Imperial College London

Classification of brain cell types from epigenetic data using machine learning.

Edward Christopher Rees

September 2023

CID: 02167694

Imperial College London, National Heart and Lung Institute, Faculty of Medicine.

Submitted in partial fulfilment of the requirements for the degree of MSc in Genomic Medicine

Supervisor: Dr Nathan Skene

Word count: 6827

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the M.Sc. in Genomic Medicine at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline.

Abstract

Recent advances in single-cell sequencing have enabled the provision of highly granular data about the characteristics of the multitude of cells found in living organisms. Combined with parallel advances in computational bioinformatics and machine learning, analysis of this data has led to insights about the behaviour of these cells. While inferring cell types from gene expression data has become routine, doing so from epigenetic data remains a challenge due to the gene centric nature of current cell typing approaches.

Here we investigated the viability of brain cell type identification directly from chromatin accessibility epigenetic data. We benchmarked performance using annotated multi-omic datasets. Despite certain constraints, the research highlighted the importance of the right reference dataset, achieving 99% accuracy in some cases. Our approach gives insight into the cell type calling performance and is adaptable to other epigenetic assays such as those measuring histone marks or transcription factor binding. Future work could refine labelling, introduce other assays, and automate dataset selection, potentially providing further tools to help in advancing Neurogenomic research.

Acknowledgements

I would like to thank Dr Nathan Skene and Alan Murphy for their supervision, for sharing their expertise in epigenetics and handling single-cell data, and for providing me with substantial opportunities to develop as a bioinformatician during and after this project. I would also like to extend this thanks to the rest of the members at Imperial College London Neurogenomic Laboratory who welcomed me and provided useful advice and feedback throughout the course of the project.

Table of Contents

1 Introduction

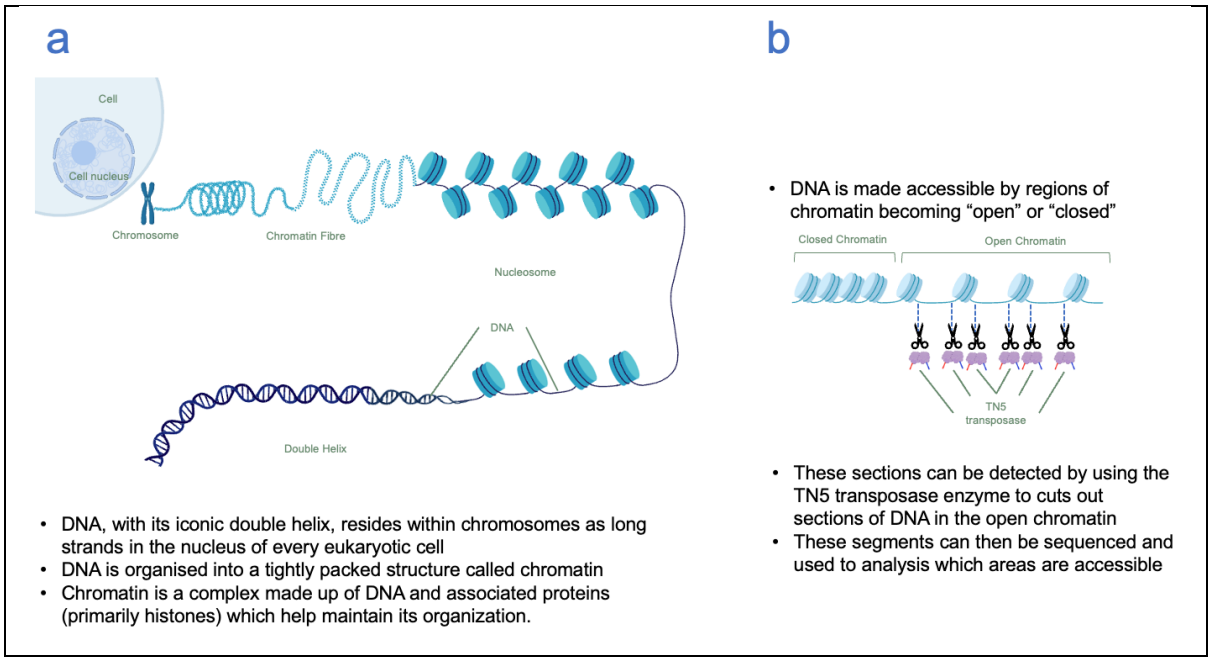
1.1 Background

The human body consists of trillions of cells which can be categorised into hundreds of different distinct major cell types residing in different tissues. Although each type of cell essentially carries the same DNA, there are significant differences in the gene expression profiles impacting each cell's behaviour. These differences are driven by the cells' regulatory system responding to extracellular, environmental cues. The regulatory system includes epigenetic factors, such as chromatin accessibility, histone modifications and non-coding RNAs, which affect the DNA in each cell and impact its function and phenotype.

1.2 Developments in sequencing methods and technology

Recent years have seen a rapid development in omics sequencing technologies which allow researchers to better understand the transcriptomic (gene expression) and epigenetic profiles of different cell types and, more recently, individual cells.

Chromatin accessibility refers to the degree to which DNA is accessible or "open" and is indicative of gene regulatory activity (Figure 1). Assays supporting investigation into chromatin accessibly have been available since 2006 using DNase-seq(1) and more recently, ATAC-seq. This was developed in 2013 (2), providing a faster and simpler method to obtain chromatin accessibility data and better understand chromatin dynamics across different cells.



Adapted from "Genomic Architecture" by BioRender.com (2023).
Retrieved from https://app.biorender.com/biorender-templates

Figure 1: Illustrative Structure of DNA and Chromatin

a. Illustrative schematic showing organisation of DNA within chromatin, chromosome and cell nucleus (not to scale)
b. Overview of process to detect open chromatin regions using TN5 transposase

Bulk RNA sequencing was first introduced in 2008 and this allowed researchers to measure the presence and quantity of gene transcripts(3) in collections of cells.

Recent advances in microfluidics combined with DNA barcoding meant that individual cells could be separated into tiny droplets and tagged for unique identification in large scale sequencing runs(4, 5, 6). This single cell technology is now widely applied to RNA-seq (7) and epigenetic assays to support the analysis of chromatin accessibility(8) and histone modification at a single cell level (9, 10).

More recently single-cell multi-omics assays now provide simultaneous transcriptomic and epigenetic information at a single cell resolution, and these have now become widely commercially available (11, 12), These have the drawback that they are more expensive, and cannot cover as many cells with sufficient read quality (13).

These tools, coupled with improvements in throughput, accuracy and automation, have allowed researchers to obtain highly detailed and granular genetic and epigenetic profiles of complex organs at a cellular level. This data has helped in deciphering the intricate biological functions and interactions between cells. Larger programmes have also been progressed to consolidate this highly detailed information into resources such as Human Brain Atlases and other Omics data repositories (14, 15).

1.3 Parallel development of supporting software tools

In response to these advances in sequencing technology and availability of increasingly rich datasets, an ecosystem of largely open-source computational bioinformatics tools has developed in parallel. These address complex tasks in handling and analysing single cell data. Examples include: -

1. Computational pipelines to manage the raw sequence data to generate gene expression matrices per cell or peaks per genomic location (examples include 10X cell ranger (16), STAR (17), featureCounts (18))

2. Quality control tools to filter out unwanted cell data or account for droplets containing more than one cell (examples include Seurat (19), SCANPY (20))

3. Integration of data from different samples, technologies and experimental batches (Seurat (19))

4. Reduction in dimensionality of data to help highlight variation across cell types (examples include UMAP (21) and t-SNE (22) integrated within the Seurat (19) toolkit)

5. Clustering techniques to group by different features such as differential gene expression (examples include Seurat (19), edgeR (23), DESeq2 (24) refs).

6. Transferring labels from previously labelled reference datasets using nearest neighbour methods and clustering. This supports cell type annotation of query single cell datasets (e.g. Seurat (19), SingleR (25), scClassify (26))

1.4 Impact of advances in Machine Learning (ML)

The wide availability of biomedical data and size of data sets has presented both a challenge and useable resource to which machine learning and data science techniques can be readily applied.

Machine learning (ML) is a field in computer science that uses statistical models to learn and make predictions or decisions based on patterns of input and output data rather than direct programming of rules. ML can be used for supervised tasks (where the output data has been labelled with expected values to be predicted from the input data) or unsupervised tasks (where hidden patterns can be established from unlabelled data).

ML and Deep Learning (DL – a specialised form of ML which uses multiple layers of inputs and outputs to model non-linear interactions) have seen considerable use in the past few years to help study complex biological systems. In particular they have been used for automatic feature extraction, feature selection and generation of predictive models. These are frequently integrated with the bioinformatics tools highlighted above (27) and an example is discussed in section 1.5.

The growth in use of these computationally intensive models has been, at least partly, driven by the increasing use of powerful parallel graphical processing units (GPUs), originally

designed for rendering graphics in video games. The availability of large GPUs via cloud computer providers has meant the ability to process large multidimensional data sets has not only been made widely available to researchers with limited local computing resources but has also enhanced reproducibility (28, 29).

Advances in open-source tools and libraries have also given bioinformaticians access to the power of ML without needing to deal with the complexities or expensive licence fees. Widely used examples include TensorFlow (30), Keras (31), PyTorch (32) and Scikit-Learn (33).

1.5 Example application: determining cell types from gene expression data.

The identification of cell types from single cell omics data (often referred to as cell typing) is one example of a task that can be supported by machine learning. Historically identifying cell types was a time consuming and laborious task. This involved manually reviewing characteristics under a microscope or use of florescent antibodies attaching to marker proteins. Accurate identification of cell types is a key requirement to improve understanding of human biology, disease pathology and ultimately develop improved therapeutic strategies (34).

The advent of scRNA-seq has seen the development of computational tools to determine cell types based on gene expression characteristics. These represent an improvement in the potential automation and throughput of cell typing methods, allowing more informed analysis of large dataset.

Several methods already exist to determine cell type from scRNA-Seq data (35), however most are based on reference datasets which have been manually curated (based on known marker genes and experts' opinions). Although the availability of annotated data sets for scRNA-Seq data are becoming more widely available, similar reference datasets are not currently available for epigenetic marks.

1.6 Challenges in identifying human brain cell types.

Classification of human brain cells into distinct types is an important step in improving the understanding of the how different parts of the brain function, how they interact and which cell types are implicated in neurodegenerative diseases such as Alzheimer’s and Parkinson’s (36). When combined with knowledge about specific genes and proteins associated with a disease this can also facilitate the development of more targeted therapies.

Brain cells present special challenges in cell typing for a number of reasons. Firstly, the availability of fresh, healthy human tissue is severely limited due to the inherent risks in obtaining brain biopsies from healthy subjects and limited supply of donated post-mortem samples. This means studies are often underpowered. Secondly the preparation of brain tissue needs to reflect the delicate and sensitive nature of the raw materials. Different cell types are often closely intermingled in brain tissue and traditional separation methods such as florescence activated cell sorting (FACS) can stress or damage the cells (37). Additionally the process to dissociate brain cells can harm the integrity of cells meaning only single nucleus rather than whole cell data is available with consequent loss in data quality (38).

Aside from manual cell typing methods, due to lack of available curated references, methods using unsupervised clustering of transcriptomic data from cells are often used. These, however, require manual annotation based on gene characteristics and expert opinion which can lead to challenges in reproducing results.

Easy to use scRNA-seq celltyping tools such as Azimuth(39), can be used via a web interface and provide a quick accurate method without requiring a programming interface. However, this method currently has only limited reference datasets available for analysing human brain cells (based on the Human Motor Cortex), so using this reference for other brain regions or tissue types could provide uncertain results.

1.7 Development of open access data resources

In the past few years, a number of studies have carried out comprehensive transcriptomic analyses of different mammalian organs and tissue types. These have provided a rich source of annotated datasets covering a wide range of tissue types and developmental stages. Examples include studies covering mouse brain (40, 41, 42) the human brain (14, 15).

When used in combination with the software tools identified above, these provide a potential source of reference material covering a wide range of tissue and cell types to allow more targeted cell typing for specific tissues.

1.8 Advantages of cell typing from epigenetic data only

Since the most commonly used methods of cell typing rely on differential gene expression, researchers investigating biological systems using epigenetic assays, such as scATAC-seq,

currently need to undertake additional steps to determine cell types or rely on rudimentary nearest marker gene approaches(2). Currently multi-omics methods (which can identify cell type from the scRNA-seq component) are more expensive and have more limited single cell coverage.

A method to accurately identify cell types from epigenetic data alone could be helpful in improving the efficiency of research experiments and also provide additional insights; ultimately the data provided from these assays combined with cell type information could provide additional understanding of the mechanisms driving cell type-specific gene expression in healthy and diseased states.

