nature portfolio

Corresponding author(s): Rongling Wu, Charles G. Danko, Zhong Wang

Last updated by author(s): Jan 18, 2026

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 or custom code was used for data collection; all datasets were obtained by manual download from public repositories.
Data analysis	Peak calling from TF ChIP-seq bam files was performed using MACS2 (version 2.2.9.1) with a q-value cutoff of 0.05. The peak calling results in BED format based on the hg38 reference genome were converted to hg19 using UCSC liftOver. Bedtools (v2.26.0) was used for processing BED files. RO-seq data were normalized using the reads per kilobase per million mapped reads (RPKM) method. DNA sequence, histone modification ChIP-seq, and RNA-seq data used in this study were not subjected to any additional processing. All prediction and training codes are available at https://github.com/zhichunlizzx/BioSeq2Seq. Pretrained model weights for the four subtasks, histone modification prediction, functional element identification, gene expression prediction, and TFBS prediction, can be downloaded from https://zenodo.org/records/18234973/files/BioSeq2Seq_model.zip?download=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 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

All datasets used in this study are publicly available from the sources described in the Methods section, with detailed infor- mation on the RO-seq, histone modification ChIP-seq, RNA-seq, and TF ChIP-seq datasets used in this study provided in Supplementary Tables S2, S3, S7, and S8, respectively. Quality control reports for the data used in this study are available for download at https://zenodo.org/records/18234973/files/BioSeq2Seq_QC.zip?download=1. The Supplementary Information includes additional figures, tables, methods, and results, all of which are provided in Supplementary_Information.pdf. The source data underlying all figures in the manuscript are provided in Source_data.zip. In addition, the model-predicted results have been deposited in Zenodo as follows: histone modification: [https://doi.org/10.5281/zenodo.18242398], functional element [https://doi.org/10.5281/zenodo.18242609], gene expression [https://doi.org/10.5281/zenodo.18243067], and TFBS [https://doi.org/10.5281/zenodo.18241447]. These data consist solely of model predictions generated from publicly available datasets and are provided to support reproducibility and reuse. All supplementary materials will be available upon publication.

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 study used publicly available datasets which include data across various sexes, but no new data were collected. The datasets provide aggregated data without individual identification or gender-related labels.
Reporting on race, ethnicity, or other socially relevant groupings	This study did not specifically collect or analyze data based on race, ethnicity, or other social groupings. The datasets used contain anonymized and aggregated information, and the focus was on genomic data analysis rather than demographic characteristics.
Population characteristics	The datasets used in this study include human cell lines and tissue samples from various sources, but no direct demographic information (e.g., race, ethnicity, age) was included in the analyses. The focus of the study was on genomic and epigenomic data.
Recruitment	No recruitment of human participants or biological materials occurred in this study. All data were obtained from publicly available repositories.
Ethics oversight	As this study used publicly available, anonymized data, no ethical approval or oversight was required. The data used were obtained from established repositories, such as ENCODE and GEO, which comply with ethical standards for the collection and sharing of genomic data.

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	For each of the four downstream tasks, the prediction target categories were first defined. Specifically, histone modification categories were selected to be consistent with those used in the dHIT study (PMID: 35273399); functional element categories were defined based on commonly studied functional element types; subcellular RNA-seq categories were determined by the subcellular types available in the corresponding database (GSE30567); and TFBSs were defined based on TF ChIP-seq datasets from ENCODE that were used in the maxATAC study (PMID: 36719906) and remain publicly available. Subsequently, for each task, datasets were collected from cell lines for which RO-seq data were available. Within a given cell line, not all predefined prediction categories were necessarily available due to the incompleteness of publicly available datasets. As a result, the final dataset composition for each task reflects both the predefined prediction categories and the availability of matched data within each cell line. The resulting datasets span multiple cell lines and regulatory contexts and are derived from established public resources, providing sufficient coverage and diversity to robustly evaluate model performance and support the study's conclusions.
Data exclusions	No data were excluded from the publicly available datasets used in this study, except for standard preprocessing steps such as filtering low-quality reads or missing values, as documented in the Methods section.
Replication	As the study is based solely on analysis of publicly available high-throughput sequencing datasets, no experimental replicates were performed.

Replication	However, replication is supported through the use of multiple datasets across different cell types and tasks, and all analysis pipelines are provided for reproducibility.
Randomization	Randomization is not applicable as this study did not involve experimental treatment groups. All analyses were computational and based on predefined datasets from public repositories.
Blinding	Blinding was not applicable because no experimental procedures or outcome assessments involving human judgment were performed. All analyses were computational and automated.

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)	All cell line data used in this study were obtained from publicly available datasets, primarily from the ENCODE and GEO databases. Detailed accession codes and sources can be found in Supplementary Tables S2, S3, S7 and S8.
Authentication	Cell line authentication was performed by the original data providers. We used the data as provided without additional authentication.
Mycoplasma contamination	Information regarding mycoplasma contamination is not applicable to this study, as we used publicly available sequencing data rather than cell cultures.
Commonly misidentified lines (See ICLAC register)	This study did not use any commonly misidentified or cross-contaminated cell lines. Data were obtained from well-curated public repositories.

Plants

Seed stocks	This study did not involve the use of plants, and therefore no seed stocks were utilized.
Novel plant genotypes	This study did not involve the use of plants or the generation of novel plant genotypes.
Authentication	This study did not involve the use of plants, so no authentication procedures for plant materials were performed.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://zenodo.org/records/18234973/files/TFBS_peak.zip?download=1

Files in database submission	The file TFBS_peak.zip contains the results of peak calling using MACS2 on multiple TF ChIP-seq datasets across the six cell lines used in this study, namely K562, GM12878, HCT116, IMR-90, MCF-7, and HepG2. A detailed list of the files is provided in Supplementary Table S8. Details of the histone modification ChIP-seq datasets used in this study are provided in Supplementary Table S3.
Genome browser session (e.g. UCSC)	No genome browser session is available for this study.

Methodology

Replicates	Publicly available data from ENCODE were used. Replicate information is described in the original ENCODE metadata and has not been modified for this study.
Sequencing depth	Sequencing depth was not determined in this study. All data were obtained from ENCODE, and the corresponding accession codes are provided in Supplementary Table S3 and S8, where sequencing depth information can be found in the associated ENCODE metadata.
Antibodies	No antibody experiments were performed in this study. Public ChIP-seq data were used as provided by ENCODE, and antibody details are available in the ENCODE metadata.
Peak calling parameters	Peak calling was performed using MACS2 (v2.2.9.1) on ENCODE-provided ChIP-seq BAM files (TF). A q-value threshold of 0.05 was used, and the genome size was set to human (-g hs).
Data quality	All ChIP-seq datasets were obtained from ENCODE, which applies rigorous QC pipelines. Quality metrics are available via the ENCODE portal and can also be downloaded from https://zenodo.org/records/18234973/files/BioSeq2Seq_QC.zip?download=1.
Software	MACS2 (v2.2.9.1) was used for peak calling. Bedtools (v2.26.0) was used for processing BED files. All other software packages and versions used for analysis are listed in the Methods section of the manuscript.