# Systematic benchmark evaluation of distance metrics for scRNA-seq data

by

Nathan Dyjack

A thesis submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Master of Science

Baltimore, Maryland
August, 2021

# Abstract

When attempting to generate statistical inference, the notion of distance or (dis)similarity among observations is a crucial for understanding the data's structure. When the data are sparse, as in single-cell RNA-sequencing (scRNA-seq), some notions of distance can give false signals regarding observation structure. Motivated by a multinomial model for scRNA-seq data, we test sought to test the performance of several dissimilarities using experimental and simulated scRNA-seq data. Methods and results for the permutations of these analyses are provided and summarized herein.

We leveraged the `minicore` package as an efficient and accurate means to compute fifteen notions of dissimilarity for experimental and simulated scRNA-seq data. Calculations were performed in experimental scRNA-seq data that had cluster and lineage structure using multiple levels of variable genes for robustness. The simulated scRNA-seq data sought to test robustness in response to experimental factors, so simulated cluster and lineage structure data was tested with multiple varying simulation settings. We provide five fitness metrics for each dissimilarity, kAcc (nearest-neighbor accurarcy), TrajCor (lineage structure accuracy), ARI (truth label concordance with simple clustering algorithm), 1-G+ (tightness of truth cluster labels), and GapStat(evidence for $k > 1$ clusters). While no single distance vastly outperforms all others, geometric (non-normalized) distances are consistently out-performed by statistical (normalized). We reiterate the suggestions of `minicore` and recommend JSD as a distance, which demonstrates strong overall performance in almost all test scenarios.

# Thesis Committee

Stephanie Hicks (Primary Advisor)
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Ben Langmead
Associate Professor
Department of Computer Science
Johns Hopkins University, Whiting School of Engineering

# Acknowledgements

I would first like to greatly thank Dr. Hicks for her excellent mentorship and insight, this journey would not have been possible without you. I would also like to thank Dr. Langmead for co-advising with this project, and Daniel Baker for writing `minicore`. I'd like to thank my family and especially my parents for supporting me throughout this endeavor. I feel the need to acknowledge the unique and welcoming faculty and environment at JHSPH Biostatistics, this environment has been significant in many ways for my continued growth. Last but not least, I'd like to thank my friends and cohort from the Biostatistics department who helped me make it through tough exams and homework.

# Table of Contents

# List of Tables

# List of Figures

# 1 Introduction

Ribonucleic Acid sequencing (RNA-seq) is a set of protocols that allows researchers to quantify genomic species at the transcript level. Notably, RNA-seq provides more accurate quantification than previous technologies (Microarray, Sanger sequencing) due to the use of reverse transcriptase to provide integer counts of RNA species [1]. RNA-seq is most commonly used to measure gene transcripts (which will eventually be translated into proteins), however, the technology is fully capable of assessing less common RNA such as micro RNA and long-noncoding RNA [2].

Advancements in RNA-seq technology have permitted RNA quantification of genes at the level of a single cell (scRNA-seq). First published by Tang et al. [3], notable advancements on the protocol include SMART-seq, CEL-seq, and droplet-based methods [4, 5]. A crucial factor which distinguishes previous generations of standard ('bulk') RNA-seq and scRNA-seq is the amount of required input material. For example, bulk RNA-seq protocols require nanogram-scale (1-500ng) amounts of RNA, typically acquired from lysis and integration of a tissue sample to form one RNA-seq observation [6]. In comparison, a similar tissue sample can generate thousands of scRNA-seq measurements by virtue of the many cells which compose it. Consequentially, scRNA-seq necessitates the input of substantially less RNA (picogram-scale, often as low as 0.1pg) than bulk RNA-seq [7]. Due to the small scale of cellular RNA quantities or technical limitations (i.e., 'dropout' events), count matrices generated with scRNA-seq protocols demonstrate severe sparsity in comparison to those created by bulk RNA-seq. Often, as many as 90% of entries in a scRNA-seq dataset will be zero-valued [8].

Bulk RNA-seq has leveraged tremendous power for studying genomic processes at the population level, i.e., cohort-level studies of development and disease. While scRNA-seq can also be used in this way, the technology differs itself from others by

means of turning a single tissue sample into thousands of single-cell observations. This level of measurement allows researchers to investigate differences in cellular populations across disease conditions at a fine level. For example, recent work has unraveled the role of cigarette smoking in the dysregulation of basal cell differentiation, leading to differential porportions and functions of airway epithelial cells such as ciliated and mucus-producing cells [9]. Intrinsic to this form of research is the capacity to quickly and accurately identify populations and their relationship to each other within scRNA-seq data. Common forms of this analysis include unsupervised clustering where the researcher seeks to group or label the observations (cells) into a discrete clusters, often, cell types. As an extension of this analysis, researches often seek to investigate the differences amongst these cell types. Often, classifying the cells into $k$ discrete clusters is an oversimplification, for example, in the case where the sampled tissue is comprised of multiple cell types that all stem from a single progenitor cell. In this case, the path from the progenitor states to the end states may follow a continuous trajectory with varying levels of cells in transitory states between end and finish. To this end, researchers also seek to identify this lineage-type structure within the data, and place the cells along a one-dimensional lineage or ordering, i.e., 'pseudotime' analysis [10, 11].

Intrinsic to all of these analytical methods is the capacity of the researcher to accurately quantify (dis)similarity amongst all observations in the dataset. RNA-seq measurements are often modeled using vectors. In this context, a vector can be described as multivariate ordering of observations where each dimension corresponds to the abundance (counts) of a given transcript (gene) within that sample. Historically, the (dis)similarity between the gene expression of two cells has been modeled using the normal distribution. This stems from near-Gaussian distribution of gene expression intensities measured with microarray technology. The most common choice of dissimilarity measure has been the squared Euclidean (L2) distance using observed

counts that have scaled by the total number of reads (i.e., more library size / sequencing depths) [12, 13]. These normalized data are then log2-transformed for variance stabilization and to make the data more Gaussian (see Section 2.1). As noted by Witten (2011) [14], another popular choice is correlation-based distance, which is equivalent to squared Euclidean distance up to a scaling of the observations [15]. In fact, the squared Euclidean distance can be derived from a hypothesis test using a simple Gaussian model for the data [14].

The analytic tools (e.g. dissimiarity measures) that assume a Gaussian distribution and use normalized and log2-transformed scRNA-seq data can be used for many downstream analyses, including classification methods based on linear discriminant analysis, clustering methods that use Euclidean distance such as $k$-means, or methods that project the high-dimensional data to a low-dimensional space using, for example, multidimensional scaling (MDS) [16]. Another example is Principal Components Analysis (PCA), a popular dimension reduction method that is implicitly based on Euclidean distance, which corresponds to maximizing a Gaussian likelihood [17]. The assumption of Gaussian data in analytic tools for the analysis of scRNA-seq data are ubiquitous and widely adopted into scRNA-seq workflows [18, 19]. However, several recent studies have shown that either explicitly or implicitly use Euclidean distance on normalized log2-transformed scRNA-seq data can induce unwanted technical artifacts and may confound the results [20, 21, 22, 14, 17].

In contrast to modeling the data with a Gaussian distribution, RNA-seq (and scRNA-seq) are nonnegative counts and can be modeled using discrete count distributions, such as the Poisson, negative binomial or multinomial distributions. Accordingly, recent work has been made to improve the performance via optimizing the dissimilarity measure itself that is designed for the count-based nature of sequencing data. Notably, Witten (2011) [14] derived a dissimilarity metric for bulk RNA-seq data, based on an empirically justified assumption a Poisson distribution [23, 24].

However, recent papers have argued that technical noise or variation from scRNA-seq data actually follow a multinomial [17] or negative binomial (NB) [25] (compared to Poisson) distribution. This idea is behind count-based dimensionality reduction methods such as generalized principal components analysis (glmpca) [17] compared to Gaussian-based dimensionality reduction methods, such as principal components analysis. Therefore, there is a need to have a distance metric based other nonnegative count distributions on that can be used for other downstream applications including dimensionality reduction, clustering, and corsets for scRNA-seq data [26].

Relevant work in this area is from Berninger et al. (2008) who proposed a dissimilarity metric for sequencing data based on the multinomial distribution [27]. Witten (2011) contrasts the Poisson dissimilarity metric [14] with the multinomial dissimilarity metric [27]:

> "Berninger et al. (2008) propose a method for computing a dissimilarity matrix using sequencing data that is also very closely related to ours. They assume that each observation is drawn from a multinomial distribution, and they test whether or not the multinomial parameters for each pair of observations are equal. This is almost identical to our Poisson model and associated hypothesis testing framework, since if the observations are distributed according to [equation] (14), then their distribution conditional on $X_i$, $X_{i'}$ is multinomial. In fact, the log-likelihood ratio statistics under our model and theirs are identical for certain very natural estimates of $N_{ij}$, $N_{i'j}$, $d_{ij}$, and $d_{i'j}$ in [equation] (17) (see the Appendix). However, there are some important differences between the two proposals. Berninger et al. (2008) place a Dirichlet prior on the parameters for the multinomial distribution, and then use a Bayes factor as a measure of the dissimilarity between two observations. Consequently, two identical observations can have nonzero dissimilarity according to Berninger et al.

(2008), and two different observations can have smaller dissimilarity than two identical observations. This leads to problems in the interpretation of their dissimilarity measure as well as in the performance of any clustering approach that is based upon it. Finally, their approach can suffer from numerical issues where the computed dissimilarity between a pair of observations rounds to zero."

In this work and in contrast to the Bayes factor approach from Berninger et al., we extend the work from Witten (2011) [14] to a novel dissimilarity metric for uniquely specified for scRNA-seq data based on the multinomial distribution. Furthermore, we test the performance of many other statistic dissimilarity metrics, which operate under the assumption of a multinomial distribution.

# 2 Methods

All distances matrices were calculated with default parameters or with a prior of 1 (see 2.1.3) and implemented in the `minicore` Python library [28] and saved as `pickles` [29]. These which were then imported into R [30] using `reticulate` [31, 32]. The rest of the Methods Section describes an overview of known and novel dissimilarity measures (Section 2.1), where the mathematical derivations for the novel dissimilarity measure based on the multinomial distribution introduced in this thesis is discussed in Section 2.1.2. Next, we describe the data used to evaluate the dissimilarity measures (Section 2.2), all scRNA-seq preprocessing performed (Section 2.3), a description of how the null scRNA-seq data were simulated (Section 2.4), and a description of performance metrics used to evaluate the dissimilarity measures (Section 2.5).

## 2.1 Overview of dissimilarity measures

Assume we have two observations (or cells) $x_j$ and $y_j$ each observations from a set of $m$ features (genes), namely $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_m)$. We denote $\hat{x}_j$ as the scaled (normalized by the total counts for that cell) version of this entry, that is $\hat{x}_j = \frac{x_j}{\sum_{j=1}^{m} x_j}$ such that $\sum_{j=1}^{m} \hat{x}_j = 1$. We define a dissimilarity $d_j(x_j, y_j)$ between $x_j$ and $y_j$ for the $j^{th}$ feature, and $D(\mathbf{x}, \mathbf{y})$ as a generalized dissimilarity between observations (cells) $\mathbf{x}$ and $\mathbf{y}$. Typically, $D(\mathbf{x}, \mathbf{y})$ will take the form a transformed sum over individual $d_j(x_j, y_j)$

$$D(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{m} d_j(x_j, y_j)$$

In the next two sections, we first describe a set of known dissimilarity measures that we used (Section 2.1.1) and a novel dissimilarity measure based on the multinomial distribution (Section 2.1.2).

### 2.1.1 Known dissimilarity measures

The absolute error loss (or $L^1$ norm – referred to here as L1) is:

$$\text{L1} = ||\mathbf{x} - \mathbf{y}||_1 = \sum_j |x_j - y_j| \tag{1}$$

The Euclidean distance (or Euclidean norm or $L^2$ norm – referred to here as L2) is:

$$\text{L2} = ||\mathbf{x} - \mathbf{y}||_2 = \sqrt{\sum_j (x_j - y_j)^2} \tag{2}$$

The squared Euclidean distance (referred to here as SqL2) is:

$$\text{SQL2} = (\text{L2})^2 = (||\mathbf{x} - \mathbf{y}||_2)^2 = \sum_j (x_j - y_j)^2 \tag{3}$$

The Multinomial Kullback-Leibler Divergence (referred to here as MKL) is:

$$\mathrm{MKL}(\mathbf{x}, \mathbf{y}) = \sum_j \hat{x}_j \times \log \frac{\hat{x}_j}{\hat{y}_j} \tag{4}$$

The (Total Variation Distance):

$$\mathrm{TVD} = \frac{1}{2} \times \sum_j (|\hat{x}_j - \hat{y}_j|) \tag{5}$$

The Jensen-Shannon Divergence (referred to here as JSD) is:

$$\mathrm{JSD} = \frac{1}{2} \times (MKL(\mathbf{x}, \mathbf{y}) + MKL(\mathbf{y}, \mathbf{x})) \tag{6}$$

The Jensen-Shannon Divergence can be converted to a metric, namely the Jensen-Shannon Metric (JSM), using a square-root transform:

$$\mathrm{JSM} = \sqrt{JSD} \tag{7}$$

The Hellinger distance (HEL) is given by:

$$\mathrm{HEL} = \sqrt{\sum_j \left(\sqrt{\hat{x}_j} - \sqrt{\hat{y}_j}\right)^2} \times \frac{1}{2} \tag{8}$$

The Bhatacharrya Distance (BCD) is:

$$\mathrm{BCD} = -\log \sqrt{\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}} \tag{9}$$

And the metric version of this distance, the Bhatacharrya Distance Metric (BCM) is:

$$\mathrm{BCM} = \sqrt{1 - \mathrm{BCD}} \tag{10}$$

The Itakura-Saito Distance (ISD) is:

$$\text{ISD(X, Y)} = \sum_j \frac{\hat{x}_j}{\hat{y}_y} - \log \frac{\hat{y}_j}{\hat{y}_j} - 1 \tag{11}$$

The Reverse Itakura-Saito Distance (RISD) simply swaps the role of $X$ and $Y$. The Symmetric Itakura-Saito Distance (SIS) is a weighted average of the ISD and RISD:

$$\text{SIS} = \frac{1}{2} \times \Big( ISD(\mathbf{x}, \mathbf{y}) + RISD(\mathbf{x}, \mathbf{y}) \Big) \tag{12}$$

### 2.1.2 A novel dissimilarity measure based on the multinomial distribution

In this section, we extend the work of Witten (2011) [14] who derived a distance metric based on the Poisson distribution for count-based sequencing data. The main idea behind this distance metric is it is defined as a Likelihood Ratio Test (LRT) comparing the null and alternative hypotheses that the parameters for the distributions associated with cell $\mathbf{x} = (x_1, \ldots, x_m)$ and cell $\mathbf{y} = (y_1, \ldots, y_m)$ are not different (null hypothesis) or there is a difference (alternative). Witten (2011) [14] used the Poisson distribution, but here we use the multinomial distribution, which has been empirically shown to model the mean-sparsity relationship in scRNA-seq data better than other count-based distributions (Poisson and Negative Binomial) [17].

Consider a single observation $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, here represented as a vector of counts where each dimension is an gene. Assuming that this cell represents a draw from a multinomial distribution, we describe the probability of observing the counts for each gene using a vector $p_x$ (e.g., $p_{x,j}$ represents the probability of observing counts for gene $j$ of cell $\mathbf{x}$, $x_j$). Assume the random variable $X$ follows a multinomial distribution. Then, the probability mass function (PMF) for $X$ conditional on its multinomial parameterization is

$$P(X|p_x) = \frac{(\sum_j x_j)!}{\prod_j x_j} \prod_j p_{x,j} \tag{13}$$

The maximum likelihood estimate (MLE) for the true relative abundance $p_x$ will be given by $\hat{p}_{x,j} = \frac{x_j}{\sum_j x_j}$. If we then consider a second cell as an independent multinomial observation $Y$, their joint PMF is given by

$$P(X, Y|p_x, p_y) = P(X|p_x)P(Y|p_y) \propto \prod_j p_{x,j} p_{y,j} \tag{14}$$

Extending the work of Witten (2011) [14], we propose a hypothesis test which acts as a measure of dissimilarity. We consider the null hypothesis $H_0$ in which cells $X$ and $Y$ have the same multinomial parameterization $\bar{p}_{xy}$. For each $\bar{p}_{j,xy}$, we simply take the mean of the MLE for gene $j$ in cells $X$ and $Y$, $\bar{p}_{j,xy} = \frac{\hat{p}_{x,j}+\hat{p}_{y,j}}{2}$. Under $H_0$, the likelihood for our observed data is then

$$P(X, Y|H_0) = P(X, Y|\bar{p}_{j,xy}) \propto \prod_j \bar{p}_{j,xy}^{x_j} \bar{p}_{j,xy}^{y_j} \tag{15}$$

We then consider an alternate hypothesis $H_1$ in which $X$ and $Y$ are parameterized by their respective MLEs, $\hat{p}_x$ and $\hat{p}_y$.

$$P(X, Y|H_1) = P(X, Y|\hat{p}_x, \hat{p}_y) \propto \prod_j \hat{p}_{j,x}^{x_j} \hat{p}_{j,y}^{y_j} \tag{16}$$

Similar to the work of Witten (2011) [14], our dissimilarity is then defined as a LRT for $H_0$ and $H_1$, explicitly

$$\Lambda = \frac{P(X, Y|H_0)}{P(X, Y|H_1)} = \prod_j \frac{\hat{p}_{j,x}^{x_j} \hat{p}_{j,y}^{y_j}}{\bar{p}_{j,xy}^{x_j} \bar{p}_{j,xy}^{y_j}} \tag{17}$$

The intuition of this test is that if the data are more likely under $H_0$, (the cells share a common parameterization), the denominator increases, and the value for $\Lambda$ approaches 0. In practice, the sparsity of scRNA-seq will result in values of $\hat{p}_{x,j}$ that are 0 for many values of $j$. Computationally, the quantity $\prod_j \hat{p}_{x,j}$, then quickly becomes smaller than the numerical tolerance $\epsilon$ of most machines. This may be

remedied by taking the log LRT (LLR), which we here denote

$$LLR = \log(\Lambda) = \sum_j x_j \log(\hat{p}_{j,x}) + y_j \log(\hat{p}_{j,y}) - x_j \log(\bar{p}_{j,xy}) - y_j \log(\bar{p}_{j,xy}) \quad (18)$$

Noting that the LLR serves to functionally weigh the estimates $p$ by their respective observations, we also introduce an unweighted log-likelihood ratio test (UWLLR),

$$UWLLR = \sum_j \log(\hat{p}_{j,x}) + \log(\hat{p}_{j,y}) - 2\log(\bar{p}_{j,xy}) \quad (19)$$

**Difference between our dissimilarity metric (Equation 18) compared to the Poisson dissimilarity measured proposed in Witten (2011) [14] and the multinomial dissimilarity in Berninger et al. [27].** Let $\|\cdot\|_1$ represent the L1 norm (vector sum) of a single cell/observation. Using our notation, the LLR calculated using the MLE (Equation 21 in [14]) for these two is given by

$$\Big(\frac{x_j}{\|\mathbf{x}\|_1}\Big)^{x_j}\Big(\frac{y_j}{\|\boldsymbol{y}\|_1}\Big)^{y_j} \Big/ \frac{1}{2}\Big(\frac{x_j}{\|\mathbf{x}\|_1} + \frac{y_j}{\|\mathbf{y}\|_1}\Big)^{xj+y_j} \quad (20)$$

We can similarly re-write down our LLR (Eq:17) with MLE estimates

$$\Big(\frac{x_j}{\|\mathbf{x}\|_1}\Big)^{x_j}\Big(\frac{y_j}{\|\mathbf{y}\|_1}\Big)^{y_j} \Big/ \Big(\frac{x_j + y_j}{\|\mathbf{x}\|_1 + \|\mathbf{y}\|_1}\Big)^{x_j+y_j} \quad (21)$$

With some manipulation it can be seen that the numerators for Equations 20-21 are the same, and the denominators of both are equal when $\|\mathbf{x}\|_1 = \|\mathbf{y}\|_1$. In other words, our estimator theoretically differs from the work of [14] and [27] except for cells of the exact same library size. In addition to this difference regarding hypothesis test formulation, [14] and [27] both utilize priors, which can be given in `minicore` [28]. In this work we calculate the LLR as per Equation 18, so any portion multiplying $x_j = 0$ or $y_j = 0$ is simply not counted toward the distance sum, essentially giving these observations zero weight. This allows us to calculate Equation 18 without a

prior or 'pseudocount', a notable difference from the work of [14] and [27].

### 2.1.3 Comparison of dissimilarity measures

In this section, we seek to group the dissimilarity measures using their mathematical properties given in Table 1. We grouped them into two categories: (i) geometric and (ii) probabilistic. The first category uses un-normalized counts $\mathbf{x}$, while the latter uses unit-normalized vectors $\hat{\mathbf{x}}$. Specifically, the distances with standard geometric interpretation (L1, L2, SQL2) are calculated using un-normalized vectors, while the assumption of a generative multinomial model implicitly normalized the observations in the other distances ([R]ISD, SIS, [UW]LLR, JSD[M], BCD[M], HEL, TVD, MKL). Some dissimilarities further standardize the observations using square-root transformations (HEL, BCD), or square-root the resulting summand (L2, JSM, BCM). Other dissimilarities standardize with a log-transformation of the distances or their ratio (MLK, [R]ISD) or their resulting summands (BCM). Several dissimilarities utilize a dot product between the two vectors (BCD[M], [UW]LLR). Several dissimilarities utilize a difference between vectors (L1, L2, SQL2, HEL, TVD). Further distances take an absolute-value transformation (TVD, L1). Other distances square entries (L2, SQL2) or their summands (SQL2).