1.9 Project objective and outline

This project seeks to establish whether automatic cell type classification methods based on non-curated epigenetic data are feasible and if they can be practically applied to tissue and cell types encountered in the brain.

To establish this, the following steps were carried out: -

- A brief review of recently documented epigenetic cell typing methods.
- Selection of suitable multi-omic scRNA-seq and scATAC-seq data relating to human brain tissue, to be used for benchmarking purposes.
- Selection of suitable reference data sets matching the tissue type of the benchmark data with suitable annotations for cell types.
- Creation of a set of benchmark results for different tissues cell types using well established methods operating on the RNA-Seq data component of the multi-omic data set and the appropriate reference data sets for that tissue type.
- A practical in-silico experiment based on a subset of the epigenetic methods using only the scATAC-seq data component of the multi-omic benchmarking data to predict cell types.
- Analysis of the results to quantify the accuracy of the predictions.

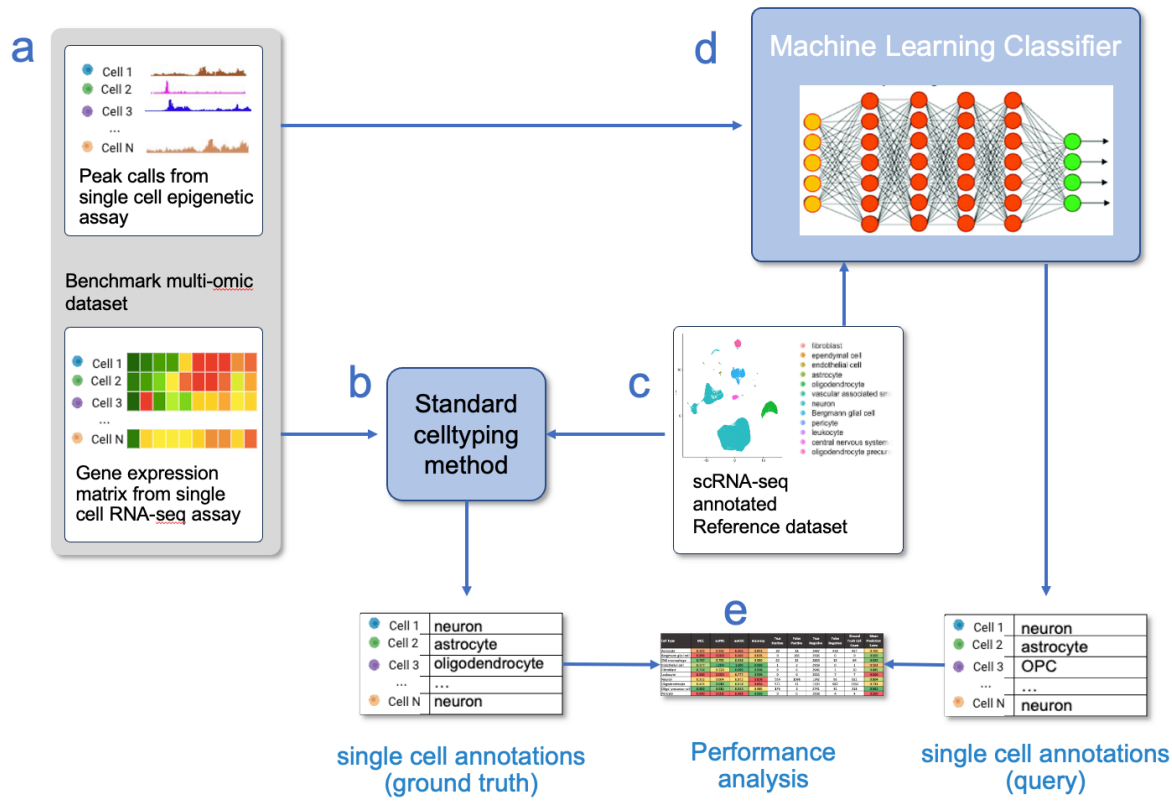


Figure 2: High level project outline

a. scRNA-seq and scATAC-seq from multi-omic benchmarking data processed separately
b. Standard scRNA-seq celltyping method applied to scRNA-seq benchmark component
c. Annotated reference data used for label transfer
d. Machine learning classifier uses scATAC-seq data to generate annotations for query data
e. Results from scRNA-seq and scATAC-seq annotations compared to assess performance

2 Materials and Methods

2.1 Pre-processing of annotated reference data

A basic coding structure was designed such that any benchmark data could be used in the form of a Seurat object containing suitable cell type labels. Although this data may have already had pre-processing steps carried out and stored, a standard process was followed based on raw counts to ensure consistency across different laboratories and data sets.

In all cases the default or recommended settings and parameters were used based on the supplied documentation and examples

Following installation of required statistical and bioinformatics packages (see section 2.6), The annotated reference data was processing using the following steps: -

a) Basic consistency checks were carried out on reference object to ensure compatibility with the Seurat (V4 or above) and Signac functions used in later steps. This included checks to ensure that the relevant RNA count data is present in the correct assay area in the object for a standard pipeline.

b) A standard pipeline was run to pre-process the data. This was based on the recommendations in the Seurat and Signac documentation (43, 44) and consisted of the following steps: -

    i. Normalisation of data. This ensures raw counts are comparable between cells. The standard LogNormalise function was used. This firstly normalises the feature expression measurements for each cell based on the total expression. It then uses a a scale factor (10,000 by default), and finally calculates and stores the result of the log1p (Natural log of 1+x) function.

    ii. Finding variable features – This selects a sub-set of gene showing the greatest variability to optimise later computation steps. The recommended setting of 2000 genes was used. The method uses a variance stabilising transformation (VST) which provides an improvement on using log normalised variance by accounting for the mean-variance relationship that is inherent to single-cell RNA-seq data (45). These most variable genes allow optimisation of later steps

by filtering out genes which have low variability and thus minimal impact on the later analysis steps(46).

iii. Scaling – this step helps to reduce bias in later stages by centring the mean and variance of the gene counts prior to principal component analysis.

c) A standard process was then followed to reduce the dimensionality of the data to support transfer of labels (for use in later benchmark and query data). This consisted of the following steps: -

iv. Running a principal component analysis (PCA) based on the variable features identified in 4b using the recommended 30 dimensions. PCA is a commonly used to reduce the dimensional complexity of single cell gene expression matrices. This helps to reduce the background noise of gene expression variability, supports clustering or grouping of cells by common features and provides suitable data for downstream processes such as clustering or trajectory analysis (47).

v. Running a Uniform Manifold Approximation and Projection (UMAP) (21) on the results of the PCA step for use in visualisation and to confirm that the labels are consistent with the clustering from these 2 steps (see supporting plots in supplementary materials 6.4)

2.2 Creation of a benchmark results based on scRNA-Seq data

For the purposes of benchmarking, a multi-Omics data set was used. This includes both scRNA-seq gene expression data and scATAC-seq chromatin accessibility information for the same set of cells. Each cell is uniquely identified by a “barcode” – a small strip of DNA attached to each genomic fragment.(48) This allows the data for each cell to be matched across the assay type.

Raw count matrices representing the scRNA-seq and scATAC-seq parts were extracted from the data output from the assay pipeline (in this case it was based on a 10X genomics cell ranger pipeline(11)). Separate data objects were created for each assay type to ensure a clean separation so that no cell typing data was present in the scATAC-seq part. The unique barcode could then be used to compare the results of cell typing for each cell using the scRNA-seq method and the scATAC-seq method.

As data from different sources held gene information in different formats, Gene-Ids were used throughout based on the Ensembl(49) standard. Gene references based on gene name were converted via a simple look up table based on 58,000 Gene-ids (including accession numbers) from the reference data. This provided a greater coverage of gene-ids than the standard Bioconductor tools and preserved approximately 95% of the genes provided in the raw data.

Processing steps for the benchmark scRNA-seq data included the same steps as 2.1 a) and b) above.

Following completion of the PCA step additional steps were included to support the transfer of Cell type labels from Reference data set to Query data set. This process followed the standard Seurat process for Cell typing from RNA-seq as follows:-

a) **Anchor identification:** This process uses an unsupervised learning approach to identify “anchors”, which are 2 cells (one from each) which have matching characteristics across the Reference and query data set.(46)

   This process reviews the nearest neighbour characteristics of the pairs of cells based on their gene expression profiles in the reduced space PCA data sets (called Weighted Nearest Neighbors (WNN) analysis)(39). Once closest matching cells across data sets have been identified (based on their nearest neighbour scores) and filtered (to remove low confidence anchors) , this allows the clusters in the reference and query sets to be linked, which then supports the transfer of other information such as cell type labels (19, 46).

b) **Label transfer:** Once the anchors have been identified, the cell type labels and associated prediction metrics were transferred from the reference data set to the query dataset based on these anchors (Figure 3). This information forms the ground truth for comparison of cell types with those derived from the scATAC-seq assay.

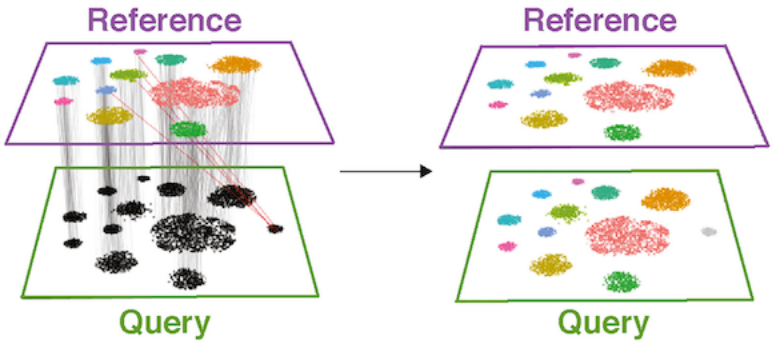


Figure 3: Creation of Anchors to support transfer of Labels from Reference to Query scRNA-seq datasets
(adapted from (46))

2.3 Creation of celltyping results from scATAC-seq data

The scATAC-seq data consists of a series of raw counts for a specific genomic range. These are based on the number of reads that have been collected for this sequence of DNA from the scATAC-seq assay. Each read represents a section of DNA where the Tn5 transposase has tagged open chromatin regions with sequencing adapters (2, 50). This data was separately processed as follows:

a) Creation of Seurat chromatin data object

The GenomicRanges Bioconductor library was used to create a standard Genomic ranges (GRanges) object from the strings containing the genomic intervals and chromosome information. Gene annotations are also prepared (using standard chromosomes only) based on annotation information derived from the Ensembl hg38 human genome (49). These are then used to together with the raw read fragments to create a Seurat Chromatin Assay data object needed for subsequent steps.

To support reduction of dimensionality of the assay data, the following standard steps were carried out. Note that these differ from the process used for scRNA-seq data.

b) Calculation of **Term Frequency – Inverse document frequency (TF-DIF)** score for each peak. This indicates the relative importance of a peak in a particular cell relative to its frequency across the wider cell population. This helps to highlight peaks unique to particular cells.

c) Feature selection

The most variable peak features are then derived from TF-DIF data in order to create a smaller data set representing the most variable regions for use in subsequent analysis.

d) LSI dimensionality reduction

A dimensionality reduction step is carried out using **Latent Semantic Indexing (LSI),** a method used in natural language processing used to discover patterns in unstructured text. This uses the identified variable peak features from the previous TF-DIF step and then applies **Singular Value Decomposition (SVD**)(51). SVD is used in preference to PCA since the ATAC-seq data is sparse (containing many empty areas) and does not require the same scaling or centring required for RNA-seq data (52).

e) UMAP and Clustering

To support visualisation and plotting, a UMAP step was carried out using the output from the LSI reduction. Note that only dimensions 2:30 were used since dimension 1 has been shown to correlate to sequencing read depth, which is not useful for cell differentiation.(52)

Next, graph-based clustering was performed on dimensions 2:30 by calculating a shared nearest-neighbour graph from the LSI space using the FindNeighbors and FindClusters functions provided by Seurat.

f) Generation of gene activity matrix

A gene activity matrix simulates the gene expression counts based on the ATAC-seq assay information prepared in the steps above. This is achieved by counting the ATAC-seq read fragments overlapping the gene body within a pre-defined range. This is generated by the Signac GeneActivity function (44).

