BMC Bioinformatics

RESEARCH Open Access

Combining single-cell ATAC and RNA sequencing for supervised cell annotation

Jaidip Gill[1], Abhijit Dasgupta[2], Brychan Manry[2] and Natasha Markuzon[3*]

*Correspondence: natasha.markuzon@astrazeneca.com

[1] School of Public Health, Imperial College London, London, England
[2] Oncology Data Science, AstraZeneca, Gaithersburg, MD, USA
[3] Oncology Data Science, AstraZeneca, Waltham, MA, USA

Abstract

Motivation: Single-cell analysis offers insights into cellular heterogeneity and individual cell function. Cell type annotation is the first and critical step for performing such an analysis. Current methods mostly utilize single-cell RNA sequencing data. Several studies demonstrated improved unsupervised annotation when combining RNA with single-cell ATAC sequencing, but improvements in supervised methods have not been explored.

Results: Single-cell 10x genomics multiome datasets containing paired ATAC and RNA from human peripheral blood mononuclear cells (PBMC) and neuronal cells with Alzheimer's Disease were used for supervised annotation. Using linear and nonlinear dimensionality reduction methods and random forest, support vector machine and logistic regression classification models, we demonstrate the improvement in supervised annotation and prediction confidence in PBMC data when using a combination of RNA seq and ATAC-seq data. No such improvement was observed when annotating neuronal cells. Specifically, F1 scores were improved when using scVI embeddings to annotate PBMC sub-types. CD4 T effector memory cells showed the largest improvement in F1 score.

Keywords: RNA, ATAC, Machine learning, Cell annotation, Single-cell sequencing

Introduction

Single-cell sequencing technologies allow for detailed profiling of individual cell phenotypes, enabling downstream analysis like differential gene and protein expression. By leveraging multimodal technologies such as the 10X multi-omics protocol [1], and computational techniques like Seurat V3 [2], it becomes possible to generate multi-modal profiles of cellular phenotypes. For instance, single-cell RNA sequencing reveals gene expression profiles, while single-cell ATAC sequencing identifies accessible chromatin regions across the genome using Tn5 transposase [3]. These RNA and ATAC profiles can be utilized to construct gene-regulatory networks, describing the interactions between genes and regulatory DNA regions. [4]. Cell annotation, the identification of cell types, is a crucial step for such analyses, with the accuracy of downstream analyses being dependent on the quality of cell annotation. While Fluorescence-Activated Cell Sorting (FACS) serves as the gold standard for cell annotation, its practicality is limited for

large single-cell datasets due to high costs and practicality issues when analyzing thousands of cells [5]. Instead, unsupervised and supervised machine learning approaches are employed to quickly provide predicted cell annotations. Unsupervised methods have been widely used for cell annotation, but supervised methods have not been explored as deeply.

Cell annotation methodologies

Accurate identification and annotation of cell types is a critical step in single-cell sequencing analysis. Cell annotation methodologies can be categorized into two main methodologies: unsupervised and supervised. Unsupervised annotation involves clustering cells based on their sequencing profile, typically using RNA profiles. Subsequently, differentially expressed genes are identified in each cluster and compared to marker gene lists for potential cell types. [6]. Some methods also incorporate the use of under-expressed/negative gene markers to enhance annotation [7]. However, there are limitations to this approach. Firstly, cell subtypes may exhibit similar levels of expression for marker genes, hindering their discrimination [8]. Moreover, annotations are not standardized, leading to reduced reproducibility between experiments [9]. Lastly, manual annotation is laborious and time-consuming due to the need for individual inspection and comparison of clusters to known literature marker genes [9].

On the other hand, supervised annotation methods utilize classification models for cell annotation. These models are trained on a reference dataset with a similar cell-type composition as the unannotated dataset. The advantage of this approach is that the model is data-driven in its selection of features for determining annotation, rather than being literature-driven as in the unsupervised approach [10]. Additionally, supervised annotation workflows have the potential to use features previously unknown to discriminate between cell types.

While ATAC datasets are typically labeled using a reference RNA dataset [11], despite the fact that RNA enables greater cell type discrimination, ATAC data has not been fully utilized to enhance RNA classification. Given that cell types are often better differentiated at the epigenetic level, there is potential for ATAC to capture cell states that cannot be captured at the RNA level, particularly in cell subtypes such as those in T cells [12]. The additional epigenetic information captured by ATAC could therefore improve supervised cell classification of cell subtypes, especially in highly heterogeneous tissues such as cancers [13].

Integrating single-cell RNA and ATAC

One approach for simultaneously sequencing RNA and ATAC profiles at the single-cell level involves the use of microfluidics to isolate cells into separate droplets. In this method, cells are tagged with unique barcodes and then lysed. The RNA undergoes reverse transcription into barcoded cDNA for transcriptomic profiling, while the DNA provides epigenomic profiles, generating RNA and ATAC profiles for each cell. Examples of this approach include SNARE-seq [14] and the 10x multiomics protocols [15].

Another methodology, combinatorial indexing, employs multiple rounds of indexing to label cells. In each round, cells are randomly distributed across multiple wells and

tagged with a unique barcode specific to that well, resulting in a unique combination of barcodes to identify cells. Examples of this methodology include SNARE-seq2 [16] and SHARE-seq [17], offering the advantage of potential higher-throughput multi-omic sequencing.

Commercial methods like the 10x multi-omics protocol have strong support and easier software integration. For instance, the single-cell sequencing Python package Scanpy [18] provides readily available functions to read data generated specifically from 10x protocols. These multi-omic datasets can then be utilized to enhance annotations for unsupervised pipelines [19], unimodal datasets [20–22], or to integrate many -omic datasets simultaneously [23]. This integration is achieved by assuming that the relationship between RNA and ATAC from the unimodal datasets can be accurately derived from the multi-modal dataset.

Improvements in supervised annotation

In the pre-processing pipeline for annotation, dimensionality reduction plays a crucial role. Neural networks can enable embeddings to capture more complex, non-linear relationships for multi-modal co-embeddings [24] or unimodal embeddings using Generative Adversarial Networks [13]. For instance, the single-cell Variational Inference (scVI) autoencoder utilizes an assumption of the zero-inflated binomial distribution of RNA data to improve the latent representations (encoding) and in the generation of new RNA samples (decoding) [25]. These latent variables are modeled using a probability distribution that conditions for batch number, allowing for correction of batch effects. However, a limitation of these approaches is that neural networks, with their greater representational capacity, are more likely to fit to noise than linear methods, which is common in single-cell sequencing datasets [26]. Denoising methods exist for ATAC [27] and RNA [28], yet autoencoders have reduced interpretability compared to linear methods such as principal component Analysis (PCA) and CCA, which hinder the identification of peaks/genes driving differentiation into rare cell sub-types.

ATAC datasets are more sparse than RNA due to the abundance of closed chromatin sites across the entire genome [11], making methods such as PCA more computationally challenging [29]. Additionally, classifiers exhibit differences in their utility for supervised annotation, with nonlinear classifiers such as the SVM showing slightly stronger classification performance than linear models when using unimodal RNA data [9, 10] and unimodal ATAC data [30]. However, whether non-linear dimensionality reduction and classification models are needed to exploit ATAC features for improved annotation has not been fully explored.

