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

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

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.

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 (Table S1). 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.

Results

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 (Fig. S1). We also confirmed that the vast majority of these IDRs are distinct from Pfam domains (Fig. S2). 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 (Table S1) 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 (Fig. 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 $p<0.01$, Fig. 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) (Fig. 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 (Fig. 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).

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 $p<0.01$, Fig. 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.

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 intrinsically disordered regions (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) (Fig. 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, Fig. 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 $p<0.05$, Fig. 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 (Fig. 2C; full micrographs in Fig. S3). 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.

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) (Fig. 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, Fig. 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 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 (Table S2; see Methods).

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 (Fig. 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 Fig. 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 (Fig. 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.

We also find functions associated with our clusters that have not been previously associated with molecular features of IDRs. For example, cluster D (Fig. 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 (Fig. 5A), which is also clear in the comparison between the real Srs2 orthologs and the simulation (Fig. 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.

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 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 (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).

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 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) (Fig. 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 (Fig. S4A) (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 (Fig. S4B). 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 $p \leq 6.5x10^{-11}$, Wilcoxon test) (Fig. 6B, red box). Interestingly, the adjacent cluster V (Fig. 6B, purple box), which we hypothesize to contain targeting sequences for the endoplasmic reticulum, is distinct from cluster W in this analysis.

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) (Fig. 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 (Fig. 6C; full micrographs in Fig. S5). We also repeated this experiment with another protein that has an experimentally verified N-terminal mitochondrial targeting sequence, Mdl2, and found the same results (Fig. S6). 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.

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 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 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.

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; 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.

Methods

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).

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 Fig. S1.

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 Fig. S2.

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 Fig. 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:

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

Strain construction and growth conditions

All strains (Table S3) 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.

Confocal microscopy and image analysis

We acquired all images with a Leica TCS SP8 microscope using standard, uncoated glass slides with a 100x objective. To quantify basal Fus1pr-GFP expression, single cells in micrographs were segmented using BeerGoggles (http://beergoggles.csb.utoronto.ca/). 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://beergoggles.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.

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 similarity 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 percent identity in all the clusters. This metric is presented for each cluster in Table S2. 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.

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 Table S2.

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 (Table S2).

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 2.

Acknowledgements

We thank Alex X Lu, 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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, Cyert MS, Moses AM. 2015. Short linear motifs – ex nihilo evolution of protein regulation. *Cell Commun Signal* **13**:43. doi:10.1186/s12964-015-0120-z

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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, 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

Figures

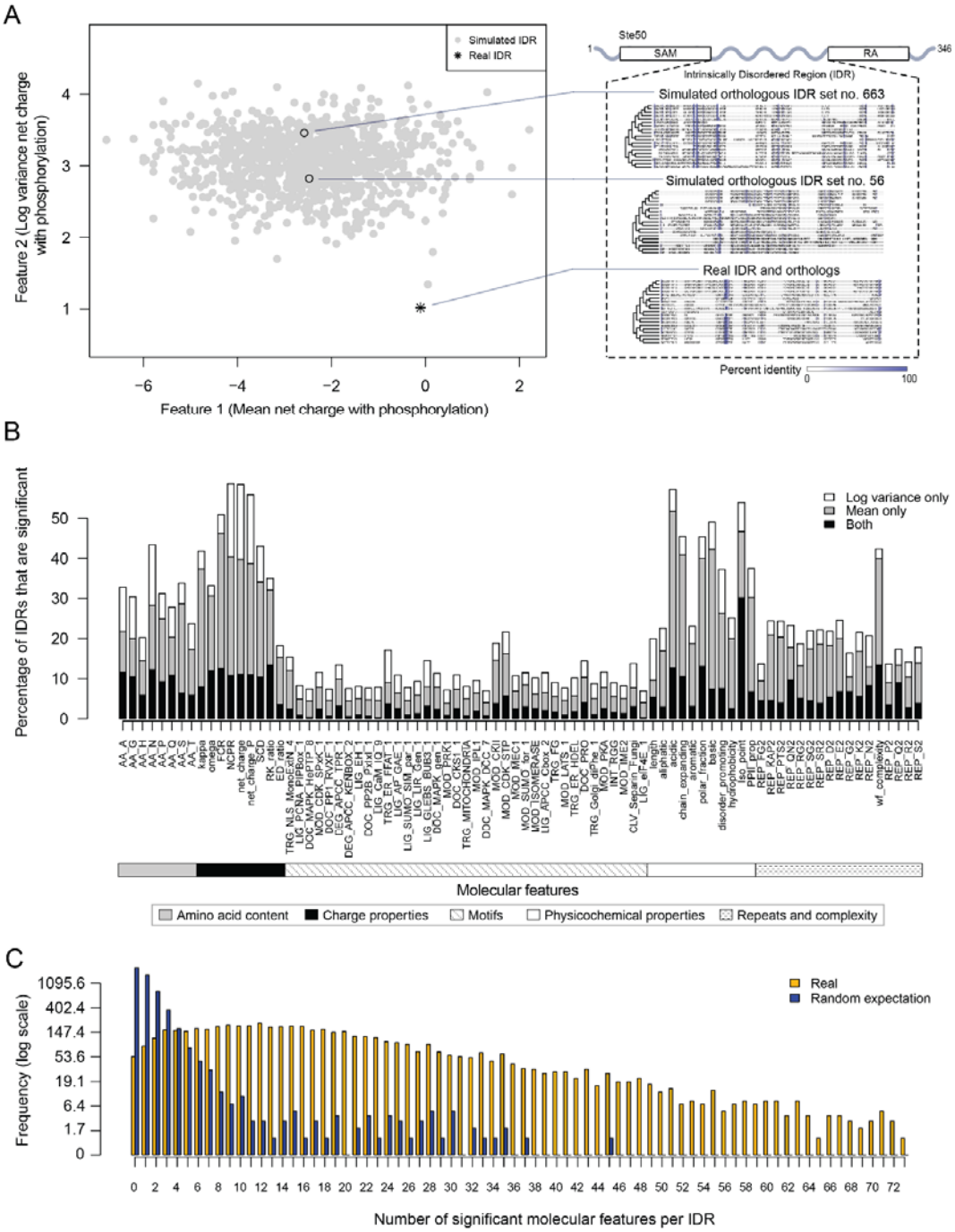

Figure 1. Proteome-wide evolutionary analysis reveals evolutionarily constrained sequence features are widespread in highly diverged intrinsically disordered regions. A) Left: Mean versus log variance of the “net charge with phosphorylation” molecular feature for the real Ste50 IDR (a.a. 152-250) ortholog 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.