The synthesized gene activity matrix is used to create an RNA-seq assay which undergoes, normalisations, scaling and dimensional reduction steps as outlined above.

g) Label transfer:

The simulated gene expression data was then used to perform cell type labelling from the reference data in a similar way to that used for a normal scRNA-seq data set, with the following key difference: the anchor cells are identified using canonical correlation analysis (CCA) rather than PCA (in line with the Seurat recommendations for cross-modality data). When transferring the cell type labels the LSI results are used to calibrate the weightings used in the nearest neighbour calculations and the 1st dimension is ignored (as explained above)

The results of these predictions were finally saved to allow comparison with the ground truth to assess the performance of the annotation process.

2.4 Analysis of results to quantify performance.

Evaluating the cell typing method and supporting reference datasets required choosing suitable performance metrics. While a simple measure of "accuracy" (based on percentage of correct predictions divided by total predictions) seems straightforward, it doesn't adequately reflect misclassifications like false positives and negatives.

More comprehensive metrics like Receiver Operator Curves (ROC) and Precision Recall (PR) curves, which consider prediction scores beyond binary outcomes, were included. However, the significance of false positives or negatives varies by application. Some situations might tolerate some false negatives for gain more true negatives or accept some false positives for high true positive coverage. Given these nuances and the presence of varying cell type populations, a balanced metric like Matthew’s correlation coefficient (MCC) (53, 54) was chosen for its comprehensive reflection of the confusion matrix.

Results, using scRNA-seq as ground truth and scATAC-seq predictions, were generated for three reference datasets and validated against standard human brain references. Low gene expression (>200 features) or peak count data (<3000 reads) were excluded in-line with recommended thresholds, leading to about 8.4% fewer cells in the results used in the analysis.

2.5 Creation of matching results for each reference dataset

To assess performance across a range of different reference data the following data sets were used. The first 3 data sets allowed a comparison of results where different populations of similar cells were present in the same tissue type as the benchmark. The 4th reference

was from a commonly used reference data set for a generic area of the brain (Table 1). This provided a sanity check for the benchmark RNA-seq data set to ensure consistency of results.

Reference	Tissue Type	Description	Cell Type Count														
			Astrocyte	Bergmann glial cell	CNS macrophage	choroid plexus epithelial	endothelial	ependymal	fibroblast	leukocyte	neuron	oligodendrocyte	oligodendrocyte precursor cell	pericyte	vascular assoc. smooth muscle	Vascular Leptomeningeal	Total
1	Cerebellum - Cerebellar Vermis	Dissection of Cerebellar Vemis from Homo Sapiens based on 10X genomic 3' v3 assay. Available on CellXgene resource based on paper by Siletti et al.	203	6,614	1,431		278	1	173	61	54,680	4,387	3,862	176	8		71,874
2	Cerebellum - Cerebellar deep nuclei	Dissection of Cerebellar deep nuclei from Homo Sapiens based on 10X genomic 3' v3 assay. Available on CellXgene resource based on paper by Siletti et al.	6,267	19	3,304	2	60	17	182	238	23,155	16,991	1,729	67	14		52,045
3	Cerebellum - Lateral hemisphere	Dissection of Cerebellum Lateral hemisphere from Homo Sapiens based on 10X genomic 3' v3 assay. Available on CellXgene resource based on paper by Siletti et al.	108	1,070	241		34		69	11	24,799	1,105	564	25	2		28,028
4	Motor Cortex	The reference as used by the Azimuth referencing tool was generated from data and annotations available from the Allen Institute. Originally described in Bakken et al, bioRxiv 2020, this data was created by integration across the two individual donors	568		108		64				72,528	2,942	283			40	76,533

Table 1: Annotated reference data sets used in cell typing comparisons

An additional reference data set from a completely different human tissue (the lung) was additionally tested to assess whether the results produced would indicate if the reference data set was not suitable.

2.6 Software and data availability

All scripts and R Code used to process and analysis benchmark results are available at https://github.com/neurogenomics/benchmark_epig_celltyping. All R package versions are listed in supplementary materials 6.6. Additionally, a Conda environment to install all these packages is available on the GitHub link above to support reproduction of the environments used. Details and links for the data used are also provided on the GitHub site.

3 Results

3.1 Results using reference set 1

We first aimed to demonstrate the basic viability of the cell typing process using a suitable initial reference data set and a benchmarking multi-omic data set (Figure 4).

A flash-frozen human healthy brain tissue (cerebellum) data set containing 3233 cells was sourced from the 10X Genomics datasets(55) . We used the Cerebellum - Cerebellar Vermis dataset sourced from the Chan/Zuckerburg resource(56) as a reference (see supplementary materials 6.2).

a

Cell Type	MCC	auPRC	auROC	Accuracy	True Positive	False Positive	True Negative	False Negative	Ground Truth Cell Count	Mean Prediction Score
Astrocyte	0.236	0.344	0.700	0.924	24	13	2713	212	236	0.521
Bergmann glial cell	0.595	0.492	0.939	0.912	192	241	2508	21	213	0.887
CNS macrophage	0.850	0.835	0.979	0.992	73	15	2864	10	83	0.901
Endothelial cell	0.577	1.000	1.000	0.999	1	2	2959	0	1	0.655
Fibroblast	0.652	0.611	0.949	0.997	8	7	2945	2	10	0.816
Leukocyte	0.000	0.067	0.996	0.999	0	0	2960	2	2	0.000
Neuron	0.388	0.733	0.893	0.649	533	981	1388	60	593	0.861
Oligodendrocyte	0.463	0.935	0.938	0.678	648	31	1360	923	1571	0.740
Oligo' precursor cell	0.857	0.908	0.946	0.979	190	3	2710	59	249	0.955
Pericyte	0.000	0.252	0.623	0.999	0	0	2958	4	4	0.000

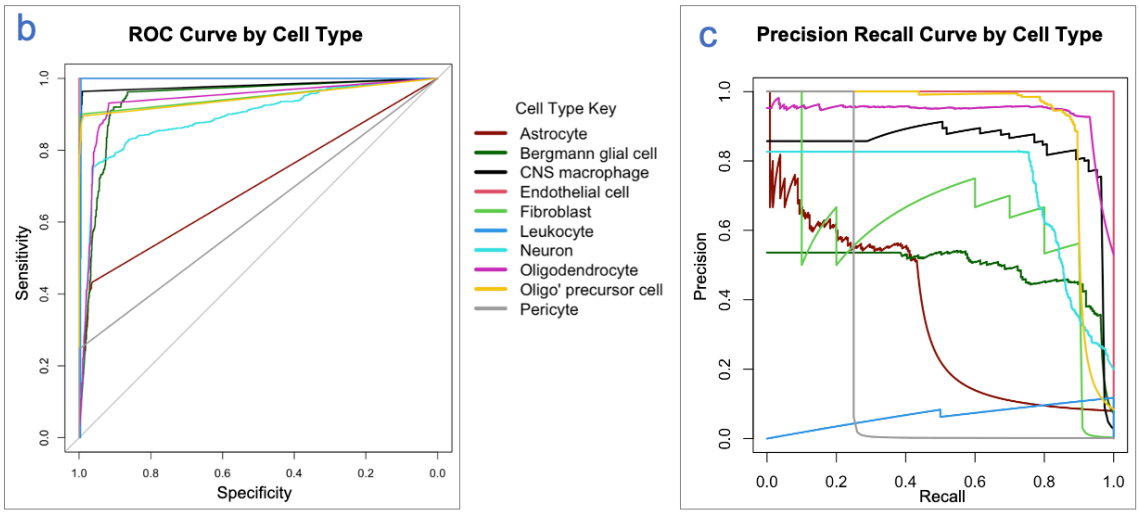


Figure 4: Results using Cerebellar Vermis reference.
(a) shows the performance of scATAC-seq prediction vs Ground Truth prediction (based on scRNA-seq data by cell type). MCC = Matthews Correlation Coefficient, auPRC= area under precision recall curve, auROC = area under receiver operator curve.
(b) shows the Receiver Operator Curve (ROC) and **(c)** gives the Precision Recall Curve for each cell type.

As can be seen, some cell types had reasonable prediction performance with oligodendrocyte precursor cells and central nervous system (CNS) macrophages matching the scRNA-seq derived cell types most closely, with MCC both > 0.85. However, oligodendrocytes, astrocytes and neurons accounted for over 80% of the cells and only had an MCC below 0.5. Overall mean prediction score (indicating consistency of anchors with the reference data sets) across all cells was 0.840.

We reviewed the poorly performing cell types and checked whether the reference dataset adequately covered these cell types. We noticed that although this dataset had more cells (>70,000) than others from the same tissue, the coverage of some cell types (such as oligodendrocytes and astrocytes – see table 1) was lower than the cerebellum reference data set for deep nuclei, so this was tested next.

3.2 Results using reference set 2

We carried out a similar process to 3.1 using the same benchmark data but using an alternative cerebellum reference from the same source covering cerebellar deep nuclei (Figure 5). This consisted of fewer overall cells but with a different balance of cell types compared to the reference used in 3.1. For example there were 30 times more astrocytes in this reference data set.

a

Cell Type	MCC	auPRC	auROC	Accuracy	True Positive	False Positive	True Negative	False Negative	Ground Truth Cell Count	Mean Prediction Score
Astrocyte	0.893	0.925	0.973	0.970	456	49	2417	40	496	0.947
Bergmann glial cell	0.000	0.000	0.000	1.000	0	0	2962	0	0	0.000
CNS macrophage	0.903	0.910	0.977	0.994	90	15	2853	4	94	0.928
Endothelial cell	0.000	1.000	1.000	1.000	0	0	2961	1	1	0.000
Fibroblast	0.638	0.668	0.999	0.997	7	8	2946	1	8	0.851
Leukocyte	0.516	0.532	0.822	0.999	2	3	2956	1	3	0.628
Neuron	0.922	0.970	0.992	0.977	489	17	2405	51	540	0.953
Oligodendrocyte	0.929	0.991	0.987	0.965	1569	68	1289	36	1605	0.981
Oligo' precursor cell	0.929	0.950	0.975	0.991	186	2	2749	25	211	0.967
Pericyte	-0.001	0.079	0.623	0.998	0	1	2957	4	4	0.381

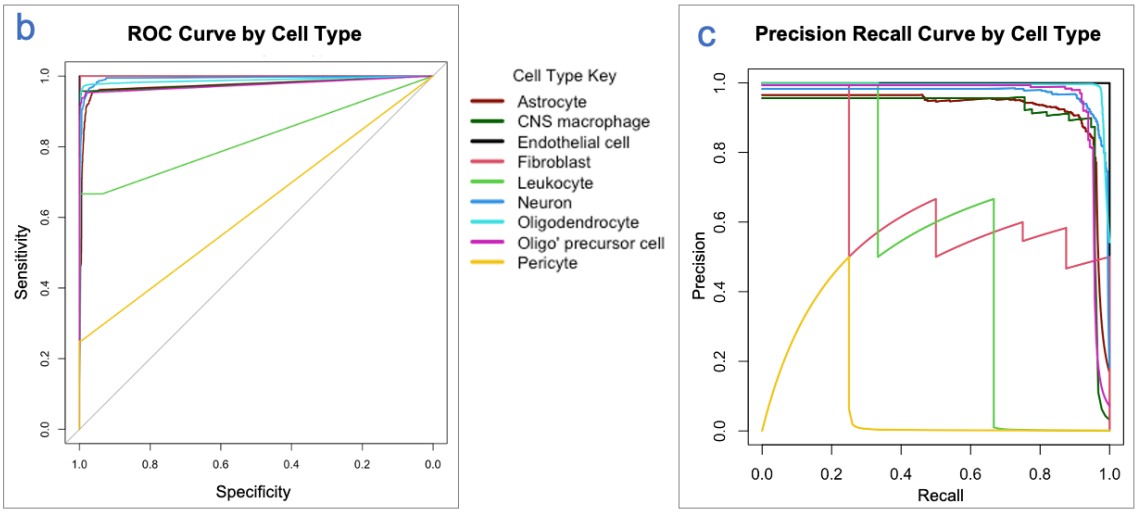


Figure 5: Results using cerebellar deep nuclei reference.
(a) shows the performance of scATAC-seq prediction vs Ground Truth prediction (based on scRNA-seq data by cell type). MCC = Matthews Correlation Coefficient, auPRC= area under precision recall curve, auROC = area under receiver operator curve.
(b) shows the Receiver Operator Curve (ROC) and **(c)** gives the Precision Recall Curve for each cell type.

Using this reference, the predictions demonstrated acceptable accuracy across a wider range of cell types; astrocytes, CNS macrophages, neurons, oligodendrocytes and oligodendrocyte precursor cells all achieved MCC's of 0.89 or better, representing over 99% of cells. Only cells with low cell counts were inadequately predicted. Overall mean prediction score across all cells was higher than achieved with the Cerebellar Vemis reference at 0.967.

