Title
Distortion in Dimensionality Reduction and Implications for the Analysis of Single Cell RNA-Sequencing Data

Author
Cooley, Shamus McKinney

Publication Date
2021

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Distortion in Dimensionality Reduction and Implications for the Analysis of Single Cell RNA-Sequencing Data

A dissertation submitted in partial satisfaction of the

Requirements for the degree Doctor of Philosophy

In Bioinformatics

by

Shamus McKinney Cooley

2021

ABSTRACT OF THE DISSERTATION

Dimensionality Reduction

for the Analysis of Single Cell

RNA-Sequencing Data

by

Shamus McKinney Cooley

Doctor of Philosophy in Bioinformatics

University of California, Los Angeles, 2021

Professor Eric Deeds, Chair

Dimensionality reduction is nearly ubiquitous in the analysis of single cell sequencing data. However, until the current work, no serious effort had been made to quantify the distortion introduced by dimensionality reduction and the effect of that distortion on the analysis. Here, I first present a method for the measurement of distortion caused by dimensionality reduction, Average Jaccard Distance. I will show that the application of this metric to data analysis workflows suggests the need for revision in the way that these methods are used for single cell RNA sequencing analysis. Next, I propose a revised methodology, and present the results of applying this revised methodology to the study of small cell lung cancer. The results include the identification of a stem-like population of cancer cells and many potential drug targets. Finally, I

present the schematic of a new, more accurate method of dimensionality reduction using deep neural networks.

The dissertation of Shamus McKinney Cooley is approved.

Alexander Hoffmann

Van Savage

Xinshu Grace Xiao

Eric Deeds, Committee Chair

University of California, Los Angeles

2021

Dedicated to Maya,

without whose love and support

this work would have been impossible

Table of Contents:

Acknowledgements

I would like to acknowledge the individuals who made this work possible: Christian Ray for helping to develop the Average Jaccard Distance metric and providing invaluable advice and feedback throughout my PhD training. Timothy Hamilton, who did a great deal of the legwork in analyzing data from small cell lung cancer, and whose contribution in terms of ideas are too numerous to list here, including the use of CytoTRACE. Samuel Aragones for his tireless work and ability to overcome even the most frustrating computing hardware issues. Serena Hughes, who collaborated closely with me in the development and implementation of neural networks for dimensionality reduction, and who solved many of the difficult problems involved with that effort, including finding a differentiable approximation of the average Jaccard distance to use as a loss function. Melanie Tu, who worked tirelessly to develop and test the implementation of the Deep Embedder algorithm. Pushpa Itagi, Anupama Kante, and Leo Lagunes for their patient attention to and valuable feedback about many, many practice presentations. Finally, Eric Deeds, for his patient mentorship and support.

Chapter 1 is a version of a preprint; A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-seq data. The authors of that work are myself, Timothy Hamilton, Samuel Aragones, Christian Ray, and Eric Deeds. Chapter 2 is a version of a soon-to-be-posted preprint, Unbiased analysis of scRNA-Seq data reveals cancer stem cells in small cell lung cancer cell lines. The authors of that work are myself, Serena Hughes, Melanie Tu, Timothy Hamilton, and Eric Deeds. Chapter 3 is also a version of a work that is yet to be posted as a preprint; Deep Embedder: Deep Neural Networks for Dimensionality Reduction. The authors of that work are myself, Timothy Hamilton, Samuel Aragones, Sarah Maddox Groves, and Eric Deeds,

Curriculum Vitae

Shamus Cooley attended the University of Kansas and received a B.A. in Mathematics from the University of Kansas in 2017. He attended graduate school at the Center for Computational Biology at the University of Kansas from 2017-2019 where he worked as a researcher in the lab of Dr. Eric Deeds. In 2019, The Deeds lab moved to UCLA, where Shamus has worked as a researcher until the date of this publication.

Chapter 1: A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data.

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[14,15,16,17], 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. This type of data typically undergoes the following process or a variation thereof: First, an arbitrary number of highest-varying genes are selected from the dataset. These genes are analyzed, while the genes that vary less between samples are disregarded. Next, the HVGs are represented in a lower-dimensional space with PCA. The number of principal components chosen for this step is also arbitrary but can be based on inspection of a scree plot. Finally, the data is visualized in 2 or 3 dimensions using t-SNE or UMAP. In our review of the literature, we have found that some groups perform further quantitative analysis on the PCA

representation of the data, while others perform analysis on the 2 or 3 dimensional embeddings given by UMAP or t-SNE. When we measure distortion by calculating AJD, we found that commonly used techniques disrupt *90-99%* of the local neighborhoods in the data *prior* to performing further quantitative analysis. 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 many pseudotime ordering algorithms[20].

Our results demonstrate that, regardless of the technique used to reduce dimensionality, most of the local structure of high-dimensional data is lost when compressed into the number of dimensions typically used for scRNA-Seq analysis. This implies that any analysis based on this kind of 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, 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).

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 between the position of a point in the original space and its embedded position[19,23], without considering the change in relative position between the point and its neighbors. However, quantifying the extent of the loss of structure caused by NDR requires consideration of neighborhoods within the data, not just changes in the positions of individual points. For example, a 2-D representation of the swiss roll might be stretched out, greatly distorting the pointwise distances, 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 an intermediate value for a representation that retains part of the information.

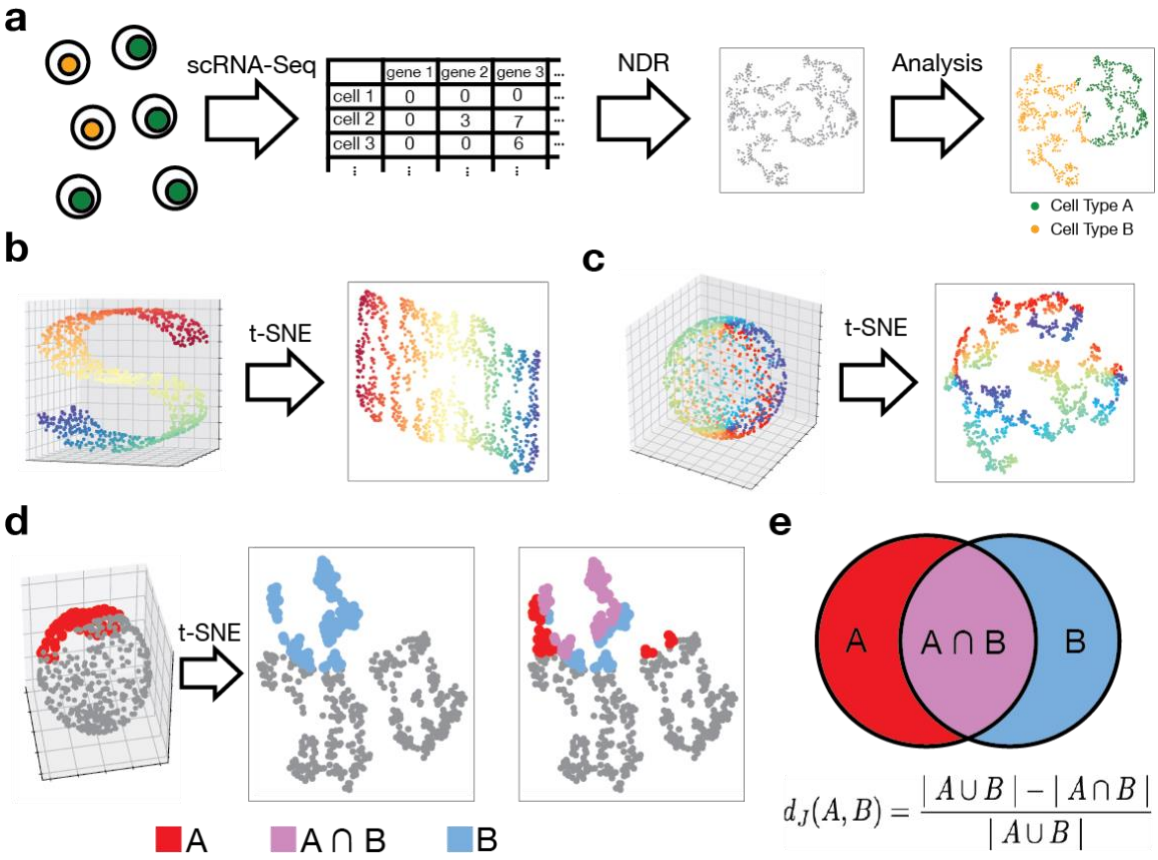


Fig. 1. ***(a)*** *A schematic of some scRNA-Seq workflows. The gene expression data are stored as a matrix, with each row corresponding to a cell, and each column correspond to a gene (after correcting for UMI swapping). The data undergo dimensionality reduction, and analysis is performed on the lower-dimensional representation of the data.* ***(b)*** *The "swiss roll" data set. t-SNE can 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.*

Testing on Synthetic Data

To test the usefulness of the 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).

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). 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 (Supplemental Figure 1). 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 (Figs. 2D & 2E). 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]. We followed a typical dimensionality reduction workflow. Namely, we first selected 5000 HVGs, reduced the dimensionality of this subset with PCA using 45 principal components. (The number of PCs was selected by inspection of a scree plot). For this dataset, we selected one of the largest cell type clusters defined in the 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.

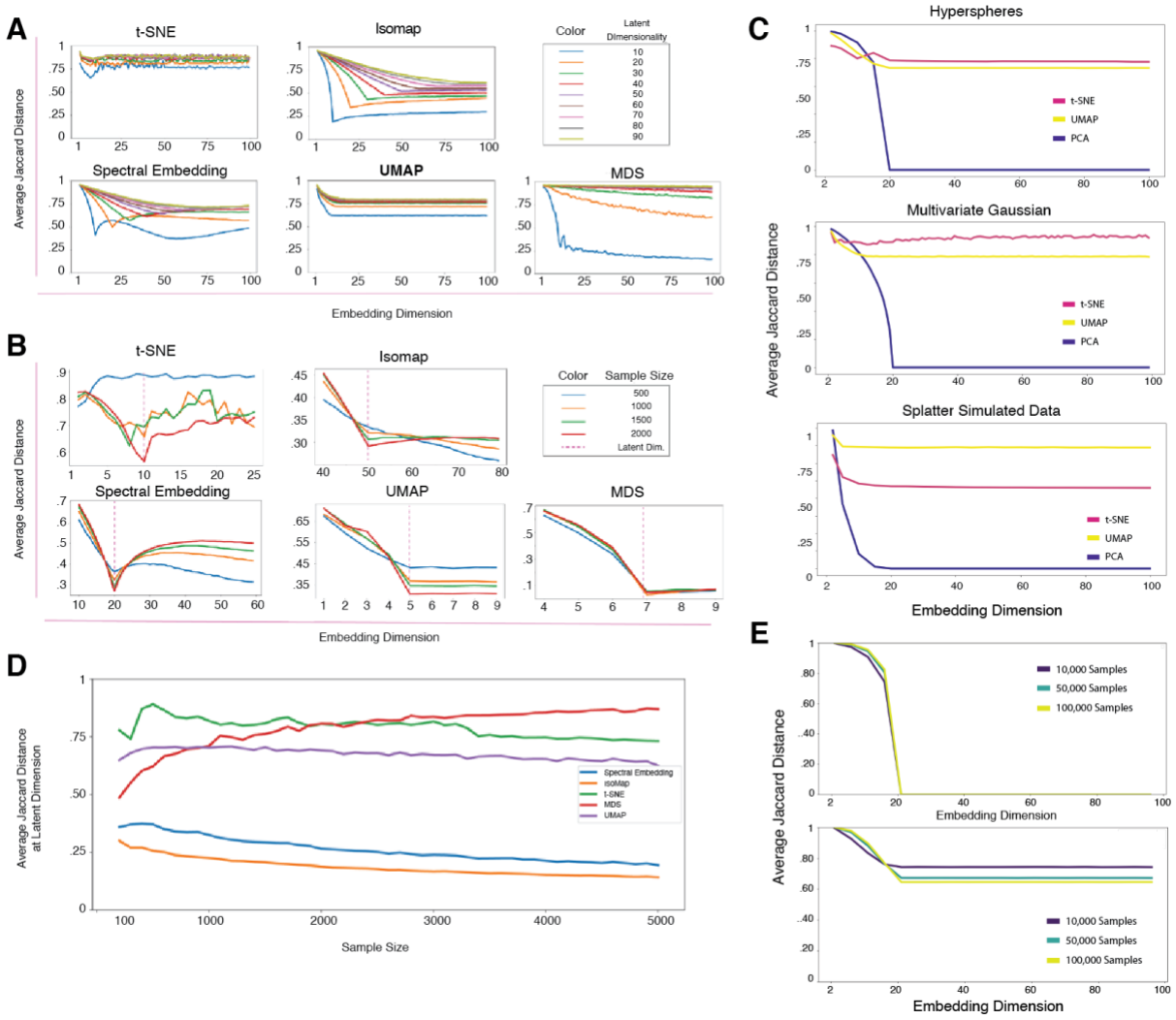


Fig. 2. (A) *The Average Jaccard Distance (AJD) for points randomly sampled from the surface of hyperspheres of varying dimension embedded in dimensions 1-100. The AJD is lowest when the latent dimensionality of the manifold is lowest.* ***(B)*** *The effect of sample size on Average Jaccard Distance. 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)*** *AJD for varying high-dimensional geometries. Three simulated 20-Dimensional datasets, hyperspheres, multivariate gaussians, and virtual scRNAseq data simulated by the Splatter package, are each embedded into spaces of dimension varying from 2-100. The AJD is calculated for each embedding.* ***(D) AJD vs. Sample size.*** *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.)* ***(E)*** *Large Sample Sizes. Datasets are sampled from a 20-dimensional hypersphere and embedded in spaces of varying dimension. Increase the size of the sample does not alleviate the distortion introduced by dimensionality reduction.*

