Bioinformatics, 38(8), 2022, 2169–2177
https://doi.org/10.1093/bioinformatics/btac084
Advance Access Publication Date: 14 February 2022
Original Paper

OXFORD

Gene expression

Optimal transport improves cell–cell similarity inference in single-cell omics data

Geert-Jan Huizing [1,2,*], **Gabriel Peyré** [2] and **Laura Cantini** [1,*]

[1]Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, INSERM, Ecole Normale Supérieure, Université PSL, 75005 Paris, France and [2]Département de Mathématiques et Applications de l'Ecole Normale Supérieure, CNRS, Ecole Normale Supérieure, Université PSL, 75005 Paris, France

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on July 12, 2021; revised on December 17, 2021; editorial decision on February 5, 2022; accepted on February 8, 2022

Abstract

Motivation: High-throughput single-cell molecular profiling is revolutionizing biology and medicine by unveiling the diversity of cell types and states contributing to development and disease. The identification and characterization of cellular heterogeneity are typically achieved through unsupervised clustering, which crucially relies on a similarity metric.

Results: We here propose the use of Optimal Transport (OT) as a cell–cell similarity metric for single-cell omics data. OT defines distances to compare high-dimensional data represented as probability distributions. To speed up computations and cope with the high dimensionality of single-cell data, we consider the entropic regularization of the classical OT distance. We then extensively benchmark OT against state-of-the-art metrics over 13 independent datasets, including simulated, scRNA-seq, scATAC-seq and single-cell DNA methylation data. First, we test the ability of the metrics to detect the similarity between cells belonging to the same groups (e.g. cell types, cell lines of origin). Then, we apply unsupervised clustering and test the quality of the resulting clusters. OT is found to improve cell–cell similarity inference and cell clustering in all simulated and real scRNA-seq data, as well as in scATAC-seq and single-cell DNA methylation data.

Availability and implementation: All our analyses are reproducible through the OT-scOmics Jupyter notebook available at https://github.com/ComputationalSystemsBiology/OT-scOmics.

Contact: laura.cantini@ens.fr or huizing@ens.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Allowing the measurement of gene expression in thousands of cells in a single experiment, single-cell RNA sequencing (scRNA-seq) has unveiled the diversity of the cells constituting human tissues (Stegle *et al.*, 2015). The possibility to assess cellular heterogeneity at a previously inaccessible resolution has profoundly impacted our understanding of development, of the immune system functioning, and of many diseases (Papalexi and Satija, 2018; Potter, 2018; Rajewsky *et al.*, 2020). While scRNA-seq is now mature, the single-cell technological development has shifted to the measurement of other omics, e.g. DNA methylation, proteome and chromatin accessibility (Lee *et al.*, 2020; Ma *et al.*, 2020).

A common goal in single-cell data analysis is the identification of the cell types and cell states present in a sample (Luecken and Theis, 2019). This is typically achieved in a data-driven fashion through unsupervised clustering (Kiselev *et al.*, 2019; Xiong *et al.*, 2019). Cells with similar transcriptional profiles are assembled into clusters, which are then annotated based on markers (Kiselev *et al.*, 2019). As a consequence, the quality of such clustering plays a critical role in the derived biological discovery. While numerous clustering algorithms have been proposed, they all rely on a similarity metric for categorizing individual cells. Popular metrics include the Euclidean and Manhattan distances, Cosine similarity and Pearson correlation (Guo *et al.*, 2015; Kim *et al.*, 2019; Macosko *et al.*, 2015; Satija *et al.*, 2015).

Optimal Transport (OT) emerged in the last decade as a promising mathematical toolkit to analyze and compare high-dimensional data using different variants of the Wasserstein distance (Peyré and Cuturi, 2019; Santambrogio, 2015). Recently, applications of OT to biology have been proposed (Bellazzi *et al.*, 2021; Cao *et al.*, 2020; Demetci *et al.*, 2020; Huizing *et al.*, 2021; Schiebinger *et al.*, 2019). Some works use OT on cells in the context of trajectory inference (Schiebinger *et al.*, 2019) and alignment of unpaired (i.e. independently profiled) scRNA-seq and scATAC-seq data (Cao *et al.*, 2020;

Demetci *et al.*, 2020). Others apply OT on genes in scRNA-seq data and perform supervised cell classification, such as malignant versus normal cells (Bellazzi *et al.*, 2021). Of note, classical OT, as applied by Bellazzi *et al.*, requires solving a costly linear optimization problem thus making computations in large single-cell data slow and sometimes computationally intractable. In addition, the use of supervised clustering in biological applications is of limited relevance. Indeed, most of the existing biological studies use single-cell data to discover and characterize the cell populations/types present in a biological sample, a task that is intrinsically unsupervised. We finally recently developed a joint metric learning method simultaneously computing OT distances between both genes and cells, but not providing relevant improvements over existing single-cell approaches (Huizing *et al.*, 2021).

Here, we propose the use of OT as a cell–cell similarity metric for single-cell data. In particular, we use OT with entropic regularization (Cuturi, 2013), expected to control the systematic noise due to the stochasticity of gene expression at single-cell level and to the presence of dropouts. In addition, using OT with entropic regularization we could efficiently analyze datasets with large numbers of cells using a Graphics Processing Unit (GPU). We further extensively benchmark OT against state-of-the-art metrics. We apply the different metrics to single-cell data with known groups (e.g. cell types, cell lines of origin) and we evaluate their ability to detect the similarity between cells belonging to the same group. We then apply different unsupervised clustering algorithms to the computed distance matrices and test the quality of the resulting clusters. Of note, all the tests are performed in three conditions: (i) simulated scRNA-seq data, where the effect of the number of cells and of the size and overlap of the clusters can be tested in-depth; (ii) real scRNA-seq data, profiled from cell lines and colorectal tissue and (iii) other omics data, including scATAC-seq and single-cell DNA methylation.

All the performed analyses are reproducible using the OT-scOmics Jupyter notebook provided on GitHub (https://github.com/ComputationalSystemsBiology/OT-scOmics). Users can also employ OT-scOmics to test the various metrics on new single-cell data and to evaluate the performances of other/new metrics.

2 Materials and methods

2.1 scRNA-seq data simulation

scRNA-seq data with 5000 genes and three underlying clusters have been simulated using the R Bioconductor package Splatter (Zappia *et al.*, 2017). Splatter simulates scRNA-seq data using the Splat model, built around a Gamma-Poisson distribution. Different parameters can be tuned in the Splatter simulation. Details on the parameters used to run Splatter are available in Supplementary Text S1.

