Constructing Cell-type Taxonomy by Optimal Transport with Relaxed Marginal Constraints

Sebastian Pena[1], Lin Lin[2], Jia Li[1]

[1]Department of Statistics, The Pennsylvania State University
Email: skp5799@psu.edu, jiali@psu.edu
[2]Department of Biostatistics and Bioinformatics, Duke University
Email: l.lin@duke.edu

Abstract

The rapid emergence of single-cell data has facilitated the study of many different biological conditions at the cellular level. Cluster analysis has been widely applied to identify cell types, capturing the essential patterns of the original data in a much more concise form. One challenge in the cluster analysis of cells is matching clusters extracted from datasets of different origins or conditions. Many existing algorithms cannot recognize new cell types present in only one of the two samples when establishing a correspondence between clusters obtained from two samples. Additionally, when there are more than two samples, it is advantageous to align clusters across all samples simultaneously rather than performing pairwise alignment. Our approach aims to construct a taxonomy for cell clusters across all samples to better annotate these clusters and effectively extract features for downstream analysis. A new system for constructing cell-type taxonomy has been developed by combining the technique of Optimal Transport with Relaxed Marginal Constraints (OT-RMC) and the simultaneous alignment of clusters across multiple samples. OT-RMC allows us to address challenges that arise when the proportions of clusters vary substantially between samples or when some clusters do not appear in all the samples. Experiments on more than twenty datasets demonstrate that the taxonomy constructed by this new system can yield highly accurate annotation of cell types. Additionally, sample-level features extracted based on the taxonomy result in accurate classification of samples.

1 Introduction

Many recent technologies, such as flow cytometry, Cytometry by time of flight (CyTOF), have led to the massive accumulation of single-cell data. Specifically, single-cell RNA sequencing

data (scRNA-seq) has expanded from considering a single cell in 2009 to considering millions of cells by 2017 [44]. Advances in single cell technologies have allowed the measurement of global gene expression profiles of individual cells and have enabled the inspection of heterogeneous cell populations in complex samples that were inaccessible in bulk sequencing data.

Numerous genes measured by scRNA-seq tools are uninformative and complicate downstream data analyses [16]. The general workflow with scRNA-seq data involves reducing the number of genes, creating a manifold representation for visualization, identifying cell types by unsupervised clustering, and finding expression signatures of cell types by differential expression analysis [27]. While unsupervised clustering is crucial for discovering new cell types as well as unknown cell-gene relationships, manual annotation of these cell clusters is time-consuming and subjective. One major reason for the importance of cell categorization is that the proportion of each cell type in a sample and the associated differentially expressed marker genes can help predict the phenotype class of the sample. Example phenotypes include different levels of resistance to drugs treating cancer or different responses to vaccines, prognosis for diabetes and other diseases [21, 24, 40, 41, 50].

Next, we introduce some useful terminologies. The data we consider are those containing multiple, usually more than two, samples. Each sample is itself a dataset comprised of measurements on many cells. Specifically, a cell corresponds to a data point in the sample and the expression levels of genes are features measured for any cell. The same collection of genes, thus features, are considered for all cells in all samples. However, certain cell types (i.e., particular groups of cells) may be present in one sample but not in the others. Usually, the samples are derived from multiple experiments on a single organism or distinct organisms within the same species. The cells in any sample are grouped into clusters, which correspond to different cell types, but the types may not have been identified.

We aim to construct a *taxonomy* of clusters derived from all samples. Specifically, this taxonomy will be a hierarchical organization of clusters that identifies which clusters correspond to the same cell type and illustrates the relationships between different cell types in terms of their similarity. As clusters are often generated separately for individual samples, their labels are not coherent across samples. The central question we investigate here is how to establish a taxonomy to ensure consistent recognition of cell clusters.

The taxonomy is valuable for multiple purposes. First, once a taxonomy is formed across

samples, if one sample contains manually labeled cell types, the unlabeled cell clusters in other samples can be classified with known cell types. More detailed discussion on this usage scenario will be provided shortly. Secondly, the taxonomy can help us conduct downstream analysis. Consider the classification of samples according to their phenotype conditions, for instance, healthy patients versus those with a disease. With coherent cluster labels given to different samples, we can derive cluster-wise features such as cluster proportions and use them to classify samples. Note that we have regarded classification in two different contexts, that of clusters by their corresponding cell types and that of samples by their phenotype categories.

We emphasize the assumption that samples are not pooled for simultaneous processing. Our algorithm operates exclusively on the characteristics of clusters extracted independently from different samples, such as cluster proportions and cluster-wise average features, rather than the original data. This assumption precludes batch effect removal across samples (which often requires pooling) and synchronized clustering of data points across all samples. The rationale for avoiding centralized analysis is multifaceted. Privacy concerns or the sheer size of the data may prohibit pooling. Additionally, clusters in different samples are often generated manually by expert researchers, and understanding the relationships between clusters across samples may be the primary objective. In such cases, pooling samples to cluster all data points together may not only be unnecessary but also meaningless. While batch effect removal can contribute to constructing a taxonomy, and our algorithm demonstrates robustness to batch variations, the new algorithm is neither designed to remove batch effects nor operates under the same framework.

2 Related Work

The task of determining cell types from scRNA-seq data typically requires several cycles of grouping, sub-grouping, and merging of clusters. This procedure heavily depends on the manual input from subject matter experts who must balance the outcomes proposed by unsupervised clustering of multivariate data [20, 22, 48]. Despite efforts to streamline this process, cell types manually created continue to serve as the benchmark for most applications [34]. A common method for automatically assigning cell type labels to clusters involves utilizing reference data with expert verified ground-truth labels [2, 15, 30, 38, 51, 52]. These reference clusters are

matched with the unlabeled clusters. There are several statistical tools for matching cell clusters between different samples by quantifying the similarity of the gene expression profiles of cells in the clusters being compared [27]. This approach, however, has several drawbacks and negatively affects subsequent biological interpretations. It is noted that the presence of many cell types remarkably lessens the performance of classification tools [27]. Furthermore, the performance of classification is sensitive to the selection of the reference set [1]. Finally, it is often a challenge to find a representative annotated training set encompassing all cell types present in the target datasets.

Cell labels for clusters are usually not available in published scRNA-seq data. One study found that nearly half of 72 datasets did not include inferred cell types [42]. However, even when the biological nature of the cell clusters are unknown, coherent labels assigned to clusters across samples are highly useful. Consistent labels can help merge clusters across samples; and the integrated clusters can better indicate marker genes. These marker genes are usually chosen based on a specific computational clustering method and are utilized to annotate clusters [37].

Even in the presence of an annotated reference sample, automatic annotation methods have had limited success. Clusters that represent identical or similar cell types may have annotations suggesting otherwise. Mislabeling can be caused by the dropout phenomenon that "a gene is observed at moderate or even high expression level in one cell but is not detected in another cell" [20] as well as batch effects [13, 47]. To the best of our knowledge, an automatic tool has not been developed to build a taxonomy of clusters in more than two samples. Instead, cluster labels in a reference sample are used to annotate clusters in other samples in a pairwise fashion [2, 3, 15, 23, 55]. It is known that compared to bulk-cell data, scRNA-seq data has a higher level of noise due to both biological and technical reasons [35]. As a result, often times, no single sample serves as a good reference for annotating cell clusters. We expect that the approach of building a taxonomy of clusters based on all samples, in contrast to a pair of samples, can potentially mitigate the effect of noise and thus yield more accurate cluster annotation, an advantage confirmed by our experiments.

By treating all samples simultaneously, our proposed method creates a hierarchy of clusters without using an annotated reference sample. Our method combines state-of-the-art optimal transport (OT) techniques and hierarchical clustering (aka, dendrogram clustering). *ClusterMap* [12] is an existing method with a similar logical framework as ours, which also considers

multiple datasets with predefined clusters. Using marker genes, this method creates a hierarchy of clusters based on cluster similarity measured by the Jaccard distance. The main drawback to this approach is the necessity of marker genes for each dataset prior to implementing the algorithm. Marker genes are useful for annotating cells that belong to some major cell types. Different marker gene databases often list different sets of marker genes for the same cell population [1]. Furthermore, the same marker genes frequently manifest in various cell clusters, and may represent diverse cell types [9, 18]. Lastly automatic methods that determine marker genes for clusters often disagree and are affected by the noisiness of scRNA-seq data [37, 39].

OT has been increasingly used in machine learning and single-cell data analysis [7, 17, 31, 46, 53, 54]. Although OT is a prominent framework for matching, there are other paradigms such as step-wise merging (similar to the way a dendrogram is formed) used by QFMatch [36] and bipartite graph partitioning (BGP) used by flowMatch [3]. The formulation of the standard OT contains two marginal constraints which reflect the implicit assumption that the proportions of matched clusters are fixed (modulo randomness in observation) across all the datasets. This assumption is improper for scRNA-seq data where one or more cell types can be present in some samples but not others. [28] proposed a new formulation, namely, Optimal Transport with Relaxed Marginal Constraints (OT-RMC), by introducing gap variables to relax the marginal constraints. They have demonstrated that OT is too rigid for various matching problems. We adopt as a core technique OT-RMC and develop a new algorithm that builds a taxonomic tree made up of clusters from all samples. For comparison, we will also experiment with Partial Optimal Transport (POT) which relaxes standard OT by transporting only a fraction of the total mass [6, 11].

An OT-based ensemble clustering algorithm, OTA, was proposed in [29] to integrate clustering results and assess their uncertainty. A key advancement of our method over OTA is its ability to perform simultaneous alignment across multiple samples, whereas OTA performs only pairwise alignment. Pairwise approaches typically rely on selecting a reference sample with known ground-truth labels, but identifying such a sample is often nontrivial. In many cases, a reference may lack the full set of relevant cell types needed to label the remaining samples—indeed, an appropriate labeled reference may not exist at all. This challenge is particularly pronounced and consequential in the applications considered in this work, more so than in the original context of OTA. Our method circumvents these limitations by jointly con-

sidering all samples. It does not require a reference; instead, it operates on clustered cells from different samples, even when the clusters lack meaningful labels. Moreover, OT-RMC (rather than OT) is used as a subroutine to compute dissimilarity scores between clusters. The novel contribution lies in using these scores to construct a dendrogram, enabling flexible resolution for grouping clusters into meta-clusters. A natural resolution level corresponds to the total number of distinct cell types across samples, allowing consistent labeling across datasets even in the absence of ground truth.

