

Transfer learning compensates limited data, batch effects and technological heterogeneity in single-cell sequencing

Youngjun Park, Anne-Christin Hauschild* and Dominik Heider^{ID*}

Data Science in Biomedicine, Faculty of Mathematics and Computer Science, Philipps-University of Marburg, Marburg 35039, Germany

Received July 31, 2021; Revised October 07, 2021; Editorial Decision October 11, 2021; Accepted October 18, 2021

ABSTRACT

Tremendous advances in next-generation sequencing technology have enabled the accumulation of large amounts of omics data in various research areas over the past decade. However, study limitations due to small sample sizes, especially in rare disease clinical research, technological heterogeneity and batch effects limit the applicability of traditional statistics and machine learning analysis. Here, we present a meta-transfer learning approach to transfer knowledge from big data and reduce the search space in data with small sample sizes. Few-shot learning algorithms integrate meta-learning to overcome data scarcity and data heterogeneity by transferring molecular pattern recognition models from datasets of unrelated domains. We explore few-shot learning models with large scale public dataset, TCGA (The Cancer Genome Atlas) and GTEx dataset, and demonstrate their potential as pre-training dataset in other molecular pattern recognition tasks. Our results show that meta-transfer learning is very effective for datasets with a limited sample size. Furthermore, we show that our approach can transfer knowledge across technological heterogeneity, for example, from bulk cell to single-cell data. Our approach can overcome study size constraints, batch effects and technical limitations in analyzing single-cell data by leveraging existing bulk-cell sequencing data.

INTRODUCTION

The development of next-generation sequencing (NGS) enables a systematic measurement and analysis of biological questions. During the last decades, a massive amount of biological data are produced. However, at the forefront of the

research field, there are still a lack of data issues. The underlying problem arising is that we cannot fully facilitate existing data for new inquiries. There are different reasons for that (i) each of the datasets is generated under the unique experimental setup, i.e. biological heterogeneity, (ii) the batch effect adds noise onto the data in the same experimental setups (1) and (iii) as new sequencing technologies are developed, technological heterogeneity is introduced (2). Consequently, there is a lack of methodology to integrate various NGS studies. We need a new methodology to facilitate integrative analysis with pre-existing large-scale biological data to newly produced data for complex biological questions. Thus, the availability of the end-to-end learning with multiple datasets can accelerate this integrative analysis of various data sources (3).

Recently, various machine learning has been introduced to handle biological and biomedical data heterogeneity and integrate different datasets to address novel medical and biological questions. Especially in single-cell omics, these new approaches are used to overcome the technical limitations in this field. Gene regulatory network inference study with single-cell sequencing data showed that the data integration with additional genomic data can overcome technical limitations and improve inference results (4,5). Stumpf *et al.* showed the potential of inter-species knowledge transfer for single-cell sequencing data analysis (6). BERMUDA introduced autoencoder-based approaches to remove batch effects between datasets in a similar context (7). Mieth *et al.* obtained improved clustering results by applying a transfer learning model to a small single-cell sequencing dataset. They transferred knowledge from a large well-annotated source dataset to an unannotated small dataset (8). In contrast, MARS applied a meta-learning scheme in a single-cell dataset (9). They trained a deep neural network model with a well-annotated large-scale single-cell sequencing dataset from the TabulaMuris consortium (10) to learn cell type classification ability. In this study the meta-training step improved cell-type classification power and enabled to annotate unknown cell types. Interestingly, the meta-training

*To whom correspondence should be addressed. Tel: +49 6421 2821579; Email: dominik.heider@uni-marburg.de
Correspondence may also be addressed to Anne-Christin Hauschild. Email: hauschild@uni-marburg.de

dataset does not have to contain the same cell types as the test dataset. Meta-training aims to pre-train a model to set initial parameters to learn efficiently from a new, small dataset (11). This meta-learning approach has already gained attention in cancer genomics, where many large-scale datasets have already been generated compared to other fields (12). Other studies utilized well-known TCGA, CCLE (13) and functional genomics datasets as a source dataset for meta-training to perform survival analysis (14) and investigate cancer-drug discovery (15).

Although various studies proposed meta-learning and transfer learning approaches to overcome technical limitations in biological and biomedical data (16), the methodology to overcome technological heterogeneity is not well addressed yet. To fully facilitate our accumulated knowledge in a public database, we must utilize various datasets across different technologies, for example, bulk-cell sequencing to single-cell sequencing. In this study, we aim to analyse the potential of meta-transfer learning approaches to overcome both data heterogeneity and extremely small sample sizes. Therefore, we train a few-shot learning model to distinguish different expression patterns with only a few examples for unique class labels, i.e. cell or organ types. Meta-training datasets can be large scale NGS datasets about different tissues or organs' gene-expression profiles or various case-control datasets. Similar to a few-shot image classification tasks, we expect that the model can recognize unique patterns in a gene-expression profile during meta-training with large-scale and multi-label datasets (17). Furthermore, we expect to transfer this learned knowledge to different classification tasks from different technologies (18). Consequently, we show that the model can learn how to handle classification tasks in high-throughput omics biology. This few-shot learning model with a meta-learning algorithm can build a reliable classifier with a marginal amount of data and overcome batch effects and technological heterogeneity in single-cell sequencing projects. In the following, we refer to this few-shot classification model as a transferable molecular patterns (TMP) recognition model.

MATERIALS AND METHODS

Related works

Few-shot learning. Few-shot learning (FSL) has been used in various ways in different fields. Wang *et al.* summarized FSL studies in three main areas: (i) training data augmentation by prior knowledge, (ii) constraining hypothesis space by prior knowledge and (iii) altering search strategies in hypothesis space by prior knowledge (19). This work will focus on few-shot learning for altering search strategies in hypothesis space by prior knowledge. Our TMP model is based on relation networks (20); however, we further fine-tuned the model with an additional training dataset in the single-cell sequencing data study (Figure 1).

