Yeast proteins reversibly aggregate like amphiphilic molecules

Pouria Dasmeh[1,3,4], Andreas Wagner[1,2,4]

[1]Institute for Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland. [2]The Santa Fe Institute, Santa Fe, New Mexico, [3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA USA 02139. [4]Swiss Institute of Bioinformatics (SIB).

Abstract:

More than a hundred proteins in yeast reversibly aggregate and phase-separate in response to various stressors, such as nutrient depletion and heat shock. We know little about the sequence and structural features behind this ability, which has not been characterized on a proteome-wide level. To identify the distinctive features of aggregation-prone regions, we apply machine learning algorithms to genome-scale limited proteolysis-mass spectrometry data from 96 yeast proteins that phase-separate upon heat shock. We find that the propensity to phase separate cannot be solely driven by disordered regions, because the aggregation-prone regions of our study proteins are not significantly disordered. Instead, the phase separation of these proteins requires contributions from both disordered and structured regions. APRs are significantly enriched in aliphatic residues and depleted in positively charged amino acids. Aggregator proteins with longer APRs show a greater propensity to aggregate, a relationship that can be explained by equilibrium statistical thermodynamics. Altogether, our observations suggest that proteome-wide reversible protein aggregation is mediated by sequence-encoded properties. Aggregating proteins resemble supra-molecular amphiphiles, where APRs are the hydrophobic parts, and non-APRs are the hydrophilic parts.

Body:

Proteins can aggregate reversibly or irreversibly. Irreversible aggregation is often pathological, indicates damaged cellular regulation[1-3], and is involved in multiple diseases such as Alzheimer's and Parkinson's diseases[4,5]. Reversible aggregation, however, can be beneficial and help cells survive stressors. For example, in yeast more than a hundred proteins in multiple subcellular compartments form reversible aggregates in response to nutrient starvation, heat shock, or chemical stress[6]. These aggregated proteins are not misfolded or tagged for degradation[7]. Instead, they help increase cells re-initiate growth during recovery from stress by protecting metabolic enzymes from degradation[6-9].

Reversible protein aggregation is a special case of a widespread phenomenon called protein phase-separation. In this process, a well-mixed solution of proteins de-mixes into two phases of high and low densities[10]. Proteins can phase separate with other proteins, but also with RNA or DNA molecules, into biomolecular condensates that regulate transcription[11], chromatin states[12,13], and RNA metabolism[14]. Phase-separating proteins often harbor intrinsically-disordered regions (domains) that do not fold into well-defined tertiary structures[15]. These regions can form a network of specific and non-specific protein interactions to promote phase separation (see ref. [16] for a review).

The role of disordered regions in stress-induced phase separation has been controversial. On the one hand, mammalian and yeast proteins that phase separate under stress are highly disordered and enriched in features that favor liquid-liquid phase separation. These features include many protein-protein interactions and low complexity regions with multiple polar and charged amino acids, as well as specific amino acids such

59 as tyrosine[17]. These observations suggest that the assembly of low-complexity and
60 disordered regions drive stress-induced phase separation of these proteins. On the other
61 hand, disordered domains are not essential for phase-separation in some well-studied
62 proteins, where they only modulate the temperature at which phase separation begins.
63 Examples include the poly-adenylate binding protein (Pab1), and the ATP-binding RNA
64 helicase (Ded1) from yeast[18,19]. To understand the importance of disordered regions for
65 stress-induced aggregation and phase separation comprehensively requires a proteome-
66 wide view. In this work, we provide such a view. We also characterize sequence features
67 that can help predict whether proteins will aggregate reversibly.

68 We took advantage of recent proteome-wide data from limited proteolysis-mass
69 spectrometry (LiP-MS)[20,21]. This method permits the detection of both pronounced and
70 subtle changes that protein and peptide structures experience in response to stressors
71 such as heat and osmotic shock. Recently, Cappelletti *et. al.* characterized structural
72 changes within 96 proteins that reversibly aggregate upon heat shock in yeast[22]. These
73 proteins are also known as aggregators. They have diverse biological functions, such as
74 telomeric DNA-binding, RNA-helicase activity, or ribosome assembly. Because they are
75 functionally diverse, they constitute an excellent dataset to examine sequence features
76 that facilitate reversible protein aggregation. We studied regions within these proteins
77 whose proteolysis resistance changes significantly upon aggregation [22]. We refer to these
78 regions as aggregation-prone regions (APRs, Figure 1A).

79 Our first analysis focuses on the question whether APRs preferentially comprise
80 intrinsically disordered regions. To find out, we first calculated the disorder score of all

81 amino acids within the aggregator proteins using the *IUpred* disorder predictor[23] (Dataset
82 S1). This algorithm assigns a score between 0 and 1 to each amino acid using its
83 neighboring sequence in the protein, and calculated from a statistical approach that
84 distinguishes globular domains from intrinsically disordered regions. Residues with the
85 score > 0.5 are predicted to be disordered. Remarkably, these amino acids were not
86 significantly enriched in APRs (p~1; Fisher's exact test). To the contrary, amino acids with
87 disorder scores less than 0.5 were enriched ($p \sim 10^{-7}$; Fisher's exact test). Thus, on
88 average, APRs do not preferentially comprise disordered regions. However, the incidence
89 of disorder in APRs substantially varied among individual proteins (Dataset S2). As shown
90 in Figure 1B, ~15% of aggregators were significantly more enriched with disordered
91 amino acids than expected by chance ($p < 0.05$; Fisher's exact test, false discovery rate
92 (FDR)-corrected for multiple testing[24]). Conversely, disordered amino acids were
93 significantly depleted in ~11% of aggregators ($p < 0.05$; Fisher's exact test, corrected for
94 multiple testing). The proteins whose APRs were significantly disordered are
95 preferentially involved in RNA and nucleic acid binding (Figure 1C; Dataset S3, GO terms
96 of GO:0003723, and GO:0003676). For example, the protein with the highest disorder
97 enrichment is the nuclear poly-adenylated RNA-binding protein 3 (NAB3; yeast gene
98 identifier YPL190C), which is required for packaging pre-mRNAs into ribonucleoprotein
99 structures for efficient RNA processing[25].

