nature research

Corresponding author(s): Grzegorz Ira, Kaifu Chen

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	All yeast genome features are downloaded from SGD database.
Data analysis	All custom codes are developed by Python with numpy, scipy, pandas. Statistical test is performed by using function within Scipy package. They are deposited into "https://github.com/fagisX/FAID". Commercial or free software used: ApE - A plasmid Editor v2.0.53c by Wayne Davis Prism 7.0d by GraphPad Software ImageQuant TL 7.0 by GE SnapGene version 4.2.6 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/reviewers upon request. 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

Raw data in a form of DNA sequences of insertion are provided within Supplemental Table 1 - excel file. They are grouped by genotype except the last page that has multiple strains but with smaller number of events per genotype. Thus anybody who wishes to repeat the analysis could upload the sequences from Excel file.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-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 events and donor DNAs per genotype is provided in Extended Data Table 1. The major observation in the manuscript relates to the frequency of insertions among cells that repaired a DSB. Insertions are not observed in wild type (0/644) and are observed in a dna2 mutant (8%, 148/1794). The sample size (number of colonies tested) that would determine whether derivatives of dna2 mutant or other mutants show similar or significantly different frequency of insertions was determined to be 113; and the lowest sample size tested here is 144. In order to ensure that the 95% confidence interval estimate with 5% margin of error, n = p x (1-p) x (Z/E)2, where Z for 95% confidence interval is 1.96, E for 5% margin of error is 0.05, p for the proportion of insertions observed in dna2 mutant is 0.08 (148/1794). The estimates of the necessary sample size is 113 to assures an adequate power to detect statistical significance.
Data exclusions	All of the sequences of inserted DNA are provided in Supplemental Table 1. Analysis of the insertion features was done with all events from all dna2 mutants (Fig 1a, 1c, 1e and extended Fig. 1b) except: -In Extended Data Fig. 2a, 2b, nearly all events are presented. The exceptions are insertions coming from telomeres and transposons because these are repetitive elements and we don't know exact loci of the origin of inserted DNA. Also 2 micron plasmid insertion events are not shown in Extended Data Fig. 2a, 2b because it is not a part of chromosome. These are shown separately in figure 3c. -All transposon insertion events from all dna2 mutants are presented in Figure 2a and 2b. -All rDNA insertion events from all dna2 mutants are presented in Figure 3a. -Insertions from pif1-m2 dna2 rad52 mutant and Cas9 induced insertions were not included in any statistical analysis. These data constitute only a small fraction of all events analyzed and were were added at revision step to address specific reviewer questions that are not related to global features of insertions. -In figures 4a, 4b, and 4c left, all insertions were analyzed with exception of repetitive elements (rDNA insertions, transposons insertions, telomere insertions). Reason - we don't know exact loci of the origin of inserted repetitive DNA. Also 2 micron plasmid insertion events were not included as the plasmid is not a part of chromosome. -In figure 4c right, all insertions were analyzed with exception of repetitive elements (rDNA insertions, transposons insertions) and non-chromosomal 2 micron plasmid insertions. In Extended Data Figure 2c, all insertions at MATa locus were analyzed with exception of rDNA insertions, transposons insertions, 2 micron plasmid insertions and telomere. Reason - we don't know exact loci of the origin of inserted repetitive DNA and 2 micron is circular extrachromosomal DNA. For all bar graphs that show insertion number per mutant strain (1b, 4d) the exact number of colonies tested is shown in Extended Data Table 1. These numbers were used to calculate p values to determine significant differences between mutant strains (chi square, confidence interval 95%).
Replication	We confirmed the major discovery, insertions of DNA fragments in dna2 mutant strains in several ways. First, we observed large insertions in all 3 independent pif1-m2 dna2 strains at MATa locus. Second, we observed this phenotype in yen1ON dna2 strain at HO site at MATa locus. Third, we observed this phenotype in pif1-m2 dna2 strain at HO induced DSB at URA3 locus. Fourth, we observed this phenotype in pif1-m2 dna2 strain at Cas9 induced DSB at LYS2 locus. To measure the NHEJ efficiency, we repeated the experiment at least three times for each mutant. Analysis of free DNA, analysis of Ty's cDNA amount and its stability and RNA amount of Ty1 was repeated 3 times or more. The rate of spontaneous transposition was measured as previously described (Fig 2d), brief description is provided in methods section.
Randomization	To test by PCR the presence of insertions at DSB after repair by nonhomologous end joining we screened all colonies or random colonies grown on YP-GAL plates.
Blinding	All or random colonies from the plate that represent survivors of DSB repair were analyzed. Colonies that carry insertion at MATa or LYS2 loci are not distinguishable from colonies that do not have insertions at DNA break and therefore can not be selected for.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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