Meta-transfer learning. Meta-transfer learning (MTL) is introduced by Sun *et al.* to leverage few-shot classification accuracy with a pre-trained deep neural network (21). They integrated meta-learning and transfer learning to search hyperparameters effectively. Sun-MTL added lightweight neuron operations for embedded feature vectors, scaling

and shifting, and fine-tuned it for effective transfer learning. In Sun-MTL, the embedding layer is fixed after pre-training with the entire data and only fine-tuned the classifier to prevent the model from overfitting (22,23).

In a 'transfer' point of view, Sun-MTL uses weights of deep neural network (DNN) models pre-trained on a large-scale dataset for another task. TMP also uses a large-scale dataset for the pre-training step on the model. Compared to the Sun-MTL, the TMP uses additional training data during a fine-tuning phase for a special purpose, transferring knowledge across different technological heterogeneity. In a 'meta' point of view, Sun-MTL has a base learner with two light-weight neuron operations. Those are hyperparameters trained on few-shot learning tasks. Instead of the base learner, TMP trains a relation network, another DNNs, as a classifier considering non-linearity on embedded feature vectors. It is an equivalent part to a base learner with the two operations in Sun-MTL. Those neural networks are also trained on few-shot learning tasks. Unlike Sun-MTL, the TMP-model fine-tunes not only the relation network but also the feature extractor.

Nevertheless, Sun-MTL and TMP have a similar concept in terms of both transfer- and meta-learning. They are, however, different in implementation details for domain-specific applications. Here, we will use Sun-MTL for the specific model of Sun *et al.* and meta-transfer learning as a general term to describe the model training scheme used in TMP and Sun-MTL (21). Our TMP is the first application of the meta-transfer learning concept on similar tasks with different datasets generated from different omics technologies.

Implementation few-shot learning model with meta- and transfer-learning

Transferable molecular pattern recognition model network structure. Our network model has two trainable networks, the feature encoder and the relation network (20). Feature encoder blocks are composed of a fully connected layer, a batch normalization layer and a ReLU layer. Feature encoder for a gene expression profile of 18 000 genes is composed of three blocks of feature encoders. Additionally, the feature encoder is a non-negative network by clamping negative weight to zero during iterations to focus on co-expressional patterns rather than other regulatory mechanisms. This approach is inspired by work with non-negative kernel autoencoders (24). The relation network block is composed of three blocks. Two blocks are composed of a fully connected layer, a batch normalization layer, and a ReLU layer. The last block is composed of a fully connected layer and a sigmoid function for the output value.

The feature encoder ($f(x)$) has three encoding blocks. Each of the fully connected layers is set to 4000, 2000 and 1000 neurons. The relation network ($r(x)$) is set to 500 and 100 neurons. Consequently, the 18 000 genes are embedded into 1000 length size vectors through the feature encoder, and the relation network gets two feature vectors for comparison.

In training and evaluation, we used the same design with a support set $S = \{x_i, y_i\}_{i=1}^{C \times K}$ and a test (or query) set $T = \{x_j, y_j\}_{j=1}^T$ from given dataset. We randomly selected five

