

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 (<i>n</i>) 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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☒ | ☐ 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ 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 <i>d</i> , Pearson's <i>r</i>), 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

1. Oxford Nanopore (ONT) long- read data
HG002–HG007 ONT datasets were downloaded from the ONT Open Data S3 repository (giab_2025.01 release, s3://ont-open-data/giab_2025.01/basecalling/sup/) using the AWS CLI (aws s3 cp, AWS CLI v2). Long read ONT data for 557 samples were obtained from the CHN100K project (<https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA012590>), with consent from the data submitter/corresponding author of DOI:10.1016/j.scib.2024.06.017. ONT reads for 35 HGSVC3 samples were downloaded from the International Genome Sample Resource (IGSR) (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20240816_JAX_ONT_guppy6_Rebasecalled/). Additionally, ONT long read data for Macaca fascicularis, Mus musculus, Drosophila melanogaster, Danio rerio and Arabidopsis thaliana were obtained from the Sequence Read Archive (SRA) or European Nucleotide Archive (ENA) under BioProjects PRJNA1037719, PRJNA1237537, PRJNA1299309, and study SAMN40876533; A. thaliana ONT data were specifically retrieved from ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436642/Nanopore_Kz-9.fastq.gz.

2. PacBio HiFi long -read data
HG002 HiFi (~30x) long read data were obtained from the Genome in a Bottle (GIAB) FTP repository ([ftp trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/alignment/HG002.Sequel.15kb.pbmm2.hs37d5.whatshap.haplotag.RTG.10x.trio.bam](ftp.trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/alignment/HG002.Sequel.15kb.pbmm2.hs37d5.whatshap.haplotag.RTG.10x.trio.bam)), provided as aligned BAM files. PacBio HiFi reads for 35 HGSVC3 samples were downloaded from the IGSR (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/working/20250307_Alignments_HiFi_T2T/). HiFi reads for M. fascicularis, M. musculus, D. melanogaster, D. rerio and A. thaliana were retrieved from the SRA/ENA under BioProjects PRJNA1037719, PRJNA1237537, PRJNA1299309, and study SAMN40876533; A. thaliana HiFi data were obtained from the Genome Sequence Archive (<https://ngdc.cncb.ac.cn/gsa/browse/CRA008584/CRR591671>).

3. Illumina short read sequencing data

HG002 60x HiSeq data were obtained from the GIAB FTP repository (NCBI). HG002 35x PCR free NovaSeq FASTQ files were downloaded from the NIST Public Data Repository. Whole genome sequencing data from the 1000 Genomes Project Phase 4 (3,202 samples) were downloaded from the IGSF FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/). Illumina short read data for *M. fascicularis*, *M. musculus* and *A. thaliana* were acquired from the SRA/ENA under BioProject PRJNA1037719, study SAMN40876533, and runs ERR11436063 (R1 and R2; ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436063/Illumina_Kz-9_R1.fastq.gz and ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436063/Illumina_Kz-9_R2.fastq.gz), respectively. DNBSEQ short read data for *D. melanogaster* were obtained from the SRA under BioProject PRJNA1237537.

4. Reference genomes and Telomere to Telomere (T2T) assemblies

Forty seven Human Pangenome Reference Consortium (HPRC) assemblies were obtained from the HPRC public S3 bucket (AWS S3, s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/). The HPRC pangenome reference (minigraph cactus graph, v1.1) was downloaded from s3-us-west-2.amazonaws.com/human-pangenomics/pangenomes/freeze/freeze1/minigraph_cactus/hprc_v1.1_mc_grch38/. Reference genome assemblies used were: *M. fascicularis* (GCA_011100615.1), *M. musculus* (GCF_000001635.27), *D. melanogaster* (GCF_000001215.4), *D. rerio* (GCF_049306965.1) and *A. thaliana* (GCF_000001735.4), all obtained from NCBI Datasets. T2T assemblies were obtained for *M. fascicularis* (GCF_037993035.2), *D. melanogaster* (from BioProject PRJNA1237537), *D. rerio* (GCA_052040795.1), and *M. musculus* (https://github.com/yulab-q/mhaESC_genome/releases/download/upd_rmvector/mouse.241018.v1.1.0.combined.fasta.gz). For *A. thaliana*, 32 ecotype T2T assemblies were downloaded from Figshare (<https://figshare.com/articles/dataset/21673895>).

5. Public variant callset

The DRAGEN official structural variant callset for HG002 was obtained from the AWS S3 bucket (s3://1000genomes-dragen-v4-4-7/giab/individuals/hg38_alt_masked.cnv.graph.hla.methyl_cg.rna-11-r5.0-2/HG002/HG002.sv.vcf.gz).

