RESEARCH ARTICLE

Figures and figure supplements

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

Taraneh Zarin *et al*

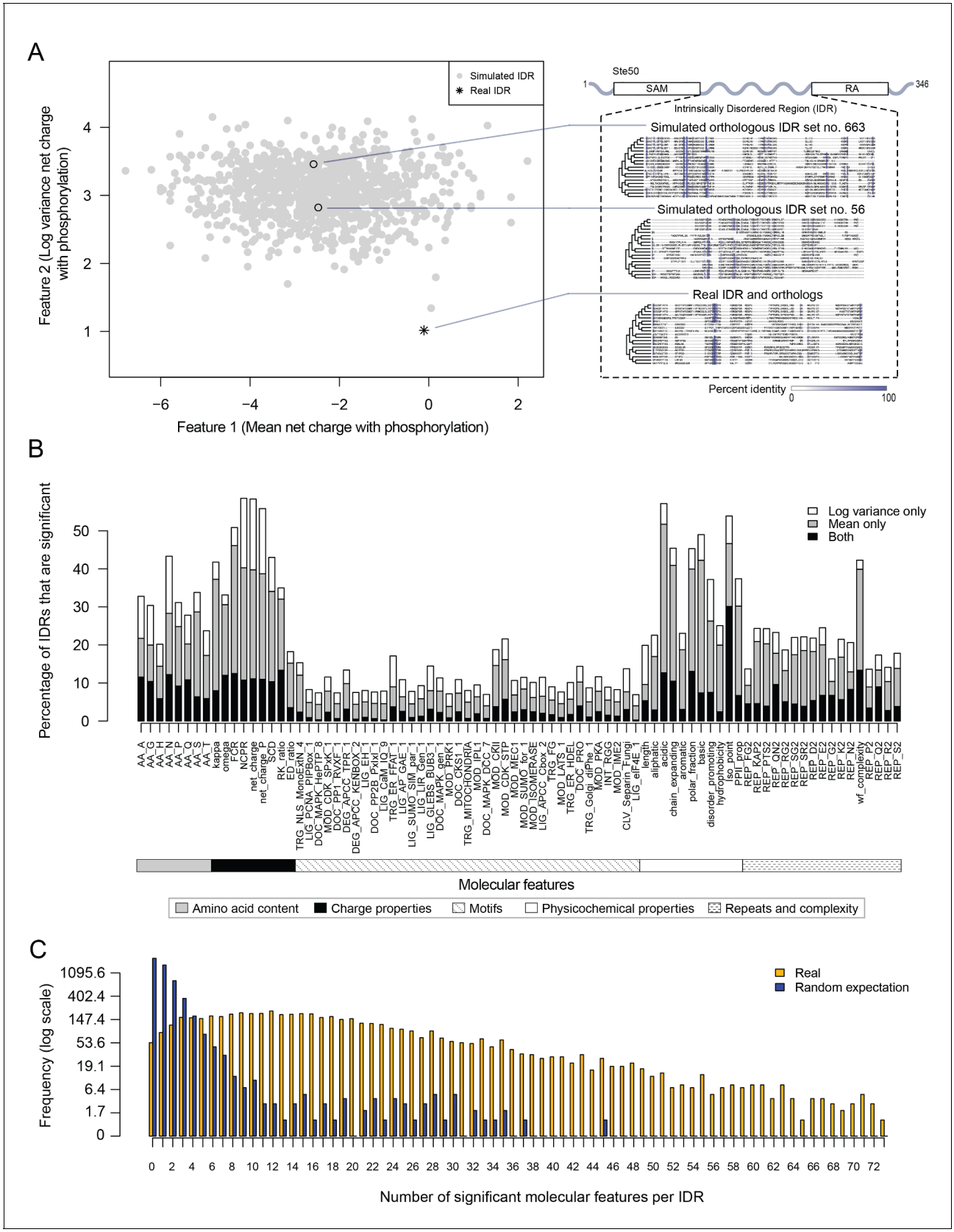

Figure 1. Proteome-wide evolutionary analysis reveals evolutionarily constrained sequence features are widespread in highly diverged intrinsically disordered regions. (**A**) Left: Mean versus log variance of the 'net charge with phosphorylation' molecular feature for the real Ste50 IDR (a.a. 152–250)

Figure 1 continued on next page

Figure 1 continued

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, colored 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.

