Cobolt: Joint analysis of multimodal single-cell sequencing data

Boying Gong[1], Yun Zhou[1], and Elizabeth Purdom[2,*]

[1]Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA
[*]Corresponding author. Email: epurdom@stat.berkeley.edu

July 14, 2021

Abstract

A growing number of single-cell sequencing platforms enable joint profiling of multiple omics from the same cells. We present Cobolt, a novel method that not only allows for analyzing the data from joint-modality platforms, but provides a coherent framework for the integration of multiple datasets measured on different modalities. We demonstrate its performance on multi-modality data of gene expression and chromatin accessibility and illustrate the integration abilities of Cobolt by jointly analyzing this multi-modality data with single-cell RNA-seq and ATAC-seq datasets.

Keywords: Single cell, Multi-omics, Integration

1 Introduction

Single-cell sequencing allows for quantifying molecular traits at the single-cell level, and there exist a wide variety of platforms that extend traditional bulk platforms, such as mRNA-seq and ATAC-seq, to the single cell. Comparison of different cellular features or *modalities* from cells from the same biological system gives the potential for a holistic understanding of the system. Most single-cell technologies require different cells as input to the platform, and therefore there remains the challenge of linking together the biological signal from the different modalities, with several computational methods proposed to estimate the linkage between the different modalities, such as LIGER [Welch et al., 2019] and Signac (Seurat) [Stuart et al., 2020, 2019].

Recently, there are a growing number of platforms that allow for measuring several modalities on a single cell. CITE-seq [Stoeckius et al., 2017] jointly sequences epitope and transcriptome; scNMT-seq [Clark et al., 2018] jointly profiles chromatin accessibility, DNA methylation, and gene expression; and sci-CAR [Cao et al., 2018], Paired-seq [Zhu et al., 2019], and SNARE-seq [Chen et al., 2019] enable simultaneous measurement of transcription and chromatin accessibility (we direct readers to Lee et al. [2020] for a comprehensive review). By directly measuring the different modalities on the same cells, these techniques greatly enhance the ability to relate the different modalities. With the emergence of joint platforms, new computational methodologies for analyzing multi-modality data have also been developed. Early methods mainly focused on CITE-seq [Gayoso et al., 2021, Wang et al., 2020, Hao et al., 2020], which jointly sequences gene expressions and at most a few hundred antibodies. Recently, more methods have been proposed to enable the modeling of cells with simultaneous measurement of gene expression and higher-dimensional modalities such as chromatin accessibility, such as MOFA+ (also known as MOFA2, Argelaguet et al. [2020]), BABEL [Wu et al., 2021], and scMM [Minoura et al., 2021].

However, single-cell datasets on a single modality are far more common, and are usually of higher throughput. Indeed it is natural that joint-modality data from the same system will be used to augment single-modality data, or vice versa. Therefore, there is a critical need for an analysis tool that is both a stand-alone application for multi-modality data as well as a tool for integration of these datasets with single-modality platforms. BABEL [Wu et al., 2021] and scMM [Minoura et al., 2021], while not directly targeting this task, do allow the use of the joint-modality data to predict one single-modality dataset into another type of

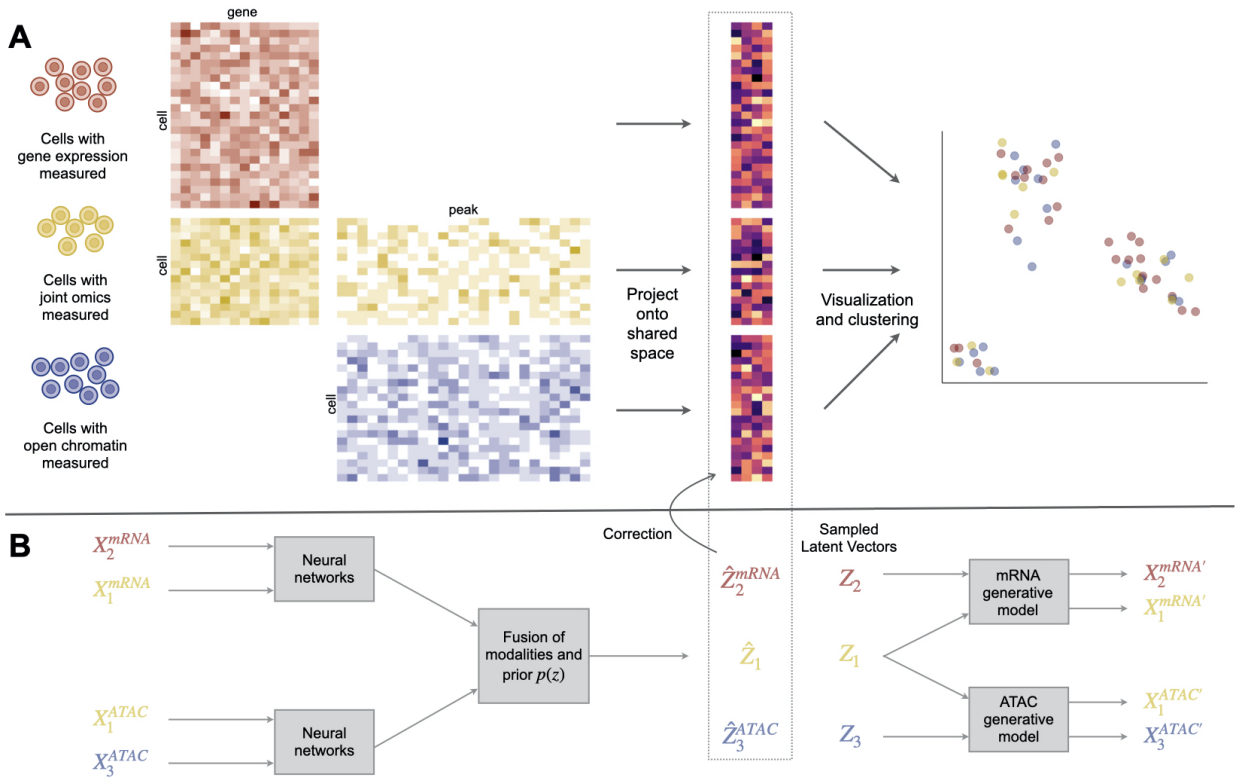

Figure 1: An overview of the `Cobolt` method. The upper panel shows the workflow of `Cobolt`, which takes as input datasets with varying modalities, projects the data into shared latent space, and then performs visualization and clustering with all datasets combined. The lower panel shows the `Cobolt` variational autoencoder model with encoders plotted on the left and decoders on the right.

modality. However, neither directly integrate the data together to allow for downstream analysis of the joint set of data regardless of modality, such as cell subtype detection.

Our method `Cobolt` (https://github.com/boyinggong/cobolt) fills this gap by providing a coherent framework for a full integrative analysis of multi-modality and single-modality platforms. The result of `Cobolt` is a single representation of the cells irrespective of modalities, which can then be used directly by downstream analyses, such as joint clustering of cells across modalities. `Cobolt` estimates this joint representation via a novel application of Multimodal Variational Autoencoder (MVAE) [Wu and Goodman, 2018] to a hierarchical generative model. The integration of the single-modalities is done by a transfer learning approach which harnesses the valuable information found by joint sequencing of the same cells and extends it to the cells in the single-cell platform. The end result is a single representation of all of the input cells, whether sequenced on a multi-modality platform or a single-modality platform. In this context, `Cobolt` gives an over-all integrative framework that is flexible for a wide range of modalities.

We demonstrate `Cobolt` on two use-cases. The first uses `Cobolt` to analyze only a multi-modality sequencing dataset from the SNARE-seq technology; we show that `Cobolt` provides a joint analysis that better distinguishes important facets of each modality, compared to existing methods. The second example demonstrates the use of `Cobolt` to integrate multi-modality data with single-modality data collected from related biological systems, where `Cobolt` creates a joint representation that can be used for downstream analysis to provide meaningful biological insights. We show that `Cobolt` also performs better than related tools in this integrative task.

2 Results

2.1 The Cobolt Model

We develop a novel method, Cobolt, that utilizes joint-modality data to enable the joint analysis of cells sequenced on separate sequencing modalities. We do this by developing a Multimodal Variational Autoencoder based on a hierarchical bayesian generative model. We briefly describe the premise of the model using the example of two modalities: mRNA-seq and ATAC-seq (for more details in greater generality, see the Methods Section 4). We assume that we have a set of cells with both mRNA-seq and ATAC-seq data collected using the joint modality platform (X_1^{mRNA} and X_1^{ATAC}), as well as (optionally) a set of cells with only mRNA-seq data (X_2^{mRNA}), and a set of cells with only ATAC-seq data (X_3^{ATAC}). Cobolt takes all of this data as input to find a representation of the cells in a shared reduced dimensionality space, regardless of modality (Figure 1A).

Cobolt models the sequence counts from the modalities inspired by Latent Dirichlet Allocation (LDA) [Blei et al., 2003] model. The LDA model is a popular Bayesian model for count data which has been successfully applied to genomics in areas such as clustering of single-cell RNA-seq data [Yotsukura et al., 2016, Sun et al., 2018], single-cell ATAC-seq analysis [González-Blas et al., 2019], and functional annotation [Backenroth et al., 2018]. Cobolt builds a hierarchical latent model to model data from different modalities and then adapts the MVAE approach to both estimate the model and allow for a transfer of learning between the joint-modality data and the single-modality data.

Cobolt assumes that there are K different types of possible categories that make up the workings of a cell. For ease of understanding, it is useful to think of these categories as biological processes of the cell, though the categories are unlikely to actually have a one-to-one mapping with biological processes. Each category will result in different distributions of features in each of the modalities – i.e., different regions of open chromatin in ATAC-seq or expression levels of genes in mRNA-seq for different categories. The features measured in a cell are then the cumulative contribution of the degree of activation of each category present in that cell. The activation level of each category is represented by the latent variable θ_c for each cell c, which gives the relative activity of each of the K categories in the cell. θ_c is assumed to be an intrinsic property of each cell representing the underlying biological properties of the cell, while the differences of data observed in each modality for the same cell are due to the fact that the categories active in a cell have different impacts in the modality measured (open chromatin in ATAC-seq versus gene expression in mRNA-seq). We assume $\theta_c = \sigma(z_c)$, where z_c is drawn from a Gaussian prior and σ is the soft-max transformation; this is an approximation to the standard Dirichlet prior for θ_c that allows use of variational autoencoders to fit the model [Srivastava and Sutton, 2017]. The mean of the posterior distribution gives us an estimate of our latent variable z_c for each cell, and the posterior distribution is estimated using variational autoencoders (VAE).

In the end, Cobolt results in an estimate of the latent variable z_c for each cell, which is a vector that lies in a K-dimensional space. This space represents the shared biological signal of the individual cells, irregardless of modality, and can be used for the common analysis tasks of single-cell data, such as visualization and clustering of cells to find subtypes. Importantly, we can predict the latent variable z_c even when a cell does not have all modalities measured. Moreover, because of the joint-modality platforms, Cobolt does not require that the different modalities measure the same features in order to link the modalities together – the fact that some of the cells were sequenced on both platforms provides the link between different types of features. Therefore, ATAC-seq peaks and mRNA-seq gene expression can be directly provided as input. This is unlike methods that do not make use of the joint-modality data and require that the different modalities be summarized on the same set of features, for example by simplifying ATAC-seq peaks to a single measurement per gene.

2.2 Cobolt as an analysis tool for multi-modality data

While the full power of Cobolt is to integrate together data from single and multiple modality datasets, in its simplest form Cobolt can be used for the analysis of data from solely a multi-modality technology. We demonstrate this usage with the SNARE-seq data of Chen et al. [2019], which consists of paired transcription and chromatin accessibility sequenced on 10,309 cells of adult mouse cerebral cortices.