100 Because disorder scores may be subject to prediction errors, we validated this
101 approach with two well-studied aggregator proteins[18,26]. Specifically, we used the
102 polyadenylate-binding protein Pab1, which binds the poly(A) tail of mRNA, regulating

103 mRNA stability and translation; and the ATP-dependent RNA helicase Ded1, an ATP-
104 binding RNA helicase that helps identify start codons and initiate translation[27]. Both
105 proteins form reversible gel-like condensates upon heat shock[18,19]. Figure 1D shows
106 APRs within the two proteins. Importantly, only one of eight APRs falls within the
107 intrinsically disordered domain (green) of Pab1, and only one of two does for Ded1.

108 We then mapped APRs on the 3D structure of Pab1 (Figure 1E; PDB ID=6R5K[28]),
109 and a homology-modeled structure of Ded1 (Figure 1F; Dataset S4). A common feature
110 of both proteins are peptides that form long loops which occur either within an APR or
111 constitute an entire APR. The regions ^{151}D-K^{156}, ^{167}G-E^{176}, ^{186}L-R^{204}, ^{242}F-K^{252}, ^{383}N-L^{392}
112 in Pab1, and ^{122}T-R^{127} in Ded1 are examples of such loop-forming peptides. In the case
113 of Ded1, the APR ^{122}T-R^{127} lies near the disordered N-terminal domain of this protein,
114 which modulates the temperature onset of phase-separation[19]. This proximity suggests a
115 possible synergy between the N-terminal domain and the APR in the reversible
116 aggregation of Ded1. Altogether, our observations suggest that reversible aggregation of
117 yeast proteins require a synergy between both disordered and structured regions.

118 What are the characteristic sequence features of APRs? To answer this question,
119 we first performed a principal component analysis on the frequency of 20 amino acids in
120 all 270 APRs of yeast aggregators (Dataset S5). As shown in Figure 2A-B, the first and
121 the second principal components only accounted for ~12% and ~10% of the variation.
122 The PCA biplot for these components shows that no specific combinations of few amino
123 acids preferentially occur in APR sequences (Figure 2B, Figure S1), a further indication
124 that APRs do not show the kind of reduced sequence complexity characteristic of

125 disordered regions. We also looked for linear sequence motifs in APRs using DALEL[29],
126 an algorithm for the exhaustive identification of degenerate sequence motifs, but found
127 no such enriched motifs.

128 We then compared the frequencies of amino acids, dipeptides, and tripeptides in
129 APRs with the rest of the protein sequence in aggregators (Dataset S6). Here, we found
130 a significant depletion of the positively charged residues Arg and Lys, and their dipeptides
131 (Figure S2; $p \sim 10^{-16}$; Wilcoxon signed-rank test). We also observed that aliphatic residues
132 Leu, Ile, Ala, Val, were significantly more frequent in APRs compared to the rest of the
133 protein sequence ($p \sim 10^{-16}$; Wilcoxon signed-rank test). The fraction of positively charged
134 to aliphatic residues was the strongest predictor of APRs compared to the fraction of all
135 other amino acids (Figure 2C; $p \sim 10^{-21}$; Wilcoxon signed-rank test).

136 To find out whether these patterns are linked to differences in physicochemical
137 properties of amino acids, we used a random forest approach, a widely-used machine
138 learning technique for the classification of two or more data sets[30]. Specifically, we used
139 all 96 aggregators with experimentally characterized APRs[22], and subdivided each
140 protein into two sequence data sets, one comprising the APRs and one comprising the
141 non-APRs. For each of these datasets, we calculated a feature matrix that consisted of
142 500 amino acid properties (Dataset S7, and S8). We took these properties from the
143 AAindex database[31], which curates various physicochemical and biochemical properties
144 of amino acids. The classifier achieved an accuracy of ~93% in 100 independent runs,
145 with the data split into a training set (80% of the data) and a testing set (20%) (Figure 2D;
146 Dataset S9). Two amino acid properties were most important in this classification. The

147 first is a low positive charge of amino acids in APRs compared to non-APRs. The second
148 is a high amphiphilic propensity of these amino acids. The amphiphilic propensity
149 measures the preference of amino acids to occur at protein-solvent interfaces, such as
150 the end regions of transmembrane helices[32]. Amino acids like arginine and lysine have
151 the highest preference for these environments, and the largest value of the amphiphilic
152 index. In contrast, hydrophobic amino acids such as leucine, isoleucine, valine, and
153 alanine prefer internal protein environments and have the lowest amphiphilic propensity[32].

154 Because of the statistically significant difference between the amphiphilic nature
155 of APRs and non-APRs, we argue that aggregator proteins resemble supramolecular
156 amphiphiles. These molecular structures have distinct hydrophobic and hydrophilic
157 parts[33]. In response to external stimuli such as temperature, pH, and ionic strength[33],
158 they can reversibly aggregate by non-covalent bonding forces, e.g., through electrostatic
159 and hydrophobic interactions.

160 We next investigated the relevance of amphiphilic aggregation to stress-induced
161 phase separation of our yeast aggregator proteins. An important feature of amphiphilic
162 molecules is that they are more likely to occur in an aggregated phase if their hydrophobic
163 chain is long[34]. We thus predicted that aggregator proteins with longer APRs will occur
164 preferentially in the aggregated phase. To test this prediction, we asked whether
165 aggregator proteins with long APRs preferentially occur in the pellet fraction of proteins
166 extracted from heat-stressed yeast cells[35]. Specifically, we used for this purpose the $\log_2$
167 ratio of the protein's abundance in the pellet fraction of yeast cells to that of supernatant
168 as a measure of protein's enrichment in the aggregated phase[35]. We found this

169 enrichment data for 31 aggregated proteins in our dataset. Indeed, aggregator proteins
170 with longer APRs preferentially accumulate in the aggregated pellet fraction (Figure 2F,
171 R = 0.53, p = 0.0021; Spearman's rank correlation). This association increases to R =
172 0.64 for aggregator proteins whose APRs have a low positive charge-to-aliphatic bias
173 (Figure 2G, p= 0.012; Spearman's rank correlation). To see whether this observation
174 extends to other stress-induced conditions, we also used the mass spectrometry dataset
175 of nutrient-starved yeast cells reported by Narayanasamy et al[36]. We identified 56
176 aggregator proteins in this dataset and checked the association between APR lengths in
177 these proteins and the Z-score of proteins' enrichment in the pellet fraction. In nutrient-
178 starved cells too, the aggregator proteins with longer APRs preferentially accumulate in
179 the aggregated pellet fraction (Figure 2F, R = 0.47, p = 0.00014; Spearman's rank
180 correlation), and the association increased for the low positive charge-to-aliphatic bias (R
181 = 0.53, p = 0.0022; Spearman's rank correlation).

