nature portfolio

Corresponding author(s): Jian Yang

Last updated by author(s): Jan 24, 2026

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	No data-collection software was used.
Data analysis	The analysis scripts used in this study: https://github.com/JianYang-Lab/1KCP-analysis PIGA: https://github.com/JianYang-Lab/PIGA CCS (v6.2.0): https://github.com/PacificBiosciences/ccs actc (v0.6.0): https://github.com/PacificBiosciences/actc DeepConsensus (v1.2.0): https://github.com/google/deepconsensus Hifiasm (v0.15.3): https://github.com/chhylp123/hifiasm minimap2 (v2.24): https://github.com/lh3/minimap2 Unicycler (v0.5.0): https://github.com/rrwick/Unicycler BEDTools (v2.30.0): https://github.com/arq5x/bedtools2 blastn (v2.13.0): https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST Meryl (v1.4.1): https://github.com/marbl/meryl Merqury (v1.3): https://github.com/marbl/merqury Inspector (v1.0.1): https://github.com/Maggi-Chen/Inspector Flagger (v0.2.0): https://github.com/mobinasri/flagger rustybam (v0.1.34): https://github.com/mrvollger/rustybam Dipcall (v0.3): https://github.com/lh3/dipcall BWA (v0.7.17): https://github.com/lh3/bwa GATK (v4.2.6.1): https://github.com/broadinstitute/gatk DeepVariant (v1.4.0): https://github.com/google/deepvariant

hap.py (v0.3.14): https://github.com/Illumina/hap.py
Truvari (v4.2.2): https://github.com/ACEnglish/truvari
WhatsHap (v1.4): https://github.com/whatshap/whatshap
Tandem Repeat Finder (v4.09): https://tandem.bu.edu/trf/trf.html
RepeatMasker (v4.1.1): https://www.repeatmasker.org/
Liftoff (v1.6.3): https://github.com/agshumate/Liftoff
HISAT2 (v2.2.1): https://github.com/DaehwanKimLab/hisat2
StringTie (v2.2.1): https://github.com/gpertea/stringtie
gffcompare (v0.12.9): https://github.com/gpertea/gffcompare
CPC2 (v1.0.1): https://github.com/gao-lab/CPC2_standalone
GeneMarkS-T (v5.1): https://genemark.bme.gatech.edu/GeneMark
InterProScan (v5.66): https://www.ebi.ac.uk/interpro
SEI: https://github.com/FunctionLab/sei-framework
Enformer (v0.8.8): https://github.com/lucidrains/enformer-pytorch
MASC3 (v3.0.0): https://github.com/macs3-project/MACS
Minigraph (v0.21): https://github.com/lh3/minigraph
Mash (v2.3): https://github.com/marbl/Mash
gaffilter: https://github.com/ComparativeGenomicsToolkit/cactus-gfa-tools
odgi (v0.8.6): https://github.com/pangenome/odgi
seqwish (v0.7.4): https://github.com/ekg/seqwish
smoothxg (v0.7.9): https://github.com/pangenome/smoothxg
GFAffix (v0.1.5): https://github.com/marschall-lab/GFAffix
GenomeScope (v2.0.1): https://github.com/tbenavi1/genomescope2.0
dna-brnn (v0.1): https://github.com/lh3/dna-nn
VG (v1.59.0): https://github.com/vgteam/vg
vcfbub (v0.1.0): https://github.com/pangenome/vcfbub
vamos (v2.1.5): https://github.com/ChaissonLab/vamos
BCFtools (v1.20): https://github.com/samtools/bcftools
GraphAligner (v1.0.20): https://github.com/maickrau/GraphAligner
Sniffles2 (v2.6.2): https://github.com/fritzsedlazeck/Sniffles
SURVIVOR (v1.0.6): https://github.com/fritzsedlazeck/SURVIVOR
Jasmine (v1.1.5): https://github.com/mkirsche/Jasmine
Manta (v1.6.0): https://github.com/Illumina/manta
GangSTR (v2.5): https://github.com/gymreklab/GangSTR
vcfwave (v1.0.14): https://github.com/vcflib/vcflib
edlib (v1.2.7): https://github.com/Martinsos/edlib
MeLoDe: https://github.com/JavenCao/MeLoDe
miniprot (v0.12): https://github.com/lh3/miniprot
pangene (v1.1): https://github.com/lh3/pangene
clusterProfiler (v4.12.6): https://github.com/YuLab-SMU/clusterProfiler
GOSemsim (v2.30.2): https://github.com/YuLab-SMU/GOSemSim
Bandage (v0.9.0): https://github.com/rrwick/Bandage
PLINK2 (v2.0.0): https://www.cog-genomics.org/plink/2.0
HiFiHLA (v0.3.1): https://github.com/PacificBiosciences/hifihla
HLA-HD (v1.7.1): https://w3.genome.med.kyoto-u.ac.jp/HLA-HD
eLD (v1.0): https://www.sg.med.osaka-u.ac.jp/tools.html
STAR (v2.7.9): https://github.com/alexdobin/STAR
RNA-SeQC (v2.4.2): https://github.com/getzlab/rnaseqc
edgeR (v3.22.5): https://bioconductor.org/packages/release/bioc/html/edgeR.html
PEER: https://github.com/PMBio/peer
tensorQTL (v1.0.9): https://github.com/broadinstitute/tensorqtl
GCTA (v1.94.1): https://yanglab.westlake.edu.cn/software/gcta
dgrm: https://github.com/PeixiongYuan/dgrm
coloc (v5.2.3): https://github.com/chr1swallace/coloc
SMR (v1.3.1): https://yanglab.westlake.edu.cn/software/smr
Minimac4 (v4.1.6): https://genome.sph.umich.edu/wiki/Minimac4