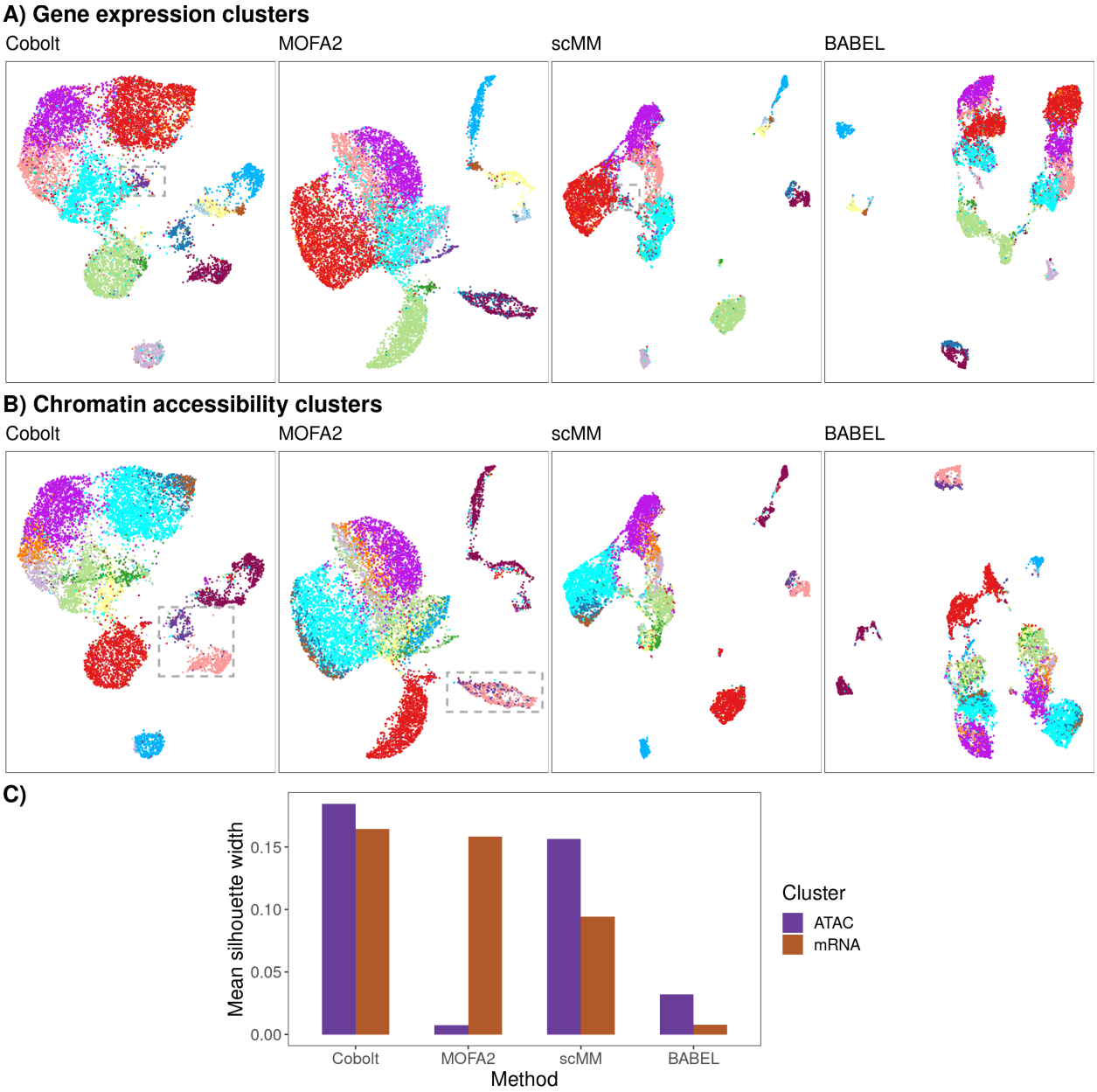

Figure 2: **Comparison of multi-modality analysis methods.** A, B) UMAP visualizations of the reduced dimensionality space created by `Cobolt`, MOFA2, scMM, and BABEL. The cells are color-coded by the cluster they are assigned to based on clustering of A) only gene expression modality and B) only chromatin accessibility modality. We note that the cluster colors are randomly and separately assigned for panel A and B. Highlighted in the panels are clusters that are well separated in the analysis of `Cobolt`, but not the other methods. C) Mean silhouette width of the gene expression clusters and chromatin accessibility clusters evaluated on the UMAP representations generated by `Cobolt`, MOFA2, scMM, and BABEL. More details on the silhouettes per cluster can be found in Figure S3.

We first compare Cobolt with a simple, but common, approach for analyzing joint-modality data: the two modalities are analyzed separately and then the results are linked together. This is the strategy of Chen et al. [2019], where the authors primarily clustered the gene expression modality to form clusters of the cells, and then performed a separate analysis on chromatin accessibility modality as a comparison. Focusing on the gene expression modality is common, since it is often assumed to have the greatest resolution in determining cell-types. However, the reduced representations and clusters created on one modality may not be representative of all the underlying cell subtypes. Indeed, when we perform clustering analysis on only the gene expression modality using Seurat [Stuart et al., 2019] and only the chromatin accessibility modality using cisTopic [González-Blas et al., 2019] (consistent with Chen et al. [2019], see Section 4.3 for details), both modalities find distinct clusters that are not reflected in the other modality. A joint analysis with `Cobolt`, on the other hand, finds the subtypes that are not detected by only one modality or the other (Figure S2).

Next we compare with other methods that explicitly analyze the two modalities jointly, like `Cobolt`. We consider the methods mentioned in the introduction: MOFA2, scMM, and BABEL. MOFA2 uses Bayesian group factor analysis for dimensionality reduction of multi-modality datasets; BABEL trains an interoperable neural network model on the paired data that translates data from one modality to the other; and scMM [Minoura et al., 2021] uses a deep generative model for joint representation learning and cross-modal generation. We apply each of these methods to the SNARE-seq data.

Since a joint analysis method should be able to reflect subtype signals captured by all modalities, we similarly evaluate the methods on how well their lower-dimensional spaces represent separate clusters identified separately in each modality, as described above. In Figure 2, we visualize the lower-dimensional space generated by `Cobolt`, MOFA2, scMM, and BABEL via UMAP (Uniform Manifold Approximation and Projection, McInnes et al. [2018]), where we color the cells based on the clusters found by clustering the gene expression modality data (Figure 2A) and those based on clustering the chromatin accessibility data (Figure 2B). We would note that BABEL does not create a single reduced-dimensionality representation for a paired cell, but rather one per modality (the two latent representations are learned jointly and are quite similar). BABEL's lower-dimensionality representation does quite poorly, separating major clusters of cells found in both modalities, such as layer 2 to 6 intratelencephalic (IT) neurons (colored red, purple, pink, and cyan in Figure 2A). Both MOFA2 and scMM capture the large clusters, such as glutamatergic excitatory neuron subtypes, which are shared between the modalities. However, we see clusters specific to a single modality not reflected on their lower-dimensional space. For example, the highlighted gene expression cluster in Figure 2A is practicably indistinguishable in the scMM UMAP but separated in the `Cobolt` UMAP. Similarly, in Figure 2B, we highlight clusters of cells which clearly separate based on chromatin profile and in the `Cobolt` analysis but are mixed with other cells in the MOFA2 analysis. To quantitatively evaluate these observations, we calculate the average silhouette widths of the modality-specific clustering on the UMAPs generated by `Cobolt`, MOFA2, scMM, and BABEL (shown in Figure 2C), where higher silhouette widths indicate that cells are closer to other cells in the same cluster. As expected from our observations, BABEL's representation results in extremely small silhouette widths, reflecting the many clusters separated in BABEL's representation. MOFA2 has the smallest silhouette width on chromatin accessibility clusters, supporting our observation that its joint space does not represent this modality well; similarly, scMM gives relatively small measures on the gene expression modality. `Cobolt` best represents both modalities with the highest silhouette width measure.

2.3 `Cobolt` for integrating multi-modality data with single-modality data

We now turn to integrating multi-modality data with single-modality data. For this use-case, we use `Cobolt` to jointly model three different datasets — the SNARE-seq of mouse cerebral cortices analyzed in the above section, together with a scRNA-seq and a scATAC-seq dataset of mouse primary motor cortex (MOp) [Yao et al., 2020]. After obtaining the lower-dimensional representation ($K = 30$), we applied the Louvain community detection algorithm [Blondel et al., 2008] for cell clustering (see details in Section 4.3).

The result is a lower-dimensional latent space that aligns the three datasets into a single representation. In Figure 3A, we visualize this low-dimensional space via UMAP, with cells colored by their dataset of origin. We see that the cells from different datasets are well aligned regardless of their source of origin for the most part. In Figure 3B, we color the MOp cells by their cellular subtype as defined by Yao et al. [2020]. For

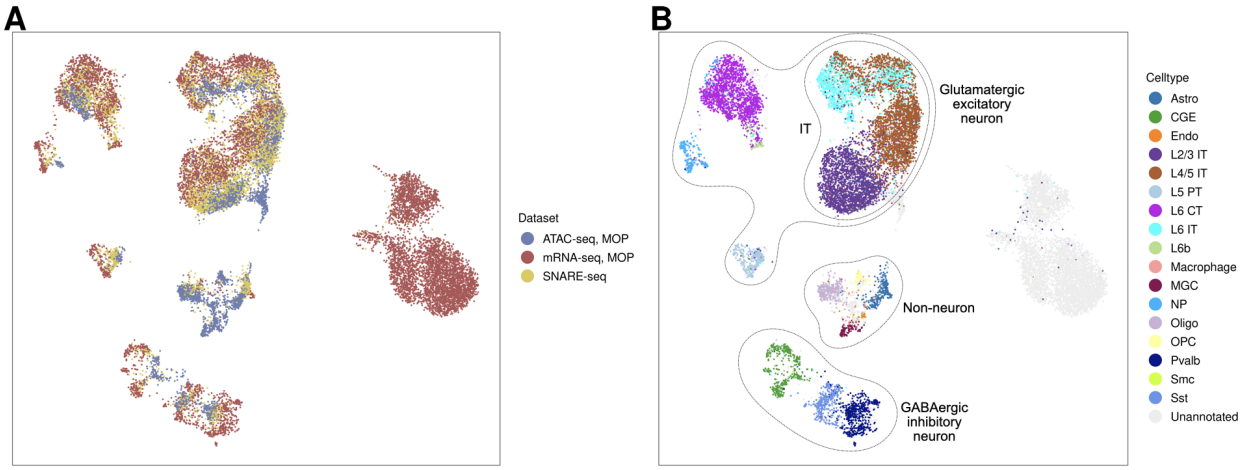

Figure 3: The UMAP visualization of integrating SNARE-seq and MOp datasets. Cells are colored by A) dataset of origin and B) cell type annotation. Both the scRNA-seq and the scATAC-seq contain a substantial fraction of cells labeled "unannotated" by the authors of the data and that do not map to known cell types. The cell type abbreviation largely followed Yao et al. [2020]: astrocytes (Astro); caudal ganglionic eminence interneurons (CGE); endothelial cells (Endo); layer 2 to layer 6 (L2-6); intratelencephalic neurons (IT); pyramidal tracts (PT); corticothalamic neurons (CT); L6b excitatory neurons (L6b); microglial cells (MGC); near-projecting excitatory neurons (NP); oligodendrocytes (Oligo); oligodendrocyte precursors (OPC); smooth muscle cells (SMC); medial ganglionic eminence interneurons subclasses based on marker genes (Sst, Pvalb).

the purposes of comparison across the modalities, we integrated some cell-types into larger groupings and modified the names so as to have comparable groups (see Section 4.3). For the SNARE-seq cells, we do not have the cell types from Chen et al. [2019], so we use the identifications found by analyzing only the SNARE-seq cells (see Section 4.3). We see that cells from the same cellular subtypes are projected closely regardless of the data source. We also see that the representation of `Cobolt` respects the larger category of cell types by grouping three major cell classes: GABAergic inhibitory neurons (CGE, Sst, Pvalb), glutamatergic excitatory neurons (IT, L5 PT, L6 CT, L6b, NP), and non-neurons.

