PRESERVED TOPOLOGY OF THE GENE EXPRESSION LANDSCAPE ACROSS SCRNA-SEQ EXPERIMENTS

Runjia Luke Li
Department of Bioinformatics
University of California, Los Angeles
Los Angeles, CA 90095
luke0321lrj@gmail.com

ABSTRACT

Single cell RNA-Seq (scRNA-Seq) is a powerful tool for cell type detection. However, for multiple datasets, especially datasets from different scRNA-Seq technologies, comparing or merging them remain challenging. In this paper, we showed that if measuring the same sample, the toplogies of gene expression pattern were preserved across experiments. We have developed a method to match clusters (cell sub-types) from multiple datasets, which only relies on distance matrices of clusters (cell sub-types).

1 Introduction

Single cell RNA-Seq (scRNA-Seq), as a very powerful tool, facilities the cell-level exploration. One fundamental usage of scRNA is to discover cell types or cell sub-types. However, combining information from multiple datasets can be very challenging. Many methods have been developed to combine different datasets by performing batch correction. For example, Haghverdi et al. [1] uses mutual nearnest neighbors to identify cells in different datasets that are likely to be from the same cell type, and uses these cells as anchors to merge the experiments by linear shifting; Butler et al. [2] utilizes canonical correlation analysis to identify combinations of genes that are correlated across experiments and aligns the experiments using these "meta-genes". Some other methods have been proposed to match the clusters or individual cells between experiments [3, 4]. These methods either rely on external annotations for marker genes of cell types to perform cluster-level matching, or require complete information from the original expression matrices to perform cell-level matching. We define a "topology" to be the spatial configurations of different cell types in the gene expression space, relative to each other. Different experiments on the same sample can be thought of as repeated measurements of the same gene expression landscape, with the differences between experiments (caused by batch effects or different sequencing technologies) acting as affine transformations of the entire landscape. Therefore, we reason that the topology should be preserved across the experiments, making it possible to match cell types across the experiments using only spatial information.

In this study, we first visually illustrate out claim through exploratory data analysis of a human PBMC dataset reported by Ding et al. [5]. We then establish *distance correlation of cell type matching* as a quantitative metric for the extent of preserved topologies between experiments. Finally, we present an algorithm to match the cell types between experiments solely based on the distance matrices of cell types, and evaluate its performance on both simulated data and the PBMC dataset.

2 Materials and methods

2.1 Exploratory data analysis

We used a subset of PBMC dataset to explore the topology across experiments. The six data we used are PBMC1 Smart-seq 2, PBMC2 Smart-seq 2, PBMC1 10x-Chromium-v2.A, PBMC2 10x-Chromium-v2.

For EDA, we basically used R package `Seurat` [6] to process the data. `Seurat` was used for filtering genes and cells, data normalization and finding variable genes. We performed principal component analysis (PCA) for dimension reduction on 2000 variable genes, and kept top 50 PCs for downstream analysis. We used K-means clustering with $k = 5$. These clusters were then labeled by one or several cell types based on the canonical maker genes [6]. We plot 2 of 3 top PCs for data visualization.

2.2 Data processing

For the PBMC dataset, we performed data preprocessing, clustering, cell type labeling and distance computation on a per-experiment (replicate-technology combination) basis, using the python package `scanpy` [7].

2.2.1 Preprocessing and clustering

For each experiment, genes with zero count across all cells were removed. The UMI counts for the genes of each cell were normalized by dividing with the total UMI count of that cell, and then log-transformed after addition of a pseudocount of 1. Highly variable genes were identified and annotated using the method proposed by Satija et al. [8]. We performed PCA (with 50 PCs) on the cells, generated a directed k-nearest neighbors (k-NN) graph using the top PCs, and then used the Louvain community detection algorithm on the k-NN graph to identify clusters of cells [9, 10]. For the four crucial hyperparameters in the clustering analysis, we used the optimal values reported by Ding et al. [5]:

1. Whether only highly-variable genes are considered.
2. The number of top PCs to generate the k-NN graph.
3. The number of nearest neighbors (k) for the k-NN graph
4. The resolution value for the Louvain algorithm

