nature portfolio

Corresponding author(s): Patrick Ellinor

Last updated by author(s): Oct 19, 2022

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

mygene (v3.2.1): https://github.com/SuLab/mygene.info
requests (v2.25.1): https://github.com/psf/requests

Data analysis

The pretrained Geneformer model and transcriptome tokenizer are available on the Huggingface Model Hub (https://huggingface.co/ctheodoris/Geneformer). All other code used in this study is available upon request from the corresponding author.

anndata (v0.7.6): https://github.com/scverse/anndata
cudatoolkit (v11.2.2): https://developer.nvidia.com/cuda-toolkit
cudnn (v8.1.0.77): https://developer.nvidia.com/cudnn
datasets (v1.6.2): https://github.com/huggingface/datasets
deepspeed (v0.3.10): https://github.com/microsoft/DeepSpeed
gseapy (v0.10.8): https://github.com/zqfang/GSEApy
harmonypy (v0.0.5): https://github.com/nasa/harmony-py
hyperopt (v0.2.5): https://github.com/hyperopt/hyperopt
loompy (v3.0.6): https://github.com/linnarsson-lab/loompy
matplotlib (v3.4.2): https://github.com/matplotlib/matplotlib
mpi4py (v3.0.3): https://github.com/mpi4py/mpi4py
multiprocess (v0.70.11.1): https://github.com/uqfoundation/multiprocess
nccl (v2.8.4.1): https://developer.nvidia.com/nccl
networkx (v2.5): https://github.com/networkx
numpy (v1.20.2): https://github.com/numpy/numpy
openmpi (v4.0.5): https://github.com/open-mpi
pandas (v1.2.4): https://github.com/pandas-dev/pandas
parquet-cpp (v1.5.1): https://github.com/apache/parquet-cpp

pyarrow (v4.0.0): https://github.com/apache/arrow/tree/master/python/pyarrow
python (v3.8.5): https://www.python.org/downloads/
pytorch (v1.8.0): https://github.com/pytorch/pytorch
ray-tune (v1.9.2): https://github.com/ray-project/ray/tree/master/python/ray/tune
scanpy (v1.7.2): https://github.com/scverse/scanpy
scikit-learn (v0.24.2): https://github.com/scikit-learn/scikit-learn
scipy (v1.6.3): https://github.com/scipy/scipy
seaborn (v0.11.1): https://github.com/mwaskom/seaborn
statsmodels (v0.12.2): https://github.com/statsmodels/statsmodels
tdigest (v0.5.2.2): https://github.com/tdunning/t-digest
tensorboard (v2.4.1): https://github.com/tensorflow/tensorboard
tokenizers (v0.10.2): https://github.com/huggingface/tokenizers
torchmetrics (v0.3.2): https://github.com/PyTorchLightning/metrics
transformers (v4.6.0): https://github.com/huggingface/transformers

R (v4.0.5): https://www.r-project.org/
LoomExperiment (v1.8.0): https://bioconductor.org/packages/release/bioc/html/LoomExperiment.html
CellRanger (v6.0.1): https://support.10xgenomics.com/single-cell-gene-expression/software

For preprocessing and differential expression analysis of the cardiomyopathy dataset:
CellRanger (v4.0.0): https://support.10xgenomics.com/single-cell-gene-expression/software
cutadapt (v1.18): https://cutadapt.readthedocs.io/en/stable/
CellBender remove-background (v0.2): https://github.com/broadinstitute/CellBender
scR-Invex (sha1:4a067c5): https://github.com/broadinstitute/scrinvex
Python (v3.7): https://www.python.org/
scanpy (v1.6.0): https://github.com/theislab/scanpy
harmony-pytorch (v0.1.4): https://github.com/lilab-bcb/harmony-pytorch
ndd (v1.6.3): https://github.com/simomarsili/ndd
Scrublet (v0.2.1): https://github.com/AllonKleinLab/scrublet
R (v3.5.0, v3.6.0): https://www.r-project.org/
DESeq2 (v1.20.0): https://bioconductor.org/packages/release/bioc/html/DESeq2.html
limma (v3.36.5): https://bioconductor.org/packages/release/bioc/html/limma.html
edgeR (v3.22.5): https://bioconductor.org/packages/release/bioc/html/edgeR.html
GOstats (v2.46.0): https://bioconductor.org/packages/release/bioc/html/GOstats.html

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

Genecorpus-30M is available on the Huggingface Dataset Hub (https://huggingface.co/datasets/ctheodoris/Genecorpus-30M).
Processed single-nuclei transcriptomic data from non-failing hearts and hearts affected by hypertrophic cardiomyopathy or dilated cardiomyopathy are available through the Broad Institute's Single Cell Portal (https://singlecell.broadinstitute.org/single_cell) under project ID SCP1303 (https://singlecell.broadinstitute.org/single_cell/study/SCP1303/). Raw sequence data are available to authorized users through dbGaP (the database of Genotypes and Phenotypes) accession number phs001539.

