Mouse-Geneformer: A Deep Learning Model for Mouse Single-Cell Transcriptome and Its Cross-Species Utility

Keita Ito[1], Tsubasa Hirakawa[2], Shuji Shigenobu [3,4]*, Hironobu Fujiyoshi [5]*, Takayoshi Yamashita [6]*

[1] Graduate School of Engineering, Chubu University, Kasugai, Aichi, Japan

[2] Chubu University, Kasugai, Aichi, Japan

[3] Trans-Scale Biology Center, National Institute for Basic Biology, Okazaki, Aichi, Japan

[4] Life Science Center for Survival Dynamics, Tsukuba Advanced Research Alliance (TARA), University of Tsukuba, Tsukuba, Ibaraki, Japan

[5] Department of Computer Science, Chubu University, Kasugai, Aichi, Japan

[6] Department of Artificial Intelligence and Robotics, Chubu University, Kasugai, Aichi, Japan

* Correspondence authors:

Takayoshi Yamashita, Hironobu Fujiyoshi, Shuji Shigenobu

emails: takayoshi@isc.chubu.ac.jp, fujiyoshi@isc.chubu.ac.jp, shige@nibb.ac.jp

Abstract

Deep learning techniques are increasingly utilized to analyze large-scale single-cell RNA sequencing (scRNA-seq) data, offering valuable insights from complex transcriptome datasets. Geneformer, a pre-trained model using a Transformer Encoder architecture and human scRNA-seq datasets, has demonstrated remarkable success in human transcriptome analysis. However, given the prominence of the mouse, *Mus musculus*, as a primary mammalian model in biological and medical research, there is an acute need for a mouse-specific version of Geneformer. In this study, we developed a mouse-specific Geneformer (mouse-Geneformer) by constructing a large transcriptome dataset consisting of 21 million mouse scRNA-seq profiles and pre-training Geneformer on this dataset. The mouse-Geneformer effectively models the mouse transcriptome and, upon fine-tuning for downstream tasks, enhances the accuracy of cell type classification. In silico perturbation experiments using mouse-Geneformer successfully identified disease-causing genes that have been validated in in vivo experiments. These results demonstrate the feasibility of analyzing mouse data with mouse-Geneformer and highlight the robustness of the Geneformer architecture, applicable to any species with large-scale transcriptome data available. Furthermore, we found that mouse-Geneformer can analyze human transcriptome data in a cross-species manner. After the ortholog-based gene name conversion, the analysis of human scRNA-seq data using mouse-Geneformer, followed by fine-tuning with human data, achieved cell type classification accuracy comparable to that obtained using the original human Geneformer. In in silico simulation experiments using human disease models, we obtained results similar to human-Geneformer for the myocardial infarction model but only partially consistent results for the COVID-19 model, a trait unique to humans (laboratory mice are not susceptible to SARS-CoV-2). These findings suggest the potential for cross-species application of the Geneformer model while emphasizing the importance of species-specific models for capturing the full complexity of disease mechanisms. Despite the existence of the original Geneformer tailored for humans, human research could benefit from mouse-Geneformer due to its inclusion of samples that are ethically or technically inaccessible for

humans, such as embryonic tissues and certain disease models. Additionally, this cross-species approach indicates potential use for non-model organisms, where obtaining large-scale single-cell transcriptome data is challenging.

Author Summary

Researchers have developed Geneformer, a powerful tool that utilizes advanced deep learning techniques and large-scale single-cell transcriptome data to analyze human cell genetic activity. However, given the extensive use of mice (*Mus musculus*) in medical and biology research, there is a need for a similar tool tailored to this model organism. To address this gap, we developed mouse-Geneformer, an adaptation of Geneformer trained on a large dataset of mouse single-cell RNA sequencing data obtained from 20 million cells. Mouse-Geneformer demonstrates high accuracy in identifying distinct cell types and predicting disease-causing genes in gene manipulation simulation experiments. Moreover, mouse-Geneformer exhibited comparable accuracy to the original human Geneformer, even when applied to human cell data, suggesting its potential for cross-species use. For instance, it performed well in studying heart disease but was less consistent with COVID-19, likely due to the differences between species in how they react to the virus. Overall, mouse-Geneformer could be a valuable resource for studying not only mice but also other animals, especially when large-scale data are challenging to obtain. Furthermore, this cross-species approach may probe beneficial in human research, especially for tissues that are difficult to access, such as embryonic samples.

Keywords: Single-cell RNA-sequence analysis, Bioinformatics, Deep learning, Geneformer

Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful technique that quantifies gene expression profiles at the individual cell level [1]. Recent technological advances in scRNA-seq have facilitated the rapid expansion of transcriptomic data and enabled the simultaneous analysis of thousands of single cells. This capability has significantly enhanced our understanding of developmental processes and disease mechanisms by revealing previously hidden heterogeneous cellular populations and novel cell types. scRNA-seq is being applied to a wide variety of experiments across different organisms, resulting in a rapid growth of scRNA-seq databases. These extensive datasets offer tremendous potential for collective use, providing valuable insights into genetic architecture and furthering our knowledge of cellular and molecular biology.

Deep learning, a recent advancement in artificial intelligence, has been successfully applied to numerous problems involving large datasets, and has emerged as a promising tool for analyzing scRNA-seq data [2–9]. This approach is particularly effective at extracting valuable information from noisy, heterogeneous, and high-dimensional transcriptome data. Among the successful methods in this domain is Geneformer, a pre-trained model that employs a Transformer Encoder architecture [10], similar to BERT [11], a widely used attention-based deep learning model in natural language processing. Geneformer utilizes the attention mechanism to calculate the relationships between genes and the context within “cell sentences”, enabling it to comprehend context-dependent genetic network dynamics. By fine-tuning Geneformer on specific downstream tasks using limited data, researchers can achieve accurate cell type classification. Additionally, Geneformer facilitates in silico simulations of gene manipulation experiments, thereby streamlining the identification of disease-causing genes and advancing our understanding of genetic networks and disease mechanisms.

The mouse, *Mus musculus*, is the foremost mammalian model for studying human biology and disease [12]. Extensive knowledge has been accumulated about mouse physiology, anatomy, and genetics. Methods for genetic manipulation, such as creating transgenic, knockout, and knockin animals, which are ethically and technically challenging in humans, have significantly advanced mouse research. These tools have led to a dramatic increase in the use of mice as model organisms. scRNA-seq experiments have also been actively applied to mouse studies, resulting in a rapid accumulation of scRNA-seq data. In this context, a deep learning model of the mouse transcriptome would greatly benefit mouse studies, and Geneformer presents a promising candidate for this purpose. However, the original Geneformer is modeled on the human transcriptome. Predicting disease-causing genes in mice using a human-based Geneformer is not straightforward. Therefore, a mouse version of Geneformer (mouse-Geneformer) is in high demand.

This paper aims to construct a mouse version of Geneformer, a deep learning model trained on mouse scRNA-seq data. We then evaluate the usefulness of mouse-Geneformer for various downstream analyses such as cell type classification and in silico perturbation experiments. We also explore the potential for cross-species application of the mouse-Geneformer. If successful, mouse-Geneformer could be used for humanstudies, where some samples are inaccessible due to ethical and technical constraints, and for non-model organisms, where large-scale scRNA-seq data are not available.

Materials and methods

Construction of the mouse-Geneformer

Architecture

We constructed mouse-Geneformer, a deep learning model aimed at predicting the gene network of normal mice. We developed the mouse-Geneformer following the original human version of Geneformer, a context-aware, attention-based deep learning model, which was pretrained on large-scale transcriptome data comprising approximately 30 million human single-cell RNA-seq data, developed by Theodoris et al [10]. The architecture and transfer learning strategy employed in our development of mouse-Geneformer are the same as those of the original human Geneformer, with minor modifications. An overview diagram of the mouse-Geneformer is shown in Fig 1. Instead of human RNA-seq data, we used mouse single-cell RNA-seq data as an input corpus. The transcriptome of each single cell was presented to the model by using the Rank Value Encoding method that was also developed in the original human Geneformer [10]. Finally, a Geneformer model constructed with Transformer Encoder using the created cell texts is pre-trained with self-supervised learning to build the mouse-Geneformer.

