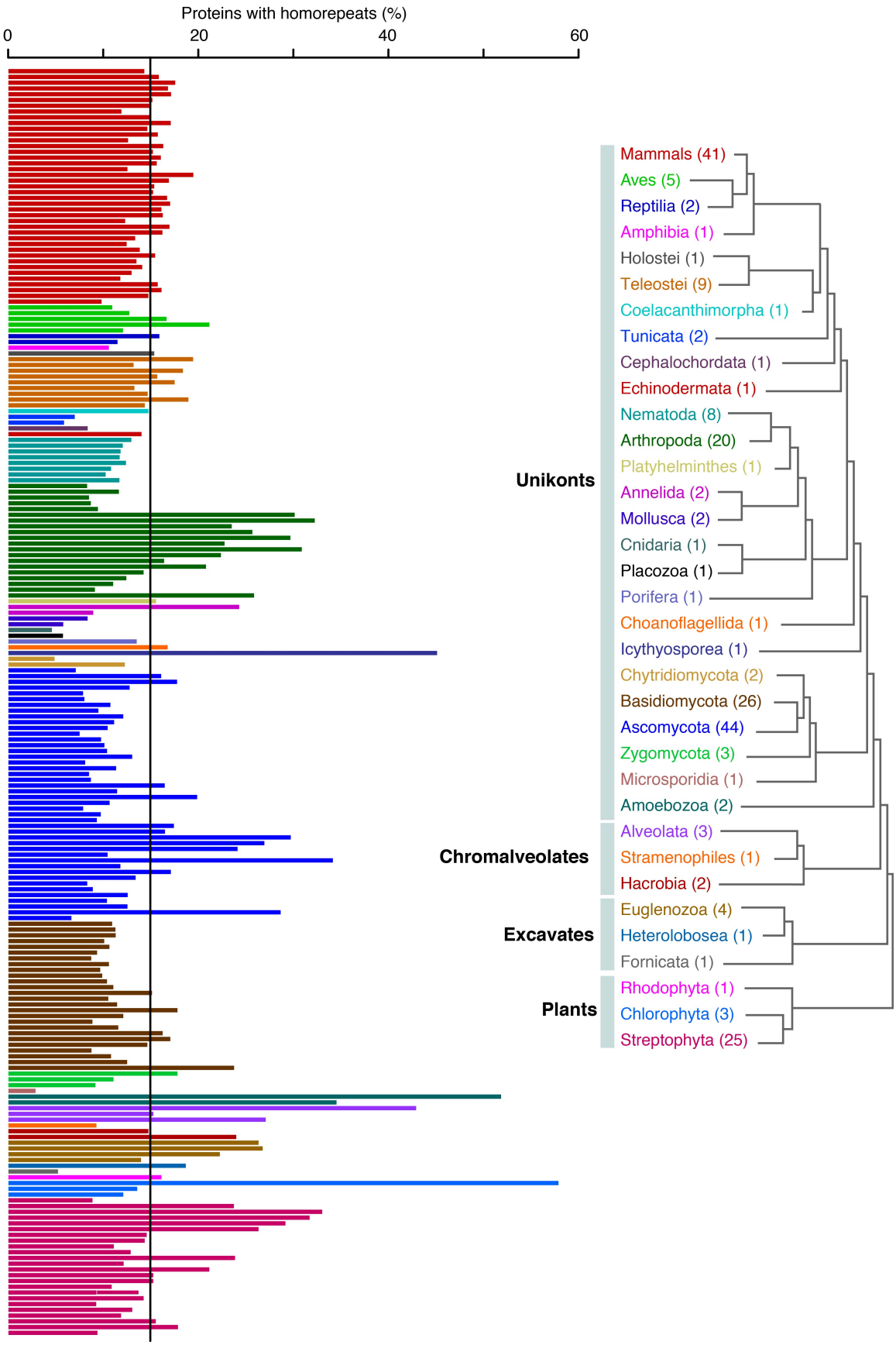
Supplementary Notes and Supplementary Tables for

Constraints and consequences of the emergence of amino acid repeats in eukaryotic proteins

Sreenivas Chavali[*], Pavithra L. Chavali[†], Guilhem Chalancon[†], Natalia Sanchez de Groot, Rita Gemayel, Natasha S. Latysheva, Elizabeth Ing-Simmons, Kevin J. Verstrepen, Santhanam Balaji, M. Madan Babu[1,*]

* Correspondence to: schavali@mrc-lmb.cam.ac.uk (S.C.), madanm@mrc-lmb.cam.ac.uk (M.M.B.)