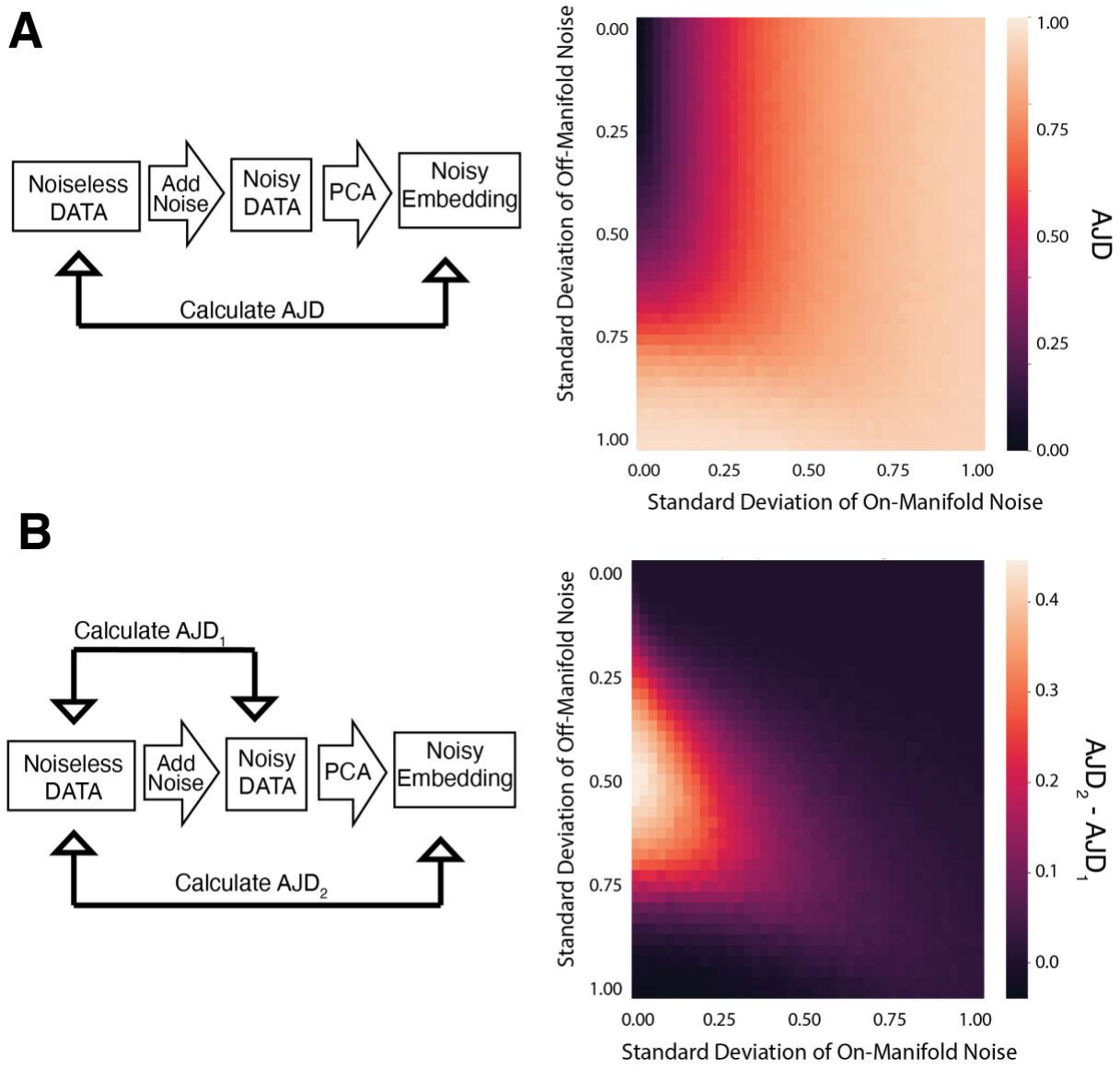


Figure 3. Does PCA Filter Noise? **(*A*)** *A 20-dimensional multivariate gaussian "cloud" is simulated in a 100-dimensional space. Noise is added to the dimensions containing the manifold (on-manifold noise) as well as to the dimensions not containing the manifold (off-manifold noise). The noisy data is embedded with PCA, and the Average Jaccard Distance is calculated between the raw data and the embedding. The experimented is repeated with a range of standard deviations for the added noise, both on and off the simulated manifold. The Average Jaccard Distance changes little, indicating that PCA is unable to remove the noise except in*

cases where the noise is very small. ***(B)*** *Again, a 20-dimensional multivariate gaussian "cloud" is simulated in a 100-dimensional space. Noise is added to the dimensions containing the manifold (on-manifold noise) as well as to the dimensions not containing the manifold (off-manifold noise). The noisy data is embedded with PCA. The Average Jaccard Distance is calculated between the raw data and the embedding, as well as between the raw data and the noisy data. The experimented is repeated with a range of standard deviations for the added noise, both on and off the simulated manifold. The heatmap displays the difference between these two measurements of distortion. The Average Jaccard Distance changes little, indicating that PCA is unable to remove the noise.*

In order to confirm that the observed distortion wasn't unique to the these two studies, we next selected a wide variety of scRNA-seq studies from a diverse set of model organisms, both vertebrate (Fig. 3e) and invertebrate (Fig. 3f) and repeated our analysis in Seurat, using the dimensionality reduction techniques PCA and UMAP (Fig. 3e). In every case, the distortion introduced by UMAP was substantial, and the technique consistently failed to find a low-distortion embedding even in higher dimensions. The performance of PCA varied from data set to data set, but often needed well over 100 dimensions to represent the data with low levels of distortion (e.g. AJD < 0.05),

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 analyses are.

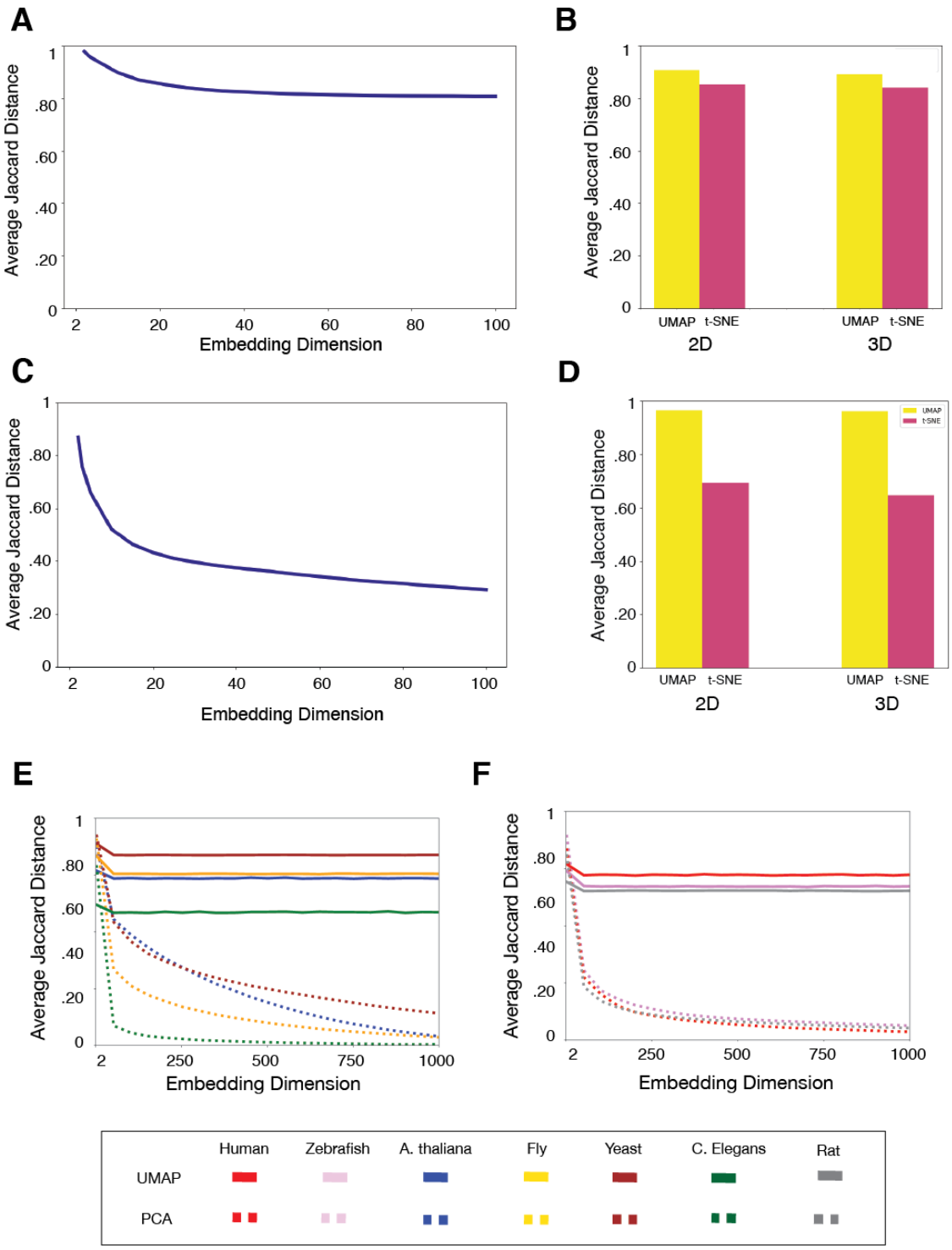


Fig. 4. Distortion introduced by dimensionality reduction in scRNA-Seq. (A) *The entire dataset from Siebert et al. undergoes typical dimensionality reduction. First, the 5000 most*

highly varying genes (HVGs) are selected. PCA is performed using a number of principal components varying from 2-100. AJD is calculated between each of these embeddings and the raw data. ***(B)*** *The "best" PCA representation of the data according to a scree plot (45 PCs) undergoes dimensionality reduction via UMAP and t-SNE. AJD is measured between these embeddings and the original, raw data.* ***(C)*** *A single cluster in the hydra dataset (as identified by the authors) undergoes selection for HVGs and the PCA into spaces of dimension 2-100. The distortion is somewhat less, but still significant.* ***(D)*** *The "best" PCA representation (45 PCs) undergoes nonlinear dimensionality reduction into 2 and 3 dimensions using t-SNE and UMAP. Again, the distortion is less for a single cluster.* ***(E)*** *Average Jaccard Distance vs. Embedding Dimension for Invertebrate scRNA-Seq studies.* ***(F)*** *Average Jaccard Distance vs. Embedding Dimension for Vertebrate scRNA-Seq studies.*

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 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. Although it is not always the case, 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 typical analysis pipelines applied to a wide variety of data sets. We used the Seurat package in R 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[2824]. For each study we used the same embedding dimension for PCA as was used by the original investigators. We then reduced the data to 2 dimensions with UMAP and computed the AJD between each step in the pipeline.