Despite detecting mostly similar cell types, the MOp datasets profile several cell types distinct to the modality. For example, microglial cells (MGC) and smooth muscle cells (SMC) are uniquely detected in scATAC-seq. The different datasets also have different cellular compositions of their shared subtypes, where astrocytes (Astro) and oligodendrocytes (Oligo) are much more abundant in the scATAC-seq (6.55% and 10.50%) than in the SNARE-seq (4.5% and 2.84%) and scRNA-seq (0.40% and 0.36%). As shown in Figure 3B, cell populations unique to one dataset are grouped in the UMAP plot and are distinguishable from the other datasets/cell types, indicating that `Cobolt` reconciles data even when one cell population is entirely absent or scarcely represented in one or more data sources.

`Cobolt` also facilitates subtype identification at a finer resolution by transferring information between modalities. Figure 4 shows a breakdown of one cell type, caudal ganglionic eminence interneurons (CGE). In the scATAC-seq MOp dataset, this was annotated as one cluster, while the scRNA-seq annotation further divided CGE cells into 3 subtypes based on marker genes –Lamp5, Vip, and Sncg. Our joint mapping of the cells relates the subtypes detected in scRNA-seq to scATAC-seq and provides a finer resolution breakdown of CGE in scATAC-seq. As shown in Figure 4A, our clustering of the results of the joint analysis of `Cobolt` resulted in a cluster (cluster 13) which composed of Lamp 5 and Sncg, while another (cluster 16) is mostly Vip cells. This subdivision is further validated by gene expression and gene activity levels of marker genes *Lamp5* and *Vip* in these clusters, which discriminate subtype Lamp5 from subtype Vip (Figure 4B). This shows that the joint model of `Cobolt` can help distinguish noisy cells in one dataset with additional information from other datasets or modalities.

Furthermore, `Cobolt` is robust to poor-quality cells and low-expressed genes by using a count-based model, which should naturally down-weight the influence of low-count cells. In the above analysis, we followed Yao et al. [2020] and filtered out 2.5% of MOp mRNA cells and 15.8% ATAC-seq cells due to low

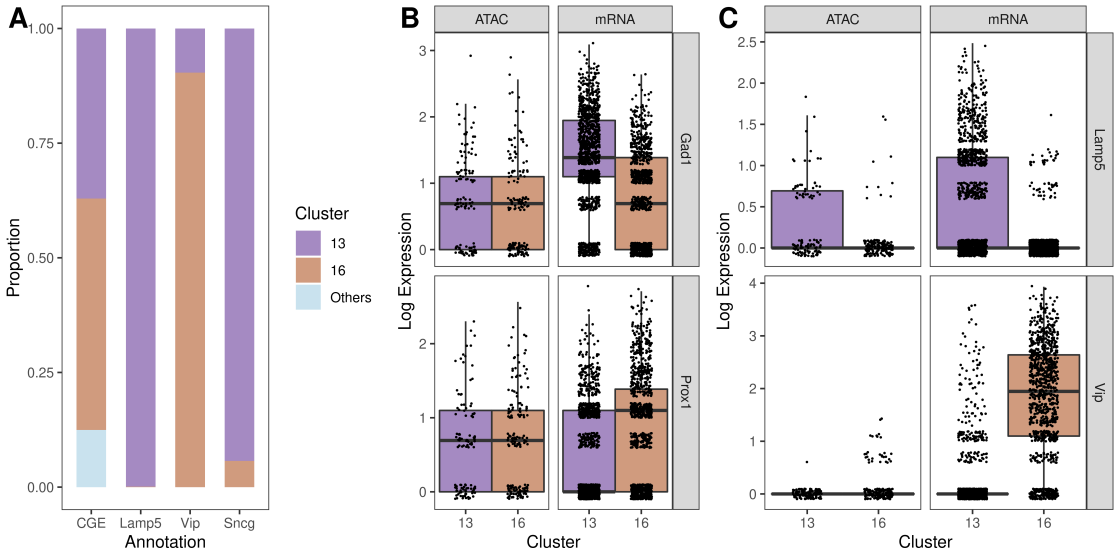

Figure 4: A) Composition of clusters found by clustering reduced dimensionality of `Cobolt`. We show cells annotated as CGE by the scATAC-seq dataset compared to the cells annotated in the subtypes Lamp5, Vip, and Sncg by the scRNA-seq dataset. (B-C) Plots of the gene expression (mRNA) and gene body assessability summaries (ATAC) in clusters 13 and 16 for the marker genes (B) *Gad1, Prox1*, and (C) *Lamp5, Vip*. *Gad1, Prox1* are marker genes for the larger CGE cluster that contains Lamp5, Vip, and Sng, while (*Lamp5, Vip*) distinguish between cell types Lamp5 and Vip [Yao et al., 2020].

counts and other quality control measures. Yet `Cobolt` is robust to these choices, even in the extreme case where there is no filtering on cells and only a minimal filter on genes is performed (Supplementary Figure S9).

2.3.1 Comparison with existing methods

As described in the introduction, there are few existing methods that allow full integration of multi-modality sequencing data with single-modality data. MOFA2 only analyzes joint-modality data; BABEL and scMM train a joint model on the paired joint-modality data, and allow the user to apply this model to single-modality datasets to predict the other "missing" modality. Unlike `Cobolt`, these two methods do not use the single-modality data in the training of the model, nor do they provide a representation of the single-modality data in a single shared representation space regardless of modality, for example, for the purposes of joint clustering of all of the cells across modalities. Therefore, to have additional points of comparison, we apply the LIGER and Signac methods, which are designed for integrating *unpaired* modalities (the implementation details can be found in Section 4.3). LIGER applies an integrative nonnegative matrix factorization (iNMF) approach to project the data onto a lower-dimensional space and then builds a shared nearest neighbor graph for joint clustering. Signac implements canonical correlation analysis (CCA) for dimensionality reduction; Signac subsequently transfers cell labels by identifying mutual nearest neighbor cell pairs across modalities.

To evaluate the performance of these methods when integrating single-modality data with multiple-modality data, we return to the SNARE-seq data and use it to create an artificial set of multiple-modality data and single-modality data. We do this by randomly assign cells the cells to either the paired dataset (20%) or the single-modality datasets (80%) and run each of the methods. In this way, we have a ground truth on how the single-modality data should be connected to the paired data and can compare the methods by evaluating whether coordinates of the reduced dimensions Z for pairs of cells assigned to the single-modality datasets were close together. Specifically, for each cell in the mRNA single-modality set we find its coordinates $\hat{Z}^{mRNA}$ and for a fixed number k locate its k nearest neighbors in the ATAC single-modality set based on the coordinates $\hat{Z}^{ATAC}$. We then calculate the percentage of mRNA single-modality cells whose paired cell in the ATAC single-modality is included in its set of nearest neighbors. The reverse analysis was done using chromatin accessibility as the query and evaluating the percentage whose nearest neighbors

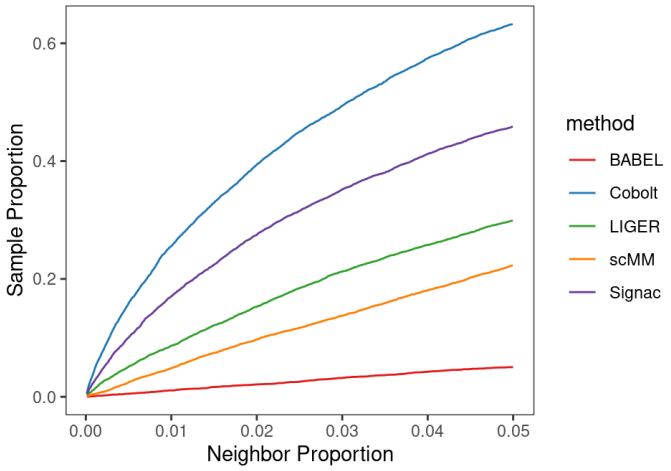

Figure 5: **Evaluation of dimensionality reduction.** A comparison of lower-dimensional representation generated by LIGER, Signac, scMM, BABEL, and `Cobolt` on the SNARE-seq data. The x-axis shows the number of neighbors considered (k) as a proportion of the total testing sample size. The y-axis shows the proportion of cells whose paired data are within their k-nearest neighbors in the other modality. The plot gives the results when using gene expression data as the query. We observe very similar results when chromatin accessibility is used as the query (Supplementary Figure S4).

include their mRNA pair. For LIGER and Signac, which are not designed for multiple-modality data, we hide the pairing information on all cells, and treat all of the cells as if they were collected on different cells. The choice of 20% and 80% were based on the relative size of the SNARE-Seq joint-modality data to the individual MOp scATAC-Seq data and scRNA-Seq datasets, and reflects the fact that single-modality datasets are much higher-throughput than paired-modality data. Many popular clustering routines use nearest-neighbor graphs for identifying clusters, so this is a metric directly related to whether the cells assigned to the single-modality data would likely correctly cluster together across modalities, but avoids having to specify cluster parameters, especially as applied to different methods (and LIGER and Signac have their own clustering techniques specific for this data).

As shown in Figure 5, the `Cobolt` joint representation does a much better job of assigning coordinates to the single-modality cells that are close to their pair. The proportion of single-modality cells that are neighbors to their pair is much larger than any of these other methods. Surprisingly, the other methods that make use of the joint-modality data to develop their model (scMM and BABEL) do much worse even than Signac and LIGER, which do not have any information linking the cells together. This points to the power of truly integrating the high-throughput single-modality data into the analysis. We see improved performance of scMM and BABEL when we increase the proportion of dual-modality cells to 80% and only 20% of cells being single-modality (Supplementary Figure S5), but still well below the performance of `Cobolt`.

Subtype detection is a critical aspect of single-cell data analysis, and integrating single-modality data with multiple-modality data gives the potential for higher resolution detection. Our nearest-neighbor metric is directly related to subtype detection, with a higher neighbor proportion corresponding roughly to finding larger clusters in a clustering algorithm. scMM and BABEL only provide cross-modality predictions, rather than a joint embedding of the single-modality data with the multiple-modality data. Downstream tasks such as clustering for subtype detection must be done on either the mRNA expression space or the chromatin expression space, and as we have seen in Section 2.2 each of which can miss important features of the data. Indeed this may be an important factor in their low performance on our nearest-neighbor analysis. LIGER and Signac do provide a joint embedding of all of the data, but do so without the use of the pairing information for the joint-modality data for training the embedding.

The previous analysis provides a comparison with a known ground-truth and an evaluation at different levels of resolution. Now we return to the joint embedding of MOp scRNA-Seq and scATAC-Seq with the SNARE-Seq data that we considered in Section 2.3 and consider the performance of LIGER and Signac, which performed comparatively well in our earlier comparison. We would note that LIGER and Signac

have their own clustering strategies, separate from their dimensionality reduction, but we focus here on the results of their dimensionality reduction. We provided LIGER and Signac only the MOp data, as there is no clear way of including the SNARE-seq into LIGER and Signac without focusing on one of its modalities and adding extra batch correction steps. We compare the results to the clusters published with the scRNA-seq and scATAC-seq data in Yao et al. [2020] (Figure S7). We see that `Cobolt` generates a UMAP visualization that well represents rare subpopulations and respects broader cell classes. LIGER gives greater separation between cell types but splits several subtypes into far-away islands, such as for L5 PT, MGC, and Astro. Signac adopts an asymmetric strategy of transferring scRNA-seq labels to scATAC-seq data, and as a result, Signac performs well on major cell types but poorly on under-represented subpopulations in scRNA-seq such as astrocytes (Astro), which accounts for only 0.4% of the cells in scRNA-seq but 6.55% in scATAC-seq. Further, `Cobolt`, unlike LIGER and Signac, not only groups together the subtypes, but appears to also represent the three broader categories major GABAergic inhibitory neurons, glutamatergic excitatory neurons, and non-neurons.

