Genome architecture and stability in the *S. cerevisiae* knockout collection

Fabio Puddu[1,2], Mareike Herzog[1,2], Alexandra Selivanova[1], Siyue Wang[1], Jin Zhu[3], Shir Klein-Lavi[4], Molly Gordon[3], Roi Meirman[4], Gonzalo Millan-Zambrano[1], Iñigo Ayestaran[1], Israel Salguero[1], Roded Sharan[5], Rong Li[3], Martin Kupiec[4], and Stephen P. Jackson[1]

Affiliations:

[1]The Wellcome Trust/CRUK Gurdon Institute and Department of Biochemistry, Cambridge, CB2 1QN, UK

[2]The Wellcome Trust Sanger Institute, Hinxton, UK

[3]Department of Cell Biology, Center for Cell Dynamics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

[4]School of Molecular Cell Biology and Biotechnology, Tel Aviv University, Ramat Aviv 69978, Israel.

[5]School of Computer Science, Tel Aviv University, Ramat Aviv 69978, Israel.

While identification and analysis of genes affecting genome stability have traditionally relied on reporter assays, whole-genome sequencing technologies now enable, in principle, direct measurements of genome instability globally and at scale. Here, we have surveyed the *Saccharomyces cerevisiae* gene knockout collection by sequencing the whole genomes of its strains, and characterized genomic changes caused by the absence of essentially any one of the non-essential yeast genes. Analysing this dataset (http://sgv.gurdon.cam.ac.uk) reveals genes affecting repetitive-element maintenance or mutagenesis, highlights cross-talks between nuclear and mitochondrial genome stability, and shows how strains have adapted to loss of non-essential genes.

Understanding gene functions is a key goal of biomedical research. A major scientific milestone was the completion of the *S. cerevisiae* gene knockout collection (YKOC), which facilitated high-throughput phenotypic screens, genetic-interaction analyses and reverse genetics[1-3]. Ensuing experimental work highlighted occasional inconsistencies, mistakes, and genome changes rebalancing the effects of specific knockouts[4-6] but a complete overview is lacking. To address how gene loss influences genomic stability/structure, we whole-genome sequenced (WGS) nearly all of the 4,732 strains of the homozygous diploid YKOC. After identifying and reassigning targeted genes when they did not correspond to the expected, we confirmed most knockout strains (**Extended Data Fig. 1a-b**; for all gene-specific information, see supplementary tables 1-4 and http://sgv.gurdon.cam.ac.uk). We then determined each strain's genome-instability profile, particularly in regard to repetitive DNA (**Fig. 1a**).

rRNA defects lead to rDNA expansion

We developed and validated tools to accurately measure repetitive DNA copy-numbers based on the concept that, when a sequenced genome contains more copies of a locus than the single copy-number reference genome, that locus displays proportionately elevated sequence coverage (**Extended Data Fig. 2 and methods)**. Global analysis of the repetitive rDNA locus that encodes large ribosomal RNAs (rRNAs) revealed copy-number alterations in ~10% of strains (**Fig. 1b**). Knockouts known to affect rDNA-locus size, such as *RTT109, DPB3* and *DPB4*[7] displayed rDNA copy numbers in line with expectations (**Fig. 1c**). Comparing genes predicted to affect rDNA locus size with a previous study[8] revealed a partial overlap as well as many new players (**Fig. 1d**). Strains with rDNA expansion included knockouts affecting rDNA transcription, such as those for RNA Pol I transcription factor Uaf30 or TORC subunits Tor1 and Tco89[9] (**Fig. 1c**), suggesting that reduced rDNA transcription fosters rDNA-locus expansion. Indeed, measuring rDNA copy numbers in strains bearing thermosensitive *RPA190* (RNA Pol I subunit) or *RRN3* (Pol I recruitment factor) alleles revealed temperature-dependent rDNA-locus size increases, in line with recent findings[10] (**Fig. 1e**). Knockout of *HAM1* or *LOG1* (*YJL055W)*, encoding (d)NTP-pool sanitising factors[11], was also associated with a larger rDNA locus, probably because incorporation of non-canonical nucleotides into rRNA impairs its functionality (**Fig. 1c**). Because knockouts affecting tandem-repeat stability could affect rDNA-locus size, we compared rDNA and repetitive *CUP1*-locus copy-numbers (**Extended Data Fig. 1c**). Strains with high positive correlations included *TEL1*, *RAD27*, *SSN8* and *ZDS2* knockouts, but lack of overall correlation suggests that different selective pressures largely drive variation at these loci (**Extended Data Fig. 1d-e**).

Genes affecting telomere length

Copy-number distributions for other genomic elements (telomeres, 2µ plasmid and Ty transposons) had lower variabilities than rDNA (**Fig. 1f and Extended Data Fig. 1c**; 2µ copy numbers did not follow a normal distribution, see **Extended Data Fig. 1d** for discussion). Telomere-length estimates highlighted strains lacking telomerase subunits (*est1*Δ, *est2*Δ, *est3*Δ) as having longest telomeres, in line with alternative telomere elongation (ALT) not being subject to normal telomere-length homeostasis. Our predicted *TLM* (telomere maintenance) genes partially overlapped with those defined previously[12,13] (**Fig. 1g-h**). To validate predictions for novel *TLM* genes, we measured telomere lengths by gel electrophoresis (**Fig. 1g-h**, bold gene names**)**, confirming telomeric phenotypes for 12 of 14 selected (**Fig. 1i and Extended Data Fig. 3a**; as shown in **Extended Data Fig 3b**, we also validated telomeric impacts for ~70% of a selection of strains failing stringent criteria but with high/low telomere-length estimates). Network analysis of predicted *TLM* genes prominently featured DNA-, RNA- and/or chromatin-metabolism, and highlighted connections to endoplasmic reticulum, Golgi and/or vesicle trafficking[14] (**Extended Data Fig 3c**).

Aneuploidies in the YKOC

Sequencing coverage also offers information on copy-number variation between and within chromosomes. Strikingly, ~52% of sequenced colonies displayed some deviation from 2n ploidy, with a minority exhibiting stable chromosome gains and/or losses (see Methods and **Fig. 2a**). Most aneuploid strains displayed "incomplete" aneuploidies (gain/loss of less than a whole chromosome unit), suggesting cell culture heterogeneity. These events affected different chromosomes differently, with the smallest chromosomes most frequently lost (I, VI, and III; in line with chromosome loss negatively correlating with chromosome size[15]) and the largest most frequently gained (chr XII; in line with its delayed segregation likely resulting in occasional non-disjunctions[16]; **Extended Data Fig. 4a-b)**. Highly-aneuploid strains included knockouts encoding proteins involved in chromosome segregation such as Sgo1[17], Eos1[18], as well as genes without documented connections to karyotype maintenance. While some of these can be explained by the deletion cassette affecting a neighbouring gene (e.g. *iwr1Δ on NUP84*)[19], the origin of aneuploidy in others requires further investigation. To determine if high aneuploidy correlates with chromosome instability (CIN), we used a mini-chromosome CIN reporter[20] in freshly-generated knockouts (**Fig. 2b**). Most of these had marginally increased CIN, with only a small subset displaying a strong CIN phenotype, suggesting that most aneuploidies do not originate from severe chromosome-segregation defects caused by the gene knockout. In line with this, when we focused on ribosomal proteins — often encoded by two near-identical gene paralogs — in 28/97 cases analysed, gene-deletion was accompanied by gain of the chromosome carrying the paralog (**Extended Data Fig. 4c**).

Approximately 2% of strains carried single or multiple sub-chromosomal deletions/amplifications, often initiating at Ty transposons or interspersed LTR sites and terminating at similar sites or chromosome ends **(Fig. 2c-d)**. As previously observed[21-23], breakpoint-localisations to repetitive regions suggests that homologous recombination (HR) processes drive such rearrangements. Indeed, mapping the remaining breakpoints revealed strains *rad2*Δ and *mrpl21*Δ to carry a single-copy gain or loss of the same chromosome IV fragment whose boundaries map to the homologous *RPL35A* and *RPL35B* genes. As with aneuploidies, we found evidence of rearrangements compensating effects of gene knockouts (**Extended Data Fig. 4d).** Together with previous works[6,24-26], our results suggest that relatively frequent chromosomal mis-segregations permit evolution of phenotypic suppression and fitness improvements in many knockout settings.

Genome instability drives increased mtDNA

Estimating mitochondrial DNA (mtDNA) copy-numbers from whole-genome sequencing data (**Extended Data Fig. 5**) indicated that ~6% of non-essential genes are required to maintain mtDNA (**Figure 3a)** and that these overlap with genes required for respiratory growth or encoding mitochondrial proteins (**Extended Data Fig. 6a)**. When we assessed connections between various features of nuclear-genome instability, and between these and mtDNA copy-number, we found correlations between aneuploidies and mtDNA copy-number changes (and between aneuploidies and altered Ty copy-numbers, since chromosome gains/losses can alter Ty-element numbers; **Fig. 3b**). Indeed, strains with higher deviations from 2n ploidy were over-represented among *rho*^{0} colonies that lack detectable mtDNA (**Fig. 3c)**. The extent to which connections between mtDNA loss and aneuploidy reflect iron-sulphur-cluster formation defects[27] and/or other mechanisms remains to be explored.

Mitochondrial DNA analyses also highlighted ~130 knockouts associated with increased mtDNA (*rho*$^{++}$*;* **Fig. 3a**). Compared to other strains, *rho*$^{++}$ strains collectively displayed higher deviation from diploidy (**Fig. 3d**) and were enriched in genes related to DNA repair/metabolism rather than to mitochondrial biology (**Extended Data Fig. 6b**). We hypothesised that DNA-repair defects might lead to *rho*$^{++}$ phenotype via persistent DNA-damage response (DDR) activation. Indeed, genotoxin-mediated DDR activation in wild-type strains triggered increased mtDNA copy-number **(Fig. 3e)**. mtDNA copy-number increases can arise from overexpression of ribonucleotide reductase (RNR) components[28] that are under DDR control[29]; and accordingly, knockouts of RNR-gene transcriptional repressors (*RFX1*, *TUP1*, *CYC8*) led to mtDNA copy-number increases (**Fig. 3f**). Furthermore, of 16 *rho*$^{++}$ strains selected for analysis, 9 displayed RNR induction, 6 displayed spontaneous Rad53 phosphorylation (another marker of DDR activation), and 13 were hypersensitive to hydroxyurea (HU), an RNR inhibitor causing replication stress (**Extended Data Fig. 6c**). Furthermore, HU-hypersensitive KO strains[30,31] were enriched among *rho*$^{++}$ strains (**Extended Data Fig. 6d**). Comparing mtDNA contents with systematic analyses of Rnr3 levels[32] revealed that many KOs deleted in genes protecting genome stability displayed both increased Rnr3 levels and a *rho*$^{++}$ phenotype (**Extended Data Fig. 6e, green**). Accordingly, overexpressing the RNR catalytic subunit Rnr3, and to a lesser extent Rnr1, increased mtDNA levels (**Fig. 3g;** overexpressing Rnr2 or Rnr4 did not). As Rnr1/Rnr3 overexpression increases dNTP levels[33,34], these results suggested that elevated dNTP production increases mtDNA copy-number, perhaps by stimulating mtDNA replication. Curiously, among *rho*$^{++}$ strains with normal Rnr3 levels were knockouts of genes encoding tryptophan biosynthesis-pathway enzymes (**Extended Data Fig. 7**). The basis for this connection requires further investigation.

Alleles suppressing mtDNA loss

Extracting information on single-nucleotide variants (SNVs) and insertion/deletions (INDELs) across the YKOC revealed genes carrying higher than expected mutation numbers (**Fig 4a**). Inspecting their coding sequences revealed microsatellite features (homopolymers and/or small degenerate tandem repeats), suggesting that these variants had arisen via imprecise DNA replication or imprecise sequencing/read-mapping (**Fig 4a**, grey bars **and Extended Data Fig 8**). Another subset of frequently-mutated genes each carried one very frequent “founder” mutation (**Fig 6a**, yellow bars) found predominantly in KO strains generated by single YKOC consortium laboratories (**Extended Data Fig 9**). The remaining frequently-mutated genes included *ATP1*, *ATP2* and *ATP3*, encoding subunits of the ATP-synthase complex recently identified as a cell-fitness hub[6]. Analysis of mutation prevalence revealed that *ATP1*, *ATP2*, *ATP3* and *SIT4* (a phosphatase acting on ATP-synthase[35]) mutations were frequent in *rho*^{0} strains but relatively rare in other strains. (**Fig 4b**). These results suggest that alteration of ATP-synthase function can promote fitness of *rho*^{0} cells.

HR defects yield a SNV signature

As YKOC strains have been cultured over many years, overall mutation numbers could highlight hyper-mutators. Thus, we estimated sequence divergence of two colonies of the same KO strain from each other and from wild-type (BY4743). Hyper-mutators are expected to show increases in both; and indeed, DNA-mismatch repair (MMR) defective strains were readily detected this way (**Fig 4c-d**; *pms1*Δ, *msh2*Δ, *msh6*Δ; only one *mlh1*Δ colony sequence was available). Many strains lacking HR components (*rad51*Δ, *rad52*Δ, *rad55*Δ, *rad57*Δ) also showed a similar phenotype, implying that HR deficiencies yield SNVs, possibly via mutagenic trans-lesion polymerases[36,37]. Distinct mutation patterns can be used to cluster genome-stability genes into functional categories[38]. Clustering of hypermutators that we identified based on their mutation profiles produced two main groups corresponding to known MMR and HR genes, and also clustered *ymr166c*Δ and *ggc1*Δ with the MMR group (**Fig 4e**). *YMR166C* is located upstream of *MLH1*, so its deletion could affect *MLH1* expression[19], while SNV abundance in *ggc1*Δ remains to be explained.

Our systematic survey of genomic and karyotypic features highlights impacts of gene knockouts on genome architecture and underscores the flexibility of the yeast genome to mitigate effects of gene loss. Overall, ~36% of YKOC strains carry some repetitive DNA/chromosomal abnormality; and among these, loss of 151 genes is associated with multiple alterations, suggesting that their gene-products protect against widespread genome instability (**Extended Data Fig. 10a**). While gene-ontology analysis revealed strong enrichments for genome biology (**Extended Data Fig. 10b**), the largest fraction of these genes encodes proteins controlling mitochondrial functions (**Extended Data Fig. 10c**). Further work is required to understand these connections and how nuclear and mitochondrial genome maintenance is intertwined. Crucially, our results highlight repetitive-DNA alterations as new types of mutation signatures, in addition to SNVs first identified in cancers[39].It will be interesting to integrate this resource with data from future studies assessing propagated lineages of knockout strains, not only for repetitive-element numbers, but also their variability, and determine whether and how repetitive-element instability might relate to other types of genome instability. Sequencing genomes of strains bearing hypomorphic alleles of essential genes and of synthetic-sick double-mutants[40] should also provide insights into processes impacting genomic stability. Such work may have medical relevance because, together with SNVs and INDELs, copy-number variations and aneuploidies are linked to developmental disorders, cancer and other diseases.

References

1. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science (New York, N.Y.)* **294,** 2364‑2368 (2001).

2. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418,** 387‐391 (2002).
3. Giaever, G. & Nislow, C. The yeast deletion collection: a decade of functional genomics. *Genetics* **197,** 451‐465 (2014).
4. Hughes, T. R. *et al.* Widespread aneuploidy revealed by DNA microarray expression profiling. *Nature genetics* **25,** 333‐337 (2000).
5. Lehner, K. R., Stone, M. M., Farber, R. A. & Petes, T. D. Ninety-six haploid yeast strains with individual disruptions of open reading frames between YOR097C and YOR192C, constructed for the Saccharomyces genome deletion project, have an additional mutation in the mismatch repair gene MSH3. *Genetics* **177,** 1951‐1953 (2007).
6. van Leeuwen, J. *et al.* Exploring genetic suppression interactions on a global scale. *Science (New York, N.Y.)* **354,** aag0839‐aag0839 (2016).
7. Ide, S., Saka, K. & Kobayashi, T. Rtt109 prevents hyper-amplification of ribosomal RNA genes through histone modification in budding yeast. *PLoS genetics* **9,** e1003410 (2013).
8. Saka, K., Takahashi, A., Sasaki, M. & Kobayashi, T. More than 10% of yeast genes are related to genome stability and influence cellular senescence via rDNA maintenance. *Nucleic acids research* **44,** 4211‐4221 (2016).
9. Claypool, J. A. *et al.* Tor pathway regulates Rrn3p-dependent recruitment of yeast RNA polymerase I to the promoter but does not participate in alteration of the number of active genes. *Molecular biology of the cell* **15,** 946‐956 (2004).
10. Mansisidor, A. *et al.* Genomic Copy-Number Loss Is Rescued by Self-Limiting Production of DNA Circles. *Molecular Cell* **72,** 583‐593.e4 (2018).
11. Carlsson, M., Gustavsson, M., Hu, G.-Z., Murén, E. & Ronne, H. A Ham1p-Dependent Mechanism and Modulation of the Pyrimidine Biosynthetic Pathway Can Both Confer Resistance to 5-Fluorouracil in Yeast. *PloS one* **8,** e52094 (2013).
12. Askree, S. H. *et al.* A genome-wide screen for Saccharomyces cerevisiae deletion mutants that affect telomere length. *Proc Natl Acad Sci U S A* **101,** 8658‐8663 (2004).
13. Gatbonton, T. *et al.* Telomere length as a quantitative trait: genome-wide survey and genetic mapping of telomere length-control genes in yeast. *PLoS genetics* **2,** e35 (2006).
14. Rog, O., Smolikov, S., Krauskopf, A. & Kupiec, M. The yeast VPS genes affect telomere length regulation. *Curr. Genet.* **47,** 18‐28 (2005).
15. Murray, A. W., Schultes, N. P. & Szostak, J. W. Chromosome length controls mitotic chromosome segregation in yeast. *Cell* **45,** 529‐536 (1986).
16. Sullivan, M., Higuchi, T., Katis, V. L. & Uhlmann, F. Cdc14 phosphatase induces rDNA condensation and resolves cohesin-independent cohesion during budding yeast anaphase. *Cell* **117,** 471‐482 (2004).
17. Indjeian, V. B., Stern, B. M. & Murray, A. W. The centromeric protein Sgo1 is required to sense lack of tension on mitotic chromosomes. *Science (New York, N.Y.)* **307,** 130‐133 (2005).