3 The Approach

3.1 Framework for Constructing Taxonomy

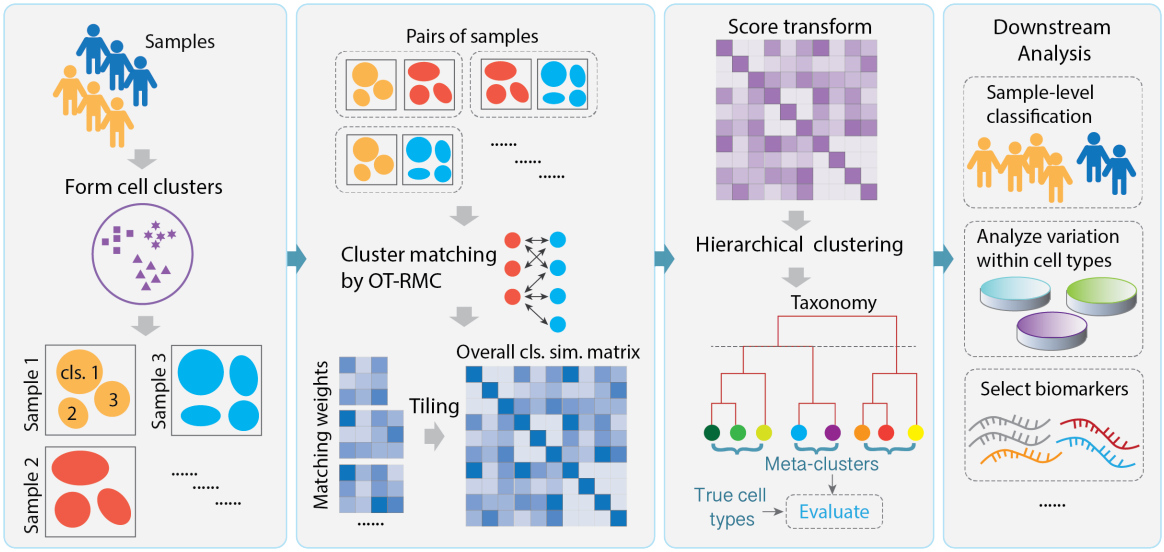


Figure 1: A schematic representation of the Multisample OT Taxonomy (MOTT) system. The taxonomy hierarchically organizes clusters across samples, identifying which clusters correspond to the same cell type and illustrating the relationships between different cell types based on their similarity.

Consider a collection of samples $\mathcal{X}_i$, $i = 1, ..., N$, each containing measurements on a set of cells. Specifically, $\mathcal{X}_i = \{x_1^{(i)}, ..., x_{m_i}^{(i)}\}$, where $x_j^{(i)} \in \mathbb{R}^d$ and d is the dimension of the cell-wise feature vector. Suppose that each sample $\mathcal{X}_i$ has been clustered. Denote clusters generated for $\mathcal{X}_i$ by $C_k^{(i)}$, $k = 1, ..., n_i$, where $C_k^{(i)}$ contains the points (aka, cells) assigned to the kth cluster. We refer to $C_k^{(i)}$'s as *cell clusters* or simply clusters if it is clear from the context. Note that $\mathcal{X}_i = \cup_{k=1}^{n_i} C_k^{(i)}$. Denote $\mathcal{P}_i = \{C_1^{(i)}, C_2^{(i)}, ..., C_{n_i}^{(i)}\}$.

We address the problem of constructing a taxonomy for all clusters in all samples when the clusters C_k^i from different samples are not labeled consistently. Inconsistent labeling can occur in various cases, the simplest being label permutation. However, in practice, a cluster in one sample may split into multiple clusters in another sample, or conversely, multiple clusters may merge into one. Additionally, a new cluster may emerge. Moreover, due to sample variations and nuances in clustering, determining relationships between clusters across samples is often challenging. We propose a system to build a taxonomy tree that reveals which clusters in different samples represent the same cell-type and how the cell-types relate to each other in a hierarchy of similarity. We employ techniques of OT and dendrogram clustering to effectively leverage information across all clusters in one sample and across all samples. We refer to the system as *Multisample OT Taxonomy* (*MOTT*). A schematic flowchart for MOTT is provided in Figure 1.

A taxonomy constructed by MOTT can help us understand relationships between cell types that are not manually annotated and allow for coherent labeling of cell clusters across samples, which is valuable for downstream analysis. Next, we describe the main components in the MOTT system, as illustrated in Figure 1.

1. For every pair of samples $\mathcal{X}_i$ and $\mathcal{X}_j$, $i, j \in \{1, ..., N\}$, compute the cost of matching cluster $C_k^{(i)}$ and $C_l^{(j)}$, denoted by $D(C_k^{(i)}, C_l^{(j)})$, $k = 1, ..., n_i$, $l = 1, ..., n_j$. We refer to the $n_i \times n_j$ matrix containing entries $D(C_k^{(i)}, C_l^{(j)})$ as the cost matrix, which is denoted by $\mathbf{D}_{i,j}$. Apply OT-RMC to compute a matching weight matrix of size $n_i \times n_j$, denoted by $\mathbf{W}_{i,j} = (w_{k,l}^{(i,j)})_{k=1,...,n_i, l=1,...,n_j}$ based on the input cost matrix $\mathbf{D}_{i,j}$. We only need to compute $\mathbf{W}_{i,j}$, $j \geq i$, since $\mathbf{W}_{j,i} = \mathbf{W}_{i,j}^T$. The weights $w_{k,l}^{(i,j)} \in [0, 1]$ indicates the extent of matching cluster $C_k^{(i)}$ in sample $\mathcal{X}_i$ with cluster $C_l^{(j)}$ in sample $\mathcal{X}_j$, a higher value corresponding to stronger matching. We then compute a normalized version of $\mathbf{W}_{i,j}$, denoted by $\widetilde{\mathbf{W}}_{i,j}$, by performing row-wise and column-wise normalization on $\mathbf{W}_{i,j}$ and then computing the average of the two matrices. We treat the matching weights in $\widetilde{\mathbf{W}}_{i,j}$ as the similarity scores between clusters. The detailed formulation of OT-RMC is explained in the next subsection.

2. Form matrix $\mathbf{B}$ by tiling $\widetilde{\mathbf{W}}_{i,j}$, $i = 1, ..., N$, $j = 1, ..., N$. Let $n = \sum_{i=1}^{N} n_i$. $\mathbf{B}$ is an $n \times n$ matrix that contains $\widetilde{\mathbf{W}}_{i,j}$ as its (i, j)th block. We refer to $\mathbf{B}$ as the *overall cluster*

similarity matrix, which contains the similarity score between any pair of clusters across all samples. To convert the similarity matrix $\mathbf{B}$ into an *overall cluster distance matrix* $\mathbf{A}$, we apply an entry-wise transform $-\log(\cdot)$ and then standardize the matrix by dividing each entry by the maximum entry. The nonlinear transform $-\log(\cdot)$ amplifies the difference between matching weights near 0, which occur frequently, thereby allowing us to better capture cluster distances. For any entry smaller than $\xi \cdot 10^{-7}$, where ξ is the smallest cluster proportion in the dataset, instead of applying $-\log(\cdot)$ transform, we set the entry to $7 - \log(\xi)$. This modification ensures that we never encounter $\log(0)$ and mitigates the impact of precision errors in the solution of OT.

3. We apply dendrogram clustering using Ward's linkage method with the *linkage()* function in *Matlab*, using $\mathbf{A}$ as the input pairwise dissimilarity matrix. The result is the taxonomy tree (a dendrogram). We then apply the *cluster()* function to the taxonomy tree to generate the final grouping of clusters. The number of groups is set equal to the number of ground-truth cell types in the original dataset.

Our final result is a partitioning of the cell clusters across all samples. Clusters assigned to the same group are considered to represent the same cell type. We call a group of clusters obtained by a taxonomy a *meta-cluster*. The MOTT system aligns clusters across all samples in a holistic manner, an alignment approach which we refer to as *Simultaneous Alignment* (SA) in the discussion below. We can assign distinct labels to each group of clusters, thus generating consistent cluster labels across samples. Although our method does not explicitly address this, we have not observed different clusters from the same sample being grouped together when the samples are from the same species and the cell types are defined at the same resolution, likely because the large distances between clusters within the same patient generally prevent it. Finally, we summarize the pseudo-code of MOTT in Algorithm 1.

3.2 Matching Clusters by OT-RMC

Consider two samples with partitions $\mathcal{P}_i = \{C_1^{(i)}, ..., C_{n_i}^{(i)}\}$, $i = 1, 2$, where $C_k^{(i)}$ is the kth cluster in the ith partition and n_i is the number of clusters. The distance matrix $\mathbf{D}$ contains entries $D(C_k^{(1)}, C_l^{(2)})$, representing the distance between any pair of clusters across the two partitions,

Algorithm 1: MOTT Algorithm

Input: For each of N samples: cluster proportion vector $\mathbf{q}_i$, $i = 1, ..., N$, covariance matrix, and mean vector of each cluster. Let the number of pairs of samples be $T = N(N-1)/2$.

Output: Taxonomy of meta-clusters

Set $\lambda = .075$;

for $i \leftarrow 1$ **to** T **do**

  Compute the cost matrix $\mathbf{D}$;

  Compute $\mathbf{W}$ using OT-RMC based on $\mathbf{D}$;

  Standardize $\mathbf{W}$ to $\widetilde{\mathbf{W}}$;

Form the similarity matrix $\mathbf{B}$ by tiling the matching weight matrices $\widetilde{\mathbf{W}}$;

Compute the dissimilarity matrix, $\mathbf{A} = -\log(\mathbf{B})$;

Build a dendrogram using the Ward's minimum-variance criterion; cut the dendrogram to generate the desired number of meta-clusters.

which is defined as the cost of matching the two clusters. Hence, $\mathbf{D}$ is the aforementioned cost matrix. To determine the matching weights between clusters in the two samples, we employ OT-RMC, formulated as follows. Let *gap vector* $\mathbf{g}_1 = (g_{1,1}, ..., g_{1,n_1})^T$ and $\mathbf{g}_2 = (g_{2,1}, ..., g_{2,n_2})^T$. Let $\mathbf{g}$ be the concatenation of $\mathbf{g}_1$ and $\mathbf{g}_2$. Roughly speaking, the gap vector $\mathbf{g}$ indicates the extent of deviation from the marginal constraints, $\mathbf{q}_1$ and $\mathbf{q}_2$, in OT (a review of basic OT is provided in Section 1 in **S1 Appendix**). Nonzero values in $\mathbf{g}$ are penalized to ensure that OT-RMC does not degenerate into trivial solutions. Depending on the penalty function used for $\mathbf{g}$ and whether and how upper and lower bounds are set up for the variables in the optimization problem, OT-RMC can have many variations. In our case here, the penalty for the gap variables is the L_1 norm denoted by $L(\mathbf{g})$. The optimization problem is presented in Problem (1) below.