These cluster identifications that we highlight from Yao et al. [2020] are relatively robust, well-known cell types, representing the large structural changes in the data, for which we expect most strategies to be able to detect reasonably well. On the other hand, our nearest-neighbor analysis emphasizes the performance at a high level of resolution. Putting both of these together points to the fact that `Cobolt` provides a superior integration of the datasets across a wide spectrum of resolutions.

3 Discussion

We have shown that `Cobolt` is a flexible tool for analyzing multi-modality sequencing data, whether separately or integrated with single-modality data. `Cobolt` synthesizes the varied data into a single representation, preserving meaningful biological signal in the different modalities and at different resolutions. Moreover, this latent variable space is appropriate for standard downstream analysis techniques commonly used for analyzing cells without any further specialized adjustments, allowing `Cobolt` to fit into standard analysis pipelines.

In this paper, we focused on reconciling cells from different sources and provide a consensus clustering. Future work could make use of the graphical model and inferred parameters to establish connections between features. For example, the probability vectors generated by $B^{(i)}$ naturally provide a reduced-dimensional space of molecular features and can potentially help the construction of gene networks.

We would note that while we have focused on the capability of `Cobolt` to analyze data from two-modalities, the underlying method can be extended to larger number of modalities and integration of different combinations of modalities, such as datasets with different pairs of modalities. Thus `Cobolt` provides a framework to integrate a wide range of varieties of multi-modality platforms as well as single-modality platforms.

4 Methods

While the most common application of joint-modality platforms consists of pairs of modalities (such as the example of mRNA-seq and ATAC-seq we described above), we will describe `Cobolt` in generality, assuming that there are M modalities.

4.1 Modeling Modality Dependency

For an individual cell c, we can (potentially) observe M vectors of data $x_c = \{x_c^{(1)}, \cdots, x_c^{(M)}\}$, each vector of dimension $d_1, \ldots, d_M$ corresponding to the number of features of each modality. We assume a Bayesian latent model, such that for each cell there is a latent variable $z_c \in \mathbb{R}^K$ representing the biological signal of the cell, where z_c is assumed drawn from a Gaussian prior distribution. Given z_c, we assume that the data $x_c^{(m)}$ for each modality has an independent generative process, potentially different for each modality. Specifically, we assume that the data $x_c^{(m)}$ from each modality are conditionally independent given the common latent

variable z_c. That is,

$$p(x_c^{(1)}, \cdots, x_c^{(M)}, z_c) = p(z_c) \prod_{i=1}^{M} p(x_c^{(i)}|z_c).$$

We use $q(z_c|x_c^{(1)}, \cdots, x_c^{(M)})$ as a variational approximation of the posterior distribution $p(z_c|x_c^{(1)}, \cdots, x_c^{(M)})$. $q(z_c|x_c^{(1)}, \cdots, x_c^{(M)})$, the encoder, is assumed Gaussian with parameters modeled as neural networks (i.e. Variational Autoencoder, VAE [Kingma and Welling, 2013]). This allows for estimation of the posterior distribution $p(z_c|x_c^{(1)}, \cdots, x_c^{(M)})$ and the underlying latent variable for each cell c. The posterior mean of this distribution $\hat{z}_c$ will be our summary of the shared representation across modalities.

Importantly, this model can be estimated even when not all of the input data contains all modalities. In this case, an individual cell c contains a subset of the modalities, $\mathcal{S}_c \subset \{1, \ldots, M\}$, and consists of data $x_c = \{x_c^{(i)}, i \in \mathcal{S}_c\}$. Without all modalities observed, the cell can contribute to the estimation of the model as its distribution can be explicitly written out:

$$p(x_c, z_c) = p(z_c) \prod_{i \in \mathcal{S}_c} p(x_c^{(i)}|z_c).$$

Further, we can estimate latent variables for such cells by using posterior distribution of z_c when conditioning only on the observed modalities, $q(z_c|\{x_c^{(i)}, i \in \mathcal{S}_c\})$. Instead of using separate neural networks for $2^M - 1$ posterior distributions of different modality combinations, we adopt a technique introduced in Multimodal Variational Autoencoder (MVAE) [Wu and Goodman, 2018], which largely reduces the number of decoders to $2M$ (See Supplementary Material S1 for inference details).

As an example, if there are two modalities, mRNA-seq and ATAC-seq, and we have n_1 cells with paired data from the joint modality platform, $X_1 = (X_1^{mRNA}, X_1^{ATAC})$; n_2 cells with only mRNA measured, $X_2 = X_2^{mRNA}$; and n_3 cells with only ATAC-seq measured, $X_3 = X_3^{ATAC}$. All $N = n_1 + n_2 + n_3$ cells can be used in the estimation of the joint distribution of the latent variables, and estimates of the latent variables can be found as the mean of the relevant approximate posterior distributions:

$$\begin{aligned} \hat{Z}_1 &= E(Z|X_1^{mRNA}, X_1^{ATAC}) && \text{(Paired Cells)} \\ \hat{Z}_2^{mRNA} &= E(Z|X_2^{mRNA}) && \text{(mRNA-seq only)} \\ \hat{Z}_3^{ATAC} &= E(Z|X_3^{ATAC}) && \text{(ATAC-seq only)} \end{aligned}$$

Correcting for missing modalities In practice we find that the distributions $q_\phi(z_c|\{x_c^{(i)}, i \in \mathcal{S}\})$ have subtle differences for different subsets $\mathcal{S}$, i.e. the latent variables $\hat{Z}_1$, $\hat{Z}_2^{mRNA}$, and $\hat{Z}_3^{ATAC}$ show distinct separations (Supplementary Figure S6). One possibility could be due to platform differences between the different datasets that remain even after our batch correction (see Section 4.2). However, we also see differences in these distributions even if we only consider the joint-modality data, where we can estimate all of these posterior distributions on the same cells, $\hat{Z}_1^{mRNA} = E(Z|X_1^{mRNA})$ or $\hat{Z}_1^{ATAC} = E(Z|X_1^{ATAC})$ (Supplementary Figure S6). Indeed there is nothing in the optimization of the posterior distribution that requires these different posterior distributions to be the same.

While the effects are small, these subtle differences can affect downstream analyses, e.g., in clustering cells for subtype discovery. Rather than directly force these posterior distributions to match in our estimation of the model, `Cobolt` fits the model as described above (using all of the data), then uses the paired data to train a prediction models that predict $\hat{Z}_1$ from the modality-specific estimates $\hat{Z}_1^{mRNA}$ and $\hat{Z}_1^{ATAC}$. We then apply these prediction models to $\hat{Z}_2^{mRNA}$ and $\hat{Z}_3^{ATAC}$ to obtain estimates $\hat{Z}_2$ and $\hat{Z}_3$ which are better aligned to be jointly analyzed in the same space. In practice, we find XGBoost [Chen and Guestrin, 2016] and k-nearest neighbors algorithm work equally well. We present results based on XGBoost (See Section 4.3 for details).

4.2 Modeling Single Modality of Sparse Counts

The choice of the generative distribution $p_\psi(x^{(i)}|z)$ should be chosen to reflect the data, and in principle can vary from modality to modality. For example, single-modality VAE models using zero-inflated negative

binomial distributions (ZINB) [Lopez et al., 2018] have been proposed for scRNA-seq datasets to account for sparse count data. However, we found ZINB models performed less well for technologies that measure modalities like chromatin accessibility, which results in data with sparser counts and larger feature sizes than scRNA-seq. Therefore, we develop a latent model for these types of modalities inspired by the Latent Dirichlet Allocation (LDA) [Blei et al., 2003].

Our generative model for a single modality i starts by assuming that the counts measured on a cell are the mixture of the counts from different latent categories. In the genomic setting, these categories could correspond to biological processes of the cell, for example. Each category has a corresponding distribution of feature counts. The cumulative feature counts for a cell c are then the result of combining the counts across its categories, i.e. a mixture of the categories' distributions. Specifically, each cell c has a latent probability vector $\theta_c \in [0,1]^K$ describing the proportion of each category that makes up cell c. Each category k has a probability vector $\sigma(\beta_k)$ that provides the distribution of its feature counts. Here σ indicates the softmax function that transforms β_k to a probability vector that sums to 1. The observed vector of counts x_c are a multinomial draw with probabilities π_c, where $\pi_c = \sigma(B)\theta_c$, and $B = (\beta_1, \ldots, \beta_K)$ is a matrix of the individual β_k vectors. To extend this model to multiple modalities, we assume a shared latent variable z_c that is common across modalities, but each modality has a different $B^{(i)}$ that transforms the shared latent class probabilities into the feature space of the modality.

Futhermore, it is well known that there can be meaningful technical artifacts ("batch effects") between different datasets on the same modality, for example due to differences between platforms or laboratory preparations. To counter this, our model also adjusts the sampling probabilities $\sigma(B^{(i)})\theta_c$ differently for data from different batches within the same modality i.

θ_c is the latent variable describing the contributions of each category to cell c, and is shared across all modalities. In LDA models, it is typically assumed to have a Dirichlet prior distribution. However, we use a Laplace approximation to the Dirichlet introduced in ProdLDA [Srivastava and Sutton, 2017], which allows for incorporation into a VAE model. This prior assumes a latent variable z_c with a Gaussian prior, and sets $\theta_c = \sigma(z_c)$, where σ is the softmax transformation. We use this approximation to the Dirichlet distribution to provide a multi-modality method appropriate for sparse sequence count data.

4.3 Data Processing and Method Implementation

SNARE-seq processing and annotation We downloaded the processed counts of the adult mouse cerebral cortex data (10,309 cells) described in Chen et al. [2019]. We applied quality filtering that retained cells having a number of genes detected greater than 20. For genes, we used the ones detected in more than 5 cells and have a total number of counts greater than 10. For peaks, we removed the ones having nonzero counts in more than 10% of cells or less than 5 cells. We performed clustering analysis using Seurat (version 3.2.2) on the gene expression modality. The data were normalized using `SCTransform` function with default parameters, followed by principal component analysis (PCA) using the default 3,000 variable features. Louvain algorithm was applied on the first 50 PCs with the resolution parameter equals 0.65. Cell type annotations are generated on the resulting 15 clusters using the marker genes provided in Yao et al. [2020]. We applied cisTopic (version 0.3.0) on the chromatin accessibility data with default parameters. Model selection was conducted based on log-likelihood using `runWrapLDAModels` and `selectModel` functions, and 30 topics are used in the results.

For the integration of SNARE-seq with the MOp data using `Cobolt`, we map the SNARE-seq counts to the peak set called on the MOp scATAC-seq data. Since peaks are typically called in a dataset-specific manner, the ideal integration strategy would be to redo the peak-calling with all datasets combined. However, in Supplementary Figure S4 we show that our simple alternative of mapping data to peaks called on a different dataset from the same system does not result in significant performance loss for `Cobolt`.

MOp Data Preprocessing We downloaded the single-nucleus 10x v3 transcriptome dataset (90,266 cells) and the open chromatin dataset (15,731 cells, sample 171206_3C) described in Yao et al. [2020]. For mRNA-seq quality control, we filtered cells that have less than 200 genes detected or have greater than 5% mitochondrial counts. For ATAC-seq, we utilized the `TSSEnrichment` and blacklist region reads calculation functionalities in Signac. We subsetted cells with the blacklist ratio less than 0.1, the number of unique molecular identifiers (UMIs) greater than 50, and the TSS enrichment score greater than 2 and less than 20.

A total of 88,021 and 13,249 were retained for mRNA-seq and ATAC-seq, separately. To make annotation in the two datasets consistent, we merged the layer 2/3 IT and layer 4/5 IT subclusters in ATAC-seq data. For mRNA-seq, we merged Lamp5, Vip, and Sncg into one CGE cluster. When integrating the MOp datasets with the SNARE-seq data, we used only genes detected in both the scRNA-seq and the SNARE-seq datasets.

Gene Activity Calculation Gene activity matrix for chromatin accessibility is generated by counting the number of reads overlapping genes and their promoters using BEDOPS [Neph et al., 2012], where a promoter is defined as the region starting from the transcription start site (TSS) to 3,000 base pairs upstream of TSS.

Cobolt Network Architecture and Training For each modality i, the encoder takes as input the log-plus-one transformed counts. We use one fully connected layer of size 128, followed by fully connected layers for mean $\tilde{\mu}^{(i)}$ and log-variance $\log \tilde{\Sigma}^{(i)}$. We tried networks with one or two hidden layers of varying sizes and found the results pretty stable. The decoders follow our probability model for sparse counts described in Section 4.2 (see also Equation S5) and do not contain neural networks. We set the parameter of the Dirichlet prior to 50 divided by the number of latent variables K. The actual parameters used for the Gaussian prior are calculated using the Laplace approximation (see also Equation S3). For the ELBO objective, we set the weighting terms $\lambda^{\mathcal{A}}$ reciprocal to the number of samples available to modality combination $\mathcal{A}$. And we set the hyperparameter weights η for conditional likelihood terms to 1. Adam optimizer is used, and we select a learning rate of 0.005 after tunning. We adopt a KL cost annealing schedule that linearly increases the weight of the KL term γ from 0 to 1 in the first 30 epoch. During training, we use a batch size of 128 and a fixed number of 100 epochs.

For features, we focus on genes and peaks that have top 30% average expression and remove the ones in the top 1%. The choice of top features is less important here, which we found to have a small effect on the results.

Notice that the softmax transformation from z_c to θ_c is not a one-to-one transformation. Therefore, we scale $\hat{z}_c$ to mean 0 before the downstream correction, followed by clustering and visualization.

In correcting for missing modalities, we predict using XGBoost, setting the objective function to regression with squared loss, the learning rate to 0.8, and the maximum depth of a tree to 3. XGBoost is applied separately to each modality (scRNA-Seq and scATAC-Seq), resulting in our corrected estimates $\hat{Z}_2, \hat{Z}_3$. .

Clustering and Visualization of Cobolt Results The results showing integration of SNARE-Seq, scRNA-Seq, and scATAC-Seq are based on combining the estimated $\hat{Z}_1$, $\hat{Z}_2$, and $\hat{Z}_3$. Clusters are generated on the corrected latent variables using Louvain algorithm [Blondel et al., 2008]. We use the implementation of naive Louvain algorithm in `FindClusters` function from the R package Seurat. All parameters, other than the resolution controlling the number of clusters identified, are set to default. UMAPs are generated using the `umap` function from the R package uwot with the number of neighbors set to 30.

MOFA2, scMM, and BABEL analysis on the SNARE-seq data Following the vignette of MOFA2, we used the top 2500 variable genes and cisTopic embeddings of the chromatin accessibility modality as input. Variable genes are selected using the `FindVariableFeatures` function from `Seurat` using selection method "vst". The number of factors was set to 10. Two factors were identified as technical factors after inspecting their correlation with the total number of reads counts per cell. The UMAP representation was then generated using the rest of the factors with the number of neighbors set to 30. scMM was run with batch size equals 32, number of epochs equals 50, learning rate equals 10^{-4}, number of latent dimensions equals 10, number of hidden dimension for gene expression equals 100, and number of hidden dimensions for chromatin accessibility equals 500. The parameters were chosen following the scMM paper. BABEL was run with the number of latent dimensions set to 10 and the batch size set to 256. Other parameters were chosen to follow the built-in SNARE-seq defaults.

LIGER and Signac Analysis on the MOp Data LIGER and Signac (Seurat) take as input gene-level count summaries from different modalities, such as gene expression or gene body methylation/chromatin accessibility measures. The input is different from `Cobolt`, which uses the peak summaries directly without needing to summarize at the gene level. Therefore, we applied these two methods on the gene expression

and gene activity matrices, where the latter is defined as the summarized chromatin accessibility counts over gene and promoter regions.

We ran LIGER (version 0.5.0) using default parameters on the filtered data. Parameter K for factorization is set to 30 after inspecting the plot generated by function `suggestK`. Louvain clustering was performed by setting the resolution such that 17 clusters were obtained. We ran Signac (version 1.1.0) on the same filtered data. We first performed clustering analysis on the gene expression modality and then transferred the cluster labels to the open chromatin modality. Both the gene expression matrix and gene activity matrix were normalized by running `NormalizeData` followed by `ScaleData`. For the gene activity matrix, the `scale.factor` parameter was set to the median of UMI distribution as suggested by the vignette. Other parameters were kept as default. We then ran `FindTransferAnchors` with the default number of dimensions equals 30, which is the same as used for LIGER. Finally, we ran `TransferData` with weight reduction set to "cca" or the Latent Semantic Indexing (LSI) from analyzing the peak matrix. Results using LSI are presented in the paper as it performed better than the "cca" option.

Test training split for method comparision For scMM and BABEL, which do not allow the single-modality data to be used in the training of the model, we assign 20% of the cells as paired modality data that is used for the training set; the trained model was then used to generate the embedding on the rest of the cells without providing the pairing information. This provided separate estimates for the unpaired mRNA and ATAC modalities, respectively, with which we evaluated whether the paired cells were close together. For `Cobolt`, which allows the use of single-modality data for the training of the model, we assign 20% of the cells as paired modality data and the other 80% were given as single-modality data, and then we trained the model on all of the cells. For LIGER and Signac, which are not designed for multiple-modality data, we hide the pairing information on all cells, and treat all of the cells as if they were collected on different cells. To be comparable with the other methods, we only evaluated their performance on the 80% of cells treated by the other methods as single-modality data.

5 Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and materials An open-source package `Cobolt`, implemented in Python, is available on GitHub (https://github.com/boyinggong/cobolt). Code scripts for reproducing results presented in the paper are publicly accessible on GitHub (https://github.com/boyinggong/cobolt_manuscript).

Datasets used in this paper are all publicly available. The SNARE-seq data can be downloaded from Gene Expression Omnibus with accession number GSE126074. The MOp scRNA-seq and scATAC-seq data can be downloaded from NeMO Archive with accession number nemo:dat-ch1nqb7.

Competing interests The authors declare that they have no competing interests.

Funding This work has been supported by the NIH grant U19MH114830.

Authors' contributions EP formulated the problem and supervised the work. BG developed the model and performed the analysis, with inputs from all authors. BG and YZ developed the model inference and coded the initial implementation. BG implemented the package. BG and EP drafted and revised the manuscript. YZ revised the manuscript and figures. All authors read and approved the final manuscript.

Acknowledgements The authors would like to thank Jean-Philippe Vert for useful early discussions on the analysis of multi-modality data, and Hector Roux de Bezieux for assistance in accessing the MOp data.

References

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21:1–17, 2020.

Daniel Backenroth, Zihuai He, Krzysztof Kiryluk, Valentina Boeva, Lynn Pethukova, Ekta Khurana, Angela Christiano, Joseph D Buxbaum, and Iuliana Ionita-Laza. Fun-lda: a latent dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. *The American Journal of Human Genetics*, 102(5):920–942, 2018.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

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.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):1–9, 2018.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, pages 1–11, 2021.

Carmen Bravo González-Blas, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nature methods*, 16(5):397–400, 2019.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie Jane Lee, Aaron J Wilk, Charlotte Darby, Michael Zagar, et al. Integrated analysis of multimodal single-cell data. *bioRxiv*, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jeongwoo Lee, Daehee Hwang, et al. Single-cell multiomics: technologies and data analysis methods. *Experimental & Molecular Medicine*, 52(9):1428–1442, 2020.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Kodai Minoura, Ko Abe, Hyunha Nam, Hiroyoshi Nishikawa, and Teppei Shimamura. Scmm: Mixture-of-experts multimodal deep generative model for single-cell multiomics data analysis. *Available at SSRN 3806072*, 2021.

Shane Neph, M Scott Kuehn, Alex P Reynolds, Eric Haugen, Robert E Thurman, Audra K Johnson, Eric Rynes, Matthew T Maurano, Jeff Vierstra, Sean Thomas, et al. Bedops: high-performance genomic feature operations. *Bioinformatics*, 28(14):1919–1920, 2012.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *International Conference on Learning Representations*, 2017.

Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Tim Stuart, Avi Srivastava, Caleb Lareau, and Rahul Satija. Multimodal single-cell chromatin analysis with signac. *bioRxiv*, 2020.

Zhe Sun, Ting Wang, Ke Deng, Xiao-Feng Wang, Robert Lafyatis, Ying Ding, Ming Hu, and Wei Chen. Dimm-sc: a dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, 34(1):139–146, 2018.

Xinjun Wang, Zhe Sun, Yanfu Zhang, Zhongli Xu, Hongyi Xin, Heng Huang, Richard H Duerr, Kong Chen, Ying Ding, and Wei Chen. Brem-sc: a bayesian random effects mixture model for joint clustering single cell multi-omics data. *Nucleic acids research*, 48(11):5814–5824, 2020.

Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

Kevin E Wu, Kathryn E Yost, Howard Y Chang, and James Zou. Babel enables cross-modality translation between multiomic profiles at single-cell resolution. *Proceedings of the National Academy of Sciences*, 118 (15), 2021.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5575–5585, 2018.

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.

Sohiya Yotsukura, Seitaro Nomura, Hiroyuki Aburatani, Koji Tsuda, et al. Celltree: an r/bioconductor package to infer the hierarchical structure of cell populations from single-cell rna-seq data. *BMC bioinformatics*, 17(1):363, 2016.

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. *Nature structural & molecular biology*, 26(11):1063–1070, 2019.

Supplementary Material

S1 Model and Inference

The Objective Function To estimate the posterior distribution $p(z|x)$ of the latent variable z and the parameters, we follow the VAE approach to maximize the evidence lower bound (ELBO):

$$\mathrm{ELBO}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\psi(x|z)] - \gamma\,\mathrm{KL}(q_\phi(z|x), p(z)),$$

where the encoder $q_\phi(z|x)$ is parameterized by neural networks and serves as an approximation to actual posterior distribution. The decoder $p_\psi(x|z)$ follows our model for single modality, which we will describe in detail later.

The above equation gives the objective function when all the modalities are observed. In case of misssing modalities, our conditional independence assumption allows us to obtain a partial ELBO. Consider a subset of modalities $\mathcal{A} \subset \{1, \cdots, M\}$ and data $x^{\mathcal{A}} = \{x^{(i)} | i \in \mathcal{A}\}$, its log-likehood given the latent variable is $\log p_\psi(x^{\mathcal{A}}|z) = \sum_{i\in\mathcal{A}} \log p_\psi(x^{(i)}|z)$. The corresponding ELBO can be written as

$$\mathrm{ELBO}(x^{\mathcal{A}}) = \mathbb{E}_{q_\phi(z|x^{\mathcal{A}})}\left[\sum_{i\in\mathcal{A}} \eta_i \log p_\psi(x^{(i)}|z)\right] - \gamma\,\mathrm{KL}\left(q_\phi(z|x^{\mathcal{A}}), p(z)\right), \tag{S1}$$

where η_i are hyperparameters allowing different weights for different modalities.