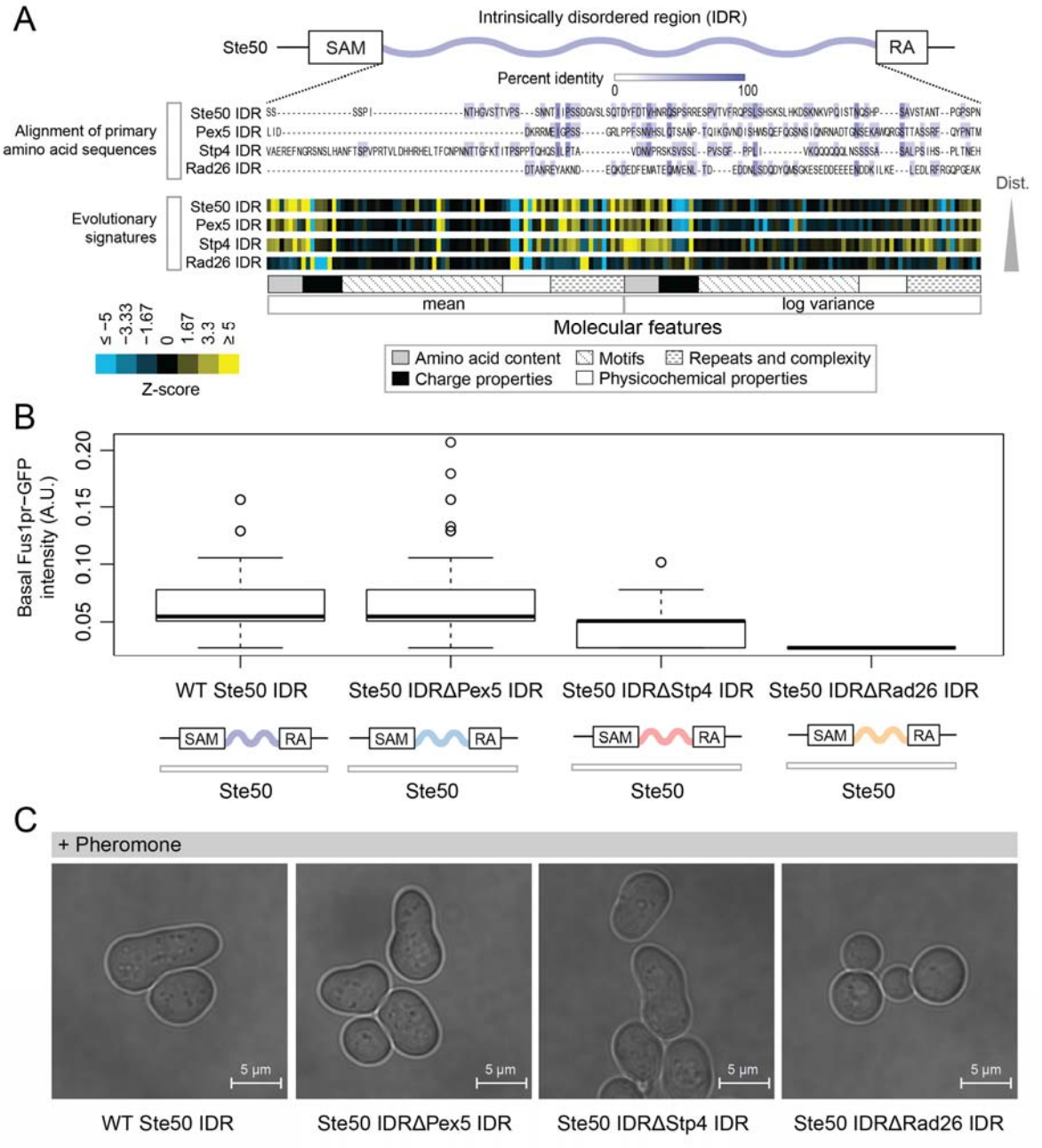

Figure 2. Intrinsically disordered regions with similar evolutionary signatures can rescue wildtype phenotypes, while those with different evolutionary signatures cannot. A) Multiple sequence alignment of Ste50 IDR (a.a. 152-250), Pex5 IDR (a.a. 77-161), Stp4 (a.a. 144-256), and Rad26 IDR (a.a. 163-239) 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 25th-75th percentile of the data, the black line represents the median, and whiskers represent 1.5*the interquartile range. Outliers are represented by unfilled circles. 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.

