Similarity metrics for binary cell clustering

How close can we get to state-of-the-art ?

Bartosz Golik

Supervisor(s): Marcel J.T. Reinders, Gerard A. Bouland

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Bartosz Golik
Final project course: CSE3000 Research Project
Thesis committee: Marcel J.T. Reinders, Gerard A. Bouland, Bart Gerritsen

Abstract

Analysing single-cell RNA sequencing data is becoming an increasingly tedious task as the size of data sets grows. As a proposed solution, recent discoveries suggest that these data sets can be binarized without losing much information. This in turn should allow for memory and time efficient methods of storage and computation. Numerous analyses techniques require cell clustering as a preliminary procedure, which suggests the need to evaluate binary representation performance under that context. In this work we present a comparison between binary clustering results and the state-of-the-art, with a focus on similarity metric choice and the impact on intermediate steps of the procedure (i.e. similarity matrices and kNN graphs). The method was evaluated on single-cell transcriptomic data sets, utilizing a combination of R and C++ as an evaluation framework. Through these means we found that some of the similarity metrics operating on continuous input can possibly be reproduced with similarity metrics operating on binary input.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful tool utilized in studying heterogeneity of cell populations. The method facilitates the examination of transcriptome for each individual cell in the population, which lead to discoveries previously unachievable. Specifically, scRNA-seq found its extensive use in studying diversity of cancer cells [1–3], as well as bringing insight into early embryo development [4,5] and emerging plenteous cell atlases projects [6–8], which contribute to understanding the cell biology of living organisms and disease development.

Numerous analyses techniques require cell clustering as a preliminary procedure. Recent years saw a 100-fold increase in the number of cells in scRNA-seq datasets [9], which resulted in an increase of time and memory requirements, including for the clustering process. As a solution, a change has been proposed to the representation of expression matrix - the magnitude of a count could be disregarded and only the absence or presence of a gene could be useful enough information on its own [10]. This idea of a cell clustering algorithm which operates on a binarized input will be referred to as binary cell clustering. Further analyses [9] prove that with larger datasets, the sparsity increases, which leads to more relative information being stored in the presence or absence of a count, rather than its magnitude. Initial experiments with the binary representation [9] show a 17-fold decrease in storage requirements, while preserving close to perfect correlation in quality of cell type identification on recent datasets. Thus, binary cell clustering already shows promising results at an early development stage, which suggests a need for further investigation into its potential optimisations.

As the clustering task is based on cell-to-cell transcriptome similarity, a numerical method is needed in order to decide the resemblance between two arbitrary cells. The choice of such a similarity metric can have a big impact on the quality of cell type identification, which was proven in multiple existing evaluations for single-cell clustering [11–13]. However, none of the studies seem to specifically target clustering with binarized input. Furthermore, existing evaluations keep their focus on the quality of final results - namely, the correspondence of clustering labels to the ground truth. Since the binary cell clustering is still a novel idea, analysing only the end outcomes might not be enough to discover all implications coming from the input change. Instead, we suggest taking a closer look at the comparison of similarity metrics with binary and non-binary input and the immediate following outcomes: 1) similarity matrices and 2) kNN graphs. As such, the main question in this work is: how similar are binary similarity metrics to continuous similarity metrics when applied on scRNAseq data ?

The problem was tackled with empirical data analysis. To the extent of our knowledge, there were no viable frameworks implementing binary cell clustering. As such, an entire clustering pipeline (Fig. 1) has been developed in the process alongside an evaluation framework which quantified the quality of results. Furthermore, the research involved two methods of comparison between binary and non-binary clustering - one for similarity matrices and one for kNN graphs.

We will first discuss the methodological approach, followed by the presentation and analysis of the results obtained with experiments. Afterwards, we will reflect on technical limitations and provide directions for future research of the topic. A short chapter will also be dedicated to ethical aspects of our work. A summary of key conclusions can be found in the very end of the document.

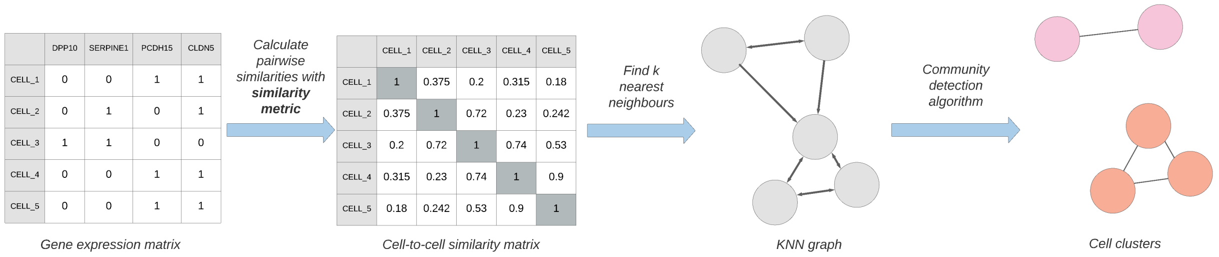
2 Methodology

2.1 Similarity metrics

It should first be noted that, formally, a similarity metric is a function that has the four properties: reflexivity, non-negativity, symmetry and triangle inequality. Some of the functions in this work do not posses all four characteristics. However, for the simplification purposes, they will still be referred to as 'similarity metrics'. Furthermore, similarity metrics that operate on a continuous scRNAseq dataset will be referred to as 'continuous metrics'. Metrics that operate on binary scRNAseq dataset will be referred to as 'binary metrics'.

2.2 Continuous metrics

In an attempt to gather a list of most popular and best performing continuous similarity metrics for non-binarized cell clustering, previous evaluations on that topic were considered. However, it seems that there is no agreement among researchers, neither with regards to which metrics perform best, nor to which of them are worth evaluating in the first place. Watson et al. [11] gathered 17 similarity metrics and showed that their performance is highly dependent on dataset characteristics. Skinnider et al. [12] also evaluated 17 metrics, with a clear recommendation for proportionality-based ones. Kim et al. [13] only included 5 metrics, arguing that correlation-based metrics outperform true distance metrics.


Calculate pairwise similarities with **similarity metric**
Find k nearest neighbours
Community detection algorithm
Gene expression matrix
Cell-to-cell similarity matrix
KNN graph
Cell clusters

	DPP10	SERPINE1	PCDH15	CLDN5
CELL_1	0	0	1	1
CELL_2	0	1	0	1
CELL_3	1	1	0	0
CELL_4	0	0	1	1
CELL_5	0	0	1	1

	CELL_1	CELL_2	CELL_3	CELL_4	CELL_5
CELL_1	1	0.375	0.2	0.315	0.18
CELL_2	0.375	1	0.72	0.23	0.242
CELL_3	0.2	0.72	1	0.74	0.53
CELL_4	0.315	0.23	0.74	1	0.9
CELL_5	0.18	0.242	0.53	0.9	1

Figure 1: Binary cell clustering pipeline assumed for the C++ implementation

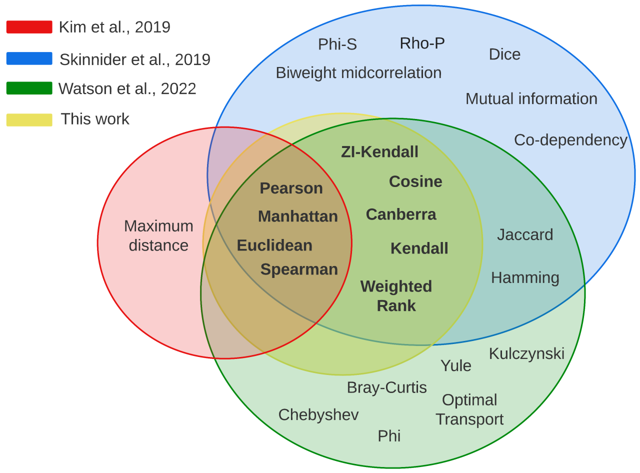


Figure 2: Venn diagram presenting continuous similarity metrics used in existing evaluations.