In actual training, even when all modalities are collected, we still subsample the modalities such that posterior likelihoods given different modality subsets are learnt. Denote $\mathcal{S}_c$ as the modalities collected for cell c and $x_c = \{x^{(i)} | i \in \mathcal{S}_c\}$ as the data, our final objective function is given by

$$\sum_{c=1}^{N}\left[\lambda\,\mathrm{ELBO}(x_c) + \sum_{\mathcal{A}\subsetneq\mathcal{S}_c} \lambda_{\mathcal{A}}\,\mathrm{ELBO}(x_c^{\mathcal{A}})\right], \tag{S2}$$

where λ's are hyperparameters allowing different weightings for the ELBO terms with different modalities.

Prior Distribution We use Gaussian distribution for the prior $p(z)$. A transformed vector $\theta = \sigma(z)$ is then calculated as the mixing probability of our latent model (Figure S1) for modeling single modalities, where σ is the softmax transformation. θ therefore follows a logistic normal distribution, which is commonly used as an approximation to a Dirichlet distribution. To approximate $Dirichlet(\theta|\xi)$, the parameters of the logistic normal $\mathcal{LN}(\theta|\mu_0(\xi), \Sigma_0(\xi))$ are calculated as a function of ξ,

$$\begin{aligned}\mu_0(\xi)_k &= \log\xi_k - \frac{1}{K}\sum_j \log\xi_j,\\ \Sigma_0(\xi)_{kk} &= \frac{1}{\xi_k}\left(1 - \frac{2}{K}\right) + \frac{1}{K^2}\sum_j \frac{1}{\xi_j}.\end{aligned} \tag{S3}$$

Therefore, given the hyperparameter ξ, we use the Gaussian distribution $\mathcal{N}(z|\mu_0(\xi), \Sigma_0(\xi))$ for the prior $p(z)$.

Decoders For cell c in modality i, our generative likelihood $p(x_c^{(i)}|\theta_c, B^{(i)}, \alpha_{0,:}^{(i)}, \alpha_{1,:}^{(i)})$ follows multinomial distribution with probability

$$\pi_c^{(i)} = \sigma\big(\,\mathrm{diag}(\alpha_{1,\ell_c}^{(i)})\sigma(B^{(i)})\theta_c + \alpha_{0,\ell_c}^{(i)}\big), \tag{S4}$$

where ℓ_c indicates the batch of cell c and $\alpha_{0,\ell_c}^{(i)}, \alpha_{1,\ell_c}^{(i)} \in \mathbb{R}^{d_i}$, and $B^{(i)} \in \mathbb{R}^{d_i\times K}$ are parameters to be estimated that are shared across samples. $\sigma(B^{(i)})$ gives the probability vectors for each latent category. $\alpha_{0,\ell_c}^{(i)}$ and $\alpha_{1,\ell_c}^{(i)}$ therefore provides scaling and shifting for adjusting dataset- or platform-specific differences in feature counts.

Notice that Equation S4 works without the softmax transformation on $B^{(i)}$, which gives us another version of the model

$$\pi_c^{(i)} = \sigma\big(\operatorname{diag}(\alpha_{0,\ell_c}^{(i)})B^{(i)}\theta_c + \alpha_{0,\ell_c}^{(i)}\big). \tag{S5}$$

In practice, we find S5 works slightly better than S4. The results we present in this paper are based on S5.

In summary, we write the joint likelihood of the latent variable z_c and the data x_c of cell c as,

$$p(x_c^{\mathcal{S}}, z_c | B^{\mathcal{S}}, \alpha_{0,:}^{\mathcal{S}}, \alpha_{1,:}^{\mathcal{S}}) = p(z_c) \prod_{i \in \mathcal{S}} p(x_c^{(i)} | \theta_c = \sigma(z_c), B^{(i)}, \alpha_{0,:}^{(i)}, \alpha_{1,:}^{(i)}), \tag{S6}$$

where $p(x_c^{(i)}|\theta_c, B^{(i)}, \alpha_{0,:}^{(i)}, \alpha_{1,:}^{(i)})$ follows multinomial distribution with probability vector given by Equation S5. Figure S1 shows a graphical model representation of this data generation process.

Encoders We would like to obtain an approximation $q(z|x^{\mathcal{A}})$ of the posterior distribution $p(z|x^{\mathcal{A}})$ given any collection of modalities $x^{\mathcal{A}}$. To do this, we first write the true posterior as

$$p(z|x^{\mathcal{A}}) \propto p(z) \prod_{i \in \mathcal{A}} \frac{p(z|x^{(i)})}{p(z)}.$$

We then approximate $\frac{p(z|x^{(i)})}{p(z)}$ with Gaussian $g(z|x^{(i)}) := \mathcal{N}(\tilde{\mu}^{(i)}, \tilde{\Sigma}^{(i)})$, where $\tilde{\mu}^{(i)}$ and $\tilde{\Sigma}^{(i)}$ are estimated by neural networks. Here we use the property that the product of two multivariate Gaussian densities is proportional to a Gaussian density. $q_\phi(z|x^{\mathcal{A}}) \propto p(z) \prod_{i \in \mathcal{A}} g(z|x^{(i)})$ is still Gaussian whose parameters can be analytically written out,

$$\begin{aligned} \Sigma^{\mathcal{A}} &= \Big(\tilde{\Sigma}_0^{-1} + \sum_{i \in \mathcal{A}} (\tilde{\Sigma}^{(i)})^{-1}\Big)^{-1}, \\ \mu^{\mathcal{A}} &= \Sigma^{\mathcal{A}}\Big(\tilde{\Sigma}_0^{-1}\tilde{\mu}_0 + \sum_{i \in \mathcal{A}} (\tilde{\Sigma}^{(i)})^{-1}\tilde{\mu}^{(i)}\Big). \end{aligned} \tag{S7}$$

With this trick, instead of using $2^M - 1$ neural networks to estimate all $\Sigma^{\mathcal{A}}$ and $\mu^{\mathcal{A}}$ individually, our model uses one network for $\tilde{\mu}^{(i)}$ and $\tilde{\Sigma}^{(i)}$ separately for each modality i, resulting in a total of $2M$ networks. The variational posteriors $q_\phi(z|x^{\mathcal{A}})$ are calculated using $\tilde{\mu}^{(i)}$ and $\tilde{\Sigma}^{(i)}$. The variational posterior $q_\phi(\theta|x^{\mathcal{A}})$ is then defined to be logistic normal $\mathcal{LN}(\mu^{\mathcal{A}}, \Sigma^{\mathcal{A}})$.

S2 Supplementary Figures

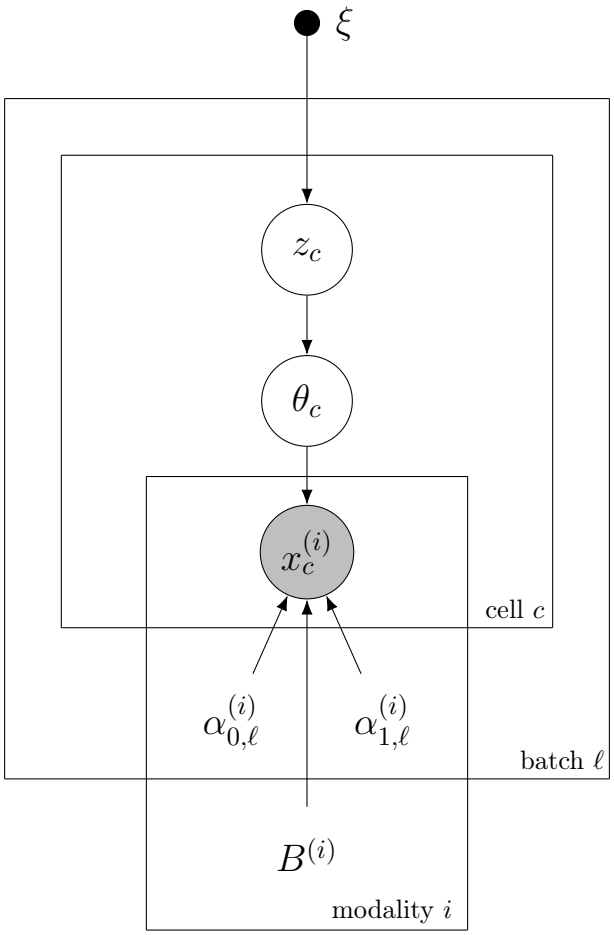

Figure S1: Graphical model representation of the generative distribution of `Cobolt`.

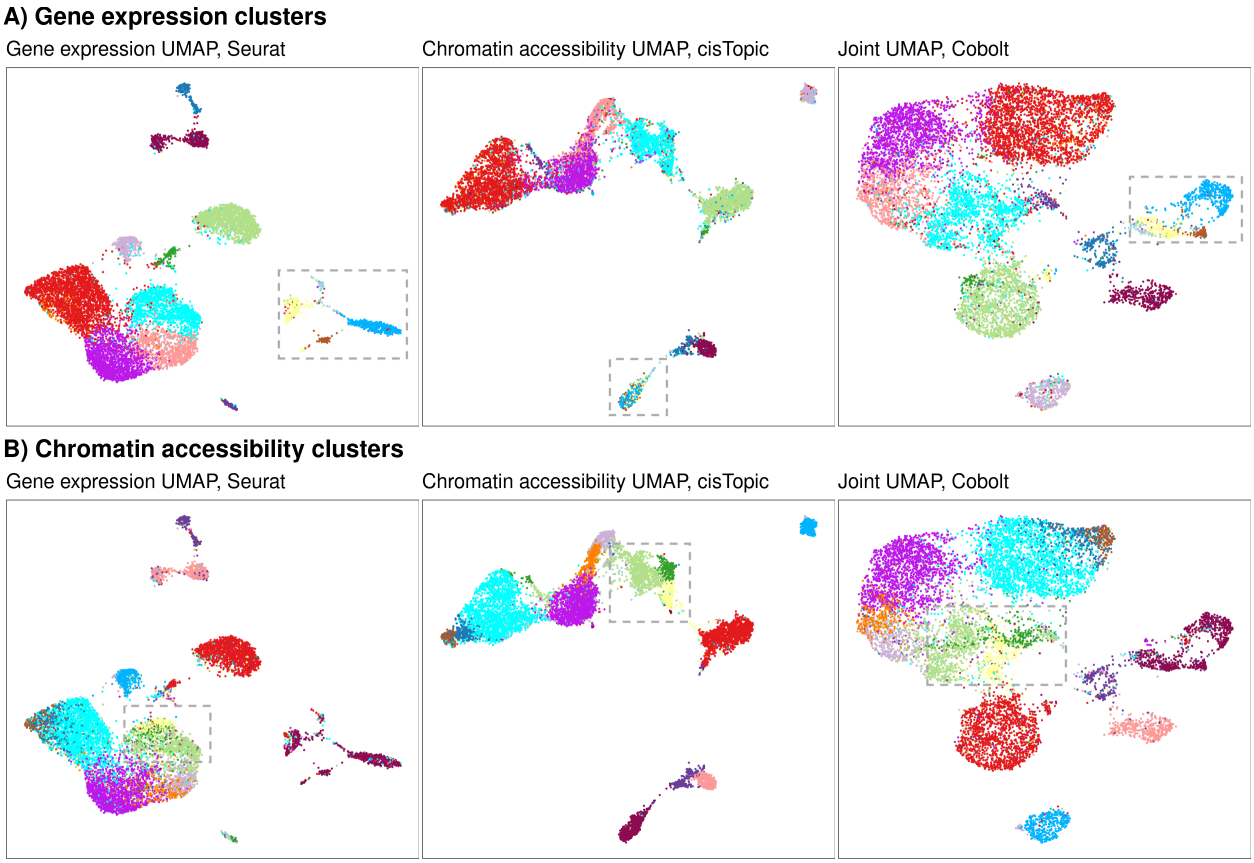

Figure S2: Plotted are UMAP visualizations of the reduced dimensionality space created by analyzing only gene expression data using Seurat [Stuart et al., 2019] and only chromatin accessibility using cisTopic [González-Blas et al., 2019], and jointly using both modalities (`Cobolt`). The cells are color-coded by the cluster they are assigned to based on clustering of A) gene expression modality and B) chromatin accessibility. We note that the cluster colors are randomly and separately assigned for panel A and B. In both cases, we see distinct clusters that are found in one modality but are not reflected in the other modality. Highlighted in the panels of A) are cells that show separated clusters in the gene expression, but not in chromatin accessibility, and highlighted in B) are cells that are well separated in chromatin accessibility but not in gene expression. Both sets of highlighted cells show separation in the joint analysis of `Cobolt`. In both these cases, the results of the `Cobolt` analysis shows that `Cobolt` integrates both modalities to find these modality-specific subtype differences: the two highlighted cell populations mentioned above are both distinguishable in the `Cobolt` reduced dimensions, and their cluster median silhouette widths are improved compared to single modality analysis (Supplementary Figure S3). These results demonstrate the power of a joint analysis of modalities using `Cobolt` for a comprehensive understanding of subtypes.

