nature portfolio

Corresponding author(s): Grzegorz Ira, Kaifu Chen

Last updated by author(s): Aug 19, 2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	All raw fastq files were generated in this study using illumina MiSeq with the system suite version 4.0.0.116. These data have been deposited in the GEO database under accession code GSE246469 with no restriction.
Data analysis	All the details of data analyses can be found in the method section. All custom codes are developed by Perl and Shell. Statistical tests were performed by R packages or Prism. All original code has been deposited on GitHub. The code for large insertion detection can be accessed at the following GitHub repository: https://github.com/gucascau/iDSBins.git. The features related to large insertion analysis have been uploaded to a dedicated GitHub repository, which can be found here: https://github.com/gucascau/LargeInsertionFeature.git. Commercial or free software used: PEAR v0.9.11 BLASTN v2.8.1 BBDuk v38.46 BEDTools shuffle (Quinlanet al., 2010) Prism 10.0d by GraphPad Software ImageQuant TL 7.0 by GE R4.2.0 SnapGene version 7.0 by GSL Biotech LLC

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The Break-ins data generated in this study have been deposited in the GEO database under accession code GSE246469 with no restriction. Locations of the confirmed origins of replication (ARSes) were collected from the OriDB database (http://cerevisiae.oridb.org/). We collected locations of reference R-loops from a published source (Wahba et al., 2016). Known tandem repeats were downloaded from UCSC genome browser as a compacted file provided at the website (https://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/bigZips/chromTrf.tar.gz). All other genomic features, i.e., tRNA, telomere, centromere, were acquired from the SGD database (https://downloads.yeastgenome.org/).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

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 ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental 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	The exact number of insertion and number of colonies tested and cells plated are provided in supplementary Data1. Considering the super low frequency of insertion in wild-type growing cells, we collected more than 5000 colonies for each sample, which is enough to get the statistically significant difference.
Data exclusions	Insertions presented in each figure are described in figure legend. In analyses only single insertions are taken into consideration. The reason is that breakpoints of complex insertions from mtDNA are harder to define precisely because of highly repetitive nature of mtDNA. We have not analyzed nsertions from proximity of Mata, which is very rare and beyond the scope of this manuscript.
Replication	We confirmed phenotypes in nuc1 mutants using at least 3 different transformation isolates. Cell viability, respiration deficiency, frequency of TRP1-mtDNA transfer to nucleus, Southern blot analysis of mtDNA and Ty and all other experiments were repeated successfully at least 3 times. We specified the times of repeat in figure legends.
Randomization	We tested all colonies grown on YP-GAL plates for insertion analysis by Break-Ins seq. We randomly picked colonies grown on YP-GAL plates to analyze insertion originating from transformed DNA fragments of different sizes.
Blinding	We analysed all NHEJ products from the plate. Each NHEJ products is represented as individual colony. We do not know whether NHEJ carries or not templated insertions. Therefore we were blinded to data collection. Similarly we analysed all sequenced events in wild-type and mutants the same.

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
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A