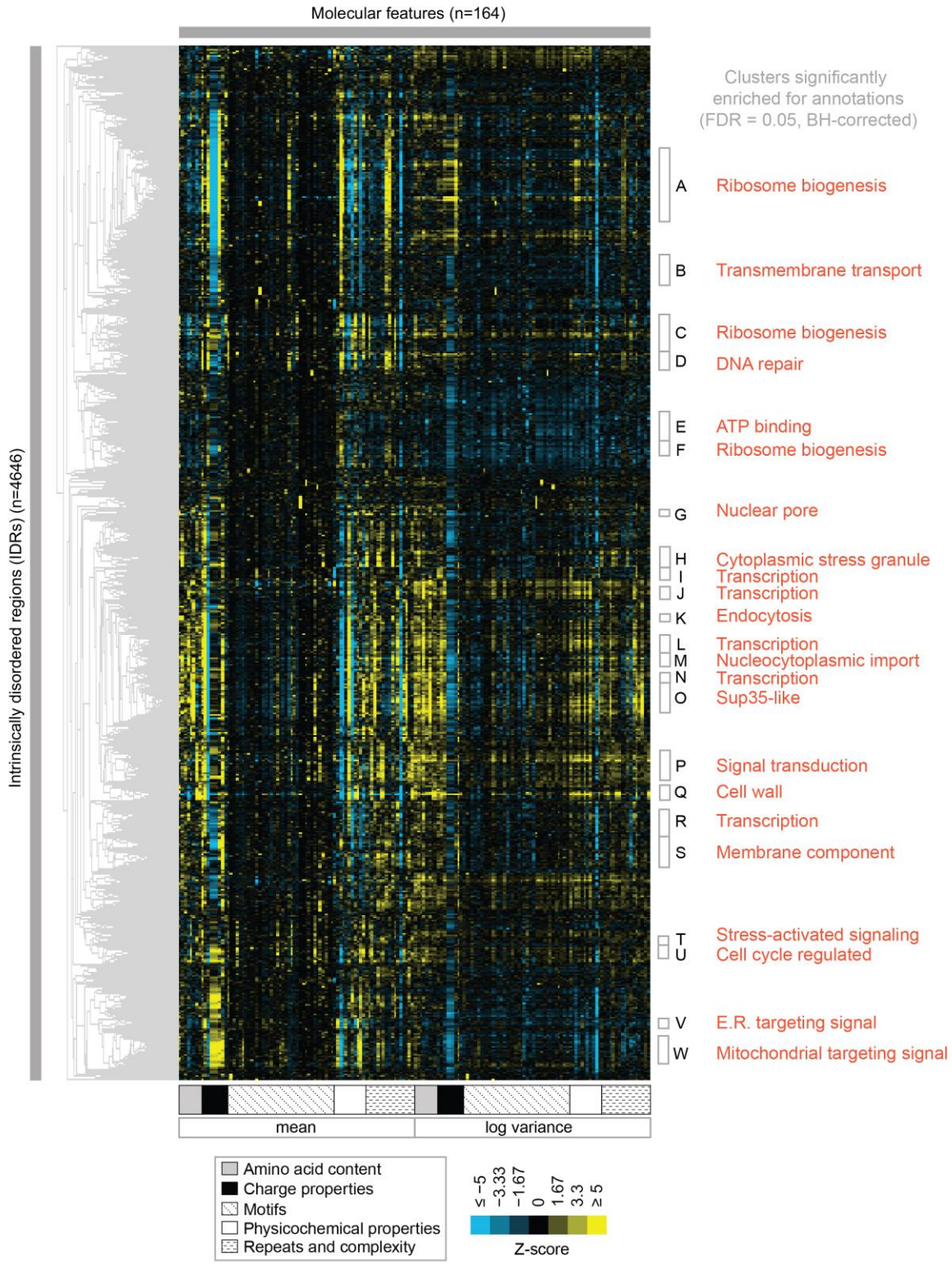


Figure 3-3. Clustering evolutionary signatures shows that IDRs in the proteome share evolutionary signatures, and that these clusters of IDRs are associated with specific biological

functions. A-W show clusters significantly enriched for annotations (see Table 3-1; full table of enrichments in supplementary data). Cluster names represent summary of enriched annotations.

Table 3-1. Top 5 enriched GO term annotations and top 3 enriched phenotype annotations for each cluster (in order of decreasing corrected p-values). Full table of >1000 significant GO term, phenotype, and literature enrichments in supplementary data.

ID	Annotations (Positive proteins in cluster/Total proteins in cluster)	Corrected P <=
A	nucleus (201/295), rRNA processing (40/295), ribosome biogenesis (39/295), nucleolus (50/295), maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (14/295), inviable (110/295), RNA.accumulation..decreased (46/295), RNA.accumulation..increased (39/295)	1.46e-03
B	amino acid transmembrane transport (8/140), amino acid transmembrane transporter activity (8/140), transmembrane transport (21/140), amino acid transport (9/140)	1.11e-02
C	nucleolus (42/159), rRNA processing (27/159), ribosome biogenesis (26/159), nucleus (107/159), preribosome, large subunit precursor (13/159), RNA.accumulation..increased (28/159), inviable (60/159), RNA.accumulation..decreased (27/159)	4.88e-03
D	nucleus (72/86), DNA repair (20/86), cellular response to DNA damage stimulus (18/86), DNA binding (28/86), damaged DNA binding (7/86), mutation.frequency..increased (14/86), chromosome.plasmid.maintenance..decreased (29/86), cell.cycle.progression.in.S.phase..increased.duration (4/86)	4.21e-02
E	motor activity (4/89), ATP binding (25/89), ASTRA complex (3/89)	4.23e-02
F	90S preribosome (11/73), rRNA processing (14/73), ribosome biogenesis (14/73), endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) (6/73), nucleolus (15/73)	2.49e-02
G	nuclear pore nuclear basket (4/35), nucleocytoplasmic transporter activity (4/35)	4.54e-02
H	nucleic acid binding (16/66), translational initiation (7/66), cytoplasmic stress granule (9/66), mRNA binding (13/66), translation initiation factor activity (6/66)	3.60e-03
I	regulation of transcription, DNA-templated (23/52), transcription, DNA-templated (22/52), positive regulation of transcription from RNA polymerase II promoter (12/52)	6.58e-03
J	RNA polymerase II transcription factor activity, sequence-specific DNA binding (10/52), positive regulation of transcription from RNA polymerase II promoter (14/52), regulation of transcription, DNA-templated (21/52), RNA polymerase II core promoter proximal region sequence-specific DNA binding (9/52), transcription, DNA-templated (19/52)	1.22e-02
K	trehalose biosynthetic process (2/19), Golgi to endosome transport (3/19), ubiquitin binding (4/19)	3.81e-02
L	sequence-specific DNA binding (21/70), RNA polymerase II core promoter proximal region sequence-specific DNA binding (13/70), DNA binding (27/70), positive regulation of transcription from RNA polymerase II promoter (17/70), regulation of transcription, DNA-templated (27/70)	6.75e-05
M	structural constituent of nuclear pore (8/54), protein targeting to nuclear inner membrane (5/54), nuclear pore central transport channel (6/54), mRNA transport (9/54), nuclear pore (8/54)	5.87e-05
N	sequence-specific DNA binding (18/39), DNA binding (19/39), zinc ion binding (11/39), regulation of transcription, DNA-templated (19/39), RNA polymerase II transcription factor activity, sequence-specific DNA binding (8/39)	6.21e-04
O	regulation of transcription, DNA-templated (53/130), transcription, DNA-templated (50/130), sequence-specific DNA binding (25/130), positive regulation of transcription from RNA polymerase II promoter (26/130), nuclear-transcribed mRNA catabolic process, deadenylation-dependent decay (8/130), endocytosis..decreased (26/130), invasive.growth..increased (37/130), cell.shape..abnormal (15/130)	1.29e-02
P	intracellular signal transduction (19/129), protein kinase activity (22/129), protein serine/threonine kinase activity (22/129), kinase activity (24/129), phosphorylation (24/129)	3.34e-06
Q	extracellular region (33/67), fungal-type cell wall (30/67), cell wall (25/67), anchored component of membrane (20/67), cell wall organization (23/67)	1.01e-20
R	positive regulation of transcription from RNA polymerase II promoter (21/119), DNA binding (32/119), RNA polymerase II core promoter proximal region sequence-specific DNA binding (12/119), transcription factor activity, sequence-specific DNA binding (10/119), transcription, DNA-templated (33/119)	1.55e-02
S	integral component of membrane (59/133), membrane (68/133), fungal-type vacuole membrane (18/133), vacuole (18/133), L-tyrosine transmembrane transporter activity (4/133)	5.48e-03
T	stress-activated protein kinase signaling cascade (4/33), regulation of apoptotic process (4/33)	3.57e-02
U	cytoskeleton (15/80), spindle (6/80), kinetochore microtubule (3/80)	1.47e-02
V	fungal-type vacuole (15/43), mannosylation (7/43), integral component of membrane (28/43), cell wall mannoprotein biosynthetic process (6/43), alpha-1,6-mannosyltransferase activity (4/43)	1.45e-05
W	mitochondrion (144/165), mitochondrial inner membrane (57/165), mitochondrial matrix (34/165), oxidation-reduction process (31/165), mitochondrial translation (22/165), respiratory.growth..decreased.rate (81/165), respiratory.growth..absent (71/165), mitochondrial.genome.maintenance..absent (25/165)	3.15e-15

Several of the functions that we find enriched within our clusters have been previously associated with molecular features of IDRs, which we recover in our analysis. For example, we find a cluster that is associated with “nucleocytoplasmic transporter activity” (cluster M) that includes IDRs from FG-NUP proteins Nup42, Nup145, Nup57, Nup49, Nup116, and Nup100 that form part of the nuclear pore central transport channel (Alber et al., 2007). In cluster M, we find molecular features such as increased asparagine content, increased polar residue content, and increased proline and charged residue demixing (“Omega” (Martin et al., 2016)) in addition to the well-known “FG” repeats that are found in the FG-NUP IDRs (reviewed in (Terry and Wente, 2009)). Another interesting example is cluster O, which contains IDRs from proteins that are enriched for a wide range of annotations such as “P-body”, “cytoplasmic stress granule”, “actin cortical patch”, and “DNA binding”. Cluster O contains IDRs from proteins associated with phase separation and membraneless organelles such as Sup35 (Franzmann et al., 2018) and Dhh1 (Protter et al., 2018). The evolutionary signatures for the IDRs in this cluster include features that are typically associated with so-called “prionogenic”, low complexity disordered regions, such as increased mean polyglutamine repeats (Alberti et al., 2009), but also indicate that there are other relevant molecular features for this set of disordered regions (Figure 3-4A). For example, in these regions, the variance of the net charge is reduced, and charged residues are depleted during evolution. These sequence features are illustrated in Figure 3-4B, where we compare the presence of glutamine and charged residues in an example disordered region from this cluster (Ccr4; a protein that is known to accumulate in P-bodies (Teixeira and Parker, 2007)) to an example from the corresponding simulation (Figure 3-4B). Taken together, these results indicate that our analysis captures molecular features that have been previously associated with IDR functions, and suggests additional molecular features in these IDRs that may be important for their functions.

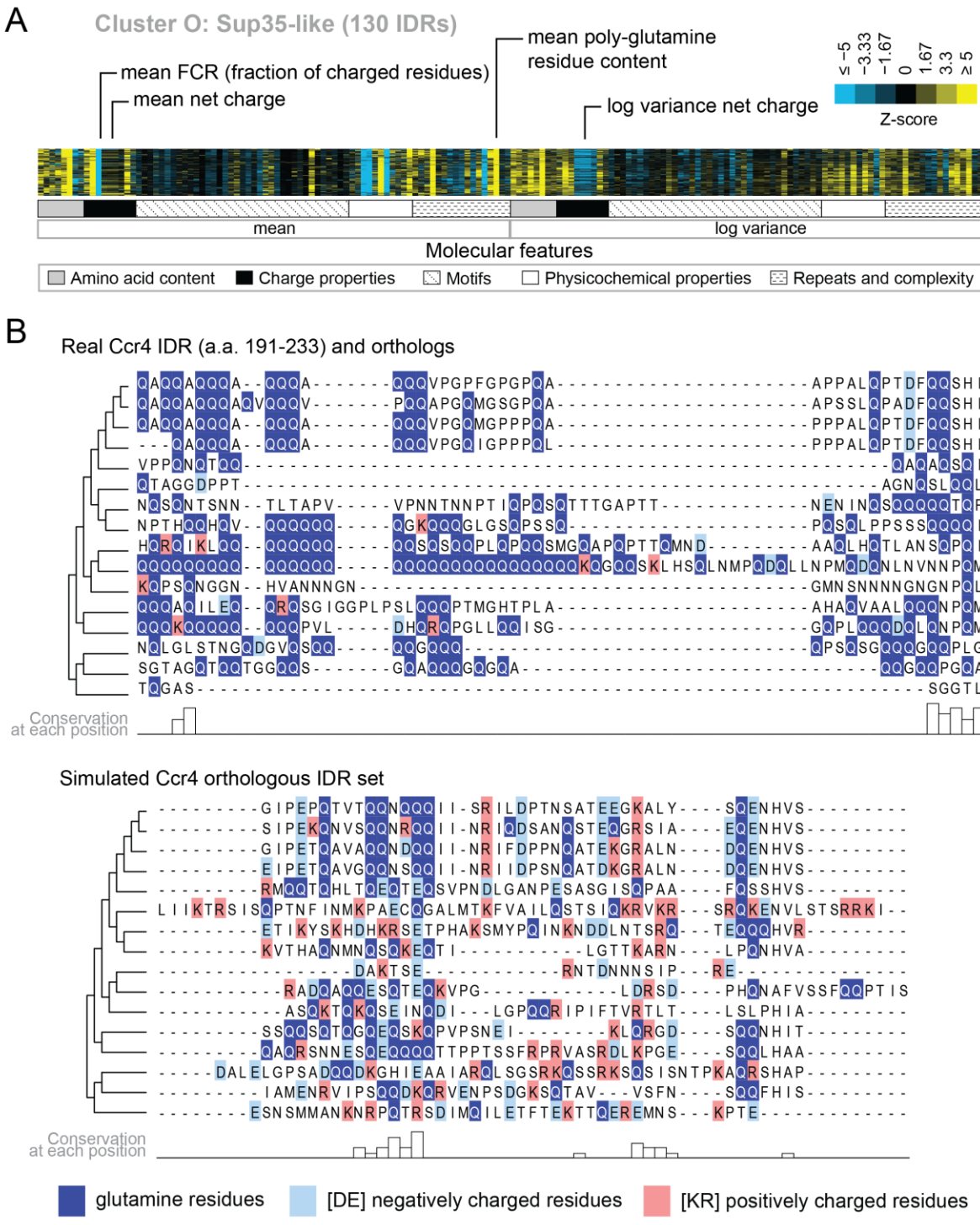


Figure 3-4. Evolutionary signatures in cluster O contain some molecular features that are typically associated with IDRs as well as some that are not. A) Pattern of evolutionary signatures in cluster O. B) Example disordered region from cluster O, Ccr4, with a subset of highlighted molecular features compared between its real set of orthologs and an example set of simulated

orthologous IDRs. Species included in phylogeny in order from top to bottom are *S.cerevisiae, Saccharomyces mikatae, Saccharomyces kudriavzevii, Saccharomyces uvarum, Candida glabrata, Kazachstania naganishii, Naumovozyma castellii, Naumovozyma dairenensis, Tetrapisispora blattae, Tetrapisispora phaffii, Vanderwaltozyma polyspora, Zygosaccharomyces rouxii, Torulaspora delbrueckii, Kluyveromyces lactis, Eremothecium (Ashbya) cymbalariae, Lachancea waltii*.

We also find functions associated with our clusters that have not been previously associated with molecular features of IDRs. For example, cluster D (Figure 3-5A) is associated with DNA repair, and its evolutionary signature contains increased mean "Kappa" (Das and Pappu, 2013) and decreased mean "Sequence Charge Decoration" (SCD) (Sawle and Ghosh, 2015), both of which indicate that there is an increased separation of positive and negatively charged residues in these IDRs compared to our null expectation. This is illustrated by the IDR from Srs2, a protein that is known to be involved in DNA repair (Aboussekhra et al., 1989; Yeung and Durocher, 2011), and shows high charge separation compared to an example corresponding simulation (Figure 5B). The evolutionary signature for this cluster also reveals an increased mean fraction of charged residues and negatively charged residues in particular (Figure 3-5A), which is also clear in the comparison between the real Srs2 orthologs and the simulation (Figure 3-5B). Although acidic stretches have been associated with IDRs in histone chaperones (Warren and Shechter, 2017), to our knowledge, the separation of oppositely charged residues has not been associated with the wider functional class of DNA repair IDRs.

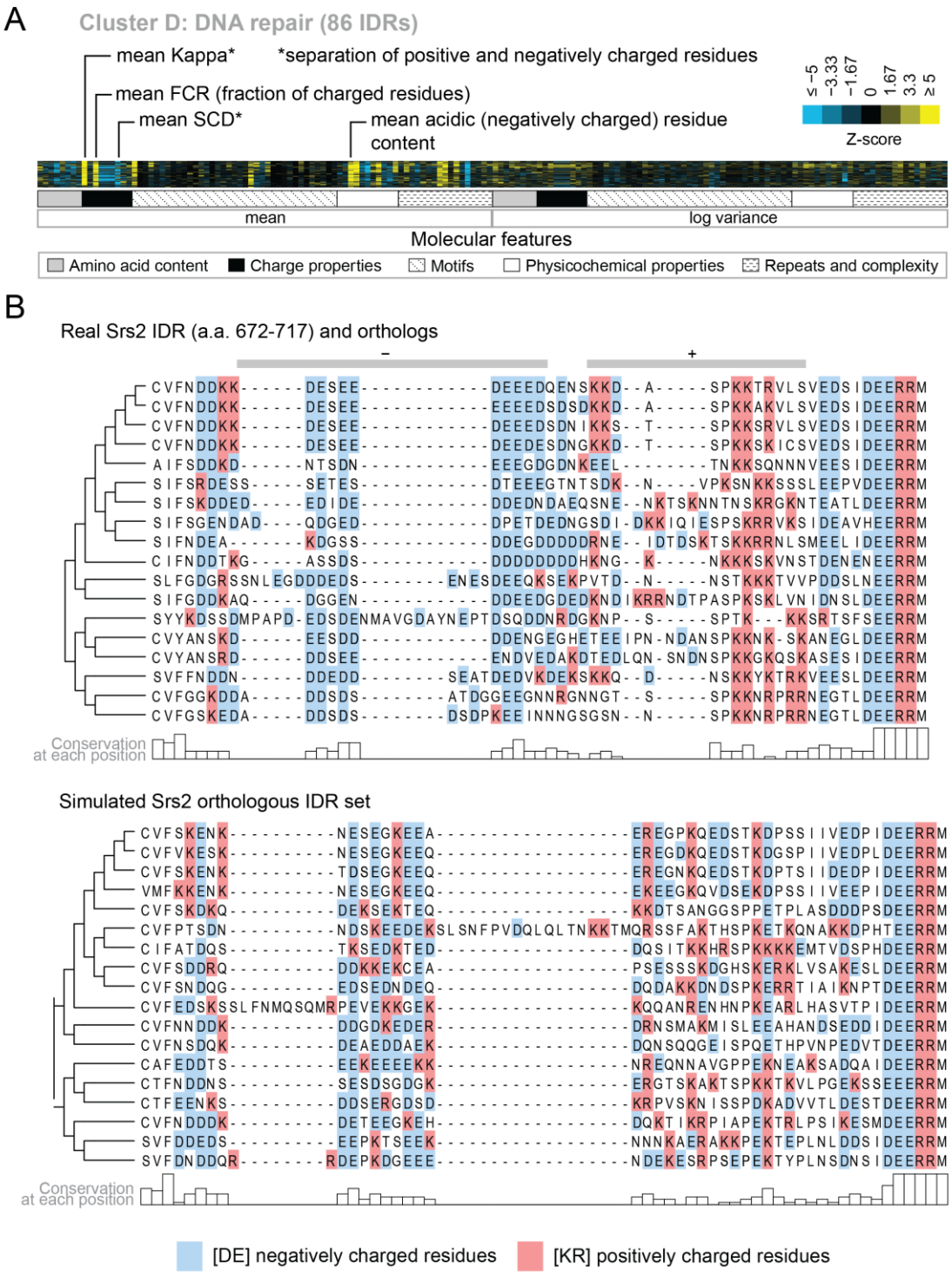


Figure 3-5. Cluster D contains disordered regions associated with DNA repair. A) Pattern of evolutionary signatures in cluster D. B) Example disordered region from cluster D, Srs2, with a subset of highlighted molecular features compared between its real set of orthologs and an example set of simulated orthologous IDRs. Species included in phylogeny in order from top to

bottom are *S.cerevisiae, S.mikatae, S.kudriavzevii, S.uvarum, C.glabrata, Kazachstania africana, K.naganishii, N.castellii, N.dairenensis, T.phaffii, Z.rouxii, T.delbrueckii, K.lactis, Eremothecium (Ashbya) gossypii, E.cymbalariae, Lachancea kluyveri, Lachancea thermotolerans, L.waltii*.

Our analysis also indicates that there is not necessarily a 1:1 mapping between IDRs with shared evolutionary signatures and current protein functional annotations. For example, we find three clusters associated with ribosome biogenesis (cluster A, C, F) that cannot be distinguished based on their enriched GO terms. The largest of these is cluster A, where 201/295 proteins have a "nucleus" annotation, and 110/295 are essential proteins ("inviable" deletion phenotype). This cluster is also enriched for several phenotypes associated with RNA accumulation (Table 3-1, cluster A; see supplementary data for full list of significant enrichments). Cluster A contains highly acidic IDRs with CKII phosphorylation consensus sites. CKII has been previously associated with nucleolar organization (Louvet et al., 2006), and a previous analysis of non-conserved consensus phosphorylation sites found ribosome biogenesis as strongly enriched in predicted CKII targets (A. C. W. Lai et al., 2012). In contrast, cluster C shares neither of these molecular features with cluster A, and cluster F shares only highly acidic residue content. Interestingly, cluster C contains increased mean polylysine repeats, and is significantly enriched for proteins that have been experimentally verified as targets for lysine polyphosphorylation (Bentley-DeSousa et al., 2018) ($p=2.7x10^{-3}$, hypergeometric test). Overall, although the IDRs in these clusters share different evolutionary signatures, they are all found in proteins associated with ribosome biogenesis. We hypothesize that these different signatures point to different functions relating to ribosome biogenesis, but we have no indication of what these might be based on current protein annotations (see Discussion).

We find similar observations in multiple clusters that have distinct evolutionary signatures enriched for terms associated with regulation of transcription (clusters I, J, L, N, O, R). These clusters are not clearly separable based on mechanistic steps of transcription (such as sequence-specific DNA binding, chromatin remodeling, etc.). Some of these clusters exhibit molecular features that have been associated with different classes of transcriptional activation domains that are based on amino acid composition (reviewed in (Frietze and Farnham, 2011)). For example,

cluster J, O and N have increased glutamine residue content, while cluster N has increased proline residue content. However, clusters I and R have no amino acid composition bias, while cluster N has increased proline-directed phosphorylation consensus sites, suggesting post-translational modifications. This indicates that our analysis reveals new sub-classifications of transcription-associated IDRs. While we hypothesize that these IDRs have different functions, once more we have no indication of what these functions could be based on current protein annotations (see Discussion).

3.3.4 A cluster of evolutionary signatures is associated with N-terminal mitochondrial targeting signals

One of our clusters of intrinsically disordered regions is exceptionally strongly associated with the mitochondrion (144/165 proteins in the cluster) and other annotations that are related to mitochondrial localization and function (for example, 81/165 proteins in the cluster have shown a decreased respiratory growth phenotype) (Table 3-1, cluster W; see supplementary data for full list of significant enrichments). The vast majority of mitochondrial proteins are synthesized with N-terminal pre-sequences (Maccecchini et al., 1979) (also known as N-terminal targeting signals) that are cleaved upon import (Vögtle et al., 2009) and are thought to sample dynamic structural configurations (Saitoh et al., 2011, 2007) (Figure 3-6A). Since 145/165 of the disordered regions in this cluster are N-terminal, we hypothesized that this cluster contains disordered regions that are associated with mitochondrial targeting signals (Vögtle et al., 2009). In line with this hypothesis, we find previously described sequence features of mitochondrial N-terminal targeting signals in our evolutionary signatures; for example, these IDRs are depleted of negatively charged residues, have an abundance of positively charged residues, and are much more hydrophobic than our null expectation (Appendix Figure 2-4A) (Garg and Gould, 2016; Vögtle et al., 2009). Examples of disordered regions in this cluster include those of the Heme A synthase Cox15 and the mitochondrial inner membrane ABC (ATP-binding cassette) transporter Atm1 (Appendix Figure 2-4B). In order to test our hypothesis that this cluster of evolutionary signatures identifies mitochondrial N-terminal targeting signals, we used a recently published tool that scores the probability that a sequence is a mitochondrial targeting signal (Fukasawa et al., 2015). Using this tool, we find that the IDRs in cluster W have a much higher probability of being mitochondrial targeting signals than any other cluster with enriched annotations in our analysis (Bonferroni-corrected $p \leq 6.5 \times 10^{-11}$, Wilcoxon test) (Figure 3-6B, red box).

Interestingly, the adjacent cluster V (Figure 3-6B, purple box), which we hypothesize to contain targeting sequences for the endoplasmic reticulum, is distinct from cluster W in this analysis.

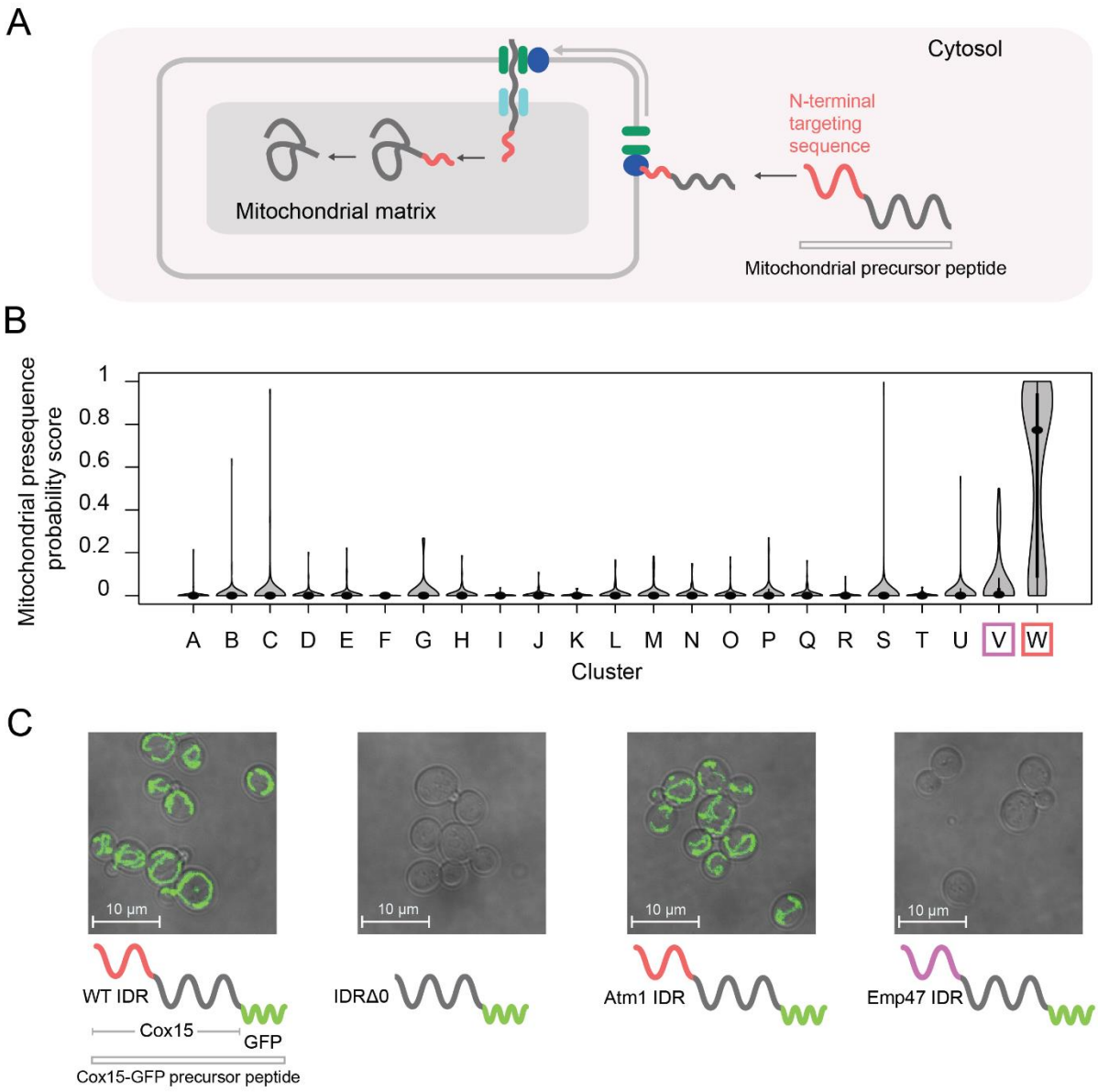


Figure 3-6. Cluster W is associated with mitochondrial N-terminal targeting signals. A) Schematic (not to scale) showing the path of a mitochondrial precursor peptide (with N-terminal targeting sequence in red) from the cytosol, where it is translated, to the mitochondrial matrix, where the peptide folds and targeting sequence is cleaved. B) Violin plots (median indicated by black dot, thick black line showing 25th-75th percentile, and whiskers showing outliers) show distributions of mitochondrial presequence probability scores for all IDRs in each cluster. The

cluster that we predict to contain mitochondrial N-terminal targeting signals is outlined in red, while the cluster that we predict to contain endoplasmic reticulum targeting signals is outlined in purple. C) Micrographs of *S.cerevisiae* strains in which Cox15 is tagged with GFP, with either the wildtype Cox15 IDR, deletion of the Cox15 IDR, replacement of the Cox15 IDR with the Atm1 IDR (also in the mitochondrial targeting signal cluster), or replacement of the Cox15 IDR with the Emp47 IDR (from the endoplasmic reticulum targeting signal cluster).