To better understand the role the reference dataset plays in cell typing performance, we reran our pipeline using a third available cerebellum reference.

3.3 Results using reference set 3

This test consisted of the same benchmark described in 3.1 and 3.2 above, but used the lateral hemisphere dataset for the cerebellum from the same source (figure 6).

a

Cell Type	MCC	auPRC	auROC	Accuracy	True Positive	False Positive	True Negative	False Negative	Ground Truth Cell Count	Mean Prediction Score
Astrocyte	0.205	0.332	0.606	0.853	39	18	2487	418	457	0.760
Bergmann glial cell	0.000	0.000	0.606	0.876	0	366	2596	0	0	0.900
CNS macrophage	0.787	0.795	0.938	0.989	65	15	2863	19	84	0.922
Endothelial cell	0.577	1.000	1.000	0.999	1	2	2959	0	1	0.563
Fibroblast	0.734	0.720	0.999	0.998	9	6	2946	1	10	0.885
Leukocyte	0.000	0.050	0.777	0.998	0	0	2955	7	7	0.000
Neuron	0.352	0.664	0.872	0.608	559	1098	1242	63	622	0.864
Oligodendrocyte	0.425	0.918	0.918	0.656	572	33	1370	987	1559	0.733
Oligo' precursor cell	0.883	0.912	0.953	0.985	176	3	2741	42	218	0.961
Pericyte	0.000	0.001	0.498	0.999	0	0	2958	4	4	0.000

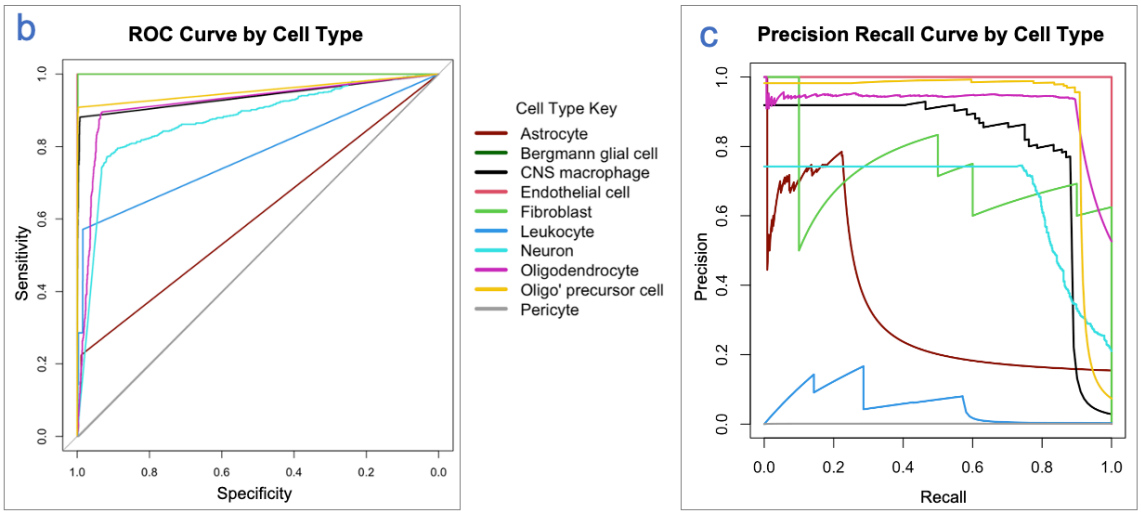


Figure 6: Results using cerebellar lateral hemisphere reference
(a) shows the performance of scATAC-seq prediction vs Ground Truth prediction (based on scRNA-seq data by cell type). MCC = Matthews Correlation Coefficient, auPRC= area under precision recall curve, auROC = area under receiver operator curve.
(b) shows the Receiver Operator Curve (ROC) and **(c)** gives the Precision Recall Curve for each cell type.

Using this reference, we found that, as with 3.1, some predictions were obtained with reasonable accuracy. CNS macrophages, fibroblasts and oligodendrocyte precursor cells were predicted with MCC's greater than 0.73 but the majority of cells were badly predicted. It was notable that 366 Bergmann glial cells were predicted from the ATAC-seq data where none were predicted from the scRNA-seq data.

Overall mean prediction score across all cells was 0.847.

3.4 Comparison of the 3 reference sets

Next, we compared results across all 3 data sets and considered possible correlations. Figure 6 below shows a comparison of the area under precision recall (auPR) obtained from each reference (Figure 7).

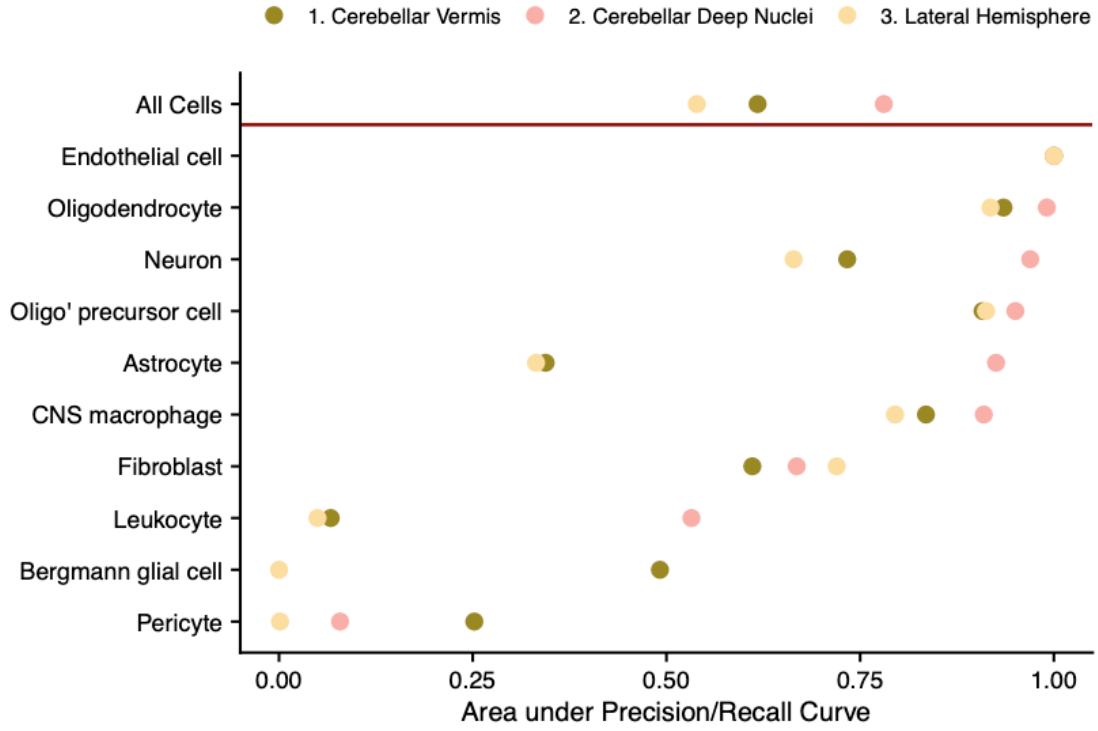


Figure 7: Plot showing comparative prediction performance (using auPR) by cell type for each of the 3 cerebellum reference datasets.
"All cells" shows the average auPR across cell types.

Clearly reference dataset 2 provided the best overall results based on comparison of the scATAC-seq prediction to the scRNA-seq cell type results, and this was most pronounced in the cell types with higher counts.

By reviewing the 3 reference data sets there intuitively appears to be a link between the accuracy of the predictions and the coverage of those cell types in the reference, so this area was analysed further.

3.5 Correlation analysis

In order to understand the relationship of the benchmark prediction performance to the annotation reference data, a correlation analysis was carried out plotting the prediction performance (shown by auPR) against the number of cells for a given cell type in the reference. This showed a correlation with a Pearson coefficient of 0.529 (Figure 8).

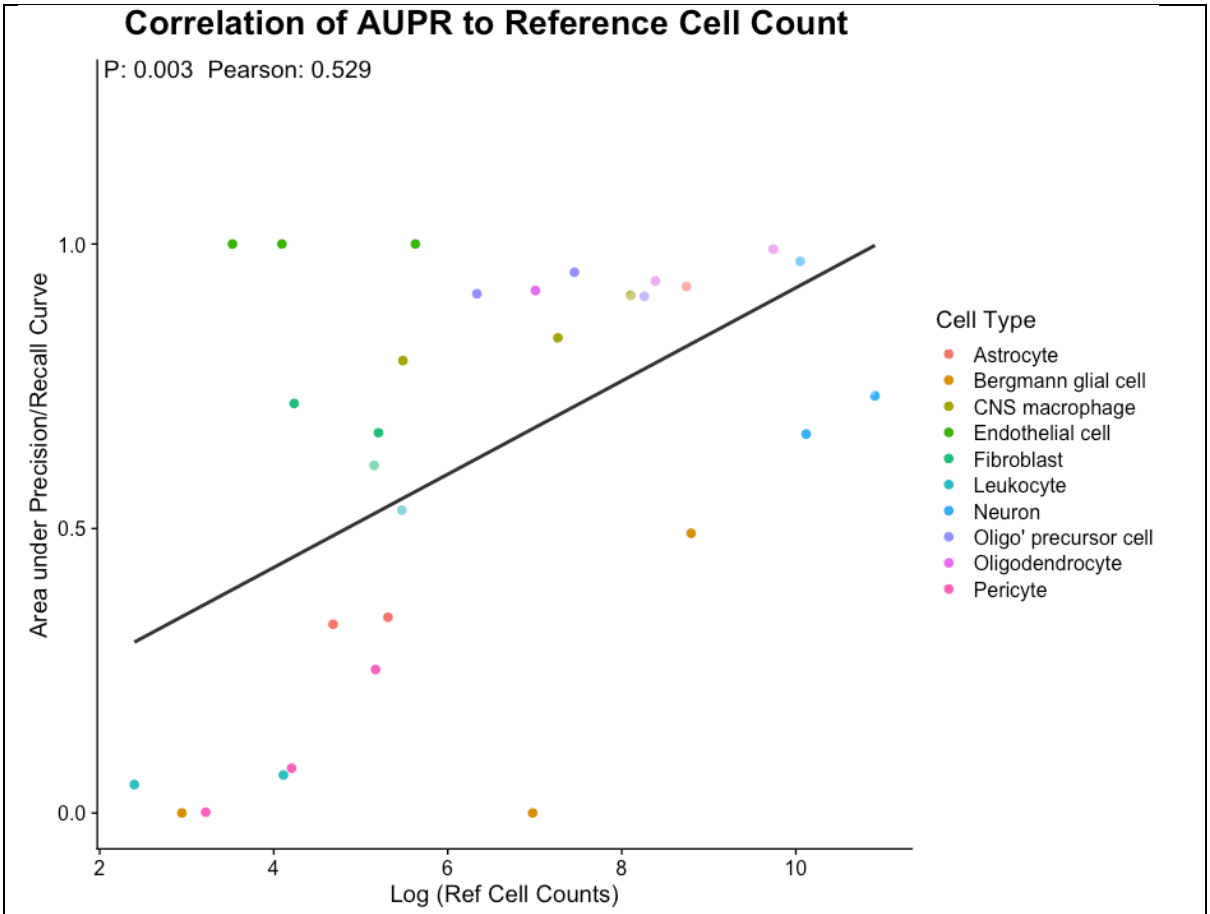


Figure 8: Correlation of benchmark auPR performance to cell type count in reference dataset

This highlights a relatively strong relationship between the number of cells and the predictive performance.

Next, we wanted to understand if a researcher only had the scATAC-seq data available, whether the prediction scores provided by the transfer labels function would be a predictor of accuracy, and thus help to identify whether a reference data set was suitable for use.

We therefore carried out an analysis of the auPR performance to average prediction scores (Figure 9). This showed a Pearson correlation of 0.568.