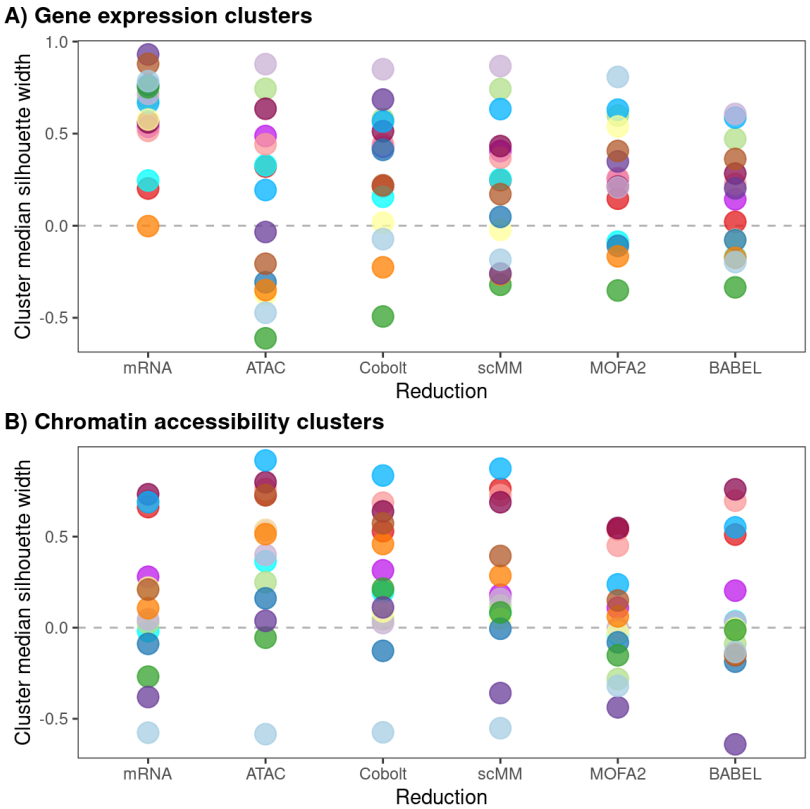

Figure S3: The median silhouette width of SNARE-seq clusters generated using A) gene expression modality and B) chromatin accessibility. For each of the clustering, we calculate the median silhouette width per cluster on reduced dimensionality space created by analyzing only gene expression data (using Seurat), only chromatin accessibility (using cisTopic), and jointly with both modalities using `Cobolt`, scMM, MOFA2, and BABEL. Cluster colors are assigned the same as in Figure 2 and Figure S2.

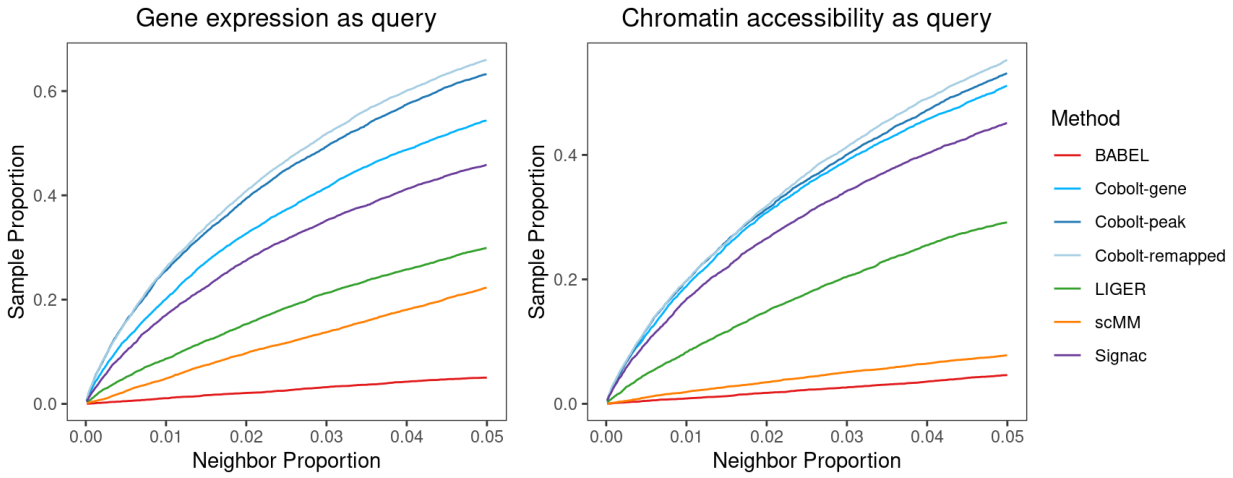

Figure S4: Full results of Figure 5. Also included is a comparison of `Cobolt` dimensionality reduction on SNARE-seq data 1) using peaks called on the SNARE-seq (`Cobolt`-peak), 2) mapping counts to the peaks called on the MOp scATAC-seq (`Cobolt`-remapped), and 3) summarizing chromatin accessibility counts to gene regions (`Cobolt`-gene). No significant performance loss is observed for `Cobolt`-remapped compared to `Cobolt`-peak, indicating our strategy of mapping data to peaks called on similar datasets a simple but effective alternative in integration analysis. Using peaks vastly improved the performance relative to using the gene activities summarization (`Cobolt`-gene), suggesting that there is a loss of information in using gene summaries, such as distant binding sites not included in the gene summarization.

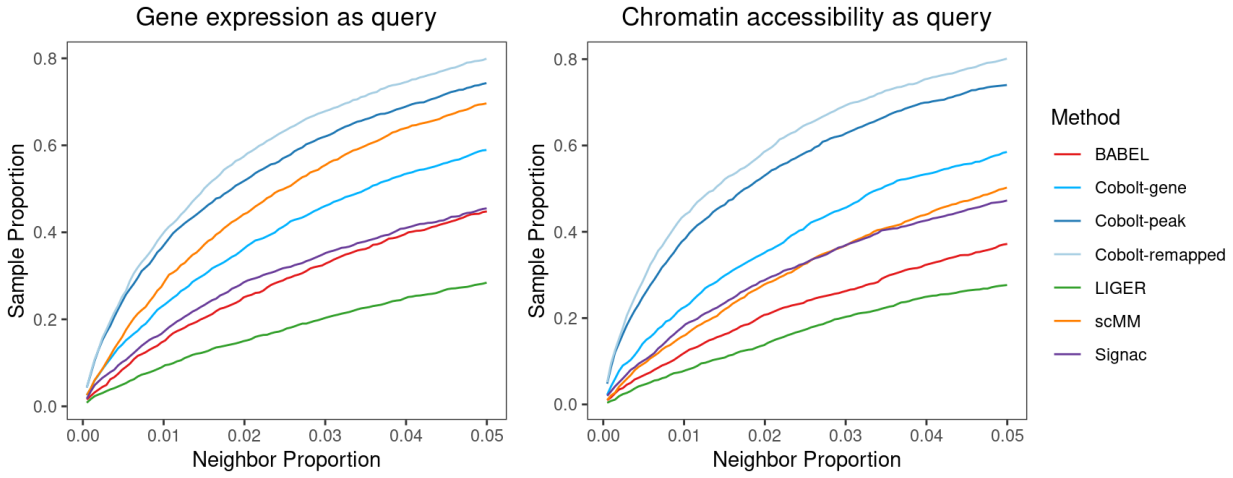

Figure S5: An analysis as in Figure 5, with a larger percentage of cells used as paired data in the training of scMM, BABEL, and Cobolt (80% rather than 20%). The x-axis shows the number of neighbors considered (k) as a proportion of the total testing sample size. The y-axis shows the proportion of cells whose paired data are within their k-nearest neighbors in the other modality.

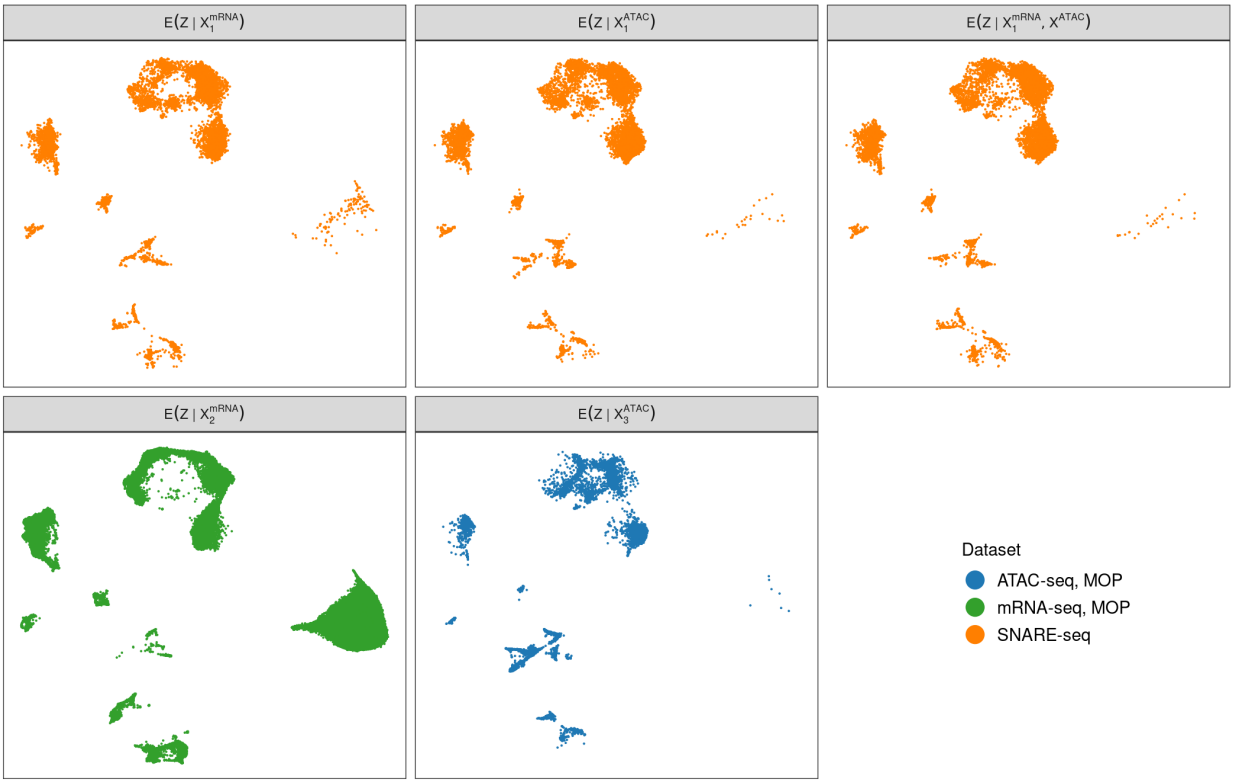

Figure S6: Estimates of the posterior means *without* missing modality correction. The upper panel shows the UMAP dimensionality reduction plots of the posterior means $\hat{Z}_1 = E(Z|X_1^{mRNA}, X_1^{ATAC})$, $\hat{Z}_1^{mRNA} = E(Z|X_1^{mRNA})$, and $\hat{Z}_1^{ATAC} = E(Z|X_1^{ATAC})$ estimated for joint cells. The lower panel shows $\hat{Z}_2^{mRNA} = E(Z|X_2^{mRNA})$ and $\hat{Z}_3^{ATAC} = E(Z|X_3^{ATAC})$ estimated for single-modality cells. Other than the unannotated cells from MOp scRNA-seq data, cell populations from different datasets are well-aligned. We noticed slight distributional shifts between $\hat{Z}_1$, $\hat{Z}_1^{mRNA}$, and $\hat{Z}_1^{ATAC}$ in the upper panel, motivating additional correction procedures.

