Genome architecture and stability in the Saccharomyces cerevisiae knockout collection

Fabio Puddu1,2,3*, Mareike Herzog1,2,3, Alexandra Selivanova1,2, Siyue Wang1,2, Jin Zhu4, Shir Klein–Lavi5, Molly Gordon4, Roi Meirman5, Gonzalo Millan–Zambrano1,6, Iñigo Ayestaran1,2, Israel Salguero1,2, Roded Sharan7, Rong Li4, Martin Kupiec5 & Stephen P. Jackson1,2,8

Despite major progress in defining the functional roles of genes, a complete understanding of their influences is far from being realized, even in relatively simple organisms. A major milestone in this direction arose via the completion of the yeast Saccharomyces cerevisiae gene-knockout collection (YKOC), which has enabled high-throughput reverse genetics, phenotypic screenings and analyses of synthetic–genetic interactions1–3. Ensuing experimental work has also highlighted some inconsistencies and mistakes in the YKOC, or genome instability events that rebalance the effects of specific knockouts4–5, but a complete overview of these is lacking. The identification and analysis of genes that are required for maintaining genomic stability have traditionally relied on reporter assays and on the study of deletions of individual genes, but whole-genome-sequencing technologies now enable—in principle—the direct observation of genome instability globally and at scale. To exploit this opportunity, we sequenced the whole genomes of nearly all of the 4,732 strains comprising the homozygous diploid YKOC. Here, by extracting information on copy-number variation of tandem and interspersed repetitive DNA elements, we describe—for almost every single non-essential gene—the genomic alterations that are induced by its loss. Analysis of this dataset reveals genes that affect the maintenance of various genomic elements, highlights cross-talks between nuclear and mitochondrial genome stability, and shows how strains have genetically adapted to life in the absence of individual non-essential genes.

To address how gene loss influences genomic structure and stability, we sequenced the whole genomes of the strains that comprise the S. cerevisiae homoyzygous diploid YKOC. After identifying and reassigning targeted genes when they did not correspond to the expected knockout, we confirmed most knockout strains (Extended Data Fig. 1a, b) (all gene-specific information is available in Supplementary Tables 1–4 and at http://sgv.gurdon.cam.ac.uk). We then determined the genome-instability profile of each strain, focusing particularly on repetitive DNA (Fig. 1a).

We developed and validated tools to accurately measure the copy numbers of repetitive DNA on the basis of the concept that when a sequenced genome contains more copies of a locus than does the single-copy-number reference genome, the locus displays proportionately increased sequence coverage (Extended Data Fig. 2, Methods). Global analysis of the repetitive ribosomal DNA (rDNA) locus that encodes large ribosomal RNAs (rRNAs) revealed copy-number alterations in about 10% of strains (Fig. 1b). Gene knockouts that are known to affect the size of the rDNA locus, probably because incorporation of non-canonical nucleotides into RNA impairs its functionality (Fig. 1c). Because gene knockouts that affect the stability of tandem repeats could affect the size of the rDNA locus, we compared the copy numbers of rDNA and the repetitive CUP1 locus (Extended Data Fig. 1c). Strains with high positive correlations included those with knockouts of TEL1, RAD27, SSN8 or ZDS2, but the lack of overall correlation suggests that variation at these loci is driven largely by different selective pressures (Extended Data Fig. 1d, e).

The distributions of copy numbers for other elements of the genome (telomeres, the 2μ plasmid and Ty transposons) had lower variations than rDNA (Fig. 1f, Extended Data Fig. 1c); copy numbers for the 2μ plasmid did not follow a normal distribution, Extended Data Fig. 1f). Estimates of telomere length highlighted strains that lack subunits of telomerase (est1Δ, est2Δ or est3Δ) as having the longest telomeres, which is consistent with alternative telomere elongation not being subject to normal homeostasis of telomere length. Our predicted telomere maintenance (TLM) genes partially overlapped with those that have previously been defined12,13 (Fig. 1g, h). To validate predictions for the newly identified TLM genes, we measured telomere lengths by gel electrophoresis, confirming telomeric phenotypes for 12 of the 14 selected (Fig. 1i, Extended Data Fig. 3a). We also validated the telomeric effects for about 70% of a selection of strains that failed stringent criteria, but which had high or low telomere-length estimates (Extended Data Fig. 3b). Network analysis of predicted TLM genes prominently featured DNA-, RNA- and/or chromatin metabolism, and highlighted connections to endoplasmic reticulum, Golgi and/or vesicle trafficking14 (Extended Data Fig. 3c).