Study	Model Organism	Number of PCs	AJD after "De-Noising" with PCA	AJD after UMAP
Siebert et al.[24]	*Hydra vulgaris*	31	0.87	0.92
Jean-Baptiste et al.[38]	*Arabidopsis thaliana*	25	0.75	0.81
Farrell et al.[40]	*Danio rerio* (Zebrafish)	97	0.90	0.92
Taylor et al.[41]	*Caenorhabditis elegans*	125	0.94	0.95
Davie et al.[42]	*Drosophila melanogaster* (Fruit Fly)	82	0.94	0.95
Ma et al.[43]	*Homo sapiens*	20	0.90	0.91
Mays et al.[44]	*Rattus norvegicus*	13	0.99	0.99

Table 1. *Average Jaccard distance (AJD) between the minimally processed (raw) scRNA-Seq datasets and the representations produced by dimensionality reduction.*

As expected based on our findings above, each step of dimensionality reduction introduced 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.

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[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 (Table 2). Because clustering only makes sense in the case where there are multiple distinct cell types, we applied this analysis only to those studies where it was computationally feasible to analyze all cells in the data set. We obtained clusters using the standard procedure in Seurat (see Methods).

Study	**Model Organism**	**ARI: PCA**	**ARI: UMAP**
Siebert et al.[24]	*Hydra vulgaris*	0.61	0.43
Jean-Baptiste et al.[38]	*Arabidopsis thaliana*	0.53	0.45
Jackson et al.[39]	*Saccharomyces cerevisiae* (Yeast)	0.25	0.14
Farrell et al.[40]	*Danio rerio* (Zebrafish)	0.12	0.09
Taylor et al.[41]	*Caenorhabditis elegans* (Worm)	0.31	0.23
Ma et al.[43]	*Homo sapiens* (Human)	0.36	0.21
Davie et al.[42]	*Drosophila melanogaster* (Fruit Fly)	0.27	0.12

Table 2. *Adjusted Rand Index (ARI) between clustering performed on the minimally processed (raw) scRNA-Seq datasets and clustering performed on representations produced by dimensionality reduction. In each case, the number of PCs used for PCA is the same as in the original study, and UMAP into 2 dimensions is performed downstream of PCA. In every case, the clustering is substantially different after PCA, and even more dissimilar after UMAP.*

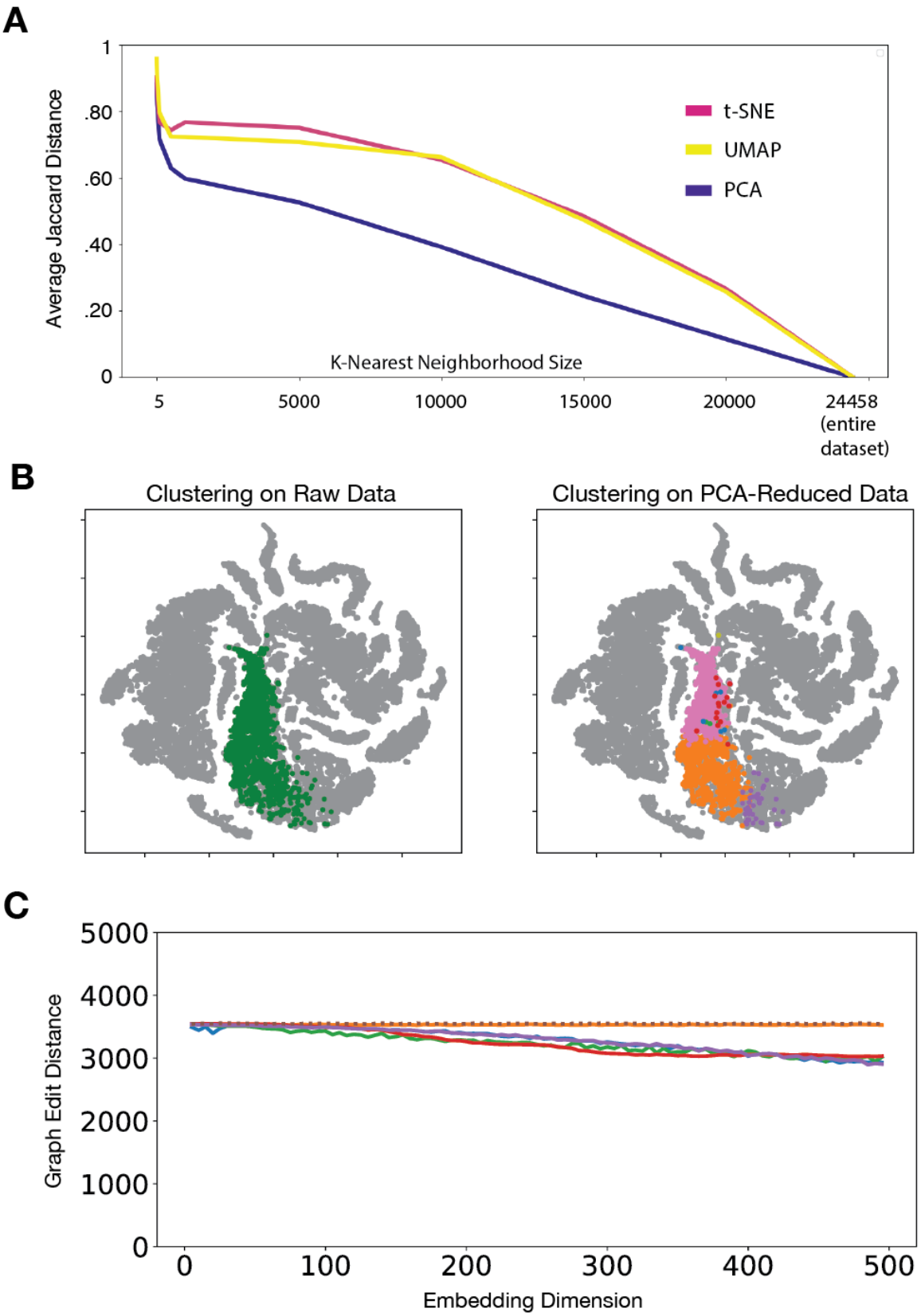

A
Average Jaccard Distance
1
.80
.60
.40
.20
0
t-SNE
UMAP
PCA
K-Nearest Neighborhood Size
5
5000
10000
15000
20000
24458
(entire
dataset)
B
Clustering on Raw Data
Clustering on PCA-Reduced Data
C
Graph Edit Distance
5000
4000
3000
2000
1000
0
0
100
200
300
400
500
Embedding Dimension

Fig. 5. (A) *Distortion vs. neighborhood size. A single cell RNA sequencing dataset is filtered for highly varying genes. The data is then embedded into a 45 dimensional space using PCA. (The choice of 45 principal components was based on inspection of a scree plot) The data is then embedded into 2 dimensions using t-SNE and UMAP. Average Jaccard Distances are calculated between the raw data and the PCA embedding, as well as between the raw data and the 2-dimensional embeddings using various values for the k-nearest neighbor search.* ***(B)*** *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.* ***(C)*** *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.*

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 values between the clusters obtained from raw data and the clusters based on the PCA-reduced data indicates significant differences between the clusters in every case. This effect is visualized in Fig. 4b 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. 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. 4c). 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. 4c). 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, our findings suggest that distortion caused by NDR could have a large effect on the results.

Finally, to determine whether the distortion that we observe is unique to scRNA-Seq data, we measured the distortion caused by dimensionality reduction on several standard machine learning data sets (Table 3). In every case, substantial distortion was observed to have been introduced by dimensionality reduction, leading us to conclude that commonly used dimensionality techniques, both linear and nonlinear, are prone to introducing distortion into local neighborhoods and thus distort the structure of the data.

Dataset	Dimensionality of Data	t-SNE	UMAP	PCA
Abalone (Sea snail) Age Prediction[50]	8	0.33	0.48	0.45
Sonar Object Classification[47]	60	0.53	0.56	0.61

Banknote Authenticity Prediction[48]	4	0.26	0.34	0.52
Human Subpopulation Diabetes Prediction[46]	8	0.46	0.51	0.63
Iris Classification[49]	4	0.26	0.33	0.24
Ionosphere[51]	34	0.53	0.60	0.61
Wine Quality Prediction[45]	11	0.42	0.54	0.57

Table 3: *Distortion caused by dimensionality reduction on some standard machine learning datasets. In every case, dimensionality reduction into two dimensions introduces substantial distortion into the data.*

K-Nearest Neighbors	Minimum Distance					
	0	0.1	0.25	0.5	0.8	0.99
10	0.785	0.779	0.778	0.797	0.819	0.830
20	0.798	0.789	0.789	0.802	0.824	0.836
50	0.814	0.805	0.807	0.813	0.833	0.847
100	0.829	0.817	0.816	0.822	0.839	0.853
200	0.842	0.831	0.827	0.831	0.848	0.860
400	0.847	0.836	0.834	0.839	0.854	0.864

Table 4. Grid Search on UMAP parameters. *UMAP was used to embed the hydra dataset from Siebert et. al. into 2 dimensions with various values for the parameters (neighborhood size and minimum distance).*

Learning Rate	Perplexity					
	5	25	50	100	200	400
12.5	0.836	0.948	0.948	0.936	0.948	0.948
25	0.838	0.774	0.716	0.654	0.948	0.948
50	0.851	0.825	0.835	0.911	0.948	0.948
100	0.855	0.794	0.76	0.763	0.795	0.814
200	0.883	0.846	0.783	0.706	0.662	0.64
400	0.899	0.925	0.908	0.885	0.903	0.896

Table 5. Grid Search on t-SNE parameters. *The hydra dataset was embedded into two dimensions using the t-SNE algorithm with various values for perplexity and learning rate.*

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:

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

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* uniformly distributed samples in 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 standard normal distribution *n* times (using the Python Numpy method numpy.random.normal(0,1)). This method ensured that the sampling on the sphere was 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.deomposition.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:

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

Where $P(\text{mst}_1, \text{mst}_2)$ 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 Workflow

To replicate a typical workflow, we used Seurat in R. 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. 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. 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 several 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.

Our findings, and the recommendations above, might at first glance seem to be in conflict with the fact that most scRNA-Seq studies ultimately produce results that are broadly consistent with orthogonal data regarding the system under study. For instance, t-SNE and UMAP plots still tend to place cells of similar type close to one another. This is often checked by coloring cells according to the expression of marker genes on that are known to be associated with certain cell types, and finding that those cells tend to cluster together, at least on visual inspection[24,25]. Similarly, pseudotime analysis often results in expression dynamics that broadly correlate with known expression dynamics obtained from other techniques[24,25].

While this agreement seems reassuring, there is a subtle issue with this kind of analysis.

Each of the dimensionality reduction techniques mentioned above are governed by one or more parameters. A small adjustment in any of these parameters can result in vastly different representations of the data (Supplementary Fig. 6). How does one decide the appropriate values

for the parameters? In practice, one first selects marker genes that they know correspond to certain cell types based on previous studies. The expectation in this case is that the analysis pipeline, which entails several steps of dimensionality reduction, will have been executed correctly when the marker genes cluster according to prior knowledge. Adjusting the parameters of the algorithm until agreement is achieved, the researcher concludes that these are the correct parameter values, and this is the correct representation because the result has been "validated" by prior knowledge. Other observed clusters can then be interpreted as representing new cell types. Popular packages, such as Seurat, include suggestions along these lines for users in their documentation.

The problem with this approach is that it is inherently biased to reproduce known aspects of the system in question. To see why, suppose that the biological ground truth doesn't agree with prior biological knowledge. The researcher will discard such a result and adjust the parameters of the analysis pipeline until the representation comes into agreement with their expectations. In other words, if prior knowledge is used to guide the analysis, the fact that one ultimately sees agreement between the result and that prior knowledge is no guarantee that the analysis itself is sound. This is true even if the marker genes used to guide clustering or other analysis are different from the ones used for "validation," since it is unlikely that any such sets of genes will be truly independent of one another. Thus, while many scRNA-Seq analysis agree with well-established prior knowledge, that in no way guarantees that distortion due to dimensionality reduction has not significantly impacted the analysis.