We note that many of the aforementioned dissimilarities are defined using ratios of observations. Due to the sparsity of scRNA-seq data, many of these ratios will be undefined. One way to circumvent this is to calculate the given dissimilarity using a multinomial distribution with a Gamma$(\beta, \beta)$ prior [14]. This practice is quite common in the analysis of scRNA-seq data, and is often referred to as adding a 'pseudocount' [25]. In-depth exploration of prior parameter choice (pseudocount value) for scRNA-seq is beyond the depth of this work, so we simply choose the accepted value of adding a single count to each observation such that all ratios and distances are well defined. The distances to which a pseudocount is required to obtain

entirely sensible dissimilarities were MKL, SID, RSID, and SIS. As JSD and JSM are symmetrizations of MKL, the undefined indices for $MKL(\mathbf{x}, \mathbf{y})$ and $MKL(\mathbf{y}, \mathbf{x})$ can be dropped from the sum to create a defined dissimilarity. We similarly address zero-valued entries in LLR and UWLLR, specifically, we choose to drop indices for which the ratios are undefined.

| Full | Short | $\frac{X}{\|X\|}$ | $\sqrt{X}$ | $\|X\|$ | $log(X)$ | $X \cdot Y$ | $X - Y$ | $X^2$ | $\frac{X}{Y}$ |
|---|---|---|---|---|---|---|---|---|---|
| L1 Distance | L1 | - | - | + | - | - | + | - | - |
| L2 Distance | L2 | - | + | - | - | - | + | + | - |
| Squared L2 | SQL2 | - | + | - | - | - | + | + | - |
| Kullback-Leibler Divergence | MKL | + | - | - | + | - | - | - | + |
| Jensen-Shannon Divergence | JSD | + | - | - | + | - | - | - | + |
| Jensen-Shannon Metric | JSM | + | + | - | + | - | - | - | + |
| Total Variation Distance | TVD | + | - | + | - | - | + | - | - |
| Hellinger Distance | HEL | + | + | - | - | - | + | + | - |
| Bhattacharyya Distance | BCD | + | + | - | + | + | - | - | - |
| Bhattacharyya Metric | BCM | + | + | - | + | + | - | - | - |
| Likelihood Ratio Test | LLR | + | - | - | + | + | + | - | - |
| Unweighted likelihood Ratio Test | UWLLR | + | - | - | + | + | + | - | - |
| Itakura Saito Distance | ISD | + | - | - | + | - | + | - | + |
| Reverse ISD | RISD | + | - | - | + | - | + | - | + |
| Symmetric Itakura Saito Distance | SIS | + | - | - | + | - | + | - | + |

Table 1: **Mathematical operations performed by each distance measure**. The rows represent the measures. Columns 1 and 2 represent the full and short name for each measure. The columns 3 through 9 represent various transformations that exist in the measures. The + represents a transformation that exists in the measure, while the - represents that this transformation does not exist in the measure.

## 2.2 Data

In practice, the underlying structure of the observations in scRNA-seq data will be unknown. To this end, we assembled both simulated and real scRNA-seq datasets order to asses performance of each distance metric in a variety of scenarios. In terms of simulated data, we used a state-of-the-art simulation package `Splatter` [33]. In terms of real scRNA-seq data, we leveraged single cells measured using the CITE-seq protocol which simultaneously quantifies gene and protein expression to provide as

an orthogonal and more biologically driven labeling of the cell types. These included datasets generated by multiple scRNA-seq protocols with both discrete and continuous similarity structures amongst the cells. Using simulated datasets, we further investigated the potential effects of observation clustering in addition to changes in technical variation such as average library size.

### 2.2.1 Single cells from scRNA-seq cell lines

1. Single cells and pseudo cells from the CellBench [34] scRNA-seq benchmarking dataset

    - All UMI-based data from five cell lines (HCC827, H1975, H2228, H838, A549) in the CellBench [34] benchmarking dataset (except for *cellmix5*, a population control) using the CEL-seq2 protocol (*sc_celseq2*, *sc_celseq2_5cl_p1*, *sc_celseq2_5cl_p2*, *sc_celseq2_5cl_p3*, *cellmix1*, *cellmix2*, *cellmix3*, *cellmix4*, *RNAmix_celseq2*), Drop-seq Dolomite protocol (*sc_dropseq*), the Sort-seq protocol (*RNAmix_sortseq*), and 10x Chromium Genomics protocol (*sc_10x*, *sc_10x_5cl*). For a description of the experimental design, GEO accession numbers, protocol parameters, see the sc_mixology GitHub repo and Additional file 4: Table S3.

2. *Jurkat* cell lines

    - *10x_293t_jurkat* (293T cells): $N$=3258 cells measured using UMIs and the droplet-based protocol from 10x Genomics [5] (`https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/jurkat`)

3. *HEK293T* cell lines

    - *10x_293t_jurkat* (jurkat cells): $N$=2885 cells measured using UMIs and the droplet-based protocol from 10x Genomics [5] (`https://support.`

10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/293t)

4. *50%Jurkat:50%HEK293T* mixture experiment

    - *10x_293t_jurkat* (jurkat cells): $N$=3400 cells measured using UMIs and the droplet-based protocol from 10x Genomics [5] (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/jurkat:293t_50:50)

5. PBMC CITE-seq

    - *10x_PBMC* (Peripheral blood mononuclear cell): $N$=7,865 cells measured using UMIs and surface protein expression with a droplet-based protocol from 10x Genomics [35], (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_protein_v3:50)

6. MALT CITE-seq

    - *10x_MALT* (mucosa-associated lymphoid tissue): $N$=8,412 cells measured using UMIs and surface protein expression with a droplet-based protocol from 10x Genomics [35] (https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/malt_10k_protein_v3)

### 2.2.2 Single cells from `splatter`

Count data intended to replicate patterns seen in scRNA-seq data was simulated using the `Splatter` package [33]. For each simulated dataset, $G$=20000 total features (genes) were generated, and the distance calculation and downstream analyses were conducted using only 1000 HVG's. Four variables of data structure were explored: $n$ (number of cells), $k$ (number of clusters), $b$ (balance of cell types, i.e., the relative proportions of cell type numbers within the dataset), and $t$ (trajectory structure).

Furthermore, for each study, three levels of 'difficulty' (i.e., clarity of signal or effect sizes) were simulated to assess possible interactions amongst these factors. These difficulty parameters included $lS$ (library size), $dP$ (probability of a gene to be differentially expressed), $lF$ (log-fold change of differentially expressed genes). Below we describe the `Splatter` parameters used for the four simulation studies:

**1. Increasing numbers of cells simulation study ('$n$ study')**

- $n$ was varied between $n = 1000, 5000, 10000$ cells
- $k$ was fixed at $k = 5$
- $b$ was fixed at 20% for each cluster
- Easy setting: $lS = 10.0$, $dP = 0.20$, $lF = 0.30$
- Medium setting: $lS = 9.5$, $dP = 0.10$, $lF = 0.20$
- Hard setting: $lS = 9.0$, $dP = 0.05$, $lF = 0.10$

**2. Increasing numbers of true clusters simulation study ('$k$ study')**

- $n$ was fixed at $n = 5000$
- $k$ was varied for increasing $k = 2, 5, 10$
- $b$ was varied at (with increasing $k$) at
  - $b = \{50\%, 50\%\}$ $(k = 2)$
  - $b = \{20\%, 20\%, 20\%, 20\%, 20\%\}$ $(k = 5)$
  - $b = \{10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%, 10\%\}$ $(k = 10)$
- Easy setting: $lS = 10.0$, $dP = 0.20$, $lF = 0.30$

- Medium setting: $lS = 9.5$, $dP = 0.10$, $lF = 0.20$
- Hard setting: $lS = 9.0$, $dP = 0.05$, $lF = 0.10$

**3. Changing the proportion of cell types (balanced vs imbalanced with rare cell types) ('$b$ study')**

- $n$ was fixed at $n = 5000$
- $k$ was fixed for $k = 5$
- $b$ was varied from
  - balanced $b = \{20\%, 20\%, 20\%, 20\%, 20\%\}$
  - slightly unbalanced $b = \{15\%, 35\%, 25\%, 15\%, 10\%\}$
  - very unbalanced $b = \{05\%, 55\%, 25\%, 05\%, 10\%\}$
- Easy setting: $lS = 10.0$, $dP = 0.20$, $lF = 0.30$
- Medium setting: $lS = 9.5$, $dP = 0.10$, $lF = 0.20$
- Hard setting: $lS = 9.0$, $dP = 0.05$, $lF = 0.10$

**4. Simulating various types of scRNA-seq trajectories ('$t$ study')**

- $n$ was fixed at $n = 5000$
- $k$ was fixed for $k = 5$
- *nsteps* (intermediate trajectory cell type parameter) was fixed at $s = 2000$
- $t$ (trajectories) varied from $t = 1$ (linear with even balance), $t = 2$ (branched with even balance, and $t = 3$ (branched with uneven end-states). (See Figure 4C for examples of simulated lineage structures)

- $b$ was varied along with $t$. For $t = 1$ and $t = 2$: $b = \{20\%, 20\%, 20\%, 20\%, 20\%\}$. For $t = 2$: $b = \{20\%, 20\%, 20\%, 35\%, 05\%\}$.
- Easy setting: $lS = 10.0$, $dP = 0.20$, $lF = 1.00$
- Medium setting: $lS = 9.5$, $dP = 0.10$, $lF = 0.50$
- Hard setting: $lS = 9.0$, $dP = 0.05$, $lF = 0.25$

## 2.3 Data preprocessing of scRNA-seq data

### 2.3.1 Quality Control for scRNA-seq data

For each dataset and quality control metric, 1% of cells which meet known thresholding for 'poor' quality were removed. We used two accepted metrics of cellular quality, percentage of reads which map to mitochondrial genes (indicative of lysed cell [36, 37]), number of total RNA reads (library size, indicative of poor RNA-seq experimental results) using `scater` [38]. Furthermore, for the CITE-seq datasets, the total number of protein counts was used as a quality control metric. The smallest and greatest half-percentiles were removed then removed from each dataset.

### 2.3.2 Normalization of scRNA-seq data

The data were potentially normalized twice in these analysis. First, to produce any PCA plots, counts were log2-normalized using [38] using `scater`. All other normalization was performed implicitly during distance matrix calculation using `minicore` [28]. See section 2.1.3 for an explanation of which distances implicitly normalize (ie., count, square-root, or log transformations) observations during calculation of the distance matrix.

### 2.3.3 Cellular Identities for CITE-seq data

Cellular surface protein abundance was used to generate the truth labels for the CITE-seq datasets. Identities for the 10x PBMC CITE-seq dataset were generated according to established markers for immune PBMC cells. Specifically, hierarchical clustering was induced on the log-normalized counts of CD8a (a T-cell coreceptor highly expressed in cytotoxic t-cells, [39]), CD14 (highly expressed in monocytes lineages [40]), CD16 (a marker for most cytotoxic cell types [41]), CD19 (a marker for B-cell lineages [42]), CD3 (T-cell co-receptor shared across all T-cell lineages [43]), and CD4 (a marker for helper T-cells [44]) as seen in Figure S1. The dendrogram induced from clustering the cells on these protein markers was then cut at a height, which gave $k = 5$ clusters to induce the labels seen in Figure S1.

Identities for the MALT CITE-seq dataset were similarly induced hierarchical clustering ($k = 2$) with log-normalized CD3 (T-cell marker) and CD19 (B-Cell marker) as can be seen in Figure S2.

## 2.4 Null simulation using scRNA-seq data

We consider the simplest case in a scRNA-seq experiment, that is, all sequenced cells belong to a homogeneous population measured using the cell transcriptome profile in a 'null' setting and ask whether the dissimilarity measures find evidence (or not) for 1 or more than one population of cells. The idea is that we assume that all dissimilarity measures perform equally in terms of being able to recover one homogeneous population of cells. However, scRNA-seq is affected by experimental artifacts (sequencing depth, normal variability of the cellular genome with a population), which may cause a subset of dissimilarity measures to induce false structure (clustering) among observations which are homogeneous in truth.

To study this 'null' scenario, we leverage two types of scRNA-seq data: (i) a pub-

licly available 10x Genomics dataset comprised solely of 293T cells [5], and (ii) simulated scRNA-seq data of variable difficulty using the `Splatter` [33] R/Bioconductor package. In this 'null' simulation study, we defined the true number of clusters ($k$) to be 1. To quantitatively evaluate whether the distance measures identify unwanted, false signal (e.g. $k \neq 1$), we utilize the Gap Statistic [45] described in Section 2.5.2. Specifically, we apply multidimensional scaling (MDS) [16] with the distance measures, followed by inducing cluster labels using Partitioning Around Medoids (PAM) [46]. Finally, we apply the Gap Statistic at $k = 1$ to 5, which we refer to Gap(k). In Section 2.5.2, we describe more details on the Gap Statistic and how to interpret the results using it as a performance metric.

## 2.5 Performance metrics to evaluate dissimilarity measures

### 2.5.1 Partitioning Around Medoids and Adjusted Rand Index

Here, we describe a performance metric that we use in our evaluation of dissimilarity measures when there is a reliable ground-truth cluster label associated with each cell. We do this in the context of a downstream analysis, namely unsupervised clustering of scRNA-seq data. First, to assess the downstream behavior of each dissimilarity measure, we utilize the $k$-medoids (also known as partitioning around medoids, PAM) family of algorithms to induce cluster labels [46] using `cluster` [47]. The $k$-medoids algorithms differ from the popular $k$-means algorithms in that they may utilize a generalized dissimilarity matrix as input, whereas $k$-means is generally defined for L2 distance. Broadly speaking, this family of algorithms seeks to minimize the sum of within-cluster distances ($W_k$), and operates as follows.

1. Greedily select $k$ of the $n$ observations as the centroid (medoid)

2. Associate each data point with its closest centroid

3. Calculate the change in $W_k$ associated with swapping every centroid and non-centroid point

4. While $W_k$ decreases, greedily select the swaps which give the greatest decrease in $W_k$

As we have curated datasets with reliable ground-truth cluster labels (see Section 2.2), we can assess the performance cluster labels induced via $k$-medoids. A popular method to non-parametrically compare the concordance of two sets of cluster labels for the same data is the Adjusted Rand Index (ARI). Briefly, ARI utilizes the values of a contingency table for two groupings or partitions (e.g. cluster labels) $A = \{A_1, A_2, \ldots, A_r\}$ and $B\{B_1, B_2, \ldots, B_s\}$ to calculate their concordance. Let $n_{ij}$ denote an entry in the contingency table which is the number in common between $A_i$ and $B_j$ or $n_{ij} = |A_i \cap B_j|$. Further let $a_i$ and $b_j$ denote the marginal sums for clusters $A_i$ and $B_j$. Then ARI is defined as follows:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \tag{22}$$

For each dissimilarity matrix, we induced PAM labels for the known value of $k$, and generated the ARI as a function of these labels versus the truth labels.

### 2.5.2 Gap Statistic

Next, we describe the gap statistic as introduced by Tibshirani et al. (2001) [45], which we use a performance metric in our evaluation of dissimilarity measures in the setting where we have only a homogeneous population of cells (Section 3.1). Specifically, we denote the observed scRNA-seq count data as $\{x_{ij}\}$ where $i = (1, 2, \ldots, n)$ and $j = (1, 2, \ldots, m)$ consists of $m$ features (genes) and $n$ observations (cells). Let $d_{ii'}$ denote an arbitrary dissimilarity measurement between observations $i$ and $i'$. In this

notation, a commonly used dissimilarity measure is the squared Euclidean distance $\Sigma_j(x_{ij} - x_{i'j})^2$. Assume the data have been clustered in to $k$ groups, $\{C_1, C_2, \ldots, C_k\}$ with $C_r$ denoting the indices for a cell which has been placed in cluster $r$ and $n_r = |C_r|$. We then define the sum of within-cluster distances as follows

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \tag{23}$$

To appropriately weight this sum by the number of clusters $r$, we then define $W_k$:

$$W_k = \sum_{r=1}^{k} \frac{D_r}{2n_r} \tag{24}$$

The Gap Statistic [45] was derived in order to quantify the evidence for a specific value of $k$ in the context of clustering. Specifically, whether a given set of cluster labels has less within-cluster distance ($W_k$) than one would expect by chance of clustering data with no intrinsic cluster structure. One way this idea may be formalized is by inducing cluster labels followed by calculating $W_k$ from simulated datasets of the same size and feature space as the original dataset. The weighted within-cluster distance for these background datasets is then denoted as $W_k^*$. In practice, $W_k^*$ is calculated using $B$ simulated reference datasets and then compared to the true values of $W_k$. Hence, the Gap Statistic is given by

$$\text{Gap(k)} = \frac{1}{B} \sum_{b} \log\left(W_{kb}^*\right) - log(W_k) \tag{25}$$

In this way, Gap(k) represents a weighted difference between the sum of within-cluster distances for our observed data ($W_k$) and uniform noise ($W_{kb}^*$). In general, large positive values of Gap(k) indicate that the observed cluster labels induce smaller values of within-cluster distances (cluster tightness) than one would expect by chance (hence better performance). As Gap(k) approaches 0 or becomes negative, this is evidence that there is no little to no cluster structure to the data that $k = 1$. In

practice, Gap(k) is calculated for for a range of $k$ values, and tested whether the highest Gap(k) is substantially greater than others values.

Furthermore, the Gap Statistic provides a useful way to quantify evidence for the presence (or absence) of clusters. If Gap(k) is non-increasing as a function of k, this indicates that no information is gained at any particular value of k, i.e., there is no evidence for any cluster structure of the data. Moreover, negative values of Gap(k) indicate that information is lost by clustering the data, in other words, this is explicit evidence of no cluster structure within the data.

To this end, we induced PAM cluster labels for $k = 1, \ldots, 5$ on utilized real and simulated datasets with no cluster structure (i.e., a single cell type). We then reported scaled and unscaled Gap(k). Briefly, for each distance and $k = 1, \ldots, 5$, the data $x = Gap(k)$ were scaled $function(x) = (x - min(x))/(max(x) - min(x))$ such that each $x$ has a range of $[0, 1]$ (Figure 1E, Figures S4-S5).

### 2.5.3 G+

Next, we describe the a performance metric that assesses the tightness of the use of a ground-truth label clusters. One potential problem with assessing the tightness of the clusters when using different dissimilarity measures is that each dissimilarity measure has different ranges of values that it spans (some are bounded, others not [28]). Therefore, when we compare the performance across dissimilarity measures, it is difficult to interpret the results.

To address this, we use the G+ performance metric [48, 49]. For $n$ observations, a dissimilarity matrix will generally have dimension $n \times n$, with the all diagonals having value 0. As the matrix is symmetric, there are then $N_t = \frac{n(n-1)}{2}$ unique (total) dissimilarity values. Given a cluster label for each observation, each of these dissimilarity values can be classified as a within-cluster ($I_w$) or between cluster ($I_B$). One can assess the tightness of these cluster labels without explicitly biasing the performance

by the numerical value of the dissimilarity measure by counting the number of times a within-cluster distance is strictly larger than a between-cluster distance. Adopting the notation of [49], we denote this quantity as $s^-$, which can be explicitly written as follows. Where $r, s$ index over the rows and columns corresponding to within-cluster distances ($I_W$) and similarly with $u, v$ between-cluster distances ($I_B$)

$$s^- = \sum_{(r,s)\in I_B} \sum_{(u,v)\in I_W} 1_{d_{uv}>d_{rs}} \tag{26}$$

As there are $N_t$ distinct dissimilarity values, this summand requires a total of $\frac{N_t(N_t-1)}{2}$ comparisons. The $G+$ index is then defined as follows

$$G+ = \frac{s^-}{N_t(N_t-1)/2} \tag{27}$$

In the case that every distance pair is discordant (every within-cluster distance is bigger than every between-cluster distance), then $G+ = 1$. In the opposite case, $G+ = 0$. Perhaps un-intuitively, the lower a value of $G+$ indicates greater cluster tightness. Thus, for the purpose of this work, we report the value $1-G+$, which retains the same bounded behavior, but indicates performance is best when $1 - G+ = 1$.

In the case where there is no cluster structure in the data, the expected value $E[G+] = \frac{1}{4}$. For intuition on this expected value, we expand on this mathematically. First, we simulate this case with Gaussian data with observations that have been randomly assigned cluster labels in (Figure S3A, left column). Theoretically, we assume that samples drawn from within-($d_W$) and between-cluster ($d_B$) distances (Figure S3B, left column) have all come from the same distribution, which we parameterize with mean $\mu$ and variance $\sigma^2$. More explicitly, $E[d_W] = E[d_B] = \mu$ and $Var[d_W] = Var[d_W] = \sigma^2$. Using fundamental properties of the indicator function, we first note

$$E[1_{d_W>d_B}] = P(d_W > d_B) \tag{28}$$

We simplify the resulting probability by defining $X = d_W - d_B$. Using definitions expectation and variance, we also note that

$$E[X] = E[d_W - d_B] = E[d_W] - E[d_B] = \mu - \mu = 0 \tag{29}$$

Assuming that $d_W$ and $d_B$ are independently drawn, we also have that

$$Var[X] = Var[d_W - d_B] = Var[d_W] + Var[d_B] - 2Cov(d_W, d_B) = 2\sigma^2 \tag{30}$$

Using these equalities, and invoking the Central Limit Theorem along with the symmetry of the Gaussian distribution, we find

$$P(d_W > d_B) = P(X > 0) = P\left(\frac{X - E[X]}{Var[X]} > \frac{E[X]}{Var[X]}\right) = P\left(\frac{X}{2\sigma^2} > 0\right) = \frac{1}{2} \tag{31}$$

Thus, we now have that

$$E[1_{d_{uv} > d_{rs}}] = P(d_{uv} > d_{rs}) = \frac{1}{2} \tag{32}$$

Conditional on the number of observations, $s^-$ can then be evaluated as

$$s^- = \sum_{(r,s) \in I_B} \sum_{(u,v) \in I_W} 1_{d_{uv} > d_{rs}} = \frac{N_t(N_t - 1)}{2} \times \frac{1}{2} = \frac{N_t(N_t - 1)}{4} \tag{33}$$

By plugging this value for $s^-$ into the formula for G+, it can be seen that the in the case of totally random data, we expect that $E[G+] = \frac{1}{4}$ or $1 - E[G+] = \frac{3}{4}$ (Figure S3C, left column).

### 2.5.4 Efficient Estimation of G+

Full calculation of G+ requires comparison of $O(n^2)$ distinct dissimilarity values to each other. This necessitates $O(n^2) \times O(n^2) \approx O(n^4)$ comparisons, a computational order which quickly becomes impractical as the number of observations increases. In

practice, the information of how often within-cluster distances $D_W$ are strictly greater than between-cluster distances $D_B$ is largely contained within the order statistics of these sets. For example, if the biggest within-cluster distance is less than the smallest between-cluster distance, or $max(D_W) < min(D_B)$, then the $G+ = 0$ for this dissimilarity matrix and label set. We leverage this observation to efficiently approximate $G+$ using $o(n)$ comparisons. Specifically, we calculate $p$ order statistics for $D_W$ and $D_B$. The estimator for G+ utilizes the same equation as described above, simply using these order statistics in place of the full sets. We demonstrate the efficiency of this estimator for simulated data and the L2 distance in Figure S3A-D. In simulated examples, our estimator approaches the true value of G+ sampling few as %0.25 of the order statistics (Figure S3C). In our work, we utilize sampling %1 of all dissimilarity values ($p = \%0.5$ for $D_W$ and $D_B$ each) to ensure the asymptotic behavior of this estimator. For simulated datasets, of 1000 observations, we observe substantial improvement in computational performance using our estimator ($\approx$ 10 minutes for the full calculation versus $<$ 10 seconds for the estimated version using %1 of all dissimilarity values).

### 2.5.5 kAccuracy

Many popular clustering methods for scRNA-seq data utilize graph-based clustering algorithms, such as Louvain or Leiden clustering [50, 51]. Intrinsic to these algorithms is the generation of an adjacency matrix (or graph). In short, an adjacency matrix is a weighted or unweighted similarity matrix which describes the strength of connections (edges) between observations (nodes). Many methods induce an adjacency matrices from a kNN/sNN (k/shared nearest neighbors) matrix. In order to test potential performance for methods generated from this series of algorithms, we calculated the kNN accuracy (kAcc) for each cell. For given $k$ (in this work, we fix $k = 50$), we define kAcc as the portion of cell $i$'s $k$ closest neighbors that reside in the same cluster as

cell $k$. For given cluster labels $c_r$

$$kAcc = \frac{\sum_{j=1}^{k} 1_{c_i=c_j}}{k} \tag{34}$$

### 2.5.6 Performance metrics for Trajectory analyses

Differences between cell types are often continuous. For example, a single progenitor cell population may differentiate into several end-states given different environmental stimuli. In this case, the differentiated end-states of the cells may be accurately described as discrete groups/clusters, but the difference between the progenitor and differentiated cells will follow a continuous trajectory. One way in which this continuous trajectory manifests with cells intermediate to the differentiated (end) and progenitor (beginning) state. In this case, clustering algorithms designed to classify the cells into discrete groups will inaccurately describe the cells in intermediate differentiation states. To account for these types of continuous biological processes, researchers have introduced 'pseudotime' analyses which seek to accurately project the cells along a single 'time' dimension [10, 11]. More explicitly, these methods seek to represent a continuous biological differentiation process by ordering cells along a single 'time' trajectory. R researchers can then study genomic regulation of differentiation processes associated with these lineages.

We sought to order test differential effects of dissimilarity measurements on the accuracy of pseudotime analyses. To this end, we utilized the `Slingshot` [52] R/Bioconductor package, as it allows the user to provide its own set of reduced dimensions. For each dissimilarity matrix, we used MDS `cmdscale` [30] to project the observations into low-dimensional ($m = 50$) space, and provided this representation to `Slingshot` for trajectory analyses. As the true lineage structure of each test dataset was known, we could provide the start and end cluster labels as further input the `Slingshot`, such that the algorithm was primarily testing the capacity of each dissimilarity measure

to capture the latent trajectory structure. Once `Slingshot` had detected lineages within the dataset, the `Slingshot` package was then used to project each cell onto this lineage, which provides a single numerical feature (i.e., its pseudotime value) that represents the place of this cell in the given lineage. For each lineage, the rank correlation for the calculated pseudotime and the truth label order of these cells was calculated and reported as the 'TrajCor' value.

# 3 Results

Using the dissimilarity measured described above, we consider three benchmark evaluations. First, we performed a null analysis where we expect no structure in the data and evaluate if the distance measures artificially induce any structure in a downstream analysis, namely unsupervised clustering (Section 3.1). Second, we use real and synthetic scRNA-seq data with real biological structure to evaluate which distance measures most accurately capture the true biological variation: (i) we consider scRNA-seq falling into discrete clusters and use unsupervised clustering (Section 3.2) and we consider scRNA-seq along a continuum and use trajectory analysis (Section 3.3).

## 3.1 Null analysis

We consider the simplest case in a scRNA-seq experiment, that is, all sequenced cells belong to a homogeneous population measured using the cell transcriptome profile. In this scenario, we imagine the possibilities that experimental artifacts (sequencing depth, normal variability of the cellular genome with a population) may induce false structure among observations which are homogeneous in truth. To study this scenario, we leverage a publicly available 10x Genomics dataset comprised solely of 293T cells [5]. We further investigate these phenomena using simulated datasets of variable diffi-

culty using the `Splatter` [33] R/Bioconductor package. In the previously mentioned cases, the true number of clusters ($k$) is 1. To quantitatively evaluate whether the distance measures identify unwanted, false signal (e.g. $k \neq 1$), we utilize the Gap Statistic [45] described in Section 2.5.2. Specifically, we supply the distance measures and fixed $k$ for inducing cluster labels using PAM [46]. Finally, we calculate the Gap Statistic for $k = 1, \ldots, 5$ which we refer to Gap(k). Large positive values of Gap(k) indicate that the given cluster labels induce smaller values of within-cluster distances than one would expect by chance (Equation 25). As Gap(k) approaches 0 or becomes negative, this is evidence that there is no little to no cluster structure to the data that $k = 1$. In practice, Gap(k) is calculated for a range of $k$ values and tested whether the highest Gap(k) is substantially greater than others.

Figure 1A provides a schematic demonstration of how plotting the cells in two-dimensional space can hint at the structure of the data. In low dimensional space, homogeneous data ($k = 1$) will often tightly cluster with few outliers (Figure 1A, green). Data with distinct cell types ($k > 2$) will often separate in low dimensional space (Figure 1A, red). For these three scenarios, scaled Gap Statistic for $k = 1, 2, 3, 4, 5$ is schematically illustrated in Figure 1B. Figure 1C shows demonstrates low-dimensional embeddings (MDS) of L2 and LLR distances matrices for the 293t only data (top row) and unimodal simulated splatter data (bottom row). In Figure 1C, the data are colored by the library size of each cell, while in Figure 1D, the data are coloured by induced PAM cluster labels with fixed $k = 2$. We find that MDS1 is more closely related to library size using L2 distances as opposed to LLR distance. In other words, this false separation is reduced in the LLR plots. In Figure 1E, we demonstrate scaled Gap Statistics for all of the tested distances (unscaled values of these gap statistics are given in Figures S4-S5). In the 293T plot (top row of Figure 1E), the plot with the least increase from $k = 1$ to $k = 2$ is indicative of the dissimilarity which gives the least false signal for $k > 2$. In the simulated data

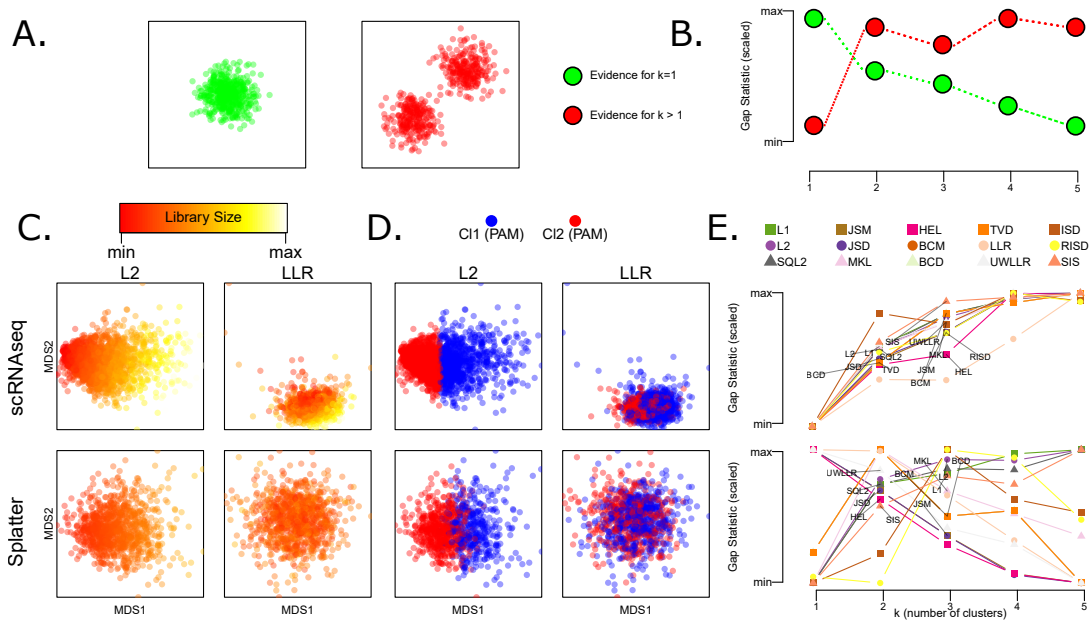


Figure 1: **Motivation for benchmark evaluation of dissimilarity measures using droplet-based scRNA-seq data**. **(A)** Schematic of two true low-dimensional representations of scRNA-seq data with one true cluster (green) and two true clusters (red). **(B)** Schematic representation of scaled Gap(k) ($k = 1, \ldots, 5$) for data with one cluster (green) and more than one cluster (red). Gap statistic plots for which $k = 1$ should peak at $G(k) = 1$, and decrease for greater $k$. Evidence for $k > 1$ would peak some value after 1. **(C)** Demonstration of how library size effects can induce false cluster structure in real (top row) and simulated (bottom row using the `Splatter` R/Bioconductor package) scRNA-seq data. The color in the first two columns represents the observed library size (defined as the total sum of counts across all relevant features) for each cell. **(D)** The color in the last two columns represents the cluster labels induced using PAM. Distances which give stronger evidence for the (true) case that $k = 1$ will have increase less (or ideally decrease) from Gap(1) to Gap($\geq 2$). **(E)** Scaled gap statistic plots for scRNA-seq (top row) and simulated (bottom row) data. Library size effects induce false signal for the scRNA-seq data, thus, distances with the least increase from $k = 1$ to $k = 2$ demonstrate the best performance. In simulated data, true evidence of $k = 1$ can be seen for several dissimilarities.

(bottom row of Figure 1E), true evidence for $k = 1$ can be seen in distances for which Gap(1) is greatest, and decreases for $k > 2$. For both of these plots, the LLR and HEL distances perform quite well, while L1 and SQL2 perform poorly. The unscaled and scaled version of the Gap statistic analyses are fully reported for the 293t-only RNA-seq data (Figures S3, Table S1) and single-cluster `Splatter` data (Figures S4, Table S2).

## 3.2 Clustering analyses

### 3.2.1 Using nine real scRNA-seq datasets

For each distance, dataset, and variable gene amount, we calculated the performance metrics described in the Methods (Section 2.5). Herein, we explore the results of the performance metrics which relate to the detection of cluster structure within the datasets. The first of these clustering metrics is 1-G+, the proportion of times within-cluster distances are strictly less than between-cluster distances. In other words, we expect standard clustering algorithms to attain better performance as 1-G+ approaches 1. The second of the clustering methods is to induce a computational set of labels using a common clustering technique (PAM), and compare the concordance of these labels with the known labels using ARI. Lastly, we assess potential of graph-based clustering algorithms in using the portion of each cell's $k$ closest neighbors which lie in the same cluster as the given cell (kAcc). For all nine of the real scRNA-

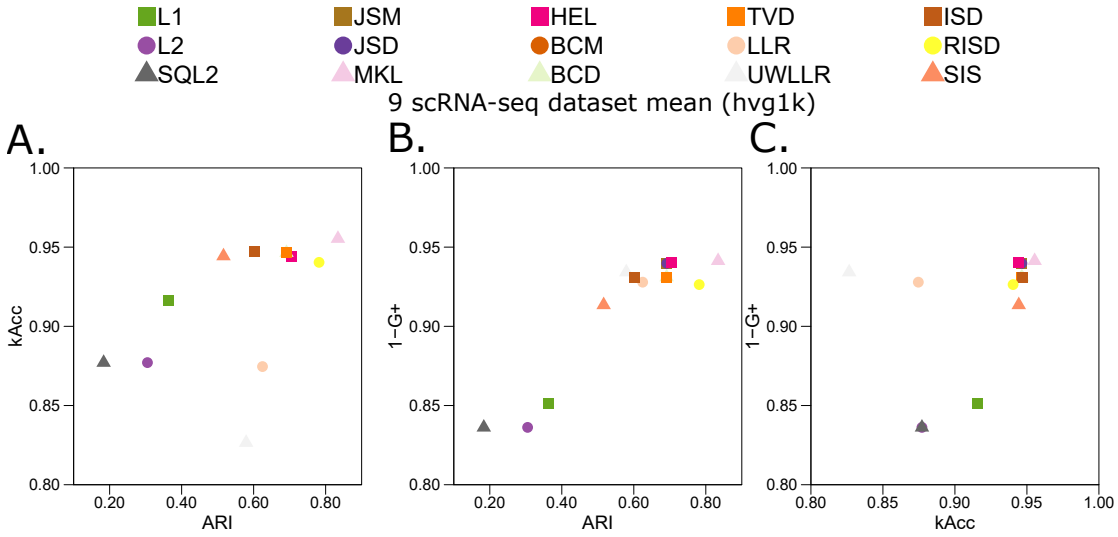


Figure 2: **Performance results for clustering analysis using real scRNA-seq data.** Paired scatter plots demonstrating the clustering performance metrics (ARI, kAccuracy, and 1-G+) in all distances tested using 1000 HVGs for all distances. **(A)** ARI (x-axis) versus kAcc (y-axis) **(B)** ARI (x-axis) versus 1-G+ (y-axis) **(C)** kAcc (x-axis) versus 1-G+ (y-axis).

seq datasets and each distance calculated with the top 1000 highly variable genes, we induced a set of cluster labels with PAM and calculated each performance metric described above. The average results across all nine scRNA-seq datasets is reported in as pairwise scatter plots (Figure 2).

For all three plots, the top performing dissimilarity measures are MKL, JSM/D, HEL, and (R)ISD. Most notably, the MKL distance is in the top-right corner in Figure 2A-C, indicating best performance for the given performance metric pair. Non-normalized distances (L1, L2, and SQL2) consistently demonstrated the worst performance in the 1-G+ metric. Greatest performance was achieved by normalized metrics in which summands and/or entries were log-transformed (MKL, JSM/D, BCM/D).

In contrast to averaging the results across all nine scRNA-seq datasets, we also provide heatmaps (and numerical values) for all three performance metrics separated out for each dataset and for considering sets of HVGs (all genes, top 5000 genes, top 1000 genes, or top 500 genes) using 1-G+ (Figure S6, Table S3), kAccuracy (Figure S7, Table S4), and ARI (Figure S8, Table S5). Notably, the performance of 1-G+ improves as fewer HVGs were used to calculated the dissimilarity metric (Figure S6), while kAccuracy and ARI are less affected by the choice of HVGs (Figures S7-S8).

### 3.2.2 Using simulated scRNA-seq data from `Splatter`

**Using 1-G+**. Using simulated scRNA-seq datasets, we found that 1-G+ is relatively stable across changes in cell type distribution (Figure S9, Table S6). Specifically, the reported distances are robust to simulated changes in cell type proportion ($b$), number of clusters ($k$), and number of cells ($n$). In terms of top performers, MKL perfromed best, with JSD/M, BCD/M, and TVD also performing well. Notably, all of these distances except TVD utilize a log-transformation, which may have contributed to the performance of these distance metrics in simulated data. Additonally, we see the

un-normalized geomtric distances (L1,L2,SQL2) consistenly demonstrate the lowest performance

**Using ARI**. We next consider the PAM-induced ARI of the distance methods for simulated scRNA-seq datasets. Similar to the 1-G+ results, non-normalized distances (L1, L2, SQL2) tend to perform quite poorly for all datasets and variable gene amounts, while normalized distances with log operations (JSM/D, MKL) perform

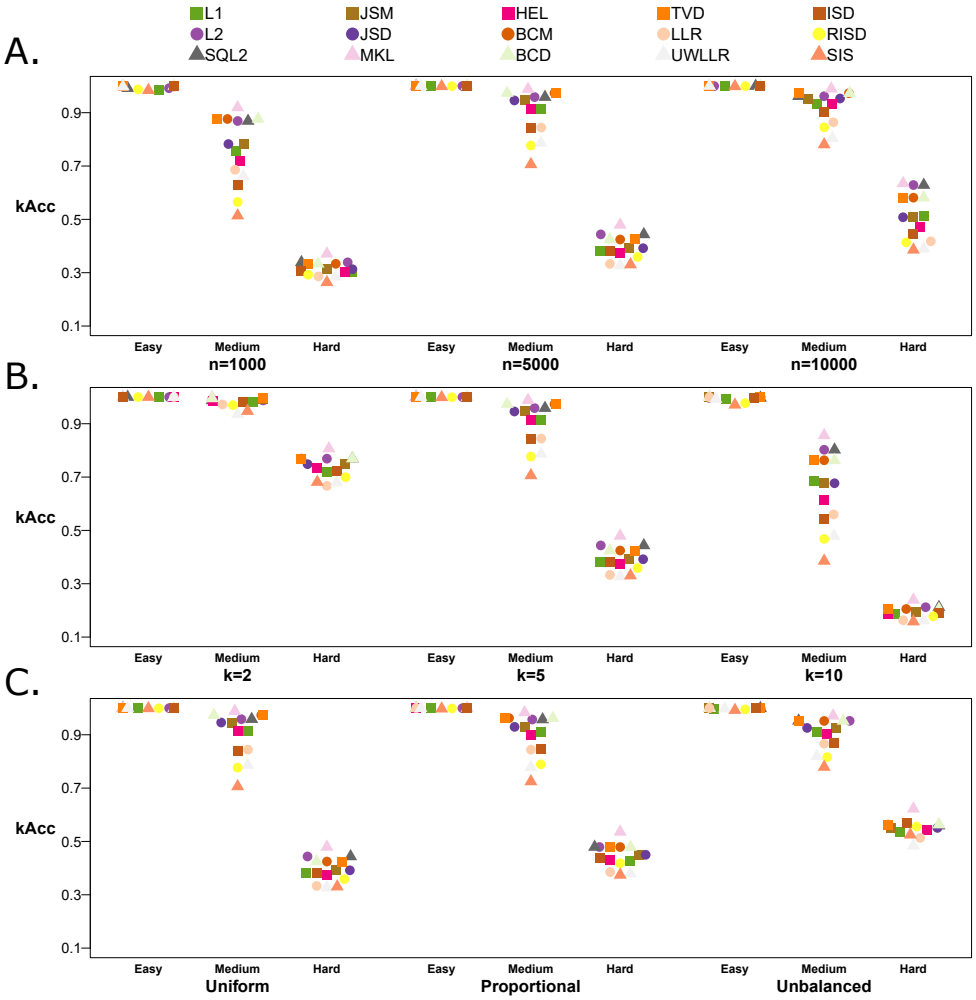


Figure 3: **Performance results with kAcc for clustering analysis using simulated scRNA-seq data.** kAcc dotplots for all distances as a function of **(A)** increasing number of cells $n$ from 1000, 5000 and 10,000 **(B)** increasing numbers of clusters $k$ from 2, 5, and 10 **(C)** and proportion of cell type balance $p$ (uniform, proportional, and unbalanced). Within each simulation framework, there are three levels of difficulties considered (see Methods Section 2.2 for details): easy, medium, and hard.

quite well (Figure S10, Table S7). Notably, the MKL distance is the only metric which demonstrates any level robustness HVG robustness in the MALT CITE-seq dataset. Furthermore, distance metrics for which summands were root-normalized (JSM, HEL) demonstrated the most robust ARI performance.

For the simulated scRNA-seq datasets, we note that ARI is largely consistent when grouped by simulation difficulty level (Figure S10). Some distances demonstrate marginally greater ARI performance for more balanced study designs or smaller $k$ (MKL, JSM/D, BCM/D). In general, however, most distances are largely robust to changes in number of cells, number of clusters, and proportional balance of these clusters.

**Using kAccuracy**. We lastly consider the kAccuracy (kAcc) of the dissimilarity measures. Broadly, we found that MKL appears as a consistent top performer within in each simulation study (Figure 3). Again, JSDM, BCD/M TVD and HEL also performed well, and suggesting that kAcc is well represented with many distances. Similar to previous results, kAcc tended to improve the distances matrices calculated with fewer HVGs (Figure S11, Table S8).

Notably, the common distinction between normalized and non-normalized distances appears less distinct in the kAcc performance. Specifically, LLR and UWLLR, both of which utilized unit-scaled entries, demonstrate worse kAcc than non-normalized distances (L1, L2, SQL2) in most of the tested scRNA-seq datasets. L2 and SQL2 were both highly robust in this metric, perhaps an artifact of the log-normal generative process from which `Splatter` counts are simulated. We also note that MKL, which performs quite well in other tests, demonstrated high-tier performance with the single exception of all using all genes in the Cellbench 10x scRNA-seq 3-cluster dataset (Figure S7).

For the simulated datasets, distances measures demonstrated increasing kAcc performance as $n$ increased from 1000 to 5000 and 10,000 (Figure S11). We hypothesize

that this behavior is partially attributable to the increased number of cells within each population, such that the odds of a given cell's $k$ closest neighbors are increased. Notably, kAcc is distinctly and negatively affected by increasing number of cell populations, particularly at more difficult simulation settings. Notably, for the hard simulation setting, the kAcc for $k = 10$ are typically one third as accurate for those of $k = 2$. We hypothesize that greater number of potential cell labels increases the odds of misclassification, which would negatively impact performance.

## 3.3 Trajectory analyses

In order to assess the manner in which the distance matrices can be accurately reduced to a one-dimensional ordering, we use MDS and `Slingshot` to construct a psuedotime ordering for each relevant trajectory dataset using all the distance measures. This analyses encompassed three RNA-seq datasets from the CellBench [34] datasets (each with four amounts of varying sets of highly variable genes) and three simulated trajectories of varying difficulty (see Section 2.2 for more details). In general, normalized and log-transformed distances (JSD/M, BCD/M) demonstrated the most robust performance in terms of real and simulated scRNA-seq datasets (Figures 4 and S12, Table S9). Similarly to the clustering data, BCD/M, HEL and TVD also appeared as top performers acrross all datasets.

Notably, these analyses appeared much less sensitive to the selection of an HVG subset (Figure S12). We hypothesize that this feature is due to the projection of these dissimlarity matrices into a lower-dimensional space via MDS. Particularly, low-dimensional embedding is a standard technique to reduce noise and improve performance of scRNA-seq analytical methods [19]. Herein, we note that in some cases, subsetting for too few HVGs results in decreased performance for some distances (for example, MKL and UWLLR in the RNAmix 10x dataset). We note that in the simulated datasets, perturbation of the trajectory structure from linear (Trajectory

1) to branched (Trajectories 2 and 3) does not drastically affect the performance. We conclude, as before, that unit-normalized and log-transformed distances tend to universally provide the most stable performance for trajectory analyses, regardless of gene selection or lineage structure.