If the specificity of the function of the IDRs in this cluster is strong, we predict that swapping an IDR from cluster W with that of a verified mitochondrial targeting sequence would result in correct localization to the mitochondria, while swapping an IDR from a different cluster would not. To test this, we first used the (uncharacterized) disordered region from Atm1 that falls into cluster W to replace that of Cox15, which also falls into cluster W and is an experimentally verified mitochondrial targeting sequence (Vögtle et al., 2009) (Figure 3-6C). In accordance with our hypothesis, we find that GFP-tagged Cox15 correctly localizes to the mitochondria when its disordered region is swapped with that of Atm1, but does not localize correctly when its disordered region is deleted (Figure 3-6C; full micrographs in Appendix Figure 2-5). We also repeated this experiment with another protein that has an experimentally verified N-terminal mitochondrial targeting sequence, Mdl2, and found the same results (Appendix Figure 2-6). Next, we replaced the Cox15 IDR with the disordered region of Emp47, which has an evolutionary signature that we predict to be associated with targeting signals for the endoplasmic reticulum (cluster V). In this case, as we predicted, we found no mitochondrial localization of Cox15-GFP. Importantly, these putative targeting signals have no detectible similarity when their primary amino acid sequences are aligned, and we therefore suggest that the similarity in their molecular features is preserved by stabilizing selection (see Discussion). These results confirm that IDRs with similar evolutionary signatures can rescue subcellular targeting functions, and suggest that the evolutionary signatures are specific enough to predict function of at least some IDRs.

3.3.5 Evolutionary signatures of function can be used for functional annotation of fully disordered proteins

A major challenge to proteome-wide analysis of IDRs is the limited applicability of homology-based sequence analysis. Proteins with a mixture of disordered regions and structured domains can be assigned function based on homology to their structured domains, but fully disordered proteins are much more difficult to classify (reviewed in (Van Der Lee et al., 2014)). We therefore asked whether hypotheses about functions of fully disordered proteins could be generated using evolutionary signatures. We identified ten yeast proteins of unknown function that are predicted to be most disordered (see Methods). To predict function according to our clustering analysis, we simply assigned them the annotation of the cluster in which they fell (Table 3-2). For example, Rnq1 has been extensively studied as a "yeast prion", but there is no clear function associated with this protein under normal conditions (Kroschwald et al., 2015; Sondheimer and Lindquist, 2000; Treusch and Lindquist, 2012). Interestingly, Rnq1 falls into our cluster of disordered regions that are associated with nucleocytoplasmic transport (cluster M) and the nuclear pore central transport channel. While Rnq1 is annotated with a cytosolic localization, an *RNQ1* deletion was recently shown to cause nuclear aggregation of the polyQ-expanded huntingtin exon1 (Httex1) in a model of Huntington's disease (Zheng et al., 2017). Therefore, we propose a role for Rnq1 in nucleocytoplasmic transport. For some of these largely disordered proteins, we obtain large disordered segments falling into multiple clusters (indicated by more than one cluster ID in Table 3-2), suggesting more than one possible function for the protein (see Discussion). This analysis illustrates how evolutionary signatures can be used to generate hypotheses of function for fully disordered proteins.

Table 3-2. Evolutionary signatures of function can be used for functional annotation of previously uncharacterized proteins and IDRs.

ID	Name	Description	% Disorder	Cluster ID
YCL028W	RNQ1	Protein whose biological role is unknown; localizes to the cytosol	96	M: Nucleocytoplasmic transport
YKL105C	SEG2	Protein whose biological role is unknown; localizes to the cell periphery	92	P: Signal transduction
YGR196C	FYV8	Protein whose biological role is unknown; localizes to the	89	A: Ribosome biogenesis

		cytoplasm in a large-scale study		R: Transcription
YGL023C	PIB2	Protein whose biological role is unknown; localizes to the mitochondrion in a large-scale study	86	R: Transcription
YOL036W		Protein whose biological role and cellular location are unknown	84	P: Signal transduction R: Transcription
YNL176C	TDA7	Protein whose biological role is unknown; localizes to the vacuole	83	Q: Cell wall organization
YFR016C		Protein whose biological role is unknown; localizes to both the cytoplasm and bud in a large-scale study	83	A: Ribosome biogenesis
YBL081W		Protein whose biological role and cellular location are unknown	82	M: Nucleocytoplasmic transport
YBR016W		Protein whose biological role is unknown; localizes to the bud membrane and the mating projection membrane	82	O: Sup35-like
YOL070C	NBA1	Protein whose biological role is unknown; localizes to the bud neck and cytoplasm and colocalizes with ribosomes in multiple large-scale studies	81	Does not fall into annotated cluster; close to ribosome biogenesis cluster

3.4 Discussion

In this work, we tested for evolutionary constraints on highly diverged intrinsically disordered regions proteome-wide. In contrast to the relative lack of constraint on primary amino acid sequence alignments (compared to folded regions, (Brown et al., 2002; Tóth-Petróczy and Tawfik, 2013)), we find that the vast majority of disordered regions contain molecular features that deviate in their evolution from our null expectation (a simulation of disordered region evolution (Nguyen Ba et al., 2014, 2012)). Our discovery that highly diverged disordered regions contain (interpretable) molecular features that are under evolutionary constraint provides researchers with testable hypotheses about molecular features that could be important for function in their proteins of interest. Furthermore, in principle, our framework for the analysis of diverged disordered regions can be extrapolated to proteins from other species.

Importantly, our choice of features was based on previous reports of important sequence features in IDRs that could be easily calculated for protein sequences and scaled to millions of simulated sets of orthologous IDRs. Thus, our evidence for constraint must represent a lower bound on the total amount of functional constraint on highly diverged IDRs: there are very likely to be sequence characteristics that were not captured by our features. Further, even when we do find evidence for constraint on a feature, we do not know whether our feature represents the actual feature required for IDR function, or is simply correlated with it. For example, we found IDRs that show constraint on glycine and arginine content, but these may reflect the real constraint on planar-pi interactions (Vernon et al., 2018) and are not fully captured by either of these features. In the future, we could exhaustively search for protein sequence features that best explain the evolutionary patterns as was done for features of activation domains that explain reporter activity (Ravarani et al., 2018).

Despite the somewhat arbitrary choice of molecular features, we found strong evidence that groups of disordered regions share “evolutionary signatures”, and that these groups of IDRs are associated with specific biological functions. To demonstrate the association of evolutionary signatures with previously known functions, we associated IDRs with protein function. However, many proteins contain multiple IDRs. In these proteins, the IDRs may perform different functions (just as multiple folded domains may perform independent functions), thus complicating the mapping of molecular functions to molecular features of IDRs. Systematic data at the level of individual IDRs would greatly facilitate future progress in this area.

Another challenge in associating specific functions with individual IDRs is that current bioinformatics predictions of IDRs at the proteome level often lead to arbitrary breaks (or merging) of IDRs, as IDR boundaries are very difficult to define precisely (even with sensitive experimental approaches (Jensen et al., 2013)). Whether or not IDRs serve as distinct functional units across a linear peptide sequence, and where the boundaries for these regions lie on a proteome-wide scale, is an area for further research. In our cluster analysis, we find that the vast majority of IDRs in multi-IDR proteins fall into different clusters, and that this matches our expectation from random chance. A small minority of IDRs from very large (>1500 amino acid) disordered proteins cluster together, suggesting that they are “broken up” pieces of larger units.

Despite the caveat of IDR boundaries in proteome-wide analyses, evolutionary signatures of selection on molecular features represent a new way to assign function to the large numbers of currently enigmatic IDRs that have been identified based on protein sequences. This approach is complementary to current bioinformatics approaches to predict IDR function that are based on presence (Edwards et al., 2007) or conservation of SLiMs (Beltrao and Serrano, 2005; Davey et al., 2012; A. C. W. Lai et al., 2012; Nguyen Ba et al., 2012), prediction of interactions (MoRFs) (Fuxreiter et al., 2004; Lee et al., 2012; Mohan et al., 2006; Oldfield et al., 2005; Vacic et al., 2007), or the recently proposed phase separation propensity score (Vernon et al., 2018).

Widespread evidence for shared functions in the highly diverged portions of IDRs also has several evolutionary implications. The lack of homology between most IDRs with similar evolutionary signatures suggests that the molecular features are preserved in each IDR independently. For example, the more than 150 IDRs that we believe represent mitochondrial N-terminal targeting signals share similar constraints on their molecular features, yet these signals have been preserved independently over very long evolutionary time as mitochondrial genes were transferred individually to the nuclear genome (Adams and Palmer, 2003). The preservation of molecular features over long evolutionary time, despite accumulation of amino acid divergence, is consistent with a model of stabilizing selection (Bedford and Hartl, 2009; Hansen, 1997; Lande, 1976), where individual amino acid sites are under relatively weak functional constraints (Landry et al., 2014). In this view, single point mutations are unlikely to dramatically impair IDR function, and therefore large evolutionary divergence can accumulate. This also suggests that disease-causing mutations in disordered regions are more likely to cause gain of function, consistent with at least one recent study (Meyer et al., 2018).

Although current models for the evolution of short linear motifs (well-characterized functional elements in IDRs) also implicate stabilizing selection (Koch et al., 2018; Landry et al., 2014), these motifs represent only a minority of the residues in disordered regions (Nguyen Ba et al., 2012). Our observation of shared evolutionary signatures associated with specific functions in highly diverged IDRs suggests that this evolutionary mechanism is shaping the proteome on a much wider scale than currently appreciated. Further, stabilizing selection stands in contrast to purifying selection, the major evolutionary mechanism thought to preserve function in stably folded regions of the proteome (Taylor and Raes, 2004). Thus, we propose that these two major

biophysical classes of protein regions (IDRs vs. folded regions) also evolve under two different functional regimes.

3.5 Methods

3.5.1 Multiple sequence alignments and visualization

We acquired orthologs of *Saccharomyces cerevisiae* from the Yeast Gene Order Browser (Byrne and Wolfe, 2005) and made multiple sequence alignments using MAFFT (Katoh and Standley, 2013) with default settings, as previously described (Nguyen Ba et al., 2014, 2012). We visualized multiple sequence alignments using Jalview (Waterhouse et al., 2009).

3.5.2 Quantification of evolutionary divergence of IDRs and ordered regions of the proteome

We identified IDRs in the *S.cerevisiae* proteome using DISOPRED3 (Jones and Cozzetto, 2015) and filtered them to include only those that are 30 amino acids or longer. We identified the non-disordered regions of the proteome as the inverse subset of the IDRs, and again only included regions that are 30 amino acids or longer. Using the multiple sequence alignments constructed for these protein regions (as above), and only including those proteins for which there at least 10 species in the alignment and at least 10 amino acids for each species, we calculated evolutionary distances for each region using PAML (Yang, 2007) using the WAG model, with an initial kappa of 2, initial omega of 0.4, and clean data set to 0. We used the sum of branch lengths for each region to estimate the evolutionary divergence, and plotted the distribution of this metric for IDRs and non-IDRs in the *S.cerevisiae* proteome in Appendix Figure 2-1.

3.5.3 Quantification of IDR overlap with Pfam annotations

We obtained the list of Pfam (El-Gebali et al., 2018) domain coordinates for *S.cerevisiae* from the Saccharomyces Genome Database (SGD) (Cherry et al., 2012). We included domain coordinates that had e-values less than or equal to 1, and which occurred in more than one protein in the *S.cerevisiae* proteome. We then computed the percentage overlap of each IDR (coordinates determined as above) with the Pfam domain coordinates, and plotted the distribution of percent overlap values for all predicted IDRs in the *S.cerevisiae* proteome in Appendix Figure 2-2.

3.5.4 Evolutionary analysis of diverged disordered regions

Evolutionary analysis of diverged disordered regions was performed as in (Zarin et al., 2017), with some modifications to facilitate proteome-wide analysis. Using the multiple sequence alignments of *S.cerevisiae* IDRs and species branch lengths (as described above), we used the previously described phyloHMM software (Nguyen Ba et al., 2012) to estimate the "local rate of evolution", "column rate of evolution", and any Short Linear Motif (SLiM) coordinates. For each IDR, we simulated 1000 orthologous sets of IDRs using the *S.cerevisiae* sequences as the root and a previously described disordered region evolution simulator (Nguyen Ba et al., 2014) that preserves SLiMs and evolves sequences according to disordered region substitution matrices. This simulator requires a scaling factor to convert evolutionary distances from substitutions per site as obtained from PAML (Yang, 2007). We chose the scaling factor such that the average distance between *S.cerevisiae* and *S.uvarum* over all the IDR alignments equals 1.

Sequences and trees were read into R using the "seqinr" (Charif and Lobry, 2007) and "ape" (Paradis and Schliep, 2018) packages, respectively. Sequences were parsed in R using the "stringr" (Wickham, 2010) and "stringi" (Gagolewski, 2019) packages. We calculated all the sequence features for the real and simulated set of IDR orthologs using custom functions in R except for "Omega" (Martin et al., 2016), "Kappa" (Das and Pappu, 2013), and Wootton-Federhen complexity (Wootton and Federhen, 1993), which were calculated using the localCider program (Holehouse et al., 2017) called through R using the "rPython" package (Bellosta, 2015). We calculated the mean and log variance of each feature for each real set of orthologous IDRs and each of the 1000 sets of orthologous IDRs. Because simulations sometimes lead to the deletion of the IDR, we did not include those IDRs that had fewer than 950 non-empty simulations. To obtain a random expectation for Figure 3-1C, we quantified the number of significant ($p<0.01$) molecular features in a set of randomly chosen simulated IDRs (one for each real IDR). To summarize the difference between each real set of orthologous IDRs and its corresponding 1000 simulated sets of orthologous IDRs, we used a standard Z-score (Z) where we subtracted the mean of the simulations (μ) from the real value (x) and divided by the standard deviation of the simulations (σ). The formula for the Z-score is as follows:

$$Z = \frac{x - \mu}{\sigma}$$

3.5.5 Strain construction and growth conditions

All strains (Appendix Table 2-3) were constructed in the *S. cerevisiae* BY4741 background. IDR transformants were constructed using the *Delitto Perfetto in vivo* site-directed mutagenesis method (Storici et al., 2001). Ste50 IDR mutants were constructed in the ssk22Δ0::HisMx3 ssk2Δ0 background as in (Zarin et al., 2017). Genomic changes in transformed strains were confirmed by Sanger sequencing. For mitochondrial strains, starting strains were acquired from the GFP collection (Huh et al., 2003). The Fus1pr-GFP reporter was constructed as in (Zarin et al., 2017) using Gibson assembly (Gibson et al., 2009), integrated at the *HO* locus using a selectable marker (URA3), and confirmed by PCR.

All experiments were done on log-phase cells grown at 30°C in rich or synthetic complete media lacking appropriate nutrients to maintain selection of markers, unless otherwise stated. Two percent (wt/vol) glucose was used as the carbon source.

3.5.6 Confocal microscopy and image analysis

We acquired all images with a Leica TCS SP8 microscope using standard, uncoated glass slides with a 100x objective. For all GFP images, 7 evenly spaced *z*-slices covering ~6 microns in the z plane were collected for each field of view, and maximum projections of these slices were quantified for Figure 3-2B, or presented as micrographs in Figure 3-6C. To quantify basal Fus1pr-GFP expression, single cells in micrographs were segmented using YeastSpotter (http://yeastspotter.csb.utoronto.ca/) (Lu et al., 2019). The segmented masks and corresponding fluorescent images were imported into R using the "EBImage" package (Pau et al., 2010), and GFP intensity for each cell was quantified using a custom R script (sample script available on http://yeastspotter.csb.utoronto.ca). To assay shmooing, log phase cells were inoculated with 1 uM alpha factor for 2 hours at 30°C (as in (Kompella et al., 2016)), at which point they were imaged in brightfield as above. We repeated each microscopy experiment at least twice on different days, and present representative results from one of those days in Figure 3-2B, 3-2C, and 3-66C.

3.5.7 Clustering of proteome-wide evolutionary signatures

Hierarchical clustering was performed using the Cluster 3.0 program (de Hoon et al., 2004). The evolutionary signature data was first filtered to include only those IDRs that had at least one Z-

score with an absolute value of 3 or more, and with at least 95% data present for the 164 features. This resulted in 4646 IDRs (filtered from the initial 5149) that were then clustered using uncentered correlation distance and average linkage, with “cluster” and “calculate weights” options selected for “genes” (i.e. IDRs), but not for arrays (i.e. molecular features). Clusters were picked manually for further analysis. The full clusterplot is available in supplementary data.

In order to ensure that the clustering was not simply due to homology between the disordered regions, for each cluster, we computed the pairwise distance of its disordered region sequences based on the BLOSUM62 substitution matrix, and compared this to the pairwise distance between all disordered regions outside of that cluster (using the Biostrings R package (Pagès et al., 2018)). We compared the pairwise distance of the IDRs in each cluster to that of the IDRs outside that cluster, and calculated the percent of disordered regions that fell in the top 1% of pairwise distance in all the clusters. This metric is presented for each cluster in Appendix Table 2-2. For example, the cluster with the highest amount of “homologous” IDRs according to this threshold (top 1% homology) is cluster Q, with 8.9% homologous IDRs. However, the vast majority of the clusters have negligibly homologous IDRs; for example, 17/23 clusters have less than 1% homology between IDRs.

3.5.8 Tests for enrichment of annotations

Annotations for Gene Ontology (GO) terms, phenotypes, and literature were acquired from SGD (Cherry et al., 2012) for the *S.cerevisiae* proteome. We included GO terms that applied to a maximum of 5000 genes in the *S.cerevisiae* proteome. A test for enrichment of annotations was done using the hypergeometric test for each cluster against all the proteins in the clustering analysis. To obtain Q-values, p-values were corrected using the Benjamini-Hochberg method. Q-values below an FDR of 5% were retained. Because there is not a 1-to-1 correspondence between IDRs and annotations, which are based on proteins, we also calculated Q-values using permutation tests. To do so, we uniformly sampled 1000 clusters of IDRs for each cluster from the 4646 IDRs included in our clusterplot, and obtained the sum of the top ten – log Q-values associated with each test for enrichment, as above. We compared this test statistic to the observed sum of top ten – log Q-values for each cluster, and reported the difference as a standard Z-score in Appendix Table 2-2.

In order to understand how our evolutionary signatures compare to information obtained only from amino acid frequencies, we computed vectors of Z-scores for each IDR that represented their amino acid frequencies normalized to the proteome-wide average. We clustered these vectors using k-means (K=25) with the Cluster 3.0 program (de Hoon et al., 2004). We performed a similar permutation test (as above), where the sample of 1000 clusters was not uniform, but drawn to create 1000 random clusters of IDRs with similar amino acid composition for each cluster. For example, for each IDR in a cluster, we found the cluster that it fell into in the amino acid frequency clusterplot, and sampled from that cluster to replace the IDR in our evolutionary signature clusterplot. We did this 1000 times for each cluster, and used the same test statistic as the above-described permutations to report the difference in enriched annotations between our clusterplot based on evolutionary signatures and the clusterplot based on amino acid frequencies (Appendix Table 2-2).

3.5.9 Identification of highly disordered proteins with unknown function

We identified proteins whose biological role is unknown according to their SGD annotation (Cherry et al., 2012). We quantified the percent of residues that were predicted to be disordered in each protein with unknown function, and present the top ten most disordered proteins in Table 3-2.

3.6 Acknowledgements

We thank Alex X Lu, Dr. Christiane Iserman, Dr. Iva Pritisanac, Shadi Zabad, and Ian S Hsu for comments on the manuscript. We thank Alex X Lu for stimulating discussions about clustering and Dr. Iva Pritisanac for suggesting analysis of completely disordered proteins. We thank Dr. Helena Friesen and Dr. Brenda Andrews for providing strains from the yeast GFP collection. We thank Canadian Institutes for Health Research (CIHR) for funding to AMM and JDF-K, Canada Foundation for Innovation (CFI) for funding to AMM, and the National Science and Engineering Research Council of Canada (NSERC) for an Alexander Graham Bell scholarship and Michael Smith Foreign Study Supplement to TZ.

3.7 Author contributions

T.Z., J.D.F-K., and A.M.M. designed research; T.Z. and B.S. performed research; A.N.N.B. contributed new reagents/analytic tools; T.Z. and A.M.M. analyzed data; S.A., J.D.F-K., and A.M.M.supervised research; and T.Z. and A.M.M. wrote the paper.

3.8 Supplementary materials

Supplementary materials are available in Appendix 2. Supplementary data is available at hershey.csb.utoronto.ca/TZ_evolsig_full_sub_data.tar.gz.

Chapter 4
Predicting function using evolutionary signatures in intrinsically disordered regions

This work has not been previously published.

Taraneh Zarin[1], Bob Strome[1], Alan M Moses[1,2,3]

1. Department of Cell and Systems Biology, University of Toronto, Toronto, Canada
2. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada
3. Department of Computer Science, University of Toronto, Toronto, Canada

4 Predicting function using evolutionary signatures in intrinsically disordered regions

4.1 Abstract

Although 30-40% of the proteome is predicted to be disordered in eukaryotes, the vast majority of these disordered regions remain uncharacterized, even in the "simple" eukaryote *Saccharomyces cerevisiae*. There is currently no broadly applicable method to predict the function of intrinsically disordered regions from their sequences. We previously extracted evolutionary information about molecular sequence features in intrinsically disordered regions, and found that these are associated with different biological functions. Here, we use these evolutionary signatures to predict different functions in the yeast proteome. We are able to use the same features to predict kinase substrates, localization to different cellular compartments, and phenotypes. We identify the molecular features that are predictive of these functions and phenotypes, and demonstrate that knocking out these features *in vivo* leads to loss of function. Finally, we use our model to distinguish different functions of intrinsically disordered regions in the same protein.

4.2 Introduction

Understanding the sequence to function relationship in intrinsically disordered regions (IDRs) is of great research interest, as these regions comprise a large portion of the proteome (Ward et al., 2004), are difficult to classify (Van Der Lee et al., 2014), and are sites of several disease-associated mutations (Andresen et al., 2012; Patel et al., 2015). In recent work, we found that bulk molecular features can be calculated for IDRs, and that the evolution of these molecular features could be quantified into an "evolutionary signature" (Zarin *et al.*, Chapter 3 of this thesis). We used an unsupervised framework to cluster these evolutionary signatures, and found that IDRs with similar evolutionary signatures could be associated with specific functions, suggesting that these signatures contain functional information. A complementary approach to understanding how sequence determines function in IDRs is to assess if we can use these evolutionary signatures to train a classifier that learns different functions or phenotypes. Such a classifier would allow us to generate functional predictions directly from IDR sequences, and would complement existing methods, which have mostly been developed for various subsets of IDRs (reviewed in (Van Der Lee et al., 2014)). For example, models have been developed to

classify glutamine and asparagine-rich prionogenic IDRs (Alberti et al., 2009), positively charged and hydrophobic mitochondrial N-terminal targeting signals (Fukasawa et al., 2015), and, most recently, transactivation domains of transcription factors (Ravarani et al., 2018). Having a unified predictor that can annotate IDR sequences would be a large step forward in understanding the relationship between sequence and function in IDRs. Using evolutionary signatures in particular could be promising as a unifier, as we should not be limited by learning sequence features of a specific sub-type of disordered region. Furthermore, training a classifier will allow us to understand which features in a given evolutionary signature are most predictive of a given function as a product of the model itself, as compared to unsupervised methods such as clustering, which would require the application of dimensionality reduction techniques (thus reducing interpretability of features).

Here, we classify disordered regions according to functions and phenotypes based on their evolutionary signatures. We use an iteratively re-weighted logistic regression model that predicts the probability that a given IDR belongs to a given class (i.e. performs a certain function). Importantly, proteins can have multiple IDRs in them, yet the functional information that we have is often at the protein level. To address this, the model also learns the "weight" of each IDR (estimated by the posterior probability) with respect to the protein function. We apply this model to IDRs in the yeast proteome, and demonstrate that we can predict several types of functions associated with IDRs, including substrates of the Cyclin Dependent Kinase Cdc28, phenotypes such as RNA accumulation, as well as localization to different cellular compartments such as the nucleolus, stress granules, and the mitochondrial matrix. We identify the molecular features that are most predictive of these functions, and show that knocking these out results in loss of localization function *in vivo*. Finally, we present examples of proteins for which multiple predicted IDRs point to different functions.

4.3 Results

4.3.1 Evolutionary signatures of IDRs can be used to predict diverse functions

In order to understand if evolutionary signatures can be used to predict IDR function, we used a statistical model in a binary classification framework (see Methods). Our target data is comprised of the mapping between the yeast IDRs for which we have evolutionary signatures (n=4545), and

different protein annotations, functions, and phenotypes (see Methods for details). We present 8/23 of the functions for which our model had the strongest predictive power in Figure 4-1, where we compare the ROC curves for the real target data (blue) to a randomized control (orange). Prediction of CDK substrates has been an active area of research interest in our lab and others' (Iakoucheva et al., 2004; A. C. W. Lai et al., 2012; Moses et al., 2007a). Since CDK phosphorylation sites are known to be abundant in IDRs (Holt et al., 2009), we used these to assess whether or not our model could classify substrates of Cdc28, and find that it can predict the majority of these substrates with high specificity (Figure 4-1, top right). Another function that has been associated with IDRs, and which has been predicted based on sequence, is mitochondrial targeting (Fukasawa et al., 2015; Vögtle et al., 2009). In recent work (Zarin et al., Chapter 3 of this thesis), we identified a cluster of IDRs as mitochondrial N-terminal targeting signals (Vögtle et al., 2009). Interestingly, we can predict the majority of mitochondrial inner membrane proteins with high specificity (Figure 4-1). This may represent a stratification in the sequence properties of mitochondrial targeting signals that further specify where they are localized beyond the general mitochondrion. In another example, the cell wall associated proteins are almost perfectly predicted by the evolutionary signatures in their IDRs (Figure 4-1, top left). This is interesting, because the cell wall annotation was also the most strongly enriched term that we associated with a set of evolutionary signatures in our previous unsupervised analysis (Zarin et al., Chapter 3 of this thesis). Another application of this method, in principle, is to associate function with unknown proteins based on the evolutionary signatures of their IDRs. We present the top five predictions of proteins that are annotated as unknown in gene description or gene summary on the Saccharomyces Genome Database (SGD) (Cherry et al., 2012) in Appendix Table 3-1. Overall, there is evidence that we can predict at least 8 and up to 23 different functions or localizations of proteins using evolutionary signatures of IDRs.