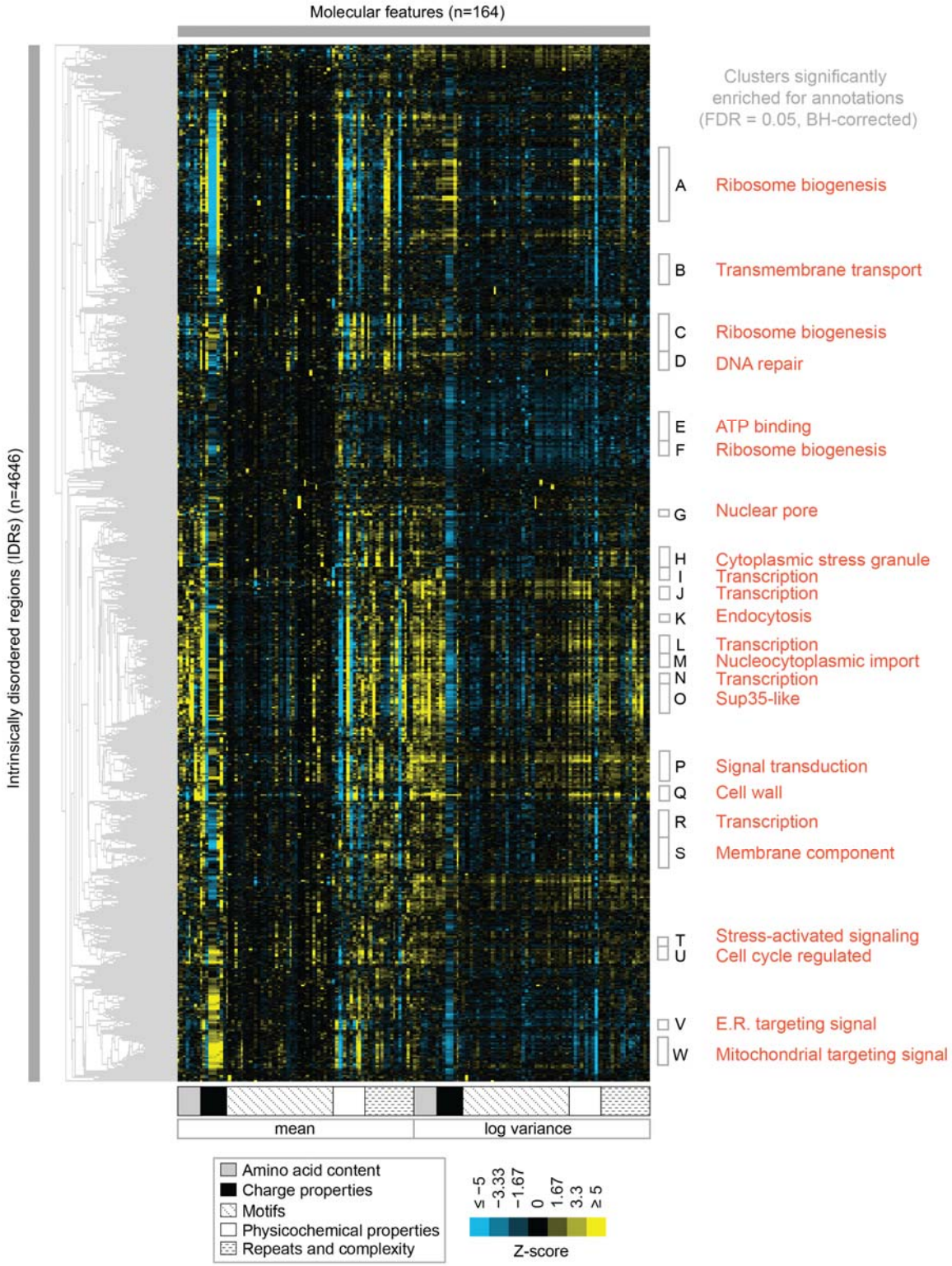

Figure 3. Clustering evolutionary signatures shows that IDRs in the proteome share evolutionary signatures, and that these clusters of IDRs are associated with specific biological functions. A-W show clusters significantly enriched for annotations (see Table 1; full table of enrichments in supplementary data). Cluster names represent summary of enriched annotations.

Table 1. Top 5 enriched GO term annotations and top 3 enriched phenotype annotations for each cluster. Full table of >1300 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