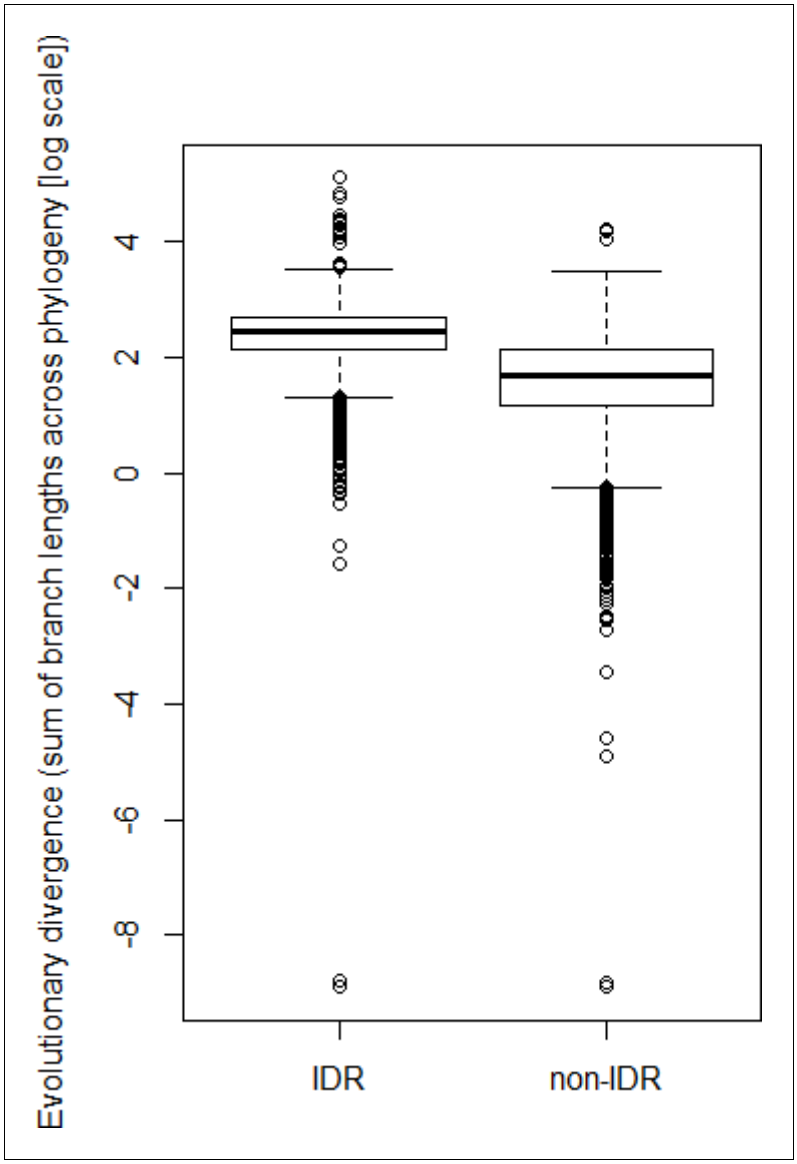

Figure 1—figure supplement 1. Predicted IDRs in the *S. cerevisiae* proteome ('IDR') are more highly diverged compared to regions that are not predicted to be disordered ('non-IDR') ($p<2.2\times10^{-16}$, Wilcoxon test). Boxplot boxes represent the 25th-75th percentile of the data, the black line represents the median, and whiskers represent 1.5*the interquartile range. Outliers fall outside the 1.5*interquartile range, and are represented by unfilled circles.

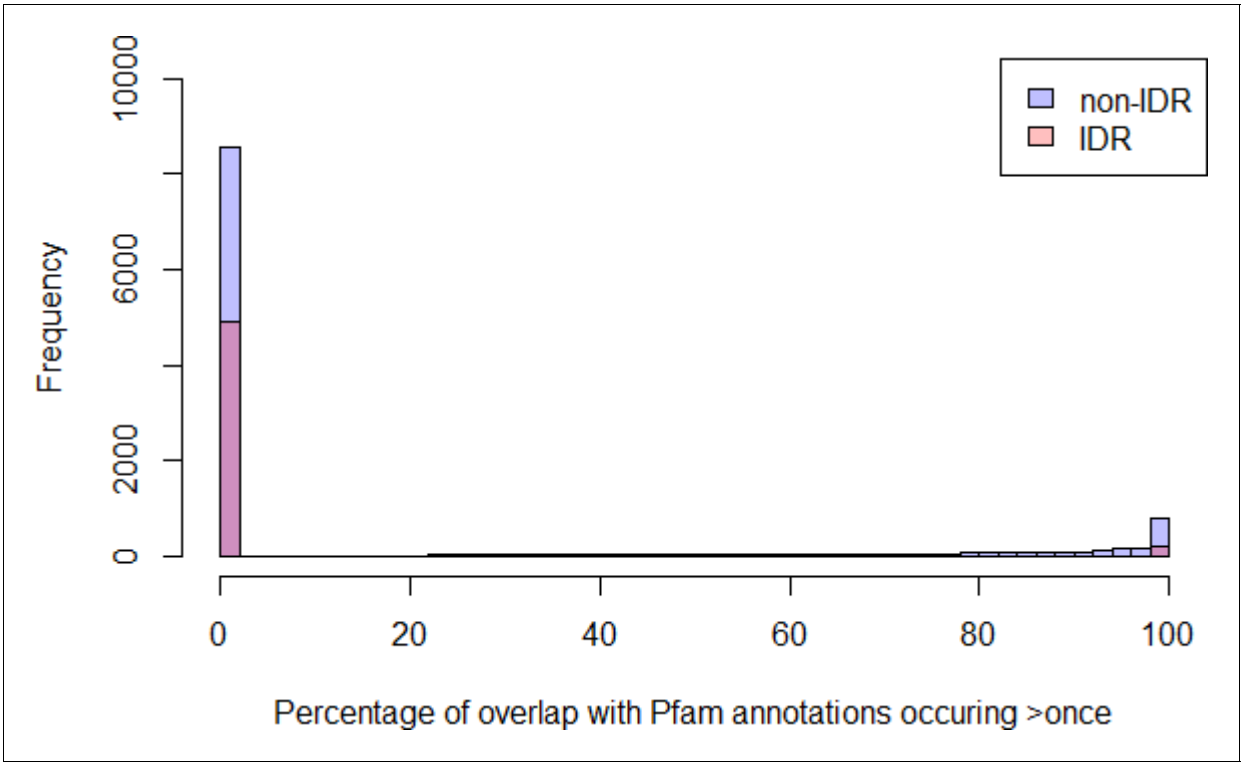

Figure 1—figure supplement 2. Percentage of overlap with Pfam domains for IDRs predicted to be disordered in the *S. cerevisiae* proteome that are >= 30 amino acids ('IDR') have less overlap with Pfam domains compared to all other regions that are >= 30 amino acids ('non-IDR') ($p<2.2 \times 10^{-16}$, Wilcoxon test). Percentage of regions with 0% Pfam overlap for IDRs is 91%, whereas for non-IDRs it is 74%.