Using simulated scRNA-seq data, we can assess in-depth the influence of different factors on the quality of the inferred cell–cell similarities. We simulated five scRNA-seq datasets (Table 1) obtained by varying three main factors:

1. The number of cells constituting the scRNA-seq dataset (batchCells parameter in Splatter). Datasets with 500, 1000 and 10 000 cells are simulated. Of note, we thereby also test whether the different metrics can be computed for large numbers of cells.
2. The overlap of the clusters (de.prob parameter in Splatter). Overlapping and well-separated clusters are simulated varying de.prob between 0.4 and 0.7, respectively.
3. The equal or unbalanced size of the clusters (group.prob parameter in Splatter). We set the probabilities of the clusters either equal ($1/3$, $1/3$, $1/3$) or unbalanced (0.75, 0.20, 0.05). The unbalanced case reflects the more realistic scenario of a tissue composed of a mixture of prevalent and rare cell types or states.

2.2 Single-cell omics data acquisition and preprocessing

Several publicly available single-cell omics datasets have been employed (Table 1). Only public datasets providing ground-truth labels for all the profiled cells were considered. The labels are intended to associate each cell to a specific group (e.g. cell type, cell line of origin). Of note, labels are only used to evaluate the quality of our results, as all the performed benchmarking is unsupervised.

For scRNA-seq data four datasets have been considered. First, the scRNA-seq data (called 'Liu scRNA') present in the scCAT-seq joint profiling of (Liu *et al.*, 2019) containing 206 cells profiled from three cancer cell lines (HCT116, HeLa-S3, K562). Next, a bigger dataset composed of 561 cells profiled from seven cell lines (A549, GM12878, H1437, HCT116, IMR90, H1, K562) from Li *et al.* (2017) (called 'Li cell lines'). Finally, two colorectal cancer (CRC) datasets, corresponding to primary CRC tumors and matched normal mucosa have been also taken into account. The first (called 'Li Tumor') contains 364 cells clustered into seven cell types: B cells, endothelial cells, epithelial cells, fibroblasts, macrophages, mast cells and T cells. The second (called 'Li NM') is composed of 266 cells clustered according to the same seven cell types.

Other single-cell omics are also included in our analysis: methylation and scATAC-seq data. For scATAC-seq data we considered: (i) the dataset included into the scCAT-seq joint profiling of Liu *et al.* (2019) (called 'Liu scATAC-seq') composed of 206 cells extracted from three cancer cell lines (HCT116, HeLa-S3, K562); (ii) the leukemia scATAC-seq data from Corces *et al.* (2016) (called 'Leukemia scATAC'), containing 391 cells and composed of monocytes and lymphoid-primed multipotent progenitors (LMPP) isolated from a healthy human donor, together with leukemia stem cells (SU070_LSC, SU353_LSC) and blast cells (SU070_Leuk, SU353_Blast), isolated from two patients with acute myeloid leukemia. To represent single-cell DNA methylation, we considered two neuronal snmC-seq datasets from Luo *et al.* (2017). The first (called 'scMethylation mouse') is composed of 3377 cells extracted from mouse frontal cortical neurons clustered into 16 neuronal subtypes. The second (called 'scMethylation human') is composed of 2740 cells, extracted from human frontal cortical neurons and clustered into 21 neuronal subtypes.

A summary of the considered datasets is available in Table 1. The downloaded datasets had already undergone standard preliminary preprocessing and, following standard practices (Luecken and Theis, 2019), we log-transformed the scRNA-seq counts and selected the 10 000 most varying features. Alternative preprocessing strategies for scRNA-seq (Hafemeister and Satija, 2019; Lun *et al.*, 2016; Yip *et al.*, 2017) have also been tested using the code provided in Chen *et al.* (2021), with no impact on the results (Supplementary Fig. S1 and Tables S1 and S2).

2.3 Baseline metrics

We consider a single-cell omics dataset as a matrix X, whose columns correspond to cells, and whose rows correspond to features (e.g. peaks, genes). Given two cells indexed by l and m as columns of X, different metrics are classically used to infer the similarity between their omic profiles $x = X[\cdot, l] = (x_1 \dots x_n)$ and $y = X[\cdot, m] = (y_1 \dots y_n)$. We here focus on four state-of-the-art metrics, henceforth called *baseline metrics*, and defined as

1. Euclidean distance (L2): $||x - y||_2 := \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{1/2}$
2. Manhattan distance (L1): $||x - y||_1 := \sum_{i=1}^{n} |x_i - y_i|$
3. Cosine similarity: $\cos(x,\ y) := \frac{x \cdot y}{||x|| \times ||y||}$, where $||\cdot||$ is the Euclidean norm
4. Pearson correlation: $\mathrm{corr}(x,\ y) := \cos(x - \overline{x},\ y - \overline{y})$, where $\overline{x}$ and $\overline{y}$ are the mean of the values of x and y, respectively.

For cosine similarity and Pearson correlation, we used the distance-like formulation $1 - \cos(x,\ y)$ and $1 - \mathrm{corr}(x,\ y)$. Of note, these are not strictly speaking distances, in particular, they do not respect the triangular inequality. The baseline metrics have been computed using functions from the Python package SciPy (scipy.spatial.distance): euclidean, cityblock, cosine, correlation.

Baseline metrics have been computed on the input data, after the data preprocessing detailed in the section above. Optional per-cell normalization and feature scaling have been also considered (Fig. 1), but they resulted to be less performant (Supplementary Table S1).

Table 1. Summary of the datasets used for our benchmark and their clustering results

Dataset name	Data description					Clustering results			
	Data type	Reference	Number of cells	Number of features after preprocessing	Number of clusters in ground-truth	Detected number of clusters for hierarchical clustering (OT)	Detected number of clusters for hierarchical clustering (Pearson)	Detected number of clusters for Leiden clustering (OT)	Detected number of clusters for Leiden clustering (PCA + Euclidean)
500 cells	Simulated scRNA-seq data	Splatter	500	5000	3	3	3	3	11
1000 cells		Splatter	1000	5000	3	3	3	3	12
10 000 cells		Splatter	10 000	5000	3	3	5	9	12
Unbalanced clusters		Splatter	1000	5000	3	3	3	5	8
Overlapping clusters		Splatter	1000	5000	3	3	3	3	12
Liu scRNA	scRNA-seq	Liu *et al.*	206	10 000	3	3	3	3	3
Li Tumor		Li *et al.*	364	10 000	7	3	6	2	2
Li NM		Li *et al.*	266	10 000	7	9	23	6	4
Li cell lines		Li *et al.*	561	10 000	7	8	10	9	7
Liu scATAC	scATAC-seq	Liu *et al.*	206	10 000	3	3	3	3	7
Leukemia scATAC		Corces *et al.*	391	7602	6	3	25	4	3
scMethylation mouse	Single-cell DNA methylation	Luo *et al.*	3377	10 000	16	3	3	14	18
scMethylation human		Luo *et al.*	2740	10 000	21	3	7	16	32

Note: In the first part of the table ('Data description'), for each dataset, we specify the name with which we denote it in the paper, the reference to its original publication, the type of data, the number of cells, the number of features after preprocessing and the ground-truth number of clusters (e.g. cell types, cell lines) present in the data. In the second part of the table ('Clustering results'), we report the number of clusters obtained by maximizing the silhouette score, with hierarchical clustering (for Pearson correlation and OT distance) and for a typical single-cell clustering workflow based on Leiden clustering (for the Euclidean distance on PCA components and for the OT distance).