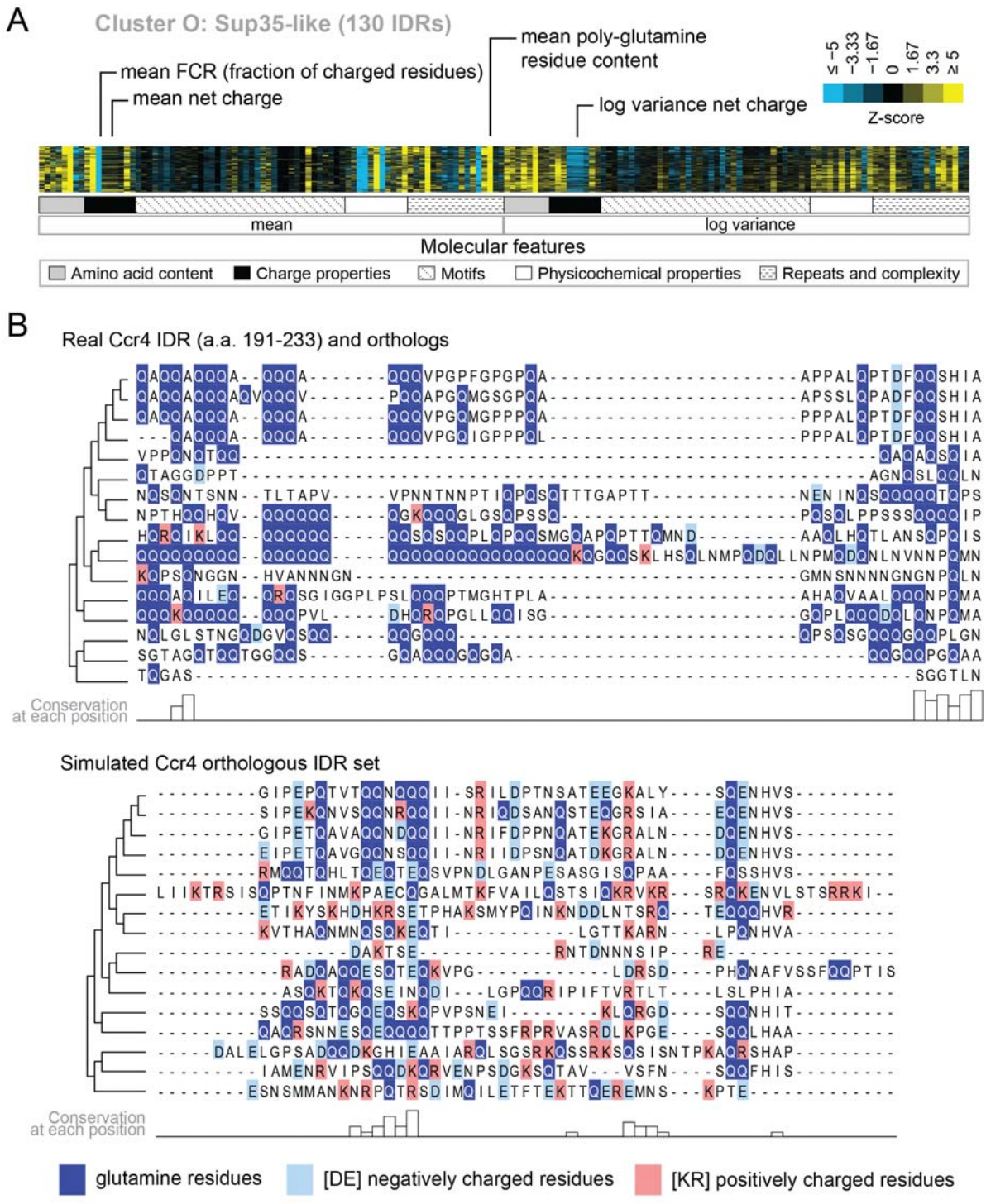

Figure 4. Evolutionary signatures in cluster O contain some molecular features that are typically associated with IDRs as well as some that are not. A) Pattern of evolutionary signatures in cluster O. B) Example disordered region from cluster O, Ccr4, with a subset of highlighted molecular features compared 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*.

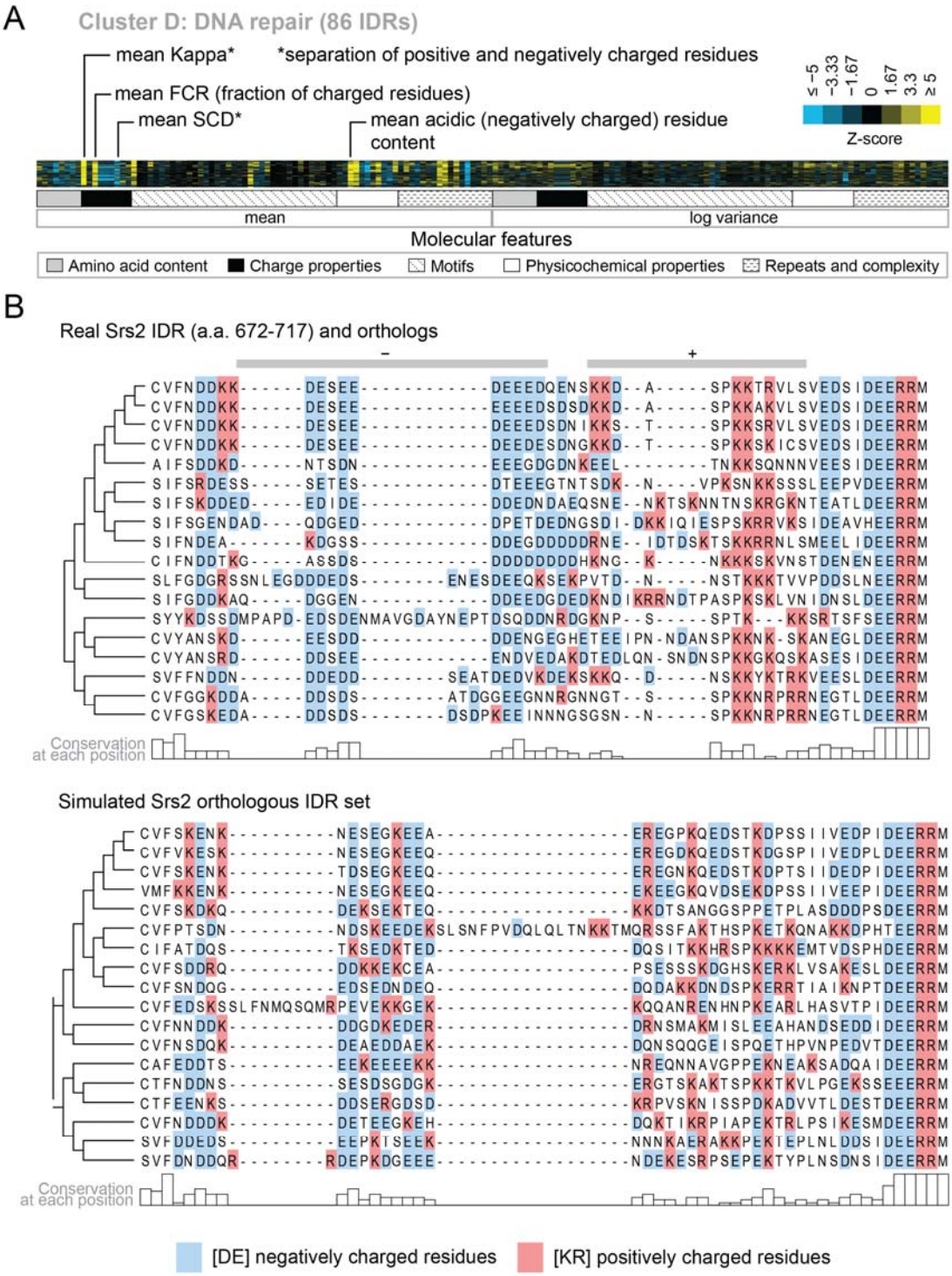

Figure 5. Cluster D contains disordered regions associated with DNA repair. A) Pattern of evolutionary signatures in cluster D. B) Example disordered region from cluster D, Srs2, with a subset of highlighted molecular features compared between its real set of orthologs and an example set of simulated 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*.