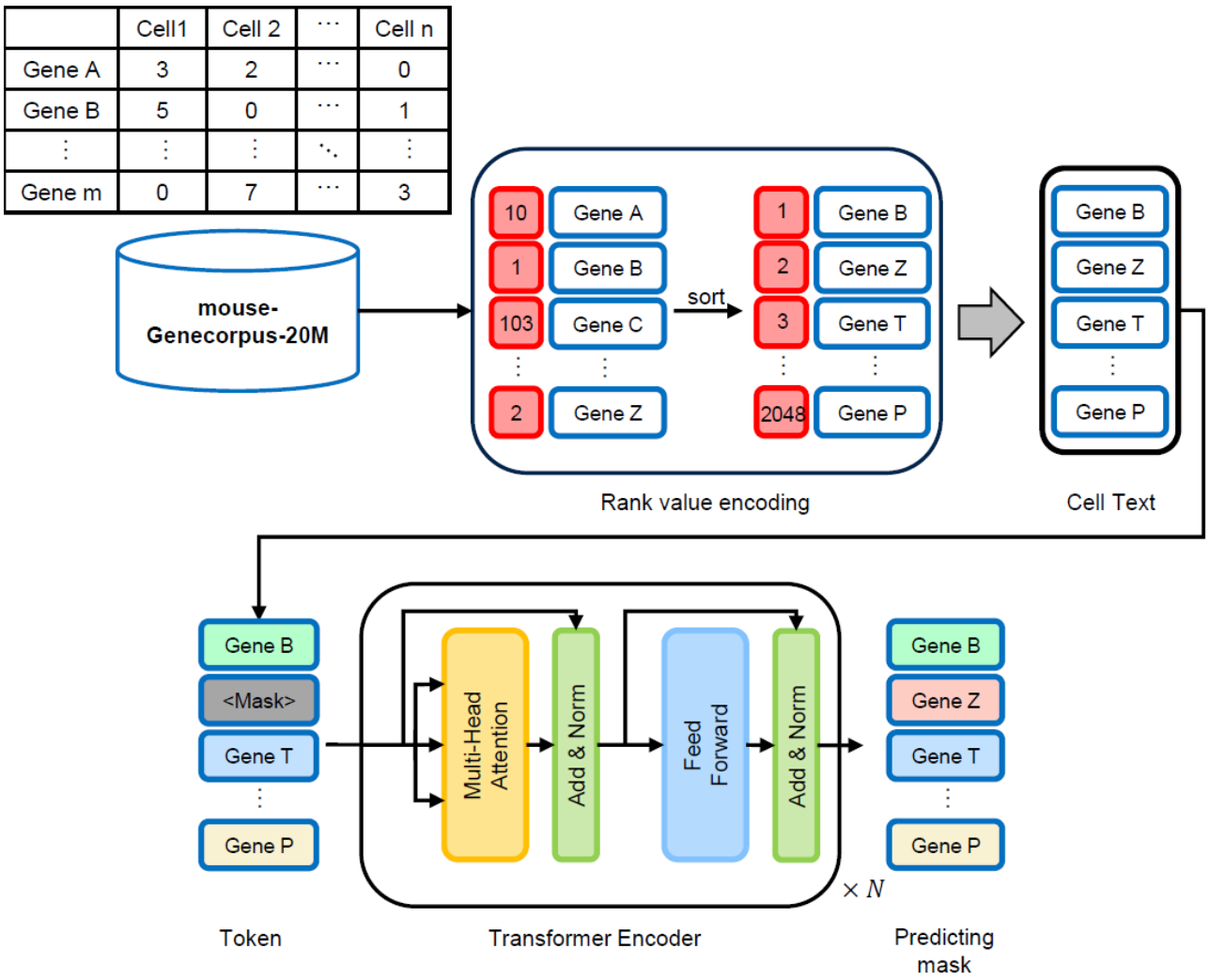


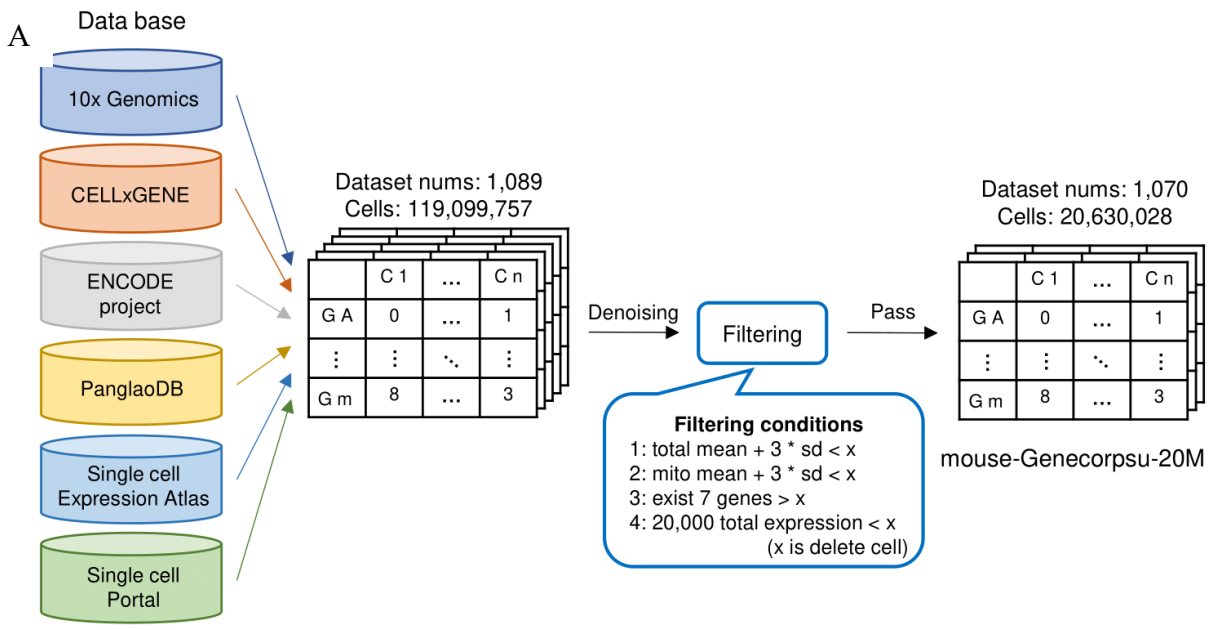
Fig 1 | Overview of the mouse-Geneformer. mouse-Genecorpus-20M is a large mouse transcriptome dataset obtained from single-cell RNA-seq experiments, Rank value encoding is a method for generating "cellular sentences", Transformer Encoder is the architecture of the mouse-Geneformer.

Construction of mouse-Genecoupus-20M

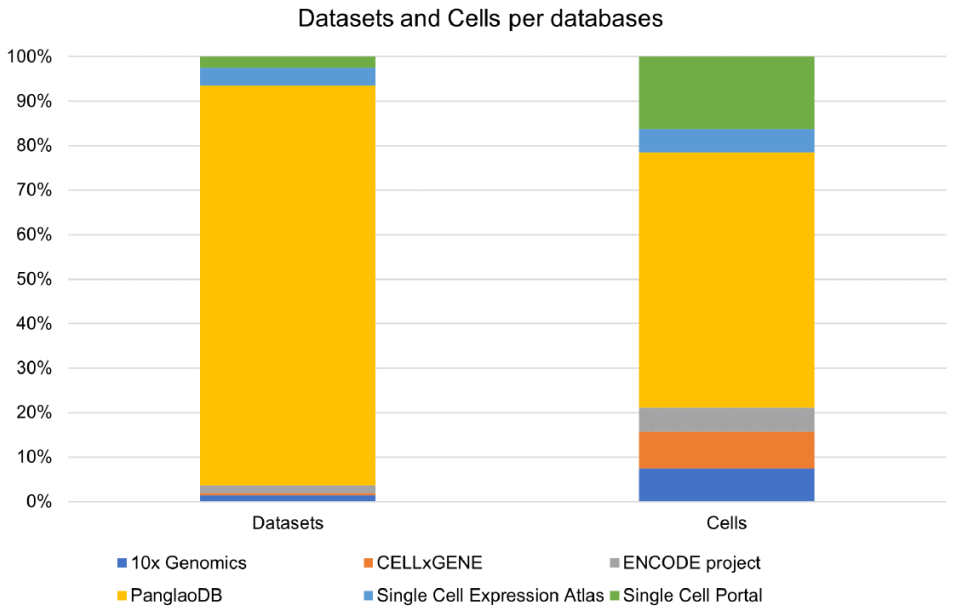
We constructed a large-scale dataset of mouse single-cell RNA-seq data, termed mouse-Genecorpus-20M, derived from healthy wild-type *Mus musculus* strains. The dataset was compiled

from publicly available mouse single-cell RNA-seq data sourced from multiple databases: PanglaoDB [13], Single Cell Expression Atlas [14], Single Cell Portal [15], ENCODE project [16], 10x Genomics [17], and CELLxGENE [18]. We used datasets generated by droplet-based sequencing platforms such as Chromium (10x Genomics), Drop-seq, DropNuc-seq, MIRACL-seq, and inDrops. In total, 1,089 RNA-seq datasets were compiled. The complete list of datasets, along with their metadata, is provided in S1 Table, and an overview is shown in Fig 2. These datasets collectively represent single-cell transcriptome data from 119 million cells before filtering (see below).

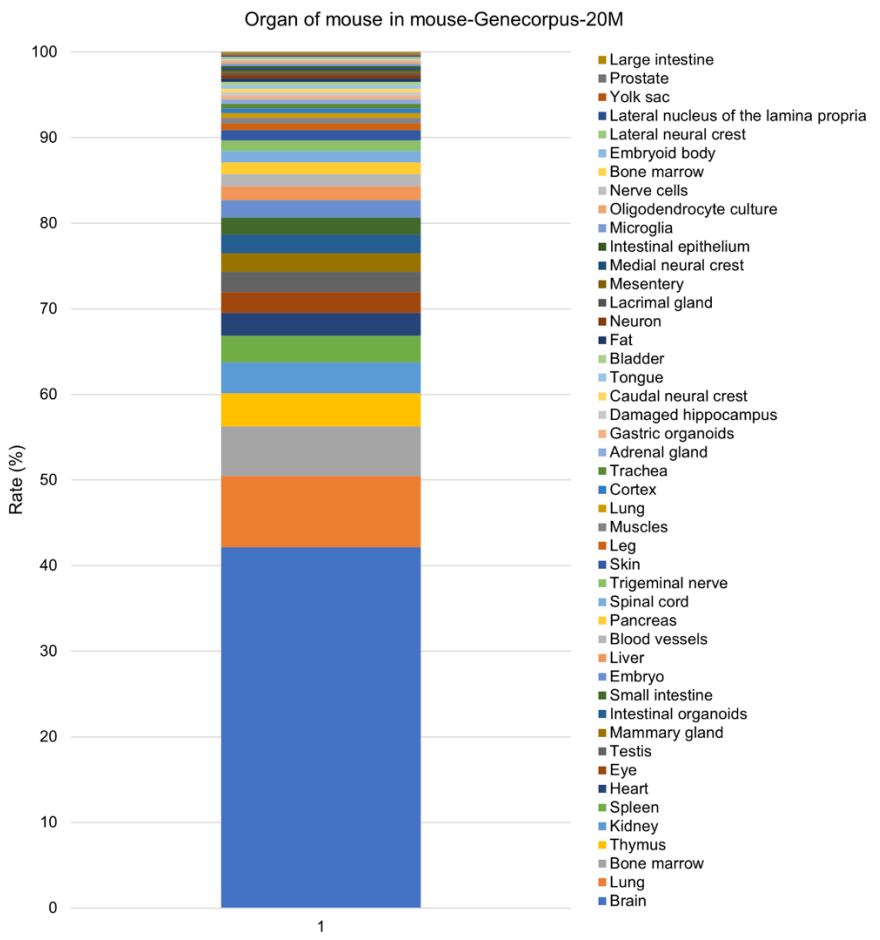
These single-cell datasets were stored and processed in HDF5 format loom files, H5AD files, and pickle files using the tools and libraries including scanpy, loompy, pandas, numpy, pickle, and scipy. The raw gene expression data is structured in matrices where rows represent genes, columns represent cells, and the gene expression levels are stored in the cells of the matrix. The dataset uses Ensembl IDs for gene names. Since the Ensembl ID version varied across datasets, the version information was ignored.



