nature portfolio

Corresponding author(s): Marie Lopez and Thomas Pierrot

Last updated by author(s): Sep 10, 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	python 3.9, pandas 1.5.2, torch 2.0.1
Data analysis	Model code and weights of the pre-trained transformer models as well as inference code in Jax are available for research purposes at https://github.com/instadeepai/nucleotide-transformer. HuggingFace versions of the models, in PyTorch, can be found at https://huggingface.co/InstaDeepAI. Example note- books are available on HuggingFace at https://huggingface.co/docs/transformers/notebooks#pytorch-bio.

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 Nucleotide Transformer pre-training sequences were obtained from publicly available resources.
The 1000 Genomes Project sequences were obtained from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage, and the

human and multispecies reference genomes from https://ftp.ncbi.nlm.nih.gov/genomes/refseq/. Gene annotations were obtained from GENCODE (https://www.gencodegenes.org/) and Ensembl databases (https://www.ensembl.org). Variants effects predictions were obtained using the Variant Effect Prediction (VEP) API from Ensembl (https://www.ensembl.org/info/docs/tools/vep/index.html). Pathogenic and regulatory variants were extracted from ClinVar (https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/), the Genome-Wide Repository of Associations Between SNPs and Phenotypes (GRASP) (https://grasp.nhlbi.nih.gov), and The Human Gene Mutation Database (HGMD) (public version 2020.4) through Ensembl Biomart. The Nucleotide Transformer downstream tasks were curated from the following publicly available resources: ENCODE (https://www.encodeproject. org/, https://screen.wenglab.org/, https://www.meuleman.org/research/dhsindex/), the Eukaryotic Promoter Database (https://epd.expasy.org/epd/) and GENCODE. We make available HuggingFace versions of our multispecies and human pre-training datasets (https://huggingface.co/collections/InstaDeepAI/nucleotide-transformer-65099cdde13ff96230f2e592) as well as our downstream tasks benchmark (https://huggingface.co/datasets/InstaDeepAI/nucleotide_transformer_downstream_tasks_revised). We have established an interactive leaderboard containing results for all models across each task to facilitate comparisons, available at https://huggingface.co/spaces/InstaDeepAI/nucleotide_transformer_benchmark. We have also created an interactive browser session with the pre-trained model token probabilities across the full chromosome 22 on the WashU Epigenome Browser at https://shorturl.at/jov28.

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	We used the 1000 Genomes Project database which contains 3202 high-coverage human genomes, originating from 27 geographically structured populations of African, East Asian and European ancestry. The multispecies database contained 850 species' whole genomes retrieved from NCBI.
Data exclusions	No data were excluded from the analyses.
Replication	We implemented a 10-fold cross-validation strategy in our benchmark.
Randomization	We splatted sequences into training and test held-out sets.
Blinding	Not applicable, we used held-out test sets.

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

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A