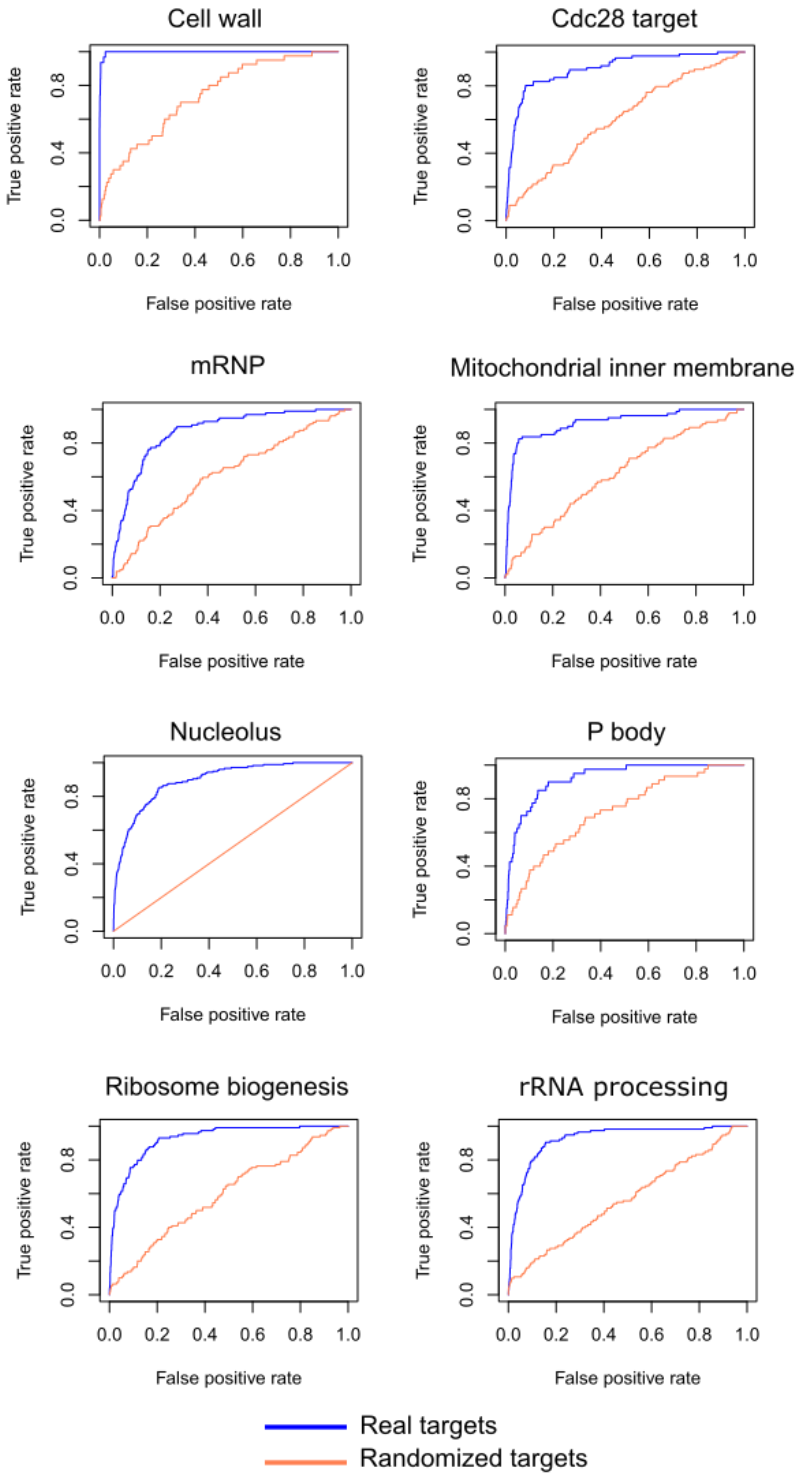


Figure 4-1. A range of functions can be predicted from evolutionary signatures in IDRs. ROC curves comparing the prediction of several different functions and localizations associated with IDRs using real targets (blue) compared to a randomized target control (orange). mRNP refers to

messenger ribonucleoprotein complexes which were identified in budding yeast using a systematic screen (Mitchell et al., 2012).

4.3.2 Different molecular features are predictive of different functions

In order to understand which molecular features underlie the predictions for each function, we extracted and clustered t-scores from the coefficients of the model (Figure 4-2, y-axis). Since the parameters of the model are penalized, many of these are null for each function ("t-score" represented as 0, coloured black in clusterplot). We found some molecular features that were previously known, as well as molecular features that to our knowledge have not been associated with specific IDR functions in the literature. We also clustered the functions (Figure 4-2, x-axis) to see if there were any patterns in their predictive features, and found that related functions clustered together (Figure 4-2, x axis label). For example, the most predictive feature for Cdc28 substrates is the strong CDK consensus motif, [ST]P.K, followed by the weaker consensus motif [ST]P, both of which are expected. The next most predictive feature is the presence of the KEN motif, which is a degradation signal. This feature also seems to be predictive of proteins associated with intracellular signal transduction and protein kinase activity.

The cell wall and extracellular region annotations cluster together, and are predicted by previously unappreciated (to our knowledge) molecular features such as a depletion in mean acidic residue repeats (D/E) and increased mean hydrophobicity. We also find that an increase in mean serine content is predictive of cell wall and extra-cellular proteins, which is encouraging as serine-rich cell wall proteins with large extracellular regions have been reported in the literature for *S.cerevisiae* (Ketela et al., 1999). Examples of these molecular features can be visualized in the top predicted cell wall protein, Aga1, which is known to localize to the cell wall (Figure 4-3). Another interesting cluster of functions that share predictive molecular features are the nucleolus, ribosome biogenesis, rRNA processing, and the phenotype of increased RNA accumulation. These are all related functions and localizations, and share the predictive molecular feature of CKII consensus motifs. CKII is a kinase that has been shown to play roles in nucleolar organization (Louvet et al., 2006). Interestingly, there is a cluster of functions that includes cytoplasmic stress granules, mRNA binding, as well as sets of proteins that were found to form reversible assemblies upon nutrient starvation (mRNA-protein complexes [mRNPs]

(Mitchell et al., 2012)) and heat stress (Wallace et al., 2015). The molecular features that are predictive of these functions include the presence of RGG motifs, a feature that has been previously appreciated as important for RNA binding proteins (P. A. Chong et al., 2018). Other features that are predictive include an increase in K/A/P residue repeats, as well as an absence of threonine residues, which, to our knowledge, have not been previously associated with these functions. Finally, yet another cluster of functions is the mitochondrion, mitochondrial inner membrane, and mitochondrial matrix, which cluster with the decreased respiratory growth phenotype. For these functions, predictive features include an increased mean isoelectric point, which could reflect the reported highly basic charge of mitochondrial targeting signals (Garg and Gould, 2016).

In order to test the hypothesis that we are identifying meaningful features in our evolutionary signatures that are predictive of function, we "knocked out" these features in an example IDR by replacing a demonstrated mitochondrial targeting signal in Cox15, a mitochondrial protein with a known targeting signaling (Vögtle et al., 2009), with a corresponding simulated IDR that does not contain these features (Figure 4-4A). As expected, while the wildtype Cox15-GFP strain has a clear mitochondrial localization (Figure 4-4B), the Cox15-GFP strain replaced with a simulated IDR does not. Thus, the molecular features that make up our evolutionary signatures seem to be directly related to the function of the IDR, as a generic, simulated IDR cannot perform the same function. Beyond the model predictions, this is further evidence that the evolutionary signatures identify molecular features that are important for protein function.

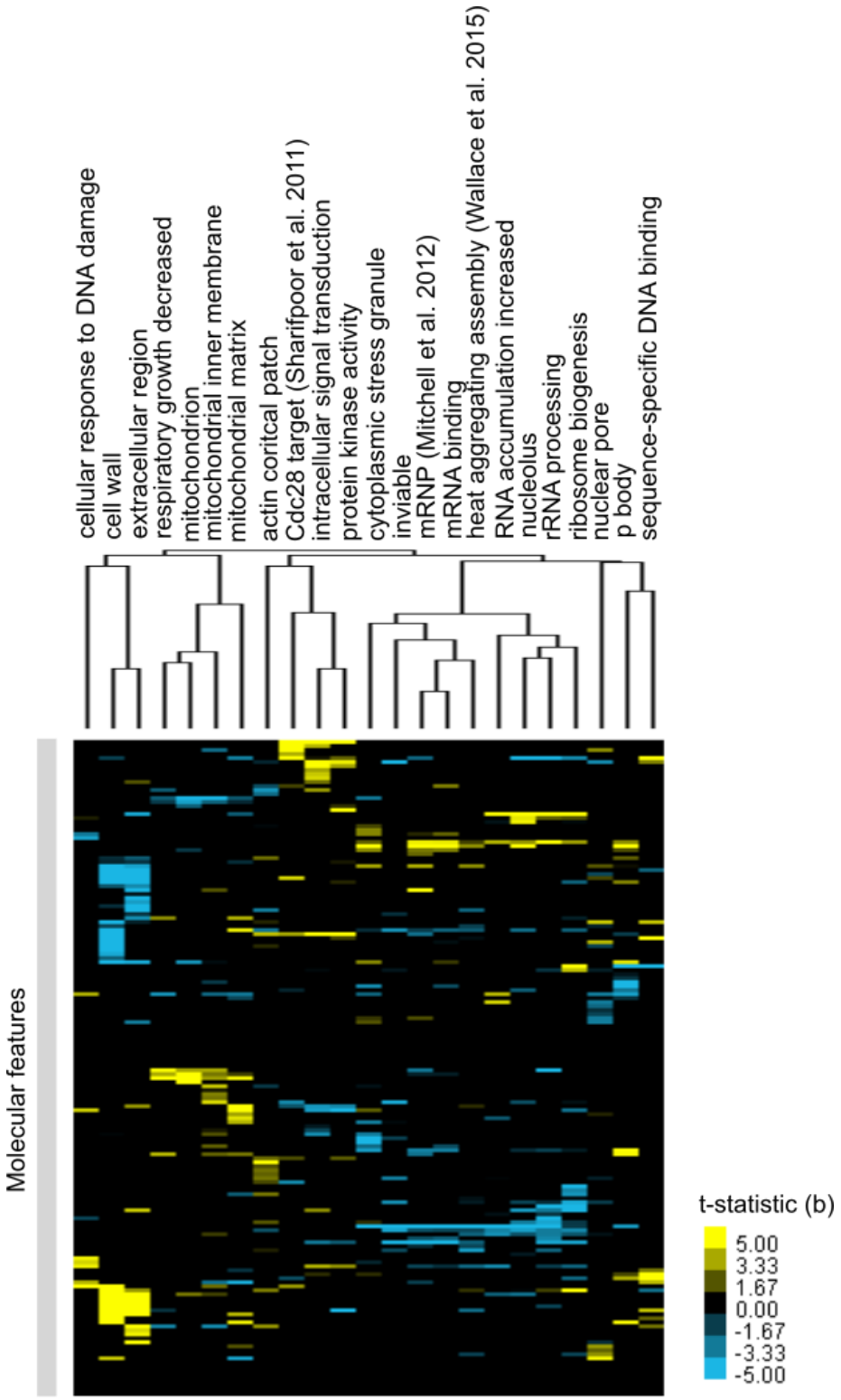


Figure 4-2. Related functions and phenotypes cluster together based on shared predictive molecular features. Clustered t-statistic representing coefficients for different molecular features (b) (n=164) for different functions (n=23).

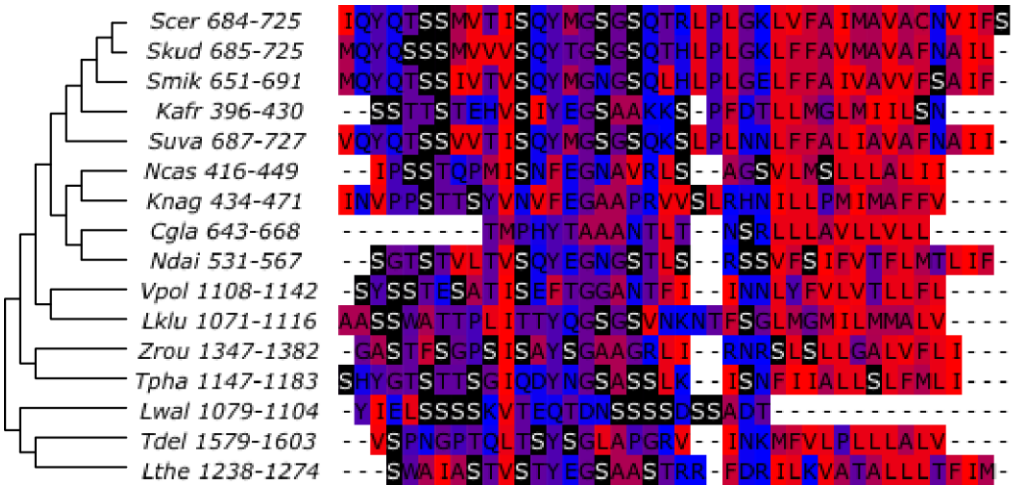


Figure 4-3. Multiple sequence alignment showing the disordered C-terminus of the Aga1 protein. Serines are highlighted with black boxes, while hydrophobicity of amino acids is indicated on a scale from blue (low hydrophobicity) to red (high hydrophobicity). Protein coordinates and species names are indicated on the left side of the alignment. Species, from top to bottom, are: *S.cerevisiae, S.kudriazevii, S.mikatae, K.africana, S.uvarum, N.castellii, K.naganishii, C.glabrata, N.dairenensis, V.polyspora, L.kluyveri, Z.rouxii, T.phaffii, L.waltii, T.delbrueckii, L.thermotolerans.* Alignment is visualized with Jalview (Waterhouse et al., 2009).

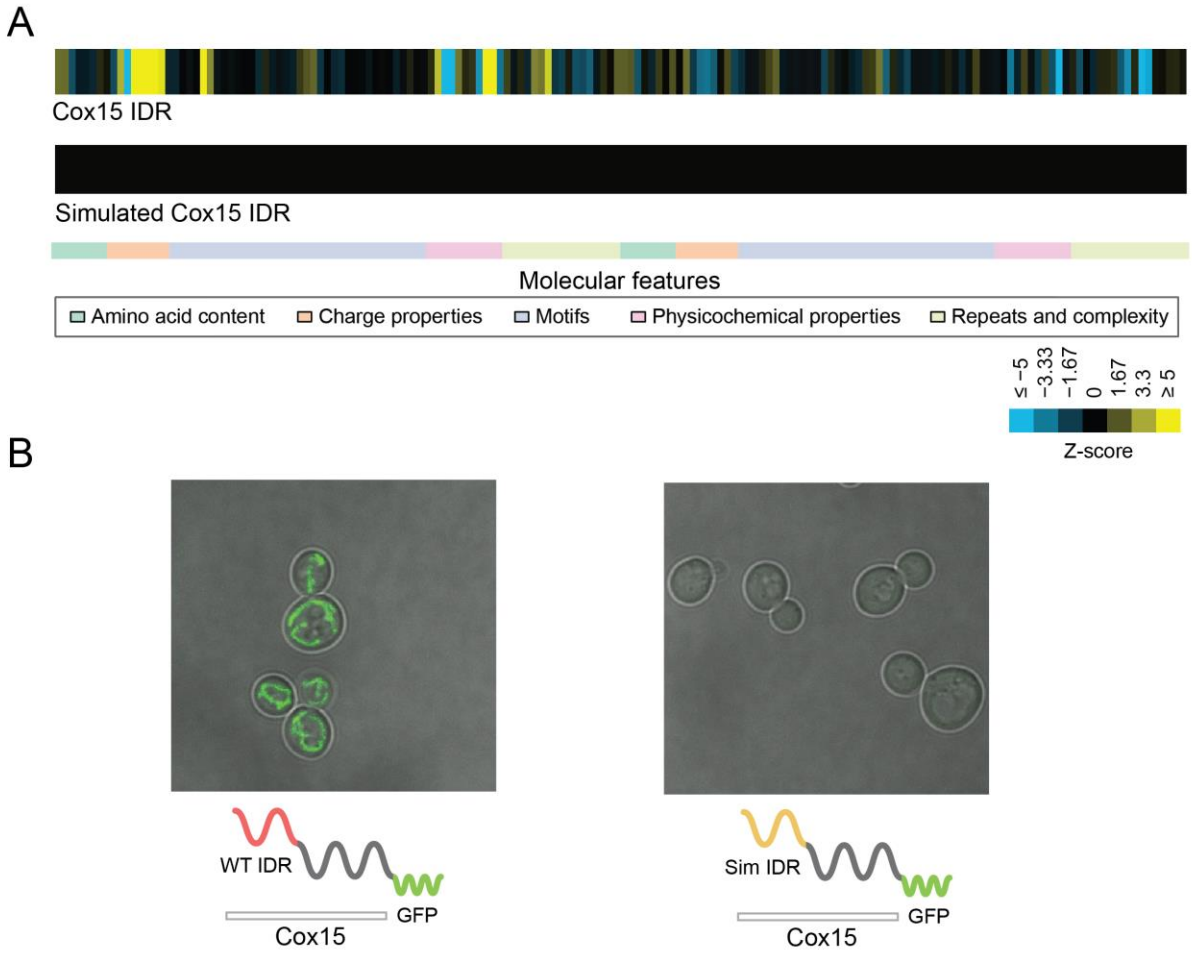


Figure 4-4. Replacement of N-terminal targeting IDR with simulated IDR impairs protein targeting function. A) Evolutionary signature of the Cox15 IDR compared to a schematic of a simulated Cox15 IDR. Z-score summarizes evolution of molecular features. B) Confocal images showing the wildtype Cox15-GFP localization in budding yeast (left) compared to that of a budding yeast strain with the Cox15 IDR replaced by a simulated IDR.

4.3.3 Association of protein function with specific IDRs

Since our model can learn the "weight" associated with each functional prediction for each IDR (i.e. the posterior probability that the IDR in question is associated with the protein function in question), we are able to assign specific functions to specific IDRs in each protein, particularly in those that have multiple IDRs (these comprise close to half of our dataset). This is given that

they perform different functions that are captured in our list of annotations. First, we are able to ascribe the correct function to the correct IDR in the case where the function and location of the IDR are known. An example of this is again demonstrated by mitochondrial targeting signals. The mitochondrial protein Mdl2, for example, has a verified targeting signal at its N-terminus (Vögtle et al., 2009). However, Mdl2 has 2 IDRs in our dataset. Thus, our model should be able to distinguish between these 2 IDRs to identify the one associated with mitochondrial localization. This is indeed the case, as the N-terminal IDR in Mdl2 has a posterior probability value of 0.99 for the mitochondrial inner membrane (0.86 for the more general mitochondrion), whereas the C-terminal Mdl2 IDR has a posterior probability of 0.01 and 0.14, respectively. Another interesting case is the protein Rpm2, which is a protein subunit of the mitochondrial RNase P. This protein has diverse roles in nuclear transcription, cytoplasmic and mitochondrial RNA processing, and is found in the cytoplasm, mitochondria, and p bodies (Morales et al., 2006; Stribinskis et al., 2005; Stribinskis and Ramos, 2007). Rpm2 has 2 IDRs in our dataset, and remarkably, each of these IDRs is strongly associated with different functions that are part of the Rpm2 repertoire of functions. One IDR in this protein has a posterior probability of 0.96 in relation to the mitochondrial matrix, whereas the other has a posterior probability of 0.89 for its association to p body localization. These examples illustrate that evolutionary signatures can be used to train a model that can learn specifically which IDRs are associated with specific functions in proteins.

4.4 Discussion

Intrinsically disordered regions of proteins have thus far been difficult to characterize based on their sequences. In previous work, we found that evolutionary signatures extracted from sequences of IDRs and their orthologs contain information that can be used to associate them with biological functions. Here, we trained a statistical model using these evolutionary signatures with the aims of predicting protein function, learning the relative contribution of each IDR to that function (if applicable), and quantifying the contribution of different molecular features to that function. We demonstrated that the model can learn to classify proteins for several different biological functions, and that we can learn interpretable information about the sequence features contributing to these functions.

One of the important assumptions of our model is that each IDR is one “unit of function”, and that the model can learn how the protein function in question is split amongst the different IDRs in the protein. Although this is a reasonable assumption, we do not know if this is the case. In our dataset, the vast majority of IDRs get weighted equally with respect to the protein function in question. This could mean that the IDRs do not contain the signal for protein function, or that they equally contribute to the protein function. It could also reflect the possibility that in many cases, one IDR alone does not contain all the information for a particular function mapped at the protein level. It is interesting that in cases where the probability for functional prediction is high, the functions in question are IDR-associated. For example, in the case where the IDR acts a site for phosphorylation, or in the case where the IDR is a targeting signal, it is easy to imagine the IDR as a functional unit.

In relation to the point above, there are currently very few types of data, experimental or otherwise, that are mapped at the IDR level. Thus, it is challenging to associate IDRs with functions without using protein annotations as a proxy. Systematic studies of IDRs are needed to address this gap. In this study, we used an iteratively re-weighted logistic regression to learn the contribution of each IDR to the protein function in hopes of addressing this issue with the data that we have on hand.

Despite concerns about IDR boundaries and experimental data, it is overall very encouraging that it is possible to classify proteins according to their functions based on evolutionary information and their sequences. It will be important to continue this work in hopes of characterizing the many thousands of predicted disordered regions in eukaryotic proteomes.

4.5 Future Directions

There are many areas for future work following this analysis. Although preliminary results with randomized data indicate that the model is informative, we need to formally test that the model is not overfitting in the cases where we have confident predictions by testing the model on held-out data. Furthermore, in the cases where similar functions have been predicted, we can formally compare our model based on evolutionary signatures with those developed by others; one example is the model Mitofates, which predicts mitochondrial localization signals (Fukasawa et al., 2015). Following this, it will be very interesting to see what, if any, signals distinguish the mitochondrial inner membrane predicted IDRs from general mitochondrial IDRs. Although it is

known that beyond mitochondrial localization, there are other cleavage signals and/or interacting partners that aid localization of mitochondrial precursors to different parts of the organelle, it is not currently clear if there are sequence signatures that delineate these groups from each other.

4.6 Methods

4.6.1 Extraction of evolutionary signatures from predicted IDRs

In order to obtain evolutionary signatures from predicted IDRs, we performed a similar method to (Zarin *et al.*, Chapter 3.5.4 of this thesis). Briefly, for each set of orthologous IDRs (obtained from the Yeast Gene Order Browser (Byrne and Wolfe, 2005), we simulated 1000 sets of orthologous IDRs (using previously described methods (Nguyen Ba et al., 2014, 2012)), and quantified the difference in evolution of molecular features in the real set of orthologous IDRs to that of the simulated orthologous IDRs using a standard z-score (as described in Zarin *et al.*, Chapter 3.5.4 of this thesis). The evolutionary signature of each IDR is comprised of a vector of 164 z-scores (mean and log variance over evolution for each of the 82 molecular features listed in Appendix Table 2-1 of this thesis).

In previous work (Zarin *et al.*, Chapter 3 of this thesis), we calculated evolutionary signatures of function for rapidly evolving IDRs by constraining the evolution of short conserved segments. In order to use all of the available potentially functional information in IDRs, including conserved short linear motifs (SLiMs), we once again calculated evolutionary signatures in predicted IDRs across the yeast proteome, but did not constrain the evolution of these short segments.

4.6.2 Compilation of functions and phenotypes for prediction

We compiled a series of functional annotations, phenotypes, and datasets to predict with our model. These fell into four broad categories:

1. Gene Ontology annotations that we previously found to be strongly enriched in our unsupervised clustering analysis (Zarin *et al.*, Chapter 3 of this thesis), acquired from the Saccharomyces Genome Database (SGD) (Cherry et al., 2012)

2. Datasets that screened the *S.cerevisiae* proteome for membraneless organelle/reversible protein assembly formation under stress (Mitchell et al., 2012; Narayanaswamy et al., 2009; Wallace et al., 2015)

3. A dataset of gold-standard Cdc28 substrate predictions (Sharifpoor et al., 2011)

4. A set of publicly available *S.cerevisiae* phenotype annotations from SGD (Cherry et al., 2012)

From this list, we include several examples of functions that we can predict with reasonable sensitivity and specificity in Figure 4-1. In total we found 23 functions for which we could get accurate predictions. Overall, we tested the model on ~120 functions and phenotypes.

4.6.3 A statistical model that accounts for multiple IDRs in one protein

The model was implemented in R with the glmnet package (Friedman et al., 2010). A schematic of the model is provided in Appendix Figure 3-1. A description of the model is as follows:

Each dataset corresponding to functions or phenotypes is provided as a set of targets in a binary format, where Y=1 if the protein is associated with that target or phenotype, and Y=0 if the protein is not associated with that target or phenotype. n is the number of proteins.

$$Y = Y_1, Y_2, \ldots, Y_n$$

We use a model to account for the fact that there is not necessarily a 1:1 mapping between IDRs and protein annotations that we are trying to predict. We assume that for each function, there is one IDR that is contributing to that function more than the other(s), and thus predict the weight of each IDR in relation to the functional prediction. We use a hidden variable, X_{ij}, whose probability is the weight of the jth IDR for the ith protein for the function in question. The weight of each IDR starts simply as 1 divided by the number of IDRs in the protein, r. Z_{ij} is the vector of Z-scores (i.e. the "evolutionary signature") for the jth IDR of the ith protein. m is the number of features.

$$Z_{ij} = Z_{ij1}, Z_{ij2}, \ldots, Z_{ijm}$$

In the framework of a linear regression, where b is the vector of coefficients, the likelihood of this model is:

$$P(Y|Z,b) = \prod_{i=1}^{n} \sum_{j=1}^{r} P(X_{ij}) \left[\frac{1}{1+e^{-z_{ij}b}}\right]^{Y_i} \left[1 - \frac{1}{1+e^{-z_{ij}b}}\right]^{1-Y_i}$$

We maximize this likelihood using the E-M algorithm. To do this, we iteratively maximize the expected complete log likelihood $\langle log\ CL \rangle$, wherein we assume that the hidden variables are observed.

$$\langle log\ CL \rangle = \sum_{i=1}^{n} \sum_{j=1}^{r} \langle X_{ij} \rangle \left(Y_i \log \frac{1}{1+e^{-z_{ij}b}} + (1 - Y_i) \log \left(1 - \frac{1}{1+e^{-z_{ij}b}} \right) \right)$$

Inspection of the expected complete log likelihood reveals exactly one term for each IDR, weighted by $\langle X_{ij} \rangle$, the expected value of the hidden variable.

This model corresponds to an iteratively re-weighted logistic regression problem (Hastie et al., 2015), where the weights are the current estimates of the hidden variables $\langle X_{ij} \rangle$. Thus, we can maximize the LASSO penalized version of the expected complete log likelihood by adding an extra set of weights to the iteratively re-weighted least squares (IRLS) algorithm (Hastie et al. 2015), such that the weights are:

$$w_{ij} = \langle X_{ij} \rangle \frac{1}{1+e^{-z_{ij}b}} \left(1 - \frac{1}{1+e^{-z_{ij}b}} \right)$$

In practice, we use 5 iterations of IRLS.

In the E-step of the E-M algorithm, these expectations are calculated using Bayes' theorem as follows:

$$\langle X_{ij} \rangle = P\left(X_{ij} = 1 | Y_i, Z_{ij}, b\right) = \frac{P\left(X_{ij} = 1 | Z_{ij}, b\right) P\left(Y_i | X_{ij} = 1, Z_{ij}, b\right)}{P\left(Y_i | Z_{ij}, b\right)}$$

$$= \frac{Y_i P\left(Y_i = 1 | X_{ij} = 1, Z_{ij}, b\right) + (1 - Y_i)\left(1 - P\left(Y_i = 1 | X_{ij} = 1, Z_{ij}, b\right)\right)}{\sum_k Y_i P\left(Y_i = 1 | X_{ik} = 1, Z_{ij}, b\right) + (1 - Y_i)\left(1 - P\left(Y_i = 1 | X_{ik} = 1, Z_{ik}, b\right)\right)}$$

Note that we have assumed no prior knowledge about which IDR in a protein is most likely to be responsible for the function (uninformative priors), i.e., $P\left(X_{ij} = 1 | Z_{ij}, b\right)$ is constant over j.

In practice, we use 20 iterations of E-M.