Fig. 1. Workflow for metrics comparison. The employed procedure, from the input preprocessed data to the performance evaluation is summarized for (**A**) baseline metrics and (**B**) OT, respectively. The graphic contents in the figure are taken from flaticon.com

2.4 Optimal Transport distance

OT, as defined by Monge (1781) and Kantorovich (1942), aims at finding a coupling between two probability distributions that minimizes transportation cost. The classical OT distance, also known as the Wasserstein distance, between two distributions $a = (a_1 \dots a_n)$ and $b = (b_1 \dots b_n)$ is defined as the minimal cost of transportation to morph a into b. Given a and b discrete probability distributions, their OT distance is thus defined as

$$W_C(a,b) = \min_{P \in \mathbb{R}_+^{n \times n}} < P, C > = \min_{P \in \mathbb{R}_+^{n \times n}} \sum_{i,j} P_{ij} C_{ij}, \qquad (1)$$

$$\text{with } P \in \mathbb{R}_+^{n \times n} \text{such that } \textstyle\sum_j P_{ij} = a_i \text{ and } \sum_i P_{ij} = b_j,$$

where P is the coupling. According to P, the mass in the discrete probability distribution a is thus moved from one bin to another one in order to transform a into b. C is called *ground cost* and it encodes the penalty for moving a unit of mass from one bin to another one. Hence, C should be chosen in such a way that similar bins i and j have a low cost C_{ij}.

We propose the use of the OT distance to capture cell–cell similarity in different single-cell omics data. For simplicity, let us consider a scRNA-seq dataset (X); the same concepts apply to other single-cell omics. Given a pair of cells l and m, as done for baseline metrics, we consider their expression profiles, corresponding to the vectors $x = X[\cdot, l] = (x_1 \dots x_n)$ and $y = X[\cdot, m] = (y_1 \dots y_n)$. Given that Equation (1) is defined only for discrete probability distributions, we transform x and y into two discrete probability distributions $a = \frac{x}{||x||_1}$ and $b = \frac{y}{||y||_1}$. In the following, we refer to such a transformation as *per-cell normalization*. After transformation, we can apply Equation (1) and compute the OT distance $W_C(a, b)$, between a and b, which corresponds to evaluating the minimal cost required to transform the gene expression discrete probability distribution (a) of the first cell into the gene expression discrete probability distribution (b) of the second cell. Based on Equation (1), OT computes the distance between a pair of cells (a, b) by taking into account the joint gene expression activity present in the two cells and encouraging genes to exchange mass according to the coupling P. This exchange of mass is expected to happen in between genes that are related to each other, such as genes involved in the same regulatory program. In consequence, we expect the OT similarity between cells to not be driven by specific genes, but by the overall activity of their regulatory programs (see Supplementary Text S2).

As discussed above, the OT distance is parametrized by the ground cost C. The ground cost reflects the cost of moving a unit of gene expression from a gene to another gene. The choice of the ground cost plays a central role in the final performances of OT and, in our case, there is no straightforward choice. Also ground costs based on prior information (e.g. pathways) could be employed. However, dataset/tissue specific ground costs are expected to be able to be more performant. It is for this reason that for all couples of genes i and j in the single-cell matrix X, we define

$$C_{ij} = d(X[i,\cdot], X[j,\cdot]), \quad (2)$$

where d corresponds to a metric among Euclidean and Manhattan distance, Cosine similarity and Pearson correlation. For a comparison of the performances of these different ground costs, see Supplementary Table S2. Of note, nonlinear distances have also been employed to compute the ground cost in applications of OT to single-cell (Bellazzi *et al.*, 2021; Yang *et al.*, 2020). We here chose linear distances as they do not require to be adjusted based on the omics under analysis.

Given that the number of features (e.g. genes, peaks) of single-cell data is in the order of tens of thousands, the classical OT problem would be computationally intractable. Indeed, solving Equation (1) relies on costly linear programming methods (Peyré and Cuturi, 2019). Hence, we considered the entropic regularization of the classical OT distance, also called Sinkhorn divergence (Cuturi, 2013; Genevay et al., 2019). The entropic-regularized OT distance between two distributions a and b is defined as

$$W_{C,\varepsilon}(a,\ b) := \min_{P \in \mathbb{R}_+^{n \times n}} < C, P > + \varepsilon \sum_{i,j} P_{ij}(\log(P_{ij}) - 1), \quad (3)$$

$$\text{with } P \in \mathbb{R}_+^{n \times n} \text{ such that } \sum_j P_{ij} = a_i \text{ and } \sum_i P_{ij} = b_j.$$

The first term of Equation (3) corresponds exactly to Equation (1) with P coupling and C ground cost. The additional term corresponds to the entropic regularization. Therefore, if the regularization parameter ε is set to zero, Equation (3) corresponds exactly to Equation (1) and classical OT is obtained. Increasing values of ε correspond to a more diffused coupling. From a biological perspective, we expect the introduction of the entropic regularization to allow to control for the systematic noise due to the stochasticity of gene expression at single-cell level and for the presence of dropouts, as motivated by the tests reported in Supplementary Table S3 and Text S3. At the same time, the entropic regularization allows a faster execution of the algorithm thus opening to the possibility of analyzing single-cell datasets bigger than those that can be analyzed with the classical OT (Supplementary Table S3 and Text S3). Finally, the entropic regularization, encouraging exchanges of mass in between features, allows OT to give more importance to the relationships between features (e.g. genes), which further motivates its application to complex data like single-cell data.

The parameter ε thus plays a central role in the final performances of Sinkhorn divergence and should be carefully chosen. The advantage of the formulation at Equation (3) is that $W_{C,\varepsilon}$ can be efficiently computed on a GPU, thereby coping with the high dimensionality of single-cell data. Not only is entropy pivotal to scale the algorithm, but it is also important to break the curse of dimensionality which makes classical OT distance extremely hard to estimate from high-dimensional single-cell data. This phenomenon, analyzed theoretically in Genevay et al. (2019) is supported by our analysis (Supplementary Table S2).

