nature portfolio

Corresponding author(s): Raul Rabadan

Last updated by author(s): Oct 3, 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

Code is avaliable in GitHub (https://github.com/GET-Foundation). LocalColabFold 1.5.1 (https://github.com/YoshitakaMo/localcolabfold) and AlphaFold-multimer 2.3.1 weights were used to predict structures.
The folllowing python packages were also used:
numpy==1.26.4
pandas==1.5.3
pyranges==0.1.2
scipy==1.14.1
PyYAML==6.0
zarr==2.14.2
numcodecs==0.11.0
pyBigWig==0.3.22
matplotlib==3.6.2
networkx==2.8.4
plotly==5.22.0
seaborn==0.11.0
tqdm==4.65.0
cdt==0.6.0
pysam==0.21.0
requests==2.32.3
seqlogo==5.29.8
MOODS-python==1.9.4.1

urllib3==2.2.2
pyliftover==0.4
biopython==1.80
gprofiler-official==1.0.0
pyfaidx==0.7.1

Data analysis

PyTorch and PyTorch Lightning was used in the modeling of this study. Code for pretraining, finetuning, data preprocessing, and analysis has been made available in GitHub (https://github.com/GET-Foundation and https://github.com/RabadanLab) . The pretrained model is avaliable in huggingface (https://huggingface.co/get-foundation). Code for the website is available at https://huggingface.co/spaces/get-foundation/GET/tree/main. Code for analysis are in Zenodo 10.5281/zenodo.13357635. Python, NumPy, cdt (Causal Discovery Toolbox), seaborn, matplotlib, pyranges, statsannot, gprofiler, tsnecuda, UMAP, MOODS-python were used in the analysis. A detailed list with version are available as a conda environment file in the root directory of the codebase.
The following python package were used.
lightning==2.3.3
numpy==1.26.4
pandas==2.1.4
scipy==1.14.1
scikit-learn==1.2.2
matplotlib==3.6.2
seaborn==0.11.0
pytorch==2.1.1
networkx==3.1
umap-learn==0.5.3
statsmodels==0.13.5
zarr==2.14.2
louvain==0.8.0
MACS2==2.2.7.1
hydra-core==1.3.2
pyranges==0.1.2
PyYAML==6.0
numcodecs==0.11.0
pyBigWig==0.3.22
plotly==5.22.0
tqdm==4.65.0
cdt==0.6.0
pysam==0.21.0
requests==2.32.3
seqlogo==5.29.8
MOODS-python==1.9.4.1
urllib3==2.2.2
pyliftover==0.4
biopython==1.80
gprofiler-official==1.0.0
pyfaidx==0.7.1
tsnecuda==3.0.1

GelAnalyzer 23.1 were used to analyze western blot gel.

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

Precomputed regulatory inference results, preprocessed data, and structure predictions can be viewed at the GET website: https://huggingface.co/spaces/get-foundation/GET. The full processed data and inference results is provided in a public AWS S3 bucket at s3://2023-get-xf2217/get_demo. Bulk RNA-sequencing of B-ALL patients published in our previous study was acquired at SRA (PRJNA534488). Human transcription factor protein interaction networks were downloaded from supplementary data of Göös et al. Fetal accessibility data was downloaded from https://descartes.brotmanbaty.org/bbi/human-chromatin-during-development/. Adult accessibility data was downloaded from http://catlas.org/humanenhancer/. Fetal expression atlas data was downloaded from https://descartes.brotmanbaty.org/bbi/human-gene-expression-during-development/. Adult expression atlas data was downloaded from Tabular Sapiens (https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219/5). HUDEP-2 HiChIP was downloaded from GEO (GSE157311). Erythroblast base-editing benchmark data was downloaded from https://github.com/YichaoOU/ABE_NonCoding_functional_score/blob/master/per_A_base_score/comparison_to_CADD_DeepSEA_GERP/Editable_A_scores.combined.scores.csv. K562 NEAT-seq was downloaded from GEO (GSE178707). K562 omniATAC was downloaded from ENCODE (ENCSR483RKN). K562 scATAC-seq was downloaded from ENCODE (ENCFF998SLH). K562 CRISPRi benchmark data was downloaded from https://github.com/EngreitzLab/CRISPR_comparison/blob/main/resources/crispr_data/EPCrisprBenchmark_ensemble_data_GRCh38.tsv.gz. K562 lentiMPRA quantification was downloaded from ENCODE (ENCFF348WYK). K562 CAGE was downloaded from FANTOM5 (CNhs12336). HTAN GBM data was downloaded from GEO (GSE240822). Astrocyte bulk RNA-seq was downloaded from GEO (GSE147870).