For L1 regularization, we used a lambda value of 0.2.

4.6.4 Model assessment

To assess the model, we plotted ROC curves to display the false positive rate vs. true positive rate of our predictions. ROC curves were plotted with the ROCR package in R (Sing et al., 2009). Real predictions were compared to those obtained using randomized targets.

4.6.5 Clustering of t-scores

We used the Cluster 3.0 program (de Hoon et al., 2004) to cluster the coefficient t-scores from the molecular features vectors (n=164) as well as the functions that we predicted. We used hierarchical clustering with correlation distance and average linkage.

4.6.6 *In vivo* elimination of molecular features comprising evolutionary signature in mitochondrial IDR

The budding yeast strains were constructed, grown, and imaged in identical conditions as in Zarin et al., Chapter 3. The Cox15 IDR was replaced with a simulated IDR (sequence: MLLRNVESSKPEAKLITRASYAVPRKMNNSYLGDNTLNNLVLKKSYLLAVPRKIPTIPAS LPQIRDKD) using the previously described *Delitto Perfetto* method (Storici et al., 2001).

4.6.7 Contributions

TZ and AMM designed research. TZ, BS, and AMM performed research. TZ and AMM analyzed data.

Chapter 5
Conclusions and future directions

5 Conclusions and future directions

5.1 Summary of conclusions

Intrinsically disordered regions (IDRs) are varied, widespread, and have important implications for human disease. Because they have been dismissed as anomalies since their discovery, the constraints and functions of these regions are only now starting to be discovered and synthesized to form a cohesive view. Although it has been appreciated that IDRs are highly diverged in evolution compared to ordered regions, the consequences of this divergence have not been explored. Thus, the first aim of my thesis has been to understand the functional consequences of highly diverged IDRs. Compared to ordered regions in the same protein, the sequence homology of orthologous IDRs can resemble that of unrelated, randomly scrambled sequences. Does this mean that these orthologous regions have diverged in function, or that they are evolving neutrally as "junk"? Could there be conservation of function despite this rapid divergence? In chapter 2 of my thesis, I started exploring these questions, and discovered that sequence divergence does not necessarily imply functional divergence or absence of constraint in IDRs. Despite the rapid evolution of amino acid sequences in IDRs, bulk molecular features in these regions can contribute to aggregate, or quantitative phenotypes, which can in turn be under stabilizing selection. This relaxed constraint on specific amino acids can give the appearance of rapid divergence in orthologs, even when underlying functions or phenotypes are conserved. In chapter 3, I asked if this is a property that is specific to the IDR I studied in chapter 2, or whether this is a general property of IDRs proteome-wide. I found that most IDRs in the yeast proteome contain not just one, but multiple molecular features that are preserved through evolution despite high divergence in their primary amino acid sequences. I also discovered that IDRs in the proteome share sets of preserved molecular features, and that groups of IDRs that share these "evolutionary signatures" are associated with a wide variety of protein functions. In chapter 4, I applied a statistical model to these evolutionary signatures, and found that they contain interpretable information that can be used to predict different protein functions.

5.2 Discussion and future directions

5.2.1 Order and disorder are differentially constrained

One of the main ideas that has laid the foundation for (and been corroborated by) my thesis work is that ordered regions and disordered regions are under vastly different evolutionary constraints (Afanasyeva et al., 2018; Brown et al., 2010, 2002; Khan et al., 2015; Light et al., 2013; Nilsson et al., 2011; Tóth-Petróczy and Tawfik, 2013). In future work, it will be important to understand how we can modify existing methods and create new methods that take this into account. For example, current state-of-the-art approaches for finding constraint in protein sequences rely on multiple sequence alignments, and best practices require the user to ignore or cut out parts of the alignment that contain deletions or insertions (e.g. Phylogenetic Analysis using Maximum Likelihood [PAML] (Yang, 2007)). This is reminiscent of the way IDRs have been handled since the advent of X-ray crystallography, whereby any flexible parts of the protein must to be cut out in order for there to be any hope for crystallization. Similar to how NMR (Nuclear Magnetic Resonance) spectroscopy has allowed researchers to study flexible proteins in solution (Forman-Kay and Mittag, 2013), we need a methodological shift to prevent us from "cutting out" one third of the eukaryotic proteome in our functional genomics analyses. Shifting to alignment-free methods that are not based on sequence similarity of amino acids will be important for achieving this goal.

Related to this point, an outstanding question about the different constraints in ordered and disordered regions is whether this should be a dichotomy at all. Should we be considering a continuum of constraints to reflect the continuum of structures (or lack thereof) employed by disordered regions and ordered regions alike? How much of our discretization of ordered and disordered regions is based on the binary classifiers that predict disordered regions for us? These are questions to think about as we continue exploring the relationship between sequence, structure, function, and evolution of these regions.

5.2.2 Convergent evolution on a massive scale

One of the most interesting hypotheses that came out of our proteome-wide study on highly diverged disordered regions (Ch. 3 of this thesis) is the idea that IDRs have undergone, and continue to undergo, convergent evolution on a large scale. This is in contrast to ordered

domains, which are thought to mainly evolve by duplication, divergence, and negative purifying selection (Taylor and Raes, 2004). The hypothesis for convergent evolution in IDRs is supported by our observation that tens to hundreds of IDRs in the yeast proteome have similarly constrained molecular features, or evolutionary signatures, despite their lack of homology. Especially in the case of mitochondrial proteins (and their disordered targeting signals), where we know their evolutionary history (i.e., that mitochondrial genes were transferred to the nuclear genome over time (Adams and Palmer, 2003)), this is not hard to imagine. Interestingly, coiled-coils are a type of protein structure with biased amino acid composition that have been shown to emerge through convergent evolution (Mistry et al., 2013; Rackham et al., 2010). Intrinsically disordered regions with biased amino acid compositions could have evolved in the same way. It will be interesting to understand the degree of convergent evolution in IDRs, and understand how molecular features in these regions arise and are preserved through evolutionary time, especially with the availability of new tools (Hu et al., 2019).

5.2.3 Understanding the effect of mutations in IDRs

Another important question, and one that is implicit in all attempts to map sequence and function, is the effect of mutations on IDR sequences. Besides the inherent interest in this question, there is also an immediately practical application, as an estimated 20% of all disease mutations occur in IDRs (Vacic et al., 2007). Although it is not hard to imagine a loss of function mutation in IDRs (given the importance of short linear motifs (Nguyen Ba et al., 2012; Tompa et al., 2014) in some IDRs), it is also clear that in many cases, single mutations can have much less prominent effects. As previously discussed, most IDRs are highly diverged, and likely evolving under stabilizing selection where single mutations would not affect fitness outcomes (Landry et al., 2014). In this light, it is interesting to note that there has also been a recent appreciation for the role of gain-of-function mutations in IDRs. For example, it was recently discovered that the emergence of a missense-aided di-leucine motif in cytosolic IDRs of transmembrane proteins can cause disease by increasing clathrin-binding (Li and Babu, 2018; Meyer et al., 2018). Understanding the effects of mutations in IDRs, and how these differ from ordered regions, will be an exciting area for future work.

5.2.4 IDR data collection and storage for validation

Related to understanding the effect of mutations in IDRs, there is a great need to perform and catalogue experimental studies on IDRs. Firstly, there is a need to catalogue and curate low-throughput experimental studies in IDRs, as there is currently no consistent terminology or method to store these. Many studies have been done on IDRs without explicit acknowledgement of them as IDRs, meaning that tens to hundreds of valuable experimental results are untapped for our general understanding, as well as for use in machine learning analyses. In addition, there is a need to expand the scale of experiments in IDRs. For example, although there have been some major advances in systematically probing single IDRs (Bolognesi et al., 2019; Ravarani et al., 2018), there is still a need for efforts to survey IDRs on a proteome-wide scale. For example, a resource akin to the budding yeast deletion collection (Giaever et al., 2002) aided with precise, high throughput genome engineering (Li et al., 2011; Roy et al., 2018) would allow for large gains in our understanding of how IDRs (and IDR deletions or mutations) impact the genome on a large scale.

5.2.5 Tools for classification of IDRs

Through the work presented in this thesis and elsewhere, it is becoming increasingly clear that IDR sequences contain functional information, and that this information can be gleaned by evolutionary analysis. This has led us to hypothesize that we can use this information to make functional predictions about IDRs based on their sequences alone. Although the methods that we use to make functional predictions about IDRs in Chapter 2 and Chapter 3 are generalizable in principle, they rely on evolutionary simulations and a confident set of orthologs, which makes them cumbersome and technically challenging in practice. In future work, it will be important to quantify the contribution of evolution as a feature when using IDR sequences to predict function. Furthermore, the current method where each evolutionary signature is comprised of the difference between real orthologous IDRs and 1000 simulated orthologous IDRs could be tested to see if there can be a reduction in the scale of simulations. Finally, developing more robust evolutionary models for quantitative traits as applied to protein sequences will be important in making predictions about functional features of IDR sequences.

Appendix 1
Supplementary material for Chapter 2

This is an author-produced PDF of an article accepted for publication in Proceedings of the National Academy of Sciences of the United States of America following peer review. The version of record:

Selection maintains signaling function of a highly diverged intrinsically disordered region

Proc Natl Acad Sci U S A. 2017 Feb 21;114(8):E1450-E1459. doi: 10.1073/pnas.1614787114. Epub 2017 Feb 6.

Zarin, T.[1], Tsai, C.[2], Nguyen Ba, A.N.[3], Moses, A.M.[1,2]

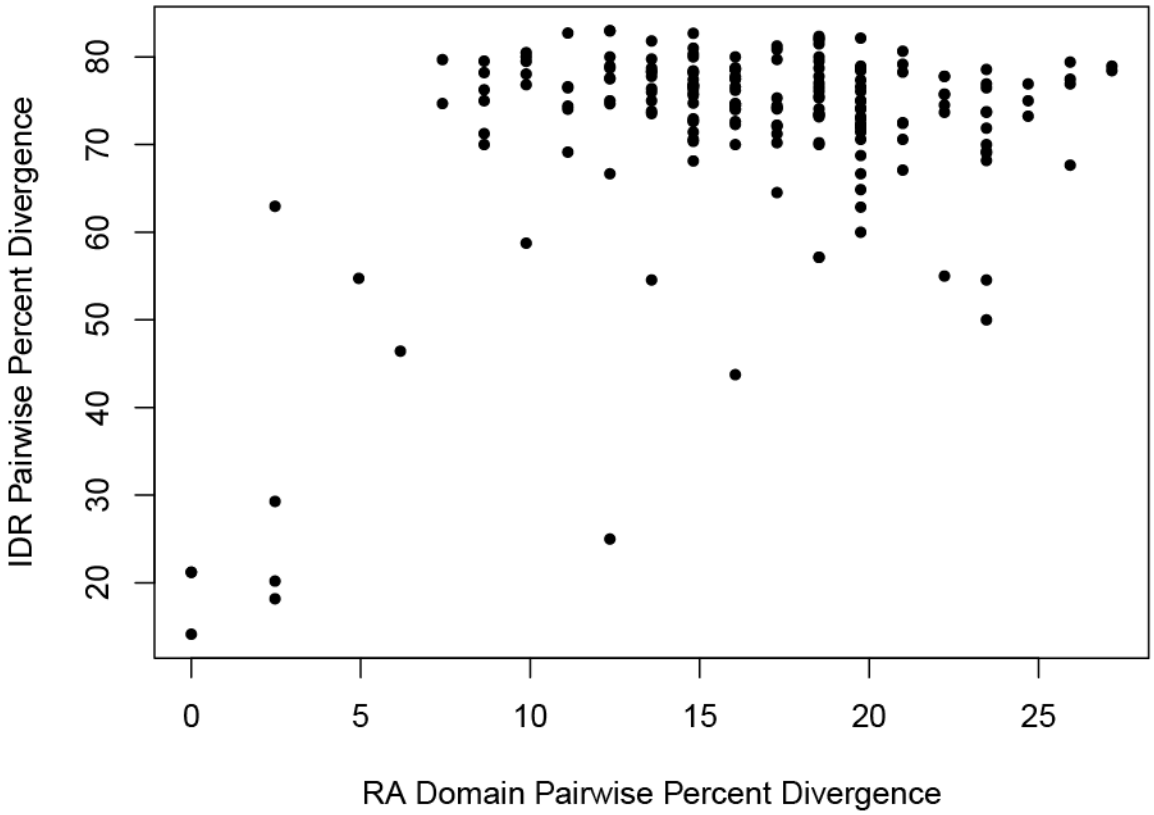
is available online at: https://www.pnas.org/content/114/8/E1450.long

4. Department of Cell and Systems Biology, University of Toronto, 25 Harbord St., Toronto, ON, Canada, M5S 3G5
5. Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks St., Toronto, ON, Canada, M5S 3B2
6. FAS Center for Systems Biology, Harvard University, 52 Oxford Street, Cambridge, MA, USA, 02138

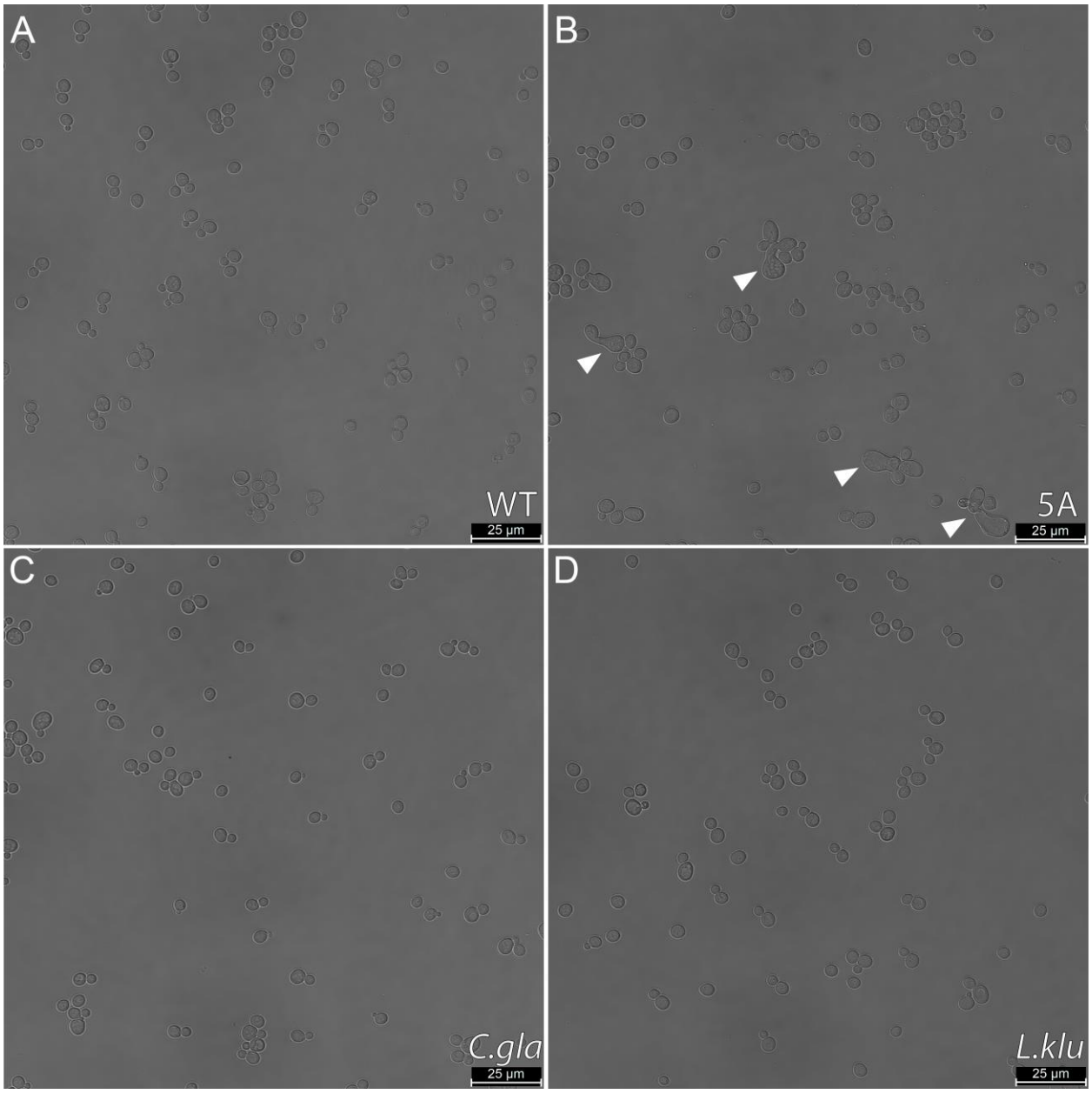
Appendix Table 1-1. List of strains used in this study

Strain	Genotype	Source	Used to assay:
DMA580	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22:KanMX4	(Giaever et al., 2002)	N/A
YTZ3	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0	This study	Morphology
YTZ44	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}A,T^{244}A,S^{248}A	This study	Morphology
YTZ45	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Cgla IDR	This study	Morphology
YTZ49	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Lklu IDR	This study	Morphology
YTZ8R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP pCAN1::pRPL39-ymCherry-LEU2	This study	Hog1 signaling dynamics
YTZ8B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP	This study	Hog1 signaling dynamics
YTZ24R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50IDR-S^{155}A,S^{196}A S^{202}AT^{244}A,S^{248}A pCAN1::pRPL39-ymCherry-LEU2	This study	Hog1 signaling dynamics
YTZ24B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}AT^{244}A,S^{248}A pCAN1::pRPL39-mTagBFP2-LEU2	This study	Hog1 signaling dynamics
YTZ25R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50 Cgla IDR pCAN1::pRPL39-ymCherry-LEU2	This study	Hog1 signaling dynamics
YTZ25B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50 Cgla IDR pCAN1::pRPL39-mTagBFP2-LEU2	This study	Hog1 signaling dynamics
YTZ29R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50 lKlu IDR pCAN1::pRPL39-ymCherry-LEU2	This study	Hog1 signaling dynamics
YTZ29B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HOG1-Cterm::yemGFP STE50IDR::STE50 lKlu IDR pCAN1::pRPL39-mTagBFP2-LEU2	This study	Hog1 signaling dynamics
YTZ60	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ62	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ62EE	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR- $SP^{155\text{-}156}$EE HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ63	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}A HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ63EE	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR- $SP^{155\text{-}156}$EE,$SP^{196\text{-}197}$EE,$SP^{202\text{-}203}$EE HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ64	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3	This	Fus3 basal

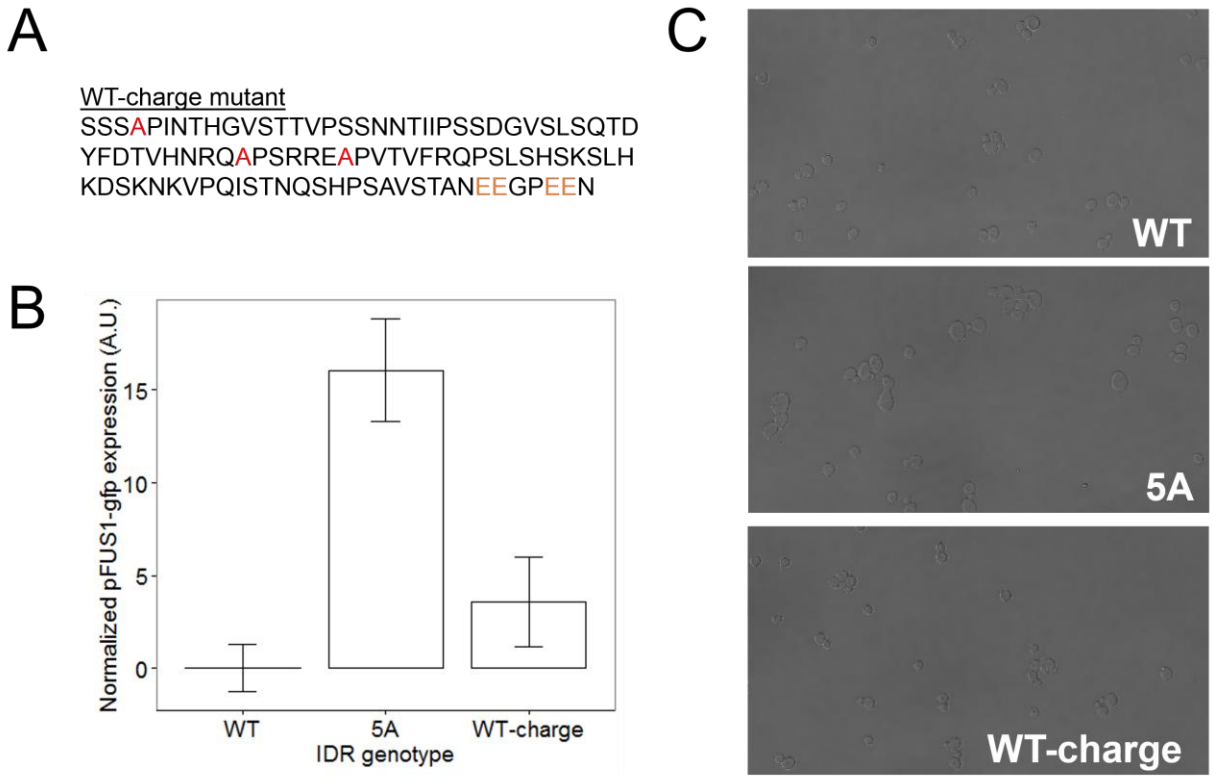
	SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}AT^{244}A,S^{248}A HO::pFUS1-yemGFP-klURA3	study	signaling
YTZ64EE	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR- $SP^{155-156}$EE,$SP^{196-197}$EE,$SP^{202-203}$EE,$TP^{244-245}$EE,$SP^{248-249}$EE HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ65	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Cgla IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ66	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Zrou IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ67	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Lwal IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ68	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Lthe IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ69	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Lklu IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ70	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Klac IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ71	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Scer-Lklu IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ72	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Scer-Zrou IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ73	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::Klac-Scer IDR HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling
YTZ3R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 pCAN1::pRPL39-ymCherry-LEU2	This study	Fitness
YTZ3B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 pCAN1::pRPL39-mTagBFP2-LEU2	This study	Fitness
YTZ44R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}AT^{244}A,S^{248}A pCAN1::pRPL39-ymCherry-LEU2	This study	Fitness
YTZ44B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}AT^{244}A,S^{248}A pCAN1::pRPL39-mTagBFP2-LEU2	This study	Fitness
YTZ45R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Cgla IDR pCAN1::pRPL39-ymCherry-LEU2	This study	Fitness
YTZ45B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Cgla IDR pCAN1::pRPL39-mTagBFP2-LEU2	This study	Fitness
YTZ49R	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Lklu IDR pCAN1::pRPL39-ymCherry-LEU2	This study	Fitness
YTZ49B	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50 Lklu IDR pCAN1::pRPL39-mTagBFP2-LEU2	This study	Fitness
YTZ54	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}A,$TP^{244-245}$EE,$SP^{248-249}$EE	This study	Morphology
YTZ74	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 STE50IDR::STE50IDR-S^{155}A,S^{196}A,S^{202}A,$TP^{244-245}$EE,$SP^{248-249}$EE HO::pFUS1-yemGFP-klURA3	This study	Fus3 basal signaling



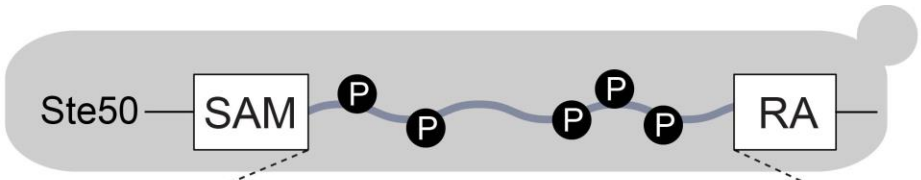
Appendix Figure 1-1. The pairwise percent divergence of the Ste50 IDR saturates with divergence time (as measured by RA domain divergence). Each point represents a species pair, where the pairwise percent divergence (i.e. 100 - pairwise percent identity) is plotted for the Ste50 RA domain versus the IDR.



Appendix Figure 1-2. Example brightfield (BF) micrographs from each assayed strain. Arrows indicate example cells with abnormal morphology.



Appendix Figure 1-3. An unphosphorylatable mutant S.cer IDR with identical basal net charge to the wildtype ("WT-charge" mutant) recapitulates wildtype morphology and pFUS1 expression. a) Amino acid sequence of the WT-charge mutant IDR, with mutated phosphorylation sites highlighted in red when mutated to alanine, and orange when mutated to glutamic acid. B) Mean pFUS1-gfp expression for wildtype (WT), 5A mutant (5A), and WT-charge mutant IDRs. Error bars represent 1.96 s.e. of 3 biological replicates. C) Example brightfield (BF) micrographs from WT, 5A, and WT-charge mutant IDR strains.