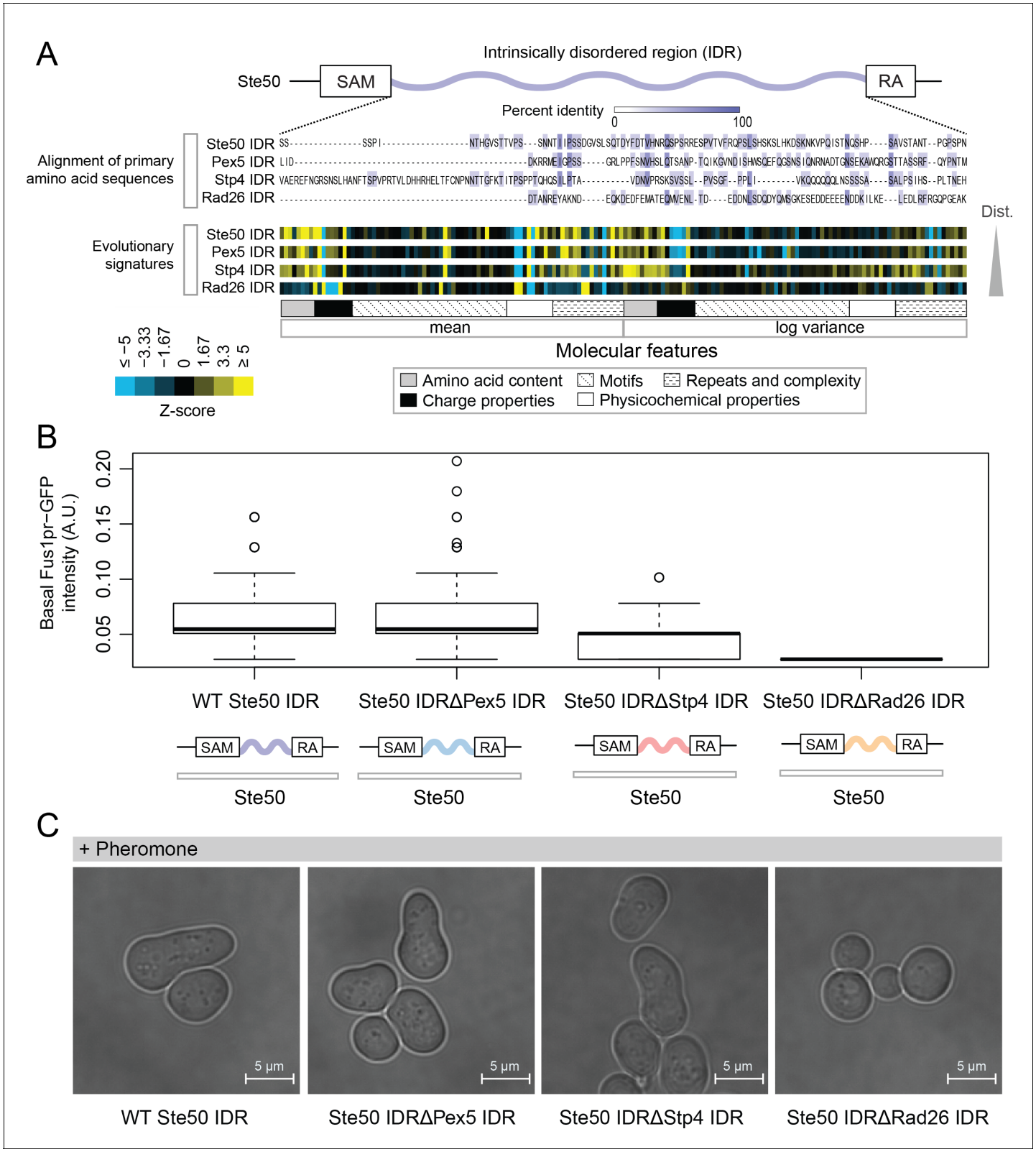

Figure 2. Intrinsically disordered regions with similar evolutionary signatures can rescue wildtype phenotypes, while those with different evolutionary signatures cannot. (**A**) Multiple sequence alignment of Ste50 IDR (a.a. 152–250), Pex5 IDR (a.a. 77–161), Stp4 (a.a. 144–256), and Rad26 IDR (a.a. 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. While the Ste50 IDR has five consensus phosphorylation sites that are implicated in its function

Figure 2 continued on next page

Figure 2 continued

(***Hao et al., 2008***; ***Yamamoto et al., 2010***; ***Zarin et al., 2017***), the Pex5 IDR and Rad26 IDR have none, and the Stp4 IDR has 4. IDRs are presented in order of increasing Euclidian distance between their evolutionary signatures, though we do not recommend using this measure to quantitate similarity between evolutionary signatures independently (see Discussion). 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 fall outside the 1.5*interquartile range, and are represented by unfilled circles. Distribution of GFP activity is based on quantification of GFP intensity in single cells pooled from four colonies (which we define as biological replicates) for each strain; sample sizes for each distribution are as follows: WT n = 588 cells, Pex5 IDR n = 196 cells, Stp4 IDR n = 228 cells, Rad26 IDR n = 271 cells. (**C**) Brightfield micrographs showing each strain from part B following exposure to pheromone. Shmooing cells are those which have elongated cell shape, representing mating projections.