We chose different metrics based on their frequency of occurrence in aforementioned papers (Fig. 2). Four metrics were included in all three papers, seven more were included in both Watson and Skinnider. Both Watson and Skinnider defined Jaccard and Hamming as operating on binarized input. As that would yield trivial results in our context, these two metrics were excluded, ending up with nine continuous metrics in total.

2.3 Binary metrics

The research on binary similarity metrics, especially in the context of clustering, seems to be sparse. There is insufficient empirical analysis to easily extract a list that could be suitable for binary cell clustering.

The main issue is that, when combined with kNN algorithm, binary metrics often produce equal results. As calculating a similarity matrix is an expensive operation, knowing which metrics would produce same results was crucial to decrease time required for computation. We first discuss the problem of 'metric duplicates' in detail, followed by the process of extraction of the final list.

	$x_i = 1$	$x_i = 0$	
$y_i = 1$	a	b	$a + b$
$y_i = 0$	c	d	$c + d$
	$a + c$	$b + d$	$p = a + b + c + d$

Table 1: Frequency table for two binary sequences x and y

Identifying monotonic binary metrics

Two binary metrics can yield exact same results even if they share differences when applied to non-binary data. This issue, however, is not as trivial as it may seem, as we can differ two possibilities for that situation to happen:

(I) Metrics yield the same output

An example could include Canberra distance and Manhattan distance. By definition, Canberra is a weighted variant of Manhattan distance with formal mathematical definition given as follows:

$$D_{\text{Manhattan}} = \sum_{i=1}^{n} |x_i - y_i|$$

$$D_{\text{Canberra}} = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

When the input is binary however, Canberra's definition simply collapses into regular Manhattan:

$$D_{\text{Binary\_Canberra}} = \sum_{i=1}^{n} |x_i - y_i|$$

and always produces the same result.

(II) Metrics yield different output, but kNN ordering remains the same - metrics are monotonic.

As an example, we can construct cell-to-cell similarity matrix for the following data:

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

where each row in the matrix represents data for a single cell. Furthermore, we can use two binary metrics, Simple Matching and Rogers & Tanimoto, defined as:

$$S_{\text{Simple Matching}} = \frac{a+d}{a+b+c+d}$$

$$S_{\text{Rogers \& Tanimoto}} = \frac{a+d}{a+2b+2c+d}$$

where a, b, c, d are true positives, false positives, false negatives and true negatives, accordingly (Table 1). When applied, similarity matrices present as follows:

$$M_{\text{Simple Matching}} = \begin{bmatrix} 1 & 0.2 & 0.6 & 0.4 & 0.6 \\ 0.2 & 1 & 0.6 & 0.4 & 0.6 \\ 0.6 & 0.6 & 1 & 0.4 & 0.2 \\ 0.4 & 0.4 & 0.4 & 1 & 0.4 \\ 0.6 & 0.6 & 0.2 & 0.4 & 1 \end{bmatrix}$$

$$M_{\text{Rogers \& Tanimoto}} = \begin{bmatrix} 1 & 0.11 & 0.43 & 0.25 & 0.43 \\ 0.11 & 1 & 0.43 & 0.25 & 0.43 \\ 0.43 & 0.43 & 1 & 0.25 & 0.11 \\ 0.25 & 0.25 & 0.25 & 1 & 0.25 \\ 0.43 & 0.43 & 0.11 & 0.25 & 1 \end{bmatrix}$$

As can be seen, nearest neighbours of each cell have the same ordering for both Simple Matching and Rogers & Tanimoto. More formally, computing Spearman's Rank Correlation row-wise for their similarity matrices, produces 1 in each row. Spearman's Rank correlation of two metric 'duplicates' will always be 1 for any binary sequence - the 'duplicates' are monotonic.

Extracting metric list

There exist multiple papers listing, evaluating and describing properties of binary metrics. Most of them, however, lacked information about monotonicity relation. Todeschini et al. [14] published in 2012 a review of 51 binary similarity coefficients for binary chemoinformatics data. Out of 51, seven turned out to have perfect Pearson correlation with other metrics and thus were excluded from further evaluation.

For the remaining 44, authors measured pairwise Spearman's Rank Correlation ρ, revealing monotonicity. Based on Todeschini et al.'s work, an arrangement was created (Table 5), containing seven groups where coefficients express a high ($\rho > 0.97$) correlation, and thus are treated as duplicates. Each group was assigned an arbitrarily chosen 'representative'. The groups contained 30 coefficients in total, leaving out 14 which don't express high correlation with none other from the initial list. Furthermore, Cole I, Cole II and Dice II were excluded due to their lack of symmetry: $s(x, y) \neq s(y, x)$, leaving out a total of 18 binary metrics in the final list (Table 6).

2.4 Comparing similarity matrices

We can measure the resemblance of two matrices with Spearman's Rank Correlation. For each unique pair (b, d) of binary metric b and continuous metric d, the similarity matrix is computed, resulting in B : matrix of binary dataset and D : matrix of continuous dataset. Both matrices are of the same size $n \times n$, with $B_1, B_2, \ldots, B_n$ and $D_1, D_2, \ldots, D_n$ representing rows in corresponding matrices. For each $i \in \{1, \ldots, n\}$, Spearman's Rank Correlation $\rho(B_i, D_i)$ was calculated, producing measurements $\rho_1, \rho_2, \ldots, \rho_n$, one for each row. Finally, we computed: 1) mean value $\hat{\rho}$ of $\rho_1, \rho_2, \ldots, \rho_n$, representing overall quality of matching between b and d and 2) standard deviation ρ_{SD} of $\rho_1, \rho_2, \ldots, \rho_n$, representing how varying the matching quality is over all cells. Fig. 3 presents the entire process. For simplification purposes, $\hat{\rho}$ will be referred to as "mean Spearman" or a "matrix matching" between binary metric b and continuous metric d. Standard deviation ρ_{SD} will be referred to as "matrix matching inaccuracy".

Name/Reference	cells	genes	sparsity
Baron et al. [15]	1886	1440	85.7%
Darmanis et al. [16]	466	1669	61.6%
Fletcher et al. [17]	616	3806	70.9%
Lawlor et al. [18]	638	7444	65.4%
PBMC dataset [19]	3500	272	84.4%
Pollen et al. [20]	366	1287	68.8%
Romanov et al. [21]	2881	3046	81.0%
Tasic et al. [22]	1809	7484	61.5%

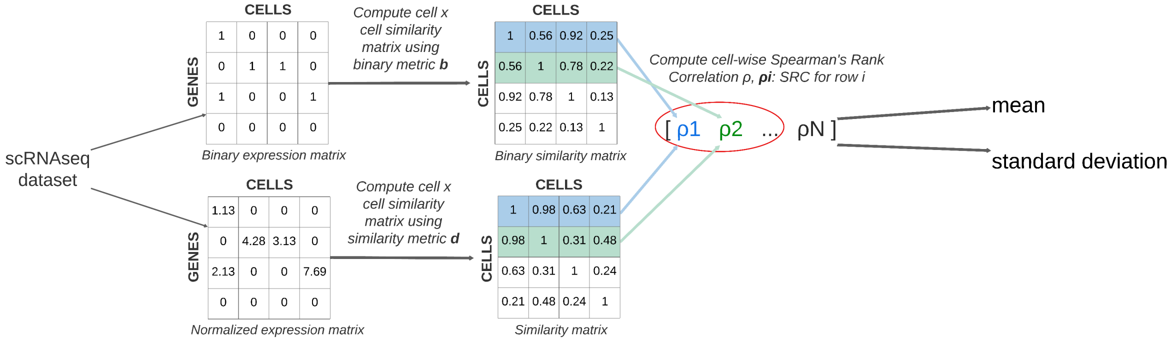
Table 2: scRNAseq datasets used in this work.