An issue with Equation (3) for $\varepsilon > 0$ is that, given two cells having the same expression distribution $(a = b)$, $W_{C,\varepsilon}(a,\ b) > 0$. We thus used the debiased Sinkhorn divergence (Feydy et al., 2019) in order to ensure a distance equal to zero for identical cells

$$\overline{W}_{C,\varepsilon} := W_{C,\varepsilon}(a,\ b) - \left(W_{C,\varepsilon}(a,\ a) + W_{C,\varepsilon}(b,\ b)\right)/2. \quad (4)$$

For sake of simplicity in the rest of the paper, we will use the term *OT distance* to refer to the debiased Sinkhorn divergence.

As discussed above, the OT distance depends on two main parameters: the regularization parameter ε and the ground cost C. For every dataset, we performed a grid search varying ε among 1000, 5, 1, 0.5, 0.1, 0.05, 0.01 and the ground cost among Euclidean, Manhattan, Cosine and Pearson correlation. As shown in Supplementary Table S2, the best performances on average across datasets and ground costs were obtained for the regularization parameter ε set at 0.5. We thus suggest 0.5 as default ε for future users. In contrast, the best performing ground cost C varied depending on the analyzed data. In scRNA-seq simulated data and in single-cell DNA methylation data, Pearson correlation achieved the best performances, while on scRNA-seq and scATAC-seq data, cosine similarity performed the best. The performances presented in Section 3 thus correspond to this choice of ε and C. The computation of the OT distance has been implemented using the Python package PyTorch and run on a GPU. Computation times are listed in Supplementary Table S4.

2.5 Performance evaluation

For all single-cell datasets, simulated and real, ground-truth labels are available. For the real single-cell omics, the ground-truth labels correspond to cell types, defined through clustering in the original publication, or to the cell lines from which the cells have been extracted.

We first use the C-index and Silhouette score to evaluate to which extent the various metrics detect the similarity between cells associated with the same label, as well as the difference between cells with different labels (Fig. 1). The C-index (Hubert and Schultz, 1976) is an internal clustering evaluation index. Given a cell-to-cell distance matrix, the C-index measures if the closest pairwise distances correspond to cells belonging to the same cluster. It is defined as

$$C = \frac{S - S_{\min}}{S_{\max} - S_{\min}},$$

where n is the number of intracluster pairwise distances, $S_{\min}$ is the sum of the n smallest distances if all pairs of cells are considered, $S_{\max}$ is the sum of the n largest distances out of all pairs and S is the sum of distances over all pairs of cells form the same cluster.

Of note, the C-index is always in the interval [0, 1] and it should be minimum in the case of a perfect clustering. To make the results easily readable, we consider $1 - C$, so that the best performances are obtained by maximizing the score. Concerning the implementation, we used our own Python implementation of the C-index.

As a complementary evaluation, we further considered the Silhouette score, defined as

$$S(x) = \frac{E(x) - e(x)}{\max(E(x), e(x))},$$

where $E(x)$ is the average distance between the cell x and the other cells of the same cluster and $e(x)$ is the average distance between x and cells in the closest different clusters. The distance used to compute $E(x)$ and $e(x)$ varies among the benchmarked metrics (OT, Euclidean, Manhattan, Cosine similarity and Pearson correlation). The global Silhouette score is then obtained by averaging $S(x)$ over all cells. We used the Silhouette score implementation of the Python package scikit-learn (sklearn; Pedregosa *et al.*, 2011).

To further test the quality of the inferred cell-to-cell distance matrices, we used them as inputs for a clustering algorithm and assessed the agreement between the inferred clusters and the ground-truth labels (Fig. 1). We considered clustering algorithms for which applications to single-cell data have been already proposed and directly applicable to the cell-to-cell distance matrices. We thus selected hierarchical clustering, with complete linkage, and spectral clustering (Von Luxburg, 2007; Zheng *et al.*, 2019), both implemented in scikit-learn. Regarding hierarchical clustering, we chose complete linkage in place of average linkage, used in other single-cell clustering works (Guo *et al.*, 2015), because this approach provided better performances for both baseline and OT distances (Supplementary Table S5). Concerning spectral clustering, since it requires in input an affinity matrix, we converted the inferred distance matrices D to affinity matrices $A = 1 - D$, with D normalized such that the maximum value is set to 1, and run the clustering algorithm with default parameters. Both clustering algorithms require the specification of the number of clusters in which the cells should be partitioned. In addition, we considered the typical single-cell clustering workflow (Luecken and Theis, 2019) composed of: (i) dimensionality reduction (DR) (PCA), (ii) kNN graph construction based on Euclidean distance and (iii) Leiden/Louvain clustering (Blondel *et al.*, 2008; Traag *et al.*, 2019) of the obtained graph using Scanpy default parameters (Wolf *et al.*, 2018). In case of OT, we only applied the two last steps of the workflow. Indeed, PCA relies on

Euclidean distance and would thus negatively affect the results of OT. The considered clustering algorithms depend on different parameters. To apply spectral and hierarchical clustering, the user needs to set a desired number of clusters (k). On the opposite, the results of Leiden and Louvain depend on the resolution (r), which indirectly affects the number of clusters. We thus set these parameters (k, r) in an unsupervised way, by optimizing the Silhouette score (Rousseeuw, 1987). For spectral and hierarchical clustering, we varied the number of clusters ($k = 3 - 25$), while for Leiden and Louvain we varied the resolution parameter ($r = 0.25 - 1.5$) and chose the values maximizing the Silhouette score of the clustering. Thereby, we can test how frequently the number of ground-truth labels present in the data are captured by the different metrics. Of note, the overall behavior observed when the number of clusters is optimized is in good agreement with the results obtained by fixing the number of clusters to the number of ground-truth labels (Supplementary Table S5). To evaluate the quality of the obtained clusters, we used Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI; Fig. 1).

Given U, clustering inferred from a distance matrix and V, ground-truth labels, the ARI is defined as

$$\mathrm{ARI(U,V)} = \frac{\mathrm{RI}(U,V) - E(\mathrm{RI})}{\max(\mathrm{RI}) - E(\mathrm{RI})},$$

where $\mathrm{RI}(U,V)$ is the Rand index, i.e. the fraction of pairs of samples that are either in the same group or in different groups in both U and V, $E(\mathrm{RI})$ is the expected Rand index between U and a random V, and $\max(\mathrm{RI})$ is the largest possible Rand index between U and any V.

To consider a complementary score, we also used the NMI, defined as

$$\mathrm{NMI(U,V)} = \frac{2\ \mathrm{MI}(U,V)}{H(U) + H(V)},$$

where $\mathrm{MI}(U,V)$ is the mutual information between U and V, i.e.

$$\mathrm{MI}(U,V) = H(U) - \mathrm{H}(U|V)$$

and $H(\cdot)$ denotes entropy. To compute ARI and NMI, we used the corresponding scikit-learn implementations.