182 We then investigated the relationship between APR length in aggregator proteins
183 and their differential enrichment in the aggregated phase using equilibrium statistical
184 thermodynamics. We considered a two-state model where proteins exist either as a
185 soluble monomer or an insoluble aggregate. The fraction of proteins occurring in the
186 aggregate at the critical concentration (cc) for aggregation ($\mathcal{F}_{cc}$) is expressed as:

187 $$\mathcal{F}_{cc}(theory) = \frac{(P_{agg})_{cc}}{(P_{agg})_{cc}+(P_{mon})_{cc}} = \frac{1}{1+e^{-\beta\Delta G_{agg}}} \quad \text{(Eq. 1)}$$

188 Here, P_{agg} and P_{mon} are the relative concentrations of proteins in the aggregated and the
189 monomer phase, $\beta = 1/k_B T$ where k_B is the Boltzmann's constant, and ΔG_{agg} is the free

energy change upon aggregation. Because the aggregation free energy increases in proportion to the length of the hydrophobic chain[37], we rearranged Equation 1 to:

$$\mathcal{F}_{cc}(theory) = \frac{1}{1+e^{-\beta.B.L}} \qquad \text{(Eq. 2)}$$

Where *B* is the proportionality constant between the aggregation free energy and the APR length. We then calculated the observed fraction of the aggregator proteins in the aggregated phase from their reported $\log_2$ enrichment ratio in the pellet fraction of heat-stressed yeast ($log_2 \frac{[Pellet]}{[Supernatant]}$) as:

$$\mathcal{F}_{cc}(experiment) = (1 + 2^{-log_2 \frac{[Pellet]}{[Supernatant]}})^{-1} \qquad \text{(Eq. 3)}$$

Equation 2 predicted the observed relationship between APR length and a protein's fraction in the aggregated phase well (Figure 2H, R=0.53, p=0.0021; Spearman's rank correlation, see supplementary information for the fitted constants). Importantly, the predictive power of Equation 2 was higher for proteins whose APRs had a low positive charge-to-aliphatic bias (R=0.64, p=0.012; Spearman's rank correlation).

In summary, our work demonstrates that both structured and disordered regions contribute to the reversible aggregation of yeast proteins upon heat shock. In the case of the two well-known aggregator proteins Pab1 and Ded1, structured aggregation-prone regions form long loops or short secondary structure elements. Our observations suggest that these elements can readily unfold in response to stressors, form hydrophobic long chains, and phase-separate in a manner akin to amphiphilic molecules. Although further structural studies will be necessary to generalize these observations, we propose that the

210 partial unfolding of a protein's tertiary structure is a critical stage in stress-induced phase
211 separation. Indeed, the secondary structure of both Ded1 and Pab1 remains largely
212 unchanged and only their tertiary structures change when these proteins form heat-
213 induces condensates[18,38]. We also observed that aggregator proteins with longer APRs
214 are more prone to aggregation. This relationship also suggests that aggregator proteins
215 may leave the aggregated phase by masking their APRs from self-assembly or from
216 interaction with other proteins. Molecular chaperones can facilitate this step: They have
217 been co-purified with stress-induced aggregated proteins and can solubilize them[39].
218 Altogether, our study demonstrates that stress-induced phase separation is a sequence-
219 encoded phenotype. It can be both explained and predicted by the self-assembly and
220 aggregation of amphiphilic molecules.

221

222 **Methods**:

223 We identified aggregation-prone regions in the 96 aggregator proteins from the dataset
224 of Cappelletti et. al[22], and downloaded their sequences from the Uniprot database[40]
225 (Dataset S13). We chose 500 amino acid properties from the AAindex database to
226 distinguish the physicochemical properties of APRs from non-APR segments in the
227 aggregator protein[31]. For mapping APRs onto the 3D structure of proteins, we used the
228 structure of Pab1 as determined by electron microscopy (protein database (PDB) id:
229 6R5K[28]). We used homology-modeling with the Phyre2 portal to generate the structure of
230 Ded1[41]. In brief, Phyre 2 constructs a 3D structure from diverse homologs of the query

231 sequence using hidden Markov models. For the generation of Ded1 structure, Phyre 2
232 used the template structures of several Ded1 homologs, including the ATP-dependent
233 RNA helicase DDX3X from humans (PDB ID= 2I4I), the ATP-dependent RNA helicase
234 VASA from *Drosophila melanogaster* (PDB ID=2DB3), and the pre-mRNA-splicing ATP-
235 dependent RNA helicase PRP28 from *Saccharomyces cerevisiae* (PDB ID=5ZWN).

236 We used a binary classification to group regions within an aggregator protein into
237 aggregation-prone regions (class A) and the remainder of the protein (class B). More
238 specifically, we used a random forest classification, splitting aggregator proteins into a
239 random subset comprising 80% of aggregator proteins for training and 20% for testing.
240 To build features for the classification, we calculated the average value of 500
241 physicochemical properties for each sequence in the APR and non-APR dataset. This
242 yielded two feature matrices for the APR and non-APR sequences. To apply random
243 forest classification, we used the randomForest package in R[42], and evaluated the best
244 number of trees (*nTree*) and the number of variables randomly sampled at each split
245 (*mtry*), in the random forest algorithm. To do so, we systematically varied the *nTree* and
246 *mtry*, and calculated the accuracy of classification with 10-fold cross-validation and 3
247 repeats. We defined accuracy as the percentage of correctly identified classes of proteins
248 (APRs and non-APRs) out of all instances. The combination of *nTree*=5000 trees and
249 *mtry*=10 variables achieved the highest accuracy of ~ 90%. We then used these
250 parameters to perform 100 random forest clusterings, in which we randomly assigned
251 proteins to the training and the testing datasets. To quantify the accuracy of classification
252 we counted the number of true positive and false positive predictions and calculated the