2.5 Comparing kNN graphs

Taking a step further in the clustering pipeline, a comparison between kNN graphs can be drawn in a similar manner to similarity matrices. The kNN graphs were compared in their adjacency list representation. Instead of Spearman's Rank Correlation, we used Jaccard index to compare the rows. Furthermore, the experiment was repeated for $k \in \{3, 5, 10, 15, 20, 30\}$. Fig. 4 presents the process. For simplification purposes, mean cell-wise Jaccard index for kNN graphs $\hat{\text{J}}$ will be referred to as "mean Jaccard" or a "graph matching" between binary metric b and continuous metric d. The standard deviation of Jaccard index across the cells J_{SD} will be referred to as "graph matching inaccuracy".

2.6 Datasets

Metric pairs were evaluated on 8 datasets available through R's *scRNAseq* package [23]. Data used in this work is a subset of data evaluated in [9]. Small size of datasets was imposed due to technical limitations (Section 5). Table 2 summarizes key information about datasets used in this work.


scRNAseq dataset
CELLS
GENES
Binary expression matrix
Compute cell x cell similarity matrix using binary metric **b**
CELLS
CELLS
Binary similarity matrix
Compute cell-wise Spearman's Rank Correlation ρ, **ρi**: SRC for row i
[ρ1 ρ2 ... ρN]
mean
standard deviation
CELLS
GENES
Normalized expression matrix
Compute cell x cell similarity matrix using similarity metric **d**
CELLS
CELLS
Similarity matrix

Binary expression matrix

1	0	0	0
0	1	1	0
1	0	0	1
0	0	0	0

Binary similarity matrix

1	0.56	0.92	0.25
0.56	1	0.78	0.22
0.92	0.78	1	0.13
0.25	0.22	0.13	1

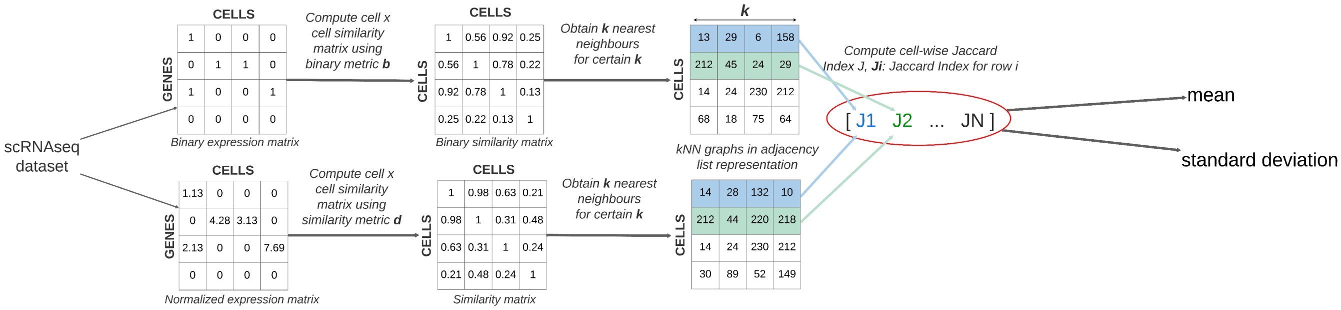
Normalized expression matrix

1.13	0	0	0
0	4.28	3.13	0
2.13	0	0	7.69
0	0	0	0

Similarity matrix

1	0.98	0.63	0.21
0.98	1	0.31	0.48
0.63	0.31	1	0.24
0.21	0.48	0.24	1

Figure 3: Computation flow for comparing similarity matrices.


scRNAseq dataset
CELLS
GENES
Binary expression matrix
Compute cell x cell similarity matrix using binary metric **b**
CELLS
CELLS
Binary similarity matrix
Obtain **k** nearest neighbours for certain **k**
k
CELLS
kNN graphs in adjacency list representation
Compute cell-wise Jaccard Index J, **Ji**: Jaccard Index for row i
[J1 J2 ... JN]
mean
standard deviation
CELLS
GENES
Normalized expression matrix
Compute cell x cell similarity matrix using similarity metric **d**
CELLS
CELLS
Similarity matrix
Obtain **k** nearest neighbours for certain **k**
CELLS

Binary expression matrix

1	0	0	0
0	1	1	0
1	0	0	1
0	0	0	0

Binary similarity matrix

1	0.56	0.92	0.25
0.56	1	0.78	0.22
0.92	0.78	1	0.13
0.25	0.22	0.13	1

kNN graphs in adjacency list representation

13	29	6	158
212	45	24	29
14	24	230	212
68	18	75	64

Normalized expression matrix

1.13	0	0	0
0	4.28	3.13	0
2.13	0	0	7.69
0	0	0	0

Similarity matrix

1	0.98	0.63	0.21
0.98	1	0.31	0.48
0.63	0.31	1	0.24
0.21	0.48	0.24	1

14	28	132	10
212	44	220	218
14	24	230	212
30	89	52	149

Figure 4: Computation flow for comparing kNN graphs.

2.7 Implementation

The entire binary cell clustering pipeline was developed in C++, which was motivated by the requirement to operate on a data type that physically implements bit sequence in computer memory, as well as time and memory efficiency considerations. The *Boost* C++ library [24] was used to physically access the bits in computer memory and *IGraph* [25] provided utility methods for creating the kNN graphs and applying community detection algorithms. Code repository is public and available at Bitbucket.org.

In order to compare the results for binary and non-binary cell clustering, a suitable implementation was also required for the latter. Conveniently, alongside the paper [12], Skinnider also published an R package, *dismay* [26] that computes similarity matrix for each metric listed in their evaluation. The package was also further used in Watson et al. [11]. To maintain consistency, *dismay*'s implementation was used for each of the 9 continuous metrics in this work. C++ and R code was integrated together using the *RCpp* package [27].

3 Results & Discussion

Based on the matrix and graph comparison results, further analysis was carried out. For each binary metric b - continuous metric d pair (b, d), computed quantities include:

- For similiarity matrices:
  - $\hat{\rho}$: mean of cell-wise Spearman's Rank Correlation, referred to as "matrix matching"
  - ρ_{SD}: standard deviation of cell-wise Spearman's Rank Correlation, referred to as "matrix matching inaccuracy"
- For kNN graphs, $k \in \{3, 5, 10, 15, 20, 30\}$:
  - $\hat{J}$: mean of cell-wise Jaccard score, referred to as "graph matching"
  - J_{SD}: standard deviation of cell-wise Jaccard score, referred to as "graph matching inaccuracy"

Furthermore, the highest matching for a particular continuous metric d across all binary metrics is referred to as "optimal matching".

It is worth noticing that *dismay* was not capable of computing ZI-Kendall metric for two of the datasets, producing "NA" as output. No other artifacts were recorded.

