nature portfolio

NCOMMS-23-59396A
Corresponding author(s): Kun QU
qukun@ustc.edu.cn
Last updated by author(s): Sep 19, 2024

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 software was used for data collection.
Data analysis	All original code has been deposited at Github (https://github.com/QuKunLab/eccDNABenchmarking). Other code or softwares used for analyzing the data was reported as follow:(name (link)) AmpliconArchitect (https://github.com/virajbdeshpande/AmpliconArchitect) ART(http://www.niehs.nih.gov/research/resources/software/art) Circle_finder(https://github.com/pk7zuva/Circle_finder) Circle-Map (https://github.com/iprada/Circle-Map) Control-FREEC (https://github.com/BoevaLab/FREEC) CReSIL (https://github.com/visanuwan/cresil) ecc_finder (https://github.com/njaupan/ecc_finder) eccDNA_RCA_nanopore (https://github.com/YiZhang-lab/eccDNA_RCA_nanopore) Eccsplorer (https://github.com/crimBubble/ECCsplorer) MUMmer3 (https://github.com/marbl/MUMmer3) NanoCircle (https://github.com/RAHenriksen/NanoCircle) PBSIM2 (https://github.com/yukiteruono/pbsim2) Prism 9 (https://www.graphpad.com) Pysam (https://github.com/pysam-developers/pysam) Python 3.9.13 (https://www.python.org/) Scipy 1.10.1 (https://scipy.org/citing-scipy/)

Seaborn 0.12.2 (https://seaborn.pydata.org/#)
SnapGene Viewer (https://www.snapgene.com/snapgene-viewer)

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 raw sequencing data (WGS-SR, WGS-LR, ATAC-Seq-SR, 3SEP-SR, 3SEP-LR, Circle-Seq-SR, and Circle-Seq-LR) generated in this study have been deposited in the Genome Sequence Archive for Human(GSA-Human) database under accession code HRA006020 [https://ngdc.cncb.ac.cn/gsa-human/browse/HRA006020]. The public data used in this study in Supplementary Figure1 to generate our simulation datasets are openly available from following study.

Dataset 1 (sperm cells): the raw sequencing data are available in the Sequence Read Archive (SRA) database under accession code PRJNA655819 [https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA655819].

Dataset 2 (EJM cell line) and Dataset 3 (JJN3 cell line): the raw sequencing data are available in the Sequence Read Archive (SRA) database under accession code PRJNA806866 [https://www.ncbi.nlm.nih.gov/sra/?term=PRJNA806866].

Dataset 4 (Kelly cell line): the raw sequencing data are available in the European Nucleotide Archive (ENA) database under accession code PRJEB50518 [https://www.ebi.ac.uk/ena/browser/view/PRJEB50518].

Dataset 5 (medulloblastoma): the raw sequencing data are available in the Gene Expression Omnibus (GEO) repository under accession code GSE205178 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205178].

Dataset 6 (muscle cells): the raw sequencing data are available in the Sequence Read Archive (SRA) database under accession code SRR6315400 [https://www.ncbi.nlm.nih.gov/sra/?term=SRR6315400].

Dataset 7 (OVCAR8 cell line): the raw sequencing data are available in the Gene Expression Omnibus (GEO) repository under accession code GSE68644 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68644].

We reanalyzed the data of EJM, JJN3, Kelly cell line, and muscle cells with corresponding pipelines in Supplementary Figure1. We used processed data of sperm cells, medulloblastoma and OVCAR8 cell line in their original paper. The processed template files can be found at https://github.com/QuKunLab/eccDNABenchmarking/tree/main/ecsim/ecsim/resource/template. All other data supporting the findings described in this paper are available in the article and its Supplementary Information files. Source data for figures are also provided with this paper.

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	This is not applicable for this study.
Reporting on race, ethnicity, or other socially relevant groupings	This is not applicable for this study.
Population characteristics	This is not applicable for this study.
Recruitment	This is not applicable for this study.
Ethics oversight	This is not applicable for this study.

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	We performed 3 replicates for each experimental method to compare their performance, which is the minimal sample size for statistical analysis. For comparing the analysis pipeline performance, we included up to 7 simulated datasets which is more than the minimal requirement for statistical analysis.
Data exclusions	No data was excluded from the study.
Replication	We performed three replicate experiments to perform the statistical analysis. All the results were found from the blinding assay.
Randomization	This is not relevant to our study because we used all the sequencing data or generated simulated data.
Blinding	In the comparison, all the analysis pipelines were blinded to the simulated data, and the outputs from the pipelines were then compared to the simulated ground truth. For experimental methods comparison, we are blind to all experimental methods and perform the experimental methods using the DNA materials from the same pool.

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HeLa
Authentication	STR analysis was done for HeLa cell line by BeNa Culture Collection Company, and the cell line was authenticated as CVCL-0030
Mycoplasma contamination	Mycoplasma contamination test was performed by the BeNa Culture Collection Company, and cell line used in this study was not contaminated by Mycoplasma
Commonly misidentified lines (See ICLAC register)	NA

Plants

Seed stocks	This is not applicable for this study.
Novel plant genotypes	This is not applicable for this study.
Authentication	This is not applicable for this study.