A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data.

Shamus M. Cooley[1,3,4], Timothy Hamilton[2,3], Eric J. Deeds[2,3,4,5,*], and J. Christian J. Ray[4,5,*]

1. Interdepartmental Program for Bioinformatics, University of California – Los Angeles

2. Department of Integrative Biology and Physiology, University of California – Los Angeles

3. Institute for Quantitative and Computational Biosciences, University of California – Los Angeles

4. Center for Computational Biology, University of Kansas – Lawrence

5. Department of Molecular Biosciences, University of Kansas – Lawrence

*Corresponding authors: EJD (deeds@ucla.edu) and JCJR (jjray@ku.edu).

Abstract

High-dimensional data are becoming increasingly common in nearly all areas of science. Developing approaches to analyze these data and understand their meaning is a pressing issue. This is particularly true for the rapidly growing field of single-cell RNA-Seq (scRNA-Seq), a technique that simultaneously measures the expression of tens of thousands of genes in thousands to millions of single cells. The emerging consensus for analysis workflows reduces the dimensionality of the dataset before performing downstream analysis, such as assignment of cell types. One problem with this approach is that dimensionality reduction can introduce substantial distortion into the data; consider the familiar example of trying to represent the three-dimensional earth as a two-dimensional map. It is currently unclear if such distortion affects analysis of scRNA-Seq data sets. Here, we introduce a straightforward approach to quantifying this distortion by comparing the local neighborhoods of points before and after dimensionality reduction. We found that popular techniques like t-SNE and UMAP introduce significant distortion even for relatively simple geometries such as simulated hyperspheres. For scRNA-Seq data, we found the distortion in local neighborhoods was greater than 95% in the 2- and 3-dimensional space typically used for downstream analysis. This high level of distortion can readily introduce important errors into cell type identification, pseudotime ordering, and other analyses that rely on local relationships. We found that principal component analysis can generate accurate embeddings of the data, but only when using dimensionalities that are much higher than typically used in scRNA-Seq analysis. We suggest approaches to take these findings into account and call for a new generation of dimensional reduction algorithms that can accurately embed high dimensional data in its true latent dimension.

Introduction

Technological advances over the past century have enabled collection and analysis of data sets of unprecedented size and complexity. In geology, a modern assay might report the concentrations for over fifty elements from a single sample[1]; in climatology, measurements of sea surface temperature and the strength of zonal winds can be obtained simultaneously from hundreds of different sensors at any given point in time[2]; in cell and molecular biology, sequencing technologies have scaled up the throughput and resolution of genome data in populations[3, 4] and gene expression levels in cells[5, 6], into many thousands of dimensions in the case of single cell RNA-Seq (scRNA-Seq). Future technologies will doubtlessly expand the numbers of dimensions detected in complex systems by orders of magnitude.

While such datasets promise to provide greater insight into the problems being studied, high-dimensional data are also more difficult to analyze. The computational complexity of many data analysis algorithms scales exponentially with the dimensionality of the dataset, statistical inference often becomes difficult as dimensionality increases, and algorithms that work in lower dimensions become intractable in higher-dimensional spaces[7, 8]. This is often referred to as the "curse of dimensionality". The aim of dimensionality reduction is to reduce the dimensionality of the problem while retaining as much of the relevant information as possible– ideally all of it. It has become an indispensable tool for the rapidly growing number of scRNA-Seq studies.

Dimensionality reduction has a long history[9, 10]. Principal Component Analysis (PCA) is perhaps the oldest and most common linear approach, but many alternative approaches to linear dimensionality reduction exist as well, such as Non-negative Matrix Factorization (NMF) and Independent Component Analysis (ICA)[9, 11]. These algorithms are useful in a broad class of problems. However, linear approaches may be insufficient when the data display significant nonlinear characteristics[12]. In such situations, one often adopts a "manifold" assumption, which posits that the data can be modeled as smoothly varying local neighborhoods of dimension significantly lower than the ambient space[13]. A large number of Nonlinear Dimensionality Reduction (NDR) techniques have been developed to approximate these manifolds[14-17], including popular visualization methods like t-distributed Stochastic Neighbor Embedding (t-SNE)[18] and Uniform Manifold Approximation and Projection (UMAP)[19]. Collectively, the use of NDR techniques is often referred to as "manifold learning"[13].

In NDR techniques, one specifies the dimension of the resulting representation of the data. For example, if we use t-SNE to reduce the dimension of scRNA-Seq data, we tell the algorithm the number of dimensions that we want in the end. Unfortunately, the appropriate (or *latent*) dimensionality needed to correctly represent any given data set is generally not known *a priori*. A natural choice for visualization purposes is to choose two dimensions, since that kind of representation is easy to reproduce in the format of a figure. In the analysis of scRNA-Seq data, two dimensions are commonly used not just for visualization but also for downstream analyses ranging from cell type clustering (Fig. 1a) to "pseudotime" ordering[20]. Currently, it is unclear just how much character of the original data is being lost in the reduction of data on the order of 20,000 dimensions, typical for scRNA-Seq in many species, to two dimensions. Even when more dimensions are employed, the amount of information preserved in the dimensionality reduction step is not obvious. Because thousands or millions of cells can be characterized using scRNA-Seq, the resulting datasets are often massive, and dimensionality reduction is generally considered a necessary step in the analysis.

In order to understand the issues that might be introduced through dimensionality reduction, consider the familiar problem of making a 2-D map of the entire surface of the Earth.

Doing this requires "slicing" the earth along some axis in order to unfold it into a map; this is commonly done in a line through the Pacific, since few landmasses are disrupted by this cut. Then, the mapmaker must either increase the relative size of landmasses near the poles or slice the map again in order to project the globe into two dimensions. Regardless of technique, the globe cannot be represented in two dimensions without slicing and distorting the map in some way, which has led, for instance, to popular criticisms of the Mercator Projection. While distortion of distance and area are of course important, perhaps more concerning is the fact that the discontinuous slices mentioned above take points that are nearby (e.g. two points in the Pacific) and place them on opposite sides of the map. This means that the local neighborhoods of many of the points on the globe are completely different between the Earth itself and the 2-D representation.

With this observation in mind, it becomes apparent that there is no guarantee that high dimensional data sets, such as those associated with single cell genomics, can be represented in two dimensions without introducing analogous discontinuous slices into the data. Even techniques that attempt to objectively find a lower-dimensional representation using more than two dimensions, such as the common scree (elbow) plot technique in PCA to choose the directions that capture most of the variation in the data[21], could also suffer from similar problems. Yet, little analysis has been done to elucidate the extent to which NDR techniques introduce discontinuities into reduced-dimensional representations.