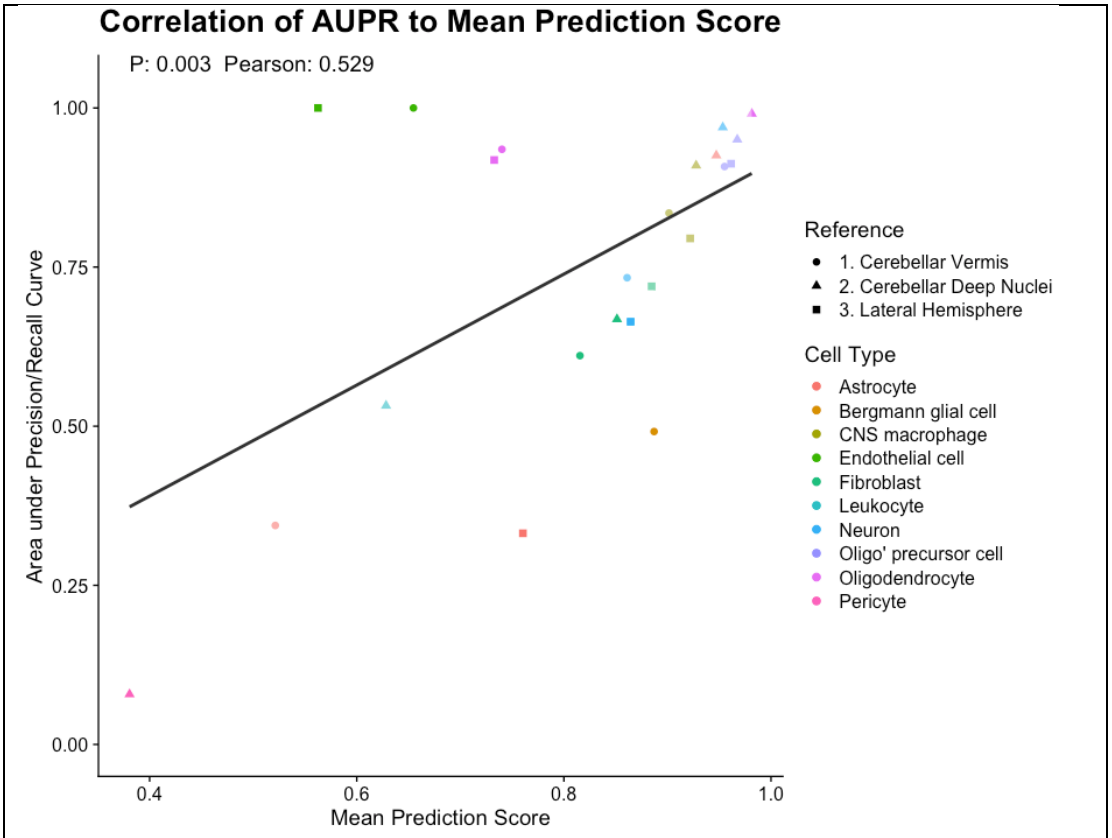


Figure 9: Correlation of benchmark auPR performance to prediction score by cell type and reference.

This also shows a reasonable correlation which is more pronounced for the cells with higher counts in the benchmark data (top right quadrant). This is consistent with our mean prediction scores per reference which show reference 2 with the highest mean score of 0.967. Overall, these findings show that the mean prediction score is a strong indicator of cell typing accuracy.

3.6 Consistency test using Azimuth cortex reference

Next, we wanted to ensure our ground truth cell type labelling was reasonable. To check this the scRNA-seq part of the benchmark data went through an annotation process using a standard Azimuth tool (39, 57) and a pre-processed reference data set (58) for the human brain motor cortex - the only brain reference data set provided with Azimuth. This uses an automated process (via a web app) to annotate cells. The results, based on comparison with the ground truth from 1.2 above, are shown in figure 7 below.

Cell Type	Matched		Not Matched	
	Count	Matching Score	Count	Matching Score
Astrocyte	436	0.999	98	0.944
CNS macrophage	95	0.995	9	0.888
Endothelial cell *	0	0.000	3	0.785
Fibroblast *	0	0.000	9	0.781
Leukocyte *	0	0.000	4	0.884
Neuron	486	0.997	77	0.956
Oligodendrocyte	1703	0.999	23	0.782
Oligo' precursor cell	201	0.975	18	0.793
Pericyte	0	0.000	4	0.823
Total	2921	0.997	245	0.909

(*These cell types are not available in the Azimuth motor cortex reference data)

Table 2: Table showing consistency of ground truth from cerebellar deep nuclei reference with Azimuth motor cortex cell type annotation

As can be seen above over 92% of cell types matched despite the different cell types available in the Azimuth reference. In all cases the matching scores for the matched cells were higher than the unmatched cells. As the Azimuth reference is based on different tissue with different cell type mix (including some missing) we would not expect a perfect match compared to the ground truth obtained in 1.2 above.

3.7 Verification test against reference from unrelated tissue with different cell types

Finally, to understand the impact of using a completely different tissue type (with non-matching cell types), the benchmark data set was tested using the Tabula Sapiens Lung scRNA-seq data set containing 35,682 cells (59). Although this successfully annotated the scRNA-seq data, the average prediction score was a very low 0.355 - indicating that this reference was unsuitable. These results (see supplementary materials 6.5), highlight how the mean predictive scores are reliable even when used with an unrelated tissue type as a reference.

4 Discussion

Finding a way to accurately identify cell types from epigenetic assays would provide numerous advantages for researchers. By understanding which cell types exhibit particular epigenetic patterns could provide insights into the functional role of the cell and its behaviour in diseased and healthy conditions. Other benefits could include simplified experimental steps by avoiding methods such as cell typing using multi-omic assays, or traditional FACS based sorting methods.

Despite the limitations of these experiments (explained below), we have demonstrated that given a suitable reference data set, brain cell types can be identified with sufficient accuracy to provide meaningful data for researching specific cell types. When using the best performing reference, 99% of cells were classified with an MCC > .89.

Our findings underscore that the provided prediction metrics reliably forecast performance, establishing a quality control framework for subsequent applications.

4.1 Limitations

There are a number of limitations to this research study which must be highlighted. Firstly, the number of benchmarking data sets was limited to only 1 multi-omic data set containing 2962 cells (after excluding low quality data). An additional data set covering human spinal cord data was also processed, but this proved to be unusable due to missing raw scATAC-seq fragment data which was only identified to be critical late in the experiment.

As more multi-omic data sets become available, it should be possible to carry out similar benchmarking for a wider set of tissue and cell types using the same processing and analysis pipeline. This will help to determine whether the results and conclusions are generalisable and/or whether classification of particular cell types have better or worse performance (which may be of interest in itself). Our pipeline has been made available (see Methods) and should only require straightforward changes to specify the updated multi-omic and reference file locations to generate new results.

Our Cell type annotation approach effectively creates a synthetic gene expression matrix which is then used to identify the cell types. Whilst this has limitations (further explained in 4.3a below), the predictions produced can be relatively easy to understand and explain,

based on this intermediate dataset. Any classification discrepancies could help further understand the complexities of deducing gene expression solely from chromatin accessibility.

Other methods were considered, including some which directly classify annotation data from scATAC-seq data using supervised deep learning models, but these were found to be unsuitable for the data and application of this project due to discrepancies in their training data (see supplementary materials 6.1). As more multi-omic data sets become available, this type of model may become more useable once pre-trained models cover a wide enough range of tissue and cell types.

4.2 Accuracy of ground truth

Since a detailed analysis of scRNA-seq methods was not the core objective of the project and it was not practical to consider expert manual curated cell typing approaches, the benchmark method was limited by the accuracy automated approach used. As demonstrated in the results, there was some variability in the cell types derived from the scRNA-seq data, which could impact the accuracy of the comparison with the scATAC-seq results.

A possible refinement would be to exclude cells where low confidence or lack of consensus in the scRNA-seq benchmark (“ground truth”) results were obtained. The Seurat method used provides two complementary scores to support this. The prediction score reflects the level of support by multiple consistent anchors between the query and reference data sets. The mapping score reflects how well the reference data set represents the data of the cell being queried. Cells could be excluded by setting a threshold in the mapping and/or prediction scores obtained during the ground truth labelling step.

A more comprehensive Quality Control (QC) process would also ensure only cells passing more stringent tests (as used in a standard bioinformatics pipeline) are included in the benchmarking dataset. This would help reduce technical noise and other unwanted artefacts inherent in wet lab processes. The reference data was assumed to have been through the documented quality control process(15), so no further quality control checks were applied.

4.3 Refinements to scATAC-seq method

A number of additional changes could be considered to improve the accuracy of the scATAC-seq method used here. Possible refinements could include: -

a) Modifications to the gene activity prediction model.
b) Filtering of benchmark data to exclude cells falling below a quality threshold.
c) Use of different levels of cell type labels
d) Exclusion of cells where mapping or prediction scores fall below a quality threshold.

a) Prediction of gene expression

The method used by Signac to determine the gene activity (and thus synthesise the expression matrix) uses the basic approach of summing all the ATAC-sec fragments intersecting the gene body and promoter region. The method uses the default of a 2Kb region upstream of the gene body to determine the promoter region. This figure is configurable and can also be extended downstream. Similarly, the length of the Gene body is limited to 500k bases by default to aid performance. A sensitivity analysis of these parameters may highlight possible prediction accuracy and compute performance impacts.

Moreover, other approaches to generate the gene activity matrix exist. The Signac authors recommend using the Cicero co-accessibility method. This analyses the co-accessibility of chromatin regions. By linking these accessibility pairs into cis co-accessibility Networks (CCANs), Cicero thus builds an association of chromatin region to genes beyond simple proximity (60). ArchR (61) also provides an alternative method to predict gene activity from chromatin accessibility data. Based on a comparison of 56 different models, they proposed a method that includes a weighting of the distance of the chromatin accessible sites from the transcription start sites (TSS).

Although the Signac method used in this project performs adequately, if a particular cell type is poorly predicted in other datasets, it is possible that key marker gene predictions are improved using these alternative methods.

b) scATAC-seq data quality assessment and control

As explained above for the scRNA-seq part, more stringent and standardised QC checks could also be applied to the scATAC-seq data. As this is normally part of a standard

pipeline we did not consider it appropriate to attempt to replicate this separately here, but using cleaner data could have impacted the results.

c) Level of cell type labels

Deciding what level and standard of cell type annotation to use would depend on the purpose the research user requires. We took the view that the highest level was the most important first step for accuracy assessment. Rather than define a particular standard ontology we directly used the cell type ontology used in the reference data source. Since 4 different levels of cell type label were available in our data, a further analysis of the lower-level cell types and respective accuracy could be useful depending on the specific requirements of the research being undertaken.

d) Use of label transfer scores

The Seurat method used for this labelling can provide both a prediction score and a mapping score for the label transfer process used here. In 4.2 we have explained how this could be used to check the validity of the scRNA-seq (ground truth) labelling.

The prediction score provides a metric which can be used to support selection of reference data (see results 3.5 & 3.7). Since the label transfer process for data derived from scATAC-seq data uses canonical correlation analysis (CCA), it is not possible to obtain a mapping score as part of the labelling process, but running a separate PCA based method would provide this, which could help in excluding cells with low mapping confidence.

In addition to the suggestions above, there are other areas where potential improvements could be further investigated. In general, we have used default or recommended parameter values at each step. However some steps (such as dimensionality reduction, clustering and identification of transfer anchors) could be further optimised by systematically modifying parameters and reviewing the impact on prediction and run-time performance.

4.4 Selection of reference dataset

As we can see from the results the selection of the reference dataset appears to be a significant factor in achieving acceptable results. Fortunately, the average prediction scores from the ATAC-seq cell annotation appears to provide a guide to help in this selection. When provided with a set of single cell data from a known tissue section and a set of relevant

potential reference datasets, it should be possible to select the most appropriate reference dataset and obtain acceptable cell type annotations.

This implies that given a wide enough available selection of reference data sets from sources such as CellxGene(56), it should be possible to achieve acceptable results for any tissue type from which the scATAC-seq data has been derived.

A useful follow on from this project would be to find a way to automate this reference selection process. Some papers suggest possible ways to better integrate reference data sets to make them smaller and more efficient, such as Symphony(62). This could provide an approach to create a more general-purpose reference data set for a given organ or set of tissue types. A recently published paper introduces CellAnn (63), which provides a way to automate the selection of reference data sets as part of an overall annotation workflow. This could provide a more automated process which integrates well with the Seurat objects we have used here.

4.5 Possible research applications

Since the method used in the project requires data concerning fragments of DNA linked to genomic location, other epigenetic assays which provide similar data at single cell resolution could also provide suitable query data. Rather than identifying regions of unbounded DNA, these assays are designed to look for bound proteins such as histone modifications or transcription binding sites. Experiments using scCUT&TAG, scCUT&RUN data for example, should in theory be able to provide similar raw data which could be used to generate gene activity matrices and ultimately for cell-type annotation.

As the current approach only considers reads at the promoter regions of genes, it may have limited success in generating predictions for assays where proteins bind solely outside these regions. For example, assays capturing histone marks which impact genes in the same vicinity (e.g. H3K4me3 which indicative of active promoters) are likely to be sufficiently well captured by the gene activity model used here. These would intuitively be expected to provide better cell typing than histone marks which act across longer stretches of DNA (e.g. H3K4me1 which is indicative of active enhancers). Sparsity of peak data from these assays compared to scATAC-seq could also negatively impact results. Nevertheless, the results would be of research interest and could help understand the cell type specific behaviour of epigenetic marks.

