nature research

Corresponding author(s): Joseph G. Gleeson, Xiaoxu Yang

Last updated by author(s): Sep 30, 2022

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 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

Visualization of mosaic variants was based on Python (v3.7.81) packages Pysam (v0.11.2.2, https://github.com/pysam-developers/pysam) and NumPy (v1.16.2, https://numpy.org/), The input for this visualization is short-read sequencing data in the format of a BAM file, processed with a GATK (v3.8.1) best-practice pipeline (with insertion/deletion, or INDEL, realignment, followed by base quality score recalibration). Codes available as https://github.com/Virginiaxu/DeepMosaic/blob/master/deepmosaic/featureExtraction.py.
Reads from deep amplicon sequencing were mapped to the GRCH37d5 reference genome by BWA mem and processed according to GATK (v3.8.2) best practices without removing PCR duplicates. Putative mosaic sites were retrieved using SAMtools (v1.9) mpileup and pileup filtering scripts described in previous TAS pipelines.
SimData1: Pysim was then used to simulate paired-end sequencing reads.
SimData2: Pysim was then used to simulate paired-end sequencing reads.
SimData3: BAMSurgeon (updated 24 Dec 2020) was used to generate this simulation dataset.
BioData4: Reads were aligned to the GRCh37d5 genome with BWA (v0.7.15) mem and duplicates were removed with sambamba (v0.6.6) and base quality recalibrated by GATK (v3.5.0).
BioData5: Reads were aligned to the GRCh37d5 genome with BWA(v0.7.17) mem and duplicates were removed and base quality recalibrated by GATK (v4.0.4) according to the established best-practice pipeline.
BioData6: Fastq files were generated using Picard SAMTOFASTQ and aligned to GRCh37d5 genome with BWA (v0.7.17) mem. Duplicates were removed, reads near INDEL regions were realigned, and base quality scores were recalibrated with GATK v3.8.1 and picard v2.20.7.

Data analysis

DeepMosaic is implemented in Python (3.7.81); the code, documentation and demos are available at https://github.com/Virginiaxu/DeepMosaic. Data analysis is described in detail in the methods, and codes used to produce the data are also provided in https://github.com/shishenyxx/Adult_brain_somatic_mosaicism and https://github.com/shishenyxx/Sperm_control_cohort_mosaicism. Specifically we used the following software for data analysis: BWA v0.7.17, GATK 3.5.0, and v3.8.1 and v3.8.2, Strelka2 v2.9.2, Mutect2 from GATK v4.0.4, MosaicForecast v8-13-2019, MosaicHunter v1.0.0, Pysim, SAMtools v1.9, BEDTools v2.27.1, efficientnet_pytorch v0.6.1.
SimData1: FASTQ files were aligned to the GRCh37d5 human reference genome with BWA (v0.7.17) mem command. Aligned data were processed by GATK (v3.8.1) and Picard (v2.18.27) for marking duplicates, sorting, INDEL realignment, base quality recalibration, and germline

variant calling.
SimData2: A total of 439,200 different variants were generated. FASTQ files were aligned and processed with BWA (v0.7.17), SAMtools (v1.9), and Picard (v2.18.27). The data were subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both paired mode and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with different models trained for different read depth (250x model for depth≥300x).
SimData3: Bam files with and without simulated data were downsampled to 500x, 400x, 300x, 200x, 100x, and 50x. The data were subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both paired mode, and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with different models trained for different read depth (250x model for depth≥300x).
BioData4: Processed BAM files were subjected to DeepMosaic as well as MuTect2 (GATK v4.0.4, both paired mode and single mode), Strelka2 (v2.9.2), MosaicHunter (v1.0.0), and MosaicForecast (v8-13-2019) with 200x models trained for the specific depth.
BioData5: Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v4.0.4) single mode as well as GATK (v4.0.4) Haplotypecaller ("polidy" 50) and previously established filters.
BioData6: Processed BAM files were subjected to the DeepMosaic pipeline followed by MuTect2 (GATK v4.0.4) single mode, then the final call set was compared with the TCGA-MC3 call set detected by MuSE(PMID: 27557938), MuTect(PMID: 23396013), SomaticSniper(PMID: 22155872), VarScan2 (PMID: 22300766), and Radia(PMID: 25405470) using the publicly released gold standard (https://gdc.cancer.gov/about-data/publications/mc3-2017) from the same dataset.
Variant processing software includes Picard v2.18.27, BCFtools v1.10.32, sambamba v0.6.6, iFish, Define. Plotting and visulization software include R v3.5.1, ggplot2 v3.3.1, Rcpp v1.03, PyTorch v 1.6.0, pysam v0.11.2.2, Python v3.7.1 and v3.7.81, SciPy v1.3.1, pandas v0.24.2, matplotlib v3.1.1, numpy v1.16.2, and seaborn v0.9.0.

