BMC Biology

METHODOLOGY

Open Access

annATAC: automatic cell type annotation for scATAC-seq data based on language model

Lingyu Cui[1†], Fang Wang[2†], Hongfei Li[3], Qiaoming Liu[4], Murong Zhou[1] and Guohua Wang[5,6*]

Abstract

Background Cell type annotation serves as the cornerstone for downstream analysis of single cell data. Nevertheless, scATAC-seq data is characterized by high sparsity and dimensionality, presenting significant challenges to its annotation process.

Results We introduce a novel method based on language model, named annATAC, which is designed for the automatic annotation of cell types in scATAC-seq data. This method primarily consists of three stages. During the pre-training stage, by training on a vast amount of unlabeled data, the model can learn the interaction relationships between peaks, thus building a preliminary understanding of the data features. Subsequently, in the fine-tuning stage, a small quantity of labeled data is utilized to conduct secondary training on the model, which enables the model to identify cell types accurately. Finally, in the prediction stage, the trained model is applied to annotate scATAC-seq data.

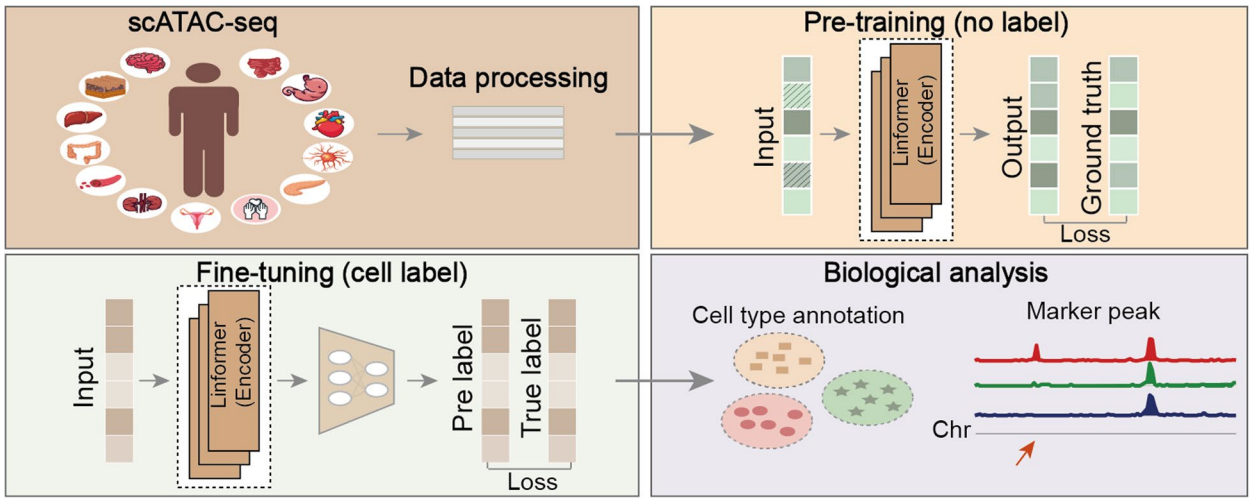
Conclusions Compared with other automatic annotation methods across multiple datasets, annATAC demonstrates superiority on the annotation performance. Further experiments have validated that annATAC holds great potential in identifying marker peaks and marker motifs. It is expected that annATAC will provide more profound and precise analysis outcomes for scATAC-seq research. As a result, it will effectively promote the progress of relevant biomedical research.

Keywords Single cell epigenomics, Automatic annotation, Language model, Pre-training, Fine-tuning

†Lingyu Cui and Fang Wang contributed equally to this work.

*Correspondence:
Guohua Wang
ghwang@nefu.edu.cn
Full list of author information is available at the end of the article

Graphical Abstract



Background

In the human body, although cells share the same DNA and gene pool, they exhibit diverse cell types [1, 2]. This diversity is attributed to the complex gene expression regulatory mechanisms, which involve interactions at multiple levels such as DNA methylation [3, 4], histone modification [5, 6], and the regulation of transcription factor activity [7–9]. These regulatory differences determine the specificity of cell functions. Moreover, chromatin accessibility plays a key role in explaining the diversity of cell types from a broader perspective. It contributes to the generation of different cell types by regulating transcription factor binding [10–12], epigenetic modifications [13–15], and spatiotemporal gene expression patterns [16–18]. Currently, the technology and data analysis of single-cell RNA sequencing (scRNA-seq) has become relatively mature, with more than a thousand available tools [19]. However, there is still a lack of specialized tools for single-cell Assay for Transposase Accessible Chromatin with high-throughput sequencing (scATAC-seq) data, which restricts the in-depth expansion of single-cell chromatin research.

Cell type annotation is the core of downstream single-cell data analysis [20, 21]. However, scATAC-seq data faces two core challenges. Firstly, high sparsity, as there are few opportunities to capture open sites in the diploid genome, and the number of reads per cell is limited, resulting in an extremely low probability of capturing specific sites [22, 23]. Secondly, it has high dimensionality caused by the complexity of chromatin structure and state. The accessibility of the chromatin genome at different positions varies among cells, and a large number of accessibility features need to be measured [24, 25]. These two characteristics significantly increase the difficulty of annotating scATAC-seq data. Currently, algorithms for annotating scATAC-seq data are categorized into (i) label transfer, whose principle is to search for similar cells between the two omics through statistical models, then transfer the labels of scRNA-seq data to scATAC-seq data. This method is currently widely used. For example, Seurat [26], ArchR [27], Signac [28], AtacAnnoR [29], and scJoint [30]. However, these two types of data cannot be fully aligned, due to various biological processes caused by distribution differences, even if they are mapped to a common latent space. (ii) Automatic annotation. With the continuous increase in the demand for annotation accuracy and the generation of massive amounts of scATAC-seq data, researchers have also developed a few automatic annotation methods without relying on the labels of scRNA-seq data. For example, Cellcano [31], RAINBOW [32], and EpiAnno [33]. Although these methods have universality when annotating cells, they do not directly utilize the information contained in the peak-cell matrix. Instead, they convert the data pattern with the help of tools like ArchR [27] or TF-IDF [34] to achieve annotation. This process not only ignores the regulatory information of peaks (genomic regions with significant enrichment of sequencing signal) on specific cell types, but also introduces additional uncertainties due to the format conversion.

In the genome, peaks that are adjacent in spatial position or close to each other due to the folding of the

three-dimensional chromatin structure can cooperatively regulate the transcription initiation of surrounding genes, thereby promoting the diversification of cell types [35, 36]. Moreover, peaks with similar functions or those involved in the same biological pathway may exhibit similar patterns of changes in accessibility, thus regulating the expression of related genes and ensuring that specific cells perform their unique functions [24, 37]. Therefore, constructing a relationship matrix between peaks and cells is crucial for the identification and classification of cell types, as well as for in-depth exploration of cell type-specific regulatory mechanisms.

Currently, the amount of unlabeled scATAC-seq data is huge, while the amount of labeled data is relatively scarce [38]. Language model have achieved remarkable progress [39–41]. Their core advantage lies in the ability to learn complex feature representations from a large amount of unlabeled data, refine potential patterns and regularities, and then fine-tune using a small amount of labeled data based on supervised learning to accurately annotate the unlabeled data [42–44]. Given this characteristic, language model has great potential for better annotation performance of scATAC-seq data.

In this study, we introduce annATAC, a method based on a language model, for automatically identifying cell types in scATAC-seq data. Firstly, pre-training. The peak-cell matrix of a large amount of unlabeled scATAC-seq data is used for pre-training, enabling the model to fully learn the interaction relationships between peaks. Secondly, fine-tuning. A small amount of labeled data is employed to conduct secondary training on the pre-trained model, so that the model is fully developed. Finally, predict cell types. Cell type prediction is performed on the unlabeled scATAC-seq data. Given the high sparsity and dimensionality of scATAC-seq data, this study proposes targeted strategies for "peak islands" and linear mapping, to effectively overcome the cell type annotation challenge caused by data structures. Through comparison and verification with existing automated annotation methods on multiple datasets, annATAC has shown a higher annotation performance. Moreover, it helps with the identification of marker peaks and the analysis of marker motifs, and also has the potential to discover novel cell types. In summary, the main contributions made by annATAC are as follows:

- ✦ Provide a pre-trained model for the interaction between features in scATAC-seq data, allowing users to fine-tune it according to different downstream tasks.
- ✦ Reliably predict cell types and cell subtypes.
- ✦ Achieve the prediction of novel cell types.

Results

The overview of annATAC

annATAC is a deep learning method based on language model (Bidirectional Encoder Representations from Transformers, BERT) and is applied to the annotation task of scATAC-seq data. The model mainly consists of five stages (Fig. 1). In stage 1, comprehensive data collection and peak island-based pre-processing. We collected scATAC-seq data covering 13 types of human tissues from the GEO database, including cancer samples and normal samples. On the premise of maximizing the preservation of the original open information of the data, these data were processed into the form of the cell-peak island to serve as the input data for the model (for detailed processing strategies, please refer to " Methods"). In stage 2, data masking. Specifically, this study divided the expression values of peak islands into five categories and then randomly masked them. During the masking process, the positions with an expression value of 0 were ignored. In stage 3, unsupervised pre-training, which was carried out with a large amount of unlabeled scATAC-seq data. annATAC takes the BERT as its basic architecture, but due to the characteristics of scATAC-seq data, it uses the multi-head attention mechanism in Linformer [45] instead of the original BERT. Through the learning of masked positions, the model can effectively learn the interaction relationships between peak islands during the pre-training stage, which plays a crucial role in the downstream cell type annotation. In stage 4, supervised fine-tuning. After the pre-training is completed, the model has already gained a relatively in-depth understanding of the interactions between peak islands. In this stage, a small amount of labeled data is used to conduct secondary training on the model, prompting the model to be further optimized so that it can accurately identify cell types and finally complete the construction of the model. In stage 5, biological analysis. The trained model is used to conduct multiple biological analyses, such as predicting novel cell types. For the specific design of the experiment, please refer to the " Methods" section for details.

Performance evaluation of cell type annotation

Comparative analysis with baseline methods

To evaluate the performance of annATAC in annotating cell types, this study conducted performance evaluations on eight tissues of human adults (Breast, Stomach, Esophagus Muscularis, Lung, Transverse Colon, Gastrocnemius Muscle, Heart Right Atrial Appendage, and Heart Left Ventricle). As shown in Fig. 2A, the Accuracy (ACC), Jaccard-weighted, and Cohen's kappa of annATAC are

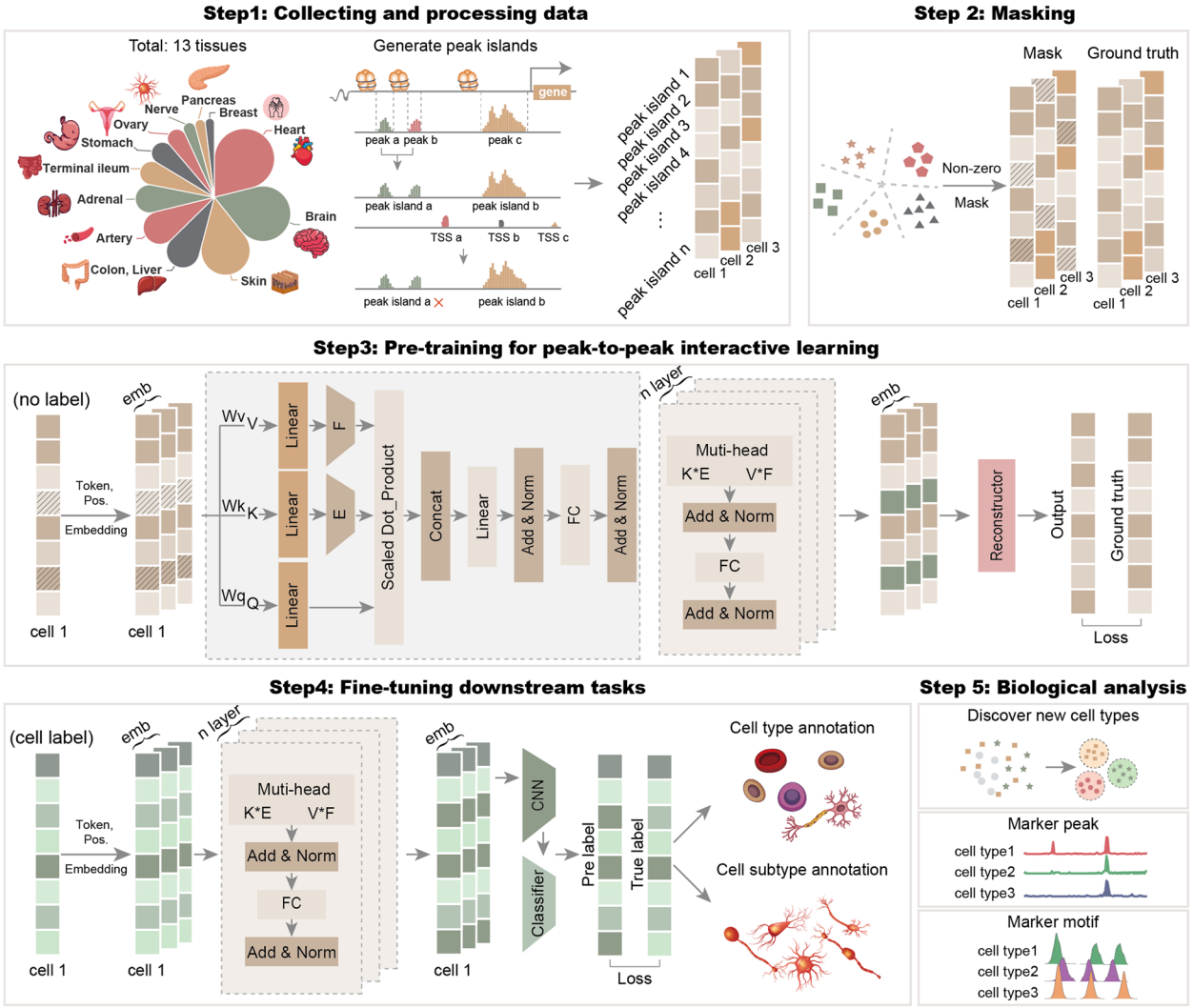


Fig. 1 Overview of the annATAC. In step 1, collect millions of scATAC-seq data, covering 13 kinds of tumor and normal tissues in the *Homo sapiens*. Then further convert these data into the form of a cell-peak island for the input of the model. In step 2, perform data masking and ignore the positions where the expression values are zero. In step 3, conduct unsupervised pre-training. Build an Encoder by introducing linear transformation into the multi-head attention mechanism. Train the model by minimizing the loss between the model's output and the ground truth before masking. In step 4, carry out supervised fine-tuning. Through secondary training on a small amount of labeled data, the model is fully converged. In step 5, conduct downstream biological analysis

superior to the comparison methods in the vast majority of tissues and comparable to Cellcano in some tissues. Specifically, on the Esophagus Muscularis data with the largest number of cells, annATAC had a significant advantage in ACC, while on the Stomach data with the smallest number of samples, its ACC was slightly lower than that of Cellcano. We believe that this phenomenon is caused by the high sparsity of small sample data. The evaluation results of the Heart Right Atrial Appendage and Heart Left Ventricle tissues are shown in Additional file 1: Fig. S1A. Compared with Cellcano, annATAC had a higher ACC in predicting T cell and Smooth Muscle cells with smaller sample sizes. To comprehensively display the comparison results with the baseline methods, this study plotted the overall distribution map of ACC on the eight tissues in Fig. 2B. In summary, annATAC's ACC performance is better that of the baseline method, with Cellcano ranking second, followed by RAINBOW. The overall distribution of Jaccard (weight) and F1 score is shown in Additional file 1: Fig. S1B.

In addition, we visualized the annotation results on the Heart and Stomach tissues (Fig. 2C). The first column of this figure shows the cell labels provided by the original study. Through experimental comparison, it was found

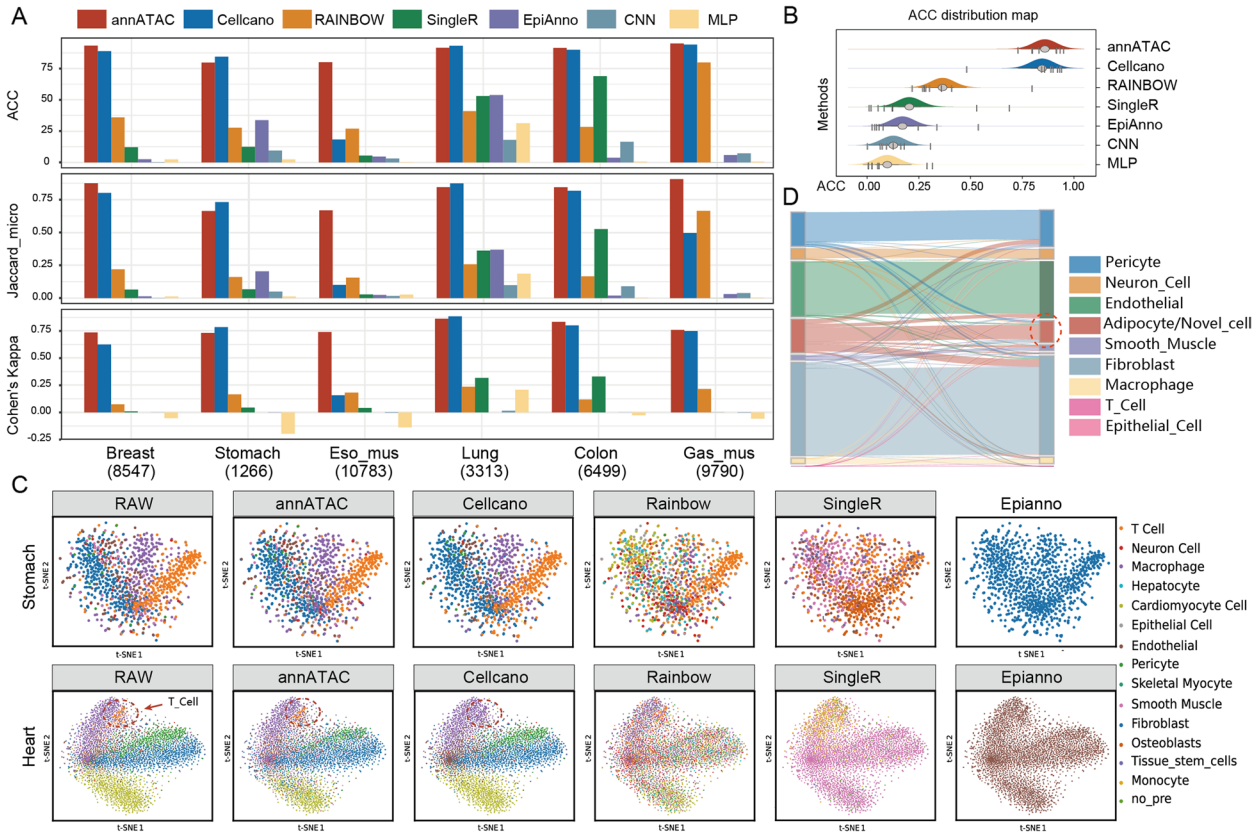


Fig. 2 Performance evaluation of annATAC in annotating cell types. **A** Evaluation of the annotation performance of ACC, Jaccard_weighted, and Jaccard_micro on the scATAC-seq data of six adult tissues. The abscissa indicates the number of cells in each tissue. **B** Overall distribution performance evaluation of ACC on the scATAC-seq data of eight adult tissues. **C** t-SNE visualization display on the Heart and Stomach Tissues, where the first column shows the annotation of the original class labels. **D** Evaluation of annATAC's performance in detecting novel cell types. The red color in the Sankey diagram represents novel cell types, with the real cell labels on the left side and the labels predicted by annATAC on the right side

that only annATAC and Cellcano could better restore the annotations of cell labels, and annATAC performed even better. It is particularly worth noting that annATAC performed excellently in predicting cell labels with a small number of cells. For example, in the Heart tissue, annATAC accurately predicted T cells. Identifying T cells is of crucial importance for gaining a deeper understanding of the immune microenvironment of cardiac tissue as well as the mechanisms of disease occurrence and development. For the visualization of the annotation results of other tissues, please refer to Additional file 1: Fig. S1 C.

annATAC can identify novel cell types

annATAC has a built-in option for identifying novel cell types. When the confidence level of a predicted cell is lower than the default confidence score, the model will label this cell as a novel cell type. This study tested the ability of annATAC to explore novel cell types in Adipose Omentum tissue. Compared to the previously tested and compared tissues, "adipocyte" in this tissue is a novel cell type. The experimental results are shown in Fig. 2D. annATAC can accurately predict common cell types. For the novel cell type Adipocyte, the model predicted most of these cells as "Novel_cell_type," but a small number were predicted as Fibroblasts. Considering that both Fibroblasts and Adipocytes originate from Mesenchymal Stem cells and there are many overlaps in their differentiation regulatory mechanisms, we speculate that this is the reason why annATAC predicts some Adipocytes as Fibroblasts. Through this experiment and combined with prior biological knowledge, it can be demonstrated that annATAC can identify novel cell types.

Performance evaluation of cell subtype annotation

The differences in chromatin accessibility among different cell subtypes are relatively subtle. Therefore,

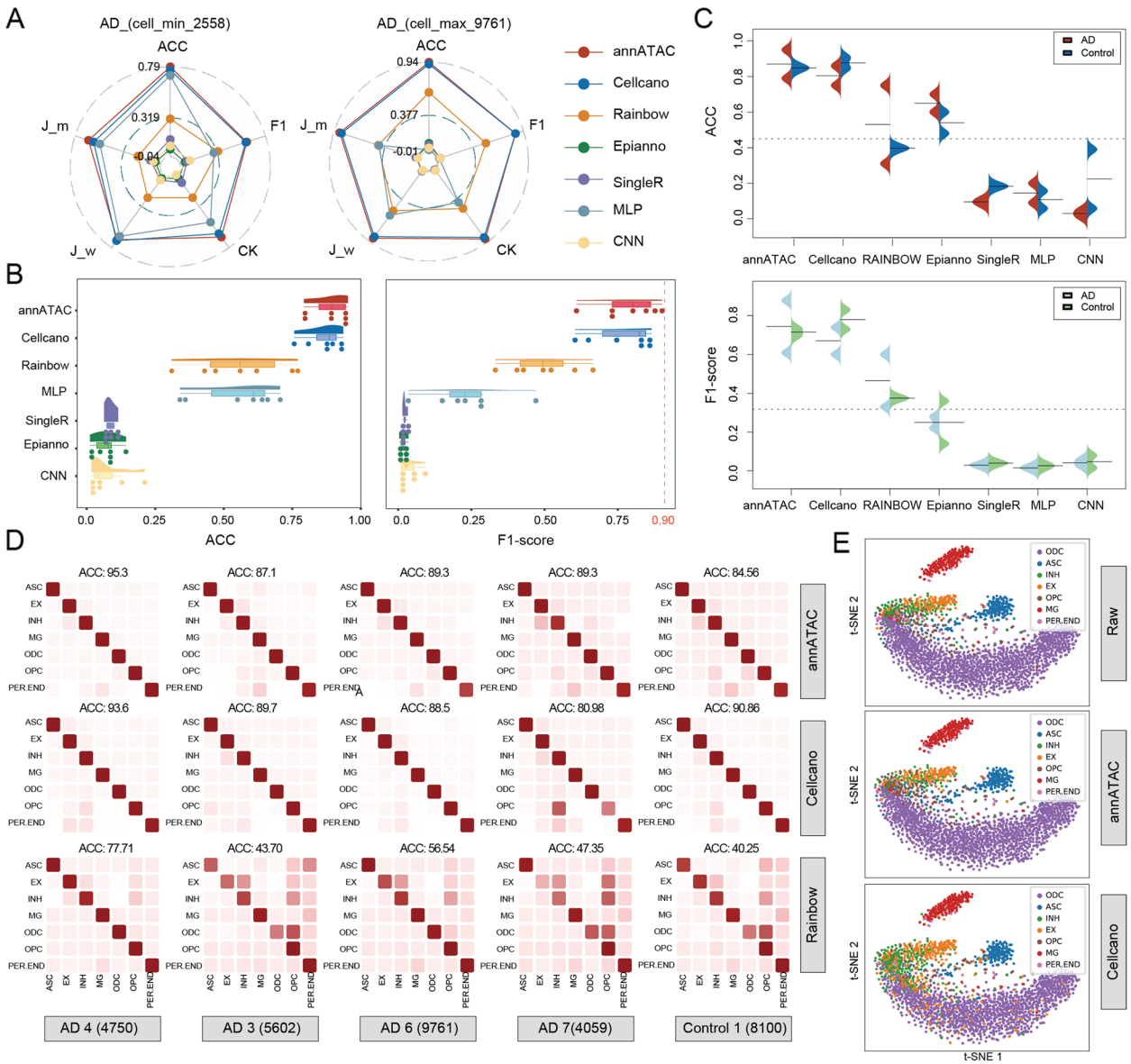


Fig. 3 Performance evaluation of cell subtype annotation. **A** Comparing annATAC with other comparative algorithms in the AD datasets with the smallest sample size (2588) and the largest sample size (9761). **B** Semi-violin plots display the performance of annATAC and other algorithms in terms of ACC and F1 score on seven AD datasets with different sample sizes. **C** The experimental performance of two AD and control groups is evaluated in terms of ACC and F1 score. **D** A comparison displays the annotation results of annATAC and those of RAINBOW and Cellcano on four AD and one control datasets. The bottom row indicates different tissues and the number of cells. **E** Performing t-SNE visualization on the AD datasets shows that different colors represent different cell types

compared with the annotation of rough cell types, the annotation of cell subtypes is more difficult. This study conducted experiments on different Neuronal cell subtypes of patients with Alzheimer's Disease (AD) (Astrocytes (ASC), Excitatory Neurons (EX), Olig.progenitors (OPC), Inhibitory Neurons (INH), Microglia (MG), Oligodendrocytes (ODC), and Pericytes/Endothelial (PER.END)). We collected a total of 9 datasets from the GEO database [46], including 7 datasets of AD patients and 2 datasets of the control group.

annATAC accurately annotates cell types across different data scales

Among the nine collected datasets, we selected two datasets with the smallest (2558) and the largest (9761) number of cells respectively to evaluate the annotation

results. As shown in the radar chart in Fig. 3A, after comprehensively evaluating with five metrics, it was found that regardless of the data scale, annATAC always maintained a relatively high annotation ACC, and Cellcano still ranked second in terms of performance. Secondly, to further and comprehensively observe the comparison results between annATAC and other algorithms, we used semi-violin plots to display the ACC and F1 score of annATAC and other algorithms (Fig. 3B). It can be seen from the figure that annATAC had an excellent performance in ACC, and its F1 score could reach 90%. It can thus be concluded that annATAC had a relatively high ACC in predicting cell subtypes. In addition, Additional file 1: Fig. S2 presents the experimental results of another three metrics, namely Cohen's kappa, Jaccard_weighted, and Jaccard_micro. Through comparison, the superiority of annATAC could also be confirmed.

annATAC has superiority in identifying cell types in Alzheimer's disease

This study collected two sets of control experiment data, which were respectively from 2 AD patients and two healthy control groups. The ages of the patients were all around 80 years old. The same sequencing technique (Illumina NovaSeq 6000) was adopted for both groups, and the sequencing sites were all in the Frontal Cortex region. The experimental results are shown in Fig. 3C. Through a comparative analysis using ACC and F1 score, we found that annATAC had a significant advantage in ACC when predicting the cell types of AD patients and was nearly 10 percentage points higher than Cellcano in ACC. While in the control group, the ACC of Cellcano was only around 5 percentage points lower than that of annATAC.

Comparative evaluation with other methods

We presented the prediction results of different algorithms on AD datasets of 5 patients and one control group dataset. Among them, Cellcano and RAINBOW algorithms with better performance were selected for comparison with our method. As shown in Fig. 3D, annATAC demonstrated superiority in most tissues. However, on the control group data, its performance was slightly lower than that of Cellcano. In addition, we selected the prediction results of patient 2 with AD for visual display, as shown in Fig. 3E compared to Cellcano, annATAC had a higher ACC in annotating cell labels. Cellcano mispredicted some ODCs as EXs. Moreover, in the t-SNE plot, it could be observed that Cellcano failed to separate well the regions where cell types were interconnected, while annATAC performed better in this regard.

Biological analysis of Alzheimer's disease

To deeply explore the analytical potential of annATAC, we carried out a series of biological analyses on patient 4 with AD. Through multi-dimensional experimental investigations and precise analyses, it is strongly demonstrated that annATAC has significant potential in identifying marker peaks and marker motifs, and is expected to provide new technical support and research perspectives for relevant biological research and disease analysis.

annATAC has the potential to identify marker peaks

This study focused on the biological analysis of patient 4 with AD. Firstly, the cell type annotation results were presented using Uniform Manifold Approximation and Projection (UMAP), as shown in Fig. 4A on the left side of the figure, the *FindClusters* function in the Seurat tool [26] was used for cell clustering. On the right side, the prediction results of annATAC were labeled for visual presentation. Through the UMAP visualization analysis, it was observed that annATAC had very distinct boundaries for the annotation and classification of cell types. In particular, it was able to accurately annotate the easily confused Neuron types EX and INH. In addition, Fig. 4B shows the proportion of regulatory elements in different cell types, with UnionPeaks presenting the proportion of overlapping regulatory elements. Subsequently, the *getMarkerFeatures* function in the ArchR tool [27] was used to identify the specific marker peaks in different cell types (Fig. 4C). From the experimental results, it could be seen that the peak enriched in the LHX families was the marker peaks of ASC. As shown in the graphs on the left side of Fig. 4C, the corresponding marker peak location information could be found at the corresponding positions on chromosome 1. Meanwhile, it was found that the peak enriched in the NEUROG families was the marker peaks of EX, which were presented in the graphs on the right side of Fig. 4C. Our findings are consistent with existing research results [47, 48]. In addition, the other peak enriched by the ARID family is the marker peak of ASC, and the other peak enriched by the TCF family is the marker peak of EX (Additional file 1: Fig. S3 A). The peak enriched in the SOX family were the marker peak of ODC, and the peak enriched in CD68 were the marker peak of MG. These findings not only deepened our understanding of the characteristics of cell types in AD patients, but more importantly, provided crucial technical support for the targeted treatment of AD, and were expected to promote the further development of the AD treatment field.

annATAC has the potential to identify marker motifs

To further explore the regulatory mechanisms specific to different cell types, we carried out a marker motif analysis

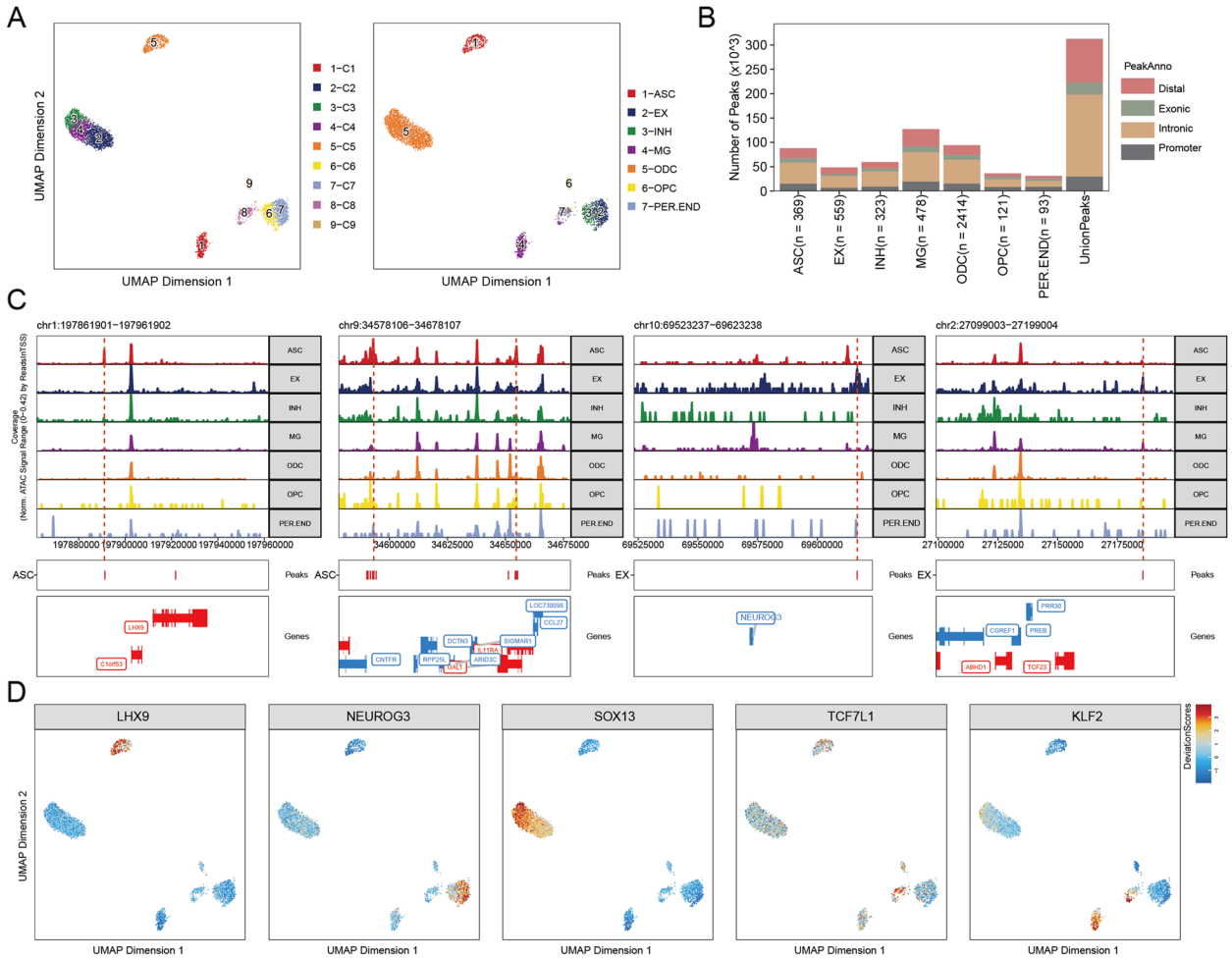


Fig. 4 Biological analysis of AD. **A** UMAP visualization of annATAC's cell type annotation for patient 4 with AD. **B** Proportions of regulatory elements contained in different cell types. **C** Display of marker peaks on different cell types, with red auxiliary lines indicating the positions of marker peaks. **D** Display of marker motifs on different cell types

(Fig. 4D). During the research process, by combining the analysis results with Fig. 4A, we confirmed that LHX9, NEUROG3, SOX13, TCF7L1, and KLF2 are the marker motifs of ASC, EX, ODC, PER.END, and MG respectively. Furthermore, to more accurately depict the specificity of the marker motifs, we used mountain plots for visual display. For specific details, please refer to Additional file 1: Fig. S3B. Through the mountain plots, the distribution characteristics and specificity of the marker motifs of different cell types can be presented intuitively, providing a clear and powerful basis for further research on cell regulatory mechanisms.

Cell type-specific analysis

To further reveal the regulatory mechanism of transcription factors, we carried out transcription factor footprint analysis using the *getFootprints* function in the ArchR tool, and the experimental results are presented in Fig. 5A. In the upper part of Fig. 5A, the transcription footprints of the SOX4 and SOX17 enriched in ODC are clearly shown. This presentation intuitively presents the action sites and patterns of these two transcription factors in the gene regulatory regions related to ODC cells. The lower part focuses on PER.END and presents the transcription footprints of the SMARCC family enriched in this cell type. Through the analysis above, we can accurately locate the binding positions of transcription factors in gene promoters and other key regulatory regions, especially the binding situation near the transcription start site (TSS), providing key information for analyzing the initiation mechanism of gene transcription. Meanwhile, this analysis can also deeply reveal the binding patterns of transcription factors to DNA and the interrelationships among different transcription factors. These findings are crucial for understanding the precise regulation of gene expression and provide valuable clues for

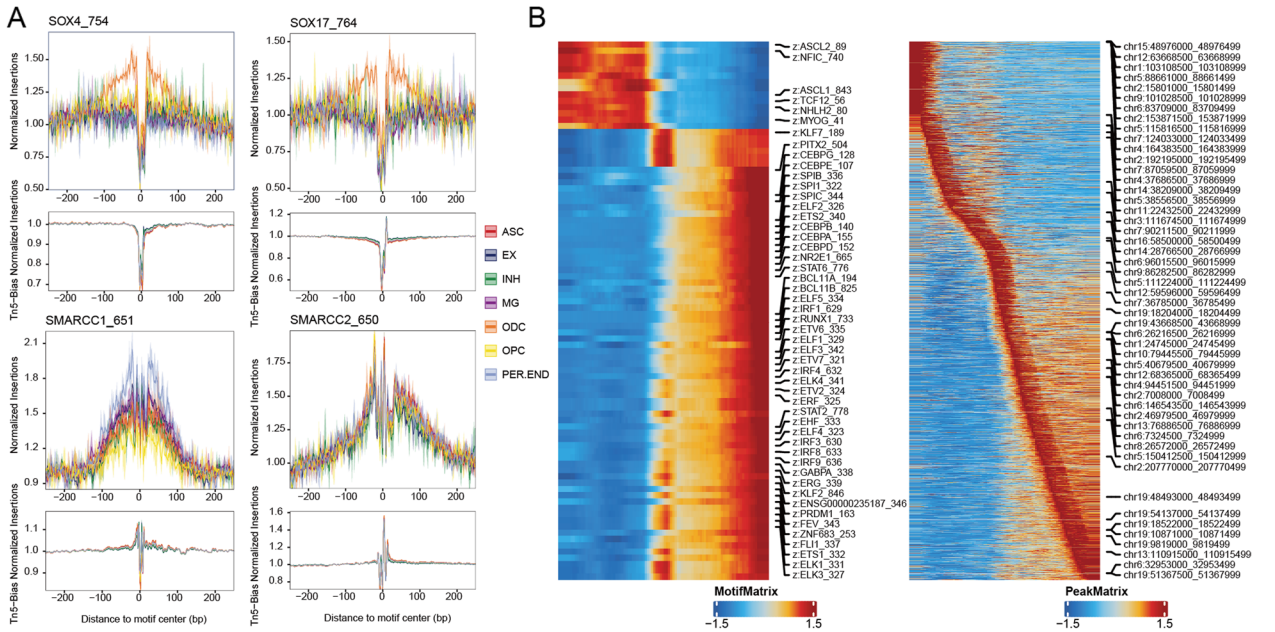


Fig. 5 **A** Analysis of cell type-specific transcription factor footprints in AD patients. **B** Pseudo-time trajectory analysis of marker motifs and marker peaks in AD patients

the study of disease pathogenesis, which are expected to promote the in-depth development of related disease research fields. In addition, we conducted cell trajectory analysis on PER.END cells in both motif and peak dimensions (Fig. 5B). In this analysis, we used MotifMatrix and PeakMatrix to construct cell trajectories. The motif matrix contains information about transcription factor binding motifs, while the peak matrix reflects the peak characteristics of chromatin accessibility. By using these two matrices to construct cell trajectories, we can deeply analyze the impact of dynamic changes in transcription factor binding behavior on cell development trajectories during processes such as cell differentiation. Starting from the key aspect of transcriptional regulation helps us to have a more comprehensive and in-depth understanding of the dynamic changes in cells at different stages, providing important clues for revealing the molecular mechanisms of cell differentiation.

Model robustness testing

annATAC has strong generalization ability

To further investigate the generalization ability of the model to noise, this study conducted experiments on five tissues: Breast Mammary tissue, Transverse Colon, Esophagus Muscularis, Lung, and Stomach (Fig. 6A). Firstly, we fine-tuned the model using the original labels (corresponding to the RAW in the first column of the figure) and tested the annotation ACC on the aforementioned 5 tissues. Subsequently, to simulate noises, we randomly shuffled the fine-tuned data labels at ratios of 5%, 10%, 15%, and 20%, and then conducted training in sequence (corresponding to columns 2, 3, 4, and 5 in the figure), finally predicted the annotation ACC for each scenario. From the experimental results, it can be observed that regardless of the proportion of label shuffling or the number of cells, the ACC of the model prediction remains relatively stable, and the trend of change shows significant consistency. This phenomenon fully demonstrates that annATAC has a strong generalization ability and can effectively resist the interference of data noise. In addition, we further expanded the scope of the experiment by performing a 10% random shuffling operation on the fine-tuned labels on 8 tissues, then comparing them with the results obtained from training with the original labels. The results show that the trends of the two are consistent, and the specific experimental results are detailed in Additional file 1: Fig. S4 A.

annATAC has a relatively strong adaptability

In order to further test the adaptability of the model to fine-tuning data volume, this study conducted specialized adaptability tests on two organizations, the Gastrocnemius Muscle and Lung. In the experimental operation, we strictly extracted 40%, 60%, and 80% of the data from the fine-tuning data for fine-tuning of the model. After completing the fine-tuning, the model was tested on the corresponding datasets of these two organizations to verify its adaptability to different amounts of data.

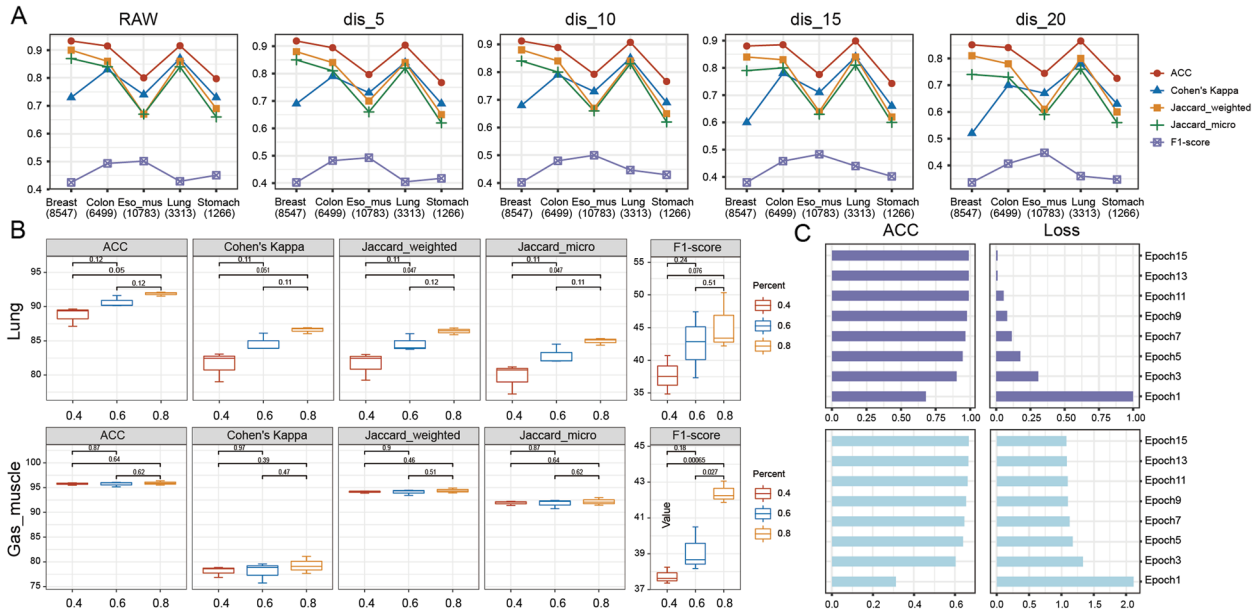


Fig. 6 Model stability testing. **A** Model robustness testing was conducted on five tissues: Breast Bammary tissue (cell: 8547), Transverse Colon (cell: 6499), Esophagus Muscularis (cell: 10,783), Lung (cell: 3313), and Stomach (cell: 1266). RAW indicates that the labels were not shuffled during the fine-tuning process, dis_5 indicates that 5% of the labels were randomly shuffled, dis_10 indicates that 10% of the labels were randomly shuffled, and so on. **B** Randomly select 40%, 60%, and 80% of the fine-tuning data for training, and test the annotation ACC on two tissues, Gastrocnemius Muscle and Lung. **C** Pre-training ablation experiments. The upper part shows ACC and loss with pre-training, while the lower part shows the results without pre-training

To ensure the reliability and stability of the experimental results, three rounds of experiments were independently repeated for each subset of data at different ratios. The experimental results in Fig. 6B comprehensively evaluate the performance of the model under different data volume adjustments using five evaluation metrics. Meanwhile, *t*-test statistical methods were adopted to accurately detect the differences in prediction ACC corresponding to different proportions of fine-tuned data. The clear trend of changes presented in the graph, as well as the statistically significant *p* values displayed, demonstrates that annATAC exhibits strong adaptive data volume capabilities. This means that the model does not overly rely on fine-tuning the amount of data when making predictions and can maintain stable and efficient performance at different scales of fine-tuning data.

Ablation experiment

To enhance the interpretability of the model, we carried out ablation experiments (Fig. 6C). In this ablation experiment, we set the training period to 15 epochs. The upper part of Fig. 6C shows the changes in ACC and loss after introducing the pre-trained model for fine-tuning, while the lower part presents the corresponding ACC and loss results when directly training on the fine-tuning dataset without introducing the pre-trained model. It can be seen that after introducing the pre-trained model, the ACC at epoch 0 is 0.67, while without introducing the pre-trained model, the ACC at epoch 0 is only 0.32. This significant difference indicates that the pre-trained model provides a good foundation for the subsequent fine-tuning. In terms of the convergence speed of the model, the fine-tuning process with the addition of the pre-trained model converges relatively quickly. To sum up, pre-training plays a crucial role in the prediction performance of the model. In addition, we also conducted a comparative analysis of the F1 score before and after the ablation (Additional file 1: Fig. S4B).

Discussion

annATAC not only demonstrates significant superiority in cell type annotation, but also shows great potential in the field of predicting novel cell types. In addition, taking the in-depth analysis of AD patients as an example, this model provides strong support for accurately identifying marker peaks and marker motifs and is expected to open up new exploration paths for AD-related research.

annATAC also has certain limitations. Firstly, although this study proposed peak islands to dramatically decrease the sparsity of scATAC-seq data, there is still a certain gap to the sparsity of omics data such as transcriptome and proteome data, which may make the model fail to

comprehensively learn all the interaction relationships between peaks during the pre-training process. Nevertheless, through pre-training ablation experiments, it can be clearly and powerfully proven that the pre-training stage plays an indispensable and crucial role in the downstream cell type annotation task. Secondly, in terms of the model architecture, we currently only provide a model version with a pre-training model version with a 4-layer Encoder. Considering the diverse needs of different users, users can flexibly modify the network depth (such as adjusting it to 6 layers, 8 layers, etc.) according to the actual situation of their research, and then conduct pre-training again to better adapt to specific research scenarios. Thirdly, from the perspective of scalability, our model currently focuses on two downstream tasks, namely cell type annotation and cell subtype annotation. There is still a broad application space to be explored around scATAC-seq data. For example, developing other downstream tasks targeting specific regulatory elements will not only help to further expand the application range of the model but also provide a new way to deeply explore the gene regulation mechanism. We have listed it as one of the important work directions in the future.

Conclusions

In the field of bioinformatics, the annotation of scATAC-seq data has always been an urgent challenge to overcome. This study focuses on this and develops annATAC, which is based on pre-training with a large amount of unlabeled scATAC-seq data. Compared to other existing scATAC-seq data annotation tools, the innovation of annATAC is mainly reflected in the following three aspects. (1) Automatic annotation. annATAC has unique automatic annotation capabilities and can independently annotate scATAC-seq data without relying on other omics information, avoiding the complexity and potential errors of cross-omics data integration. (2) Predicting new cell types. It can use its own algorithm and model architecture to identify and predict novel cell types, breaking through the limitations of traditional tools that can only annotate known cell types, and opening up a broader exploration space for cell research. (3) Targeted fine-tuning. Users can fine-tune the model according to specific downstream tasks, enabling it to accurately adapt to different application scenarios, greatly improving the practicality and targeting of the tool. In the pre-training stage, the model deeply extracts the complex interactions between peaks by restoring the expression of mask positions, thus initially constructing a model a certain generalization ability. After pre-training, the model further utilizes a small amount of labeled data for downstream fine-tuning. During this process, the model continuously optimizes its own parameters and gradually converged to accurately annotate different cell types. To fully validate the effectiveness of annATAC, this study carefully designed two downstream fine-tuning tasks, cell type annotation and cell subtype annotation. Through comprehensive and systematic evaluation using multiple dimensions and metrics on multiple organizational samples, the results clearly demonstrate that annATAC exhibits significant superiority in both tasks, providing an advanced and reliable solution for the annotation of scATAC-seq data.

Methods

Comprehensive data collection and peak island-based pre-processing

In this study, scATAC-seq data of 13 human tissues were collected for pre-training (without cell labels), involving a total of 1,011,883 cells, including data of cancer samples and normal samples. The specific information is shown in Fig. 1 and Additional file 2: Table S1. Secondly, the data used for the cell type annotation task in the downstream analysis came from the single-cell reference atlas of human genome chromatin accessibility [49], with a total of 98,703 cells (with cell labels). The data for the cell subtype annotation task came from a dataset of AD, which was composed of different Nerve cells [46], with a total of 41,796 cells (with cell labels).

To maximally preserve the open information of different peaks among cells in scATAC-seq data, annATAC adopts a concise pre-processing method to construct the feature matrix (cell-peak) for the data of pre-training and two fine-tuning tasks. Firstly, fastq-dump is used to convert the format of the original data. Then, cellranger-atac is utilized to process and generate the data in the form of cell-peak as the input for the model. Besides, in response to the highly sparse nature of scATAC-seq data, annATAC further divides each chromosome into bins, each with the length of 5000 base pairs. All the peaks within each bin are gathered into a "peak island." A unique openness value is assigned to each peak island by averaging the inside peaks. Finally, the peak islands that contain transcription start sites are selected as the input features for the model.

The architecture of annATAC

annATAC adopts similar architecture as BERT, which begins with the masking and input embedding, through multi-head attention, and finally ends with feedforward networks. To deal with the high dimensionality of scATAC-seq data, multi-head attention in Linformer [45] is adopted to significantly reduce the computational complexity. The design of each component within annATAC is discussed below.

Masking

The peak island-based pre-processing results of scATAC-seq serve as the input. As the positions with expression values of 0 did not reflect open information, these positions were ignored during the masking process. It is worth noting that annATAC is designed to evenly distribute the number of masks into different categories during the masking process, which avoids model preference learning and benefits the stability and reliability of model learning.

Input embedding

Input embedding takes the masking are firstly divided into five categories, each of which is represented as a token embedding. Besides, an additional position embedding generated by TruncatedSVD [50] is adopted to better learn the interaction relationships among peaks. Input embedding is the sum of the token embedding and position embedding.

Multi-head attention

Firstly, for the input embedding results X, linear transformations are performed on it to obtain the Q (query), K (key), and V (value) vectors. Specifically, for each head h (there are H heads in total), there are $_{Q_h = XW_q^h}$, $K_h = XW_k^h$, and $V_h = XW_v^h$, where W_q^h, W_k^h, and W_v^h are three learnable weight matrices. The calculation formula for the attention score of the *h*th head in the original Transformer is shown as follows:

$$\text{Attention}(Q_h, K_h, V_h) = \text{Softmax}(\frac{Q_h K_h^T}{\sqrt{d_k}})V_h \quad (1)$$

In the above formulas, as the length of the input features increases, the computational workload will grow quadratically, resulting in extremely low computational efficiency when dealing with high dimensional data. In Linformer, two projection matrices E_h and F_h are introduced to significantly reduce the dimensionality of K_h and V_h, and the calculation formula for the attention score is updated as follows:

$$\text{Attention}(Q_h, K_h, V_h) = \text{Softmax}(\frac{Q_h(E_h K_h^T)}{\sqrt{d_k}})F_h V_h \quad (2)$$

Through this step of linear mapping, the computational complexity can be significantly reduced, and thus long sequence data can be effectively processed.

Feedforward neural network

The output of multi-head attention is used as the input of the feedforward neural network, and layer normalization and residual connection are carried out respectively after these two modules. In addition, pre-training and downstream fine-tuning tasks have different modules following the feedforward neural network. Firstly, pre-training is connected to a fully connected layer, while fine-tuning tasks are connected to a convolutional neural network.

The workflow of annATAC

Pre-training

Pre-training on a large number of unlabeled datasets enables the model to fully learn the interaction relationships among peak islands by using the unmasked openness information to predict the information at masked positions, which plays a crucial role in the annotation of downstream cell types. During the training process, DistributedDataParallel is utilized to accelerate the training process of the model. In addition, we choose cross-entropy as the loss function of the model:

$$L = -\sum_{i=1}^{n}\sum_{j=1}^{m} y_{i,j}\log(p_{i,j}) \quad (3)$$

Here, n represents the number of cells, m represents the number of masks, and $y_{i,j}$ and $p_{i,j}$ respectively represent the actual and predicted expression values of the *j*th peak island in the *i*th cell.

Fine-tuning

This study first uses the scATAC-seq data provided by the single-cell reference atlas of human genome chromatin accessibility as the first downstream fine-tuning task. After the model has been pre-trained on a large number of unlabeled datasets, it has already had an in-depth learning of the interaction relationships among peak islands and the model has been preliminarily converged. Therefore, in the fine-tuning stage, we apply the model to a small number of labeled scATAC-seq datasets for secondary training so that it can comprehensively capture the distribution characteristics of different cell types. The fine-tuning stage shares the model parameters with the pre-training stage, and its process is quite similar to that of the training stage. However, there are mainly three differences as follows: Firstly, there is no need to perform masking processing on the data during the fine-tuning process. Secondly, in the fine-tuning stage, the loss between the predicted cell labels and the actual cell labels is calculated, while in the pre-training stage, the loss between the predicted masked expression values and the actual expression values is

calculated. Finally, in the fine-tuning stage, a convolutional neural network (CNN) is adopted as the classifier.

Secondly, given that cell subtypes may exhibit significant differences in drug responses, accurate annotation and classification of cell subtypes are of great significance for predicting the progression of diseases and prognosis. Therefore, the second fine-tuning task carried out in this study focuses on the annotation of cell subtypes. Compared with the rough cell type annotation in the first downstream fine-tuning task, the cell subtype annotation task is more complicated because the differences in openness on chromosomes among different cell subtypes are relatively subtle. Similar to the first fine-tuning task, the second fine-tuning task also shares the model parameters with the pre-training stage. However, to enhance the model's ability to distinguish different cell subtypes, we have increased the number of fully connected layers of the CNN classifier in the second fine-tuning task.

Prediction

After pre-training and fine-tuning, the model has successfully converged and can annotate both rough cell types and cell subtypes. To verify its performance, this study comprehensively compared annATAC with other counterparts on a total of 18 datasets covering rough cell type annotation and cell subtype annotation.

annATAC can identify novel cell types

annATAC has a built-in flexible option for identifying novel cell types, allowing users to configure according to their actual needs. Specifically, when the confidence level of a certain cell type predicted by the model is lower than the preset confidence score threshold (the default value is 0.5), this cell will be automatically marked as "Novel_cell_type." By introducing this function, annATAC not only has the potential to explore and identify novel cell types but also can improve the ACC and interpretability of predictions in a broader dimension.

Comparative algorithms

This study mainly compared the methods for automatically annotating scATAC-seq data. Meanwhile, it is also compared with SingleR [20], a commonly used annotation method for single-cell transcriptome. For the sake of fairness and interpretability, we converted the data into the form of a cell-gene score and input it into SingleR for prediction. In addition, we also compared annATAC with the conventional convolutional neural network and feedforward neural network in deep learning.

Cellcano [31]

Cellcano is a method for automatically annotating scATAC-seq data based on a two-round supervised learning algorithm. Firstly, Cellcano trains a multi-layer perceptron on the reference dataset. Then, it selects the cells with better predictions as anchors to form a new training set. Subsequently, it trains a knowledge distillation (KD) model to predict the cell types of non-anchor cells.

RAINBOW [32]

RAINBOW is a reference-guided automatic annotation method based on the contrastive learning framework. It mainly consists of three modules, namely training, integrating prior knowledge, and prediction. After the transformation by TF-IDF [34], it utilizes contrastive learning to learn the latent representations of the training set, and then integrates prior knowledge to conduct the prediction of cell types.

EpiAnno [33]

A probabilistic generative model integrated with Bayesian neural networks (BNN), automatically annotates scATAC-seq data in a supervised manner. EpiAnno first performs feature selection, TF-IDF transformation, and z-score normalization on the matrix, and then derives the latent representation of cells from the Gaussian distribution using prior knowledge. It projects the latent representation into the original feature space of scATAC-seq data through the nonlinear Bayesian neural network. EpiAnno uses the trained model to obtain cell embeddings and infer the probabilities of cells belonging to various cell types.

SingleR [20]

SingleR is based on the gene expression profiles of cells with known types. By calculating the gene expression similarity (such as the Pearson correlation coefficient) between the single cells to be annotated and the reference cells, it assigns them to the cell type with the highest similarity, thus completing the cell type annotation. For the sake of fairness, the data input into SingleR in this study is in the form of a cell-gene score.

Convolutional neural network

For the sake of fairness in comparison, the convolutional neural network compared in this study is the same as the classifier in the fine-tuning task. It consists of one convolutional layer and two fully connected layers, and all are trained on the same training set. The trained model is then used to predict cell types.

Feedforward neural network

The feedforward neural network compared in this study was created using the MLPClassifier in sklearn [51]. Two hidden layers were set, with the ReLU activation function and the Adam optimization algorithm selected.

Additionally, similar to the aforementioned convolutional neural network, for the sake of fairness in comparison, they were all trained on the same training set and the trained models were utilized to make predictions.

Evaluation metrics

This work mainly adopts the following five metrics to comprehensively evaluate the performance of cell annotation: Acc, F1 score, Cohen's kappa coefficient, Jaccard_micro, and Jaccard_weighted. All of them are calculated using the functions in sklearn [51] of Python. Please refer to Additional file 2: Supplementary File 2 for specific information.

Abbreviations

scATAC-seq	Single-cell Assay for Transposase Accessible Chromatin with high-throughput sequencing
scRNA-seq	Single-cell RNA sequencing
BERT	Bidirectional Encoder Representations from Transformers
ACC	Accuracy
AD	Alzheimer's disease
UMAP	Uniform Manifold Approximation and Projection
TSS	Transcription start site
CNN	Convolutional neural network
KD	Knowledge distillation
ASC	Astrocytes
EX	Excitatory neurons
OPC	Olig.progenitors
INH	Inhibitory neurons
MG	Microglia
ODC	Oligodendrocytes
PER.END	Pericytes/endothelial
BNN	Bayesian neural networks

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02244-5.

Additional file 1: Fig. S1. Performance evaluation of annotations on different data. Fig. S2. Evaluation of the annotation performance of annATAC compared with RAINBOW and Cellcano on seven AD datasets and two control datasets. Fig. S3. Display of marker peaks and marker motifs in different cell types. Fig. S4. Model stability testing and ablation experiment on pre-training.

Additional file 2: Table S1. Detailed information of the pre-training datasets. Supplementary File 2. The specific calculation formulas for all evaluation indicators used in this study.

Acknowledgements

Not applicable.

Authors' contributions

L.Y.C. and F.W. conceived the problem and designed the study. G.H.W. and F.W. supervised the work. L.Y.C. and Q.M.L. performed deep learning experiments. H.F.L. and M.R.Z. performed bioinformatics analysis. L.Y.C. and F.W. wrote the manuscript, and other authors made modifications.

Funding

This work was supported by the National Natural Science Foundation of China (62450112, 62225109, 32400546, 62302342). We would like to thank the anonymous reviewers for their constructive suggestions.

Data availability

All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. The code and datasets of annATAC are freely available at the repository Zenodo (https://doi.org/10.5281/zenodo.15377860) and Github (https://github.com/Cuily-v/annATAC/tree/master).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]College of Life Science, Northeast Forestry University, Harbin 150040, China. [2]The Quzhou Affiliated Hospital of Wenzhou Medical University, Quzhou People's Hospital, Quzhou 324000, China. [3]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324003, China. [4]College of Artificial Intelligence, Henan University, Zhengzhou 450000, China. [5]College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China. [6]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China.

Received: 14 February 2025 Accepted: 12 May 2025
Published online: 28 May 2025

References

1. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9.
2. Sindhu C, Samavarchi-Tehrani P, Meissner A. Transcription factor-mediated epigenetic reprogramming*. J Biol Chem. 2012;287(37):30922–31.
3. Pashos ARS, Meyer AR, Bussey-Sutton C, O'Connor ES, Coradin M, Coulombe M, et al. H3K36 methylation regulates cell plasticity and regeneration in the intestinal epithelium. Nat Cell Biol. 2025;27:202–217.
4. Martino D, Kresoje N, Amenyogbe N, Ben-Othman R, Cai B, Lo M, et al. DNA methylation signatures underpinning blood neutrophil to lymphocyte ratio during first week of human life. Nat Commun. 2024;15(1):8167.
5. Yeung J, Florescu M, Zeller P, de Barbanson BA, Wellenstein MD, van Oudenaarden A. scChIX-seq infers dynamic relationships between histone modifications in single cells. Nat Biotechnol. 2023;41(6):813–23.
6. Preissl S, Gaulton KJ, Ren B. Characterizing cis-regulatory elements using single-cell epigenomics. Nat Rev Genet. 2023;24(1):21–43.
7. Westmann CA, Goldbach L, Wagner A. The highly rugged yet navigable regulatory landscape of the bacterial transcription factor TetR. Nat Commun. 2024;15(1):10745.
8. Teo AYY, Squair JW, Courtine G, Skinnider MA. Best practices for differential accessibility analysis in single-cell epigenomics. Nat Commun. 2024;15(1):8805.
9. Corces MR, Shcherbina A, Kundu S, Gloudemans MJ, Frésard L, Granja JM, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet. 2020;52(11):1158–68.
10. Myers BL, Brayer KJ, Paez-Beltran LE, Villicana E, Keith MS, Suzuki H, et al. Transcription factors ASCL1 and OLIG2 drive glioblastoma initiation and co-regulate tumor cell types and migration. Nat Commun. 2024;15(1):10363.
11. Sekine K, Onoguchi M, Hamada M. Transposons contribute to the acquisition of cell type-specific cis-elements in the brain. Communications Biology. 2023;6(1):631.
12. Zhang F-L, Feng Y-Q, Wang J-Y, Zhu K-X, Wang L, Yan J-M, et al. Single cell epigenomic and transcriptomic analysis uncovers potential transcription factors regulating mitotic/meiotic switch. Cell Death Dis. 2023;14(2):134.

13. Pastore A, Gaiti F, Lu SX, Brand RM, Kulm S, Chaligne R, et al. Corrupted coordination of epigenetic modifications leads to diverging chromatin states and transcriptional heterogeneity in CLL. Nat Commun. 2019;10(1):1874.
14. Schiroli G, Kartha V, Duarte FM, Kristiansen TA, Mayerhofer C, Shrestha R, et al. Cell of origin epigenetic priming determines susceptibility to Tet2 mutation. Nat Commun. 2024;15(1):4325.
15. Tedesco M, Giannese F, Lazarević D, Giansanti V, Rosano D, Monzani S, et al. Chromatin velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. Nat Biotechnol. 2022;40(2):235–44.
16. Zeng Y, Luo M, Shangguan N, Shi P, Feng J, Xu J, et al. Deciphering cell types by integrating scATAC-seq data with genome sequences. Nature Computational Science. 2024;4(4):285–98.
17. Li Y, Zhang D, Yang M, Peng D, Yu J, Liu Y, et al. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration. Nat Commun. 2023;14(1):6045.
18. Ziyani C, Delaneau O, Ribeiro DM. Multimodal single cell analysis infers widespread enhancer co-activity in a lymphoblastoid cell line. Communications Biology. 2023;6(1):563.
19. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biol. 2021;22(1):301.
20. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163–72.
21. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc. 2021;16(6):2749–64.
22. Tayyebi Z, Pine AR, Leslie CS. Scalable and unbiased sequence-informed embedding of single-cell ATAC-seq data with Cell Space. Nat Methods. 2024;21(6):1014–22.
23. Hu H, Quon G. scPair: boosting single cell multimodal analysis by leveraging implicit feature selection and single cell atlases. Nat Commun. 2024;15(1):9932.
24. Rachid Zaim S, Pebworth M-P, McGrath I, Okada L, Weiss M, Reading J, et al. MOCHA's advanced statistical modeling of scATAC-seq data enables functional genomic inference in large human cohorts. Nat Commun. 2024;15(1):6828.
25. Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nat Methods. 2023;20(9):1355–67.
26. Hao Y, Stuart T, Kowalski MH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42(2):293–304.
27. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet. 2021;53(3):403–11.
28. Stuart T, Srivastava A, Madad S, LareFigau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18(11):1333–41.
29. Tian L, Xie Y, Xie Z, Tian J, Tian W. AtacAnnoR: a reference-based annotation tool for single cell ATAC-seq data. Brief Bioinform. 2023;24(5):bbad268.
30. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Wang YXR. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10.
31. Ma W, Lu J, Wu H. Cellcano: supervised cell type identification for single cell ATAC-seq data. Nat Commun. 2023;14(1):1864.
32. Li S, Tang S, Wang Y, Li S, Jia Y, Chen S. Accurate cell type annotation for single-cell chromatin accessibility data via contrastive learning and reference guidance. Quantitative Biology. 2024;12(1):85–99.
33. Chen X, Chen S, Song S, Gao Z, Hou L, Zhang X, et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nature Machine Intelligence. 2022;4(2):116–26.
34. Das B, Chakraborty S. An improved text sentiment classification model using TF-IDF and next word negation. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv180606407D. Accessed 01 June 2018
35. Quinodoz SA, Ollikainen N, Tabak B, Palla A, Schmidt JM, Detmar E, et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell. 2018;174(3):744-57.e24.
36. Wang S, Luo Z, Liu W, Hu T, Zhao Z, Rosenfeld MG, et al. The 3D genome and its impacts on human health and disease. Life Med. 2023;2(2):lnad012.
37. Han M-H, Park J, Park M. Advances in the multimodal analysis of the 3D chromatin structure and gene regulation. Exp Mol Med. 2024;56(4):763–71.
38. Xiong L, Xu K, Tian K, Shao Y, Tang L, Gao G, et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. Nat Commun. 2019;10(1):4576.
39. Nahali S, Safari L, Khanteymoori A, Huang J. StructmRNA a BERT based model with dual level and conditional masking for mRNA representation. Sci Rep. 2024;14(1):26043.
40. Zeng R, Li Z, Li J, Zhang Q. DNA promoter task-oriented dictionary mining and prediction model based on natural language technology. Sci Rep. 2025;15(1):153.
41. Liu F, Yuan C, Chen H, Yang F. Prediction of linear B-cell epitopes based on protein sequence features and BERT embeddings. Sci Rep. 2024;14(1):2464.
42. 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.
43. Hao M, Gong J, Zeng X, Liu C, Guo Y, Cheng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024;21(8):1481–91.
44. 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.
45. Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: self-attention with linear complexity. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv200604768W. Accessed 01 June 2020.
46. Morabito S, Miyoshi E, Michael N, Shahin S, Martini AC, Head E, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. Nat Genet. 2021;53(8):1143–55.
47. Lareau CA, Duarte FM, Chew JG, Kartha VK, Burkett ZD, Kohlway AS, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol. 2019;37(8):916–24.
48. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. Cell. 2018;174(5):1309-24.e18.
49. Zhang K, Hocker JD, Miller M, Hou X, Chiou J, Poirion OB, et al. A single-cell atlas of chromatin accessibility in the human genome. Cell. 2021;184(24):5985-6001.e19.
50. Deng H, Yang Y, Li J, Chen C, Jiang W, Pu S. Fast Updating Truncated SVD for Representation Learning with Sparse Matrices. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240109703D. Accessed 01 Jan 2024.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.

Publisher's Note