Sequencing coverage also offers information on copy-number variation between and within chromosomes. Notably, about 52% of the chromosomes that we sequenced displayed some deviation from 2n ploidy, and a minority exhibited stable chromosome gains and/or losses (Fig. 2a, Methods). Most of the aneuploid strains displayed ‘incomplete’ aneuploidy (the gain or loss of less than a whole chromosome unit), which suggests cell–culture heterogeneity. These events affected different chromosomes in different ways, with the smallest chromosomes (I, VI, and III) being the most frequently lost (consistent with chromosome loss negatively correlating with chromosome size15) and the largest
Fig. 1 | Assessment of repetitive-DNA alterations in the YKOC. a, Screen schematics. b, Distribution of rDNA copy number for YKOC strains (details in the legend of Extended Data Fig. 1c). c, Extract from b, showing that gene knockouts that affect functional rRNA synthesis display increased rDNA copy number. Median from $n = 2–8$ biologically independent samples. WT, wild type. d, Overlaps between genes that affect rDNA length identified in this study (red circles) and a previous study (blue circles). Asterisks indicate that some genes (for which we have no data) were removed from the hits identified in the previous study. e, Estimates of rDNA copy number of strains that carry temperature-sensitive ($ts$) 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, Distribution of telomere length for YKOC strains (rpm, reads per million; details in the legend of Extended Data Fig. 1c). g, h, Overlap between gene knockouts associated with telomere lengthening or shortening identified in this work (red circles) and previous studies (yellow circles, ref. 12; blue circles, ref. 13). EST-knockout strains (grey) have previously been identified as having shorter telomeres. Gene symbols in bold indicate gene knockouts that were selected for validation of their telomere-length phenotype by gel electrophoresis. i, Validation of 14 predicted newly identified TLM genes.

Fig. 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 the legend of Extended Data Fig. 1c). b, Chromosome instability (CIN) estimates for 106 fresh knockout strains selected from 1,006 candidates. Each dot represents a strain and indicates the deviation of chromosome content from the expected diploid status ($2n$). Red dot, average; green band, wild-type sample s.d.; blue column, knockout sample s.d. c, Distribution of chromosome rearrangement breakpoints detected in the YKOC (details in the legend of Extended Data Fig. 1c). d, Localization of chromosome rearrangement breakpoints in strains analysed. $MAT/HML/HMR$, $MAT$ and silent mating-type loci.
that most aneuploidies do not originate from severe chromosome-segregation defects caused by a gene knockout. Consistent with this, when we focused on ribosomal proteins—which are often encoded by two near-identical gene paralogues—in 28 out of 97 cases that we analysed, gene deletion was accompanied by gain of the chromosome that carries the parologue (Extended Data Fig. 4c).

Approximately 2% of strains carried single or multiple subchromosomal deletions or amplifications, often initiating at Ty transposons or interspersed long-terminal-repeat sites and terminating at similar sites or chromosome ends (Fig. 2c, d). As previously observed\(^{21-23}\), breakpoint localizations to repetitive regions suggests that homologous-recombination processes drive rearrangements such as these. Mapping the remaining breakpoints revealed two strains carrying a single-copy gain (rad2Δ) or loss (mrp21Δ) of the same chromosome IV fragment, the boundaries of which map to the homologous RPL35A and RPL35B genes. As with aneuploidy, we found evidence that rearrangements compensate for the effects of gene knockouts (Extended Data Fig. 4d).

Together with previous work\(^{24-26}\), our results suggest that relatively frequent chromosomal mis-segregations enable the evolution of aneuploidy-driven phenotypic suppression and fitness improvements in many gene-knockout settings.

Estimating mitochondrial DNA (mtDNA) copy numbers from whole-genome sequencing data (Extended Data Fig. 5) indicated that about 5% of non-essential genes are required to maintain mtDNA (Fig. 3a), and that these genes overlap with genes that are required for respiratory growth or that encode mitochondrial proteins (Extended Data Fig. 6a). When we assessed connections between various features of nuclear-genome instability, and between these features and mtDNA copy number, we found correlations between aneuploidies and changes in mtDNA copy number (Fig. 3b). We also found expected correlations between aneuploidies and altered Ty copy numbers, as chromosome gains or losses can alter the number of Ty elements (Fig. 3b).