3.1 Not all continuous similarity metrics have a matching binary counterpart

It seems that not all continuous metrics can be consistently reproduced with binary counterparts (Fig. 6). Correlation metrics (Spearman, Weighted Rank, Kendall & ZI-Kendall), excluding Pearson, found at least one near-perfect ($\hat{\rho} > 0.95$) matrix matching in each dataset. Reproducing Pearson, Cosine & Canberra produced mixed results. In some datasets,

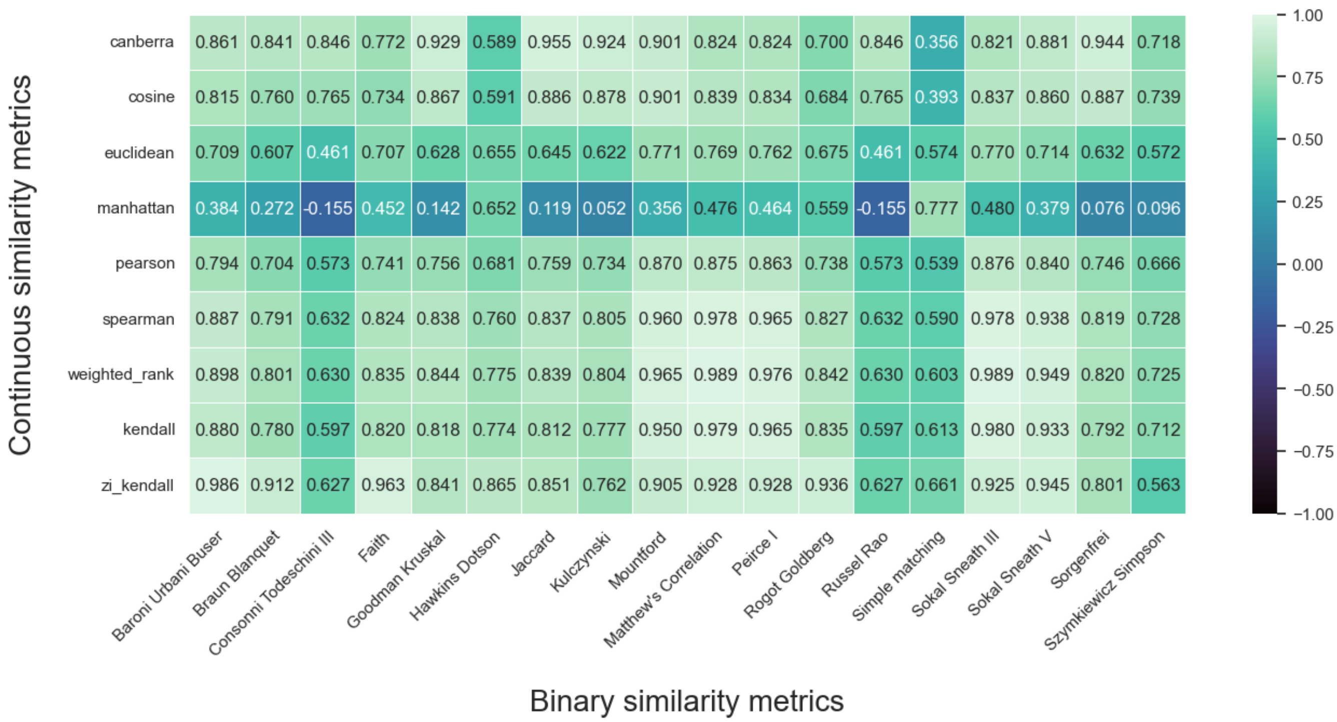


Figure 5: Matrix matching $\hat{\rho}$ between each metric pair for DarmanisBrain dataset. Other computed quantities are not included, but could be presented in the same way. This figure does not provide any conclusions by itself. Its purpose is to help the reader visualize the results of matrix & graph comparison study.

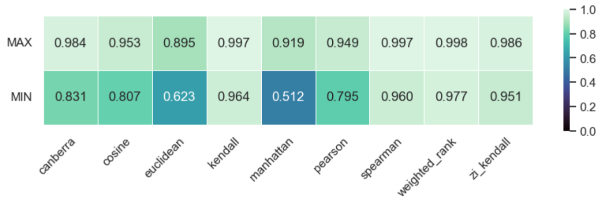


Figure 6: Lowest and upper bounds of the optimal matrix matching for each continuous similarity metric across all datasets.

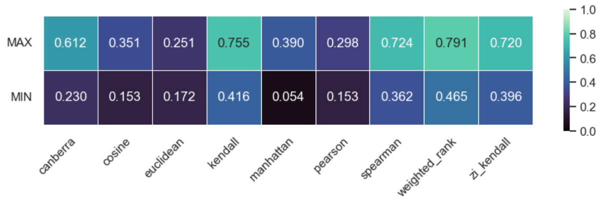


Figure 7: Lowest and upper bounds of the optimal graph matching for each continuous similarity metric across all datasets.

the metrics found a very high ($\hat{\rho} > 0.94$) optimal matrix matching, but this behaviour was not consistent across different inputs. Optimal matrix matchings for Euclidean and Manhattan distance were significantly lower than for any other continuous metric.

3.2 Optimal matchings stay consistent across datasets

Optimal matchings seem to stay consistent with different inputs - if b is an optimal matching for d in one dataset, it is highly likely to also be close to optimal in a different dataset. For instance, weighted rank metric consistently ranked Matthew's Correlation, Sokal Sneath III and Mountford amongst its top five most optimal matrix matchings (Table 3). This behaviour, to a certain degree, was observable with all continuous metrics in both matrix and graph matchings. However, some exceptions to this rule have also been observed. For instance, Sorgenfrei was the optimal matrix matching for Euclidean distance in Tasic dataset, but it was excluded from the top five in almost all other datasets (in Romanov, it placed 4th). This perhaps indicates that the choice of optimal matching could be related to dataset characteristics, but the diversity of datasets was insufficient to discover the true pattern.

3.3 kNN graphs are not reproducible with binary metrics

Despite promising results for matrix matchings (Fig. 6), none of the continuous metrics found a high-quality, consistent graph matching with any of the binary metrics (Fig. 7). This observation could be explained by multiple factors, including insufficient size of datasets or their relatively low sparsity. However, it is also worth noticing that Spearman's rank correlation is much less punishing for the order mismatch of

Dataset	1	2	3	4	5
Baron	Matthew's Correlation 0.9983	Sokal Sneath III 0.9953	Sokal Sneath V 0.9718	Rogot Goldberg 0.9613	Kulczynski 0.9572
Darmanis	Matthew's Correlation 0.9983	Sokal Sneath III 0.9892	Peirce I 0.9757	Mountford 0.9645	Sokal Sneath V 0.9488
Fletcher	Sokal Sneath III 0.9950	Matthew's Correlation 0.9949	Mountford 0.9865	Sokal Sneath V 0.9572	Peirce I 0.9389
Lawlor	Sokal Sneath III 0.9932	Matthew's Correlation 0.9928	Mountford 0.9901	Sokal Sneath V 0.9811	Baroni Urbani Buser 0.9279
PBMC	Matthew's Correlation 0.9963	Sokal Sneath III 0.9923	Mountford 0.9595	Sokal Sneath V 0.9573	Rogot Goldberg 0.9353
Pollen	Matthew's Correlation 0.9903	Sokal Sneath III 0.9896	Mountford 0.9756	Peirce I 0.9466	Sokal Sneath V 0.9318
Romanov	Matthew's Correlation 0.9975	Sokal Sneath III 0.9969	Mountford 0.9693	Sokal Sneath V 0.9676	Sorgenfrei 0.9340
Tasic	Mountford 0.9768	Sokal Sneath III 0.9734	Matthew's Correlation 0.9710	Jaccard 0.9315	Sokal Sneath V 0.9279

Table 3: Top five optimal matrix matchings for weigthed rank similarity metric across different datasets. Table cells include the name of binary metric with its corresponding $\hat{\rho}$ value.

nearest neighbours than Jaccard score, because it depends on the magnitude of rank difference, while Jaccard penalizes every rank difference in the same manner.