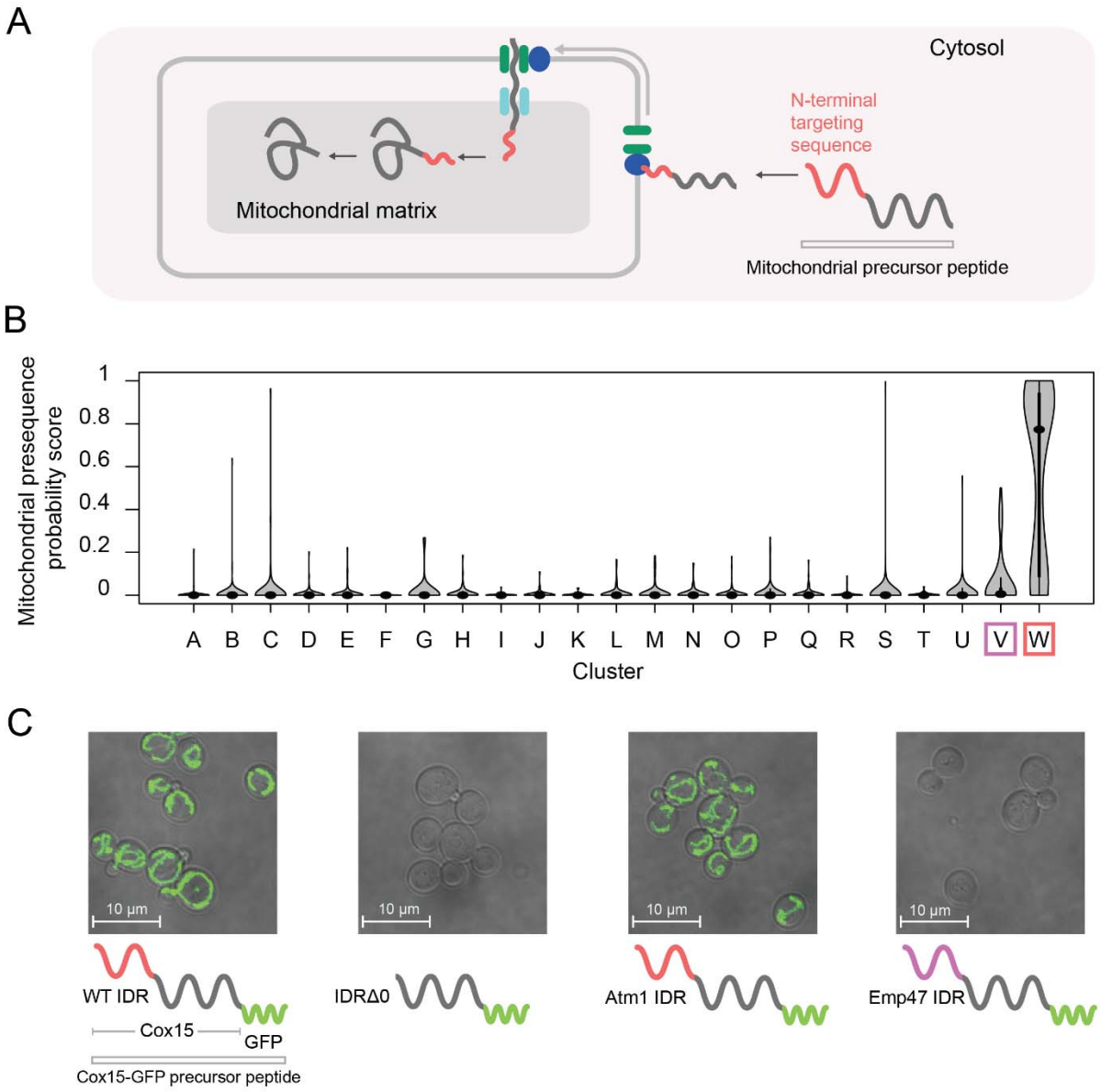

Figure 6. Cluster W is associated with mitochondrial N-terminal targeting signals. A) Schematic (not to scale) showing the path of a mitochondrial precursor peptide (with N-terminal targeting sequence in red) from the cytosol, where it is translated, to the mitochondrial matrix, where the peptide folds 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).

Table 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 cytoplasm in a large-scale study	89	A: Ribosome biogenesis 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

Supplementary figures

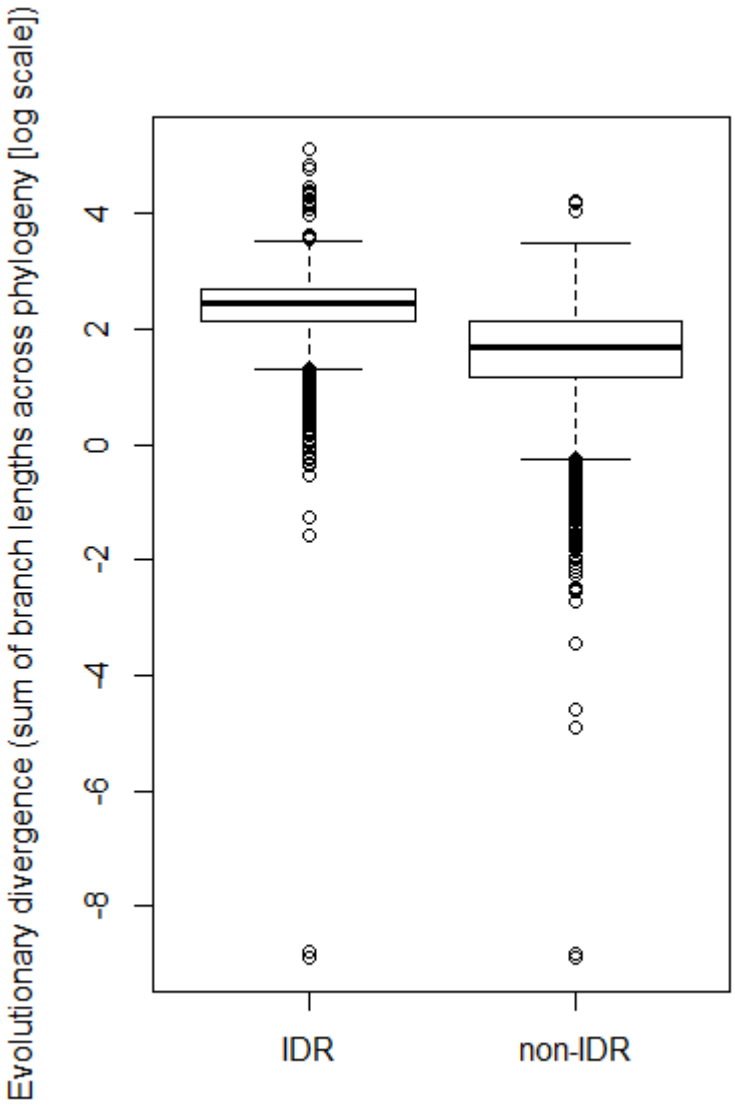

Figure S1. Predicted IDRs in the *S.cerevisiae* proteome ("IDR") are more highly diverged compared to regions that are not predicted to be disordered ("non-IDR") ($p<2.2x10^{-16}$, Wilcoxon test).