$$\begin{aligned} R(\mathbf{D}, \mathbf{q}_1, \mathbf{q}_2) \quad &= \min_{\mathbf{W},\mathbf{g}} \langle \mathbf{D}, \mathbf{W} \rangle + \lambda L(\mathbf{g}) \\ \text{s.t.} \quad & \mathbb{1}_{n_1}^T \cdot \mathbf{W} \cdot \mathbb{1}_{n_2} = 1\,, \quad \mathbf{W} \geq 0 \\ & \mathbf{q}_1 - \mathbf{g}_1 \leq \mathbf{W} \cdot \mathbb{1}_{n_2} \leq \mathbf{q}_1 + \mathbf{g}_1 \\ & \mathbf{q}_2 - \mathbf{g}_2 \leq \mathbf{W}^T \cdot \mathbb{1}_{n_1} \leq \mathbf{q}_2 + \mathbf{g}_2 \end{aligned} \tag{1}$$

Denote the optimal $\mathbf{W}$ for Problem (1) above by the matching weight matrix $\mathbf{W}^*$. Matrix

$\mathbf{W}^* = (w^*_{k,l})$ specifies the matching proportion assigned to each pair of clusters $C_k^{(1)}$ and $C_l^{(2)}$ after optimizing the matching cost. The proportion $\tilde{q}_k^{(1)} = \sum_{l=1}^{n_2} w^*_{k,l}$ is called induced proportion given to $C_k^{(1)}$ and $\tilde{q}_l^{(2)} = \sum_{k=1}^{n_1} w^*_{k,l}$ is the induced proportion given to $C_l^{(2)}$. If it is overly costly to match a cluster with any cluster in the other partition, this cluster can be assigned with zero proportion, that is, $\tilde{q}_k^{(1)} = 0$ or $\tilde{q}_l^{(2)} = 0$.

To define the distance $D(C_k^{(1)}, C_l^{(2)})$, we first compute the squared Wasserstein distance between two Gaussian distributions characterizing clusters $C_k^{(1)}$ and $C_l^{(2)}$ respectively. Suppose the Gaussian distribution representing $C_k^{(i)}$ is $N(\mu_k^{(i)}, \Sigma_k^{(i)})$, where $\mu_k^{(i)}$ is the mean vector and $\Sigma_k^{(i)}$ is the covariance matrix. Then

$$D_W^2(C_k^{(1)}, C_l^{(2)}) = ||\mu_k^{(1)} - \mu_l^{(2)}||_2^2 + \mathbf{Tr}\left[\Sigma_k^{(1)} + \Sigma_l^{(2)} - 2\left(\Sigma_k^{(1)\frac{1}{2}}\Sigma_l^{(2)}\Sigma_k^{(1)\frac{1}{2}}\right)^{\frac{1}{2}}\right].$$

For any cluster, we estimate the mean $\mu_k^{(i)}$ by its sample mean and $\Sigma_k^{(i)}$ by its sample covariance matrix. We then standardize $D_W(C_k^{(1)}, C_l^{(2)})$ by dividing each entry indexed by $k = 1, ..., n_1$ and $l = 1, ..., n_2$ by the maximum entry. Define the standardized matrix as the cost matrix $\mathbf{D} = (D(C_k^{(1)}, C_l^{(2)}))_{k=1,...,n_1, l=1,...,n_2}$. The standardization used to define $\mathbf{D}$ makes λ less dependent on the data. How to select λ in our experiments is explained in the Results section.

We assume each cluster can be reasonably approximated by a Gaussian distribution for analytical tractability, an assumption shown to be adequate in many practical settings. Although not universally accurate, this assumption is often empirically supported in single-cell data following standard preprocessing steps such as normalization and dimensionality reduction (e.g., PCA), which tend to produce clusters with roughly ellipsoidal shapes, especially when clusters contain many cells. We acknowledge that for small clusters, Gaussian estimation—especially of the covariance matrix—can be unstable. In such cases, regularized estimators, such as shrinkage toward a diagonal matrix, can be used. More generally, this is a problem of estimating multivariate densities, for which a rich literature exists. Although we adopt Gaussian modeling in this work, the broader taxonomy framework is compatible with non-Gaussian distributions; the only required modification is in how inter-cluster distances are computed. For example, if sufficient data are available, clusters can be modeled as Gaussian mixtures, and distances computed using minimized aggregated Wasserstein (MAW) distances between GMMs [8, 31].

3.3 Labeling based on Alignment with a Reference

In this subsection, we introduce an alternative approach to label the clusters of multiple samples based on alignment with a pre-chosen reference. We refer to this method as *Reference Alignment* (RA).

The standard approach for labeling clusters in a sample involves pairwise matching [15, 23]. With pairwise matching, clusters in one sample are treated as the ground truth, and clusters in another sample are labeled by aligning them with the clusters in the reference sample. The sample providing the cluster labels is called the reference. By aligning the clusters of every other sample with those of the reference sample, coherent labels can be assigned across all samples, thereby constructing a taxonomy. This process can be repeated with different reference samples. Specifically, we implement the following process to provide a baseline for comparison with the MOTT system.

The samples chosen as references are those with the highest number of unique cell types. A reference sample may not contain all the ground-truth clusters present in the other samples. We then use a reference sample to label the clusters of the rest of the samples. Suppose the reference sample has n_1 clusters and the sample to be aligned has n_2 clusters. OT, OT-RMC, and POT all yield a $n_1 \times n_2$ weight matrix, $\mathbf{W}$. According to the way we select a reference sample, $n_1 \geq n_2$. The assigned label for the lth cluster in the non-reference sample is determined by $k^* = \arg\max_k W_{k,l}$.

For evaluation, we can repeat the process using different samples as the reference and measure the performance by averaging results obtained using different references.

3.4 Sample-level Classification

We evaluate a taxonomy by comparing it with the true annotations of cell types, if provided in the data, and by computing the accuracy of classifying samples based on feature vectors extracted using the taxonomy. The latter is an indirect approach, assessing the effect of the taxonomy on a downstream analysis task. Once a taxonomy is established, we assign coherent labels to clusters in all samples. We then compute the proportions of cells belonging to each cluster label and use these proportions as features for classifying samples. Thus, the dimension of the feature vector corresponds to the number of distinct cluster labels. Note that even

if the true types of cell clusters are unavailable, for instance, in the case of computationally generated clusters, an automatically generated taxonomy will still allow for coherent labeling of cell clusters and, consequently, enable sample-level classification. In our experiments, we test different approaches to constructing the taxonomy, some serving as baselines for comparison, and apply Random Forest (RF) to classify samples.

4 Results

4.1 Experimental Setups

We evaluate MOTT in terms of cell type accuracy for cell clusters and classification accuracy for samples. We tested it on 11 scRNA-seq datasets. Most datasets do not divide samples into categories or provide labels for all samples. Therefore evaluation by sample-level classification is conducted only on 4 datasets. For the tested datasets, different samples typically correspond to different patients, though in some cases, they correspond to replicates recorded at different times. The task of constructing a taxonomy becomes more interesting and challenging with a large number of samples. Unfortunately, many scRNA-seq datasets lack metadata to allocate cells to different samples. In other cases, while the dataset might include sample division information, the number of samples might be too few (2 to 5). To effectively examine our new method, we generated simulated partitions of cell-level data to create "*simulated samples*" (SS). Since simulated samples are not subject to batch effects (which commonly occur in scRNA-seq data), differences in results by various methods can be better attributed to the characteristics of the methods.

For baseline methods, we consider alternatives to MOTT from two perspectives: using OT and POT instead of OT-RMC for computing matching weights between clusters, and using RA instead of SA for aligning all clusters. Hence, we compared a total of six methods. If cluster matching by OT-RMC in MOTT is replaced by OT or POT, we name the method OT-SA and POT-SA respectively. If SA is replaced by RA, we name the method OT-RMC-RA, POT-RA, or OT-RA, depending on which cluster matching method is used. Finally, the method OT-RMC-SA corresponds to MOTT. The only natural baseline method from existing literature is OT-RA. However, since the novelty of MOTT comes from employing both OT-RMC and SA,

we have investigated all six methods.

To apply MOTT, we start with the cell clusters provided for each sample. The 11 datasets we tested include ground-truth cell types, which are used to form cell clusters within each sample. To perform cluster matching using OT or OT-RMC, as explained in the previous section, we estimate a Gaussian distribution in a reduced-dimensional space for each cluster based on all the cells it contains. To reduce dimension, we first select highly variable genes using the R tool *HVGs*. We then apply Principal Component Analysis (PCA) and select the number of principal components needed to account for 90% of the variability.

4.2 Evaluation Measures

Suppose a taxonomy for cell clusters has been established, specifically generating meta-clusters. All clusters in the same meta-cluster are regarded as coming from the same cell type. To compare with the ground truth, we assign a cell type to each meta-cluster by majority vote. In particular, the most common cell type among the clusters in a meta-cluster is taken as the type of the meta-cluster. Ideally, if all clusters in one meta-cluster are of the same type, the taxonomy would have consistently labeled them. Otherwise, errors occur in labeling the cell types of some clusters.

We then compute the accuracy of labeling cell clusters in two ways. The cluster-level accuracy, denoted by ζ_{cls}, is defined as the percentage of clusters (across all samples) for which the ground-truth cell type matches the cell type indicated by the meta-cluster. The cell-level accuracy, denoted by ζ_{cell}, is defined as the percentage of cells that are assigned with correct cell types based on the meta-clusters. Essentially, ζ_{cell} is a weighted average version of ζ_{cls}, where the weight of a cluster is proportional to the number of cells contained in it, meaning that larger clusters have a greater impact on the overall accuracy calculation.

In addition, we compute the Adjusted Rand Index (ARI) for cell clusters based on ground-truth cell labels and the partition of cells based on meta-clusters. While calculating the ARI, cells from all samples are pooled together. Specifically, cells from all samples with the same ground-truth cell labels or meta-cluster labels are grouped into the same cluster.