It could be assumed that adding additional task-related information should improve model performance. However, the belief that ATAC features are task-related is an assumption that should be tested because the amount of relevant signal will depend on the tissue that cells belong to. Importantly, these benefits may be tissue-dependent. This means that the common-sense assumption may only hold true for certain biological contexts due to differences between tissues in the amount of cell-type distinction occurring at the epigenetic (ATAC) level. Similarly, there may be differences in ATAC utility for labelling specific cells within a tissue, again due to differences in the level of epigenetic encoding occurring in cells within a tissue. These findings are relevant to investigate as

they have implications for determining the contexts where RNA + ATAC annotation methods should be employed, or where RNA-only methods can be used for sequencing and/or annotation, potentially saving costs (via cheaper RNA-only sequencing), reducing dataset size and simplifying the annotation process. This would serve to guide use-cases for RNA + ATAC-based supervised annotation.

Research objectives

Our study aimed to compare the impact of different models for dimensionality reduction and classification on the utility of adding ATAC features. We also aimed to investigate how the predictive power of ATAC differed between immune cells and neuronal cells, due to the difference in epigenetic distinction occurring between tissue types. This would serve to guide use-cases for RNA + ATAC-based supervised annotation. Additionally, we sought to investigate epigenetic distinction occuring how the utility of ATAC features varied across different cell types and the level of granularity used to differentiate between these cell types.

Methods

Dataset description

Figure 1 shows an overview of the supervised annotation workflow. The datasets utilized in this study consisted of single-cell 10x genomics multiome datasets containing paired ATAC and RNA measurements. The first dataset comprised 11,909 human peripheral

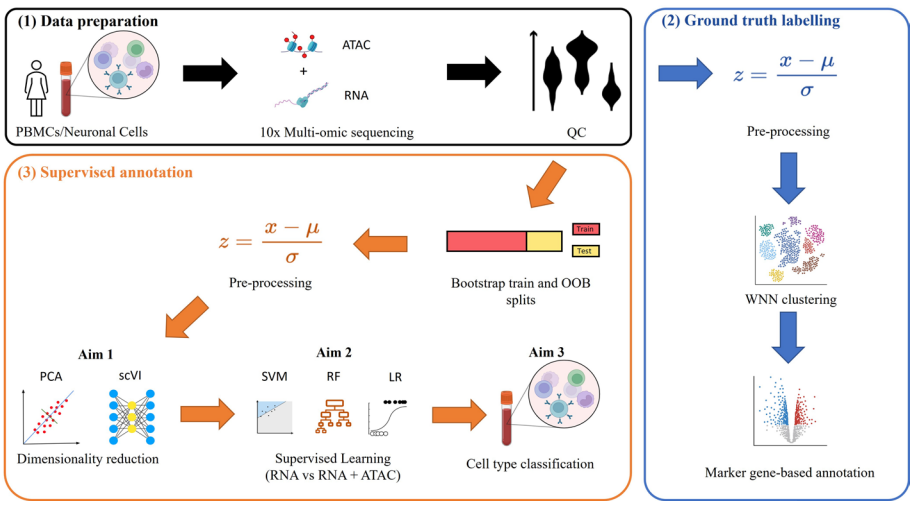

Fig. 1 Overview of supervised annotation workflow. Fig. 1 illustrates an overview of the supervised annotation workflow, which involved three primary steps. Initially, a 10x Genomics multi-omic (RNA + ATAC) sequencing dataset detailing peripheral blood mononuclear cells (PBMC) or neuronal cells underwent rigorous quality control (QC) procedures. Subsequently, ground truth labels were derived through a pre-processing pipeline, followed by weighted nearest neighbors (WNN) clustering and marker gene-based annotation. These labels served as the ground truth for the supervised classification models. The final step involved generating 10 bootstrapped train and out-of-bag (OOB) test sets. This was followed by pre-processing and dimensionality reduction using principal component analysis (PCA) or single-cell Variational Inference (scVI). Classification was then performed using support vector machine (SVM), random forest (RF), or logistic regression (LR) models. Models utilizing RNA-only embeddings were compared to those utilizing RNA and ATAC

blood mononuclear (PBMC) cells from a healthy female donor aged 25. Granulocytes were excluded via cell sorting. Sequencing was conducted using the Illumina Novaseq 6000 v1 kit and subsequent quality control and pre-processing (Supplementary Methods). The second dataset encompassed 10,530 neuronal cells obtained from seven patients diagnosed with Alzheimer's Disease (AD) (mean age 78 years) and eight unaffected controls (mean age 63 years). Quality control and annotations for the AD dataset were derived from a prior study.

Ground truth labelling

Using a third-party dataset would have reduced the amount of subjective bias in the manual annotation process used to generate the ground truth labels. However, the third-party labelled datasets available at the time of this study were labelled using only the RNA profiles of the cells. Such labels would be biased towards RNA-only models since any signals in the ATAC features would not have been considered when generating ground truth labels. Therefore, the only way to study the impact of ATAC features was to self-generate labels. This introduces some subjective bias in the ground truth labels, which we aimed to minimize by using best practices for manual annotation. Multi-modal clustering was used to take advantage of signals from both RNA and ATAC modalities. First, PCA was applied to the entire dataset (as annotations would serve as the ground truth). Then, the WNN method for calculating cell-cell distances [19] was implemented using the 'neighbours' function in Muon [50]. This was used with the Leiden clustering method [31]. Clusters were then annotated using manual annotation (Supplementary Methods).

Supervised learning pipeline

Bootstrapping

Bootstrapping was performed to control for variability in the training and test set data splits. Out-of-bag validation was performed by generating 10 training datasets via sampling with replacement from the original dataset, where each bootstrap sample was 100% of the original sample size.

Dimensionality reduction

Embedding was performed on pre-processed data using two methods: PCA and scVI autoencoder. The Scikit-learn [32] implementation of PCA was fit to each training dataset and used to transform each train-test bootstrap pair. Scvi-tools [33] was used to implement the scVI autoencoder as another method for embedding. This was performed independently for each modality, with the parameters discussed in Supplementary Methods.

Classification models

Classification models were implemented using the Scikit-learn implementation of random forest classifier (RF), support vector machine (SVM) and logistic regression (LR). Each model was trained on 10 bootstrapped datasets of the embedded data, and tested on corresponding out-of-bag datasets, with separate pre-processing pipelines. Balanced class weights were used for each model to adjust for class imbalance. Within each

bootstrap dataset, hyperparameter tuning was performed using five-fold cross-validation. Table S1 shows the hyperparameter grid used for each model.

Model evaluation

Unsupervised metrics

PCA and scVI-embedded training datasets were visualized in two dimensions using Uniform Manifold Approximation and Projection (UMAP) [34] and labelled using ground truth labels. Silhouette scores were calculated using Euclidian distances calculated from the PCA and scVI embeddings for each of the bootstrap training and out-of-bag test sets. Silhouette scores were reported as the median and IQR due to a small sample size (10) and non-normal distribution.