Of course, one question raised by our results is whether 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. We note, however, that NDR techniques failed to find meaningful embeddings even for non-scRNA-Seq data (Table 3), strongly suggesting that the issue here lies with the techniques themselves, rather than representing limitations of the individual data sets. The only technique that we found to provide something close to a "true" embedding, PCA, does so only at dimensionalities that are much larger than those typically used. Indeed, PCA sometimes only finds a true embedding at the largest possible dimension 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 not only in cell biology, but across all scientific disciplines that rely on the analysis of high-dimensional data. 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.

Chapter 2: Unbiased analysis of scRNA-Seq data reveals cancer stem cells in small cell lung cancer cell lines

Introduction

It is commonly accepted that that heterogeneity of cancer cells is a key factor by which tumors resist treatment[36,37]. It has been shown that certain transcriptionally distinct subtypes of SCLC are more likely to survive chemical treatment and proliferate[38,39]. Understanding the transcriptional variation between subpopulations of cells can is likely to suggest new treatments as well as further our understanding of the biology of cancer. Therefore, how to best classify SCLC subtypes has been a much-debated question in the field.

The first observation of SCLC heterogeneity was published in 1985, when biochemical characterization of 50 SCLC cell lines revealed that SCLC cell lines could be subdivided into two distinct classes[40,41]. Little progress was made until 2013, when gene expression profiling of the treated tumors revealed that one of these classes showed higher expression of ASCL1 and the other displayed high expression of NeuroD1[42]. Soon after, a third classification was added to the scheme that was characterized by a non-neuroendochrine nature and low expression of both ASCL1 and NeuroD1[43,44]. This third subtype was also found to be heterogenous, with one subpopulation demonstrating high expression of YAP1[45] and the other high expression of POU2F3[46]. Additionally, the MYC family of oncogenes have been found to be over-expressed in a mutually exclusive manner, with MYC being highly expressed in one population, and MYCL being overexpressed in another. These observations were combined into a proposed classification scheme in 2020 by Poirier and colleagues[47].

Another, competing classification scheme exists, and was proposed by Wooten et al. This scheme draws on RNA-sequencing data to construct a Boolean network model of gene regulatory network reconstruction[48]. This classification relies on global characteristics of the gene regulatory network as reconstructed from single cell data. This scheme was further modified by Groves et al. to consist of a continuum of "archetypes" rather than static classifications. This scheme is largely based on results of Archetypal Analysis[49] and RNA velocity[50].

Here, we attempt to characterize heterogeneity in an unbiased way from single cell RNA sequencing data without gene network reconstruction. The most appropriate way to analyze single cell data is still an active area of debate[51,52]. Our previous studies have shown that mainstream methods of analysis for single-cell RNA sequencing suffer from serious shortcomings[53]. Namely, most analysis workflows employ one or more methods of dimensionality reduction to reduce the computational expense of analysis[51]. As described in Chapter 1 of this dissertation, we developed an objective measure to quantify this distortion. This metric, Average Jaccard Distance (AJD), is calculated by comparing the k-

nearest neighbors in the original space with the k-nearest neighbors in the lower-dimensional representation produced by dimensionality reduction. An AJD of zero indicates that the neighborhoods are the same, and that no distortion has taken place. An AJD of one indicates that all of the neighbors in the lower-dimensional representation are different, and that the information contained in the original data is lost. When we applied this metric to several recent published studies, we found that the distortion introduced by dimensionality reduction was often as high as .95, and rarely better than .70, meaning that the majority of information contained in the data is mostly lost when these techniques are employed.

Here, we analyze single-cell RNA sequencing (scRNAseq) data from eight immortalized SCLC cell lines. To avoid the high level of distortion introduced by dimensionality reduction, we removed dimensionality reduction from our quantitative analysis pipeline, and instead use dimensionality reduction for visualization only. Although this approach is more computationally expensive, it results in an undistorted view of the data. We used Louvain clustering to provide an unsupervised classification of each of our eight SCLC cell lines. In each cell line, we observed 4-6 clusters. On further analysis of these clusters, we found that each had a transcriptionally distinct gene expression profile. We analyzed gene expression with respect to signaling pathways suspected to be involved in progression of SCLC and found that the signaling pathways were expressed in a highly heterogenous way across clusters. We also investigated expression of genes associated with proliferation and found that at least one cluster in each cell line expressed these genes more highly. Finally, we employed an algorithm, CytoTRACE[54],that is trained to identify the stem-like quality of cells, we found that our more proliferative cluster also had a strong stem-like quality, and that this subpopulation may be more proliferative than the population as a whole, and could be a promising target for new treatments.

Results

Canonical Marker Genes

Since an active area of debate in the study of SCLC is how to best characterize sub-types, we began our analysis by examining the distributions of previously proposed sub-type marker genes in each

of our cell lines (Fig. 1). Interestingly, POU2F3 and YAP1 are expressed similarly, despite being previously identified as a marker for distinguishing subtypes[45,46]. MYCL is typically highly expressed as well in these cell lines, despite being previously thought to correspond only to the ASCL1 subtype. MYC, which was previously thought to correspond to the NEUROD1, POU2F3, and YAP1 subtypes is not here expressed in high levels with those marker genes. Overall, the distributions of each gene are highly multimodal, which indicates that the diversity in cancer cells is not adequately described by the marker gene model.

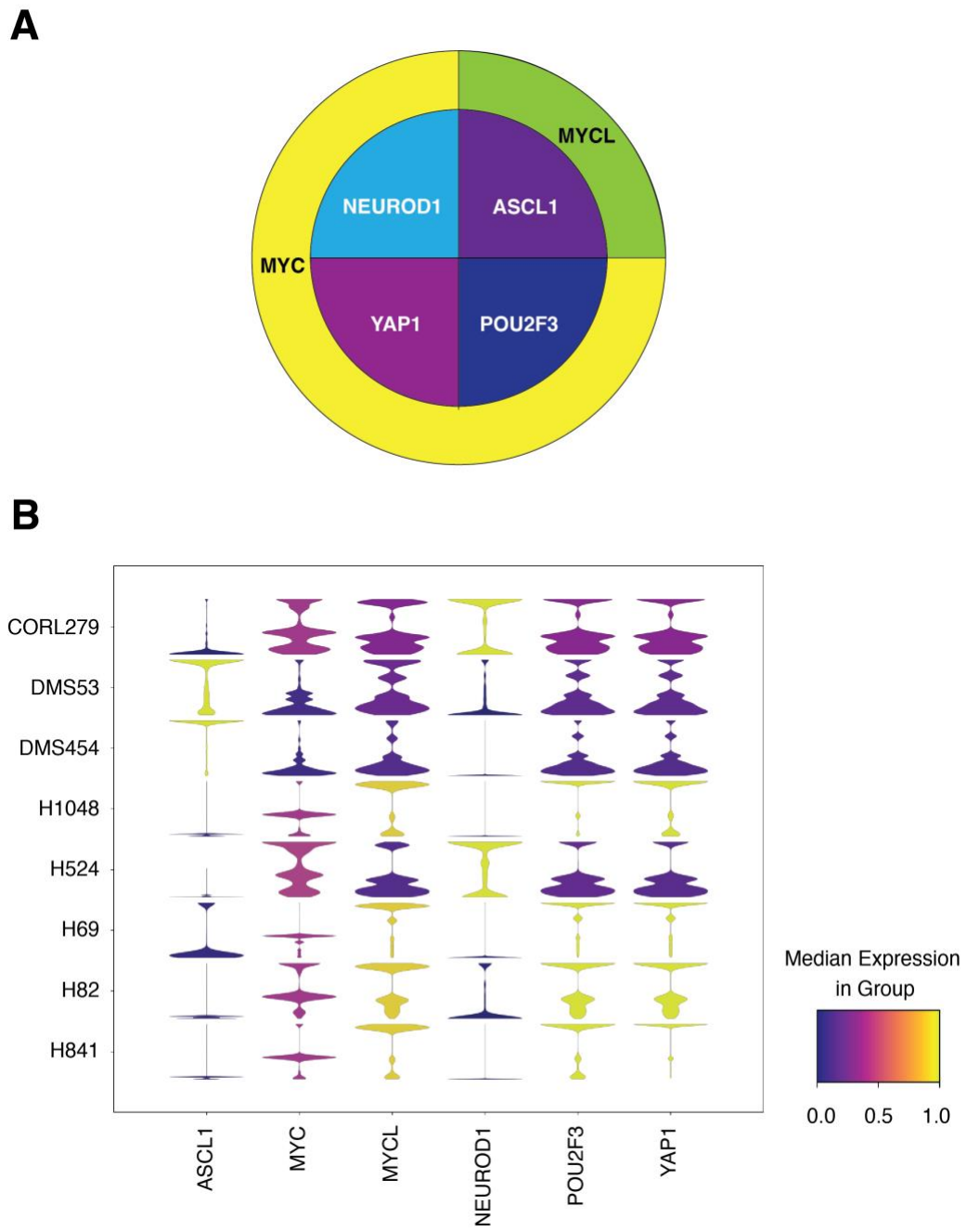


Figure 1. *Subtype marker genes in SCLC cell lines.* ***(A)*** *Four subtypes have been proposed by previous studies, corresponding to NEUROD1, ASCL1, POU2F3, and YAP1. MYC is thought to correspond to the NEUROD1, POU2F3, and YAP1 subtypes, and MYCL is thought to correspond to the ASCL1 subtype.* ***(B)*** *The distribution of previously identified subtype markers in eight SCLC cell lines.*

Unbiased Clustering Reveals Previously Uncharacterized Heterogeneity

Previous results indicated that most forms of dimensionality reduction do not preserve the local relationships between points in high dimensional space[53], thereby confounding any downstream analyses which relies on a nearest-neighbor search. To avoid this, we started our analysis by clustering in the full high dimensional space, upstream of any form of feature selection or dimensionality reduction. For each of the 8 cell lines we analyzed, we clustered the cells in the full high dimensional space using the Louvain Clustering algorithm implemented in scanpy[55,56]. In each of the 8 cell lines, 4-6 clusters were observed. We then used dimensionality reduction via UMAP and t-SNE, solely for visualization (Fig. 1). We observed that in each cell line, there are one or two clusters that take a unique shape whether visualized with UMAP or t-SNE. We also find that the cell lines, when combined, are clearly classified with the unsupervised clustering approach, indicating the validity of this clustering approach.

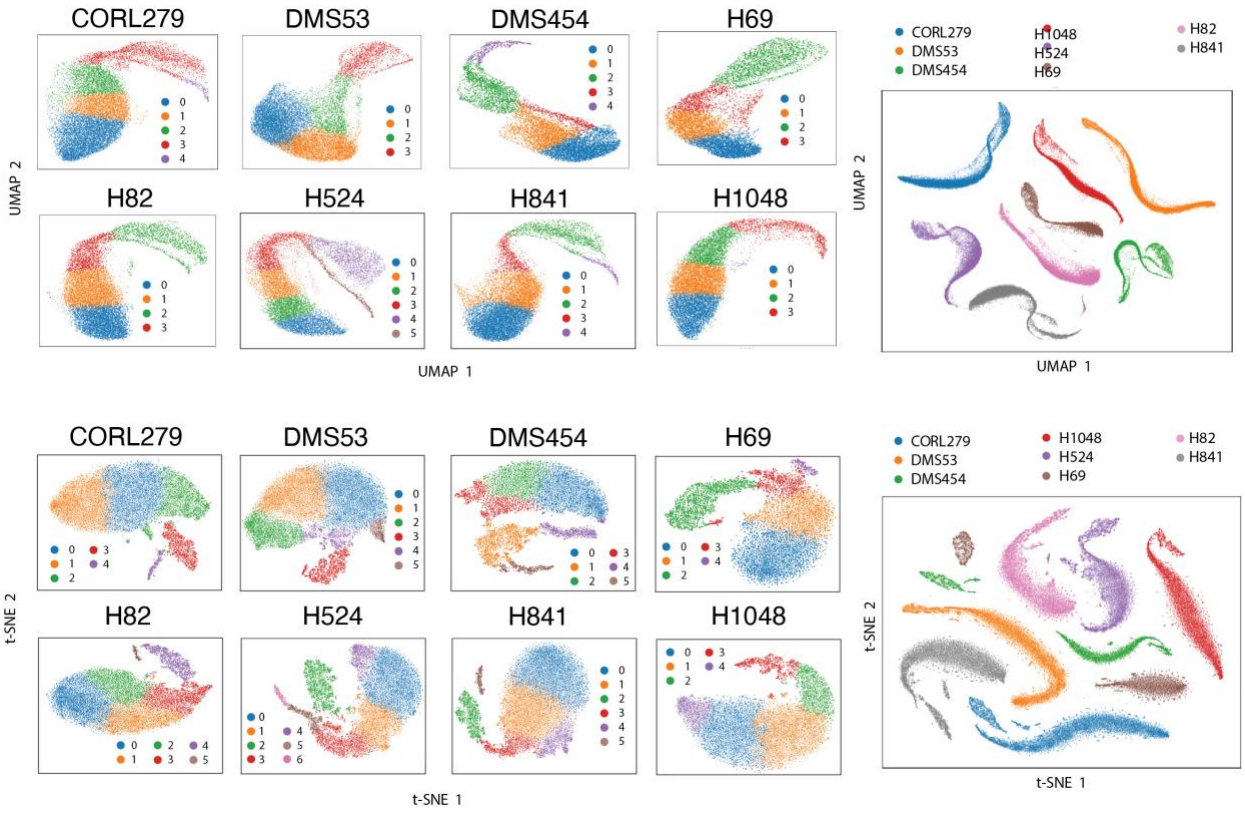


Figure 2. *Louvain Clustering reveals previously unrecognized heterogeneity. (A) Each cell line undergoes clustering using the Louvain clustering algorithm upstream of dimensionality reduction. Clusters are then visualized with (A) UMAP and (B) t-SNE. Data from all cell lines is combined and undergoes Louvain clustering before visualization with (C) UMAP and (D) t-SNE*