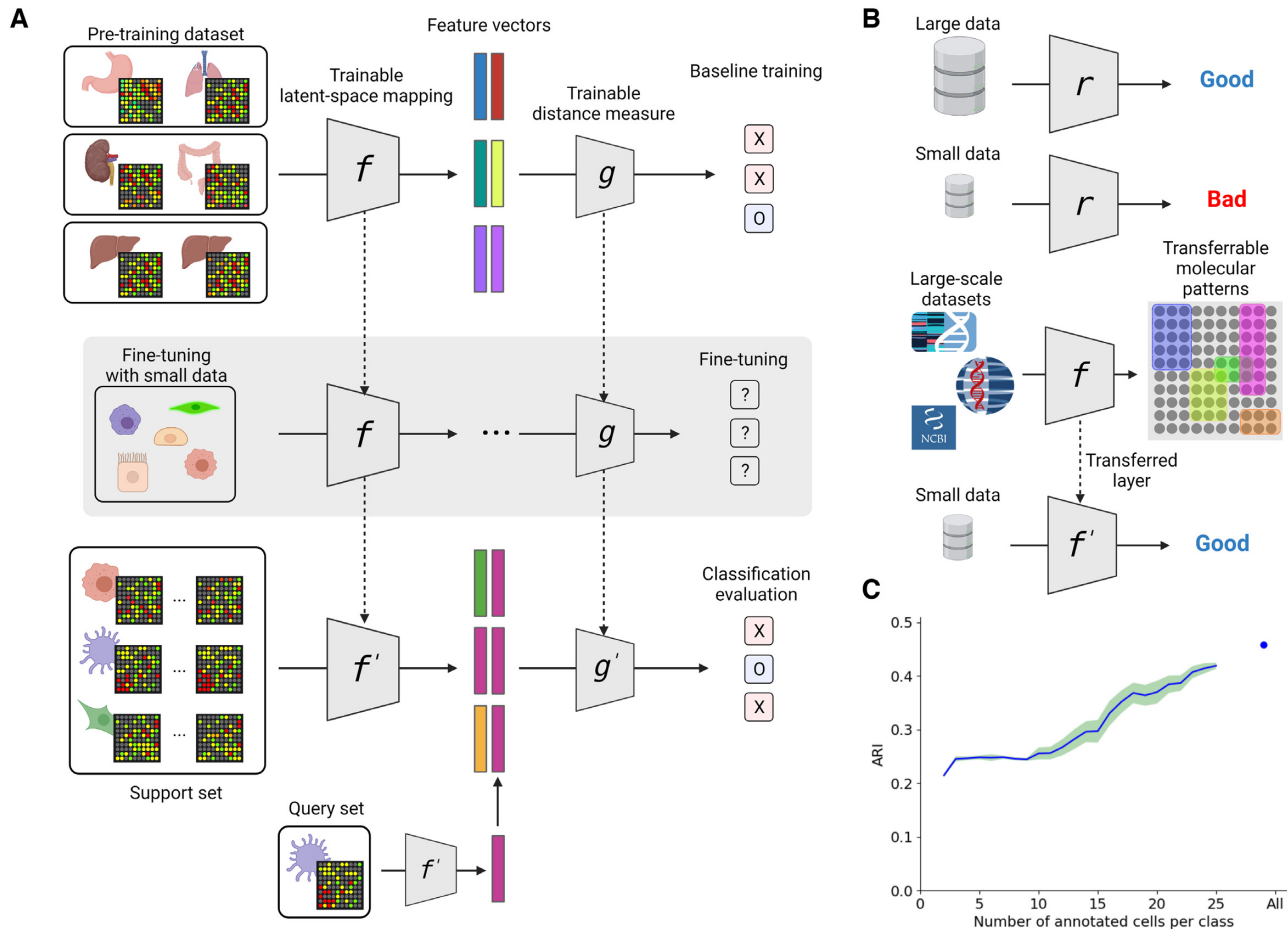


Figure 1. Overview of our TMP-model scheme. (A) Training and testing schemes of our model. Pre-training dataset is a large scale bulk-cell sequencing dataset and fine-tuning dataset is a single-cell sequencing dataset. (B) Hypothesis of this study. Pre-training with large-scale datasets can improve models based on small datasets and different sequencing technologies. (C) Adjusted rand indexes (ARI) for PCA-based clustering in cell type classification problems in single-cell sequencing datasets. The blue line is an average score, and the light green area is 95% confidence interval. Created with [BioRender.com](https://www.biorender.com/).

classes ($C = 5$) for each training iteration and five samples for each class ($K = 5$) for a support set. In the same class set, we randomly selected ten samples ($T = 10$; training and testing batch sizes) for a training/testing batch set. With these datasets, loss is defined as follows:

$$L = \sum_i^{C \cdot K} \sum_j^T mse[r(f(x_i), f(x_j)), 1(y_i == y_j)] \quad (1)$$

After each iteration, we clamped negative weights in the feature encoder to zero. We use the Adam optimizer with a learning rate of 0.0005, and every 100 000 epochs decreased the learning rate by half. We stopped pre-training the model at 300 000 epochs and transferred this network onto different datasets. For fine-tuning, we used the Adam optimizer with a learning rate of 0.0001. The results reflect the evaluation scores at 100 000 epochs of fine-tuning.

Few-shot training with GTEx. The GTEx dataset was used as a pre-training dataset in both TCGA data analysis and single-cell pancreas datasets analysis. To ensure that pre-training has been carried out with sufficient samples among different labels, we held out tissue with <100 samples, i.e.

bladder (21), cervix ectocervix (9), cervix endocervix (10), fallopian tube (9), kidney medulla (4) and kidney cortex (85). We randomly selected five tissues during the training phase and selected five examples as support set and ten samples as training batch query set.

Few-shot training with TCGA. To investigate a few-shot learning task in omics data, we used a pre-trained model with the GTEx dataset and fine-tuned it with the subset of TCGA data. The fine-tuning dataset (15 samples per class) was randomly selected and tested with the others. The accuracy is obtained from 5000 episodes with 10 test batch sizes. To further investigate the data quantity issue, we evaluated various classes and numbers of data out of 33 cancer types and their samples. Because some of the classes in the TCGA dataset has <75 samples, we excluded the cancer types with <100 samples in the training step. Training with the TCGA dataset was done until 50 000 epochs with step size 1500. In every 2500 epochs, we checked batch accuracy and early stopped the training. If the average batch accuracy of 2500 epochs is >0.99, the fine-tuning was stopped. Evaluation is done with 1000 episodes with 10 test batch sizes. We did ten repeats with random seed numbers from 1 to 10.

Few-shot training with single-cell pancreas data. For the batch effect and technological heterogeneity task, five datasets were split such that one dataset was used for training, and the remaining datasets were held for evaluation. We removed cells which have not enough samples for training and testing in a 5-way 5-shot task, namely t_cell (7), schwann (13), epsilon (18) in Baron, epsilon (3), unclear (4) in Muraro, MHC class II cell (5), mast cell (7), epsilon cell (7), unclassified cell (2) in Segerstolpe, acinar (6), delta (9) in Wang, PP.contaminated(8) and delta.contaminated(9) in Xin. In fine-tuning, we selected 15 samples per class in the training datasets. The evaluation has been carried out by 5000 episodes with 10 test batch sizes. Average and 95% confidence interval are represented. Raw values are available in the GitHub repository. Each sample was randomly selected, thus we repeated this training ten times and reported the distribution of the average value over these replicates. Additionally, the average value and 95% confidence interval for each run are available in the GitHub repository.

Data preprocessing

We downloaded processed data of TCGA, GTEx and five single-cell datasets (section in Data Availability). We removed the ambiguous label, 'not applicable', in the Segerstolpe dataset. Depending on its technological origin, each dataset has a different set of genes. Thus, we need to fix the gene list for further analysis. A total of 18 000 genes are selected from the intersection of TCGA, GTEx and StringDB. Because our goal is not mining novel genes for cell-type annotation, StringDB is applied as the simplest feature selection method to control memory usage and training speed. Different gene-IDs are converted to gene symbols with the Biomart package, and missing genes in the independent dataset are filled with 0. After filtering the gene list, we Log2-normalized the gene expression levels and re-scaled it to a 0-1 scale.