2.2.2 cell type labeling

We annotated each cluster with a cell type using a similar approach as in Ding et al. [5]. First, we obtained a curated list of marker genes for each cell type. For cell type s, let M_s^+ be the set of up-regulated marker genes and M_s^- be the set of down-regulated marker genes. Cell i was assigned a score $f_{i,s}$ for cell type s, computed as:

$$f_{i,s} = \log\left(10^4 * \frac{(\sum_{j\in M_s^+} x_{ij} - \sum_{j\in M_s^-} x_{ij})}{\sum_j x_{ij}}\right) \tag{1}$$

where x_{ij} is the UMI count for gene j of cell i. Each cluster was annotated with the cell type whose score best distinguishes cells in the cluster from cells not in the cluster. For each pair of (cluster C, cell type s), we computed of area under the receiver operating characteristic curves (AUROCs) by treating the score for s as a binary classifier of whether an arbitrary cell belongs to C, using $f_{0,s}, f_{1,s}, ...$ as the predictions for all cells and $I(0 \in C), I(1 \in C), ...$ as the true cluster assignments for all cells, where $I(i \in C)$ is the binary indicator of whether cell i is in cluster C. The assigned cell type label for each cluster was the cell type with maximum AUROC for that cluster.

2.3 Distance correlation of cell type matching

2.3.1 Generalized formulation

Let $\mathcal{S}_A = \{S_{A1}, S_{A2}, ..., S_{An}\}$ be a partition of the cells in experiment A into n sets, each representing a cell type. Then, a *cell type matching* between experiments A and B is defined as $H = \{(S_{Aa_1}S_{Bb_1}), ..., (S_{Aa_k}, S_{Bb_k})\} \subset \mathcal{S}_A \times \mathcal{S}_B$, a set of bijective pairings between k different cell types in experiment A, indexed by $a_1, ..., a_k$ and k different cell types in experiment B, indexed by $b_1, ..., b_k$. Given a cell type distance metric d, the distance matrix of A's cell types in H is just $\mathbf{D}_A = (D_{Aij}) \in \mathbb{R}^{k\times k}$, where $D_{Aij} = d(S_{Aa_i}, S_{Aa_j})$; similarly we have $\mathbf{D}_B$, and an important fact is that the cell type in experiment A corresponding to the i^{th} row & column of $\mathbf{D}_A$ and the cell type in experiment B corresponding to the i^{th} row & column of $\mathbf{D}_B$ are matched in H. The above process is intuitively illustrated in figure 1. The distance correlation [11, 12] of H is then defined as:

$$\mathrm{dCor}(H) = \frac{(\frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k D'_{Aij}D'_{Bij})^{\frac{1}{2}}}{\sqrt{(\frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k D'_{Aij})^{\frac{1}{2}}(\frac{1}{k^2}\sum_{i=1}^k\sum_{j=1}^k D'_{Bij})^{\frac{1}{2}}}} \in [0,1] \tag{2}$$

where

$$D'_{Aij} = D_{Aij} - \frac{1}{k}\sum_{j'=1}^{k} D_{Aij'} - \frac{1}{k}\sum_{i'=1}^{k} D_{Ai'j} + \frac{1}{k^2}\sum_{i'=1}^{k}\sum_{j'=1}^{k} D_{Ai'j'} \tag{3}$$

and similarly,

$$D'_{Bij} = D_{Bij} - \frac{1}{k}\sum_{j'=1}^{k} D_{Bij'} - \frac{1}{k}\sum_{i'=1}^{k} D_{Bi'j} + \frac{1}{k^2}\sum_{i'=1}^{k}\sum_{j'=1}^{k} D_{Bi'j'} \tag{4}$$

In essence, $\mathrm{dCor}(H)$ is just the Pearson correlation of the ordered entries of the two distance matrices, $\mathbf{D}_A$ and $\mathbf{D}_B$. We reason that $\mathrm{dCor}(H)$ will take a large value if the spatial configurations of cell types are indeed preserved across experiments and the cell types shared by the two experiments are correctly matched to each other by H (in other words S_{Aa_i} and S_{Bb_i} represent the same cell type, and so forth), because in this case, the entries at the same positions in the distance matrices of the two experiments should be highly correlated to each other. For example, let I, J and K be three cell types corresponding to sets $S_{Aa_i}, S_{Aa_j}, S_{Aa_k}$ and $S_{Bb_i}, S_{Bb_j}, S_{Bb_k}$ in the two experiments respectively. If cell type I is more similar to J than to K in terms of gene expression, then due to the preservation of gene expression topology across experiments we have both $d(S_{Aa_i}, S_{Aa_j}) > d(S_{Aa_i}, S_{Aa_k})$ and $d(S_{Bb_i}, S_{Bb_j}) > d(S_{Bb_i}, S_{Bb_k})$. Therefore, if these three cell types are correctly matched for the two experiments, that is $\{(S_{Aa_i}, S_{Bb_i}), (S_{Aa_j}, S_{Bb_j}), (S_{Aa_k}, S_{Bb_k})\} \subset H$, apparently D_{Aij}, D_{Aik} should be correlated with D_{Bij}, D_{Bik}.

