nature portfolio

Corresponding author(s): Inanc Birol

Last updated by author(s): Oct 17, 2025

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	Data from the Human Pangenome Reference Consortium (HPRC) and the 1000 Genomes Project ONT Sequencing Consortium were downloaded by AWS. Data from the Human Genome Structural Variation Consortium were downloaded by the Unix utility wget.
Data analysis	Hifiasm (v0.19.5) was used to assemble PacBio HiFi sequencing data from the HPRC. minimap2 (v2.24) was used to align the sequencing reads for tandem repeat (TR) detection by Straglr (v1.4.1). TRF (v4.09) was used for TR detection in sequence assemblies. A Snakemake (v7.32.4) pipeline with custom Python (v3.9.4) scripts was built to extract TRs from the assemblies and map them to the reference genomes. Code of the pipeline is available in the public GitHub repository https://github.com/bcgsc/tr_catalog and archived under DOI: https://doi.org/10.5281/zenodo.17336276

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

BED files of TR genotypes with sample origins, sizes, repeat copies, and annotated features for both GRCh38 and T2T-CHM13 genome references are archived https://doi.org/10.5281/zenodo.1523764678. Unfiltered (v1) and subsets with only genic (v2.genic) and polymorphic (v2.polymorphic) loci, respectively, are available in addition to the final version (v2) of the catalogs. Online locations of LRS read and assembly sequences used in this study can be found in Supplementary Table 1.

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	Sex and gender analysis was not performed in this study.
Reporting on race, ethnicity, or other socially relevant groupings	Genetic ancestries of samples were used to test if tandem repeat genotypes are sufficient for ancestry classification. Ancestry labels (African, European, South Asian, East Asian, and American) were available from the 1000 Genome Consortium.
Population characteristics	No other population characteristics of the samples were used.
Recruitment	Recruitment is handled by the 1000 Genome Consortium. We are only using data that is publicly available.
Ethics oversight	There is no ethics requirement in this study as only publicly available anonymous data is used.

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	272 samples from the 1000 Genome Consortium with publicly available long-read sequencing data
Data exclusions	One Oxford Nanopore (ONT) sequence assembly from one sample was excluded because an assembly of the same sample using PacBio HiFi sequencing data (lower error rate) was available
Replication	Results from all the software employed in this study were deterministic and therefore all tandem repeat genotypes reported in the catalog should be reproducible
Randomization	Samples were not allocated into experimental groups in this study
Blinding	No group allocation was performed in this study

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	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*