Clearly, graph matching increases for higher matrix matching (Fig. 8). However, there seems to be a non-linear relationship between the two (Fig. 9). This confirms that high matrix matching is a necessary, but insufficient requirement for high graph matching between binary and continuous metrics. We can also observe that graph similarity tends to increase for higher k (Fig. 9). This, perhaps, indicates that differences in neighbour ordering happen more often for the closest (lower k) neighbours.

3.4 Matching accuracy within cells

Matrix matchings with a high cell-wise Spearman mean tend to have lower cell-wise Spearman standard deviation (Fig. 11). This implies that the high quality of matrix matching is reflected among all cells. For the graph matching, there has been insufficient high quality results to show increasing cell-wise accuracy. However, a similar trend was observed with low quality graph matchings reflecting the low quality over all cells (Fig. 11).

3.5 Dataset characteristic impact

In an attempt to predict matrix and graph matching quality on larger datasets, we plotted these quantities against the three data characteristics considered in this work - cell count, gene count and sparsity. However, the results were too varied to draw any specific conclusion. There seems to be no pattern between matching quality and the mentioned characteristics for datasets chosen in this work. It does seem very likely that this is caused by insufficient diversity of datasets, as distance metric performance in cell-clustering was shown to be dependant from the input features [11].

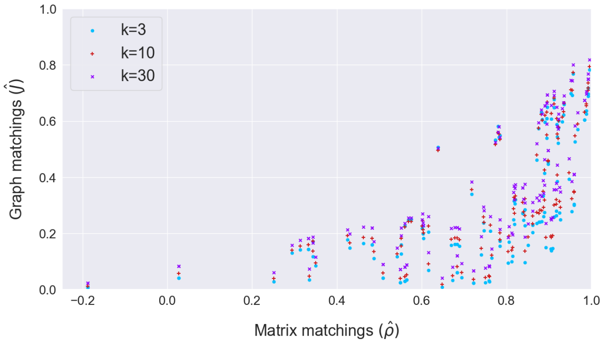


Figure 8: Relation between matrix matching and graph matching in PBMC dataset. Each point represents a result for a single continuous metric-binary metric pair.

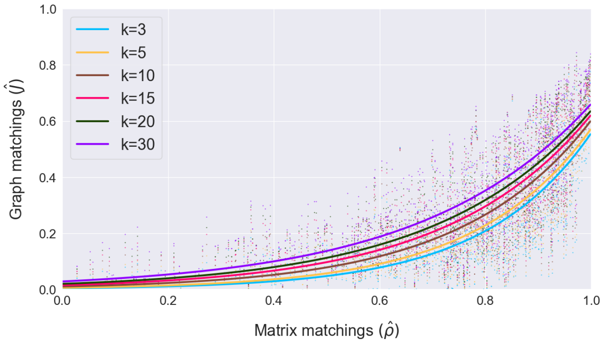


Figure 9: Best exponential curve fit for relation between matrix matching and graph matching across all datasets. Each point represents a result for a single continuous metric-binary metric pair.

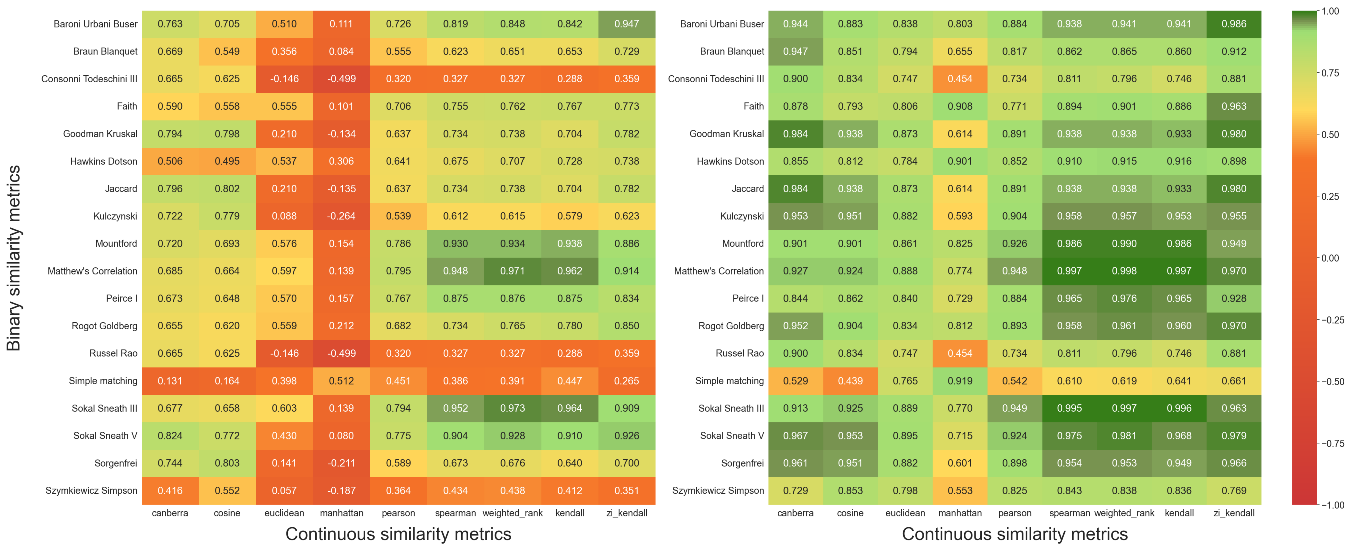


Figure 10: Minimal (left) and maximal (right) matrix matchings of each of the metric pair across all datasets.

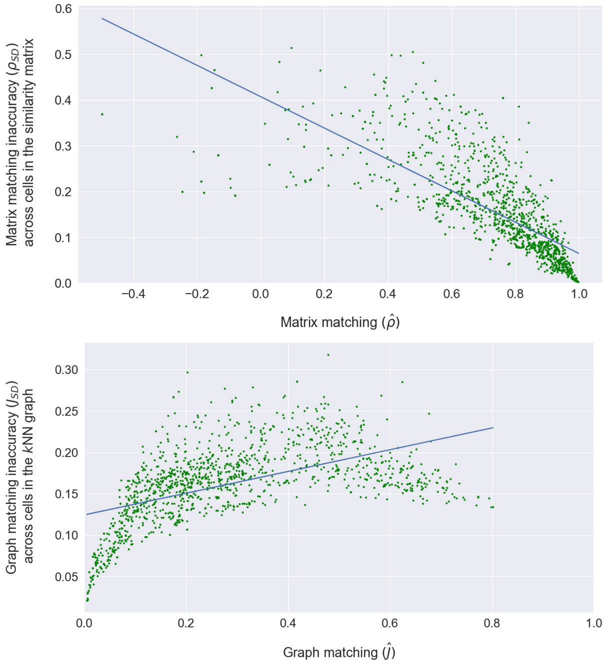


Figure 11: Inaccuracy of matrix matching (top) and graph matching (bottom) vs. their overall quality. Each dot symbolizes a metric pair in particular dataset. Lines of best fit plotted in blue.

3.6 Noteable matchings

Aiming to summarize the results, for each of the continuous metrics we tried to find the most suitable binary metrics that can consistently reproduce their results. We found the minimal and maximal matrix matchings for each of the metric pair across all datasets (Fig. 10) and chose pairs with highest and most consistent scores. Discovered metric pairs are summarized in Table 4. We hope that the following list can serve as a useful reference for future evaluations on larger datasets.

4 Responsible Research

There is an ongoing crisis of result reproducibility in scientific research, with more than 70% of researchers trying and failing to reproduce another researcher's experiments [28]. To overcome this issue, we consulted Stodden et al.'s article [29], which presents six recommendations for reproducible research in computational science. Adhering to these recommendations, we made all of our code and datasets publicly available at Bitbucket.org. Incremental changes in research progress were documented with version-control system (Git). All data is accessible in non-proprietary .csv format. Code repository also contains a manual discussing the computing environment and software version used for implementation.

5 Conclusions and Future Work

In this work, we aimed to answer how close can we get with binary similarity metrics to the output produced by continuous similarity metrics, when applied on scRNAseq data. We discovered that reproducing similarity matrices is possible, but high resemblance of kNN graphs is not achievable.

Each continuous metric have its set of matching binary metrics, which achieve the highest resemblance of results. These matchings seem to stay consistent across different datasets. Furthermore, highest consistent resemblance was achieved for correlation-based metrics and lowest for true distance metrics. Best performing metric matchings also stay more accurate in quality across different cells.

The relationship between matrix similarity and kNN graph similarity is non-linear. In order to achieve high kNN graph resemblance, binary similarity matrix has to be very close to its counterpart computed with continuous metric. This is a necessary, but insufficient requirement. kNN graphs also become more similar for higher k, suggesting that mismatches between binary and continuous metrics happen more often for closest cell neighbours.

Continuous metric	**Binary metric(s)**	$\hat{\rho}_{\min}$	$\hat{\rho}_{\max}$
Spearman's Rank Correlation	Sokal Sneath III	0.952	0.995
	Matthew's Correlation	0.948	0.997
	Mountford	0.930	0.986
Weighted Rank Correlation	Sokal Sneath III	0.973	0.997
	Matthew's Correlation	0.971	0.998
	Mountford	0.934	0.990
Kendall Rank Correlation	Sokal Sneath III	0.964	0.996
	Matthew's Correlation	0.962	0.997
	Mountford	0.938	0.986
Zero-inflated Kendall	Baroni Urbani Buser	0.947	0.986
Pearson's Correlation	Matthew's Correlation	0.795	0.948
	Sokal Sneath III	0.794	0.949
	Mountford	0.786	0.926
Canberra distance	Sokal Sneath V	0.824	0.967
	Jaccard	0.796	0.984
	Goodman Kruskal	0.794	0.984
Cosine similarity	Sorgenfrei	0.803	0.951
	Jaccard	0.802	0.938
	Goodman Kruskal	0.798	0.938
	Kulczynski	0.779	0.951
	Sokal Sneath V	0.772	0.953
Euclidean distance	Sokal Sneath III	0.603	0.889
	Matthew's Correlation	0.597	0.888
Manhattan distance	Simple matching	0.512	0.919

Table 4: Sets of binary metrics with highest similarity of results for each of the continuous metrics. $\hat{\rho}_{\min}$, $\hat{\rho}_{\max}$ are lowest and highest matrix matchings found across the datasets.

The quality of matrix and graph matching seems to be independent from dataset characteristics. However, we believe that these results could change for larger and sparser datasets. The size especially could have a major influence on graph similarity, due to the change of proportion between k and all possible neighbours. Unfortunately, conducting evaluation for larger datasets is limited for technical reasons. As the size for similarity matrix grows quadratically with the number of cells, evaluating bigger datasets requires large amount of computer memory. Regular machines have inadequate computational capabilities for such evaluation.

As such, we recommend future researchers of this topic to equip themselves with machines that are capable of handling large amounts of data. For the most recent scRNAseq datasets ($>$ 100.000 cells), a supercomputer might be necessary to obtain results in feasible time. We also believe that the implementation of evaluation framework could be improved. Our experience shows that most of the memory errors were caused during the computation of similarity matrix for a continuous similarity metric using R. This also proved to be the most time consuming part of the empirical studies. Perhaps, some of the memory problems could be mitigated by developing this part of implementation in a different programming language, such as C or C++.

6 Acknowledgements

I would like to dedicate this last section to thank my supervisor, Gerard A. Bouland and the responsible professor, Marcel J.T. Reinders, for the invaluable help given throughout the research process, feedback provided during weekly progress sessions and the inspiration to further delve into the field of bioinformatics in the future years of my education.

A Binary metrics

GROUP	METRIC	GROUP	METRIC
1	**Simple Matching (SM)**	5	**Jaccard-Tanimoto (JT)**
	Rogers-Tanimoto (RT)		Jaccard (Ja)
	Sokal-Sneath II (SS2)		Gleason (Gle)
	Austin-Colwell (AC)		Sokal-Sneath (SS1)
	Consonni-Todeschini I (CT1)	6	**Sokal-Sneath III (SS3)**
	Consonni-Todeschini II (CT2)		Consonni-Todeschini V (CT5)
	Driver-Kroeber (DK)	7	**Sokal-Sneath IV (SS4)**
	Forbes (For)		Harris-Lahey (HL)
	Fossum in Holiday (Fos)	8	**Russel-Rao (RR)**
	Consonni-Todeschini IV (CT4)	9	**Kulczynski (Kul)**
2	**Peirce I (Pe1)**	10	**Simpson (Sim)**
	Maxwell-Pilliner (MP)	11	**Braun-Blanquet (BB)**
	Michael (Mic)	12	**Baroni-Urbani-Buser I (BUB)**
	Dennis in Holiday (Den)	13	**Faith (Fai)**
	dispersion in Choi et al. (dis)	14	**Mountford (Mou)**
	Cohen (Coh)	15	**Rogot-Goldberg (RG)**
	Peirce II (Pe2)	16	**Goodman-Kruskal (GK)**
3	**Sorgenfrei (Sor)**	17	**Hawkins-Dotson (HD)**
	Dice I (Di1)	18	**Consonni-Todeschini III (CT3)**
4	**Pearson-Heron (Phi)**	19	Cole I (Co1)
	Yule I (Yu1)	20	Cole II (Co2)
	Yule II (Yu2)	21	Dice II (Di2)

Table 5: Arrangement of binary metrics based on their mutual Spearman's Rank Correlation (ρ) as in [14]. Metrics in a single group with mutual $\rho > 0.97$. Metrics included in this work shown in bold. Metric names presented with abbreviation in brackets consistent with [14].