253 area under the curve (AUC). These values are shown as a receiver operating
254 characteristic curve (ROC) in Figure 2D. The most important physicochemical properties
255 were the ones whose Gini index in the classification decreased the most compared to all
256 other properties. This index is a widely-used measure of dispersion that reflects inequality
257 in the values of a frequency distribution. Within the random forest framework, this index
258 is calculated as the average probability that each of 500 physicochemical properties
259 wrongly classifies APRs and non-APRs in the random forest algorithm. We used the
260 g:profiler server for Gene Ontology enrichment analysis[43]. We performed all models and
261 statistical analyses using R. Scripts are available at:
262 https://github.com/dasmeh/yeast_aggregators

263 **References**

264 1. Aguzzi, A. & Calella, A.M. Prions: protein aggregation and infectious diseases.
265 *Physiological reviews* **89**, 1105-1152 (2009).
266 2. Aguzzi, A. & O'connor, T. Protein aggregation diseases: pathogenicity and therapeutic
267 perspectives. *Nature reviews Drug discovery* **9**, 237-248 (2010).
268 3. Ross, C.A. & Poirier, M.A. Protein aggregation and neurodegenerative disease. *Nature*
269 *medicine* **10**, S10-S17 (2004).
270 4. Scheibel, T. & Buchner, J. Protein aggregation as a cause for disease. *Molecular*
271 *chaperones in health and disease*, 199-219 (2006).
272 5. Shulman, J.M., De Jager, P.L. & Feany, M.B. Parkinson's disease: genetics and
273 pathogenesis. *Annual Review of Pathology: Mechanisms of Disease* **6**, 193-222 (2011).
274 6. O'Connell, J.D. *et al.* A proteomic survey of widespread protein aggregation in yeast.
275 *Molecular BioSystems* **10**, 851-861 (2014).
276 7. Wallace, E.W. *et al.* Reversible, specific, active aggregates of endogenous proteins
277 assemble upon heat stress. *Cell* **162**, 1286-1298 (2015).
278 8. Saad, S. *et al.* Reversible protein aggregation is a protective mechanism to ensure cell
279 cycle restart after stress. *Nature cell biology* **19**, 1202-1213 (2017).
280 9. Cereghetti, G., Saad, S., Dechant, R. & Peter, M. Reversible, functional amyloids: towards
281 an understanding of their regulation in yeast and humans. *Cell Cycle* **17**, 1545-1558
282 (2018).

10. Hyman, A.A., Weber, C.A. & Jülicher, F. Liquid-liquid phase separation in biology. *Annual review of cell and developmental biology* **30**, 39-58 (2014).
11. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K. & Sharp, P.A. A phase separation model for transcriptional control. *Cell* **169**, 13-23 (2017).
12. Erdel, F. & Rippe, K. Formation of chromatin subcompartments by phase separation. *Biophysical journal* **114**, 2262-2270 (2018).
13. Valsecchi, C.I.K. *et al.* RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* **589**, 137-142 (2021).
14. Youn, J.-Y. *et al.* Properties of Stress Granule and P-Body Proteomes. *Molecular Cell* **76**, 286-294 (2019).
15. Uversky, V.N. Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Current opinion in structural biology* **44**, 18-30 (2017).
16. Posey, A.E., Holehouse, A.S. & Pappu, R.V. Phase separation of intrinsically disordered proteins. in *Methods in Enzymology*, Vol. 611 1-30 (Elsevier, 2018).
17. Kuechler, E.R., Budzyńska, P.M., Bernardini, J.P., Gsponer, J. & Mayor, T. Distinct features of stress granule proteins predict localization in Membraneless organelles. *Journal of molecular biology* **432**, 2349-2368 (2020).
18. Riback, J.A. *et al.* Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* **168**, 1028-1040. e19 (2017).
19. Iserman, C. *et al.* Condensation of Ded1p Promotes a Translational Switch from Housekeeping to Stress Protein Production. *Cell* (2020).
20. Gao, X. *et al.* High-throughput limited proteolysis/mass spectrometry for protein domain elucidation. *Journal of structural and functional genomics* **6**, 129-134 (2005).
21. Schopper, S. *et al.* Measuring protein structural changes on a proteome-wide scale using limited proteolysis-coupled mass spectrometry. *Nature protocols* **12**, 2391 (2017).
22. Cappelletti, V. *et al.* Dynamic 3D proteomes reveal protein functional alterations at high resolution in situ. *Cell* **184**, 545-559. e22 (2021).
23. Dosztányi, Z. Prediction of protein disorder based on IUPred. *Protein Science* **27**, 331-340 (2018).
24. Bonferroni, C.E., Bonferroni, C. & Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita'. (1936).
25. Sugimoto, K., Matsumoto, K., Kornberg, R.D., Reed, S.I. & Wittenberg, C. Dosage suppressors of the dominant G1 cyclin mutant CLN3-2: identification of a yeast gene encoding a putative RNA/ssDNA binding protein. *Molecular and General Genetics MGG* **248**, 712-718 (1995).
26. Altmeyer, M. *et al.* Liquid demixing of intrinsically disordered proteins is seeded by poly (ADP-ribose). *Nature communications* **6**, 8088 (2015).
27. Jamieson, D.J., Rahe, B., Pringle, J. & Beggs, J.D. A suppressor of a yeast splicing mutation (prp8-l) encodes a putative ATP-dependent RNA helicase. *Nature* **349**, 715-717 (1991).
28. Schäfer, I.B. *et al.* Molecular basis for poly (A) RNP architecture and recognition by the Pan2-Pan3 deadenylase. *Cell* **177**, 1619-1631. e21 (2019).

29. Kelil, A., Dubreuil, B., Levy, E.D. & Michnick, S.W. Exhaustive search of linear information encoding protein-peptide recognition. *PLoS computational biology* **13**, e1005499 (2017).
30. Ho, T.K. Random decision forests. in *Proceedings of 3rd international conference on document analysis and recognition* Vol. 1 278-282 (IEEE, 1995).
31. Kawashima, S. & Kanehisa, M. AAindex: amino acid index database. *Nucleic acids research* **28**, 374-374 (2000).
32. Mitaku, S., Hirokawa, T. & Tsuji, T. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane–water interfaces. *Bioinformatics* **18**, 608-616 (2002).
33. Zhang, X. & Wang, C. Supramolecular amphiphiles. *Chemical Society Reviews* **40**, 94-101 (2011).
34. Shaul, B. & Gelbart, W. Theory of chain packing in amphiphilic aggregates. *Annual Review of Physical Chemistry* **36**, 179-211 (1985).
35. Zhu, M. *et al.* Proteomic analysis reveals the direct recruitment of intrinsically disordered regions to stress granules in S. cerevisiae. *Journal of Cell Science* **133**(2020).
36. Narayanaswamy, R. *et al.* Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings of the National Academy of Sciences* **106**, 10147-10152 (2009).
37. Budin, I., Prywes, N., Zhang, N. & Szostak, J.W. Chain-length heterogeneity allows for the assembly of fatty acid vesicles in dilute solutions. *Biophysical journal* **107**, 1582-1590 (2014).
38. Iserman, C. *et al.* Condensation of Ded1p promotes a translational switch from housekeeping to stress protein production. *Cell* **181**, 818-831. e19 (2020).
39. Cherkasov, V. *et al.* Coordination of translational control and protein homeostasis during severe heat stress. *Current biology* **23**, 2452-2462 (2013).
40. Consortium, U. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204-D212 (2015).
41. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. & Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**, 845-858 (2015).
42. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22 (2002).
43. Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47**, W191-W198 (2019).