4.6 Conclusions

This project explored the potential of identifying distinct brain cell types using only epigenetic assay data. Despite certain limitations, notably the restricted number of benchmarking datasets available, the research demonstrated that good prediction performance is achievable but underscores the significance of selecting the correct reference dataset.

We have built an approach which should be adaptable for benchmarking of wider multi-omic datasets as they become available. Anomalies in cell classification on new benchmark datasets could in themselves provide helpful insights into the interaction of gene expression and chromatin activity.

Future improvements could consider refining the level of cell type labelling, introducing other assay types, and exploring automated reference dataset selection. While further exploration and validation is required, the preliminary results hold promise for simplifying cell typing and furthering our understanding of cell specific epigenetic characteristics in Neurogenomic studies.

5 Glossary

The following terms and abbreviations have been used in the text above.

ATAC-seq	An assay that measures chromatin accessibility across the genome. It stands for Assay for Transposase-Accessible Chromatin using sequencing.
auPRC	Area under the precision recall curve. A comprehensive statistic to evaluate a model's performance based on precision and recall values over different thresholds.
auROC	Area under the receiver operator curve. It represents the probability that a model will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
Azimuth	A reference tool for mapping and characterizing single cells.
Bergmann glial cells	Specialized astrocytes found in the cerebellum, a part of the brain.
CCA	Canonical Correlation Analysis. A way of measuring the linear relationship between two sets of multidimensional variables.
CNS	Central Nervous System. The part of the nervous system consisting primarily of the brain and spinal cord.
Conda	An open-source package management system and environment management system.
LSI	Latent Semantic Indexing. A technique in natural language processing, in particular in vectorial semantics, of analyzing relationships between a set of documents and the terms they contain.
MCC	Matthews Correlation Coefficient. A statistic used in machine learning as a measure of the quality of classifications.
Oligodendrocyte	A type of neuroglial cell. Its main function is to provide support and insulation to axons in the CNS.

PCA	Principal Component Analysis. A method used for reducing the dimensionality of datasets, increasing interpretability but at the same time minimizing information loss.
Pearson coefficient	A measure of the linear correlation between two variables. It ranges from -1 to 1, with 1 being total positive linear correlation, 0 no linear correlation, and -1 total negative linear correlation.
PR curves	Precision Recall curves. A metric used to evaluate the precision and recall of a binary classification.
ROC	Receiver Operator Curves. A statistical measure that evaluates the performance of a binary classification system.
scATAC-seq	Single-cell Assay for Transposase-Accessible Chromatin using sequencing. A method to determine chromatin accessibility at the single-cell level.
scRNA-seq	Single-cell RNA sequencing. A method used to sequence the RNA within individual cells to give insights into cellular functions.
Seurat	A tool for single cell genomics to identify and interpret sources of cellular heterogeneity from massive single-cell datasets.
Signac	An extension of Seurat, designed for the analysis of chromatin-based data.
TF-DIF	Term Frequency – Inverse document frequency. A method used to reflect how important a word is to a document in a collection or corpus.
UMAP	Uniform Manifold Approximation and Projection. A dimension reduction technique that can be used for visualization similar to t-SNE, but can also support general non-linear dimension reduction.
VST	Variance stabilising transformation. A method that stabilizes the mean-variance relationship in RNA-seq data.
WNN	Weighted Nearest Neighbors. A method for identifying the nearest cells based on their gene expression profiles.

6 Supplementary Materials

6.1 Review of epigenetic cell typing methods

A brief review of current methods was carried out by searching for suitable terms in documents published in the past 4 years (since the advent of standard scATAC-seq methods).

A recent benchmarking exercise(64) provided a useful starting point for possible candidates.

A short list of required features was drawn up following discussion with the principal investigator and secondary supervisor. The following key features were determined to be important: -

a) The tool should include a suitably pre-trained model. Due to limited amounts of data, it was considered undesirable to use the same multi-modal data set for both training and benchmarking. In addition, the time taken, and assumptions used for training could influence the results and lead to less effective generalisable use due to overfitting
b) As the lab is focussed on neurogenomics, the models should be suitable for use on single cell data derived from human brain cells
c) The models should work on Single Mode Epigenetic data and not require multi-mode data either for training or generating results (The multi modal data should be used purely for benchmarking)
d) The tool should ideally provide confidence or scoring information to support choice of reference and help indicate reliability of results when only scATAC-seq data is provided
e) The method should be practical to use and have sufficient documentation to support straightforward set and reproducibility of results. It should ideally integrate with existing automated processing pipelines for single cell data.

6.2 Selection of Reference data sets for label transfer

Selection of suitable reference data sets covering the same tissue and cell types in the benchmarking data sets.

Although a number of annotated reference data sets are already provided for scRNA-seq methods such as Azimuth(65) (a resource developed as part of the NIH Human Biomolecular Atlas Program / Hubmap consortium), these only cover a few tissue types and it is unclear how well these references provide granular or specific enough data to correctly annotate cells for scRNA seq from other tissue types.

A recent study undertook a transcriptomic study across the cell types of the human brain(15). This data was additionally provided as part of a wider data resource as part of the Chan/Zuckerburg initiative (56). This provides full transcriptomic data for a wide range of tissue and dissections of the human brain in a format which fits into the common toolset. A number of suitably annotated files for the spinal cord and cerebellum were identified and used. Annotations included in these file covered a number of levels including basic cell type and annotation of 461 clusters and 3313 subclusters.(15)

6.3 Choice of benchmarking methods – acting on scRNA-seq data

In order to create a ground truth of cell types for each cell in the benchmarking data, a standard method using scRNA-seq data needed to be employed.

As this was not the primary aim of the project only a brief review of available methods was undertaken based on published and pre-published papers. The following criteria were applied in making this choice: -

i. Method should be applicable to widely available scRNA-seq datasets produced from common assay methods.

ii. Method should be well documented with evidence of general use, regular releases, and some level of support (such as regularly updated GitHub issues list)

iii. It should be straightforward to integrate into commonly used pipelines in the Lab environment.

iv. Method should be referenced in benchmarking studies and be in top performing methods.

Bases on these factors Seurat was chosen as the preferred tool for this step. This conformed to the criteria as follows: -

i. Tools are provided to read data from common output of assays including 10X cell ranger, single cell experiment formats

ii. Regular releases, frequently updated issues in GitHub, widely used

iii. Seurat tools already used in single cell pipelines in the lab for various tasks such as QC

iv. Included in 3 out of the 5 benchmarking studies and was always in the top performing group (see Table x below)

Benchmarking Paper	Authors / Journal	Methods Reviewed	Findings Summary

1.Automated methods for cell type annotation on scRNA-seq data	Pasquini, Giovanni ; Rojo Arias, Jesus Eduardo ; Schäfer, Patrick ; Busskamp, Volker Computational & Structural Biotechnology Journal -19/1/2021	25	• Breakdown into 4 types
			  ○ Marker Gene DB
			  ○ Correlation
			  ○ Supervised Classification
			  ○ Other
			• Used 11 reference sources (inc. Allen Brain data)
			• 10 methods benchmarked against brain data
			• References used in other studies below but no preferred options highlighted

2. A comparison of automatic cell identification methods for single-cell RNA sequencing data	Abdelaal, T.R.M ; Michielsen, L.C.M ; Cats, Davy ; Hoogduin, Dylan ; Mei, Hailiang ; Reinders, M.J.T ; Mahfouz, A.M.E.T.A Genome Biology, 2019	22	• Performance evaluated using 27 publicly available datasets of different sizes, features (but no human brain samples) • Included 6 general-purpose classifiers from the scikit- learn library • Overall, several classifiers performed accurately across different datasets and experiments. SVMrejection, SVM, singleCellNet appear best overall • Used Seurat to provide marker genes using MAST method
3. Evaluation of single-cell classifiers for single-cell RNA sequencing data sets	Zhao, Xinlei ; Wu, Shuang ; Fang, Nan ; Sun, Xiao ; Fan, Jue Briefings in bioinformatics, 2020	9	• 8 pairs of datasets in evaluation (no Human brain) • Demonstrated that Seurat, SingleR and CaSTLe outperformed the rest of tools.

4. Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data	Huang, Qianhui ; Liu, Yu ; Du, Yuheng ; Garmire, Lana X. Genomics, proteomics & bioinformatics, 2021	10	• Seurat, SingleR, CP, RPC, and SingleCellNet performed well overall. Seurat was best for major cell type annotation. • Wide variety of public scRNA-seq datasets and some simulation data used for comparison tests
5. A comprehensive comparison of supervised and unsupervised methods for cell type identification in single-cell RNA-seq	Xiaobo Sun 1, Xiaochu Lin 2, Ziyi Li 3, Hao Wu 2 Briefings in Bioinformatics . 2022 Mar	18	• Performance of 8 supervised and 10 unsupervised cell type identification methods reviewed using 14 public scRNA-seq datasets with different tissues, sequencing protocols and species. (174K Human Brain Cells) • Generally the supervised methods outperformed the unsupervised methods, exception - unknown cell types • Seurat v3 mapping/singleR and Seurat v3 clustering were best overall supervised and unsupervised methods respectively