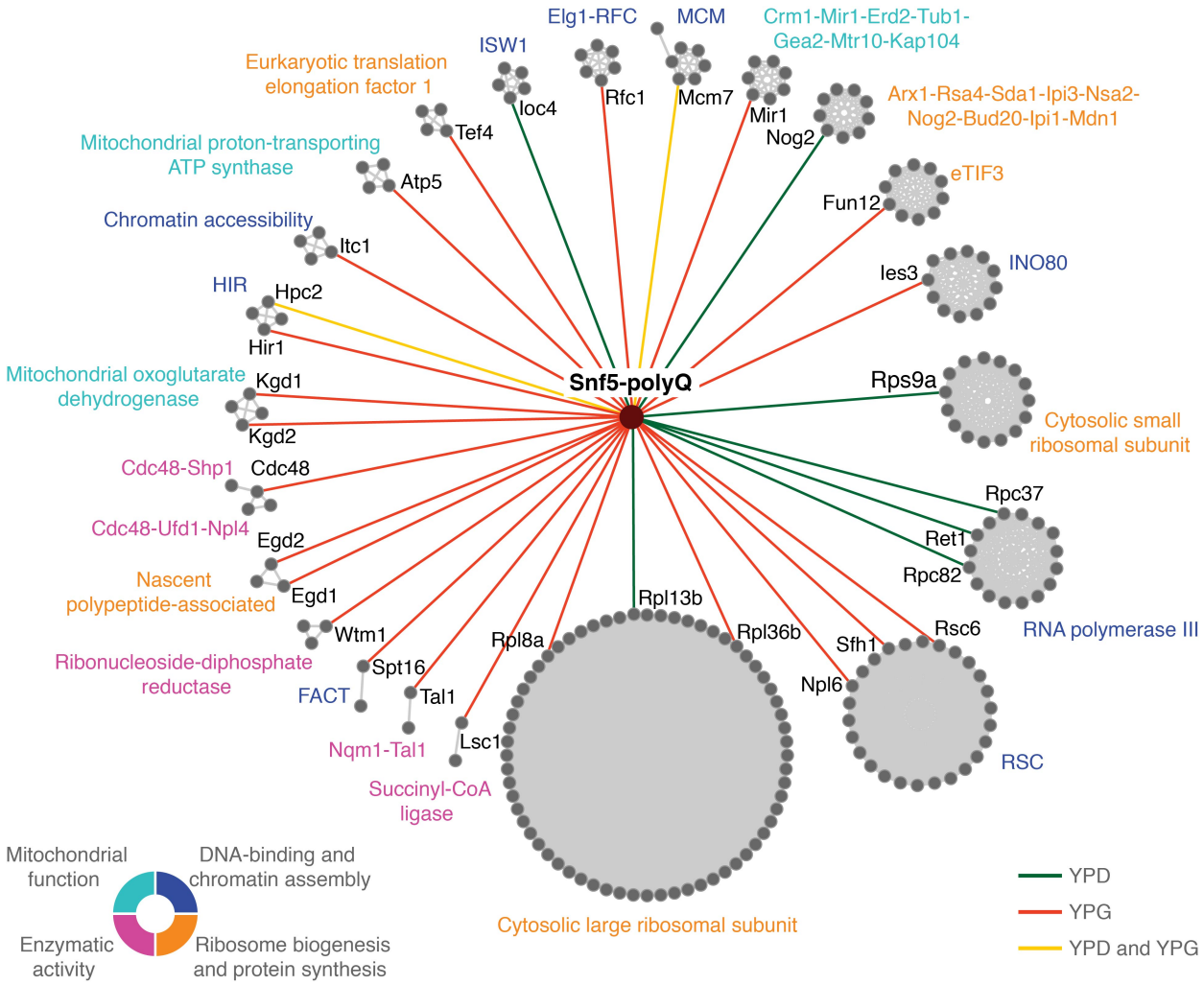
† These authors contributed equally to this work


Proteins with homorepeats (%)
0
20
40
60
Unikonts
Chromalveolates
Excavates
Plants
Mammals (41)
Aves (5)
Reptilia (2)
Amphibia (1)
Holostei (1)
Teleostei (9)
Coelacanthimorpha (1)
Tunicata (2)
Cephalochordata (1)
Echinodermata (1)
Nematoda (8)
Arthropoda (20)
Platyhelminthes (1)
Annelida (2)
Mollusca (2)
Cnidaria (1)
Placozoa (1)
Porifera (1)
Choanoflagellida (1)
Icythyosporea (1)
Chytridiomycota (2)
Basidiomycota (26)
Ascomycota (44)
Zygomycota (3)
Microsporidia (1)
Amoebozoa (2)
Alveolata (3)
Stramenophiles (1)
Hacrobia (2)
Euglenozoa (4)
Heterolobosea (1)
Fornicata (1)
Rhodophyta (1)
Chlorophyta (3)
Streptophyta (25)

