nature portfolio

Corresponding author(s): Luca Pinello

Last updated by author(s): Sep 21, 2025

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	We developed a script using Python 3.8, available at https://github.com/pinellolab/EPInformer, to acquire data from the ENCODE Encyclopedia Version 2, 4D Nucleome (4DN), and the Roadmap Epigenomics Project.
Data analysis	The ABC software (v1.1.2) for promoter-enhancer link analysis is available at https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction/tree/master. Data analysis is conducted using Python 3.8, while PyTorch 2.4 is utilized for training, validating, and testing deep learning models. Samtools (v1.20) indexes BAM files, and deepTools (v3.5.6) converts BAM files to bigWig format. PyGenomeTracks (v3.1.2) visualizes genome tracks from bigWig files, and SciPy (v1.14.1) performs correlation analysis. All code for data processing and analysis can be found at https://github.com/pinellolab/EPInformer.

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 genomic datasets analyzed during the current study are available in the ENCODE Project repository (https://www.encodeproject.org/) under the following accession codes: DNase-seq (K562: ENCFF425WDA, ENCFF205FNC; GM12878: ENCFF020WZB, ENCFF729UYK; H1: ENCFF761ZRE; HepG2: ENCFF691HJY; HUVEC: ENCFF091KTX; NHEK: ENCFF117RNM); H3K27ac ChIP-seq (K562: ENCFF600THN, ENCFF232RQF, ENCFF704LGA; GM12878: ENCFF269GKF, ENCFF201OHW; H1: ENCFF693IFG, ENCFF860ABR; HepG2: ENCFF745JCH, ENCFF862NDZ, ENCFF926NHE; HUVEC: ENCFF374DGO, ENCFF609TUB; NHEK: ENCFF051NTC, ENCFF770JWP); and reference Hi-C (ENCFF134PUN). Additional Hi-C contact matrices are available from the 4D Nucleome Data Portal (https://data.4dnucleome.org/) under accession codes 4DNFITUOMFUQ and 4DNFI1UEG1HD. CAGE data are available from FANTOM5 (https://fantom.gsc.riken.jp/5/sstar/Main_Page) under accession codes CNhs11250 and CNhs12333. RNA-seq expression profiles are available from the Roadmap Epigenomics Consortium (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz). The enhancer-gene linkage benchmarking datasets are available in the Engreitz Lab GitHub repositories (https://github.com/EngreitzLab/CRISPR_comparison and https://github.com/EngreitzLab/eQTLEnrichment).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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 collected 305,746 promoter-enhancer pairs for the K562 cell line and 325,999 for the GM12878 cell line, along with CAGE and RNA-seq data for 18,377 protein-coding genes, to develop gene expression prediction models. We collected 1,575 enhancer-gene links across 244 genes assayed by CRISPR perturbations from Engreitz Lab for benchmarking.
Data exclusions	Promoter-enhancer pairs located more than 100 kb from the transcription start site (TSS) of target genes were excluded from the analysis.
Replication	The training and validation of our models can be reproduced using the code at https://github.com/pinellolab/EPInformer
Randomization	The weights of models was randomly initialized before training.
Blinding	Blinding was not relevant to data collection as this study utilizes publicly available genomic datasets. During data analysis, traditional blinding was not applicable; however, to ensure unbiased evaluation and prevent data leakage, we implemented a strict 12-fold cross-chromosome validation strategy. In each fold, two chromosomes were allocated for testing and two for validation, while the remaining chromosomes were used for training; thus, the model was effectively blinded to the test and validation sets during the training phase.

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

Methods

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