18. Daniel, J. A., Keyes, B. E., Ng, Y. P. Y., Freeman, C. O. & Burke, D. J. Diverse functions of spindle assembly checkpoint genes in Saccharomyces cerevisiae. *Genetics* **172,** 53-65 (2006).
19. Ben-Shitrit, T. *et al.* Systematic identification of gene annotation errors in the widely used yeast mutation collections. *Nature methods* **9,** 373-378 (2012).
20. Zhu, J. *et al.* Single-Cell Based Quantitative Assay of Chromosome Transmission Fidelity. *G3 (Bethesda)* **5,** 1043-1056 (2015).
21. Argueso, J. L. *et al.* Double-strand breaks associated with repetitive DNA can reshape the genome. *Proc Natl Acad Sci U S A* **105,** 11845-11850 (2008).
22. Vernon, M., Lobachev, K. & Petes, T. D. High rates of ‘unselected’ aneuploidy and chromosome rearrangements in tel1 mec1 haploid yeast strains. *Genetics* **179,** 237-247 (2008).
23. Lemoine, F. J., Degtyareva, N. P., Lobachev, K. & Petes, T. D. Chromosomal translocations in yeast induced by low levels of DNA polymerase a model for chromosome fragile sites. *Cell* **120,** 587-598 (2005).
24. Yona, A. H. *et al.* Chromosomal duplication is a transient evolutionary solution to stress. *Proc Natl Acad Sci U S A* **109,** 21010-21015 (2012).
25. Rancati, G. *et al.* Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135,** 879-893 (2008).
26. Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519,** 349-352 (2015).
27. Veatch, J. R., McMurray, M. A., Nelson, Z. W. & Gottschling, D. E. Mitochondrial Dysfunction Leads to Nuclear Genome Instability via an Iron-Sulfur Cluster Defect. *Cell* **137,** 1247-1258 (2009).
28. Taylor, S. D. *et al.* The conserved Mec1/Rad53 nuclear checkpoint pathway regulates mitochondrial DNA copy number in Saccharomyces cerevisiae. *Molecular biology of the cell* **16,** 3010-3018 (2005).
29. Huang, M., Zhou, Z. & Elledge, S. J. The DNA replication and damage checkpoint pathways induce transcription by inhibition of the Crt1 repressor. *Cell* **94,** 595-605 (1998).
30. Parsons, A. B. *et al.* Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature biotechnology* **22,** 62-69 (2004).
31. Woolstencroft, R. N. *et al.* Ccr4 contributes to tolerance of replication stress through control of CRT1 mRNA poly(A) tail length. *J. Cell. Sci.* **119,** 5178-5192 (2006).
32. Hendry, J. A., Tan, G., Ou, J., Boone, C. & Brown, G. W. Leveraging DNA damage response signaling to identify yeast genes controlling genome stability. *G3 (Bethesda)* **5,** 997-1006 (2015).
33. Sabouri, N., Viberg, J., Goyal, D. K., Johansson, E. & Chabes, A. Evidence for lesion bypass by yeast replicative DNA polymerases during DNA damage. *Nucleic acids research* **36,** 5660-5667 (2008).
34. Huang, M. & Elledge, S. J. Identification of RNR4, encoding a second essential small subunit of ribonucleotide reductase in Saccharomyces cerevisiae. *Molecular and cellular biology* **17,** 6105-6113 (1997).

35. Pereira, C., Pereira, A. T., Osório, H., Moradas-Ferreira, P. & Costa, V. Sit4p-mediated dephosphorylation of Atp2p regulates ATP synthase activity and mitochondrial function. *Biochim. Biophys. Acta* **1859,** 591-601 (2018).
36. Endo, K., Tago, Y.-I., Daigaku, Y. & Yamamoto, K. Error-free RAD52 pathway and error-prone REV3 pathway determines spontaneous mutagenesis in Saccharomyces cerevisiae. *Genes Genet. Syst.* **82,** 35-42 (2007).
37. Liefshitz, B., Parket, A., Maya, R. & Kupiec, M. The role of DNA repair genes in recombination between repeated sequences in yeast. *Genetics* **140,** 1199-1211 (1995).
38. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nature communications* **9,** 1744 (2018).
39. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149,** 979-993 (2012).
40. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science (New York, N.Y.)* **353,** aaf1420-aaf1420 (2016).

Data availability: The primary sequencing data have been deposited in the European Nucleotide Archive in the study ERP109205 under the accession numbers detailed in Supplementary Table 1. Gene-specific information is available interactively at the address http://sgv.gurdon.cam.ac.uk.

Code availability. The custom code used for the analysis of primary sequencing data is available at the address https://github.com/fabiopuddu/augur-fermentorum (v0.5) and it relies on Slurm workload manager (version 15.08.13); Samtools Version: 1.3.1 (using htslib 1.3.1); VCFtools (v0.1.13); Bcftools Version: 1.3.1 (using htslib 1.3.1); BWA Version: 0.7.12-r1039; Python 2.7.12; Perl (v5.22.1); Gnuplot (Version 5.0 patchlevel 3). Code for secondary analyses is available on request.

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Warringer for providing the YKOC and discussing results; E. Alonso-Perez for help with YKOC management; the Cancer Genome Project and DNA Sequencing pipelines at the Wellcome Sanger Institute for help with sample submission, tracking, library preparation and sequencing; C. Bradshaw for help with HPC cluster and nanopore sequencing; R. Wellinger for providing mutant strains; V. Mustonen, I. Vazquez-Garcia, M. Stratton, D. Adams, S. Nik-Zainal, and members of the SPJ laboratory for advice. This research was supported by: Wellcome Strategic Award 101126/Z/13/Z (COMSIG); Wellcome Investigator Award 206388/Z/17/Z; Wellcome PhD Fellowship 098051 to MH; Cancer Research UK Programme Grant C6/A18796; Cancer Research UK C6946/A24843 and Wellcome WT203144 Institute Core Funding; and the National Institutes of Health R35 GM118172 to RL.

Author contributions The project was conceived by F.P. and S.P.J.; YKOC duplication was carried out by I.S. and F.P.; single colony isolation and DNA extractions were carried out by A.S. and S.W.; data analyses were carried out by F.P., M.H. and I.A.; logistic regression and telomere length analysis were carried out by R.S., R.M., S.K. and M.K.; CIN measurements were performed by J.Z. and M.G. under R.L.'s supervision; validation experiments were designed by F.P. and carried out by F.P. and G.M.-Z.; the manuscript was written by F.P. and S.P.J., with contributions by G.M.-Z., M.H., M.K., I.S., J.Z., I.A., and R.L .

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial or non-financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.P. (f.puddu@gurdon.cam.ac.uk) or S.P.J. (s.jackson@gurdon.cam.ac.uk).

Legends to figures

Figure 1. Assessment of repetitive-DNA alterations in the YKOC. (**a**) Screen schematics. (**b**) rDNA copy-number distribution for YKOC strains (details in Extended Data Fig 1c legend). (**c**) Extract from (b) showing that gene-knockouts affecting functional rRNA synthesis display increased rDNA copy number (median from n=2-8 biologically independent samples). (**d**) Overlaps between genes affecting rDNA length identified in this study and the literature. Asterisks indicate that some genes, for which we have no data, were removed from the "hits" identified in that study. (**e**) rDNA copy-number estimates of strains carrying *ts* alleles of *RRN3* or *RPA190* grown at permissive or semi-permissive temperatures (the average of n=2 biologically independent samples taken on different days of growth is shown). (**f**) Telomere length distribution for YKOC strains (details in Extended Data Fig 1c legend). (**g-h**) Overlap between gene KOs associated with telomere lengthening or shortening identified in this work and the literature; EST KO strains were previously identified as having shorter telomeres. (**i**) Validation of 14 predicted novel *TLM* genes.

Figure 2. Identification of knockout strains with aberrant karyotypes. (a) Distribution of total deviation from expected ploidy (2n) for YKOC strains in chromosome units (details in Extended Data Fig 1c legend). (**b**) CIN estimates for 106 fresh KO strains selected from those with highest deviation from diploidy (n=4 biologically independent samples for most strains; red dot: average; green band: wild-type sample SD; blue column: KO sample SD). **(c)** Distribution of chromosome rearrangements (CRs) detected in the YKOC (details in Extended Data Fig 1c legend). **(d)** Localisation of CR breakpoints in strains analysed.

Figure 3. mtDNA copy number and its links to genome instability and elevated RNR expression. (a) Distribution of mtDNA copy number for YKOC strains (details in Extended Data Fig 1c legend). **(b)** Correlation between pairs of abnormal genomic parameters measured in different colonies. Every circle is a comparison between two parameters; diameter = number of colonies found as abnormal in both parameters. Holm-Bonferroni corrected p-values and fold-change from the corresponding hyper-geometric distribution (n=8843 biologically independent samples). **(c)** Frequency of aneuploidy types in all YKOC and in *rho^0* strains; hyper-geometric p-values for under/over-enrichment in *rho^0* strains (n^{all}=8843 or n^{rho0}=303 biologically independent samples). **(d)** Distribution of mtDNA copy number in strains with increasing deviation from diploidy. **(e)** DNA-damage induction leads to increased mtDNA copy number (mean from n=2 biologically independent samples). **(f)** Strains lacking transcriptional repression of RNR genes have increased mtDNA (mean from n^{wt}=8 or n^{KO}=2 biologically independent samples; p = two-tailed unpaired t-test). **(g)** *RNR1* or *RNR3* overexpression increases mtDNA levels.

Figure 4. HR defects yield a SNV signature.

(a) Number of "independent" mutations in the YKOC by gene: grey = genes with short repetitive regions; yellow = genes carrying "founder" mutations; green = other frequently-mutated genes. (**b**) Enrichment of *ATP1-3* and *SIT4* mutations in *rho^0* strains. (**c-d**) Average number of SNVs/INDELs in YKOC strains versus the wild-type (BY4743) and between different colonies of the same KO strain. **(e)** Clustering of hypermutator KO strains based on their SNV and INDEL patterns and schematic of potential underlying mutagenic processes.

Methods (online-only)

Duplication of the YKOC and colony isolation. The homozygous diploid YKOC was acquired from EUROSCARF in 2001, propagated for two batch expansions (10-20 cell generations) in liquid CSM (complete synthetic medium; 0.14% YNB, 0.5% ammonium sulphate, 0.077% complete supplement mixture [ForMedium], 2% (w/v) glucose and pH buffered to pH 5.8 with 1% (w/v) succinic acid) before being stored in 20% glycerol at –80°C for 13 years. A further duplication was made by thawing one copy of the collection and pinning it into new 96-well plates with CSM using a RoToR HDA robot (Singer Instruments Ltd). Inoculated plates were incubated at 30°C for three days and glycerol was added to a 25% final concentration before freezing at -80°C. From each well, single colonies were isolated on YPAD medium and grown for 2–8 days at 30°C. Two colonies per knockout were inoculated in 1.8ml YPAD and grown to saturation for 2–5 days, before collecting by centrifugation. Other strains and plasmids used in this work are described in Supplementary Table 5.

Genomic DNA extractions, library preparations and sequencing. Genomic DNA extractions and library preparations were carried out as previously described[41]. DNA from up to 96 colonies was multiplexed in each sequencing lane. Sequencing of the YKOC was carried out on Illumina HiSeqX machines. All other sequencing was carried out on Illumina HiSeqX or HiSeq2500 platforms.

Sequencing read alignment. Sequencing reads were aligned to the yeast reference genome (S288c_R64-1-1_Ensembl) using *bwa mem* with the following options: *-t 16 -p -T 0*; duplicates were marked with *bamstreamingmarkduplicates* (biobambam2 2.0.50).

Gene knockout validation. For each sequenced DNA sample (corresponding to one isolated colony), sequencing coverage across the ORF expected to be deleted in that sample was computed. An ORF coverage <15% of the relevant chromosome-wide median (CWM) was deemed to indicate full gene deletion. We noticed that some knockout strains (made by different laboratories) were not deleted for the entire ORF, but rather of a region of varying dimensions within the ORF itself, usually disrupting the start codon and likely producing ORF inactivation. To account for this, ORFs failing the previous criteria were evaluated for the presence of at least 9 consecutive bases with coverage <5% of the CWM. Samples matching these criteria are indicated as "partial deletions". In each sample, UP and DN barcodes were read from reads matching, in normal and in reverse complement directions, the patterns /U1[ACGT]+U2/ and /D1[ACGT]+D2/ respectively. U1, U2, D1, D2 are the relevant unique flanking sequences (U1: GATGTCCACGAGGTCTCT; U2: CGTACGCTGCAGGTCGAC; D1: CGGTGTCGGTCTCGTAG; D2: ATCGATGAATTCGAGCTCG). For each sample the frequency of barcodes, observed at least 5 times and for a frequency > 20%, is reported in Supplementary Table 6, and noted if it does not match with the expected.

Gene knockout re-assignment. Any sample not carrying the expected deletion was reassigned to the correct knockout strain using information encoded in the unique barcodes, when possible. Samples re-assigned in this way were subsequently validated again for the presence of the deletion as described in the previous paragraph before any further work.

Development and validation of tools to study repetitive DNA instability. To study repetitive regions, we developed tools to measure copy-number variation from sequencing data, based on the concept that the number of sequencing reads generated by a certain locus is proportional to its copy number. This produces an apparent coverage increase which can be used for copy number estimation (**Extended Data Fig. 2a**). A strong correlation between rDNA copy-number measures derived from WGS and from gel electrophoresis indicates that WGS-based methods are at least as accurate as traditional methods, and re-sequencing the same DNA preparations revealed highly consistent measures (**Extended**

Data Fig. 2b-c). This approach was employed to determine copy numbers of the repetitive *CUP1* locus, the five yeast transposon (Ty1–Ty5) and the 2µ plasmid **(Extended Data Fig. 1a)**. To extract information on telomere length, we measured the abundance of sequencing reads arising from telomeric DNA and demonstrated the ability of this method to detect biologically relevant differences by sequencing strains with normal, short or long telomeres (wild-type, *tel1*Δ and *rif1*Δ) and observing the expected telomere-length differences (**Extended Data Fig. 1d-e**). We determined the variability of these measures in a wild-type context by sequencing single colonies from two widely used genetic backgrounds. In addition to differences in the number of some Ty elements, this analysis revealed that compared to BY4743, W303 strains carry a larger rDNA locus, and a shorter *CUP1* array, consisting of 4 repeats only. This last number was confirmed using long-read sequencing and identifying one read spanning the *CUP1* locus and containing the expected 4 repeats (**Extended Data Fig. 1f-g**).

Quantification of rDNA and *CUP1* repeat copy number. Quantification of rDNA and *CUP1* copy-number was carried out using the S288C genome as a reference. Copy-number estimates were obtained by dividing the median coverage across the appropriate genomic region (rDNA: XII:452000-459000; *CUP1*: VIII:212986-213525) by the genome-wide median coverage and multiplying by 2 (this is because the S288C reference genome contains two copies of the minimal repeat unit).

Oxford Nanopore Technology (ONT) library preparation and sequencing. Yeast genomic DNA was prepared by using standard molecular biology techniques (spheroplasting, SDS lysis, K acetate precipitation, RNAse A digestion) without phenol/chloroform extraction. A sequencing library was prepared from purified gDNA with the rapid sequencing kit (SQK-RAD004, ONT) and following the manufacturer's instructions. Sequencing data were acquired with MinKNOW 1.15.6 with live base-calling, and reads were aligned to the yeast reference genome with minimap2 (2.14-r883)[42]. Long reads aligning in the region of interest (VIII:213000-215000) were extracted and graphically aligned to the reference genome using FlexiDot[43].