We approached this problem by applying a simple metric, inspired by the above metaphor of the globe, to quantify the extent to which any given dimensionality reduction technique discontinuously slices or folds the data in some way. This metric is based on comparing the *local neighborhood* of a point in the original data with the local neighborhood of that same point in the reduced-dimensional space using the Jaccard distance[22]. We first applied this approach to the simple problem of embedding points on the surface of a hypersphere (which is a straightforward generalization of the sphere to more than three dimensions) into the appropriate latent dimension from a higher-dimensional space. We found that many popular techniques, such as t-SNE and UMAP, not only introduced discontinuous slices into the data when trying to embed hyperspheres into two dimensions, but also when trying to embed into the correct latent dimension. Indeed, we failed to identify an NDR technique currently in widespread use for analysis or visualization of scRNA-Seq data that could successfully embed hyperspheres above approximately 10 dimensions.

We then used our metric to analyze how dimensionality reduction affects analysis of scRNA-Seq data. When embedding into 2 dimensions, we found that commonly-used techniques disrupt *95-99%* of the local neighborhoods in the data. Even when embedding into higher dimensions, NDR techniques generally introduced substantial discontinuity into the data. These discontinuities have important consequences for any approach that uses local neighborhoods for inference in scRNA-Seq data, including clustering and pseudotime ordering[20]. We found that PCA could find a true embedding for some data sets by using many more dimensions than are typically obtained through analysis of scree/elbow plots.

Our results demonstrate that, regardless of the technique used to reduce dimensionality, the majority of the local structure of high-dimensional data is lost when compressed into two dimensions. This implies that any analysis based on a 2-dimensional representation of the data introduces substantial bias into interpretations of the results. We show that NDR techniques do not generate valid embeddings even for simple manifolds, and that the distortion introduced by NDR techniques applied to existing scRNA-Seq datasets can significantly alter the results of downstream analyses like cell type clustering and pseudotime ordering. Our findings suggest straightforward guidelines for evaluating the quality of a lower-dimensional representation of scRNA-Seq data. Nevertheless, it is clear that new NDR techniques are needed that can reliably produce true topological embeddings, or, at least, closer approximations than current techniques can produce. We expect that the metric and approach introduced here will be helpful in evaluating and developing more effective approaches to the problem of manifold learning and analysis of scRNA-Seq or other high-dimensional data.

Results

Quantifying discontinuities introduced by dimensionality reduction

The goal of NDR is to learn a representation of a data set that has fewer features, but still retains the bulk of the information contained in the data. The extent to which the representations created by dimensionality reduction techniques actually preserve information is often illustrated with toy datasets such as the swiss roll (Fig. 1b). This example tests the ability of NDR techniques to represent the three-dimensional swiss roll data set in two dimensions while preserving the local structure of the original dataset (as can be seen here by the preservation of the "rainbow" pattern in the t-SNE representation). Most NDR techniques perform well on this task because a swiss roll is just a "rolled up" two-dimensional plane – a relatively simple transformation of a plane into a three-dimensional object. However, many objects, like the sphere in Fig. 1c, cannot be represented in 2-D without introducing significant distortion in local neighborhoods. This results in a notable scattering of the rainbow pattern (Fig. 1c).

Mathematically, a mapping from a high dimension to a lower dimension that (locally) preserves the structure of the data is called an *embedding*: technically, this a bijective map that is continuous in both directions (also called a *homeomorphism*). For topological spaces, a key mathematical property of an embedding is that it is *continuous*, and a consequence of that continuity is that local neighborhoods (e.g. the rainbow pattern in Fig. 1c) are preserved. For a swiss roll, NDR techniques like t-SNE can usually find an embedding, or something close to one. For a sphere, however, NDR finds a representation of the data in two dimensions that is not, strictly speaking, an embedding.

It is clear from the simple example in Fig. 1c that a major problem with trying to embed a sphere in 2-D is that this is impossible to do without introducing discontinuities into the resulting representation. In the context of experimental scRNA-Seq data, this means that the local structure of the data may be lost in the dimensionality reduction, and error (possibly large error) could be introduced into any analysis that happens downstream of NDR. This is particularly problematic because we do not know *a priori* what the true dimension of a particular scRNA-Seq data set might be. Previous work on quantifying distortion in NDR has focused on the notion of Euclidean distance[19, 23], which is the formulation of distance most of us are familiar with. However, quantifying the extent of the loss of structure caused by NDR is difficult using Euclidean distance, because it is not necessarily correlated with distortion in the local structure of the data. For example, a 2-D representation of the swiss roll might be stretched out, greatly distorting the Euclidean distance, while still maintaining the rainbow structure depicted in Fig. 1c and thus providing a true embedding. This suggests the need to develop alternative approaches to quantifying distortion in NDR, particularly focused on characterizing discontinuities that may be introduced by dimensionality reduction techniques.

For any point in the swiss roll, the neighborhood of other points that are nearest to it are roughly the same in three dimensions and in the t-SNE representation in two dimensions (Fig. 1b). The two-dimensional representation of the sphere, on the other hand, gives noticeably different sets of nearest neighbors to many points (Fig. 1c). We thus developed a straightforward metric based on quantifying how similar the sets of neighbors are around each point between the original, high-dimensional data in the ambient space, and the low-dimensional representation. First, we find the *k*-nearest neighbors for each point in the original data. We call this set A (see Fig. 1d). Next, we find the *k*-nearest neighbors in the lower-dimensional space. We call this set B. We compare these two sets using a measure of dissimilarity called the Jaccard distance (Fig. 1e). Calculating the Jaccard distance involves computing the size (or *cardinality*) of the *symmetric difference* between A and B: the symmetric difference is just the set of points that are in A or B, but not both. This is equivalent to subtracting the number of points in the intersection between A and B from the number of points in the union (Fig. 1e). The Jaccard distance is the ratio of the size of this symmetric difference to the total number of points in A and B together (i.e. the number of points in the union between A and B).

If A and B are identical sets, meaning the neighbors of the point in the high-dimensional data and the low-dimensional representation are the same, then the Jaccard distance is 0. If A and B are completely different sets (i.e. the neighbors around this point completely change) then the Jaccard distance is 1. It is easy to prove that, for a true topological embedding the Jaccard distance will be zero for every point in the dataset (Supplemental Info); in other words, in a true embedding all local information is preserved. To characterize the global "distance" of any low-dimensional representation from this ideal, we first compute the Jaccard distance for all the points in the data set and then average these values. We refer to this quantity as the Average Jaccard Distance (AJD), and it gives a value of 0 for a true embedding, 1 for a representation that retains none of the information about the local structure of the data for any point in the data set, and intermediate values for a representation that retains part of the information.