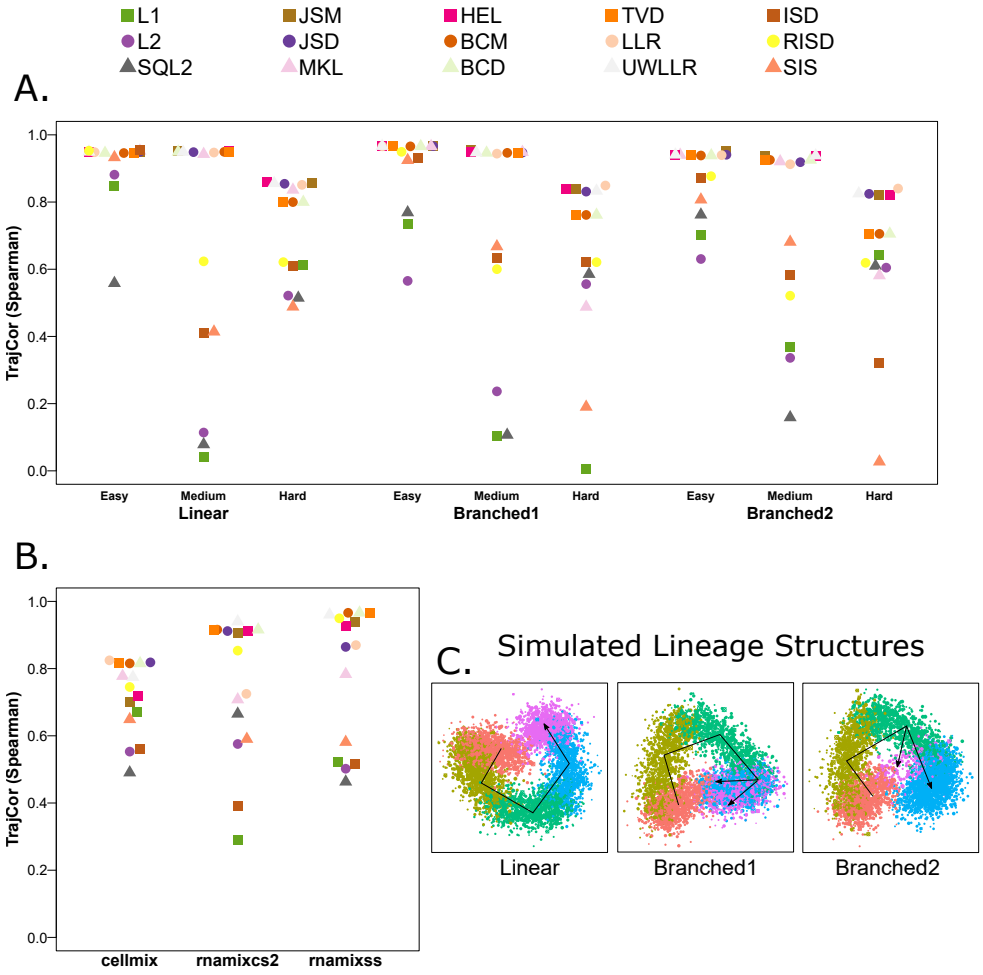


Figure 4: **Performance results with TrajCor in trajectory analysis using real and simulated scRNA-seq data.** Dotplots for trajectory correlations (TrajCor) analysis **(A)** in simulated scRNA-seq datasets and **(B)** real scRNA-seq data with three known-lineages. **(C)** PCA plots demonstrating the simulated scRNA-seq lineage structures evaluated in (A).

# 4 Discussion

In order to assess the overall performance of each distance across different datasets and computational experiments, we ranked each distance's mean behavior within performance measures using data pooled into: (i) real scRNA-seq data (Figure 5A) and

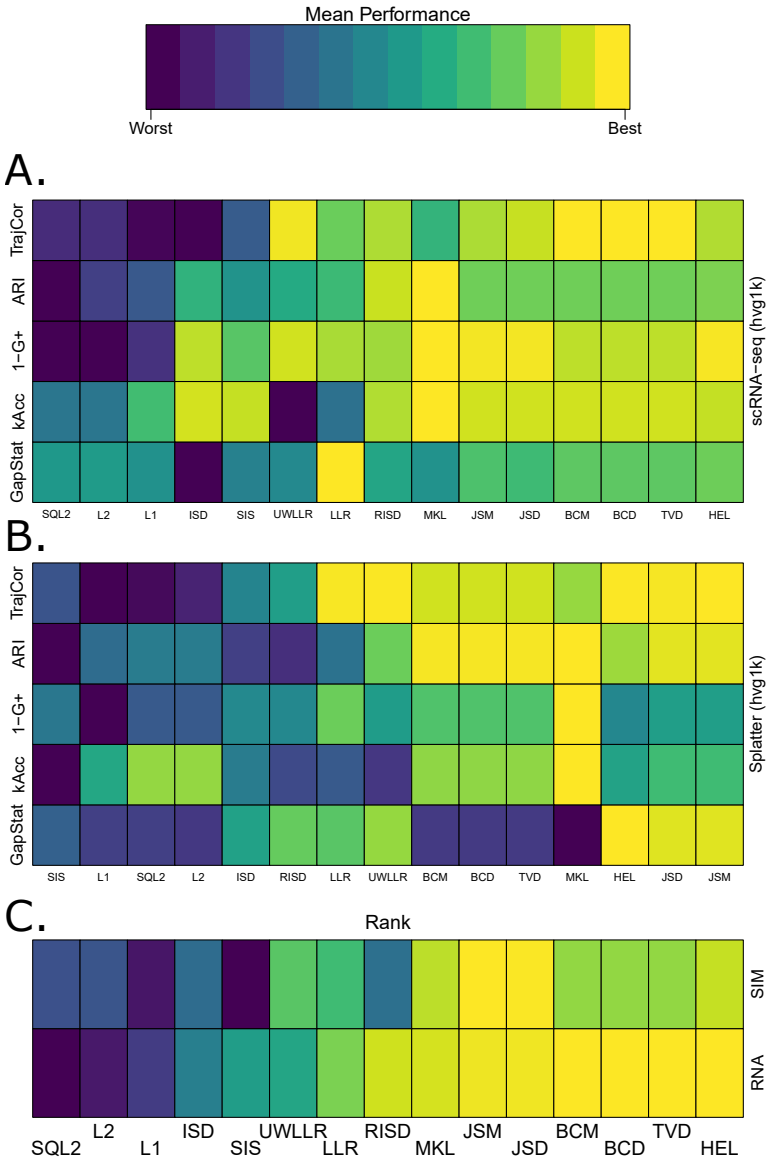


Figure 5: **Overall performance results across all analyses with each distance and performance metric. (A)** Ranked mean performance using real scRNA-seq data. **(B)** Ranked mean performance for simulated scRNA-seq data. **(C)** Final ranked performance (ordered by performance in the columns in (A)).

(ii) simulated scRNA-seq data (Figure 5B). Specifically, for the scRNA-seq datasets, mean performance for each distance was computed across each dataset and corresponding set of highly variable genes. For the simulated datasets, mean performance for each distance was computed by pooling mean performance for each difficulty and simulation setting. Each distance then had a two mean performance value for each performance metric, one calculated with simulated data and one with real scRNA-seq data. Within each data category, these preference values were then ranked according to their performance in the scRNA-seq data (Figure 5C).

Based on the results of our studies, distances which implicitly normalize data garner improved performance. Specifically, L1, L2, and SQL2 all rank amongst the bottom three (real scRNA-seq data) or bottom four (simulated scRNA-seq data) of all tested distances (Figure 5C.) This corroborates many previous studies and standard practice in analysis of RNA-seq in which normalizing data is known to reduce noise in downstream analysis [19, 18, 13]. We note that HEL demonstrates most consistent performance as the best in real scRNA-seq data, and third-best in simulated scRNA-seq data. We also draw attention to JSD/M performing in the top 5 (Figure 5A) or top 2 (Figure 5B) for each metric. Furthermore, both BCD/M are in the top 3 for real scRNA-seq data, and top 6 for simulated scRNA-seq data. Notably, all of these except HEL utilize a log transformation of the data. These findings indicate that implicit normalization (i.e., lowering the overall scale of the data) serves to decrease noise in downstream analyses. Moreover, the variety of mathematical operations amongst the top-performing distances we cannot currently determine if using one operand for (dis)similarity (dot-product in BCD/M versus ratio in JSD/M or difference in TVD) will yield drastically improved results over any other.

Based on Figures 2-4 and related supplemental figures (Figures S6-S11), strong evidence is presented that unit-normalizing data improves clustering performance. While there are fallacies to be seen with comparing gap statistics due to the bound-

edness of different distances, Figures S5-S6 show that these same distances do give overall good performance, especially in the unscaled data (bottom row) of show that these same distances tend to be very low, even negative, which does suggest no clustering (Figure S6) [45]. We refer to previous publications using `minicore` [28], in particular, JSD and BCM appear to strike a balance of performance in our work and computational speed in [28]. We note that in this plot, JSM/D perform within the top 2/3 of scRNA-seq, and the two best for simulated data. This, along with positive ARI results for JSD in [28] could indicate that that this distance is the most robust. Given the performance of JSD in [28] and the its performance in our studies, we feel comfortable suggesting the `minicore` implementation of JSD for most scRNA-seq analyses.

# Supplemental Material

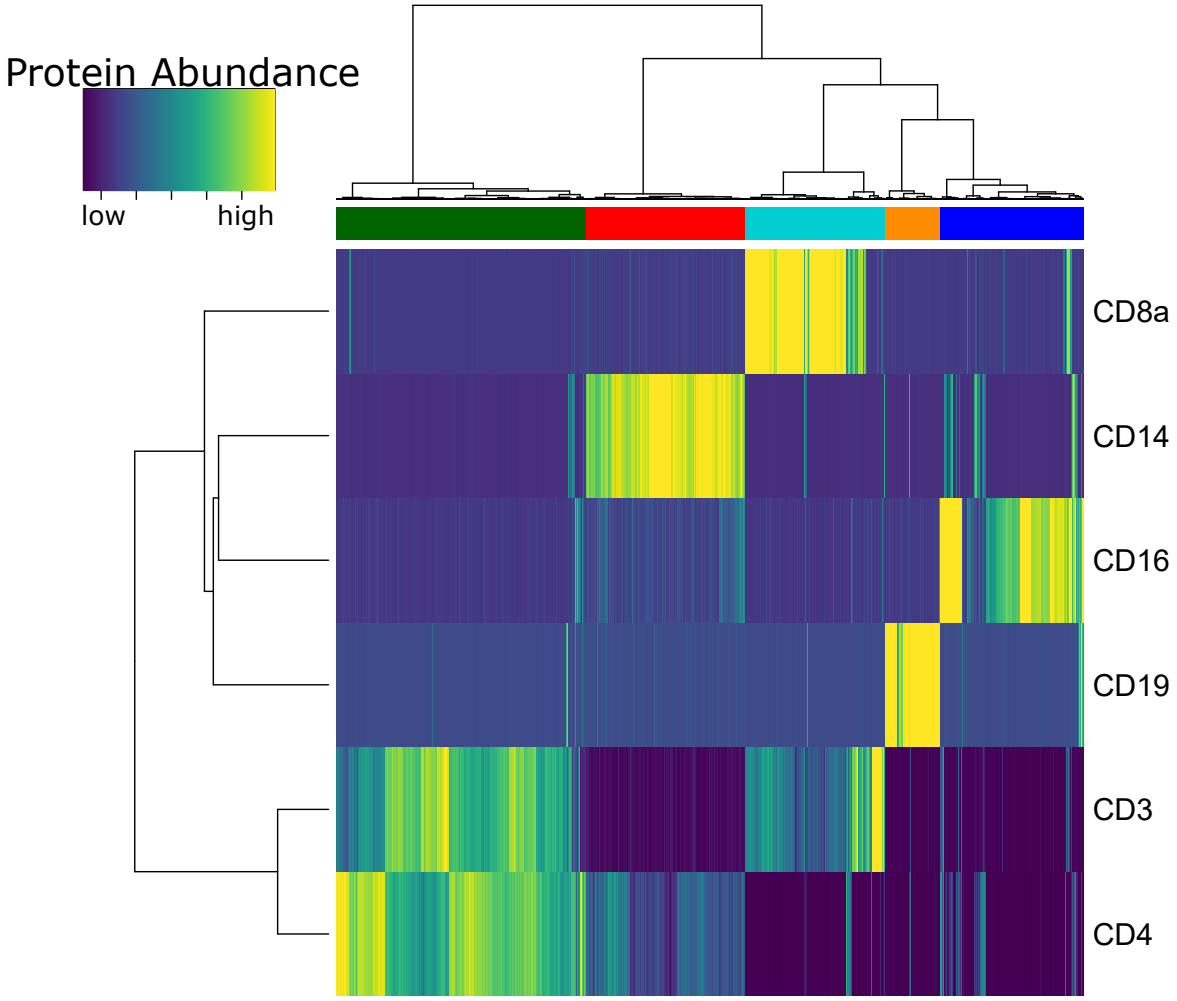


Figure S1: Heatmap of protein abundance and associated labels generated from hierarchical clustering in the 10x PBMC CITE-seq dtataset.

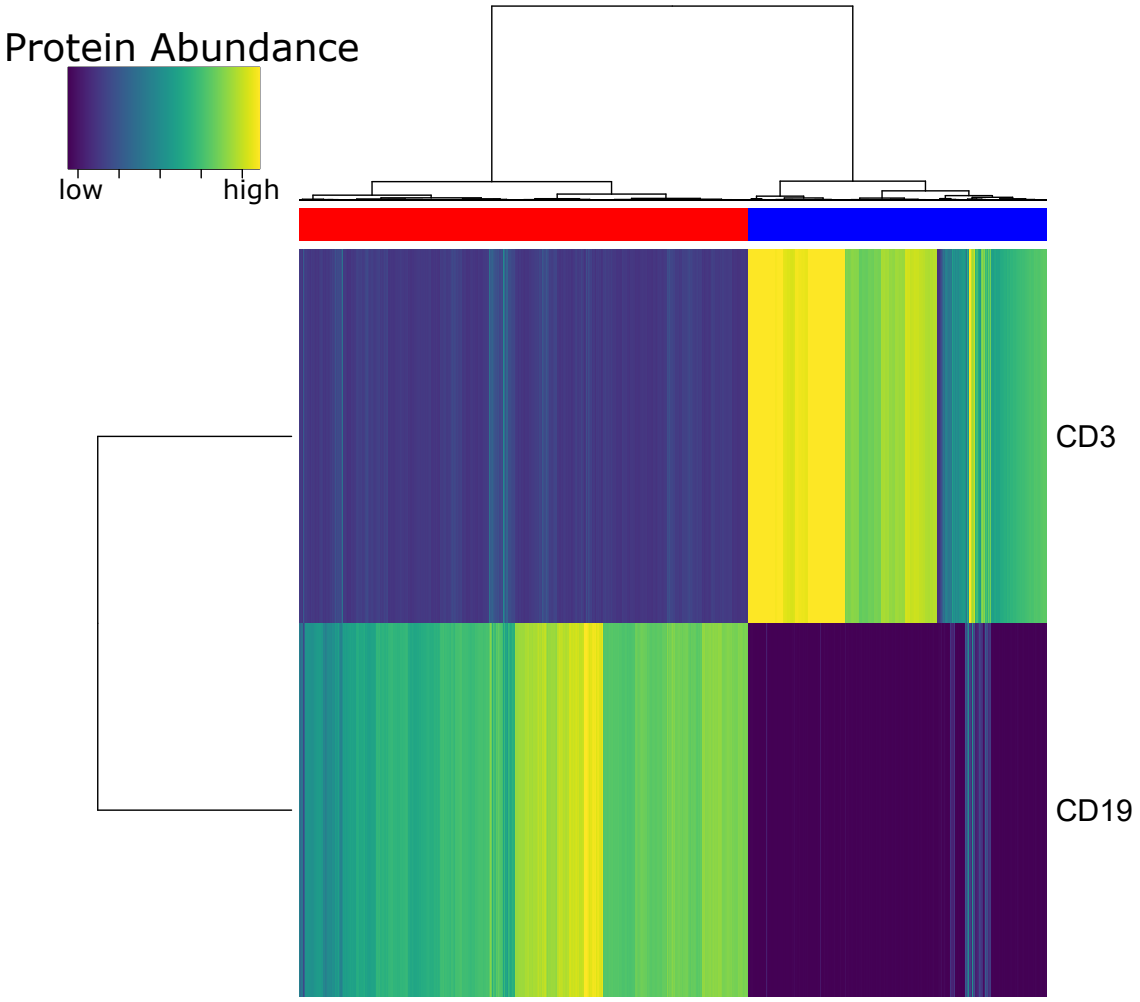


Figure S2: Heatmap of protein abundance and associated labels generated from hierarchical clustering in the 10x MALT CITE-seq dtataset.

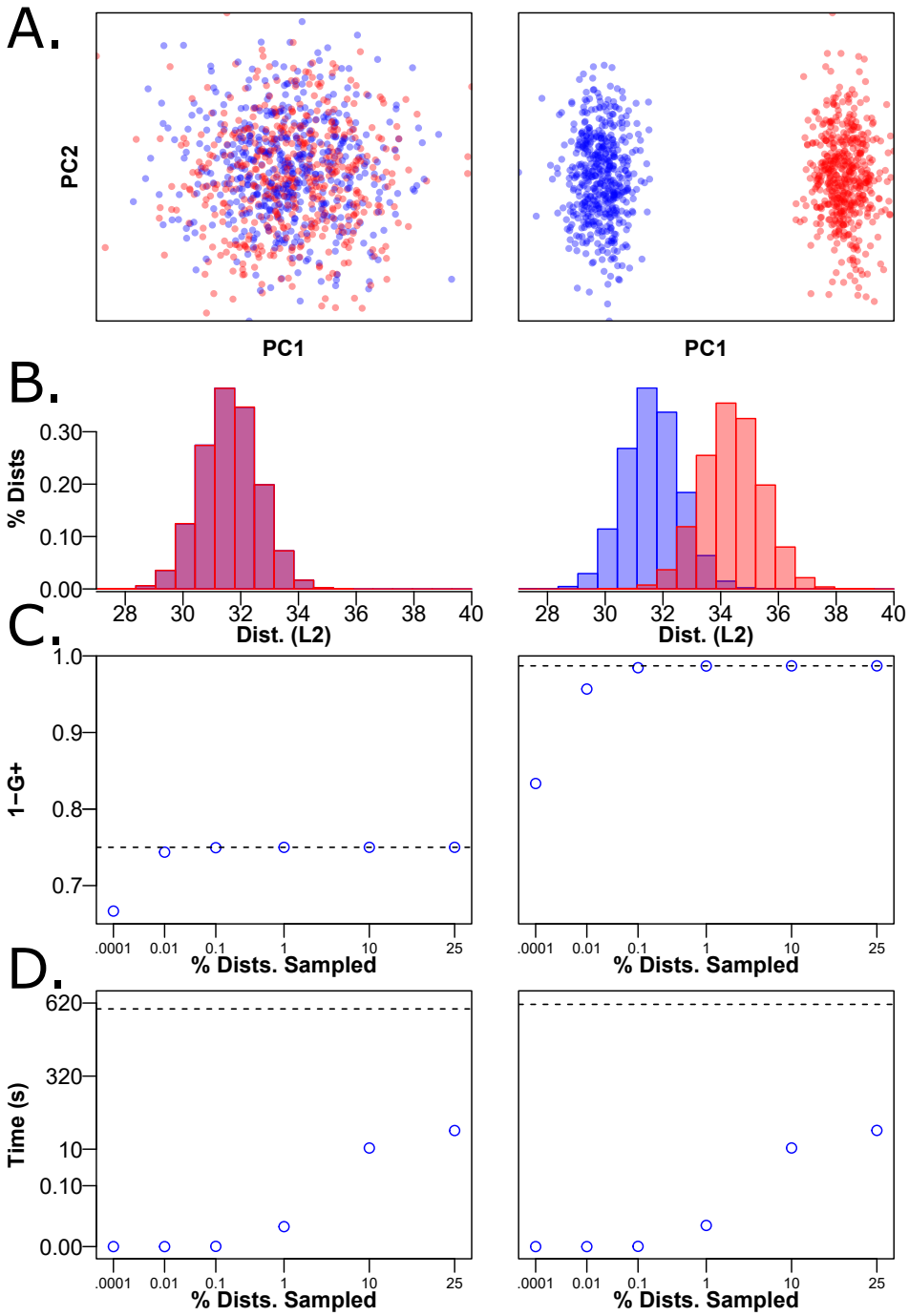


Figure S3: **(A)** PCA plots of simulated data with no structure (randomly assigned labels, left column) and two-cluster structure (right column). **(B)** Histograms of witin-cluster distances (blue) and between-cluster distances. **(C)** 1-G+ plots for the true G+ value (dashed line) and estimated values (blue circles) as a function of varying number of order statistics sampled, **(D)** Time for associated calculations in row (C) with log10-scaled y-axes for visualization.

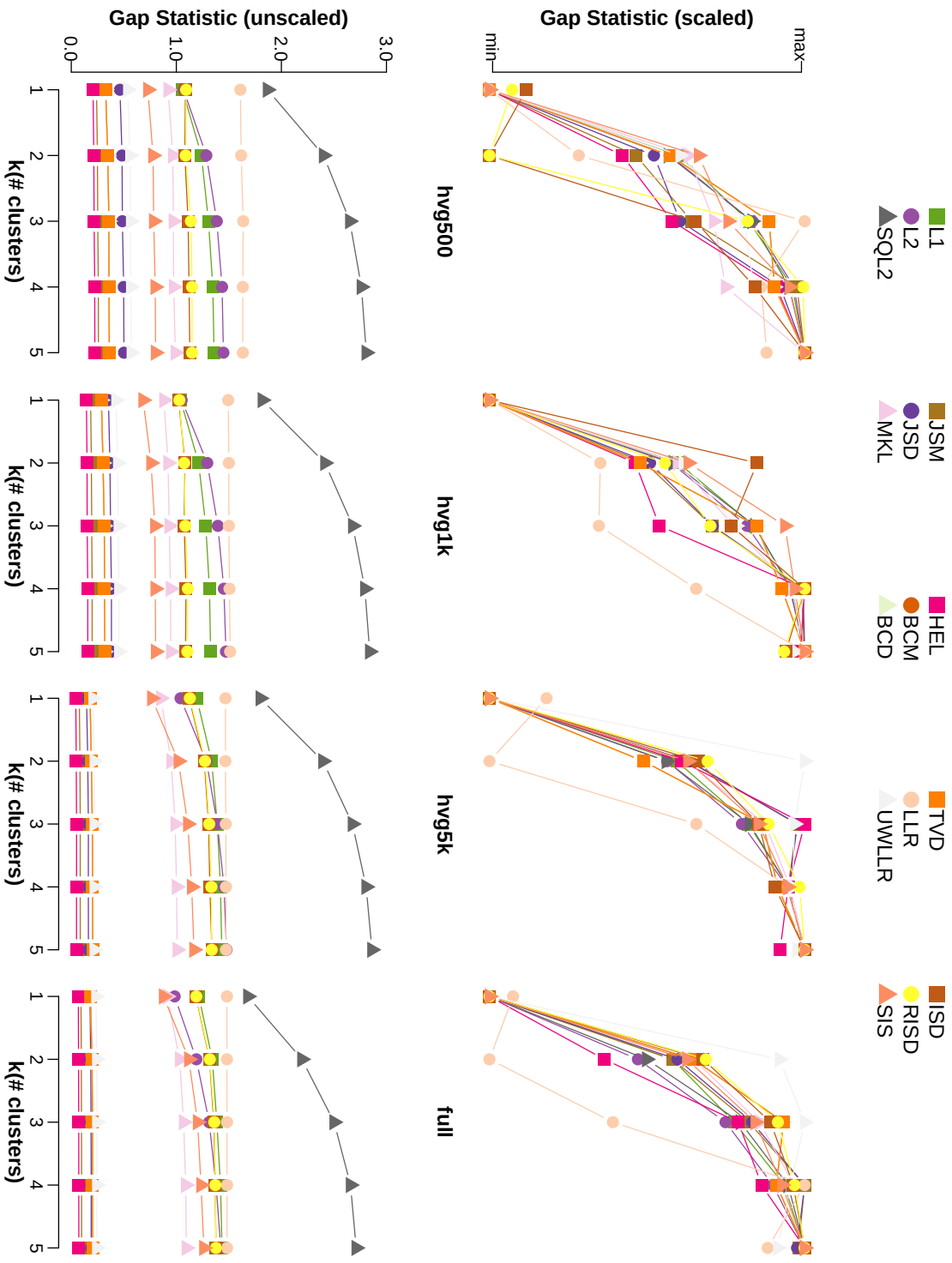


Figure S4: Gap statistic for each distance as a function of $k$ for the simulated one-cluster dataset. Gap statistics scaled to (0,1) are given in the top row, and unscaled gap statistics are given in the bottom row. Columns from left to right indicate increasing numbers of HVGs used for distance calculation.

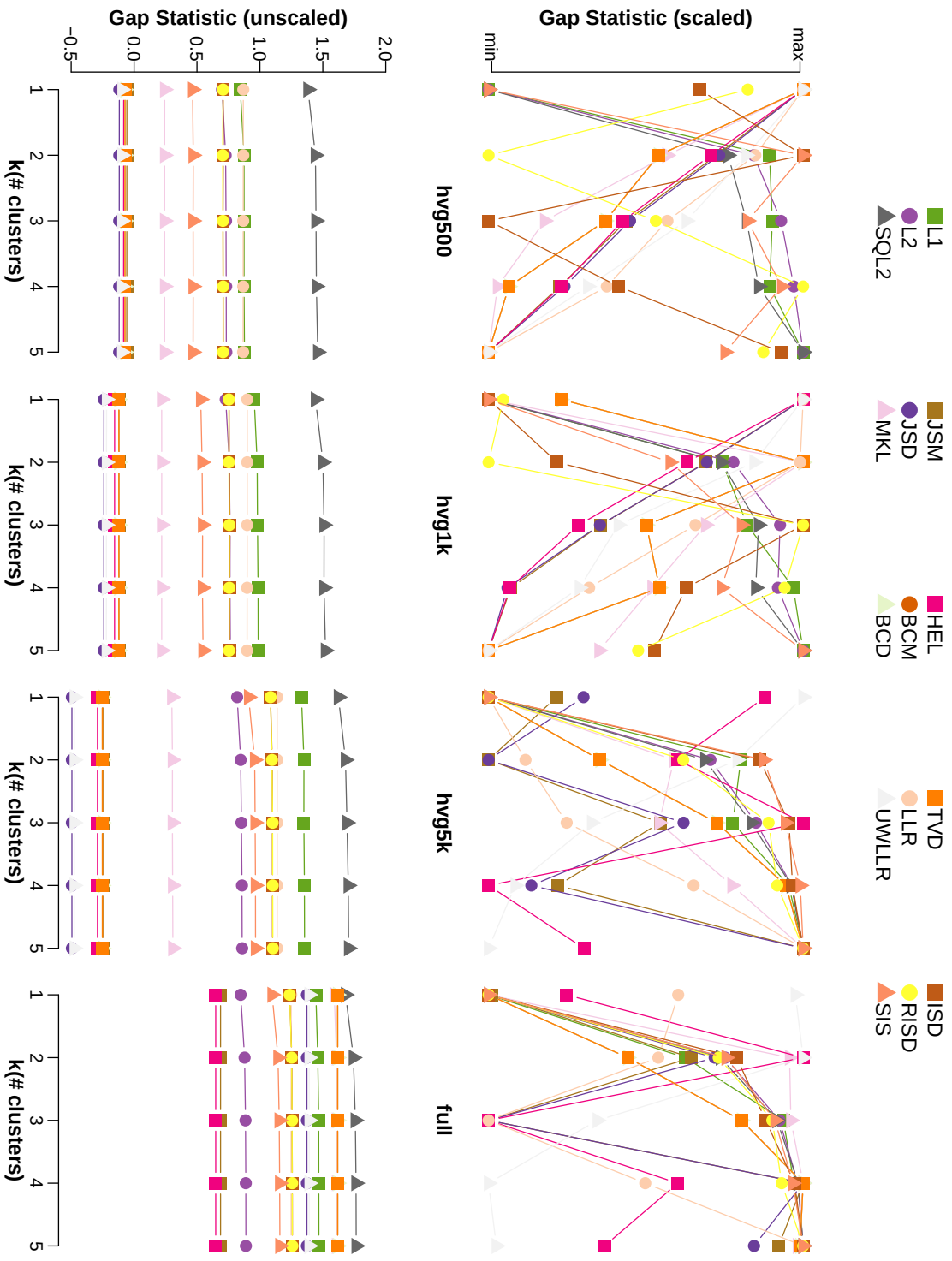


Figure S5: Gap statistic for each distance as a function of $k$ for the 293t only scRNA-seq dataset. Gap statistics scaled to (0,1) are given in the top row, and unscaled gap statistics are given in the bottom row. Columns from left to right indicate increasing numbers of HVGs used for distance calculation.

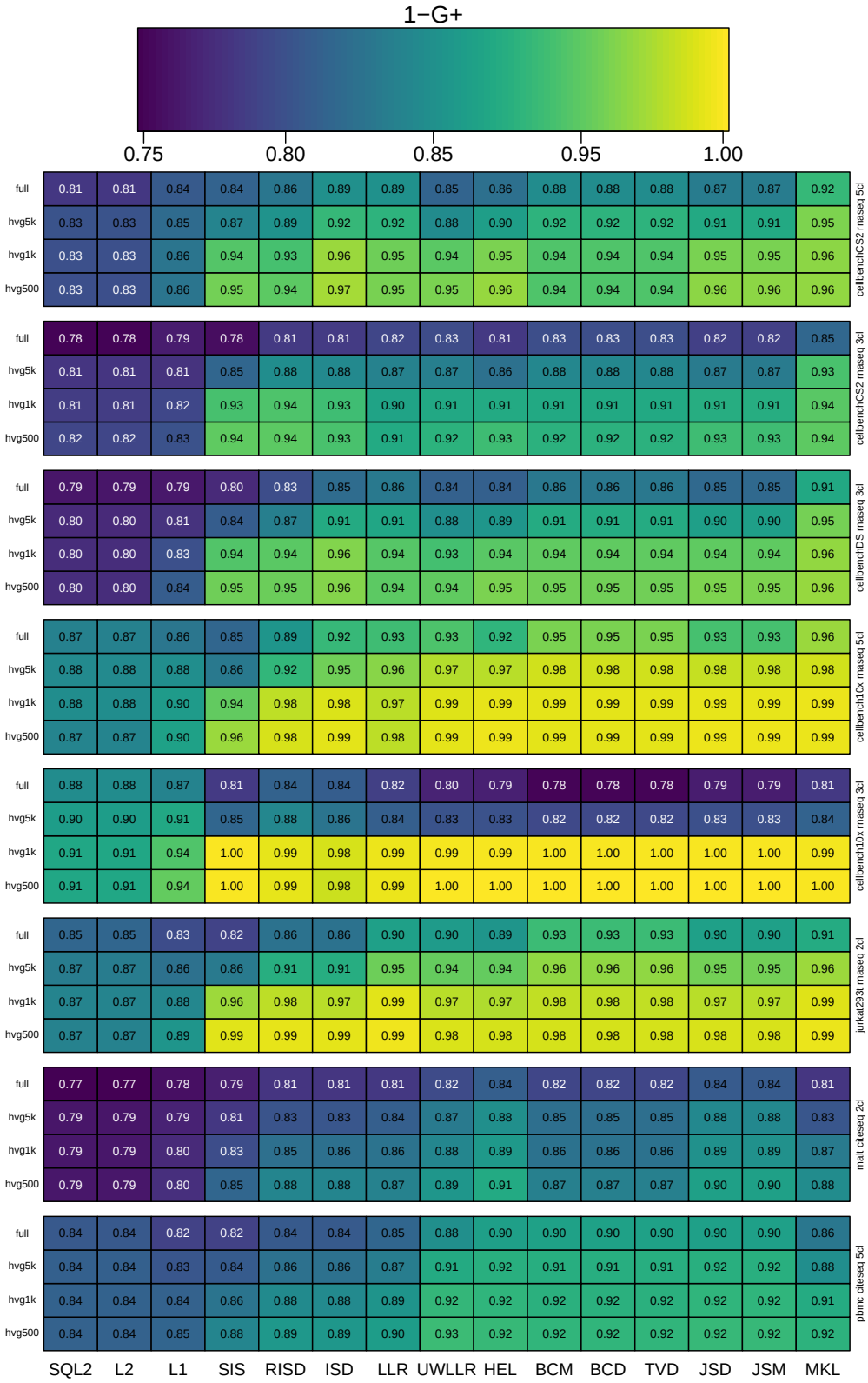


Figure S6: Heatmap of 1-G+ for all distances and scRNA-seq cluster datasets. Each test scRNA-seq dataset is grouped in a set of four rows, with number of HVGs used increasing from 500 to full.

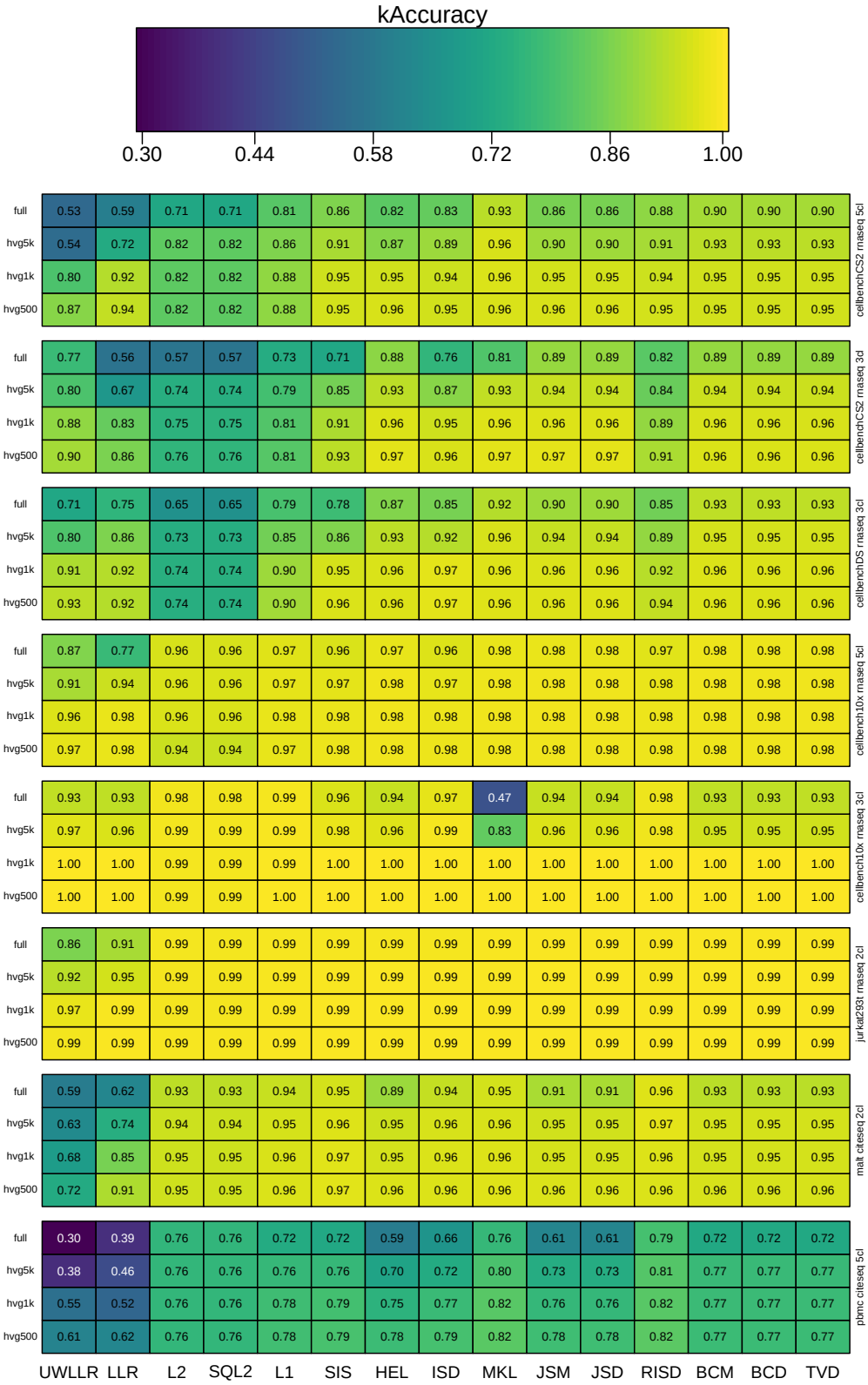


Figure S7: Heatmap of kAccuracy for all distances and scRNA-seq cluster datasets. Each test scRNA-seq dataset is grouped in a set of four rows, with number of HVGs used increasing from 500 to full.

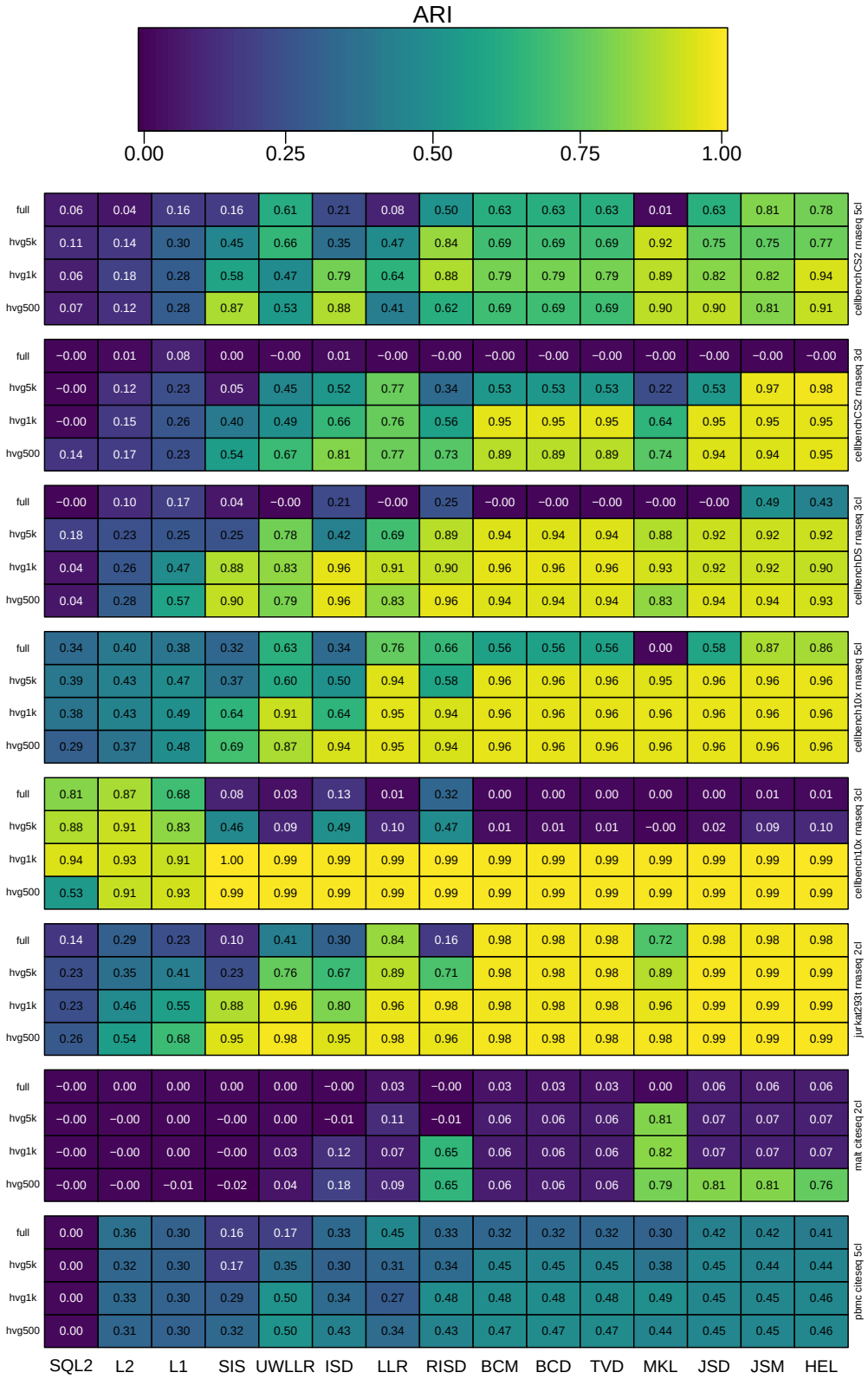


Figure S8: Heatmap of ARI for all distances and scRNA-seq cluster datasets. Each test scRNA-seq dataset is grouped in a set of four rows, with number of HVGs used increasing from 500 to full.

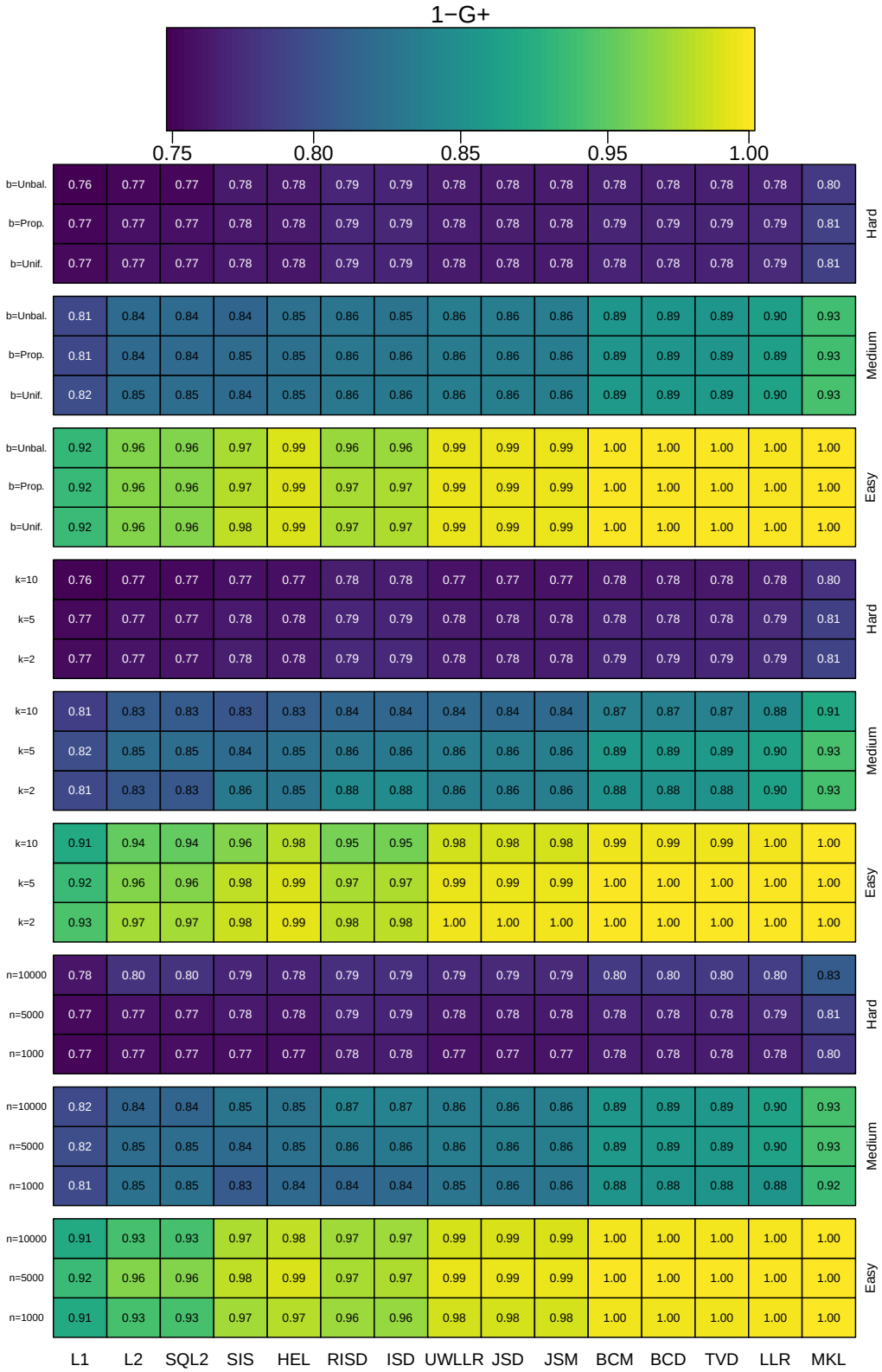


Figure S9: Heatmap of 1-G+ for all distances and simulated cluster datasets. Each set of three rows compares performance accross varying celltype balance ($b$, top nine rows), number of clusters ($k$, middle nine rows), or nubmer of cells ($n$, bottom nine rows). Within each set of nine rows are groups of three rows for easy, medium, and hard simulation settings.

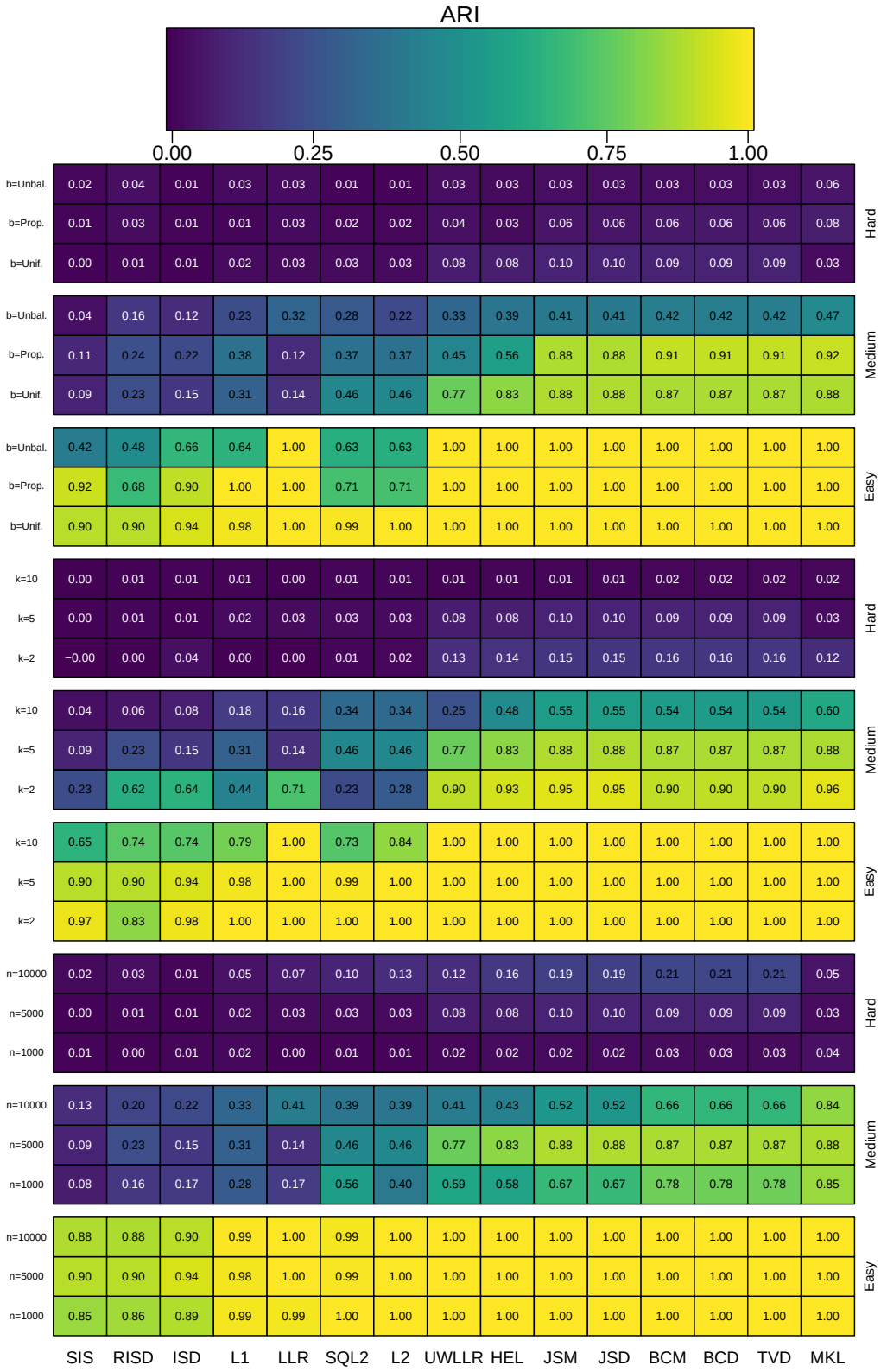


Figure S10: Heatmap of ARI for all distances and simulated cluster datasets. Each set of three rows compares performance accross varying celltype balance ($b$, top nine rows), number of clusters ($k$, middle nine rows), or nubmer of cells ($n$, bottom nine rows). Within each set of nine rows are groups of three rows for easy, medium, and hard simulation settings.

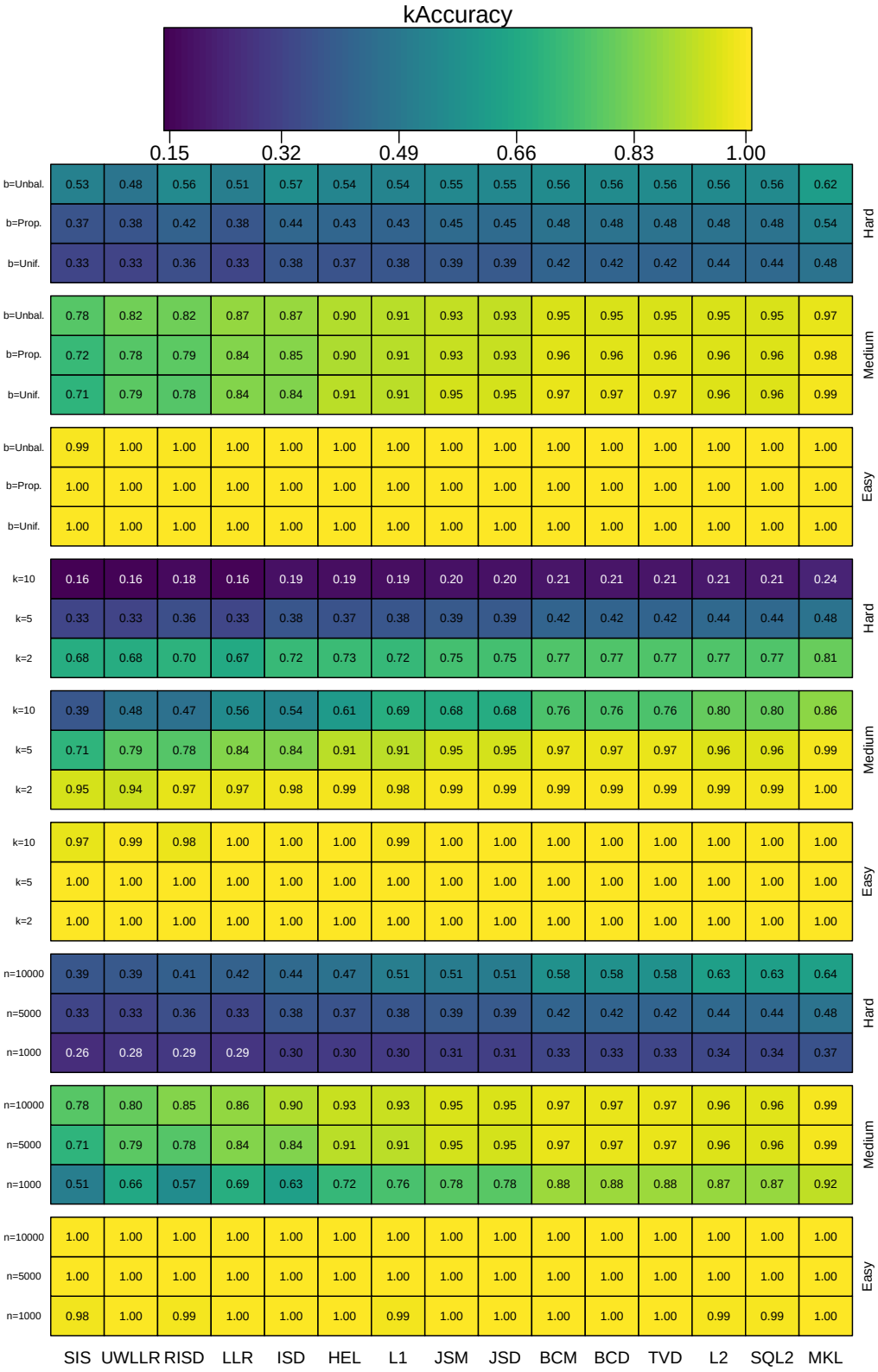


Figure S11: Heatmap of kAccuracy for all distances and simulated cluster datasets. Each set of three rows compares performance accross varying celltype balance ($b$, top nine rows), number of clusters ($k$, middle nine rows), or nubmer of cells ($n$, bottom nine rows). Within each set of nine rows are groups of three rows for easy, medium, and hard simulation settings.

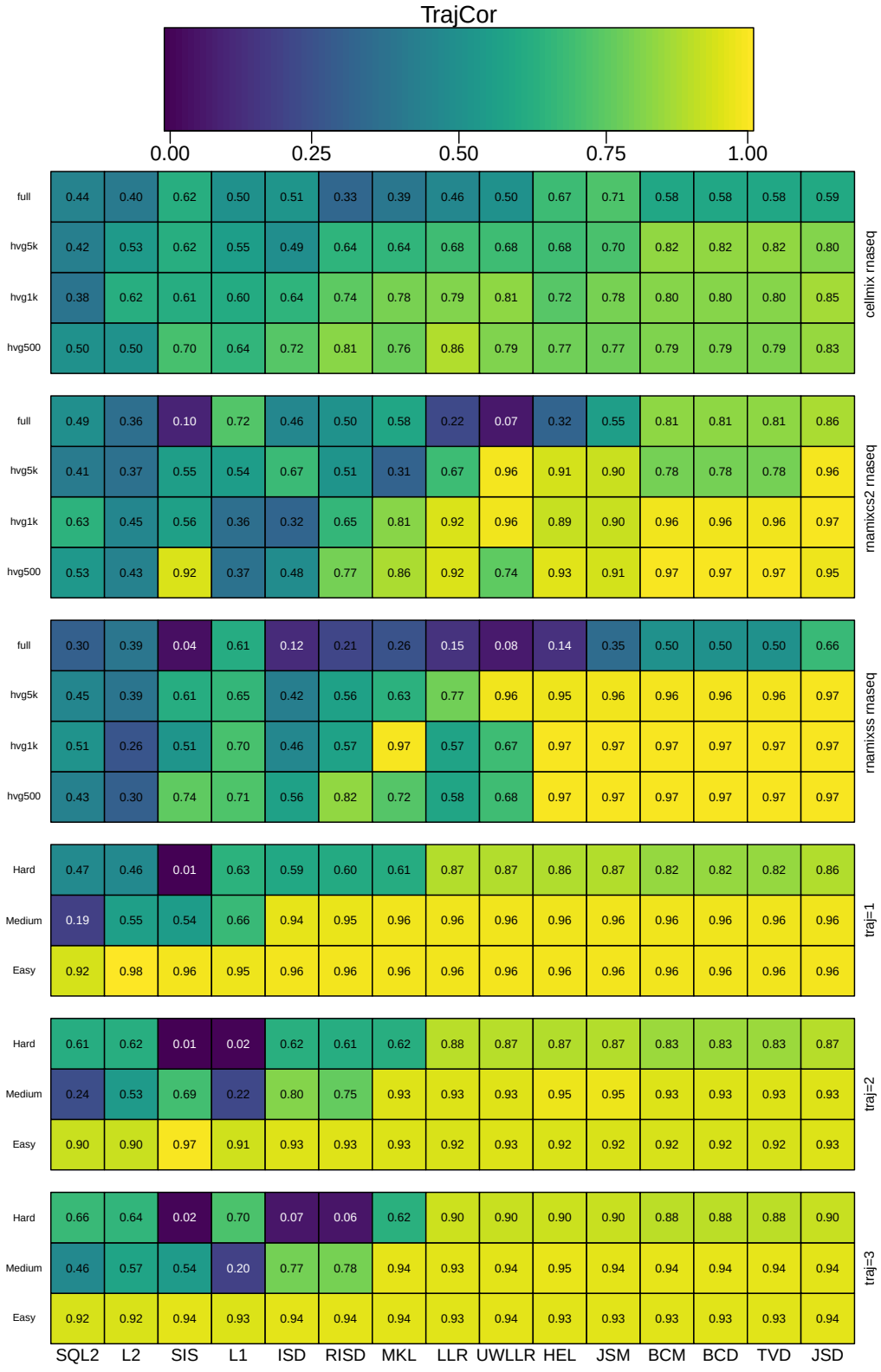


Figure S12: Trajectory correlation results for all distances and scRNA-seq (top twelve rows) and simulated (bottom nine rows) datasets. Each set of four rows within the scRNA-seq datasets represents varying amounts of HVGs for the same trajectory dataset. Each set of three rows with the simulated datasets represents varying levels of simulation difficulty.

Table S1: Corresponding scRNA-seq Gap statistic data for Figure S4

| | L1 | L2 | SQL2 | JSM | JSD | MKL | HEL | BCM | BCD | TVD | LLR | UWLLR | ISD | RISD | SIS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hvg500_k=1 | 1.06 | 1.07 | 1.87 | 0.24 | 0.46 | 0.92 | 0.21 | 0.33 | 0.33 | 0.33 | 1.61 | 0.54 | 1.09 | 1.09 | 0.73 |
| hvg500_k=2 | 1.24 | 1.29 | 2.40 | 0.25 | 0.48 | 0.97 | 0.21 | 0.35 | 0.35 | 0.35 | 1.62 | 0.56 | 1.08 | 1.09 | 0.78 |
| hvg500_k=3 | 1.31 | 1.39 | 2.65 | 0.25 | 0.49 | 0.97 | 0.22 | 0.36 | 0.36 | 0.36 | 1.64 | 0.57 | 1.11 | 1.14 | 0.78 |
| hvg500_k=4 | 1.35 | 1.44 | 2.76 | 0.26 | 0.50 | 0.97 | 0.22 | 0.36 | 0.36 | 0.36 | 1.63 | 0.57 | 1.12 | 1.15 | 0.80 |
| hvg500_k=5 | 1.36 | 1.45 | 2.81 | 0.26 | 0.50 | 0.99 | 0.22 | 0.36 | 0.36 | 0.36 | 1.63 | 0.57 | 1.13 | 1.15 | 0.80 |
| hvg1k_k=1 | 1.04 | 1.05 | 1.82 | 0.18 | 0.35 | 0.88 | 0.14 | 0.29 | 0.29 | 0.29 | 1.49 | 0.43 | 1.02 | 1.03 | 0.68 |
| hvg1k_k=2 | 1.21 | 1.29 | 2.41 | 0.19 | 0.37 | 0.92 | 0.15 | 0.30 | 0.30 | 0.30 | 1.50 | 0.44 | 1.08 | 1.07 | 0.76 |
| hvg1k_k=3 | 1.28 | 1.40 | 2.68 | 0.19 | 0.37 | 0.93 | 0.15 | 0.31 | 0.31 | 0.31 | 1.50 | 0.45 | 1.08 | 1.09 | 0.80 |
| hvg1k_k=4 | 1.31 | 1.45 | 2.79 | 0.20 | 0.38 | 0.94 | 0.16 | 0.32 | 0.32 | 0.32 | 1.51 | 0.45 | 1.09 | 1.11 | 0.80 |
| hvg1k_k=5 | 1.33 | 1.47 | 2.84 | 0.20 | 0.38 | 0.95 | 0.16 | 0.32 | 0.32 | 0.32 | 1.51 | 0.45 | 1.09 | 1.10 | 0.80 |
| hvg5k_k=1 | 1.19 | 1.04 | 1.80 | 0.08 | 0.15 | 0.85 | 0.04 | 0.18 | 0.18 | 0.18 | 1.47 | 0.21 | 1.12 | 1.13 | 0.77 |
| hvg5k_k=2 | 1.34 | 1.29 | 2.40 | 0.08 | 0.16 | 0.95 | 0.05 | 0.19 | 0.19 | 0.19 | 1.47 | 0.22 | 1.27 | 1.27 | 1.02 |
| hvg5k_k=3 | 1.39 | 1.39 | 2.68 | 0.09 | 0.16 | 0.99 | 0.05 | 0.20 | 0.20 | 0.20 | 1.47 | 0.22 | 1.31 | 1.31 | 1.11 |
| hvg5k_k=4 | 1.42 | 1.45 | 2.81 | 0.09 | 0.16 | 1.00 | 0.05 | 0.20 | 0.20 | 0.20 | 1.47 | 0.22 | 1.32 | 1.33 | 1.15 |
| hvg5k_k=5 | 1.43 | 1.48 | 2.86 | 0.09 | 0.16 | 1.01 | 0.05 | 0.20 | 0.20 | 0.20 | 1.47 | 0.22 | 1.34 | 1.34 | 1.17 |
| full_k=1 | 1.21 | 0.98 | 1.68 | 0.09 | 0.18 | 0.91 | 0.07 | 0.18 | 0.18 | 0.18 | 1.48 | 0.23 | 1.19 | 1.19 | 0.88 |
| full_k=2 | 1.34 | 1.19 | 2.20 | 0.10 | 0.18 | 1.03 | 0.07 | 0.20 | 0.20 | 0.20 | 1.48 | 0.24 | 1.32 | 1.32 | 1.12 |
| full_k=3 | 1.39 | 1.31 | 2.50 | 0.10 | 0.19 | 1.07 | 0.07 | 0.20 | 0.20 | 0.20 | 1.48 | 0.24 | 1.36 | 1.36 | 1.20 |
| full_k=4 | 1.42 | 1.38 | 2.66 | 0.10 | 0.19 | 1.09 | 0.07 | 0.20 | 0.20 | 0.20 | 1.48 | 0.24 | 1.37 | 1.37 | 1.24 |
| full_k=5 | 1.43 | 1.42 | 2.71 | 0.10 | 0.19 | 1.10 | 0.07 | 0.20 | 0.20 | 0.20 | 1.48 | 0.24 | 1.38 | 1.38 | 1.26 |

Table S2: Corresponding simulated Gap statistic data for Figure S5

| | L1 | L2 | SQL2 | JSM | JSD | MKL | HEL | BCM | BCD | TVD | LLR | UWLLR | ISD | RISD | SIS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| hvg500_k=1 | 0.84 | 0.70 | 1.38 | -0.06 | -0.12 | 0.24 | -0.08 | -0.07 | -0.07 | -0.07 | 0.87 | -0.10 | 0.71 | 0.71 | 0.47 |
| hvg500_k=2 | 0.87 | 0.73 | 1.44 | -0.06 | -0.12 | 0.24 | -0.08 | -0.07 | -0.07 | -0.07 | 0.87 | -0.10 | 0.71 | 0.70 | 0.47 |
| hvg500_k=3 | 0.87 | 0.73 | 1.45 | -0.06 | -0.12 | 0.24 | -0.08 | -0.07 | -0.07 | -0.07 | 0.87 | -0.10 | 0.71 | 0.71 | 0.47 |
| hvg500_k=4 | 0.87 | 0.73 | 1.45 | -0.06 | -0.12 | 0.24 | -0.08 | -0.07 | -0.07 | -0.07 | 0.87 | -0.10 | 0.71 | 0.71 | 0.47 |
| hvg500_k=5 | 0.88 | 0.73 | 1.46 | -0.06 | -0.12 | 0.24 | -0.08 | -0.07 | -0.07 | -0.07 | 0.87 | -0.10 | 0.71 | 0.71 | 0.47 |
| hvg1k_k=1 | 0.95 | 0.73 | 1.44 | -0.12 | -0.24 | 0.22 | -0.15 | -0.12 | -0.12 | -0.12 | 0.90 | -0.22 | 0.76 | 0.75 | 0.53 |
| hvg1k_k=2 | 0.98 | 0.76 | 1.50 | -0.12 | -0.24 | 0.22 | -0.15 | -0.12 | -0.12 | -0.12 | 0.90 | -0.22 | 0.76 | 0.75 | 0.54 |
| hvg1k_k=3 | 0.98 | 0.76 | 1.51 | -0.12 | -0.24 | 0.22 | -0.15 | -0.12 | -0.12 | -0.12 | 0.90 | -0.22 | 0.76 | 0.76 | 0.54 |
| hvg1k_k=4 | 0.99 | 0.76 | 1.51 | -0.12 | -0.24 | 0.22 | -0.15 | -0.12 | -0.12 | -0.12 | 0.90 | -0.22 | 0.76 | 0.76 | 0.54 |
| hvg1k_k=5 | 0.99 | 0.77 | 1.52 | -0.12 | -0.24 | 0.22 | -0.15 | -0.12 | -0.12 | -0.12 | 0.90 | -0.22 | 0.76 | 0.76 | 0.55 |
| hvg5k_k=1 | 1.33 | 0.82 | 1.62 | -0.25 | -0.49 | 0.30 | -0.29 | -0.25 | -0.25 | -0.25 | 1.14 | -0.48 | 1.08 | 1.08 | 0.91 |
| hvg5k_k=2 | 1.35 | 0.85 | 1.68 | -0.25 | -0.49 | 0.30 | -0.29 | -0.25 | -0.25 | -0.25 | 1.14 | -0.48 | 1.10 | 1.10 | 0.96 |
| hvg5k_k=3 | 1.35 | 0.85 | 1.69 | -0.25 | -0.49 | 0.30 | -0.29 | -0.25 | -0.25 | -0.25 | 1.14 | -0.48 | 1.10 | 1.10 | 0.96 |
| hvg5k_k=4 | 1.35 | 0.86 | 1.70 | -0.25 | -0.49 | 0.31 | -0.29 | -0.25 | -0.25 | -0.25 | 1.14 | -0.48 | 1.10 | 1.10 | 0.96 |
| hvg5k_k=5 | 1.35 | 0.86 | 1.71 | -0.25 | -0.49 | 0.31 | -0.29 | -0.25 | -0.25 | -0.25 | 1.14 | -0.48 | 1.10 | 1.10 | 0.97 |
| full_k=1 | 1.44 | 0.85 | 1.68 | 0.69 | 1.37 | 1.59 | 0.65 | 1.62 | 1.62 | 1.62 | 1.25 | 1.39 | 1.24 | 1.24 | 1.10 |
| full_k=2 | 1.46 | 0.88 | 1.74 | 0.69 | 1.37 | 1.61 | 0.65 | 1.62 | 1.62 | 1.62 | 1.25 | 1.39 | 1.25 | 1.25 | 1.14 |
| full_k=3 | 1.47 | 0.89 | 1.76 | 0.69 | 1.37 | 1.61 | 0.65 | 1.62 | 1.62 | 1.62 | 1.25 | 1.39 | 1.26 | 1.26 | 1.15 |
| full_k=4 | 1.47 | 0.89 | 1.76 | 0.69 | 1.37 | 1.61 | 0.65 | 1.62 | 1.62 | 1.62 | 1.25 | 1.39 | 1.26 | 1.26 | 1.16 |
| full_k=5 | 1.47 | 0.89 | 1.77 | 0.69 | 1.37 | 1.61 | 0.65 | 1.62 | 1.62 | 1.62 | 1.25 | 1.39 | 1.26 | 1.26 | 1.16 |

Table S3: Corresponding scRNA-seq 1-G+ data for Figure S6

| | SQL2 | L2 | L1 | SIS | RISD | ISD | LLR | UWLLR | HEL | BCM | BCD | TVD | JSD | JSM | MKL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pbmc_citeseq_5cl_hvg500 | 0.84 | 0.84 | 0.85 | 0.88 | 0.89 | 0.89 | 0.90 | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 |
| pbmc_citeseq_5cl_hvg1k | 0.84 | 0.84 | 0.84 | 0.86 | 0.88 | 0.88 | 0.89 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.91 |
| pbmc_citeseq_5cl_hvg5k | 0.84 | 0.84 | 0.83 | 0.84 | 0.86 | 0.86 | 0.87 | 0.91 | 0.92 | 0.91 | 0.91 | 0.91 | 0.92 | 0.92 | 0.88 |
| pbmc_citeseq_5cl_full | 0.84 | 0.84 | 0.82 | 0.82 | 0.84 | 0.84 | 0.85 | 0.88 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 | 0.86 |
| malt_citeseq_2cl_hvg500 | 0.79 | 0.79 | 0.80 | 0.85 | 0.88 | 0.88 | 0.87 | 0.89 | 0.91 | 0.87 | 0.87 | 0.90 | 0.90 | 0.90 | 0.88 |
| malt_citeseq_2cl_hvg1k | 0.79 | 0.79 | 0.80 | 0.83 | 0.85 | 0.86 | 0.86 | 0.88 | 0.89 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.87 |
| malt_citeseq_2cl_hvg5k | 0.79 | 0.79 | 0.79 | 0.81 | 0.83 | 0.83 | 0.84 | 0.87 | 0.88 | 0.85 | 0.85 | 0.88 | 0.88 | 0.88 | 0.83 |
| malt_citeseq_2cl_full | 0.77 | 0.77 | 0.78 | 0.79 | 0.81 | 0.81 | 0.81 | 0.82 | 0.84 | 0.82 | 0.82 | 0.84 | 0.84 | 0.84 | 0.81 |
| jurkat293t_rnaseq_2cl_hvg500 | 0.87 | 0.87 | 0.89 | 0.99 | 0.99 | 0.99 | 0.99 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg1k | 0.87 | 0.87 | 0.88 | 0.96 | 0.98 | 0.97 | 0.99 | 0.97 | 0.97 | 0.98 | 0.98 | 0.98 | 0.97 | 0.97 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg5k | 0.87 | 0.87 | 0.86 | 0.86 | 0.91 | 0.91 | 0.95 | 0.94 | 0.94 | 0.96 | 0.96 | 0.96 | 0.95 | 0.95 | 0.96 |
| jurkat293t_rnaseq_2cl_full | 0.85 | 0.85 | 0.83 | 0.82 | 0.86 | 0.86 | 0.90 | 0.90 | 0.89 | 0.93 | 0.93 | 0.93 | 0.90 | 0.90 | 0.91 |
| cellbench10x_rnaseq_3cl_hvg500 | 0.91 | 0.91 | 0.94 | 1.00 | 0.99 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| cellbench10x_rnaseq_3cl_hvg1k | 0.91 | 0.91 | 0.94 | 1.00 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 |
| cellbench10x_rnaseq_3cl_hvg5k | 0.90 | 0.90 | 0.91 | 0.85 | 0.88 | 0.86 | 0.84 | 0.83 | 0.83 | 0.82 | 0.82 | 0.82 | 0.83 | 0.83 | 0.84 |
| cellbench10x_rnaseq_3cl_full | 0.88 | 0.88 | 0.87 | 0.81 | 0.84 | 0.84 | 0.82 | 0.80 | 0.79 | 0.78 | 0.78 | 0.78 | 0.79 | 0.79 | 0.81 |
| cellbench10x_rnaseq_5cl_hvg500 | 0.87 | 0.87 | 0.90 | 0.96 | 0.98 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| cellbench10x_rnaseq_5cl_hvg1k | 0.88 | 0.88 | 0.90 | 0.94 | 0.98 | 0.98 | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| cellbench10x_rnaseq_5cl_hvg5k | 0.88 | 0.88 | 0.88 | 0.86 | 0.92 | 0.95 | 0.96 | 0.97 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| cellbench10x_rnaseq_5cl_full | 0.87 | 0.87 | 0.86 | 0.85 | 0.89 | 0.92 | 0.93 | 0.93 | 0.92 | 0.95 | 0.95 | 0.95 | 0.93 | 0.93 | 0.96 |
| cellbenchDS_rnaseq_3cl_hvg500 | 0.80 | 0.80 | 0.84 | 0.95 | 0.95 | 0.96 | 0.94 | 0.94 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.96 |
| cellbenchDS_rnaseq_3cl_hvg1k | 0.80 | 0.80 | 0.83 | 0.94 | 0.94 | 0.96 | 0.94 | 0.93 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.96 |
| cellbenchDS_rnaseq_3cl_hvg5k | 0.80 | 0.80 | 0.81 | 0.84 | 0.87 | 0.91 | 0.91 | 0.88 | 0.89 | 0.91 | 0.91 | 0.91 | 0.90 | 0.90 | 0.95 |
| cellbenchDS_rnaseq_3cl_full | 0.79 | 0.79 | 0.79 | 0.80 | 0.83 | 0.85 | 0.86 | 0.84 | 0.84 | 0.86 | 0.86 | 0.86 | 0.85 | 0.85 | 0.91 |
| cellbenchCS2_rnaseq_3cl_hvg500 | 0.82 | 0.82 | 0.83 | 0.94 | 0.94 | 0.93 | 0.91 | 0.92 | 0.93 | 0.92 | 0.92 | 0.92 | 0.93 | 0.93 | 0.94 |
| cellbenchCS2_rnaseq_3cl_hvg1k | 0.81 | 0.81 | 0.82 | 0.93 | 0.94 | 0.93 | 0.90 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.94 |
| cellbenchCS2_rnaseq_3cl_hvg5k | 0.81 | 0.81 | 0.81 | 0.85 | 0.88 | 0.88 | 0.87 | 0.87 | 0.86 | 0.88 | 0.88 | 0.88 | 0.87 | 0.87 | 0.93 |
| cellbenchCS2_rnaseq_3cl_full | 0.78 | 0.78 | 0.79 | 0.78 | 0.81 | 0.81 | 0.82 | 0.83 | 0.81 | 0.83 | 0.83 | 0.83 | 0.82 | 0.82 | 0.85 |
| cellbenchCS2_rnaseq_5cl_hvg500 | 0.83 | 0.83 | 0.86 | 0.95 | 0.94 | 0.97 | 0.95 | 0.95 | 0.96 | 0.94 | 0.94 | 0.94 | 0.96 | 0.96 | 0.96 |
| cellbenchCS2_rnaseq_5cl_hvg1k | 0.83 | 0.83 | 0.86 | 0.94 | 0.93 | 0.96 | 0.95 | 0.94 | 0.95 | 0.94 | 0.94 | 0.94 | 0.95 | 0.95 | 0.96 |
| cellbenchCS2_rnaseq_5cl_hvg5k | 0.83 | 0.83 | 0.85 | 0.87 | 0.89 | 0.92 | 0.92 | 0.88 | 0.90 | 0.92 | 0.92 | 0.92 | 0.91 | 0.91 | 0.95 |
| cellbenchCS2_rnaseq_5cl_full | 0.81 | 0.81 | 0.84 | 0.84 | 0.86 | 0.89 | 0.89 | 0.85 | 0.86 | 0.88 | 0.88 | 0.88 | 0.87 | 0.87 | 0.92 |

Table S4: Corresponding scRNA-seq kAccuracy data for Figure S7

| | UWLLR | LLR | L2 | SQL2 | L1 | SIS | HEL | ISD | MKL | JSM | JSD | RISD | BCM | BCD | TVD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pbmc_citeseq_5cl_hvg500 | 0.61 | 0.62 | 0.76 | 0.76 | 0.78 | 0.79 | 0.78 | 0.79 | 0.82 | 0.78 | 0.78 | 0.82 | 0.77 | 0.77 | 0.77 |
| pbmc_citeseq_5cl_hvg1k | 0.55 | 0.52 | 0.76 | 0.76 | 0.78 | 0.79 | 0.75 | 0.77 | 0.82 | 0.76 | 0.76 | 0.82 | 0.77 | 0.77 | 0.77 |
| pbmc_citeseq_5cl_hvg5k | 0.38 | 0.46 | 0.76 | 0.76 | 0.76 | 0.76 | 0.70 | 0.72 | 0.80 | 0.73 | 0.73 | 0.81 | 0.77 | 0.77 | 0.77 |
| pbmc_citeseq_5cl_full | 0.30 | 0.39 | 0.76 | 0.76 | 0.72 | 0.72 | 0.59 | 0.66 | 0.76 | 0.61 | 0.61 | 0.79 | 0.72 | 0.72 | 0.72 |
| malt_citeseq_2cl_hvg500 | 0.72 | 0.91 | 0.95 | 0.95 | 0.96 | 0.97 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| malt_citeseq_2cl_hvg1k | 0.68 | 0.85 | 0.95 | 0.95 | 0.96 | 0.97 | 0.95 | 0.96 | 0.96 | 0.95 | 0.95 | 0.96 | 0.95 | 0.95 | 0.95 |
| malt_citeseq_2cl_hvg5k | 0.63 | 0.74 | 0.94 | 0.94 | 0.95 | 0.96 | 0.95 | 0.96 | 0.96 | 0.95 | 0.95 | 0.97 | 0.95 | 0.95 | 0.95 |
| malt_citeseq_2cl_full | 0.59 | 0.62 | 0.93 | 0.93 | 0.94 | 0.95 | 0.89 | 0.94 | 0.95 | 0.91 | 0.91 | 0.96 | 0.93 | 0.93 | 0.93 |
| jurkat293t_rnaseq_2cl_hvg500 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg1k | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg5k | 0.92 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_full | 0.86 | 0.91 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| cellbench10x_rnaseq_3cl_hvg500 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| cellbench10x_rnaseq_3cl_hvg1k | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| cellbench10x_rnaseq_3cl_hvg5k | 0.97 | 0.96 | 0.99 | 0.99 | 0.99 | 0.98 | 0.96 | 0.99 | 0.83 | 0.96 | 0.96 | 0.98 | 0.95 | 0.95 | 0.95 |
| cellbench10x_rnaseq_3cl_full | 0.93 | 0.93 | 0.98 | 0.98 | 0.99 | 0.96 | 0.94 | 0.97 | 0.47 | 0.94 | 0.94 | 0.98 | 0.93 | 0.93 | 0.93 |
| cellbench10x_rnaseq_5cl_hvg500 | 0.97 | 0.98 | 0.94 | 0.94 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| cellbench10x_rnaseq_5cl_hvg1k | 0.96 | 0.98 | 0.96 | 0.96 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| cellbench10x_rnaseq_5cl_hvg5k | 0.91 | 0.94 | 0.96 | 0.96 | 0.97 | 0.97 | 0.98 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| cellbench10x_rnaseq_5cl_full | 0.87 | 0.77 | 0.96 | 0.97 | 0.96 | 0.97 | 0.96 | 0.96 | 0.98 | 0.98 | 0.98 | 0.97 | 0.98 | 0.98 | 0.98 |
| cellbenchDS_rnaseq_3cl_hvg500 | 0.93 | 0.92 | 0.74 | 0.74 | 0.90 | 0.96 | 0.96 | 0.97 | 0.96 | 0.96 | 0.96 | 0.94 | 0.96 | 0.96 | 0.96 |
| cellbenchDS_rnaseq_3cl_hvg1k | 0.91 | 0.92 | 0.74 | 0.74 | 0.90 | 0.95 | 0.96 | 0.97 | 0.96 | 0.96 | 0.96 | 0.92 | 0.96 | 0.96 | 0.96 |
| cellbenchDS_rnaseq_3cl_hvg5k | 0.80 | 0.86 | 0.73 | 0.73 | 0.85 | 0.86 | 0.93 | 0.92 | 0.96 | 0.94 | 0.94 | 0.89 | 0.95 | 0.95 | 0.95 |
| cellbenchDS_rnaseq_3cl_full | 0.71 | 0.75 | 0.65 | 0.65 | 0.79 | 0.78 | 0.87 | 0.85 | 0.92 | 0.90 | 0.90 | 0.85 | 0.93 | 0.93 | 0.93 |
| cellbenchCS2_rnaseq_3cl_hvg500 | 0.90 | 0.86 | 0.76 | 0.76 | 0.81 | 0.93 | 0.97 | 0.96 | 0.97 | 0.97 | 0.97 | 0.91 | 0.96 | 0.96 | 0.96 |
| cellbenchCS2_rnaseq_3cl_hvg1k | 0.88 | 0.83 | 0.75 | 0.75 | 0.81 | 0.91 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.89 | 0.96 | 0.96 | 0.96 |
| cellbenchCS2_rnaseq_3cl_hvg5k | 0.80 | 0.67 | 0.74 | 0.74 | 0.79 | 0.85 | 0.93 | 0.87 | 0.93 | 0.94 | 0.94 | 0.84 | 0.94 | 0.94 | 0.94 |
| cellbenchCS2_rnaseq_3cl_full | 0.77 | 0.56 | 0.57 | 0.57 | 0.73 | 0.71 | 0.88 | 0.76 | 0.81 | 0.89 | 0.89 | 0.82 | 0.89 | 0.89 | 0.89 |
| cellbenchCS2_rnaseq_5cl_hvg500 | 0.87 | 0.94 | 0.82 | 0.82 | 0.88 | 0.95 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 |
| cellbenchCS2_rnaseq_5cl_hvg1k | 0.80 | 0.92 | 0.82 | 0.82 | 0.88 | 0.95 | 0.95 | 0.94 | 0.96 | 0.95 | 0.95 | 0.94 | 0.95 | 0.95 | 0.95 |
| cellbenchCS2_rnaseq_5cl_hvg5k | 0.54 | 0.72 | 0.82 | 0.82 | 0.86 | 0.91 | 0.87 | 0.89 | 0.96 | 0.90 | 0.90 | 0.91 | 0.93 | 0.93 | 0.93 |
| cellbenchCS2_rnaseq_5cl_full | 0.53 | 0.59 | 0.71 | 0.71 | 0.81 | 0.86 | 0.82 | 0.83 | 0.93 | 0.86 | 0.86 | 0.88 | 0.90 | 0.90 | 0.90 |

Table S5: Corresponding scRNA-seq ARI data for Figure S8

| | SQL2 | L2 | L1 | SIS | UWLLR | ISD | LLR | RISD | BCM | BCD | TVD | MKL | JSD | JSM | HEL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pbmc_citeseq_5cl_hvg500 | 0.00 | 0.31 | 0.30 | 0.32 | 0.50 | 0.43 | 0.34 | 0.43 | 0.47 | 0.47 | 0.44 | 0.45 | 0.45 | 0.45 | 0.46 |
| pbmc_citeseq_5cl_hvg1k | 0.00 | 0.33 | 0.30 | 0.29 | 0.50 | 0.34 | 0.27 | 0.48 | 0.48 | 0.48 | 0.49 | 0.45 | 0.45 | 0.45 | 0.46 |
| pbmc_citeseq_5cl_hvg5k | 0.00 | 0.32 | 0.30 | 0.17 | 0.35 | 0.30 | 0.31 | 0.34 | 0.45 | 0.45 | 0.45 | 0.38 | 0.45 | 0.44 | 0.44 |
| pbmc_citeseq_5cl_full | 0.00 | 0.36 | 0.30 | 0.16 | 0.17 | 0.33 | 0.45 | 0.33 | 0.32 | 0.32 | 0.32 | 0.30 | 0.42 | 0.42 | 0.41 |
| malt_citeseq_2cl_hvg500 | 0.00 | 0.00 | -0.01 | -0.02 | 0.04 | 0.18 | 0.09 | 0.65 | 0.06 | 0.06 | 0.06 | 0.79 | 0.81 | 0.81 | 0.76 |
| malt_citeseq_2cl_hvg1k | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.12 | 0.07 | 0.65 | 0.06 | 0.06 | 0.06 | 0.82 | 0.07 | 0.07 | 0.07 |
| malt_citeseq_2cl_hvg5k | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.01 | 0.11 | -0.01 | 0.06 | 0.06 | 0.06 | 0.81 | 0.07 | 0.07 | 0.07 |
| malt_citeseq_2cl_full | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.03 | 0.03 | 0.03 | 0.00 | 0.06 | 0.06 | 0.06 |
| jurkat293t_rnaseq_2cl_hvg500 | 0.26 | 0.54 | 0.68 | 0.95 | 0.98 | 0.95 | 0.98 | 0.96 | 0.98 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg1k | 0.23 | 0.46 | 0.55 | 0.88 | 0.96 | 0.80 | 0.96 | 0.98 | 0.98 | 0.98 | 0.98 | 0.96 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_hvg5k | 0.23 | 0.35 | 0.41 | 0.23 | 0.76 | 0.67 | 0.89 | 0.71 | 0.98 | 0.98 | 0.98 | 0.89 | 0.99 | 0.99 | 0.99 |
| jurkat293t_rnaseq_2cl_full | 0.14 | 0.29 | 0.23 | 0.10 | 0.41 | 0.30 | 0.84 | 0.16 | 0.98 | 0.98 | 0.98 | 0.72 | 0.98 | 0.98 | 0.98 |
| cellbench10x_rnaseq_3cl_hvg500 | 0.53 | 0.91 | 0.93 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| cellbench10x_rnaseq_3cl_hvg1k | 0.94 | 0.93 | 0.91 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| cellbench10x_rnaseq_3cl_hvg5k | 0.88 | 0.91 | 0.83 | 0.46 | 0.09 | 0.49 | 0.10 | 0.47 | 0.01 | 0.01 | 0.00 | 0.02 | 0.09 | 0.10 | |
| cellbench10x_rnaseq_3cl_full | 0.81 | 0.87 | 0.68 | 0.08 | 0.03 | 0.13 | 0.01 | 0.32 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | |
| cellbench10x_rnaseq_5cl_hvg500 | 0.29 | 0.37 | 0.48 | 0.69 | 0.87 | 0.94 | 0.95 | 0.94 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| cellbench10x_rnaseq_5cl_hvg1k | 0.38 | 0.43 | 0.49 | 0.64 | 0.91 | 0.64 | 0.95 | 0.94 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| cellbench10x_rnaseq_5cl_hvg5k | 0.39 | 0.43 | 0.47 | 0.37 | 0.60 | 0.50 | 0.94 | 0.58 | 0.96 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 |
| cellbench10x_rnaseq_5cl_full | 0.34 | 0.40 | 0.38 | 0.32 | 0.63 | 0.34 | 0.76 | 0.66 | 0.56 | 0.56 | 0.00 | 0.58 | 0.87 | 0.86 | |
| cellbenchDS_rnaseq_3cl_hvg500 | 0.04 | 0.28 | 0.57 | 0.90 | 0.79 | 0.96 | 0.83 | 0.96 | 0.94 | 0.94 | 0.83 | 0.94 | 0.94 | 0.93 | |
| cellbenchDS_rnaseq_3cl_hvg1k | 0.04 | 0.26 | 0.47 | 0.88 | 0.83 | 0.96 | 0.91 | 0.90 | 0.96 | 0.96 | 0.93 | 0.92 | 0.92 | 0.90 | |
| cellbenchDS_rnaseq_3cl_hvg5k | 0.18 | 0.23 | 0.25 | 0.25 | 0.78 | 0.42 | 0.69 | 0.89 | 0.94 | 0.94 | 0.88 | 0.92 | 0.92 | 0.92 | |
| cellbenchDS_rnaseq_3cl_full | 0.00 | 0.10 | 0.17 | 0.04 | 0.00 | 0.21 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.49 | 0.43 | |
| cellbenchCS2_rnaseq_3cl_hvg500 | 0.14 | 0.17 | 0.23 | 0.54 | 0.67 | 0.81 | 0.77 | 0.73 | 0.89 | 0.89 | 0.74 | 0.94 | 0.94 | 0.95 | |
| cellbenchCS2_rnaseq_3cl_hvg1k | 0.00 | 0.15 | 0.26 | 0.40 | 0.49 | 0.66 | 0.76 | 0.56 | 0.95 | 0.95 | 0.64 | 0.95 | 0.95 | 0.95 | |
| cellbenchCS2_rnaseq_3cl_hvg5k | 0.00 | 0.12 | 0.23 | 0.05 | 0.45 | 0.52 | 0.77 | 0.34 | 0.53 | 0.53 | 0.22 | 0.53 | 0.97 | 0.98 | |
| cellbenchCS2_rnaseq_3cl_full | 0.00 | 0.01 | 0.08 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| cellbenchCS2_rnaseq_5cl_hvg500 | 0.07 | 0.12 | 0.28 | 0.87 | 0.53 | 0.88 | 0.41 | 0.62 | 0.69 | 0.69 | 0.90 | 0.90 | 0.81 | 0.91 | |
| cellbenchCS2_rnaseq_5cl_hvg1k | 0.06 | 0.18 | 0.28 | 0.58 | 0.47 | 0.79 | 0.64 | 0.88 | 0.79 | 0.79 | 0.89 | 0.82 | 0.82 | 0.94 | |
| cellbenchCS2_rnaseq_5cl_hvg5k | 0.11 | 0.14 | 0.30 | 0.45 | 0.66 | 0.35 | 0.47 | 0.84 | 0.69 | 0.69 | 0.92 | 0.75 | 0.75 | 0.77 | |
| cellbenchCS2_rnaseq_5cl_full | 0.06 | 0.04 | 0.16 | 0.16 | 0.61 | 0.21 | 0.08 | 0.50 | 0.63 | 0.63 | 0.01 | 0.63 | 0.81 | 0.78 | |

Table S6: Corresponding simulated 1-G+ data for Figure S9

| | L1 | L2 | SQL2 | SIS | HEL | RISD | ISD | UWLLR | JSD | JSM | BCM | BCD | TVD | LLR | MKL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| splatsim_n1_d1 | 0.91 | 0.93 | 0.93 | 0.97 | 0.97 | 0.96 | 0.96 | 0.98 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n2_d1 | 0.92 | 0.96 | 0.96 | 0.98 | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n3_d1 | 0.91 | 0.93 | 0.93 | 0.97 | 0.98 | 0.97 | 0.97 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n1_d2 | 0.81 | 0.85 | 0.85 | 0.83 | 0.84 | 0.84 | 0.84 | 0.85 | 0.86 | 0.86 | 0.88 | 0.88 | 0.88 | 0.88 | 0.92 |
| splatsim_n2_d2 | 0.82 | 0.85 | 0.85 | 0.84 | 0.85 | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.90 | 0.93 |
| splatsim_n3_d2 | 0.82 | 0.84 | 0.84 | 0.85 | 0.85 | 0.87 | 0.87 | 0.86 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.90 | 0.93 |
| splatsim_n1_d3 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.78 | 0.78 | 0.80 |
| splatsim_n2_d3 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.79 | 0.81 |
| splatsim_n3_d3 | 0.78 | 0.80 | 0.80 | 0.79 | 0.78 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 | 0.80 | 0.80 | 0.80 | 0.80 | 0.83 |
| splatsim_k1_d1 | 0.93 | 0.97 | 0.97 | 0.98 | 0.99 | 0.98 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k2_d1 | 0.92 | 0.96 | 0.96 | 0.98 | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k3_d1 | 0.91 | 0.94 | 0.94 | 0.96 | 0.98 | 0.95 | 0.95 | 0.98 | 0.98 | 0.98 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 |
| splatsim_k1_d2 | 0.81 | 0.83 | 0.83 | 0.86 | 0.85 | 0.88 | 0.88 | 0.86 | 0.86 | 0.86 | 0.88 | 0.88 | 0.88 | 0.90 | 0.93 |
| splatsim_k2_d2 | 0.82 | 0.85 | 0.85 | 0.84 | 0.85 | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.90 | 0.93 |
| splatsim_k3_d2 | 0.81 | 0.83 | 0.83 | 0.83 | 0.83 | 0.84 | 0.84 | 0.84 | 0.84 | 0.84 | 0.87 | 0.87 | 0.87 | 0.88 | 0.91 |
| splatsim_k1_d3 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.79 | 0.79 | 0.79 | 0.79 | 0.81 |
| splatsim_k2_d3 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.79 | 0.81 |
| splatsim_k3_d3 | 0.76 | 0.77 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.77 | 0.77 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.80 |
| splatsim_b1_d1 | 0.92 | 0.96 | 0.96 | 0.98 | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b2_d1 | 0.92 | 0.96 | 0.96 | 0.97 | 0.99 | 0.97 | 0.97 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b3_d1 | 0.92 | 0.96 | 0.96 | 0.97 | 0.99 | 0.96 | 0.96 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b1_d2 | 0.82 | 0.85 | 0.85 | 0.84 | 0.85 | 0.86 | 0.86 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.89 | 0.90 | 0.93 |
| splatsim_b2_d2 | 0.81 | 0.84 | 0.84 | 0.85 | 0.85 | 0.86 | 0.86 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.93 |
| splatsim_b3_d2 | 0.81 | 0.84 | 0.84 | 0.84 | 0.85 | 0.86 | 0.85 | 0.86 | 0.86 | 0.89 | 0.89 | 0.89 | 0.89 | 0.90 | 0.93 |
| splatsim_b1_d3 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.79 | 0.81 |
| splatsim_b2_d3 | 0.77 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.79 | 0.79 | 0.79 | 0.79 | 0.81 |
| splatsim_b3_d3 | 0.76 | 0.77 | 0.77 | 0.78 | 0.78 | 0.79 | 0.79 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.78 | 0.80 |

