nature research

Corresponding author(s): Justin M Zook

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	No software used for data collection.
Data analysis	All code for analyzing genome sequencing data to generate high-confidence variants and regions developed for this manuscript are available in a GitHub repository at https://github.com/jzook/genome-data-integration. Publicly available software used to generate input callsets includes novoalign version 3.02.07, samtools version 0.1.18, GATK v3.5, Freebayes 0.9.20, Complete Genomics tools v2.5.0.33, Torrent Variant Caller v4.4, LifeScope v2.5.1, LongRanger v2.1, GenomeWarp, rtg-tools v3.7.1, and sentieon version 201611.rc1.

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 sequence data were previously published in Scientific Data (DOI: 10.1038/sdata.2016.25), and were deposited in the NCBI SRA with the accession codes SRX1049768 to SRX1049855, SRX847862 to SRX848317, SRX1388368 to SRX1388459, SRX1388732 to SRX1388743, SRX852932 to SRX852936, SRX847094, SRX848742 to SRX848744, SRX326642, SRX1497273, and SRX1497276. 10x Genomics Chromium bam files used are at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/10XGenomics_ChromiumGenome_LongRanger2.0_06202016/. The high-confidence vcf and bed files resulting from work in this manuscript are available in the NISTv3.3.2 directory under each genome on the GIAB FTP release folder ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/, and in the future updated calls will be in the "recent" directory under each genome. The data used in this manuscript and other datasets for these genomes are available in ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/, and in the NCBI BioProject PRJNA200694.

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	7 human genomes were characterized as benchmarks because these were all of the samples currently chosen for extensive characterization by the Genome in a Bottle Consortium
Data exclusions	No data excluded
Replication	Overall statistics of the benchmark sets were compared across all seven GIAB genomes to determine reproducibility across samples.
Randomization	Randomization is not relevant to our study, as there were not distinct experimental groups
Blinding	These benchmark samples are an open science resource, so no information is blinded.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Coriell NIGMS Cell Line Repository (GM24385, GM24149, GM24143, GM24631, GM24694, GM24695, GM12878)
Authentication	Whole genome sequencing and variant calling was performed on all specimens
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination

Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.