Conventional machine learning methods

We used the Python package scikit-learn <http://scikit-learn.org/stable/index.html> to implement conventional machine learning methods.

For PCA and k-means clustering, we randomly selected five classes and a small subset of data per class from 2 to 25 for training. All other data points are projected on the latent space, and we did k-means clustering to find five clusters. Because k-means is starting with random initialization, we did 100 repeats for each data point. The performance is evaluated with an adjusted rand index (ARI), and average ARI with 95% confidence intervals are shown.

RESULTS

Meta-transfer learning for recognizing transferable molecular patterns and handling of small-size data

In this study, we used a two-step learning phase, namely the pre-training phase and the fine-tuning phase in Figure 1A. In the pre-training phase, the model learns how to recognize gene expression patterns from large-scale transcriptome data. The data for fine-tuning phase does not have to

be in the same context or the same technology compared to the data used for pre-training phase. However, it needs to be within a similar context or similar sequencing technology for adjusting batch effects or technological heterogeneity.

The machine learning model requires a reasonable amount of data for the training phase to ensure a good performance in terms of accuracy (Figure 1B). In the single-cell sequencing data field, many approaches use dimensionality reduction methods to overcome the technical limitation of single-cell sequencing itself. For example, recent studies employed PCA (principal component analysis), tSNE (t-distributed stochastic neighbor embedding) or UMAP (uniform manifold approximation and projection). Other recent methods are based on autoencoders and intend to reduce noise and batch effects and embed single-cell gene expression profiles into feature vectors with reduced size. Those approaches can be very effective when the given dataset has well-balanced and large enough samples. In our study, we used PCA and k-means clustering with single-cell data to evaluate the impact of data size on a model. We aimed at keeping this analysis as simple as possible and thus we did not use transductive inferences such as tSNE or UMAP. We used a randomly picked small dataset to calculate the principal components based on human pancreas single-cell data. After that, all samples are projected to the latent space and clustered subsequently (details in the Materials and Methods). Until data size reaches ten samples per class, the ARI (Adjusted rand indexes) score did not change. When more data are used for finding the latent space, the k-means clustering results gradually improved (Figure 1C). In the following, we will focus on the 'small-size data' condition and show that meta-transfer learning can improve a model having limited data, which is a crucial limiting factor in many areas of biomedical research.

Overview of our TMP

TMP is based on a relation network model for few-shot image classification (20) that comprises two trainable DNN. The first network projects the given gene expression profile into the feature vector, and the second network measures the distance between two given feature vectors. Both DNNs are trained by backpropagation to find proper latent spaces and relevant distance measures for the latent spaces. Given large-scale multi-class gene expression profile datasets, TMP learns how to embed given gene expression data and measure distances between data points in latent space. Afterward, this pre-trained model is fine-tuned with a small amount of data from the target task (Figure 1A). Details about the deep neural networks can be found in the Materials and Methods section.

In this study, we pre-trained the model with GTEx data, i.e. human tissue-level gene expression profiles. Based on this pre-trained model, we adapted it for two different datasets: (i) TCGA cancer dataset to analyze the few-shot learning model concerning its ability to recognize transferable molecular patterns and (ii) to human single-cell pancreas datasets to show the possibility of cross-modal data integrative analysis. TMP can learn gene expression patterns from GTEx and thus can be directly applied for other datasets and tasks. Furthermore, with these adjusted pa-