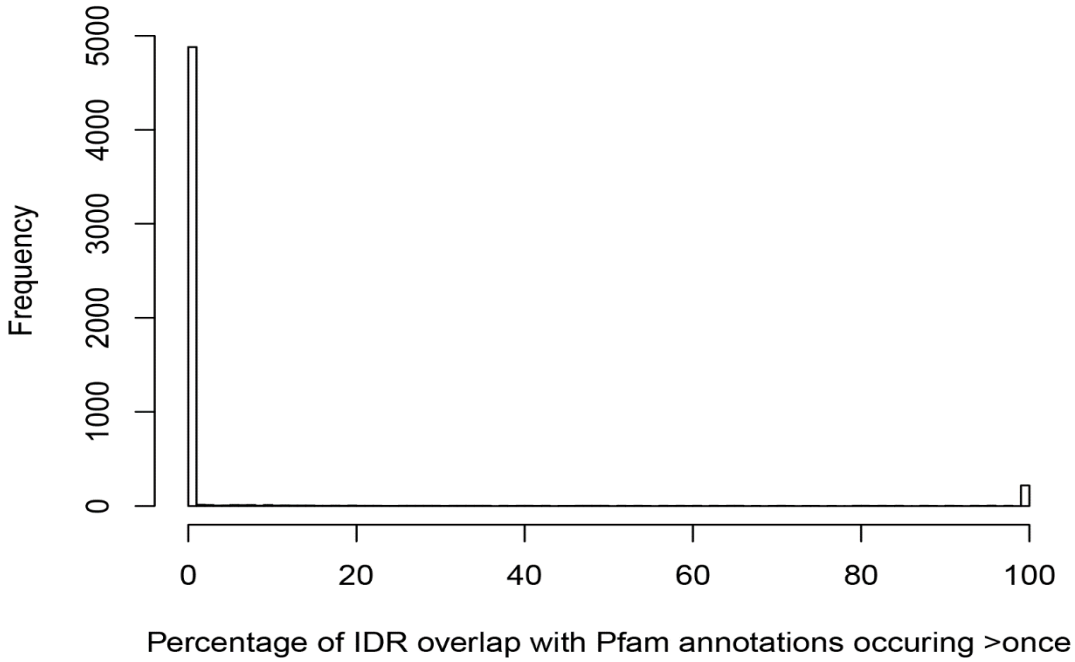

Figure S2. The vast majority of predicted IDRs in the *S.cerevisiae* proteome do not overlap with Pfam domains.

Table S1. 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 content	Q	Amino acid content	NA	Fraction of Q residues	(Alberti et al., 2009; 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_IQ_9	Helical calmodulin binding motif	[ACLIVTM][^P][^P][ILVMFCT]Q[^P][^P][^P][RK][^P]{4,5}[RKQ][^P][^P]	Motifs	ELM	Helical peptide motif responsible for Ca2+-independent binding of the 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	(Dinkel et al., 2016)
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_par_1	SUMO interaction site	[DEST]{0,5}.[VILPTM][VIL][DESTVILMA][VIL].{0,1}[DEST]{1,10}	Motifs	ELM	Motif for the parallel beta augmentation mode of non-covalent binding to SUMO protein.	(Dinkel et al., 2016)

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_4	NLS classical Nuclear Localization Signals	(([PKR].{0,1}[^DE])|([PKR]))((K[RK])|(RK))(([^DE][KR])|([KR][^DE]))[^DE]	Motifs	ELM	Monopartite variant of the classical basically charged NLS. N-extended version.	(Dinkel et al., 2016)
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 motif	[ST][DE].[DE]	Motifs	Condens	NA	(A. C. W. Lai et al., 2012; Meggio and 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	(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 content	[FYW]	Physicochemical properties	Literature /localCIDER	NA	(Holehouse et al., 2017)
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_E2	E repeat	E{2,}	Repeats and complexity	Literature	Fraction of 2 or more E in a row	(Chavali et 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	(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)