We also propose that, if a number of shared cell types are identified for a pair of experiments, we can match each of the shared cell types in one experiment to the other and compute the distance correlation of this correct matching $\mathrm{dCor}(H_{correct})$ as a similarity metric between the two experiments. A large value for $\mathrm{dCor}(H_{correct})$ indicates preserved gene expression topology of these shared cell types whereas a small value for $\mathrm{dCor}(H_{correct})$ implies distortions of the topology.

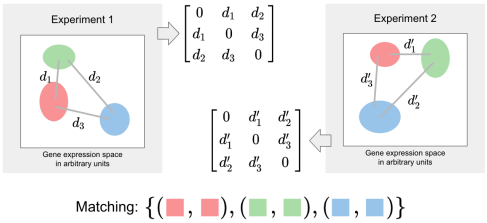


Figure 1: Schematic of the gene expression space of two experiments and their corresponding distance matrices, generated by a matching. The two experiments share three cell types, color-coded as red, green and blue respectively. The distances between cell types in each experiment are also shown, under an arbitrary (and, for simplicity, symmetric) distance measure. In this particular matching, the same cell types are matched to each other across the two experiments, but it is not necessarily the case.

2.3.2 Cell type distance metrics

The distance correlation computation described above is clearly dependent on the choice of d, the distance metric between two cell types. Let S_{Ai}, S_{Aj} be two sets of cells in a same experiment, each corresponding to a cell type and let $\boldsymbol{x}$,$\boldsymbol{y}$ be the feature vectors of two cells. The features can either be the gene expression of the coordinates in the principal component space. We considered the following distance metrics:

1. The average Euclidean distance between cells in the two cell types:

$$d(S_{Ai}, S_{Aj}) = \frac{1}{|S_{Ai}||S_{Aj}|}\sum_{\boldsymbol{x}\in S_{Ai}}\sum_{\boldsymbol{y}\in S_{Aj}} \|\boldsymbol{x}-\boldsymbol{y}\|_2 \tag{5}$$

2. The average cosine distance between cells in the two cell types:

$$d(S_{Ai}, S_{Aj}) = \frac{1}{|S_{Ai}||S_{Aj}|}\sum_{\boldsymbol{x}\in S_{Ai}}\sum_{\boldsymbol{y}\in S_{Aj}} 1 - \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{||\boldsymbol{x}||_2||\boldsymbol{y}||_2} \tag{6}$$

3. The average correlation distance (one minus Pearson correlation) between cells in the two cell types:

$$d(S_{Ai}, S_{Aj}) = \frac{1}{|S_{Ai}||S_{Aj}|}\sum_{\boldsymbol{x}\in S_{Ai}}\sum_{\boldsymbol{y}\in S_{Aj}} 1 - \frac{(\boldsymbol{x}-\bar{\boldsymbol{x}})\cdot(\boldsymbol{y}-\bar{\boldsymbol{y}})}{||\boldsymbol{x}-\bar{\boldsymbol{x}}||_2||\boldsymbol{y}-\bar{\boldsymbol{y}}||_2} \tag{7}$$

4. The average Manhattan (cityblock) distance between cells in the two cell types:

$$d(S_{Ai}, S_{Aj}) = \frac{1}{|S_{Ai}||S_{Aj}|} \sum_{\boldsymbol{x} \in S_{Ai}} \sum_{\boldsymbol{y} \in S_{Aj}} \|\boldsymbol{x} - \boldsymbol{y}\|_1 \tag{8}$$

5. The Euclidean distance between the centroids of the two cell types:

$$d(S_{Ai}, S_{Aj}) = \left\| \frac{1}{|S_{Ai}|} \sum_{\boldsymbol{x} \in S_{Ai}} \boldsymbol{x} - \frac{1}{|S_{Aj}|} \sum_{\boldsymbol{x} \in S_{Aj}} \boldsymbol{x} \right\|_2 \tag{9}$$