4.3 scRNA-seq Datasets

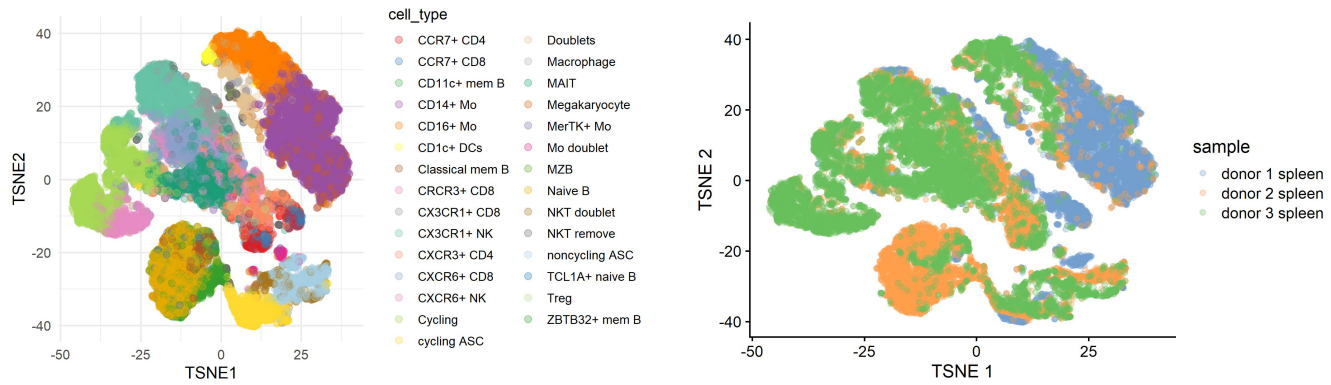
We used 11 data sources: *Segerstolpe* [41], *Bacher* [4], *Baron* [5], *Zhao* [56], *Tasic* [45], *Zilionis Human* [57], *Zilionis Mouse* [57], *Lawlor* [26], *He Organ* [14], *Jessa Brain* [19], and *Wu* [49]. Two datasets were formed from the Jessa Brain data based on different regions of the brain (pons versus forebrain). Table 1 provides basic information about the datasets. Notably, two datasets, Zhao and Zillions Human, include two sets of ground-truth labels: one coarse (representing major cell types) and one fine (representing cell subtypes). We experimented with both sets of ground-truth labels. More detailed descriptions of the datasets and data preprocessing are provided in Section 2 of **S1 Appendix**. Figure 2 presents a pair of t-SNE plots for two example datasets, where the data points are colored according to their ground-truth cell types and the samples they belong to. The plots show the complexity of the datasets, as the cell population sizes vary significantly and batch effects are present in both.

As aforementioned, some datasets do not divide cells into samples. For this reason, and to facilitate performance comparisons, we generated simulated samples by randomly splitting the entire dataset. Unless specified otherwise, an original dataset was randomly partitioned into 20 subsets, each treated as one sample. For the Segerstolpe, Zilionis Mouse, and Wu datasets, cells were collected under multiple conditions, and we generated samples separately by randomly dividing cells under each individual condition. We refer to such simulated samples as "simulated by label" (SL). Note that the cell condition is different from the cell type; instead, the cell condition corresponds to the class label assigned to each sample.

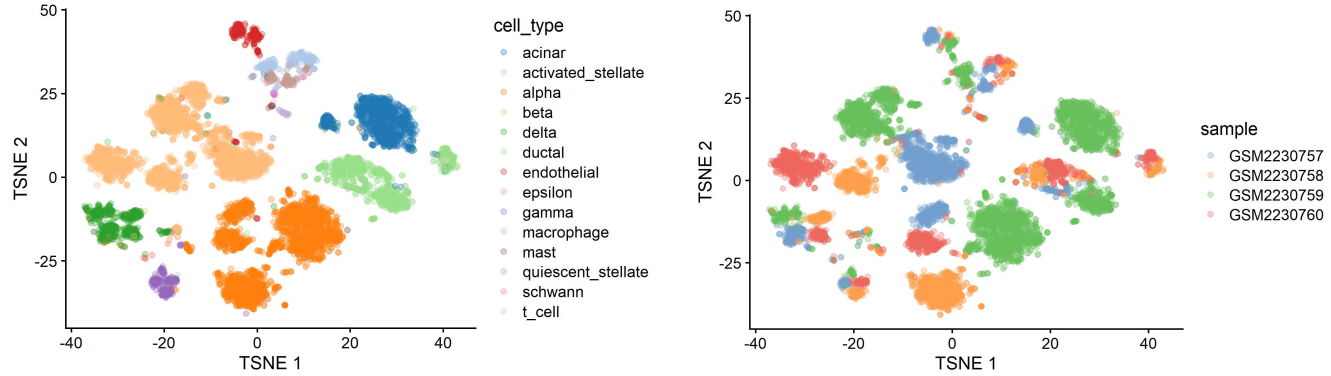
Sample-level classification was conducted on the Segerstolpe, Bacher, Zilionis Mouse, and Wu datasets. Specifically, samples in Segerstolpe were labeled in the original data with two categories: patients with type II diabetes mellitus and healthy donors (4 versus 6). Similarly, the Bacher samples were originally labeled as patients previously exposed to a virus or those unexposed (14 versus 6). For the Zilionis Mouse and Wu datasets, sample division was not indicated in the original data, but cells come from two conditions: healthy versus lung tumor or diseased. Simulated samples were generated by randomly dividing either diseased or healthy cells (the SL scheme), and the samples were thus labeled accordingly as either diseased or healthy (10 versus 10). Additionally, the Wu dataset lacks ground-truth cell types; instead, hierarchical clustering into 10 clusters was used to simulate cell types.

Dataset	#Smp	C. Types	Cells	Genes	PCs	Run Time (sec)
Segerstolpe	10	11	2,056	24,454	50	51.655
Bacher	20	6	104,417	33,538	50	58.795
Baron	4	14	8,569	18,845	50	13.882
Zhao	9	4 & 29	60,672	33,694	8	.952 & 16.523
Tasic	(20)	7	1,727	24,058	50	68.302
Zilionis Human	7	21 & 35	54,773	41,861	50	80.272 & 194.854
Zilionis Mouse	(20)	6	15,905	28,205	50	66.672
Lawlor	8	7	617	26,616	50	11.261
He Organ	15	42	76,037	12,021	20	81.899
Jessa Pons	(20)	27	27,954	27,998	2	44.039
Jessa Forebrain	(20)	27	33,641	27,998	2	37.559
Wu	(20)	(10)	18,249	17,542	9	7.251

Table 1: Summary information on the scRNA-seq datasets. For each dataset, we list the numbers of samples, cell types, cells, genes, and principal components (PCs) retained in the analysis. In addition, the computation time for generating the taxonomy is reported. For some datasets, although cells might have been collected from different donors or organs, information on how to divide the cells into samples is lacking. In such cases, we artificially generated a certain number of samples, which is indicated in parenthesis. For two of the datasets, cells are categorized at both a coarse and a fine level, with more cell types annotated at the fine level. For Wu data, the original dataset provided no cell types. We applied hierarchical clustering to this dataset to generate 10 clusters. Reported time assumes precomputed cluster means and covariances; it includes only OT-RMC computation and dendrogram clustering.



(a)



(b)

Figure 2: t-SNE plots showing data grouped by ground-truth cell types (left column) and by sample (right column). (a) Zhao Spleen dataset, (b) Baron Pancreas dataset. The Zhao Spleen dataset, which only includes cells from the spleen, was used instead of the entire Zhao dataset because the latter contains over $60,000$ points, making it too large for effective visualization. The composition of cell types in different samples, both in terms of existing cell types and their population sizes, varies substantially. This observation highlights the challenge of constructing a taxonomy for cell clusters across samples.

4.4 Accuracy of Taxonomy

4.4.1 Standard Usage Scenario

As an example, Figure 3 presents the taxonomy created by MOTT for all the ten samples from the Segerstolpe dataset. For clarity, we do not display results for all samples, as the resulting dendrogram would be too large.

To numerically assess the taxonomy results, we calculate the cluster-level accuracy ζ_{cls} and

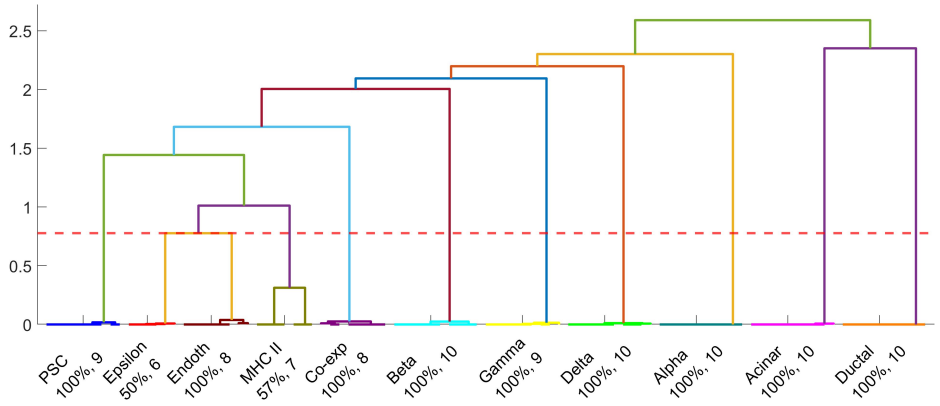


Figure 3: An example taxonomy created by MOTT for the Segerstolpe dataset. The Segerstolpe dataset includes ten samples and, collectively, 11 cell types. Correspondingly we use the taxonomy to generate 11 meta-clusters. Each meta-cluster is labeled according to the most common cell type among its constituent clusters. Beneath each meta-cluster, we label its chosen cell type, and in the second line, show the number of clusters in this meta-cluster and the percentage of clusters assigned with the correct cell type. Cell type "Endoth" stands for Endothelial, "Co-exp." stands for Co-expression, and "MHC II" stands for "MHC Class II". Only two meta-clusters, Epsilon and MHC Class II, include cluster cell types other than their chosen labels. The horizontal dashed line indicates the cut-off level at which eleven meta-clusters are formed.

cell-level accuracy ζ_{cell} in terms of labeling cell clusters. Figure 4 presents the results achieved by the four methods: OT-RMC-SA, OT-SA, OT-RMC-RA, and OT-RA. When RA is used, the results shown are average values over all valid reference samples, specifically those containing the maximum number of cell clusters. For clarity, we present the results obtained from datasets with real samples in one plot (Figure 4(a)) and those from simulated samples in another plot (Figure 4(b)). In Figure 4(a), we use fine ground-truth labels for cell types in the Zhao and Zillions Human datasets. To evaluate the impact of coarse versus fine labels on the taxonomy's performance, we compare the results obtained from these two sets of labels in Figure 4(c).