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	No human participants/data are newly collected in this study.
Reporting on race, ethnicity, or other socially relevant groupings	No human participants/data are newly collected in this study.
Population characteristics	No human participants/data are newly collected in this study.
Recruitment	No human participants/data are newly collected in this study.
Ethics oversight	No human participants/data are newly collected in 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	The annotation of the 213 cell types used in pretraining followed the original cell type classification provided in the fetal and adult accessibility atlases. This classification was achieved through clustering of ATAC-seq count profiles, with subsequent labeling based on the expression of known marker genes. We used the entire genome for training unless in leave-out-chromosome settings. Number of region per sample is determined to achieve a genomic span of ~2-4Mbp. In general, for a peak set with 100,000 peaks, we use 200 peak per sample.
Data exclusions	The comprehensive list of tissues and cell types, along with their annotations, can be found at the Human Cell Atlas. The original atlas included 222 fetal and adult cell types, but was further filtered to remove cell types with low sequencing coverage (number of cells < 600). This approach ensured that our model was trained across diverse cellular contexts while ensuring enough coverage in the chromatin accessibility pseudobulk tracks. The data is not further balanced or curated. Expression performance are evaluated on genes with accessible promoters in each cell type due to the model architecture design, which helps to focus more on cell type specific genes.
Replication	Two independent experiments were performed for the TFAP2A-ZFX Co-IP assay. Three independent experiments were performed for the Proximity ligation assay to detect PAX5-NR2C2 interactions. All attempts at replication were successful
Randomization	In erythroblast and K562 enhancer-target identification benchmark, n=1000 bootstrapping with 80% subsampling ratio were used. In causal network construction with LiNGAM, n=5 independent random seed were used to run the algorithm and results were aggregated.
Blinding	Blinding was not relevant to our study. Given the nature of the study, we sought to compare the predictive capabilities of the GET foundation model across different cell types, experimental assays, and computational tasks. For regulatory element identification, we aimed to benchmark GET against existing methods using published experimental datasets. For transcription factor interaction prediction, we compared GET's performance to other computational approaches and experimental data. Our experimental validations involved quantitative measurements and analyses that did not require qualitative assessment or blinding.

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

Antibodies

Antibodies used

streptavidin-HRP antibody (Life Technologies S911) : 1/1000
anti-PAX5 (Cell Signaling 8970): 1/500
anti-HA (Cell Signaling 3724); 1/1000
anti-NR2C2 (Cell Signaling 31646) : 1/500
anti-NCOR1 (Cell Signaling 5948): 1/500
anti-NRIP1-HRP (Santa Cruz Biotechnology sc-518071) : 1/200
anti-NR3C1 (Cell Signaling 12041): 1/500
anti-TFAP2A (ABclonal A2294): 1/750
anti-ZFX (ThermoFisher PA5-34376): 1/500
anti-β-ACTIN (Santa Cruz Biotechnology sc-47778): 1/10000
anti-SRF (ABclonal A16718): 1/750
anti-β-ACTIN (Cell Signaling Technology 4967): 1/10000

Validation

All antibodies used are commercial so they have a validation statement on their website :

Cell signaling : "To ensure product performance, we validate all CST® antibodies, assay kits, and reagents to ensure optimal performance in the approved applications shown on our product web pages"
https://www.cellsignal.com/about-us/our-approach-process/cst-antibody-performance-guarantee
Santa-Cruz : "Primary antibodies directed to mammalian target proteins have been characterized for reactivity against mouse, rat and human proteins. Our antibodies are recommended for use in most assays including Western blot, immunoprecipitation, immunostaining, and flow cytometry." https://www.scbt.com/browse/antibodies?srsltid=AfmBOoqM1X9UXE3ARiVFIHrYvrpyL-rQ0wQkM4PXW_eIhJX0lRvJ4mqh
Abclonal:
"At ABclonal, we purify our antibodies using antigen affinity purification. Validation is completed using various applications and samples from different species. To find out more, visit us at ABclonal.com or contact us. "
https://blog.abclonal.com/blog/a-quick-guide-to-antibody-validation
Invitrogen:
Invitrogen antibodies are validated following comprehensive strategy for antibody validation as stated on vendor website.
https://www.thermofisher.com/us/en/home/references/newsletters-and-journals/bioprobes-journal-of-cell-biology-applications/bioprobes-75/bioprobes-75-antibody-validation.html
Each antibody has a datasheet where they state for "WB" use with WB images validating the antibody.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)

Initial HeLa cells was purchased from ATCC #CCL-2. HeLa cells were cultured in DMEM (Gibco, 11965) supplemented with 10% defined FBS (HyClone, SH30070), at 37°C/ 5% CO2. REH cells were purchased from ATCC #CRL-8286. REH stable cell lines with control vector pCDH_13Xlinker-BioID2-HA-GFP, pCDH_PAX5-WT-13Xlinker-BioID2-HA-GFP and pCDH_PAX5-G183S-13Xlinker-BioID2-HA-GFP were incubated with 100uM biotin (Sigma Aldrich #B4501) for 24h.

Authentication

Cell lines purchased from certificated cell line bank were not further authenticated.

Mycoplasma contamination

All cell lines were tested nagative for mycoplasma.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in the study

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*