Accordingly, strains with higher deviations from 2\(H\) biologically independent samples, \(n = 8,843\) (total) or 303 (\(rho^9\))

mtDNA analyses also highlighted about 130 gene knockouts that were associated with increased mtDNA copy number (here referred to as \(rho^{++}\) phenotype) (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 and other DNA metabolic processes, rather than to mitochondrial biology (Extended Data Fig. 6b). We hypothesized that defects in DNA repair might lead to the \(rho^{++}\) phenotype via persistent activation of the DNA-damage response. Consistent with this, genotoxin-mediated activation of the DNA-damage response in wild-type strains triggered an increased mtDNA copy number (Fig. 3e). Increases in mtDNA copy number can arise from the overexpression of components of ribonucleotide reductase (RNR)\(^{28}\) that are under the control of the DNA-damage response\(^{29}\). Accordingly, knockouts of transcriptional repressors (RFX1, TUP1 or CYC8) of RNR genes led to increases in mtDNA copy numbers (Fig. 3f).

Moreover, the 16 \(rho^{++}\) strains that we selected for analysis, 9 displayed RNR induction, 6 displayed spontaneous Rad53 phosphorylation (which is a marker of activation of the DNA-damage response) and 13 were hypersensitive to hydroxyurea, which is an RNR inhibitor
that causes replication stress (Extended Data Fig. 6c). Furthermore, knockout strains that are hypersensitive to hydroxyurea\textsuperscript{36,31} were enriched among \textit{rho}^0 strains (Extended Data Fig. 6d). Comparing mtDNA contents with systematic analyses of Rnr3 levels\textsuperscript{32} revealed that many knockout strains in which genes that protect genome stability were deleted displayed increased levels of both Rnr3 and mtDNA (labelled in green in Extended Data Fig. 6e). Accordingly, overexpressing the Rnr3 catalytic subunit of RNR (and, to a lesser extent, the Rnr1 catalytic subunit) increased mtDNA levels (Fig. 3g), whereas overexpressing Rnr2 or Rnr4 did not. As overexpression of Rnr1 or Rnr3 increases dNTP production increases mtDNA copy numbers, perhaps by stimulating mtDNA replication. Unexpectedly, among \textit{rho}^0 strains with normal levels of Rnr3 were those with knockouts of genes that encode enzymes of the tryptophan biosynthesis pathway (Extended Data Fig. 7). The basis for this connection requires further investigation.

Extracting information on single-nucleotide variants (SNVs), and insertions and deletions (indels), across the YKOC revealed genes that are frequently mutated (Fig. 4a). Inspecting the coding sequences of these genes revealed a subset containing microsatellite features (homopolymers and/or small degenerate tandem repeats), which suggests that variants in these genes had arisen via imprecise DNA replication, or imprecise sequencing or read-mapping (grey bars in Fig. 4a, Extended Data Fig. 8). In another subset of frequently mutated genes, each gene carried one very frequent ‘founder’ mutation (yellow bars in Extended Data Fig. 6a) that was found predominantly in knockout strains generated by a single YKOC consortium laboratory (Extended Data Fig. 9). The remaining frequently mutated genes included \textit{ATP1}, \textit{ATP2} and \textit{ATP3}, which encode subunits of the ATP-synthase complex that has recently been identified as a cell-fitness hub\textsuperscript{56}. Analysis of the prevalence of mutations revealed that \textit{ATP1}, \textit{ATP2}, \textit{ATP3} and \textit{SIT4} (which encodes a phosphatase that acts on ATP synthase\textsuperscript{35}) mutations were frequent in \textit{rho}^0 strains, but relatively rare in other strains. (Fig. 4b). These results suggest that alteration of the function of ATP synthase can promote the fitness of \textit{rho}^0 cells.

As the YKOC strains have been cultured over many years, their overall mutation numbers could highlight hyper-mutators. Thus, we estimated the sequence divergence of two colonies of the same knockout strain from each other, and from the wild-type strain BY4743. Hyper-mutators are expected to show increases in both measures and, accordingly, strains defective in DNA-mismatch repair (\textit{pms1}\textsuperscript{Δ}, \textit{msl2}\textsuperscript{Δ} and \textit{msl6}\textsuperscript{Δ}; only one \textit{mll1}\textsuperscript{Δ} colony sequence was available) were readily detected this way (Fig. 4c, d). Many strains that lack components of homologous recombination (\textit{rad51}\textsuperscript{Δ}, \textit{rad52}\textsuperscript{Δ}, \textit{rad55}\textsuperscript{Δ} and \textit{rad57}\textsuperscript{Δ}) showed a similar phenotype, which implies that deficiencies in homologous recombination yield SNVs—possibly via mutagenic trans-lesion DNA polymerases\textsuperscript{36,37}. Distinct patterns of mutation can be used to cluster genome-stability genes into functional categories\textsuperscript{38}. The clustering of hyper-mutators that we identified on the basis of their mutation profiles produced two main groups (which correspond to known DNA-mismatch repair and homologous recombination genes), and also clustered \textit{ymr166c}\textsuperscript{Δ} and \textit{ggc1}\textsuperscript{Δ} with the group of DNA-mismatch repair genes (Fig. 4e). \textit{YMR166C} is located upstream of \textit{MLH1}, so its deletion could affect \textit{MLH1} expression\textsuperscript{19}; SNV abundance in \textit{ggc1}\textsuperscript{Δ} remains to be explained.

Our systematic survey of genomic and karyotypic features highlights the effects of gene knockouts on genome architecture, and underscores the flexibility of the yeast genome to mitigate effects of gene loss. Overall, about 36% of YKOC strains carry some repetitive DNA or chromosomal abnormality. Among these, the loss of any one of 151 genes is associated with multiple alterations, which suggests that the products of these genes protect against widespread genome instability (Extended Data Fig. 10a). Although Gene Ontology analysis revealed strong enrichments for chromosome biology (Extended Data Fig. 10b), the largest fraction of these genes encodes proteins that control mitochondrial functions (Extended Data Fig. 10c). Further work
is required to understand these connections, as well as how maintenance of the nuclear and mitochondrial genomes are intertwined. Our results also highlight alterations in repetitive DNA as types of mutation signatures, in addition to SNVs first identified in cancers. It will be interesting to integrate our findings with data from future studies that assess propagated lineages of knockout strains for numbers of repetitive elements and for repetitive-element copy-number variability, and to determine whether and how instability of repetitive elements might relate to other types of genome instability. Sequencing genomes of strains that bear hypomorphic alleles of essential genes and of synthetic-sick double mutants should also provide insights into processes that affect genomic stability. Such future work may have medical relevance because—together with SNVs and indels—variations in DNA copy number and aneuploidies are linked to developmental disorders, cancer, and other diseases.

Online content
Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1549-9.

Received: 13 June 2018; Accepted: 7 August 2019; Published online 11 September 2019.
Quantification of rDNA and CUP1-repeat copy number. Quantification of rDNA and CUP1 copy numbers was carried out using the S288C genome as a reference. Estimates of copy numbers were obtained by dividing the median coverage across the appropriate genomic region (rDNA, chromosome XIII: 452000–459000; CUP1, chromosome VIII: 212986–213525) by the genome-wide median coverage and multiplying by two (this is because the S288C reference genome contains two copies of the minimal repeat unit).

Oxford Nanopore Technology library preparation and sequencing. Yeast genomic DNA was prepared by using standard molecular biology techniques (spheroplasting, SDS lysis, K acetate precipitation and RNase A digestion) without phenol–chloroform extraction. A sequencing library was prepared from purified genomic (g)DNA with the rapid sequencing kit (SQK-RAD004, Oxford Nanopore Technology) 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 (2.1.8-r883)5. Long reads that aligned in the region of interest (VIII:213000–213500) were extracted, and graphically aligned to the reference genome using FlexiDot6.

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 = chromosome IV: 1203704–1215621; YLRTWyT-1 = chromosome XXI: 938192–950150; YILWTy3-1 = chromosome IX: 202220–213647; YHLWTy4-1 = chromosome VIII: 82539–94761; YCLWTy3-1 = chromosome III: 679–7322; MATα_HMR = chromosome III: 292388–295034; MATa_HML = chromosome III: 11146–14849; 2μ = 1J03471.

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 bamsammpairingduplicates (biobambam2 2.0.50).

Validation of gene knockout. For each sequenced DNA sample (corresponding to one isolated colony), sequencing coverage across the open reading frame (ORF) expected to be deleted in that sample was computed. An ORF coverage <15% of the relevant chromosome-wide median 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 for a region of varying dimensions within the ORF itself that usually disrupts the start codon, and probably produces ORF inactivation. To account for this, ORFs that failed the previous criteria were evaluated for the presence of at least 9 consecutive bases with coverage <5% of the chromosome-wide median. Samples that matched these criteria are indicated as ‘partial deletions’. In each sample, UP and DN barcodes were read from reads mapping—

in normal and in reverse complement directions—the patterns /U1[ACGCT]+U2/ and /D1[ACGCT]+D2/, respectively. U1, U2, D1, D2 are the relevant unique flanking sequences (U1, GATGTCACAGGTCCTC; U2, CGTACGTCACAGGTC; D1, CGGTTGTCAGGTCGAT; D2, ATCGATGAATTCGAGCTCG). For each sample, the detected barcodes—observed at least 3 times and for a frequency >20%—are reported in Supplementary Table 6, and noted if they do not match with the expected barcodes.

Re-assignment of gene knockouts. Any sample that did not carry the expected deletion was reassigned to the correct knockout strain using information encoded in the unique barcodes, when possible. Samples that were re-assigned in this way were subsequently validated again for the presence of the deletion, as described in ‘Validation of gene knockout’ before any further work.

Development and validation of tools to study instability of repetitive DNA. To study repetitive regions, we developed tools to measure copy-number variation from sequencing data, on the basis of the concept that the number of sequencing reads generated by a certain locus is proportional to its copy number. This last number was confirmed using long-read sequencing and identifying regions that spanned the CUP1 locus and containing the expected four repeats (Extended Data Fig. 1f, g).
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 for which 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 2 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 that these might be localized to transposons. To assess this, breakpoint positions were compared with positions of known transposons, long terminal repeats (LTRs), rDNA and subtelomeric regions (<15 kb from chromosome end). A breakpoint was assigned to 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 (that is, when a breakpoint mapped to both a transposon and LTR it was considered only as a transposon, as transposons also have nearby LTRs but LTRs do not necessarily have a nearby transposon).

Quantification of mtDNA copy number. mtDNA copy number was determined using the S288C genome as reference. Estimates of copy number 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 probably 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 chromosome II was obtained as described above (Extended Data Fig. 5c). RNA was purified using RNA Clean & Concentrator Kit (Zymo; cat. no. R1016). Quantification of RNR expression.

For comparison purposes, the estimation of mtDNA copy number from WGS data was also carried out using the following genomic regions, corresponding to qPCR amplitons: mito: 25574–25686 for mtDNA; chromosome II: 280382–280459 for GAL1 (qPCR control, instead of genome-wide 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-damaging drugs (MMS, 0.01%; CPT, 20μM; HU, 100 μM). Remaining cells were collected and frozen. The same process 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. The hot phenol extraction method was used to isolate total RNA. Quantity and purity were analysed using a NanoDrop 1000. Ten micrograms of total RNA were treated with Turbo DNase (Invitrogen; cat. no. AM2238). RNA was purified using RNA Clean & Concentrator Kit (Zymo; cat. no. R1016). cDNA was then prepared using Superscript III reverse transcriptase (Invitrogen; cat. no. 18080). The expression level of individual transcripts was determined by qPCR using Fast SYBR Green Master Mix (Thermo; cat. no. A385612). Relative levels were determined by normalization to the ACT1 mRNA in each sample. Primers used were: RNR1, up: GGTGAGAAGGTTTTACAC; low: TCCCTTTGGAACACAGAA. RNR2, up: TATCCA TGACGGAACACAGAAT; low: GTGGAGAAGGTTTTACAC. RNR3, up: AGGTCGTGAAACAAACAAAGC; low: TGTGGTGTTGCTCCGTGTGACAT; RNR4, up: GGTAACTTGTTACCG TTTGCTTAC; low: CTAATCCTGGACCCCGTGAACAC; and, ACT1, up: GAATGGCAAAGTCCGCTCTG; low: TACCCGGCAAGTCCCACAA.

RNR overexpression experiments. Yeast strains that carry genes that encode subunits of ribonucleotide reductase (RNR1, RNR2, RNR3 or RNR4) under the control of the GAL1–10 promoter were grown at 30°C overnight in YPD 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 YPA + 2% raffinose or YPA + 2% raffinose + galactose. The same procedure was repeated on the following days. For each set of wells, samples for DNA extraction and sequencing were collected about 48 hours 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 -d DV -C 0 -pm 3 -F 0.2 -d 10000, and bcftools call with the following options: -v -f GQ. Mutations were initially filtered using vcf tools with the following options: -H -f -/d = 15/q = 25/SnpGap = 7. 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 vcf tools excluding variants with genotype quality < 30 or depth < 10. Prediction of the effects of variants was carried out using the Ensembl variant effect predictor47 as previously described41.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

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 at http://sgv.gordon.cam.ac.uk.

Code availability
The custom code used for the analysis of primary sequencing data is available at https://github.com/fahbudda/sugar-fermentation (v.0.5) and it relies on Slurm workload manager (v.15.08.13); Samtools v.1.3.1 (using htslib 1.3.1); VCFtools v.0.2.13); Bcftools v.1.3.1 (using htslib 1.3.1); BWA v.0.7.12-r1039; Python 2.7.12; Perl (v.5.22.1); Snapgene (v.8.0). The code for secondary analyses is available at: https://github.com/brubinste/RNR.

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 34, 3594–3600 (2018).
43. Seibt, K. M., Schmidt, T. & Heitkamp, T. FlexDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. Bioinformatics 34, 3575–3577 (2018).
44. Shampay, J., Szostak, J. W. & Blackburn, E. H. DNA sequences of telomeres. Nature 310, 154–157 (1984).
45. Rubinstein, L. et al. Telomere length kinetics assay (TELKA) sorts the telomere length maintenance (telom) mutants into functional groups. Nucleic Acids Res. 42, 6314–6325 (2014).

R(LY8492) with the Frozen-EZ Yeast Transformation II Kit (Zymo; cat. no. T2001). qCFT assays were performed as previously described39, with minor adjustments: cells were not sonicated and were analysed on an Attune NxT Acoustic Focusing Cytometer with Autosampler (Thermo; cat. no. A24861).

(GAL1) was used for normalization to gDNA. Primers used were: gDNA (GAL1), up: TGGTTGGTGAACACAGAA. mtDNA, up: CACCACTAATTGAAACCT GTCTG; low: GATTATTCTGTGCTATTTCGCA.
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 laboratory of S.P.J. 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 M.H.; 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 R.L.

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.-L. and M.K. Chromosome instability measurements were performed by J.Z. and M.G. under the supervision of R.L. 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 from G.M.-Z., M.H., M.K., I.S., J.Z., I.A. and R.L.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1549-9.

Correspondence and requests for materials should be addressed to F.P. or S.P.J.

