nature research

Corresponding author(s): Evan E. Eichler
Mark J.P. Chaisson

# 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<br>*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<br>*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<br>*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 | No software was used. |
| Data analysis | Code for analyzing the resolved and unresolved segmental duplications in a de novo assembly can be found at https://github.com/mvollger/segDupPlots. Code for processing de novo assemblies to find collapses and running SDA can be found at https://github.com/mvollger/SDA.<br><br>Main software used:<br>canu version 1.5<br>whatshap version 0.16<br>blasr version rc46<br>minimap2 version 2.11<br>quiver version 1.1.0<br>mashmap version 2.0<br>miropeats version 1.0<br>RepeatMasker version 2004/03/06<br>samtools version 1.9<br>bedtools version 2.27<br>gephi version 0.9.2<br>miniasm version 0.3 |

wtdbg version 1.2.8
snakemake version 5.2.2

For a complete list software used by the distributed version of SDA see:
https://github.com/mvollger/SDA/blob/master/ymls/python3.yml, and https://github.com/mvollger/SDA/blob/master/ymls/python2.yml

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

SMRT WGS for CHM1, CHM13, and NA12940 from this study are available at the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRP044331 for CHM1; SRX818607, SRX825542, and SRX825575-SRX825579 for CHM13; and SRX1093000, SRX1093555, SRX1093654, SRX1094289, SRX1094374, SRX1094388, and SRX1096798 for NA19240. ONT WGS data are available at https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md. De novo assemblies of CHM1, CHM13, NA12940, and NA12878 from this study are available at the NCBI Assemblies database (Assembly; https://www.ncbi.nlm.nih.gov/assembly/) under accession numbers GCA_001297185.1, GCA_000983455.2, GCA_001524155.4, and GCA_900232925.1, respectively. Assembled CHORI-17 BACs are available at the NCBI Clone database (Clone; https://www.ncbi.nlm.nih.gov/clone/) under the accession numbers listed in Table S4.

# 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 | We analyzed four genome assemblies using our method (CHM1, CHM13, NA19240, and NA12878). We believe this is a sufficient sample size since these samples represent: diverse individuals, different sequencing technologies, and different genomic architectures (hydatidiform moles and true diploids), all while still showing the utility and generalizability of the method. |
| Data exclusions | No data were excluded |
| Replication | Our method is computational and non-random. Rerunning with the same input always produced the same output. |
| Randomization | This is not applicable since we make no claims on covariation between the genomes we analyzed. |
| Blinding | Not applicable, there was no group allocation done during data analysis, so no blinding was required. |

# 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 |