As shown in Figure 4(a) and (b), MOTT (i.e., OT-RMC-SA) consistently outperforms the other methods across both real and simulated sample datasets, regardless of the evaluation measures used. The only exceptions are for the Segerstolpe and Lawlor pancreas data where POT-SA does only slightly better than MOTT. In most cases when MOTT performs best, OT-RMC-RA follows as the second-best performing method, followed by POT-SA. When comparing the cluster matching schemes, OT-RMC versus OT, the former consistently yields better results than the latter in both RA and SA cases, with the disparity in performance often being pronounced. This indicates that the marginal constraints imposed by OT can be detrimental for

these datasets, which frequently include samples with different cell types, as OT fundamentally eliminates the possibility of new clusters. When comparing the cluster alignment schemes, SA versus RA, OT-RMC-SA almost always outperforms OT-RMC-RA, with the difference being remarkable for many datasets. This suggests that it is advantageous to simultaneously align clusters rather than performing alignment with a pre-selected reference. Figure 4(a) and (b) also demonstrate that building a taxonomy with OT-RMC or POT often significantly improves performance. Indeed the only instance when a taxonomy does not help is with the Jessa Pons data (having simulated samples) and using POT.

The type of dataset significantly affects the performance of the methods. For instance, the performance gain of OT-RMC-SA is weakest and marginal on the He Organ dataset. This dataset is unique in that the samples correspond to different human organs from a single patient, in contrast to samples from different patients. Additionally, the four methods perform better on simulated samples than on real samples. One likely reason is that the simulation of samples sidesteps the issue of batch effects [33]. As shown by Figure 4(b), unlike the real sample datasets, and the Tasic and Wu datasets, Z. Mouse, Jessa Pons and Jessa Forebrain yield less pronounced performance differences among the methods. Except for the Jessa Pons and Jessa Forebrain datasets, OT-SA yielded the worst performance across all three metrics by significant margins.

In Figure 4(c), we investigate the impact of smaller cluster sizes on the performance. Specifically, for three datasets, there are two sets of cell type labels: coarse versus fine. The fine cell types are obtained by dividing coarse cell types into subgroups. As a result, cell clusters are smaller under the fine type labels. These three datasets are the Zhao, Zilionis Human, and Zilionis Human SS-40. The Zilionis Human SS-40 dataset contains the same cells from the Zilionis Human dataset but the division into 40 samples is randomly generated. In all three datasets we see the performance across the three metrics being higher when coarse cell types are used. Figure 4(c) also shows that the gap in performance between the fine and coarse cell types to be much greater for the OT-RA and OT-SA methods relative to the other methods. For all three datasets we observe that OT-RMC-SA and POT-SA are the least sensitive methods, with OT-RMC-SA notably more robust on the Zhao dataset.

It is interesting to examine whether OT-RMC-SA is more robust than other methods when samples in a dataset deviate from ideal conditions. As previously mentioned, simulated samples

created from one dataset represent ideal samples, with variation attributed only to randomness rather than any systematic cause. However, non-ideal conditions in real samples arise due to the presence of batch effects and small cell populations. We formed five pairs of datasets for comparison. Each pair consists of a real sample dataset and a simulated sample dataset created from the same source, such as the Segerstolpe dataset and its version with 20 simulated samples. We computed the differences in ARI, cell-level accuracy ζ_{cell}, and cluster-level accuracy ζ_{cls} between the results from the simulated and real sample datasets. A positive difference indicates better performance on the simulated data. In Figure 5, the boxplots illustrate the differences obtained from the five pairs of datasets for each method and each performance measure. We observe that for all methods, the performance difference is predominantly positive. This aligns with our expectation: it is easier to construct a taxonomy when variation between samples is solely due to randomness. Additionally, we note that OT-RMC-SA, OT-RMC-RA and POT-SA are substantially less sensitive to the different ways samples are generated, with OT-RMC-SA performing slightly more robustly than OT-RMC-RA and POT-SA according to the median and range of ζ_{cell} and ζ_{cls}.

The Bacher dataset exhibits pronounced batch effects between its 20 samples. To verify whether MOTT is compromised, we compare the performance on the uncorrected data and then on data corrected with the established *fastMNN* algorithm from the *Batchelor* package [32]. Only marginal differences are observed: the Adjusted Rand Index (ARI) rose from 0.853 to 0.863, while the cell-level accuracy ζ_{cell} decreased slightly from 0.993 to 0.949; the cluster-level accuracy ζ_{cls} increased from 0.932 to 0.940.

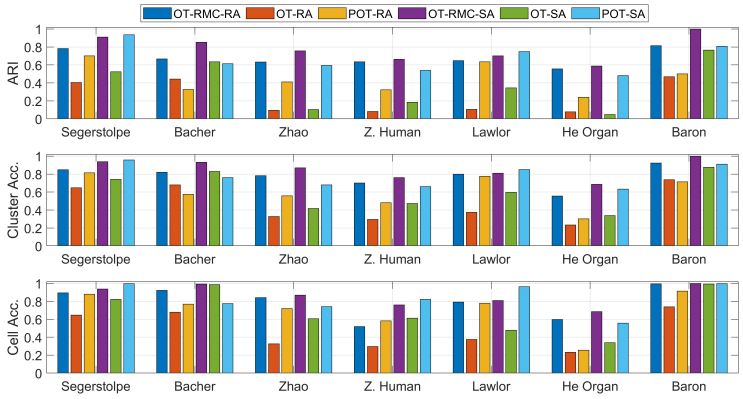
For results in Figure 4 and 5, a total of 20 datasets are used in the alignment experiments. To determine whether OT-RMC-SA demonstrates superior overall performance across all datasets and metrics (ARI, ζ_{cls}, ζ_{cell}), we conducted a non-parametric analysis of 60 unique dataset-metric cases (20 datasets $\times$3 metrics). Within each case, we ranked all six methods (rank 1 means best), then performed a Friedman test that rejected the global null hypothesis of equal performance with p-value ($< 10^{-10}$). Subsequent one-sided Wilcoxon signed-rank tests (with Holm-Bonferroni correction) specifically compared OT-RMC-SA against each competitor (the alternative hypothesis is that OT-RMC-SA is better than any other method). The result demonstrates universal statistical significance: all five adjusted p-values are below 10^{-4}, confirming OT-RMC-SA's significantly superior performance against every alternative method.

The average ranks for the methods OT-RMC-SA, OT-RMC-RA, POT-SA, OT-SA, POT-RA, and OT-RA are respectively 1.30, 2.64, 2.68, 4.35, 4.65, 5.38.

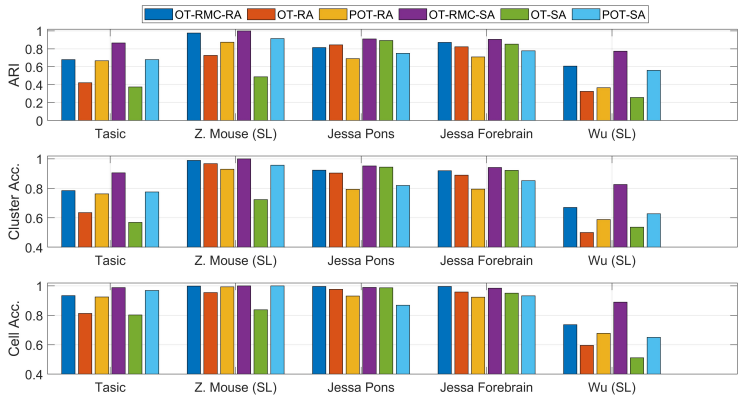
4.4.2 Alignment of Cell Types at Different Resolutions

Next, we examine how OT-RMC-SA performs when samples are clustered at different resolutions. For two datasets—Zilionis Human and Zhao—each sample is grouped into clusters at both a coarse and a fine resolution. An ideal outcome for a taxonomy would yield meta-clusters that each contain a single unique coarse-level cluster type, along with all of its corresponding fine-level clusters as sub-types. We applied MOTT to these two datasets, using all samples with cluster labels at both resolutions. For the Zhao data, all fine-level cell-type labels are fully nested within the coarse-level cell-type labels; that is, each fine-level label is associated with exactly one coarse-level label. This was not the case for the Zilionis Human data, where some fine-level labels did not correspond clearly to any single coarse-level label. To avoid confusion in the evaluation, we excluded cells from all samples that had fine-level labels not predominantly associated with a single coarse-level label (defined as fewer than 70% of their cells being assigned to the same coarse-level label).

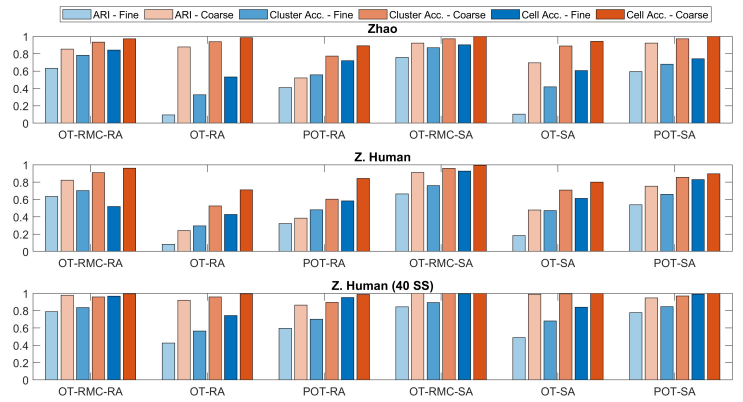
We set the number of meta-clusters in the taxonomy to match the known number of cell types at the coarse resolution. To measure performance, we assigned a coarse-level label to each meta-cluster using majority vote based on the coarse-level clusters it contains. We then calculated the proportion of clusters (coarse or fine) in the meta-cluster that had a coarse-level label different from the assigned label. Since each fine-level label is associated with a corresponding coarse-level label, we refer to this proportion as the error rate. Finally, we averaged the error rate across all meta-clusters. OT-RMC-SA achieves average error rates of 0.053 and 0.049 for the Zilionis Human and Zhao data, respectively. In comparison, POT-SA produces average error rates of 0.122 and 0.032. OT-SA failed to associate fine-level clusters with their respective coarse-level clusters, resulting in meta-clusters containing either all coarse-level clusters or all fine-level clusters.



(a) Performance on datasets with real samples



(b) Performance on datasets with simulated samples



(c) Comparison of Performance: fine versus coarse cell types

Figure 4: Accuracy achieved by four methods for identifying cell types based on the taxonomy. OT-RMC-SA corresponds to the MOTT system. The accuracy is measured by ARI, cluster-level accuracy ζ_{cls}, and cell-level accuracy ζ_{cell}. "SS" stands for simulated samples. The number of simulated samples in a dataset has a default value of 20 unless specified in the parenthesis. If "SL" is indicated, the samples were simulated by randomly dividing cells within different cell conditions separately; otherwise, the samples were generated by randomly dividing the entire data. Z. Human stands for Zilionis Human and Z. Mouse stands for Zilionis Mouse. Some datasets had two sets of ground-truth labels; coarse indicates major cell types and fine cell subtypes.