362 **Figure captions:**
363
364 **Figure 1. Intrinsically-disordered and structured regions both contribute to**
365 **reversible aggregation in yeast aggregators.** A) Structure of data set and study design.
366 We identified aggregation-prone regions within 96 aggregator proteins in yeast using a
367 dataset of limited proteolysis-mass spectrometry by Cappelletti et. al[22]. We performed
368 several sequence-based statistical and predictive analyses and investigated the
369 enrichment of protein disorder, specific sequence composition, and distinctive features of
370 aggregation-prone regions in these proteins. B) Volcano plot showing the preferential
371 enrichment of intrinsically-disordered regions in aggregation-prone regions of the
372 aggregator proteins. The y-axis represents the logarithm of the p-value calculated from
373 Fisher's exact test and the x-axis shows the difference between the expected and
374 observed fraction of intrinsically disordered amino acids. Proteins with significant
375 enrichment and depletion of disordered amino acids in their aggregation-prone regions
376 are shown in red, and blue, respectively. C) The enrichment of Gene Ontology terms for
377 molecular function in proteins whose aggregation-prone regions are enriched for (red),
378 and depleted in (blue) disordered residues. D) The sequences of Pab1 and Ded1 with
379 their intrinsically disordered domains shown in green. Aggregation-prone regions are
380 labeled P1 to P8 for Pab1, and D1 and D2, for Ded1. E) The crystal structure of Pab1
381 (PDB ID=6R5K[28]). F) The homology-modeled structure of Ded1 using the Phyre2 protein
382 modeling web portal [41]. In both panels D and E, aggregation-prone regions are shown in
383 red and labeled according to their definition in panel D. The regions P7 in Pab1 and D2

384 in Ded1 fall in the low complexity domains which are missing from the PDB X-ray
385 structure.
386
387 **Figure 2. Sequence composition and physicochemical properties of aggregation-**
388 **prone regions in the yeast aggregators.** A) The percentage of explained variance for
389 10 dimensions from a principal component analysis on the frequencies of 20 amino acids
390 in 270 aggregation-prone regions of 96 yeast proteins. B) Biplot for the first and the
391 second principal components. Grey circles, labeled by numbers from 1 to 270, represent
392 aggregation-prone regions. Red arrows represent the weight of different amino acid
393 frequencies. C) The amino acid frequencies in APRs (shown in red) versus the rest of the
394 protein sequence (shown in green) in 96 yeast aggregators. D) The receiver-operating
395 characteristic (ROC) curves for the clustering of APRs from non-APRs using random
396 forest classification. ROC curves in grey are from 100 random forest clustering and their
397 average is shown in black. E) Ranked importance of physicochemical variables quantified
398 as the average decrease in the Gini index for different physicochemical properties. F) The
399 log2 ratio of the protein's abundance in the pellet fraction to that of supernatant in heat-
400 stressed yeast cells (30 out of 96 aggregator proteins[35]) versus their APR length. G) The
401 aggregator proteins' enrichment in the pellet fraction upon nutrient starvation (59 out of
402 96 proteins[35]) versus their APR length in log-scale. The red and the grey circles in both
403 panels F and G represent all aggregator proteins, and those where the positive charge-
404 to-aliphatic-bias in APRs was lower than its median value ~ 0.25. H) The fraction of
405 aggregator proteins in the aggregated phase calculated from Equation 3 versus the length

406 of APRs. I) The fraction of aggregator proteins with a low positive charge-to-aliphatic-bias
407 in their APRs versus the length of APRs calculated from Equation 3. For both panels H
408 and I, the fitted curve to the observed fraction of proteins in the aggregated phase using
409 Equation 2 is shown in red. R values are the Spearman's rank correlation between the
410 expected (Equation 2) and observed (Equation 3) fractions of proteins in the aggregated
411 phase (See Dataset S12). J) Schematic of our two main findings. First, both disordered
412 and structured regions contribute to stress-induced phase separation in yeast. Second,
413 proteins whose aggregation-prone regions are longer or have a lower positive charge-to-
414 aliphatic bias, accumulate preferentially in the aggregated phase.
415
416
417
418
419