Supplementary notes 1

Proteins with homorepeats in eukaryotic proteomes

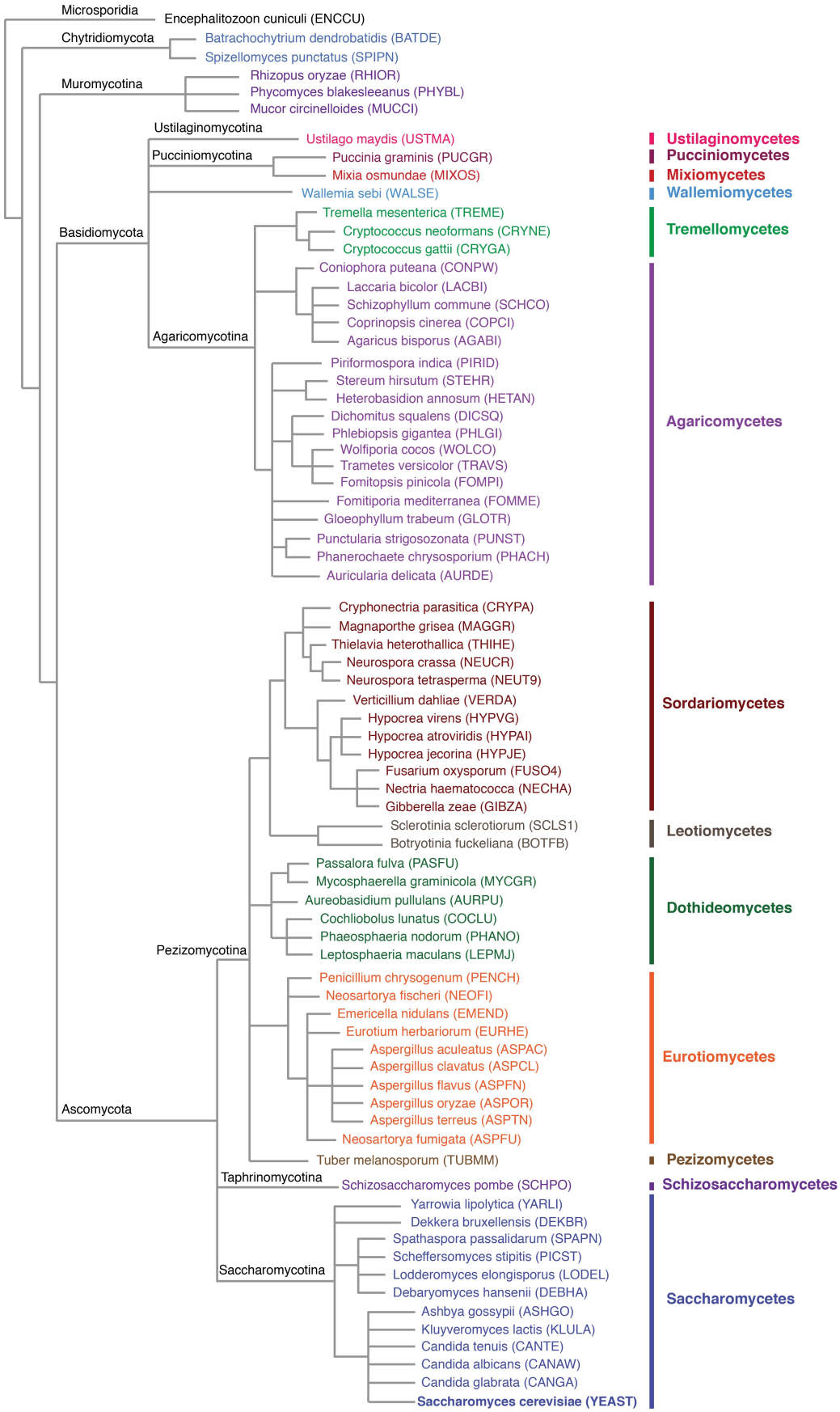
Distribution of proteins with homorepeats (HRPs) in eukaryotic proteomes. The solid black vertical line represents the average number of HRPs found in eukaryotic proteomes. Proteome sequences were obtained from OMA browser (Altenhoff, A.M., *et al., Nucleic Acids Res* **39**, D289–94, 2011). The color of the bars corresponds to the classes given in the phylogenetic tree. The phylogeny was reconstructed based on taxonomic classifications and the branch lengths are just illustrative.



Supplementary notes 2

Snf5 polyQ-mediated interactions with diverse protein complexes

Network showing polyQ mediated Snf5 interactions with multiple protein complexes across diverse molecular functions, obtained by integrating Snf5 interactions with yeast protein complex network (Benschop, J.J. *et al., Molecular cell* **38**, 916-28, 2010). Edge colors depict the carbon source in which the interaction was identified. This result suggests that polyQ in Snf5 influences inter-complex interactions, and our observation that all the Swi/Snf complex members could be detected in all conditions implies that polyQ in Snf5 does not affect its binding to the Swi/Snf members. This is consistent with the identification of the Swi-Snf assembly residue E582 within the SNF5 domain (Geng, F., *et al., Mol Cell Biol* **21**, 4311-20, 2001).

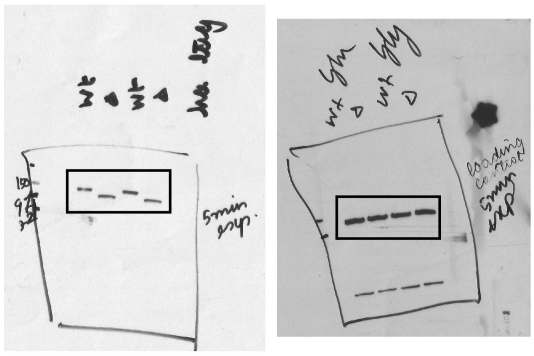

Microsporidia
Encephalitozoon cuniculi (ENCCU)
Chytridiomycota
Batrachochytrium dendrobatidis (BATDE)
Spizellomyces punctatus (SPIPN)
Muromycotina
Rhizopus oryzae (RHIOR)
Phycomyces blakesleeanus (PHYBL)
Mucor circinelloides (MUCCI)
Ustilaginomycotina
Ustilago maydis (USTMA)
Pucciniomycotina
Puccinia graminis (PUCGR)
Mixia osmundae (MIXOS)
Wallemia sebi (WALSE)
Basidiomycota
Tremella mesenterica (TREME)
Cryptococcus neoformans (CRYNE)
Cryptococcus gattii (CRYGA)
Coniophora puteana (CONPW)
Laccaria bicolor (LACBI)
Schizophyllum commune (SCHCO)
Coprinopsis cinerea (COPCI)
Agaricus bisporus (AGABI)
Agaricomycotina
Piriformospora indica (PIRID)
Stereum hirsutum (STEHR)
Heterobasidion annosum (HETAN)
Dichomitus squalens (DICSQ)
Phlebiopsis gigantea (PHLGI)
Wolfiporia cocos (WOLCO)
Trametes versicolor (TRAVS)
Fomitopsis pinicola (FOMPI)
Fomitiporia mediterranea (FOMME)
Gloeophyllum trabeum (GLOTR)
Punctularia strigosozonata (PUNST)
Phanerochaete chrysosporium (PHACH)
Auricularia delicata (AURDE)
Cryphonectria parasitica (CRYPA)
Magnaporthe grisea (MAGGR)
Thielavia heterothallica (THIHE)
Neurospora crassa (NEUCR)
Neurospora tetrasperma (NEUT9)
Verticillium dahliae (VERDA)
Hypocrea virens (HYPVG)
Hypocrea atroviridis (HYPAI)
Hypocrea jecorina (HYPJE)
Fusarium oxysporum (FUSO4)
Nectria haematococca (NECHA)
Gibberella zeae (GIBZA)
Sclerotinia sclerotiorum (SCLS1)
Botryotinia fuckeliana (BOTFB)
Passalora fulva (PASFU)
Mycosphaerella graminicola (MYCGR)
Aureobasidium pullulans (AURPU)
Cochliobolus lunatus (COCLU)
Phaeosphaeria nodorum (PHANO)
Leptosphaeria maculans (LEPMJ)
Pezizomycotina
Penicillium chrysogenum (PENCH)
Neosartorya fischeri (NEOFI)
Emericella nidulans (EMEND)
Eurotium herbariorum (EURHE)
Aspergillus aculeatus (ASPAC)
Aspergillus clavatus (ASPCL)
Aspergillus flavus (ASPFN)
Aspergillus oryzae (ASPOR)
Aspergillus terreus (ASPTN)
Neosartorya fumigata (ASPFU)
Ascomycota
Tuber melanosporum (TUBMM)
Taphrinomycotina
Schizosaccharomyces pombe (SCHPO)
Yarrowia lipolytica (YARLI)
Dekkera bruxellensis (DEKBR)
Spathaspora passalidarum (SPAPN)
Scheffersomyces stipitis (PICST)
Lodderomyces elongisporus (LODEL)
Debaryomyces hansenii (DEBHA)
Saccharomycotina
Ashbya gossypii (ASHGO)
Kluyveromyces lactis (KLULA)
Candida tenuis (CANTE)
Candida albicans (CANAW)
Candida glabrata (CANGA)
Saccharomyces cerevisiae (YEAST)
Ustilaginomycetes
Pucciniomycetes
Mixiomycetes
Wallemiomycetes
Tremellomycetes
Agaricomycetes
Sordariomycetes
Leotiomycetes
Dothideomycetes
Eurotiomycetes
Pezizomycetes
Schizosaccharomycetes
Saccharomycetes

Supplementary notes 3

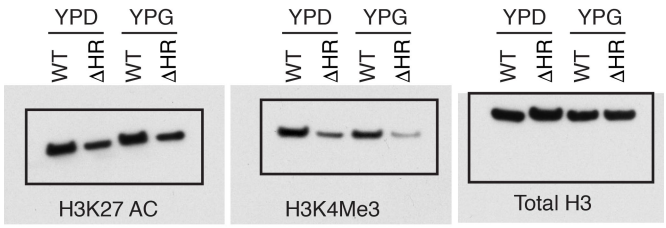
Taxonomic and phylogenetic details of fungal species investigated in this study

The abbreviation provided in the parenthesis for each organism corresponds to the codes used in **Fig. 7d, Supplementary Figs. 9h and 10j**. The phylogeny was reconstructed based on taxonomic classifications and the branch lengths are simply illustrative and do not represent any evolutionary attribute such as divergence times.

