nature portfolio

Corresponding author(s): Brian Hie (brianhie@stanford.edu), Patrick Hsu (patrick@arcinstitute.org)

Last updated by author(s): Dec 18, 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	Python 3.11, Mash v2.3, Prodigal v2.6.3, MMseqs 15, pybloom 1.1
Data analysis	Python 3.11 (package list on github), sklearn 1.3, Prodigal v2.6.3, Mitoz 3.6, AlphaFold Server (2 2025), MMseqs (release 16), Foldseek (release 9), Lovis4u 0.1.1, ESMFold, HHpred 3.3.0, Homer v5, TomTom v5.5, YGAP, Promoter 2.0

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 OpenGenome2 dataset used to train Evo 2 is available at:https://huggingface.co/datasets/arcinstitute/opengenome2.

We also include the following supplementary data files:

Data S1: Human variant effect prediction results. Scores and metadata for human variant effect prediction analysis.

Data S2: Mechanistic interpretability TF motif report. A comprehensive list of promoter-enriched motifs from the HOCOMOCO v12 CORE database and associated SAE feature hits.

Data S3: DNA sequences for experimental chromatin accessibility designs. Experimentally tested DNA sequences in the Morse code mESC and HEK293T/K562 chromatin accessibility design tasks.

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	N/A
Reporting on race, ethnicity, or other socially relevant groupings	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	Sample sizes for downstream benchmarking analyses were predetermined by the relevant database resources. Experimental sample sizes were predetermined to ensure technical and biological reproducibility.
Data exclusions	To enable consistent comparison, we filtered the results for ClinVar coding SNVs to variants that have a reported AlphaMissense score. No data exclusions for experimental results were performed.
Replication	For mESC chromatin accessibility designs, we created triplicate clonal cell lines for "square wave long," "square wave medium," and "Morse code LO". We created duplicate clonal cell lines for "square wave medium" and "Morse code EVO2". We were only able to obtain a single clonal cell line for "Morse code ARC." We created duplicate clonal cell lines for all HEK293T and K562 designs (Figure sS10H-L). We observed strong agreement in chromatin accessibility profiles measured by ATAC-seq across replicate clonal cell lines.
Randomization	Cells were randomly allocated into experimental conditions for our chromatin accessibility designs.
Blinding	Investigators were not blinded to the computational design information during experimental testing.

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)	BL6xCAST mESCs were obtained according to Eckersley-Maslin et al. (Developmental Cell, 2014). We obtained K562 and HEK293T from ATCC.
Authentication	The BL6xCAST mESC cell line has been verified by a variety of next-generation sequencing assays. We did not perform additional authentication for the K562 and HEK293T cell lines.
Mycoplasma contamination	Each of the cell lines with the Evo 2 constructs were not individually tested for mycoplasma, but we have no indication of infection and we do not expect mycoplasma infection to affect any of the results.
Commonly misidentified lines (See ICLAC register)	N/A

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.*