Ste50 IDR Δ

>S.cer IDR WT (a.a. 152-250)
SSSSPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQSPSRRESPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

>S.cer IDR 1A
SSSAPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQSPSRRESPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

>S.cer IDR 1EE
SSSEEINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQSPSRRESPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

>S.cer IDR 3A
SSSAPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQAPSRREAPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

>S.cer IDR 3EE
SSSEEINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQEESRREEEVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

>S.cer IDR 5A
SSSAPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQAPSRREAPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANAPGPAPN

>S.cer IDR 5EE
SSSEEINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTVHNRQEESRREEEVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANEEGPEEN

>C.gla IDR (a.a. 137-223)
KTAEHGMSSPQLTPSTASSKPSGARTDYFDHGQLRQSPRAKEVRPTLLYNKSTPNSNSNSNSKYHDQIIHANSSLTIPNIKSNLSPI

>Z.rou IDR (a.a. 141-221)
NSSPKPPSQQQFAYQQQAAGSDYFDTTHHGGEFATGQATPHGISRKMSGHSVKPANPSRSTSSHLVNETLAPSQASQQVSS

>L.wal IDR (a.a. 133-212)
ATPVYPQLQMNQQQQSQQHQPPQHDYFEGHKALTPGSPNTGTLRQVPNRSQSSGASQQTFGIGGLTTPGAQHPPSSGAGN

>L.the IDR (a.a. 133-211)
PHASFPPIQLNTQHQQPQHDYFEGQKIVTPGSSNNTFVRQAPARSHSNSTSQQTLGIGGSTTSTPGVAPQQTPSGGSGT

>L.klu IDR (a.a. 151-229)
NGNINTTSPSFGTQPQPTGDYFDQQKHPLIINGSSGTTNNLGSNGSKSSVLRSGSSTASVPALASSNSFGGTPGGNSTN

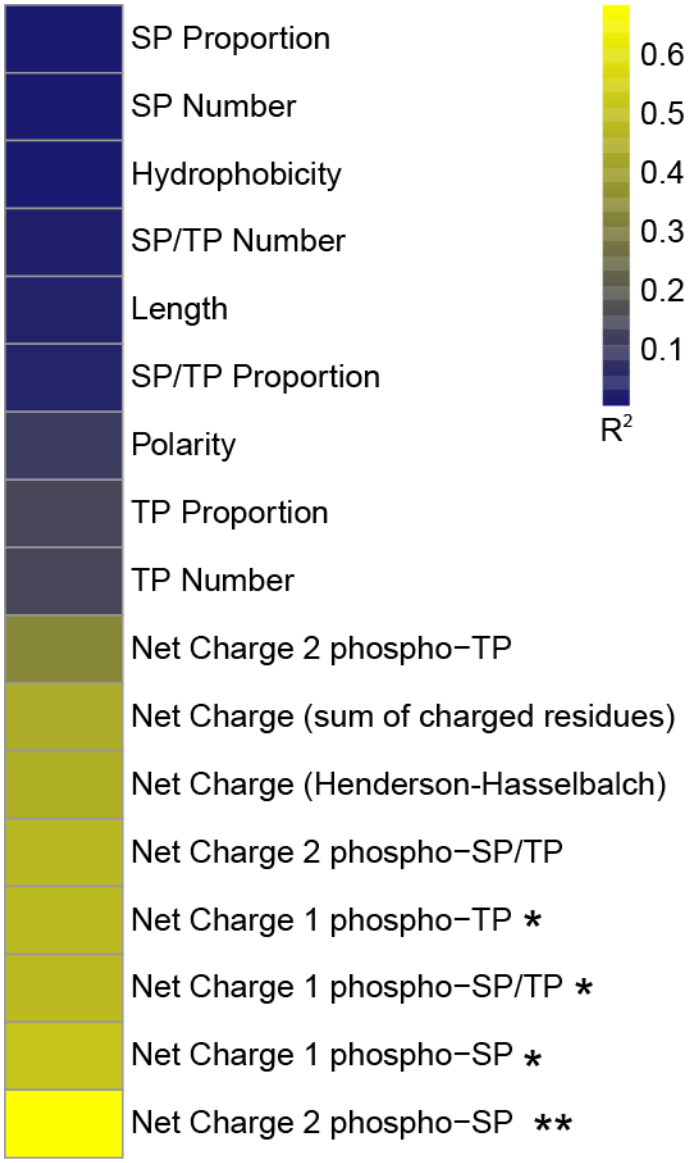
>K.lac IDR (a.a. 120-203)
SGSGSGSSQSSQQLQHQPSSQDYFDRAPHHSSHTPSSPKYRSNSNPGLPPQSSSTRNVSSSHIATTPGSANTPGGAGVSSAPSS

>S.cer (a.a. 152-186)-L.klu (a.a.170-229) chimaeric IDR
SSSSPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDQQKHPLIINGSSGTTNNLGSNGSKSSVLRSGSSTASVPALASSNSFGGTPGGNSTN

>S.cer (a.a. 152-186)- Z.rou (a.a. 162-221) chimaeric IDR
SSSSPINTHGVSTTVPSSNNTIIPSSDGVSLSQTDYFDTTHHGGEFATGQATPHGISRKMSGHSVKPANPSRSTSSHLVNETLAPSQASQQVSS

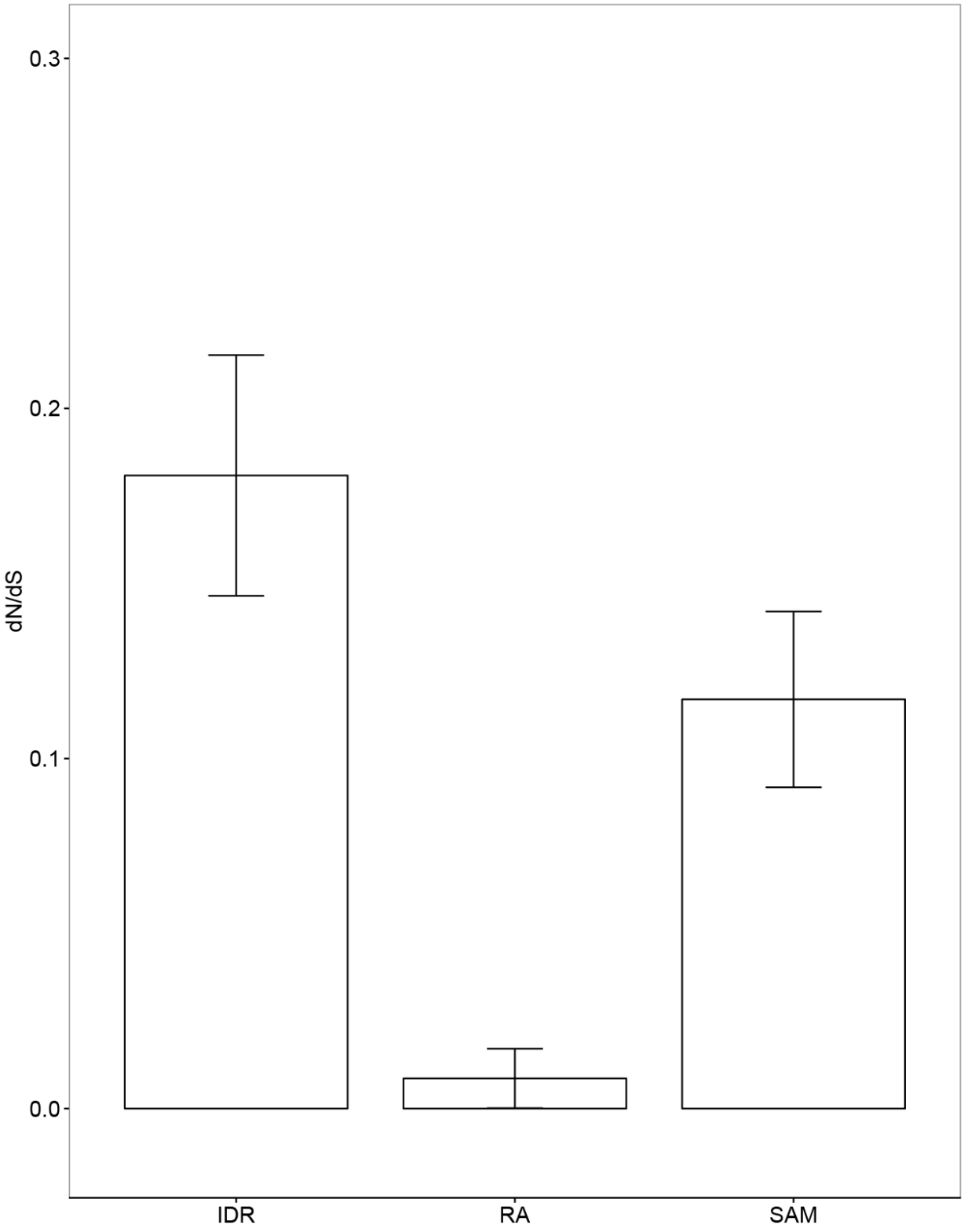
>K.lac (a.a. 120-144)-S.cer (a.a.190-250) chimaeric IDR
SGSGSGSSQSSQQLQHQPSSQDYFDTVHNRQSPSRRESPVTVFRQPSLSHSKSLHKDSKNKVPQISTNQSHPSAVSTANTPGPSPN

Appendix Figure 1-4. List of engineered Ste50 IDRs used in correlation study (Figure 2-4).



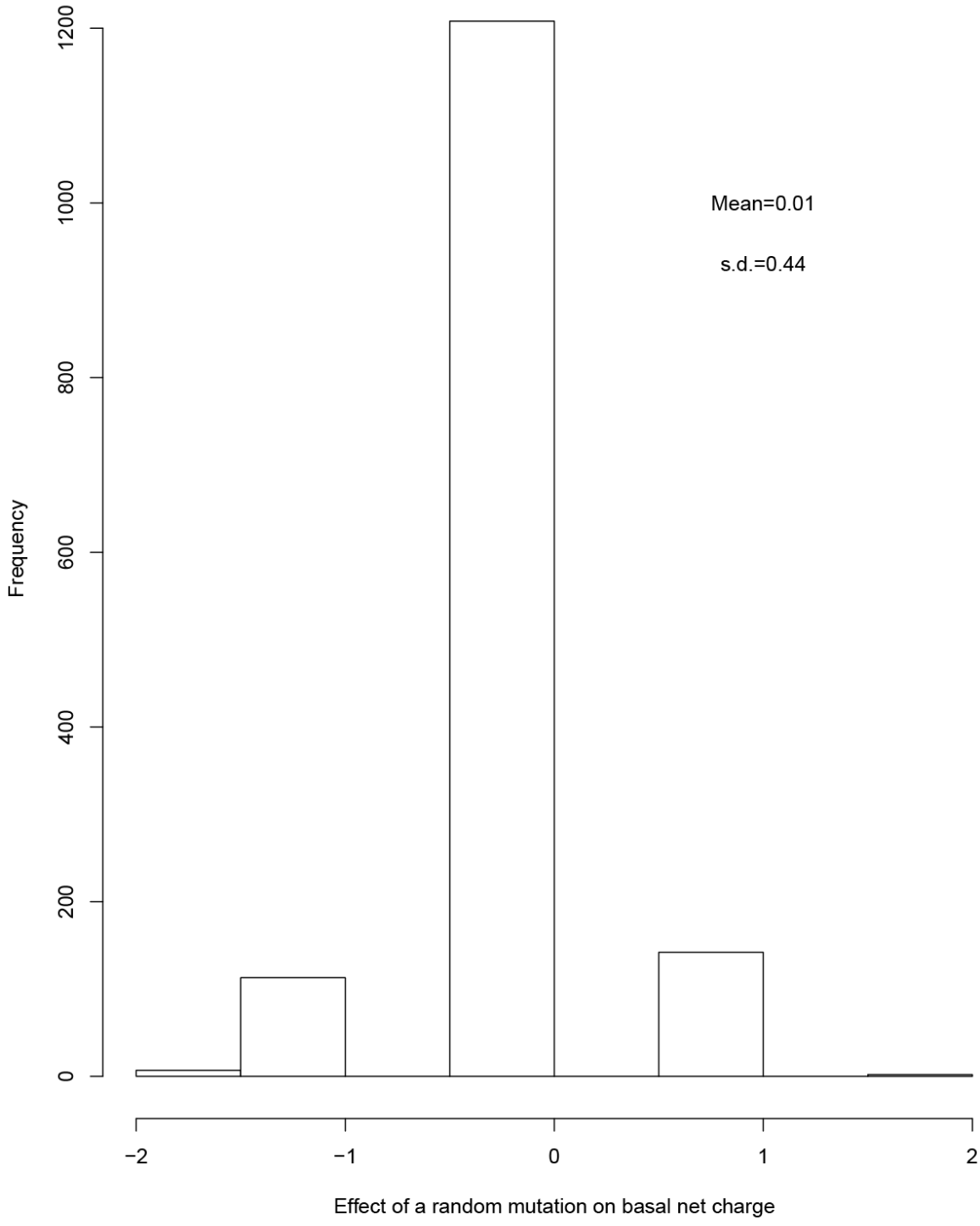
Appendix Figure 1-5. Heat map of sequence features correlated with functional output (pFUS1-GFP expression) of the Ste50 IDRs tested. * indicates Bonferroni-corrected P<0.05, ** indicates Bonferroni-corrected P<0.01. In order of appearance in the figure: SP proportion refers to the number of "SP" phosphorylation consensus motifs divided by the total number of amino acids in the IDR, SP number refers to number of "SP" phosphorylation sites regardless of IDR length, Hydrophobicity refers to the GRAVY (grand average of hydropathy) index score of each IDR, SP/TP number is the number of "SP" or "TP" phosphorylation consensus motifs in the IDR, length is the total number of amino acid residues in the IDR, SP/TP proportion is the number of "SP" or "TP" phosphorylation consensus motifs divided by the total number of amino acids in the IDR, polarity is the average polarity score of the IDR, TP proportion is the number of "TP"

consensus phosphorylation sites divided by the total number of amino acids in the IDR, TP number is the total number of "TP" consensus phosphorylation sites in the IDR, Net Charge (sum of charged residues) is the number of positively charged residues in the IDR minus the number of negatively charged residues in the IDR, Net Charge (Henderson-Hasselbach) is the net charge of the IDR as calculated by the Henderson-Hasselbalch equation at pH 7 and pKa determined by the Lehninger scale, Net Charge 1 or 2 phospho-SP or TP or SP/TP refers to the Net Charge (sum of charged residues) in the IDR with the potential for basal phosphorylation of 1 or 2 "SP", "TP", or "SP/TP" phosphorylation consensus motifs (see methods for more details on calculations).



Appendix Figure 1-6. dN/dS values compared between the Ste50 IDR, the RA domain, and the SAM domain. The dN/dS value for the IDR is higher (0.18) compared to the SAM (0.12) and

RA (0.01) domains. However, much of the sequence variation in IDRs comes from the high rates of non-frameshifting insertions and deletions that we find in these regions (9-13), which would not be captured in the dN/dS analysis. Therefore, dN/dS is likely to overestimate the constraint in disordered regions. Error bars represent 1.96 s.e.



Appendix Figure 1-7. Distribution of effects of a random nucleotide mutation on basal net charge. N=1472.

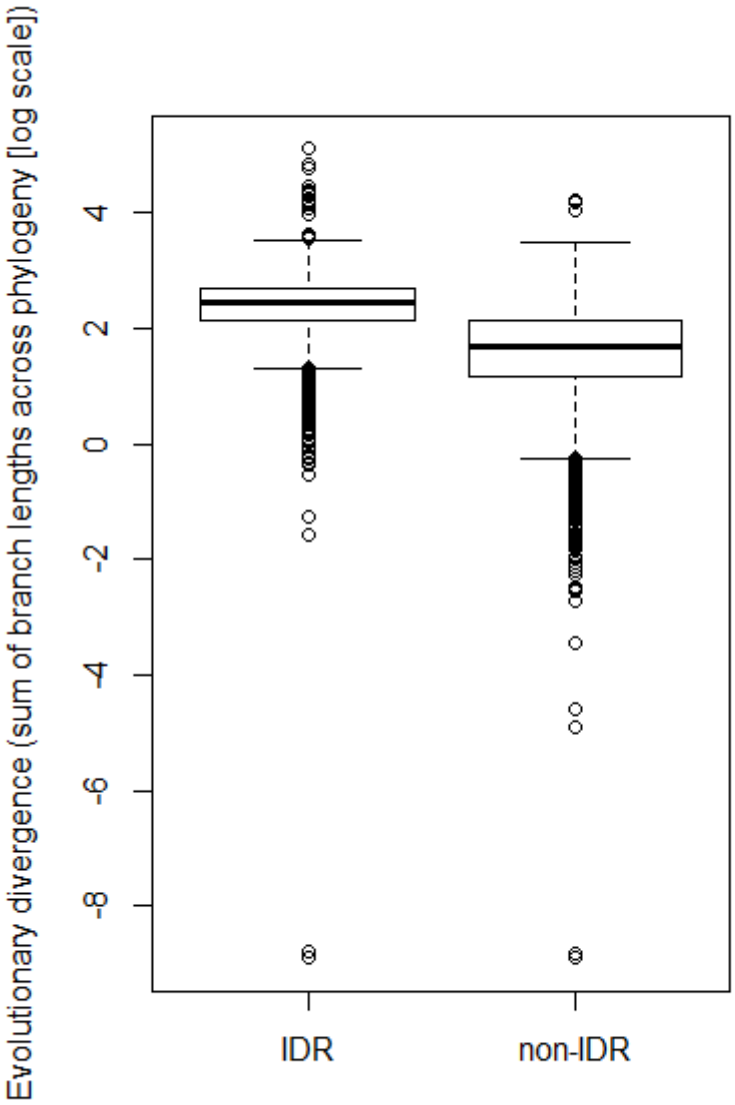
Appendix 2
Supplementary material for Chapter 3

This work has been submitted as: Proteome-wide signatures of function in highly diverged intrinsically disordered regions.

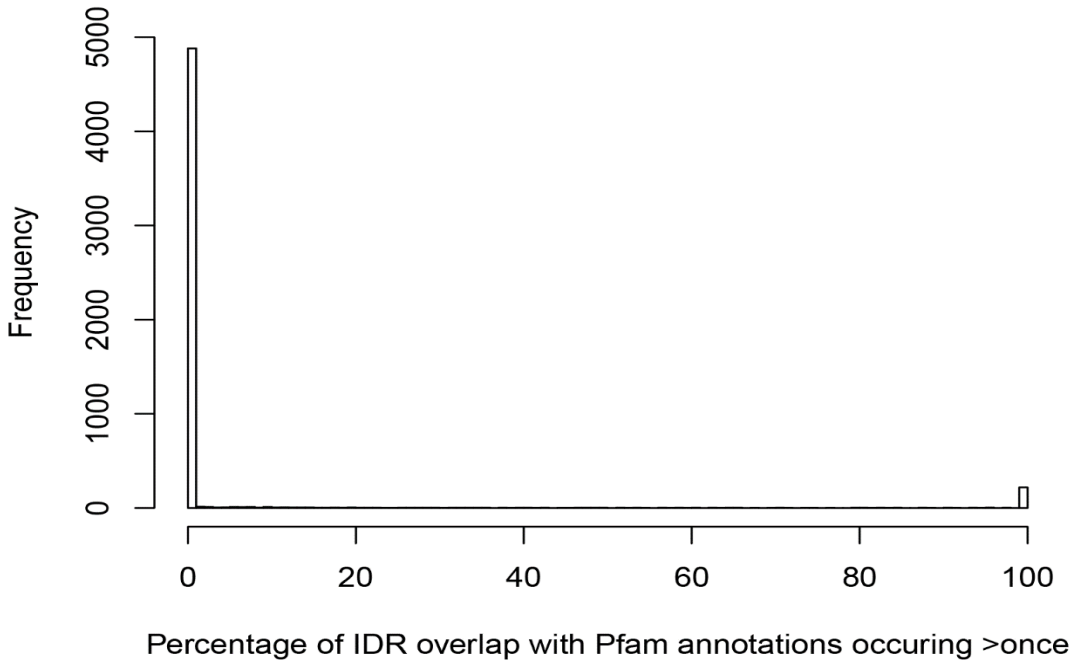
Submitted to eLife.

Taraneh Zarin[1], Bob Strome[1], Alex N Nguyen Ba[2], Simon Alberti[3,4], Julie D Forman-Kay[5,6], Alan M Moses[1,7,8]

1. Department of Cell and Systems Biology, University of Toronto, Toronto, Canada
2. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge MA 02138
3. Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
4. Technische Universität Dresden, Center for Molecular and Cellular Bioengineering, Biotechnology Center, Dresden, Germany
5. Program in Molecular Medicine, Hospital for Sick Children, Toronto, Canada
6. Department of Biochemistry, University of Toronto, Toronto, Canada
7. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada
8. Department of Computer Science, University of Toronto, Toronto, Canada



Appendix Figure 2-1. Predicted IDRs in the *S.cerevisiae* proteome ("IDR") are more highly diverged compared to regions that are not predicted to be disordered ("non-IDR") ($p<2.2x10^{-16}$, Wilcoxon test). Boxplot boxes represent the 25^{th}-75^{th} percentile of the data, the black line represents the median, and whiskers represent 1.5*the interquartile range. Outliers fall outside the 1.5*interquartile range, and are represented by unfilled circles



Appendix Figure 2-2. The vast majority of predicted IDRs in the *S.cerevisiae* proteome do not overlap with Pfam domains. N=5351 IDRs.

Appendix Table 2-1. Molecular features that have been shown or are hypothesized to be important in IDRs. All motif features are calculated as the fraction of motifs in the IDR normalized to the proteome-wide average. Some motif descriptions taken from Eukaryotic Linear Motif (ELM) resource (Dinkel et al., 2016) – refer to the ELM website for more details: http://elm.eu.org.

	ID	Name	Regular expression (regex)	Type	Source	Description	Reference
1	AA_S	S content	S	Amino acid content	NA	Fraction of S residues	(Haynes et al., 2006)
2	AA_P	P content	P	Amino acid content	NA	Fraction of P residues	(Marsh and Forman-Kay, 2010; Neduva and Russell, 2005; Simon and Hancock, 2009)
3	AA_T	T content	T	Amino acid content	NA	Fraction of T residues	Reviewed in (Van Der Lee et al., 2014)
4	AA_A	A content	A	Amino acid content	NA	Fraction of A residues	(Perez et al., 2014)
5	AA_H	H content	H	Amino acid content	NA	Fraction of H residues	(Marsh and Forman-Kay, 2010)
6	AA_Q	Q	Q	Amino acid	NA	Fraction of Q residues	(Alberti et al., 2009;

		content		content			Halfmann et al., 2011)
7	AA_N	N content	N	Amino acid content	NA	Fraction of N residues	(Alberti et al., 2009; Halfmann et al., 2011)
8	AA_G	G content	G	Amino acid content	NA	Fraction of G residues	(Elbaum-Garfinkle et al., 2015)
9	kappa	Kappa	NA	Charge properties	localCIDER	Measure of separation between positively versus negatively charged residues	(Das and Pappu, 2013; Holehouse et al., 2017)
10	omega	Omega	NA	Charge properties	localCIDER	Measure of separation between charged residues and prolines versus all other residues	(Holehouse et al., 2017; Martin et al., 2016)
11	FCR	Fraction of charged residues	NA	Charge properties	localCIDER	FCR: basic fraction + acidic fraction	(Holehouse et al., 2017; Mao et al., 2013)
12	NCPR	Net charge per residue	NA	Charge properties	localCIDER	NCPR: basic fraction - acidic fraction	(Holehouse et al., 2017; Mao et al., 2013, 2010)
13	net_charge	net charge	NA	Charge properties	Literature /localCIDER	Net charge (# [RK] - # [DE])	(Daughdrill et al., 2007; Strickfaden et al., 2007; Zarin et al.,

							2017)
14	net_charge_P	net charge with phosphorylation of [ST]P consensus sites	NA	Charge properties	Literature	Net charge as influenced by phosphorylation of consensus sites	(Strickfaden et al., 2007; Zarin et al., 2017)
15	SCD	Sequence charge decoration	NA	Charge properties	Literature	Measure of separation between positively versus negatively charged residues	(Sawle and Ghosh, 2015)
16	RK_ratio	R/K ratio	NA	Charge properties	Literature	Ratio of arginine to lysine residues (#R + 1) /(#K + 1)	(Vernon et al., 2018)
17	ED_ratio	E/D ratio	NA	Charge properties	NA	Ratio of glutamic acid to aspartic acid residues (#E + 1)/(#D + 1)	NA
18	CLV_Separin_Fungi	Separase cleavage motif	S[IVLMH]E[IVPFMLYAQR]GR.	Motifs	ELM	Separase cleavage site, best known in sister chromatid separation. Also involved in stabilizing the anaphase spindle and centriole disengagement.	(Dinkel et al., 2016)
19	DEG_APCC_KENBOX_2	APCC-binding Destruction motif	.KEN.	Motifs	ELM	Motif conserving the exact sequence KEN that binds to the APC/C subunit Cdh1 causing the protein to be targeted for 26S proteasome mediated degradation.	(Dinkel et al., 2016)

20	DEG_APCC_TPR_1	APCC_TPR-docking motif	.[ILM]R	Motifs	ELM	This short C-terminal motif is present in co-activators, the Doc1/APC10 subunit and some substrates of the APC/C and mediates direct binding to TPR-containing APC/C core subunits.	(Dinkel et al., 2016)
21	DOC_CKS1_1	Cks1 ligand	[MPVLIFWYQ].(T)P..	Motifs	ELM	Phospho-dependent motif that mediates docking of CDK substrates and regulators to cyclin-CDK-bound Cks1.	(Dinkel et al., 2016)
22	DOC_MAPK_DCC_7	MAPK docking motif	[RK].{2,4}[LIVP]P.[LIV].[LIVMF]|[RK].{2,4}[LIVP].P[LIV].[LIVMF]	Motifs	ELM	A kinase docking motif mediating interaction towards the ERK1/2 and p38 subfamilies of MAP kinases	(Dinkel et al., 2016)
23	DOC_MAPK_gen_1	MAPK docking motif	[KR]{0,2}[KR].{0,2}[KR].{2,4}[ILVM].[ILVF]	Motifs	ELM	MAPK interacting molecules (e.g. MAPKKs, substrates, phosphatases) carry docking Motifs that help to regulate specific interaction in the MAPK cascade. The classic Motifs approximates (R/K)xxxx#x# where # is a hydrophobic residue.	(Dinkel et al., 2016)
24	DOC_MAPK_HePTP_8	MAPK docking motif	([LIV][^P][^P][RK]....[LIVMP].[LIV].[LIVMF])|([LIV][^P][^P][RK][RK]G.{4,7}[LIVMP].[LIV].[LIVMF])	Motifs	ELM	A kinase docking motif that interacts with the ERK1/2 and p38 subfamilies of MAP kinases.	(Dinkel et al., 2016)

25	DOC_PP1_RVXF_1	PP1-docking motif RVXF	..[RK].{0,1}[VIL][^P][FW].	Motifs	ELM	Protein phosphatase 1 catalytic subunit (PP1c) interacting Motifs binds targeting proteins that dock to the substrate for dephosphorylation. The motif defined is [RK]{0,1}[VI][^P][FW].	(Dinkel et al., 2016)
26	DOC_PP2B_PxIxI_1	Calcineurin (PP2B)-docking motif PxIxI	.P[^P]I[^P][IV][^P]	Motifs	ELM	Calcineurin substrate docking site, leads to the effective dephosphorylation of serine/threonine phosphorylation sites.	(Dinkel et al., 2016)
27	LIG_APCC_Cbox_2	APC/C_Apc2-docking motif	DR[YFH][ILFVM][PA]..	Motifs	ELM	Motifs in APC/C co-activators that mediates binding to the APC/C core, possibly the catalytic Apc2 subunit. This second variant defines the motif in APC/C co-activators from TAXON:4751 and TAXON:554915.	(Dinkel et al., 2016)
28	LIG_AP_GAE_1	Gamma-adaptin ear interaction motif	[DE][DES][DEGAS]F[SGAD][DEAP][LVIMFD]	Motifs	ELM	The acidic Phe motif mediates the interaction between a set of accessory proteins and the gamma-ear domain (GAE) of GGAs and AP-1. Proposed roles: in clathrin localization and assembly on TGN/endosome membranes and in traffic between the TGN and endosome.	(Dinkel et al., 2016)
29	LIG_CaM_	Helical calmodulin binding	[ACLIVTM][^P][^P][ILVMFCT]Q[^P][^	Motifs	ELM	Helical peptide motif responsible for Ca2+-independent binding of the	(Dinkel et al., 2016)

	IQ_9	motif	P][^P][RK][^P]{4,5}[RKQ][^P][^P]			CaM . The motif is manly characterized by a hydrophobic residue at position 1, a highly conserved Gln at position 2, basic charges at positions 6 and 11, and a variable Gly at position 7	
30	LIG_EH_1	EH ligand	.NPF.	Motifs	ELM/PhyloHMM	NPF motif interacting with EH domains, usually during regulation of endocytotic processes	(Dinkel et al., 2016)
31	LIG_eIF4E_1	eIF4E binding motif	Y....L[VILMF]	Motifs	ELM	Motif binding to the dorsal surface of eIF4E.	(Dinkel et al., 2016)
32	LIG_GLEBS_BUB3_1	GLEBS motif	[EN][FYLW][NSQ].EE[ILMVF][^P][LIVMFA]	Motifs	ELM	Gle2-binding-sequence motif	(Dinkel et al., 2016)
33	LIG_LIR_Gen_1	Atg8 protein family ligands	[EDST].{0,2}[WFY]..[ILV]	Motifs	ELM	Canonical LIR motif that binds to Atg8 protein family members to mediate processes involved in autophagy.	(Dinkel et al., 2016)
34	LIG_PCNA_PIPBox_1	PCNA binding PIP box	((^.{0,3})|(Q)).[^FHWY][ILM][^P][^FHILVWYP][HFM][FMY]..	Motifs	ELM/PhyloHMM	The PCNA binding PIP box motif is found in proteins involved in DNA replication, repair and cell cycle control.	(Dinkel et al., 2016)
35	LIG_SUMO_SIM_pa	SUMO interaction site	[DEST]{0,5}.[VILPTM][VIL][DESTVILMA][VIL].{0,	Motifs	ELM	Motif for the parallel beta augmentation mode of non-covalent binding to SUMO	(Dinkel et al., 2016)

	r_1		1}[DEST]{1,10}			protein.	
36	MOD_CDK_SPxK_1	CDK Phosphorylation Site	...([ST])P.[KR]	Motifs	ELM/Condens	Canonical version of the CDK phosphorylation site which shows specificity towards a lysine/arginine residue at the [ST]+3 position.	(Dinkel et al., 2016)
37	MOD_LATS_1	LATS kinase phosphorylation motif	H.[KR]..([ST])[^P]	Motifs	ELM	The LATS phosphorylation motif is recognised by the LATS kinases for Ser/Thr phosphorylation. Substrates are often found toward the end of the Hippo signalling pathway.	(Dinkel et al., 2016)
38	MOD_SUMO_for_1	Sumoylation site	[VILMAFP](K).E	Motifs	ELM	Motif recognised for modification by SUMO-1	(Dinkel et al., 2016)
39	TRG_ER_FFAT_1	FFAT motif	[DE].{0,4}E[FY][FYK]D[AC].[ESTD]	Motifs	ELM	VAP-A/Scs2 MSP-domain binding FFAT (diphenylalanine [FF] in an Acidic Tract) motif	(Dinkel et al., 2016)
40	TRG_Golgi_diPhe_1	ER export signals	Q.{6,6}FF.{6,7}	Motifs	ELM	ER to Golgi anterograde transport signal found at the C-terminus of type I ER-CGN integral membrane cargo receptors (cytoplasmic in this topology), it binds to COPII.	(Dinkel et al., 2016)
41	TRG_NLS_MonoExtN_	NLS classical Nuclear Localization	(([PKR].{0,1}[^DE])|([PKR]))((K[RK])|(RK))(([^DE][KR])|([KR][^	Motifs	ELM	Monopartite variant of the classical basically charged NLS. N-extended version.	(Dinkel et al., 2016)