2.3.3 Distance correlation computation for the PBMC dataset

Foe each experiment, we calculated the distances between each pair of cell types using the five aforementioned cell type distance metrics respectively, in the space spanned by the top principal components of that experiment. Then, for each pair of experiments in the PBMC dataset with more than 3 shared cell types identified through clustering, we computed the distance correlation with correct matching of shared cell types ($\mathrm{dCor}(H_{correct})$) using each of the five cell type distance metrics. A total of 5×88 distance correlations were obtained. Each distance metric was having 88 distance correlations being computed, and for each pair of distance metrics we computed the Pearson correlation of the two sets of 88 distance correlations.

2.4 Cell type matching

We next investigated whether we can correctly match the shared cell types between two experiments by finding a matching that maximizes the distance correlation. The generalized approach is shown by algorithm 1, with the same notations as in section 2.3.

Algorithm 1 Cell type matching ($\mathcal{S}_A$, $\mathcal{S}_B$)

```
d_max ← 0
H_max ← {}
for each matching H of S_A, S_B do
    d ← dCor(H)
    if d > d_max then
        d_max ← d
        H_max ← H
    end if
end for
return H_max
```

An obvious issue is that there are a total of $\sum_{n=1}^{\min(|\mathcal{S}_A|,|\mathcal{S}_B|)} \binom{|\mathcal{S}_A|}{n} \binom{|\mathcal{S}_B|}{n} n!$ unique matchings. We therefore add one constraint to the algorithm, that is $|\mathcal{S}_A| = |\mathcal{S}_B| = |H| = k$. Stated otherwise, the two experiments should identify the same number of cell types and we force each of the cell types in one experiment to be matched to a cell type in another, leaving no singlets. The number of possible matchings is therefore reduced to $k!$, which can still be substantial but is relatively manageable for our dataset.

We first tested this matching algorithm on simulated data which satisfied the constraint. For the PBMC dataset, we conformed to the constraint by only testing the matching algorithm on pairs of experiments with more than 3 shared cell types identified and only using these shared cell types as candidates for generating a matching. We also tested each cell type distance metric for each of the selected pairs of experiments and counted the number of correctly matched cell types in the output matching for each test. A total of 5×88 tests on the PBMC dataset were conducted.

2.5 Simulation of scRNA-seq data

We employed a similar approach as in Haghverdi et al. [1] to simulate scRNA-seq data, detailed by the following steps:

1. We specified six 2-d Gaussians with different means and covariances, each representing a cell type.
2. We generated a matrix $\mathbf{M} \in \mathbb{R}^{2 \times 100}$ with each entry being drawn independently from a standard Gaussian.

3. We specified six mixture weights, one for each Gaussian in step 1, and draw 1000 cells from the resulting six-component Gaussian mixture model. We then projected the cells into 100 dimensions (to simulate 100 genes) using $\mathbf{M}$.
4. We generated a 100-d standard Gaussian random vector as the gene-specific batch effect and added it to each cell.
5. For each gene of each cell, we added a random noise drawn independently from a standard Gaussian.

Using steps 1-5, we simulated a scRNA-seq experiment with 6 cell types. We repeated steps 3-5 with different mixture weights, batch effect and random noise to generate another experiment with the same cell types. We calculated the distances between cell types (using average Euclidean distance as metric) in the 100-d gene expression space for each experiment, and checked how many cell types among the six are correctly matched by the matching algorithm using distance correlations. We repeated the matching analysis for 100 pairs of simulated experiments.

3 Results

3.1 Similar topology of expression pattern between experiments

We visualized the PCA spaces of each data by choosing two PCs among top 3 PCs (figure 2). Based on the PCA plots, we clearly saw that the cell patterns are similar to each other. The CD 14+ cell group, CD 8+ cell group and B cell group tend to show a triangle shape in every group. As our expectation, the similarity with in one scRNA technology is much higher than cross technology. This EDA showed us the property that the expression pattern can be preserved across multiple experiments in a PCA space.