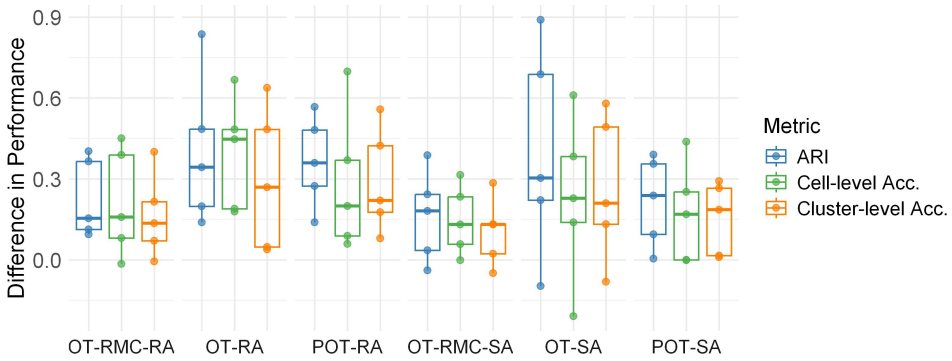


Figure 5: Boxplots showing the differences in performance between pairs of real sample and simulated sample datasets. Five datasets are analyzed: Segerstolpe, Zhao (fine cell types), Zilionis Human (fine cell types), He Organ, and Baron. Each simulated dataset (generated by the SS scheme) consists of 20 samples, except for the Zilionis Human dataset, which consists of 40 samples. Performance is evaluated using three metrics: ARI, cluster-level accuracy ζ_{cls}, and cell-level accuracy ζ_{cell}. A positive difference indicates better performance on the simulated sample datasets. In nearly all cases across different dataset pairs and methods, the simulated sample datasets show superior performance compared to their real sample counterparts. Notably, OT-RMC-RA and OT-RMC-SA exhibit smaller medians and ranges, indicating that these methods are less sensitive to how samples are generated.

4.4.3 Alignment across Species

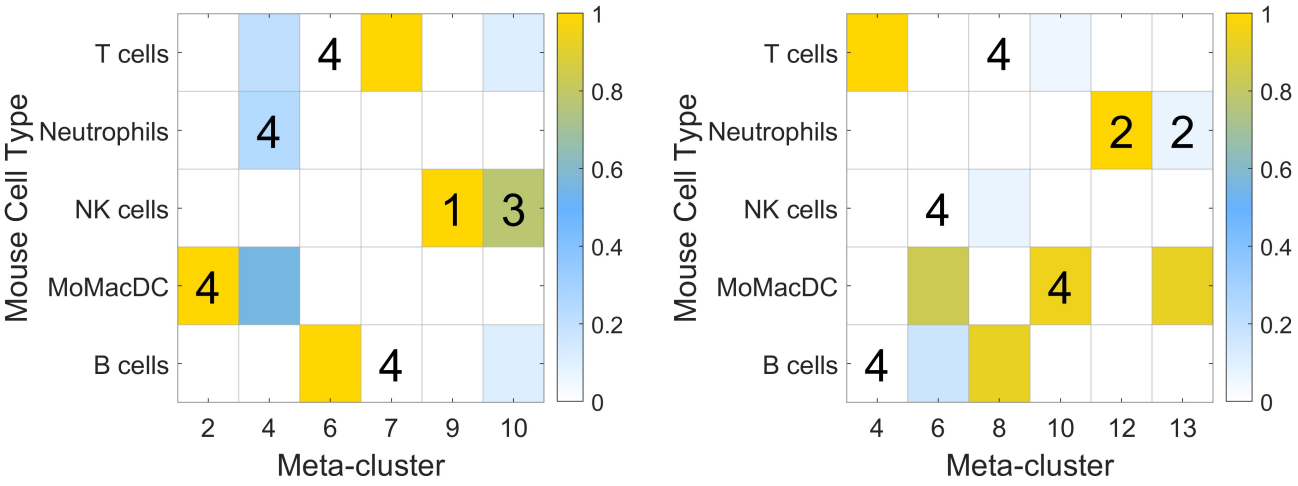
We also evaluated how OT-RMC-SA performs when aligning samples from two different species. The Zilionis Human and Zilionis Mouse datasets, which contain human and mouse lung cells respectively, are suitable for this experiment. The mouse dataset shares five coarse-level cell types with the human dataset: B cells, T cells, MoMacDC, NK cells, and Neutrophils. While we use the coarse human labels, they are still more refined than the labels provided for the mouse cells—21 human cell labels in total, with subtypes such as two B cell subtypes and six T cell subtypes. In total, there are 14 unique cell types across the two species: 7 appear only in human, 2 only in mouse, and 5 are shared. Thus, this alignment task spans both cross-species and cross-resolution scenarios.

The mouse dataset includes four samples, so each mouse cell type is represented by four clusters. With 14 unique cell types, we apply OT-RMC-SA to produce 14 meta-clusters. To visualize and compare results, we present heatmaps for OT-RMC-SA and POT-SA in Figure 6. Each row corresponds to a mouse cell type and each column to a meta-cluster. For example, OT-RMC-SA places all four mouse MoMacDC clusters into meta-cluster 11, indicated by a '4' in the corresponding heatmap cell. To assess cross-species alignment, we also show—for each

mouse cell type—the proportion of human clusters of the same type in each meta-cluster using a color scale. Ideally, each mouse cell type would align to a meta-cluster containing only human clusters of the same type.

The results indicate that T cells and B cells are easily confused across species—a phenomenon also noted in prior studies [1, 43]. For both OT-RMC-SA and POT-SA, mouse T cells are grouped with meta-clusters dominated by human B cells, and vice versa. However, OT-RMC-SA aligns NK cells and MoMacDC across species more accurately than POT-SA. For instance, POT-SA assigns all mouse NK cell clusters to meta-cluster 6, which contains no human NK cells. In contrast, OT-RMC-SA places all four mouse NK cell clusters into meta-clusters 9 and 10, which also contain a high proportion of human NK cell clusters.

We also ran alignment using OT-SA. Because OT assumes all cell types are shared across samples, it is poorly suited to this scenario with species-specific cell types. Indeed, OT-SA collapses all 20 mouse clusters corresponding to the five shared cell types into a single meta-cluster, demonstrating its limitation in handling unshared cell types.



(a) OT-RMC-SA (b) POT-SA

Figure 6: Cross-species alignment results using OT-RMC-SA and POT-SA, visualized as heatmaps. Each row corresponds to a mouse cell type, and each column to one of the 14 meta-clusters. The number in each heatmap cell indicates how many of the four mouse clusters for that cell type are assigned to the corresponding meta-cluster. The background color represents the proportion of human clusters within that meta-cluster that belong to the same cell type. The color scale is shown as a vertical bar on the right-hand side.

4.4.4 Alignment of Unsupervised Clusters

The experimental protocol described above relied on ground-truth annotations to delineate the cell clusters present in each sample. As a complementary investigation, we adopted a fully unsupervised pipeline in which an independent clustering model is fitted to every sample and the resulting partitions are subsequently aggregated with *MOTT* to obtain meta-clusters. Purity for each meta-cluster is then evaluated against the ground-truth label of every constituent cell. We carried out this study on the Baron and Segerstolpe datasets, employing a finite Gaussian-mixture model implemented in the *Mclust* package for R. The number of components for each sample was selected by maximizing the Bayesian Information Criterion (BIC). In the absence of prior knowledge of the number of cell types, the meta-cluster budget was chosen as the maximum number of mixture components observed across all samples, which is 13 for Baron and 12 for Segerstolpe. We experimented with replacing *Mclust* by linkage-based (*hclust*) and graph-based (*igraph*) clustering methods, and obtained similar results (omitted here for brevity).

Since the clusters generated by Mclust do not correspond directly to true cell types, evaluating the resulting meta-clusters is less straightforward. We adopt the previously defined measures ζ_{cls} and ζ_{cell}. Specifically, we assign each unsupervised cluster a cell-type label by majority vote based on the types of the cells it contains, and evaluate the resulting meta-clusters as in prior analyses. In addition, recognizing that unsupervised clusters may not align perfectly with the true cell types, we introduce a third measure, Purity Score (denoted by ζ_{unsup}). For each cluster, the purity score is defined as the proportion of its cells that belong to the majority cell type. To assess the overall clustering, we compute a weighted average of the purity scores across all clusters, where weights are proportional to cluster sizes. As an oracle upper bound, we consider the purity score obtained by directly applying majority vote to each cluster produced by the unsupervised algorithm. This value represents the highest possible purity achievable by any meta-clustering method and depends on the clustering accuracy of the unsupervised algorithm for a given dataset. The oracle upper bounds for the Baron and Segerstolpe datasets are 0.91 and 0.69, respectively. Results are shown in Table 2, which indicate that MOTT performs worse than POT on Baron and similarly on Segerstolpe.

Dataset	MOTT			POT		
	ζ_{cls}	ζ_{cell}	ζ_{unsup}	ζ_{cls}	ζ_{cell}	ζ_{unsup}
Baron	0.89	0.86	0.85	0.94	0.95	0.89
Segerstolpe	0.77	0.88	0.65	0.76	0.87	0.66

Table 2: Performance of MOTT and POT, evaluated using ζ_{cls}, ζ_{cell}, and ζ_{unsup}, on two datasets with clusters generated by Mclust, an unsupervised clustering algorithm.

4.5 Hyperparameter Selection, Stability Analysis, and Computational Complexity

MOTT requires one pre-selected hyperparameter, λ. We determined this hyperparameter empirically for OT-RMC by varying λ from 0.005 to 0.15, finding that the optimal value consistently falls around 0.075. Consequently, we set $\lambda = 0.075$ for the results reported above.