Data analysis

All tools used in this study are listed in Supplementary Table 2 and list as below:

1. BWA-MEM v0.7.15-r1140 (<https://github.com/lh3/bwa>) was used for short-read sequencing (SRS) data alignment, while pbmm2 v1.17.0 (<https://github.com/PacificBiosciences/pbmm2>) and minimap2 v2.17-r941 (<https://github.com/lh3/minimap2>) were used for long-read sequencing (LRS) data alignment.
2. Manta v1.6.0 (<https://github.com/Illumina/manta>), Delly v1.1.6 (<https://github.com/dellytools/delly>), Lumpy v0.2.13 (<https://github.com/arq5x/lumpy-sv>), GATK-SV (<https://github.com/broadinstitute/gatk-sv>), DRAGEN v4.4 (https://github.com/srbehera/DRAGEN_Analysis) and Blend-Seq (https://github.com/broadinstitute/blend_seq_paper) were used for short-read based SV calling.
3. Sniffles2 v2.2 (<https://github.com/fritzsiedlazeck/Sniffles>), SVDS2 v2.0.0 (<https://github.com/Parsoa/SVDSS>), cuteSV2 v2.1.1 (<https://github.com/tjiangHIT/cuteSV>), sawfish v0.12.7 (<https://github.com/PacificBiosciences/sawfish>) and SVision-pro v1.8 (<https://github.com/songbowang125/SVision-pro>) were used for long-read based SV calling.
4. Truvari v5.4.0 (<https://github.com/ACEnglish/truvari>) was used for variant benchmarking.
5. AnnotSV v3.0.9 (<https://github.com/lgmgeo/AnnotSV>), TRF v4.10.0-rc.2 (<https://github.com/Benson-Genomics-Lab/TRF>) and RepeatMasker v4.1.2-p1 (<https://github.com/Dfam-consortium/RepeatMasker>) were used for variant annotation and repeat analysis.
6. samtools v1.1 (<https://github.com/samtools/samtools>), bedtools v2.27.1 (<https://github.com/arq5x/bedtools2>), bcftools v1.10.2 (<https://github.com/samtools/bcftools>), bgzip v1.17 and tabix v1.17 (<https://github.com/samtools/htslib>) were used for file processing and manipulation.

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](#)

Data availability:

All datasets used in this study are listed in Supplementary Table 1 and summarized below:

1. HG002 HiFi 30X (PacBio HiFi reads): https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_CCS_15kb/alignment/HG002.Sequel.15kb.pbmm2.hs37d5.whatshap.haplotag.RTG.10x.trio.bam
2. HG002 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG002/PAW70337/calls.sorted.bam (download by aws s3 cp)
3. HG003 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG003/PAY87794/calls.sorted.bam
4. HG004 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG004/PAY87778/calls.sorted.bam
5. HG005 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG005/PAW87816/calls.sorted.bam
6. HG006 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG006/PAY77227/calls.sorted.bam
7. HG007 ONT (Nanopore long reads): s3://ont-open-data/giab_2025.01/basecalling/sup/HG007/PAY12990/calls.sorted.bam
8. 557 ONT samples from CHN100K (Nanopore long reads): <https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA012590>
9. HG002 SRS 60X (Illumina short reads): https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_HiSeq_HG002_Homogeneity-10953946/NHGRI_Illumina300X_AJtrio_novalign_bams/HG002.hs37d5.60x.1.bam
10. HG002 SRS 35X (Illumina short reads): https://opendata.nist.gov/pdrsv/mds2-2336/input_fastqs/HG002.novaseq.pcr-free.35x.R1.fastq.gz
11. 47 HPRC samples (T2T assemblies): s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/
12. HPRC callset (HPRC sample VCF): [s3://human-pangenomics/working/HPRC/*/assemblies/year1_freeze_assembly_v2.1/assembly_qc/dipcall_v0.2/](https://s3-us-west-2.amazonaws.com/human-pangenomics/working/HPRC/*/assemblies/year1_freeze_assembly_v2.1/assembly_qc/dipcall_v0.2/)
13. HPRC pan-reference (HPRC Pangenome Resources): <https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=pangenomes/freeze1/minigraph-cactus/hprc-v1.1-mc-grch38/>
14. 1KGP phase4 whole genome sequencing (3202 samples, Illumina short reads): https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/data/
15. 35 HGVC3 HiFi samples (PacBio HiFi reads): https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGVC3/working/20250307_Alignments_HiFi_T2T/

16. 35 HGSC3 ONT samples (Nanopore long reads): https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSC3/working/20240816_JAX_ONT_guppy6_Rebasecalled/

17. Macaca fascicularis reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_011100615.1/

18. Macaca fascicularis T2T assembly: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_037993035.2/

19. Macaca fascicularis sequencing reads (PacBio HiFi / Nanopore / Illumina): https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=1037719

20. Mus musculus reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001635.27/

21. Mus musculus T2T assembly: https://github.com/yulab-ql/mhaESC_genome/releases/download/upd_rmvector/mouse.241018.v1.1.0.combined.fasta.gz

22. Mus musculus sequencing reads (PacBio HiFi / Nanopore / Illumina): https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SAMN40876533&o=acc_s%3Aa

23. Drosophila melanogaster reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001215.4/

24. Drosophila melanogaster T2T assembly: <https://www.ncbi.nlm.nih.gov/datasets/genome/?bioproject=PRJNA1237537>

25. Drosophila melanogaster sequencing reads (PacBio HiFi / Nanopore / DNBSEQ): https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=1237537