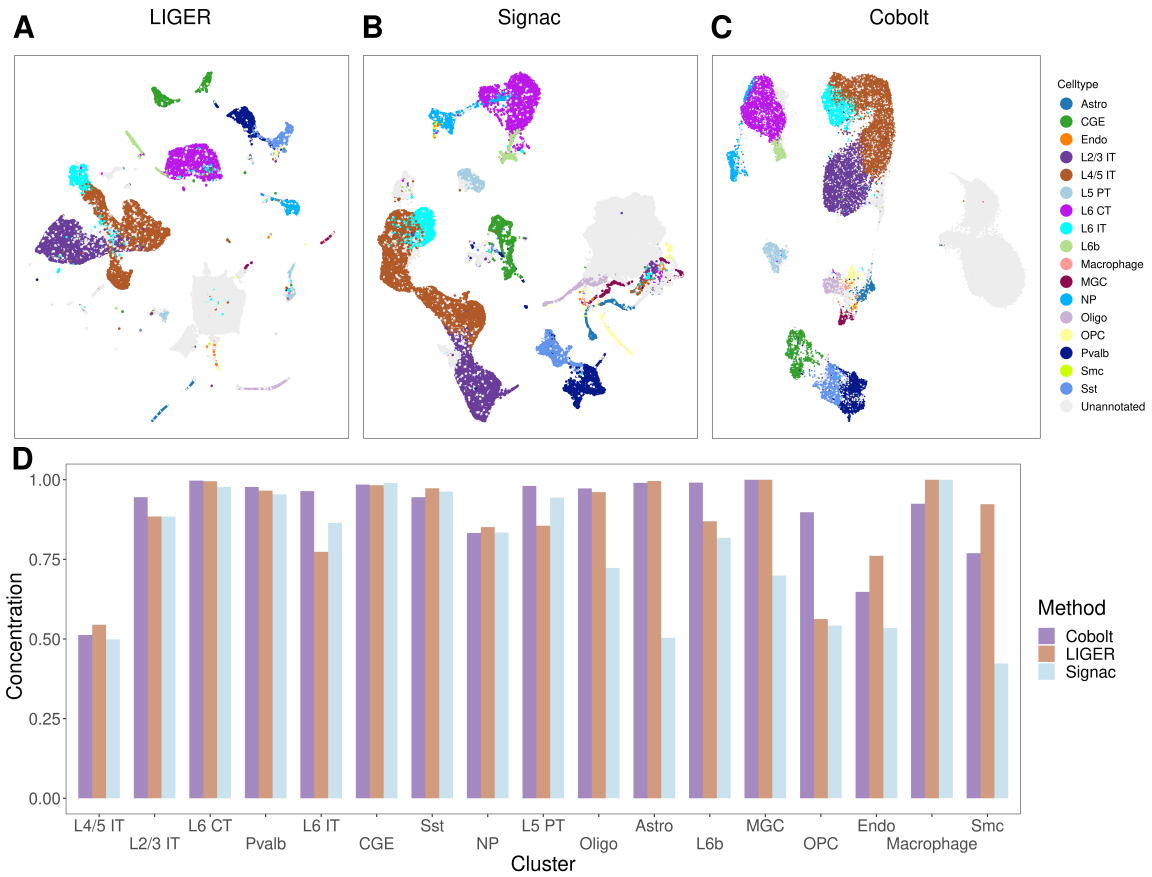

Figure S7: A, B, C) UMAP visualizations generated by (A) LIGER, (B) Signac, and (C) `Cobolt` aligning MOp scATAC-seq and scRNA-seq, colored by cell type annotations. The three major classes of cell types—GABAergic inhibitory neurons, glutamatergic excitatory neurons, and non-neurons—are indicated on the `Cobolt` UMAP plot. D) Barplot of maximum overlap proportion between identified clusters and each cell type. For the complete concordance between clustering results and *a priori* cell type labels, see Supplementary Figure S8).

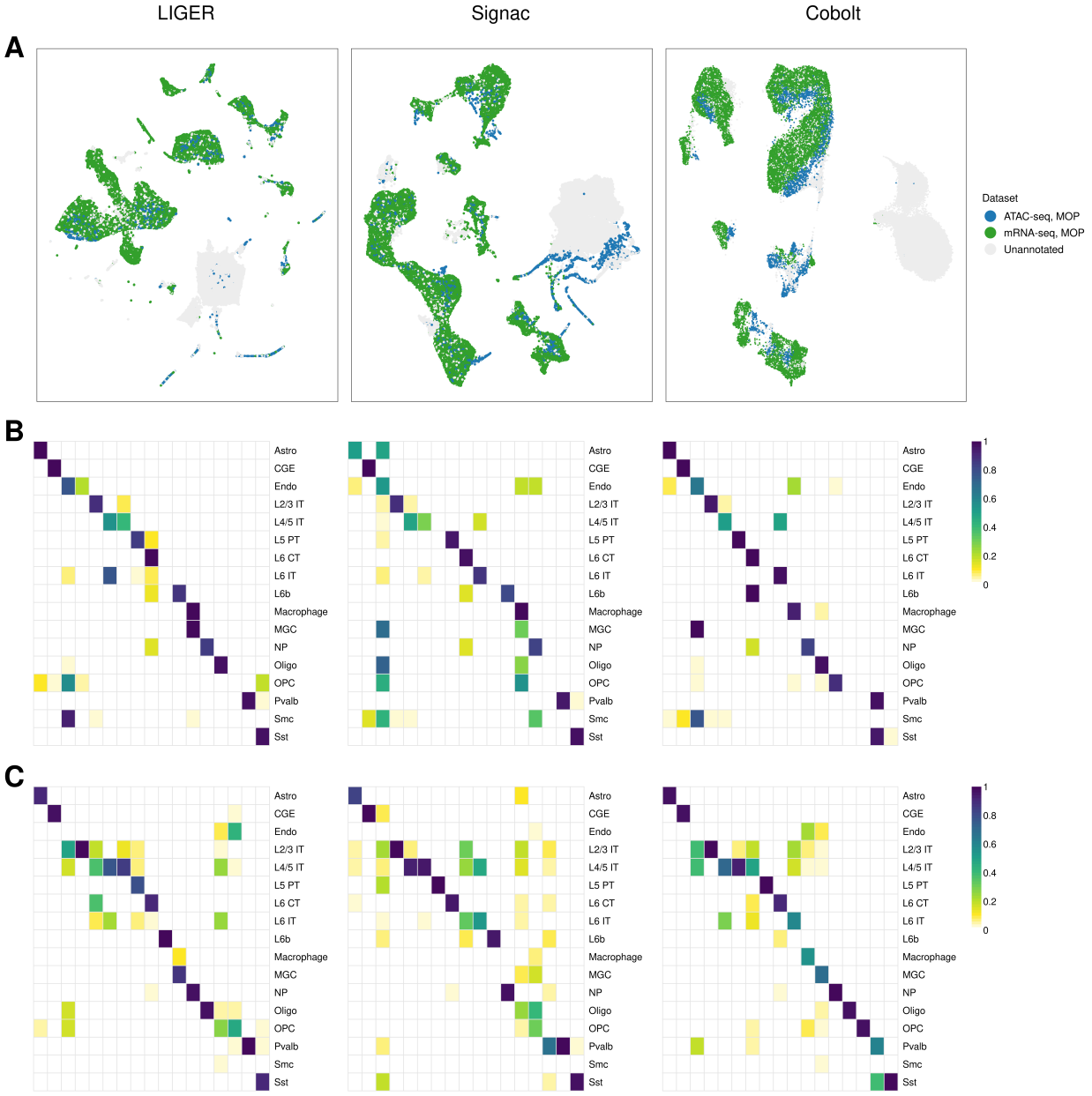

Figure S8: A) UMAP visualizations generated by LIGER, Signac, and `Cobolt` for MOp scATAC-seq and scRNA-seq, colored by datasets. B, C) Concordance matrices between clustering results and annotations, scaled to sum 1 by (B) rows and (C) columns.

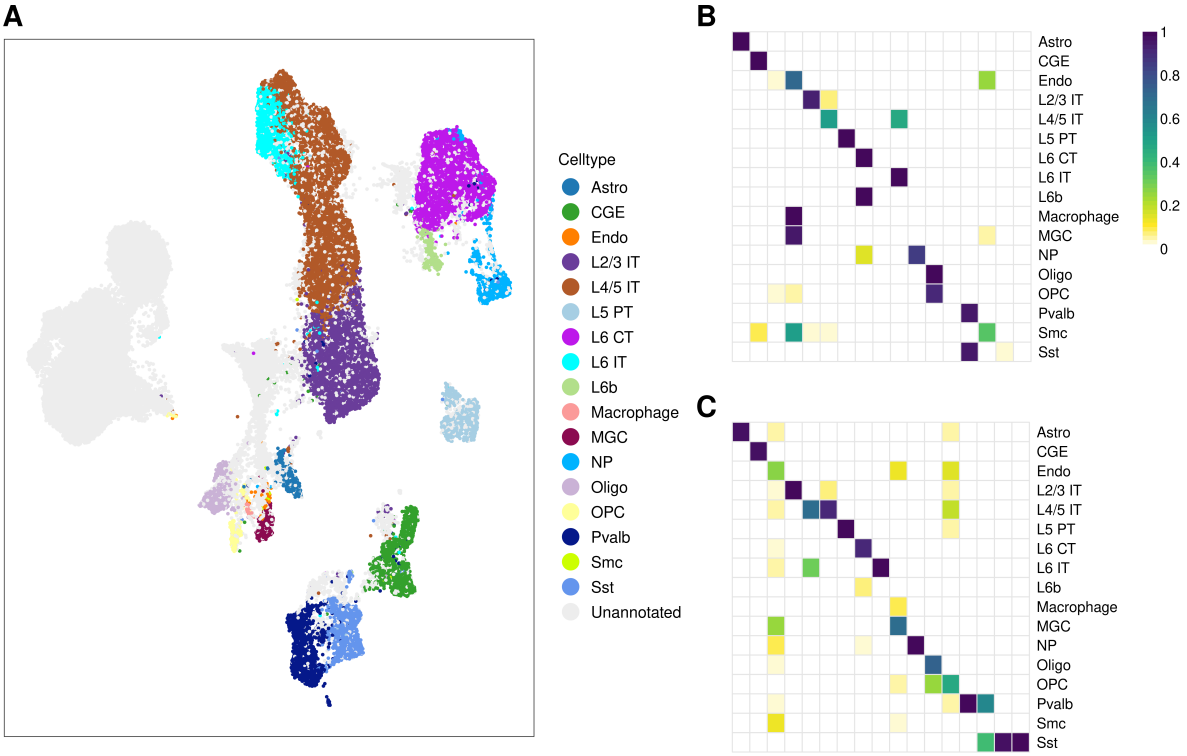

Figure S9: `Cobolt` results with no filtering on cells and minimal filtering on features. Genes and peaks with total counts larger than 10 and non-zero counts larger than 5 are retained in the model. A) UMAP visualization colored by cell type annotation. B, C) Concordance matrices between clustering results and annotations, scaled to sum 1 by (B) rows and (C) columns.