Quantification of Ty1–Ty5, 2µ and definition of genetic mating type. An artificial reference genome was constructed in which each "chromosome" contained the following genomic regions:

YDRWTy1-5= IV: 1203704-1215621;

YLRWTy2-1= XII: 938192-950150;

YILWTy3-1= IX: 202220-213647;

YHLWTy4-1= VIII: 82539-94761;

YCLWTy5-1= III: 679-7322;

MATa_HMR= III: 292388-295034;

MATalpha_HML= III: 11146- 14849;

2-micron= J01347.1.

Sequencing reads, extracted from bam files, were re-aligned to this reference genome using bwa mem with the following options: *-t 16 -T 0*; duplicates were marked with *bamsormadup* (biobambam2 2.0.87) with the following options: *threads=16 SO=coordinate level=0 verbose=0 fixmate=1 adddupmarksupport=1*; median coverage was calculated across the following regions and normalized by the genome wide median to obtain the copy number estimate (in number of copies per haploid genome; in the case of the 2-micron plasmid these estimates were multiplied by the expected ploidy to obtain the number of copies per cell; see also quantification of mtDNA copy number).

YDRWTy1-5: 4000-6999;

YLRWTy2-1: 4000-6999;

YILWTy3-1: 4000-6999;

YHLWTy4-1: 4000-6999;

YCLWTy5-1: 2000-2999;

MATa_HMR: 1400-2000;

MATalpha_HML: 1700-2700;

2-micron: 2000-4500;

Quantification of telomere length. Telomeric sequences were identified as reads containing a minimum number of instances of the following degenerate repeat unit (regex pattern: /(TG){1,3}TG{2,3}|C{2,3}A(CA){1,3}/[44]). A synthetic measure of telomere length was obtained by counting the number of telomeric reads in both orientations/read pairs and normalizing by the total number of reads. The best minimum required number of repeat units in a read was determined by testing every possibility between 1 and 20 and choosing the one (5 repeats) that maximised the difference between wild-type, *tel1*Δ and *rif1*Δ strains (**Extended Data Figure 1d**).

Southern blot telomere length analysis. A logistic regression model was used to prioritize strains with the longest and shortest telomeres. DNA was extracted and telomeric Southern blots were carried out as previously described[45].

Quantification of chromosome aneuploidy. Chromosome ploidy was obtained by estimating the copy number of the centromere of each chromosome. This was obtained by dividing the coverage across a 10 kb window centred on the centromere by the genome-wide median and multiplying by 2 (assuming a largely diploid samples). Ploidy estimates <0.2 and >0.8 units were rounded to the closest unit (see next paragraph for details about how these thresholds were calculated); all others were maintained as rational numbers. Plots were obtained using Circos[46] and masking all repetitive regions. Settings to reproduce these graphs are available in the software repository.

Determination of thresholds for "normal" ploidy. Analysis of centromere copy number for 255 diploid centromeres from wild type BY4743 strains (16 colonies, 16 centromeres/colony, excluding one triploid chromosome) revealed that virtually all estimates were between the values of 1.8 and 2.2. Statistical analysis indicates that by defining the ploidy of a centromere as altered when this is <1.8 or >2.2 incurs in a ~ 0.07% chance of false positive (type I error, p=0.000704, two tailed z-test). We define a strain as deviating from diploidy when any of the 16 chromosomes deviates from diploidy; this corresponds to a ~1.1% chance of wrongly defining a diploid strain as deviating from diploidy when it is not.

Quantification of chromosome instability. CIN rates were measured by the quantitative chromosome transmission fidelity (qCTF) assays[20]. To generate strains for measurement, deletion cassettes were amplified from the Yeast Knockout MATa Collection with primers annealing 400 bp up- and down-stream of the KanMX4 module, and then transformed into the qCTF strain (RLY8492) with the Frozen-EZ Yeast Transformation II Kit (Zymo; Catalog number T2001). qCTF assays were performed as described previously with minor adjustments: cells were not sonicated and were analyzed on an Attune™ NxT Acoustic Focusing Cytometer with Autosampler (Thermo; Catalog number A24861).

Quantification of chromosomal rearrangements. To determine the number and positions of chromosomal rearrangements (regions with ploidy differing from the ploidy of the chromosome at the centromere), we proceeded in two steps. First, we determined the average ploidy across the chromosome using 400 bp bins and noted regions in which ploidy was largely different from the ploidy of the chromosome (difference >0.4 units) and largely similar to the ploidy of the previous data point (difference <0.8 units). Second, we merged all regions where the end of one abnormal ploidy region was less than 20 kb from the beginning the next one and the ploidy of the two regions was similar (difference <0.8 units; this step is necessary to avoid one single rearrangement being split by a masked region within itself: the largest masked regions in the reference genome correspond to two consecutive transposons (20 kb). Finally, the number of rearrangements was calculated by counting the number of regions with abnormal ploidy (greater than [centromere ploidy + 0.5] or smaller than [centromere ploidy – 0.5]) larger than 20 kb. These data were also used to determine the position of the breakpoints of each rearrangement.

Mapping of breakpoints of chromosomal rearrangements. Observation of random breakpoint positions suggested these might be localised to transposons. To assess this, breakpoint positions were compared with positions of known transposons, LTRs, rDNA and sub-telomeric regions (<15 kb from chromosome end). A breakpoint was assigned one or more of the relevant categories if it mapped within 10 kb of any of these features. The list was simplified by removing redundant categories (i.e. when a breakpoint mapped to both a "transposon" and "LTR" it was just considered as a "transposon" since transposons also have nearby LTRs, but LTRs do not necessarily have a nearby transposon).

Quantification of mtDNA copy number. Mitochondrial DNA copy number was determined using the S288C genome as reference. Copy number estimates were obtained by dividing the median coverage across a region of the *COX1* gene (Mito:14000-20000) by the genome-wide median and multiplying by the expected ploidy of the sample (haploid = 1; diploid = 2; this is because the same median coverage reflects different copy numbers in haploid and diploid samples). *COX1* was chosen primarily because it represents the largest continuous stretch of DNA in which the sequencing coverage is stable (**Extended Data Fig. 5a**). Coverage instability most likely resulted from increased DNA breakage in AT-rich regions of the mitochondrial genome during DNA purification and/or preparation of sequencing libraries (**Extended Data Fig. 5b**). An independent estimate using a much smaller stable region corresponding to part of the *COX3* gene (Mito:79213-80022) yielded similar results (**Extended Data Fig. 5c**). *rho0* colonies were defined as having less than 1 mtDNA copy.

Correction of batch effects. To correct for batch effects caused by different library preparations and sequencing runs, we calculated for every genomic feature (rDNA, *CUP1*, Ty1-5, mtDNA, 2µ or telomeres) and for each sequencing run/lane the median of the copy-number estimates of the ~80-96 strains sequenced in that lane. All the estimates in each lane were then normalized by a factor obtained by dividing the median of that lane by the median of the lane containing, among other strains, the 8 wild-type samples.

rDNA and CUP1 copy number correction. Copy number of *CUP1* and rDNA was corrected for the observed ploidy of chromosome VIII and XII respectively, multiplying the raw copy number value by the ratio between the expected and observed ploidy of the relevant chromosome.

Correlation study between *CUP1* and rDNA. After correction the values for *CUP1* and rDNA were scaled to have a mean = 0 and SD = 1. The correlation (*C*) for each deletion strain (with *n* different samples) was calculated as the mean of the product of *CUP1* (X_i) and rDNA (Y_i) normalized values, with a further penalisation so that:

$$C = \frac{1}{n}\sum_{i=1}^{n} \frac{X_i Y_i}{\left(\frac{e}{2}\right)^{d_i}},$$

where d_i is the Euclidean distance between the vector (X_i, Y_i) and the $|X| = |Y|$ diagonal. This factor introduces an exponential decay on the final correlation value when the *CUP1* and rDNA values are very different from each other, controlling the cases where only one of the two elements shows extreme values.

mtDNA quantification by qPCR and comparison with WGS data. Genomic DNA obtained as described above was analyzed by qPCR using Fast SYBR™ Green Master Mix (Thermo; Catalog number 4385612). Relative mtDNA levels were determined by normalization to genomic DNA. Primers used were:

gDNA (*GAL1*)	Up: TGCTTTGTCAAATGGATCATATGG
	Low: CCTGGAACCAAGTGAACAGTACAA
mtDNA	Up: CACCACTAATTGAAAACCTGTCTG
	Low: GATTTATCGTATGCTCATTTCCAA

For comparison purposes, estimation of mtDNA copy number from WGS data was also carried out using the following genomic regions corresponding to qPCR amplicons:

Mito:25574-25686	for mtDNA
II:280382-280459	for *GAL1* (qPCR control, instead of genomewide median)

DNA damage treatments. Wild-type cells were inoculated in 1.8 ml cultures and grown overnight to saturation in YPD. The following day (day 1), 10 µl of each culture was used to inoculate new mini-cultures in YPD or YPD supplemented with DNA damage inducing drugs (MMS: 0.01%; CPT: 20µM; HU: 100 mM). Remaining cells were harvested and frozen. The same was repeated for the following days. DNA extractions were carried out in parallel at the end of the experiment.

Quantification of RNR expression. Relevant strains were grown to exponential phase. Hot phenol extraction method was used to isolate total RNA. Quantity and purity were analyzed using a NanoDrop 1000. 10µg of total RNA were treated with TURBO DNAse (Invitrogen; Catalog number AM2238). RNA was purified using RNA Clean & Concentrator Kit (Zymo; Catalog number R1016). cDNA was then prepared using Superscript III reverse transcriptase (Invitrogen; Catalog number 18080). The expression level of individual transcripts was determined by qPCR using Fast SYBR™ Green Master Mix (Thermo; Catalog number 4385612). Relative levels were determined by normalization to the *ACT1 mRNA* in each sample. Primers used were:

RNR1	Up: GCTCTATATACCCGCTACGAGAAA
	Low: TCCTTGTAAACAACGAAAGGTGTA
RNR2	Up: TATCCATGACTGGAACAACAGAAT
	Low: GTGGAGAAGTTTTCAACCAAGTTT
RNR3	Up: AGGTCGTGGTAAAACAATTAAAGC
	Low: TGTTGGTTTGTCTTCCTGTTACAT

RNR4	Up: GGTAACTTGTTAGCCTTGTCCATT
	Low: CATAATCTGGAACCCGTAGAAACT
ACT1	Up: GAAATGCAAACCGCTGCTCA
	Low: TACCGGCAGATTCCAAACCC

RNR overexpression experiments. Yeast strains carrying genes encoding for ribonucleotide reductase subunits (*RNR1*, *RNR2*, *RNR3*, *RNR4*) under the control of the *GAL1-10* promoter were grown at 30°C overnight in YPAD in 96-well plate format (deep well, 1.8 ml culture). Cells were then pelleted, resuspended in fresh YPA + 2% raffinose, diluted 1:20 in a new well, and incubated overnight. The following day, cells were diluted 1:20 in a new well in either YPA + 2% raffinose or YPA + 2% raffinose + 2% galactose. The same procedure was repeated on the following days. For each set of wells, samples for DNA extraction and sequencing were collected ~48hrs after inoculation.

Analysis of SNVs and INDELs. Mutation calling was performed against the EF4.69 revision of the yeast genome with samtools mpileup with the following options -g -t DP,DV -C0 -pm3 -F0.2 -d10000 and bcftools call with the following options -vm -f GQ; mutations were initially filtered using vcftools with the following options: -H -f +/d=15/q=25/SnpGap=7 and mutations present in the wild-type BY4743 strains were removed with bedtools intersect using all the wild-type samples as reference. Mutations were subsequently filtered with vcftools excluding variants with genotype quality (GQ) <91; mapping quality (MQ) <30; or depth (DP) <10. Prediction of the effects of variant was carried out using the Ensembl variant effect predictor[47] as previously described [41].

Methods References

41. Herzog, M. *et al.* Detection of functional protein domains by unbiased genome-wide forward genetic screening. *Sci Rep* **8,** 6161 (2018).
42. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34,** 3094‑3100 (2018).
43. Seibt, K. M., Schmidt, T. & Heitkam, T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics (Oxford, England)* **34,** 3575‑3577 (2018).
44. Shampay, J., Szostak, J. W. & Blackburn, E. H. DNA sequences of telomeres maintained in yeast. *Nature* **310,** 154‑157 (1984).
45. Rubinstein, L. *et al.* Telomere length kinetics assay (TELKA) sorts the telomere length maintenance (tlm) mutants into functional groups. *Nucleic acids research* **42,** 6314‑6325 (2014).
46. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19,** 1639‑1645 (2009).
47. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics (Oxford, England)* **26,** 2069‑2070 (2010).
48. Huh, W.-K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686‑691 (2003).

Extended Data

Extended Data Figures

Extended Data Figure 1. Statistics of YKOC analyses and distribution of repetitive DNA estimates across the YKOC. (a) Colonies that did not carry the expected deletion (red) were reassigned by reading the barcode inserted with the deletion marker; deletion of an alternative gene was then confirmed by loss of sequencing coverage. **(b)** Number of strains proceeding through the steps of the pipeline for data generation, and analysis used to create the dataset on which this work is based. **(c)** Distribution of copy-number estimates for the indicated repeats across the YKOC. Strains are sorted by the average across the colonies sequenced, and the estimate of each colony is shown; red zones represent values >3 standard deviations of the wild-type distribution (n=8 biological independent samples for wild-type strains; n=1 biological sample for 258 KO strains; n=2 biologically independent samples for 4093 KO strains; n=3 biologically independent samples for 30 KO strains; n=4 biologically independent samples for 72 KO strains; n=5 biologically independent samples for 1 KO strain). (**d**) Correlations between relative copy-number changes at rDNA and *CUP1* tandem-repeat loci; average correlations in all colonies of each KO strain are shown.(n=8 biological independent samples for wild-type strains; n=1 biological sample for 258 KO strains; n=2 biologically independent samples for 4093 KO strains; n=3 biologically independent samples for 30 KO strains; n=4 biologically independent samples for 72 KO strains; n=5 biologically independent samples for 1 KO strain). **(e)** No overall correlation for rDNA and *CUP1* copy-number estimates across all colonies sequenced. (**f**) Distribution of copy-number estimates for the 2μ plasmid across the YKOC and gene ontology analysis of the hits. 2μ copy numbers did not follow a normal distribution with the maximum standard deviation increasing linearly with the mean copy number; this is consistent with the mode of 2μ amplification, which is activated by expression of the *in cis* gene *FLP1* when the copy number crosses a lower threshold (we estimate this at 20-25 copies); different durations of *FLP1* expression will result in different copy number increases. Also in line with this amplification mechanism, we detected an enrichment in gene-knockouts connected to gene silencing is strains with high 2μ copy number.

Extended Data Figure 2. Development of tools to study repetitive DNA instability. (**a**) Left: schematic of yeast chromosome XII. The apparent increase in sequencing coverage maps to rDNA repeats. Right: similar apparent coverage increases mapped to *CUP1* and Ty transposon loci. (**b**) Whole-genome sequencing (WGS) estimates of rDNA-repeat copy number linearly correlate with estimates obtained by pulsed-field gel electrophoresis (n=3-

4 biologically independent samples per strain). (**c**) Percent deviation of two sequencing technical replicates from their average (n=2 technical replicates derived from n=89 biologically independent samples; median: yellow line; quartiles: blue line). measurements were within 5% of their average. Notable exceptions were Ty5 and *CUP1,* probably due to their relatively low repeat numbers. **(d)** Relative estimated content of telomeric repeats in indicated strains, normalized to estimated content of one wild-type colony, is plotted as a function of the minimum number of telomeric repeats in a sequencing read required to classify that read as telomeric (n=2 biologically independent samples per strain). (**e**) Telomere length estimations for wild-type, *tel1*Δ and *rif1*Δ strains obtained calculating the relative abundance of telomeric reads (mean from n=4-8 biologically independent samples per strain). (**f**) Estimations of rDNA, *CUP1*, Ty1, Ty2, and mtDNA copy numbers, and telomeric DNA content for *MATa*, *MATα* and diploid strains in W303 and BY4743 backgrounds (median from n=8 biologically independent samples per strain). (**g**) A long read spanning the *CUP1* locus derived from ONT (Oxford nanopore) sequencing of a W303 (K699) genomic library. (**h**) Comparison of *CUP1* copy number estimated by qPCR or WGS; the same DNA samples (as indicated in labels) were analyzed. Two estimates were extracted from WGS data: "from *CUP1*" indicates estimation using a large region of the *CUP1* locus and the genome-wide median for reference (the same method used for the entire YKOC); "from qPCR amplicon" indicates a small region of *CUP1* and a small region of *GAL1* for reference (the same regions used for qPCR).

Extended Data Figure 3. Southern blot analysis and functional connections between *TLM* genes. (a) Gel electrophoresis and Southern blot analysis of telomeres for 14 novel predicted *TLML* strains (hits = strains with two or more colonies with measures >3 times the SD of the wild-type distribution) and 21 strains failing such stringent hit-selection criteria (non–hits) but still displaying relatively high or low telomere length estimates (representative images from two independent experiments).. Purple lines: location of molecular weight markers; orange line: average telomere length for wild-type samples; green dashes: average telomere lengths for strains predicted to have longer telomeres; white dashes: average telomere lengths for strains predicted to have shorter telomeres **(b)** Validation of KOs failing *TLM* selection criteria but still displaying high or low telomere counts. **(c)** Network-graph analysis of KOs affecting telomere length highlighting novel genes validated by Southern blotting.

Extended Data Figure 4. Examples of aneuploidies and chromosomal rearrangements in the YKOC. (a) Example of a strain with fractional aneuploidy of chromosome XII, likely reflecting clonal heterogeneity. (**b**) Distribution of fractional and non-fractional aneuploidies per chromosome (n=8843 biologically independent samples). **(c)** Knockout of genes encoding ribosomal protein subunits frequently leads to gain of the chromosome carrying the paralog gene. **(d)** Ploidy plots of chromosome II for two different colonies of the *hta1*Δ, *swi4*Δ, and *spt10*Δ KO strains: *hta1*Δ cells (deleted in one of the two genes encoding histone H2A) accumulate a specific amplification of a genome region containing the paralog *HTA2,* a centromere, and two origins of replication). This is most likely transmitted as a circular genetic element formed by recombination between two adjacent transposon sequences. Only two other YKOC strains were found to carry the same genetic element and these were *spt10*Δ and *swi4*Δ, encoding factors controlling the transcription of cell-cycle regulated genes, including histones.

Extended Data Figure 5. Calculation of mtDNA copy number. (a) Sequencing coverage across the mitochondrial genome of a wild-type haploid (BY4741; accession ERS616991). Shaded areas indicate regions (loosely corresponding to *COX1* and *COX3* genes) used to estimate total mtDNA content. **(b)** mtDNA regions of low sequence coverage correspond to regions with strongly reduced GC content. **(c)** Comparison of mtDNA content estimated by qPCR and by WGS. The same DNA samples (as indicated by labels) were analyzed by qPCR and WGS. Two estimates were extracted from WGS data: "from *COX1*" indicates estimation using a large region of the *COX1* gene and the genome-wide median for reference (the same method used for the entire YKOC); "from qPCR amplicon" indicates a small region of *COX1* and a small region of *GAL1* for reference (the same regions used for qPCR). **(d)** Correlation between estimates of mtDNA content using *COX1* or *COX3* region on all sequenced strains belonging to the YKOC (Pearson R^2=0.7596).

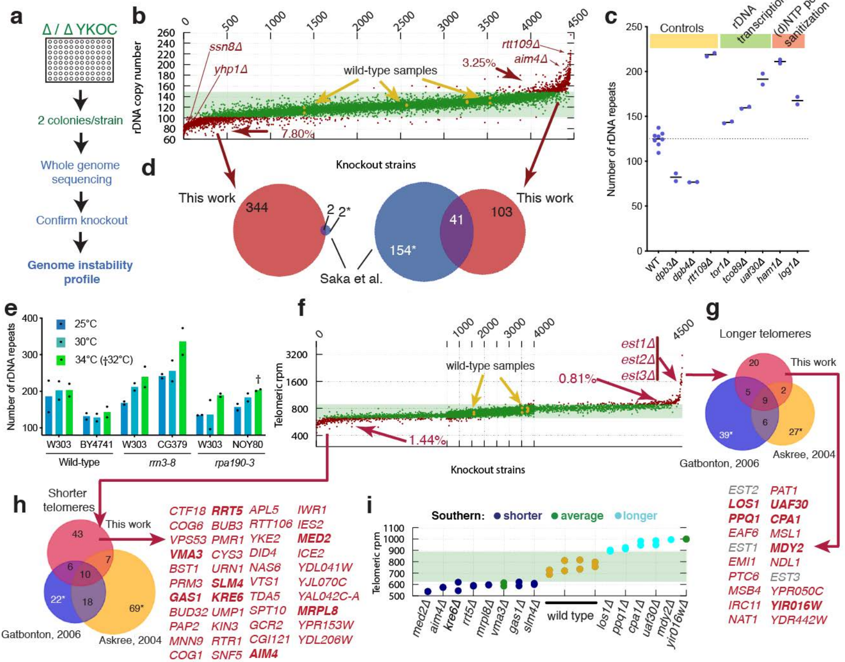
Extended Data Figure 6. Connections between mtDNA and nuclear genome alterations. (a) Venn diagram showing overlap between genes identified as *rho^0* by our sequencing, genes encoding mitochondria proteins (source[48]), and gene knockouts for which respiratory growth was annotated as 'absent' (source SGD: http://www.yeastgenome.org). **(b)** Gene-ontology of *rho^0* strains (estimated mtDNA copy number <1) and *rho^{++}* strains (estimated mtDNA copy number >20.3; Bonferroni corrected p-values). **(c)** Sixteen gene-knockouts from the top end of the mtDNA distribution were assessed for spontaneous DDR activation by Rad53 and histone H2A phosphorylation (representative images from two technical replicates, source data in Supplementary Figure 1), and RNR expression (average from 3 technical replicates, one biological sample per strain). Strains with increased RNR expression (violet) or increased RNR expression and Rad53 hyperphosphorylation (yellow) are highlighted. Serial dilutions of the same cultures were also tested for hydroxyurea (HU) sensitivity. (**d**) Comparisons of mtDNA estimates with systematic analysis of HU sensitivity; HU-sensitive strains are highlighted in different colours depending on the study (Parsons: n=62 and Woolstencroft: n=33 biologically independent samples). (**e**) Comparison of predicted mtDNA copy-number and RNR3 expression levels: KOs with increased Rnr3 protein levels (blue, Z-score >2); KOs with increased mtDNA (yellow, mtDNA >22.2); KOs with both measures increased (green); n=4436 by KO averages of n=8843 biologically independent samples.

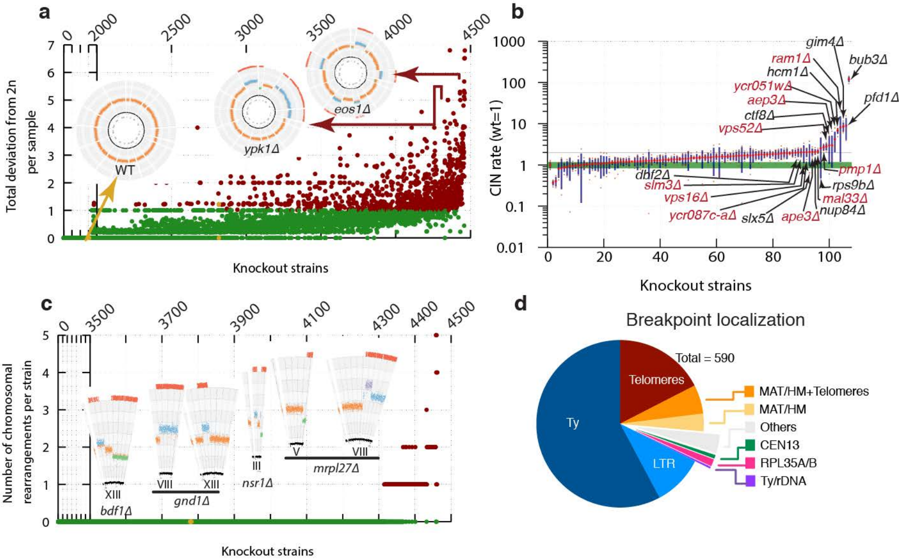
Extended Data Figure 7. mtDNA in KOs for genes encoding tryptophan metabolism enzymes. Pathway for tryptophan biosynthesis from phosphoenolpyruvate, tryptophan import, and NAD biosynthesis from tryptophan are depicted along with mtDNA copy number estimates for strains lacking each of the enzymes in the pathways (mean from $n^{wt} = 8$ or $n^{KO} = 2$ biologically independent samples).

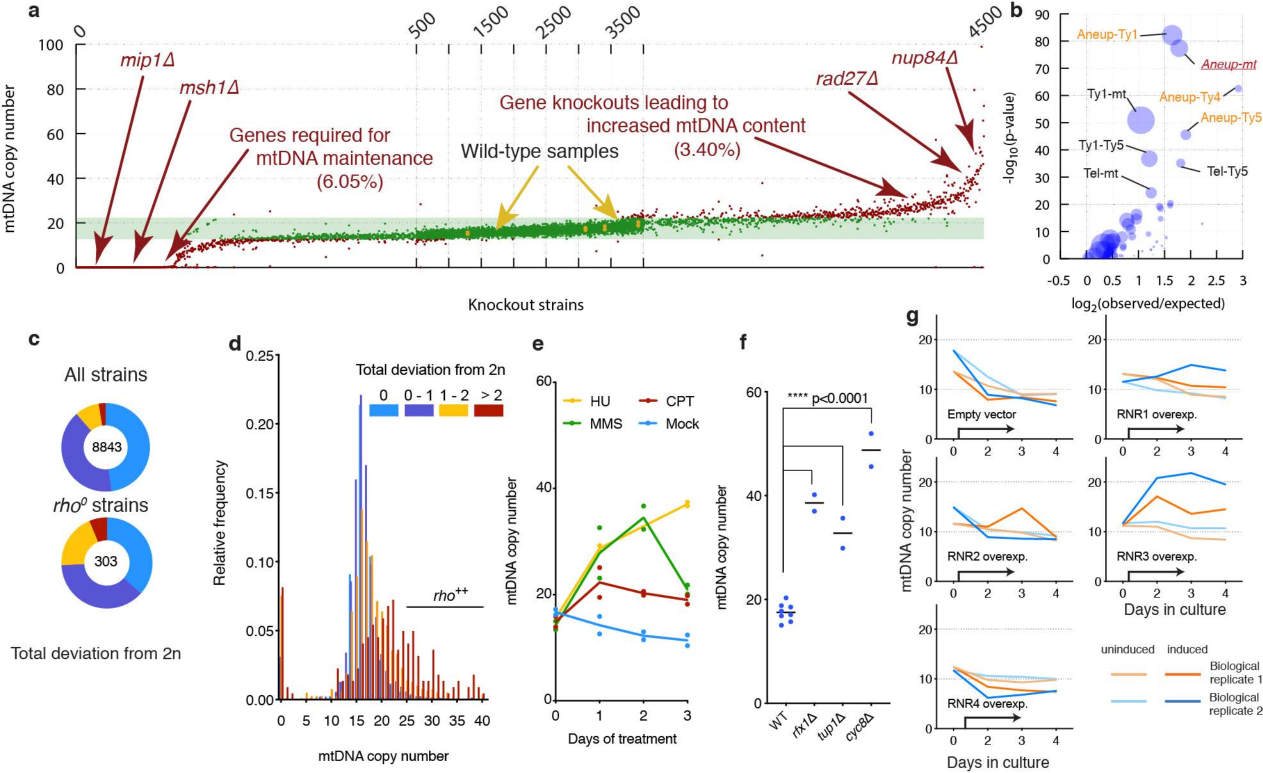
Extended Data Figure 8. Genes frequently carrying mutations contain repetitive regions. Self dot-plots highlighting degenerate repetitive regions in the DNA sequence of genes found to be frequently mutated in the YKOC. Plots were obtained using FlexiDot.

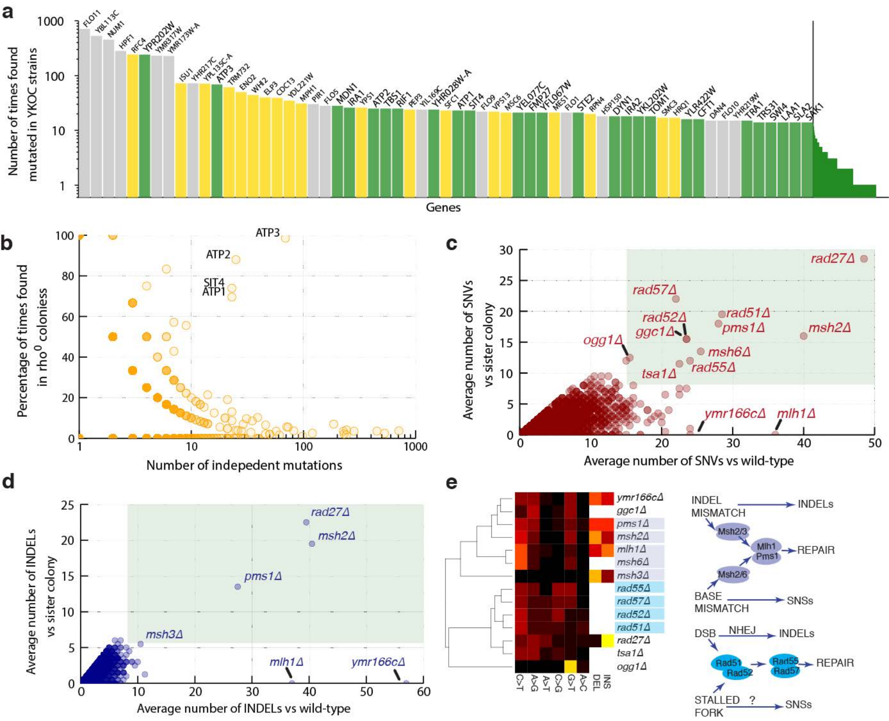
Extended Data Figure 9. Most frequent YKOC mutations and their distributions between different source laboratories. Most frequent mutations, with predicted effects on genes, detected in the YKOC (top 200) and their distribution among different source laboratories. **(a)** Left: the mutation is indicated by its predicted effect, and the background indicates whether it is a mutation in a gene with degenerate repeats (grey), a mutation coming from founder effect (yellow), or a frequently mutated site (green); boldface indicates homozygous mutation. Centre: heatmap of the distribution of the most frequent mutations by laboratory in which the strain carrying that mutation was produced (100% indicates that all the strains with a certain mutation were generated in the same laboratory). Right: number of strains carrying the mutation. **(b)** Not all strains derived from each laboratory share founder mutations: as in (a), but a value of 100% in the heatmap indicates that all the strains generated by a certain laboratory have that mutation.

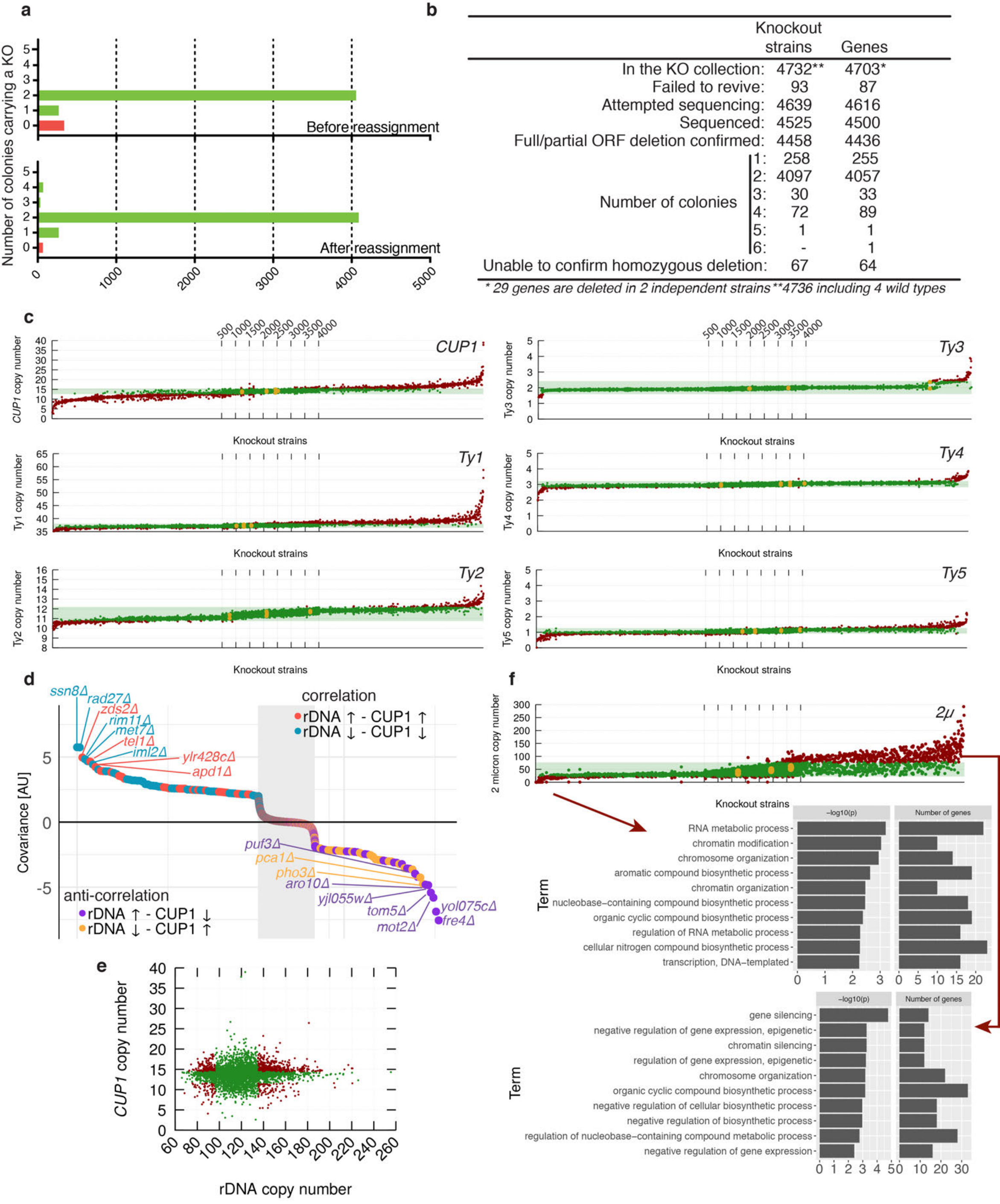
Extended Data Figure 10. Overview of genomic instability caused by non-essential gene knockouts. (a) Overview of results from our genome instability screens. Strains with an abnormal copy number for different genomic features, aneuploidies and chromosomal rearrangements (CR) are represented by coloured boxes. **(b)** Gene ontology analysis for 154 GI genes, defined as KOs showing three or more abnormal features. The number of genes in each GO category as well as Holm-Bonferroni corrected p-values are reported. **(c)** GI genes were manually sorted into classes based on their function, inferred from annotations in SGD.


