CS 502 - Project report

Hanrong Hu Nearchos Potamitis Karim Abi Said

Abstract

Few-shot learning is a machine learning method aimed at classifying new data with minimal initial information, proving highly beneficial in biomedicine where annotated examples are scarce and costly. This report explores the efficacy of various few-shot learning algorithms on a novel benchmark: the ATAC-seq dataset - a technique for mapping chromatin accessibility. Our primary aim is to develop a predictive tool capable of annotating cell types based on single-cell chromatin accessibility profiles.

1. Introduction

ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) utilizes a transposase enzyme to insert sequencing adapters into open chromatin regions, indicative of active genome areas where gene transcription occurs (Grandi et al., 2022). These regions, known as cis-regulatory elements (CREs), provide insights into gene regulation, epigenetics, and cellular differentiation. However, the variability of open chromatin regions and the lesser extent of available databases and annotations compared to RNA-seq data make ATAC-seq analysis more challenging for cell type annotation.

In this report, we evaluate several few-shot learning models on a newly compiled ATAC-seq dataset of human cells (Zhang et al., 2021), encompassing fetal and adult samples. The dataset includes 30 different tissues from adult samples with 111 cell types, and 16 tissues from fetal samples with the same number of cell types. Overall, over 1.3 million single nuclei were sequenced, identifying approximately 1.2 million candidate CREs. The candidate CREs were classified in to promoter (-200 to 200 bp to transciption starting site), promoter proximal, and distal.

The models evaluated include ProtoNet (Snell et al., 2017), MatchingNet (Vinyals et al., 2016), Baseline-Finetuning (Chen et al., 2019), MAML (Finn et al., 2017), and a conventional neural network. On the class level, we tested different splits of tissues, resulting in different cell type composition in train, validation and test samples. On the feature level, we performed ablation studies on CRE classes to achieve acceptable accuracies, as well as to extract the most important information from our cells through different methods.

2. Method

2.1. Preprocesssing the data

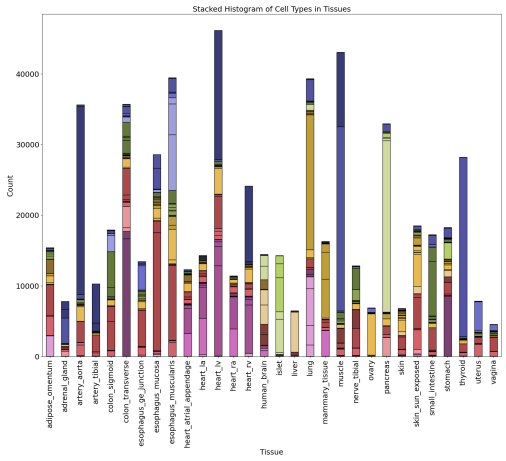

Figure 1. Distribution of cell types across 30 tissues in preprocessed adult cells. Histogram is colored by 111 cell types (legend not shown).

The initial dataset comprises a raw count matrix of 1.3 million cells by 1.2 million CREs. Due to time constraints and computational limitations, our analysis was confined to adult cells, yielding a refined matrix of 615,998 cells and 890,130 CREs. The data pre-processing follows the recommended best practice (Heumos et al., 2023). The preprocessed adult samples have 600,307 cells and 7,829 CREs (Figure.1).

1. Filtering low-quality cells and noise correction: Cells with an insufficient number of detected CREs (nCREs) and low count depth (nCounts) were removed. This step ensures the removal of data that could introduce noise or skew results.

2. Normalization and variance stabilization: Count normalization makes cellular profiles comparable. Subsequent variance stabilization ($log1p$ transformation)

ensures that outlier profiles have limited effect on the overall data structure.

3. Selection of highly variable CREs: To focus the analysis on biologically significant CREs and manage the dataset's size, the count matrix was refined to include only the most informative features.

2.2. Ablation

The atlas data categorizes the CREs into three categories: *Promoter* (ranging from -200 to +200 of the Transcription Start Site or TSS), *Promoter Proximal*, and *Promoter Distal* regions (Figure.2). Traditionally, promoters are considered highly indicative in predicting gene expression, and consequently, in determining cell type annotation. In this study, we conducted ablation studies to compare the efficacy of cell type prediction using all three CRE categories versus using promoters only. These ablation studies aim to discern the relative importance of each CRE category in accurately predicting cell types, with a particular focus on the traditionally emphasized role of promoters.

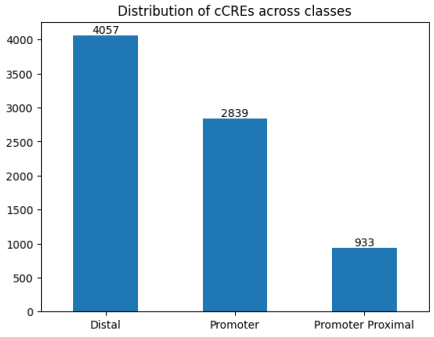

Figure 2. Distribution of candicate CRE across feature classes in pre-processed adult cells.

2.3. Splits for Training, Validation, and Testing

In this study, adult samples are collected from 30 tissues of different donors, each with a unique cell type composition (Figure.1). To implement the few-shot learning approach, we partitioned the cell types according to their tissue of origin while trying to group the cells based on similar organs/systems, i.e. the heart or endocrine system (Figure.3 and Table.1). It is important to note that after this partitioning, the testing and validation datasets included cell types for which the training dataset had limited samples, or in some cases, cell types that were completely unseen in the training set (Figure.4). This approach was specifically chosen to test the robustness of our few-shot learning models in scenarios where they encounter sparse or novel data.

Table 1. Split of pre-processed adult cells

SPLIT	TISSUES	CELLS	CELL TYPES
TRAIN	20	401,908	101
VALIDATION	5	108,316	51
TEST	5	90,083	64

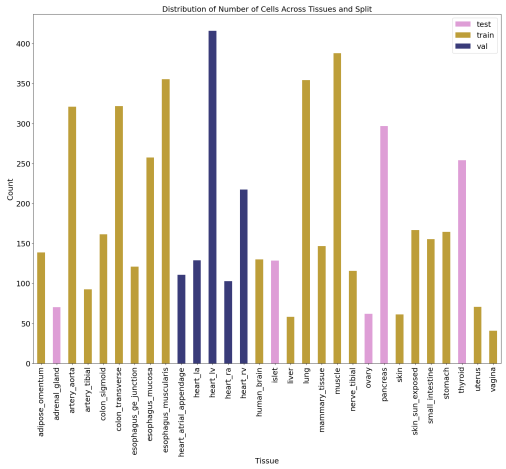

Figure 3. Splits of pre-processed adult cells based on tissue.

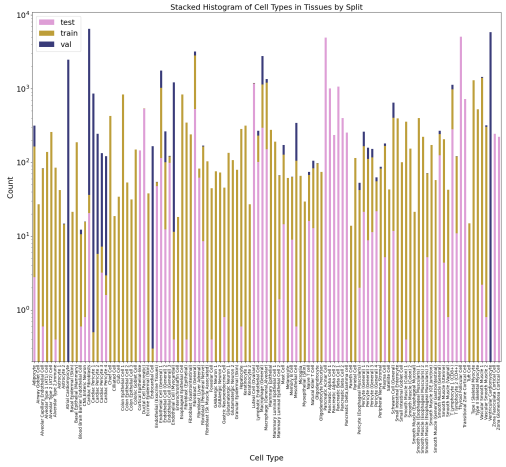

Figure 4. Cell types in tissues by splits in adult samples.

3. Results

3.1. Comparison of algorithms

We first evaluated 5 machine learning algorithms on adult cells with all feature classes (promoter, promoter proximal and distal). Results are shown in Table.2.

We compared the five benchmarks we have employed, in terms of loss and training, validation, and test accuracy (Table.2 and Figure.5). For the data containing all the adult cells, we notice that Baseline++ and Baseline perform the best out of the 5 methods, and they are typically used as the benchmark comparison for testing different methods. We note that MAML comes in last place, and this could be since it requires task-specific adaptation, as it is model-agnostic. ProtoNet and MatchingNet perform comparably well, but ProtoNet does usually perform better, since it is able to create a prototypical or "general" image of the class, but that causes it to be susceptible to outliers or to a small query set. However since we have a large number of samples for every single query set (tissue), we notice that its performance will be quite high.

It is important to note that while Baseline and Baseline++ demonstrate strong results, they are marked by significantly longer run times compared to other models. This factor becomes particularly relevant when processing extensive biological sequencing data, suggesting the potential utility of faster algorithms in such scenarios.

Table 2. Table for experiment results of adult cells all feature classes. 5-way, 5-shot.

METHOD	LOSS	TRAIN ACC.	VAL ACC	TEST ACC
BASELINE	0.769	94.65	76.56	85.18
BASELINE++	0.783	95.00	77.30	84.42
PROTONET	0.372	89.62	68.60	79.74
MATCHINGNET	0.468	83.73	64.91	75.01
MAML	0.729	74.66	56.44	67.09

3.2. Comparison of prediction using all features and only promoter

Traditionally, ATAC-seq data annotation primarily focuses on reads at promoter loci, given their strong correlation with gene expression prediction. Building on this, our study evaluated five machine learning algorithms on adult cells, specifically using the promoter feature class (Table.3 and Figure.5).

As expected, in the training process, utilizing only promoter data led to higher training accuracy along with comparable or even improved validation accuracy. However, the findings from the testing data were unexpected: the test accuracy was either similar to or lower than anticipated. This outcome suggests that while promoters are undoubtedly crucial for cell type annotation in ATAC-seq data, incorporating other regulatory elements, such as enhancers, is essential for a more comprehensive understanding and accurate prediction.

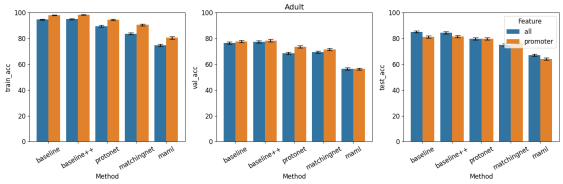

Figure 5. Results on adult cells using all features and only promoter. From left to right: training accuracy, validation accuracy and testing accuracy.

3.3. N way and N shot analysis

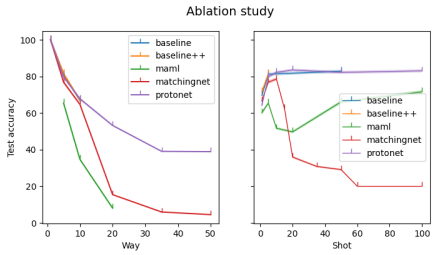

Figure 6. Analysis for ways and shots.

A fundamental aspect of few-shot learning involves the number of shots, or known samples used for training, and the number of ways, representing the classes for classification. Typically, increasing the number of shots tends to enhance model accuracy, as the model gains more examples to learn from. This trend is evident in Figure 6 (right), illustrating improved generalization with around 5 shots. Conversely, increasing the number of ways generally leads to a decrease in accuracy. This is because the model faces a more complex task, needing to discriminate among a wider array of classes, as also demonstrated in Figure 6 (left). Such trends underscore the balance required in few-shot learning between the number of examples and the diversity of classes.

3.4. Comparison of model performance on adult and fetal samples

Throughout development, the epigenetic and chromatin accessibility profile of cells can undergo significant changes. While adult cells in mature tissues are specialized with limited regenerative capacity, fetal cells are more primitive and possess greater plasticity, crucial for rapid fetal development. Therefore, we set out to evaluate the performance of the computational pipeline on fetal samples.

After preprocessing (using the same parameters as for adult cells), we observed that fetal samples include 696,626 cells, spanning 16 unique tissues and 111 cell types, with 21,829

candidate CREs identified (details in supplements). Notably, while the cell count in fetal samples was comparable to adult samples, there was a significant increase in the number of CREs. This might be due to the higher plasticity of fetal cells, likely linked to a more open chromatin structure facilitating diverse gene program activation. However, we also observed a decrease in tissue diversity, possibly due to the primitive nature of fetal cells or challenges in sample collection, which could also impact the quality of training data.

What we see in Figure.7 is that for training and validation datasets, the accuracy for fetal data is similar to that of adult data (Figure.5, also see supplementary Table.5 and 6). However, the testing accuracy for fetal data is markedly lower. This could be attributed to the initial, less expressive state of fetal cells and the incomplete nature of their differentiation, raising challenges for accurate classification, especially in few-shot learning scenarios and when only a limited number of samples are considered.

Like with the adult samples, focusing solely on promoter regions within the CREs generally improved training and validation accuracy for fetal samples. However, this approach did not yield better performance on the test dataset, often resulting in reduced accuracy.

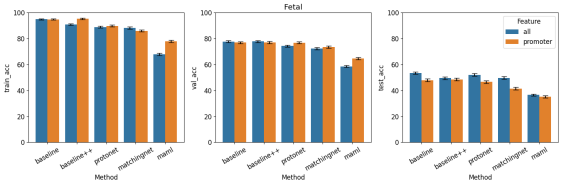

Figure 7. Results on fetal cells using all features and only promoter. From left to right: training accuracy, validation accuracy and testing accuracy.

3.5. Fine Tuning For MAML

As previously observed, MAML exhibited the lowest accuracy across the dataset, and its learning curve did not reach a plateau using the default settings. Consequently, we attempted to fine-tune the model by adjusting the number of dimensions in the fully connected layers of the backbone, as well as modifying the inner learning rate.

We tested 6 inner learning rates and 4 different number of dimensions on the adult data using only promoter (Figure.8). We can see that increase in learning rates and dimensionalities of the backbones can both increase the accuracy in both training and testing processes.

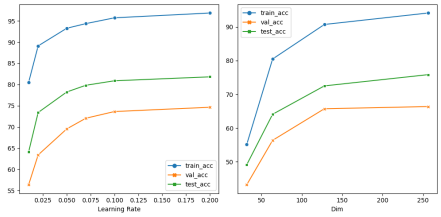

Figure 8. Fine-tuning of MAML on adult samples with only promoter. Left: experiments of inner learning rate (dim=64). Right: experiments of dimensionalities in the fully connected layers (lr=0.01). Both fine-tuning are done with 5-way, 5-shot.

4. Conclusion

In conclusion, our study has developed and tested a new dataset across various methods for a few-shot learning benchmark, demonstrating the viability of using few-shot learning for classifying cell types in the human body via cis-regulatory elements (CREs) assessed through ATAC-sequencing. We optimized training, validation, and test splits to extract maximal information from adult and fetal datasets and provided flexible code for users to experiment with different splits. Our evaluation revealed that fetal cells, with their complexity, pose more challenges compared to adult cells. We also found that while promoters are crucial for predicting cell type, other regulatory elements might offer additional insights.

Interestingly, Baseline and Baseline++ algorithms consistently outperformed others. However, baseline approaches require significantly more computing power and time, a crucial consideration for large-scale biological data. Further analysis, including N way/N shot and fine-tuning experiments, suggests that customizing meta-learning algorithms could enhance predictions while conserving processing resources.

However, this study has limitations. First, we focused solely on fully connected layer backbones, leaving the exploration of other backbones for different data types (like images or graphs) for future research. Additionally, due to time and resource constraints, in-depth ablation studies and fine-tuning were only conducted on MAML using adult data (promoter only). Lastly, incorporating more detailed metadata about CREs, such as their gene proximity, could potentially improve model performance.

A. Data and code availability

The data used in this project can be downloaded from the database: http://catlas.org/catlas_downloads/humantissues. Codes are available on https://github.com/Potamitisn/DLBIO.

B. Supplement tables

Table 3. Table for experiment results of adult cells with only feature class promoters. 5-way, 5-shot.

METHOD	LOSS	TRAIN ACC.	VAL ACC	TEST ACC
BASELINE	0.292	98.04	77.72	81.20
BASELINE++	0.427	98.24	78.40	81.54
PROTONET	0.226	94.42	73.48	79.71
MATCHINGNET	0.265	90.54	71.63	76.84
MAML	0.544	80.49	56.36	64.09

Table 4. Split of pre-processed fetal cells.

SPLIT	TISSUES	CELLS	CELL TYPES
TRAIN	10	522,355	110
VALIDATION	3	73,039	95
TEST	3	101,232	77

Table 5. Table for experiment results of fetal cells with all feature classes. 5-way, 5-shot.

METHOD	LOSS	TRAIN ACC.	VAL ACC	TEST ACC
BASELINE	0.362	95.01	77.75	53.56
BASELINE++	0.380	90.99	77.88	49.57
PROTONET	0.333	88.98	74.34	52.04
MATCHINGNET	0.389	88.30	72.47	49.76
MAML	0.781	67.89	58.62	36.57

Table 6. Table for experiment results of fetal cells with only feature class promoters. 5-way, 5-shot.

METHOD	LOSS	TRAIN ACC.	VAL ACC	TEST ACC
BASELINE	0.387	94.89	76.99	47.94
BASELINE++	0.407	95.42	77.17	48.60
PROTONET	0.339	89.83	76.94	46.60
MATCHINGNET	0.381	86.08	73.51	41.53
MAML	0.671	78.07	64.65	35.27

C. Supplement figures

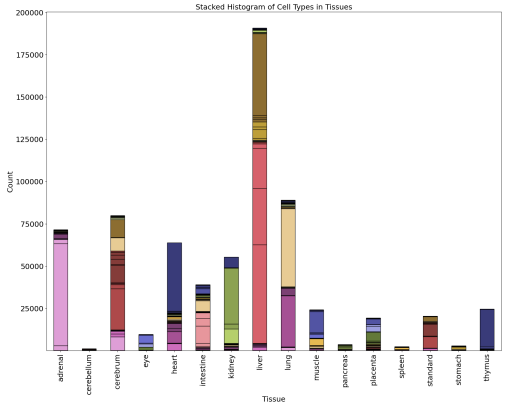

Figure 9. Cell types across tissues fetal samples. Histogram is colored by 111 cell types (legend not shown).

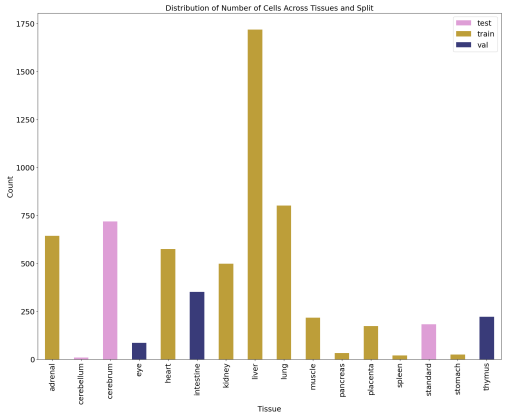

Figure 10. Cell types across tissues by splits in fetal samples.

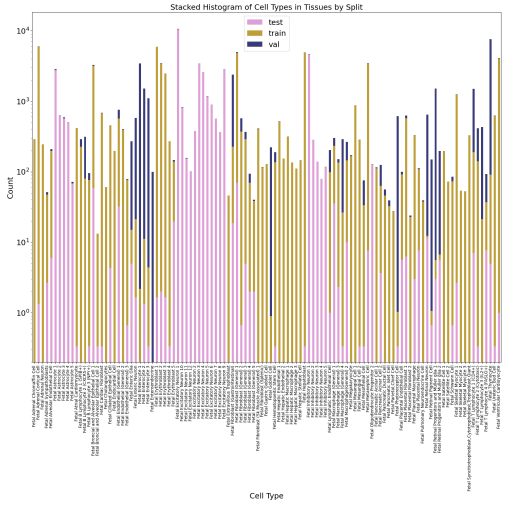

Figure 11. Cell types across tissues by splits in fetal samples.

References

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A closer look at few-shot classification, 2019.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks, 2017.

Grandi, F. C., Modi, H., Kampman, L., and Corces, M. R. Chromatin accessibility profiling by atac-seq. *Nature Protocols*, 17(6):1518–1552, Jun 2022. ISSN 1750-2799. doi: 10.1038/s41596-022-00692-9. URL https://doi.org/10.1038/s41596-022-00692-9.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Aliee, H., Ansari, M., Badia-i Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., Hediyeh-zadeh, S., Hetzel, L., Ibarra, I. L., Jones, M. G., Lotfollahi, M., Martens, L. D., Müller, C. L., Nitzan, M., Ostner, J., Palla, G., Patro, R., Piran, Z., Ramírez-Suástegui, C., Saez-Rodriguez, J., Sarkar, H., Schubert, B., Sikkema, L., Srivastava, A., Tanevski, J., Virshup, I., Weiler, P., Schiller, H. B., and Theis, F. J. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, March 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00586-w. URL http://dx.doi.org/10.1038/s41576-023-00586-w.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning, 2016.

Zhang, K., Hocker, J. D., Miller, M., Hou, X., Chiou, J., Poirion, O. B., Qiu, Y., Li, Y. E., Gaulton, K. J., Wang, A., Preissl, S., and Ren, B. A single-cell atlas of chromatin accessibility in the human genome. *Cell*, 184(24):5985–6001.e19, 2021. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2021.10.024. URL https://www.sciencedirect.com/science/article/pii/S0092867421012794.