3 Results

Debiased entropic-regularized OT distance (see Section 2), henceforth called OT distance, is here proposed as a metric to infer cell–cell similarity across different single-cell omics data. The performances of the OT distance are then benchmarked with respect to state-of-the-art metrics, henceforth called baseline metrics, namely the Euclidean and Manhattan distances, Cosine similarity and Pearson correlation.

The benchmark is performed in three main contexts (Table 1). First, simulated scRNA-seq data are considered. Data composed of different numbers of cells are generated to test whether the metrics scale also to high-dimensional data, as the currently available single-cell data, and if the number of cells impacts performances. Then the overlap and size of the clusters underlying the simulated scRNA-seq data are also varied to challenge the various metrics in detecting less clear and rare groupings of cells. Second, four real scRNA-seq data,

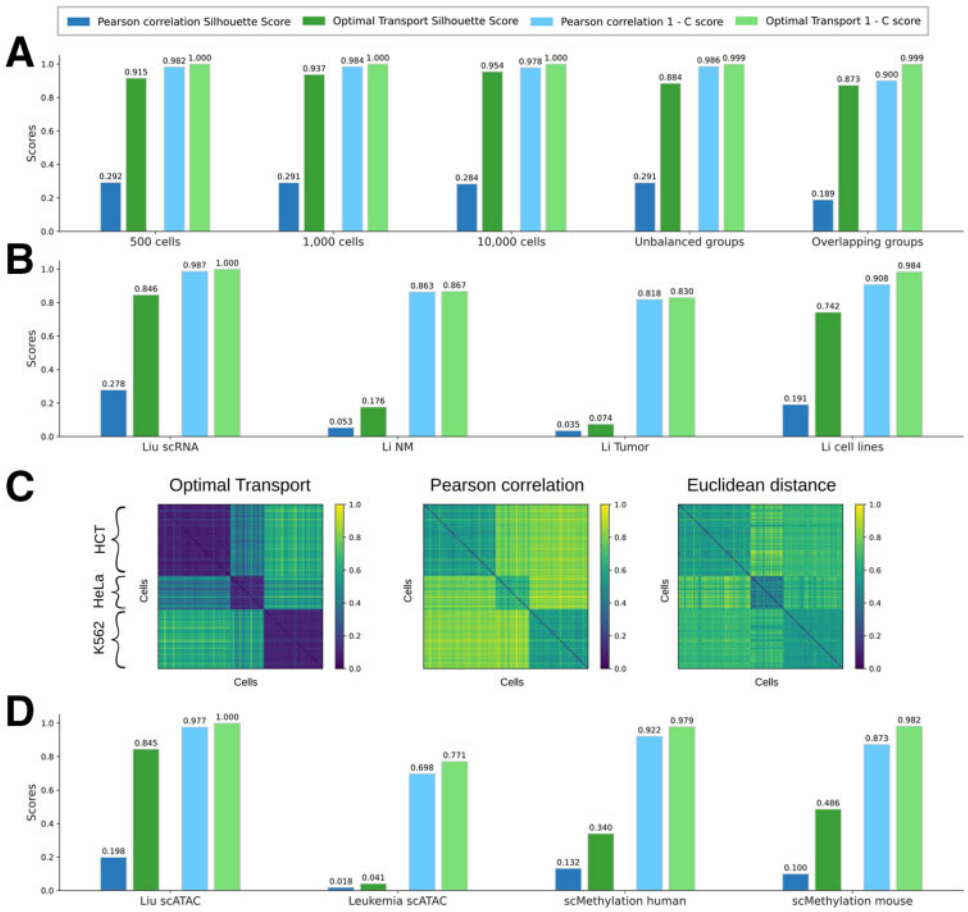

Fig. 2. Comparison of OT against Pearson correlation in cell–cell similarity inference. Barplots for C-index and Silhouette score are reported for (**A**) simulated scRNA-seq data composed of 500, 1000 and 10 000 cells, with unbalanced groups and overlapping clusters; (**B**) four scRNA-seq datasets; (**D**) two single-cell DNA methylation and two scATAC-seq data. Examples of the distance matrices obtained with OT, Pearson correlation and Euclidean distance in Liu scRNA-seq are reported in (**C**)

profiled from CRC and cell lines, are considered. Finally, we further challenged the distance measures on other single-cell omics: scATAC-seq and single-cell DNA methylation. See Section 2 and Table 1 for details concerning the data.

In all three contexts, ground-truth labels were available for all cells. In simulated scRNA-seq data, the ground-truth labels have been imposed during the data simulation (see Section 2). In real single-cell omics, the ground-truth labels are extracted from the original publications. For the four profilings of cell lines (Liu scRNA, Li cell lines, Liu scATAC and Leukemia scATAC), labels correspond to the cell line of origin and thus reflect strong transcriptional/epigenetic differences. In contrast, in the case of the neuronal single-cell DNA methylation (scMethylation mouse and scMethylation human) and scRNA-seq from CRC samples (Li Tumor and Li NM), the ground-truth labels reflect clusters previously identified in the data, based on the activity of predefined markers. These last applications are clearly more challenging, as much weaker differences exist between different cell types or states.

As summarized in Figure 1, we first tested with C-index and Silhouette score whether the various metrics are able to detect the similarity between cells belonging to the same group. We then applied different unsupervised clustering algorithms to the computed distance matrices and tested the quality of the resulting clusters using the ARI and NMI.

3.1 Comparing OT against baseline metrics based on cell–cell similarity detection

Figure 2 summarizes the results of the comparison between OT and baseline metrics. Pearson correlation outperformed the other baseline metrics on all data types. We thus used it as representative of baseline metrics in Figure 2 and in the following. The results of alternative baseline metrics are available in Supplementary Table S1. Of note, better performances of Pearson correlation with respect to other state-of-the-art metrics had been previously observed (Kim *et al.*, 2019).

In all simulated data, OT outperforms all baseline metrics (Fig. 2A), both in terms of C-index and Silhouette score. In particular, the results of OT are not impacted by the number of cells, given that it shows a consistently performant behavior for 500, 1000 and 10 000 cells. These results also suggest that OT is scalable to high-dimensional datasets, a crucial feature for the analysis of single-cell data. Finally, OT achieves superior performance with highly unbalanced clusters, which is more realistic as biological samples are often composed of a mixture of rare and common populations of cells, and also with overlapping clusters, which reflects the scenario of subpopulations of cells sharing similar transcriptional patterns, as cell populations tracked over different developmental phases.