Peer review information
Nature thanks Grant Brown, Sergei Mirkin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Extended Data Fig. 1 | See next page for caption.
Extended Data Fig. 1 | Statistics of YKOC analyses, and distribution of estimates of repetitive DNA across the YKOC. a, Colonies that did not carry the expected deletion (red) were re-assigned 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 estimates of copy numbers 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 s.d. of the wild-type distribution. n = 8 biologically independent samples for wild-type strains; n = 1 sample for 258 of the knockout strains; n = 2 biologically independent samples for 4,097 of the knockout strains; n = 3 biologically independent samples for 30 of the knockout strains; n = 4 biologically independent samples for 72 of the knockout strains; and n = 5 biologically independent samples for 1 of the knockout strains. d, Correlations between changes in relative copy numbers at rDNA and CUP1 tandem-repeat loci. Average correlations in all colonies of each knockout strain are shown. n = 8 biologically independent samples for wild-type strains; n = 1 sample for 258 of the knockout strains; n = 2 biologically independent samples for 4,097 of the knockout strains; n = 3 biologically independent samples for 30 of the knockout strains; n = 4 biologically independent samples for 72 of the knockout strains; and n = 5 biologically independent samples for 1 of the knockout strains. e, No overall correlation for estimates of rDNA and CUP1 copy numbers across all colonies sequenced. f, Distribution of estimates of copy number for the 2μ plasmid across the YKOC, and Gene Ontology analysis of the hits. The 2μ copy numbers did not follow a normal distribution. Instead, the maximum s.d. increases 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 to be at 20–25 copies). Different durations of FLP1 expression will result in different increases in copy number. We detected an enrichment in gene knockouts connected to gene silencing in strains with high 2μ copy numbers, which is also consistent with this 2μ amplification mechanism.
Extended Data Fig. 2 | See next page for caption.
Extended Data Fig. 2 | Development of tools to study the instability of repetitive DNA. 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, WGS estimates of copy numbers of rDNA repeats linearly correlate with estimates obtained by pulsed-field gel electrophoresis. n = 3 or 4 biologically independent samples per strain. c, Per cent 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 owing to their relatively low repeat numbers. d, Relative estimated content of telomeric repeats in indicated strains (normalized to the 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, Estimates of telomere length for wild-type, tel1Δ and rif1Δ strains obtained by calculating the relative abundance of telomeric reads. Mean from n = 4–8 biologically independent samples per strain. f, Estimates of rDNA, CUP1, Ty1, Ty2 and mtDNA copy numbers, and telomeric DNA content for MATα, 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 Oxford Nanopore Technology sequencing of a W303 (K699) genomic library. h, Comparison of CUP1 copy number estimated by qPCR or by WGS. The same DNA samples (as indicated by the labels) were analysed. 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 Fig. 3 | Southern blot analysis and functional connections between TLM genes. a, Gel electrophoresis and Southern blot analysis of telomeres for 14 newly identified predicted TLM strains (‘hits’). Hits denote strains with two or more colonies with measures >3× the s.d. of the wild-type distribution. Twenty-one strains that failed these stringent hit-selection criteria (non-hits), but which displayed relatively high or low estimates of telomere length, are also shown. Representative images from two independent experiments. Purple lines, location of molecular mass 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 gene knockouts that failed the TLM selection criteria but that displayed relatively high or low telomere counts. c, Network-graph analysis of gene knockouts that affect telomere length, highlighting the newly identified genes that were validated by Southern blotting.
Extended Data Fig. 4 | Examples of aneuploidies and chromosomal rearrangements in the YKOC. a, Example of a strain with fractional aneuploidy of chromosome XII, which probably reflects clonal heterogeneity. b, Distribution of fractional and non-fractional aneuploidies per chromosome. \( n = 8,843 \) biologically independent samples. c, Knockout of genes that encode subunits of ribosomal proteins frequently leads to gain of the chromosome that carries the paralogue gene. d, Ploidy plots of chromosome II for two different colonies of the hta1\( \Delta \), swi4\( \Delta \) and spt10\( \Delta \) knockout strains. hta1\( \Delta \) cells (in which one of the two genes that encodes histone H2A is deleted) accumulate a specific amplification of a genome region that contains the paralogue HTA2, a centromere and two origins of replication. This is most probably 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: these were spt10\( \Delta \) and swi4\( \Delta \), which encode factors that control the transcription of cell-cycle-regulated genes (including histones).
**Extended Data Fig. 5 | Calculation of mtDNA copy number.**

a, Sequencing coverage across the mitochondrial genome of a wild-type haploid strain (BY4741; accession ERS616991). Shaded areas indicate regions (that loosely correspond to COX1 and COX3 genes) that were 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 the labels) were analysed by qPCR and WGS. Two estimates were extracted from WGS data: ‘from COX1’ indicates estimates 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 Fig. 6 | Connections between mtDNA and nuclear genome alterations. 

a. Venn diagram showing the overlap between genes identified as rho\(^0\) by our sequencing, genes that encode mitochondria proteins (ref. 44) and gene knockouts for which respiratory growth was annotated as 'absent' (in the Saccharomyces Genome Database (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 activation of the DNA-damage response by Rad53 and histone H2A phosphorylation (representative images from two technical replicates; source data in Supplementary Fig. 1) and RNR expression (average from three 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 HU sensitivity. 

parsons, 2004' denotes ref. 30, n = 62 biologically independent samples; 'woolstencroft, 2006' denotes ref. 31, n = 33 biologically independent samples. 

d. Comparison of predicted mtDNA copy number and RNR3 expression levels. Gene knockouts with increased levels of Rnr3 protein (blue, Z-score > 2); gene knockouts with increased mtDNA (yellow, mtDNA > 22.2); gene knockouts with both measures increased (green). 