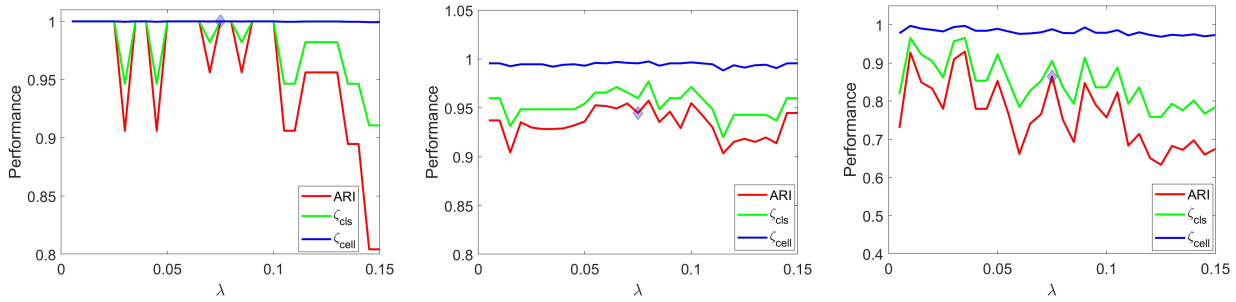
If the ground-truth number of cell types is unknown, we must select the number of meta-clusters. In practice, one may set this number to the average number of clusters across samples, or to the maximum number of clusters observed in any single sample. Additionally, established methods for determining optimal dendrogram cut points can be employed. For example, the dynamic tree cut method by [25] provides a data-driven approach for selecting the number of clusters based on the dendrogram structure.

In Fig. 7, we show the cell type labeling accuracy achieved by OT-RMC-SA across three datasets as a function of λ, evaluated using ζ_{cls}, ζ_{cell}, and ARI. The performance remains stable as long as λ is not too large. We attribute the low sensitivity to λ to the normalization of the cost matrix, which makes the algorithm more robust to this parameter. Therefore, we recommend normalizing the cost matrix and then setting $\lambda = 0.075$.

To assess the stability of MOTT, we generated 15 bootstrap samples of the Baron Pancreas dataset and evaluated MOTT on each. For the original dataset, ARI, ζ_{cls}, and ζ_{cell} all equal 1. For the 15 bootstrap samples, ζ_{cls} is consistently 1; ζ_{cell} equals 1 for 10 samples and 0.956 for the other 5; ARI is 1 for 10 samples and 0.982 for the other 5. For the more challenging data of Zhao with fine labels, ARI, ζ_{cls}, and ζ_{cell} are .757, .869, and .903 respectively. For the 15 bootstrap samples the average ARI, ζ_{cls}, and ζ_{cell} are .728, .847, and .891 respectively; and the

corresponding standard deviations are .021, .014, and .03 respectively.

To evaluate the computational complexity of MOTT, we report the computation time for each dataset used in our experiments (Table 1). The computation time scales quadratically with the total number of clusters across all samples and linearly with the data dimensionality involved in computing distances between clusters. The reported time assumes that the mean vector and covariance matrix for each cluster are precomputed, and includes only the time required for performing OT-RMC and generating meta-clusters via dendrogram clustering. All experiments were conducted using MATLAB R2024a on an Acer Aspire VN7-791 laptop with a 2.6 GHz Intel Core i7 CPU running Windows.



(a) Baron (b) Segerstolpe (c) Tasic

Figure 7: Cell type labeling performance achieved by OT-RMC-SA, as measured by cluster-level accuracy ζ_{cls}, cell-level accuracy ζ_{cell}, and ARI, with varying λ. Three datasets are analyzed: (a) Baron, (b) Segerstolpe, (c) Tasic.

4.6 Accuracy of Sample-level Classification

Next, we examine the effect of taxonomy on sample-level classification using the Segerstolpe, Bacher, Zilionis Mouse, and Wu datasets. With the exception of the Bacher dataset, we utilized simulated samples generated through the SL scheme. As explained in the previous section, the features used for classification are the proportions of cells in each cluster. An established taxonomy ensures the coherency of cluster labels across different samples. As a performance benchmark, we also compute the proportions of cells in each ground-truth cell type and use them as classification features. In practice, ground-truth labels are not expected to be available for every sample, which is the primary reason for constructing a taxonomy.

We experimented with the classification algorithms RF. Due to the small number of available

Method	Metric	Segers.	Bacher	Z. Mouse	Wu
OT-RMC-RA	Acc.	.474	.529	.931	.947
	AUC	.522	.507	.939	.95
OT-RA	Acc.	.711	.449	.951	.947
	AUC	.522	.507	.939	.95
POT-RA	Acc.	.579	.486	.938	.974
	AUC	.583	.414	.938	1
OT-RMC-SA	Acc.	.9	.6	1	1
	AUC	.99	.526	1	1
OT-SA	Acc.	.85	.5	.95	.95
	AUC	.87	.507	.939	.95
POT-SA	Acc.	.9	.45	1	1
	AUC	.98	.604	1	1
Ground Truth	Acc.	.85	.6	1	1
	AUC	.98	.531	1	1

Table 3: Sample-level classification accuracy on three datasets. Except for Bacher, we used simulated samples via the SL scheme. For the "Ground Truth" method, ground-truth cell type labels are used as the taxonomy. Classification algorithm RF is applied. Accuracy is computed based on the leave-one-out scheme and is reported in terms of the percentage of correctly classified instances (Acc.) and Area Under the Curve (AUC).

samples (no greater than 20), we computed classification accuracy using the leave-one-out scheme. For scRNA-seq datasets found in the literature, although the number of cells per sample can be large, the number of samples is consistently low. The results are shown in Table 3. The results demonstrate that, compared with the other methods for building a taxonomy, OT-RMC-SA (i.e., MOTT) and POT-SA yield the most accurate classification across datasets. The advantage of OT-RMC-SA and POT-SA is most prominent with the Segerstolpe dataset. It is also notable that the ground-truth taxonomy rarely outperforms OT-RMC-SA. For the Wu dataset, the classification task is straightforward, such that OT-RMC-SA, POT-SA, and the ground-truth taxonomy all achieve perfect accuracy. For the Bacher data, these three methods also achieve similar classification accuracies, with POT-SA slightly worse than the other two. However, POT-SA has a higher AUC compared to the other two methods. For the Segerstolpe dataset, however, OT-RMC-SA and POT-SA perform better than the ground-truth taxonomy. It is difficult to pinpoint why OT-RMC-SA or POT-SA can outperform the ground-truth taxonomy. It is possible that while constructing a taxonomy, OT-RMC-SA and POT-SA capture valuable information for subsequent classification.

5 Discussion

In this paper, we introduced a new method, called Multisample OT Taxonomy (MOTT), for constructing a hierarchical taxonomy of cell clusters from different samples. One of the major challenges in single-cell RNA sequencing is achieving consistent labeling of cell clusters across samples due to variations caused by random data noise or batch effects. Inconsistent annotations hinder our ability to accurately understand cellular diversity and to extract meaningful features for tasks such as disease classification. MOTT addresses these challenges by creating a taxonomy that identifies which cell clusters across samples represent the same intrinsic type and reveals similarities between cell types. Even without employing a batch effect removal tool, MOTT performs effectively and exhibits lower sensitivity to batch effects compared to other methods. Our experiments show that MOTT accurately annotates cell clusters and offers valuable insights for downstream analysis. While POT-SA achieves comparable performance to MOTT, it is often slightly outperformed. Notably, beyond the choice of the matching method, we observe that incorporating a taxonomy enhances performance on sample-level classification.

One limitation of our study is that we have only assessed the taxonomy's ability to identify cell clusters within the same cell type. While MOTT organizes these clusters hierarchically, establishing a foundation for defining similarity measures between cell types, our data lacks ground truth for evaluating any specific definition of similarity. As a result, we have not tested the taxonomy's effectiveness in quantifying cell type similarity.

Extending MOTT to other single-cell omics, such as scATAC-seq or spatial transcriptomics, is a promising direction. The main challenge lies in defining meaningful pairwise distances between clusters across different modalities, particularly when their feature spaces differ significantly. For multi-omics data collected from the same set of cells, one possible strategy is to define cluster-level dissimilarity based on shared cell membership across modalities. This shared structure can serve as a bridge for computing distances between clusters, enabling alignment. More broadly, adapting MOTT for multi-omics integration would require developing modality-agnostic or harmonized representations of clusters to ensure comparability. We believe this is a valuable extension.

The datasets considered in this investigation had a maximum of 100,000 cells across all samples. In this context, simpler non-foundational models are typically preferred for aligning clusters across scRNA-seq samples due to their computational efficiency, ease of use, and interpretability, while achieving comparable integration accuracy compared with foundational models at this scale. MOTT can be executed in a matter of seconds or minutes and requires no parameter tuning. Foundational models (e.g., scGPT/scBERT) become advantageous when handling massive datasets ($> 500,000$ cells), extreme batch effects, or rare cell types. However, for routine analyses involving fewer than 100,000 cells, they introduce unnecessary overhead with diminishing returns on accuracy [10, 51]. An extension of our method that might lead to better results involves using a foundational model to generate a robust embedding of the scRNA-seq data and then implementing MOTT.

Furthermore, exploring dynamic contexts beyond static sample collections presents an interesting avenue for future work. For instance, our current approach treats samples as independent, whereas they could be related over time (e.g., samples from the same patient collected at different periods) or through other factors. Incorporating such contextual information into the cluster alignment process—using methods like OT-RMC—could enhance performance by more effectively capturing the changes in cell clusters across related samples.

Data Availability

The source data used are in public domain in other publications in the reference. The processed data generated and analyzed during this study and the code implementation are available on Github: https://github.com/sebastian5799/Multisample-OT-Taxonomy

Acknowledgments

J. Li's research is supported by the National Science Foundation under grant CCF-2205004.

References

[1] T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. T. Reinders, and A. Mahfouz. "A comparison of automatic cell identification methods for single-cell RNA sequencing data". *Genome Biology* 2019 (2019), p. 194.

[2] J. Alquicira-Hernandez, A. Sathe, H. P. Ji, Q. Nguyen, and J. E. Powell. "scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data". *Genome Biology* 20 (2019).

[3] A. Azad and A. Pothen. "flowMatch: Cell population matching and meta-clustering in Flow Cytometry". *Bioconductor* (2022).

[4] P. Bacher, E. Rosati, D. Esser, G. R. Martini, C. Saggau, E. Schiminsky, J. Dargvainiene, I. Schröder, I. Wieters, Y. Khodamoradi, et al. "Low-avidity CD4+ T cell responses to SARS-CoV-2 in unexposed individuals and humans with severe COVID-19". *Immunity* 53.6 (2020), pp. 1258–1271.

[5] M. Baron, A. Veres, S. L. Wolock, A. L. Faust, R. Gaujoux, A. Vetere, J. H. Ryu, B. K. Wagner, S. S. Shen-Orr, A. M. Klein, D. A. Melton, and I. Yanai. "A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure." *Cell Systems* 3 (2016), pp. 346–360.

[6] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. "Iterative Bregman projections for regularized transportation problems". *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[7] K. Cao, Q. Gong, Y. Hong, and L. Wan. "A unified computational framework for single-cell data integration with optimal transport". *Nature Communications* 13.1 (2022), p. 7419.