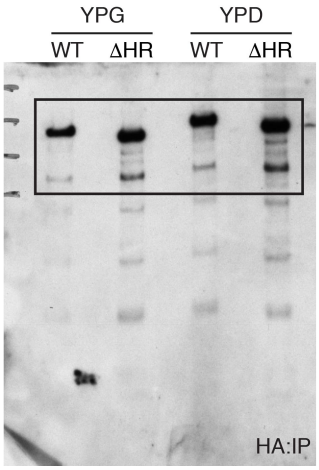
Blots related to Fig. 4c



Blots related to Fig. 4f



Blot related to Supplementary Fig. 5b



Supplementary notes 4

Original blots related to Fig. 4 and Supplementary Fig. 5

Supplementary Table 1: Compendium of datasets used in this study

Type of information (Source)	Description of the data
Genome-scale datasets of *Saccharomyces cerevisiae* (Budding yeast)	
Features related to physiological relevance of proteins (phenotypes)	
Essential genes[1]	Each of the 5916 yeast genes was systematically deleted by mitotic recombination and its effect on growth was examined. About 1105 genes were found be essential for growth on rich glucose medium.
Over-expression toxic genes[2]	In this study, an ordered array of 5280 yeast strains was constructed, with each conditionally overexpressing a unique yeast gene. Each strain carried different yeast ORF and expressed from the inducible GAL1/10 promoter on a multicopy plasmid. Effect of each one of the genes on cellular fitness when overexpressed, was tested in a medium containing galactose, by examining corresponding strains for defects in colony formation. This study identified 769 over-expression toxic genes, as their over-expression resulted in significantly slower growth.
Cell size genes (large and small)[3]	Using a set of yeast ORF deletion strains, 4812 viable haploid deletion strains were surveyed for alterations in the cell size distributions of exponentially growing cultures using Coulter principle. About 200 genes when deleted resulted in large cell size and 190 gene deletions resulted in small cell size.
Slow growth genes[4]	Homozygous gene deletion strains were profiled for fitness defects in rich medium. This study identified 891 non-essential genes whose gene deletion resulted in slow growth.
Haploinsufficient genes[4]	Fitness profiling of heterozygous gene deletion mimicking loss-of-function allele of ~5900 yeast genes in rich medium (YPD) identified 184 haploinsufficient genes, with significant growth defect.

Type of information (Source)	Description of the data
Modulators of aggregation[5]	About 4850 yeast haploid gene deletion mutants of non-essential genes were transformed with constructs that express mutant huntingtin fragment (HD53Q) or a-synuclein under the control of inducible promoters. From these, 137 genes were identified as modulators of aggregation based on the sensitivity (synthetic lethality or sickness) to HD53Q or α-synuclein.
Filamentous growth: Hypo- and hyper-invasive growth genes, hypo- and hyper-biofilm development genes, hypo-and hyper-pseudohyphal growth genes[6]	Genome-wide screening of targeted gene deletion alleles introduced into a filamentous yeast strain S1278b identified 577, 700 and 688 genes involved in hypo- or hyper- (a) haploid invasive growth, (b) pseudohyphal growth and (c) biofilm development, respectively.
Morphological phenotypes[7]	Fluorescence-based imaging of yeast gene-deletion mutants (4786 strains) was undertaken to obtain high-dimensional, quantitative data spanning several morphological features such as cell size, bud size and nucleus location (*Saccharomyces cerevisiae* morphological database). This study identified genes that affect 247 morphological parameters of yeast.
Response to small molecules[8]	About 5000 homozygous non-essential gene deletion strains were tested for growth response in 174 unique conditions, representing various small molecules and environmental stresses. A gene deletion strain was defined as sensitive to a treatment if it showed a statistically significant growth defect in the treatment compared to its growth in control (without treatment) conditions. Sensitivity of a gene deletion to a condition implies that its presence is required for resistance towards that particular environmental perturbation. Of the 174 conditions investigated, we chose 136 small molecule conditions, representing a spectrum of chemical insults.

Type of information (Source)	Description of the data
Features related to functional relevance of proteins	
Genetic interactions[9]	Using a synthetic genetic array and fitness profiling, a genome-scale genetic interaction map for ~75% of all genes in the budding yeast was constructed by examining 5.4 million gene-gene pairs for synthetic genetic interactions. A genetic interaction between a pair of genes was assigned if the double mutant shows significant deviation in fitness compared to the expected multiplicative effect of combining the two single mutants. Based on this, the interactions between two genes were classified as (i) negative genetic interactions: if a more severe fitness defect was observed than expected, with the extreme case being synthetic lethality or (ii) positive genetic interactions: if double mutants showed a less severe fitness defect than expected. This network consisted of 73825 interactions among 4273 genes.
Protein-protein interactions[10]	Protein-protein interactions were obtained from BioGRID database. Interactions that were identified using biochemical approaches such as affinity capture, protein-fragment complementation assay were considered. The network consisted of 73429 interactions among 5398 proteins.
Protein complex network[11]	Using the physical interactome map of yeast, consensus protein complexes (PC) integrating predictions from three different protein complex prediction algorithms were obtained. Briefly, PCs from these different PC prediction sets were mapped against each other using hypergeometric testing. Based on the reciprocal matches between the different PCs, PCs from the different sets are grouped together as a PC cluster, if they have significant overlap of members. The PCs are then aligned within each cluster, and only those PC members that are present in at least two originating PCs are retained in the consensus PC.

Type of information (Source)	Description of the data
	The protein-complex network consisted of 494 complexes with 1890 members.
Transcriptional regulatory network (Protein-DNA interactions)[12]	Yeast transcriptional regulatory network (TRN) was reconstructed by combining a previously published TRN[13] with the recent genome-wide *in vivo* binding map of yeast regulatory proteins[14]. For the latter, promoter occupancy cutoff of at least 3-fold higher than background was considered. This network consisted of 29426 interactions between 169 transcription factors and 5621 targets.
RNA binding protein network (Protein-RNA interactions)[15]	In this study, RNAs bound to RNA binding proteins (RBPs) were identified using a two-step approach. Each RBP was TAP-tagged, expressed under control of their native promoters. Two sets of RNAs were extracted (i) RBP-bound mRNA by immunoprecipitation of messenger-ribonucleoproteins using affinity purification, and (ii) cellular RNA representing the whole set of transcripts in the cell. Subsequently, hybridization of the two isolated RNA samples using dual-color microarrays was done and analyzed for enriched transcripts to detect the bound targets of a RBP. This network consisted of 13514 interactions between 41 RBPs and 4416 transcripts.
Gene perturbation network[16]	Expression profiling of strains with single gene deletion of each of 1481 putative regulators was undertaken using DNA microarrays. A common reference experiment design was adopted in which cRNA from replicate mutant cultures was cohybridized in dye-swap with cRNA from a batch of common reference WT RNA, against which comparisons were drawn to measure alterations of gene expression. By considering only robust expression changes, the gene perturbation network (GPN) was reconstructed and consists of 50294 edges between 700 regulators and 3014 targets.

Type of information (Source)	Description of the data
Protein solubility[17]	Lysates from unstressed yeast cells in SILAC light or heavy medium were subjected to high-speed centrifugation. Solubility of proteins was determined by mass spectrometry based comparison of the light labeled supernatant and the heavy labeled pellet. Linear regression of heavy/light ratio on a log scale was done and the proteins with values along the flattest portion of the curve were classified as those with normal solubility. The intercept of the linear regression at x = 0 and x=1 were used to delimit the low solubility and high solubility proteins respectively.
Stress granule proteins and P-body proteins[18]	Yeast stress granule cores expressing Pab1-GFP1 were affinity purified using a TAP-tagged eIF4G1. A minimum of 2-fold enrichment of peptides over unstressed control and complete absence from untagged control was used to identify members of stress granule proteome. Through literature mining, Jain et al., further collated members of P-body proteome.
Heritable proteins[19]	In this study, yeast proteome was screened for the ability to elicit to stable biological traits by transient over-expression of each of ~5300 yeast ORFs. About 50 proteins showed heritable epigenetic states, persistent over hundreds of generations, after expression levels returned to normal and were inherited from mother to daughter cells.
Post-translational modifications[20]	We compiled a list of experimentally identified PTM sites from curated Swissprot subsection of the Uniprot database and by literature curation. The list included 12470 post-translational modification sites across 2648 yeast proteins.
Putative linear motifs[21]	Putative linear motifs were identified using ANCHOR, which predicts protein-binding regions that are disordered in isolation but can undergo disorder-to-order transition upon binding using estimated energy calculations.

Type of information (Source)	Description of the data
Features related to Proteostasis	
Protein abundance[22]	Endogenous protein levels during log-phase of growth in rich medium were obtained by measuring the intensity of GFP tagged proteins using flow cytometry.
Relative translational rate[23]	In this study, polysome fractionation using velocity sedimentation, followed by a quantitative microarray analysis of several fractions across the gradient was undertaken to measure the translational status of each mRNA. Translational status was measured as a function of (i) Ribosomal density - the number of ribosomes per unit ORF length and (ii) Ribosomal occupancy - the fraction of transcripts of each gene engaged in translation. Relative translational rate was estimated as the product of ribosomal density and ribosomal occupancy. Since Ingolia et al.,[24] focused on genes that were relatively free of repetitive sequences, we did not consider this dataset for our analyses.
Protein half-life[25]	*In vivo* protein half-lives were determined by inhibiting protein synthesis with cycloheximide and then monitoring the abundance of each C-terminally TAP-tagged protein in the yeast genome by quantitative western blotting at three time points. We disregarded 366 proteins with a half-life of exactly 300 minutes, as these values were assigned to stable proteins for which degradation curves could not be fitted by an exponential decay function. Further, we disregarded seven proteins with extremely long half-lives of >6000 minutes.
Factors influencing Protein synthesis	
Translation initiation	
Poly(A) tail length[26]	Using polyadenylation state array (PASTA) analysis, which combines separation of cellular mRNA on poly(U) Sepharose with subsequent microarray analyses, the yeast transcriptome