WNT and Notch Signaling Heterogenous Across Clusters

Previously, in a whole-exome sequencing study, Wagner et al. found transcriptional up-regulation of wingless-related integration (WNT) pathway genes in chemotherapy-resistant tumors[57]. WNT activation was enriched in patient tumors with low levels of ASCL1 expression. Here, we examined the distribution of WNT pathway genes in each of our eight cell lines (Fig 3). In each cell line, the expression of WNT signaling is heterogeneous across clusters.

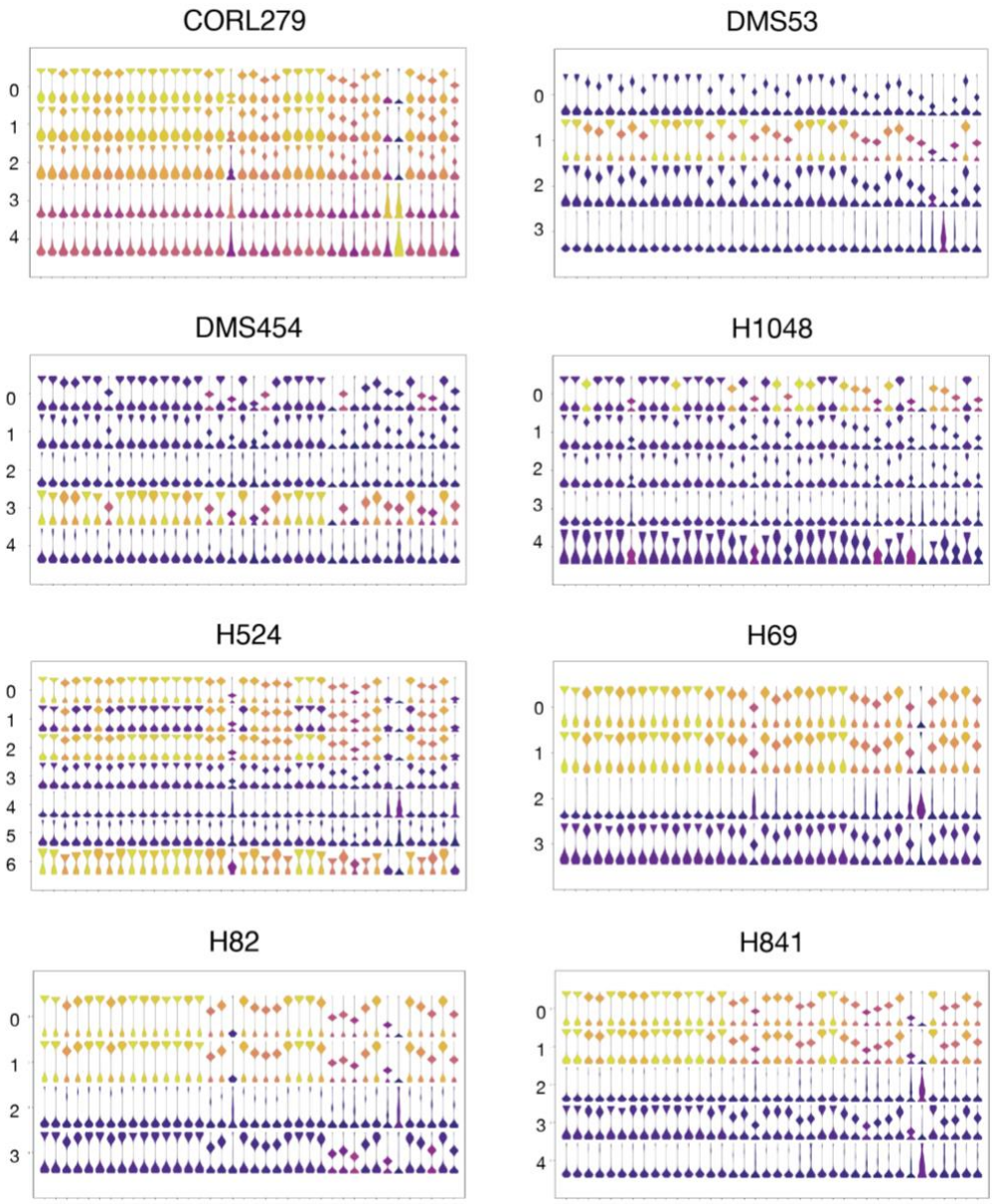


Figure 3. *Distributions of WNT signaling genes. WNT signaling pathway genes are expressed heterogeneously in each cell line.*

Inactivating mutations in the NOTCH family of genes has been found to correlate with SCLC[58]. On the other hand, activation is found in a subset of human and mouse tumors[59]. Additionally, a special class of neuroendocrine stem cells that have a unique capacity for self-renewal are characterized

by NOTCH2 expression[60]. We examined the distributions of NOTCH signaling genes across all clusters in each of our cell lines (Fig 4.)

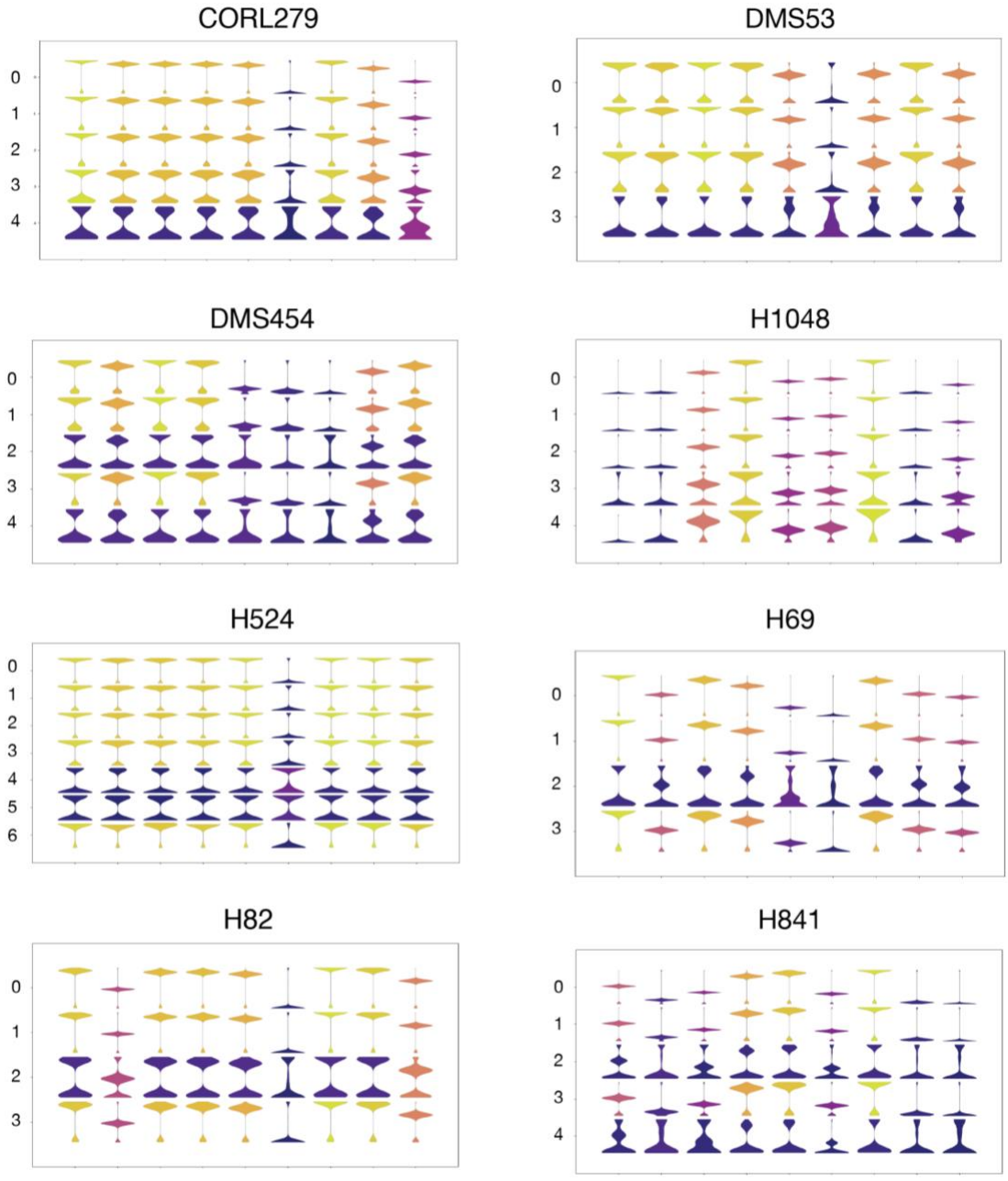


Figure 4. *Distributions of NOTCH signaling genes. NOTCH signaling pathway genes are expressed heterogeneously in each cell line.*

Principal Component Analysis

To determine whether these clusters were transcriptionally distinct, we used Principal Component Analysis (PCA) to embed each of our Louvain-determined clusters into lower dimensions ranging from 1-1000. As a control, we also performed PCA on a random subsample of size equal to each cluster. If our clusters were artifacts of the clustering algorithm, we would expect the values of AJD to be the same for the sample and the control. However, if each of the clusters is transcriptionally distinct, we would expect the AJD at each dimension to be significantly different from one another. We found that each of our clusters was transcriptionally distinct, with one or more clusters in each cell line deviating strongly from the random control. (Fig 5)

For two different sets of data, if the PCA determined components are different between the two sets of data, then one set of components will not easily embed the data of another without significant distortion. To further determine whether the orientation of the local neighborhoods of these specified clusters were unique relative to the rest of the dataset, we determined the PCA components using the data from our "special" clusters. This provided us with *loadings* (the weights that serve as the points' coordinates in the embedding space). At each embedding dimension, we measured the Average Jaccard Distance between the full gene expression space and the embedding dimension space. We repeated the experiment, switching which cluster was used to make the PCA components and which cluster was used to make the loadings in the low dimensional space. In both cases, we found AJDs that were consistently higher than those found by finding the PCA components of a cluster and applying those components to the same cluster to find the loadings. This implies that the Principal Components of each cluster are unique to that cluster, and this in turn implies that each cluster must be transcriptionally distinct from the others.