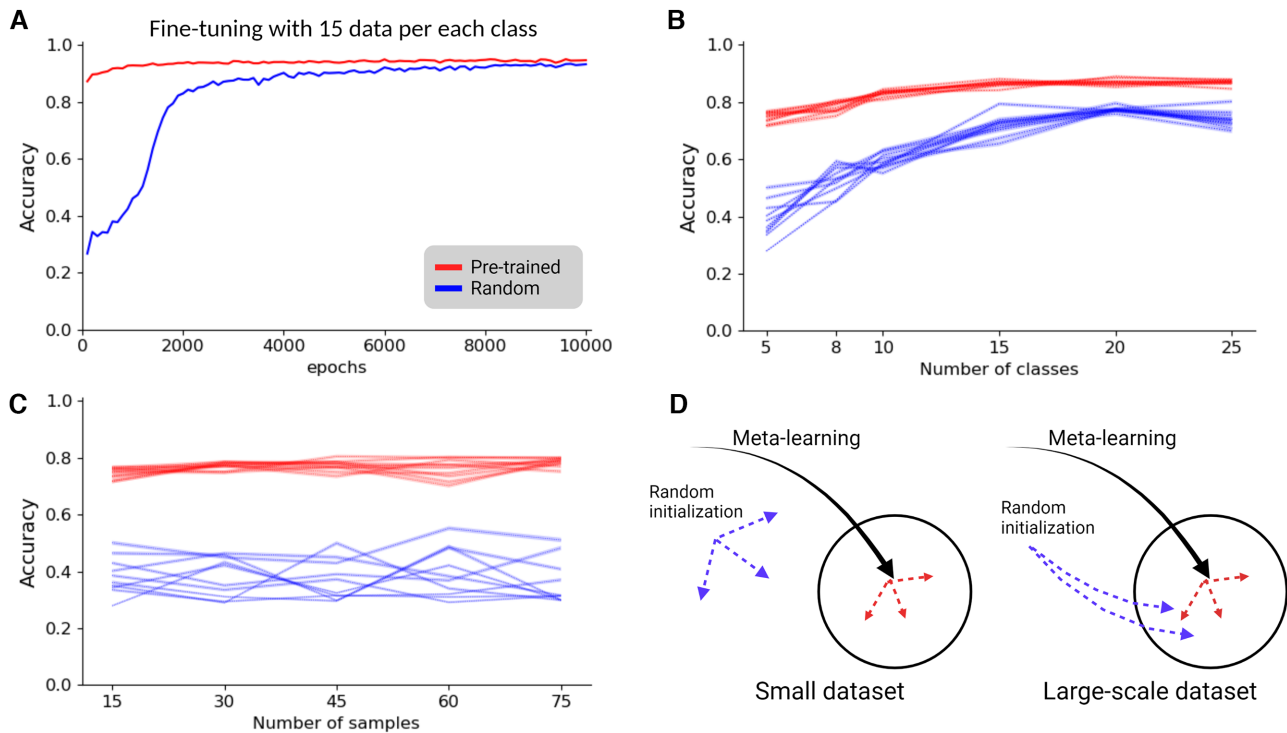


Figure 2. An accuracy chart of 5way-5shot task with TCGA dataset. (A) An 5-way 5-shot accuracy chart during fine-tuning with TCGA pan-cancer dataset. Both red and blue lines used 15 samples for each of 33 different cancer types. The red line is pre-trained with GTEx dataset, and the blue line is randomly initialized. (B) Accuracy score varying number of classes during fine-tuning. Red lines are pre-trained models and blue lines are randomly initialized models. Number of samples is fixed to 15. Dotted line is average score of tests and colored are is a 95% confidence interval. (C) Accuracy score varying number of samples during fine-tuning. Red lines are pre-trained models and blue lines are randomly initialized models. Number of classes is fixed to five. Dotted line is average score of tests and colored are is a 95% confidence interval. (D) Illustration to compare learning processes for both variables, pre-training and given data sizes in fine-tuning. Created with [BioRender.com](https://www.biorender.com).

rameters, the model can quickly be adapted to a new task. Our approach demonstrates that batch effects or technological heterogeneity in single-cell omics can be handled and compensated with TMP.

Transfer learning can improve classifiers built with small-size data

Our few-shot learning model aims to solve a variety of classification problems with limited data. We transferred knowledge from a large-scale dataset that is not directly related to the classification problem to achieve the goal. To investigate the general performance of the few-shot learning model on the transcriptome dataset, we used TCGA and GTEx datasets which are large-scale omics with multi-class labels datasets. TCGA has a human organ-level class label, and GTEx has a human tissue-level class label. Thus, the two datasets are in a similar context but have different class label sets, which makes them ideal candidates for few-shot learning (19). First, we trained our model with the GTEx dataset. With this pre-training step, the model learns to classify different types of tissues by gene expression patterns. Without any additional fine-tuning, the GTEx pre-trained model shows $78.91\% \pm 0.76\%$ accuracy on TCGA. The TCGA pre-trained model shows $84.57\% \pm 0.66\%$ accuracy on the GTEx data in a 5-way 5-shot task (Supplementary Table S3). Next, we compared the cancer type classification

problem with recently published methods. We fine-tuned the model with only 15 samples per class and tested with the remaining (495 samples out of 9781; 5% of the dataset). Although the accuracy of the model can be affected by the randomly chosen 15 samples, we could obtain an average accuracy of 94%+ in 10 repeated training runs, even though we used only 5% of the TCGA data for fine-tuning (Figure 2A). Although accuracy in few-shot learning is not directly comparable to standard models, our approach shows similar performance (95.6% and 95.7% accuracy) compared to recent CNN-based classification models that used 80% of the data for training (25).

When using all 33 classes in the TCGA cancer dataset, the pre-trained model converged early in comparison to the randomly initialized model which needs more epochs to converge. However, fully trained both models have similar accuracy (Figure 2A). With this fully supervised setup, we can see that transferred knowledge can reduce search space and help to reach early convergence. To investigate transferred knowledge as general information that can enhance the few-shot learning model, we trained the model with a subset of the samples of the 33 classes and tested with all remaining samples. We compared two baseline models: pre-trained with GTEx (i) and a randomly initialized model (ii) and further trained them with 5, 10, 15, 20 and 25 classes out of the 33 classes. This fine-tuning is stopped when batch accuracy reached 99% to avoid overfitting (Materials and

Methods section). With this result, we can observe that transferred knowledge is very useful when fine-tuning data are limited to a small number of different classes. With a pre-trained model, we can build a few-shot model with 80%+ accuracy with only 1.5% of the TCGA dataset (10 out of 33 classes and 15 samples per class). Because the meta-learning algorithm is efficient with small data sizes, in the cancer datasets, a more critical factor in better accuracy was the number of classes than the number of samples. When we fixed the number of classes to five and varied the number of samples from 15 to 75, the accuracy change was not significant (Figure 2B,C). With the ARI measure, it shows similar patterns (Supplementary Figure S1). In the case of the randomly initialized models, model parameters are determined by a small size dataset randomly selected during fine-tuning. Thus, they show higher variance than a pre-trained model fine-tuned with a small data size in the same way. However, if the number of classes is 25, we can see marginal improvement when the number of samples is increased (Supplementary Figure S2). The analysis of the TCGA and GTEx datasets, demonstrates that few-shot learning with a meta-learning algorithm can be a new method to facilitate existing large-scale datasets and transfer knowledge (see Supplementary Figure S3). Model initialization with transferred knowledge can improve model accuracy and also reduce training times (Figure 2D).

Meta-transfer learning across technological heterogeneity

TCGA and GTEx can not fully reflect the current technological heterogeneity issue in omics biology. Thus, we further investigated a single-cell sequencing transcriptome dataset. Various methods have been introduced for effective batch correction and integrative analysis with multiple datasets from various techniques in this field. In our study, we used five different single-cell datasets from the human pancreas and the GTEx and TCGA datasets as pre-training datasets. Pre-trained model with both large-scale datasets shows accuracies of <50% accuracy on single-cell sequencing datasets (Supplementary Table S3). Based on this test results from few-shot baseline models with GTEx and TCGA datasets, we could show that the models can identify different cell types from single-cell sequencing datasets even though a model has been trained only as a organ/tissue level classification task. However, test classification result shows that the pre-trained models recognize same class label in different dataset to different class label (Supplementary Figure S4).