[8] Y. Chen, J. Ye, and J. Li. "Aggregated Wasserstein distance and state registration for hidden Markov models". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.9 (2019), pp. 2133–2147.

[9] Z. A. Clarke, T. S. Andrews, J. Atif, D. Pouyabahar, B. T. Innes, S. A. MacParland, and G. D. Bader. "Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods". *Nature Protocols* 16 (2021), pp. 2749–2764.

[10] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". *Nature Methods* 21.8 (2024), pp. 1470–1480.

[11] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. "POT: Python Optimal Transport". *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8. URL: `http://jmlr.org/papers/v22/20-451.html`.

[12] X. Gao, D. Hu, M. Gogol, and H. Li. "ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions". *Bioinformatics* 35.17 (2019), pp. 3038–3045.

[13] L. Haghverdi, A. T. Lun, M. D. Morgan, and J. C. Marioni. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". *Nature Biotechnology* 36.5 (2018), pp. 421–427.

[14] S. He, L.-H. Wang, Y. Liu, Y.-Q. Li, H.-T. Chen, J.-H. Xu, W. Peng, G.-W. Lin, P.-P. Wei, B. Li, et al. "Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs". *Genome Biology* 21 (2020), pp. 1–34.

[15] R. Hou, E. Denisenko, and A. R. Forrest. "scMatch: a single-cell gene expression profile annotation tool using reference datasets". *Bioinformatics* 35.22 (2019), pp. 4688–4695.

[16] H. Huang, C. Liu, M. M. Wagle, and P. Yang. "Evaluation of deep learning-based feature selection for single-cell RNA sequencing data analysis". *Genome Biology* 24.1 (2023), p. 259.

[17] G.-J. Huizing, G. Peyré, and L. Cantini. "Optimal Transport improves cell-cell similarity inference in single-cell omics data". *Bioinformatics* (2022). ISSN: 1367-4803.

[18] A. Ianevski, A. K. Giri, and T. Aittokallio. "Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data". *Nature Communications* 13.1 (2022), p. 1246.

[19] S. Jessa, A. Blanchet-Cohen, B. Krug, M. Vladoiu, M. Coutelier, D. Faury, B. Poreau, N. De Jay, S. Hébert, J. Monlong, et al. "Stalled developmental programs at the root of pediatric brain tumors". *Nature Genetics* 51.12 (2019), pp. 1702–1713.

[20] P. V. Kharchenko, L. Silberstein, and D. T. Scadden. "Bayesian approach to single-cell differential expression analysis". *Nature Methods* 11.7 (2014), pp. 740–742.

[21] B. Kim, E. Lee, and J. K. Kim. "Analysis of Technical and Biological Variability in Single-Cell RNA Sequencing". *Computational Methods for Single-Cell Data Analysis*. Ed. by G.-C. Yuan. Humana New York, NY, Jan. 2019. Chap. 3, pp. 25–44.

[22] V. Y. Kiselev, T. S. Andrews, and M. Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". *Nature Reviews Genetics* 20.5 (2019), pp. 273–282.

[23] V. Y. Kiselev, A. Yiu, and M. Hemberg. "scmap - A tool for unsupervised projection of single cell RNA-seq data". *bioRxiv* (2017).

[24] Y. Kotliarov, R. Sparks, A. J. Martins, M. P. Mulè, Y. Lu, M. Goswami, L. Kardava, R. Banchereau, V. Pascual, A. Biancotto, et al. "Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus". *Nature Medicine* 26.4 (2020), pp. 618–629.

[25] P. Langfelder, B. Zhang, and S. Horvath. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". *Bioinformatics* 24.5 (2008), pp. 719–720.

[26] N. Lawlor, J. George, M. Bolisetty, R. Kursawe, L. Sun, V. Sivakamasundari, I. Kycia, P. Robson, and M. L. Stitzel. "Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes". *Genome Research* 27.2 (2017), pp. 208–222.

[27] H. Le, B. Peng, J. Uy, D. Carrillo, Y. Zhang, B. D. Aevermann, and R. H. Scheuermann. "Machine learning for cell type classification from single nucleus RNA sequencing data". *Plos One* 17.9 (2022), e0275070.

[28] J. Li and L. Lin. "Optimal Transport with Relaxed Marginal Constraints". English (US). *IEEE Access* 9 (2021), pp. 58142–58160.

[29] J. Li, B. Seo, and L. Lin. "Optimal transport, mean partition, and uncertainty assessment in cluster analysis". *Statistical Analysis and Data Mining: The ASA Data Science Journal* 12.5 (2019), pp. 359–377.

[30] Z. Li and H. Feng. "A neural network-based method for exhaustive cell label assignment using single cell RNA-seq data". *Scientific Reports* 12.1 (2022), p. 910.

[31] L. Lin, W. Shi, J. Ye, and J. Li. "Multisource single-cell data integration by MAW barycenter for Gaussian mixture models". *Biometrics* 79.2 (2023), pp. 866–877.

[32] A. T. L. Lun and L. Haghverdi. *batchelor: Single-Cell Batch Correction Methods*. R package version 1.25.0. Bioconductor. 2025.

[33] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, et al. "A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data". *The Pharmacogenomics Journal* 10.4 (2010), pp. 278–291.

[34] Z. Miao, P. Moreno, N. Huang, I. Papatheodorou, A. Brazma, and S. A. Teichmann. "Putative cell type discovery from single-cell gene expression data". *Nature Methods* 17.6 (2020), pp. 621–628.

[35] T. Mou, W. Deng, F. Gu, Y. Pawitan, and T. N. Vu. "Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing". *Frontiers in Genetics* 10 (2020).

[36] D. Y. Orlova, S. Meehan, D. Parks, W. A. Moore, C. Meehan, Q. Zhao, E. E. B. Ghosn, L. A. Herzenberg, and G. Walther. "QFMatch: multidimensional flow and mass cytometry samples alignment". *Scientific Reports* 8.3291 (2018).

[37] J. M. Pullin and D. J. McCarthy. "A comparison of marker gene selection methods for single-cell RNA sequencing data". *Genome Biology* 25.1 (2024), p. 56.

[38] Q. Qi, Y. Wang, Y. Huang, Y. Fan, and X. Li. "PredGCN: a Pruning-enabled Gene-Cell Net for automatic cell annotation of single cell transcriptome data". *Bioinformatics* 40.7 (June 2024), btae421. ISSN: 1367-4811.

[39] Y. Qiu, J. Wang, J. Lei, and K. Roeder. "Identification of cell-type-specific marker genes from co-expression patterns in tissue samples". *Bioinformatics* 37.19 (2021), pp. 3228–3234.

[40] X. Ruan, Y. Cheng, Y. Ye, Y. Wang, X. Chen, Y. Yang, T. Liu, and F. Yan. "PIPET: predicting relevant subpopulations in single-cell data using phenotypic information from bulk data". *Briefings in Bioinformatics* 25.4 (2024), bbae260.

[41] Å. Segerstolpe, A. Palasantza, P. Eliasson, E.-M. Andersson, A.-C. Andréasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, et al. "Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes". *Cell metabolism* 24.4 (2016), pp. 593–607.

[42] M. A. Skinnider, J. W. Squair, and G. Courtine. "Enabling reproducible re-analysis of single-cell data". *Genome biology* 22.1 (2021), p. 215.

[43] M. J. Stubbington, O. Rozenblatt-Rosen, A. Regev, and S. A. Teichmann. "Single-cell transcriptomics to explore the immune system in health and disease". *Science* 358.6359 (2017), pp. 58–63.

[44] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. "Exponential scaling of single-cell RNA-seq in the past decade". *Nature Protocols* 13 (2018), pp. 599–604.

[45] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, et al. "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics". *Nature Neuroscience* 19.2 (2016), pp. 335–346.

[46] L. C. Torres, L. M. Pereira, and M. H. Amini. "A Survey on Optimal Transport for Machine Learning: Theory and Applications". *arXiv preprint arXiv:2106.01963* (2021). arXiv: 2106.01963 [cs.LG].

[47] P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad. "Batch effects and the effective design of single-cell gene expression studies". *Scientific Reports* 7.1 (2017), p. 39921.

[48] Z. Wang, H. Ding, and Q. Zou. "Identifying cell types to interpret scRNA-seq data: how, why and more possibilities". *Briefings in Functional Genomics* 19.4 (2020), pp. 286–291.

[49] Z. Wu and H. Wu. "Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering". *Genome Biology* 21.1 (2020), pp. 1–14.

[50] X. Xu, X. Hua, H. Mo, S. Hu, and J. Song. "Single-cell RNA sequencing to identify cellular heterogeneity and targets in cardiovascular diseases: from bench to bedside". *Basic Research in Cardiology* 118.1 (2023), p. 7.

[51] F. Yang, W. Wang, F. Wang, Y. Fang, D. Tang, J. Huang, H. Lu, and J. Yao. "scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data". *Nature Machine Intelligence* 4.10 (2022), pp. 852–866.

[52] M. Yuan, L. Chen, and M. Deng. "scMRA: a robust deep learning method to annotate scRNA-seq data with multiple reference datasets". *Bioinformatics* 38.3 (Oct. 2021), pp. 738–745. ISSN: 1367-4803.

[53] L. Zhang, L. Lin, and J. Li. "Multi-view clustering by CPS-merge analysis with application to multimodal single-cell data". *PLoS Computational Biology* 19.4 (2023), e1011044.

[54] M. Zhang and F. R. Guo. "BSDE: barycenter single-cell differential expression for case–control studies". *Bioinformatics* 38.10 (2022), pp. 2765–2772.

[55] Y. Zhang, B. D. Aevermann, T. E. Bakken, J. A. Miller, R. D. Hodge, E. S. Lein, and R. H. Scheuermann. "FR-Match: robust matching of cell type clusters from single cell RNA sequencing data using the Friedman–Rafsky non-parametric test". *Briefings in Bioinformatics* 22.4 (Nov. 2020).

[56] J. Zhao, S. Zhang, Y. Liu, X. He, M. Qu, G. Xu, H. Wang, M. Huang, J. Pan, Z. Liu, Z. Li, L. Liu, and Z. Zhang. "Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human". *Cell Discovery* 6 (2020).

[57] R. Zilionis, C. Engblom, C. Pfirschke, V. Savova, D. Zemmour, H. D. Saatcioglu, I. Krishnan, G. Maroni, C. V. Meyerovitz, C. M. Kerwin, et al. "Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species". *Immunity* 50.5 (2019), pp. 1317–1334.

Supporting information

S1 Appendix. This file includes a description of the basic formulation of optimal transport and detailed information about the datasets and the data processing steps.