B


Datasets and Cells per databases
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Datasets
Cells
10x Genomics
CELLxGENE
ENCODE project
PanglaoDB
Single Cell Expression Atlas
Single Cell Portal

C


Organ of mouse in mouse-Genecorpus-20M
Rate (%)
100
90
80
70
60
50
40
30
20
10
0
1
Large intestine
Prostate
Yolk sac
Lateral nucleus of the lamina propria
Lateral neural crest
Embryoid body
Bone marrow
Nerve cells
Oligodendrocyte culture
Microglia
Intestinal epithelium
Medial neural crest
Mesentery
Lacrimal gland
Neuron
Fat
Bladder
Tongue
Caudal neural crest
Damaged hippocampus
Gastric organoids
Adrenal gland
Trachea
Cortex
Lung
Muscles
Leg
Skin
Trigeminal nerve
Spinal cord
Pancreas
Blood vessels
Liver
Embryo
Small intestine
Intestinal organoids
Mammary gland
Testis
Eye
Heart
Spleen
Kidney
Thymus
Bone marrow
Lung
Brain

D

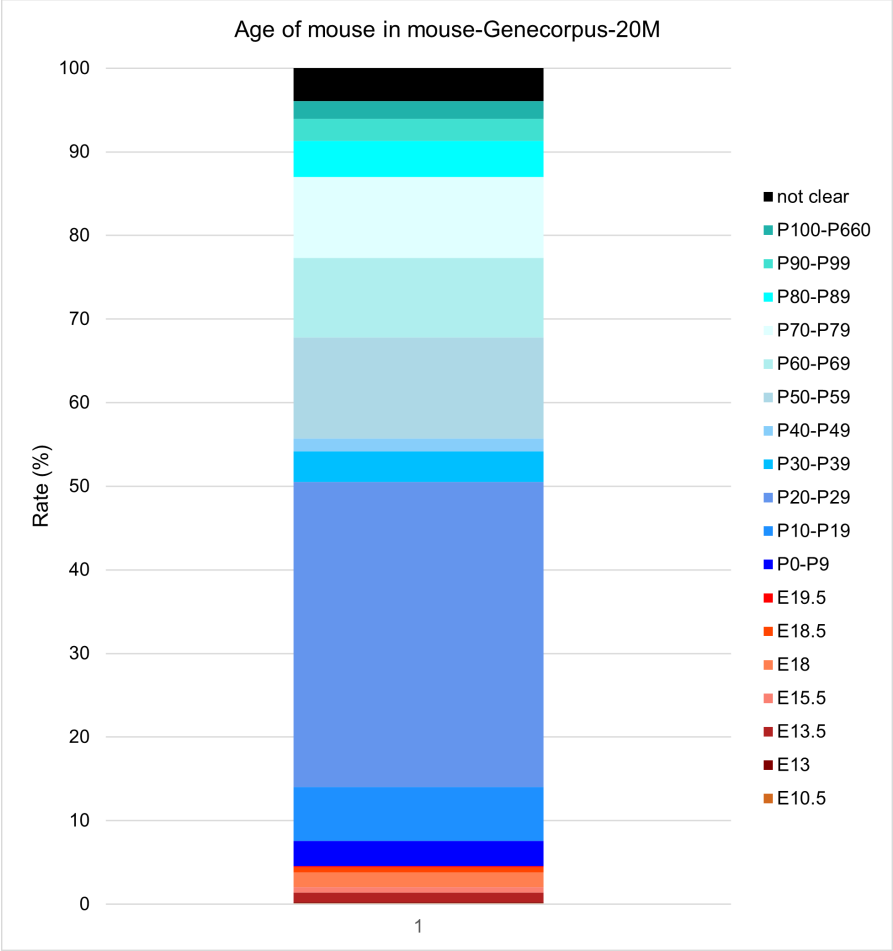


Fig 2 | Building mouse-Genecourpus-20M from single-cell RNA-seq datasets. A): Overview of how to create the mouse-Genecorpus-20M. B): Source of databases of the datasets used for building the mouse-Genecorpus-20M. The vertical axis shows the percentage of the number of datasets and cells derived from each database included in the dataset. C): Variety mouse organs included in the mouse-Genecorpus-20M. The vertical axis indicates the cell number percentage for each organ. D): Age and developmental stages of mice used in the mouse-Genecorpus-20M. The vertical axis indicates the percentage of cell numbers corresponding to mouse age or embryonic stage. Labels starting with E represents embryonic and P represents postnatal.

165 Starting with 119 million cells from the compiled 1,089 datasets, we applied filtering based
166 on metadata and quality. Since mouse-Genecorpus-20M is intended for learning the gene network
167 of normal mice to build a reference genetic architecture, we excluded single-cell data with high
168 mutation loads that could cause restructuring of gene networks, such as those from cancer cells or
169 immortalized cell lines. Quality filtering also addressed artifacts common in droplet-based single-
170 cell data, such as ambient RNA [19,20], doublets [21,22], and data derived from empty droplets.
171 The specific filtering conditions applied were as follows: 1) For each dataset, cells with total gene
172 expression levels more than three standard deviations from the mean were excluded. 2) Cells with
173 mitochondrial RNA expression levels more than three standard deviations from the mean were
174 excluded. 3) Cells with fewer than seven detected genes per cell were removed. 4) Cells with more
175 than 20,000 total gene expression levels were excluded. After applying these filters, the final dataset
176 was 1,070 RNA-seq datasets and encompassed 20,630,028 cells. The mouse-Genecorpus-20M
177 dataset is formatted in Apache Arrow and is available for download from the repository
178 (https://huggingface.co/datasets/MPRG/Mouse-Genecorpus-20M). It can be accessed using the
179 Python library "datasets" for downstream applications.

180 *Pretraining of the mouse-Geneformer*

181 Pretraining was conducted following the procedures described previously for the original human
182 Geneformer [10], with some modifications. Details are described in the Results section. Briefly, the
183 mouse-Genecorpus-20M dataset was processed using Rank Value Encoding to extract genes that
184 capture cell features, forming cell sequences with these genes as tokens. A special token called
185 [CLS] is added for classifying cell sequences. The resulting data, consisting of cell sequences with
186 token IDs, is used as input for pre-training the mouse-Geneformer. The pretraining employed a
187 masked token prediction task, where 15% of the tokens in the input sequences were randomly
188 masked, and the model was trained to predict these masked tokens using the remaining tokens. This

task optimizes the network, allowing the mouse-Geneformer to learn the relationships and expression patterns of genes in mice.

We conducted the pretraining of the mouse-Geneformer using mouse-Genecorpus-20M under the conditions shown in Table 1. The training utilized 8 NVIDIA V100 GPUs, each with 32 GB of memory, and the process took approximately 2 days

Table 1 | Summary of Experiment conditions

Pretrianing is the mouse-Geneformer prior learning conditions. Eval 1 is fine-turning conditions of the mouse-Geneformer using Cell type classification using the mouse-Geneformer and conventional methods. Eval 2 is its using Cell type classification of mouse-Geneformer with and without prior learning. Eval 3 is its using in silico perturbation experiments using the mouse-Geneformer. Eval 4 is its using Cell type classification with conversion of human and mouse genes. Eval 5 is its using in silico perturbation experiments with conversion of human and mouse genes.

Experiment conditions	**Pretrain**	**Eval 1**	**Eval 2**	**Eval 3**	**Eval 4**	**Eval 5**
Input max dimensions	2048	2048	2048	2048	2048	2048
Transformer Encoder Brocks	6	6	6	6	6	6
Attention Heads	4	4	4	4	4	4
Embedding dimensions	256	256	256	256	256	256
Activation function	SiLU	SiLU	SiLU	SiLU	SiLU	SiLU
Dropout rate	0.02	0.02	0.02	0.02	0.02	0.02
Epochs	10	10	10	10	10	10
Max learning rate	0.001	5.00E-05	5.00E-05	5.00E-05	5.00E-05	5.00E-05
Warm up steps	10000	500	500	500	500	500
Scheduler	cosine	cosine	cosine	cosine	cosine	cosine
Batch sizes	12	12	12	12	12	12
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Weight decay	0.001	0.001	0.001	0.001	0.001	0.001
Fleeze Brocks	0	0	6	0	0	0

Fine-turning of the mouse-Geneformer

Fine-tuning the mouse-Geneformer enables accurate classification of cell types and disease types, even with limited mouse single-cell data, such as data containing rare cells, disease-specific data, or organ-specific data. This process leverages the knowledge gained from pretraining on mouse-Genecorpus-20M to predict gene networks accurately. Fine-tuning involves adding a classification layer to classify the [CLS] token in the final layer of the mouse-Geneformer. The model is initialized with the weights from the pre-trained mouse-Geneformer and further trained using a small amount of specific data. When the fine-tuning task is similar to the pretraining task, some layers of the Transformer Encoder Block are fixed to enhance generalization performance. To mitigate overfitting, which can occur due to the limited amount of data, the number of training iterations is reduced.