	4	Signals	DE]))[^DE]				
42	MOD_CDK_STP	CDK phosphorylation motif	[ST]P	Motifs	Condens	NA	(Holt et al., 2009; A. C. W. Lai et al., 2012)
43	MOD_MEC1	Mec1 phosphorylation motif	[ST]Q	Motifs	Condens	NA	(A. C. W. Lai et al., 2012; Schwartz et al., 2002)
44	MOD_PRK1	Prk1 phosphorylation motif	[LIVM]….TG	Motifs	Condens	NA	(Huang et al., 2003; A. C. W. Lai et al., 2012)
45	MOD_IPL1	Ipl1 phosphorylation motif	[RK].[ST][LIV]	Motifs	Condens	NA	(Cheeseman et al., 2002; A. C. W. Lai et al., 2012)
46	MOD_PKA	Pka phosphorylation motif	R[RK].S	Motifs	Condens	NA	(Budovskaya et al., 2005; Kemp and Pearson, 1990; A. C. W. Lai et al., 2012; Townsend et al., 1996)
47	MOD_CKII	Ckii phosphorylation	[ST][DE].[DE]	Motifs	Condens	NA	(A. C. W. Lai et al., 2012; Meggio and

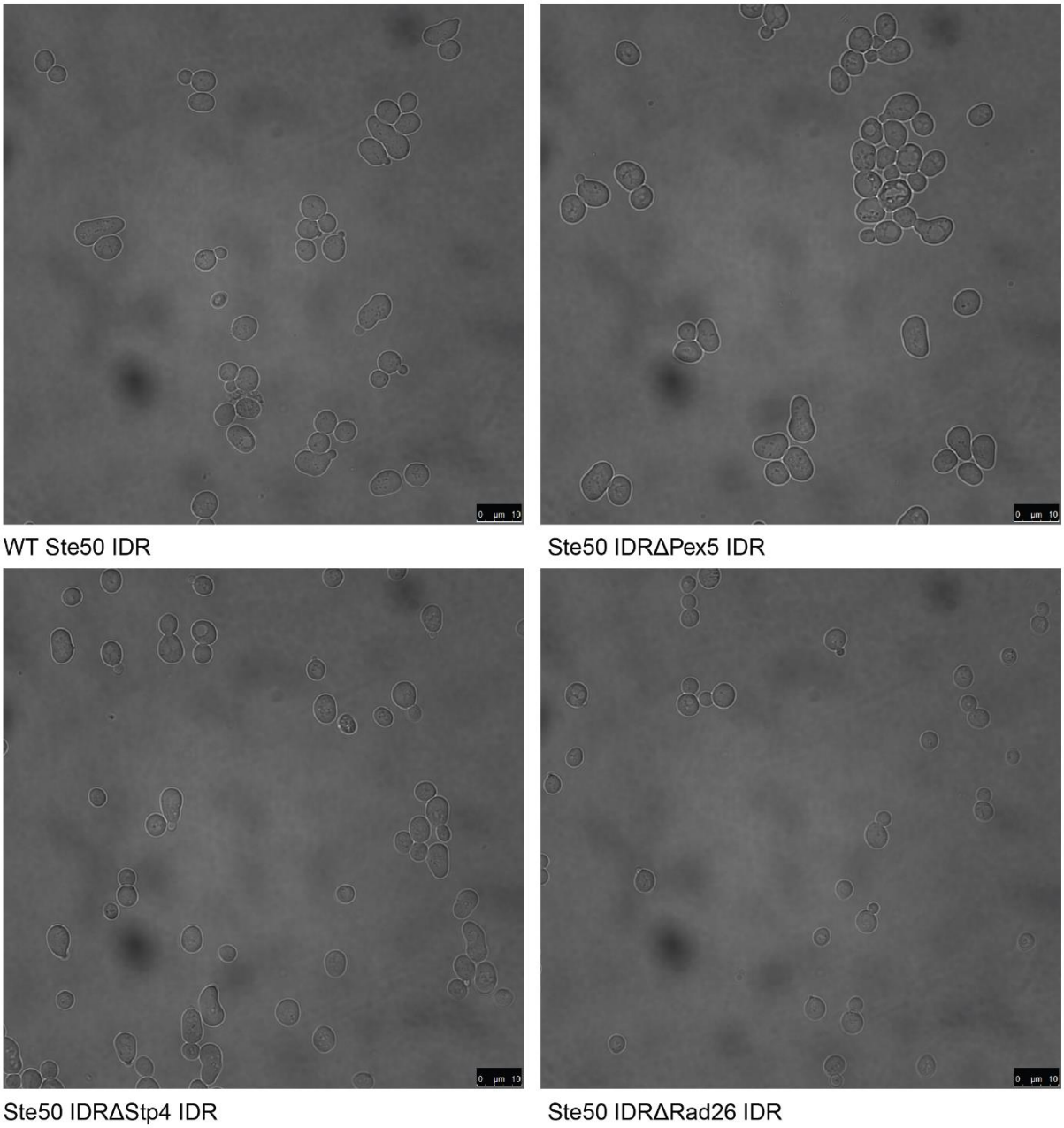
		motif					Pinna, 2003; Niefind et al., 2007)
48	MOD_IME2	Ime2 phosphorylation motif	RP.[ST]	Motifs	Condens	NA	(Holt et al., 2007; J. Lai et al., 2012)
49	DOC_PRO	proline-rich motif	P..P	Motifs	PhyloHMM	NA	(Nguyen Ba et al., 2012)
50	TRG_ER_HDEL	ER localization motif	HDEL	Motifs	PhyloHMM	NA	(Nguyen Ba et al., 2012)
51	TRG_MITOCHONDRIA	Mitochondrial localization motif	[MR]L[RK]	Motifs	PhyloHMM	NA	(Nguyen Ba et al., 2012)
52	MOD_ISOMERASE	Disulfide isomerase motif	C..C	Motifs	PhyloHMM	NA	(Nguyen Ba et al., 2012)
53	TRG_FG	FG nucleoporin motif	F.FG|GLFG	Motifs	PhyloHMM	NA	(Frey and Görlich, 2009; Nguyen Ba et al., 2012)
54	INT_RGG	RGG motif	RGG | RG	Motifs	Literature	NA	(P. A. Chong et al., 2018)

55	length	Length	NA	Physicochemical properties	Literature	Length in log scale	Reviewed in van der Lee et al. 2014
56	acidic	Acidic residue content	[DE]	Physicochemical properties	Literature /localCIDER	NA	(Warren and Shechter, 2017)
57	basic	Basic residue content	[RK]	Physicochemical properties	Literature /localCIDER	NA	(Fukasawa et al., 2015)
58	hydrophobicity	Hydrophobicity	NA	Physicochemical properties	Literature /localCIDER	Kyte-Doolittle scale	(Kyte and Doolittle, 1982)
59	aliphatic	Aliphatic residue content	[ALMIV]	Physicochemical properties	Literature /localCIDER	NA	(Holehouse et al., 2017)
60	polar_fraction	Polar residue content	[QNSTGCH]	Physicochemical properties	Literature /localCIDER	NA	(Holehouse et al., 2017)
61	chain_expanding	Chain expanding residue content	[EDRKP]	Physicochemical properties	Literature /localCIDER	NA	(Holehouse et al., 2017)
62	aromatic	Aromatic residue	[FYW]	Physicochemical	Literature /localCI	NA	(Holehouse et al., 2017)

		content		properties	DER		
63	disorder_promoting	Disorder promoting residue content	[TAGRDHQKSEP]	Physicochemical properties	Literature /localCIDER	NA	(Holehouse et al., 2017)
64	Iso_point	Isoelectric point	NA	Physicochemical properties	Literature /localCIDER	pH where charge of peptide is neutral	(Holehouse et al., 2017; Marsh and Forman-Kay, 2010; Tomasso et al., 2016)
65	PPII_prop	PPII propensity	NA	Physicochemical properties	Literature /localCIDER	Propensity for proline to form left-handed helices	(Elam et al., 2013; Holehouse et al., 2017)
66	REP_Q2	Q repeat	Q{2,}	Repeats and complexity	Literature	Fraction of 2 or more Q in a row	(Chavali et al., 2017)
67	REP_N2	N repeat	N{2,}	Repeats and complexity	Literature	Fraction of 2 or more N in a row	(Chavali et al., 2017)
68	REP_S2	S repeat	S{2,}	Repeats and complexity	Literature	Fraction of 2 or more S in a row	(Chavali et al., 2017)
69	REP_G2	G repeat	G{2,}	Repeats and complexity	Literature	Fraction of 2 or more G in a row	(Chavali et al., 2017).
70	REP_	E repeat	E{2,}	Repeats	Literatu	Fraction of 2 or more E in a	(Chavali et

	E2			and complexity	re	row	al., 2017)
71	REP_D2	D repeat	D{2,}	Repeats and complexity	Literature	Fraction of 2 or more D in a row	(Chavali et al., 2017)
72	REP_K2	K repeat	K{2,}	Repeats and complexity	Literature	Fraction of 2 or more K in a row	(Matsushima et al., 2009; Simon and Hancock, 2009)
73	REP_R2	R repeat	R{2,}	Repeats and complexity	Literature	Fraction of 2 or more R in a row	(Matsushima et al., 2009; Simon and Hancock, 2009)
74	REP_P2	P repeat	P{2,}	Repeats and complexity	Literature	Fraction of 2 or more P in a row	(Chavali et al., 2017; Matsushima et al., 2009; Simon and Hancock, 2009)
75	REP_QN2	Q/N repeat	[QN]{2,}	Repeats and complexity	Literature	Fraction of 2 or more Q/N in a row	(Alberti et al., 2009; Van Der Lee et al., 2014)
76	REP_RG2	R/G repeat	[RG]{2,}	Repeats and complexity	Literature	Fraction of 2 or more R/G in a row; aka "GAR" regions	(P. A. Chong et al., 2018; Matsushima et al., 2009)

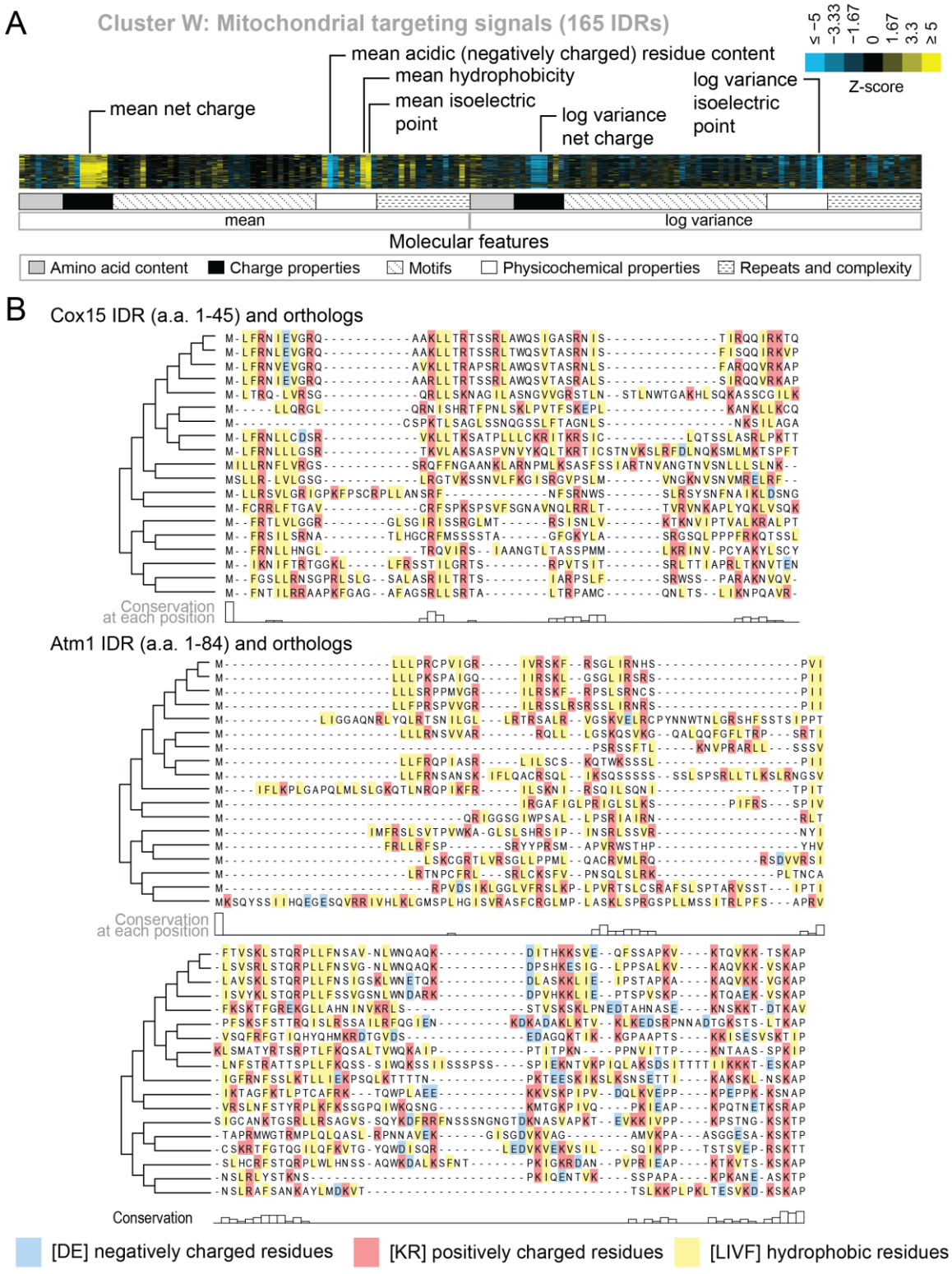
77	REP_FG2	F/G repeat	[FG]{2,}	Repeats and complexity	Literature	Fraction of 2 or more F/G in a row	Reviewed in (Van Der Lee et al., 2014)
78	REP_SG2	S/G repeat	[SG]{2,}	Repeats and complexity	Literature	Fraction of 2 or more S/G in a row	(Matsushima et al., 2009; Simon and Hancock, 2009)
79	REP_SR2	S/R repeat	[SR]{2,}	Repeats and complexity	Literature	Fraction of 2 or more S/R in a row	Reviewed in (Van Der Lee et al., 2014)
80	REP_KAP2	K/A/P repeat	[KAP]{2,}	Repeats and complexity	Literature	Fraction of 2 or more K/A/P in a row	Reviewed in (Van Der Lee et al., 2014)
81	REP_PTS2	P/T/S repeat	[PTS]{2,}	Repeats and complexity	Literature	Fraction of 2 or more P/T/S in a row	Reviewed in (Van Der Lee et al., 2014)
82	wf_complexity	Wootton-Federhen sequence complexity	NA	Repeats and complexity	Literature /localCIDER	Complexity based on SEG algorithm (Wootton and Federhen, 1993), blob length=IDR length, step size = 1	(Wootton and Federhen, 1993)



Appendix Figure 2-3. Full field-of-view micrographs of pheromone-exposed *S.cerevisiae* strains from Figure 3-2C.

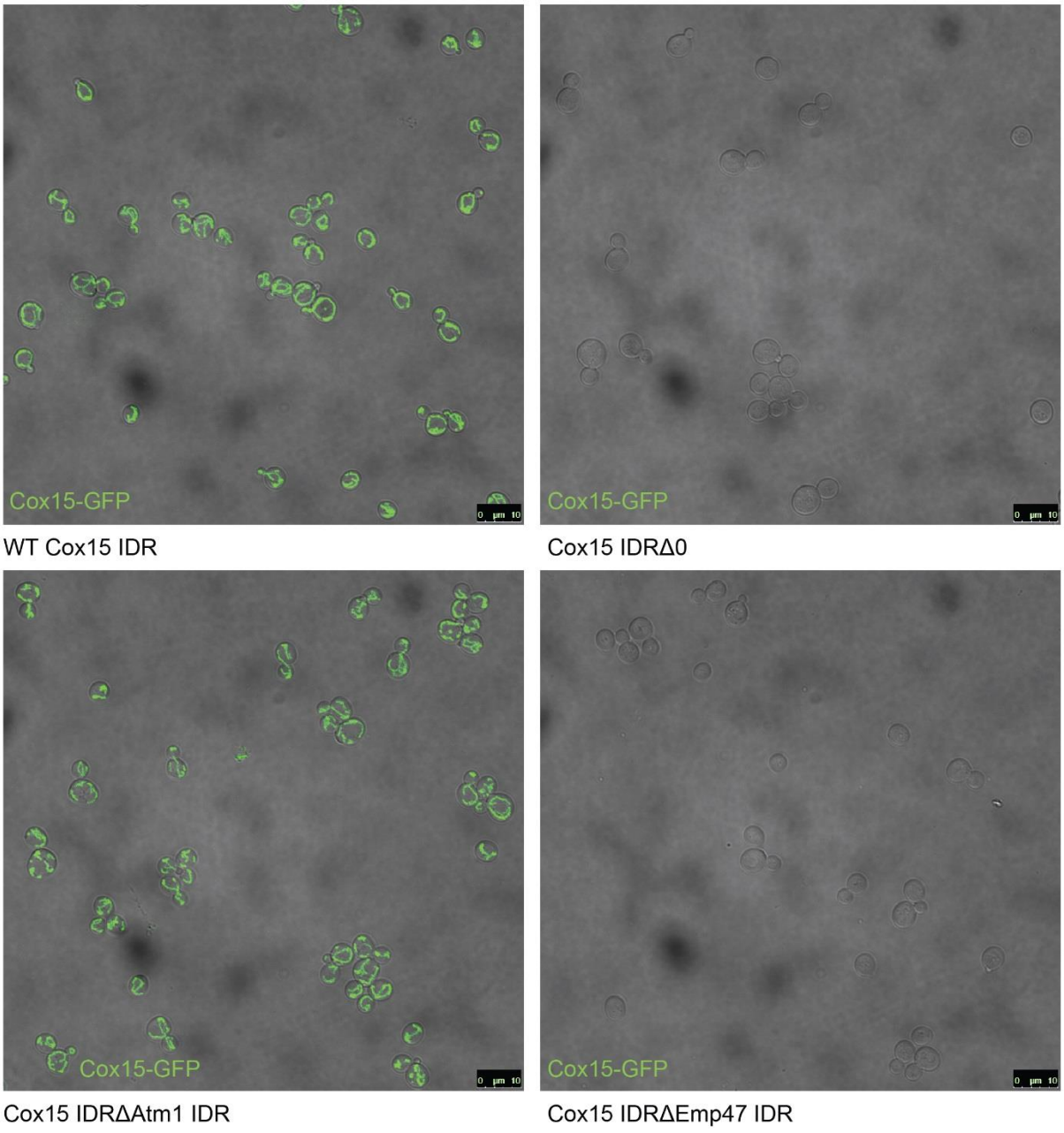
Appendix Table 2-2. Controls for clustering results.

Cluster ID	Random permutation z-score	Amino acid permutation z-score	Percent of homologous IDRs (top 1% homology in proteome)
A	24.94	6.13	1.47
B	10.21	8.99	0
C	30.77	10.74	0
D	38.02	22.07	1.23
E	7.87	6.54	0
F	15.45	12.74	0
G	12.99	9.41	5.87
H	29.01	14.35	0
I	19.88	11.37	0
J	28.05	8.62	0
K	7.9	9.95	0
L	45.49	11.62	0.43
M	47.5	15.31	2.84
N	55.28	23.98	0
O	46.42	7.16	0.6
P	50.78	22.18	0.26
Q	230.85	50.28	8.86
R	16.51	5.97	0.94
S	19.74	21.84	0
T	13.77	10.43	0
U	16.81	8.15	0.03
V	44.33	10.41	0
W	187.24	39.2	0

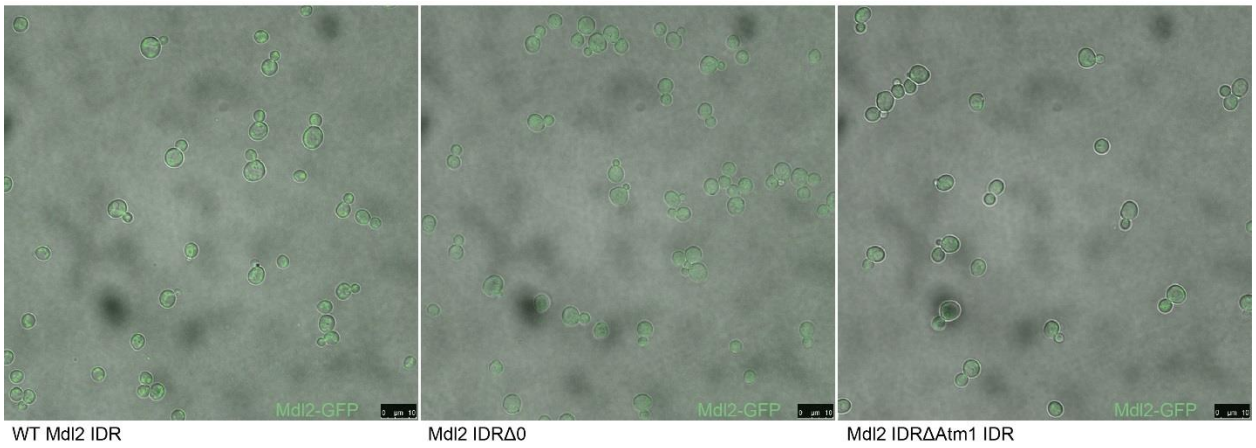


Appendix Figure 2-4. Evolutionary signatures in cluster W contain molecular features that have been previously reported for mitochondrial N-terminal targeting signals. A) Pattern of evolutionary signatures in cluster W. B) Multiple sequence alignments of example disordered

regions from Cox15 (top) and Atm1 (bottom) from cluster W, showing a subset of highlighted molecular features. Species included in phylogeny in order from top to bottom are *S.cerevisiae*, *S.mikatae, S.kudriavzevii, S.uvarum, C.glabrata, K.africana, K.naganishii, N.castellii, N.dairenensis, T.phaffii, V.polyspora, Z.rouxii, T.delbrueckii, K.lactis, E.gossypii, E.cymbalariae, L.kluyveri, L.thermotolerans, L.waltii*.



Appendix Figure 2-5. Full field-of-view micrographs of *S.cerevisiae* strains from Figure 3-6C.



Appendix Figure 2-6. Micrographs of *S.cerevisiae* strains with three different genotypes. From left to right: Mdl2-GFP has a mitochondrial localization in the wildtype (WT) strain, knocking out the Mdl2 IDR abolishes wildtype localization, and replacing the Mdl2 IDR with that of Atm1 rescues mitochondrial localization.

Appendix Table 2-3. List of strains used in this study.

Strain	Genotype	Source
YTZ113	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Cox15-GFP-His3	Huh et al., courtesy of Brenda Andrews' lab
YTZ115	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Mdl2-GFP-His3	Huh et al., courtesy of Brenda Andrews' lab
YBS270	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Cox15-GFP-His3 Cox15 IDR (a.a. 1-45)::0	This study
YBS271	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Cox15-GFP-His3 Cox15 IDR (a.a. 1-45)::Atm1 IDR (a.a. 1-84)	This study
YBS272	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Mdl2-GFP-His3 Mdl2 IDR (a.a. 1-99)::0	This study
YBS273	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Mdl2-GFP-His3 Mdl2 IDR (a.a. 1-99)::Atm1 IDR (a.a. 1-84)	This study
YBS278	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 Cox15-GFP-His3 Cox15 IDR (a.a. 1-45)::Emp47 IDR (a.a. 1-37)	This study
YTZ127	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 HO::pFUS1-yemGFP-klURA3	This study
YTZ129	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 Ste50 IDR (a.a. 152-250)::Pex5 IDR (a.a.77-161) HO::pFUS1-yemGFP-klURA3	This study
YTZ130	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 Ste50 IDR (a.a. 152-250)::Rad26 IDR (a.a. 163-269) HO::pFUS1-yemGFP-klURA3	This study
YTZ131	MATa his3Δ1 leu2Δ0 ura3Δ0 met15Δ0 SSK22::HisMX3 SSK2Δ0 Ste50 IDR (a.a. 152-250)::Stp4 IDR (a.a. 144-	This study

	256) HO::pFUS1-yemGFP-klURA3	

Appendix 3
Supplementary material for Chapter 4

This work has not been previously published.

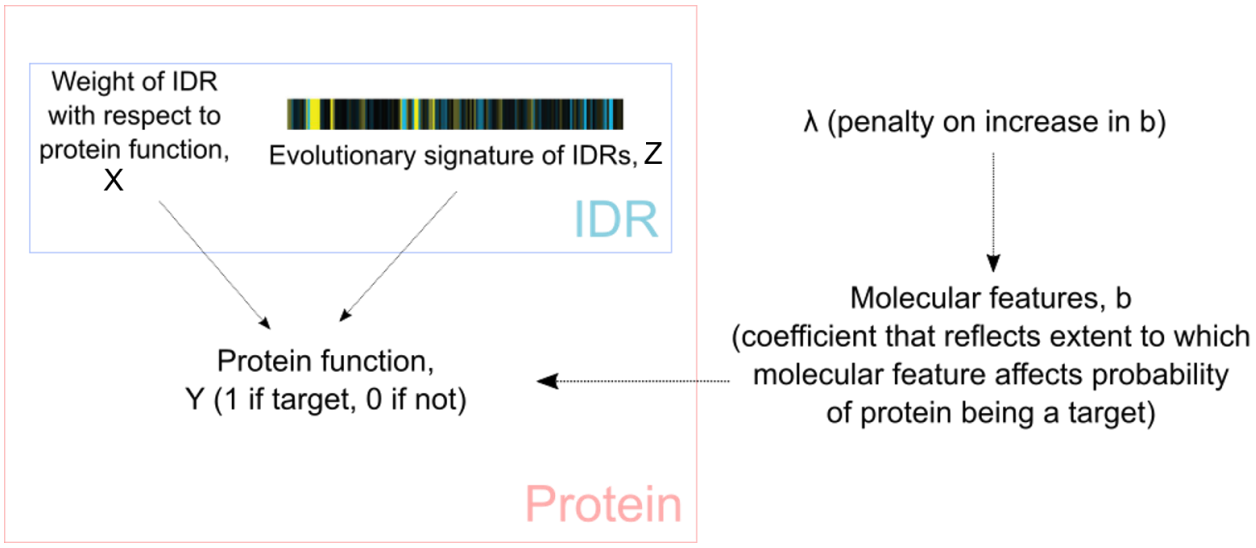
Taraneh Zarin[1], Bob Strome[1], Alan M Moses[1,2,3]

4. Department of Cell and Systems Biology, University of Toronto, Toronto, Canada
5. Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, Canada
6. Department of Computer Science, University of Toronto, Toronto, Canada

Appendix Table 3-1. Top 5 functional predictions for proteins that have unknown gene function or gene description annotations.

	ID	Name	Gene function summary	Gene description	Predicted function	Probability
1	YIL011W	TIR3	Protein whose biological role is unknown; localizes to the cell wall	Cell wall mannoprotein	Extracellular region (GO.0005576)	0.913026
2	YIL011W	TIR3	Protein whose biological role is unknown; localizes to the cell wall	Cell wall mannoprotein	Cell wall (GO.0005618)	0.864549
3	YER080W	AIM9	Protein whose biological role and cellular location are unknown	Protein of unknown function	Mitochondrion (GO.0005739)	0.781754
4	YER080W	AIM9	Protein whose biological role and cellular location are unknown	Protein of unknown function	Mitochondrial inner membrane	0.666776
5	YHR059W	FYV4	Component of the small subunit of the mitochondrial ribosome, which mediates translation in the	Protein of unknown function	Mitochondrion	0.62196

			mitochondrion			



Appendix Figure 3-1. Schematic of model used to predict protein function (and weight of each IDR's contribution) from evolutionary signatures of IDRs.