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

The release of the 1KCP data is approved by the Human Genetic Resources Administration of China (permission numbers 2025BAT00888 and 2025BAT00720). The summary-level data, including the 1KCP pangenome graph, pangenome annotations, variant sites, and eQTL summary statistics, are available at the 1KCP data portal (https://yanglab.westlake.edu.cn/1kcp). The individual-level data, including raw sequencing reads, genome assemblies, and variant genotypes, are available in the repositories of the National Genomics Data Center (NGDC), specifically in the Genome Sequence Archive for Human (GSA-Human: https://ngdc.cncb.ac.cn/gsa-human), Genome Warehouse (GWH: https://ngdc.cncb.ac.cn/gwh), and Genome Variation Map (GVM: https://ngdc.cncb.ac.cn/gvm), under BioProject accession numbers PRJCA036466 and PRJCA039156. Public datasets used in this study include: GRCh38 genome (https://gdc.cancer.gov/about-data/gdc-data-

processing/gdc-reference-files); CHM13 genome (https://github.com/marbl/CHM13); HPRC v1 assemblies (https://github.com/human-pangenomics/HPP_Yead_Assemblies); CPC assemblies (https://github.com/Shuhua-Group/Chinese-Pangenome-Consortium-Phase-I); GENCODE annotation (https://www.gencodegenes.org/human); NCBI RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release); 1000G high-coverage Illumina variants (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV); 1000G ONT SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1KG_ONT_VIENNA/release/v1.0); GnomAD variants (http://www.gnomad-sg.org); HGSVC3 SVs (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC3/release/Variant_Calls/1.0); OMIM database (https://www.omim.org); COSMIC database (https://cancer.sanger.ac.uk/cosmic); STRchive database (https://strchive.org/resources); IPD-IMGT/HLA database (https://www.ebi.ac.uk/ipd/imgt/hla); GWAS catalog (https://www.ebi.ac.uk/gwas); BioBank Japan GWAS (https://humandbs.dbcls.jp/en/hum0197-v23); EnsembleTR panel (https://github.com/gymrek-lab/EnsembleTR).

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 was recorded by self-reporting during sample collection and confirmed by calculating the coverage depth ratio of X and Y chromosomes from WGS data. The final cohort comprised 737 males and 642 females. Sex is considered a biological variable in this study. No sex-based exclusion criteria were applied.
Reporting on race, ethnicity, or other socially relevant groupings	The first phase of the 1KCP project enrolled 1,379 Chinese participants in Wenzhou, China.
Population characteristics	The cohort consists of 1,379 individuals ranging in age from 18 to 76 years. The cohort represents a general population sample derived from individuals undergoing routine health examinations, rather than a disease-specific case-control cohort; therefore, participants were not selected based on past or current diagnostic categories. All samples underwent Illumina short-read WGS and RNA-seq. Among these participants, 55 samples were selected for high-coverage PacBio HiFi long-read WGS and Hi-C sequencing, and 1,099 were selected for modest-coverage PacBio long-read WGS.
Recruitment	Participants were recruited from volunteers undergoing routine health examinations at the Health Examination Center of the Second Affiliated Hospital of Wenzhou Medical University. While there is a potential self-selection bias towards individuals with higher health awareness, germline genomic architecture is independent of such behavioral traits, and thus the results regarding variant detection remain robust.
Ethics oversight	Informed consent was obtained from all participants. The study was approved by the Ethics Committee of Westlake University (approval no. 20201127YJ001) and the Ethics Committee of Wenzhou Medical University (approval no. 2019-107).

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	In this study, we analyzed 1,144 Chinese individuals from the 1KCP project who had both short-read and long-read whole-genome sequencing (WGS) data. The sample size was determined to capture the vast majority of common and low-frequency variants, and the cohort represents the largest long-read datasets for the Chinese population to date.
Data exclusions	During data preprocessing, 28 samples were excluded due to mismatched short-read and long-read WGS data.
Replication	We selected 10 samples subjected to both modest-coverage and high-coverage PacBio HiFi long-read WGS. Among these, 3 samples were utilized to compare the assembly quality between PIGA and Hifiasm assemblies. The variant detection performance demonstrated that the vast majority of variants were reproducibly identified across different assemblies of the same samples. Variants in extremely repetitive regions could not be reliably resolved across different assemblies due to their complex graph structure and the inherent noise of the long-read data used in the study.
Randomization	Randomization is not required for this study because all experiments were conducted under identical conditions.
Blinding	Blinding is not required for this study because all experiments were conducted under identical conditions.

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.