a
Δ / Δ YKOC
2 colonies/strain
Whole genome sequencing
Confirm knockout
Genome instability profile
b
0 500 1000 1500 2000 2500 3000 3500 4000 4500
rDNA copy number
260 240 220 200 180 160 140 120 100 80 60
ssn8Δ
yhp1Δ
wild-type samples
rtt109Δ
aim4Δ
3.25%
7.80%
Knockout strains
c
250 200 150 100 50 0
Number of rDNA repeats
Controls
rDNA transcription
(d)NTP pool sanitization
WT
dpb3Δ
dpb4Δ
rtt109Δ
tor1Δ
tco89Δ
uaf30Δ
ham1Δ
log1Δ
d
This work
344
2
2*
41
103
154*
This work
Saka et al.
e
400 300 200 100
Number of rDNA repeats
25°C
30°C
34°C (†32°C)
†
W303
BY4741
W303
CG379
W303
NOY80
Wild-type
rrn3-8
rpa190-3
f
0 1000 2000 3000 4000 4500
Telomeric rpm
3200 1600 800 400
est1Δ
est2Δ
est3Δ
wild-type samples
0.81%
1.44%
Knockout strains
g
Longer telomeres
This work
20
5
2
9
6
39*
27*
Gatbonton, 2006
Askree, 2004
EST2
LOS1
PPQ1
EAF6
EST1
EMI1
PTC6
MSB4
IRC11
NAT1
PAT1
UAF30
CPA1
MSL1
MDY2
NDL1
EST3
YPR050C
YIR016W
YDR442W
h
Shorter telomeres
This work
43
7
6
10
22*
18
69*
Gatbonton, 2006
Askree, 2004
CTF18
COG6
VPS53
VMA3
BST1
PRM3
GAS1
BUD32
PAP2
MNN9
COG1
RRT5
BUB3
PMR1
CYS3
URN1
SLM4
KRE6
UMP1
KIN3
RTR1
SNF5
APL5
RTT106
YKE2
DID4
NAS6
VTS1
TDA5
SPT10
GCR2
CGI121
AIM4
IWR1
IES2
MED2
ICE2
YDL041W
YJL070C
YAL042C-A
MRPL8
YPR153W
YDL206W
i
Southern: ● shorter ● average ● longer
Telomeric rpm
1100 1000 900 800 700 600 500
med2Δ
aim4Δ
kre6Δ
rrt5Δ
mrpl8Δ
vma3Δ
gas1Δ
slm4Δ
wild type
los1Δ
ppq1Δ
cpa1Δ
uaf30Δ
mdy2Δ
yir016wΔ


a
0
2000
2500
3000
3500
4000
4500
7
6
5
4
3
2
1
0
Total deviation from 2n per sample
WT
ypk1Δ
eos1Δ
Knockout strains
b
1000
100
10
1
0.1
0.01
CIN rate (wt=1)
gim4Δ
ram1Δ
bub3Δ
hcm1Δ
ycr051wΔ
pfd1Δ
aep3Δ
ctf8Δ
vps52Δ
pmp1Δ
dbf2Δ
slm3Δ
rps9bΔ
vps16Δ
mal33Δ
ycr087c-aΔ
slx5Δ
ape3Δ
nup84Δ
0
20
40
60
80
100
Knockout strains
c
0
3500
3700
3900
4100
4300
4400
4500
5
4
3
2
1
0
Number of chromosomal rearrangements per strain
XIII
bdf1Δ
VIII
XIII
gnd1Δ
III
nsr1Δ
V
VIII
mrpl27Δ
Knockout strains
d
Breakpoint localization
Total = 590
Telomeres
Ty
LTR
MAT/HM+Telomeres
MAT/HM
Others
CEN13
RPL35A/B
Ty/rDNA

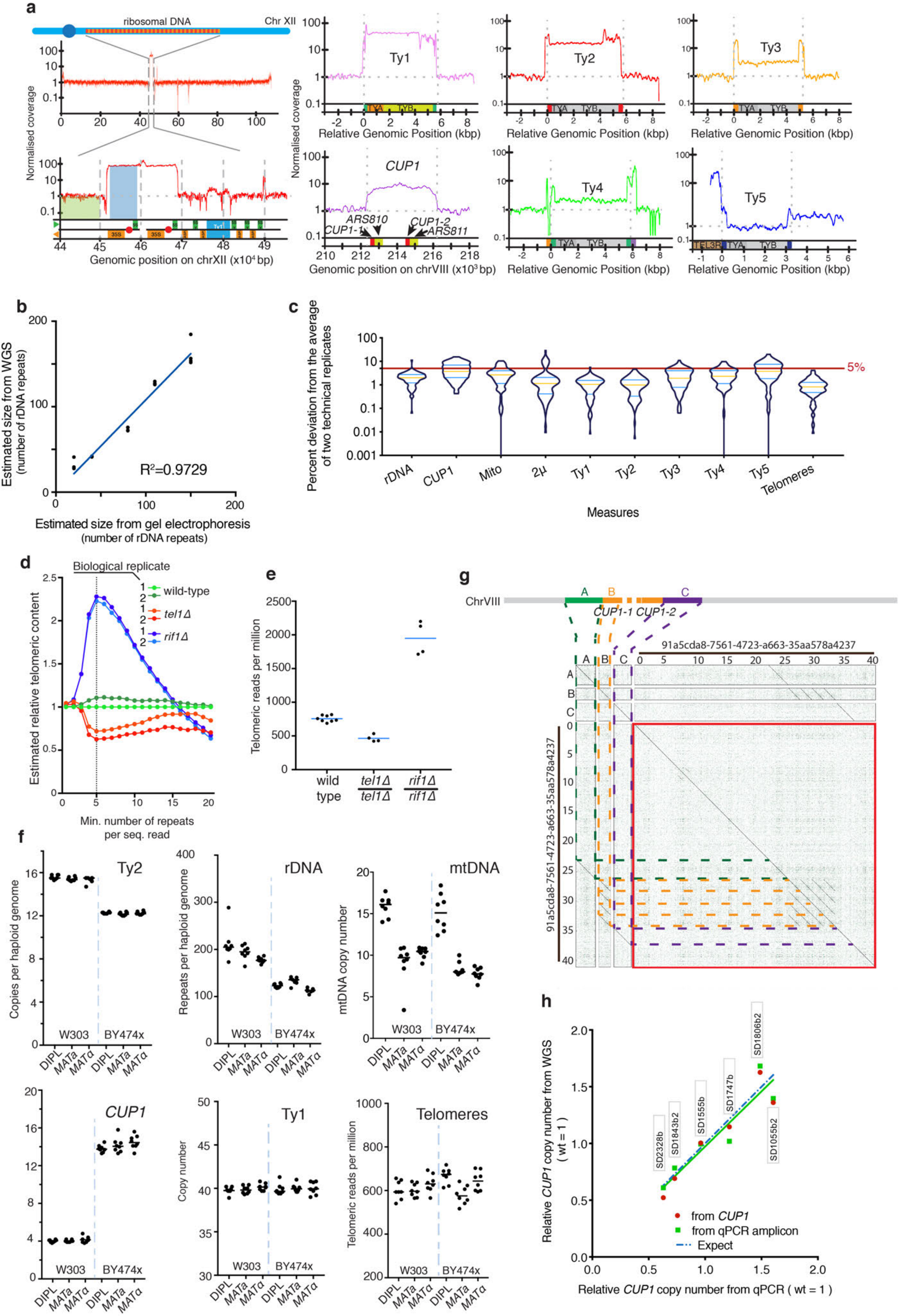

a
0
500
1500
2500
3500
4500
100
80
60
40
20
0
mtDNA copy number
mip1Δ
msh1Δ
Genes required for mtDNA maintenance (6.05%)
Wild-type samples
Gene knockouts leading to increased mtDNA content (3.40%)
rad27Δ
nup84Δ
Knockout strains
b
90
80
70
60
50
40
30
20
10
0
-0.5
0
0.5
1
1.5
2
2.5
3
$-\log_{10}$(p-value)
$\log_2$(observed/expected)
Aneup-Ty1
Aneup-mt
Ty1-mt
Aneup-Ty4
Aneup-Ty5
Ty1-Ty5
Tel-Ty5
Tel-mt
c
All strains
8843
rho^0 strains
303
Total deviation from 2n
d
Relative frequency
0.25
0.20
0.15
0.10
0.05
0.00
Total deviation from 2n
0
0 - 1
1 - 2
> 2
rho^{++}
0
5
10
15
20
25
30
35
40
mtDNA copy number
e
60
40
20
0
HU
CPT
MMS
Mock
mtDNA copy number
0
1
2
3
Days of treatment
f
60
40
20
0
**** p<0.0001
mtDNA copy number
WT
rfx1Δ
tup1Δ
cyc8Δ
g
20
10
0
Empty vector
RNR1 overexp.
RNR2 overexp.
RNR3 overexp.
RNR4 overexp.
0
2
3
4
mtDNA copy number
Days in culture
Days in culture
uninduced
induced
Biological replicate 1
Biological replicate 2

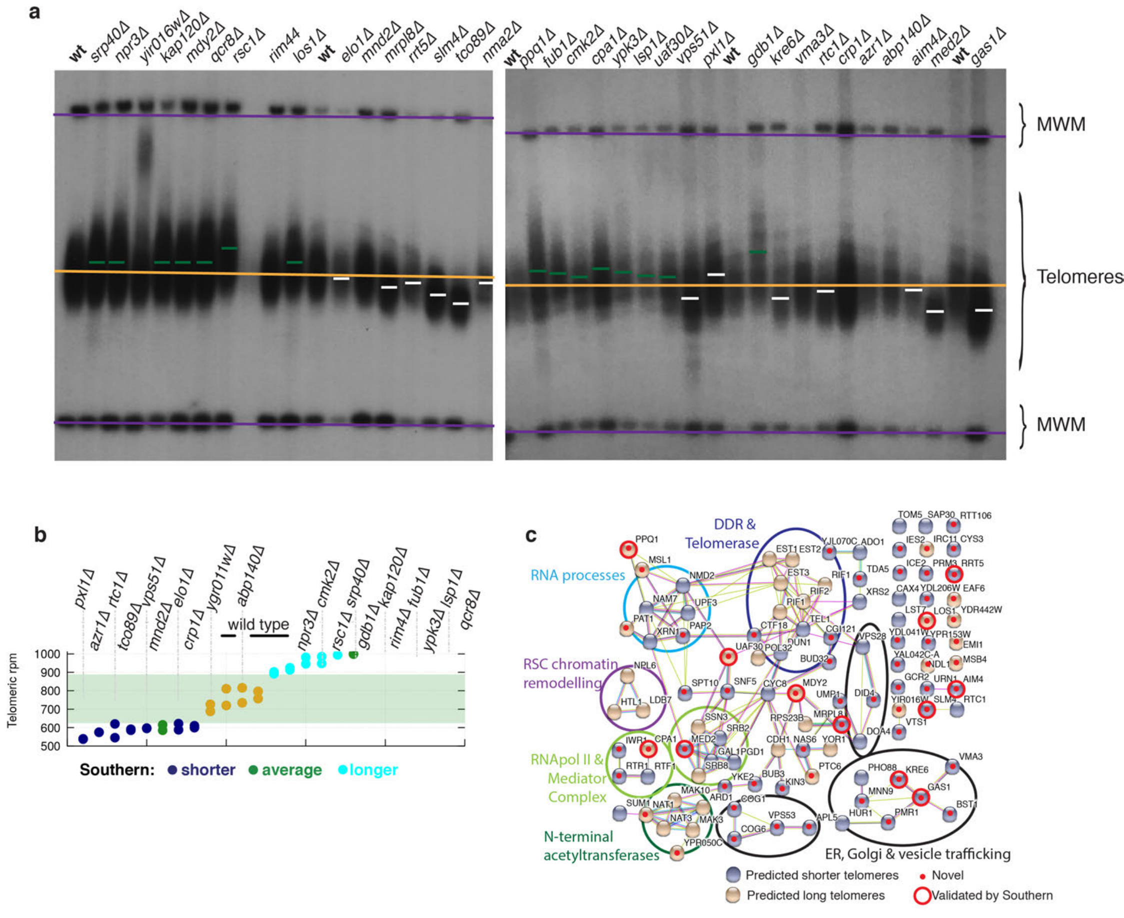

a
Number of times found mutated in YKOC strains
1000
100
10
1
FLO11
YBL113C
NUM1
HPF1
RFC4
YPR202W
YMR317W
YMR173W-A
ISU1
YHR217C
YPL135C-A
ATP3
TRM732
ENO2
WHI2
ELP3
CDC13
YDL221W
MPH1
PIR1
FLO5
MDN1
IRA1
YPS1
ATP2
TBS1
RIF1
PEP3
YIL169C
YHR028W-A
SFC1
ATP1
SIT4
FLO9
VPS13
MSC6
YEL077C
FMP27
YFL067W
MES1
FLO1
STE2
RPN4
HSP150
DYN1
IRA2
YKL202W
TOM1
SMC3
HRQ1
YLR422W
CFT1
DAN4
FLO10
YHR219W
TRA1
TRS31
SWI4
LAA1
SLA2
SAK1
Genes
b
Percentage of times found in rho^0 coloniess
100
80
60
40
20
0
ATP3
ATP2
SIT4
ATP1
1
10
100
1000
Number of indepedent mutations
c
Average number of SNVs vs sister colony
30
25
20
15
10
5
0
rad27Δ
rad57Δ
rad52Δ
ggc1Δ
rad51Δ
pms1Δ
msh2Δ
ogg1Δ
msh6Δ
tsa1Δ
rad55Δ
ymr166cΔ
mlh1Δ
0
10
20
30
40
50
Average number of SNVs vs wild-type
d
Average number of INDELs vs sister colony
25
20
15
10
5
0
rad27Δ
msh2Δ
pms1Δ
msh3Δ
mlh1Δ
ymr166cΔ
0
10
20
30
40
50
60
Average number of INDELs vs wild-type
e
ymr166cΔ
ggc1Δ
pms1Δ
msh2Δ
mlh1Δ
msh6Δ
msh3Δ
rad55Δ
rad57Δ
rad52Δ
rad51Δ
rad27Δ
tsa1Δ
ogg1Δ
C>T
A>G
A>T
C>G
G>T
A>C
DEL
INS
INDEL MISMATCH
INDELs
Msh2/3
Mlh1 Pms1
REPAIR
Msh2/6
BASE MISMATCH
SNSs
DSB
NHEJ
INDELs
Rad51 Rad52
Rad55 Rad57
REPAIR
STALLED FORK
?
SNSs

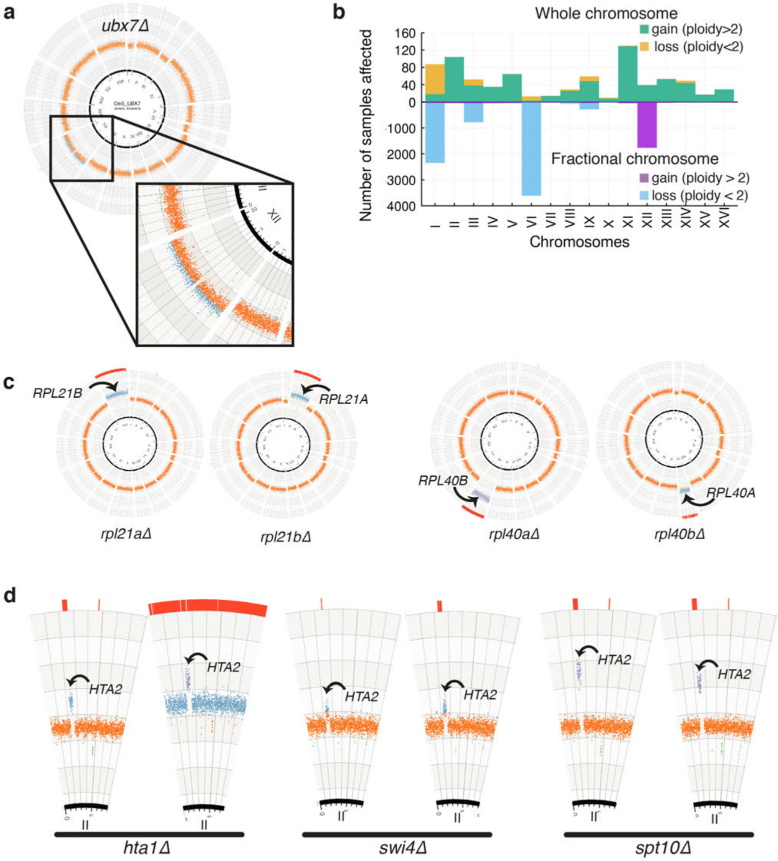


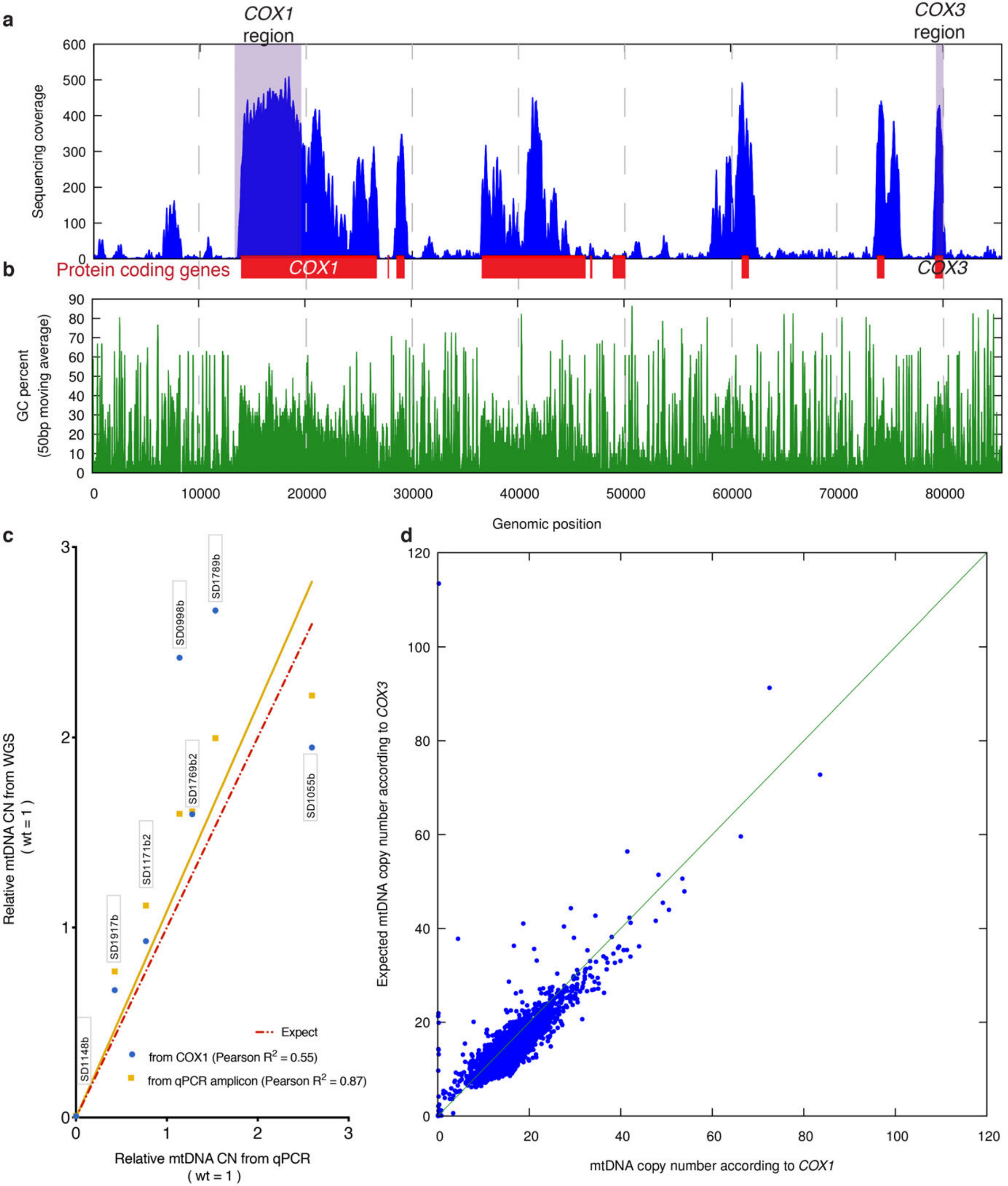
	Knockout strains	Genes
In the KO collection:	4732**	4703*
Failed to revive:	93	87
Attempted sequencing:	4639	4616
Sequenced:	4525	4500
Full/partial ORF deletion confirmed:	4458	4436
Number of colonies 1:	258	255
2:	4097	4057
3:	30	33
4:	72	89
5:	1	1
6:	-	1
Unable to confirm homozygous deletion:	67	64

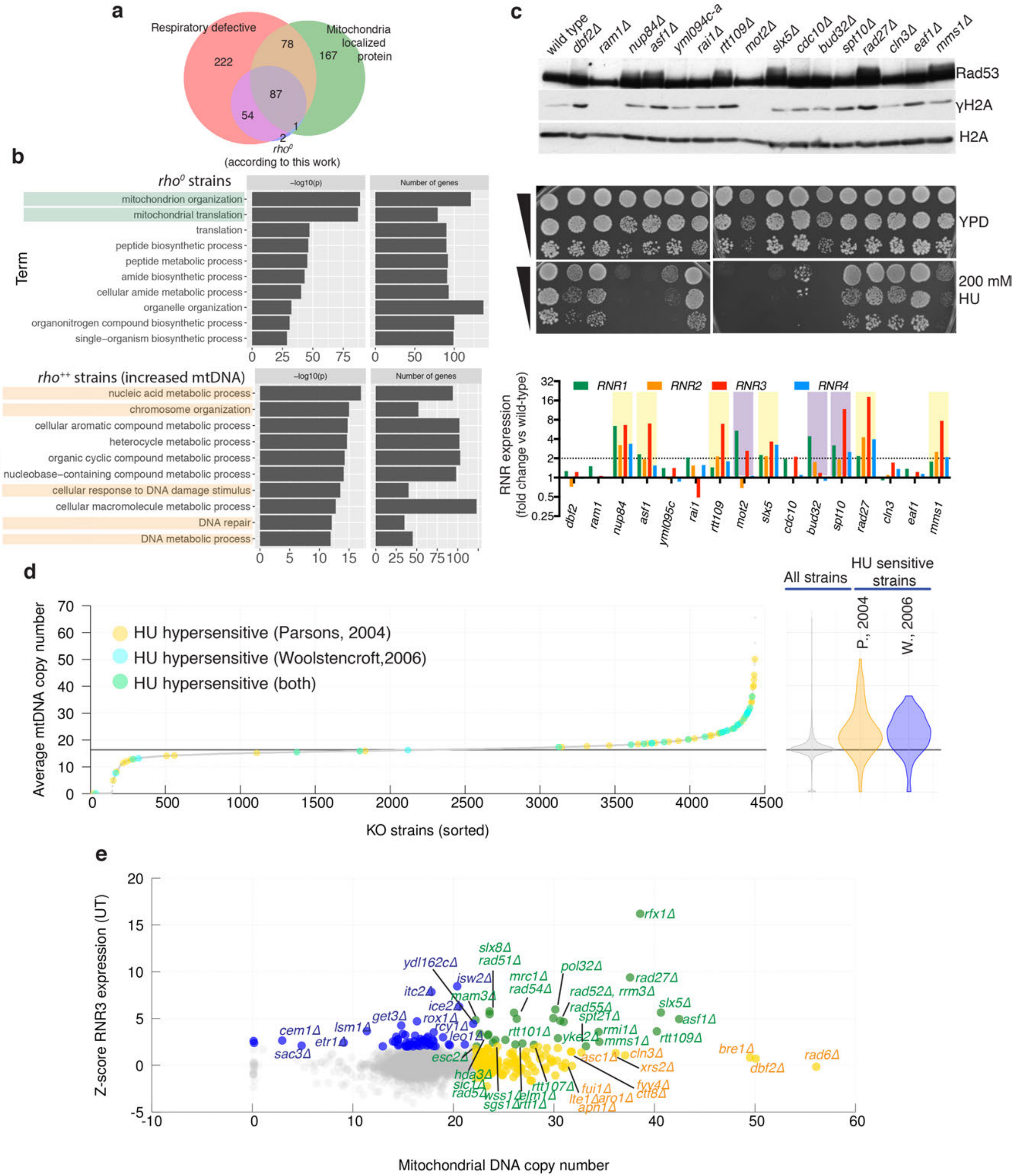
** 29 genes are deleted in 2 independent strains **4736 including 4 wild types*

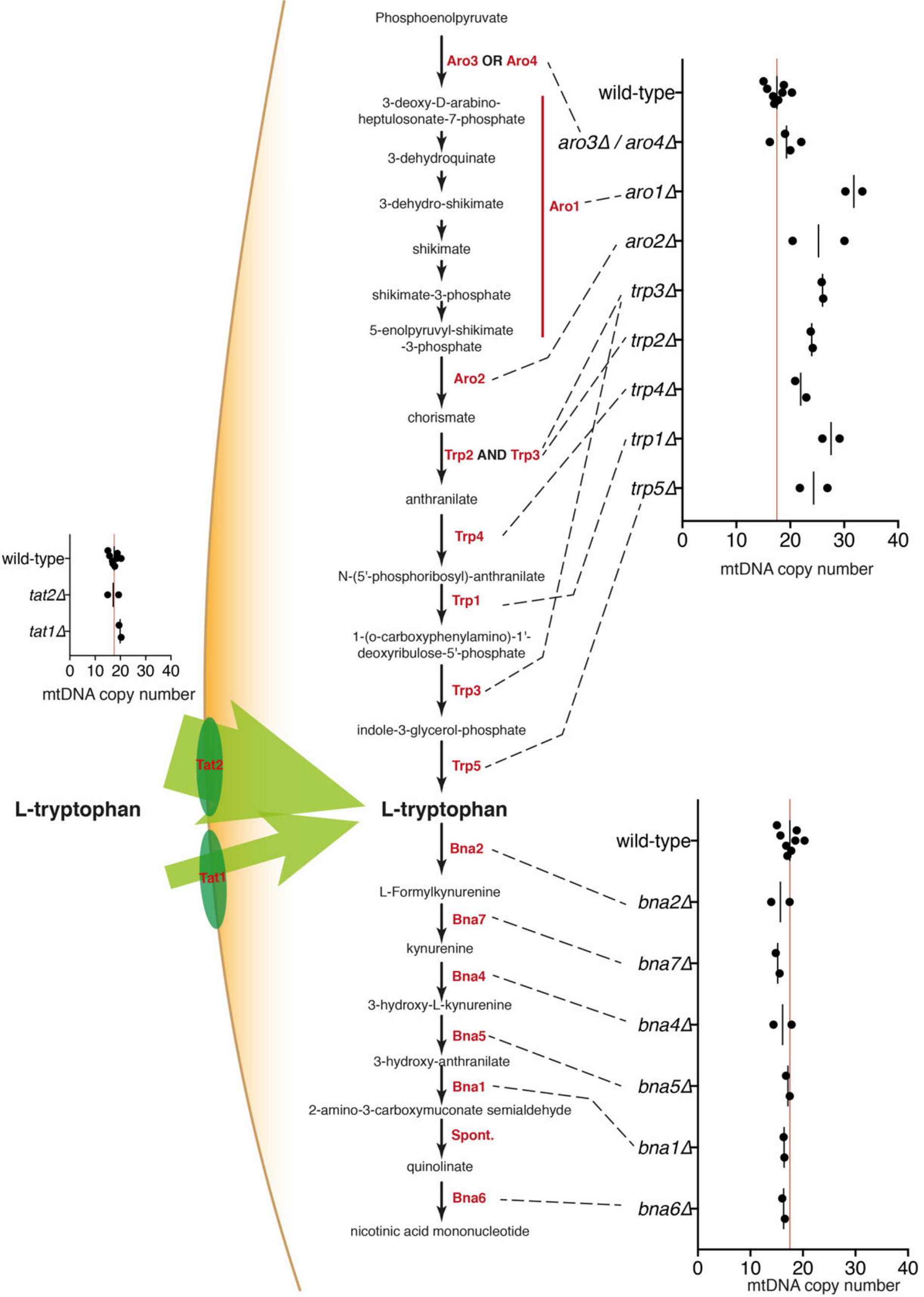

a
ribosomal DNA
Chr XII
Normalised coverage
Genomic position on chrXII (x10⁴ bp)
35S
Ty1
Ty1
Ty2
Ty3
Ty4
Ty5
TYA
TYB
TEL3R
CUP1
ARS810
ARS811
CUP1-1
CUP1-2
Relative Genomic Position (kbp)
Genomic position on chrVIII (x10³ bp)
b
Estimated size from WGS
(number of rDNA repeats)
R²=0.9729
Estimated size from gel electrophoresis
(number of rDNA repeats)
c
Percent deviation from the average
of two technical replicates
5%
rDNA
CUP1
Mito
2µ
Ty1
Ty2
Ty3
Ty4
Ty5
Telomeres
Measures
d
Biological replicate
wild-type
tel1Δ
rif1Δ
Estimated relative telomeric content
Min. number of repeats
per seq. read
e
Telomeric reads per million
wild type
tel1Δ/tel1Δ
rif1Δ/rif1Δ
f
Ty2
Copies per haploid genome
rDNA
Repeats per haploid genome
mtDNA
mtDNA copy number
CUP1
Ty1
Copy number
Telomeres
Telomeric reads per million
W303
BY474x
DIPL
MATa
MATα
g
ChrVIII
A
B
C
CUP1-1
CUP1-2
91a5cda8-7561-4723-a663-35aa578a4237
h
Relative CUP1 copy number from WGS
(wt = 1)
Relative CUP1 copy number from qPCR (wt = 1)
SD2328b
SD1843b2
SD1555b
SD1747b
SD1806b2
SD1055b2
from CUP1
from qPCR amplicon
Expect

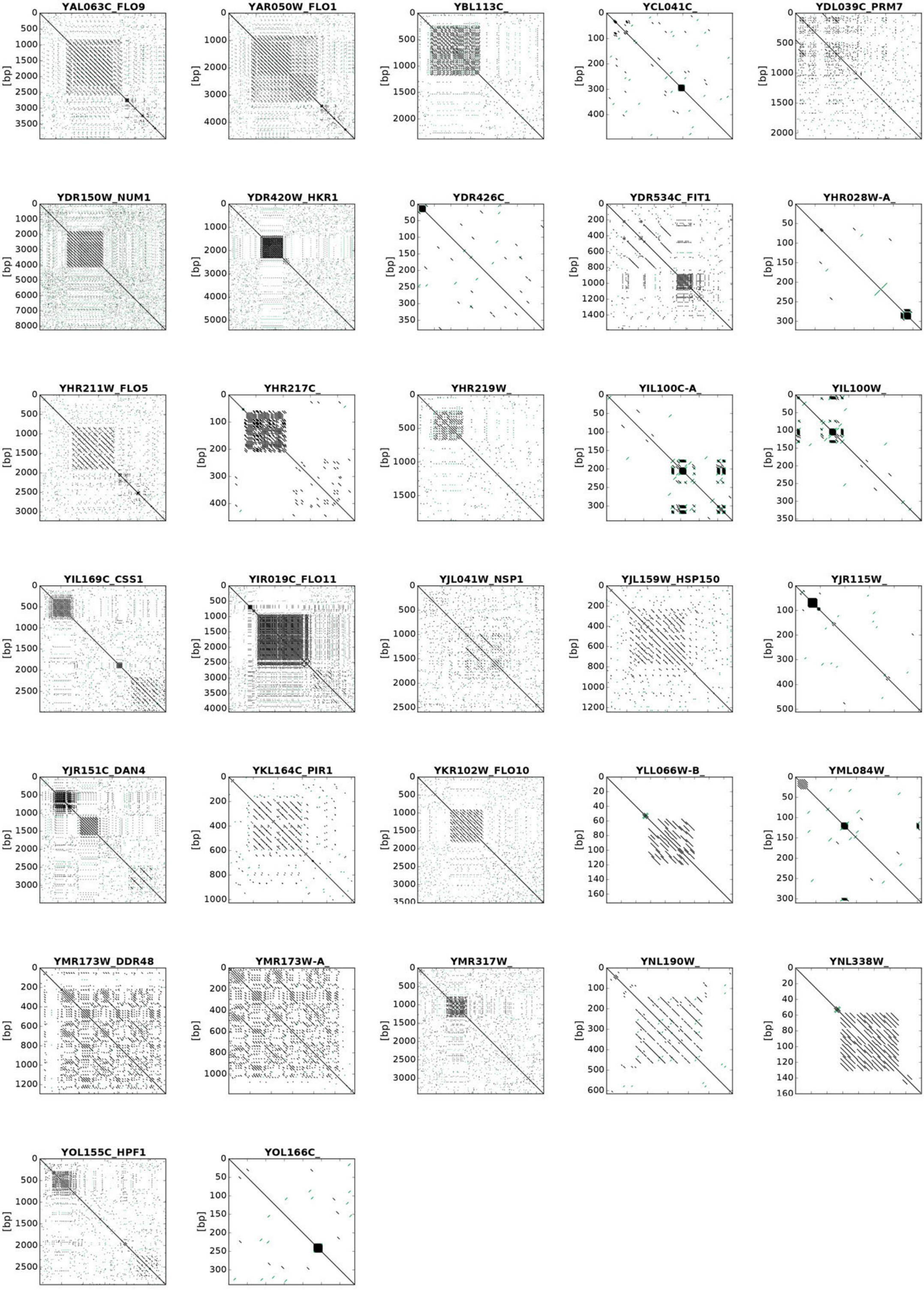

a
wt srp40Δ npr3Δ yir016wΔ kap120Δ mdy2Δ qcr8Δ rsc1Δ rim44 los1Δ wt elo1Δ mnd2Δ mrpl8Δ rrt5Δ slm4Δ tco89Δ nma2Δ
wt ppq1Δ fub1Δ cmk2Δ cpa1Δ ypk3Δ lsp1Δ uaf30Δ vps51Δ pxl1Δ wt gdb1Δ kre6Δ vma3Δ rtc1Δ crp1Δ azr1Δ abp140Δ aim4Δ med2Δ wt gas1Δ
MWM
Telomeres
MWM
b
pxl1Δ azr1Δ rtc1Δ tco89Δ vps51Δ mnd2Δ elo1Δ crp1Δ ygr011wΔ abp140Δ npr3Δ cmk2Δ rsc1Δ srp40Δ gdb1Δ kap120Δ rim4Δ fub1Δ ypk3Δ lsp1Δ qcr8Δ
wild type
Telomeric rpm
1000
900
800
700
600
500
Southern: ● shorter ● average ● longer
c
DDR & Telomerase
RNA processes
RSC chromatin remodelling
RNApol II & Mediator Complex
N-terminal acetyltransferases
ER, Golgi & vesicle trafficking
PPQ1 MSL1 NMD2 NAM7 UPF3 PAT1 XRN1 PAP2
EST1 EST2 EST3 RIF1 RIF2 PIF1 TEL1 CTF18 DUN1 POL32 UAF30 BUD32 CGI121 YJL070C ADO1 TDA5 XRS2
NPL6 HTL1 LDB7
SPT10 SNF5 CYC8 MDY2 UMP1 MRPL8 DID4 VPS28 DOA4
SSN3 SRB2 MED2 GAL1 PGD1 SRB8 RPS23B CDH1 NAS6 YOR1 PTC6 BUB3 KIN3 YKE2
IWR1 CPA1 RTR1 RTF1
MAK10 ARD1 SUM1 NAT1 NAT3 MAK3 YPR050C
COG1 VPS53 COG6 APL5
PHO88 KRE6 VMA3 MNN9 GAS1 HUR1 PMR1 BST1
TOM5 SAP30 RTT106 IES2 IRC11 CYS3 ICE2 PRM3 RRT5 CAX4 YDL206W EAF6 LST7 LOS1 YDR442W YDL041W YPR153W EMI1 YAL042C-A NDL1 MSB4 GCR2 URN1 AIM4 YIR016W SLM4 RTC1 VTS1
Predicted shorter telomeres
Predicted long telomeres
Novel
Validated by Southern

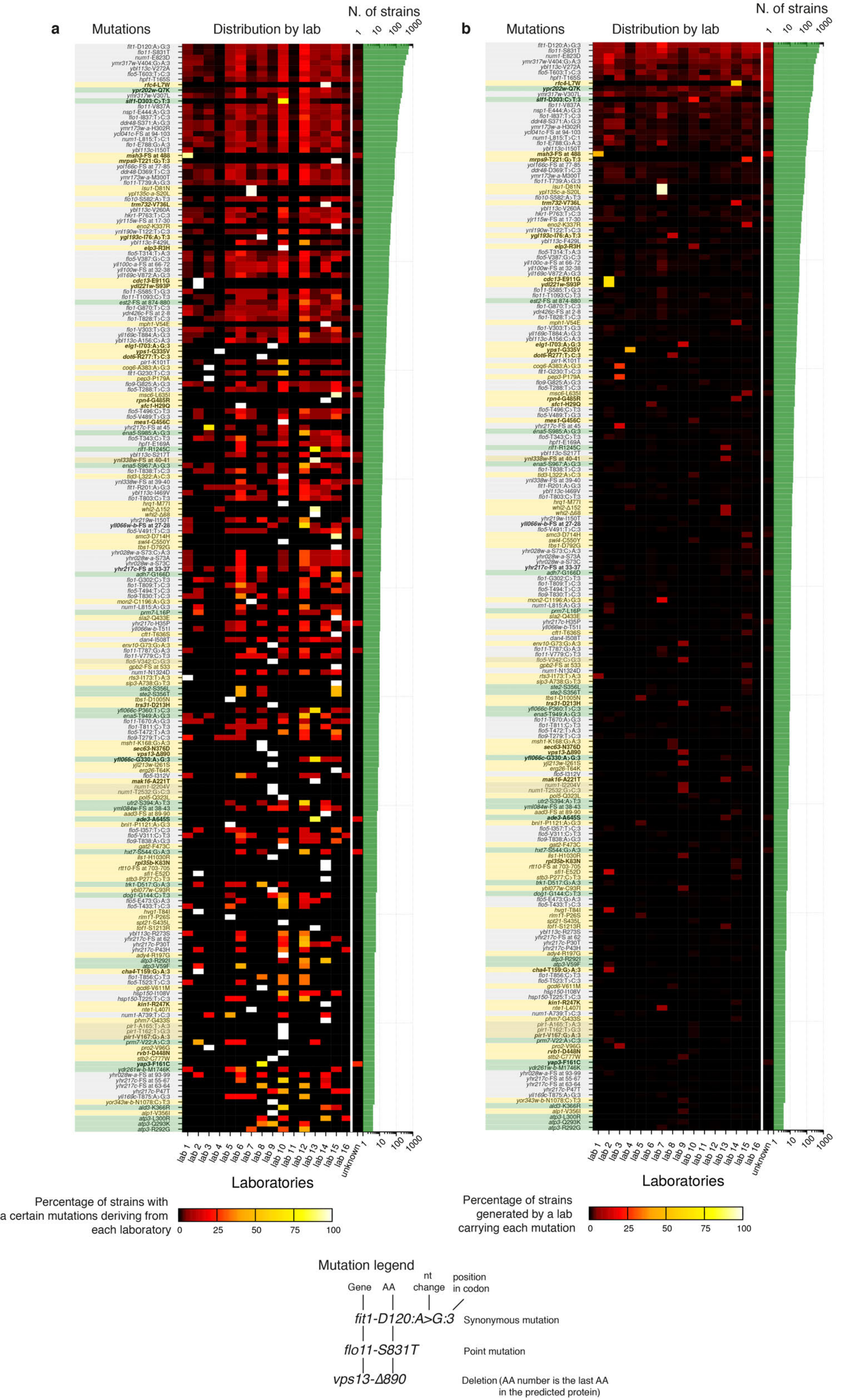

a
ubx7Δ
XII
b
Whole chromosome
gain (ploidy>2)
loss (ploidy<2)
Fractional chromosome
gain (ploidy > 2)
loss (ploidy < 2)
Number of samples affected
160
120
80
40
0
1000
2000
3000
4000
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
Chromosomes
c
RPL21B
RPL21A
RPL40B
RPL40A
rpl21aΔ
rpl21bΔ
rpl40aΔ
rpl40bΔ
d
HTA2
HTA2
HTA2
HTA2
HTA2
HTA2
II
II
II
II
II
II
hta1Δ
swi4Δ
spt10Δ

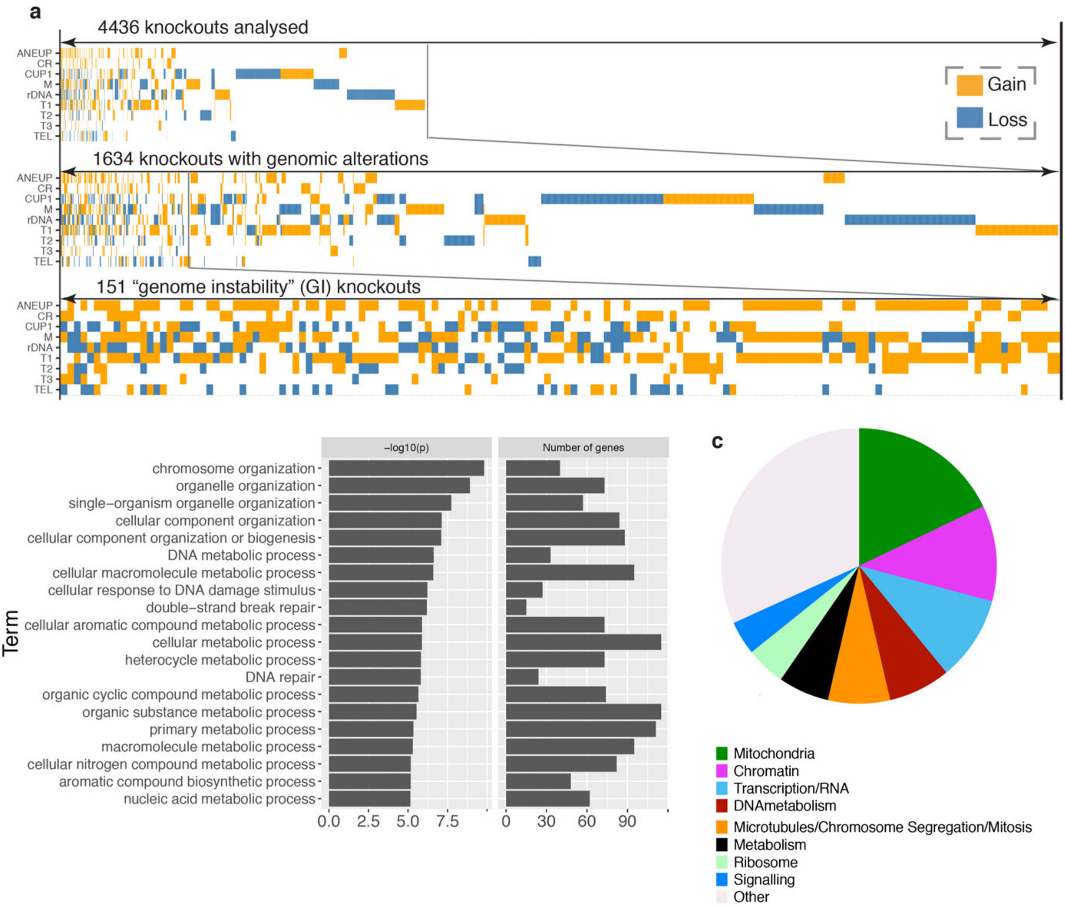

a
COX1 region
COX3 region
Sequencing coverage
600
500
400
300
200
100
0
b
Protein coding genes
COX1
COX3
GC percent (50bp moving average)
90
80
70
60
50
40
30
20
10
0
0
10000
20000
30000
40000
50000
60000
70000
80000
Genomic position
c
Relative mtDNA CN from WGS (wt = 1)
Relative mtDNA CN from qPCR (wt = 1)
0
1
2
3
SD1789b
SD0998b
SD1769b2
SD1055b
SD1171b2
SD1917b
SD1148b
Expect
from COX1 (Pearson R^2 = 0.55)
from qPCR amplicon (Pearson R^2 = 0.87)
d
Expected mtDNA copy number according to COX3
mtDNA copy number according to COX1
0
20
40
60
80
100
120


a
Respiratory defective
Mitochondria localized protein
222
78
167
87
54
1
2
rho⁰ (according to this work)
b
rho⁰ strains
−log10(p)
Number of genes
Term
mitochondrion organization
mitochondrial translation
translation
peptide biosynthetic process
peptide metabolic process
amide biosynthetic process
cellular amide metabolic process
organelle organization
organonitrogen compound biosynthetic process
single−organism biosynthetic process
0 25 50 75
0 50 100
rho⁺⁺ strains (increased mtDNA)
nucleic acid metabolic process
chromosome organization
cellular aromatic compound metabolic process
heterocycle metabolic process
organic cyclic compound metabolic process
nucleobase−containing compound metabolic process
cellular response to DNA damage stimulus
cellular macromolecule metabolic process
DNA repair
DNA metabolic process
0 5 10 15
0 25 50 75 100
c
wild type dbf2Δ ram1Δ nup84Δ asf1Δ yml094c-a rai1Δ rtt109Δ mot2Δ slx5Δ cdc10Δ bud32Δ spt10Δ rad27Δ cln3Δ eaf1Δ mms1Δ
Rad53
γH2A
H2A
YPD
200 mM HU
RNR expression (fold change vs wild-type)
RNR1 RNR2 RNR3 RNR4
32 16 8 4 2 1 0.5 0.25
dbf2 ram1 nup84 asf1 yml095c rai1 rtt109 mot2 slx5 cdc10 bud32 spt10 rad27 cln3 eaf1 mms1
d
Average mtDNA copy number
0 10 20 30 40 50 60 70
HU hypersensitive (Parsons, 2004)
HU hypersensitive (Woolstencroft,2006)
HU hypersensitive (both)
0 500 1000 1500 2000 2500 3000 3500 4000 4500
KO strains (sorted)
All strains
HU sensitive strains
P., 2004
W., 2006
e
Z-score RNR3 expression (UT)
-5 0 5 10 15 20
-10 0 10 20 30 40 50 60
Mitochondrial DNA copy number
rfx1Δ
slx8Δ
rad51Δ
ydl162cΔ
isw2Δ
mrc1Δ
rad54Δ
pol32Δ
rad27Δ
itc2Δ
mam3Δ
rad52Δ, rrm3Δ
ice2Δ
rad55Δ
slx5Δ
get3Δ
rox1Δ
spt21Δ
asf1Δ
lsm1Δ
rcy1Δ
rmi1Δ
cem1Δ
leo1Δ
rtt101Δ
yke2Δ
rtt109Δ
etr1Δ
mms1Δ
sac3Δ
esc2Δ
asc1Δ
cln3Δ
bre1Δ
rad6Δ
xrs2Δ
dbf2Δ
hda3Δ
sic1Δ
rad5Δ
wss1Δ
sgs1Δ
elm1Δ
rtt1Δ
rtt107Δ
fui1Δ
lte1Δ
aro1Δ
apn1Δ
fyv4Δ
ctf8Δ


Phosphoenolpyruvate
Aro3 OR Aro4
3-deoxy-D-arabino-heptulosonate-7-phosphate
3-dehydroquinate
3-dehydro-shikimate
Aro1
shikimate
shikimate-3-phosphate
5-enolpyruvyl-shikimate-3-phosphate
Aro2
chorismate
Trp2 AND Trp3
anthranilate
Trp4
N-(5'-phosphoribosyl)-anthranilate
Trp1
1-(o-carboxyphenylamino)-1'-deoxyribulose-5'-phosphate
Trp3
indole-3-glycerol-phosphate
Trp5
L-tryptophan
L-tryptophan
Tat2
Tat1
Bna2
L-Formylkynurenine
Bna7
kynurenine
Bna4
3-hydroxy-L-kynurenine
Bna5
3-hydroxy-anthranilate
Bna1
2-amino-3-carboxymuconate semialdehyde
Spont.
quinolinate
Bna6
nicotinic acid mononucleotide
wild-type
aro3Δ / aro4Δ
aro1Δ
aro2Δ
trp3Δ
trp2Δ
trp4Δ
trp1Δ
trp5Δ
0
10
20
30
40
mtDNA copy number
wild-type
tat2Δ
tat1Δ
0
10
20
30
40
mtDNA copy number
wild-type
bna2Δ
bna7Δ
bna4Δ
bna5Δ
bna1Δ
bna6Δ
0
10
20
30
40
mtDNA copy number


YAL063C_FLO9
0 500 1000 1500 2000 2500 3000 3500
[bp]
YAR050W_FLO1
0 1000 2000 3000 4000
[bp]
YBL113C_
0 500 1000 1500 2000
[bp]
YCL041C_
0 100 200 300 400
[bp]
YDL039C_PRM7
0 500 1000 1500 2000
[bp]
YDR150W_NUM1
0 1000 2000 3000 4000 5000 6000 7000 8000
[bp]
YDR420W_HKR1
0 1000 2000 3000 4000 5000
[bp]
YDR426C_
0 50 100 150 200 250 300 350
[bp]
YDR534C_FIT1
0 200 400 600 800 1000 1200 1400
[bp]
YHR028W-A_
0 50 100 150 200 250 300
[bp]
YHR211W_FLO5
0 500 1000 1500 2000 2500 3000
[bp]
YHR217C_
0 100 200 300 400
[bp]
YHR219W_
0 500 1000 1500
[bp]
YIL100C-A_
0 50 100 150 200 250 300
[bp]
YIL100W_
0 50 100 150 200 250 300 350
[bp]
YIL169C_CSS1
0 500 1000 1500 2000 2500
[bp]
YIR019C_FLO11
0 500 1000 1500 2000 2500 3000 3500 4000
[bp]
YJL041W_NSP1
0 500 1000 1500 2000 2500
[bp]
YJL159W_HSP150
0 200 400 600 800 1000 1200
[bp]
YJR115W_
0 100 200 300 400 500
[bp]
YJR151C_DAN4
0 500 1000 1500 2000 2500 3000
[bp]
YKL164C_PIR1
0 200 400 600 800 1000
[bp]
YKR102W_FLO10
0 500 1000 1500 2000 2500 3000 3500
[bp]
YLL066W-B_
0 20 40 60 80 100 120 140 160
[bp]
YML084W_
0 50 100 150 200 250 300
[bp]
YMR173W_DDR48
0 200 400 600 800 1000 1200
[bp]
YMR173W-A_
0 200 400 600 800 1000
[bp]
YMR317W_
0 500 1000 1500 2000 2500 3000
[bp]
YNL190W_
0 100 200 300 400 500 600
[bp]
YNL338W_
0 20 40 60 80 100 120 140 160
[bp]
YOL155C_HPF1
0 500 1000 1500 2000 2500
[bp]
YOL166C_
0 50 100 150 200 250 300
[bp]


a
Mutations
Distribution by lab
N. of strains
1
10
100
1000
fit1-D120:A>G:3
flo11-S831T
num1-E823D
ymr317w-V404:G>A:3
ybl113c-V272A
flo5-T603:T>C:3
hpf1-T165S
rfc4-L7W
ypr202w-Q7K
ymr317w-V307L
slf1-D303:C>T:3
flo11-V837A
nsp1-E444:A>G:3
flo1-I837:T>C:3
ddr48-S371:A>G:3
ymr173w-a-H302R
ycl041c-FS at 94-103
num1-L815:T>C:1
flo1-E788:G>A:3
ybl113c-I150T
msh3-FS at 488
mrps9-T221:G>T:3
yol166c-FS at 77-85
ddr48-D369:T>C:3
ymr173w-a-M300T
flo11-T739:A>G:3
isu1-D81N
ypl135c-a-S20L
flo10-S582:A>T:3
trm732-V736L
ybl113c-V260A
hkr1-P763:T>C:3
yjr115w-FS at 17-30
eno2-K337R
ynl190w-T122:T>C:3
ygl193c-I76:A>T:3
ybl113c-F429L
elp3-R3H
flo5-T314:T>A:3
flo5-V387:G>C:3
yil100c-a-FS at 66-72
yil100w-FS at 32-38
yil169c-V872:A>G:3
cdc13-E911G
ydl221w-S93P
flo11-S585:T>G:3
flo11-T1093:C>T:3
est2-FS at 874-880
flo1-G870:T>C:3
ydr426c-FS at 2-8
flo1-T828:T>C:3
mph1-V54E
flo1-V303:T>G:3
yil169c-T884:A>G:3
ybl113c-A156:C>A:3
elg1-I703:A>G:3
yps1-G335V
dot6-R277:T>C:3
pir1-K101T
coq6-A383:A>G:3
fit1-G230:T>C:3
pep3-P179A
flo9-G825:A>G:3
flo5-T288:T>C:3
msc6-L635I
rpn4-G485R
sfc1-H29Q
flo5-T496:C>T:3
flo5-V489:T>G:3
mes1-G456C
yhr217c-FS at 45
ena5-S985:A>G:3
flo5-T343:C>T:3
hpf1-E169A
rif1-R1245C
ybl113c-S217T
ynl338w-FS at 40-41
ena5-S967:A>G:3
flo1-T838:T>C:3
tid3-L322:A>C:3
ynl338w-FS at 39-40
fit1-R201:A>G:3
ybl113c-I469V
flo1-T803:C>T:3
hrq1-M77I
whi2-Δ152
whi2-Δ68
yhr219w-I150T
yll066w-b-FS at 27-28
flo5-V491:T>C:3
smc3-D714H
swi4-C550Y
tbs1-D792G
yhr028w-a-S73:C>A:3
yhr028w-a-S73A
yhr028w-a-S73C
yhr217c-FS at 33-37
adh7-G166D
flo1-G302:C>T:3
flo1-T809:T>C:3
flo5-T494:T>C:3
flo9-T830:T>C:3
mon2-C1196:A>G:3
num1-L815:A>G:3
prm7-L16P
sla2-Q433E
yhr217c-H35P
yll066w-b-T51I
cft1-T636S
dan4-I508T
env10-G73:G>A:3
flo11-T787:G>A:3
flo11-V779:C>T:3
flo5-V342:C>G:3
gpb2-FS at 533
num1-N1324D
rts3-I173:T>A:3
slp3-A738:G>T:3
ste2-S356L
ste2-S356T
tbs1-D1005N
trs31-D213H
yfl066c-P360:T>C:3
ena5-T949:A>G:3
flo11-T670:A>G:3
flo1-T811:C>T:3
flo5-T472:T>A:3
flo9-T279:T>C:3
msh1-K168:G>A:3
sec63-N376D
vps13-Δ890
yfl066c-G330:A>G:3
yjl213w-I261S
erg26-T64K
flo5-I312V
mak16-A221T
num1-I2204V
num1-T2532:G>C:3
pol5-Q323L
utr2-S394:A>T:3
yml084w-FS at 38-43
aad3-FS at 89-90
ade3-A645S
bnl1-P1121:A>G:3
flo5-I357:T>C:3
flo5-V311:C>T:3
flo9-T838:A>G:3
gat2-F473C
hxt7-S544:G>A:3
ils1-H1030R
rpl35b-K83N
rtt10-FS at 703-705
sfi1-E52D
stb3-P277:C>T:3
trk1-D517:G>A:3
ybl077w-C93R
dog1-G144:C>T:3
flo5-E473:G>A:3
flo5-T433:T>C:3
hvg1-T84I
rim11-P26S
spt21-S435L
tof1-S1213R
ybl113c-R273S
yhr217c-FS at 62
yhr217c-P30T
yhr217c-P43H
ady4-R197G
atp3-R292I
atp3-V59F
cha4-T159:G>A:3
flo1-T856:C>T:3
flo5-T523:T>C:3
gcd6-V611M
hsp150-I108V
hsp150-T225:T>C:3
kin1-R247K
nte1-L407I
num1-A739:T>C:3
phm7-G433S
pir1-A165:T>A:3
pir1-T162:T>G:3
pir1-V167:G>A:3
prm7-V22:A>C:3
pro2-V96G
rvb1-D448N
stb2-C777W
yap3-F161C
ydr261w-b-M1746K
yhr028w-a-FS at 93-99
yhr217c-FS at 55-67
yhr217c-FS at 63-64
yhr217c-P47T
yil169c-T875:A>G:3
yor343w-b-N1078:C>T:3
ald3-K366R
alp1-V356I
atp3-L300R
atp3-Q293K
atp3-R292G
lab 1
lab 2
lab 3
lab 4
lab 5
lab 6
lab 7
lab 8
lab 9
lab 10
lab 11
lab 12
lab 13
lab 14
lab 15
lab 16
unknown
1
10
100
1000
Laboratories
Percentage of strains with
a certain mutations deriving from
each laboratory
0
25
50
75
100
b
Mutations
Distribution by lab
N. of strains
1
10
100
1000
lab 1
lab 2
lab 3
lab 4
lab 5
lab 6
lab 7
lab 8
lab 9
lab 10
lab 11
lab 12
lab 13
lab 14
lab 15
lab 16
unknown
1
10
100
1000
Laboratories
Percentage of strains
generated by a lab
carrying each mutation
0
25
50
75
100
Mutation legend
Gene
AA
nt
change
position
in codon
fit1-D120:A>G:3
Synonymous mutation
flo11-S831T
Point mutation
vps13-Δ890
Deletion (AA number is the last AA
in the predicted protein)


a
4436 knockouts analysed
ANEUP
CR
CUP1
M
rDNA
T1
T2
T3
TEL
Gain
Loss
1634 knockouts with genomic alterations
151 "genome instability" (GI) knockouts
−log10(p)
Number of genes
Term
chromosome organization
organelle organization
single−organism organelle organization
cellular component organization
cellular component organization or biogenesis
DNA metabolic process
cellular macromolecule metabolic process
cellular response to DNA damage stimulus
double−strand break repair
cellular aromatic compound metabolic process
cellular metabolic process
heterocycle metabolic process
DNA repair
organic cyclic compound metabolic process
organic substance metabolic process
primary metabolic process
macromolecule metabolic process
cellular nitrogen compound metabolic process
aromatic compound biosynthetic process
nucleic acid metabolic process
0.0
2.5
5.0
7.5
0
30
60
90
c
Mitochondria
Chromatin
Transcription/RNA
DNAmetabolism
Microtubules/Chromosome Segregation/Mitosis
Metabolism
Ribosome
Signalling
Other

Supplementary Information Guide

Supplementary Table 1. Accession numbers of YKOC colonies sequenced.

Supplementary Table 2. Statistical measurements of the wild-type distribution and number of genes/colonies differing from the wild-type in each screen.

Supplementary Table 3. Repetitive DNA estimates for each YKOC colony.

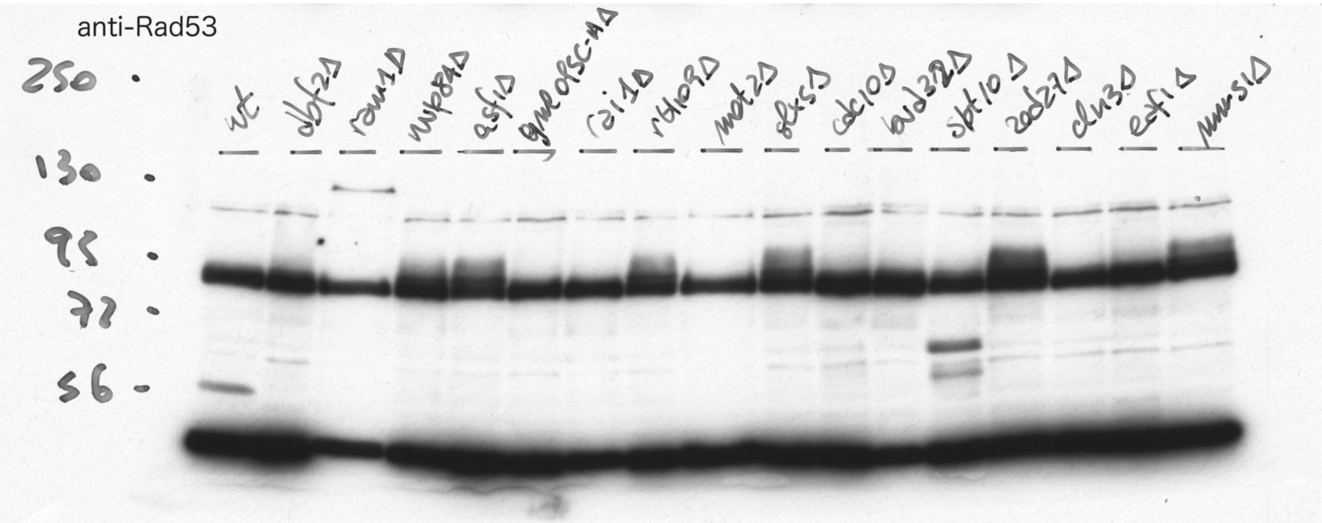
Supplementary Table 4. List of all gene-knockouts analysed highlighting KOs with copy number deviating from wild-type for different elements and associated copy number estimates.

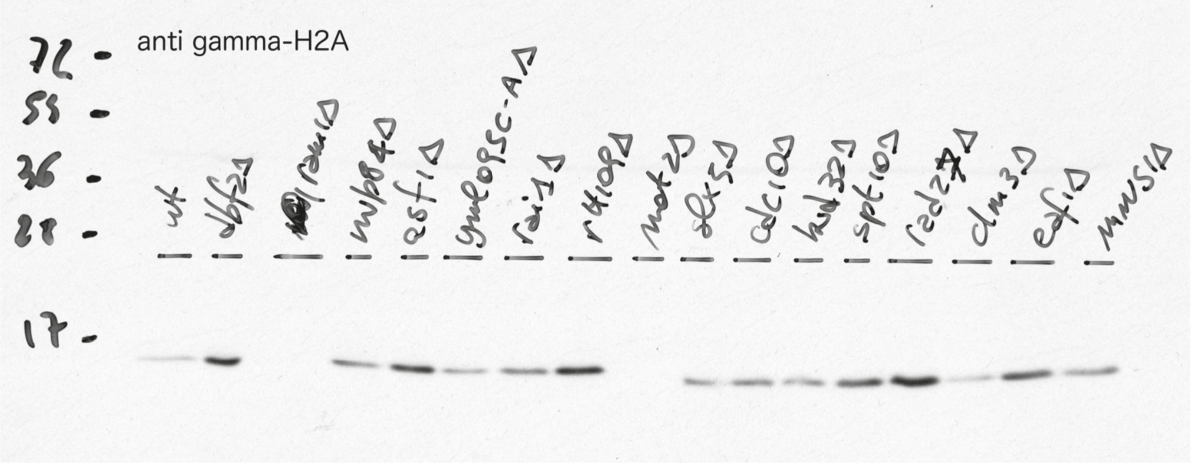
Supplementary Table 5. Other yeast strains used in this study.

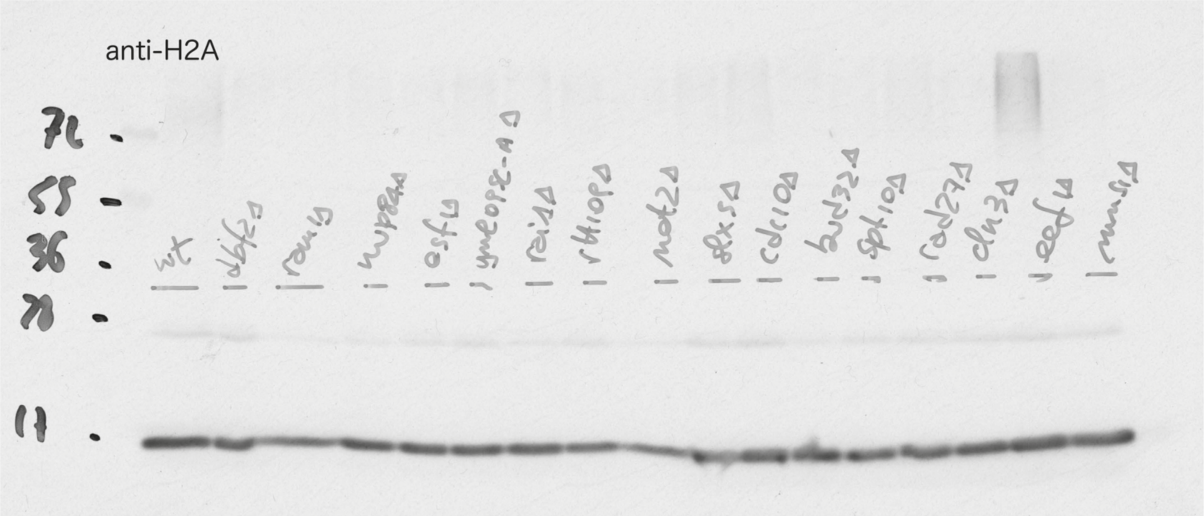
Supplementary Table 6. Deletion barcodes detected in YKOC strains.

Supplementary Figure 1. Spontaneous checkpoint activation in different KO strains.

Source data for western blot in figure ED6C.


anti-Rad53
250
130
95
72
56
wt
dbf2Δ
ram1Δ
nup84Δ
asf1Δ
yml095c-AΔ
rai1Δ
rtt109Δ
mot2Δ
slx5Δ
cdc10Δ
bud32Δ
spt10Δ
rad27Δ
clm3Δ
eaf1Δ
mms1Δ


anti gamma-H2A
72
55
36
28
17
wt
dbf2Δ
ram1Δ
nup84Δ
asf1Δ
yml095c-AΔ
rai1Δ
rtt109Δ
mot2Δ
slx5Δ
cdc10Δ
bud32Δ
spt10Δ
rad27Δ
clm3Δ
eaf1Δ
mms1Δ


anti-H2A
72
55
36
28
17
wt
dbf2Δ
ram1Δ
nup84Δ
asf1Δ
yml095c-AΔ
rai1Δ
rtt109Δ
mot2Δ
slx5Δ
cdc10Δ
bud32Δ
spt10Δ
rad27Δ
clm3Δ
eaf1Δ
mms1Δ