Supervised metrics

Proportion of ambiguous predictions

The Proportion of Ambiguous Predictions (PAP) score metric was developed as a method for measuring supervised model confidence based on the polarity of predicted probabilities. This was because supervised annotation methods typically leave low-confidence predictions unannotated as they may be of a cell type not represented in the reference/training dataset. Therefore, PAP scores would indicate the tendency of models to leave cells unannotated.

Models were trained on each bootstrap training set and prediction probabilities for each class were extracted for the corresponding OOB test sets. The empirical cumulative distribution function (ECDF) was calculated using these prediction probabilities. PAP scores for the i^{th} class/cell type were then calculated as the proportion of predictions that fall between the 0.1 and 0.9 quantile points:

$$\mathrm{PAP}_i = \mathrm{ECDF_i}(0.9) - \mathrm{ECDF_i}(0.1). \tag{1}$$

PAP scores for the i^{th} class were reported as the median and IQR over 10 bootstrap samples for a given model. Other classification metrics are described in the Supplementary Methods.

Statistical analysis

Since metrics did not show sufficient evidence of normality, non-parametric tests were used to compare models using RNA and ATAC (RNA + ATAC) to RNA alone. RNA-based models have been shown to outperform ATAC-based models for cell annotation due to the higher predictive power of the transcriptome, so RNA + ATAC models were compared to RNA-alone rather than ATAC-alone [11]. The Wilcoxon-signed rank test was used as model performance metrics were paired by the given bootstrap sample. This approach has been previously recommended for supervised classification because it has increased power where paired t-test assumptions are violated due to the lack of normality that is often seen when comparing classifiers [35]. Multiple testing adjustment was then performed (Supplementary Methods). The importance of RNA and ATAC components was investigated using a lasso regression model [51], and correlations within scVI components were calculated using Pearson correlation coefficients (Supplementary Methods).

Results

ATAC increases separation between cell type embeddings

To visualize how ATAC features alter the discriminability of cell populations, UMAPs were applied to the PCA and scVI embeddings of the bootstrap training samples with and without the ATAC features. Figure 2(A-D) shows a scatter plot of the first two UMAP components with and without the ATAC features applied to the PCA embeddings (A - B) and scVI embeddings (C - D) for the first bootstrap training set. Panels A and B show that CD4 TEM, CD8 TEM and CD8 naive populations appear better separated from the CD4 naive and CD4 TCM populations when ATAC features are included (B) in the PCA embeddings than compared to using RNA alone (A). On the other hand, scVI embeddings showed little difference between RNA only (C) and combined embeddings (D). Quantitative analysis using silhouette scores (E) showed that combined embeddings led to significantly increased silhouette scores compared to RNA alone for training and test sets for both PCA and scVI methods. The scVI test set showed the greatest difference between the RNA-only embeddings (median silhouette score = 0.068, IQR = [0.067, 0.071] and the combined embeddings (median silhouette score = 0.095, IQR = [0.093, 0.098], $p < 0.01$).

ATAC improves classification performance for scVI embeddings

Since the choice of embedding method leads to changes in the separability of cell-type clusters in the UMAP feature space, we next investigated whether this led to differences

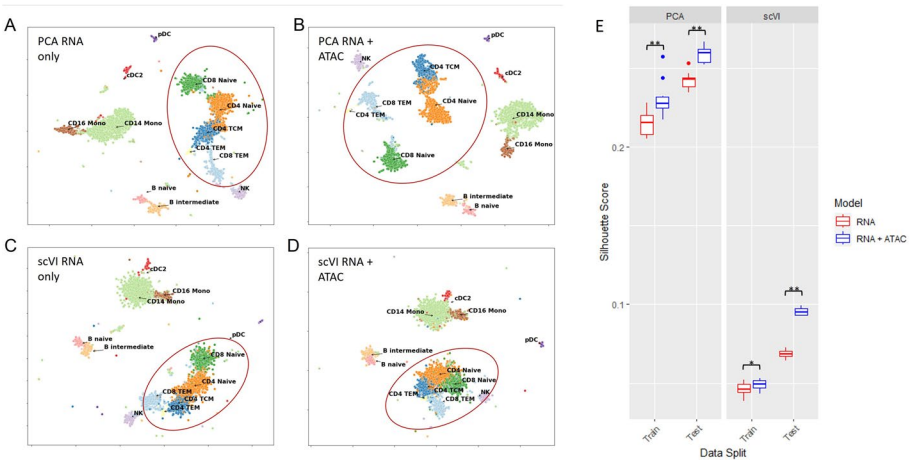

Fig. 2 Qualitative and quantitative representations of cell type embeddings. UMAP **A**–**D** transformations were applied to cell embeddings generated using PCA on RNA alone (**A**) or RNA and ATAC (**B**), and using scVI on RNA alone (**C**) or RNA and ATAC (**D**). Cells are labelled according to manual annotations generated using the weighted nearest neighbours method. The scatterplots shown are representative examples from a single bootstrap training sample. Red circles indicate T cell embeddings, which showed the largest difference upon inclusion of ATAC features for PCA. Silhouette scores (**E**, **F**) were calculated using Euclidean distances calculated in the PCA (**E**) or scVI (**F**) feature space. Silhouette scores are represented using boxes indicating the median, lower and upper quartile over 10 bootstrap samples. Whiskers indicate 1.5 times the interquartile range and dots indicate outliers beyond the whiskers. Wilcoxon-signed rank test with Bonferroni correction was used to compare RNA-alone to RNA and ATAC models (* = $p < 0.05$, ** = $p < 0.01$). Abbreviations: PCA = Principal Component Analysis, scVI = single-cell Variational Inference autoencoder, TEM = T Effector Memory, TCM = T Central Memory, Mono = Monocyte, NK = Natural Killer cell, cDC = Conventional Dendritic Cell, pDC = Plasmacytoid Dendritic Cells

in supervised model prediction quality. Figure S3 shows the macro f1 scores for different combinations of embedding and classification models, using RNA alone or RNA and ATAC (full results shown in Table S3). PCA-based embeddings led to weaker classification performance compared to scVI-based embeddings. The best performance using PCA was found using the RF with RNA + ATAC (median macro F1 score = 0.919, IQR = [0.894, 0.929]). In contrast, the best performance from scVI embeddings (RNA + ATAC features) showed improved performance (median macro F1 score = 0.946, IQR = [0.940, 0.949]). RNA + ATAC features yielded much smaller advantages in PCA-based embeddings compared to scVI. No significant differences in macro F1 score were detected between using RNA alone compared to RNA and ATAC for PCA-based embeddings. In contrast, scVI embeddings led to differences in macro F1 score between RNA alone and RNA + ATAC for every classification model. The best-performing scVI model (SVM and RNA + ATAC) showed significantly improved classification performance compared to RNA alone (median macro F1 score = 0.907, IQR = [0.902, 0.910], $p < 0.05$). Overall, advantages in RNA + ATAC classification performance compared to RNA alone required the scVI embeddings rather than PCA. However, the effect of RNA + ATAC did not show dependence on the choice of classification model.