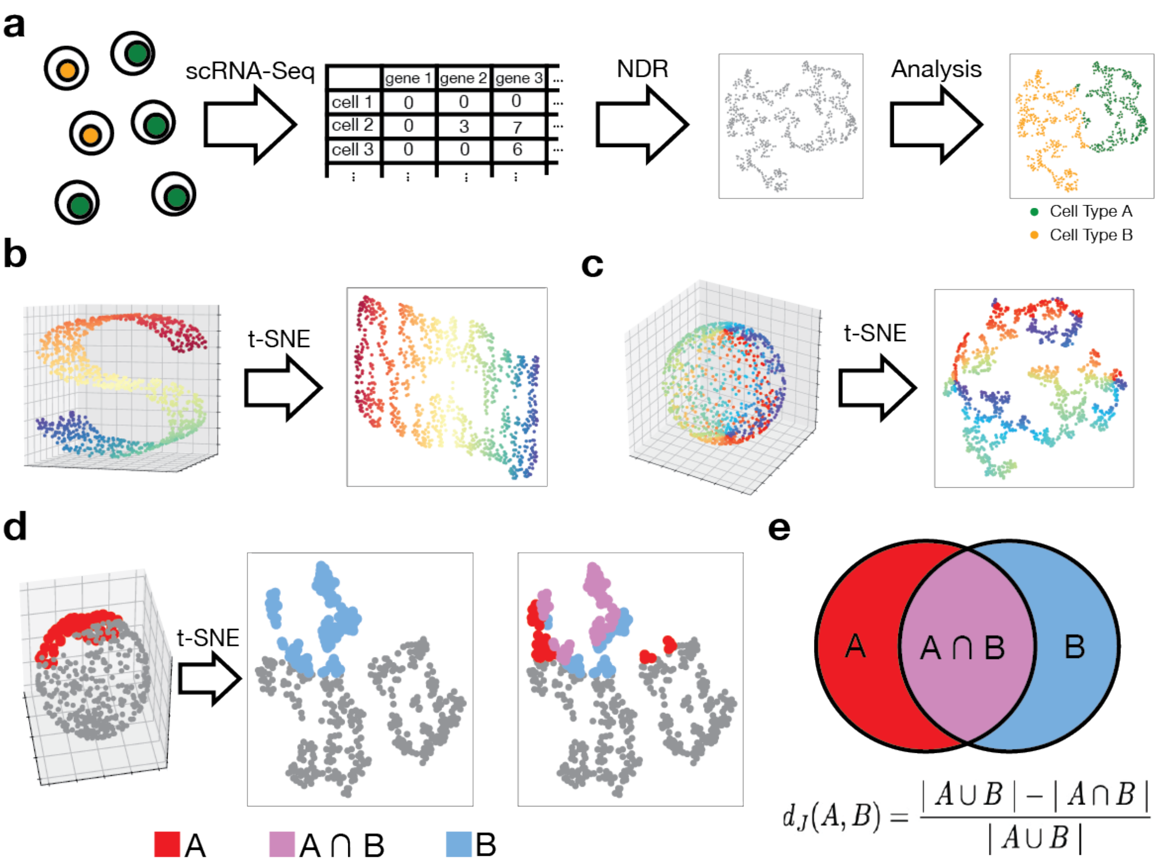

Fig. 1. (a) A schematic of some scRNA-Seq workflows. The gene expression data are stored as a matrix, with each row corresponding to a cell, and each column correspond to a gene (after correcting for UMI swapping). The data undergo dimensionality reduction, and analysis is performed on the lower-dimensional representation of the data. **(b)** The "swiss roll" data set. t-SNE is able to reduce the data into two dimensions without altering the local structure of the data. **(c)** A sphere data set. t-SNE is unable to represent the 3-dimensional object in 2 dimensions without disrupting the local structure of the data. **(d)** An illustration of how NDR distorts local neighborhoods. The red points are the *k*-nearest neighbors of a single point in the 3-dimensional space. The blue points are the *k*-nearest neighbors of the same point in the t-SNE-generated 2-dimensional representation. The violet points are the intersection between the red points and the blue points. **(e)** The Jaccard Distance is a method for quantifying the disruption in local neighborhoods pictured in **d**.

Testing on Synthetic Data

To test the usefulness of AJD, we first applied the metric to a problem where we know *a priori* the appropriate embedding dimension for the data set. Specifically, we created synthetic data for hyperspheres of varying dimension. A hypersphere is a manifold that represents a straightforward generalization of the standard 3-dimensional sphere to higher numbers of dimensions; it is just a collection of points in some n-dimensional space that are all the same distance from a central point (that distance is the radius of the sphere). In two dimensions this is a circle, in three dimensions a sphere, and in higher dimensions a hypersphere. We used a simple algorithm to sample uniformly from the surface of a hypersphere in n dimensions; for simplicity we used the origin of the space as the central point, and we set the radius of the hypersphere to 1 (see Methods). It is mathematically impossible to embed an n-dimensional sphere generated this way in less than n dimensions, so we called n the "latent dimension" of the data. To see if NDR techniques could generate a true embedding of the data into n dimensions, we first embedded our hyperspheres into a 100-dimensional ambient space. To demonstrate how we did this, take the case of a 20-dimensional hypersphere. If we sample points from that hypersphere, each one of those points is characterized by a vector of 20 numbers. We can trivially embed those points into a 100-dimensional space by just adding 80 zeroes to the end of those vectors (see Methods and Supporting Info).

We used the approach above to generate synthetic 100-dimensional datasets with 1000 points sampled from hyperspheres of known latent dimension. We then used multiple NDR techniques to embed this dataset into each lower dimension from 1 to 100. We hypothesized that the AJD would be zero for every dimension above the latent dimensionality n of the manifold that we had generated. Surprisingly, however, we found that the AJD did not reach 0 for hyperspheres with $n \geq 10$ for any NDR technique that we tried when we used a neighborhood size of $k = 20$ (see Fig. 2a and Supporting Info). In the case of the popular technique t-SNE, for instance, the embeddings it produced generally had AJDs of greater than 0.75, regardless of both the latent dimension of the hypersphere and the embedding dimension used for the t-SNE algorithm. Other techniques, such as Isomap and Spectral Embedding[12, 14] exhibited clear minima in the AJD at the appropriate latent dimension, but still produced embeddings with significant distortion. Changing the size of the neighborhood between 10 and 100 points did not significantly alter these findings (Supporting Info). This result is particularly striking because we know that it is possible to embed a 20-dimensional hypersphere into a 20-dimensional space without any distortion at all (corresponding to an AJD of 0). Indeed, for the case of this particular synthetic dataset there is a trivial mapping that results in a true embedding and an AJD of zero in the latent dimension, but none of the commonly used techniques that we tested successfully recovered it.

We hypothesized that the datasets were too small, and that an increased sample size might allow the algorithms to find a proper embedding. Although increasing the sample size created a more pronounced local minimum at the latent dimension for some techniques (Fig. 2b), the AJD at the latent dimension never dropped below a certain level: this minimum was invariant to increases in sample size of points on the sphere (Fig. 2c). In the case of MDS, increasing sample size resulted in *more* distorted representations at the latent dimension. Again, these simulated datasets represent what should be a relatively trivial problem for manifold learning. The fact that no nonlinear dimensionality reduction technique could find even this simple mapping raises questions about the accuracy of the approximate “embeddings” generated by NDR and the effects that distortion might have on the analysis of scRNA-Seq and other high-dimensional data.