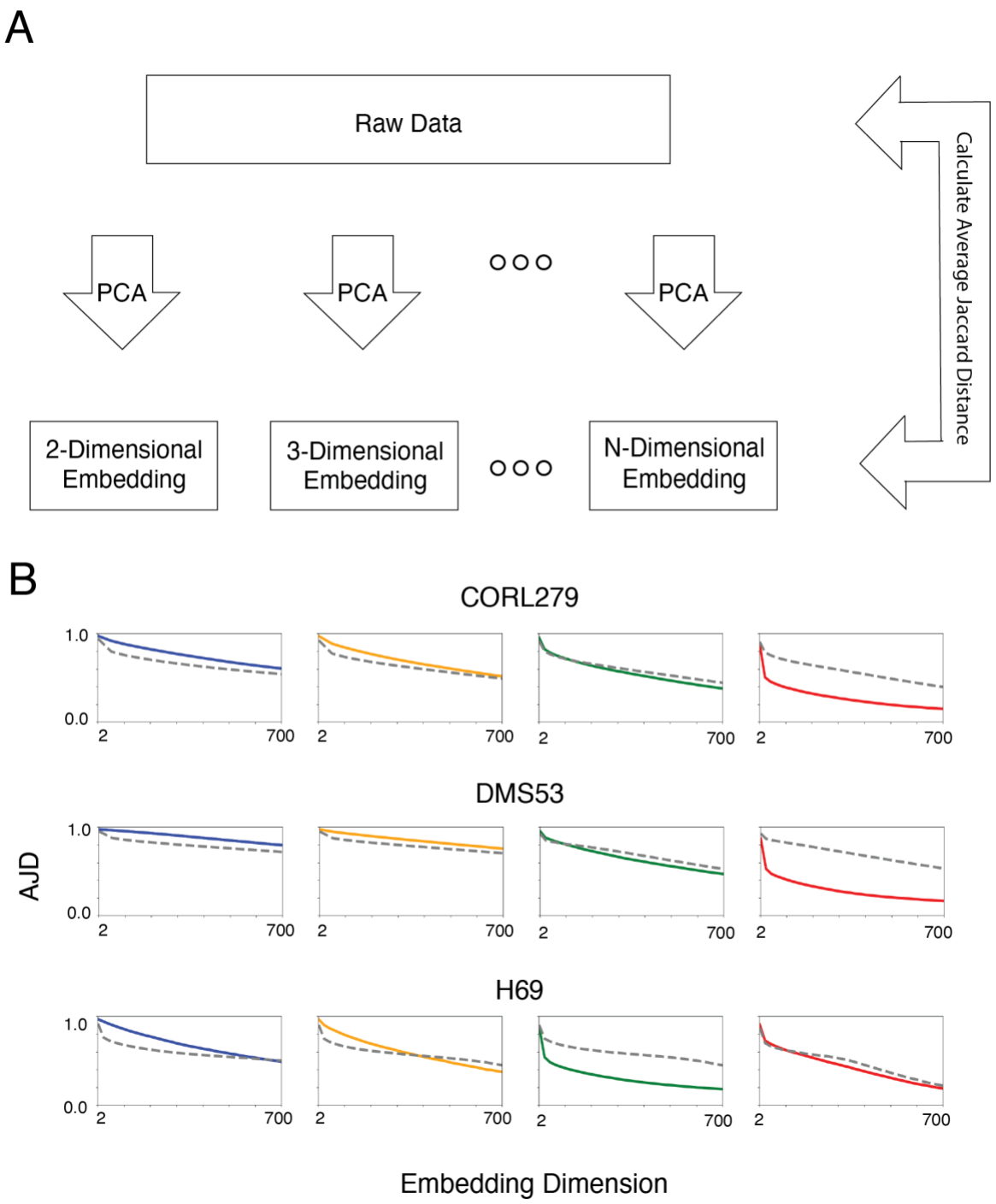


Figure 5. *Principal Component Analysis of SCLC subtype clusters. (A) The raw data is embedded into varying lower dimensional spaces using PCA. Distortion, as measured by AJD, is calculated between the raw space and the embedded representation. (B) In each cell line one or two clusters have a lower*

distortion when compared to a random subsample of the same size. This indicates unique gene expression profiles for each cluster.

Tests for Epithelial-Mesenchymal Transistion, Cell Cycle, and Chromatin Remodeling

Cells *in vitro* have been commonly observed undergoing Epithelial Mesenchymal Transition (EMT), wherein cells change from being square, planar epithelial cells to spherical motile mesenchymal cells. This transition results in a global change in gene expression structure[61]. In order to be certain that our clusters were the result of natural differences in gene expression state, and not a side-effect of cell culturing, we viewed the expression of EMT-associated genes in each of the cell lines, there was no significant difference between the gene expression of the clusters we identified as unique and the rest of the cells from that respective cell line. Nor was there any significant difference between these special clusters and the rest of the dataset for genes associated with Epithelial and Mesenchymal Cell States. These results implied that at the transcriptomic single cell level, the phenomenon that was responsible for our unique clusters was not correlated with EMT nor the Epithelial or Mesenchymal Cell State.

We considered that these unique clusters may differ from the rest of the cell line as a result of being in a different phase of the cell cycle. We compared the gene expression of the unique clusters with that of the other clusters in each respective cell line, and looked for high expression levels of genes correlated with the four phases of the cell cycle (G1, S,G2, M)[62]. We found no significant difference in this set of genes between the clusters identified as unique and the remainder of the sample. We concluded that the unique topology of these clusters was not a result of these cells occupying a distinct point on the cell cycle.

One of the most well documented phenomena that result in persistent changes to gene expression structure is open chromatin. This is when the DNA that normally is tightly bound around histones is loosened, allowing for significant changes in gene expression that remain so long as the chromatin is open. This process is mediated by chromatin remodeling proteins that modify the histones to weaken the interactions between them and the DNA double helix. While open chromatin can occur during DNA

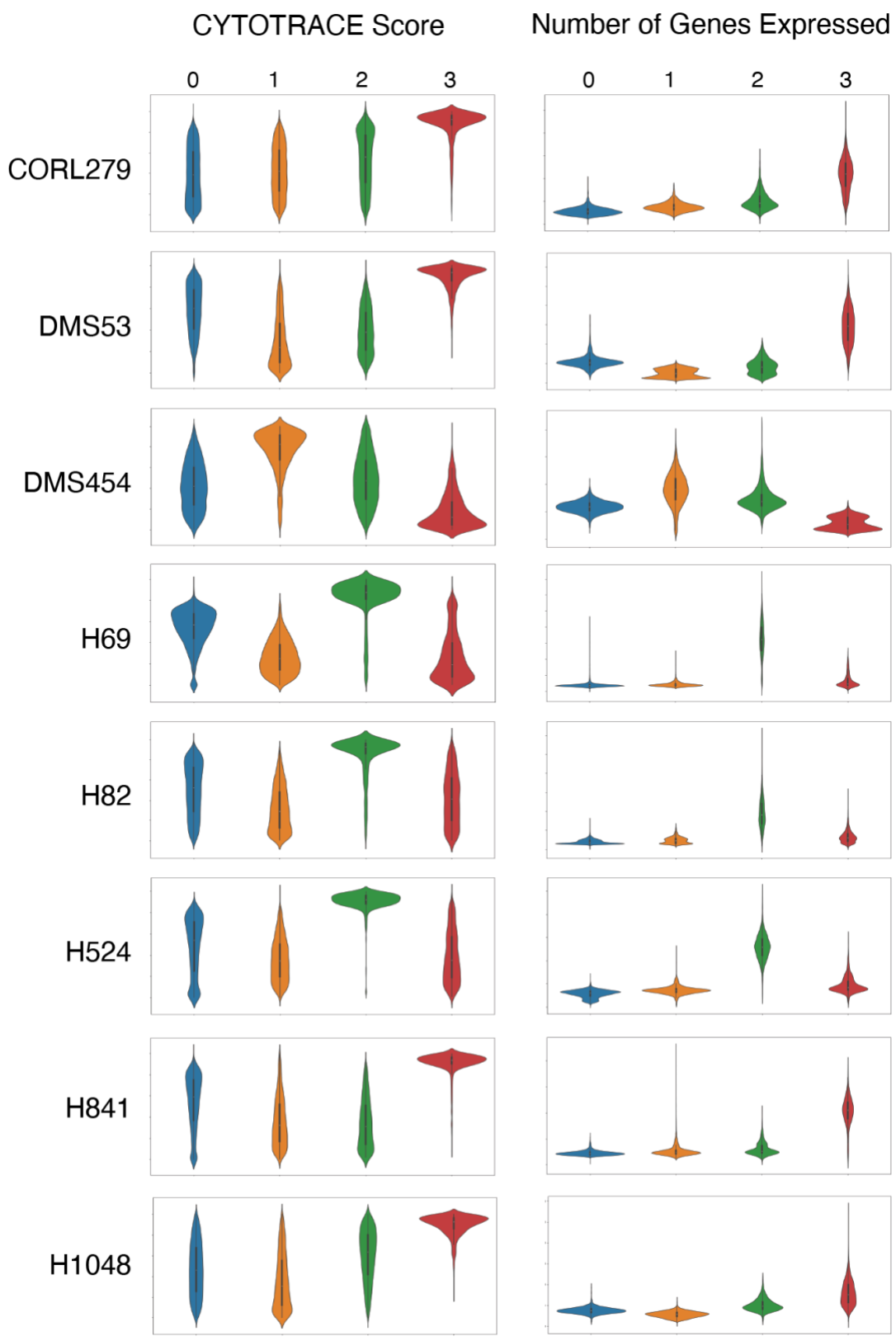
synthesis, it is more persistent in stem and stem-like cells and is responsible for causing the unique gene expression pattern of stem cells[63].

Unfortunately, since scRNAseq is a destructive assay[51], conventional methods to measure the modifications to the chromatin, such as ATAC-Seq, cannot be done on the exact same cells as those used for scRNAseq. Instead, we decided to measure the gene expression for genes involved in Chromatin Remodeling (Fig 5). We found that the clusters demonstrating unique behavior on the AJD Assay had higher expression of genes involved in chromatin remodeling. This over-expression implies that chromatin modification is underway in the cell. To test whether this result was significant, we compared the expression of the chromatin remodeling in the clusters identified as special with the rest of the data set using a Wilcoxon-Rank Sum test. With very few exceptions, we found that the chromatin remodeling genes were always expressed at a statistically significant higher level (p values< 0.00005 after Bonferroni Correction) in the "special" cluster when compared with the rest of the cells from their respective cell line. This result persisted even after we attempted to take into account the sparsity of scRNAseq data, by replacing each UMI count of 0 with the value corresponding to its cell's average gene expression. In addition, the chromatin genes that were significantly more expressed were also expressed by cells in the unique clusters across cell lines. This implies that the chromatin remodeling program is persistent and shared across all eight cell lines.

Stem-Like Cell Clusters

Given these results, we postulated that our cells identified by their unique Average Jaccard Distance vs Embedding structure were in a unique chromatin state. Cells with high potency, which are known as stem cells, have been associated with persistent open chromatin and more exposed DNA[64]. Recent findings also associate a cell's potentiality or "stemness" with the number of genes expressed in its transcriptome[54]. We applied this knowledge in two ways: first, by finding the distribution of the number of genes expressed in each cell in our clusters and by finding the CytoTRACE (an algorithm that measures differentiation potential for a cell based on its transcriptome) score assigned to it (Fig. 6). The

clusters that deviated strongly from the random control in the AJD assay demonstrated more gene expressed per cell and higher CytoTRACE scores.. These results imply that these cells may be stem-like in nature.


CYTOTRACE Score
Number of Genes Expressed
0
1
2
3
CORL279
DMS53
DMS454
H69
H82
H524
H841
H1048

Discussion

Our results from our analysis indicate three key insights regarding Small Cell Lung Cancer and Single-Cell RNA-Sequencing. The first is that there is strong evidence that the current understanding of established sub-types within Small Cell Lung Cancer fails to explain the complexity in heterogenous gene expression. Our results are not fully consistent with either of the previously proposed classification schemes. Our results indicate that our understanding of cell types from Small Cell Lung Cancer must broaden to account for heterogeneity within these cell types. Given that some clusters within each cell line correspond to stem-like behavior and hint at a presence of a unique chromatin state, it is possible that the cells determined to be in the "special" subcluster are stem-like proliferative cells, a difference that could affect treatment plans for patients.

The second insight that our analysis revealed was the importance of clustering before performing dimensionality reduction. As we have shown[53]. The standard for Single Cell RNA Sequencing analysis is to only cluster in a reduced PCA space[51], ostensibly to reduce unwanted variation that might cloud further analysis while retaining the information necessary for those analyses. Yet, doing so introduces the investigator's own biases as to what variation is important or not and thus results could be tainted by confirmation bias, limiting the power of the analysis. Our results demonstrate that significant variation exists that can differentiate cells and those differences could have biological meanings; these biological implications would at the very least, would have been distorted if we clustered in a reduced space.