6.4 UMAP plots of reference and benchmark data

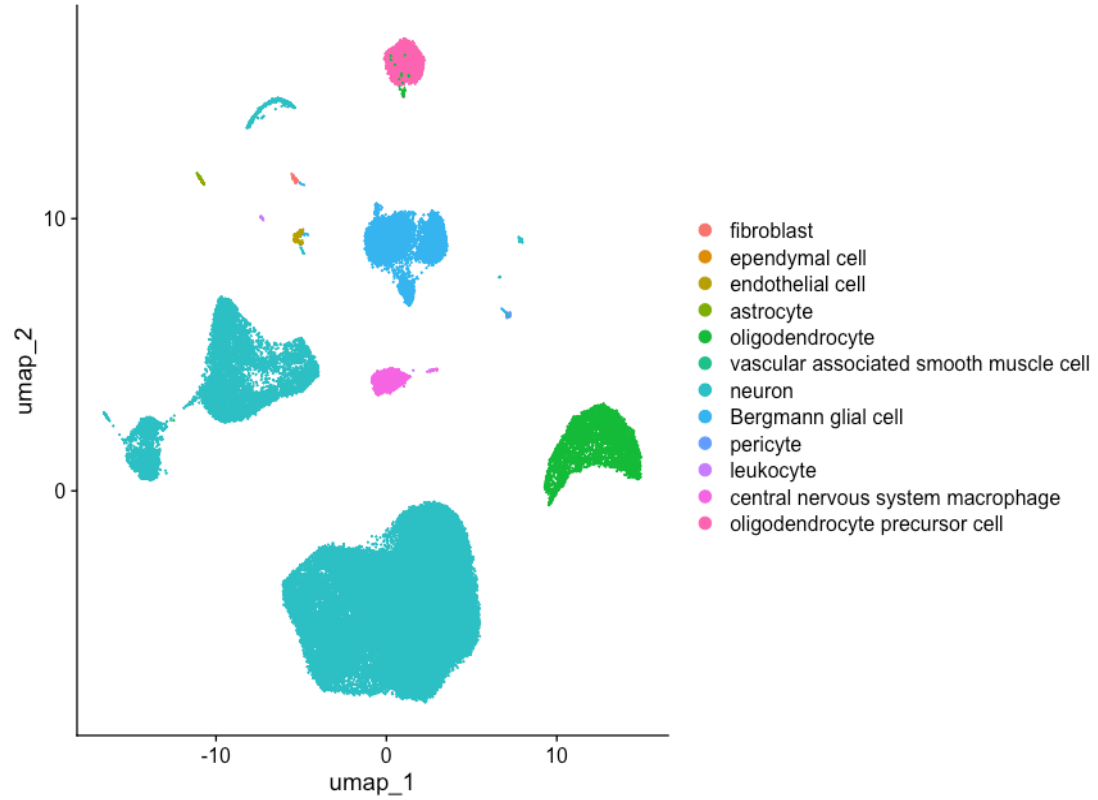


Figure S1: Example UMAP plot of Cerebellar Vermis reference dataset

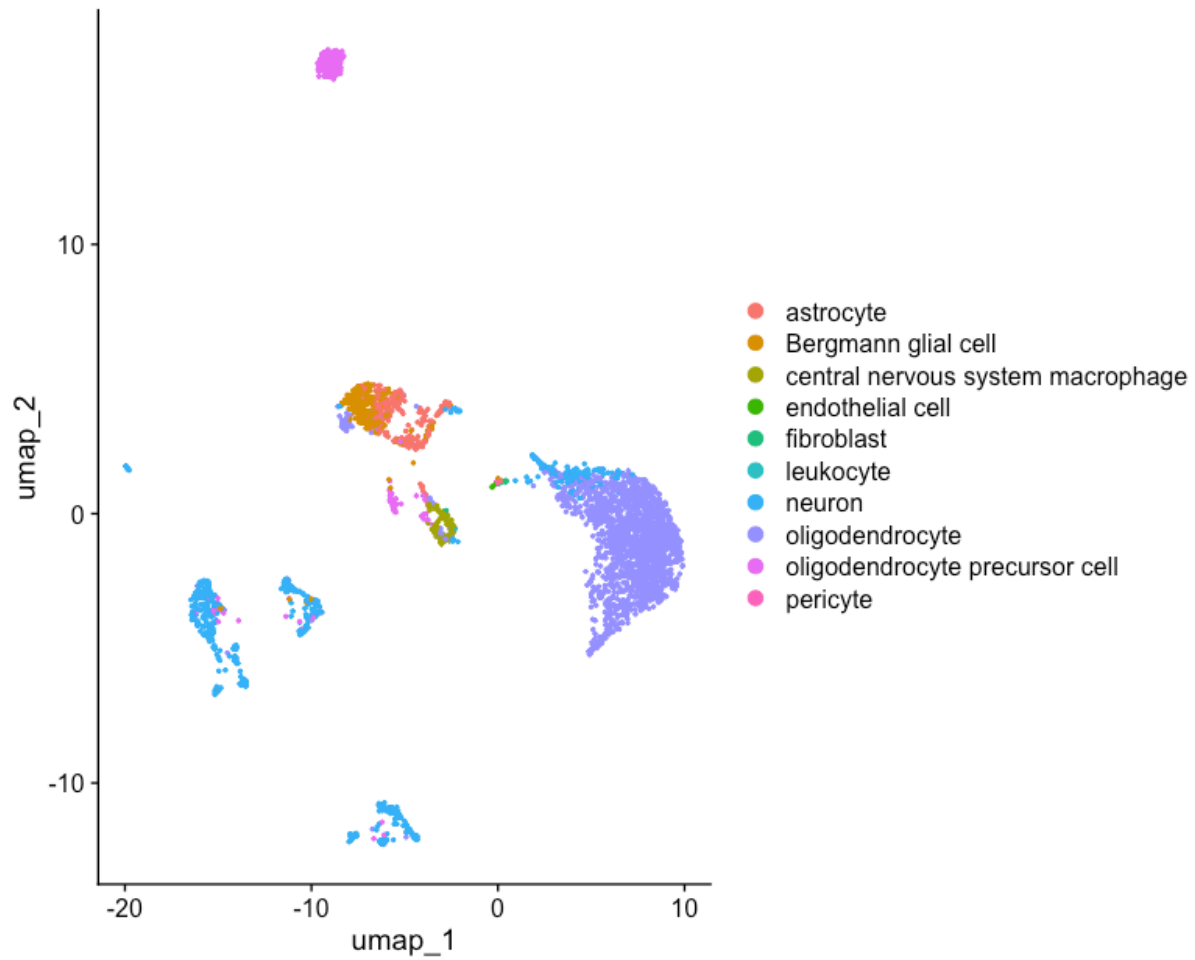


Figure S2: Example UMAP plot of Benchmark scRNA-seq data following label transfer

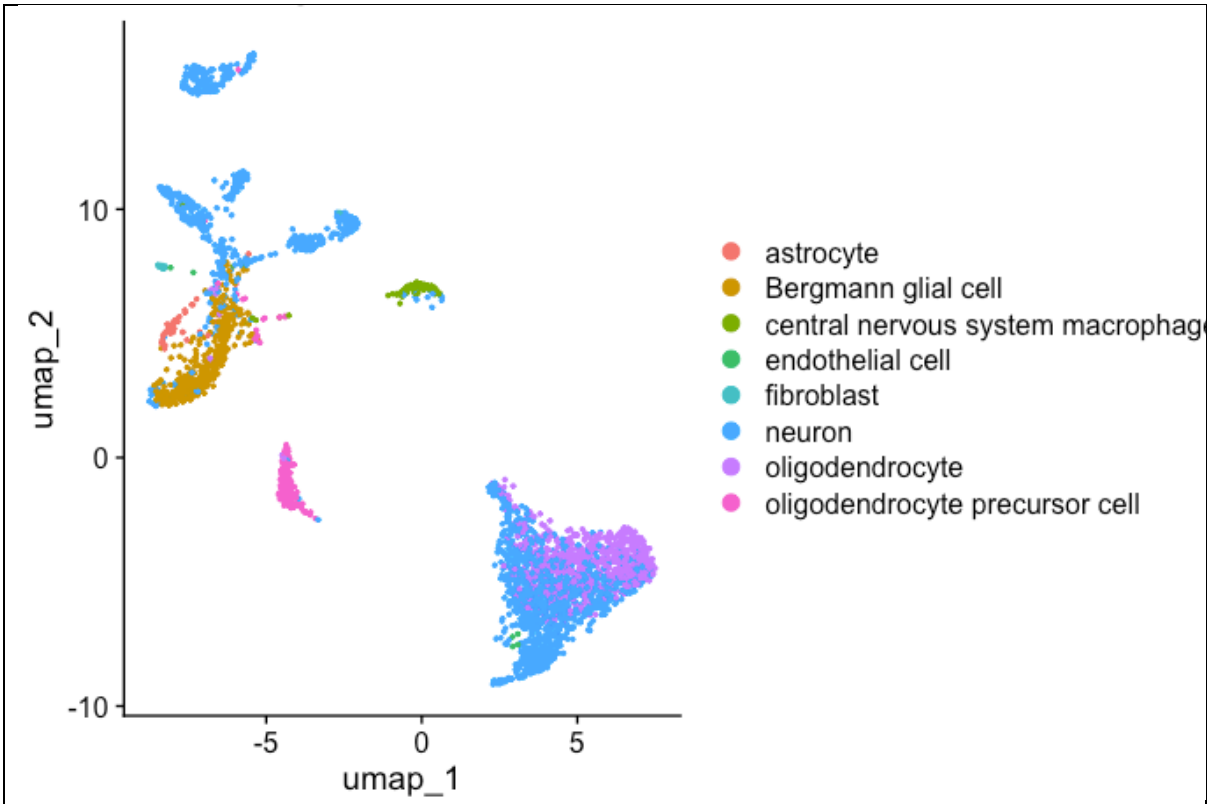


Figure S3: Example UMAP plot of query scATAC-seq data following label transfer

6.5 Results from tests against unrelated tissue

Cell type label	Prediction score average	Count of Cells
adventitial cell	0.2979	18
B cell	0.2145	3
basal cell	0.2928	332
basophil	0.2679	148
blood vessel endothelial cell	0.2000	1
capillary endothelial cell	0.4387	66
CD4-positive, alpha-beta T cell	0.3629	1
CD8-positive, alpha-beta T cell	0.2468	34
classical monocyte	0.3208	15
club cell	0.2133	47
endothelial cell of artery	0.2697	24
fibroblast	0.6779	21
intermediate monocyte	0.2876	2
macrophage	0.3597	1133
mature NK T cell	0.3360	1
non-classical monocyte	0.3489	1
plasma cell	0.2119	4
respiratory goblet cell	0.2342	237
type II pneumocyte	0.4104	1109
vein endothelial cell	0.3115	36
Grand Total	**0.3550**	**3233**