Measuring Distortion in scRNA-Seq Studies

To address these questions, we identified state-of-the-art scRNA-Seq studies[24, 25] and analyzed the effect of NDR on the analysis of these data. First, we looked at a study of Hydra cells by Siebert et al.[24]. For this dataset, we selected one of the largest cell type clusters defined in the

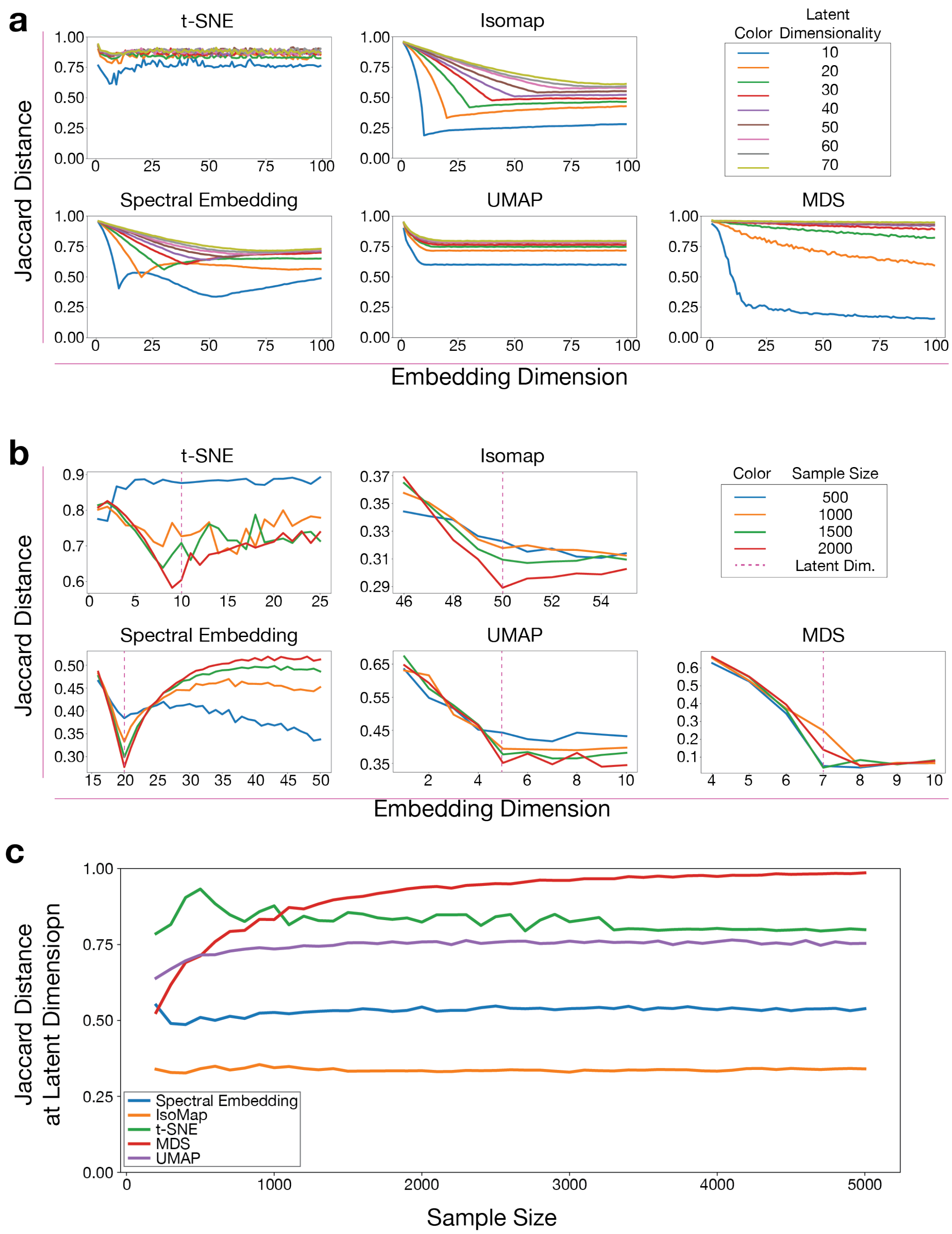

Fig. 2. **(a)** The Average Jaccard Distance (AJD) for points randomly sampled from the surface of hyperspheres of varying dimension embedded in dimensions 1-100. The AJD is lowest when the latent dimensionality of the manifold is lowest. **(b)** The effect of sample size on Average Jaccard Distance. Although the shape of the curve more clearly indicates the latent dimensionality of the manifold, the distortion in local structure (AJD) does not improve with increased sample size. **(c)** The Average Jaccard Distance as the sample size increases from 100-5000 points. The distortion created by the embedding is mostly independent of sample size. (The latent dimension of these datasets was 20, and the ambient dimension of these datasets was 100.)

study (1,778 cells), an endodermal epithelial stem cell, and reduced the gene expression data corresponding to these cells into dimensions ranging from 1 to 100 (Fig. 3 a, b). The AJD for these low-dimensional representations never dropped below 0.5, and for the most commonly used number of dimensions for analysis and visualization, 2 and 3, the AJD was close to one, regardless of the technique employed. In other words, mapping the data down to 2 or 3 dimensions introduces so much distortion that nearly every point in the dataset has a *completely different* neighborhood in the NDR representation compared to the original data. Above 100 dimensions, many techniques, such as Spectral Embedding, exhibited numerical instabilities and could not be used. For those NDR techniques that consistently worked above 100 dimensions, we attempted embedding the data in dimensions ranging up to 1400 (Fig. 3b) but did not find any indication of approaching a true embedding (AJD≈0). As a control, we used PCA and found that the AJD only approached zero when the embedding dimension approached the number of cells in the cluster (~1,750 see Fig. 3b). The number of cells sets the absolute limit of the number of dimensions that PCA can find, indicating that even PCA cannot find a meaningful reduction of the dimensionality in this particular case (see Supporting Info).

We next looked at a large study conducted by Cao et al.[25] in mice. We again selected one of the largest cell type clusters, in this case corresponding to a particular subcluster of excitatory neurons with around 10,000 total cells and used common NDR algorithms to represent the data in dimensions ranging from 1-100 (Fig. 3c). We found that NDR representations of the data demonstrated even higher AJD values than the Hydra case, and that AJD only approached zero with PCA when the embedding dimension was approximately 10,000 (Fig. 3d), which again was close to the number of cells in the cluster.

These results indicate that dimensionality reduction likely introduces significant distortion into data not only reduced to two dimensions, which is commonly used for visualization and some data analysis, but even in higher-dimensional representations of the data. As some degree of dimensionality reduction is an integral part of essentially every scRNA-Seq data analysis pipeline, it is unclear how accurate the results of most scRNA-Seq analysis are.