Figure 1

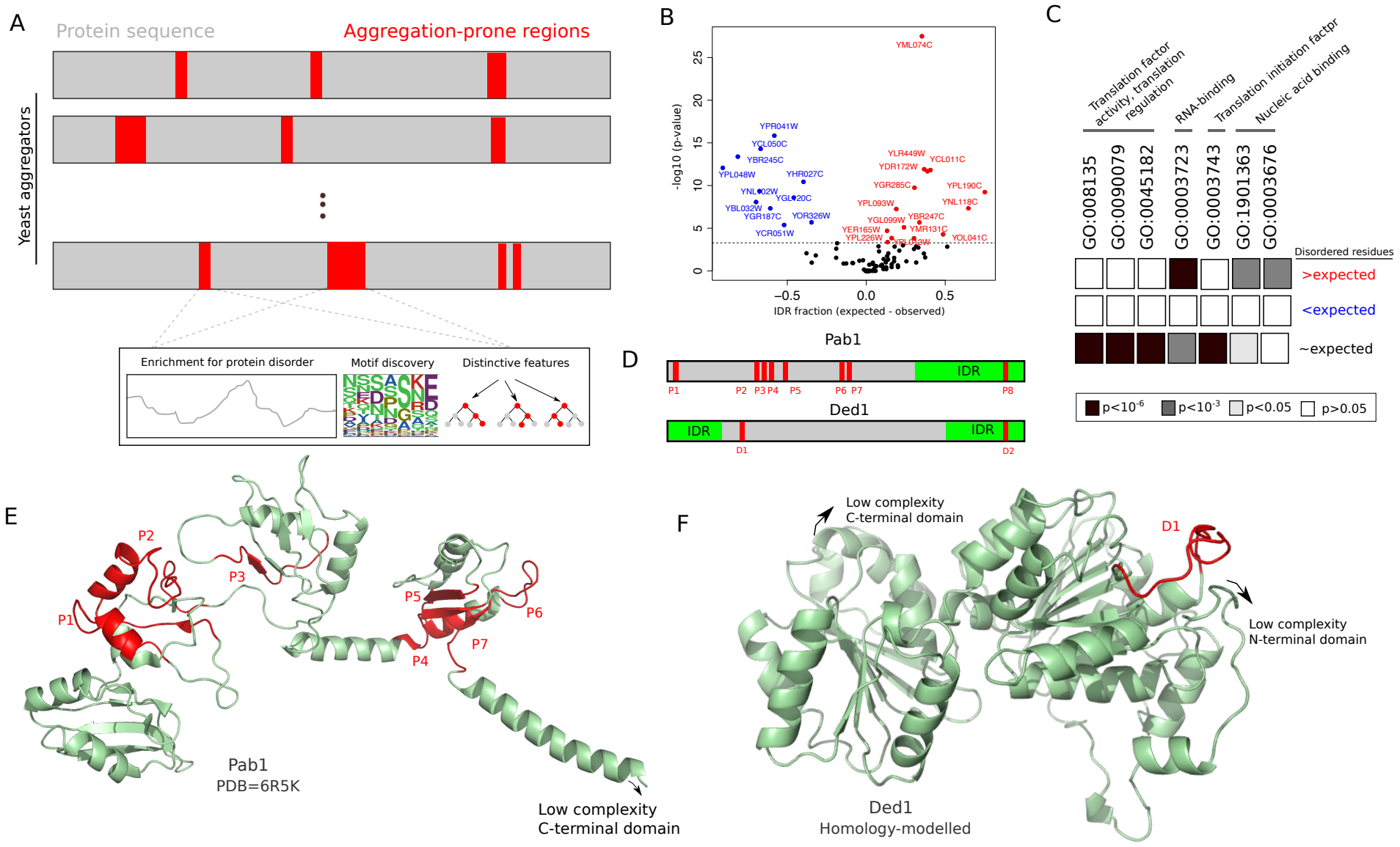

A
Protein sequence
Aggregation-prone regions
Yeast aggregators
Enrichment for protein disorder
Motif discovery
Distinctive features
B
-log10 (p-value)
IDR fraction (expected - observed)
YML074C
YPR041W
YCL050C
YBR245C
YPL048W
YHR027C
YNL102W
YGL120C
YBL032W
YGR187C
YOR326W
YCR051W
YLR449W
YCL011C
YDR172W
YGR285C
YPL190C
YPL093W
YNL118C
YGL099W
YBR247C
YER165W
YMR131C
YPL226W
YOL041C
C
Translation factor activity, translation regulation
RNA-binding
Translation initiation factor
Nucleic acid binding
GO:008135
GO:0090079
GO:0045182
GO:0003723
GO:0003743
GO:1901363
GO:0003676
Disordered residues
>expected
<expected
~expected
p<10-6
p<10-3
p<0.05
p>0.05
D
Pab1
IDR
P1
P2
P3 P4
P5
P6 P7
P8
Ded1
IDR
IDR
D1
D2
E
P1
P2
P3
P4
P5
P6
P7
Pab1
PDB=6R5K
Low complexity
C-terminal domain
F
Low complexity
C-terminal domain
D1
Low complexity
N-terminal domain
Ded1
Homology-modelled