n = 4,436 by knockout averages, of n = 8,843 biologically independent samples.
Extended Data Fig. 7 | mtDNA content in knockout strains for genes that encode enzymes involved in tryptophan metabolism. Pathway for tryptophan biosynthesis from phosphoenolpyruvate, tryptophan import and NAD biosynthesis from tryptophan are depicted along with estimates of mtDNA copy number for strains that lack each of the enzymes in the pathways. Mean from \( n = 8 \) (wild type) or 2 (knockout) biologically independent samples.
Extended Data Fig. 8 | Genes that frequently carry mutations contain repetitive regions. Self dot plots that highlight degenerate repetitive regions in the DNA sequence of genes that are found to be frequently mutated in the YKOC. Plots were obtained using FlexiDot.
Extended Data Fig. 9 | The most-frequent YKOC mutations, and their distributions between different source laboratories. The 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 that comes from a founder effect (yellow) or a frequently mutated site (green). Genes in bold indicate a homozygous mutation.

b. Not all strains derived from each laboratory share founder mutations. As in a, but a value of 100% in the heat map indicates that all the strains generated by a particular laboratory have the mutation.
Extended Data Fig. 10 | Overview of genomic instability caused by knockouts of non-essential genes. 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 151 genome instability genes, defined as knockouts that show 3 or more abnormal features. The numbers of genes in each Gene Ontology category, as well as Holm–Bonferroni-corrected $P$ values, are reported. c, Genome instability genes were manually sorted into classes on the basis of their function, inferred from annotations in the SGD.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a □ Confirmed

☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☐ The statistical test(s) used AND whether they are one- or two-sided

☐ Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ A description of all covariates tested

☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) and variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☐ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

☐ Give P values as exact values whenever suitable.

☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☐ Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

HiSeq Control Software (3.3.76); RTA basecalling (2.7.6); Illumina2bam (V1.19); biobambam (2.0.50); BamIndexDecoder (V1.19); bwa (0.7.13-r1128); scramble (1.14.8); spatial_filter (V0.0.26-dirty);

Data analysis

Augur Fermentorium (v0.5; a pipeline and collection of custom tools created for this work, available on GitHub) it relies on 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); Snuplot (Version 5.0 patchlevel 3);

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the primary sequencing data has been deposited in the European Nucleotide Archive in the study ERP109205 under the accession number detailed in Supplementary Table 1. Gene-specific information is available interactively at the address http://agv.gordon.cam.ac.uk. The software used for the analysis is available at the address https://github.com/fabiopuddu/augur-fermentorum.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size
No sample size calculation was performed; The number of 2 colonies analyzed per knockout strain was the minimum to address reproducibility and the maximum allowed by reasons of economics; this sample size is sufficient to detect strong differences between KO and wild-type samples; stringent criteria were used to minimize detection of false positives.

Data exclusions
Exclusion criteria were pre-established; samples not carrying a verifiable gene deletion were excluded from the data analysis, because it would be impossible to determine which gene KO was responsible for their phenotypes.

Replication
Two colonies [biological replicates] from each knockout strain were analyzed. Knockout strains with non-reproducible phenotypes between the two colonies were not considered hits (unless only one colony was sequenced); All other attempts at replication were successful.

Randomization
No formal randomization procedure was used to sort strain in different sequencing lanes / libraries; batch effects arising from different sequencing lanes were detected and corrected by adjusting the measures so that the median of the 96 samples in each sequencing lane is the same.

Blinding
Samples were managed through the pipeline using barcode identifiers not immediately attributable to gene knockouts. Exactly the same bioinformatic analysis was run on every sample, therefore we feel formal blinding is not relevant.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | Antibodies            |
| ☑   | Eukaryotic cell lines |
|     | Paleontology          |
| ☑   | Animals and other organisms |
| ☑   | Human research participants |
| ☑   | Clinical data        |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | ChIP-seq              |
|     | Flow cytometry        |
|     | MRI-based neuroimaging|

Antibodies

Antibodies used

Rad53 [Abcam; ab104232; polyclonal 1:4000], H2A-S129P [Abcam; ab15083; polyclonal 1:4000], H2A(Active Motif; #39235; polyclonal; 1:5000]

Validation

All validated for S. cerevisiae by the manufacturer and used extensively in other publications

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

The yeast homozygous diploid YKOC was acquired from EUROSCARF in 2001

Authentication

Samples were included in the analysis when a "clean" gene deletion could be confirmed (median coverage across the gene < 15% of the chromosome-wide median or a continuous stretch of at least 9 nucleotides with coverage < 5% of the chromosome-wide median coverage). All other samples were excluded.

Mycoplasma contamination

Not applicable to yeast
Commonly misidentified lines
[See ICLAC register]

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.