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 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

WGS data used to generate the training set are available at the Sequence Read Archive (SRA, Accession No. SRP028833 and SRP100797, BioData1). The gold standard WGS data and validated capstone project data are available at the National Institute of Mental Health Data Archive (NIMH Data Archive ID 792 and 919: https://nda.nih.gov/study.html?id=792, BioData2, and https://nda.nih.gov/study.html?id=919 BioData3) and the Brain Somatic Mosaicism Consortium Data Portal, independent benchmark brain genotyping is also part of the SRA accession PRJNA736951 (BioData3). Simulated data generated from NA24385 (HG002) are available at https://humanpangenome.org/hg002/. The independent sperm and blood deep WGS data are available at SRA (Accession No. PRJNA588332 and PRJNA660493, BioData4). Independent WES data from brain, blood, and saliva samples were available in NIMH Data Archive under study number 1484 (https://nda.nih.gov/study.html?id=1484, BioData5). TCGA-MC3 data are available on the GDC portal (https://portal.gdc.cancer.gov/, sample IDs provided with variants in Supplementary Table 3).

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	Eight different group of biological data and experimentally generated data were used in this study to justify the performance of the same software for biological and simulated data. Size of training data in each epoch was determined by the processing capacity of computational nodes we used. Total number of variants used for training were based on all available training data and the all possible combinations of depth and expected allelic fractions. As described in the main text, 180,000 variants were used in training and 20,265 for validation (BioData1 and SimData1), 400 variants were used for model evaluation (BioData2), 619,740 simulated variants were used for benchmark (SimData2 and SimData3). 40,848 variants from BioData1, BioData3, SimData1, and SimData2 were used for the additional model evaluation. Sixteen samples sequenced at 300x WGS were used for biological benchmark (BioData4) for DeepMosaic. One-hundred and eighty one samples sequenced at 300x WES were used as independent biological benchmark for non-cancer WES (BioData5). Data from the TCGA-MC3 collection were collected for 2430 samples from 1215 individual were used as independent biological benchmark for cancer WES (BioData6).
Data exclusions	We included all available data for this study.
Replication	1 simulation dataset and 2 biological datasets are included in the model training, 2 simulation and 4 biological datasets are included for the validation of performance, thus we did not include further complete biological replication experiments.
Randomization	No experimental groups were allocated by the scientists. Training and validation data were chosen as described in the Methods.
Blinding	Randomization and cross-validation in model training was designed in the codes, public available data are used for model training and evaluation, no further experimental blinding was assigned by the researchers.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The participants of BioData4 were previously recruited and the study included males at different ages (Breuss et al. 2020). The participants of BioData5 were patients with focal cortical dysplasia and underwent surgical treatment affected brain samples after surgery as well as control saliva or blood samples were collected. The participants of BioData6 were de-identified according to GDC regulations.
Recruitment	For BioData4, the procedure followed the Human Research Protections Programs at University of California, San Diego approval #161151. All recruited males are healthy but with at least one child diagnosed with ASD; see also Breuss et al. 2020. For BioData5, the procedure followed the Human Research Protections Programs at University of California, San Diego approval #140028. For BioData6, the data were de-identified according to GDC regulations.
Ethics oversight	IRB at UC, San Diego

Note that full information on the approval of the study protocol must also be provided in the manuscript.