Type of information (Source)	Description of the data
	was surveyed and mRNA groups with tendencies toward either long or short tails during steady state growth were identified.
RNA secondary structure (5'UTR and coding region)[27]	Using parallel analysis of RNA structure (PARS) *in vitro* profiling of the secondary structure of yeast mRNAs at single nucleotide resolution was carried out. This involves deep sequencing the transcript fragments that were treated with RNase V1 (preferentially cleaves at double-stranded RNA) and S1 nuclease (preferentially cleaves single-stranded RNA). From this PARS score was estimated, which is a $\log_2$ of the ratio between the number of times the nucleotide immediately downstream of the inspected nucleotide was observed as the first base when treated with RNase V1 and the number of times it was observed in the S1 nuclease treated sample. PARs score represents the likelihood of each nucleotide in a single- or double-stranded conformation. From this, we computed average PARS scores across 5' UTRs and coding regions.
5'UTR sequences[28]	Using a combination of 5' rapid amplification of cDNA ends (5'RACE) and RNA-sequencing, 5'UTR sequences of yeast protein coding genes was mapped.
Translational efficiency[29]	Optimality of codon usage was estimated using normalized translational efficiency (nTE). The nTE reflects codon optimality based on the relative tRNA abundance (supply) over the cognate codon usage (demand). Therefore, codons whose cognate tRNA availability exceeds their relative usage are considered optimal. We used nTE averaged over the entire length of the coding sequence and individual nTE estimates of first 400 codons.
Factors influencing Protein degradation	
Long N-terminal disorder[30]	Yeast proteins with long disordered tails of length >30 residues at the N-terminus were identified by inferring disordered status

Type of information (Source)	Description of the data
	of every residue using three different disorder predictors. Minor stretches (upto three consecutive residues) of structured residues were allowed within the N-terminus disordered segment. This means that the disordered stretch ended when encountering a minimum of four consecutive structured residues.
Endoproteolytic sites[30]	Internal disordered segments of length >40 residues were identified as endoproteolytic sites. Similar to the identification of long N-terminal disorder, upto three consecutive structured residues were permitted in the identification of endoproteolytic sites.
PEST motif	We predicted PEST regions using epestfind with default parameters, as included in EMBOSS 6.5.7[31].
Destruction box and KEN box[30]	KEN box and destruction box motifs in yeast proteins were predicted using GPS-ARM 1.0 with default parameters[32].
Features related to evolution of protein sequences	
Protogenes[33]	Annotated yeast ORFs found only in yeast or only in the four closely related *Saccharomyces sensu stricto* species but not in the rest of the tested Ascomycota class were classified as protogenes.
Paralogs	Paralog proteins pairs were obtained by performing an all-against-all pairwise alignments and subsequently clustering them using BLASTClust[34]. Sequences with at least 30% identity, covering 60% of length were classified as paralog pairs. Paralog pairs identified to have arisen from the yeast whole-genome duplication event were also added[35].
Orthologs[36,37]	One-to-one orthologs of budding yeast proteins from 74 species, spanning 16 fungal classes (**Fig. 7c; Supplementary Notes 3**), were obtained from OMA browser (July 2013

Type of information (Source)	Description of the data
	release). Orthologs were defined by OMA's inference algorithm, which first infers homologous sequences by performing an all-against-all Smith-Waterman alignments between all sequences and retain significant matches. Subsequently, orthologous pairs (the subset of homologs related by speciation events) were inferred using mutually closest homologs based on evolutionary distances, taking into account distance inference uncertainty and the possibility of hidden paralogy due to differential gene losses.
Yeast natural variation[38]	Missense variations identified by whole genome-sequences of 39 diverse yeast strains (Sanger sequencing) were compiled to catalog amino acid substitutions. The yeast strains included the reference strain S288c and other lab strains, pathogenic strains and those involved in baking, wine, food spoilage, natural fermentation, sake and strains obtained from probiotic and plant isolates.
Genome-scale datasets for *Schizosaccharomyces pombe* (Fission yeast)	
Protein abundance[39]	Protein abundance was estimated in *S.pombe* quiescent cells, 24h after nitrogen removal using mass spectrometry (MS) analysis. By using absolute abundance of 39 proteins, which was determined by spiked-in heavy reference peptides, the summed MS-intensities of all peptides were translated to copies/cells for all proteins analysed.
Ribosomal density[40]	Genome-wide translational profiling of vegetatively growing *S. pombe* cells was obtained using polysome fractionation, followed by a quantitative microarray analysis with RNA fractions representing different numbers of associated ribosomes. Ribosomal density was measured as the number of ribosomes per unit of transcript length.
Protein half-life[41]	Using a label switch approach, yeast cells labelled with heavy

Type of information (Source)	Description of the data
	isotopes were diluted in media with an excess of normal lysine. The decay of the heavy lysine signal in the proteome over time measured using high-resolution mass spectrometry-based proteomics was used to estimate protein half-lives.

References

1. Giaever, G. et al. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387-91 (2002).

2. Sopko, R. et al. Mapping pathways and phenotypes by systematic gene overexpression. *Molecular cell* **21**, 319-30 (2006).

3. Jorgensen, P., Nishikawa, J.L., Breitkreutz, B.J. & Tyers, M. Systematic identification of pathways that couple cell growth and division in yeast. *Science* **297**, 395-400 (2002).

4. Deutschbauer, A.M. et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915-25 (2005).

5. Willingham, S., Outeiro, T.F., DeVit, M.J., Lindquist, S.L. & Muchowski, P.J. Yeast genes that enhance the toxicity of a mutant huntingtin fragment or alpha-synuclein. *Science* **302**, 1769-72 (2003).

6. Ryan, O. et al. Global gene deletion analysis exploring yeast filamentous growth. *Science* **337**, 1353-6 (2012).

7. Ohya, Y. et al. High-dimensional and large-scale phenotyping of yeast mutants. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 19015-20 (2005).