Next, we used the Baron dataset out of five different single-cell datasets for fine-tuning of the GTEx-pretrained model to validate that our model can transfer knowledge from a large-scale bulk-cell sequencing dataset to a single-cell problem. We obtained accuracy from five different training conditions (Figure 3A): pre-trained and fine-tuned with the Baron dataset (filled circle), randomly initialized and trained with the Baron dataset (hollow circle), pre-trained and fine-tuned with subset of the Baron dataset (filled square), randomly initialized and trained with subset of the Baron dataset (hollow square) and pre-trained (filled diamond). Similar to the results from the TCGA data, pre-training does not significantly affect the accuracy if a fine-

tuning dataset has enough data. However, if the fine-tuning data are limited, pre-training can improve the quality of the model significantly. During training with Baron data, we observed that the accuracy converges within 100 000 epochs (Figure 3B). In these single-cell datasets, the classification accuracy in other similar single-cell sequencing datasets is improved if a fine-tuning training uses >15 samples from Baron (Supplementary Table S4). This improvement of accuracy was directly observable in the detailed result of test episodes. After learning the cell types classification problem with the baron dataset, the few-shot model was able to distinguish different cell types more accurately. Furthermore, we can identify groups of similar cell types across different datasets (Supplementary Figure S5). Notably, we could observe interesting classification results in the heatmap such as 'alpha-contaminated' and 'beta-contaminated' in the Xin dataset, or 'dropped' in Wang dataset, 'co-expression cell' and 'unclassified endocrine cell' in the Segerstolpe dataset. Because few-shot model can compare query sample to more class labels at one time in a 20-way and 5-shot task, we can see distinct clusters of similar labels and cell clusters that are not clearly annotated.

We further investigated the fine-tuning with small-size data, here 15 data per class (see Figure 3A). Thus, we compared all possible scenarios with single-cell datasets. We chose one dataset out of five as a fine-tuning dataset and tested its 5-way 5-shot task accuracy with the others. Pre-trained models and randomly initialized models are fine-tuned with 15 samples per class from the chosen dataset. Because the data quality of randomly picked 15 cells can significantly affect the training, we repeated the training 10 times and tested for significant differences. It turned out that the pre-trained model is significantly better than the randomly initialized model when the available training data are limited (Figure 3C).

DISCUSSION

As next-generation sequencing is routinized in experimental biology, the speed of data accumulation has been accelerated. However, most studies can utilize only partial data because the biological heterogeneity, batch effects and technological heterogeneity, which hinder integrative analysis. Thus, every study has to produce large amounts of data to obtain a significant result. Few-shot learning with meta-learning algorithms could solve this issue in biomedical research and has already been used in other areas, for example, computer vision (19). Thus, we propose that a few-shot learning model with a meta-learning algorithm can be used as a new methodology to exploit existing public resources and data to build reliable models for small sample size. This approach is not only cost-efficient but also time-efficient, and thus, very important for biomedical studies, for example, for cancer. In a recent cancer-drug discovery study (15), meta-learning has been used for functional genomics data to extract biological knowledge from very small sizes of screening data. Consequently, by using this new approach, we can investigate biological systems in higher resolution and lower costs (12).

Similar to Sun-MTL (21) in the computer vision field, the TMP also showed that knowledge transfer from large-