Table S7: Corresponding simulated ARI data for Figure S10

| | SIS | RISD | ISD | L1 | LLR | SQL2 | L2 | UWLLR | HEL | JSM | JSD | BCM | BCD | TVD | MKL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| splatsim_n1_d1 | 0.85 | 0.86 | 0.89 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n2_d1 | 0.90 | 0.90 | 0.94 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n3_d1 | 0.88 | 0.88 | 0.90 | 0.99 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n1_d2 | 0.08 | 0.16 | 0.17 | 0.28 | 0.17 | 0.56 | 0.40 | 0.59 | 0.58 | 0.67 | 0.67 | 0.78 | 0.78 | 0.78 | 0.85 |
| splatsim_n2_d2 | 0.09 | 0.23 | 0.15 | 0.31 | 0.14 | 0.46 | 0.46 | 0.77 | 0.83 | 0.88 | 0.88 | 0.87 | 0.87 | 0.87 | 0.88 |
| splatsim_n3_d2 | 0.13 | 0.20 | 0.22 | 0.33 | 0.41 | 0.39 | 0.39 | 0.41 | 0.43 | 0.52 | 0.52 | 0.66 | 0.66 | 0.66 | 0.84 |
| splatsim_n1_d3 | 0.01 | 0.00 | 0.01 | 0.02 | 0.00 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 | 0.03 | 0.04 |
| splatsim_n2_d3 | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.08 | 0.08 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.03 |
| splatsim_n3_d3 | 0.02 | 0.03 | 0.01 | 0.05 | 0.07 | 0.10 | 0.13 | 0.12 | 0.16 | 0.19 | 0.19 | 0.21 | 0.21 | 0.21 | 0.05 |
| splatsim_k1_d1 | 0.97 | 0.83 | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k2_d1 | 0.90 | 0.90 | 0.94 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k3_d1 | 0.65 | 0.74 | 0.74 | 0.79 | 1.00 | 0.73 | 0.84 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k1_d2 | 0.23 | 0.62 | 0.64 | 0.44 | 0.71 | 0.23 | 0.28 | 0.90 | 0.93 | 0.95 | 0.95 | 0.90 | 0.90 | 0.90 | 0.96 |
| splatsim_k2_d2 | 0.09 | 0.23 | 0.15 | 0.31 | 0.14 | 0.46 | 0.46 | 0.77 | 0.83 | 0.88 | 0.88 | 0.87 | 0.87 | 0.87 | 0.88 |
| splatsim_k3_d2 | 0.04 | 0.06 | 0.08 | 0.18 | 0.16 | 0.34 | 0.34 | 0.25 | 0.48 | 0.55 | 0.55 | 0.54 | 0.54 | 0.54 | 0.60 |
| splatsim_k1_d3 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.01 | 0.02 | 0.13 | 0.14 | 0.15 | 0.15 | 0.16 | 0.16 | 0.16 | 0.12 |
| splatsim_k2_d3 | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.08 | 0.08 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.03 |
| splatsim_k3_d3 | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 |
| splatsim_b1_d1 | 0.90 | 0.90 | 0.94 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b2_d1 | 0.92 | 0.68 | 0.90 | 1.00 | 1.00 | 0.71 | 0.71 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b3_d1 | 0.42 | 0.48 | 0.66 | 0.64 | 1.00 | 0.63 | 0.63 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b1_d2 | 0.09 | 0.23 | 0.15 | 0.31 | 0.14 | 0.46 | 0.46 | 0.77 | 0.83 | 0.88 | 0.88 | 0.87 | 0.87 | 0.87 | 0.88 |
| splatsim_b2_d2 | 0.11 | 0.24 | 0.22 | 0.38 | 0.12 | 0.37 | 0.37 | 0.45 | 0.56 | 0.88 | 0.88 | 0.91 | 0.91 | 0.91 | 0.92 |
| splatsim_b3_d2 | 0.04 | 0.16 | 0.12 | 0.23 | 0.32 | 0.28 | 0.22 | 0.33 | 0.39 | 0.41 | 0.41 | 0.42 | 0.42 | 0.42 | 0.47 |
| splatsim_b1_d3 | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.08 | 0.08 | 0.10 | 0.10 | 0.09 | 0.09 | 0.09 | 0.03 |
| splatsim_b2_d3 | 0.01 | 0.03 | 0.01 | 0.01 | 0.03 | 0.02 | 0.02 | 0.04 | 0.03 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.08 |
| splatsim_b3_d3 | 0.02 | 0.04 | 0.01 | 0.03 | 0.03 | 0.01 | 0.01 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.06 |

Table S8: Corresponding simulated kAccuracy data for Figure S11

| | SIS | UWLLR | RISD | LLR | ISD | HEL | L1 | JSM | JSD | BCM | BCD | TVD | L2 | SQL2 | MKL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| splatsim_n1_d1 | 0.98 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 |
| splatsim_n2_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n3_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_n1_d2 | 0.51 | 0.66 | 0.57 | 0.69 | 0.63 | 0.72 | 0.76 | 0.78 | 0.78 | 0.88 | 0.88 | 0.88 | 0.87 | 0.87 | 0.92 |
| splatsim_n2_d2 | 0.71 | 0.79 | 0.78 | 0.84 | 0.84 | 0.91 | 0.91 | 0.95 | 0.95 | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.99 |
| splatsim_n3_d2 | 0.78 | 0.80 | 0.85 | 0.86 | 0.90 | 0.93 | 0.93 | 0.95 | 0.95 | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.99 |
| splatsim_n1_d3 | 0.26 | 0.28 | 0.29 | 0.29 | 0.30 | 0.30 | 0.30 | 0.31 | 0.31 | 0.33 | 0.33 | 0.33 | 0.34 | 0.34 | 0.37 |
| splatsim_n2_d3 | 0.33 | 0.33 | 0.36 | 0.33 | 0.38 | 0.37 | 0.38 | 0.39 | 0.39 | 0.42 | 0.42 | 0.42 | 0.44 | 0.44 | 0.48 |
| splatsim_n3_d3 | 0.39 | 0.39 | 0.41 | 0.42 | 0.44 | 0.47 | 0.51 | 0.51 | 0.51 | 0.58 | 0.58 | 0.58 | 0.63 | 0.63 | 0.64 |
| splatsim_k1_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k2_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k3_d1 | 0.97 | 0.99 | 0.98 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_k1_d2 | 0.95 | 0.94 | 0.97 | 0.97 | 0.98 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 1.00 |
| splatsim_k2_d2 | 0.71 | 0.79 | 0.78 | 0.84 | 0.84 | 0.91 | 0.91 | 0.95 | 0.95 | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.99 |
| splatsim_k3_d2 | 0.39 | 0.48 | 0.47 | 0.56 | 0.54 | 0.61 | 0.69 | 0.68 | 0.68 | 0.76 | 0.76 | 0.76 | 0.80 | 0.80 | 0.86 |
| splatsim_k1_d3 | 0.68 | 0.68 | 0.70 | 0.67 | 0.72 | 0.73 | 0.72 | 0.75 | 0.75 | 0.77 | 0.77 | 0.77 | 0.77 | 0.77 | 0.81 |
| splatsim_k2_d3 | 0.33 | 0.33 | 0.36 | 0.33 | 0.38 | 0.37 | 0.38 | 0.39 | 0.39 | 0.42 | 0.42 | 0.42 | 0.44 | 0.44 | 0.48 |
| splatsim_k3_d3 | 0.16 | 0.16 | 0.18 | 0.16 | 0.19 | 0.19 | 0.19 | 0.20 | 0.20 | 0.21 | 0.21 | 0.21 | 0.21 | 0.21 | 0.24 |
| splatsim_b1_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b2_d1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b3_d1 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| splatsim_b1_d2 | 0.71 | 0.79 | 0.78 | 0.84 | 0.84 | 0.91 | 0.91 | 0.95 | 0.95 | 0.97 | 0.97 | 0.97 | 0.96 | 0.96 | 0.99 |
| splatsim_b2_d2 | 0.72 | 0.78 | 0.79 | 0.84 | 0.85 | 0.90 | 0.91 | 0.93 | 0.93 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.98 |
| splatsim_b3_d2 | 0.78 | 0.82 | 0.82 | 0.87 | 0.87 | 0.90 | 0.91 | 0.93 | 0.93 | 0.95 | 0.95 | 0.95 | 0.95 | 0.95 | 0.97 |
| splatsim_b1_d3 | 0.33 | 0.33 | 0.36 | 0.33 | 0.38 | 0.37 | 0.38 | 0.39 | 0.39 | 0.42 | 0.42 | 0.42 | 0.44 | 0.44 | 0.48 |
| splatsim_b2_d3 | 0.37 | 0.38 | 0.42 | 0.38 | 0.44 | 0.43 | 0.43 | 0.45 | 0.45 | 0.48 | 0.48 | 0.48 | 0.48 | 0.48 | 0.54 |
| splatsim_b3_d3 | 0.53 | 0.48 | 0.56 | 0.51 | 0.57 | 0.54 | 0.54 | 0.55 | 0.55 | 0.56 | 0.56 | 0.56 | 0.56 | 0.56 | 0.62 |

Table S9: Corresponding scRNA-seq and simulated TrajCor data for Figure S12

| | SQL2 | L2 | SIS | L1 | ISD | RISD | MKL | LLR | UWLLR | HEL | JSM | BCM | BCD | TVD | JSD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cellmix_traj_hvg500 | 0.50 | 0.50 | 0.70 | 0.64 | 0.72 | 0.81 | 0.76 | 0.86 | 0.79 | 0.77 | 0.77 | 0.79 | 0.79 | 0.79 | 0.83 |
| cellmix_traj_hvg1k | 0.38 | 0.62 | 0.61 | 0.60 | 0.64 | 0.74 | 0.78 | 0.79 | 0.81 | 0.72 | 0.78 | 0.80 | 0.80 | 0.80 | 0.85 |
| cellmix_traj_hvg5k | 0.42 | 0.53 | 0.62 | 0.55 | 0.49 | 0.64 | 0.64 | 0.68 | 0.68 | 0.68 | 0.70 | 0.82 | 0.82 | 0.82 | 0.80 |
| cellmix_traj_full | 0.44 | 0.40 | 0.62 | 0.50 | 0.51 | 0.33 | 0.39 | 0.46 | 0.50 | 0.67 | 0.71 | 0.58 | 0.58 | 0.58 | 0.59 |
| rnamixcs2_traj_hvg500 | 0.53 | 0.43 | 0.92 | 0.37 | 0.48 | 0.77 | 0.86 | 0.92 | 0.74 | 0.93 | 0.91 | 0.97 | 0.97 | 0.97 | 0.95 |
| rnamixcs2_traj_hvg1k | 0.63 | 0.45 | 0.56 | 0.36 | 0.32 | 0.65 | 0.81 | 0.92 | 0.96 | 0.89 | 0.90 | 0.96 | 0.96 | 0.96 | 0.97 |
| rnamixcs2_traj_hvg5k | 0.41 | 0.37 | 0.55 | 0.54 | 0.67 | 0.51 | 0.31 | 0.67 | 0.96 | 0.91 | 0.90 | 0.78 | 0.78 | 0.78 | 0.96 |
| rnamixcs2_traj_full | 0.49 | 0.36 | 0.10 | 0.72 | 0.46 | 0.50 | 0.58 | 0.22 | 0.07 | 0.32 | 0.55 | 0.81 | 0.81 | 0.81 | 0.86 |
| rnamixss_traj_hvg500 | 0.43 | 0.30 | 0.74 | 0.71 | 0.56 | 0.82 | 0.72 | 0.58 | 0.68 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| rnamixss_traj_hvg1k | 0.51 | 0.26 | 0.51 | 0.70 | 0.46 | 0.57 | 0.97 | 0.57 | 0.67 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 | 0.97 |
| rnamixss_traj_hvg5k | 0.45 | 0.39 | 0.61 | 0.65 | 0.42 | 0.56 | 0.63 | 0.77 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.97 |
| rnamixss_traj_full | 0.30 | 0.39 | 0.04 | 0.61 | 0.12 | 0.21 | 0.26 | 0.15 | 0.08 | 0.14 | 0.35 | 0.50 | 0.50 | 0.50 | 0.66 |
| splatsim_t1_d1 | 0.92 | 0.98 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| splatsim_t1_d2 | 0.19 | 0.55 | 0.54 | 0.66 | 0.94 | 0.95 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| splatsim_t1_d3 | 0.47 | 0.46 | 0.01 | 0.63 | 0.59 | 0.60 | 0.61 | 0.87 | 0.87 | 0.86 | 0.87 | 0.82 | 0.82 | 0.82 | 0.86 |
| splatsim_t2_d1 | 0.90 | 0.90 | 0.97 | 0.91 | 0.93 | 0.93 | 0.93 | 0.92 | 0.93 | 0.92 | 0.92 | 0.92 | 0.92 | 0.92 | 0.93 |
| splatsim_t2_d2 | 0.24 | 0.53 | 0.69 | 0.22 | 0.80 | 0.75 | 0.93 | 0.93 | 0.93 | 0.95 | 0.95 | 0.93 | 0.93 | 0.93 | 0.93 |
| splatsim_t2_d3 | 0.61 | 0.62 | 0.01 | 0.02 | 0.62 | 0.61 | 0.62 | 0.88 | 0.87 | 0.87 | 0.87 | 0.83 | 0.83 | 0.83 | 0.87 |
| splatsim_t3_d1 | 0.92 | 0.92 | 0.94 | 0.93 | 0.94 | 0.94 | 0.94 | 0.93 | 0.94 | 0.93 | 0.93 | 0.93 | 0.93 | 0.93 | 0.94 |
| splatsim_t3_d2 | 0.46 | 0.57 | 0.54 | 0.20 | 0.77 | 0.78 | 0.94 | 0.93 | 0.94 | 0.95 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 |
| splatsim_t3_d3 | 0.66 | 0.64 | 0.02 | 0.70 | 0.07 | 0.06 | 0.62 | 0.90 | 0.90 | 0.90 | 0.90 | 0.88 | 0.88 | 0.88 | 0.90 |

# References


[1] Fatih Ozsolak et al. “Direct RNA sequencing”. In: *Nature* 461.7265 (2009), pp. 814–818.

[2] Lina Ma, Vladimir B Bajic, and Zhang Zhang. “On the classification of long non-coding RNAs”. In: *RNA biology* 10.6 (2013), pp. 924–933.

[3] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature methods* 6.5 (2009), pp. 377–382.

[4] Jillian J Goetz and Jeffrey M Trimarchi. “Transcriptome sequencing of single cells with Smart-Seq”. In: *Nature biotechnology* 30.8 (2012), pp. 763–765.

[5] Grace X Y Zheng et al. “Massively parallel digital transcriptional profiling of single cells”. In: *Nat Commun* 8 (2017), p. 14049.

[6] Sven Schuierer et al. “A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples”. In: *BMC genomics* 18.1 (2017), pp. 1–13.

[7] Xiang-tao Huang et al. “Technical advances in single-cell RNA sequencing and applications in normal and malignant hematopoiesis”. In: *Frontiers in oncology* 8 (2018), p. 582.

[8] Stephanie C Hicks et al. “Missing data and technical variability in single-cell RNA-sequencing experiments”. In: *Biostatistics* 19.4 (2018), pp. 562–578.

[9] Katherine C Goldfarbmuren et al. “Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium”. In: *Nature communications* 11.1 (2020), pp. 1–21.

[10] Zhicheng Ji and Hongkai Ji. “TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis”. In: *Nucleic Acids Res* 44.13 (2016), e117.

[11] Alexander B Rosenberg et al. “Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding”. In: *Science* 360.6385 (2018), pp. 176–182.

[12] John Quackenbush. “Microarray data normalization and transformation”. In: *Nature genetics* 32.4 (2002), pp. 496–501.

[13] Nicholas Lytal, Di Ran, and Lingling An. “Normalization methods on single-cell RNA-seq data: an empirical survey”. In: *Frontiers in genetics* 11 (2020), p. 41.

[14] Daniela M Witten. “Classification and clustering of sequencing data using a Poisson model”. In: *The Annals of Applied Statistics* 5.4 (2011), pp. 2493–2518.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[16] Joseph B Kruskal. *Multidimensional scaling*. 11. Sage, 1978.

[17] F William Townes et al. “Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model”. In: *bioRxiv* (2019).

[18] Michael B Cole et al. “Performance assessment and selection of normalization procedures for single-cell RNA-seq”. In: *Cell systems* 8.4 (2019), pp. 315–328.

[19] Malte D Luecken and Fabian J Theis. “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Molecular systems biology* 15.6 (2019), e8746.

[20] Taiyun Kim et al. “Impact of similarity metrics on single-cell RNA-seq data clustering”. In: *Briefings in bioinformatics* 20.6 (2019), pp. 2316–2326.

[21] Catalina A Vallejos et al. “Normalizing single-cell RNA sequencing data: challenges and opportunities”. In: *Nature methods* 14.6 (2017), pp. 565–571.

[22] Geng Chen, Baitang Ning, and Tieliu Shi. “Single-cell RNA-seq technologies and related computational data analysis”. In: *Frontiers in genetics* 10 (2019), p. 317.

[23] Vanessa M Kvam, Peng Liu, and Yaqing Si. “A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data”. In: *American journal of botany* 99.2 (2012), pp. 248–256.

[24] Günter P Wagner, Koryu Kin, and Vincent J Lynch. “A model based criterion for gene expression calls using RNA-seq data”. In: *Theory in Biosciences* 132.3 (2013), pp. 159–164.

[25] Christoph Hafemeister and Rahul Satija. “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome biology* 20.1 (2019), pp. 1–15.

[26] Sariel Har-Peled and Soham Mazumdar. “On coresets for k-means and k-median clustering”. In: *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*. 2004, pp. 291–300.

[27] Philipp Berninger et al. “Computational analysis of small RNA cloning data”. In: *Methods (San Diego, Calif.)* 44 (Feb. 2008), pp. 13–21.

[28] Daniel N Baker et al. “Fast and memory-efficient scRNA-seq k-means clustering with various distances”. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2021, pp. 1–8.

[29] Guido Van Rossum. *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021.

[31] Kevin Ushey, JJ Allaire, and Yuan Tang. *reticulate: Interface to 'Python'*. R package version 1.20. 2021.

[32] Dirk Eddelbuettel and Wush Wu. “RcppCNPy: Read-Write Support for NumPy Files in R”. In: *Journal of Open Source Software* 1 (5 Sept. 2016).

[33] Luke Zappia, Belinda Phipson, and Alicia Oshlack. “Splatter: simulation of single-cell RNA sequencing data”. In: *Genome Biol* 18.1 (2017), p. 174. DOI: 10.1186/s13059-017-1305-0.

[34] Luyi Tian et al. “Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments”. In: *Nat Methods* 16.6 (June 2019), pp. 479–487.

[35] Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. In: *Nature methods* 14.9 (2017), pp. 865–868.

[36] Robert A Amezquita et al. “Orchestrating single-cell analysis with Bioconductor”. In: *Nature Methods* (2019), pp. 1–9.

[37] Ariel A. Hippen et al. “miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data”. In: *bioRxiv* (2021).

[38] Davis J McCarthy et al. “Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R”. In: *Bioinformatics* 33.8 (2017), pp. 1179–1186.

[39] SK Sanders, PA Giblin, and P Kavathas. “Cell-cell adhesion mediated by CD8 and human histocompatibility leukocyte antigen G, a nonclassical major histocompatibility complex class 1 molecule on cytotrophoblasts.” In: *The Journal of experimental medicine* 174.3 (1991), pp. 737–740.

[40] HWL Ziegler-Heitbrock. “Definition of human blood monocytes”. In: *Journal of leukocyte biology* 67.5 (2000), pp. 603–606.

[41] Ofer Mandelboim et al. “Human CD16 as a lysis receptor mediating direct natural killer cell cytotoxicity”. In: *Proceedings of the National Academy of Sciences* 96.10 (1999), pp. 5640–5644.

[42] Dennis C Otero, Amy N Anzelon, and Robert C Rickert. “CD19 function in early and late B cell development: I. Maintenance of follicular and marginal zone B cells requires CD19-dependent survival signals”. In: *The Journal of Immunology* 170.1 (2003), pp. 73–83.

[43] Christopher A Smith et al. “Antibodies to CD3/T-cell receptor complex induce death by apoptosis in immature T cells in thymic cultures”. In: *Nature* 337.6203 (1989), pp. 181–184.

[44] Derk Amsen et al. “Instruction of distinct CD4 T helper cell fates by different notch ligands on antigen-presenting cells”. In: *Cell* 117.4 (2004), pp. 515–526.

[45] Robert Tibshirani, Guenther Walther, and Trevor Hastie. “Estimating the number of clusters in a data set via the gap statistic”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

[46] Leonard Kaufman and Peter J Rousseeuw. “Partitioning around medoids (program pam)”. In: *Finding groups in data: an introduction to cluster analysis* 344 (1990), pp. 68–125.

[47] Martin Maechler et al. *cluster: Cluster Analysis Basics and Extensions*. 2021.

[48] Leo A Goodman and William H Kruskal. "Measures of association for cross classifications. II: Further discussion and references". In: *Journal of the American Statistical Association* 54.285 (1959), pp. 123–163.

[49] Bernard Desgraupes. *clusterCrit: Clustering Indices*. 2018.

[50] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". In: *Scientific reports* 9.1 (2019), pp. 1–12.

[51] Rahul Satija et al. "Spatial reconstruction of single-cell gene expression data". In: *Nat Biotechnol* 33.5 (2015), pp. 495–502.

[52] Kelly Street et al. "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19.1 (2018), p. 477.