We then considered four scRNA-seq datasets: two of them correspond to cancer cell lines (Li *et al.*, 2017; Liu *et al.*, 2019), while the remaining two correspond to colorectal tumor tissue and matched normal mucosa (Li *et al.*, 2017). All metrics tend to perform better in cancer cell lines than CRC samples. This result is most probably the consequence of the stronger transcriptional difference existing between cell lines. Overall, in all the four scRNA-seq datasets, OT outperformed baseline metrics (Fig. 2B). The improvement provided by OT in cell lines is important, especially according to Silhouette score (+0.6 Silhouette score). To show to which extent C-index and Silhouette score reflect a clear clustering structure in the cell-to-cell distance matrices, we focused on the smallest dataset, Liu scRNA (Liu *et al.*, 2019). Figure 2C reports the cell-to-cell distance matrices obtained for this dataset with OT distance, Pearson correlation and Euclidean distance. Cells in the matrices are sorted based on their cell line of origin. OT powerfully detected the similarity between cells belonging to the same cancer cell line, showing three clear blocks of cells at distance close to zero. In contrast, the blocks corresponding to the three cell lines are less marked with Pearson correlation. Finally, with Euclidean distance, the values outside the three blocks tend to be less close to one, indicating a less clear separation between cells belonging to different cell lines.

Finally, we challenged the various metrics on other single-cell omics data (Fig. 2D). Also in this case, OT shows better performance than Pearson correlation. Overall, OT performed better than existing metrics in the detection of cell–cell similarities in all considered datasets. See Supplementary Table S1 for the performances of other state-of-the-art baseline metrics.

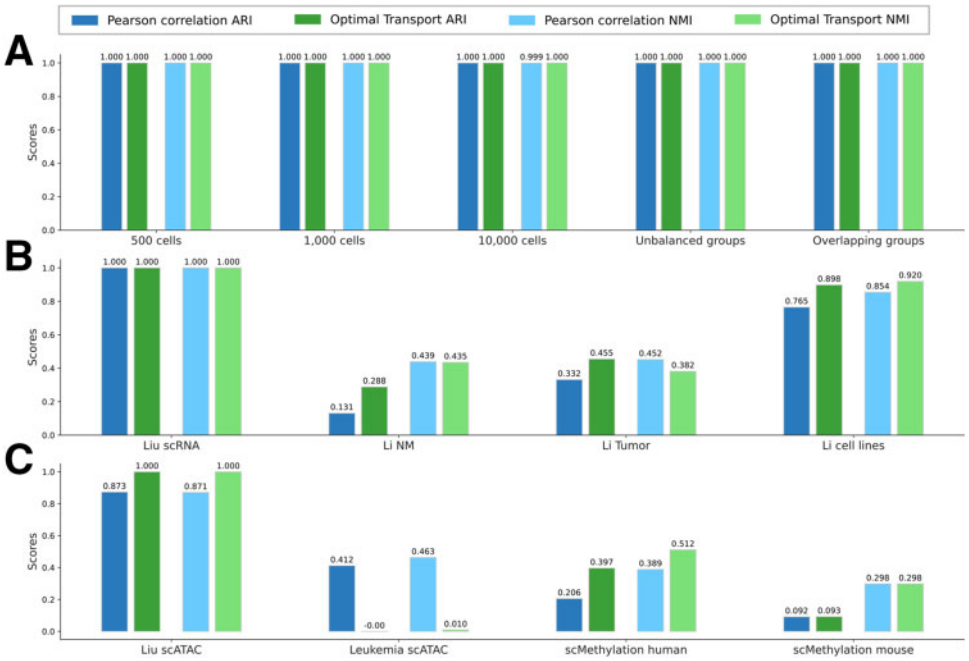

Fig. 3. Comparison of OT against Pearson correlation in hierarchical clustering. Barplots for ARI and NMI are reported for (**A**) simulated scRNA-seq data composed of 500, 1000 and 10 000 cells, with unbalanced groups and overlapping clusters; (**B**) four scRNA-seq datasets; (**C**) two single-cell DNA methylation and two scATAC-seq data

3.2 Comparing OT against baseline metrics based on hierarchical clustering

Hierarchical clustering was applied to the cell-to-cell distance matrices computed with OT and Pearson correlation. In all datasets, the number of clusters was optimized based on the Silhouette score (Section 2). Table 1 summarizes the number of clusters obtained across all datasets and compares them with those defined in the original publications. The resulting clusters were then compared with the ground-truth labels based on ARI and NMI (Fig. 3).

In simulated data, all the five datasets contain three underlying clusters. As described in Table 1, in four out of five cases, both OT and Pearson correlation identified the correct number of clusters. The only exception is represented by the dataset composed of 10 000 cells, where OT correctly detects the presence of three clusters, while Pearson correlation overclustered the data, by subdividing the three groups into five clusters. According to ARI and NMI, both Pearson correlation and OT give rise to perfect results (Fig. 3A). Pearson correlation also led to optimal ARI and NMI scores in the dataset of 10 000 cells, as overclustering is not captured by these scores.

Turning to real scRNA-seq data, in Liu scRNA data, both Pearson correlation and OT detected the correct number of clusters (Table 1). Concerning the two CRC datasets, for Li Tumor, containing seven cell types, OT only detected three clusters while Pearson predicted the presence of six clusters, thus missing only the rare population of mast cells, represented by one cell. In contrast, for Li NM, Pearson correlation dramatically overclustered the data and inferred 23 clusters, while OT suggested the presence of nine clusters. Finally, for Li cell lines, composed of seven cell lines, OT suggested the presence of eight clusters, while Pearson identified 10. Overall, in all scRNA-seq datasets, excepting Liu *et al.* (2019), no distance captured exactly the numbers of clusters reported in the corresponding publications. However, the number of clusters inferred by OT tend to be closer to the ground-truth, while Pearson correlation tended to highly overcluster the data. Regarding the NMI and ARI scores, in two out of four datasets, OT performed better than Pearson correlation (Fig. 3B), while for Liu scRNA and Li Tumor both OT and Pearson achieved comparable performances.

In other single-cell omics data, OT still performs better than Pearson correlation in the majority of the datasets. Regarding the numbers of clusters, in Liu scATAC, both Pearson correlation and OT correctly detected the presence of three clusters. In Leukemia scATAC, composed of six cell lines, OT and Pearson correlation predicted 3 and 25 clusters, respectively. Finally, in methylation data, both OT and Pearson underestimated the real number of clusters. As shown in Figure 3C, according to ARI and NMI values, OT showed better performances in all datasets except Leukemia data. Of note, Pearson correlation performed well according to ARI and NMI for Leukemia data, but this performance is obtained considering 25 clusters instead of the five real clusters presumably present in the data. This is a practical demonstration of the interest of inferring the optimal number of clusters based on the Silhouette score, rather than fixing it to the value reported in their original publication.