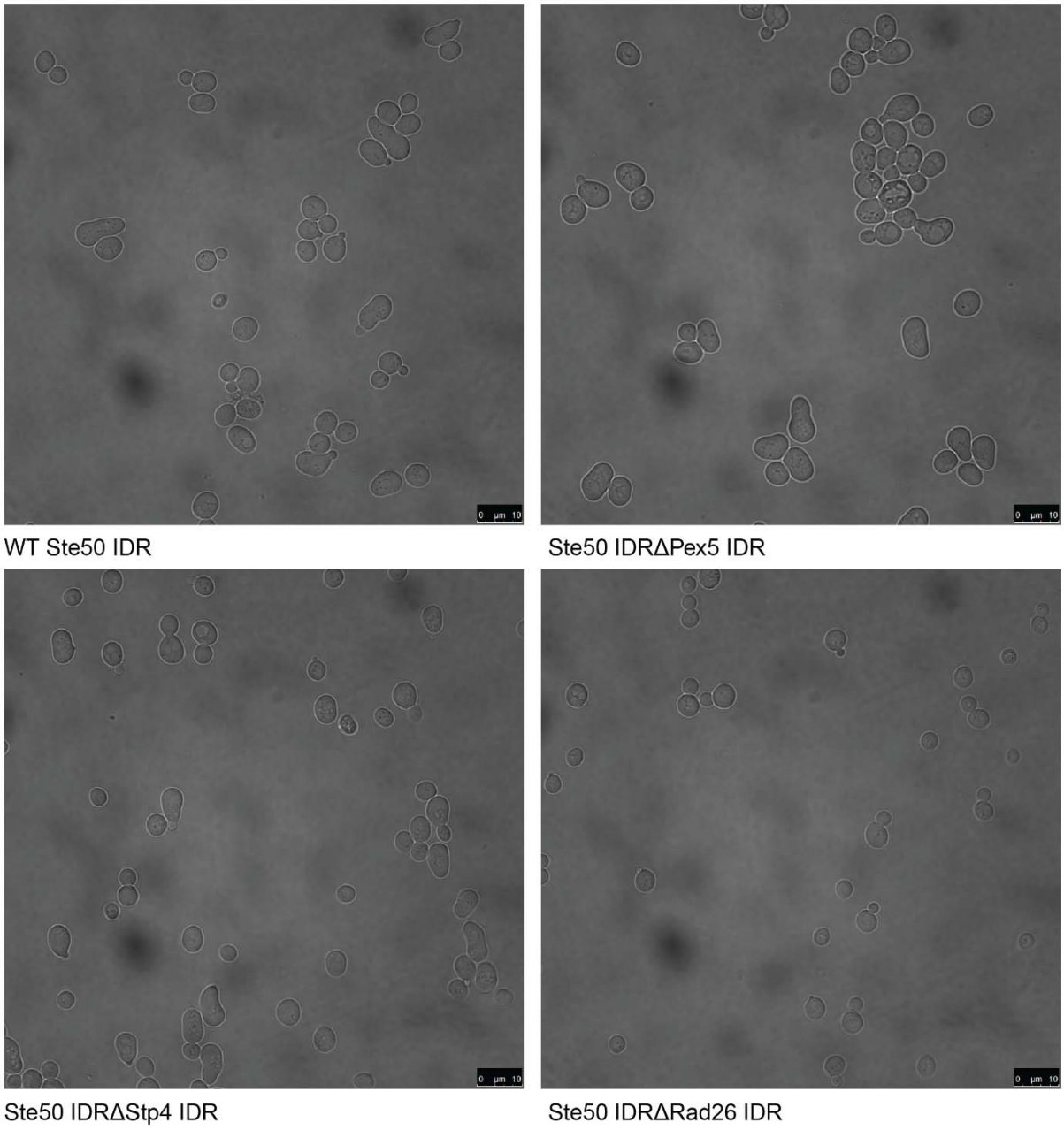

Figure S3. Full field-of-view micrographs of pheromone-exposed *S.cerevisiae* strains from Fig. 2C.

Table S2. 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

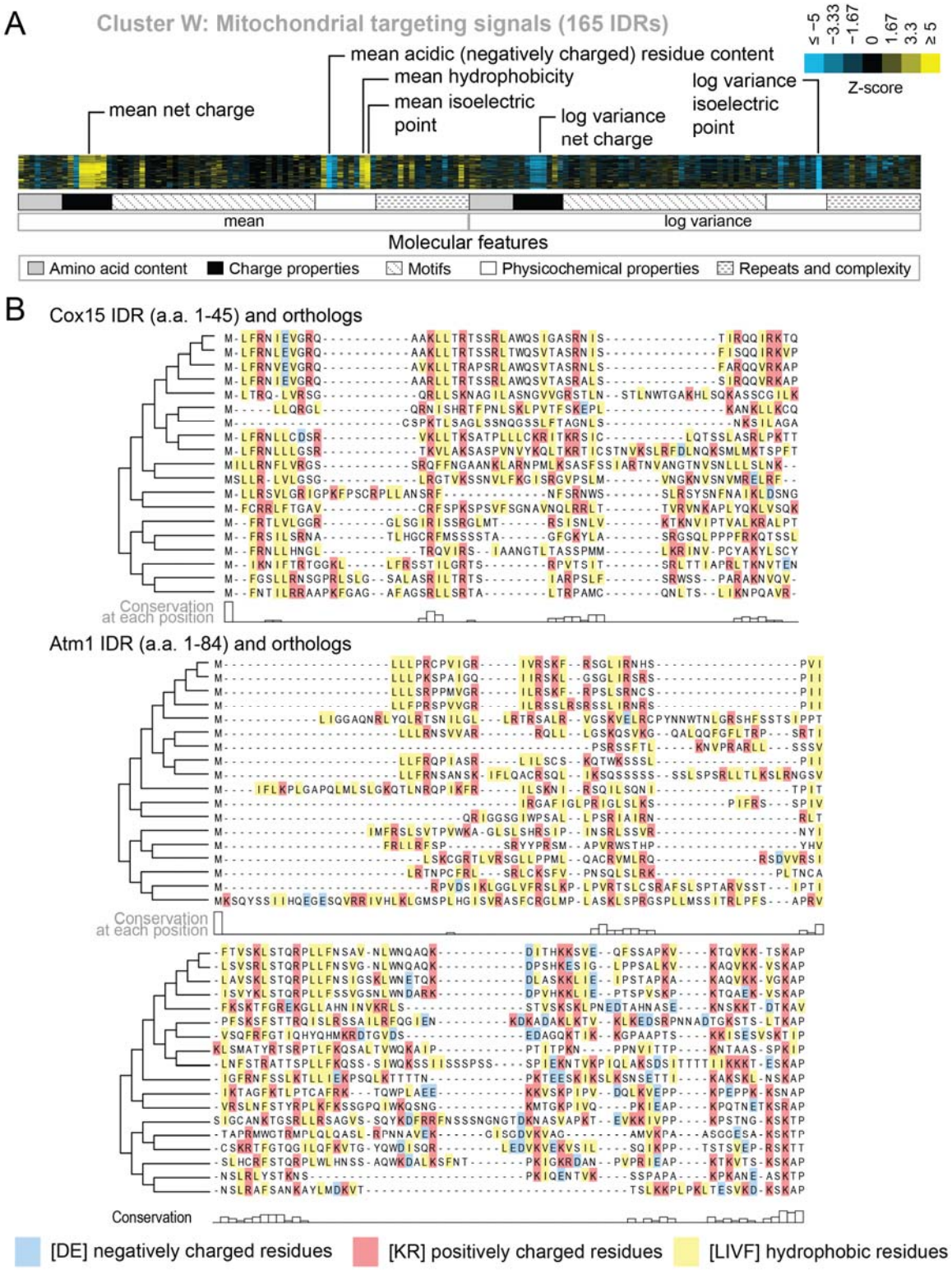

Figure S4. Evolutionary signatures in cluster W contain molecular features that have been previously reported for mitochondrial N-terminal targeting signals. A) Pattern of evolutionary signatures in cluster W. B) Multiple sequence alignments of example disordered regions from Cox15 (top) and Atm1 (bottom) 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*.