Figure 2

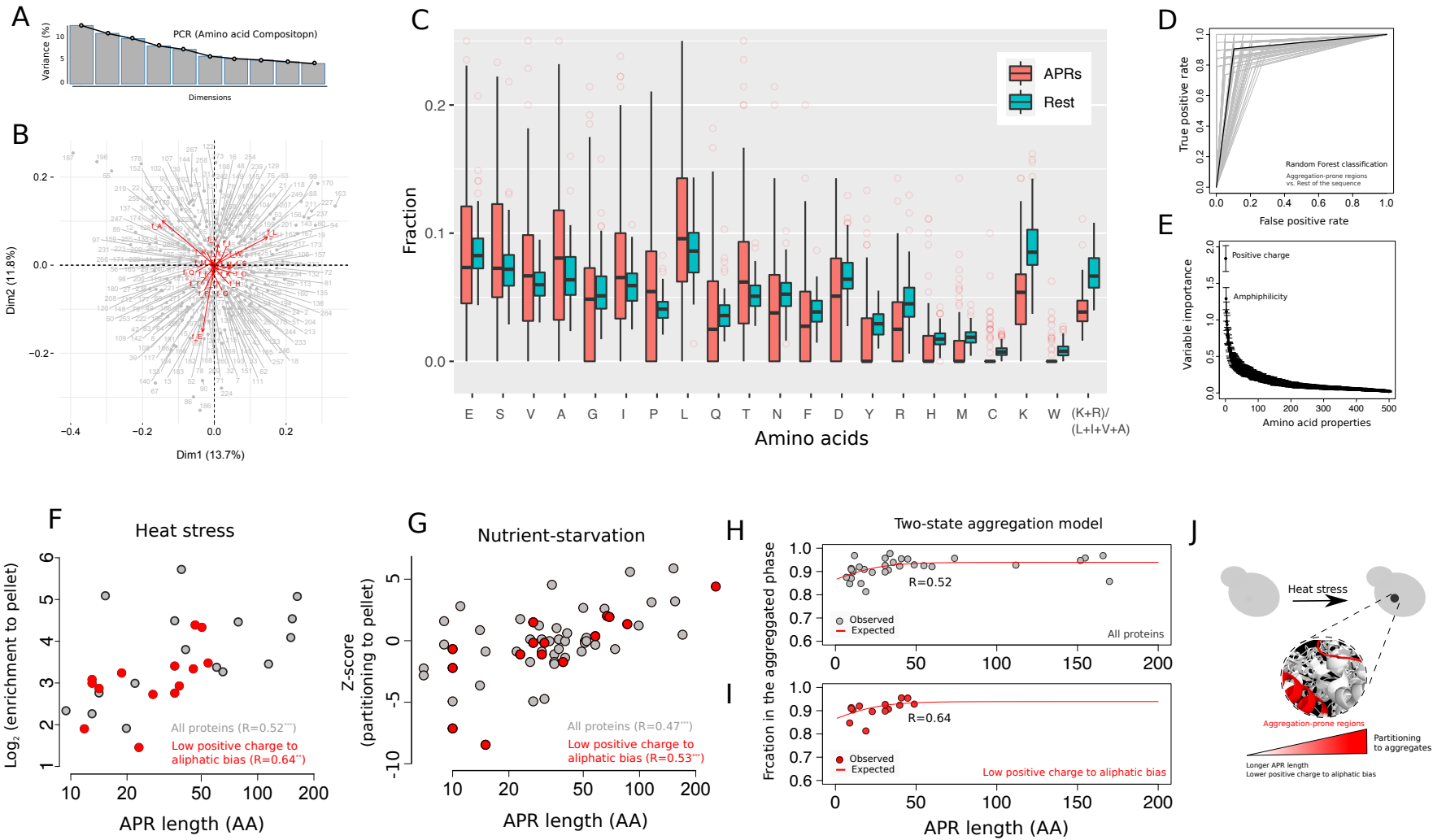

A
PCR (Amino acid Compositopn)
Variance (%)
Dimensions
B
Dim2 (11.8%)
Dim1 (13.7%)
C
APRs
Rest
Fraction
Amino acids
E S V A G I P L Q T N F D Y R H M C K W (K+R)/ (L+I+V+A)
D
True positive rate
False positive rate
Random Forest classification
Aggregation-prone regions vs. Rest of the sequence
E
Variable importance
Positive charge
Amphiphilicity
Amino acid properties
F
Heat stress
Log2 (enrichment to pellet)
All proteins (R=0.52***)
Low positive charge to aliphatic bias (R=0.64**)
APR length (AA)
G
Nutrient-starvation
Z-score (partitioning to pellet)
All proteins (R=0.47***)
Low positive charge to aliphatic bias (R=0.53***)
APR length (AA)
H
Two-state aggregation model
Frcation in the aggreggated phase
R=0.52
Observed
Expected
All proteins
I
R=0.64
Low positive charge to aliphatic bias
APR length (AA)
J
Heat stress
Aggregation-prone regions
Partitioning to aggregates
Longer APR length
Lower positive charge to aliphatic bias

