nature research

Corresponding author(s): Jie Liu

Last updated by author(s): Jul 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	No software was used for data collection
Data analysis	Source code of EPCOT and trained EPCOT models are available in a GitHub repository https://github.com/liu-bioinfo-lab/EPCOT. For EPCOT model training and testing, we used pytorch (v1.10.1). For data analysis and visualization, we used the following software in python 3.9: einops (v0.3.2), numpy (v1.19.5), scipy (v1.7.3), scikit-learn (v1.0.2), pybigwig (v0.3.17), hic-straw (v1.0.0.1), TF-MoDISco (v0.5.16.0), pygenometracks (v3.6), pyliftover (v0.4), samtools (v1.7), and deeptools (v3.5.1).

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

All the genomic data used in EPCOT is in the reference genome version hg38. DNase-seq, human CAGE-seq, ChIP-seq and STARR-seq are all downloaded from ENCODE. Mouse CAGE-seq profiles are downloaded from FANTOM. RNA-seq gene expression level data is from REMC database. Hi-C, Micro-C, and ChIA-PET contact maps of cell types are downloaded from 4DN, and GTEx tissue Hi-C contact maps are downloaded from ENCODE. Generated TF sequence binding patterns along with their motif comparison results are provided in a web page that is available in our GitHub repository https://github.com/liu-bioinfo-lab/EPCOT.

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 used all the datasets that we can collect. To train the EPCOT pre-training model, we used 656 human ChIP-seq across four cell lines that had most abundant ChIP-seq profiles on ENCODE with available DNase-seq. The ChIP-seq of each transcription factor or histone mark was available in at least two of the four cell line to provide cell-type specific information for training. To train and test the downstream models, we used 8 human RNA-seq, 11 human CAGE-seq, 4 mouse CAGE-seq, 8 human Hi-C, 2 mouse Hi-C, 2 human Micro-C, and 5 human STARR-seq profiles, because these profiles with their associated cell/tissue type specific profiles were available on public datasets including 4DN, ENCODE, and GTEx.
Data exclusions	No data were excluded.
Replication	No experimental findings were disclosed, hence no replication was performed
Randomization	Randomization was not relevant for this study as it involved a reanalysis of published datasets
Blinding	Blinding was not applicable because no new data were collected

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