Cell type classification using the mouse-Geneformer and other methods

To evaluate mouse-Geneformer, we compared its performance on cell type classification tasks with that of conventional methods, ~~,~~ Single-cell VAE (scVAE) [23] and scDeepSort [24]. The mouse data used for comparison comprised single-cell data from nine organs of mice: tongue [25], thymus [25], mammary gland [25], large intestine [25], limb muscle [25], spleen [25], heart [25], brain [25], and kidney [25] (Gene Expression Omnibus (GSE): GSE132042). These datasets were downloaded from CELLxGENE (https://cellxgene.cziscience.com/datasets). Each method was employed to classify the data, and their accuracies are compared. For cell type classification with mouse-Geneformer, the model was fine-tuned for each organ data for this task. The fine-tuning conditions are detailed in Table 1. For cell type classification with scVAE, a method applying the deep generative model Gaussian-mixture VAE (GMVAE) [26] to single-cell RNA-seq analysis [23], we

used the GMVAE with a three-layer neural network for the encoder, a Gaussian mixture distributions for the latent space, a three-layer neural network for the decoder, and a three-layer neural network for the class classification network. Additionally, the categorical distribution used in this GMVAE employs the Gumbel-Softmax [27,28]. For cell type classification with scDeepSort, a pretrained deep learning model utilizing weighted Graph Neural Networks (GNNs) for single-cell RNA-seq analysis [29], we employed scDeepSort, which consists of a three-layer graph neural network [24]. The data were randomly split into two groups, one for training (80%) and the other for testing, (20%). The evaluation metric for this experiment was accuracy, which was calculated by using the 'accuracy_score' function from the scikit-learn library in Python, and the highest accuracy was recorded. Cell distribution was visualized using UMAP [30] performed using the scanpy library in Python.

Cell type classification using mouse-Geneformer with and without prior learning

To evaluate the effect of pretraining, we compared the performance of pretrained mouse-Geneformer with non-pretrained one for the task of classification of cell types. The data used for evaluation included single-cell RNA-seq data from the mouse urethra and prostate mixed dataset [31] (Gene Expression Omnibus (GEO): GSE145929), embryos [32] (Gene Expression Omnibus (GEO): GSE197353), and kidneys [33] (Gene Expression Omnibus (GSE): GSE190094). These datasets were downloaded from CELLxGENE (https://cellxgene.cziscience.com/datasets). This embryo dataset is whole mouse embryo at embryonic age E9.5. These datasets were not included in mouse-Genecorpus-20M. The mouse-Geneformer models with and without pretraining were employed to classify the data, and their accuracies are compared. To perform cell type classification with mouse-Geneformer, the models are fine-tuned for each organ data for the cell type classification task. The fine-tuning conditions are presented in Table1. The data splitting method,

evaluation metrics, and cell distribution visualization methods are the same as those used in cell type classification experiments (see above).

In silico perturbation experiments

In in silico perturbation experiments using mouse-Geneformer, gene expression data was first normalized, and then each gene's expression was ranked using Rank Value Encoding to create the dataset. The mouse-Geneformer was then fine-tuned for disease classification tasks, ensuring that the disease classification accuracy of the test data exceeds 90%. Subsequently, the genes in the dataset were randomly perturbed multiple times, and inference was performed using the fine-tuned model. When deleting a gene, the gene was removed from the dataset, and the rank values of other genes were increased. Conversely, when activating a gene, the rank value of that gene was increased, and the rank values of other genes were decreased. Cosine similarity was used to quantify the distances between the perturbed cell state and a specific cell state.

To evaluate the effect of in silico perturbation experiments using mouse-Geneformer, we compered genes predicted in silico perturbation experiments and ones identified in vivo experiments. The data used for evaluation included single-cell data from diabetic nephropathy, UMOD nephropathy, and normal kidney data [33] (Gene Expression Omnibus (GSE): GSE190094), as well as data from cells with knocked out COP1 protein and normal cells [34] (Gene Expression Omnibus (GSE): GSE147559). These datasets were not included in mouse-Genecorpus-20M. To perform disease typing with mouse-Geneformer, the model was fine-tuned for each disease data for this task. Fine tuning conditions are shown in Table 1. Only models achieving disease subtype classification accuracies above 90.00% on the test data were used for evaluation. The data were randomly split into training (80%) and testing (20%) sets. For the in silico perturbation experiments for diabetic nephropathy, we perturbed the cell states of normal kidney

cells to resemble those of diabetic nephropathy. For UMOD nephropathy experiments, we perturb the cell states of cells from UMOD nephropathy to resemble normal kidney cells. In COP1 KO experiments, we perturb the cell states of cells with COP1 knocked out to resemble normal cells.

Gene name conversion between mouse and human for cross-species application

To convert human genes to mouse genes for cross-species application, we used the databases of Mouse Genome Informatics (MGI) [35] and the HUGO Gene Nomenclature Committee (HGNC) [36]. HGNC BioMart application (https://biomart.genenames.org/martform/#!/default/HGNC?datasets=hgnc_gene_mart_2024_03_26) serves the function to provide MGI IDs for mouse homologs of queried human genes. The MGI Batch Summary (https://www.informatics.jax.org/batch/summary) provides a feature to convert Ensembl IDs to corresponding mouse MGI IDs queried. Specifically, the conversion from human genes to mouse genes involved the following steps: 1) converting human Ensembl IDs to mouse MGI IDs using HGNC; 2) converting mouse MGI IDs to mouse Ensembl IDs and Gene Symbols using MGI. When converting human Ensembl IDs to mouse Ensembl IDs, we use a conversion table that ensures a one-to-one correspondence. In cases where the human Ensembl ID to the mouse homolog MGI ID and MGI ID to the mouse Ensembl ID don't have a one-to-one correspondence, the following steps are taken to ensure a one-to-one correspondence between human Ensembl IDs and mouse Ensembl IDs: 1) Sort the gene symbols corresponding to human Ensembl IDs in alphabetical order, and then map the last gene symbol to the mouse homolog MGI ID for the human Ensembl ID; 2) Map the mouse homolog MGI ID to the final mouse Ensembl ID in the output. After converting human Ensembl IDs to mouse Ensembl IDs, we obtain 18,269 genes. There are 910 human Ensembl IDs that have a one-to-many correspondence, and there are 687 mouse homolog MGI IDs that have a one-to-many correspondence.

298

299 **Human cell type classification using mouse-Geneformer**

300 To explored whether mouse-Geneformer could be used for the analyses of other organisms, we
301 analyzed three single-cell RNA-seq datasets derived from three human organs: breast [37] (Gene
302 Expression Omnibus (GEO): GSE195665), thymus [38] (Gene Expression Omnibus (GEO):
303 GSE144870), and cerebral cortex [39] (Brain Initiative Cell Census Network (BICCN):
304 SCR_016152). These datasets were downloaded from CELLxGENE
305 (https://cellxgene.cziscience.com/datasets) and were not included in Genecorpus-30M. To perform
306 human cell type classification using mouse-Geneformer, human genes in the original datasets were
307 converted to their mouse homologs as described in the previous section. The mouse-Genformer and
308 Geneformer were employed to classify each data, and their accuracies are compared. The
309 Geneformer was a human-Geneformer model that we pre-trained using Genecorpus-30M
310 (https://huggingface.co/datasets/ctheodoris/Genecorpus-30M). For cell type classification with
311 mouse-Geneformer and Geneformer, models were fine-tuned for each organ data for this task. The
312 fine-tuning conditions are detailed in Table 1. The procedures of fine-tuning and classification task
313 were the same for both the mouse-Geneformer and human-Geneformer models. The data were
314 randomly split into training (80%) and testing (20%) sets. The evaluation metric for this experiment
315 was accuracy. Cell distribution was visualized using UMAP [30] as described above.

316

317 **In silico perturbation of human data using mouse-Geneformer**

318 To explored whether mouse-Geneformer could be used for the analyses of other organisms, we
319 analyzed human single-cell RNA-seq data from myocardial infarction cells [40] (European
320 Genome-phenome Archive: EGAS00001006330) and COVID-19 human blood cells [41] (Gene
321 Expression Omnibus (GEO), European Genome-phenome Archive:
322 GSE150728,GSE155673,GSE150861,GSE149689,EGAS00001004571). These datasets were

downloaded from CELLxGENE (https://cellxgene.cziscience.com/datasets). To perform in silico perturbation experiments on human cells using mouse-Geneformer, human genes in the original datasets were converted to their mouse homologs as described in the previous section. To perform disease typing with mouse-Geneformer, the model was fine-tuned for each disease data for this task. Fine tuning conditions are shown in Table 1. Only models achieving disease subtype classification accuracies above 90.00% on the test data were used for evaluation. Each dataset was randomly split into training (80%) and testing (20%) sets. The in silico perturbation experiments aimed to alter the cell expression profiles of normal cells to resemble abnormal cells and vice versa. For comparison, these human datasets were also analyzed using the original human-Geneformer model.

The number of cells, cell types, and types of diseases used in this experiment with single-cell data of mice are shown in S2 Table.

Results

Development of mouse-Geneformer

Theodoris et al. originally developed Geneformer, a context-aware, attention-based deep learning model, which was pretrained on large-scale transcriptome data comprising approximately 30 million human single-cell RNA-seq data. While the original Geneformer was tailored for human transcriptome analysis, our study focuses on creating a mouse version, referred to mouse-Genefomer, from mouse transcriptome data. By leveraging the mouse-Geneformer, we aim to harness the capabilities of large-scale deep learning models to enhance research involving mice, the most extensively studied model organism. The architecture and transfer learning strategy employed in developing mouse-Geneformer are the same as those of the original human Geneformer, with

minor modifications. The primary but critical difference lies in the input corpus, which was constructed from mouse single-cell RNA-seq data.

We constructed a large-scale dataset of mouse single-cell RNA-seq data, termed mouse-Genecorpus-20M. The mouse-Genecorpus-20M comprises approximately 21 million raw single-cell data from healthy mice, encompassing a wide variety of organs and developmental stages. The breakdown of the dataset is shown in Fig 2B, 2C and 2D and detailed information is provided in Table S1. It is noteworthy that mouse-Genecopus include types of samples that are difficult to obtain from humans due to ethical or technical constraints (e.g. embryos). Since mouse-Genecorpus-20M is intended for learning the gene network of normal mice, we excluded single-cell data with high mutation loads, such as cancer cells or immortalized cell lines, that could restructure gene networks. We also omitted low-quality RNA-seq data derived from doublets and damaged cells. After the filtration, single-cell transcriptome data from 20,630,028 cells remained, constituting mouse-Genecorpus-20M.