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 pretraining, we collected as much data as was publicly available and contributing towards a breadth of represented human tissues. For fine-tuning, we selected the largest sample set available within single datasets relevant to each application to test the efficacy of fine-tuning the model with limited data. For the cardiomyopathy disease modeling, we selected the largest sample set available to us of non-failing (n=13), hypertrophic cardiomyopathy (n=15), and dilated cardiomyopathy (n=11) left ventricles for single-nuclei RNA-sequencing. No statistical methods were used to predetermine sample size.
Data exclusions	For all pretraining and fine-tuning, single cells were excluded if single cell RNA-sequencing was low quality based on 1) total read counts not

Data exclusions	within three standard deviations of the mean within that dataset, 2) mitochondrial reads not within three standard deviations of the mean within that dataset, or 3) less than seven detected Ensembl-annotated protein-coding or miRNA genes. The publicly available data used in this study were collected as count matrices so initial filtering for sequencing quality control had already been performed by the original authors. However, for the cardiomyopathy disease modeling, samples were excluded if single-nuclei RNA-sequencing was low quality based on 1) <50% of reads in cells, 2) < 65% of reads confidently mapping to transcriptome, 3) < 90% valid barcodes, 4) abnormally low Q30, or 5) no ambient plateau in the unique molecular identifier (UMI) decay curve. Additionally, we only included non-failing heart samples that had a documented normal ejection fraction. Nuclei were removed if they 1) were enriched for mitochondrial reads, 2) were enriched for the proportion of reads mapping exclusively to exonic regions, 3) had a high prediction for being a doublet, 4) had extremely high or low UMI, 5) had extremely high or low number of genes detected, or 6) had low entropy. For experimental validation in engineered cardiac microtissues, only samples with confirmed CRISPR-mediated knockout of the intended target were analyzed.
Replication	For all AUC evaluations, 5 replicates were performed for each model by 5-fold cross-validation. For cardiomyopathy disease modeling, we performed single-nuclei RNA-seq in duplicate on left ventricles from 44 hearts. At least 1 replicate was retained for 39 samples after quality control. For experimental validation in engineered cardiac microtissues, at least 7 replicates were included for each condition (see figure legends for exact number of replicates for each experiment).
Randomization	For gene classification fine-tuning, labeled genes were randomized into training data (80%) and test data (20%), repeating for 5 cross-validation folds. For cardiomyopathy disease modeling, patients were randomized into training data (non-failing n=9, hypertrophic cardiomyopathy n=11, dilated cardiomyopathy n=9) and test data (non-failing n=4, hypertrophic cardiomyopathy n=4, dilated cardiomyopathy n=2).
Blinding	Blinding was not relevant to our study. Given the nature of the study, we sought to compare the predictive potential of the pretrained Geneformer model to alternative modeling approaches and to training with smaller and less diverse pretraining corpuses. For cardiomyopathy disease modeling, we sought to compare the transcriptional networks of cardiomyopathy and non-failing left ventricles. For experimental validation in engineered cardiac microtissues, measurements were quantified through software that did not involve qualitative assessment.

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
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Patrick Ellinor Lab
Authentication	Cells were confirmed by genotyping.
Mycoplasma contamination	Cells are tested for mycoplasma contamination prior to use in stem cell culture facility.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	All samples came from the Myocardial Applied Genomics Network (MAGNet) repository. Left ventricle tissue was obtained from a total of 44 individuals, including 12 dilated cardiomyopathy donors, 16 hypertrophic cardiomyopathy donors, and 16 non-failing donors. All samples were of European ancestry. 5/12 dilated cardiomyopathy donors were female, 5/16 hypertrophic cardiomyopathy donors were female, and 10/16 non-failing donors were female. Average age of dilated cardiomyopathy patients was 55, hypertrophic cardiomyopathy 49, and non-failing 57.
Recruitment	Dilated and hypertrophic cardiomyopathy left ventricle tissue was collected at time of heart transplantation. Non-failing left ventricle tissue was collected from deceased donors with no overt cardiovascular disease.
Ethics oversight	Written informed consent for research use of donated tissue was obtained from next of kin in all cases. Research use of

Ethics oversight	tissues were approved by the relevant institutional review boards at the Gift-of-Life Donor Program, the University of Pennsylvania, Massachusetts General Hospital, and the Broad Institute.

Note that full information on the approval of the study protocol must also be provided in the manuscript.