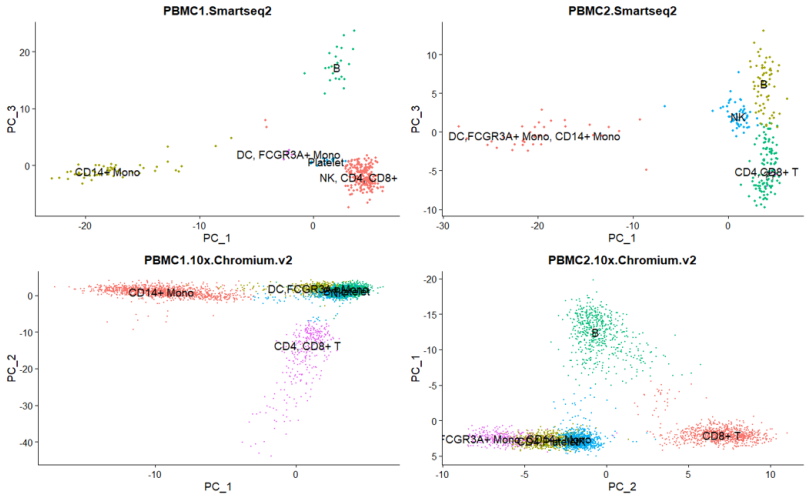


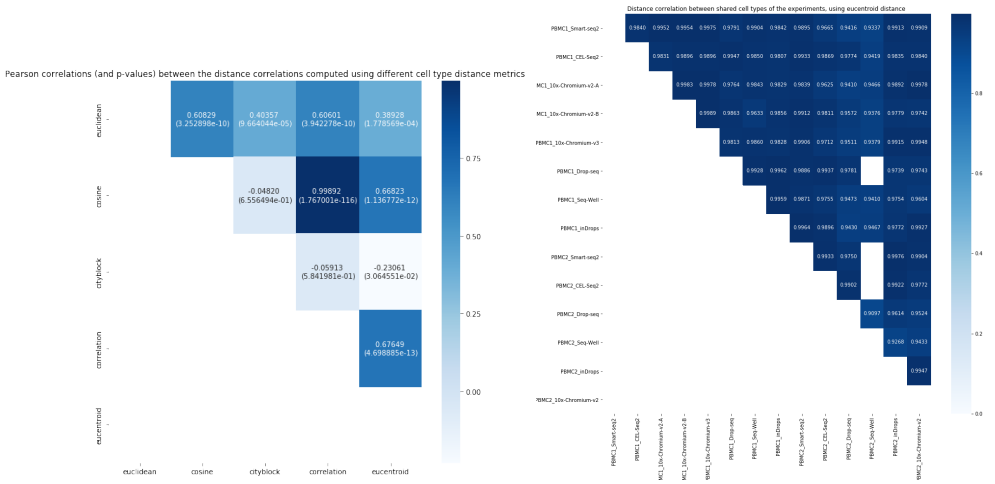
Figure 2: PCA plots of each EDA dataset. Two PCs among top 3 PCs were chosen for visualization, and clusters were labeled by cell types. The patterns are clearly similar between each pair of datasets.

3.2 Distance correlations between pairs of experiments

We computed the distance correlations of correct matching ($\mathrm{dCor}(H_{correct})$) between all pairs of experiments with > 3 shared cell types using all five cell type distance metrics. While all pairs of experiments were having very large distance correlations under all distance metrics, the different distance metrics did not entirely correlate well with each other, except for correlation distance and its unscaled version that is cosine distance (figure 3). Still, under each distance metric being used, the topology of human PBMC cell types was highly invariant across different replicates and sequencing technologies, affirming our hypothesis.

3.3 Performance of the cell type matching algorithm

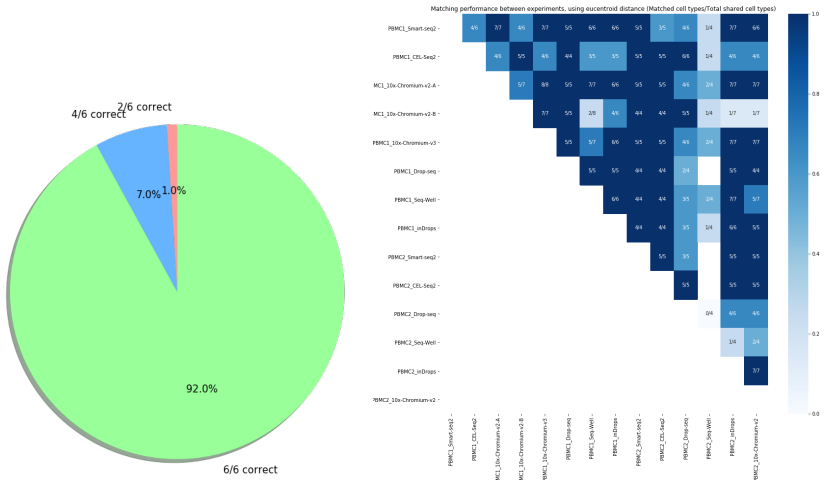
The performance of the matching algorithm on simulated data using average Euclidean distance is shown in figure 4(a). For above 90% of the time our algorithm was able to correctly match all six cell types between two experiments. In