Evaluating the Effect of NDR Distortion

Although the distortion in local neighborhoods caused by NDR is quite high when the techniques are applied to scRNA-Seq data, it is unclear if these effects are mostly local, or if the problem is more global in nature. In other words, it is possible that, within some local region of the data, NDR is essentially moving points around within the region. This would lead to an AJD near one with a neighborhood size of ~20 but may not significantly affect analyses like cell type clustering. Alternatively, the distortion caused by NDR might move points over large distances, as in the example with the sphere discussed above (Fig. 1c). More global changes like this could introduce more significant errors into cell type clustering and other analyses.

To test this, we first considered how the AJD changes as a function of the neighborhood size used to calculate the Jaccard distances. If the distance goes to 0 at a relatively small neighborhood size (say, around 100 or so), this would imply that the distortion due to NDR is primarily local. If not, it implies that the distortion is more global. We applied this analysis to hyperspheres, and found that, for many techniques including t-SNE and UMAP, the AJD did not approach 0 until we included the majority of the data set in the neighborhood even at the latent dimension, indicating that the distortion in the case of hyperspheres is global in nature (see Supporting Info). We applied a similar analysis to the endothelial cell cluster from the Siebert et al. Hydra dataset[24]. Because we do not know the “true” latent dimension for this dataset, we chose

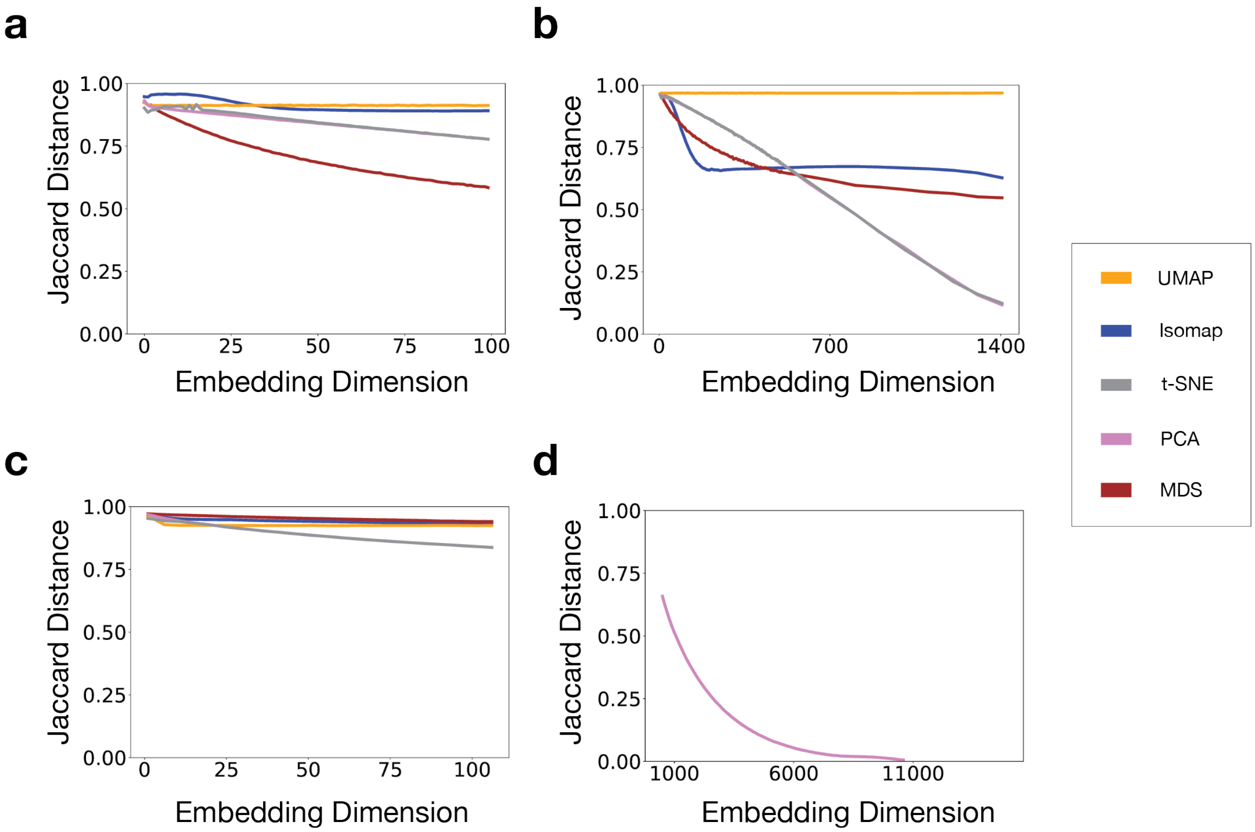

Fig. 3. scRNA-Seq data from Hydra (Siebert et al.[24]) and mouse (Cao et al.[25]). **(a)** The Average Jaccard Distance of representations in embedding dimensions from 1-100 of Hydra data using various techniques. **(b)** The Average Jaccard Distance of representations in embedding dimensions from 1-1400 in Hydra data using various techniques. Note that the t-SNE and PCA results are essentially identical; this is likely because t-SNE begins with a PCA embedding and the subsequent steps of t-SNE do not alter the embedding much in this case. **(c)** The Average Jaccard Distance of representations in embedding dimensions from 1-100 in mouse data using various techniques. **(d)** The Average Jaccard Distance of representations in embedding dimensions from 1-14,000 in mouse data using PCA (other techniques were too computationally costly at these dimensionalities).

---

to use two dimensions, the typical dimensionality for visualization and, frequently, data analysis[20]. Here we also found that the AJD did not fall to 0 until we computed the Jaccard Distance using the entire cell type cluster, which indicates that the distortion due to NDR is global in nature (Fig. 4a).

The above analyses were performed on minimally-processed scRNA-Seq data where the raw counts were just corrected for doublets, batch effects, and other common sources of technical noise in the scRNA-Seq experiment. In practice, NDR is rarely used on this type of relatively unprocessed scRNA-Seq data. In particular, transcript counts for each cell are often reduced to a subset of "Highly Variable Genes" (HVGs) that display significantly more variability between cells in the experiment than one would expect according to some null model. Reduction of the gene set to HVGs is itself a form of dimensionality reduction. Next, the data are subjected to linear dimensionality reduction. Often a scree plot is used to select the embedding dimension for PCA. Clustering is performed after this linear reduction, and nonlinear reduction is used for visualization of the results. It is common for developmental "pseudotime trajectories" to then be derived from the data after NDR[26, 27]. This is done by constructing a minimum spanning tree across the reduced data set and ordering cells using this tree[20].

Such analysis pipelines clearly entail several dimensionality reduction steps, and our results above indicate that severe distortion is likely introduced at each step. We thus sought to analyze the consequences of this distortion on the results of a typical analysis pipeline applied to the Siebert et al. Hydra data. We used the Seurat package in R[28] to perform these analyses, partially because of the popularity of the package and partially because the original analysis of the data was performed using Seurat[24]. Using the scree plot, we estimated the "elbow" in the amount of variance explained occurred at a dimension of 12 in PCA, so we choose that as the embedding dimension for the linear dimensionality reduction step in the pipeline (see Supporting Info). We first computed the AJD between each step in the standard pipeline. As expected based on our findings above, each step of dimensionality reduction introduces significant distortion, with AJD values between the original data and the processed data above 0.9 for almost every step (Table 1). Clearly, the local structure of the data is almost entirely lost downstream of the final NDR step.