The transcriptome of mouse-Genecorpus-20M was modeled using the rank value encoding method and then processed through six transformer encoder units, as employed in the original Geneformer. This process yielded "cell sequences" represented by gene names as tokens. Pretraining was conducted using a masked language model under the conditions shown in Table 1. We made several modifications to the original Geneformer parameters: 1) We changed the activation function from ReLU to SiLU (Sigmoid-weighted Linear Unit, also known as Swish) because SiLU provides smoother gradients than ReLU. It also provides gradients when the input values are less than or equal to 0, and increases the gradient when the input values are greater than or equal to 1. Therefore, it is less prone to the vanishing gradient problem and can improve the performance in deep neural network models. 2) We increased the number of epochs from three to ten to allow the number of training times on the model, reduced learning loss and improved performance. 3) In comparing linear and cosine scheduling modes for the learning rate scheduler,

we found that pre-training with the cosine schedule showed better performance by reducing learning loss more effectively. Therefore, we chose the cosine schedule. As a result of these procedures, the mouse-Geneformer was constructed.

Cell type classification: mouse-Geneformer is robust and outperforms conventional methods

We investigated the performance of mouse-Geneformer in cell-type classification across various organs. Nine experiments were conducted using different organs to compare the accuracy of mouse-Geneformer with conventional methods such as scDeepSort and scVAE (Table 2). Our findings demonstrated that mouse-Geneformer greatly improves the classification accuracy of cell types in all cases, achieving an average accuracy of 96.73%. In comparison, the GNN-based method scDeepSort and Autoencoder-based method scVAE showed accuracy scores of 66.34% and 72.95%, respectively. Notably, mouse-Geneformer constantly achieved classification accuracies exceeding 93%, regardless of target organs or the number of cell types. UMAP visualization of the tongue and limb muscle samples further supported the good performance of mouse-Geneformer in cell-type annotation (Fig 3A and 3B). Each cell type was clearly separated, although the overlap of the distribution of Langerhans cells with epithelial basal cells posed a challenge in the tongue dataset. These results indicate that mouse-Geneformer is more accurate than existing methods and robust across different organs and varying cellular complexity. This suggests that mouse-Geneformer can be applicable effectively to a wide variety of mouse organs.

397 **Table 2 | Cell type classification results using the mouse-Geneformer and two conventional**

398 **methods.**

399 Cell types is the number of cell types in each data. mouse-Geneforemr, scDeepSort and Single-cell

400 VAE are nine mice tissues classification accuracy. Classification accuracies are shown in

401 percentage (%).

402

Tissue	Cell type	mouse-Geneformer	scDeepSort	single-cell VAE
Tongue	3	94.87	76.79	80.44
Thymius	6	96.65	54.94	74.95
Mammary Gland	7	98.85	47.76	74.25
Large Intestine	7	93.08	49.78	59.00
Limb Muscle	9	99.58	90.82	79.58
Spleen	10	98.56	81.01	76.47
Heart	11	97.56	79.55	79.42
Brain	15	96.64	58.46	76.19
Kidney	18	94.86	58.01	56.25

403

404

405

406

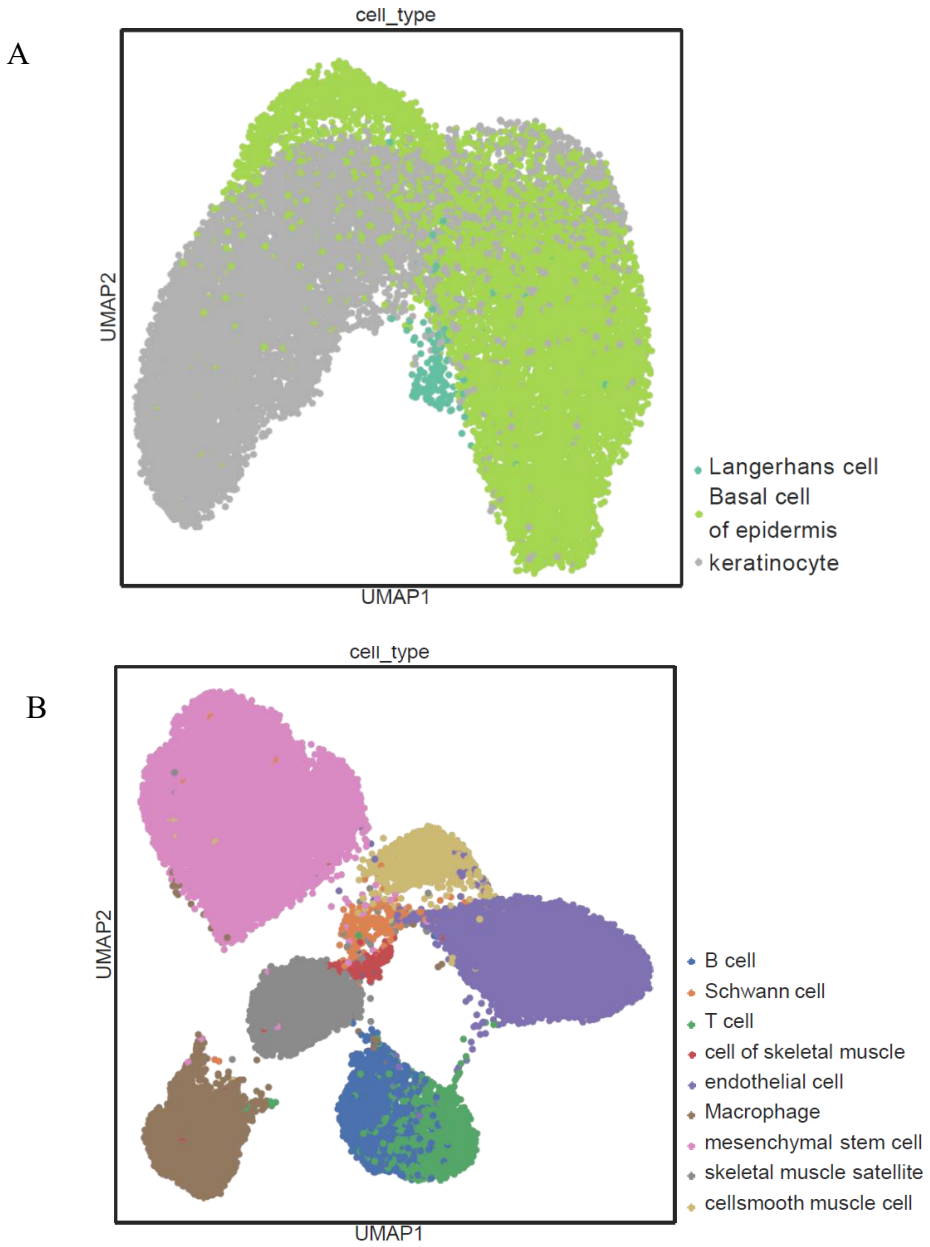


Fig 3 | Visualization of tongue and limb muscle cell distribution. A): Visualization of tongue cell distribution using the mouse-Geneformer. B): Visualization of limb muscle cell distribution using the mouse-Geneformer.

Cell type classification: pretraining is effective

To evaluate the effect of pretraining, we compared the cell type classification results of the mouse-Geneformer with and without pretraining, as shown in Table 3. We conducted three experiments using samples from 1) the prostate gland and urethra, 2) embryo that is whole mouse embryo at embryonic age E9.5, and 3) kidney. Our findings indicated that the classification accuracy of the mouse-Geneformer with pretraining was higher than that without pretraining in all cases. The maximum difference in classification accuracy was 8.22% for the embryo data, while the improvements of the other cases were marginal. UMAP visualization of the cell distribution in the embryo data demonstrated that the pretrained mouse-Geneformer depicted more discrete cell clusters compared to the non-pretrained version as shown in Fig 4A and 4B. For example, primitive erythrocytes and cardiac valve cells are distributed among multiple clusters in the non-pretrained mouse-Geneformer classification, whereas in the pretrained version, they formed distinct clusters. These results suggests that pretraining is effective in enhancing mouse-Geneformer, leading to improved classification accuracy.

Table 3 | Cell type classification results with and without prior learning. Cell types in the number of cell types in each data. w/ prior and w/o prior are mouse-Geneformer with and without prior learning. These methods are three mice tissues classification accuracy. Classification accuracies are shown in percentage (%).

Tissue	Cell type	w/ prior	w/o prior
Prostate gland & Urethra	7	94.68	93.73
Enbryo	9	74.78	66.56
Kidney	10	79.02	78.91