References

Abdel-Sater F, Iraqui I, Urrestarazu A, André B. 2004. The External Amino Acid Signaling Pathway Promotes Activation of Stp1 and Uga35/Dal81 Transcription Factors for Induction of the AGP1 Gene in Saccharomyces cerevisiae. *Genetics* **166**:1727–1739. doi:10.1534/genetics.166.4.1727

Aboussekhra A, Chanet R, Zgaga Z, Cassier-Chauvat C, Heude M, Fabre F. 1989. RADH , a gene of Saccharomyces cerevisiae encoding a putative DNA helicase involved in DNA repair. Characteristics of radH mutants and sequence of the gene. *Nucleic Acids Res* **17**:7211–7219. doi:10.1093/nar/17.18.7211

Adams KL, Palmer JD. 2003. Evolution of mitochondrial gene content: Gene loss and transfer to the nucleus. *Mol Phylogenet Evol* **29**:380–395. doi:10.1016/S1055-7903(03)00194-5

Afanasyeva A, Bockwoldt M, Cooney CR, Heiland I, Gossmann TI. 2018. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res* **28**:975–982. doi:10.1101/gr.232645.117

Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP. 2007. The molecular architecture of the nuclear pore complex. *Nature* **450**:695–701. doi:10.1038/nature06405

Alberti S, Halfmann R, King O, Kapila A, Lindquist S. 2009. A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. *Cell* **137**:146–158. doi:10.1016/j.cell.2009.02.044

Albuquerque CP, Smolka MB, Payne SH, Bafna V, Eng J, Zhou H. 2008. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol Cell Proteomics* **7**:1389–1396. doi:10.1074/mcp.M700468-MCP200

Andresen C, Helander S, Lemak A, Farès C, Csizmok V, Carlsson J, Penn LZ, Forman-Kay JD, Arrowsmith CH, Lundström P, Sunnerhagen M. 2012. Transient structure and dynamics in the disordered c-Myc transactivation domain affect Bin1 binding. *Nucleic Acids Res* **40**:6353–6366. doi:10.1093/nar/gks263

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MDS, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJK, Doggett N, Zharkikh A, Tavtigian S V, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**:1301–10. doi:10.1126/science.1072104

Bah A, Forman-Kay JD. 2016. Modulation of intrinsically disordered protein function by post-translational modifications. *J Biol Chem* **291**:6696–6705. doi:10.1074/jbc.R115.695056

Baker CR, Hanson-Smith V, Johnson AD. 2013. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science* **342**:104–8. doi:10.1126/science.1240810

Banani SF, Lee HO, Hyman AA, Rosen MK. 2017. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**:285–298. doi:10.1038/nrm.2017.7

Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution (N Y)* **66**:2369–2383. doi:10.1111/j.1558-5646.2012.01619.x

Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection. *Proc Natl Acad Sci U S A* **106**:1133–8. doi:10.1073/pnas.0812009106

Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, Boone C, Bader GD, Myers CL, Kim PM. 2011. Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* **12**:R14. doi:10.1186/gb-2011-12-2-r14

Bellosta CJG. 2015. rPython: Package Allowing R to Call Python. https://cran.r-project.org/package=rPython

Beltrao P, Albanèse V, Kenner LR, Swaney DL, Burlingame A, Villén J, Lim W a, Fraser JS, Frydman J, Krogan NJ. 2012. Systematic functional prioritization of protein posttranslational modifications. *Cell* **150**:413–25. doi:10.1016/j.cell.2012.05.036

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. doi:10.1002/msb.201304521

Beltrao P, Serrano L. 2005. Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. *PLoS Comput Biol* **1**:e26. doi:10.1371/journal.pcbi.0010026

Beltrao P, Trinidad JC, Fiedler D, Roguev A, Lim W a, Shokat KM, Burlingame AL, Krogan NJ. 2009. Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species. *PLoS Biol* **7**:e1000134. doi:10.1371/journal.pbio.1000134

Bentley-DeSousa A, Holinier C, Moteshareie H, Tseng YC, Kajjo S, Nwosu C, Amodeo GF, Bondy-Chorney E, Sai Y, Rudner A, Golshani A, Davey NE, Downey M. 2018. A Screen for Candidate Targets of Lysine Polyphosphorylation Uncovers a Conserved Network Implicated in Ribosome Biogenesis. *Cell Rep* **22**:3427–3439. doi:10.1016/j.celrep.2018.02.104

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res* **11**:1335–45. doi:10.1101/gr.178701

Bodenmiller B, Wanka S, Kraft C, Urban J, Campbell D, Pedrioli PG, Gerrits B, Picotti P, Lam H, Vitek O, Brusniak M, Roschitzki B, Zhang C, Shokat KM, Schlapbach R, Colman-Lerner A, Nolan GP, Nesvizhskii AI, Peter M, Loewith R, von Mering C, Aebersold R. 2010. Phosphoproteomic analysis reveals interconnected system-wide responses to perturbations of kinases and phosphatases in yeast. *Sci Signal* **3**:rs4. doi:10.1126/scisignal.2001182

Boehning M, Dugast-Darzacq C, Rankovic M, Hansen AS, Yu T, Marie-Nelly H, McSwiggen DT, Kokic G, Dailey GM, Cramer P, Darzacq X, Zweckstetter M. 2018. RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat Struct Mol Biol* **25**:833–840. doi:10.1038/s41594-018-0112-y

Boeke JD, Trueheart J, Natsoulis G, Fink GR. 1987. 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods Enzymol* **154**:164–75.

Boija A, Klein IA, Sabari BR, Dall'Agnese A, Coffey EL, Zamudio A V., Li CH, Shrinivas K, Manteiga JC, Hannett NM, Abraham BJ, Afeyan LK, Guo YE, Rimel JK, Fant CB, Schuijers J, Lee TI, Taatjes DJ, Young RA. 2018. Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**:1842-1855.e16. doi:10.1016/j.cell.2018.10.042

Bolognesi B, Faure AJ, Seuma M, Schmiedel JM. 2019. The mutational landscape of a prion-like domain.

Boothby TC, Tapia H, Brozena AH, Piszkiewicz S, Smith AE, Giovannini I, Rebecchi L, Pielak GJ, Koshland D, Goldstein B. 2017. Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. *Mol Cell* **65**:975-984.e5. doi:10.1016/j.molcel.2017.02.018

Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, Fernandes CB, Sottini A, Soranno A, Buholzer KJ, Nettels D, Kragelund BB, Best RB, Schuler B. 2018. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**:61–66. doi:10.1038/nature25762

Breslow DK, Cameron DM, Collins SR, Schuldiner M, Stewart-ornstein J, Newman HW, Braun S, Madhani HD, Krogan NJ, Weissman JS. 2008. A comprehensive strategy enabling high-resolution functional analysis of the yeast genome **5**. doi:10.1038/NMETH.1234

Brown CJ, Johnson AK, Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol* **27**:609–21. doi:10.1093/molbev/msp277

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Keith Dunker A. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55**:104–110. doi:10.1007/s00239-001-2309-6

Buckling A, Craig Maclean R, Brockhurst M a, Colegrave N. 2009. The Beagle in a bottle. *Nature* **457**:824–9. doi:10.1038/nature07892

Budovskaya Y V., Stephan JS, Deminoff SJ, Herman PK. 2005. An evolutionary proteomics approach identifies substrates of the cAMP-dependent protein kinase. *Proc Natl Acad Sci* **102**:13933–13938. doi:10.1073/pnas.0501046102

Burger VM, Nolasco DO, Stultz CM. 2016. Expanding the range of protein function at the far end of the order-structure continuum. *J Biol Chem* **291**:6706–6713. doi:10.1074/jbc.R115.692590

Busch DJ, Houser JR, Hayden CC, Sherman MB, Lafer EM, Stachowiak JC. 2015. Intrinsically disordered proteins drive membrane curvature. *Nat Commun* **6**:7875. doi:10.1038/ncomms8875

Butler MAA, King AAA. 2004. Phylogenetic Comparative Analysis: A Modeling Approach for Adaptive Evolution. *Am Nat* **164**:683–695. doi:10.1086/426002

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. doi:10.1101/gr.3672305

Chao L, Cox EC. 1983. Competition Between High and Low Mutating Strains of Escherichia coli. *Evolution (N Y)* **37**:125. doi:10.2307/2408181

Charif D, Lobry JR. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis In: Bastolla U, Porto M, Roman HE, Vendruscolo M, editors. Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 207–232. doi:10.1007/978-3-540-35306-5_10

Charlesworth B. 2013. Stabilizing selection, purifying selection, and mutational bias in finite populations. *Genetics* **194**:955–71. doi:10.1534/genetics.113.151555

Chavali S, Chavali PL, Chalancon G, De Groot NS, Gemayel R, Latysheva NS, Ing-Simmons E, Verstrepen KJ, Balaji S, Babu MM. 2017. Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins. *Nat Struct Mol Biol* **24**:765–777. doi:10.1038/nsmb.3441

Cheeseman IM, Anderson S, Jwa M, Green EM, Kang J-S, Yates Iii JR, Chan CSM, Drubin DG, Barnes G. 2002. Phospho-Regulation of Kinetochore-Microtubule Attachments by the Aurora Kinase Ipl1p will require the identification of any remaining kineto- chore proteins.

Given the central role that kinetochore-microtubule. *Cell* **111**:163–172.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006a. Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder. *J Proteome Res* **5**:888–898. doi:10.1021/pr060049p

Chen JW, Romero P, Uversky VN, Dunker AK. 2006b. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res* **5**:879–887. doi:10.1021/pr060048x

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. 2012. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res* **40**:700–705. doi:10.1093/nar/gkr1029

Chong PA, Vernon RM, Forman-Kay JD. 2018. RGG/RG Motif Regions in RNA Binding and Phase Separation. *J Mol Biol* **430**:4650–4665. doi:10.1016/j.jmb.2018.06.014

Chong S, Dugast-Darzacq C, Liu Z, Dong P, Dailey GM, Cattoglio C, Heckert A, Banala S, Lavis L, Darzacq X, Tjian R. 2018. Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science (80-)* **361**. doi:10.1126/science.aar2555

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen B a, Johnston M. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**:71–6. doi:10.1126/science.1084337

Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, Myers CL, Blencowe BJ, Kim PM. 2013. Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput Biol* **9**:e1003030. doi:10.1371/journal.pcbi.1003030

Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biol J Linn Soc* **118**:64–77. doi:10.1111/bij.12701

Cumberworth A, Lamour G, Babu MM, Gsponer J. 2013. Promiscuity as a functional trait:

intrinsically disordered regions as central players of interactomes. *Biochem J* **454**:361–9. doi:10.1042/BJ20130545

Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu R V. 2016. Cryptic sequence features within the disordered protein p27 ^{Kip1} regulate cell cycle signaling. *Proc Natl Acad Sci* 201516277. doi:10.1073/pnas.1516277113

Das RK, Pappu R V. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **110**:13392–7. doi:10.1073/pnas.1304749110

Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ. 2007. Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol Evol* **65**:277–288. doi:10.1007/s00239-007-9011-2

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. 2012. Attributes of short linear motifs. *Mol BioSyst* **8**:268–281. doi:10.1039/C1MB05231D

de Hoon MJL, Imoto S, Nolan J, Miyano S. 2004. Open source clustering software. *Bioinformatics* **20**:1453–1454. doi:10.1093/bioinformatics/bth078

de la Chaux N, Messer PW, Arndt PF. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol* **7**:191. doi:10.1186/1471-2148-7-191

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* **8**:675–88. doi:10.1038/nrg2160

DeForte S, Uversky V. 2016. Order, Disorder, and Everything in Between. *Molecules* **21**:1090. doi:10.3390/molecules21081090

Dennig J, Beato M, Suske G. 2018. An inhibitor domain in Sp3 regulates its glutamine-rich activation domains. *EMBO J* **15**:5659–5667. doi:10.1002/j.1460-2075.1996.tb00950.x

Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, Milchevskaya V, Schneider M, Kühn H, Behrendt A, Dahl SL, Damerell V, Diebel S, Kalman S, Klein S, Knudsen AC, Mäder C, Merrill S, Staudt A, Thiel V, Welti L, Davey NE, Diella F, Gibson TJ. 2016. ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res* **44**:D294–D300. doi:10.1093/nar/gkv1291

Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**:827–839. doi:10.1016/j.jmb.2005.01.071

Dujon B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* **22**:375–87. doi:10.1016/j.tig.2006.05.007

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud J-M, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard G-F, Straub M-L, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet J-L. 2004. Genome evolution in yeasts. *Nature* **430**:35–44. doi:10.1038/nature02579

Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, Dyson HJ, Forman-Kay J, Fuxreiter M, Gsponer J, Han K-H, Jones DT, Longhi S, Metallo SJ, Nishikawa K, Nussinov R, Obradovic Z, Pappu R V., Rost B, Selenko P, Subramaniam V, Sussman JL, Tompa P, Uversky VN. 2013. What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord Proteins* **1**:e24157. doi:10.4161/idp.24157

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

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. doi:10.1371/journal.pone.0000967

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2018. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**:427–432. doi:10.1093/nar/gky995

Elam WA, Schrank TP, Campagnolo AJ, Hilser VJ. 2013. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. *Protein Sci* **22**:405–417. doi:10.1002/pro.2217

Elbaum-Garfinkle S, Kim Y, Szczepaniak K, Chen CC-H, Eckmann CR, Myong S, Brangwynne CP. 2015. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc Natl Acad Sci* **112**:7189–7194. doi:10.1073/pnas.1504822112

Elion E a, Satterberg B, Kranz JE. 1993. FUS3 phosphorylates multiple components of the mating signal transduction cascade: evidence for STE12 and FAR1. *Mol Biol Cell* **4**:495–510. doi:10.1091/mbc.4.5.495

English JG, Shellhammer JP, Malahe M, Mccarter PC, Elston TC, Dohlman HG. 2015. MAPK feedback encodes a switch and timer for tunable stress adaptation in yeast **8**:1–14.

Erdmann R, Blobel G. 1996. Identification of Pex13p, a peroxisomal membrane receptor for the PTS1 recognition factor. *J Cell Biol* **135**:111–121. doi:10.1083/jcb.135.1.111

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Genet* **25**:471–492.

Ferrigno P, Posas F, Koepp D, Saito H, Silver P a. 1998. Regulated nucleo/cytoplasmic exchange of HOG1 MAPK requires the importin beta homologs NMD5 and XPO1. *EMBO J* **17**:5606–14. doi:10.1093/emboj/17.19.5606

Forman-Kay JD, Mittag T. 2013. From sequence and forces to structure, function, and evolution

of intrinsically disordered proteins. *Structure* **21**:1492–9. doi:10.1016/j.str.2013.08.001

Fraley C, Raftery AE, Murphy TB, Scrucca L. 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *Tech Rep 597, Univ Washingt* 1–50.

Franzmann TM, Jahnel M, Pozniakovsky A, Mahamid J, Holehouse AS, Nüske E, Richter D, Baumeister W, Grill SW, Pappu R V., Hyman AA, Alberti S. 2018. Phase separation of a yeast prion protein promotes cellular fitness. *Science (80-)* **359**. doi:10.1126/science.aao5654

Fraser HB, Levy S, Chavan A, Shah HB, Perez JC, Zhou Y, Siegal ML, Sinha H. 2012. Polygenic cis-regulatory adaptation in the evolution of yeast pathogenicity. *Genome Res* **22**:1930–1939. doi:10.1101/gr.134080.111.1930

Freschi L, Courcelles M, Thibault P, Michnick SW, Landry CR. 2011. Phosphorylation network rewiring by gene duplication. *Mol Syst Biol* **7**:504. doi:10.1038/msb.2011.43

Frey S, Görlich D. 2009. FG/FxFG as well as GLFG repeats form a selective permeability barrier with self-healing properties. *EMBO J* **28**:2554–2567. doi:10.1038/emboj.2009.199

Friedman J, Hastie T, Tibshirani R. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**:1–22.

Frietze S, Farnham PJ. 2011. Transcription factor effector domains. *Subcell Biochem* **52**:261–77. doi:10.1007/978-90-481-9069-0_12

Fukasawa Y, Tsuji J, Fu S-C, Tomii K, Horton P, Imai K. 2015. MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites. *Mol Cell Proteomics* **14**:1113–1126. doi:10.1074/mcp.M114.043083

Fuxreiter M, Simon I, Friedrich P, Tompa P. 2004. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **338**:1015–1026. doi:10.1016/j.jmb.2004.03.017

Gagolewski M. 2019. R package stringi: Character string processing facilities.

Garg SG, Gould SB. 2016. The Role of Charge in Protein Targeting Evolution. *Trends Cell Biol* **26**:894–905. doi:10.1016/j.tcb.2016.07.001

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science (80-)* **263**:808–811. doi:10.1126/science.8303297

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Véronneau S, Dow S, Lucau-Danila A, Anderson K, André B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian K-D, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Güldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kötter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang C, Ward TR, Wilhelmy J, Winzeler E a, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**:387–91. doi:10.1038/nature00935

Gibson DG, Young L, Chuang R, Venter JC, Iii CAH, Smith HO, America N, Hutchison C a, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases (supp). *Nat Methods* **6**:343–5. doi:10.1038/NMETH.1318

Gnad F, De Godoy LMF, Cox J, Neuhauser N, Ren S, Olsen J V., Mann M. 2009. High-accuracy identification and bioinformatic analysis of in vivo protein phosphorylation sites in yeast. *Proteomics* **9**:4642–4652. doi:10.1002/pmic.200900144

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG. 1996. Life with 6000 Genes. *Science (80-)* **274**:546–567. doi:10.1126/science.274.5287.546

Gregory SM, Sweder KS. 2001. Deletion of the CSB homolog, RAD26, yields Spt(-) strains with proficient transcription-coupled repair. *Nucleic Acids Res* **29**:3080–3086.

Gresham D, Desai MM, Tucker CM, Jenq HT, Pai DA, Ward A, DeSevo CG, Botstein D, Dunham MJ. 2008. The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast. *PLoS Genet* **4**. doi:10.1371/journal.pgen.1000303

Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. 2015. Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**:3356–3358. doi:10.1093/bioinformatics/btv359

Guzder SN, Habraken Y, Sung P, Prakash L, Prakash S. 1996. RAD26, the yeast homolog of human Cockayne's syndrome group B gene, encodes a DNA-dependent ATPase. *J Biol Chem* **271**:18314–18317. doi:10.1074/jbc.271.31.18314

Hagen DC, McCaffrey G, Sprague GF. 1991. Pheromone response elements are necessary and sufficient for basal and pheromone-induced transcription of the FUS1 gene of Saccharomyces cerevisiae. *Mol Cell Biol* **11**:2952–61. doi:10.1128/MCB.11.6.2952.Updated

Halfmann R. 2016. A glass menagerie of low complexity sequences. *Curr Opin Struct Biol* **38**:9–16. doi:10.1016/j.sbi.2016.05.002

Halfmann R, Alberti S, Krishnan R, Lyle N, O'Donnell CW, King OD, Berger B, Pappu R V., Lindquist S. 2011. Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins. *Mol Cell* **43**:72–84. doi:10.1016/j.molcel.2011.05.013

Handfield LF, Chong YT, Simmons J, Andrews BJ, Moses AM. 2013. Unsupervised Clustering of Subcellular Protein Expression Patterns in High-Throughput Microscopy Images Reveals Protein Complexes and Functional Relationships between Proteins. *PLoS Comput Biol* **9**. doi:10.1371/journal.pcbi.1003085

Hansen TF. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. *Evolution (N Y)* **51**:1341–1351. doi:10.2307/2411186

Hao N, Zeng Y, Elston TC, Dohlman HG. 2008. Control of MAPK specificity by feedback phosphorylation of shared adaptor protein Ste50. *J Biol Chem* **283**:33798–802. doi:10.1074/jbc.C800179200

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: Investigating evolutionary radiations. *Bioinformatics* **24**:129–131. doi:10.1093/bioinformatics/btm538

Hastie T, Tibshirani R, Wainwright M. 2015. Statistical Learning with Sparsity. Chapman and Hall/CRC. doi:10.1201/b18401

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**:0890–0901. doi:10.1371/journal.pcbi.0020100

Hegreness M, Shoresh N, Hartl D, Kishony R. 2006. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science (80-)* **311**:1615–1617. doi:10.1126/science.1122469

Hietpas RT, Jensen JD, Bolon DN a. 2011. Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A* **108**:7896–7901. doi:10.1073/pnas.1016024108

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**:677–681. doi:10.1038/nature06151

Holehouse AS, Das RK, Ahad JN, Richardson MOG, Pappu R V. 2017. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* **112**:16–21. doi:10.1016/j.bpj.2016.11.3200

Holt LJ, Hutti JE, Cantley LC, Morgan DO. 2007. Evolution of Ime2 phosphorylation sites on Cdk1 substrates provides a mechanism to limit the effects of the phosphatase Cdc14 in meiosis. *Mol Cell* **25**:689–702. doi:10.1016/j.molcel.2007.02.012

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**:1682–6. doi:10.1126/science.1172867

Hu Z, Sackton TB, Edwards S V, Liu JS. 2019. Bayesian Detection of Convergent Rate Changes of Conserved Noncoding Elements on Phylogenetic Trees. *Mol Biol Evol* 1–15. doi:10.1093/molbev/msz049

Huang B, Zeng G, Ng AYJ, Cai M. 2003. Identification of novel recognition motifs and regulatory targets for the yeast actin-regulating kinase Prk1p. *Mol Biol Cell* **14**:4871–84. doi:10.1091/mbc.e03-06-0362

Hughes AL. 2005. Gene duplication and the origin of novel proteins. *Proc Natl Acad Sci* **102**:8791–8792. doi:10.1073/pnas.0503922102

Huh, K. W, Falvo, V. J, Gerke, C. L, Carroll, S. A, Howson, W. R, Weissman, S. J, O'Shea, K. E. 2003. Global analysis of protein localization in budding yeast. *Nature* **425**:686–691.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker a. K. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* **32**:1037–1049. doi:10.1093/nar/gkh253

Jansen G, Bühring F, Hollenberg CP, Ramezani Rad M. 2001. Mutations in the SAM domain of STE50 differentially influence the MAPK-mediated pathways for mating, filamentous growth and osmotolerance in Saccharomyces cerevisiae. *Mol Genet Genomics* **265**:102–117. doi:10.1007/s004380000394

Jensen MR, Ruigrok RWH, Blackledge M. 2013. Describing intrinsically disordered proteins at atomic resolution by NMR. *Curr Opin Struct Biol* **23**:426–435. doi:10.1016/j.sbi.2013.02.007

Jones DT, Cozzetto D. 2015. DISOPRED3: Precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**:857–863. doi:10.1093/bioinformatics/btu744

Kanshin E, Bergeron-Sandoval L-P, Isik SS, Thibault P, Michnick SW. 2015. A Cell-Signaling Network Temporally Resolves Specific versus Promiscuous Phosphorylation. *Cell Rep* **10**:1202–1214. doi:10.1016/j.celrep.2015.01.052

Kato M, Han TW, Xie S, Shi K, Du X, Wu LC, Mirzaei H, Goldsmith EJ, Longgood J, Pei J, Grishin N V., Frantz DE, Schneider JW, Chen S, Li L, Sawaya MR, Eisenberg D, Tycko R, McKnight SL. 2012. Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell* **149**:753–767. doi:10.1016/j.cell.2012.04.017

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

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**:241–54. doi:10.1038/nature01644

Kemp BE, Pearson RB. 1990. Protein kinase recognition sequence motifs. *Trends Biochem Sci* **15**:342–346. doi:10.1016/0968-0004(90)90073-K

Kendrew J, Bodo G, Dintzis H, Parrish R, Wyckoff H, Phillips D. 1958. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181**:662–666. doi:10.1038/181662a0

Ketela T, Green R, Bussey H. 1999. Saccharomyces cerevisiae Mid2p is a potential cell wall stress sensor and upstream activator of the PKC1-MPK1 cell integrity pathway. *J Bacteriol* **181**:3330–3340.

Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM. 2015. Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions. *Genome Biol Evol* **7**:1815–26. doi:10.1093/gbe/evv105

Kim Y, Furman CM, Manhart CM, Alani E, Finkelstein IJ. 2018. Intrinsically disordered regions regulate both catalytic and non-catalytic activities of the MutL ␣ mismatch repair complex 1–13. doi:10.1093/nar/gky1244

Kimura M, Ohta T. 1974. On some principles governing molecular evolution. *Proc Natl Acad Sci U S A* **71**:2848–52.

Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, Gastinel LN, Habuka N, Chen X, Maldonado F, Barker JE, Bacquet R, Villafranca JE. 1995. Crystal structures of human calcineurin and the human FKBP12–FK506–calcineurin complex. *Nature* **378**:641–644. doi:10.1038/378641a0

Kjaergaard M, Kragelund BB. 2017. Functions of intrinsic disorder in transmembrane proteins. *Cell Mol Life Sci* **74**:3205–3224. doi:10.1007/s00018-017-2562-5

Kjaergaard M, Teilum K, Poulsen FM. 2010. Conformational selection in the molten globule state of the nuclear coactivator binding domain of CBP. *Proc Natl Acad Sci* **107**:12535–12540. doi:10.1073/pnas.1001693107

Klemsz MJ, Maki RA. 1996. Activation of transcription by PU.1 requires both acidic and glutamine domains. *Mol Cell Biol* **16**:390–7.

Koch V, Otte M, Beye M. 2018. Evidence for Stabilizing Selection Driving Mutational Turnover of Short Motifs in the Eukaryotic Complementary Sex Determiner (Csd) Protein. *G3: Genes|Genomes|Genetics* g3.200527.2018. doi:10.1534/g3.118.200527

Kompella PS, Moses AM, Peisajovich SG. 2016. Introduction of Premature Stop Codons as an Evolutionary Strategy To Rescue Signaling Network Function. *ACS Synth Biol* acssynbio.6b00142. doi:10.1021/acssynbio.6b00142

Kroschwald S, Maharana S, Mateju D, Malinovska L, Nüske E, Poser I, Richter D, Alberti S. 2015. Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules. *Elife* **4**:1–32. doi:10.7554/eLife.06807

Kunz H. 2002. Emil Fischer - Unequalled classicist, master of organic chemistry research, and inspired trailblazer of biological chemistry. *Angew Chemie - Int Ed* **41**:4439–4451. doi:10.1002/1521-3773(20021202)41:23<4439::AID-ANIE4439>3.0.CO;2-6

Kwok RPS, Lundblad JR, Chrivia JC, Richards JP, Bächinger HP, Brennan RG, Roberts SGE, Green MR, Goodman RH. 1994. Nuclear protein CBP is a coactivator for the transcription factor CREB. *Nature* **370**:223–226. doi:10.1038/370223a0

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**:105–132. doi:10.1016/0022-2836(82)90515-0

Lai ACW, Nguyen Ba AN, Moses AM. 2012. Predicting kinase substrates using conservation of local motif density. *Bioinformatics* **28**:962–9. doi:10.1093/bioinformatics/bts060

Lai J, Koh CH, Tjota M, Pieuchot L, Raman V, Chandrababu KB, Yang D, Wong L, Jedd G. 2012. Intrinsically disordered proteins aggregate at fungal cell-to-cell channels and regulate intercellular connectivity. *Proc Natl Acad Sci* **109**:15781–15786. doi:10.1073/pnas.1207467109

Lande R. 1976. Natural Selection and Random Genetic Drift in Phenotypic Evolution. *Evolution (N Y)* **30**:314. doi:10.2307/2407703

Landry CR, Freschi L, Zarin T, Moses AM. 2014. Turnover of protein phosphorylation evolving under stabilizing selection. *Front Genet* **5**:1–6. doi:10.3389/fgene.2014.00245

Lee S-H, Kim D-H, J. Han J, Cha E-J, Lim J-E, Cho Y-J, Lee C, Han K-H. 2012. Understanding Pre-Structured Motifs (PreSMos) in Intrinsically Unfolded Proteins. *Curr Protein Pept Sci* **13**:34–54. doi:10.1016/j.ijheatmasstransfer.2013.04.020

Lemas D, Lekkas P, Ballif BA, Vigoreaux JO. 2016. Intrinsic disorder and multiple phosphorylations constrain the evolution of the flightin N-terminal region. *J Proteomics* **135**:191–200. doi:10.1016/j.jprot.2015.12.006

Li Jianzong, Feng Y, Wang X, Li Jing, Liu W, Rong L, Bao J. 2015. An overview of predictors for intrinsically disordered proteins over 2010–2014. *Int J Mol Sci* **16**:23446–23462. doi:10.3390/ijms161023446

Li XH, Babu MM. 2018. Human Diseases from Gain-of-Function Mutations in Disordered Protein Regions. *Cell* **175**:40–42. doi:10.1016/j.cell.2018.08.059

Li Z, Vizeacoumar FJFS, Bahr S, Li J, Warringer J, Min R, Vandersluis B, Bellay J, Devit M, Fleming J a, Stephens A, Haase J, Lin Z-Y, Baryshnikova A, Lu H, Yan Z, Jin K, Barker S, Datti A, Giaever G, Nislow C, Bulawa C, Myers CL, Costanzo M, Gingras A-C, Zhang Z, Blomberg A, Bloom K, Andrews B, Boone C. 2011. Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat Biotechnol* **29**:361–7. doi:10.1038/nbt.1832

Light S, Sagit R, Ekman D, Elofsson A. 2013. Long indels are disordered: A study of disorder and indels in homologous eukaryotic proteins. *Biochim Biophys Acta - Proteins Proteomics*

1834:890–897. doi:10.1016/j.bbapap.2013.01.002

Liu J, Faeder JR, Camacho CJ. 2009. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proc Natl Acad Sci U S A* **106**:19819–23. doi:10.1073/pnas.0907710106

Louvet E, Junéra HR, Berthuy I, Hernandez-Verdun D. 2006. Compartmentation of the nucleolar processing proteins in the granular component is a CK2-driven process. *Mol Biol Cell* **17**:2537–46. doi:10.1091/mbc.e05-10-0923

Lu AX, Zarin T, Hsu IS, Moses AM. 2019. YeastSpotter: Accurate and parameter-free web segmentation for microscopy images of yeast cells. *Bioinformatics*. doi:10.1093/bioinformatics/btz402

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564–7. doi:10.1038/35000615

Ludwig MZ, Patel NH, Kreitman M. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**:949–58.