Figure 1

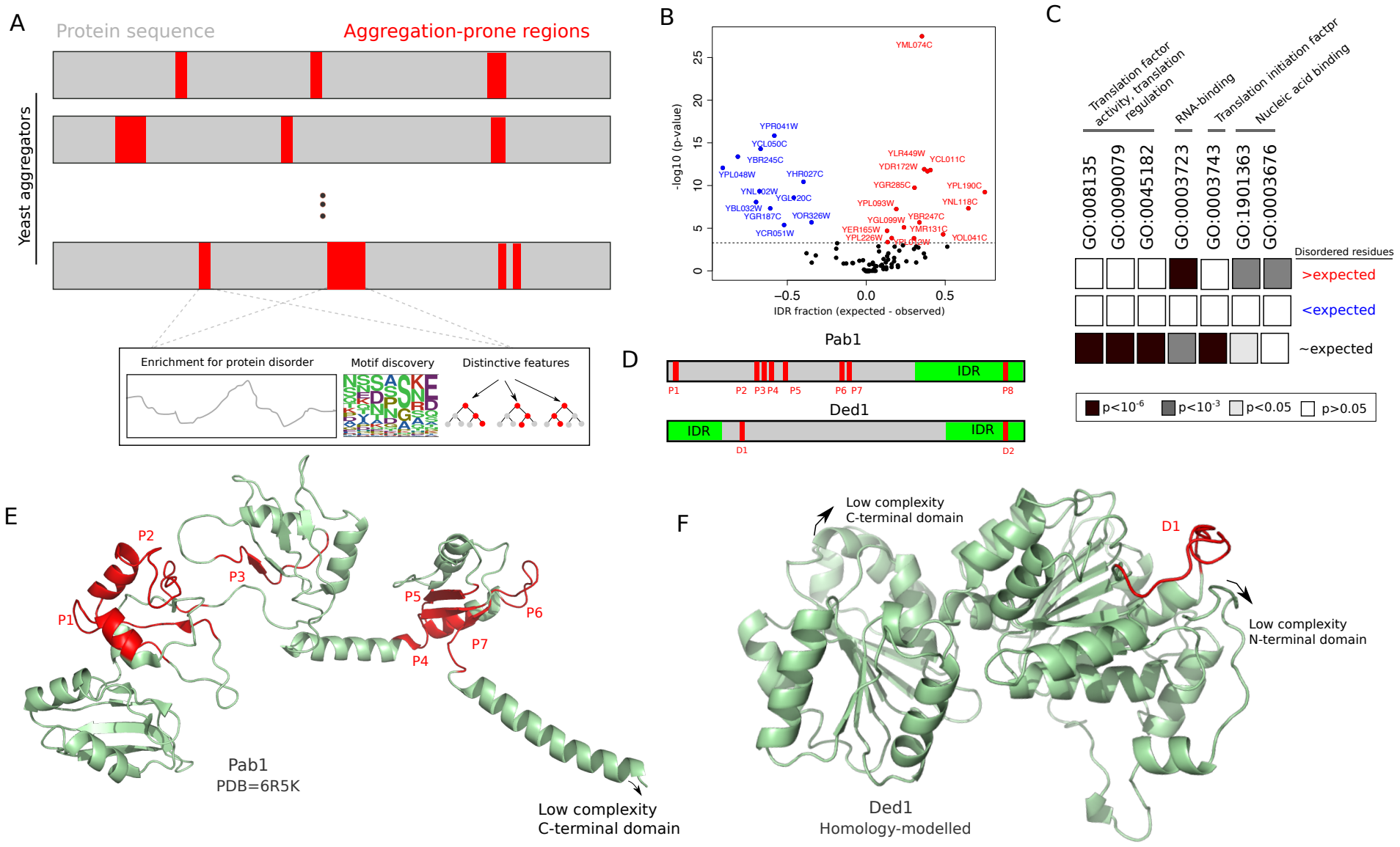
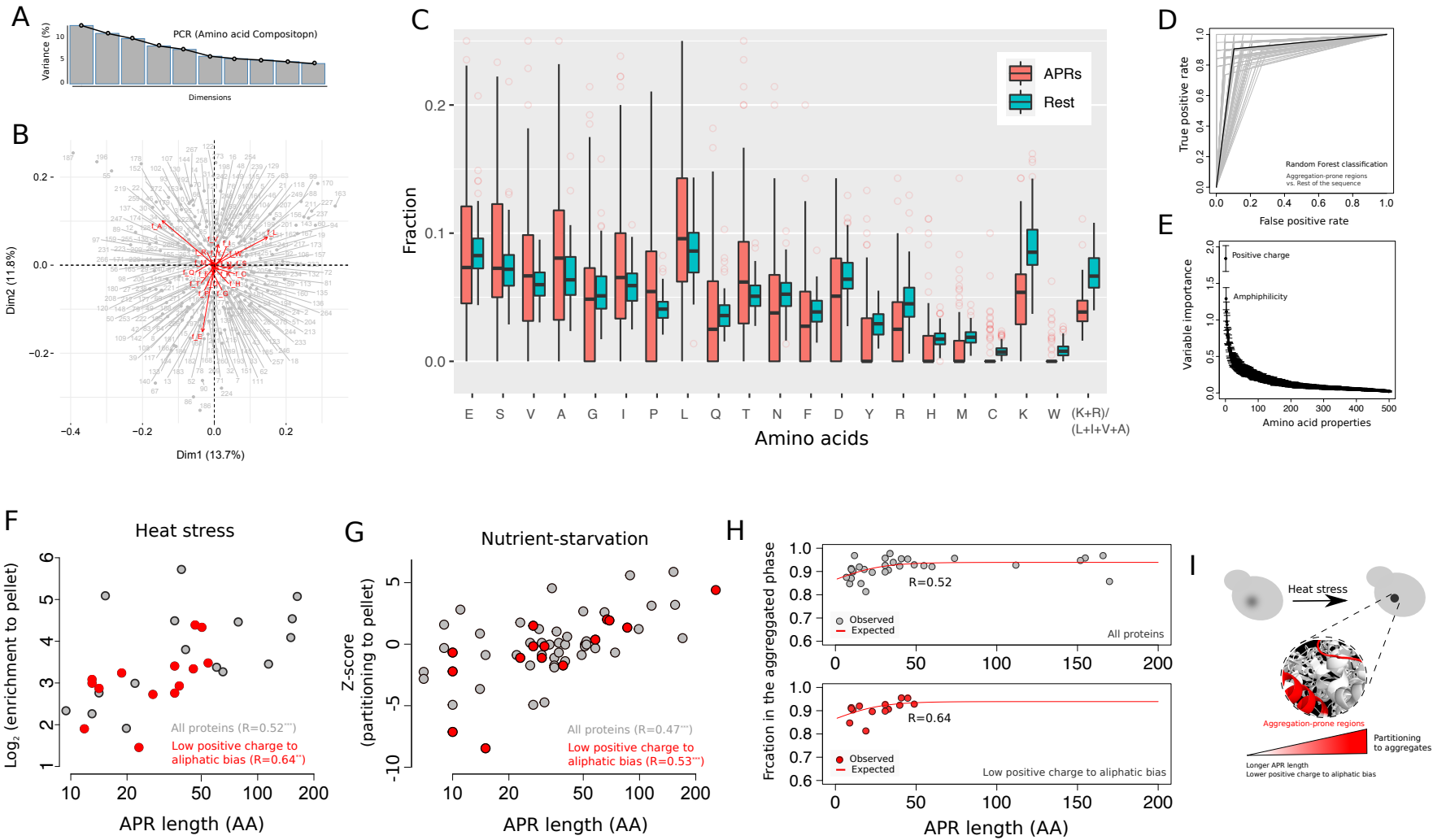

A
Protein sequence
Aggregation-prone regions
Yeast aggregators
Enrichment for protein disorder
Motif discovery
Distinctive features
B
-log10 (p-value)
IDR fraction (expected - observed)
YML074C
YPR041W
YCL050C
YBR245C
YPL048W
YHR027C
YNL102W
YGL120C
YBL032W
YGR187C
YOR326W
YCR051W
YLR449W
YCL011C
YDR172W
YGR285C
YPL190C
YPL093W
YNL118C
YGL099W
YBR247C
YER165W
YMR131C
YPL226W
YOL041C
C
Translation factor activity, translation regulation
RNA-binding
Translation initiation factor
Nucleic acid binding
GO:008135
GO:0090079
GO:0045182
GO:0003723
GO:0003743
GO:1901363
GO:0003676
Disordered residues
>expected
<expected
~expected
p<10-6
p<10-3
p<0.05
p>0.05
D
Pab1
IDR
P1
P2
P3 P4
P5
P6 P7
P8
Ded1
IDR
IDR
D1
D2
E
P1
P2
P3
P4
P5
P6
P7
Pab1
PDB=6R5K
Low complexity
C-terminal domain
F
Low complexity
C-terminal domain
D1
Low complexity
N-terminal domain
Ded1
Homology-modelled

Figure 2


A
PCR (Amino acid Compositopn)
Variance (%)
Dimensions
B
Dim2 (11.8%)
Dim1 (13.7%)
C
APRs
Rest
Fraction
Amino acids
E S V A G I P L Q T N F D Y R H M C K W (K+R)/ (L+I+V+A)
D
True positive rate
False positive rate
Random Forest classification
Aggregation-prone regions vs. Rest of the sequence
E
Variable importance
Positive charge
Amphiphilicity
Amino acid properties
F
Heat stress
Log2 (enrichment to pellet)
All proteins (R=0.52***)
Low positive charge to aliphatic bias (R=0.64**)
APR length (AA)
G
Nutrient-starvation
Z-score (partitioning to pellet)
All proteins (R=0.47***)
Low positive charge to aliphatic bias (R=0.53***)
APR length (AA)
H
Frcation in the aggreggated phase
R=0.52
Observed
Expected
All proteins
R=0.64
Low positive charge to aliphatic bias
APR length (AA)
I
Heat stress
Aggregation-prone regions
Partitioning to aggregates
Longer APR length
Lower positive charge to aliphatic bias