The third insight we gathered from this analysis was the validation of the idea to use the Average Jaccard Distance with PCA to analyze scRNA-sequencing data and mathematically classify cells. Since PCA is a linear tool, it's behavior on data is well characterized, as it builds a set of basis vectors to capture the most variance. Combined with AJD, PCA can then be used to determine the relative orientation of the basis vectors and thus classify cells on whether they exist on the same manifold in transcriptional space. This opens the door to new approaches for unsupervised analysis of scRNA-seq data that would reduce confirmation bias and allow for new pathways to be discovered and analyzed.

While work needs to be done to develop this method of manifold learning, our results thus far show that this approach is a viable way to resolve hitherto unobserved heterogeneity in transcriptomic data.

Methods

Clustering

Clustering was done in the full gene space after quality control steps were taken using the scanpy package in python. First a k-nearest neighbors graph was made using the scanpy.pp.neighbors() function, with the nearest neighbors parameter set to 20. Then Louvain clustering was done with the scanpy.tl.louvain() function, with the resolution set at 0.2. Each of the cells were labeled with the cluster label ranging from 0 to n-1 for n clusters determined by the algorithm. Each of the clusters was separated form the total dataset and saved separately.

Visualization.

For the t-SNE visualizations, the data matrices generated from the previous step were imported into the python package scanpy, for analysis. The data matrices were normalized with the normalize_total() function in scanpy with the counts_total parameter set to 10000. The data was then log transformed using the log1p() method in scanpy. 2000 highly variable genes were selected with scanpy's highly_variable_gene() function, which uses a binning method to bin the mean and standard deviations before scaling(bin size was set to 20). The data matrix that remained after removing highly variable genes was saved and used to calculate the unsupervised marker genes. After that, PCA was performed on the filtered data matrix, with 50 components selected using scanpy.tl.pca() method. Finally t-SNE was performed on the 50 principal components using scanpy.tl.tsne(), resulting in a 2-dimensional representation. The points representing each cell were colored based upon the obtained with Louvain clustering

Marker Gene Selection:

Supervised marker genes were determined based upon existing literature and data from the UniProt gene ontology database**.** The list of marker genes came for the canonical Small Cell Lung Cancer types were obtained from existing literature[47]. The list of genes involved in Chromatin Remodeling came from the UniProt Gene ontology database, out of which we chose 236.

Unsupervised Marker Genes were identified using scanpy's rank_gene_groups(), applied to the individual datasets after log-transformation and scaling. This method uses a overestimation-corrected t-test to determine if the log fold changes between the cluster and the rest of the changes were significant. These genes were then filtered using scanpy's filter_rank_gene_groups() method, which only selected marker genes based upon the function's default parameters, and stored in the AnnData object associated with the dataset

Heatmap Generation:

For each of the sets of marker genes established through literature search, we created a heatmap using scanpy's heatmap() with default parameters. This utilized the filtered and scaled data matrix that was then normalized where each gene was subtracted by the minimum value for that gene and divided by the maximum value for that gene. The heatmaps for the unsupervised marker gene determination were made using the rank_gene_groups_heatmap() function in scanpy, which used the same set of parameters and implementation as the heatmap() function, only instead using the list of genes marked as significant in the AnnData object associated with the dataset.

Wilcoxon Rank Sum:

The Wilcoxon Rank Sum tests were done utilizing the scipy implementation of the test. First, Each dataset was loaded into a pandas data frame using the read_csv() function. For each dataset, the cluster identified as special was removed from the rest of the dataset. For each individual gene in the chromatin remodeling gene set, the expression values for the special cluster and the rest of the data set. This is done using the loc slicing function in pandas. Then, the ranksums() function in the scipy package

was used, with ties being counted as 0.5. Finally, a Bonferroni collection was applied with a threshold p-value of 0.05 divided by the number of chromatin remodeling genes was used.

Average Jaccard Distance:

The Average Jaccard Distance test was done using the python packages NumPy, sklearn, and pandas. Each cluster for each dataset was loaded into the pandas DataFrame using the read_csv() function. The nearest, neighbor graph for 20 nearest neighbors was then determined using sklearn() neighbors() function with the neighborhoods being determined by the sklearn.NearestNeighbors() function using the ball-tree method. Afterward, the PCA loading was determined using the sklearn.decomposition implementation of PCA (the SVD solver was set to arpack). The PCA was fitted using the specified cluster of the dataset, and applied to that cluster as well, with the number components used set the specified dimension. Once the loadings were calculated, the nearest neighbor graph was calculated for the low- dimensional representation. Finally, the Average Jaccard Distance was calculated between high dimensional KNN graph and the low dimensional representation's KNN graph. This was repeated for each dimension from 1 to the smaller number between the number of genes and the number of cells.

As a control, in each of the datasets, a random sample of cells was drawn without replacement from the total dataset using the sample() method in the pandas package. The size of the sample was set as the number of cells in the cluster being tested. Once the sample was generated, a high dimensional KNN graph (k=20) was generated as before. The data was then embedded into the specified dimension using PCA that was trained and applied on the sample. In the low dimensional space, the 20 KNN graph was constructed as before and the Average Jaccard distance between the 2 nearest neighbor graphs was calculated. This was repeated for each dimension from 1 to the smaller number between the number of genes and the number of cells in the sample.

To compare the PCA components between the clusters, the nearest neighbor graph for 20 nearest neighbors was once again found for the specified cluster of a particular dataset. Then PCA was trained on

the largest cluster of the data and applied to the specified cluster to find it's loadings. In the low dimensional space, the 20 KNN graph was constructed as before and the Average Jaccard distance between the 2 nearest neighbor graphs was calculated. This was repeated for each dimension from 1 to the smaller number between the number of genes and the number of cells in the cluster. This test was repeated on the largest cluster for each dataset but with one significant change: the PCA components were determined using the clusters identified as special and were then applied to the largest clusters in the specified dataset. Average Jaccard Distance was then measured between the KNN graph (k= 20) of the high dimensional data and the low dimensional representation This test was once again repeated form for each dimension from 1 to the smaller number between the number of genes and the number of cells in the special cluster.

Stemness Analysis:

To determine the likelihood of the cells being stem-like, we utilized two known metrics: the number of genes expressed (AKA the gene count score) and the CytoTRACE algorithim[54]. For the gene count score each dataset was loaded into a pandas DataFrame. Then the DataFrame was casted as Boolean with the astype() function and 'bool' as the parameter. This converted the values to 1s if the raw gene expression value was greater than 0 and 0 if the raw gene expression value in the matrix was 0. Then using the sum() function in the pandas packages, the total number of genes that were expressed were added up. The distribution of the totals for each cell was plotted using the Seaborn[65] package using the violinplot() parameter, where the cell scores were separated by cluster identity and colored based upon whether the cluster they were a member of was considered special or not. All other parameters were set to default values.

For CytoTrace, we utilized the eponymous package in R to conduct the analysis. We loaded each of the datasets into and R DataFrame using the read.csv() function in the readr package. After transposing the DataFrame, we used the CytoTrace() function in the CytoTrace Package to conduct the test (enableFast was set to False to allow the whole dataset to be used). The CytoTrace scores were saved and

plotted using the Seaborn package in python with the violinplot() function. The cell scores were separated by cluster identity and colored based upon whether the cluster containing this cell was considered special or not. All other parameters were set to default values.

Chapter 3: Deep Neural Networks for Dimensionality Reduction

Introduction

Our previous work has shown that existing techniques of dimensionality reduction introduce large amounts of distortion, even in situations where a zero-distortion embedding is possible. For example, a 5-dimensional hypersphere in a 20 dimensional space can be easily represented in a five dimensional space by removing the empty dimensions. However, most nonlinear dimensionality reduction algorithms are unable to find this optimal solution, and instead give a sub-optimal solution with a large amount of distortion. Linear embedding techniques such as PCA are able to find the optimal solution in this case but are unable to resolve nonlinear data in a meaningful way. (Figure 1)

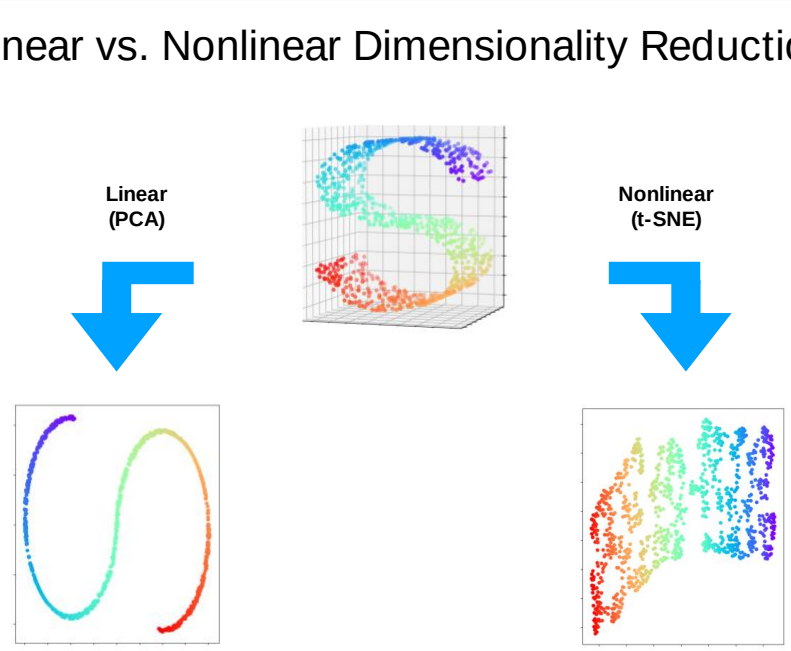


***Figure 1**: Linear vs. nonlinear techniques*

In the case of nonlinear techniques, the distortion is lower, but the algorithm makes no attempt to approximate the embedding function. This results in poor out-of-sample generalizability of the embedding. To illustrate, suppose that 5,000 cells are sequenced and a t-SNE embedding is generated to visualize the data. Then suppose then that an additional 2,000 cells are sequenced to supplement the data. For a new visualization to be generated, the t-SNE algorithm must be run again on the entire set of 7,000 cells, instead of just the new set. The optimal dimensionality reduction solution would be able to solve this problem of generalizability.

The previous chapters have shown that one way to minimize distortion in analysis of scRNAseq data is to minimize the use of dimensionality reduction and to compensate by increasing the computational cost of the analysis. However, in a time when the size and

dimensionality of datasets is increasing exponentially faster, but the annual increase in computational power of computer chips is growing at a lower rate, it is becoming increasingly apparent that this approach also has its limits. Dimensionality reduction is already an irreplaceable step in those datasets where the computational resources don't exist to analyze the data in raw form, and the number of situations will only grow. Dimensionality reduction is therefore an important problem both within the field of genomics, and within the field of biology and other sciences.

Methods

Deep neural networks have found application in a wide variety of fields[66,67,68]. The technique is a method of learning a function that takes a set of data points as input and maps them into the desired output space. In the case of classification, the output space is a set of categories, and the probability that an arbitrarily sample will fall into that category. In the case of regression, the output space is a continuous value of one or more variables. For example, one might train a network to take as input several observations of a population of organisms such as height, weight, etc., and to output the predicted age of each organism.

A neural network can be depicted as a graph of nodes and edges or as a series of matrices. Although the depiction as a graph is more popular, and no less accurate, the matrix representation is often more useful, since it more closely resembled the way that the idea is implemented in the computer. The model consists of matrices the weights and the biases. Each point in the data is multiplied by the first matrix in the set of weight matrices. Then, the first bias matrix is added to the result. Finally, the activation function is applied to each component of the point. These three operations make up the first layer of the neural network. A network can have

any number of layers, each with its own set of weights, biases, and activation functions. After each layer has conducted its operations on the datapoint, the output of the network is compared with the input. This comparison comes in the form of an objective function. The output of the objection is the loss, and this value is minimized or maximized to train the model. There are many variations on this basic architecture, such as recurrent neural networks and convolutional neural networks, but these architectures are outside the scope of this document.

In the context of our problem, the inputs for the neural network are vectors, each corresponding to a cell. Just as in previous chapters, the components of the vectors correspond to genes, with the value of each component corresponding to the count number of that gene. The output space is likewise a vector, of arbitrary dimension, with each value corresponding to some nonlinear combination of the original components. Our strategy is to use the calculation of our metric of distortion, Average Jaccard Distance, as the objective function for the a deep neural network. By minimizing the AJD, we can train a network that gives an embedding that faithfully maps the original manifold from which the data was sampled. However, there are several technical challenges to overcome.