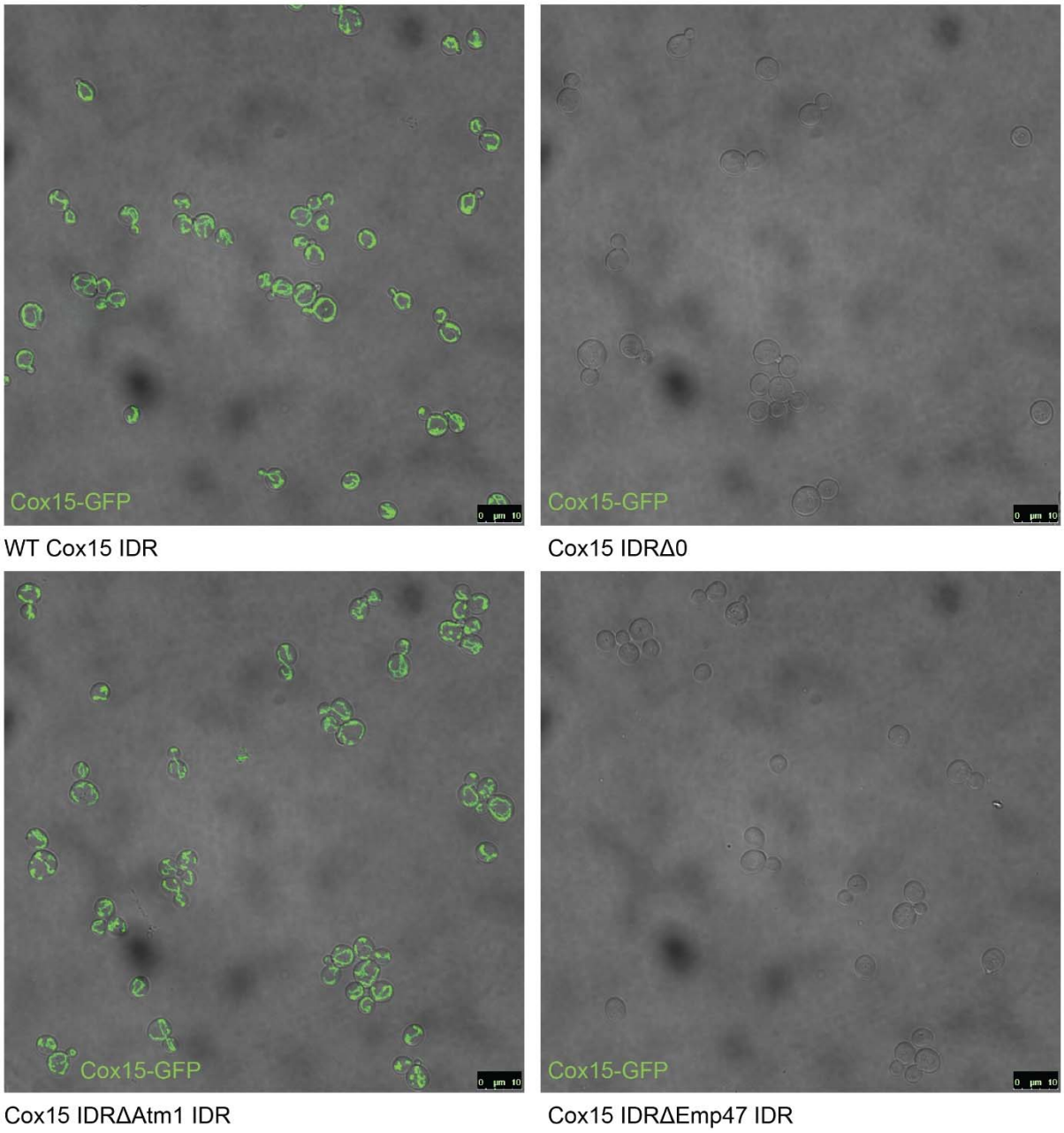

Figure S5. Full field-of-view micrographs of *S.cerevisiae* strains from Fig. 6C.

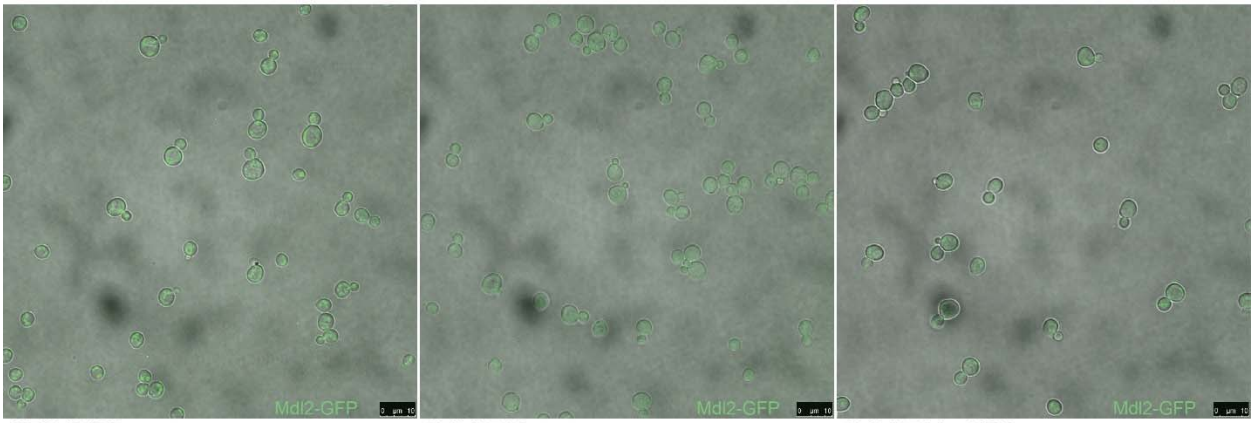

Figure S6. Micrographs of *S.cerevisiae* strains with three different genotypes. From left to right: Mdl2-GFP has a mitochondrial localization in the wildtype (WT) strain, knocking out the Mdl2 IDR abolishes wildtype localization, and replacing the Mdl2 IDR with that of Atm1 rescues mitochondrial localization.

Table S3. 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. 144-256) HO::pFUS1-yemGFP-klURA3	This study

Supplementary references

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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 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

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, 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

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

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

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

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

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

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

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

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

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

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

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

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

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