The improvement provided by OT for cell–cell similarity inference (Fig. 2) has an impact on clustering performances. Of note, these results are in agreement with those observed when the number of clusters is fixed (Supplementary Table S5). Finally, similar results are obtained when substituting hierarchical clustering with spectral clustering (Supplementary Text S4 and Fig. S2), suggesting that the choice of clustering algorithm does not affect conclusions.

3.3 Comparing a typical single-cell clustering workflow against its counterpart based on OT

Clustering in single-cell is classically performed following a typical workflow composed of: (i) DR (PCA), (ii) kNN graph construction based on Euclidean distance and (iii) Leiden/Louvain clustering of the obtained graph (Luecken and Theis, 2019). We here compare the results obtained with this typical single-cell clustering workflow with respect to its counterpart based on OT (see Section 2 for details). Figure 4 and Table 1 report the results obtained once the resolution parameter for the Leiden clustering is optimized based on the Silhouette score (Section 2). Of note, the results are not affected by the choice of the resolution parameter (Supplementary Fig. S3).

In simulated data, the estimation of the number of clusters is less precise than with hierarchical and spectral clustering. Overall, the typical clustering workflow tends to more frequently overcluster the data (Table 1). Regarding ARI and NMI scores, OT shows better performances, reaching perfect score for three out of five datasets (Fig. 4A).

For scRNA-seq data, both the typical clustering workflow and OT correctly predicted the correct amount of clusters present in Liu scRNA. The two CRC datasets contain seven cell types, corresponding to our ground-truth labels. For Li Tumor, both the typical clustering workflow and OT predicted the presence of two clusters. In contrast, for Li NM, the typical clustering workflow inferred four clusters, while OT suggested the presence of six clusters. Finally, for Li cell lines, OT suggested the presence of nine clusters, while the typical clustering workflow correctly identified seven clusters. Concerning the NMI and ARI scores, OT outperformed the typical clustering workflow in all four scRNA-seq datasets (Fig. 4B).

In other single-cell omics, OT correctly identified three clusters for Liu scATAC data, while the typical clustering workflow found seven clusters. Regarding Leukemia scATAC data, involving six cell lines, both OT and the typical clustering workflow under-clustered the data. Finally, in methylation data OT tend to be closer to the correct number of clusters. Regarding ARI and NMI (Fig. 4C), OT always outperforms the typical clustering workflow and obtains perfect performances in Liu ATAC data.

Of note, similar results are obtained when substituting the Leiden clustering with the Louvain clustering (Supplementary Text S5 and Fig. S4), indicating that the choice of clustering algorithm should not affect conclusions.

3.4 Open-source implementation and distribution

To foster the reproducibility of all the results presented in this work, we provide a Python package and Jupyter notebook covering all the analyses performed, both available at https://github.com/ComputationalSystemsBiology/ot-scOmics, together with all the preprocessed single-cell data used in this study. Since computing OT distances is computationally intensive, the code is designed to be run on a GPU, taking advantage of the PyTorch library. For users who do not have access to a GPU, extensive explanations are provided to run the Jupyter notebooks on the Google Collaboratory platform (http://colab.research.google.com/). Note that our code also allows for CPU computations.

4 Discussion

In this study, we assessed the potential of OT to infer cell–cell similarities from single-cell omics data. We extensively benchmarked OT performances against state-of-the-art metrics. Interestingly, OT outperformed alternative metrics in capturing cell–cell similarities in all the 13 considered datasets. The biological relevance of this improvement was assessed by performing cell clustering. In all cases, the use of OT distance resulted in improved clustering results. Of note, different clustering algorithms have been used to test whether the observed performances were affected by the choice of the algorithm. Finally, we further challenged the metrics to detect cell–cell similarity in other single-cell omics: scATAC-seq and single-cell DNA methylation data. The improvement provided by OT is conserved also when other single-cell omics are considered, despite the high sparsity and the close-to-binary nature of scATAC-seq data (de Souza *et al.*, 2020; Xiong *et al.*, 2019).

Single-cell datasets composed of a number of cells higher than that here employed are currently available (e.g. single-cell atlases). The scope of our analysis is to test the performances of OT and baseline measures in inferring cell–cell similarity from single-cell data. In this context, we are interested in reconstructing the distance matrices comparing all possible couples of cells present in the dataset. Such computations are quadratic in the number of cells, both in

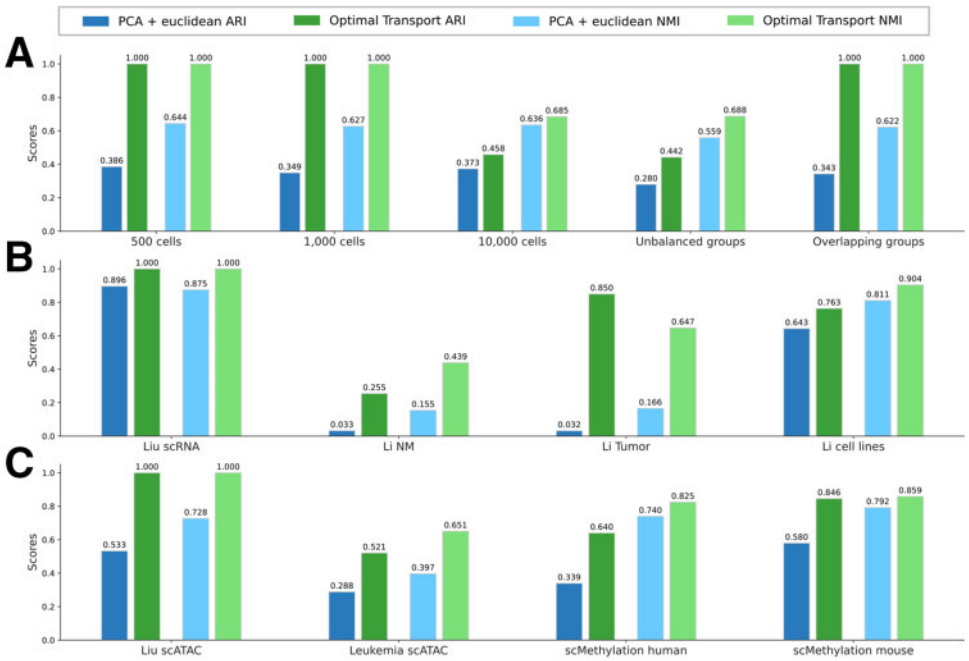

Fig. 4. Comparison of a typical single-cell clustering workflow against its counterpart based on OT. Barplots for ARI and NMI are reported for (**A**) simulated scRNA-seq data composed of 500, 1000 and 10 000 cells, with unbalanced groups and overlapping clusters; (**B**) four scRNA-seq datasets; (**C**) two single-cell DNA methylation and two scATAC-seq data

terms of memory and time complexity, thus preventing us to test performances on single-cell atlases for both baseline and OT distances. However, the computation of the complete cell–cell distance matrices is not required to cluster single-cell data, thus assuring that a clustering based on OT, or other baseline measures, can be scaled to high number of cells (e.g. single-cell atlases).