(a) (b)

Figure 3: (a) Pearson correlations (p-values shown in parentheses) between distance correlations of the 88 pairs of experiments using different cell type distance metrics. (b) Distance correlations of correct matching ($\mathrm{dCor}(H_{correct})$) between pairs of experiments with more than 3 shared cell types, using Euclidean distance between cell type centroids as cell type distance metric. Eucentroid: Euclidean distance between cell type centroids.

other words, under a simple condition where the batch effect manifested as a translation in the gene expression space, our algorithm captured the invariant nature of the cell type topology.

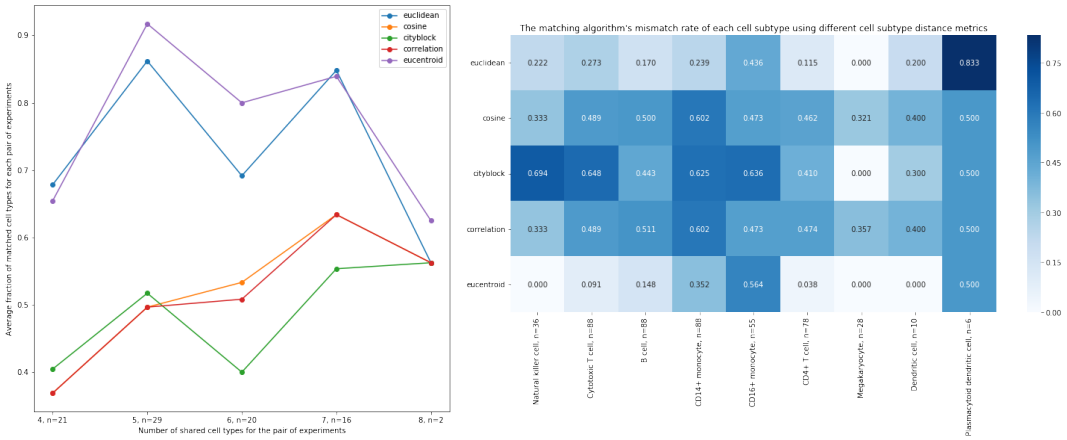


(a) (b)

Figure 4: (a) Performance of the matching algorithm on simulated data using average Euclidean distance as metric, showing the portions of the number of cell types correctly matched among a total of six cell types. (b) Performance of the matching algorithm between pairs of experiments with more than 3 shared cell types, using Euclidean distance between cell type centroids as cell type distance metric.

The performance of the matching algorithm on the PBMC dataset using each of the distance metrics is shown in figure 5(a). We observed that the best performing distance metrics are Euclidean distance between cell type centroids and average Euclidean distance between cell types, both of which consistently reaching above 65% accuracy (if not

considering the rare case with two experiments sharing 8 cell types). The other three metrics had $\sim 50\%$ accuracy on average, still performing significantly better than random guessing (whose expected accuracy is $1/k$ where k is the number of shared candidate cell types for matching). Specifically, using Euclidean distance between cell type centroids allowed us to perfectly match the cell types in more than 60% of the pairs (figure 4(b)). In addition, CEL-Seq2 for PBMC replicate 1, Drop-Seq for PBMC replicate 2 and Seq-Well for PBMC replicate 2 were always poorly matched to other experiments under all distance metrics.



(a) (b)

Figure 5: (a) Performance of the matching algorithm using different cell type distance metrics, showing the average fraction of cell types correctly matched for a pair of experiments vs. the number of shared cell types (maximum possible number of cell types that can be correctly matched) between a pair of experiments. (b) The mismatch rate (number of times a cell type is not correctly matched / number of times a cell type is shared between two experiments) of the matching algorithm for each cell type, using different distance metrics. Eucentroid: Euclidean distance between cell type centroids.

We next investigated the cell-type specific performances of our algorithm using each of the distance metrics. Figure 5(b) shows the mismatch rate for each cell type for the distance metrics. We understood that the poor performance on the plasmacytoid dendritic cell was due to the fact that this cell type was not confidently identified through clustering, and thus it only appeared in very few ($n = 6$) pairs of experiments. CD16+ and 14+ monocytes were also not well matched for any of the distance metrics because these two cell types are relatively similar in terms of biology, so they might be occupying similar spatial locations in the gene expression landscape relative to other PBMC cell types. This can be further illustrated by figure 6, showing the row-normalized cell type confusion matrix of the top performing distance metric that is Euclidean distance between cell type centroids. Apparently the algorithm often confused CD14+ with 16+ monocytes and vise versa; it was also less confident about the plasmacytoid dendritic cells. All the other cell types, however, were nearly always correctly matched for this distance metric.

4 Discussion

In this study, we first showed the invariant nature of cell type topology in gene expression space across different replicates and different sequencing platforms when measuring the same samples. To best capture the topology, we explored different cell cluster distance metrics, and showed that Euclidean distance yielded the best result. based on our clustering matching algorithm, we achieved the average accuracy $> 65\%$ in real data, and $> 90\%$ in simulation data. The highlight of our research can be summarized into two points: (1) Using PCA and appropriate distance, the distances between cell types will be consistent across different experimental conditions; this result makes us more confident in current distance-based scRNA analysis, e.g., scRNA pseudotime analysis. (2) It is possible to match clusters without original expression information.

However, there are still several limitations of our methods. The most obvious limitation is that we can only match two samples sharing exact same cell types. In other words, we are only able to achieve a "one-to-one" match currently. The future work would be generalizing this methods to variable samples and constructing a more flexible match.

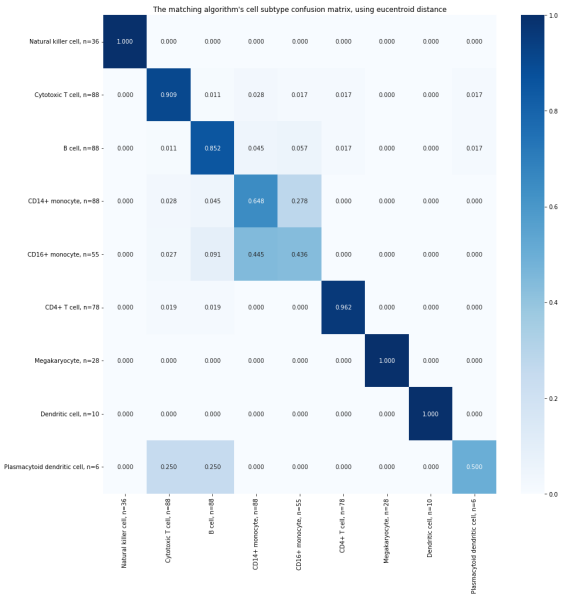


Figure 6: Row-normalized cell type confusion matrix of the matching algorithm using Euclidean distance between cell type centroids as cell type distance metric.

References

[1] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421, 2018.

[2] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411, 2018.

[3] Xin Gao, Deqing Hu, Madelaine Gogol, and Hua Li. Clustermap: compare multiple single cell rna-seq datasets across different experimental conditions. *Bioinformatics*, 2019.

[4] Erica AK DePasquale, Daniel Schnell, Phillip Dexheimer, Kyle Ferchen, Stuart Hay, Kashish Chetal, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, and Nathan Salomonis. cellharmony: cell-level matching and holistic comparison of single-cell transcriptomes. *Nucleic acids research*, 47(21):e138–e138, 2019.

[5] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparative analysis of single cell rna-sequencing methods. *bioRxiv*, 2019. doi: 10.1101/632216. URL `https://www.biorxiv.org/content/early/2019/05/09/632216`.

[6] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single cell data. *bioRxiv*, 2018. doi: 10.1101/460147. URL `https://www.biorxiv.org/content/10.1101/460147v1`.

[7] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[8] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495, 2015.

[9] 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.

[10] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[11] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[12] Gábor J Székely, Maria L Rizzo, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4): 1236–1265, 2009.