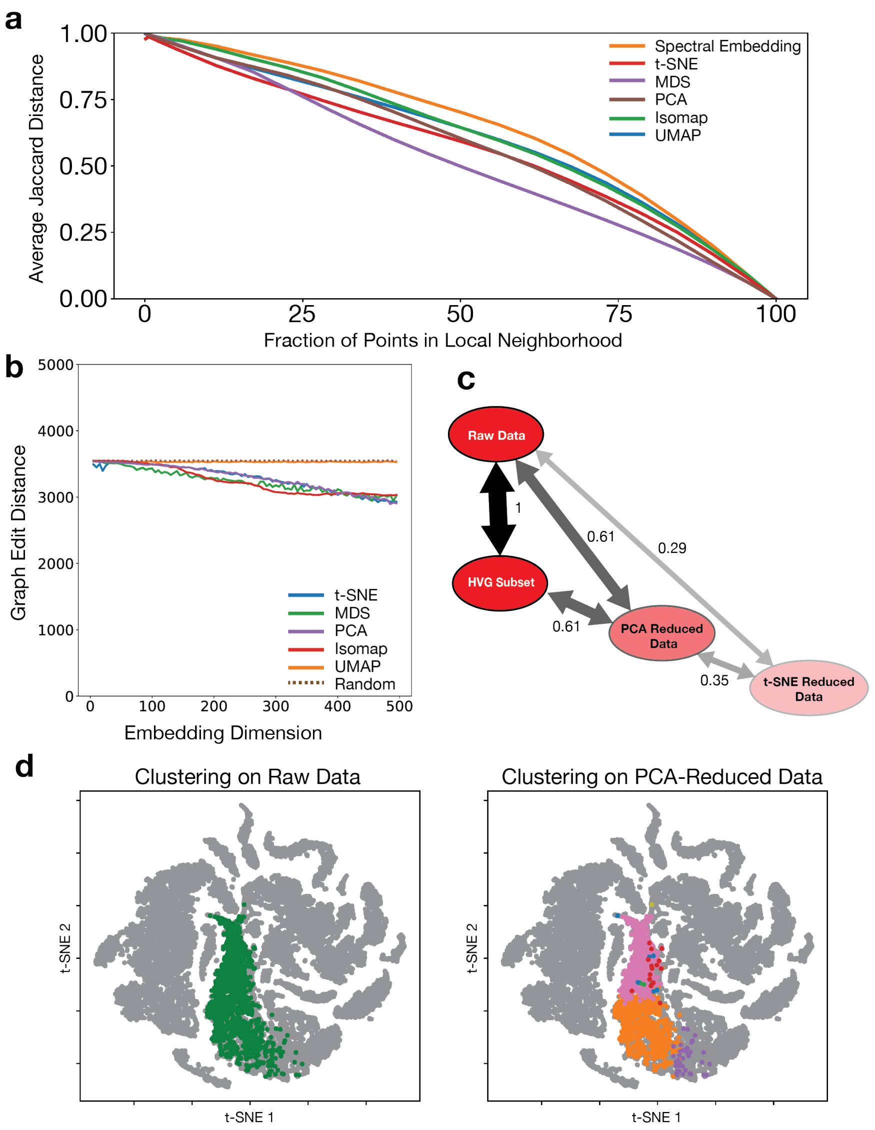

Fig. 4. (a) The Average Jaccard Distance as a function of *k*-nearest neighbors used to compute Jaccard distance for the Siebert et al. data[24]. The effect of distortion is not just limited to local neighborhoods. **(b)** The Graph Edit Distance between a minimum spanning tree constructed in the ambient space and a minimum spanning tree constructed in the NDR-reduced representation. The dotted line corresponds to a random embedding that retains none of the original information. **(c)** Adjusted Rand Index values between each stage of a typical scRNA-Seq analysis applied to the Hydra data from Siebert et al. Note that here we employ all the cells in their data set, not just the cells from the largest cluster that they identified. The opacity of the ellipse corresponds to the ARI and hence the similarity of the lower-dimensional representation to the raw data. The thickness and opacity of the arrows correspond to the value of the ARI. **(d)** The result of clustering of scRNA-Seq data in the original, ambient dimension (left), and the result using the same clustering algorithm with the same parameters on PCA-reduced representation of the data. Only a subset of the points is colored for clarity. The graphs were produced using t-SNE for the purpose of visualization only, as the t-SNE embedding loses much of the structure of the data.

One of the most common applications of scRNA-Seq analysis is in the identification of distinct cell types in the data, which is usually done by clustering the cells after dimensionality reduction has been performed[24, 25, 29]. We used the standard Adjusted Rand Index (ARI) to quantify the similarity of the clusters obtained from each step along the data analysis pipeline. Because clustering only makes sense in the case where there are multiple distinct cell types, we applied this analysis to all 24,458 cells in the Hydra data set rather than the ~1,700 endothelial cells we focused on above. We obtained clusters using the standard procedure in Seurat (see Methods). As can be seen from Fig. 4c the clusters obtained from the minimally processed data and clusters obtained from Seurat's HVGs completely agree (ARI = 1). Thus, while reducing the dataset to HVGs does disrupt local neighborhoods (Table 1), this disruption is not sufficient to change the clustering of cells into cell types. Reducing to HVGs therefore likely introduces local, but not global, distortions.

Table 1. Average Jaccard distance (AJD) between the minimally processed (raw) Hydra scRNA-Seq dataset and the data after various processing steps.

Analysis Step	**AJD from Raw Data**
Highly-Varying Genes	0.96
PCA	0.88
t-SNE	0.94

Clustering is not usually performed directly after identification of HVGs. Instead, it is common to use the elbow/scree plot to choose a number of dimensions for PCA and cluster based on the PCA-transformed data. We see that the ARI between the HVG data set and the PCA-based clusters is ≈0.6, indicating significant divergence between the clusters produced in both cases. This effect is visualized in Fig. 4d, where a cluster obtained in the HVG data is visualized using t-SNE, demonstrating a notable difference in how cells are classified into different cell types. Clustering after using t-SNE to reduce to 2 dimensions results in even greater changes in clusters, with an ARI of ≈0.3 (Fig. 4c). Overall, these results suggest that distortion introduced by both linear and non-linear dimensionality reduction can significantly change the classification of cells into specific cell types based on clustering in scRNA-Seq data.