No.	Name	Definition
1	Simple matching	$S_{\text{SM}} = \frac{a+d}{p}$
2	Jaccard	$S_{\text{Jac}} = \frac{a}{a+b+c}$
3	Matthew's Correlation	$S_{\text{MCC}} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$
4	Sorgenfrei	$S_{\text{Sor}} = \frac{a^2}{(a+b)(a+c)}$
5	Peirce I	$S_{\text{Pe1}} = \frac{ad-bc}{(a+b)(c+d)}$
6	Sokal Sneath III	$S_{\text{SS3}} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d})$
7	Sokal Sneath V	$S_{\text{SS5}} = \frac{ad}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$
8	Russel-Rao	$S_{\text{RR}} = \frac{a}{p}$
9	Kulczynski	$S_{\text{Kul}} = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$
10	Szymkiewicz-Simpson	$S_{\text{Sim}} = \frac{a}{\min(a+b,a+c)}$
11	Braun-Blanquet	$S_{\text{BB}} = \frac{a}{\max(a+b,a+c)}$
12	Baroni-Urbani-Buser	$S_{\text{BUB}} = \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$
13	Faith	$S_{\text{Fai}} = \frac{a+0.5d}{p}$
14	Mountford	$S_{\text{Mou}} = \frac{2a}{ab+ac+2bc}$
15	Rogot-Goldberg	$S_{\text{RG}} = \frac{a}{2a+b+c} + \frac{d}{2d+b+c}$
16	Hawkins-Dotson	$S_{\text{HD}} = \frac{1}{2}(\frac{a}{a+b+c} + \frac{d}{d+b+c})$
17	Goodman-Kruskal	$S_{\text{GK}} = \frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$
18	Consonni-Todeschini III	$S_{\text{CT3}} = \frac{\ln(1+a)}{\ln(1+p)}$

Table 6: Full binary metric list used in this work with their corresponding definitions. a, b, c, d, p are defined as in Table 1. Some amendmentments have been made with regards to Todeschini et al. [14]: 1) "Jaccard-Tanimoto" was renamed to "Jaccard" 2) "Phi-Heron" was renamed to "Matthew's Correlation" 3) "Sokal Sneath IV" was renamed to "Sokal Sneath V".

References

[1] Y. Zhang, D. Wang, M. Peng, L. Tang, J. Ouyang, F. Xiong, C. Guo, Y. Tang, Y. Zhou, Q. Liao, X. Wu, H. Wang, J. Yu, Y. Li, X. Li, G. Li, Z. Zeng, Y. Tan, and W. Xiong, "Single-cell rna sequencing in cancer research," 12 2021.

[2] G. Sun, Z. Li, D. Rong, H. Zhang, X. Shi, W. Yang, W. Zheng, G. Sun, F. Wu, H. Cao, W. Tang, and Y. Sun, "Single-cell rna sequencing in cancer: Applications, advances, and emerging challenges," 6 2021.

[3] A. Maynard, C. E. McCoach, J. K. Rotow, L. Harris, F. Haderk, D. L. Kerr, E. A. Yu, E. L. Schenk, W. Tan, A. Zee, M. Tan, P. Gui, T. Lea, W. Wu, A. Urisman, K. Jones, R. Sit, P. K. Kolli, E. Seeley, Y. Gesthalter, D. D. Le, K. A. Yamauchi, D. M. Naeger, S. Bandyopadhyay, K. Shah, L. Cech, N. J. Thomas, A. Gupta, M. Gonzalez, H. Do, L. Tan, B. Bacaltos, R. Gomez-Sjoberg, M. Gubens, T. Jahan, J. R. Kratz, D. Jablons, N. Neff, R. C. Doebele, J. Weissman, C. M. Blakely, S. Darmanis, and T. G. Bivona, "Therapy-induced evolution of human lung cancer revealed by single-cell rna sequencing," *Cell*, vol. 182, pp. 1232–1251.e22, 9 2020.

[4] Z. Xue, K. Huang, C. Cai, L. Cai, C. Y. Jiang, Y. Feng, Z. Liu, Q. Zeng, L. Cheng, Y. E. Sun, J. Y. Liu, S. Horvath, and G. Fan, "Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing," *Nature*, vol. 500, pp. 593–597, 2013.

[5] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang, "Single-cell rna-seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural and Molecular Biology*, vol. 20, pp. 1131–1139, 9 2013.

[6] L. Sikkema, D. Strobl, L. Zappia, E. Madissoon, N. Markov, L. Zaragosi, M. Ansari, M. Arguel, L. Apperloo, C. Bécavin, M. Berg, E. Chichelnitskiy, M. Chung, A. Collin, A. Gay, B. H. Kashani, M. Jain, T. Kapellos, T. Kole, C. Mayr, M. von Papen, L. Peter, C. Ramírez-Suástegui, J. Schniering, C. Taylor, T. Walzthoeni, C. Xu, L. Bui, C. de Donno, L. Dony, M. Guo, A. Gutierrez, L. Heumos, N. Huang, I. Ibarra, N. Jackson, P. K. L. Murthy, M. Lotfollahi, T. Tabib, C. Talavera-Lopez, K. Travaglini, A. Wilbrey-Clark, K. Worlock, M. Yoshida, L. B. N. Consortium, T. Desai, O. Eickelberg, C. Falk, N. Kaminski, M. Krasnow, R. Lafyatis, M. Nikolíc, J. Powell, J. Rajagopal, O. Rozenblatt-Rosen, M. Seibold, D. Sheppard, D. Shepherd, S. Teichmann, A. Tsankov, J. Whitsett, Y. Xu, N. Banovich, P. Barbry, T. Duong, K. Meyer, J. Kropski, D. Pe, H. Schiller, P. Tata, J. Schultze, A. Misharin, M. Nawijn, and F. Theis, "An integrated cell atlas of the human lung in health and disease," 2022.

[7] Y. Bai, H. Liu, H. Lyu, L. Su, J. Xiong, and Z. M. M. Cheng, "Development of a single-cell atlas for woodland strawberry (fragaria vesca) leaves during early botrytis cinerea infection using single-cell rna-seq," *Horticulture Research*, vol. 9, 2022.

[8] L. W. Plasschaert, R. Žilionis, R. Choo-Wing, V. Savova, J. Knehr, G. Roma, A. M. Klein, and A. B. Jaffe, "A single-cell atlas of the airway epithelium reveals the cftr-rich pulmonary ionocyte," *Nature*, vol. 560, pp. 377–381, 8 2018.

[9] G. A. Bouland, A. Mahfouz, and M. J. T. Reinders, "The rise of sparser single-cell rnaseq datasets; consequences and opportunities," 2023. An overview of why binarizing gene expression is possible and what opportunities it presents.

[10] P. Qiu, "Embracing the dropouts in single-cell rna-seq analysis," *Nature Communications*, vol. 11, 12 2020.

[11] E. R. Watson, A. Mora, A. T. Fard, and J. C. Mar, "How does the structure of data impact cell-cell similarity? evaluating how structural properties influence the performance of proximity metrics in single cell rna-seq data," *Briefings in bioinformatics*, vol. 23, 11 2022.

[12] M. A. Skinnider, J. W. Squair, and L. J. Foster, "Evaluating measures of association for single-cell transcriptomics," *Nature Methods*, vol. 16, pp. 381–386, 5 2019.

[13] T. Kim, I. R. Chen, Y. Lin, A. Y. Y. Wang, J. Y. H. Yang, and P. Yang, "Impact of similarity metrics on single-cell rna-seq data clustering," *Briefings in Bioinformatics*, vol. 20, pp. 2316–2326, 11 2019.

[14] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett, "Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets," *Journal of Chemical Information and Modeling*, vol. 52, pp. 2884–2901, 11 2012.

[15] M. Baron, A. Veres, S. Wolock, A. Faust, R. Gaujoux, A. Vetere, J. Ryu, B. Wagner, S. Shen-Orr, A. Klein, and et al., "A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure," *Cell Systems*, vol. 3, no. 4, 2016.

[16] S. Darmanis, S. A. Sloan, Y. Zhang, M. Enge, C. Caneda, L. M. Shuer, M. G. Hayden Gephart, B. A. Barres, and S. R. Quake, "A survey of human brain transcriptome diversity at the single cell level," *Proceedings of the National Academy of Sciences*, vol. 112, no. 23, p. 7285–7290, 2015.

[17] R. B. Fletcher, D. Das, L. Gadye, K. N. Street, A. Baudhuin, A. Wagner, M. B. Cole, Q. Flores, Y. G. Choi, N. Yosef, and et al., "Deconstructing olfactory stem cell trajectories at single-cell resolution," *Cell Stem Cell*, vol. 20, no. 6, 2017.