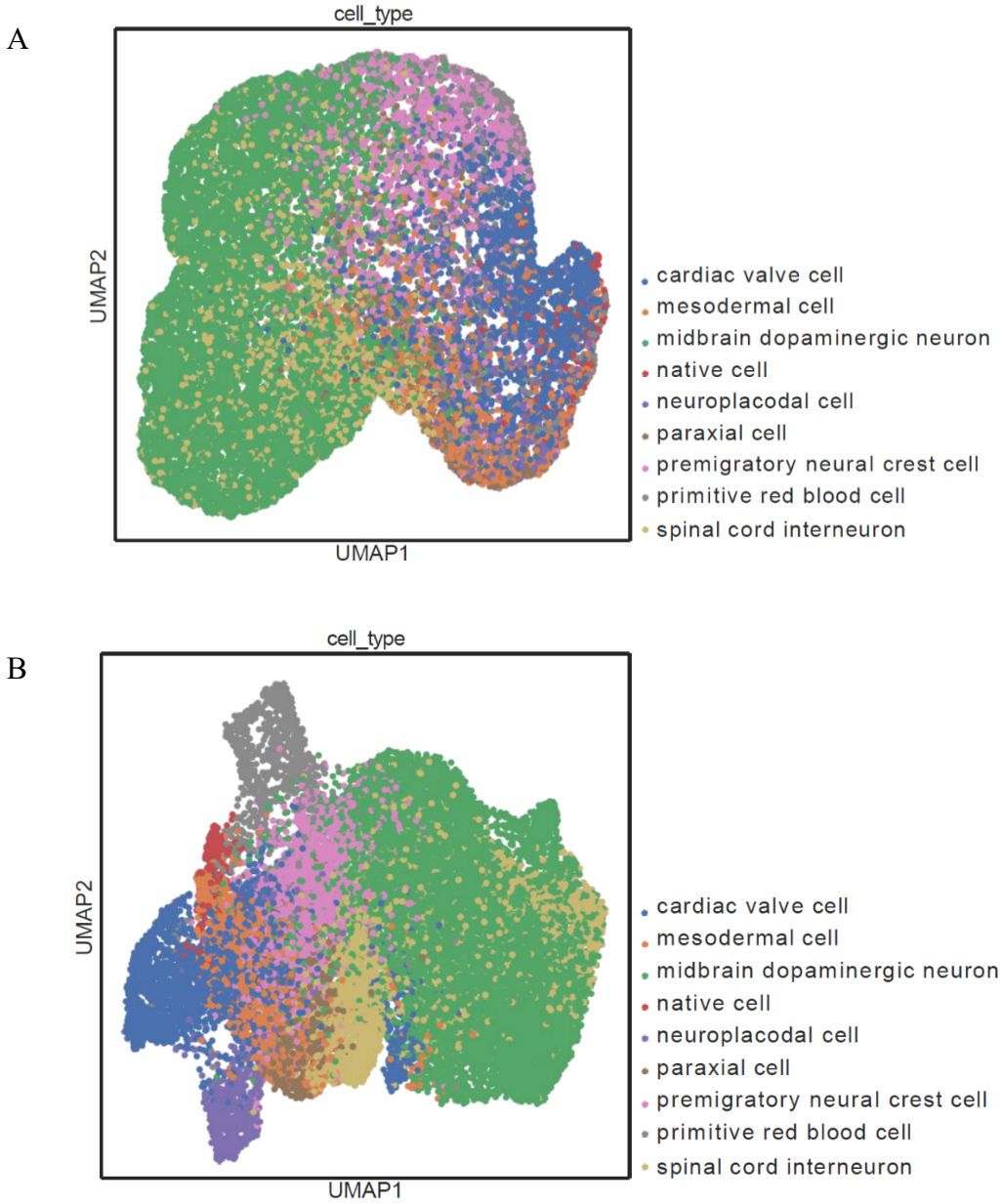


Fig 4 | Visualization of embryo cell distribution. A): mouse-Geneformer without prior learning. B): mouse-Geneformer with prior learning.

In silico perturbation experiments

We next investigated the applicability of mouse-Geneformer to in silico perturbation experiments for mice. In silico perturbation experiments simulate genetic manipulation experiments on computers, allowing us to mimic gene deletions or activations by editing the gene list within each cell. This enables the prediction of how these mutations will impact the gene network and elucidate the functions of the mutated genes. Notably, this approach provides a powerful method for candidate gene screening, as it allows for the repeated simulation of experiments for multiple target genes of interest, or even all mouse genes, in silico. By comparing the effects of these perturbations, we can identify the most influential genes. Detailed information about in silico perturbation experiment is provided in Methods. We here analyzed three disease models using mouse-Geneformer.

In silico perturbation experiment in diabetic kidney disease

We analyzed diabetic nephropathy using a set of single-cell transcriptome data collected from disease kidney exhibiting diabetes and from normal kidneys as a control. Detailed information about data and fine turning are provided in Methods. The UMAP visualization of the cell distribution shown in Fig 5 showed a clear separation between normal kidney cells and disease cells.

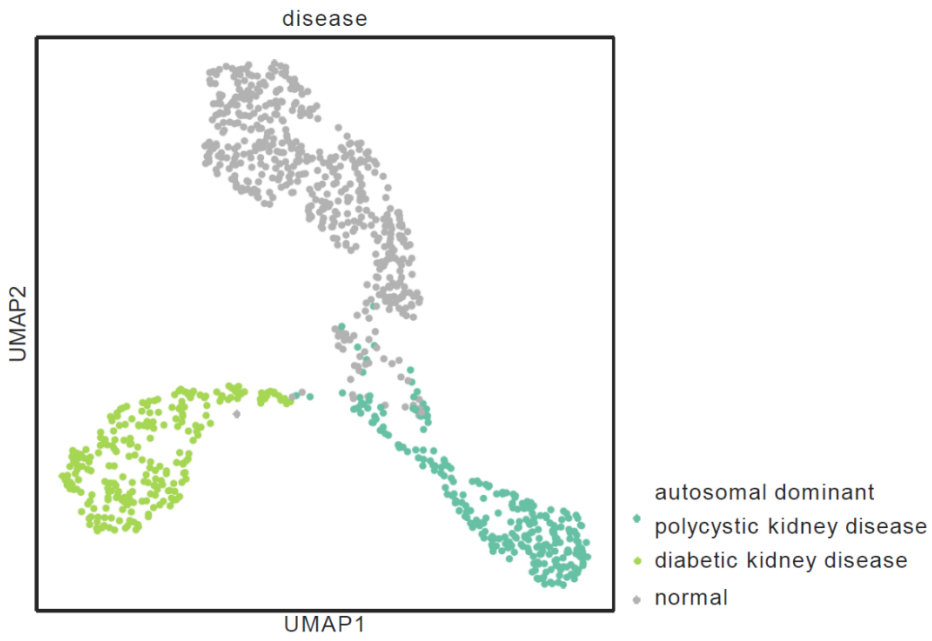


Fig 5 | Visualization of disease cell distribution. In three kidney diseases using mouse-Geneformer.

Subsequently, we conducted in silico perturbation experiments by randomly and repeatedly choosing target genes. Detailed information about in silico perturbation experiment is provided in Methods.

The analysis revealed that deleting the gene *Slc12a3* from normal kidney cells brought the cells closest to those of diabetic kidney disease, with a cosine similarity of 0.018. This in silico outcome is consistent with observations from in vivo experiments, where granular cells of the glomerular apparatus suffered from diabetic kidney disease showed diminished expression of *Slc12a3* [33].

In silico perturbation experiment in UMOD kidney disease

Following the similar approach, we analyzed UMOD kidney disease. We modeled the cell status of UMOD kidney disease by fine-tuning our mouse-Geneformer using the single-cell transcriptome

471 data (Fig 5). In silico perturbation of the disease cells by altering gene expressions revealed that
472 deleting the gene *Slc35b1*, [abundantly] expressed in cells of UMOD kidney disease, had the
473 greatest impact in moving the disease cell status closer to normal kidney cells. In in vivo
474 experiments, it has been reported that the accumulation of mutant UMOD protein in the kidney
475 alters the Unfolded Protein Response (UPR), leading to changes in the expression of 77 genes,
476 including *Slc35b1* and *Slc3a2* that activate the UPR [33]. These observations indicate that our in
477 silico perturbation experiments for UMOD kidney disease successfully identified one of the two
478 genes validated by the in vivo experiment.

479 *In silico perturbation experiment in COP1 KO microgria*

480 Using a similar approach, we analyzed a model of neuroinflammation disease. We utilized single-
481 cell RNA-seq data of microglia cells with COP1 knocked out, a gene known for suppressing
482 neuroinflammation, and normal cells [34]. Detailed information about data and fine turning are
483 provided in Methods. Fig 6 exhibits complete separation between normal cells (Cop1 WT) and
484 COP1 knocked-out cells (Cop1 KO). Then, we conducted in silico perturbation experiments.
485 Deletion of *Apoe* gene from COP1 KO cells demonstrated the closest resemblance to normal cells,
486 with a cosine similarity of 0.36. Additionally, deleting the genes *Fth1*, *Itgax*, and *Cst7* also showed
487 significant impacts in altering the cells closer to normal cells. These observations are consistent
488 with in vivo experiments, which reported that neuroinflammation occurs due to the accumulation
489 of c/EBPβ in the brain when COP1 is knocked out, and that knocking out COP1 increases the
490 expression levels of genes *Apoe* and *Fth1* and affects the expression levels of neurodegeneration-
491 related genes *Itgax* and *Cst7* [34].

492

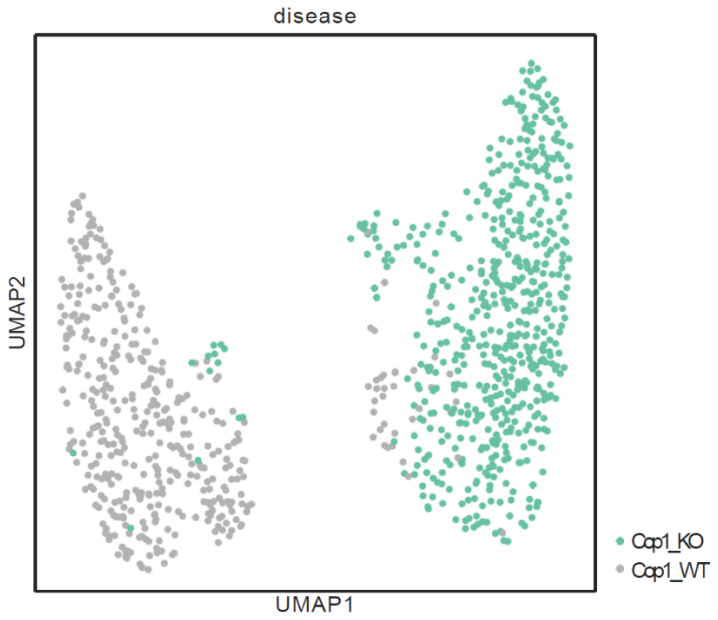


Fig 6 | Visualization of disease cell distribution. In two cop1 microglia diseases using the mouse-Geneformer.

Altogether, in silico perturbation experiments, following the fine-tuning of mouse-Geneformer model using single-cell transcriptome data of target disease model, effectively recapture the results of in vivo experiments.