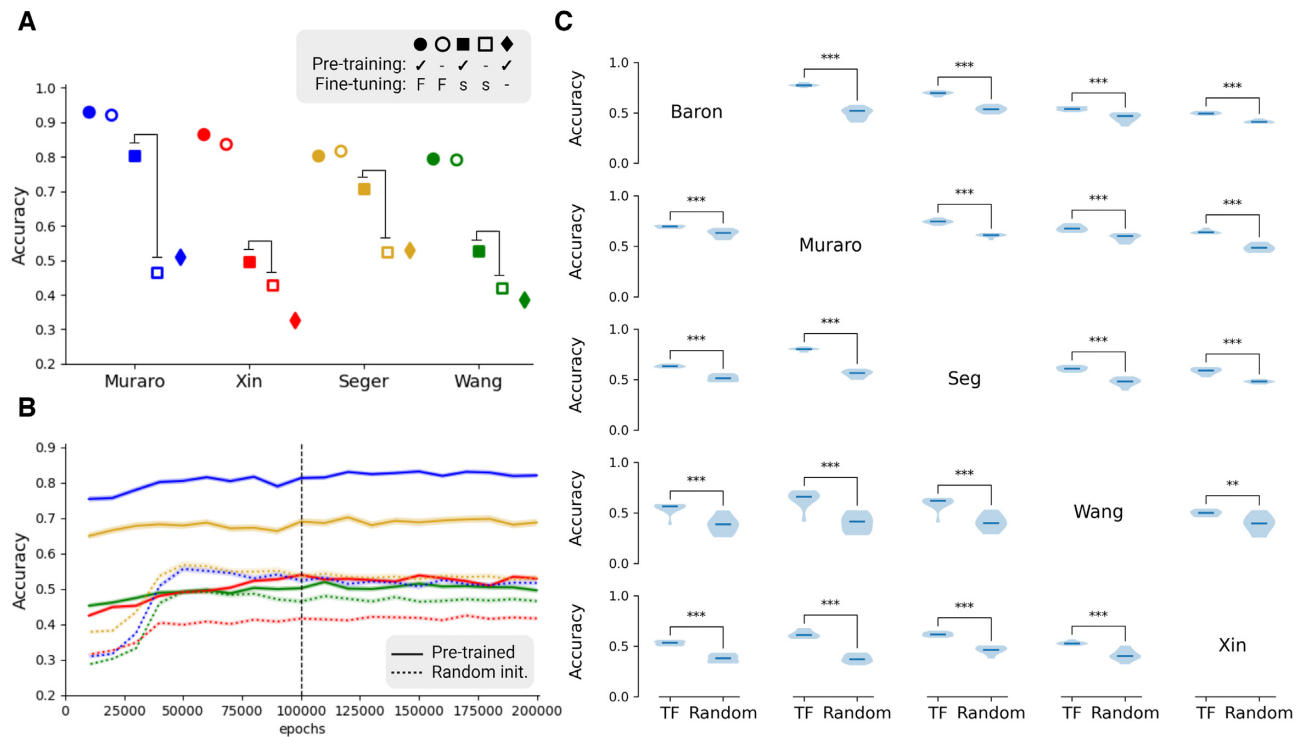


Figure 3. An accuracy chart of 5way-5shot task with five different single-cell pancreas datasets. (A) Comparisons between different training conditions show that transfer learning makes a significant difference in small-size data conditions. Pre-training refers to pre-training with the GTEx dataset, and it is represented with filled markers. Fine-tuning refers to additional training with the Baron dataset. 'F' indicates that the complete dataset of Baron was used for the fine-tuning training step, and it is represented with a circle. 's' indicates a data scarcity condition, and it is represented with a square. '-' indicates no fine-tuning training, and it is represented with a diamond. (B and C) Figures are about the data scarcity condition (filled and hollow square cases). (B) Accuracy chart during training for fine-tuning with subset of the Baron dataset (15 samples per class) in case for two squares in (A). After 50 000 epochs, performance of the model converged. Solid lines represent models based on pre-trained networks with GTEx dataset and further trained with a subset of the Baron dataset. Dotted lines are randomly initialized networks trained with 15 samples per class from the Baron dataset. (C) Pair-wise comparisons of 5 different single-cell pancreas datasets show accuracy differences between pre-trained and randomly initialized networks. Each of the violin plots is composed of 10 repeats of training. 'TF' indicates fine-tuning started with pre-trained network and 'Random' indicates randomly initialized models. The rows represent the dataset used for fine-tuning, the column the dataset used for testing. Statistical significance evaluated by Student's *t*-test is indicated by * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Created with [BioRender.com](https://www.biorender.com).

scale bulk-cell sequencing data and meta-learning with fine-tuning dataset makes model training very time efficient (Figures 2A and 3B) and shows less variability in model performance (Figures 2C and 3C). The application of meta-transfer learning for building a few-shot classifier with heterogeneous omics datasets is very efficient in our model. When training a randomly initialized model with a limited dataset, we can observe that the outcome performance of the model has higher variance in both analyses, in the TCGA cancer datasets (Figure 2C) and single-cell pancreas datasets (Figure 3C). Those high variances are originated in a randomly chosen dataset and random initialization of the model. Our results show that transferring knowledge from a large-scale bulk-cell sequencing dataset can consistently guide single-cell sequencing models during training with a small subset of the data.

Compared to Sun-MTL, Our TMP needs to train feature extractor and classifier during the fine-tuning phase. The fixed feature extractor and training additional classifier are key features of transfer learning with baseline model (26). However, freezing weights of feature extractors only hindered the performance of the TMP. This may be because of the disparity of two different omics datasets, bulk-cell

and single-cell sequencing. The major source of disparity is of biological origin in the selective sequencing of individual cell types. Typically, single-cell data contain more zero-values than bulk-cell sequencing. In addition, the technical limitations of single-cell sequencing worsen the quality of data compared to bulk-cell sequencing (16,27). Because of these challenges, the quality and feature characteristics of data are different in our given tasks with bulk- and single-cell sequencing data, even if the objective of the task is similar.

In this work, we used a relatively simple structure of networks for the omics data compared to applications in computer vision in order to demonstrate that transfer learning is a reliable method for handling of biological heterogeneity, batch effects and technological heterogeneity in omics analyses. Thus there is a room for improving the models for various data integration analyses. We expect that a more complex but well-optimized structure for biological data can improve knowledge transfer and expand the applications also to other sequencing variations. Here, we analyzed single-cell sequencing datasets as an example of a technological heterogeneity issue. There are already many successful models for the integration of single-cell sequenc-

ing datasets: Seurat (28), scVI (29), scVAE (30), Scanorama (31), MARS (9) and scETM (32). Those methods, however, focused on only single-cell data integration. Our findings indicate that besides single-cell datasets, other types of data, for example, existing bulk-cell datasets, are useful for interpreting single-cell sequencing datasets. Moreover, we expect that our cross-technology transfer learning scheme can also be applied to those recent state-of-the-art batch correction models. Based on our findings, we are hypothesize that meta-transfer learning on large bio-molecular datasets in combination with fine-tuning can improve model performance in many applications struggling with small sample sizes.

DATA AVAILABILITY

The datasets generated during and/or analyzed during this study are all public data: TCGA (<https://portal.gdc.cancer.gov/>); GTEx (<https://gtexportal.org/home/index.html>); Size of TCGA dataset is summarized in Supplementary Table S1. Single-cell pancreas datasets are summarized in Supplementary Table S2. It can be downloaded at Hemberg-lab's github (<https://github.com/hemberg-lab/scRNA.seq.datasets>).

All the code used in this study is written in Python using scikit-learn and PyTorch library. The source code, figures, and code for figures are available on Github at https://github.com/iron-lion/tmp_model.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon2020 research and innovation programme under the grant agreement No 826078. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

FUNDING

EU Framework Programme for Research and Innovation H2020 [826078]; Project FeatureCloud. Funding for open access charge: EU Framework Programme for Research and Innovation H2020 [826078].

