nature research

Corresponding author(s): Andrew Carroll

Last updated by author(s): July 20, 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	Data generated for this study was produced by PacBio instrument sequencing an analysis with pbccs v4.2.0 (https://github.com/PacificBiosciences/ccs)
Data analysis	Full commands and versions for all programs run are found in the Software Commands section of supplementary material

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

Sequencing data, predictions, and analysis files are available at:
https://console.cloud.google.com/storage/browser/brain-genomics-public/research/deepconsensus/publication

Sequencing data is available from the following sources:
Sequel II data from Novogene : https://console.cloud.google.com/storage/browser/brain-genomics-public/research/sequencing
15kb HG002 and 24kb HG002 reads from PacBio: https://console.cloud.google.com/storage/browser/brain-genomics-public/research/deepconsensus/publication/sequencing

Accession identifiers for non-human PacBio SMRT sequencing:
Rana muscosa: SRR11606868, Mus musculus: SRR11606870, Zea mays: SRR11606869

Sequel II data from PacBio: https://downloads.pacbcloud.com/public/dataset/HG002_SV_and_SNV_CCS/
HG002 diploid assembly:
https://obj.umiacs.umd.edu/marbl_publications/hicanu/hg002_hifi_hicanu_combined.fasta.gz

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	All available sequence datasets were used for HG002-HG007. (an initial HG002 PacBio sequencing run from earlier publications), 3 flowcells of new, long insert HG002 was provided by PacBio. We contracted with Novogene for 3 flowcells each of HG003, HG004, HG006, and HG007 in an earlier study (described in: https://www.biorxiv.org/content/10.1101/2020.12.11.422022v1).
Data exclusions	A single flowcell of HG007 was excluded from analysis due to a file corruption issue in the file received from the sequencing vendor. The file corruption issue prevented all downstream analysis from this single flowcell.
Replication	Results were evaluated across the full genome for every available human sample not used in model training (HG003, HG004, HG006, HG007) with concordant findings for genome assembly and variant calling. Results were evaluated across three non human species for which PacBio sequencing data was publicly available at the subread level (mouse, frog, and maize) with concordant findings for genome assembly.
Randomization	Randomization was not relevant for this study. The machine learning training followed standard practices for train-tune-and test data sets. Training is only conducted with Sequel II, Chemistry V1 of HG002. All other samples evaluated (HG003, HG004, HG006, and HG007) were never trained on.
Blinding	Investigators were not blind to groups, as all data was pooled together and publicly available. The machine learning training followed standard practices for separating train, tune, and test data sets.

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HG002-HG007 cell lines from the Coriell Institute
Authentication	The full genome of the cell lines were sequenced and aligned back to a truth set for
Mycoplasma contamination	The cell lines were not tested for Mycoplasma contamination. However, only the germline DNA content of the cell lines are required, not any transcriptional or other cell phenotype.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.