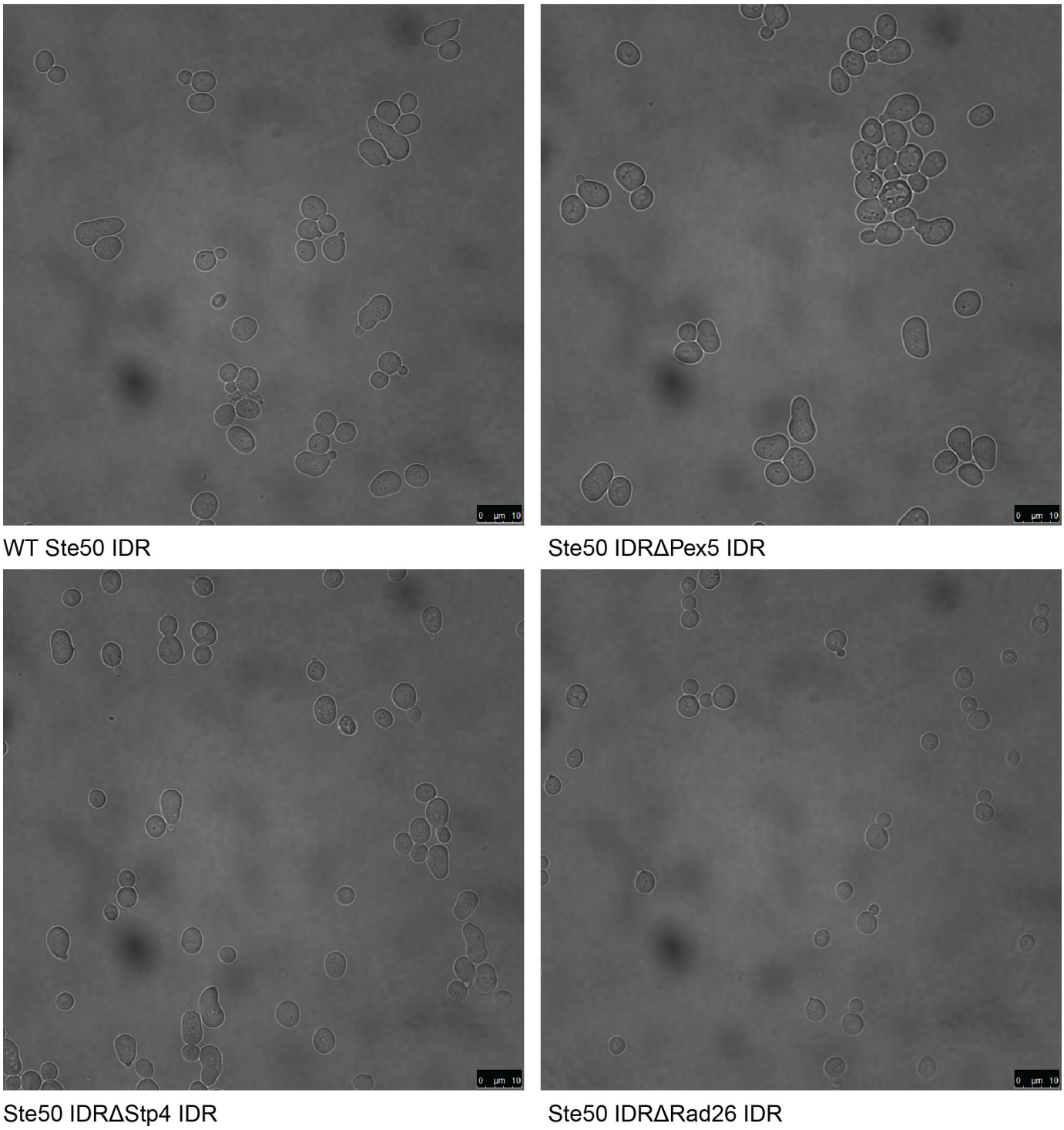

Figure 2—figure supplement 1. Full field-of-view micrographs of pheromone-exposed *S. cerevisiae* strains from ***Figure 2C***.

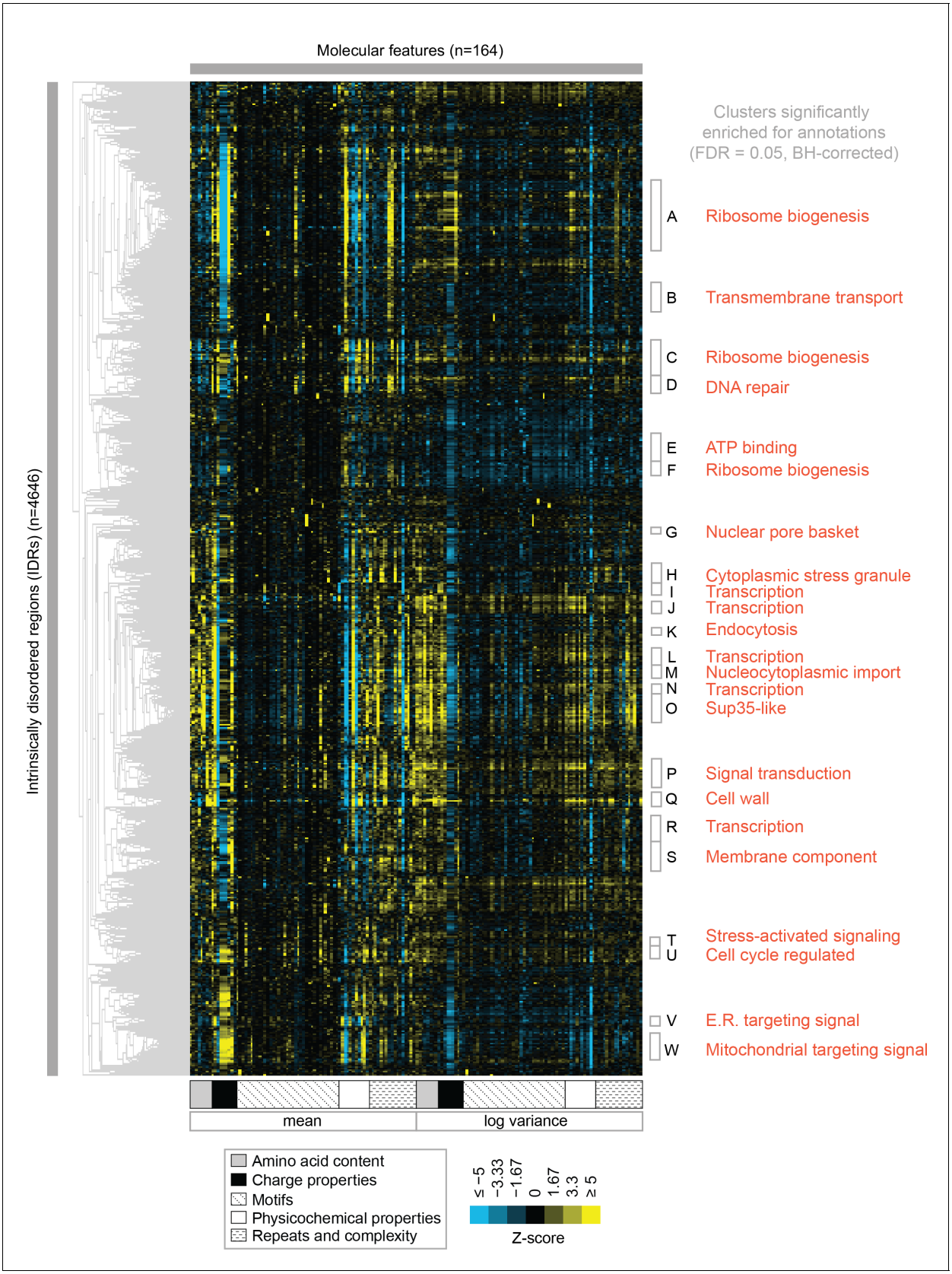

Figure 3. Clustering evolutionary signatures shows that IDRs in the proteome share evolutionary signatures, and that these clusters of IDRs are associated with specific biological functions. A-W show clusters significantly enriched for annotations (see ***Table 1***; full table of enrichments in supplementary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

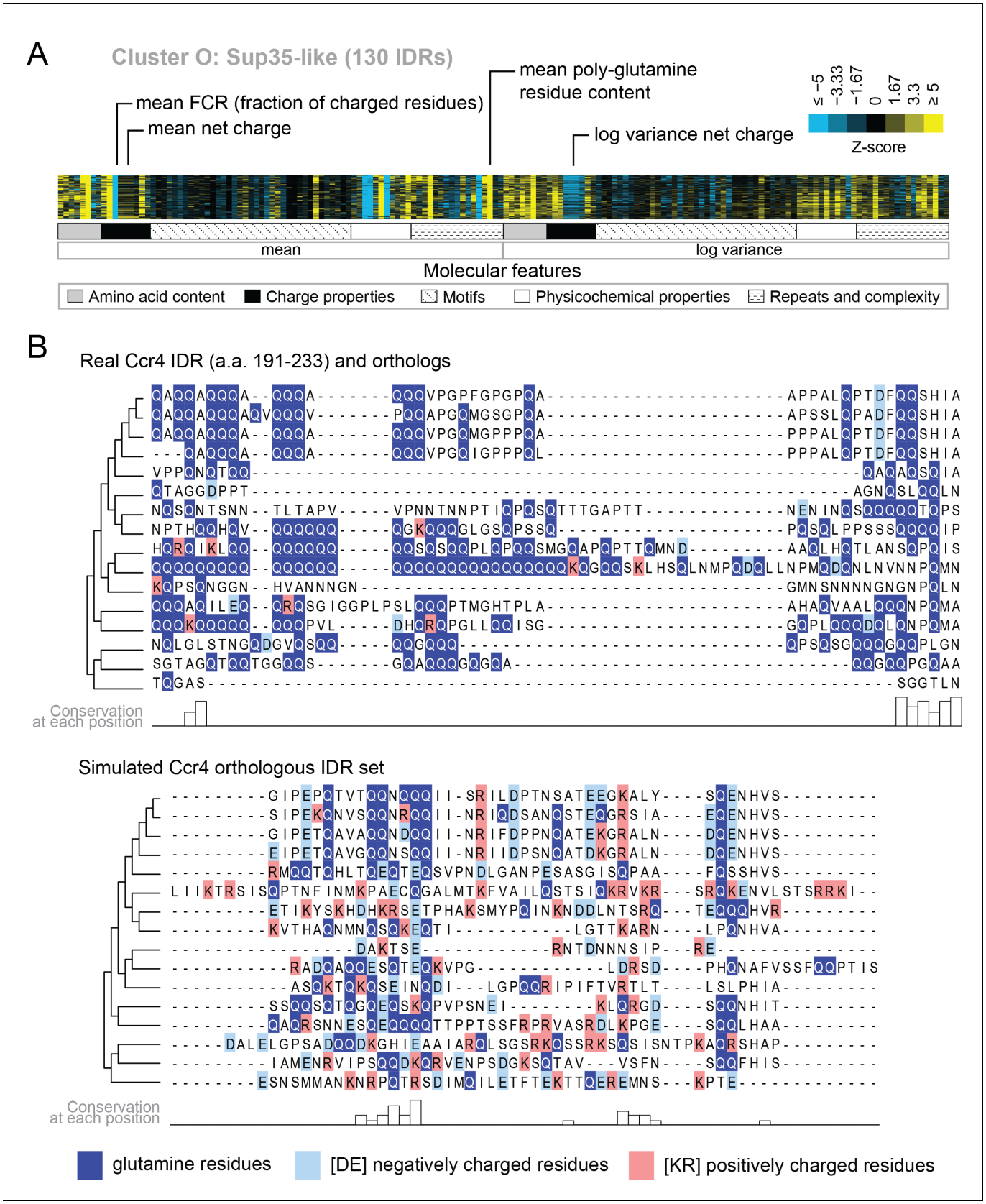

Figure 4. Evolutionary signatures in cluster O contain some molecular features that are typically associated with IDRs as well as some that are not. (**A**) Pattern of evolutionary signatures in cluster O. (**B**) Example disordered region from cluster O, Ccr4, with a subset of highlighted molecular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

Figure 4 continued on next page

Figure 4 continued

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*.

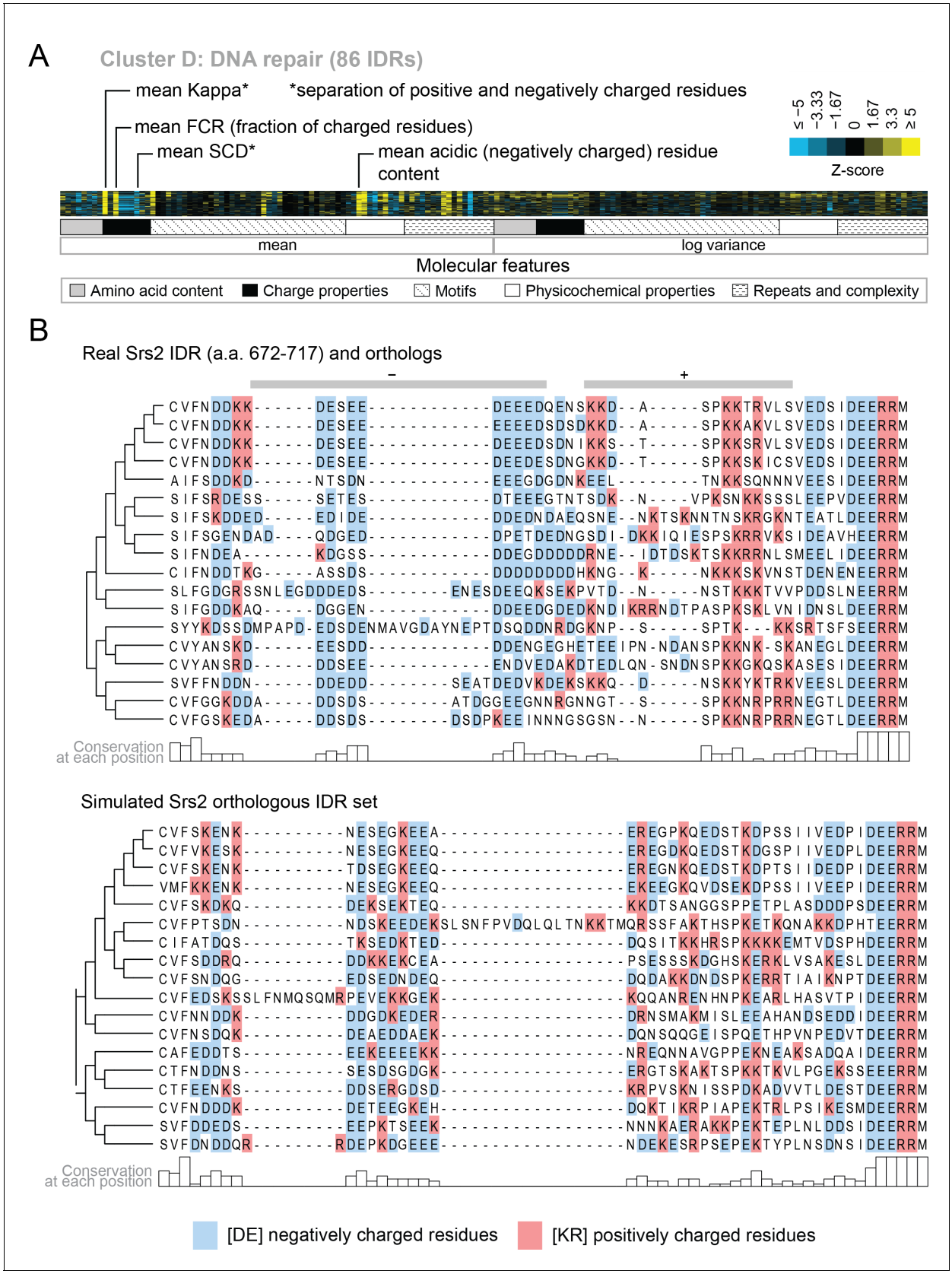

Figure 5. Cluster D contains disordered regions associated with DNA repair. (**A**) Pattern of evolutionary signatures in cluster D. (**B**) Example disordered region from cluster D, Srs2, with a subset of highlighted molecular features compared between its real set of orthologs and an example set of

Figure 5 continued on next page

Figure 5 continued

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*.

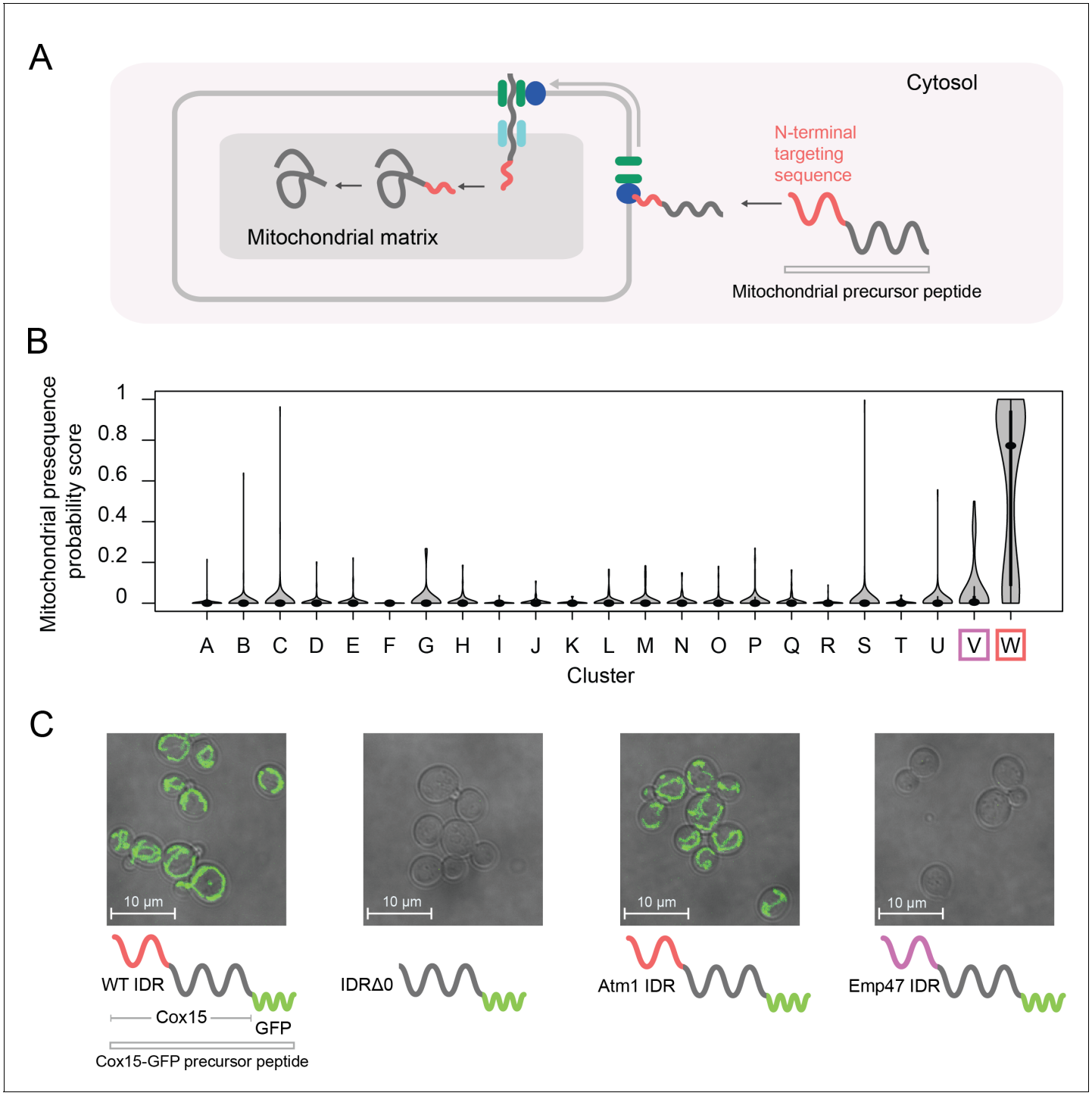

Figure 6. Cluster W is associated with mitochondrial N-terminal targeting signals. (**A**) Schematic (not to scale) showing the path of a mitochondrial precursor peptide (with N-terminal targeting sequence in red) from the cytosol, where it is translated, to the mitochondrial matrix, where the peptide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

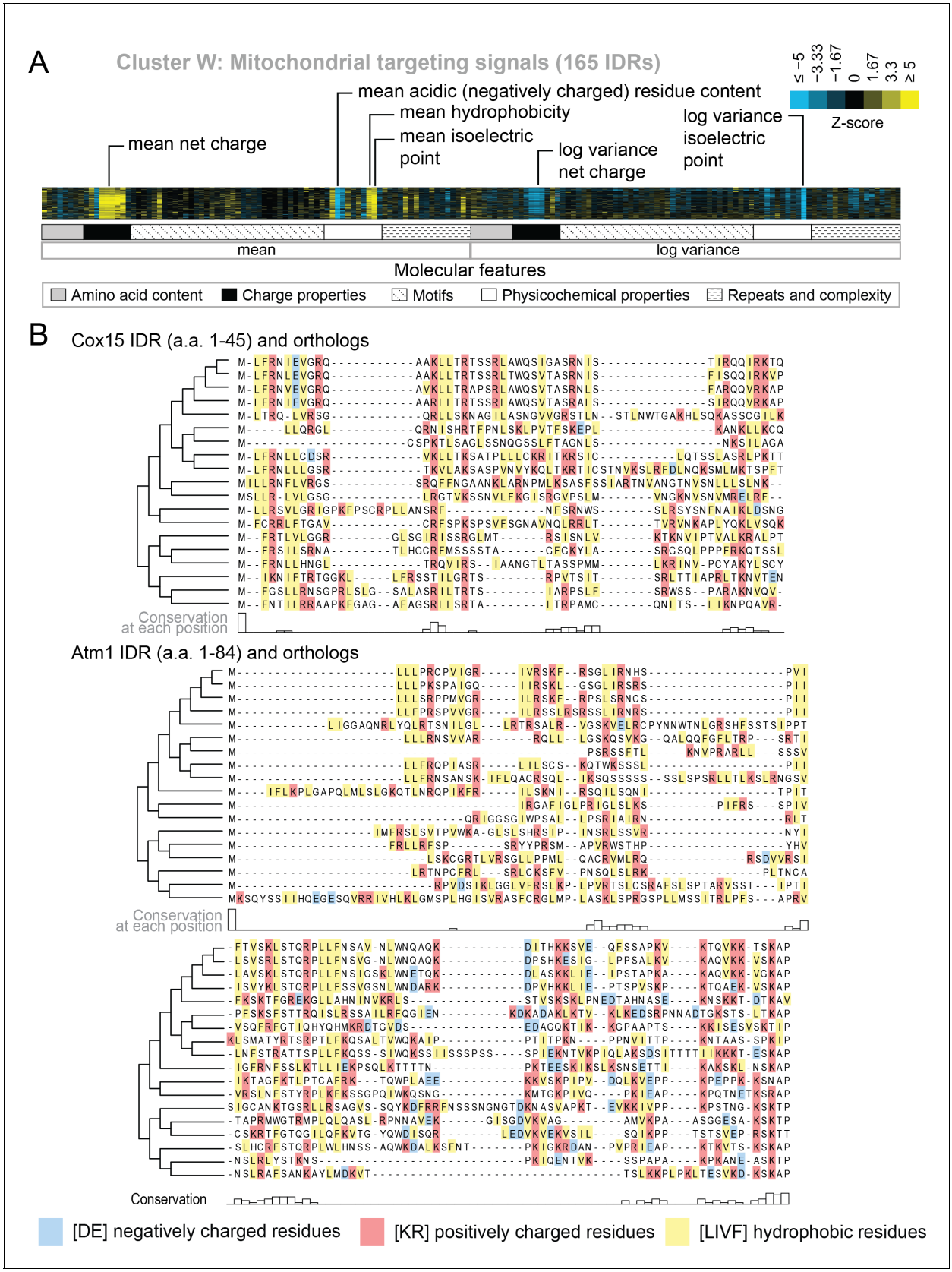

Figure 6—figure supplement 1. Evolutionary signatures in cluster W contain molecular features that have been previously reported for mitochondrial N-terminal targeting signals. (**A**) Pattern of evolutionary signatures in cluster W. (**B**) Multiple sequence alignments of example disordered regions from

Figure 6—figure supplement 1 continued on next page

Figure 6—figure supplement 1 continued

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*.