T cells exhibit the highest improvements in annotation quality

Having established the effect of model choice on ATAC utility, we next investigated the biological context of ATAC utility using the embedding and classifier combination with the strongest classification performance (scVI-SVM). To characterise the cell-specific utility of ATAC features, class-specific F1 and PAP scores were recorded for each model using RNA-only or RNA + ATAC.

Figure 4A shows that RNA + ATAC features led to significantly increased F1 scores compared to RNA alone in every cell type. T cell subtypes showed the largest gains. Specifically, CD4 TEM cells showed the largest increase, from the RNA-only model (median F1 score = 0.815, IQR = [0.800, 0.868]) compared to the RNA + ATAC model (median F1 score = 0.926, IQR = [0.861, 0.937], $p < 0.01$). T cells also showed the largest reductions in PAP scores as well, with CD4 naive biggest reduction from a median of 0.118 (IQR = [0.113, 0.120]) using RNA alone to a median of 0.066 (IQR = [0.061, 0.0691], $p < 0.01$) using RNA + ATAC. The exception to this was the CD4 TEM, which did not show any significant reduction in PAP. In contrast, dendritic cell subtypes (pDC and cDC2) only showed significant improvements in F1 score, with cDC2 showing a very small (but statistically significant) increase when moving from RNA alone (median F1 score = 0.901 [IQR = [0.848, 0.915]) to RNA + ATAC (median F1 score = 0.907, IQR = [0.887, 0.926], $p < 0.05$). Overall, T cell subtypes were most improved by the use of RNA + ATAC features, in particular CD4 TEM, whereas dendritic cells showed the least improvement.

ATAC only confers advantages when annotating at the subtype level

Since ATAC utility depends on cell type, we then investigated whether the utility of ATAC is sensitive to the granularity of the ground truth labels. The scVI-SVM model was used to classify cells using a ground truth annotated with less granular cell subtypes (WNN-L1). Model performance using the previously utilized WNN-L2 labels is

also shown for comparison. As previously shown, the scVI-SVM model using WNN-L2 labels led to significant increases in macro precision, recall and F1-score. For WNN-L1, Fig. 5 shows that macro precision was significantly increased for WNN-L1 ground truth labels from a median of 0.972 (IQR = [0.969, 0.975] using RNA alone to a median of 0.9855 (IQR = [0.985, 0.986], $p < 0.05$). However, there was no significant increase in macro recall. Thus, no significant increase in overall classification performance was observed when comparing RNA alone (median macro F1 score = 0.901, IQR = [0.892, 0.922]) to RNA + ATAC (median macro F1 score = 0.913, IQR = [0.896, 0.942]). Overall, RNA + ATAC features only lead to significant improvements in classification performance for the more granular WNN-L2 cell subtype labels.

ATAC embeddings are associated with B and T cell subtypes

Since ATAC features improved supervised annotation quality in a cell type-dependent manner, we next investigated how each modality was contributing to classification prediction. Pearson correlation coefficients were calculated for the scVI-encoded RNA + ATAC feature set, and a multinomial lasso model was fit to the entire dataset to estimate the linear contributions from each modality to each class. Figure 6A shows that correlations within the feature set were generally low, indicating low structure within the feature set. Between modalities, 98.9% of Pearson correlation coefficients were between -0.3 and 0.3. Figure 6B shows the total coefficient contribution of ATAC components to each class relative to RNA components. The cell type which showed the highest proportional association with ATAC components was CD4 TCM (66.6%) followed by CD4 TEM (57.4%) and CD8 Naive (55.1%). The cell types showing the least relative association with ATAC components were CD8 TEM (36.3%), pDC (40.3%) and cDC2 (43.3%). Overall, ATAC components showed little linear association with RNA components, and T cells, except for CD8 TEM, showed high associations with ATAC components relative to RNA.

ATAC does not confer an advantage in neuronal cell annotation

Because ATAC utility may depend on the biological context of the annotation task, we next evaluated ATAC utility for annotation in a different tissue. Figure 7 shows UMAP visualizations of RNA (A) and RNA + ATAC (A-B) scVI embeddings, as well as comparing performance for different classifiers (C) and cell-types (D-E). UMAP clusters appear more distinct when adding ATAC features (B) than without (A). This effect is strongest in the inhibitory and excitatory cell types. However, classification performance was not significantly different between RNA and RNA + ATAC models for LR, and was significantly lower when adding ATAC features for SVM models (Fig. 7C). The decrease in F1 score was greatest in the PCA-SVM model (RNA-only median macro F1 score = 0.871, IQR = [0.826, 0.894], RNA + ATAC median macro F1 score = 0.812, IQR = [0.801, 0.824], $p < 0.01$). For the best-performing model (scVI-LR), there were no significant changes in F1 or PAP scores for any cell types (Fig. 7D-E). Overall, ATAC features do not confer an advantage in classification performance for neuronal cell types compared to RNA alone.

Discussion

This study aimed to explore the utility of combining ATAC with RNA in supervised annotation and its interaction with model choice and cell types. We found that adding ATAC features improved annotation quality for all classifiers tested (RF, LR and SVM) for PBMCs when using scVI autoencoder. Increases in annotation quality were highest in T cell subtypes, whereas dendritic cells showed the smallest improvement. These improvements in annotation quality were lost if the task only required annotation of major PBMC cell types and when annotating neuronal cells.

The effects of dimensionality reduction and classifier on ATAC utility

UMAP visualisations showed incongruency with classification scores when evaluating the impact of dimensionality reduction and classifier choices. For the PBMCs, PCA appeared to improve the gain in the separation of cell types by ATAC in the UMAP feature space compared to scVI in representative train (Fig. 2) and test sets (Figure S1). However, this did not result in improvements in supervised classification performance. For the neuronal cells, adding ATAC appeared to improve cluster separation, despite no improvements in the classification task (Fig. 7). One reason for these incongruencies could be that the UMAP method used was unsuitable for the dataset. Some research has shown UMAP to outperform other visualisation methods, such as t-distributed Stochastic Neighbor Embedding (t-SNE), in the visualisation of PBMC datasets [36]. However, other research indicates that differences in performance between UMAP and t-SNE are attributable to the initialization used in the gradient descent algorithm for fitting each algorithm [37]. More specifically, effective visualizations depend on using an informative initialization. This means using PCA or scVI before fitting UMAP (or t-SNE) to improve the optimization of the UMAP fit. Since our use of UMAP utilized informative initialization, the performance of the UMAP should have been effective in capturing cell-cell relationships. A more likely reason for the lack of correspondence between visualisation and prediction results could be that UMAP distances are not directly interpretable since UMAP preserves the topological relationships between cells, rather than the real geometrical distances [38]. This focus on topology generates effective visualisations but can lead to arbitrary differences. UMAP is sensitive to parameters such as n_components (balance between local and global structure) and min_dist (minimum distance between points in the latent space) [34]. These directly affect the distances in the latent representation of cells and cannot be objectively tuned. This would also explain why silhouette scores were significantly improved for scVI embeddings but did not appear to improve in the UMAP visualization - because silhouette scores were calculated using distances in a feature space (PCA and scVI) that preserved real differences between cells. Overall, the addition of ATAC appears to improve cluster separation in both PBMCs and neuronal cells. However, these improvements do not translate to the cell classification problem.

ATAC features were only beneficial when using scVI rather than PCA for PBMCs (Fig. 3, Figure S3). The superiority of scVI to PCA has been demonstrated before over varying numbers of components [36]. Autoencoders like scVI are capable of capturing non-linear relationships within the ATAC feature matrix [25], whereas PCA is only capable of linear representations. Therefore, scVI has a greater representational capacity than PCA and thus retains more information from the original ATAC feature matrix.

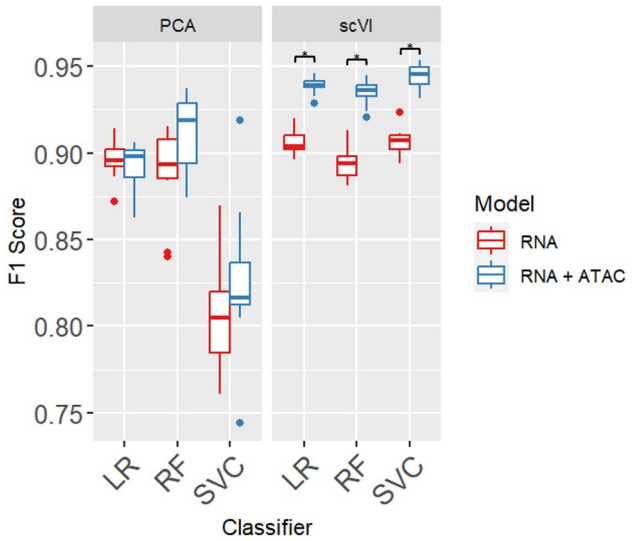

Fig. 3 Embedding and classifier-dependent effects of ATAC on annotation quality. Boxes indicate the median, lower and upper quartile over 10 bootstrap out-of-bag test samples for macro F1 scores. Whiskers indicate 1.5 times the interquartile range and dots indicate outliers beyond the whiskers. Wilcoxon-signed rank test with Bonferroni correction was used to compare RNA-alone to RNA and ATAC models (* = $p <$ 0.05). Abbreviations: LR = Logistic Regression, RF = Random Forest, SVC = SVM Classifier, PCA = Principal Component Analysis, scVI = single-cell Variational Inference autoencoder

The lack of representational capacity of PCA was further demonstrated by the fact that the proportion of variance explained was less than 12% for each bootstrap sample for RNA and ATAC embeddings. However, ATAC-driven improvements in annotation quality did not depend on the classifier used. This means the relationship between ATAC signals captured in the embeddings and cell type can be captured in non-linear (RF and SVM with polynomial kernel) and linear models (LR). For RNA-based annotations, non-linear methods (neural networks and non-linear SVMs) show advantages in classification accuracy but not macro F1 scores when compared to linear methods [10], which is a less biased metric in imbalanced datasets such as the one used in that study. In this study, the difference in median macro F1 score between the scVIRNA + ATAC SVM and LR models was only 0.024 (Fig. 3). Since linear classifiers show comparable performance to non-linear classifiers whilst also being able to take advantage of ATAC features, linear methods may be preferred when parsimony and interpretability are more important than marginal gains in annotation quality.

The effect of ATAC on the annotation quality of PBMC cell types

The effect size of improvement using ATAC showed cell-type dependency (Fig. 4). T cells showed the greatest changes in F1 score in response to the addition of ATAC features, and showed the highest associations with ATAC embeddings (Fig. 6B). One reason for this may be due to the abundance of 'poised/bivalent states' in T cells. Although RNA poising also occurs (genes are transcribed but not translated) in T cells [39], NK cells [40] and B cells [41], epigenetic poising is an additional mechanism where gene promoters exhibit both activating and repressing histone modifications, enabling rapid gene promotion upon external signalling [42]. These domains have been observed in CD4 and CD8 T cells, where 20–30% of silent genes are poised [12]. CD8 memory cells

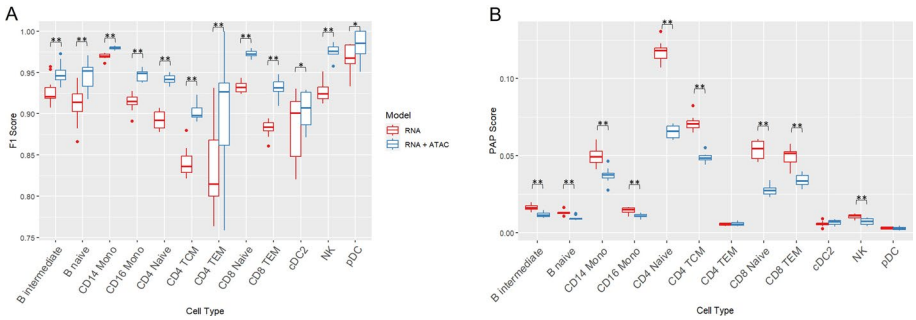

Fig. 4 The effect of ATAC embeddings stratified by cell type. T cells exhibit the highest improvements in annotation quality when adding ATAC. Dendritic cells showed the least improvement. Boxplots show F1 (**A**) and PAP scores (**B**) per class using scVI embeddings with an SVM classifier. Boxes indicate the median, lower and upper quartile over 10 bootstrap out-of-bag test samples. Whiskers indicate 1.5 times the interquartile range and dots indicate outliers beyond the whiskers. Wilcoxon-signed rank test with Benjamini-Hochberg correction was used to compare RNA-alone to RNA and ATAC models (* = $p < 0.05$, ** = $p < 0.01$). Abbreviations: PAP = Proportion of Ambiguous Predictions, TEM = T Effector Memory, TCM = T Central Memory, Mono = Monocyte, NK = Natural Killer cell, cDC = Conventional Dendritic Cell, pDC = Plasmacytoid Dendritic Cells

have been shown to have a naive-like RNA profile, but an effector-like ATAC profile [43]. Furthermore, poised domains are especially apparent in memory cells to enable rapid activation of the secondary immune response [12]. These domains are undetected at the transcriptomic level since they occur before transcriptomic changes that lead to cell differentiation. This may explain why RNA-only models showed the greatest misclassification between CD4 TCM and CD4 naive cells, whereas the addition of ATAC reduced this misclassification (Figure S2). It would also explain why CD4 TEM cells showed the highest gain in median macro F1 score (0.112).

A limitation to the view of poised domains driving enhanced T cell classification is that improvements in CD4 TEM cells also showed high variation between bootstrap samples (Fig. 4), no increase in prediction confidence and CD8 TEM showed the lowest association with ATAC components relative to RNA components (Fig. 6B). However, these can be explained by technical issues: the reason for the high variation in scores for CD4 TEM is likely due to the low sample frequency (1%, Table S2) combined with CD4 TEM being a T cell subtype, making classification harder than for other low sample but distinct cell types (cDC2 and pDC). Another mechanism could be distal regulatory histone modifications driving the detection of differences between cell types because distal modifications are more dynamic throughout T-cell development than proximal modifications [44]. Additionally, promoter modifications are more stable than corresponding RNA levels in immune cells [44]. Both mechanisms may be relevant signals to capture using ATAC to improve immune cell classification, but poised domains may provide an additional signal that enhances T cell classification further.

The improvements in PAP scores would have unique consequences in cell labelling. This is because many supervised annotation models label low-confidence cell types as 'unknown' since low confidence may be indicative of a cell type not present in the reference dataset [45, 46]. The inclusion of ATAC features increased prediction confidence of B and T cell types (Fig. 4B) whilst also improving the accuracy of their annotation.

This indicates that the additional confidence is not misplaced and may lead to reduced unnecessary labelling of T and B cell types as 'unknown'.

Dendritic cells showed the smallest improvement in classification and no improvement in prediction confidence (Fig. 4). This is unlikely to be due to sample size since CD4 TEM cells also only represented 1% of cell types (Table S2). A better explanation may be that F1 and PAP scores were already saturated using RNA alone, making further improvements difficult for this dataset. This would explain the small improvement in the median macro F1 score when adding ATAC for pDC (0.018) since it was already 0.968 with RNA alone. Without this saturation, pDC cells may have shown greater improvements since they contain distinct peaks, such as accessible regions near TSPAN13 [47]. However, this theory may not apply to cDC2 cells. This is because they showed the smallest improvement in the median macro F1 score (0.006) despite room to improve in the baseline median F1 score of 0.901 using RNA alone. This is a lower baseline than other cell types that showed much larger improvements. Therefore, ATAC signals may not carry as much importance in the identification of dendritic cell subtypes, specifically cDC2, compared to B and T cells. This would be because epigenetic differences (such as poising) are important for marking the different stages of B and T cell development. Since dendritic cells are antigen-presenting cells, they do not form memory cells that require epigenetic poising for rapid activation as B and T cells do. [48].

The effect of ATAC on the annotation quality of neuronal cell types

ATAC features did not improve classification quality for any cell types when using the best performing model (LR-scVI), and actually led to significantly worse classification quality when using SVM for both PCA and scVI embeddings (Fig. 7). One reason for this could be technical since the neuronal dataset utilized samples from 15 individuals (rather than the one used for PBMCs), along with two different health states (healthy vs Alzheimer's). The additional inter-individual and disease-state variation could have diluted cell-type signals in the ATAC features, since inter-individual variation can lead to decreases in annotation accuracy [10]. However, this wouldn't explain the significantly worse performance from ATAC observed in the SVM models (Fig. 7C). This instead indicates that ATAC introduced more noise than cell-type signals. This would have a stronger effect on the SVM since its non-linearity provides a greater representational capacity than the linear LR, leading to a greater likelihood of overfitting to the noise. ATAC may provide less signal in neuronal cells than PBMCs due to the highly dynamic and rapid differentiation seen in immune cells as previously highlighted compared to the slower and more constant cell types observed in neuronal cells. Since the signal provided by ATAC depends on the granularity of cell type labels (Fig. 5), further work should investigate whether ATAC may be useful for annotating more granular neuronal cell type labels. Overall, ATAC may only improve supervised annotation in tissues with high epigenetic diversity between cell subtypes, such as PBMCs.

Cell annotation

The cluster annotation step that occurs after clustering remains a limitation of most single-cell studies that rely on manually annotated ground truths. This is because

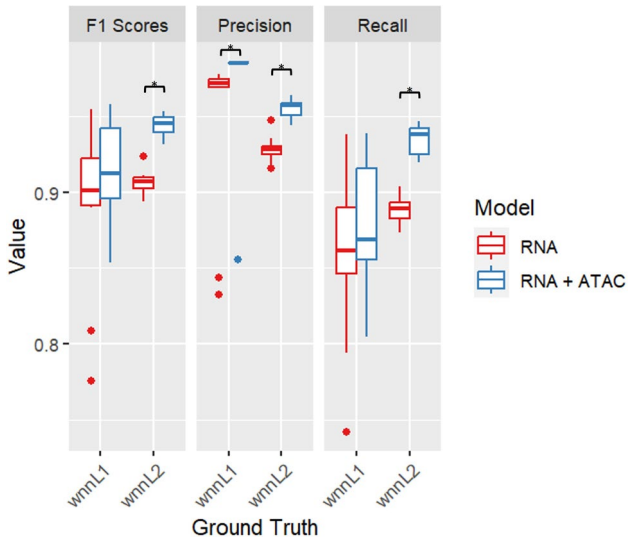

Fig. 5 Sensitivity of ATAC-derived improvements to ground truth cell type granularity.. The addition of ATAC features only leads to significant improvements in classification performance for the more granular WNN-L2 cell subtype labels. Boxplots show macro F1, precision and recall scores per class using scVI embeddings with an SVM classifier. Results are stratified by ground truth containing cell subtypes (WNNL2) or major cell types only (WNNL1). Boxes indicate the median, lower and upper quartile over 10 bootstrap out-of-bag test samples for macro-average metrics. Whiskers indicate 1.5 times the interquartile range and dots indicate outliers beyond the whiskers. Wilcoxon-signed rank test with Bonferroni correction was used to compare RNA-alone to RNA and ATAC models (* = $p < 0.05$)

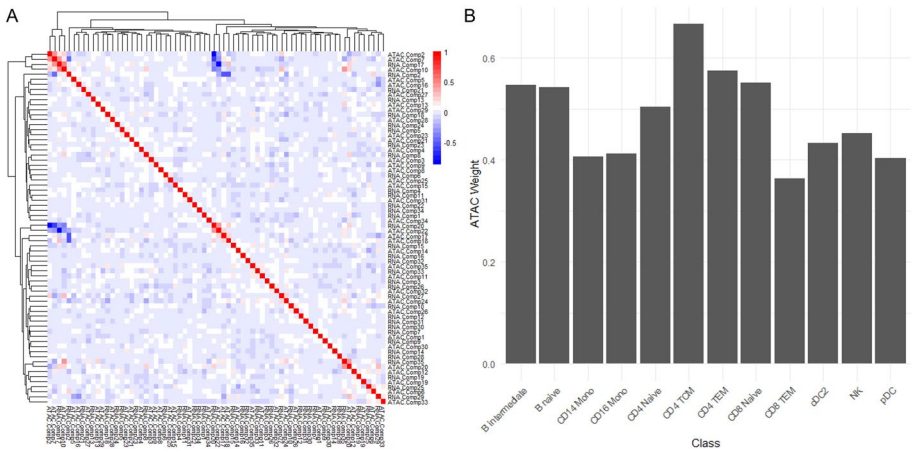

Fig. 6 Relationships between embedded features and cell types. (**A)** Pearson correlation heatmap showing correlations between scVI-embedded features of RNA and ATAC modalities. (**B)** Bar plot showing the ratio of the sum of absolute coefficients for ATAC components to the sum of absolute coefficients for both ATAC and RNA components for each cell type. Coefficient values were extracted from a multinomial lasso model fit to the entire embedded dataset

reproducibility is limited by the choice of marker gene lists (since this is not standardized) and annotations do not account for negative gene markers [7] or cell type markers from other modalities (such as protein signatures). Indeed, the weakness of any ground truth annotation generated using a manual annotation method is that it

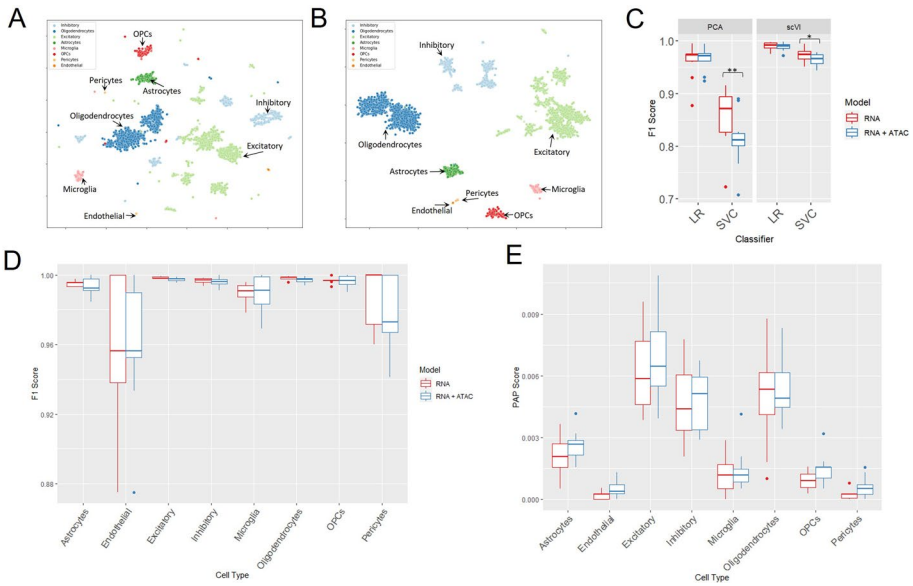

Fig. 7 UMAP visualizations and classification scores in a neuronal cell dataset. (**A**, **B)** UMAP visualisations of scVI embeddings using RNA alone (**A**) or RNA and ATAC (**B**). The scatterplots shown are representative examples from a single bootstrap training sample. (**C)** Embedding and classifier-dependent effects of adding ATAC on annotation quality. (**D**–**E)** Cell-specific effects of ATAC on classification quality and prediction confidence on the best performing model (LR-scVI). Boxplots indicate the median, lower and upper quartile over 10 bootstrap out-of-bag test samples. Whiskers indicate 1.5 times the interquartile range and dots indicate outliers beyond the whiskers. Wilcoxon-signed rank test with Bonferroni (**C**) or Benjamini-Hochberg (**D**–**E)** correction was used to compare RNA-alone to RNA and ATAC models (* = $p < 0.05$, ** = $p < 0.01$). Abbreviations: LR = Logistic Regression, OPC = Oligodendrocyte progenitor cell, PAP = Proportion of Ambiguous Predictions, SVC = Support Vector Machine

is less reliable compared to the gold-standard method of cell annotation using FACS sorting. Therefore, calculated model accuracy in this analysis is not completely representative of actual model accuracy. This study also only utilized two datasets, which constrains the generalizability of the findings, as these datasets may not comprehensively represent PBMCs and neuronal cells. Future studies will explore broader datasets to validate and expand upon these results.

Embedding methods

Since ATAC utility depended on the method of embedding used, future studies could explore the use of improved embedding methods to enhance the latent representation of the ATAC features, such as deep adversarial learning [13]. Such non-conventional machine learning methods have previously demonstrated the ability to extract additional signals from RNA data compared to conventional models, suggesting they could similarly improve annotation performance even further from the combined analysis of RNA and ATAC data than the conventional methods used in this study. Since autoencoders may fit to noise in sequencing data due to their high representational capacity, future work could also explore

the utility of denoising models in multi-omic supervised annotation methods. Sequencing data is noisy due to overdispersion and zero inflation from dropout events during sequencing [26]. Deep learning methods can enhance denoising workflows to denoise ATAC sequencing data [27]. Also, co-embeddings could be used to reduce redundancy between modalities. For example, Multi-omic Factor Analysis extends PCA to multiple modalities simultaneously [49]. This could be necessary for ATAC to confer advantages using PCA rather than an autoencoder.

Conclusion

Overall, this study demonstrated that combining ATAC with RNA embeddings generated using the scVI autoencoder substantially improve the quality of supervised annotation and prediction confidence in PBMCs for both linear (like LR) and non-linear (RF and SVM) classifiers. A primary driver behind these improvements was the heightened ability to distinguish between cell subtypes, particularly within the T cell category. However, improvements were not observed for neuronal cell types. Improved capabilities in cell type identification can reduce dataset variation, paving the way for a heightened capacity to uncover novel -omic signatures in single-cell research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06084-6.

Additional file 1.

Acknowledgements
We thank Asaf Rotem and Grant Duclos for their helpful suggestions with data interpretation.

Author contributions
J.G., A.D.G. and N.M. have made contributions to the conception and the design of this study. J.G., A.D.G., B.M., and N.M. contributed to the analysis and interpretation of the data. J.G. conducted the experiments and created new software used in this work. J.G. and N.M. have drafted and revised the manuscript.

Data availability
The 10x multi-omics PBMC dataset used in this study is available on the 10x Genomics website: https://www.10xgenomics.com/resources/datasets. The pre-processed AD dataset, along with annotations, is available on the Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214979. The code used for the analysis is available: https://github.com/JaidipGill/supervised-single-cell.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
A.D.G., B.M., and N.M. were employees of AstraZeneca when this work was performed and may have stock ownership, options, or interests in the company.

Received: 20 November 2024 Accepted: 13 February 2025
Published online: 26 February 2025

References

1. Belhocine K, DeMare L, Habern O. Single-cell multiomics: simultaneous epigenetic and transcriptional profiling: 10x genomics shares experimental planning and sample preparation tips for the chromium single cell multiome atac+ gene expression system. Genetic Eng & BiotechNews. 2021;41(1):66–8.
2. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902.e21.

3. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486–90.
4. Kartha VK, Duarte FM, Yan H, Ma S, Chew JG, Lareau CA, Earl A, Burkett ZD, Kohlway AS, Lebofsky R, Buenrostro JD. Functional inference of gene regulation using single-cell multi-omics. Cell Genom. 2022;2(9):100166.
5. Davey HM, Kell DB. Flow cytometry and cell sorting of heterogeneous microbial populations: the importance of single-cell analyses. Microbiol Rev. 1996;60(4):641–96.
6. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. Nat Methods. 2019;16(10):983–6.
7. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. Nat Commun. 2022;13(1):1246.
8. Zhang X, Lan Y, Jinyuan X, Quan F, Zhao E, Deng C, Luo T, Liwen X, Liao G, Yan M, et al. Cellmarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res. 2019;47(D1):D721–8.
9. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. 2019;20(1):194.
10. Ma W, Kenong S, Hao W. Evaluation of some aspects in supervised cell type identification for single-cell RNA-seq: classifier, feature selection, and reference construction. Genome Biol. 2021;22(1):264.
11. Pott S, Lieb JD. Single-cell atac-seq: strength in numbers. Genome Biol. 2015;16:1–4.
12. Cuddapah S, Barski A, Zhao K. Epigenomics of t cell activation, differentiation, and memory. Curr Opin Immunol. 2010;22(3):341–7.
13. Yang H, Wei Q, Li D, Wang Z. Cancer classification based on chromatin accessibility profiles with deep adversarial learning model. PLoS Comput Biol. 2020;16(11): e1008405.
14. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat Biotechnol. 2019;37(12):1452–7.
15. Belhocine K, DeMare L, Habern O. Single-Cell multiomics: simultaneous epigenetic and transcriptional profiling. Genetic Eng & Biotechnol News. 2021;41(1):66–8.
16. Plongthongkum N, Diep D, Chen S, Lake BB, Zhang K. Scalable dual-omics profiling with single-nucleus chromatin accessibility and mrna expression sequencing 2 (snare-seq2). Nat Protoc. 2021;16(11):4992–5029.
17. Kim Samuel H, Marinov Georgi K, Bagdatli S Tansu, Higashino Soon Il, Shipony Zohar, Kundaje Anshul, Greenleaf William J (2022) Simultaneous single-cell profiling of the transcriptome and accessible chromatin using share-seq. In *Chromatin Accessibility: Methods and Protocols*, pages 187–230. Springer.
18. Wolf FA, Angerer P, Theis FJ. Scanpy: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:1–5.
19. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587. e29.
20. Ashuach Tal, Gabitto Mariano I, Koodli Rohan V, Saldi Giuseppe-Antonio, Jordan Michael I, Yosef Nir. Multivi: deep generative model for the integration of multimodal data. *Nature Methods*, pages 1–10, 2023.
21. Gong B, Zhou Y, Purdom E. Cobolt: integrative analysis of multimodal single-cell sequencing data. Genome Biol. 2021;22(1):1–21.
22. Hao Y, Stuart T, Kowalski Madeline H, Choudhary S, Hoffman P, Hartman A, Srivastava A, Molla G, Madad S, Fernandez-Granda C, Satija R. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, pages 1–12, May 2023.
23. Ghazanfar S, Guibentif C, Marioni John C. Stabilized mosaic single-cell data integration using unshared features. *Nature Biotechnology*, pages 1–9, May 2023.
24. Lin Y, Wu T-Y, Wan S, Yang JYH, Wong WH, Rachel Wang YX. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. Nat Biotechnol. 2022;40(5):703–10.
25. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8.
26. Luecken MD, Theis FJ. Current best practices in single-cell rna-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746.
27. Lal A, Chiang ZD, Yakovenko N, Duarte FM, Israeli J, Buenrostro JD. Deep learning-based enhancement of epigenomics data with AtacWorks. Nat Commun. 2021;12(1):1507.
28. Chu S-K, Zhao S, Shyr Yu, Liu Q. Comprehensive evaluation of noise reduction methods for single-cell rna sequencing data. Brief Bioinform. 2022;23(2):bbab565.
29. Ilin A, Raiko T. Practical approaches to principal component analysis in the presence of missing values. J Mach Learn Res. 2010;11:1957–2000.
30. Guo H, Yang Z, Jiang T, Liu S, Wang Y, Cui Z. Evaluation of classification in single cell atac-seq data with machine learning methods. BMC Bioinform. 2022;23(5):249.
31. Traag VA, Waltman L, Eck NJV. From louvain to leiden: guaranteeing well-connected communities. Sci Rep. 2019;9(1):5233.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–30.
33. Gayoso A, Lopez R, Xing G, Boyeau P, Amiri VVP, Hong J, Katherine W, Jayasuriya M, Mehlman E, Langevin M, et al. A python library for probabilistic analysis of single-cell omics data. Nat Biotechnol. 2022;40(2):163–6.
34. McInnes L, Healy J, Umap MJ.: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426, 2018.
35. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006;7:1–30.
36. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, Ginhoux F, Newell EW. Dimensionality reduction for visualizing single-cell data using umap. Nat Biotechnol. 2019;37(1):38–44.
37. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-sne and umap. Nat Biotechnol. 2021;39(2):156–7.

38. Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of umap in population genetics. J Hum Genet. 2021;66(1):85–91.
39. Tobias W, Wenjie J, Zoppi G, Vogel IA, Akhmedov M, Bleck CKE, Beltraminelli T, Rieckmann JC, Ramirez NJ, Benevento M, et al. Dynamics in protein translation sustaining t cell preparedness. Nat Immunol. 2020;21(8):927–37.
40. Stetson DB, Mohrs M, Reinhardt RL, Baron JL, Wang Z-E, Gapin L, Kronenberg M, Locksley RM. Constitutive cytokine mrnas mark natural killer (nk) and nk t cells poised for rapid effector function. J Exp Med. 2003;198(7):1069–76.
41. Salerno F, Howden AJM, Matheson LS, Gizlenci Ö, Screen M, Lingel H, Brunner-Weinzierl MC, Turner M. An integrated proteome and transcriptome of b cell maturation defines poised activation states of transitional and mature b cells. Nat Commun. 2023;14(1):5116.
42. Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006;125(2):315–26.
43. Akondy RS, Fitch M, Edupuganti S, Yang S, Kissick HT, Li KW, Youngblood BA, Abdelsamed HA, McGuire DJ, Cohen KW, et al. Origin and differentiation of human memory cd8 t cells after vaccination. Nature. 2017;552(7685):362–7.
44. Zhang JA, Mortazavi A, Williams BA, Wold BJ, Rothenberg EV. Dynamic transformations of genome-wide epigenetic marking and transcriptional control establish t cell identity. Cell. 2012;149(2):467–82.
45. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scpred: accurate supervised method for cell-type classification from single-cell rna-seq data. Genome Biol. 2019;20(1):1–17.
46. Kiselev Vladimir Y, Yiu A, Hemberg M. scmap-a tool for unsupervised projection of single cell rna-seq data. *BioRxiv*, page 150292, 2017.
47. Wang C, Sun D, Huang X, Wan C, Li Z, Han Y, Qin Q, Fan J, Qiu X, Xie Y, Meyer CA, Brown M, Tang M, Long H, Liu T, Liu XS. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. Genome Biol. 2020;21(1):198.
48. Banchereau J, Briere F, Caux C, Davoust J, Lebecque S, Liu Y-J, Pulendran B, Palucka K. Immunobiology of dendritic cells. Annu Rev Immunol. 2000;18(1):767–811.
49. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni John C, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis–a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), June 2018 .
50. Bredikhin D, Kats I, Stegle O. Muon: multimodal omics analysis framework. Genome Biol. 2022;23(1):42.
51. Friedman J. Glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

Publisher's Note