[18] 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*, vol. 27, no. 2, p. 208–222, 2016.

[19] T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. Reinders, and A. Mahfouz, "A comparison of automatic cell identification methods for single-cell

rna sequencing data," *Genome Biology*, vol. 20, no. 1, 2019.

[20] A. Pollen, T. Nowakowski, J. Chen, H. Retallack, C. Sandoval-Espinosa, C. Nicholas, J. Shuga, S. Liu, M. Oldham, A. Diaz, and et al., "Molecular identity of human outer radial glia during cortical development," *Cell*, vol. 163, no. 1, p. 55–67, 2015.

[21] R. A. Romanov, A. Zeisel, J. Bakker, F. Girach, A. Hellysaz, R. Tomer, A. Alpár, J. Mulder, F. Clotman, E. Keimpema, and et al., "Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes," *Nature Neuroscience*, vol. 20, no. 2, p. 176–188, 2016.

[22] B. Tasic, V. Menon, T. N. Nguyen, T. K. Kim, T. Jarsky, Z. Yao, B. Levi, L. T. Gray, S. A. Sorensen, T. Dolbeare, and et al., "Adult mouse cortical cell taxonomy revealed by single cell transcriptomics," *Nature Neuroscience*, vol. 19, no. 2, p. 335–346, 2016.

[23] D. Risso and M. Cole, *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets*, 2023. R package version 2.14.0.

[24] B. Schäling, *The Boost C++ Libraries*. Boris Schäling.

[25] G. Csárdi, T. Nepusz, S. Horvát, V. Traag, F. Zanini, and D. Noom, "igraph," Jan 2023.

[26] *dismay: dismay: distance metrics for matrices*, 2023. R package version 0.0.1.

[27] D. Eddelbuettel, *Seamless R and C++ Integration with Rcpp*. New York: Springer, 2013. ISBN 978-1-4614-6867-7.

[28] A. A. Aarts, J. E. Anderson, C. J. Anderson, P. R. Attridge, A. Attwood, J. Axt, M. Babel, Štěpán Bahník, E. Baranski, M. Barnett-Cowan, E. Bartmess, J. Beer, R. Bell, H. Bentley, L. Beyan, G. Binion, D. Borsboom, A. Bosch, F. A. Bosco, S. D. Bowman, M. J. Brandt, E. Braswell, H. Brohmer, B. T. Brown, K. Brown, J. Brüning, A. Calhoun-Sauls, S. P. Callahan, E. Chagnon, J. Chandler, C. R. Chartier, F. Cheung, C. D. Christopherson, L. Cillessen, R. Clay, H. Cleary, M. D. Cloud, M. Conn, J. Cohoon, S. Columbus, A. Cordes, G. Costantini, L. D. Alvarez, E. Cremata, J. Crusius, J. DeCoster, M. A. DeGaetano, N. D. Penna, B. D. Bezemer, M. K. Deserno, O. Devitt, L. Dewitte, D. G. Dobolyi, G. T. Dodson, M. B. Donnellan, R. Donohue, R. A. Dore, A. Dorrough, A. Dreber, M. Dugas, E. W. Dunn, K. Easey, S. Eboigbe, C. Eggleston, J. Embley, S. Epskamp, T. M. Errington, V. Estel, F. J. Farach, J. Feather, A. Fedor, B. Fernández-Castilla, S. Fiedler, J. G. Field, S. A. Fitneva, T. Flagan, A. L. Forest, E. Forsell, J. D. Foster, M. C. Frank, R. S. Frazier, H. Fuchs, P. Gable, J. Galak, E. M. Galliani, A. Gampa, S. Garcia, D. Gazarian, E. Gilbert, R. Giner-Sorolla, A. Glöckner, L. Goellner, J. X. Goh, R. Goldberg, P. T. Goodbourn, S. Gordon-McKeon, B. Gorges, J. Gorges, J. Goss, J. Graham, J. A. Grange, J. Gray, C. Hartgerink, J. Hartshorne, F. Hasselman, T. Hayes, E. Heikensten, F. Henninger, J. Hodsoll, T. Holubar, G. Hoogendoorn, D. J. Humphries, C. O. Hung, N. Immelman, V. C. Irsik, G. Jahn, F. Jäkel, M. Jekel, M. Johannesson, L. G. Johnson, D. J. Johnson, K. M. Johnson, W. J. Johnston, K. Jonas, J. A. Joy-Gaba, H. B. Kappes, K. Kelso, M. C. Kidwell, S. K. Kim, M. Kirkhart, B. Kleinberg, G. Knežević, F. M. Kolorz, J. J. Kossakowski, R. W. Krause, J. Krijnen, T. Kuhlmann, Y. K. Kunkels, M. M. Kyc, C. K. Lai, A. Laique, D. Lakens, K. A. Lane, B. Lassetter, L. B. Lazarević, E. P. L. Bel, K. J. Lee, M. Lee, K. Lemm, C. A. Levitan, M. Lewis, L. Lin, S. Lin, M. Lippold, D. Loureiro, I. Luteijn, S. MacKinnon, H. N. Mainard, D. C. Marigold, D. P. Martin, T. Martinez, E. J. Masicampo, J. Matacotta, M. Mathur, M. May, N. Mechin, P. Mehta, J. Meixner, A. Melinger, J. K. Miller, M. Miller, K. Moore, M. Möschl, M. Motyl, S. M. Müller, M. Munafo, K. I. Neijenhuijs, T. Nervi, G. Nicolas, G. Nilsonne, B. A. Nosek, M. B. Nuijten, C. Olsson, C. Osborne, L. Ostkamp, M. Pavel, I. S. Penton-Voak, O. Perna, C. Pernet, M. Perugini, R. N. Pipitone, M. Pitts, F. Plessow, J. M. Prenoveau, R. M. Rahal, K. A. Ratliff, D. Reinhard, F. Renkewitz, A. A. Ricker, A. Rigney, A. M. Rivers, M. Roebke, A. M. Rutchick, R. S. Ryan, O. Sahin, A. Saide, G. M. Sandstrom, D. Santos, R. Saxe, R. Schlegelmilch, K. Schmidt, S. Scholz, L. Seibel, D. F. Selterman, S. Shaki, W. B. Simpson, H. C. Sinclair, J. L. Skorinko, A. Slowik, J. S. Snyder, C. Soderberg, C. Sonnleitner, N. Spencer, J. R. Spies, S. Steegen, S. Stieger, N. Strohminger, G. B. Sullivan, T. Talhelm, M. Tapia, A. T. Dorsthorst, M. Thomae, S. L. Thomas, P. Tio, F. Traets, S. Tsang, F. Tuerlinckx, P. Turchan, M. Valášek, A. E. V. Veer, R. V. Aert, M. V. Assen, R. V. Bork, M. V. D. Ven, D. V. D. Bergh, M. V. D. Hulst, R. V. Dooren, J. V. Doorn, D. R. V. Renswoude, H. V. Rijn, W. Vanpaemel, A. V. Echeverría, M. Vazquez, N. Velez, M. Vermue, M. Verschoor, M. Vianello, M. Voracek, G. Vuu, E. J. Wagenmakers, J. Weerdmeester, A. Welsh, E. C. Westgate, J. Wissink, M. Wood, A. Woods, E. Wright, S. Wu, M. Zeelenberg, and K. Zuni, "Estimating the reproducibility of psychological science," *Science*, vol. 349, 8 2015.

[29] V. Stodden, Y. Law, and R. Subramanian, "Reproducible research. addressing the need for data and code sharing in computational science.," 2010.