Table S1 – Test results using human lung dataset

6.6 Code library versions used in analysis

Package	Version	Package	Version	Package	Version	Package	Version	Package	Version	Package	Version
abind	1.4-5	datasets	4.3.0	globals	0.16.2	methods	4.3.0	RcppProgress	0.4.2	sp	2.0-0
annotate	1.78.0	DBI	1.1.3	glue	1.6.2	mgcv	1.9-0	RcppRoll	0.3.0	spam	2.9-1
AnnotationDbi	1.62.2	dbplyr	2.3.3	GO.db	3.17.0	mime	0.12	RcppTOML	0.2.2	SparseM	1.81
AnnotationFilter	1.24.0	DDRTree	0.1.5	goftest	1.2-3	miniUI	0.1.1.1	RCurl	1.98-1.12	sparseMatrixStats	1.12.2
askpass	1.1	DelayedArray	0.26.7	googledrive	2.1.1	minqa	1.2.5	readr	2.1.4	sparsesvd	0.2-2
assertthat	0.2.1	DelayedMatrixStats	1.22.5	googlesheets4	1.1.1	MLmetrics	1.1.1	readxl	1.4.3	spatial	7.3-17
backports	1.4.1	deldir	1.0-9	gplots	3.1.3	modelr	0.1.11	rematch	1.0.1	spatstat.data	3.0-1
base	4.3.0	desc	1.4.2	graph	1.78.0	monocle	2.28.0	rematch2	2.1.2	spatstat.explore	3.2-1
base64enc	0.1-3	devtools	2.4.5.9000	graphics	4.3.0	monocle3	1.3.3	remotes	2.4.2.1	spatstat.geom	3.2-4
batchelor	1.15.1	dichromat	2.0-0.1	grDevices	4.3.0	munsell	0.5.0	reprex	2.0.2	spatstat.random	3.1-5
beachmat	2.16.0	diffobj	0.3.5	grid	4.3.0	nlme	3.1-163	reshape2	1.4.4	spatstat.sparse	3.0-2
beeswarm	0.4.0	digest	0.6.33	gridExtra	2.3	nloptr	2.0.3	ResidualMatrix	1.10.0	spatstat.utils	3.0-3
BH	1.81.0-1	DirichletMultinomial	1.42.0	grr	0.9.5	nnet	7.3-19	restfulr	0.0.15	spData	2.3.0
biglm	0.9-2.1	docopt	0.7.1	gtable	0.3.3	numDeriv	2016.8-1.1	reticulate	1.31	spdep	1.2-8
Biobase	2.60.0	dotCall64	1.0-2	gtools	3.9.4	openssl	2.1.0	rhdf5	2.44.0	speedglm	0.3-5
BiocFileCache	2.8.0	downlit	0.4.3	Gviz	1.44.0	parallel	4.3.0	rhdf5filters	1.12.1	splines	4.3.0
BiocGenerics	0.46.0	dplyr	1.1.2	haven	2.5.3	parallelly	1.36.0	Rhdf5lib	1.22.0	stats	4.3.0
BiocIO	1.10.0	dqrng	0.3.0	HDF5Array	1.28.1	patchwork	1.1.3	RhpcBLASctl	0.23-42	stats4	4.3.0
BiocManager	1.30.22	DT	0.28	hdf5r	1.3.8	pbapply	1.7-2	Rhtslib	2.2.0	stringi	1.7.12
BiocNeighbors	1.18.0	dtplyr	1.3.1	here	1.0.1	pbkrtest	0.5.2	rjson	0.2.21	stringr	1.5.0
BiocParallel	1.34.2	e1071	1.7-13	hexbin	1.28.3	pbmcapply	1.5.1	rlang	1.1.1	SummarizedExper	1.30.2
BiocSingular	1.16.0	ellipsis	0.3.2	highr	0.1	pheatmap	1.0.12	rmarkdown	2.24	survival	3.5-7
BiocVersion	3.17.1	EnsDb.Hsapiens.v75	2.99.0	Hmisc	5.1-0	pillar	1.9.0	ROCR	1.0-11	sys	3.4.2
biocViews	1.68.1	EnsDb.Hsapiens.v86	2.99.0	hms	1.1.3	pkgbuild	1.4.2	roxygen2	7.2.3	systemfonts	1.0.4
biomaRt	2.56.1	ensembldb	2.24.0	HSMMSingleCell	1.20.0	pkgconfig	2.0.3	rpart	4.1.19	tcltk	4.3.0
Biostrings	2.68.1	evaluate	0.21	htmlTable	2.4.1	pkgdown	2.0.7	rprojroot	2.0.3	tensor	1.5
biovizBase	1.48.0	fansi	1.0.4	htmltools	0.5.6	pkgload	1.3.2.1	rsample	1.1.1	terra	1.7-39
bit	4.0.5	farver	2.1.1	htmlwidgets	1.6.2	plogr	0.2.0	Rsamtools	2.16.0	testthat	3.1.10
bit64	4.0.5	fastDummies	1.7.3	httpuv	1.6.11	plotly	4.10.2	RSpectra	0.16-1	textshaping	0.3.6
bitops	1.0-7	fastICA	1.2-3	httr	1.4.7	plyr	1.8.8	RSQLite	2.3.1	TFBSTools	1.38.0
blob	1.2.4	fastmap	1.1.1	httr2	0.2.3	png	0.1-8	rstatix	0.7.2	TFMPvalue	0.0.9
boot	1.3-28.1	fastmatch	1.1-3	ica	1.0-3	polyclip	1.10-4	rstudioapi	0.15.0	tibble	3.2.1
brew	1.0-8	filelock	1.0.2	ids	1.0.1	polynom	1.4-1	rsvd	1.0.5	tidyr	1.3.0
brio	1.1.3	fitdistrplus	1.1-11	igraph	1.5.1	poweRlaw	0.70.6	rtracklayer	1.60.0	tidyselect	1.2.0
broom	1.0.5	FNN	1.1.3.2	ini	0.3.1	pracma	2.4.2	Rtsne	0.16	tidyverse	2.0.0
BSgenome	1.68.0	fontawesome	0.5.1	interp	1.1-4	praise	1.0.0	RUnit	0.4.32	timechange	0.2.0
BSgenome.Hsapiens.UCSC.hg19	1.4.3	forcats	1.0.0	IRanges	2.34.1	presto	1.0.0	rversions	2.1.2	tinytex	0.46
BSgenome.Hsapiens.UCSC.hg38	1.4.5	foreign	0.8-84	irlba	2.3.5.1	prettyunits	1.1.1	rvest	1.0.3	tools	4.3.0
bslib	0.5.1	formatR	1.14	isoband	0.2.7	pROC	1.18.4	s2	1.1.4	tzdb	0.4.0
cachem	1.0.8	Formula	1.2-5	JASPAR2020	0.99.10	processx	3.8.2	S4Arrays	1.0.5	units	0.8-3
Cairo	1.6-0	fs	1.6.3	jpeg	0.1-10	profvis	0.3.8	S4Vectors	0.38.1	urlchecker	1.0.1
callr	3.7.3	furrr	0.3.1	jquerylib	0.1.4	progress	1.2.2	sass	0.4.7	usethis	2.2.2
car	3.1-2	futile.logger	1.4.3	jsonlite	1.8.7	progressr	0.14.0	ScaledMatrix	1.8.1	utf8	1.2.3
carData	3.0-5	futile.options	1.0.1	KEGGREST	1.40.0	promises	1.2.1	scales	1.2.1	utils	4.3.0
caTools	1.18.2	future	1.33.0	KernSmooth	2.23-22	ProtGenerics	1.32.0	scattermore	1.2	uuid	1.1-0
cellranger	1.1.0	future.apply	1.11.0	knitr	1.43	proxy	0.4-27	sctransform	0.3.5	uwot	0.1.16
checkmate	2.2.0	gargle	1.5.2	labeling	0.4.2	PRROC	1.3.1	scuttle	1.9.4	VariantAnnotation	1.46.0
cicero	1.18.0	generics	0.1.3	lambda.r	1.2.4	ps	1.7.5	sdamr	0.2.0	vctrs	0.6.3
class	7.3-22	GenomeInfoDb	1.36.1	later	1.3.1	pscl	1.5.5.1	selectr	0.4-2	VGAM	1.1-8
classInt	0.4-9	GenomeInfoDbData	1.2.10	lattice	0.21-8	purrr	1.0.2	seqLogo	1.66.0	vipor	0.4.5
cli	3.6.1	GenomicAlignments	1.36.0	latticeExtra	0.6-30	pwr	1.3-0	sessioninfo	1.2.2	viridis	0.6.4
clipr	0.8.0	GenomicFeatures	1.52.1	lazyeval	0.2.2	qlcMatrix	0.9.7	Seurat	4.9.9.9050	viridisLite	0.4.2
cluster	2.1.4	GenomicRanges	1.52.0	leiden	0.4.3	quantreg	5.96	SeuratData	0.2.2.9001	vroom	1.6.3
CNEr	1.36.0	gert	1.9.3	leidenbase	0.1.25	R.methodsS3	1.8.2	SeuratObject	4.9.9.9091	waldo	0.5.1
codetools	0.2-19	ggbeeswarm	0.7.2	lifecycle	1.0.3	R.oo	1.25.0	SeuratWrappers	0.3.19	warp	0.2.0
colorspace	2.1-0	gghalves	0.1.4	limma	3.56.2	R.utils	2.12.2	sf	1.0-14	whisker	0.4.1
combinat	0.0-8	ggplot2	3.4.3	listenv	0.9.0	R6	2.5.1	shiny	1.7.5	withr	2.5.0
commonmark	1.9.0	ggpubr	0.6.0	lme4	1.1-34	ragg	1.2.5	shinyBS	0.61.1	wk	0.7.3
compiler	4.3.0	ggrastr	1.0.2	lmtest	0.9-40	RANN	2.6.1	shinydashboard	0.7.2	xfun	0.4
conflicted	1.2.0	ggrepel	0.9.3	lubridate	1.9.2	rappdirs	0.3.3	shinyjs	2.1.0	XML	3.99-0.14
corrplot	0.92	ggridges	0.5.4	magrittr	2.0.3	RBGL	1.76.0	Signac	1.11.9000	xml2	1.3.5
cowplot	1.1.1	ggsci	3.0.0	markdown	1.7	rcmdcheck	1.4.0	SingleCellExperiment	1.22.0	xopen	1.0.0
cpp11	0.4.6	ggsignif	0.6.4	MASS	7.3-60	RColorBrewer	1.1-3	sitmo	2.0.2	xtable	1.8-4
crayon	1.5.2	ggstance	0.3.6	Matrix	1.6-1	Rcpp	1.0.11	slam	0.1-50	XVector	0.40.0
credentials	1.3.2	gh	1.4.0	MatrixGenerics	1.12.3	RcppAnnoy	0.0.21	slider	0.3.0	yaml	2.3.7
crosstalk	1.2.0	gitcreds	0.1.2	MatrixModels	0.5-2	RcppArmadillo	0.12.6.1.0	smplot2	0.1.0	zip	2.3.0
curl	5.0.2	glasso	1.11	matrixStats	1.0.0	RcppEigen	0.3.3.9.3	snow	0.4-4	zlibbioc	1.46.0
data.table	1.14.8	glmGamPoi	1.12.2	memoise	2.0.1	RcppHNSW	0.4.1	sourcetools	0.1.7-1	zoo	1.8-12

References

1. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16(1):123-31.

2. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods. 2013;10(12):1213-8.

3. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods. 2008;5(7):621-8.

4. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017;8(1):14049.

5. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods. 2014;11(2):163-6.

6. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature communications. 2017;8(1):14049-.

7. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377-82.

8. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523(7561):486-90.

9. Bartlett DA, Dileep V, Handa T, Ohkawa Y, Kimura H, Henikoff S, et al. High-throughput single-cell epigenomic profiling by targeted insertion of promoters (TIP-seq). J Cell Biol. 2021;220(12).

10. Ku WL, Pan L, Cao Y, Gao W, Zhao K. Profiling single-cell histone modifications using indexing chromatin immunocleavage sequencing. Genome Res. 2021;31(10):1831-42.

11. 10X_Genomics. User Guide - Chromium Next GEM Single Cell Multiome ATAC + Gene Expression. 2022.

12. Zhang S, Cooper-Knock J, Weimer AK, Shi M, Kozhaya L, Unutmaz D, et al. Multiomic analysis reveals cell-type-specific molecular determinants of COVID-19 severity. Cell Syst. 2022;13(8):598-614.e6.

13. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nature Reviews Molecular Cell Biology. 2023;24(10):695-713.

14. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018;36(1):70-80.

15. Kimberly S, Rebecca H, Alejandro Mossi A, Lijuan H, Ka Wai L, Peter L, et al. Transcriptomic diversity of cell types across the adult human brain. bioRxiv. 2022:2022.10.12.511898.

16. 10X_Genomics. 10X Genomics: Cell Ranger support website 2023.

17. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.

18. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.

19. Stuart T, Satija R. Integrative single-cell analysis. Nature Reviews Genetics. 2019;20(5):257-72.

20. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology. 2018;19(1):15.

21. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018.

22. van der Maaten L, Hinton G. Viualizing data using t-SNE. Journal of Machine Learning Research. 2008;9:2579-605.

23. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.

24. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

25. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-72.

26. Lin Y, Cao Y, Kim HJ, Salim A, Speed TP, Lin DM, et al. scClassify: sample size estimation and multiscale classification of cells using single and multiple reference. Mol Syst Biol. 2020;16(6):e9389.

27. Auslander N, Gussow AB, Koonin EV. Incorporating Machine Learning into Established Bioinformatics Frameworks. Int J Mol Sci. 2021;22(6).

28. Dudley JT, Pouliot Y, Chen R, Morgan AA, Butte AJ. Translational bioinformatics in the cloud: an affordable alternative. Genome Medicine. 2010;2(8):51.

29. O'Connell KA, Yosufzai ZB, Campbell RA, Lobb CJ, Engelken HT, Gorrell LM, et al. Accelerating genomic workflows using NVIDIA Parabricks. BMC Bioinformatics. 2023;24(1):221.

30. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467. 2016.

31. Chollet F. Keras: The python deep learning library. Astrophysics source code library. 2018:ascl: 1806.022.

32. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems: Curran Associates Inc.; 2019. p. Article 721.

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.

34. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, et al. Cell type ontologies of the Human Cell Atlas. Nature Cell Biology. 2021;23(11):1129-35.

35. Xie B, Jiang Q, Mora A, Li X. Automatic cell type identification methods for single-cell RNA sequencing. Computational and Structural Biotechnology Journal. 2021;19:5874-87.

36. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nature Neuroscience. 2014;17(9):1156-63.

37. Soverchia L, Ubaldi M, Leonardi-Essmann F, Ciccocioppo R, Hardiman G. Microarrays - The Challenge of Preparing Brain Tissue Samples. Addiction Biology. 2005;10(1):5-13.

38. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955-8.

39. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87.e29.

40. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007;445(7124):168-76.

41. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, Van Der Zwan J, et al. Molecular architecture of the mouse nervous system. Cell. 2018;174(4):999-1014. e22.

42. La Manno G, Siletti K, Furlan A, Gyllborg D, Vinsland E, Mossi Albiach A, et al. Molecular architecture of the developing mouse brain. Nature. 2021;596(7870):92-6.

43. Hao Y, Stuart T, Kowalski M, Choudhary S, Hoffman P, Hartman A, et al. Seurat - R toolkit for single cell genomics 2023 [Available from: https://satijalab.org/seurat/.

44. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Signac -R package for the analysis of single-cell chromatin data 2023 [Available from: https://stuartlab.org/signac/.

45. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology. 2019;20(1):296.

46. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, III, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902.e21.

47. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nature biotechnology. 2015;33(5):495-502.

48. Lee J, Hyeon DY, Hwang D. Single-cell multiomics: technologies and data analysis methods. Experimental & Molecular Medicine. 2020;52(9):1428-42.

49. Rainer J, Gatto L, Weichenberger CX. ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics. 2019;35(17):3151-3.

50. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21.9.1-.9.9.

51. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348(6237):910-4.

52. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nature Methods. 2021;18(11):1333-41.

53. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017;10(1):35.

54. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.

55. 10X_Genomics. 10X Genomics: Sample Datasets 2023 [Available from: https://www.10xgenomics.com/resources/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0.

56. Initiative CZ. CZ CELLxGENE Datasets 2023 [Available from: https://cellxgene.cziscience.com/datasets.

57. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Azimuth - App for reference-based single cell analysis [Available from: https://azimuth.hubmapconsortium.org/references/#Human%20-%20Motor%20Cortex.

58. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598(7879):111-9.

59. initiative CZ. CZ CELLxGENE Discover - Tabula Sapiens 2023 [Available from: https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5.

60. Pliner HA, Packer JS, McFaline-Figueroa JL, Cusanovich DA, Daza RM, Aghamirzaie D, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell. 2018;71(5):858-71.e8.

61. Granja JM, Corces MR, Pierce SE, Bagdatli ST, Choudhry H, Chang HY, et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature Genetics. 2021;53(3):403-11.

62. Kang JB, Nathan A, Weinand K, Zhang F, Millard N, Rumker L, et al. Efficient and precise single-cell reference atlas mapping with Symphony. Nature Communications. 2021;12(1):5890.

63. Lyu P, Zhai Y, Li T, Qian J. CellAnn: a comprehensive, super-fast, and user-friendly single-cell annotation web server. Bioinformatics. 2023;39(9).

64. Wang Y, Sun X, Zhao H. Benchmarking automated cell type annotation tools for single-cell ATAC-seq data. Frontiers in Genetics. 2022;13.

65. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. bioRxiv. 2020:2020.03.31.016972.