Pseudotime ordering attempts to use cells captured at various points along a differentiation or developmental trajectory to infer the underlying trajectory itself[20]. A key step in this analysis is the calculation of a *minimum spanning tree* that connects the beginning and end point in the trajectory. This tree is formed by linking cells in close proximity to each other to form a graph, typically after NDR is performed. Because NDR readily changes both the local and global relationships between cells in the data set (Fig. 3 and 4a), we hypothesized that the trees produced by analyzing data after NDR would not closely resemble trees formed using the original data. To test this, we calculated the graph edit distance between trees formed from the raw data and after various NDR techniques were used to project the data into a variety of different dimensions (Fig. 4b). For comparison, we also generated a random embedding by simply assigning each cell to a random point in the reduced-dimensional space (see Methods). The graph edit distances obtained from the NDR techniques and from the random embedding are similar until embedding dimensions of ~100 are reached (Fig. 4b). Even above 100 dimensions, the improvement in the graph edit distance relative to a random embedding is not very large. Because pseudotime trees are usually built using 2- or 3-dimensional representations based on t-SNE, UMAP or similar techniques[20], our findings suggest that distortion caused by NDR could have a large effect on the results.

Methods

Average Jaccard Distance

For each data point, the neighborhood consisting of the nearest k-neighbors were found in the ambient space, call this set A, and the NDR-reduced space, call this set B, using sklearn.neighbors.NearestNeighbors. We employed the ball-tree algorithm in both cases. To calculate the Jaccard distance between A and B, we used the usual definition:

$$J_d(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

The Average Jaccard Distance was calculated by taking the arithmetic mean of the Jaccard distance for every point.

Sampling of Hyperspheres

To create a synthetic dataset consisting of m samples an n-dimensional spherical manifold in d-dimensional space, we used the following method: For each of the m data points, we sampled from a uniform distribution over (-1,1) n times (using the Python method random.uniform). These samples became the first n coordinates of a vector. The remaining n+1 to d coordinates were filled with zeros. We then normalized each vector to length 1.

Dimensionality Reduction

We executed dimensionality reduction with t-SNE, Isomap, PCA, Spectral Embedding, Multidimensional Scaling, LLE, and LTSA using the implementations in Scikit-learn[30]. For the methods UMAP and diffusion maps, we used umap-learn[19] and pydiffmap[31], respectively. We implemented PCA using sklearn.decomposition.PCA. We used default parameters except where otherwise noted.

scRNA-Seq Data

The study from Siebert et. al. is published on the Broad Institute's single cell portal: https://portals.broadinstitute.org/single_cell/study/SCP260/stem-cell-differentiation-trajectories-in-hydra-resolved-at-single-cell-resolution.

The study from Cao et. al. is published on The Gene Expression Omnibus: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119945

The .txt files were converted to .csv files corresponding to individual clusters, and the data were loaded into Python pandas (https://pandas.pydata.org/) dataframes for dimensionality reduction.

Minimum Spanning Tree and Graph Edit Distance

The minimum spanning tree in the ambient space, mst_1, and the minimum spanning tree in the NDR-reduced space, mst_2, were constructed using the Python function scipy.sparse.csgraph.minimum_spanning_tree. The graph edit distance was calculated in Python according to the following equation:

$$GED(mst_1, mst_2) = \min_{\{e_1,\ldots,e_k\} \in P(\mathrm{mst}_1, \mathrm{mst}_2)} \sum_{i=1}^{k} c(e_i)$$

Where $P(\mathrm{mst}_1, \mathrm{mst}_2)$ is the set of edit paths transforming mst_1 into mst_2 and $c(e_i)$ is the cost of each graph edit operation e_i. The cost of deleting a vertex and the cost of adding a vertex were both weighted as 1.

As a control, a random embedding was created by sampling coordinates from a uniform distribution between -1 and 1. The minimum spanning tree was then computed on this random embedding and the Graph Edit Distance was calculated between this tree and the minimum spanning tree constructed in the ambient space.

Adjusted Rand Index

The Rand index quantifies the similarity between clusters in two partitions U and V (say, cell clusters in the ambient dimension and in a reduced dimension) through a contingency table that classifies pairs of points into four cases: pairs in the same cluster in both partitions (a), pairs in the same cluster in U but not V (b), pairs in the same cluster in V but not U (c), or pairs in different clusters in both partitions (d). It takes a value between 0 and 1. The adjusted Rand index corrects the value by accounting for coincidental/chance clustering and avoiding the tendency of the unadjusted Rand index to approach 1 as the number of clusters increases. It is given by $ARI = \frac{\binom{n}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)]}{\binom{n}{2}^2-[(a+b)(a+c)+(c+d)(b+d)]}$ where n is the number of points and $\binom{n}{2}$ is the total number of possible point pair combinations[32].

Replicating scRNA-Seq Workflows

To replicate a typical workflow, we used Seurat in R[28]. To isolate highly variable genes, we used the data from the function FindVariableFeatures() in Seurat with default parameters. For PCA reduction, we used the ElbowPlot function, with the “elbow” observed to be at 12 PCs. Our clustering was done in Seurat using the function FindNeighbors() on the specified dimensional space to compute the Shared Nearest Neighbor Graph, followed by the FindClusters() function. We set the resolution at 0.8, number of random starts at 10, random seed at 0, maximum number of iterations at 10 and we used the standard modularity function.

Discussion

The capacity to generate high-dimensional data is currently in the process of revolutionizing scientific inquiry. scRNA-seq, for example, has the potential to drive significant advances in our understanding of the evolution and differentiation of cell types, the progression of cellular state during development and disease, and a host of other critical biological phenomena[13, 33, 34]. Yet the very thing that makes this technique so powerful – the ability to simultaneously measure the expression level of tens of thousands of genes within a single cell – also entails the curse of dimensionality and thus complicates the analyses needed to extract meaning from it. As such, dimensionality reduction has become an indispensable part of scRNA-Seq data analysis[13]. It is currently unclear, however, to what extent dimensionality reduction disrupts the underlying structure of the data itself.

Distortion from dimensionality reduction can take several forms. Much of the previous work on this problem has focused on the extent to which the process changes the distances between points[18, 19]. Our work highlights that there are even larger problems with dimensionality reduction than just distortion of distances. For one, even in possession of a perfect technique, one cannot reduce the dimensionality of the data to arbitrarily low dimensions without creating large numbers of discontinuities in local neighborhoods and other distortions in the data. In the case of points taken from the surface of a 3-D sphere, it is mathematically impossible to project those points into a 2-D representation without introducing discontinuities into the data (e.g. the scattering of the rainbow pattern in Fig. 1c). Many analyses commonly performed with scRNA-Seq data, including cell type clustering, RNA velocity[35], and pseudotime ordering, rely at least in part on the local relationships between data points. The introduction of discontinuities thus has the potential to significantly impact the results of that kind of analysis.

A second problem is the fact that, even if it is theoretically possible to represent the data in a given dimension, available techniques may not be capable of finding that representation. Unfortunately, it is currently impossible to evaluate the extent to which either of these issues have an impact on the analysis of scRNA-Seq data (or, indeed, any high-dimensionality data).