Maccecchini ML, Rudin Y, Blobel G, Schatz G. 1979. Import of proteins into mitochondria: precursor forms of the extramitochondrially made F1-ATPase subunits in yeast. *Proc Natl Acad Sci U S A* **76**:343–7. doi:10.1073/pnas.76.1.343

Maeda T, Takekawa M, Saito H. 1995. Activation of yeast PBS2 MAPKK by MAPKKKs or by binding of an SH3-containing osmosensor. *Science* **269**:554–8.

Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu R V. 2010. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci U S A* **107**:8183–8. doi:10.1073/pnas.0911107107

Mao AH, Lyle N, Pappu R V. 2013. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem J* **449**:307–318. doi:10.1042/BJ20121346

Marsh JA, Forman-Kay JD. 2010. Sequence determinants of compaction in intrinsically

disordered proteins. *Biophys J* **98**:2374–2382. doi:10.1016/j.bpj.2010.02.012

Martin-Yken H, François JM, Zerbib D. 2016. Knr4: a disordered hub protein at the heart of fungal cell wall signalling. *Cell Microbiol* **18**:1217–1227. doi:10.1111/cmi.12618

Martin EW, Holehouse AS, Grace CR, Hughes A, Pappu R V, Mittag T. 2016. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. *J Am Chem Soc* jacs.6b10272. doi:10.1021/jacs.6b10272

Matsushima N, Tanaka T, Kretsinger R. 2009. Non-Globular Structures of Tandem Repeats in Proteins. *Protein Pept Lett* **16**:1297–1322. doi:10.2174/092986609789353745

Meggio F, Pinna LA. 2003. One-thousand-and-one substrates of protein kinase CK2? *FASEB J* **17**:349–68. doi:10.1096/fj.02-0473rev

Meyer K, Kirchner M, Uyar B, Cheng J-Y, Russo G, Hernandez-Miranda LR, Szymborska A, Zauber H, Rudolph I-M, Willnow TE, Akalin A, Haucke V, Gerhardt H, Birchmeier C, Kühn R, Krauss M, Diecke S, Pascual JM, Selbach M. 2018. Mutations in Disordered Regions Can Cause Disease by Creating Dileucine Motifs. *Cell* **0**:239–253. doi:10.1016/j.cell.2018.08.019

Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, Gruca A, Plewczynski D, Grynberg M, Bernadó P, Gáspári Z, Ouzounis CA, Promponas VJ, Kajava A V, Hancock JM, Tosatto SCE, Dosztanyi Z, Andrade-Navarro MA. 2019. Disentangling the complexity of low complexity proteins. *Brief Bioinform* **00**:1–15. doi:10.1093/bib/bbz007

Minezaki Y, Homma K, Kinjo AR, Nishikawa K. 2006. Human Transcription Factors Contain a High Fraction of Intrinsically Disordered Regions Essential for Transcriptional Regulation. *J Mol Biol* **359**:1137–1149. doi:10.1016/j.jmb.2006.04.016

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**. doi:10.1093/nar/gkt263

Mitchell SF, Jain S, She M, Parker R. 2012. Global analysis of yeast mRNPs. *Nat Struct Mol*

Biol **20**:127–133. doi:10.1038/nsmb.2468

Mittag T, Kay LE, Forman-Kay JD. 2010. Protein dynamics and conformational disorder in molecular recognition. *J Mol Recognit* **23**:105–16. doi:10.1002/jmr.961

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* **105**:17772–17777. doi:10.1073/pnas.0809222105

Moesa HA, Wakabayashi S, Nakai K, Patil A. 2012. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol Biosyst* **8**:3262. doi:10.1039/c2mb25202c

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. 2006. Analysis of Molecular Recognition Features (MoRFs). *J Mol Biol* **362**:1043–1059. doi:10.1016/j.jmb.2006.07.087

Molliex A, Temirov J, Lee J, Coughlin M, Kanagaraj AP, Kim HJ, Mittag T, Taylor JP. 2015. Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell* **163**:123–133. doi:10.1016/j.cell.2015.09.015

Morales MJ, Dang YL, Lou YC, Sulo P, Martin NC. 2006. A 105-kDa protein is required for yeast mitochondrial RNase P activity. *Proc Natl Acad Sci* **89**:9875–9879. doi:10.1073/pnas.89.20.9875

Moses AM, Hériché J-K, Durbin R. 2007a. Clustering of phosphorylation site recognition motifs can be exploited to predict the targets of cyclin-dependent kinase. *Genome Biol* **8**:R23. doi:10.1186/gb-2007-8-2-r23

Moses AM, Landry CR. 2010. Moving from transcriptional to phospho-evolution: generalizing regulatory evolution? *Trends Genet* **26**:462–7. doi:10.1016/j.tig.2010.08.002

Moses AM, Liku ME, Li JJ, Durbin R. 2007b. Regulatory evolution in proteins by turnover and lineage-specific changes of cyclin-dependent kinase consensus sites. *Proc Natl Acad Sci U S A* **104**:17713–8. doi:10.1073/pnas.0700997104

Müller-Späth S, Soranno A, Hirschfeld V, Hofmann H, Rüegger S, Reymond L, Nettels D, Schuler B. 2010. From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci U S A* **107**:14609–14. doi:10.1073/pnas.1001743107

Narayanaswamy R, Levy M, Tsechansky M, Stovall GM, O'Connell JD, Mirrielees J, Ellington AD, Marcotte EM. 2009. Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proc Natl Acad Sci* **106**:10147–10152. doi:10.1073/pnas.0812771106

Neduva V, Russell RB. 2005. Linear motifs: Evolutionary interaction switches. *FEBS Lett* **579**:3342–3345. doi:10.1016/j.febslet.2005.04.005

Nguyen Ba AN, Moses AM. 2010. Evolution of characterized phosphorylation sites in budding yeast. *Mol Biol Evol* **27**:2027–37. doi:10.1093/molbev/msq090

Nguyen Ba AN, Strome B, Hua JJ, Desmond J, Gagnon-Arsenault I, Weiss EL, Landry CR, Moses AM. 2014. Detecting Functional Divergence after Gene Duplication through Evolutionary Changes in Posttranslational Regulatory Sequences. *PLoS Comput Biol* **10**:e1003977. doi:10.1371/journal.pcbi.1003977

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. doi:10.1126/scisignal.2002515

Nido GS, Méndez R, Pascual-García a, Abia D, Bastolla U. 2012. Protein disorder in the centrosome correlates with complexity in cell types number. *Mol Biosyst* **8**:353–67. doi:10.1039/c1mb05199g

Niefind K, Yde CW, Ermakova I, Issinger OG. 2007. Evolved to Be Active: Sulfate Ions Define Substrate Recognition Sites of CK2α and Emphasise its Exceptional Role within the CMGC Family of Eukaryotic Protein Kinases. *J Mol Biol* **370**:427–438. doi:10.1016/j.jmb.2007.04.068

Nilsson J, Grahn M, Wright AP. 2011. Proteome-wide evidence for enhanced positive Darwinian

selection within intrinsically disordered regions in proteins. *Genome Biol* **12**:R65. doi:10.1186/gb-2011-12-7-r65

Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD, Baldwin AJ. 2015. Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Mol Cell* **57**:936–947. doi:10.1016/j.molcel.2015.01.013

Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. 2005. Coupled folding and binding with α-helix-forming molecular recognition elements. *Biochemistry* **44**:12454–12470. doi:10.1021/bi050736e

Ondrechen MJ, Clifton JG, Ringe D. 2001. THEMATICS: A simple computational predictor of enzyme function from structure. *Proc Natl Acad Sci* **98**:12473–12478. doi:10.1073/pnas.211436698

Osorio D, Rondon-Villarreal P, Torres R. 2015. Peptides: A Package for Data Mining of Antimicrobial Peptides. *R J* **7**:4–14.

Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2018. Biostrings: Efficient manipulation of biological strings.

Pak CWW, Kosno M, Holehouse ASS, Padrick SBB, Mittal A, Ali R, Yunus AAA, Liu DRR, Pappu RV V., Rosen MKK. 2016. Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Mol Cell* **63**:72–85. doi:10.1016/j.molcel.2016.05.042

Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}. *Bioinformatics* **xx**:xxx–xxx.

Parts L, Cubillos FA, Jain K, Warringer J, Zia A, Simpson J, Quail MA, Moses A, Edward J, Durbin R, Liti G. 2011. Genetic structure of a yeast trait revealed by sequencing a population under selection. *Genome Res* **21**:1131–1138. doi:10.1101/gr.116731.110.Freely

Patel A, Lee HO, Jawerth L, Maharana S, Jahnel M, Hein MY, Stoynov S, Mahamid J, Saha S, Franzmann TM, Pozniakovski A, Poser I, Maghelli N, Royer LA, Weigert M, Myers EW,

Grill S, Drechsel D, Hyman AA, Alberti S. 2015. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**:1066–77. doi:10.1016/j.cell.2015.07.047

Pau G, Fuchs F, Sklyar O, Boutros M, Huber W. 2010. EBImage-an R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**:979–981. doi:10.1093/bioinformatics/btq046

Pearlman SM, Serber Z, Ferrell JE. 2011. A mechanism for the evolution of phosphorylation sites. *Cell* **147**:934–946. doi:10.1016/j.cell.2011.08.052

Pemberton LF. 2014. Preparation of yeast cells for live-cell imaging and indirect immunofluorescenceMethods in Molecular Biology. pp. 79–90. doi:10.1007/978-1-4939-1363-3_6

Peng Z, Mizianty MJ, Kurgan L. 2013. Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 1–14. doi:10.1002/prot.24348

Perez RB, Tischer A, Auton M, Whitten ST. 2014. Alanine and proline content modulate global sensitivity to discrete perturbations in disordered proteins. *Proteins Struct Funct Bioinforma* **82**:3373–3384. doi:10.1002/prot.24692

Protter DSW, Rao BS, Van Treeck B, Lin Y, Mizoue L, Rosen MK, Parker R. 2018. Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly. *Cell Rep* **22**:1401–1412. doi:10.1016/j.celrep.2018.01.036

Protter DSWW, Rao BS, Van Treeck B, Lin Y, Mizoue L, Rosen MK, Parker R, Treeck B Van, Lin Y, Mizoue L, Rosen MK, Parker R. 2017. Intrinsically disordered regions contribute promiscuous interactions to RNP granule assembly. *CellReports* **22**:1401–1412. doi:10.1016/j.celrep.2018.01.036

Qian W, Zhang J. 2009. Protein Subcellular Relocalization in the Evolution of Yeast Singleton and Duplicate Genes. *Genome Biol Evol* **1**:198–204. doi:10.1093/gbe/evp021

Rackham OJL, Madera M, Armstrong CT, Vincent TL, Woolfson DN, Gough J. 2010. The Evolution and Structure Prediction of Coiled Coils across All Genomes. *J Mol Biol*

403:480–493. doi:10.1016/j.jmb.2010.08.032

Ravarani CNJ, Erkina TY, Baets G De, Dudman DC, Erkine AM, Babu MM. 2018. High-throughput discovery of functional disordered regions : investigation of transactivation domains 1–14. doi:10.15252/msb.20188190

Redfern OC, Dessailly B, Orengo CA. 2008. Exploring the structure and function paradigm. *Curr Opin Struct Biol* **18**:394–402. doi:10.1016/j.sbi.2008.05.007

Reed BJ, Locke MN, Gardner RG. 2015. A conserved deubiquitinating enzyme uses intrinsically disordered regions to scaffold multiple protein interaction sites. *J Biol Chem* **290**:20601–20612. doi:10.1074/jbc.M115.650952

Riback JA, Katanski CD, Kear-Scott JL, Pilipenko E V, Rojek AE, Sosnick TR, Drummond DA. 2017. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**:1028-1040.e19. doi:10.1016/j.cell.2017.02.027

Rokas A, Williams BI, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**:798–804. doi:10.1038/nature02053

Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. 1997. Identifying disordered regions in proteins from amino acid sequence. *IEEE Int Conf Neural Networks - Conf Proc* **1**:90–95. doi:10.1109/ICNN.1997.611643

Romero P, Obradovic Z, Kissinger CR, Villafranca JE, Garner E, Guilliot S, Dunker AK. 1998. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput* 437–48.

Romero PRM, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001. Sequence complexity of disordered protein. *Proteins* **42 1**:38–48.

Roy KR, Smith JD, Vonesch SC, Lin G, Tu CS, Lederer AR, Chu A, Suresh S, Nguyen M, Horecka J, Tripathi A, Burnett WT, Morgan MA, Schulz J, Orsley KM, Wei W, Aiyar RS, Davis RW, Bankaitis VA, Haber JE, Salit ML, St Onge RP, Steinmetz LM. 2018. Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat Biotechnol* **36**:512–520. doi:10.1038/nbt.4137

Saitoh T, Igura M, Miyazaki Y, Ose T, Maita N, Kohda D. 2011. Crystallographic snapshots of Tom20-mitochondrial presequence interactions with disulfide-stabilized peptides. *Biochemistry* **50**:5487–5496. doi:10.1021/bi200470x

Saitoh T, Igura M, Obita T, Ose T, Kojima R, Maenaka K, Endo T, Kohda D. 2007. Tom20 recognizes mitochondrial presequences through dynamic equilibrium among multiple bound states. *EMBO J* **26**:4777–4787. doi:10.1038/sj.emboj.7601888

Sawle L, Ghosh K. 2015. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys* **143**. doi:10.1063/1.4929391

Schiestl RHR, Gietz RDR. 1989. High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr Genet* **16**:339–346. doi:10.1007/BF00340712

Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. 2011. Protein disorder-a breakthrough invention of evolution? *Curr Opin Struct Biol* **21**:412–418. doi:10.1016/j.sbi.2011.03.014

Schneider C a, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**:671–675. doi:10.1038/nmeth.2089

Schwartz MF, Duong JK, Sun Z, Morrow JS, Pradhan D, Stern DF. 2002. Rad9 Phosphorylation Sites Couple Rad53 to the Saccharomyces cerevisiae DNA Damage Checkpoint. *Mol Cell* **9**:1055–1065. doi:10.1016/S1097-2765(02)00532-4

Schweers O, Schönbrunn-Hanebeck E, Marx A, Mandelkow E. 1994. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for β-structure. *J Biol Chem* **269**:24290–24297.

Sharifpoor S, Nguyen Ba AN, Young JY, van Dyk D, Friesen H, Douglas AC, Kurat CF, Chong YT, Founk K, Moses AM, Andrews BJ. 2011. A quantitative literature-curated gold standard for kinase-substrate pairs. *Genome Biol* **12**. doi:10.1186/gb-2011-12-4-r39

Sigler PB. 1988. Acid blobs and negative noodles. *Nature* **333**:210–212. doi:10.1038/333210a0

Simon M, Hancock JM. 2009. Tandem and cryptic amino acid repeats accumulate in disordered

regions of proteins. *Genome Biol* **10**:1–16. doi:10.1186/gb-2009-10-6-r59

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2009. ROCR: Visualizing the performance of scoring classifiers. *R Packag version* **1**:4.

Smolka MB, Albuquerque CP, Chen S, Zhou H. 2007. Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc Natl Acad Sci U S A* **104**:10364–10369. doi:10.1073/pnas.0701622104

Sondheimer N, Lindquist S. 2000. Rnq1: an epigenetic modifier of protein function in yeast. *Mol Cell* **5**:163–72. doi:10.1016/S1097-2765(00)80412-8

Soskine M, Tawfik DS. 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* **11**:572–582. doi:10.1038/nrg2808

Soulard A, Cremonesi A, Moes S, Schütz F, Jenö P, Hall MN. 2010. The rapamycin-sensitive phosphoproteome reveals that TOR controls protein kinase A toward some but not all substrates. *Mol Biol Cell* **21**:3475–86. doi:10.1091/mbc.E10-03-0182

Storici F, Lewis LK, Resnick M a. 2001. In vivo site-directed mutagenesis using oligonucleotides. *Nat Biotechnol* **19**:773–6. doi:10.1038/90837

Stribinskis V, Heyman H-C, Ellis SR, Steffen MC, Martin NC. 2005. Rpm2p, a Component of Yeast Mitochondrial RNase P, Acts as a Transcriptional Activator in the Nucleus. *Mol Cell Biol* **25**:6546–6558. doi:10.1128/mcb.25.15.6546-6558.2005

Stribinskis V, Ramos KS. 2007. Rpm2p, a protein subunit of mitochondrial RNase P, physically and genetically interacts with cytoplasmic processing bodies. *Nucleic Acids Res* **35**:1301–1311. doi:10.1093/nar/gkm023

Strickfaden SC, Winters MJ, Ben-Ari G, Lamson RE, Tyers M, Pryciak PM. 2007. A mechanism for cell-cycle regulation of MAP kinase signaling in a yeast differentiation pathway. *Cell* **128**:519–31. doi:10.1016/j.cell.2006.12.032

Sumner J. 1926. The Isolation and Crystallization of the Enzyme Urease. *J Biol Chem* **69**:435–441.

Sun MGF, Sikora M, Costanzo M, Boone C, Kim PM. 2012. Network evolution: Rewiring and signatures of conservation in signaling. *PLoS Comput Biol* **8**. doi:10.1371/journal.pcbi.1002411

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**:609–612. doi:10.1093/nar/gkl315

Szabo, Horvath, Schad, Murvai, Tantos, Kalmar, Chemes, Han, Tompa. 2019. Intrinsically Disordered Linkers Impart Processivity on Enzymes by Spatial Confinement of Binding Domains. *Int J Mol Sci* **20**:2119. doi:10.3390/ijms20092119

Tang X, Orlicky S, Mittag T, Csizmok V, Pawson T, Forman-Kay JD, Sicheri F, Tyers M. 2012. Composite low affinity interactions dictate recognition of the cyclin-dependent kinase inhibitor Sic1 by the SCFCdc4 ubiquitin ligase. *Proc Natl Acad Sci* **109**:3287–3292. doi:10.1073/pnas.1116455109

Tatebayashi K, Tanaka K, Yang H-Y, Yamamoto K, Matsushita Y, Tomida T, Imai M, Saito H. 2007. Transmembrane mucins Hkr1 and Msb2 are putative osmosensors in the SHO1 branch of yeast HOG pathway. *EMBO J* **26**:3521–33. doi:10.1038/sj.emboj.7601796

Tatebayashi K, Yamamoto K, Tanaka K, Tomida T, Maruoka T, Kasukawa E, Saito H. 2006. Adaptor functions of Cdc42, Ste50, and Sho1 in the yeast osmoregulatory HOG MAPK pathway. *EMBO J* **25**:3033–44. doi:10.1038/sj.emboj.7601192

Taylor JS, Raes J. 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. *Annu Rev Genet* **38**:615–643. doi:10.1146/annurev.genet.38.072902.092831

Teixeira D, Parker R. 2007. Analysis of P-body assembly in Saccharomyces cerevisiae. *Mol Biol Cell* **18**:2274–87. doi:10.1091/mbc.e07-03-0199

Terry LJ, Wente SR. 2009. Flexible gates: Dynamic topologies and functions for FG nucleoporins in nucleocytoplasmic transport. *Eukaryot Cell* **8**:1814–1827. doi:10.1128/EC.00225-09

Tomasso ME, Tarver MJ, Devarajan D, Whitten ST. 2016. Hydrodynamic Radii of Intrinsically

Disordered Proteins Determined from Experimental Polyproline II Propensities. *PLoS Comput Biol* **12**:1–22. doi:10.1371/journal.pcbi.1004686

Tompa P. 2014. Multisteric regulation by structural disorder in modular signaling proteins: An extension of the concept of allostery. *Chem Rev* **114**:6715–6732. doi:10.1021/cr4005082

Tompa P, Davey NE, Gibson TJ, Babu MM. 2014. A Million peptide motifs for the molecular biologist. *Mol Cell* **55**:161–169. doi:10.1016/j.molcel.2014.05.032

Tompa P, Schad E, Tantos A, Kalmar L. 2015. Intrinsically disordered proteins: Emerging interaction specialists. *Curr Opin Struct Biol* **35**:49–59. doi:10.1016/j.sbi.2015.08.009

Tóth-Petróczy A, Tawfik DS. 2013. Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol Biol Evol* **30**:761–71. doi:10.1093/molbev/mst003

Townsend RR, Lipniunas PH, Tulk BM, Verkman AS. 1996. Identification of protein kinase a phosphorylation sites on NBD1 and R domains of CFTR using electrospray mass spectrometry with selective phosphate ion monitoring. *Protein Sci* **5**:1865–1873. doi:10.1002/pro.5560050912

Treusch S, Lindquist S. 2012. An intrinsically disordered yeast prion arrests the cell cycle by sequestering a spindle pole body component. *J Cell Biol* **197**:369–379. doi:10.1083/jcb.201108146

Truckses DM, Bloomekatz JE, Thorner J. 2006. The RA domain of Ste50 adaptor protein is required for delivery of Ste11 to the plasma membrane in the filamentous growth signaling pathway of the yeast Saccharomyces cerevisiae. *Mol Cell Biol* **26**:912–28. doi:10.1128/MCB.26.3.912-928.2006

Tusnády GE, Dobson L, Tompa P. 2015. Disordered regions in transmembrane proteins. *Biochim Biophys Acta - Biomembr* **1848**:2839–2848. doi:10.1016/j.bbamem.2015.08.002

Uversky VN. 2013. A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Sci* **22**:693–724. doi:10.1002/pro.2261

Uversky VN. 2011. Intrinsically disordered proteins from A to Z. *Int J Biochem Cell Biol*

43:1090–103. doi:10.1016/j.biocel.2011.04.001

Uversky VN. 2002. Natively unfolded proteins: A point where biology waits for physics. *Protein Sci* **11**:739–756. doi:10.1110/ps.4210102

Uversky VN, Gillespie JR, Fink AL. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins Struct Funct Genet* **41**:415–427. doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7

Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK. 2007. Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* **6**:2351–2366. doi:10.1021/pr0701411

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu R V., Tompa P, Uversky VN, Wright PE, Babu MM. 2014. Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**:6589–6631. doi:10.1021/cr400525m

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. doi:10.1016/j.cell.2009.04.029

Vernon RM, Chong PA, Tsang B, Kim TH, Bah A, Farber P, Lin H, Forman-Kay JD. 2018. Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**:1–48. doi:10.7554/eLife.31486

Vögtle FN, Wortelkamp S, Zahedi RP, Becker D, Leidhold C, Gevaert K, Kellermann J, Voos W, Sickmann A, Pfanner N, Meisinger C. 2009. Global Analysis of the Mitochondrial N-Proteome Identifies a Processing Peptidase Critical for Protein Stability. *Cell* **139**:428–439. doi:10.1016/j.cell.2009.07.045

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**. doi:10.1371/journal.pbio.1001446

Wallace EWJ, Kear-Scott JL, Pilipenko E V., Schwartz MH, Laskowski PR, Rojek AE, Katanski CD, Riback JA, Dion MF, Franks AM, Airoldi EM, Pan T, Budnik BA, Drummond DA. 2015. Reversible, Specific, Active Aggregates of Endogenous Proteins Assemble upon Heat Stress. *Cell* **162**:1286–1298. doi:10.1016/j.cell.2015.08.041

Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, Maharana S, Lemaitre R, Pozniakovsky A, Drechsel D, Poser I, Pappu R V, Alberti S, Hyman AA. 2018. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**:688-699.e16. doi:10.1016/j.cell.2018.06.006

Wang Y, Chu X, Longhi S, Roche P, Han W, Wang E, Wang J. 2013. Multiscaled exploration of coupled folding and binding of an intrinsically disordered molecular recognition element in measles virus nucleoprotein. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1308381110

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**:635–45. doi:10.1016/j.jmb.2004.02.002

Warren C, Shechter D. 2017. Fly Fishing for Histones: Catch and Release by Histone Chaperone Intrinsically Disordered Regions and Acidic Stretches. *J Mol Biol* **429**:2401–2426. doi:10.1016/j.jmb.2017.06.005

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:1189–1191. doi:10.1093/bioinformatics/btp033

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A,

Fewell G a, Flicek P, Foley K, Frankel WN, Fulton L a, Fulton RS, Furey TS, Gage D, Gibbs R a, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves T a, Green ED, Gregory S, Guigó R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones T a, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe B a, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang S-P, Zdobnov EM, Zody MC, Lander ES. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**:520–62. doi:10.1038/nature01262

Weirauch MT, Hughes TR. 2010. Dramatic changes in transcription factor binding over evolutionary time. *Genome Biol* **11**:122. doi:10.1186/gb-2010-11-6-122

Wickham H. 2010. Stringr: Modern, Consistent String Processing. *R J* **2**:38–40.

Wohlbach DJ, Thompson DA, Gasch AP, Regev A. 2009. From elements to modules: regulatory evolution in Ascomycota fungi. *Curr Opin Genet Dev* **19**:571–8. doi:10.1016/j.gde.2009.09.007

Wolfe KH, Li WH. 2003. Molecular evolution meets the genomics revolution. *Nat Genet*

33:255–265. doi:10.1038/ng1088

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**:708–13.

Wootton JC. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**:269–85.

Wootton JC, Federhen S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* **17**:149–163. doi:10.1016/0097-8485(93)85006-X

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* **8**:206–16. doi:10.1038/nrg2063

Wright PE, Dyson HJ. 2014. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**:18–29. doi:10.1038/nrm3920

Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J Mol Biol* **293**:321–331. doi:10.1006/jmbi.1999.3110

Xiao N, Cao DS, Zhu MF, Xu QS. 2015. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics* **31**:1857–1859. doi:10.1093/bioinformatics/btv042

Xue B, Dunker AK, Uversky VN. 2012. Orderly order in protein intrinsic disorder distribution: Disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**:137–149. doi:10.1080/07391102.2012.675145

Yamamoto K, Tatebayashi K, Tanaka K, Saito H. 2010. Dynamic control of yeast MAP kinase network by induced association and dissociation between the Ste50 scaffold and the Opy2 membrane anchor. *Mol Cell* **40**:87–98. doi:10.1016/j.molcel.2010.09.011

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Yeung M, Durocher D. 2011. Srs2 enables checkpoint recovery by promoting disassembly of

DNA damage foci from chromatin. *DNA Repair (Amst)* **10**:1213–1222. doi:10.1016/j.dnarep.2011.09.005

Zarin T, Moses AM. 2014. Insights into molecular evolution from yeast genomics. *Yeast* **31**:233–41. doi:10.1002/yea.3018

Zarin T, Tsai CN, Nguyen Ba AN, Moses AM. 2017. Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc Natl Acad Sci* **114**:E1450–E1459. doi:10.1073/pnas.1614787114

Zheng J, Yang J, Choe YJ, Hao X, Cao X, Zhao Q, Zhang Y, Franssens V, Hartl FU, Nyström T, Winderickx J, Liu B. 2017. Role of the ribosomal quality control machinery in nucleocytoplasmic translocation of polyQ-expanded huntingtin exon-1. *Biochem Biophys Res Commun* **493**:708–717. doi:10.1016/j.bbrc.2017.08.126