Cross-species application of mouse-Geneformer through orthologous gene name conversion

We explored whether mouse-Geneformer could be used for the analyses of other organisms. If successful, this would allow non-model mammals, for which it is difficult, costly, and technically challenging to obtain sufficient single-cell transcriptome data to construct species-specific Geneformer models, to benefit from transfer learning using the mouse gene network model. Given the core metabolic and physiological features are conserved among closely related species, we hypothesize that the core genetic architectures would also be conserved. Therefore, mouse-Geneformer could potentially elucidate a significant portion of genetic architectures beyond mice.

509 Even humans could benefit from mouse-Geneformer, despite the existence of the original
510 Geneformer. This is because mouse-Genecopus can include types of samples that are difficult to
511 obtain from humans due to ethical or technical constraints.

512 As a proof of concept for the cross-species application of mouse-Geneformer, we
513 investigated its applicability to human transcriptome analysis. The outline of the procedure was as
514 follows: First, to analyze human genes in mouse-Geneformer, each human gene name was
515 converted to mouse ortholog based on the ortholog table. We then fine-tuned the mouse-
516 Geneformer model using human transcriptome data converted to mouse otrholog. Using this fine-
517 tuned model, we conducted in silico perturbation experiments.

518

519 *Human cell type classification using mouse-Geneformer*

520 We conducted human cell type classification using human single-cell RNAseq data with
521 mouse-Geneformer. Three experiments were conducted using different human organs. Detailed
522 information about how to convert gene and fine tuning is provided in Methods. For comparison, we
523 also analyzed the human data using original human-Geneformer model.

524 The cell type classification results are shown in Table 4. The results demonstrated that mouse-
525 Geneformer accurately classified human cell types based on the human transcriptoem data for all
526 cases, achieving accuracy scores of 95.44%, 99.98% and 88.12% for human thymus, cerebral
527 cortex, and breast, respectively. In addition, the classification accuracy of the ortholog-converted
528 mouse-Geneformer for human data was nearly equivalent to that of the original human-Geneformer,
529 with marginal differences ranging from 0.01 to 0.30%. This indicates that mouse-Geneformer is
530 effective for cell type classification of human cells and suggests that the Geneformer model can
531 work effectively across species following orthology-based data conversion.

532

533 **Table 4 | Cell type classification results after conversion of human and mouse genes**

534 Cell types in the number of cell types in each data. Geneformer and mouse-Geneformer are three
535 human and three mice tissues classification accuracy. h/ is human data. Classification accuracies are
536 shown in percentage (%).

Tissue	Cell type	mosue-Geneformer	Geneformer
h/ Thymus	4	95.59	95.44
h/ Cerebral Cortex	6	99.97	99.98
h/ Breast	12	87.82	88.12

537

538

539 *In silico perturbation of human data using mouse-Geneformer 1: myocardial infarction*
540 Encouraged by the success of applying mouse-Geneformer to human cell type classification, we
541 next investigated its utility for in silico perturbation experiments on human data. We analyzed
542 human single-cell RNA-seq data from myocardial infarction cells to evaluate the model's
543 effectiveness. The original human-Geneformer predicted that activation of *NPPB* and *ANKRD1*, as
544 well as deletion of *MYH7*, could drive normal heart cells towards myocardial infarction state. The
545 mouse-Geneformer similarly indicated that activation of *Ankrd1* and deletion of *Myh7* could lead to
546 a comparable transition from normal heart cells towards myocardial infarction cells. These result
547 aligns with findings from single-cell RNA sequencing analyses of human myocardial infarction and
548 normal heart cells [40]. Conversely, human-Geneformer model predicted that deleting *ANKRD1*
549 and *NPPB* from myocardial infarction cells would revert them to a more normal cellular state.
550 Similarly, the mouse-Geneformer model predicted that deletion of genes *Nppb*, *Ankrd1*, and *Myh7*
551 from myocardial infarction cells would also lead to a transition towards a normal state. Thus, these

552 results demonstrate that mouse-Geneformer can effectively perform in silico perturbation
553 experiments on human disease models like myocardial infarction.

554 *In silico perturbation of human data using mouse-Geneformer 2: COVID-19 human blood*

555 Next, we investigated the effectiveness of mouse-Geneformer in analyzing a human-specific
556 disease. We conducted perturbation experiments on COVID-19 human blood cells using both
557 human-Geneformer and mouse-Geneformer for comparison. Notably, SARS-CoV-2, the
558 coronavirus responsible for COVID-19, does not naturally infect wild-type laboratory mice [42,43].
559 Using the human Geneformer, we predicted that the activation of genes *CCR4*, *IL6*, and *CCR20* in
560 normal human blood cells would lead to a transition towards a state resembling SARS-CoV-2
561 infection. This prediction is consistent with findings from single-cell RNA sequencing analysis of
562 normal blood and COVID-19 blood [41]. The mouse-Geneformer predicted that the activation of
563 genes *Cxcl3* and *Ccr4* would induce a similar transition towards COVID-19-infected cells.
564 Conversely, in silico perturbation of COVID-19 human blood cells using human-Geneformer model
565 predicted that deletion of genes *CXCL2*, *IFITM3*, and *CCL20* would revert the COVID-19 cells
566 towards a normal state. The mouse-Geneformer model predicted that deleting genes *Ccr4* and *Il6*
567 would lead to a transition towards normal cells. These results suggest that both human-Geneformer
568 and mouse-Geneformer predict the involvement of similar genes, such as *CCR4* and *IL6*, in the
569 process of SARS-CoV-2 infection in human blood cells. However, the overlap of the genes
570 predicted by both models was not large. This may indicate limitations in cross-species application
571 of Geneformer model for species-specific traits. Thus, while mouse-Geneformer demonstrates
572 potential in cross-species application for understanding human diseases like COVID-19, it also
573 highlights the importance of species-specific models for capturing the full complexity of disease
574 mechanisms.

575

Discussion

Geneformer is an innovative context-aware, attention-based deep learning model pretrained on large-scale human single-cell RNA-seq data. It enables fine-tuning for a vast array of downstream tasks with limited task-specific data. Building on the successful development of Geneformer using human scRNA-seq data, we here developed a mouse version of Geneformer using mouse scRNA-seq data. The architecture and transfer learning strategy of the original human version were followed with minor modifications. The evaluation of our mouse-Geneformer indicated successful development, as fine-tuning for downstream tasks improved the accuracy of cell type classification in mouse data, and in silico simulations of gene manipulation in mouse disease models detected genes identified in in vivo experiments. This suggests that the architecture of the original human Geneformer, including key components such as rank value encoding, is robust and applicable to species beyond human. It is expected that this strategy and architecture can be applied to any species to build species-specific Geneformer models, provided that large-scale transcriptome data are available.

The mouse, *Mus musculus*, is the foremost mammalian model for studying human biology and disease. Extensive knowledge about mouse physiology, anatomy, and genetics, along with well-developed methods for genetic manipulation – such as creating transgenic, knockout, and knockin animals – has positioned the mouse as a crucial model in various biological and medical fields. In this context, the mouse-Geneformer developed in this study would greatly benefit mouse studies. We constructed a large-scale dataset of mouse single-cell RNA-seq data, termed mouse-Genecorpus-20M, comprising approximately 21 million single-cell RNA-seq profiles from healthy mice, encompassing a wide variety of organs and developmental stages. Pretrained with this dataset, mouse-Geneformer gained a fundamental understanding of the genetic network dynamics of the mouse transcriptome. By leveraging the prior knowledge, the accuracy of cell type classification with mouse-Geneformer has significantly improved compared to traditional methods.

501 Furthermore, in silico perturbation experiments using mouse-Geneformer successfully identified
502 disease-causing genes that have been validated in in vivo experiments. Thus, the mouse-
503 Geneformer not only enhances our ability to understand the genetic network of mice but also
504 enables in silico screening of key genetic factors in disease models. Specifically, in silico prediction
505 using mouse-Geneformer can help prioritize the genes to analyze before conducting animal
506 experiments, thereby avoiding *ad hoc* gene knockouts, saving time, and reducing the need for
507 sacrificing animals.

508 We found that the mouse-Geneformer can be used for the analyses of the other animal
509 species in a cross-species manner. In this study, after the ortholog-based gene name conversion, the
510 analysis of human scRNA-seq data using mouse-Geneformer followed by the fine-tuning with
511 human data achieved cell type classification accuracy comparable to that obtained using the original
512 human Geneformer. Also, mouse-Geneformer effectively performed in silico perturbation
513 experiments on human disease models of myocardial infarction. Given the core metabolic and
514 physiological features are conserved among mammals, the core genetic architectures should also be
515 conserved, thereby the mouse-Geneformer worked for human transcriptome. Despite the existence
516 of the original Geneformer tailored for human, human research could benefit from mouse-
517 Geneformer. This is because mouse-Genecopus can include types of samples that are ethically or
518 technically inaccessible for humans, such as embryonic tissues and certain disease models. As the
519 amount and variety of mouse scRNA-seq data continues to increase, the inclusion of additional
520 datasets into the current mouse-Geneorpus-20M to create an expanded mouse-Genecoupus will
521 enable mouse-Geneformer to learn more accurate gene networks. This enhanced model could serve
522 as a reference not only for mouse but also human studies.

523 The cross-species application of Geneformer holds great potential for the analysis of non-
524 model organisms, for which it is difficult, costly, and technically challenging to obtain sufficient
525 single-cell transcriptome data to construct species-specific Geneformer models. However, since we

investigated only a single combination of cross-species application between human and mouse, it remains unclear how closely related species can be analyzed with Geneformer models in a cross-species manner. The success of such applications may also depend on the traits or genes of interest. We predict that housekeeping metabolic processes involving conserved genes can be analyzed across species, whereas lineage-specific traits may not be as effectively analyzed, as revealed by the COVID-19 experiment in this study. Further research is needed to explore the full potential and limitations of cross-species applications of Geneformer models.