26. Danio rerio reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_049306965.1/

27. Danio rerio T2T assembly: https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_052040795.1/

28. Danio rerio sequencing reads (PacBio HiFi / Nanopore): https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&from_uid=1299309

29. Arabidopsis thaliana reference genome: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001735.4/

30. Arabidopsis thaliana T2T assemblies (32 ecotypes): https://figshare.com/articles/dataset/32_ecotypes_Arabidopsis_thaliana_genomes_gene_annotation_pan-TE_library_graph_pan-genome_gene_family_and_gene_presence_absence_matrices_files_/21673895

31. Arabidopsis thaliana ONT sequencing reads (Nanopore long reads): ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436642/Nanopore_Kz-9.fastq.gz

32. Arabidopsis thaliana HiFi sequencing reads (PacBio HiFi long reads): <https://ngdc.cncb.ac.cn/gsa/browse/CRA008584/CRR591671>

33. Arabidopsis thaliana Illumina sequencing reads (Illumina short reads): ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436063/Illumina_Kz-9_R1.fastq.gz; ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR114/ERR11436063/Illumina_Kz-9_R2.fastq.gz

34. DRAGEN official callset of HG002 (DRAGEN VCF): s3://1000genomes-dragen-v4-4-7/giab/individuals/hg38-alt_masked.cnv.graph.hla.methyl_cg.rna-11-r5.0-2/HG002/HG002.sv.vcf.gz

35. NIST HG002 Q100 benchmark (V1.1) (Ground truth): https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.018-20240716/

36. CMRG benchmark (Genomic regions): https://ftp.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/CMRG_v1.00/GRCh38/StructuralVariant/

37. RepeatMasker regions (Genomic regions): <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.out.gz>

38. EASY and HARD regions for short reads (Genomic regions): <https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/GRCh38@all/Union/>

39. GENCODE annotation (Annotation): ftp://ftp.ebi.ac.uk/pub/databases/genencode/Gencode_human/release_44/genencode.v44.basic.annotation.gtf.gz

40. Regulatory features (Annotation): https://ftp.ensembl.org/pub/current_regulation/homo_sapiens/GRCh38/annotation/Homo_sapiens.GRCh38.regulatory_features.v115.gff3.gz

41. RepeatMasker annotation (Annotation): <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.out.gz>

42. Tandem repeat annotation (Annotation): <https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.trf.bed.gz>

All variant callsets generated in this study are publicly available at the HitSV_call_results repository (https://github.com/hitbc/HitSV_call_results) and have been archived on Zenodo (DOI: 10.5281/zenodo.22106614).

Code availability:

HitSV is implemented in C/C++. The source code (v2.0.4), executable binaries, auxiliary scripts, and precomputed resources are publicly available at <https://github.com/hitbc/HitSV> under the GNU General Public License v3.0.

All other data supporting the findings of this study are available from the corresponding author upon reasonable request.

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	Not relevant to our study, which is a computational benchmarking study. No sex- or gender-specific analyses were performed.
Reporting on race, ethnicity, or other socially relevant groupings	Not relevant to our study. Data used in this study are mainly focus on GIAB, 1000 genome project and CHN100K which are well reported. No new race or ethnicity classifications were generated in this study.
Population characteristics	Not relevant to our study. Data used in this study are mainly focus on GIAB, 1000 genome project and CHN100K which are well reported. No additional demographic information was collected or analyzed in this study.
Recruitment	No new participants were recruited for this study.
Ethics oversight	Not relevant to our study. Data used in this study are mainly focus on GIAB, 1000 genome project and CHN100K which are well reported and fully consented. No additional ethical approval was required for this secondary analysis.

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

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined by the availability of publicly released benchmark and population-scale sequencing datasets, including GIAB samples (HG002–HG007), 557 ONT samples, 47 HPRC assemblies, 35 HGSVC3 samples, and 3,202 samples from the 1000 Genomes Project Phase 4.
Data exclusions	No data were excluded in this study.
Replication	Reproducibility was assessed through benchmarking against established reference callsets (e.g., GIAB benchmark regions), cross-platform comparisons (short-read, long-read, and hybrid data) and large population-scale cohorts (1000 genome project and CHN100K).
Randomization	Randomization was not applicable to this study. All data were derived from publicly available sequencing datasets, and no experimental allocation of samples to treatment or control groups was performed.
Blinding	Blinding was not applicable. The study involved computational analysis of publicly available genomic data, and no subjective outcome assessment or experimental group allocation was performed.

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

Methods

n/a	Involved in the study	n/a	Involved in the study
☒	☐ Antibodies	☒	☐ ChIP-seq
☒	☐ Eukaryotic cell lines	☒	☐ Flow cytometry
☒	☐ Palaeontology and archaeology	☒	☐ MRI-based neuroimaging
☒	☐ Animals and other organisms		
☒	☐ Clinical data		
☒	☐ Dual use research of concern		
☒	☐ Plants		

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable