Exploiting marker genes for robust classification and characterization of single-cell chromatin accessibility

Risa Karakida Kawaguchi[1], Ziqi Tang[1], Stephan Fischer[1], Rohit Tripathy[1], Peter K. Koo[1], and Jesse Gillis[1]✉

[1]Cold Spring Harbor Laboratory, 1 Bungtown Rd, Cold Spring Harbor, New York, 11724, USA.

Background: Single-cell Assay for Transposase Accessible Chromatin using sequencing (scATAC-seq) measures genome-wide chromatin accessibility for the discovery of cell-type specific regulatory networks. ScATAC-seq combined with single-cell RNA sequencing (scRNA-seq) offers important avenues for ongoing research, such as novel cell-type specific activation of enhancer and transcription factor binding sites as well as chromatin changes specific to cell states. On the other hand, scATAC-seq data is known to be challenging to interpret due to its high number of zeros as well as the heterogeneity derived from different protocols. Because of the stochastic lack of marker gene activities, cell type identification by scATAC-seq remains difficult even at a cluster level.

Results: In this study, we exploit reference knowledge obtained from external scATAC-seq or scRNA-seq datasets to define existing cell types and uncover the genomic regions which drive cell-type specific gene regulation. To investigate the robustness of existing cell-typing methods, we collected 7 scATAC-seq datasets targeting mouse brain for a meta-analytic comparison of neuronal cell-type annotation, including a reference atlas generated by the BRAIN Initiative Cell Census Network (BICCN). By comparing the area under the receiver operating characteristics curves (AUROCs) for the three major cell types (inhibitory, excitatory, and non-neuronal cells), cell-typing performance by single markers is found to be highly variable even for known marker genes due to study-specific biases. However, the signal aggregation of a large and redundant marker gene set, optimized via multiple scRNA-seq data, achieves the highest cell-typing performances among 5 existing marker gene sets, from the individual cell to cluster level. That gene set also shows a high consistency with the cluster-specific genes from inhibitory subtypes in two well-annotated datasets, suggesting applicability to rare cell types. Next, we demonstrate a comprehensive assessment of scATAC-seq cell typing using exhaustive combinations of the marker gene sets with supervised learning methods including machine learning classifiers and joint clustering methods. Our results show that the combinations using robust marker gene sets systematically ranked at the top, not only with model based prediction using a large reference data but also with a simple summation of expression strengths across markers. To demonstrate the utility of this robust cell typing approach, we trained a deep neural network to predict chromatin accessibility in each subtype using only DNA sequence. Through model interpretation methods, we identify key motifs enriched about robust gene sets for each neuronal subtype.

Conclusions: Through the meta-analytic evaluation of scATAC-seq cell-typing methods, we develop a novel method set to exploit the BICCN reference atlas. Our study strongly supports the value of robust marker gene selection as a feature selection tool and cross-dataset comparison between scATAC-seq datasets to improve alignment of scATAC-seq to known biology. With this novel, high quality epigenetic data, genomic analysis of regulatory regions can reveal sequence motifs that drive cell type-specific regulatory programs.

Correspondence: ***JGillis@cshl.edu***

scATAC-seq | Cell typing | Marker genes | Benchmark | Meta analysis | Deep learning | Motif analysis

Background

High-throughput single-cell RNA sequencing (scRNA-seq) data has emerged as a major tool for cell type discovery and characterization in the mammalian brain (1–4). In complement, the analysis of epigenetic profiles has attracted attention because of its potential to elucidate the regulatory network underlying cell-type dependent expression differences. Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) is a primary method used to detect the epigenetic footprint of chromatin location (5). Through the insertion of barcode sequences by the Tn5 transposase, ATAC-seq can detect accessible genomic regions and infer the strength of gene activity or the location of transcriptional regulators, such as enhancers, promoters, or transcription factor binding sites. As a result, the comparison of single cell ATAC-seq (scATAC-seq) peak regions across different cell types provides information about cell-type specific regulatory networks (6–9). Due to its high throughput and feasibility, large-scale reference ATAC-seq atlases have been constructed for diverse targets such as immune or neuronal cells (10–12). ScATAC-seq for mouse brain data has been also successfully applied to identify candidate cell-type specific enhancer regions (13, 14). While most scATAC-seq studies detected only a limited number of cell types compared to scRNA-seq studies because of coverage and throughput limitations, Li, et al. recently determined 160 subtypes from scATAC-seq profiles alone, suggesting the potential of scATAC-seq for the analysis of highly heterogeneous and diverse brain cell types (15). To infer the cell-type specific regulatory network from scATAC-seq profiles, it is essential to assign each chromatin accessibility profile to a cell type, either through known marker genes or by mapping to cell types inferred from transcriptome data. For that purpose, the gene activity profile is estimated in the scATAC-seq analysis by summing the read counts around the transcription start site (TSS) or gene body of each gene to make it comparable to the transcriptome reference. However, the gene activity is known to imperfectly match to transcriptome profiles because it ignores complex regulation mechanisms by distant or condition-specific enhancers (11) and it is simultaneously affected by experimental limitations such as sparseness and binary-like signals. These latter problems arise directly from scATAC-seq's principle, which measures the insertion of barcode sequences into accessible regions on the entire genome. While bulk ATAC-seq aggregates the signals for a group of cells, scATAC-seq measures each cell independently. As each cell contains a limited number of chromosome copies (e.g., 2 for diploid organisms), the barcode insertions for each genomic locus are observed as nearly binary (0, 1, or 2) signals across entire genomic regions for scATAC-seq analysis. Alternatively, cell-type specific profiles can be obtained directly by using recombinase driver lines (1), potentially coupled with microdissection of a specific region (16, 17). However, such approaches require laborious work for each driver line and each sample measures only one cell type, resulting in an inevitable large batch effect problem. Therefore, to analyze a cell-type specific regulatory system from scATAC-seq data, we need to find a computational mapping of associating gene activity profiles of each single cell with known biological references.

Another challenge for the robust identification of cell types for epigenetic profiles is the heterogeneity of scATAC-seq datasets due to the diversity of experimental protocols. There are several protocols developed for the scATAC-seq analysis, such as single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq), droplet single-cell assay for transposase-accessible chromatin using sequencing (dscATAC-seq), or further dscATAC-seq with combinatorial indexing (dsciATAC-seq). Each dataset is affected by a study-specific batch effect due to technology, as evidenced by the difference of the distribution in terms of library quality and proportion of TSS fragments (18). Because of such differences in protocol, the performances of computational pipelines for scATAC-seq clustering have been observed be varied (19). It remains an unmet challenge to simply characterize factors driving performance.

In this study, we carried out a comprehensive benchmark of cell-type classification based on seven scATAC-seq data obtained from mouse brain with a variety of protocols. We collected marker sets from individual studies, as well as a set of robust markers inferred from multiple scRNA-seq datasets. In a broad evaluation of marker sets, learning methods, and datasets, we found that careful selection of marker genes largely drives performance; usefully, this occurs to such a degree that if an adequately strong marker set is selected, simple aggregation of the gene-specific scATAC-seq signal characterizes cell-type remarkably well. This finding provides an important basis for future data integration and downstream applications. In order to demonstrate the utility of marker-based selection, we used the pseudo-bulk scATAC-seq profiles of jointly labeled cells to train a deep convolutional neural network (CNN) to classify which cell types are accessible for an input DNA. Through model interpretation, we highlight learned motifs that are enriched about robust biomarker gene sets for each cell type, revealing a novel view of cell type-specific regulatory programs in the motor cortex.

Results

Individual marker gene activity cannot produce practical or reproducible predictions at cluster or single-cell level. Figure 1A shows a general scATAC-seq analysis workflow, in which the process of cell typing at either cluster or smaller resolution level is essential to analyze a cell-type specific regulatory network based on the knowledge about existing cell types. To infer the cell-type information for scATAC-seq profiles, the chromatin accessibility around the TSS of each gene, called gene activity, is compared with a list of known marker genes or associated with an existing reference data that is already well characterized. We consider two levels of cell typing: at the cluster level or at the individual cell level. The cell-typing process is considered to be more robust and reliable when applied to averaged profiles over clustered cells because the averaging of gene activity profiles reduces the influence of stochastic noise, the sparsity of the dataset, and the requirement of computational resources. When a cell type is inferred for each cluster, however, the resolution of cell typing is limited to the size of clusters. This matters particularly for brain scATAC-seq analyses, which intrinsically contain potentially hundreds of cell types and clusters are expected to contain several finer grained cell types.

In fact, the disparity in terms of the number of cells and clusters for the heterogeneous datasets suggests that cluster-level annotation is inadequate since it is quite likely to depend on pipeline variation which is not held constant. Table 1 is the summary of our compendium of mouse brain datasets (mainly from cortical regions). Because the datasets are relying on different experimental protocols and computational pipelines, the number of clusters is highly variable (9 to 36 clusters) and the clusters vary in their granularity, from major cell types, such as inhibitory (IN), excitatory (EX), and non-neuronal (NN) cell types (Fig. 1B), to more detailed cell types, such as Pvalb or Vip within inhibitory neuronal cells.

To evaluate the cell-typing performance beyond the difference of scATAC-seq dataset properties, we first evaluated the cell-type classification accuracy among the most well defined categories: IN, EX, and NN cell types for the seven scATAC-seq datasets. To measure consistency (and define "true" labels), we used the annotations for each cluster of each dataset provided by the authors. Since information with respect to these cell-type information is essential to know the function of a variety of neuronal cells, 6 out of 7 datasets are published or personally provided with the metadata about these cell types at cluster-level. Among the seven datasets, two datasets from Chen et al. (20) and Zhu et al. (21) are obtained by a joint profiling methodology for chromatin accessibility and transcriptome, and coupled with scRNA-seq data. The BICCN dataset is also accompanied with the scRNA-seq and snRNA-seq reference datasets (4). We selected the scRNA-seq data based on SMART-seq v4 as a transcriptome refer-

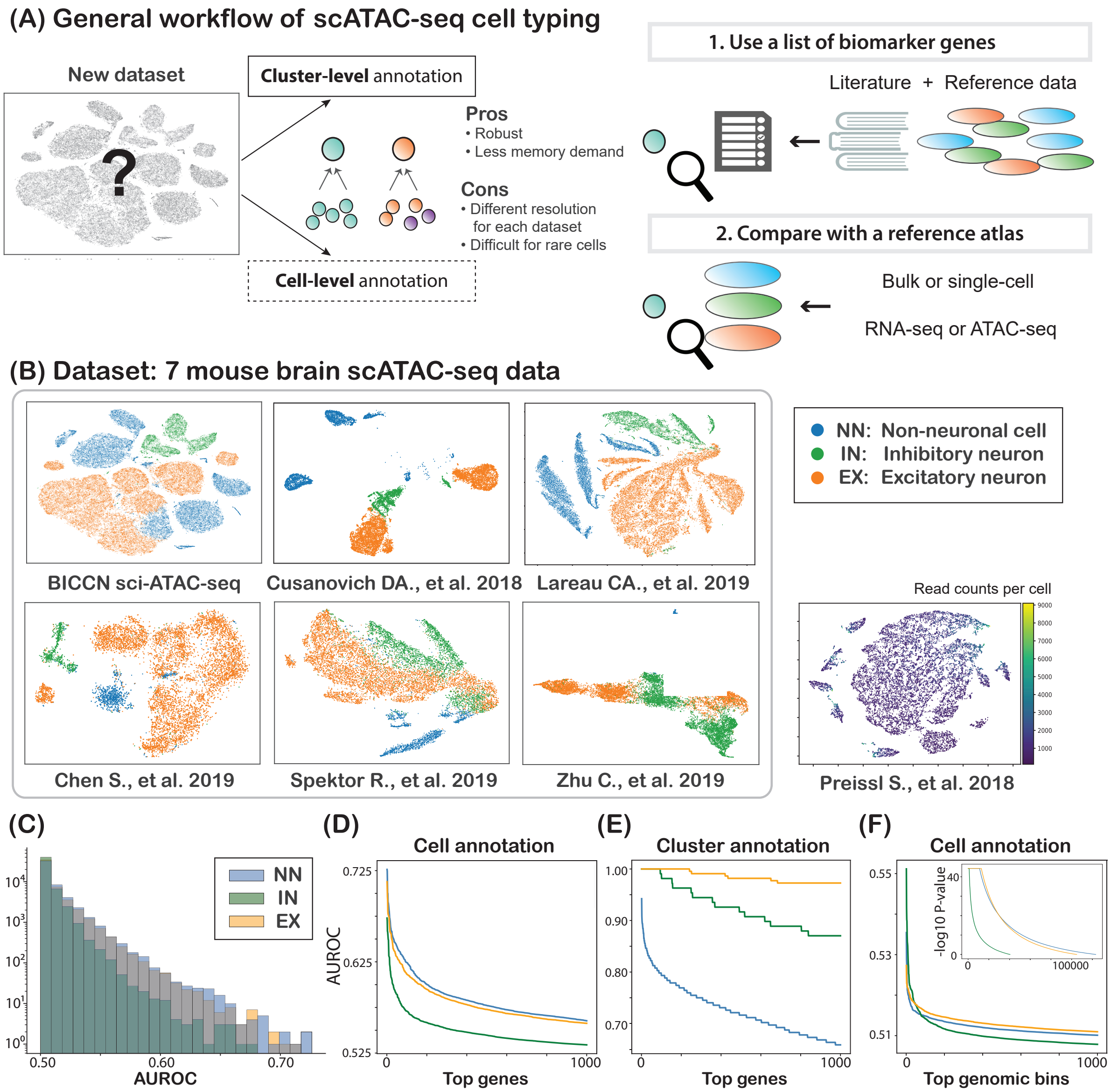


Fig. 1. A workflow of scATAC-seq analysis and the seven scATAC-seq dataset used in this study. (A) A general workflow of scATAC-seq cell typing using a reference gene list or another omics dataset. (B) tSNE mapping of 7 scATAC-seq datasets used in this study. Each cell is colored depending on the assigned cell type from three major cell types; blue for non-neuronal cell (NN), green for inhibitory neuron (IN), and orange for excitatory neuron (EX). For the dataset without a cell-type annotation from (13), the total read count of each cell is shown by a different color. (C) The AUROC distribution of major cell-type classification for the BICCN scATAC-seq dataset using the gene activity of each single gene. The AUROCs lower than 0.5 are converted into the opposite direction so that all AUROCs are no less than 0.5 by the formula $\max(1.0-\text{AUROC}, \text{AUROC})$. (D), (E), and (F) The top 1,000 features for cell-type classification of the BICCN dataset based on different features or different problems. (D) Cell-typing for each cell using a single gene activity. (E) Cell-typing for each cluster using a single gene activity from the averaged profile. (F) Cell-typing for each cell using an accessibility of each 5kb genomic bin at cell-level annotation. The inset plot shows the $-\log_{10}$(p-value) after Bonferroni multiple correction computed by Fisher's exact test from the binarized accessibility for each genomic bin.

ence atlas in this study.

Figure 1C shows the performance of major cell-type classification using the activity of each single gene for the BICCN dataset, the largest scATAC-seq atlas in our collection, at the individual cell level. The mode of the AUROC distributions ranges from 0.5 to 0.6 with a heavy tail and this pattern is consistent for all three cell types. This indicates that most genes do not substantially and consistently increase their activity in a specific cell type. In Figure 1D, we extracted the top 1,000 genes for each cell-type classification and fewer than 200 genes achieved an AUROC greater than 0.625. On the other hand, for predictions at the cluster-level, the AUROCs of top 1,000 genes are substantially higher than those for individual cell prediction (Fig. 1E). It should be noted cell-level and cluster-level classification have different sample sizes (number of cells vs number of clusters), leading to a step-wise aspect for cluster-level performance. Nevertheless, this difference in tendency is consistent with previous studies of scRNA-seq in which a few hundred genes could be obtained as reliable markers (1) while the number of genes detected in neuronal and non-neuronal cells are known to be different in thousands of genes at bulk-level because of multiple factors such as real marker genes and their co-expressed genes (22).

In addition to gene-level accessibility analysis, the read counts of each genomic bin can be also used as a feature for cell-type prediction as used in the previous studies (14). By computing the AUROCs for cell-typing based on each genomic bin activity at cell-level classification, the number of features whose AUROC is substantially higher than random (0.5) is much smaller than that based on gene activity (Fig. 1F). Since the read count data is sparser and almost binary for each genomic bin, we also computed p-values for Fisher's exact test instead of AUROC to evaluate the enrichment of the binarized read counts in each cell type. While the most of AUROCs are around 0.5, more than 6.39 % of bins had p-values smaller than 0.05 after Bonferroni correction for three major cell types (inset in Fig. 1F). Our results suggest that aggregating across either bins or cells can improve the potential of cell-type prediction accuracy by decreasing the sparsity and the noise underlying the dataset. Moreover, none of the single features predicts cell types with a satisfactory precision at the individual cell level.

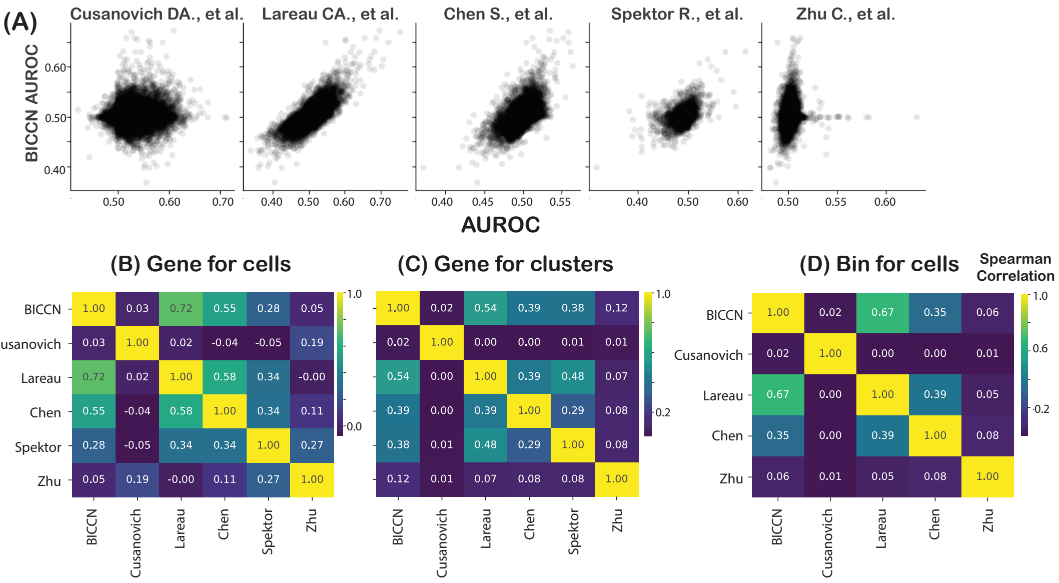
Next, by comparing the prediction performance in other datasets, we examined the reproducibility of cell-type prediction performances using AUROC and confirmed whether the tendency of cell-type classification performance is preserved for each gene beyond a study-specific batch effect. Figure 2A shows the AUROCs of each gene activity from the BICCN dataset and the 5 other author-annotated datasets for the case of IN cell-type prediction at individual cell-level. Overall, the scatter plots show a positive correlation with a large variance and specific outlier signals in some of the comparisons. Such dataset-specific dropouts potentially contribute to the study-specific batch effects.

We then computed the Spearman's correlation coefficients of AUROC scores for all pairs of datasets as a measure of reproducibility of cell-typing based on each single feature, focusing on IN cell-type classification (Fig. 2B-D). In most comparisons, we found that AUROC scores were positively correlated, with some notable exceptions, in particular for the *Cusanovich* and *Zhu* datasets. The BICCN, *Lareau*, and *Chen* dataset produced only a non-negative correlation with all other datasets, indicating the potential of a large scATAC-seq atlas. On the other hand, the *Cusanovich* and *Zhu* datasets showed a lower correlation compared to others. While we did not apply any correction for batch effects in this study, two datasets may require a special normalization to estimate gene activity enrichment. It may be possible that the mixture of different cell types are more frequent and the heterogeneity of each cluster is higher in those datasets because they contain a smaller number of clusters. In fact, the *Cusanovich* dataset was sampled from whole brain tissues and contains more non-neuronal cell types compared to other datasets. Although the AUROCs computed in this study are only for selected cell types related with brain function such as neuronal cells, microglia, or oligodendrocytes, the results of the clustering may be affected by the existence of a variety of non-neuronal cells, such as B cells or T cells.

Compared to the cell-level classification, the correlation coefficients of cluster-level classification are comparable or even lower (e.g., comparisons involving BICCN dataset), suggesting that the enrichment of high-performance features at cluster-level is only weakly reproducible across the datasets. We also tested the reproducibility of the AUROCs for each genomic bin. After selecting the bins where signals are detected in both datasets, the correlation coefficients are computed in the same way as for gene activities. We found that all pars whose correlation coefficients for gene activities were substantially positive showed lower correlation coefficients for genomic bin activities. In addition, compared to gene activity, most of the AUROCs for the classification based on genomic bins were closer to 0.5, implying that random correlation can be easily introduced if a small number of features of higher AUROCs produced irreproducible tendencies. In conclusion, the performance of cell-typing by a single feature such as gene activity or genomic bin is highly variable at both individual cell and cluster-level across scATAC-seq datasets.

A redundant marker set constructed from multiple scRNA-seq data enables robust cell typing for heterogeneous scATAC-seq datasets. While cell typing is generally performed using the expression profile of only a limited number of marker genes, performance is unstable for multiple single-cell sequencing data, as the gene activity may be stochastically missing, regardless of the importance of the gene for cell-type specific functions. This leads to the idea that a redundant marker gene set including the genes co-expressed with the marker genes would be able to capture the subtle signal from scATAC-seq datasets, effectively overcoming stochastic dropout by aggregating information over functionally related genes. To examine the efficiency of redundant marker gene sets for the meta-analytic integration of brain scATAC-seq data, we collected five marker gene sets established for single-cell sequencing data, named *SF, CU, TA, TN*, and *SC*. TA and TN are marker sets constructed in previous studies of mouse brain scRNA-seq analysis, (23) and (1), respectively. CU is defined in one of the previous scATAC-seq analyses used in this study (9). SF and SC are constructed as a robust marker gene set learned from multiple scRNA-seq data in the BICCN collection (4). SC is a subset of the SF marker set to have the same number of genes as the CU marker set to assess the importance of the size of gene sets independently of the marker gene selection process. Figure 3A shows the overlap of the various marker gene sets. We found that smaller gene sets such as CU and SC do not have any exclusive marker genes and all markers are contained in at least another marker set. Note that the CU marker set was defined based on one of our test datasets, while the construction of SF and SC marker sets did not use the information of any of our test scATAC-seq datasets.

To investigate the efficiency of individual genes included in the marker lists for the cell-type classification, we computed the AUROCs of IN cell-type prediction for the genes of each marker set. In theory, these marker genes are expected to be up-regulated and result in AUROCs higher than 0.5 (called



(B) Gene for cells

	BICCN	Cusanovich	Lareau	Chen	Spektor	Zhu
BICCN	1.00	0.03	0.72	0.55	0.28	0.05
Cusanovich	0.03	1.00	0.02	-0.04	-0.05	0.19
Lareau	0.72	0.02	1.00	0.58	0.34	-0.00
Chen	0.55	-0.04	0.58	1.00	0.34	0.11
Spektor	0.28	-0.05	0.34	0.34	1.00	0.27
Zhu	0.05	0.19	-0.00	0.11	0.27	1.00

(C) Gene for clusters

	BICCN	Cusanovich	Lareau	Chen	Spektor	Zhu
BICCN	1.00	0.02	0.54	0.39	0.38	0.12
Cusanovich	0.02	1.00	0.00	0.00	0.01	0.01
Lareau	0.54	0.00	1.00	0.39	0.48	0.07
Chen	0.39	0.00	0.39	1.00	0.29	0.08
Spektor	0.38	0.01	0.48	0.29	1.00	0.08
Zhu	0.12	0.01	0.07	0.08	0.08	1.00

(D) Bin for cells

	BICCN	Cusanovich	Lareau	Chen	Zhu
BICCN	1.00	0.02	0.67	0.35	0.06
Cusanovich	0.02	1.00	0.00	0.00	0.01
Lareau	0.67	0.00	1.00	0.39	0.05
Chen	0.35	0.00	0.39	1.00	0.08
Zhu	0.06	0.01	0.05	0.08	1.00

Fig. 2. Consistency of cell-type classification by each single feature across the datasets. (A) Scatter plots of the AUROCs of IN cell-type classification within the BICCN dataset (y-axis) and other 5 scATAC-seq datasets (x-axis). (B), (C), and (D) Spearman's correlation coefficients of the AUROCs based on each gene activity for cell-level annotation (B), averaged gene activity for each cluster (C), and signal activity for each 100kb genomic bin (D).

positive, hereafter) within a corresponding cell-type sample while marker genes for other cell types are likely to be downregulated and tend to show a lower AUROC (negative). As shown in Figure 3B, the AUROCs of genes are substantially higher than 0.5 except for several negative markers showing an AUROC around 0.5. These distributions are drastically different from those of all genes as they are highly skewed at 0.5. To evaluate the reproducibility of the AUROCs for marker genes and others, we computed the average AUROCs across the datasets for each cell-type classification, along with the minimum and maximum AUROC (Fig. 3C). For all cell-type predictions, a long plateau is observed at the average AUROC 0.5 with a large variance across the datasets, suggesting that most of the genes are not repeatedly observed to be either positive or negative for a specific cell type in scATAC-seq data. By focusing on the marker genes, however, two distinctive groups appear in the average AUROC rankings; one is ranked at the top and another is at the bottom in the average AUROC ranking. Although all genes are expected to be positive markers and have an AUROC higher than 0.5, the distribution of the bottom group ranges across 0.5, indicating that these genes effectively act as either a positive or a negative marker depending on the dataset. This result suggests that this group suffers from study-specific batch effects or weak and ambiguous signals, which stochastically leads to a negative prediction performance. Moreover, we examined the significance of the AUROC consistency by comparing the BICCN and *Lareau* datasets, whose correlation coefficient for the entire gene set is highest (0.72) among all pairs of the datasets (Fig. 2B). To evaluate the consistency of marker gene activities, the gene sets for the three cell types are aggregated to one set so that the positive and negative marker genes should be included. We then computed Spearman's correlation coefficients of the AUROCs for each marker set between the BICCN and *Lareau* datasets. Although all marker sets showed a higher correlation coefficient compared to the set of all genes (SF: 0.931, CU: 0.921, TA: 0.779, TN: 0.851, and SC: 0.880), each marker set consists of a different number of genes. Therefore, we also computed a p-value for each marker set by comparing the correlation coefficient with correlation coefficients obtained for 10,000 randomly sampled gene sets of the same size. As a result, only the p-value of SF (p-value<5e-5) is significantly lower after multiple correction ($n = 5$), indicating that the correlation of SF marker gene set is significantly higher than a random gene selection across the datasets. In conclusion, by focusing on marker genes, we can greatly increase the reproducibility of cell-type classification in scATAC-seq data.

Measuring the enrichment of marker gene activity enables a robust and practical cell-type classification. To further improve cell typing, we now consider the integration of information across multiple-gene activities, such as marker gene sets. To address this problem in a practical workflow of scATAC-seq analysis, we evaluated the performance of cluster-level annotation based on two cell-typing strategies. The first and qualitative way utilizes a cluster-specific gene list obtained by comparing each cluster with all other clusters. The second and quantitative way aggregates the signals of gene activity.

To carry out a cell-type classification using a list of cluster-specific genes, we computed the Jaccard index for each marker set with cluster-specific genes obtained by Wilcoxon's rank sum test using a Scanpy library (24). For each dataset, we scaled the Jaccard scores into the range of $[0, 1]$ for each cell type first, and within each cluster across three cell types. Using the vector of normalized Jaccard scores, we computed AUROCs for each cell type classification against the reference true cell-type labels for the clusters. In Figure 4A, the AUROCs of each marker set are shown as a

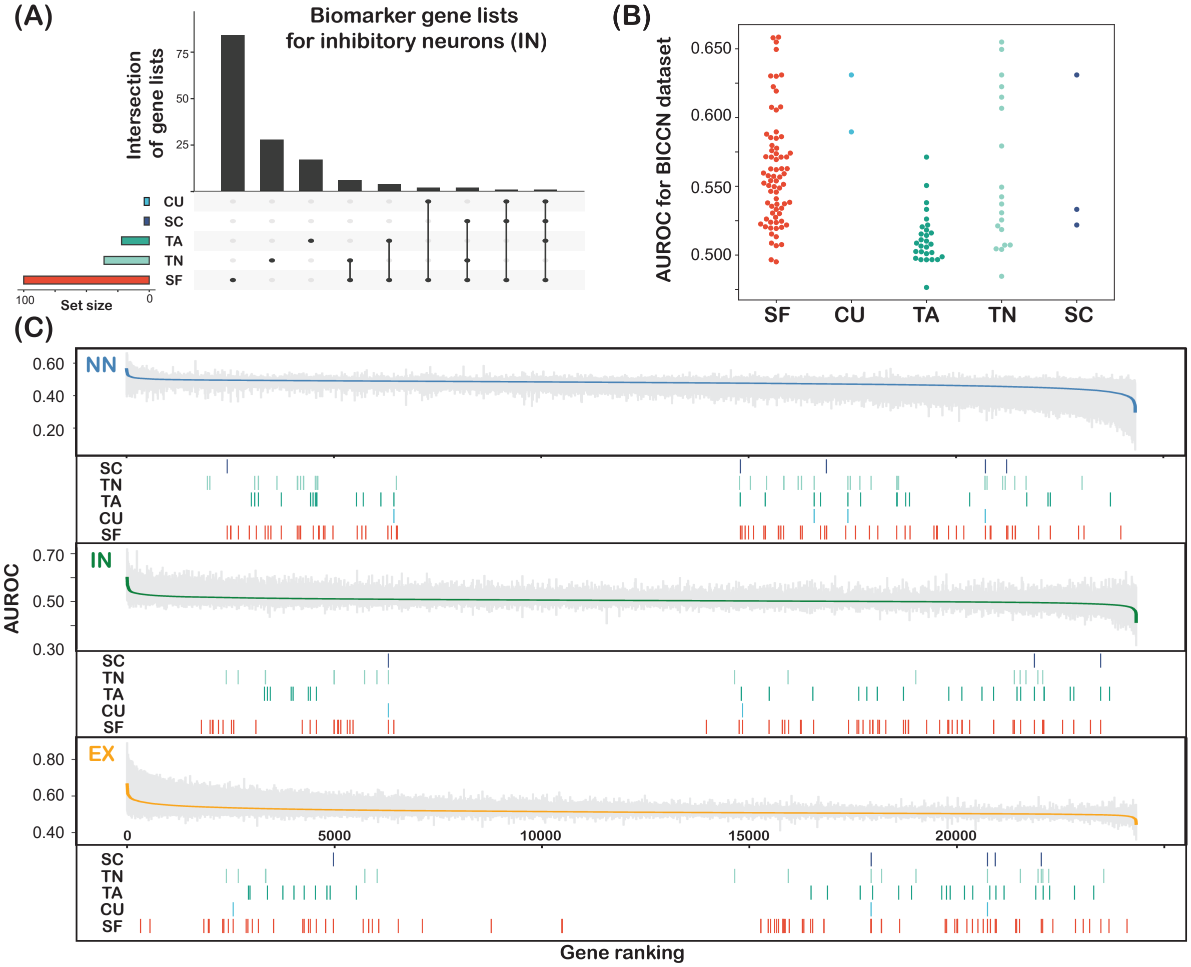


Fig. 3. Performance and reproducibility of marker gene activity for cell-type classification. (A) The size and overlap of 5 marker gene lists for the IN cell type used in this study. The biomarker set determined in (9) (referred to as CU) is used as a scATAC-seq oriented biomarker set while those from (23) and (1) (TA and TN) are selected as the representatives of scRNA-seq oriented biomarker sets. Additionally, the gene sets SF and SC are newly constructed using six scRNA-seq datasets obtained in a BICCN project detailed in (4). The SF set contains top 100 genes that were robust for scRNA-seq cell typing as a biomarker for each cell type while the SC marker only contains the top genes as much as each CU set. (B) The distribution of AUROCs of the IN cell-type classification within the BICCN dataset using the gene activity of 5 marker gene sets. (C) The distribution of mean, min, and max AUROCs for three cell-type classification using the estimated gene activity across the six scATAC-seq datasets with the information of marker gene annotation. The solid line shows the average of the AUROCs for the six datasets while the top and bottom of gray lines correspond to the maximum and minimum of AUROCs for each gene. The top, middle, and bottom panel indicate the AUROCs for NN, IN, and EX cell-type classification. In the bottom rectangle of each panel, short vertical bars are shown at the location of genes listed in the marker set for each cell type.

function of the number of top cluster-specific genes selected (shown in the x-axis). The larger the number of cluster-specific genes is, the higher the AUROCs of prediction are for all marker sets. However, the classification based on the SF marker set shows a sudden increase of the AUROCs to 0.8-0.9 only with around top 100 cluster-specific genes. Interestingly, most of the marker sets except for TA reached around 0.8 for the classification of IN cell type. This result suggests that even a few genes are enough to accurately annotate the IN cell-type group while additional genes can further improve the accuracy of cell typing. We also compared the AUROCs with those based on marker sets inferred from each dataset (Supplementary Fig. 2). Although the dataset-derived gene sets produced higher performances in some settings, the SF marker set produced higher AUROCs compared to most of the gene sets and is observed to show the most stable applicability independent from the cell-type difference.

Although cell-type annotation using a cluster-specific gene list is a simple and effective approach which is widely used for single-cell datasets, it is possible that important marker genes cannot be properly detected as top cluster-specific genes if cells from the same cell type are distributed into multiple clusters. To avoid such failure, we also made quantitative cell-type predictions by computing the enrichment of marker gene activities by averaging the gene activities for each marker set, then summarized performance as AUROCs (Fig. 4B). For cell-level annotations, we computed AUROCs from individual cell profiles, while for cluster-level annotations we used the average profiles of the cells that belong to the same cluster. As a result, we expect the performance of cell-level annotation to be generally lower than that of cluster-level annotation. While there are only small differences in AUROCs for the cluster-level classification, the SF marker set outperformed at individual cell-level classification with a median AUROC around 0.85. Moreover, to show the robustness of marker set performance without a class imbal-

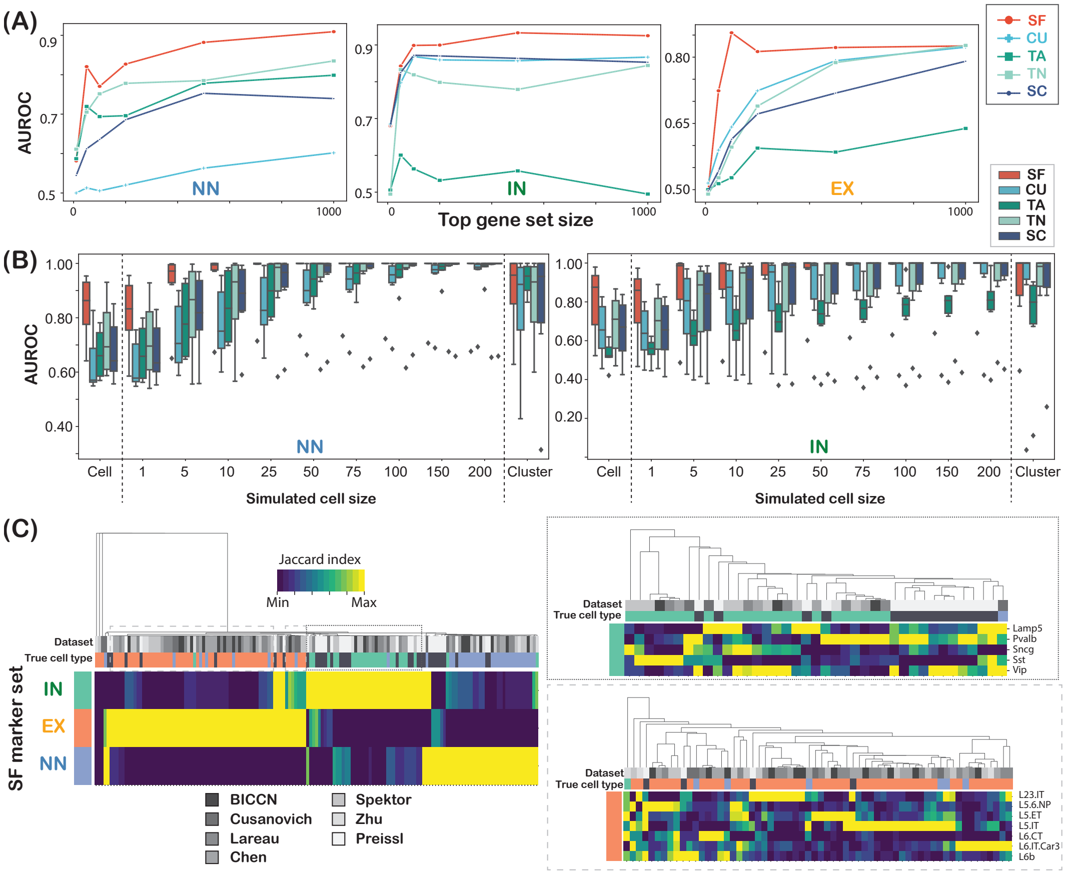


Fig. 4. Performance of cluster-level annotation based on qualitative and quantitative cell-typing strategies. (A) AUROCs of major cell-type classification for six scATAC-seq datasets at cluster-level. For each cluster, a top group of cluster-specific genes is selected based on Wilcoxon's rank sum test within each dataset (shown in x-axis). Then a Jaccard index is computed for three cell types by calculating the intersection of the top cluster-specific genes and each biomarker set and normalized into the range (0, 1) among the three cell types. Finally, the normalized Jaccard indices obtained from all annotated clusters are used as a feature vector to compute AUROCs for NN, IN, and EX cell types (corresponding to the left, center, and right panel, respectively). (B) Boxplots of AUROCs of NN (shown at left) and IN (right) cell-type classification within each dataset for the 6 datasets using an aggregated gene-activity signal of the genes listed in each biomarker set. In each panel, the left-most and right-most boxplots represent the raw prediction accuracy of cell-level and cluster-level annotation while other boxplots showed the results of down-sampling for 300 samples, in which 100 average gene activities of each cell type are obtained for randomly sampled cells of the size shown in x-axis (also see the Method section). The average AUROCs of 100 trials are shown for down-sampled simulation data with replacement as a representative performance. (C) The heatmaps of normalized Jaccard indices for SF marker sets and all clusters from seven scATAC-seq datasets. The left panel shows the scores for the SF sets of major three cell types. Two panels shown at right represent the scores for the subtypes of inhibitory neurons (top) and excitatory neurons (bottom).

ance problem, we constructed simulation data of 100 average profiles for each major cell type over a specific number of cells randomly sampled from the original datasets. To reduce random effects, the average results of cell-type classification over 100 trials were shown for each dataset. In Figure 4B, the AUROCs are shown to gradually improve with the number of cells used to construct the average profiles. Along with the increase of the AUROCs, the order of marker sets is almost consistent and the SF marker set is found to show the most stable prediction accuracy for both cell-type classification. In summary, we found that a redundant and robust marker gene set constructed in a meta-analytic way could improve the robustness of the major cell-type classification at a variety of cluster levels.

In addition to the three major cell types, we can use the SF marker sets to perform rare cell- or sub-type classification. As shown in Figure 4A, the up-regulation of normalized Jaccard scores are consistently associated with the correct cell type annotation. For the prediction of the subtypes of inhibitory and excitatory neurons, we first extracted two groups with the higher Jaccard scores for either of IN or EX marker sets (Fig. 4C). Then, we computed the number of overlapped genes again between each SF subtype marker set and cluster-specific genes. While each subtype marker set appears to be exclusively enriched in some part of the clusters, we could not validate the predictions in all datasets since some lack annotation at this level of specificity (highlighting the utility of classifiers that can be applied uniformly to these data). Thus, the cell-typing performance of several inhibitory subtypes (e.g., Sst, Pvalb, Sncg, Vip, and Lamp5) were validated for the BICCN and *Chen* datasets, which both contain the clusters associated with those subtypes. In the BICCN dataset, the AUROCs of Sst and Pvalb subtypes are 1.0 and these clusters are considered to be distinctive within the BICCN dataset by comparinig with SF subtype marker sets. On the other hand, the AUROCs for Sncg, Vip, and Lamp5 result in relatively low AUROCs (0.454, 0.722, and 0.685 respectively) because there was only one BICCN cluster corresponding to the three inhibitory cell subtypes of Sncg, Vip, and Lamp5 at the resolution we used. In the *Chen* dataset, the AUROCs of

Sst, Pvalb, and Vip of inhibitory subtypes are 0.9545, 1.0, and 0.97727 using SF subtype marker sets. Since the SF marker sets were constructed independently from the scATAC-seq datasets, the use of subtype marker sets is a promising to enable robust cell typing even for neuronal subtypes at the cluster-level. Furthermore, by carefully examining the consistency of the signals in Figure 4C, some clusters with the cell-type labels are observed to show enrichment for multiple marker sets. This suggests the heterogeneity of those clusters and our cell typing at individual cell level would be applicable to detect such rare cell populations. Finally, the clusters of the *Preissl* dataset, whose "true" labels are not available in this study, also show an exclusive signal enrichment for the SF major cell- and sub-type marker sets. This, too, indicates their applicability to labelling unknown clusters.

Comprehensive assessment of supervised cell-type classification and marker sets reveals the efficiency of robust marker gene sets and consensus prediction across multiple datasets. To determine the degree to which robust markers facilitate cell type annotation when combined with more sophisticated prediction methods, we performed a comprehensive assessment of scATAC-seq cell-type classification at the individual cell level. This assessment was to address the question whether the suitable feature selection based on marker genes is still critical, past differences in the datasets, training dataset, or prediction methods. We applied a variety of supervised learning methods for scATAC-seq, such as raw signal aggregation (as used in the previous section), machine learning (ML) classifiers, and joint clustering methods. Importantly, raw signal aggregation of the marker set is the only method that does not require a reference dataset as training data. On the one hand, this provides scope for methods with more parameter optimization to improve performance; on the other hand, this may reduce robustness. As ML classifiers, we applied four different classifiers (e.g., Logistic regression, support vector machine(SVM)) and trained them using the BICCN scRNA-seq data (*RNA atlas*) or other scATAC-seq datasets (*Consensus*). As joint clustering methods developed for the integration of scRNA-seq (and scATAC-seq), BBKNN (25) and Seurat (26) were selected. Further details on optimization and evaluation are described in Method section (also see Supplementary Fig. 3).

Figure 5A shows the summary of the AUROCs for NN cell-type classification at individual cell level. The prediction performance highly depends on the dataset quality or similarity: when the quality is low, no single method or training condition seemed to work at a practical level. In Figure 5B, the top 10 combinations in terms of average AUROCs are extracted. Most of the combinations are based on ML optimized on the Consensus although also included are some combinations trained on RNA atlas. The two best methods are the Logistic regression classifier trained on all genes, and Alternate Lasso trained on SF marker genes. With respect to the marker sets selected, SF and TN gene sets are dominant within the top 10 combination, suggesting the utility of larger marker sets as a feature selection method against the dataset-dependent variability.

Next, the prediction performances for IN cell-type classification are visualized (Fig. 5C). Unlike the result of NN cell-type classification, the AUROCs of the top and bottom combinations are clearly distinct due to differences in training datasets. The combinations based on Consensus training or raw signal aggregation show apparently higher AUROCs than those using RNA atlas. As previously, the top 10 combinations exploit combinations involving the SF and TN marker sets as well as all genes with ML classifiers. Indeed, even simple raw signal aggregation method from the SF marker gene sets is ranked as the second best method, broadly in the range defined by the best-performing methods (Fig. 5D). Additionally, we examined the classification performance using a transcriptome-based reference from the same sample (named *RNA* training) for IN cell-type classification. By applying two joint profiling data as a test dataset, the potential of the scRNA-seq reference can be characterized in more detail. Figure 5E shows the top 10 combinations that performed best joint profiling datasets. The top 5 ranks are occupied by methods of ML classifiers optimized by Consensus training data, in addition to raw signal aggregation for the SF marker gene set.

In summary, our comprehensive assessment strongly suggests that consensus training using other scATAC-seq data and simple aggregation of large marker sets are comparably powerful for major cell-type classification. Although optimization based on the reference scRNA-seq was less powerful for the classification of neuronal cells, training on the joint-profiled scRNA-seq shows a comparable prediction performance. More importantly, in all cases, the choice of marker genes most strong characterized the performance of a method/data/feature combination, suggesting the wide-applicability of robust marker gene sets for integrative analyses and interpretation of the resultant cell-type specific ATAC-seq profiles for regulatory inference, as described next.

Robust cell annotations enhance the specificity of motif analysis for rare cell population. To investigate the motifs associated with *cis*-regulation of each cell-type, we performed a deep CNN analysis on the BICCN scATAC-seq data and then interpreted the model to identify motifs enriched near robust biomarker gene sets for each cell-type. Specifically, we generated a dataset that consists of cell-type specific pseudo-bulk profiles by aggregating the scATAC-seq signals for each cell-type. This bolsters the statistics often lacking in individual cells, but maintains the same accessible chromatin sites required for good cell type-specific inference. The pseudo-bulk profiles for each cell-type was used to generate a dataset that consists of 5kb DNA with a corresponding label that specifies whether the DNA is accessible or not for cell-types identified at the finest cluster-level in the BICCN dataset (see Methods). We constructed a custom CNN with a Basset-like architecture (27), consisting of 3 convolutional layers followed by a fully-connected hidden layer, and trained it to take DNA as input and simultaneously predict chromatin accessibility across each cell-type. We

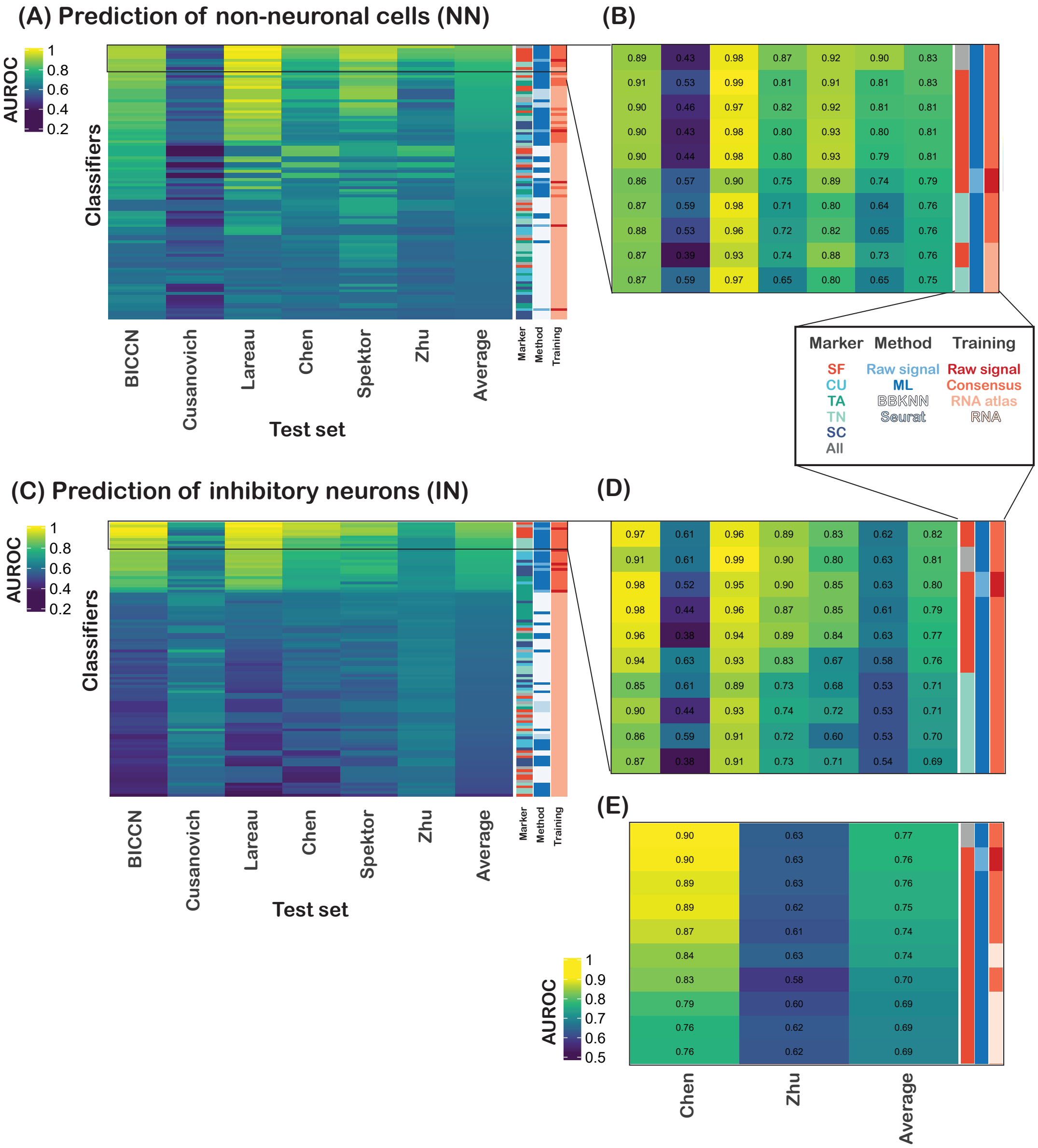


Fig. 5. Comprehensive comparison of cell-type classification at individual cell level using a combination of marker genes and supervised learning methods. (A) and (B) AUROCs of six datasets and those average for NN cell-type classification at each cell level for all demonstrated combinations (A) and top 10 combinations according to the average AUROCs (B). The columns shown at right represent the marker set, supervised learning method, and training dataset used to construct each classifier. (C) and (D) AUROCs of 6 datasets and those average for IN cell-type classification at each cell level for all demonstrated combinations (D) and top 10 combinations according to the average AUROCs (E). (E) AUROCs of two joint profiling datasets and those average for IN cell-type classification at each cell level for top 10 combinations according to the average AUROCs. This result contains the combinations using available scRNA-seq data of each dataset as a training data.

found that the CNN's classification performance on test data (i.e. all data from held-out chromosomes: 1, 3, and 5) had good predictive power with an area under the precision-recall curve (AUPR) of 0.539 on average across each cell-type – this is a significant improvement upon DeepSea's AUPR of 0.444 across 125 chromatin accessibility datasets (28), most of which derive from cell lines.

For model interpretability, we performed filter visualization and attribution methods, both of which are common techniques in genomics (29). To identify statistically significant matches to known motifs, we compared filter representations against the 2020 JASPAR vertebrates database (30) using Tomtom, a motif comparison search tool (31). We found that 36% of the filters match known motifs, which is seemingly a low number considering that when applying a similar network to the Basset dataset, our CNN yields a higher match fraction of about 62% (32). Since the Basset dataset consists of 161 chromatin accessibility datasets, which are mainly from well-studied cell lines and tissues, one possibility is that some of the motifs learned by the CNN are not found in the JASPAR database, which are comprehensive but not complete.

Attribution methods reveal the importance of each nucleotide in a given sequence on model predictions. Using an attribution method called saliency analysis (33), which takes the the gradient of a given class prediction with respect to the in-

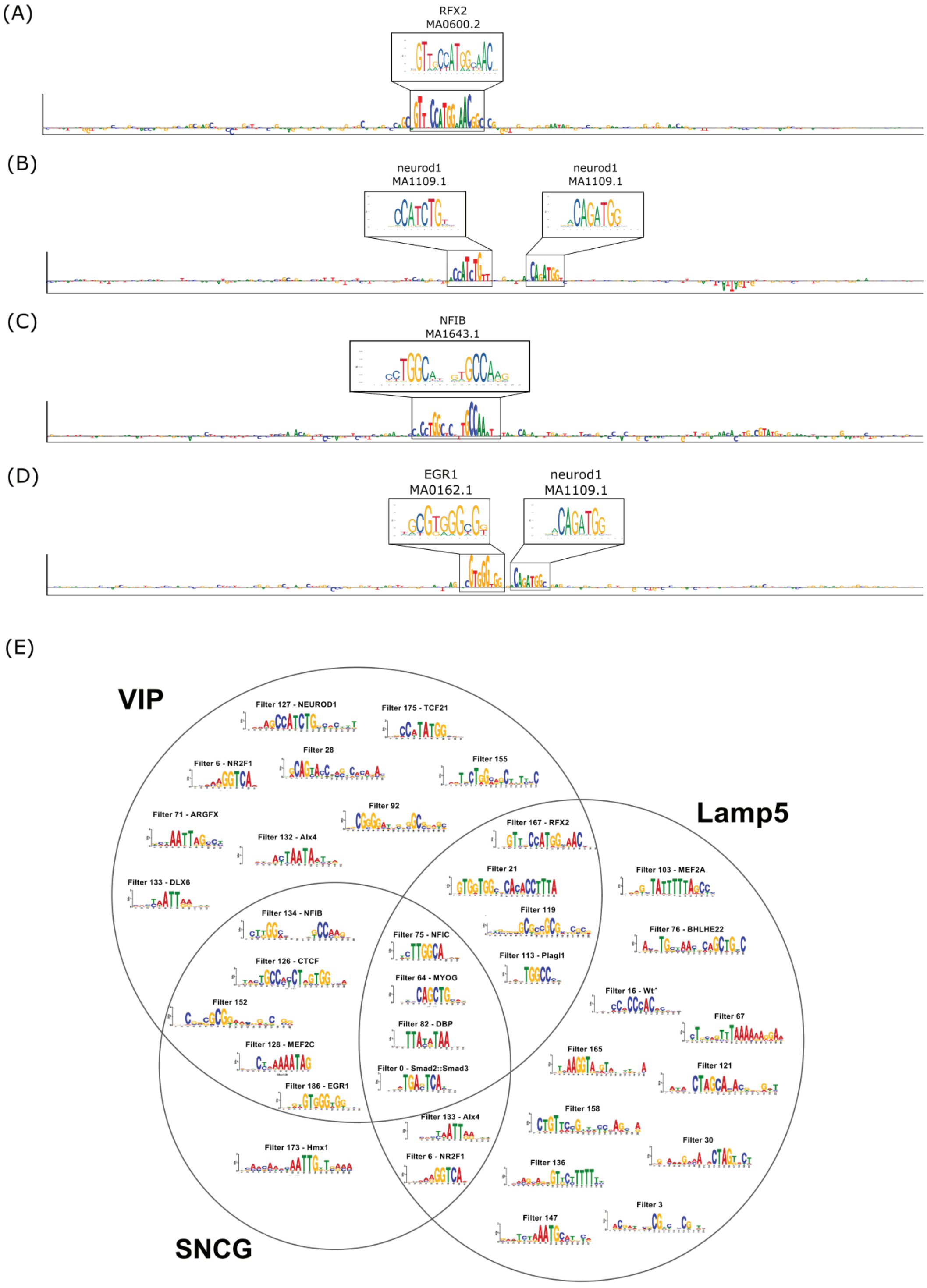


Fig. 6. Motif analysis of BICCN scATAC-seq pseudo-profiles. (A-E) Sequence logos of saliency maps from a CNN model trained at the sub-cell type level: (A) Lamp5, (B) Vip, (C) Sncg, and (D) Vip. Only 210 positions out of 5kb is shown for visual clarity. The known motifs from the JASPAR database are annotated with a box above each saliency plot, labelled with a putative motif name and JASPAR ID. (E) Venn diagram of motifs enriched in each cell type. The filter representations are shown, while the cell type-specific motif enrichment was determined with TF-MoDISCo and the motif annotations were given by statistically significant matches to the JASPAR database using Tomtom.

puts, we can generate sequence logos of the importance of each nucleotide in a given sequence (see Methods). Within the 5kb binned sequences, we often find that small patches within the attribution maps highlight known motifs either alone (Fig. 6A), in combinations with their reverse compliments (Fig. 6B-C), and with other partners (Fig. 6D). Attribution methods can footprint learned motifs that are important for model predictions at a single-nucleotide resolution. However, they have to be observed on an individual basis, which requires manually curating the recurring patterns genome-wide. To aggregate enriched patterns within the attribution maps, specifically nearby marker gene sets, we used TF-MoDISCo (34), a clustering tool that splits attribution maps into smaller segments called seqlets, clusters these seqlets, and provides an averaged representation for each cluster. We performed a separate TF-MoDISCo analysis using default parameters for each cell type. The TF-MoDISCo representations cannot be directly compared to the motifs in the JASPAR database, so we manually compared the TF-MoDISCo representations with the filter representations to link them to known motifs. The high correspondence of TF-MoDISCo-derived motifs to the convolutional filters provided further validation that our CNN has learned robust motif representations.

Figure 6E highlights a Venn diagram of the motifs enriched in different cell types: Lamp5, Vip, and Sncg. There are many motif representations that were shared between all 3 cell types, including NFIC, MYOG, DBP, and FOSL1. Vip and Lamp5 had many unique motifs enriched near marker genes, while Sncg only had a single enriched motif identified by TF-MoDISCo. This is consistent with the strong overlaps among these cell-types within the observed transcriptional hierarchy, where there is mixing across types when defined purely by expression clusters (35). We note that filter representations have many hits to known motifs and there are many proteins that bind to similar binding sites. The names of the motifs that are used in Figure 6E represent the best matching motif hits. A full list of the motif matches for each filter is provided as Supplementary data 5. We also found that many TF-MoDISCo cluster representations, which were also supported by convolutional filter representations, do not have any correspondence to a known motif in the JASPAR database – these were labelled by just their filter name. This was expected to an extent as the JASPAR database is not complete and the ability to analyze cell type-specific regulatory regions with in the brain emerged recently with the advent of scATAC-seq data.

Discussion

ScRNA-seq has proven to be a remarkably effective technology for the characterization of cell-types within the brain, shedding new light on decades old questions regarding the form, function, and organization of cell-types (36). In turn, the complexity of the brain has made it one of the strongest use cases for single cell technologies. While typing and characterization have been major success stories, understanding the regulatory basis of the observed cell-types remains an important challenge (6). Epigenetic profiling technologies, such as scATAC-seq present a likely route forward, but important questions remain. Our work offers answers to some of these questions through a wide-ranging meta-analytic evaluation of scATAC-seq data and a careful demonstration of the practical value of marker gene feature selection, finally yielding cell-type specific motif representations for cell-types “purified” from within heterogeneous scATAC-seq data.

One of our important contributions is to demonstrate just how effective marker gene selection can be. This has long been a mainstay of wet-lab biology but typically focused on specificity, rather than comprehensiveness (37). In contrast, high-throughput single-cell methods have generally preferred methods that rely on information distributed across a large fraction of genes. Our analysis suggests that a middle ground of picking redundant marker sets satisfies a number of important constraints: high performance, simple generalization, and straightforward interpretability. While we see dramatic differences from dataset to dataset, feature selection appears to be the critical determinant for accurate cell-typing, as opposed to more complicated modeling of the way those features interact (which is less likely to generalize). Because marker sets can be derived from high performing scRNA-seq data, we exploit all the existing success of cell-typing efforts there to inform the interpretation of scATAC-seq data. Importantly, the utility of feature selection for consistent annotation is likely to remain even as wet-lab technology improvements (such as paired scRNA-seq and scATAC-seq) will make clustering cells within a given dataset less challenging. The importance of marker set selection is also highlighted by the improved interpretability it offered when we turned to modeling the cell-type specific regulatory programs through deep learning.

To predict chromatin accessibility across different cell-types from just the DNA sequence, CNNs have demonstrated a remarkable ability (27, 28, 38). ScATAC-seq provides an opportunity to study cell-type specific regulatory programs in heterogeneous tissues, such as the immune cells (39) and the brain (this study), using CNNs. By “purifying” scATAC-seq data using robust cell typing, the accessibility signal becomes more reliable. This may explain why our CNN yields improved performance both in classification accuracy and interpretable motif representations. Since there are many accessible sites that are shared across different cell-types, these “overlapping” regions may not necessarily contain the information we desire, that is to know which motifs drive cell-type specific regulation. Hence, it remains a challenge to decipher which motifs are relevant for cell-type specificity. Our approach was to explore the enrichment of motifs nearby robust marker gene sets that are cell-type specific, which shifts the distribution of transcription factors that are learned genome-wide to the ones that regulate genes of a given cell-type. Indeed this approach reveals many known motifs (and some putative novel ones). Moving forward, it would be beneficial to follow up this work to try to decipher which proteins bind to these motifs in each cell-type and explore which other genes they regulate (that are not in the robust gene set).

Conclusions

In this study, we examined the usability of biomarker gene sets for scATAC-seq cell-typing at a variety of levels of granularity. We found that a redundant and robust marker gene set produced high performance at resolutions from cell to cluster level. Moreover, our comprehensive assessment of mouse brain scATAC-seq data revealed that careful feature selection via marker gene sets could improve neuronal and non-neuronal cell-type prediction when incorporated with more sophisticated supervised learning methods. We also demonstrated the potential power of this approach once heterogenous data has been partitioned into "cleaned" pseudo-bulk profiles. The resultant cell-type specific pseudo-bulk profiles can be used to train a CNN model to learn a relationship between input DNA sequence and its accessibility for a given cell-type – we showed this with the BICCN dataset. Interpreting the trained CNN revealed learned motifs that were enriched near the biomarker gene sets in a subtype-specific manner, suggesting the existence of cell-type specific regulation in the motor cortex. The straightforward feasibility of using robust biomarker gene sets for accurate subtype cell-typing within scATAC-seq data opens up many important downstream possibilities, most clearly condition- and subtype-specific regulatory network discovery, as demonstrated in our own deep learning analysis.

Methods

Mouse brain scATAC-seq datasets. From the BICCN collection, we obtained sci-ATAC-seq data which consists of 4 batches with a transcriptome reference of SMART-seq v4 scRNA-seq data (the cell number after filtering is 6,278) for mouse primary motor cortex region (4). Both datasets are available from the BICCN data portal `https://biccn.org`. Moreover, we collected scATAC-seq datasets of mouse brain published on Gene Expression Omnibus (GEO). Specifically, read count matrices and metadata of 6 scATAC-seq studies were downloaded from GEO. The corresponding GEO ids of the collected studies are GSE100033 (13), GSE111586 (9), GSE123576 (18), GSE127257 (40), GSE126074 (20), and GSE130399 (21). From the Paired-seq datasets of GSE130399, the one for an adult mouse cerebral cortex sample is only applied in this study. To convert read counts to gene activities, we used the gene structure information from an Ensembel GTF file for GRCm38 as of Nov. 2018. Each genomic feature in the original study was then assigned to the closest TSS found in the GTF file. A gene activity estimation was carried out by summing the read counts of all assigned features within the 10kb upstream or downstream from the TSS of all transcripts of the same gene id. For the datasets whose feature is peak-based, the locations of each peak center were used to associate each feature and gene. A general pre-processing, filtering, clustering, and detection of cluster-specific genes was performed on a SCANPY platform (24).

Marker gene set. We collected the existing marker gene sets established for single-cell sequencing data and additionally constructed a new robust marker set using multiple scRNA-seq data from the BICCN for sparse scATAC-seq data. TA and TN are the marker sets constructed in the previous studies of mouse brain scRNA-seq analysis, (23) and (1), respectively. CU is defined in one of the previous scATAC-seq analyses (9). Note that this dataset is also used in this study and this may give the advantage for the CU marker set during the performance assessment. SF and SC are constructed as a robust marker gene set learned from multiple scRNA-seq data in the BICCN collection (see also (4)). The SF marker set is made to select top 100 genes which are optimized to predict cell types accurately based on the training datasets of multiple scRNA-seq datasets including a reference scRNA-seq used in this study. On the other hand, SC is a subset of SF but limited to have the same number of genes with that of CU to assess the importance of the number genes, not the way of marker gene selection. The overlap of each marker gene is shown in Supplementary Figure 1.

Assessment of cell-typing for scATAC-seq. We performed a comprehensive assessment of cell-typing for six well-annotated scATAC-seq datasets using a different combination of supervised learning method, training set, and marker gene set. A graphical outline is shown in Supplementary Figure 3.

Supervised learning methods. The methods used in this study are classified into three categories: *raw expression*, *ML classifiers*, and *joint clustering methods*. Raw expression methods predict each cell type based on the Raw expression scores computed by summing the read counts for the genes included in each biomarker gene set. This method is the only method that does not require any training dataset except for a marker gene set. ML classifier methods consist of four popular ML classifiers applicable to a supervised learning of scATAC-seq cell-typing. Specifically, SVM, random forest, logistic regression with L1 regularization (Logistic regression), and a variant of logistic LASSO "Alternate Logistic LASSO" (41) are included in this category, which is expected to be more robust for sparse data. Due to the limitation of computational resources, only Logistic regression was carried out for the prediction using all genes and other three classifiers were applied with the feature selection based on the marker gene set. The last category is a joint clustering method, in which a test and training dataset is reanalyzed independently, then jointly clustered two datasets to associate each cell in the test set with the annotated cells in the training dataset. We chose Seurat (26) and BBKNN (25) for a comparison referring the results of the previous study of an integrated analysis of single-cell atlases (42). To compute AUROCs from the results of BBKNN, we implemented own script to compute the scores for a cell-type prediction by counting the nearest-neighbor cells for each cell type.

Training set. The training set is applied in four different ways. Raw expression methods use gene activity profiles from the test scATAC-seq dataset only for selected biomarker genes. Consensus methods use the scATAC-seq datasets except for the one used as a test set. For this prediction, the prediction scores are computed by the classifiers trained on each training set. Those scores are averaged to compute the final prediction scores after normalization within each dataset. RNA atlas methods use the BICCN scRNA-seq data as a training set to optimize the parameters or infer the nearest-neighbor cells. For the datasets based on joint profiling methods, we also carried out an ML-based supervised learning using scRNA-seq from the same dataset, named "RNA" training.

Gene set selection. In addition to the five marker gene sets collected from the previous studies, we performed the supervised learning with all detected genes if an optimization process is feasible. Specifically, supervised learning based on all genes were demonstrated for Logistic regression from ML classifiers and both joint clustering methods.

Supervised learning by machine learning classifiers.

To carry out supervised learning, we implemented a workflow of optimization of ML classifiers using a scikit-learn library. The parameters used for each classifier are as follows: degree is 3 and kernel is set to an rbf kernel for SVM, n_estimators is set to 100 for RF, and C is set to 1.0 (default) for Logistic regression. Other parameters are set to the default values. For Alternate Lasso, we implemented an original classification function in which the best and alternate predictors are averaged with different weights. In this study, we extracted top n predictors at maximum and summed their predictions with the weight $1/n$, where n is set to 5.

Joint clustering methods for an integrative analysis of single-cell omics datasets.

BBKNN was applied to the pair of scATAC-seq and the BICCN scRNA-seq dataset after applying a general normalization for the scRNA-seq dataset by a Scanpy function “normalize_per_cell”. The parameter for k-nearest neighbor used in the BBKNN algorithm was set to $k = 5, 10, 20, 30$ with and without a graph trimming option. We also run Seurat v3.2.2 to align the same dataset combinations as used for BBKNN. The alignment of two datasets was done via FindTransferAnchors and TransferData functions using canonical correlation analysis by setting a parameter reduction=‘cca’.

Sequence analysis of accessible sites with deep learning.

Data for binary classification. Given the 5kb bins, each of which contain a single bulk accessibility profile value, we binarized each label with a threshold of 0.02, above which is given a positive label and below is given a negative label. We then filtered sequences with no positive labels across all classes. Sequences were converted to a one-hot representation with 4 channels (one for each nucleotide: A, C, G, T) and a corresponding label vector with either a 0 for negative labels or 1 for positive labels. We split the data into a validation set (chromosomes 7 and 9; $N = 24,908$), test set (chromosomes 1, 3, and 5; $N = 14,670$), and training set (all other chromosomes; $N = 274,689$).

Model. Our CNN model takes as input a 1-dimensional one-hot-encoded sequence with 4 channels, then processes the sequence with three convolutional layers, a fully-connected hidden layer, and a fully-connected output layer that have sigmoid activations for binary predictions. Each convolutional layer consists of a 1D cross-correlation operation followed by batch normalization (43), and a non-linear activation function. The first layer used an exponential activation, which was previously found to encourage first layer filters to learn interpretable motif representations and also improves the overall interpretability with attribution methods (32); while the rest used a rectified linear unit. The first convolutional layer employs 200 filters each with a size of 19 and a stride of 1. The second convolutional layer employs 300 filters each with a size of 9 and a stride of 1. And the third convolutional layer employs 300 filters each with a size of 7 and a stride of 1. All convolutional layers incorporate zero-padding to achieve the same output length as the inputs. First two convolutional layers are followed by max-pooling layer of window size 10, and the last one followed by a global average pooling layer. The fully-connected hidden layer employs 512 units with rectified linear unit activations. Dropout (44), a common regularization technique for neural networks, is applied during training after each convolutional layer, with a dropout probability set to 0.2 for convolutional layers and 0.5 for fully-connected hidden layers.

Training. All models were trained with mini-batch stochastic gradient descent (mini-batch of 100 sequences) with Adam updates (45) with a decaying learning rate using a binary cross-entropy loss function. The initial learning rate was set to 0.001 and decayed by a factor of 0.3 if the model performance on a validation set (as measured by the Pearson correlation) did not improve for 7 epochs. Training was stopped when the model performance on the validation set does not improve for 25 epochs. Optimal parameters were selected by the epoch which yields the highest Pearson correlation on the validation set. The parameters of each model were initialized according to Glorot initialization (46).

Filter visualisation. To visualise first layer filters, we scanned each filter across every sequence in the test set. Sequences whose maximum activation was less than a cutoff of 50% of the maximum possible activation achievable for that filter in the test set were removed (27, 47). A subsequence the size of the filter centred about the max activation for each remaining sequence and assembled into an alignment. Subsequences that are shorter than the filter size due to their max activation being too close to the ends of the sequence were also discarded. A position frequency matrix was then created from the alignment and converted to a sequence logo using Logomaker (48).

Saliency analysis. To test interpretability of trained models, we generate saliency maps (33) by computing the gradients of the predictions with respect to the inputs. Saliency maps were multiplied by the query sequence (times inputs) and visualised as a sequence logo using Logomaker (48).

Abbreviations

scATAC-seq: single-cell assay for transposase accessible chromatin using sequencing, sci-ATAC-seq: single-cell combinatorial indexing ATAC-seq, dscATAC-seq: droplet single-cell assay for transposase-accessible chromatin using sequencing, dsciATAC-seq: dscATAC-seq with combinatorial indexing, BICCN: BRAIN initiative cell census network, IN: inhibitory neuron, EX: excitatory neuron, NN: non-neuronal cell, ROC: receiver operating characteristics, AUROC: area under the receiver operating characteristics curve, AUPR: area under the precision-recall curve, TSS: transcriptional start site, TF: transcription factor, SVM: support vector machine, ML: machine learning, CNN: convolutional neural network, and GEO: gene expression omnibus.

Ethics approval and consent to participate

Ethics approval was not needed for the study.

Availability of data and materials

Source codes and marker gene sets are available at `https://github.com/carushi/Catactor`.

Competing interests

The authors declare that they have no competing interests.

Funding

J.G. and RKK were supported by NIH grants R01MH113005 and R01LM012736. RKK was additionally supported by the Uehara Memorial Foundation Postdoctoral Fellowship. SF was supported by NIH grant U19MH114821. This work was also supported in part by funding from the NCI Cancer Center Support Grant (CA045508) and the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

Authors' contributions

JG conceived the project. JG and RKK designed computational experiments. RKK performed computational experiments. PK, ZT, and RT developed and interpreted deep learning models. RKK wrote the initial draft of the manuscript with assistance from JG, SF and PK. SF generated meta-analytic marker lists. All the authors read and approved the final manuscript.

Acknowledgements

We thank Dr. Rongxin Feng, Dr. Song Chen, and Dr. Chenxu Zhu for providing us with a script and metadata for their studies used in our meta-analysis. We also would like to thank Dr. Sukalp Muzumdar, Jonathan Werner, and other lab members for their thoughtful comments.

1. Bosiljka Tasic, Zizhen Yao, Lucas T Graybuck, Kimberly A Smith, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729): 72–78, 2018.
2. Megan Crow, Anirban Paul, Sara Ballouz, Z Josh Huang, and Jesse Gillis. Characterizing the replicability of cell types defined by single cell rna-sequencing data using metaneighbor. *Nature communications*, 9(1):1–12, 2018.
3. Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, 574(7778):418–422, 2019.
4. BRAIN Initiative Cell Census Network (BICCN), et al. A multimodal cell census and atlas of the mammalian primary motor cortex. *bioRxiv*, 2020. doi: 10.1101/2020.10.19.343129.
5. Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods*, 10(12):1213, 2013.
6. Ulrike M Litzenburger, Jason D Buenrostro, Beijing Wu, Ying Shen, Nathan C Sheffield, Arwa Kathiria, William J Greenleaf, and Howard Y Chang. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome biology*, 18(1):1–12, 2017.
7. John F Fullard, Mads E Hauberg, Jaroslav Bendl, Gabor Egervari, Maria-Daniela Cirnaru, Sarah M Reach, Jan Motl, Michelle E Ehrlich, Yasmin L Hurd, and Panos Roussos. An atlas of chromatin accessibility in the adult human brain. *Genome research*, 28(8):1243–1252, 2018.
8. Nicholas Pervolarakis, Quy H Nguyen, Justice Williams, Yanwen Gong, Guadalupe Gutierrez, Peng Sun, Darisha Jhutty, Grace XY Zheng, Corey M Nemec, Xing Dai, et al. Integrated single-cell transcriptomics and chromatin accessibility analysis reveals regulators of mammary epithelial cell identity. *Cell Reports*, 33(3):108273, 2020.
9. Darren A Cusanovich, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, Xingfan Huang, Lena Christiansen, William S DeWitt, et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell*, 174(5):1309–1324, 2018.
10. M Ryan Corces, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, Wanding Zhou, Tiago C Silva, Clarice Groeneveld, Christopher K Wong, Seung Woo Cho, et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413), 2018.
11. Hideyuki Yoshida, Caleb A Lareau, Ricardo N Ramirez, Samuel A Rose, Barbara Maier, Aleksandra Wroblewska, Fiona Desland, Aleksey Chudnovskiy, Arthur Mortha, Claudia Dominguez, et al. The cis-regulatory atlas of the mouse immune system. *Cell*, 176(4): 897–912, 2019.
12. Chuanyu Liu, Mingyue Wang, Xiaoyu Wei, Liang Wu, Jiangshan Xu, Xi Dai, Jun Xia, Mengnan Cheng, Yue Yuan, Pengfan Zhang, et al. An atac-seq atlas of chromatin accessibility in mouse tissues. *Scientific data*, 6(1):1–10, 2019.
13. Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U Gorkin, Yanxiao Zhang, Brandon C Sos, Veena Afzal, Diane E Dickel, et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nature neuroscience*, 21(3):432–439, 2018.
14. Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K Shiau, Xinzhu Zhou, Fangming Xie, Eran A Mukamel, Kai Zhang, Yanxiao Zhang, M. Margarita Behrens, Joseph R Ecker, and Bing Ren. Comprehensive analysis of single cell atac-seq data with snapatac. *Nature communications*, 12(1):1–15, 2021.
15. Yang Eric Li, Sebastian Preissl, Xiaomeng Hou, Ziyang Zhang, Kai Zhang, Rongxin Fang, Yunjiang Qiu, Olivier Poirion, Bin Li, Hanqing Liu, Xinxin Wang, Jee Yun Han, Jacinta Lucero, Yiming Yan, Samantha Kuan, David Gorkin, Michael Nunn, Eran A. Mukamel, M. Margarita Behrens, Joseph Ecker, and Bing Ren. An atlas of gene regulatory elements in adult mouse cerebrum. *bioRxiv*, 2020. doi: 10.1101/2020.05.10.087585.
16. Yijing Su, Jaehoon Shin, Chun Zhong, Sabrina Wang, Prith Roychowdhury, Jongseuk Lim, David Kim, Guo-li Ming, and Hongjun Song. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nature neuroscience*, 20(3):476, 2017.
17. John K. Mich, Lucas T. Graybuck, Erik E. Hess, Joseph T. Mahoney, Yoshiko Kojima, Yi Ding, Saroja Somasundaram, Jeremy A. Miller, Natalie Weed, Victoria Omstead, Yemeserach Bishaw, Nadiya V. Shapovalova, Refugio A. Martinez, Olivia Fong, Shenqin Yao, Marty Mortrud, Peter Chong, Luke Loftus, Darren Bertagnolli, Jeff Goldy, Tamara Casper, Nick Dee, Ximena Opitz-Araya, Ali Cetin, Kimberly A. Smith, Ryder P. Gwinn, Charles Cobbs, Andrew. L. Ko, Jeffrey G. Ojemann, C. Dirk Keene, Daniel. L. Silbergeld, Susan M. Sunkin, Viviana Gradinaru, Gregory D. Horwitz, Hongkui Zeng, Bosiljka Tasic, Ed S. Lein, Jonathan T. Ting, and Boaz P. Levi. Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. *bioRxiv*, 2020. doi: 10.1101/555318.
18. Caleb A Lareau, Fabiana M Duarte, Jennifer G Chew, Vinay K Kartha, Zach D Burkett, Andrew S Kohlway, Dmitry Pokholok, Martin J Aryee, Frank J Steemers, Ronald Lebofsky, et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(8):916–924, 2019.
19. Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E Vinyard, Sara P Garcia, Kendell Clement, Miguel A Andrade-Navarro, Jason D Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome biology*, 20(1):1–25, 2019.
20. Song Chen, Blue B Lake, and Kun Zhang. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology*, 37(12):1452–1457, 2019.
21. Chenxu Zhu, Miao Yu, Hui Huang, Ivan Juric, Armen Abnousi, Rong Hu, Jacinta Lucero, M Margarita Behrens, Ming Hu, and Bing Ren. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. Technical report, Nature Publishing Group, 2019.
22. Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.
23. Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience*, 19(2):335–346, 2016.
24. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.
25. Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, 08 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz625.
26. Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.
27. David R Kelle, Jasper Snoek, and John L Rinn. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990–999, 2016.
28. J Zhou and O. G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–4, 2015.
29. Peter K Koo and Matt Ploenzke. Deep learning for inferring transcription factor binding sites. *Current Opinion in Systems Biology*, 19:16–23, 2020.

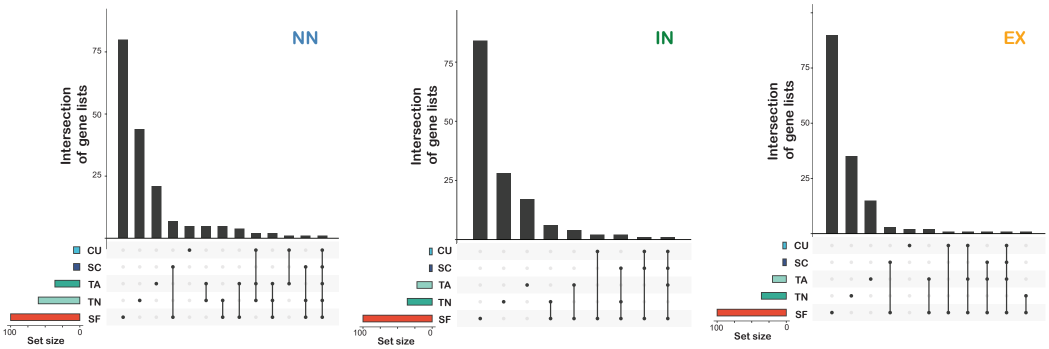
30. Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl_1):D91–D94, 2004.
31. Shobhit Gupta, John A Stamatoyannopoulos, Timothy L Bailey, and William Stafford Noble. Quantifying similarity between motifs. *Genome biology*, 8(2):R24, 2007.
32. Peter K Koo and Matt Ploenzke. Improving representations of genomic sequence motifs in convolutional networks with exponential activations. *Nature Machine Intelligence*, 3(3): 258–266, 2021.
33. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*, 1312.6034, 2013.
34. Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (tf-modisco) version 0.5. 1.1. *arXiv*, 1811.00416, 2018.
35. Zizhen Yao, Hanqing Liu, Fangming Xie, Stephan Fischer, Ricky S Adkins, Andrew I Aldrige, Seth A Ament, Ann Bartlett, M Margarita Behrens, Koen Van den Berge, et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *Biorxiv*, 2020.
36. Rebecca D Hodge, Trygve E Bakken, Jeremy A Miller, Kimberly A Smith, Eliza R Barkan, Lucas T Graybuck, Jennie L Close, Brian Long, Nelson Johansen, Osnat Penn, et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature*, 573(7772): 61–68, 2019.
37. Ken Sugino, Chris M Hempel, Mark N Miller, Alexis M Hattox, Peter Shapiro, Caizi Wu, Z Josh Huang, and Sacha B Nelson. Molecular taxonomy of major neuronal classes in the adult mouse forebrain. *Nature neuroscience*, 9(1):99–107, 2006.
38. Daniel Quang and Xiaohui Xie. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11): 107, 2016.
39. Alexandra Maslova, Ricardo N Ramirez, Ke Ma, Hugo Schmutz, Chendi Wang, Curtis Fox, Bernard Ng, Christophe Benoist, Sara Mostafavi, et al. Deep learning of immune cell differentiation. *Proceedings of the National Academy of Sciences*, 117(41):25655–25666, 2020.
40. Roman Spektor, Jee Won Yang, Seoyeon Lee, and Paul D Soloway. Single cell atac-seq identifies broad changes in neuronal abundance and chromatin accessibility in down syndrome. *bioRxiv*, page 561191, 2019.
41. Satoshi Hara and Takanori Maehara. Finding alternate features in lasso. *arXiv preprint arXiv:1611.05940*, 2016.
42. Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, A Danese, Marta Interlandi, Michaela Fee Mueller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 2020. doi: 10.1101/2020.05.22.111161.
43. Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.
44. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
45. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
46. Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.
47. Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.
48. Ammar Tareen and Justin B Kinney. Logomaker: beautiful sequence logos in python. *Bioinformatics*, 36(7):2272–2274, 2020.

Tables

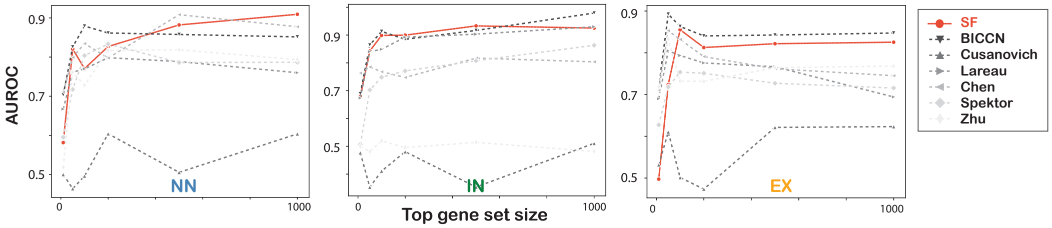
Table 1. Seven scATAC-seq datasets. C represents a combinatorial indexing method while Joint does a joint profiling method for scATAC-seq and scRNA-seq. *: A novel cluster assignment is estimated by Leiden clustering using Scanpy library. †: Inhibitory and excitatory clusters are inferred according to Slc17a7 and Gad2 activity following the description in (18).

	BICCN	Preissl	Cusanovich	Lareau	Chen	Spektor	Zhu
Cell number (after filtering)	110,013	16,767	5,081	46,653	5,081	13,766	15,191
Cluster	21	36*	12	27	23	26	9
Pipeline	SnapATAC	snATAC	TF-IDF +Seurat	chromVAR	cisTopic	Monocle	Seurat
Original DR	Yes	No	Yes	No	Yes	No	Yes
Annotation	Yes	No	Yes	Yes †	Yes	Yes	Yes
Peak or bin	1,000 bp	Peak	5,000 bp	Peak	Peak	Gene	1,000 bp
Protocol	C	C	C	C + Droplet	Joint	C	Joint

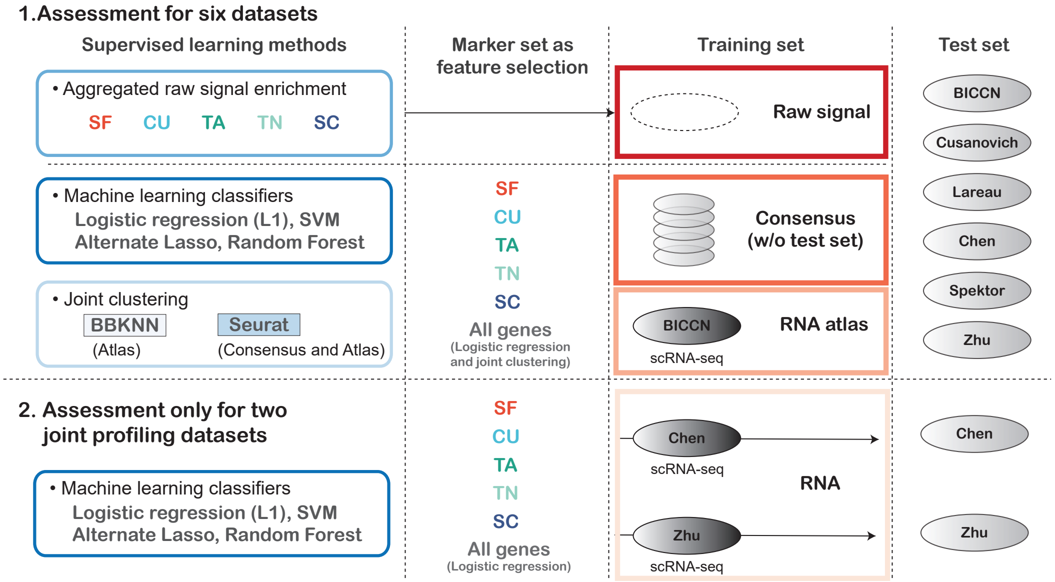
Additional Files



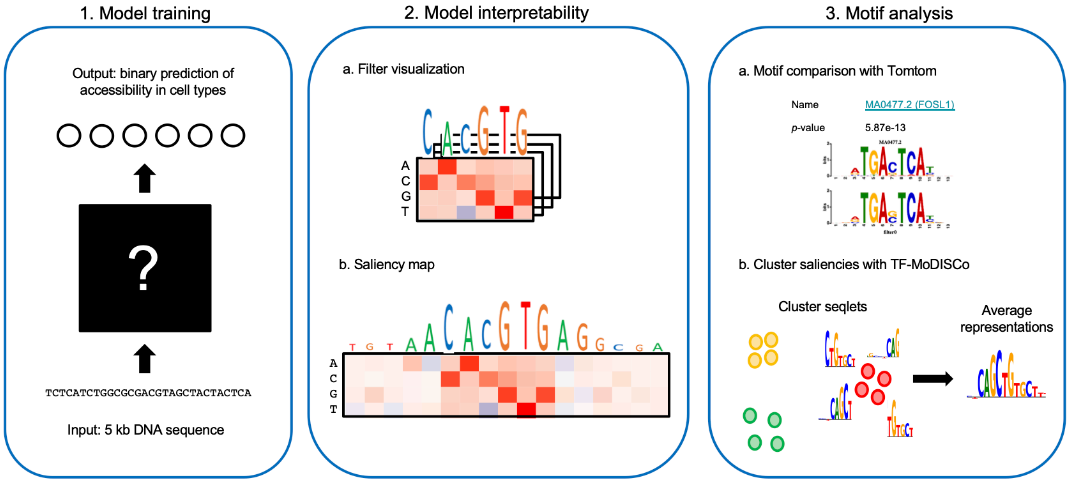
Supplementary Figure 1. Histograms of 5 marker gene sets used in this study and their intersection sizes.



Supplementary Figure 2. AUROCs of cluster-level cell-type classification by computing normalized Jaccard scores for top cluster-specific genes and marker gene sets. As a marker set, the AUROCs of top 1,000 cell-type specific genes derived from each dataset for non-overlapping datasets are compared with those of the SF marker gene sets across all well-annotated datasets.


1.Assessment for six datasets
Supervised learning methods
Marker set as feature selection
Training set
Test set
• Aggregated raw signal enrichment
SF CU TA TN SC
Raw signal
BICCN
Cusanovich
Lareau
Chen
Spektor
Zhu
• Machine learning classifiers
Logistic regression (L1), SVM
Alternate Lasso, Random Forest
SF
CU
TA
TN
SC
All genes
(Logistic regression and joint clustering)
Consensus (w/o test set)
• Joint clustering
BBKNN
(Atlas)
Seurat
(Consensus and Atlas)
BICCN
scRNA-seq
RNA atlas
2. Assessment only for two joint profiling datasets
• Machine learning classifiers
Logistic regression (L1), SVM
Alternate Lasso, Random Forest
SF
CU
TA
TN
SC
All genes
(Logistic regression)
Chen
scRNA-seq
RNA
Zhu
scRNA-seq
Chen
Zhu

Supplementary Figure 3. A workflow of a comprehensive assessment of scATAC-seq cell-type classification at an individual cell level.


1. Model training
Output: binary prediction of accessibility in cell types
?
TCTCATCTGGCGCGACGTAGCTACTACTCA
Input: 5 kb DNA sequence
2. Model interpretability
a. Filter visualization
A
C
G
T
b. Saliency map
A
C
G
T
3. Motif analysis
a. Motif comparison with Tomtom
Name
MA0477.2 (FOSL1)
p-value
5.87e-13
b. Cluster saliencies with TF-MoDISCo
Cluster seqlets
Average representations

Supplementary Figure 4. Deep learning analysis workflow. The CNN model was trained to take DNA sequence as input and predict the sequence's accessibility in each cell type. Model interpretability is accomplished by visualizing filters layer convolutional filters and by cell-type specific saliency analysis for a given sequence. Each can be used to visualize motif features extracted by the model. Motif analysis is accomplished by employing a motif comparison search against a database of known motifs (i.e. JASPAR) using Tomtom or via TF-MoDISCo, which splits saliency maps, clusters them, and provides an averaged representation.

Supplementary Data 5. An html file including the results of a Tomtom motif comparison search between 1st layer filters and the 2020 JASPAR vertebrates database.