Code and data availability

The code of mouse-Geneformer is available on GitHub at the following URL: [GitHub] https://github.com/machine-perception-robotics-group/Mouse-Geneformer. The mouse-Genecorpus-20M is also publicly available on Hugging Face at: [Hugging Face Repository] https://huggingface.co/datasets/MPRG/Mouse-Genecorpus-20M

Supporting information

- S1 Table: Detailed information of mosue-Genecorpus-20M.
- S2 Table: Overview of data using evaluation experiments

Acknowledgments

We are grateful to all members of the Yamashita Laboratory and the Fujiyoshi Laboratory for their assistance, advice, consultation, and cooperation in obtaining experimental data for this research. This work was supported in part by NIBB Collaborative Research Program (24NIBB462).

Reference

1. Saliba A, Westermann A, Gorski S, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic acids research. 2014;42(14):8845–60. doi: 10.1093/nar/gku555.

2. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic acids research. 2007;35(suppl_2):W339–44. doi: 10.1093/nar/gkm368.

3. Lieberman Y, Rokach L, Shay T. CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. PloS one. 2018;13(10):e0205499. doi: 10.1371/journal.pone.0205499.

4. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nature methods. 2018;15(12):1053–8. doi: 10.1038/s41592-018-0229-2.

5. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nature methods. 2019;16(8):715–21. doi: 10.1038/s41592-019-0494-8.

6. Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nature Machine Intelligence. 2019;1(4):191–8. doi: 10.1038/s42256-019-0037-0.

7. Yang F, Wang W, Wang F, Fang Y, Tang D, Huang J, et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. Nature Machine Intelligence. 2022;4(10):852–66. doi: 10.1038/s42256-022-00534-z.

8. Jia S, Lysenko A, Boroevich KA, Sharma A, Tsunoda T. scDeepInsight: a supervised cell-type identification method for scRNA-seq data with deep learning. Briefings in Bioinformatics. 2023;24(5):bbad266. doi: 10.1093/bib/bbad266.

9. Li S, Guo H, Zhang S, Li Y, Li M. Attention-based deep clustering method for scRNA-seq cell type identification. PLOS Computational Biology. 2023;19(11):e1011641. doi: 10.1371/journal.pcbi.1011641.

10. Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature. 2023;618(7965):616–24. doi: 10.1038/s41586-023-06139-9.

11. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 [cited 2024 Aug 6]. p. 4171–86. Available from: https://aclanthology.org/N19-1423

584 12. Hirose M, Hasegawa A, Mochida K, Matoba S, Hatanaka Y, Inoue K, et al. CRISPR/Cas9-
585 mediated genome editing in wild-derived mice: generation of tamed wild-derived strains by
586 mutation of the a (nonagouti) gene. Scientific Reports. 2017;7(1):42476. doi: 10.1038/srep42476.

587 13. Franzén O, Gan LM, Björkegren JL. PanglaoDB: a web server for exploration of mouse and
588 human single-cell RNA sequencing data. Database. 2019;2019:baz046. doi:
589 10.1093/database/baz046.

590 14. Moreno P, Fexova S, George N, Manning JR, Miao Z, Mohammed S, et al. Expression
591 Atlas update: gene and protein expression in multiple species. Nucleic acids research.
592 2022;50(D1):D129–40. doi: 10.1093/nar/gkab1030.

593 15. Tarhan L, Bistline J, Chang J, Galloway B, Hanna E, Weitz E. Single Cell Portal: an
594 interactive home for single-cell genomics data. BioRxiv [Internet]. 2023 [cited 2024 Aug 6];
595 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10370058/. doi:
596 10.1101/2023.07.13.548886.

597 16. Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, Wetterstrand K, et al. The
598 ENCODE (ENCyclopedia of DNA elements) project. Science. 2004;306(5696):636–40.

599 17. 10x Genomics [Internet]. [cited 2024 Aug 6]. Datasets. Available from:
700 https://www.10xgenomics.com/datasets

701 18. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. Cellxgene: a performant,
702 scalable exploration platform for high dimensional sparse matrices. bioRxiv. 2021;2021–04. doi:
703 10.1101/2021.04.05.438318.

704 19. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based
705 single-cell RNA sequencing data. Gigascience. 2020;9(12):giaa151. doi:
706 10.1093/gigascience/giaa151.

20. Yang S, Corbett SE, Koga Y, Wang Z, Johnson WE, Yajima M, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. Genome Biol. 2020 Dec;21(1):57. doi: 10.1186/s13059-020-1950-6.

21. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. Cell systems. 2019;8(4):329–37. doi: 10.1016/j.cels.2019.03.003.

22. Weber LL, Sashittal P, El-Kebir M. doubletD: detecting doublets in single-cell DNA sequencing data. Bioinformatics. 2021;37(Supplement_1):i214–21. doi: 10.1093/bioinformatics/btab266.

23. Grønbech CH, Vording MF, Timshel PN, Sønderby CK, Pers TH, Winther O. scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics. 2020;36(16):4415–22. doi: 10.1093/bioinformatics/btaa293.

24. Shao X, Yang H, Zhuang X, Liao J, Yang P, Cheng J, et al. scDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. Nucleic acids research. 2021;49(21):e122–e122. doi: 10.1093/nar/gkab775.

25. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature. 2020;583(7817):590–5. doi: 10.1038/s41586-020-2496-1.

26. Dilokthanakul N, Mediano PAM, Garnelo M, Lee MCH, Salimbeni H, Arulkumaran K, et al. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders [Internet]. arXiv; 2017 [cited 2024 Aug 6]. Available from: http://arxiv.org/abs/1611.02648

27. Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax. In 2022 [cited 2024 Aug 6]. Available from: https://openreview.net/forum?id=rkE3y85ee

28. Maddison CJ, Mnih A, Teh YW. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In 2022 [cited 2024 Aug 6]. Available from: https://openreview.net/forum?id=S1jE5L5gl

29. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In 2022 [cited 2024 Aug 6]. Available from: https://openreview.net/forum?id=SJU4ayYgl

30. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction [Internet]. arXiv; 2020 [cited 2024 Aug 6]. Available from: http://arxiv.org/abs/1802.03426

31. Joseph DB, Henry GH, Malewska A, Iqbal NS, Ruetten HM, Turco AE, et al. Urethral luminal epithelia are castration‐insensitive cells of the proximal prostate. The Prostate. 2020 Aug;80(11):872‐84. doi: 10.1002/pros.24020.

32. Sampath Kumar A, Tian L, Bolondi A, Hernández AA, Stickels R, Kretzmer H, et al. Spatiotemporal transcriptomic maps of whole mouse embryos at the onset of organogenesis. Nature Genetics. 2023;55(7):1176–85. doi: 10.1038/s41588-023-01435-6.

33. Marshall JL, Noel T, Wang QS, Chen H, Murray E, Subramanian A, et al. High-resolution Slide-seqV2 spatial transcriptomics enables discovery of disease-specific cell neighborhoods and pathways. Iscience [Internet]. 2022 [cited 2024 Aug 6];25(4). Available from: https://www.cell.com/iscience/fulltext/S2589-0042(22)00367-4. doi: 10.1016/j.isci.2022.104097.

34. Ndoja A, Reja R, Lee SH, Webster JD, Ngu H, Rose CM, et al. Ubiquitin ligase COP1 suppresses neuroinflammation by degrading c/EBPβ in microglia. Cell. 2020;182(5):1156–69. doi: 10.1016/j.cell.2020.07.011.

35. Shaw DR. Searching the Mouse Genome Informatics (MGI) Resources for Information on Mouse Biology from Genotype to Phenotype. CP in Bioinformatics [Internet]. 2016 Dec [cited

752 2024 Aug 6];56(1). Available from:
753 https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpbi.18

754 36. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene
755 Nomenclature Committee (HGNC). Human genetics [Internet]. 2001 [cited 2024 Aug 6];109(6).
756 Available from: https://www.genenames.org/files/PMID11810281.pdf. doi: 10.1007/s00439-
757 001-0615-0.

758 37. Kumar T, Nee K, Wei R, He S, Nguyen QH, Bai S, et al. A spatially resolved single-cell
759 genomic atlas of the adult human breast. Nature. 2023;620(7972):181–91. doi: 10.1038/s41586-
760 023-06252-9.

761 38. Lavaert M, Liang KL, Vandamme N, Park JE, Roels J, Kowalczyk MS, et al. Integrated
762 scRNA-Seq identifies human postnatal thymus seeding progenitors and regulatory dynamics of
763 differentiating immature thymocytes. Immunity. 2020;52(6):1088–104. doi:
764 10.1016/j.immuni.2020.03.019.

765 39. Siletti K, Hodge R, Mossi Albiach A, Lee KW, Ding SL, Hu L, et al. Transcriptomic
766 diversity of cell types across the adult human brain. Science. 2023 Oct 13;382(6667):eadd7046.
767 doi: 10.1126/science.add7046.

768 40. Kuppe C, Ramirez Flores RO, Li Z, Hayat S, Levinson RT, Liao X, et al. Spatial multi-omic
769 map of human myocardial infarction. Nature. 2022;608(7924):766–77. doi: 10.1038/s41586-022-
770 05060-x.

771 41. Jin K, Bardes EE, Mitelpunkt A, Wang JY, Bhatnagar S, Sengupta S, et al. An interactive
772 single cell web portal identifies gene and cell networks in COVID-19 host responses. Iscience
773 [Internet]. 2021 [cited 2024 Aug 6];24(10). Available from:

774 https://www.cell.com/iscience/fulltext/S2589-0042(21)01083-X. doi:
775 10.1016/j.isci.2021.103115.

776 42. Jiang RD, Liu MQ, Chen Y, Shan C, Zhou YW, Shen XR, et al. Pathogenesis of SARS-
777 CoV-2 in transgenic mice expressing human angiotensin-converting enzyme 2. Cell.
778 2020;182(1):50–8. doi: 10.1016/j.cell.2020.05.027.

779 43. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak
780 associated with a new coronavirus of probable bat origin. nature. 2020;579(7798):270–3. doi:
781 10.1038/s41586-020-2012-7.

782