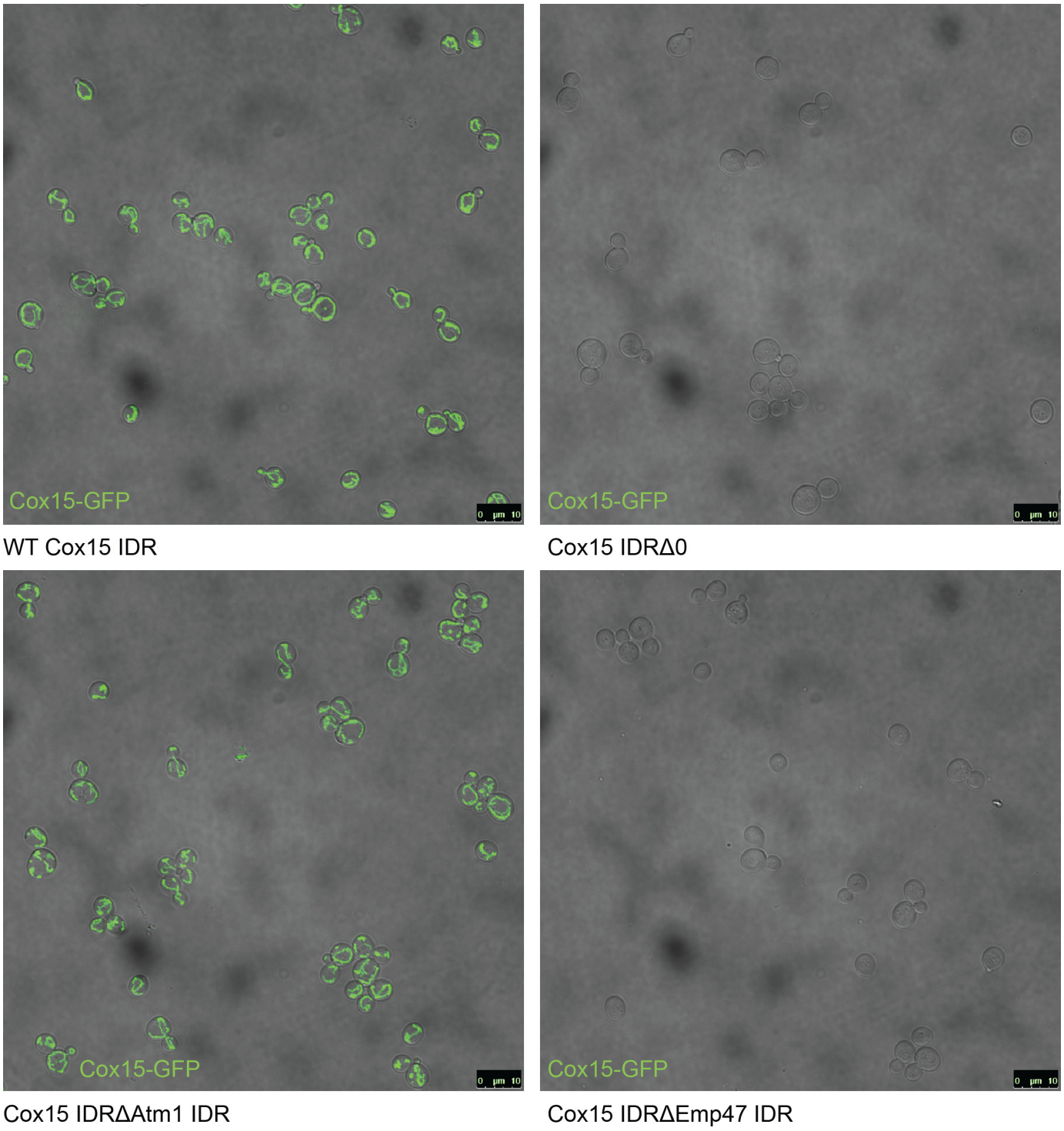

Figure 6—figure supplement 2. Full field-of-view micrographs of *S. cerevisiae* strains from ***Figure 6C***.

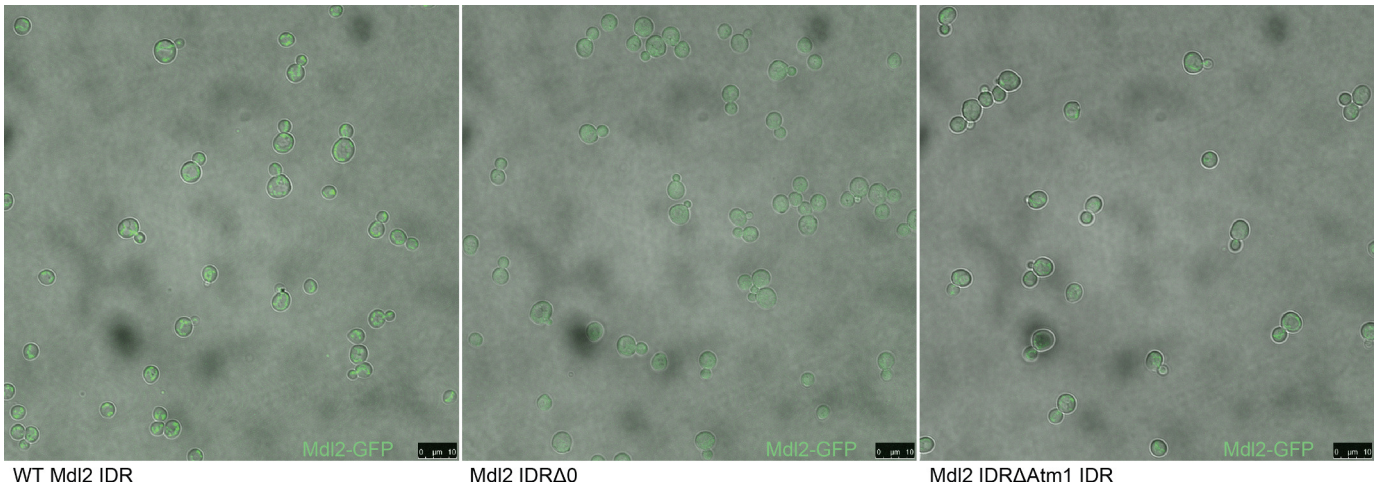

Figure 6—figure supplement 3. Micrographs of *S. cerevisiae* strains with three different genotypes. From left to right: Mdl2-GFP has a mitochondrial localization in the wildtype (WT) strain, knocking out the Mdl2 IDR abolishes wildtype localization, and replacing the Mdl2 IDR with that of Atm1 rescues mitochondrial localization.

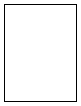

Figure 6—figure supplement 4. Reverse transformation of GFP-tagged Cox15 IDRΔ0 and Cox15ΔEmp47 strains to wildtype Cox15 IDR rescues mitochondrial localization of Cox15-GFP. Scale bars represent 10 micrometers. (**A**) GFP-tagged Cox15 IDRΔEmp47 reverted to wildtype Cox15-GFP. (**B**) GFP-tagged Cox15 IDRΔ0 reverted to wildtype Cox15-GFP.