Here, we developed a straightforward metric that quantifies the extent to which discontinuities of the type exemplified in Fig. 1c would impact the analysis of any given data set.

One immediate application of this metric is in the discovery of the appropriate latent dimension of a given data set. In testing this use case on data sampled from hyperspheres, however, we found that a large number of NDR techniques currently in widespread use are far from perfect (Fig. 2). Indeed, none of the techniques we tested could find a true embedding for even a 20-dimensional hypersphere, despite a complete lack of noise in the data and the fact that the embedding in this case was rather trivial (and known *a priori*). This finding suggests that fundamental work is needed to develop new and more effective NDR techniques. We expect that both the AJD metric we developed and the hypersphere example we explored will prove useful in the design and testing of these algorithms.

Application of our metric to scRNA-Seq data revealed that the problem there is even worse than for hyperspheres (Fig. 3). For instance, it is currently common to use t-SNE or UMAP to reduce scRNA-Seq data to two dimensions for visualizations and, in many cases, downstream data analysis[20, 24, 25]. Our work revealed that nearly 100% of the local neighborhood structure is disrupted by this kind of dimensionality reduction. We found that this level of distortion has a significant effect on the results of common analyses such as cell type clustering and pseudotime ordering (Fig. 4).

There are several practical implications of our findings for routine scRNA-Seq analysis. For one, it seems likely productive to perform cell-type clustering using a set of “Highly Variable Genes” provided by popular packages like Seurat, because this preserves the resulting clusters while reducing dimensionality (and thus the computational resources required) by about an order of magnitude (Fig. 4). Another straightforward recommendation flowing from this work is to exercise caution when analyzing data in dimensions that are significantly smaller than the ambient space of the original measurements, particularly the 2-D representations generated by t-SNE or UMAP. We recommend that practitioners use the AJD to track the distortion they introduce into their dimensionally-reduced data and report it so that others can understand potential biases and errors that may affect the results of analyses that rely on local relationships between cells in the dataset.

Of course, one question raised by our results is whether or not meaningful dimensionality reduction of scRNA-Seq data is possible at all. The poor performance of NDR techniques on the simple hypersphere tests makes it difficult to say whether the results we obtained for scRNA-Seq data are due to the limitations of available techniques or because the data do not actually lie on a low-dimensional manifold. The only technique that we found to provide something close to a "true" embedding, PCA, does so only at dimensionalities that are close to the maximum possible dimensionality that can be obtained by the technique (Fig. 3). The development of new NDR techniques that are more effective at finding true embeddings thus represent a critical step in answering central questions in cell biology. Until such techniques are developed, the relentless expansion of single-cell genomics to larger and larger scales may provide a wealth of new data that cannot be optimally mined for its biological insights.

Acknowledgments

We thank members of the Ray lab and Tom Kolokotrones for useful discussions. This work was supported in part by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM103638 to JCJR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Horrocks, T., Holden, E.J., Wedge, D., Wijns, C. & Fiorentini, M. Geochemical characterisation of rock hydration processes using t-SNE. *Comput Geosci* **124**, 46-57 (2019).

2. Chalupka, K., Bischoff, T., Perona, P. & Eberhardt, F. Unsupervised discovery of El Nino using causal feature learning on microlevel climate data. *arXiv:1605.09370 [stat.ML]* (2016).

3. Lemmon, E.M. & Lemmon, A.R. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol S* **44**, 99-121 (2013).

4. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**, 87-98 (2011).

5. Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V. & Zhang, K. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).

6. Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133-145 (2015).

7. Indyk, P. & Motwani, R. in STOC '98 Proceedings of the thirtieth annual ACM symposium on Theory of computing. (ed. J. Vitter) (ACM, Dallas, TX, USA; 1998).

8. Friedman, J.H. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* **1**, 55-77 (1997).

9. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572 (1901).

10. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* **24**, 417-441 (1933).

11. Cichocki, A. & Phan, A.-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE T Fund Electr* **E92-A**, 708-721 (2009).

12. DeMers, D. & Cottrell, G. in Advances in neural information processing systems 5 [NIPS Conference]. (eds. S.J. Hanson, J.D. Cowan & C.L. Giles) 580-587 (Morgan Kaufmann Publishers Inc., San Francisco, CA; 1993).

13. Moon, K.R., Stanley, J.S., Burkhardt, D., van Dijk, D., Wolf, G. & Krishnaswamy, S. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr Opin Syst Biol* **7**, 36-46 (2018).

14. Tenenbaum, J.B., de Silva, V. & Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323 (2000).

15. Kruskal, J.B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* **29**, 115-129 (1964).

16. Knyazev, A.V. Preconditioned eigensolvers - An oxymoron? *Electron Trans Numer Anal* **7**, 104-123 (1998).

17. Roweis, S.T. & Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323-2326 (2000).

18. van der Maaten, L. & Hinton, G.E. Visualizing data using t-SNE. *J Mach Learn Res* **164**, 2579-2605 (2008).

19. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Software* **3**, 861-861 (2018).

20. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. & Rinn, J.L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).

21. Cattell, R.B. The scree test for the number of factors. *Multivariate Behav Res* **1**, 245-276 (1966).

22. Levandowsky, M. & Winter, D. Distance between sets. *Nature* **234**, 34-35 (1971).

23. Zhang, Z.-Y. & Zha, H.-Y. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *J Shanghai University (English Edition)* **8**, 406-424 (2004).

24. Siebert, S., Farrell, J.A., Cazet, J.F., Abeykoon, Y., Primack, A.S., Schnitzler, C.E. & Juliano, C.E. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *bioRxiv: 460154* (2018).

25. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., Trapnell, C. & Shendure, J. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

26. Zhong, S., Zhang, S., Fan, X., Wu, Q., Yan, L., Dong, J., Zhang, H., Li, L., Sun, L., Pan, N., Xu, X., Tang, F., Zhang, J., Qiao, J. & Wang, X. A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524-528 (2018).

27. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A. & Schier, A.F. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, 1-15 (2018).

28. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).

29. Kim, T., Chen, I.R., Lin, Y., Wang, A.Y., Yang, J.Y.H. & Yang, P. Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief Bioinform* **bby076**, 1-11 (2018).

30. Pedregosa, F., Varoquax, G. & Gramfort, A. Scikit-learn: Machine learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

31. Berry, T. & Harlim, J. Variable bandwidth diffusion kernels. *Appl Comput Harmon A* **40**, 68-96 (2016).

32. Santos, J.M. & Embrechts, M. in Artificial Neural Networks–ICANN 2009, Vol. 5769 175-184 (2009).

33. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., Pun, S.H., Sellers, D.L., Tasic, B. & Seelig, G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).

34. Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A. & Lander, E. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *Cell* **176**, 928-943.e922 (2017).

35. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lonnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundstrom, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S. & Kharchenko, P.V. RNA velocity of single cells. *Nature* **560**, 494-498 (2018).