8. Hillenmeyer, M.E. et al. The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362-5 (2008).

9. Costanzo, M. et al. The genetic landscape of a cell. *Science* **327**, 425-31 (2010).

10. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2013 update. *Nucleic acids research* **41**, D816-23 (2013).

11. Benschop, J.J. et al. A consensus of core protein complex compositions for Saccharomyces cerevisiae. *Molecular cell* **38**, 916-28 (2010).

12. Gemayel, R. et al. Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. *Mol Cell* **59**, 615-27 (2015).

13. Balaji, S., Babu, M.M., Iyer, L.M., Luscombe, N.M. & Aravind, L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol* **360**, 213-27 (2006).

14. Venters, B.J. et al. A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. *Mol Cell* **41**, 480-92 (2011).

15. Hogan, D.J., Riordan, D.P., Gerber, A.P., Herschlag, D. & Brown, P.O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS biology* **6**, e255 (2008).

16. Kemmeren, P. et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**, 740-52 (2014).

17. Albu, R.F. et al. A feature analysis of lower solubility proteins in three eukaryotic systems. *J Proteomics* **118**, 21-38 (2015).

18. Jain, S. et al. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487-98 (2016).

19. Chakrabortee, S. et al. Intrinsically Disordered Proteins Drive Emergence and Inheritance of Biological Traits. *Cell* **167**, 369-381 e12 (2016).

20. UniProt, C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* **40**, D71-5 (2012).

21. Dosztanyi, Z., Meszaros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25**, 2745-6 (2009).

22. Newman, J.R. et al. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature* **441**, 840-6 (2006).

23. Arava, Y. et al. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3889-94 (2003).

24. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218-23 (2009).

25. Belle, A., Tanay, A., Bitincka, L., Shamir, R. & O'Shea, E.K. Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 13004-9 (2006).

26. Beilharz, T.H. & Preiss, T. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. *RNA* **13**, 982-97 (2007).

27. Kertesz, M. et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103-7 (2010).

28. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-9 (2008).

29. Pechmann, S. & Frydman, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature structural & molecular biology* **20**, 237-43 (2013).

30. van der Lee, R. et al. Intrinsically disordered segments affect protein half-life in the cell and during evolution. *Cell Rep* **8**, 1832-44 (2014).

31. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-7 (2000).

32. Liu, Z. et al. GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-boxes and KEN-boxes. *PLoS One* **7**, e34370 (2012).

33. Carvunis, A.R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370-4 (2012).

34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).

35. Wolfe, K.H. & Shields, D.C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-13 (1997).

36. Altenhoff, A.M., Schneider, A., Gonnet, G.H. & Dessimoz, C. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res* **39**, D289-94 (2011).

37. Schneider, A., Dessimoz, C. & Gonnet, G.H. OMA Browser--exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23**, 2180-2 (2007).

38. Liti, G. et al. Population genomics of domestic and wild yeasts. *Nature* **458**, 337-41 (2009).

39. Marguerat, S. et al. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151**, 671-83 (2012).

40. Lackner, D.H. et al. A network of multiple regulatory layers shapes gene expression in fission yeast. *Molecular cell* **26**, 145-55 (2007).

41. Christiano, R., Nagaraj, N., Frohlich, F. & Walther, T.C. Global proteome turnover analyses of the Yeasts S. cerevisiae and S. pombe. *Cell Rep* **9**, 1959-65 (2014).

Supplementary Table 2: List of primers used in this study

Primer name	Primer sequence	Amplicon
Primers for SNF5 repeat sequencing		
2324-TRsz-SNF5-F2	AGAGGCAATTGCTGGTTCAG	PolyQ deletion check
2325-TRsz-SNF5-R2	AGTTGAGGAAGTTGGCCAATAGT	PolyQ deletion check
Primers for SNF5 repeat deletion construction		
2355-dTR-SNF5-HYG-F	AAAAAGAGATATATAACTTTTTT AAGTGATCGGCTGGTAAATAGAC TTATAACGAGTAGATGCCATCTTT GTACAGCTTGCCT	Hygromycin resistance cassette
2356-dTR-SNF5-HYG-R	AAGGTCTTAGTTATGCTAGCCTGA AATGTATTATTGTGTACAATATAT CATCTAATGTTTCGCAGAGCCGTG GCAGG	Hygromycin resistance cassette
2357-dTR-SNF5-F	CAACTTAGAAACCAAATACAGCG ACAACAGCAACAACAGTTTAGGC ATCATGTGCAAATAGGACAAATA CCGCAATCTCAGCAAGTTCCT	Snf5 poly Q deletion
Primers for tagged SNF5 construction		
4860-RG-SNF5-6xHA-F	GAAGCGACATTGTTGACTAATAG CAATAATGGTAGCAGTAACAATA ACACACAGAATACATCCGGTTCT GCTGCTAGATACC	6X HA –KanR Tag
4861-RG-SNF5-6xHA-R	TACAAATTCTTCCACGGTTATTTA CATCTCCGGTATATTTTATATATG TGTATATATTTTGCATAGGCCACT AGTGGATCTG	6X HA –KanR Tag