Conflict of interest statement. None declared.

REFERENCES

- Goh, W. B., Wang, W. and Wong, L. (2017) Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.*, **35**, 498–507.
- Bernasconi, A., Canakoglu, A., Masseroli, M. and Ceri, S. (2021) The road towards data integration in human genomics: players, steps and interactions. *Brief. Bioinform.*, **22**, 30–44.
- Eraslan, G., Avsec, Z., Gagneur, J. and Theis, F. J. (2019) Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.*, **20**, 389–403.
- Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J. and et.al. (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
- Yuan, Y. and Bar-Joseph, Z. (2019) Deep learning for inferring gene relationships from single-cell expression data. *Proc. Natl. Acad. Sci. USA*, **116**, 27151–27158.
- Stumpf, P. S., Du, X., Imanishi, H., Kunisaki, Y., Semba, Y., Noble, T., Smith, R. C., Rose-Zerili, M., West, J. J., Oreffo, R. O. et al. (2020) Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell RNA sequencing. *Commun. Biol.*, **3**, 1–11.
- Wang, T., Johnson, T. S., Shao, W., Lu, Z., Helm, B. R., Zhang, J. and Huang, K. (2019) BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. *Genome Biol.*, **20**, 1–15.
- Mieth, B., Hockley, J. R., Görnitz, N., Vidovic, M. M.-C., Müller, K.-R., Gutteridge, A. and Ziemek, D. (2019) Using transfer learning from prior reference knowledge to improve the clustering of single-cell RNA-Seq data. *Sci. Rep.-UK*, **9**, 1–14.
- Brčić, M., Zitnik, M., Wang, S., Pisco, A. O., Altman, R. B., Darmanis, S. and Leskovec, J. (2020) MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods*, **17**, 1200–1206.
- The Tabula Muris Consortium (2018) Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, **562**, 367–372.
- Finn, C., Abbeel, P. and Levine, S. (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: *International Conference on Machine Learning*. PMLR, Sydney, Australia, pp. 1126–1135.
- Gevaert, O. (2021) Meta-learning reduces the amount of data needed to build AI models in oncology. *Brit. J. Cancer*, **125**, 309–310.
- Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., McDonald, E. R., Barretina, J., Gelfand, E. T., Bielski, C. M., Li, H. and et.al. (2019) Next-generation characterization of the cancer cell line encyclopedia. *Nature*, **569**, 503–508.
- Qiu, Y. L., Zheng, H., Devos, A., Selby, H. and Gevaert, O. (2020) A meta-learning approach for genomic survival analysis. *Nat. Commun.*, **11**, 1–11.
- Ma, J., Fong, S. H., Luo, Y., Bakkenist, C. J., Shen, J. P., Mourragui, S., Wessels, L. F., Hafner, M., Sharan, R., Peng, J. et al. (2021) Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer*, **2**, 233–244.
- Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. and Teichmann, S. A. (2015) The technology and biology of single-cell RNA sequencing. *Mol. Cell*, **58**, 610–620.
- Koch, G., Zemel, R., Salakhutdinov, R. et al. (2015) Siamese neural networks for one-shot image recognition. In: *ICML deep learning workshop*. Vol. 37, PMLR, Lille, France.
- Tseng, H.-Y., Lee, H.-Y., Huang, J.-B. and Yang, M.-H. (2020) Cross-domain few-shot classification via learned feature-wise transformation. In: *8th International Conference on Learning Representations*. OpenReview.net, Addis Ababa, Ethiopia.
- Wang, Y., Yao, Q., Kwok, J. T. and Ni, L. M. (2020) Generalizing from a few examples: a survey on few-shot learning. *ACM Comput. Surveys (CSUR)*, **53**, 1–34.
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. and Hospedales, T. M. (2018) Learning to compare: Relation network for few-shot learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. San Francisco, CA, USA, pp. 1199–1208.
- Sun, Q., Liu, Y., Chua, T.-S. and Schiele, B. (2019) Meta-transfer learning for few-shot learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, CA, USA, pp. 403–412.
- Lopez-Paz, D. and Ranzato, M. (2017) Gradient episodic memory for continual learning. *Adv. Neur. In.*, **30**, 6467–6476.
- McCloskey, M. and Cohen, N. J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. In: *Psychology of learning and motivation*. Elsevier, Academic Press, Vol. **24**, pp. 109–165.
- Tran, D., Nguyen, H., Tran, B., La Vecchia, C., Luu, H. N. and Nguyen, T. (2021) Fast and precise single-cell data analysis using a hierarchical autoencoder. *Nat. Commun.*, **12**, 1–10.

25. Mostavi,M., Chiu,Y.-C., Huang,Y. and Chen,Y. (2020) Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. Genomics*, **13**, 1–13.
26. Chen,W.-Y., Liu,Y.-C., Kira,Z., Wang,Y.-C.F. and Huang,J.-B. (2019) A closer look at few-shot classification. In: *7th International Conference on Learning Representations*. OpenReview.net, New Orleans, LA, USA.
27. Haque,A., Engel,J., Teichmann,S.A. and Lönnberg,T. (2017) A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.*, **9**, 1–12.
28. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck III,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
29. Lopez,R., Regier,J., Cole,M.B., Jordan,M.I. and Yosef,N. (2018) Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, **15**, 1053–1058.
30. Grønbech,C.H., Vording,M.F., Timshel,P.N., Sønderby,C.K., Pers,T.H. and Winther,O. (2020) scVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics*, **36**, 4415–4422.
31. Hie,B., Bryson,B. and Berger,B. (2019) Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.*, **37**, 685–691.
32. Zhao,Y., Cai,H., Zhang,Z., Tang,J. and Li,Y. (2021) Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat Commun*, **12**, 5261.