Our evaluation of metrics performances is dependent on the ground-truth labels associated with the input datasets. In cell lines, the ground-truth is well established and wide transcriptional differences exist between different cell lines. Biological samples instead consist of cell types and states having less pronounced transcriptional differences. What is a ground-truth in this case is thus less clear. We here used as ground-truth the clusters identified in the original publication of each dataset. However, such labels could be improved. In addition, given that the original publications employ Euclidean distance or Pearson correlation to define the labels, the usage of such labels as ground-truth is expected to advantage these metrics compared to alternative ones.

Of note, DR is frequently applied to reduce the noise in single-cell data before cell clustering and further downstream analyses. Here, we only applied DR in the context of Leiden/Louvain clustering for Euclidean distance, but not for the OT distance. Indeed, the most popular DR approaches, such as PCA, diffusion maps, NMF and ICA, rely on Euclidean geometry and would thus not be a good choice prior to OT computation. The good performances provided by OT in this work suggest that further efforts should be devoted to design DR methods for single-cell data based on other metrics.

Finally, we applied OT to different single-cell omics in isolation. However, different omics data presumably provide complementary information on individual cellular states. Combining different single-cell omics with appropriate metrics thus represent a critical challenge in computational biology. Our results suggest that OT could be a valuable metric for the integration of different omics data.

Acknowledgements

We thank Denis Thieffry for the scientific feedback on the work and Jean-Philippe Vert for the insightful discussion during the design of this project.

Funding

This work was supported by the Agence Nationale de la Recherche (ANR)—JCJC project scMOmix and Sanofi iTech Awards. This work was performed using HPC resources from GENCI-IDRIS [Grant 2021-AD011012285]. The work of G. Peyré was supported by the European Research Council (ERC project NORIA) and the French government under management of Agence Nationale de la Recherche as part of the 'Investissements d'avenir' program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

Conflict of Interest: none declared.

Data availability

Data from Liu et al. can be accessed through the supplementary data 3 and 4 of the original article or at the link https://github.com/ComputationalSystemsBiology/momixnotebook/tree/master/data/single-cell. Data from Li et al., Corces et al. and Luo et al. can be accessed through the Gene Expression Omnibus (GEO) accession numbers: GSE81861, GSE65360 and GSE97179 respectively. The code to reproduce the analyses is available at https://github.com/ComputationalSystemsBiology/OT-scOmics.

References

Bellazzi,R. *et al.* (2021) The gene mover's distance: single-cell similarity via optimal transport. arXiv:2102.01218.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Cao,K. *et al.* (2022) Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. Bioinformatics 38.**1**: 211–219.

Chen,W. *et al.* (2021) A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. *Nat. Biotechnol.*, **39**, 1103–1114.

Corces,M.R. *et al.* (2016) Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.*, **48**, 1193–1203.

Cuturi,M. (2013) Sinkhorn distances: lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.*, **26**, 2292–2300.

Demetci,P. *et al.* (2020) Gromov-Wasserstein optimal transport to align single-cell multi-omics data. BioRxiv.

Feydy,J. *e.t al.*. (2019) Interpolating between optimal transport and MMD using Sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019 Naha, Okinawa, Japan. PMLR, pp. 2681–2690.

Genevay,A. *e.t al.*. (2019) Sample complexity of Sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019 Naha, Okinawa, Japan. PMLR, pp. 1574–1583.

Guo,M. *et al.* (2015) SINCERA: a pipeline for single-cell RNA-seq profiling analysis. *PLoS Comput. Biol.*, **11**, e1004575.

Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 1–15.

Hubert,L. and Schultz,J. (1976) Quadratic assignment as a general data analysis strategy. *Br. J. Math. Stat. Psychol.*, **29**, 190–241.

Huizing,G.-J. *et al.* (2021) Unsupervised ground metric learning using wasserstein eigenvectors. *arXiv*:2102.06278.

Kantorovich,L. (1942) On the transfer of masses (in Russian). *Dokl. Akad. Nauk.*, **37**, 227–229.

Kim,T. *et al.* (2019) Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.*, **20**, 2316–2326.

Kiselev,V.Y. *et al.* (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.

Lee,J. *et al.* (2020) Single-cell multiomics: technologies and data analysis methods. *Exp. Mol. Med.*, **52**, 1428–1442.

Li,H. *et al.* (2017) Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.*, **49**, 708–718.

Liu,L. *et al.* (2019) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat. Commun.*, **10**, 470.

Luecken,M.D. and Theis,F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, **15**, e8746.

Lun,A.T. *et al.* (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, **5**, 2122.

Luo,C. *et al.* (2017) Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, **357**, 600–604.

Ma,A. *et al.* (2020) Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol.*, **38**, 1007–1022.

Macosko,E.Z. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Monge,G. (1781) *Mémoire sur la théorie des déblais et des remblais*. Histoire de l'Académie Royale des Sciences de Paris , Paris, France.

P E de Souza,C. *et al.* (2020) Epiclomal: probabilistic clustering of sparse single-cell DNA methylation data. *PLoS Comput. Biol.*, **16**, e1008270.

Papalexi,E. and Satija,R. (2018) Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, **18**, 35–45.

Pedregosa,F. *et al.* (2011) scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Peyré,G. and Cuturi,M. (2019) Computational optimal transport: with applications to data science. *Found. Trends Mach. Learn.*, **11**, 355–607.

Potter,S.S. (2018) Single-cell RNA sequencing for the study of development, physiology and disease. *Nat. Rev. Nephrol.*, **14**, 479–492.

Rajewsky,N. *et al.* (2020) LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature*, **587**, 377–386.

Rousseeuw,P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Santambrogio,F. (2015) *Optimal Transport for Applied Mathematicians*, Vol. **55**. Birkäuser, NY, p. 94.

Satija,R. *et al.* (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.

Schiebinger,G. *et al.* (2019) Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, **176**, 928–943.e22.

Stegle,O. *et al.* (2015) Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.*, **16**, 133–145.

Traag,V.A. *et al.* (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 1–12.

Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.

Wolf,F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 1–5.

Xiong,L. *et al.* (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**, 4576.

Yang,K.D. *et al.* (2020) Predicting cell lineages using autoencoders and optimal transport. *PLoS Comput. Biol.*, **16**, e1007828.

Yip,S.H. *et al.* (2017) Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.*, **45**, e179–e179.

Zappia,L. *et al.* (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.

Zheng,R. *et al.* (2019) SinNLRR: a robust subspace clustering method for cell type detection by non-negative and low-rank representation. *Bioinformatics*, **35**, 3642–3650.