First, Average Jaccard Distance is a measure of similarity between sets of points rather than points. Pre-packaged neural network packages are unable to accommodate this form of loss function. In fact, they are unable to accommodate any objective function that takes more than a single point and its output into consideration. To overcome this obstacle, we were obliged to implement the method ourselves, without relying on existing packages.

Second, the necessity of implementing our algorithm from scratch means that interpreted data analysis languages such as Python or R are not viable, due to the fact that the computational cost of training the model is high, and these languages are geared more towards ease of use than

efficiency. We tested an initial version of our algorithm in Julia and C++ and found the C++ implementation to be as many as two orders of magnitude faster than the comparable Julia implementation. We therefore chose C++ for our final prototyping language, and will likely develop wrappers for Python and Julia, so that other researchers have easy access to our tool.

Third, the Average Jaccard Distance is not a differentiable function, hence analytical calculation of a gradient is impossible. However, we can use the finite difference method of differentiation to approximate the gradient. Our initial results suggest that this method of approximation may also result in more accurate results, as the looser approximation of the gradient somewhat alleviates the problem of getting "stuck" in local minima.

Finally, there is the problem of step size. To efficiently determine the best step size at each epoch, we implemented a training protocol that we've dubbed "spectral step size. This simply means that for each training epoch, after the gradient is calculated, a wide spectrum of step sizes is evaluated, and the training algorithm chooses the step that results in the lowest AJD. Although evaluating each step is more computationally expensive, the resulting increase in accuracy more than makes up for the increased computational cost.

Results

Test Datasets

For some datasets, the deep embedder approach can achieve a lower distortion embedding than either t-SNE or UMAP, which are the most commonly used dimensionality reduction techniques. For example, a common dataset for testing and benchmarking machine learning techniques is the Boston housing dataset[69]. This dataset consists of thirteen observations about homes in the Boston area, as well as the appraised value for the home. To

test our method, we embedded the dataset using t-SNE, UMAP, and the deep embedder approach. (Fig. 1) The deep embedder approach can achieve a lower distortion representation of the data in any of the dimensions as well as higher dimensions.

The wheat seeds dataset is a classification dataset that is also widely used for testing and benchmarking machine learning tools[70]. The dataset consists of seven observations for each weed seed, as well as a species classification. Again, we embedded the dataset using t-SNE, UMAP, and the deep embedder approach. (Fig. 2) The deep embedder approach can achieve a lower distortion representation of the data in any of the dimensions as well as higher dimensions.

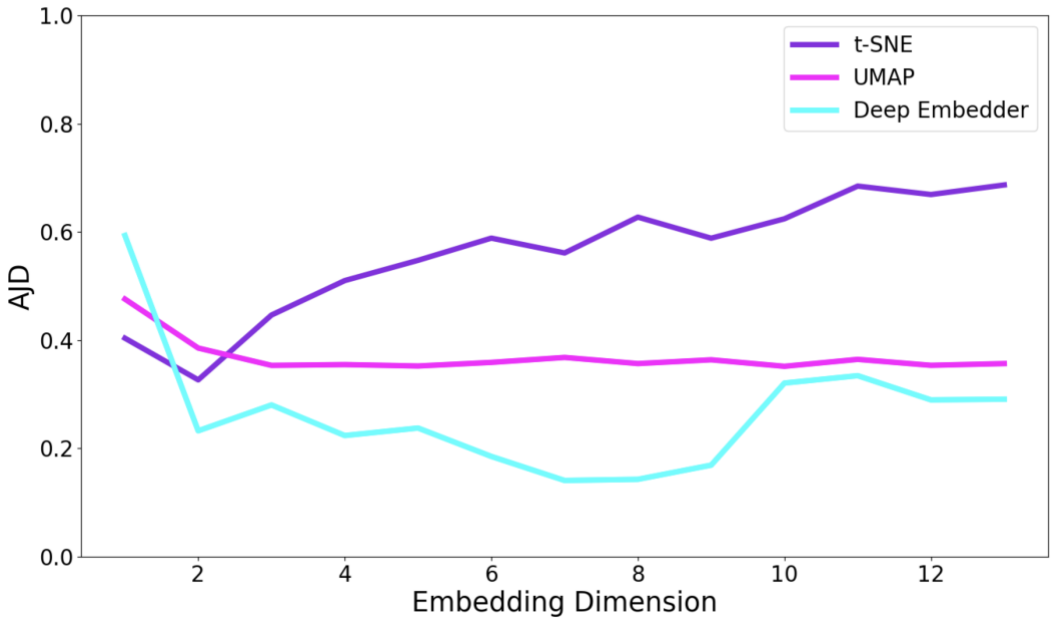


Figure 1. *Dimensionality Reduction benchmarking on Boston housing dataset. Thirteen observations are given of houses in the Boston area. When embedded in dimensions 2-12, the deep embedder approach gives a lower distortion embedding than either t-SNE or UMAP.*

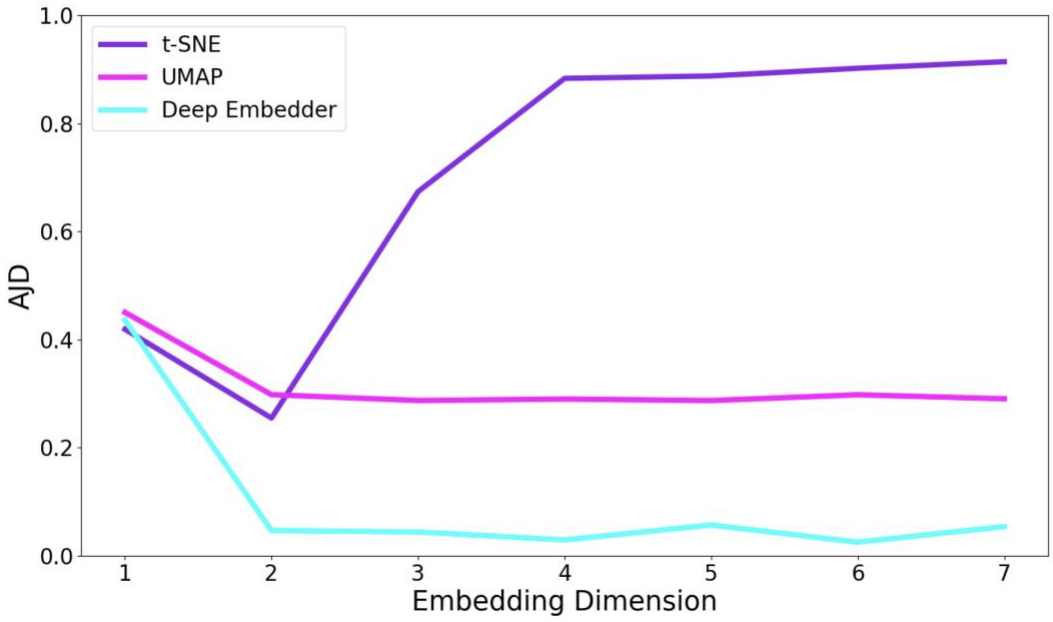


Figure 2. *Dimensionality Reduction benchmarking on wheat seeds dataset. Seven observations are given of wheat seeds. When embedded in dimensions 2-7, the deep embedder approach gives a lower distortion embedding than either t-SNE or UMAP.*

scRNA-seq Data

In a typical scRNA-seq analysis workflow, each cell is modeled as a point in an n-dimensional space, where n is the number of genes that is observed across the entire dataset. A subspace of this original space is typically chosen based on those genes thought to be most significant. The highest varying genes are a popular choice[51]. After this reduction, PCA is usually used as a next step to get the data to a manageable level. Although this linear method most likely destroys the ability of nonlinear dimensionality reduction techniques to recover nonlinear characteristics and portray them in the visualization.

Because this is the case, we compared our deep neural network approach to the performance of t-SNE and UMAP on a subsample of scRNA-seq data from Siebert et al.[71] This sample corresponds to the endodermal epithelial stem cell cluster (as identified by the authors). After being reduced to 40 dimensions via PCA. Interestingly, our approach obtained the lowest distortion results for 3 dimensions. Although t-SNE still produced the lowest distortion visualization in this scenario, we believe that with further refinement of the technique, the deep embedder approach will be able to achieve the lowest distortion in visualization dimensions, while maintaining out-of-sample generalizability that is lacking in t-SNE and UMAP.

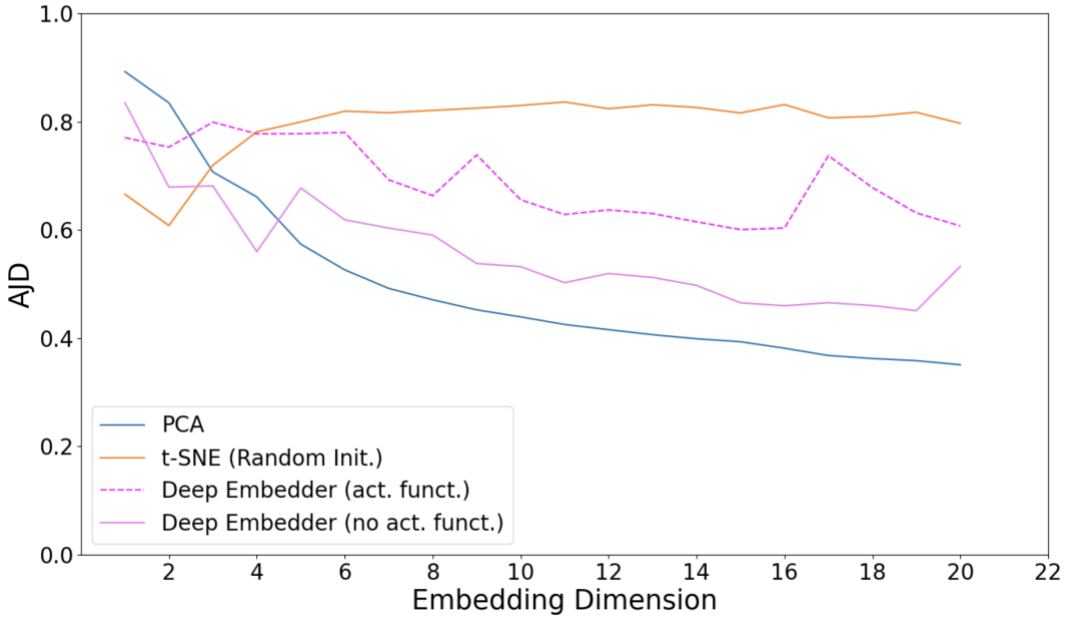


Figure 3. *Dimensionality Reduction benchmarking on scRNA-seq dataset. Single cell sequencing data from Siebert et al. The data is reduced to 40 dimensions with PCA, then that representation is further reduced with t-SNE, UMAP, and the Deep Embedder approach. When*

no activation function is applied, the deep embedder approach gives a lower distortion embedding in 3 and 4 dimensions than either t-SNE or UMAP.

Discussion

Our current implementation of the deep embedder approach uses finite difference method of approximating derivatives. This is necessary, since the function that maps a data set and its embedding to a value of AJD is not continuous or differentiable. Backpropagation, which is a more efficient means of estimating derivatives, requires a differentiable objective function. A major thrust of our future work is to find an approximation of the AJD that is continuous and differentiable. We have made some initial progress in that we have implemented backpropagation for a handful of approximations, and the algorithm's speed is improved by an order of magnitude. However, none of the approximations correlates perfectly with AJD, and we have yet to achieve results that are comparable in terms of distortion to our original implementation.

Another challenge is the fact that our implementation is randomly initialized. More specifically, all the weights and biases in the model are initialized with a random number generator. In other words, our model starts at a random location in the space of possible parameters. Because of the complexity of this landscape, the starting point affects the overall result of the dimensionality reduction. This means that some starting points are more ideal than others. Practically, this means there is a large variance in results when the deep embedder approach is used. It is always possible to repeat the algorithm and select the best result, but it would be ideal to systematically explore the landscape and find a global optimal solution. Broadly speaking, there are two ways to attack this problem: training and initialization.

Initialization refers to the place where we start, and training refers to how well we can navigate our landscape.

First, we can use an initialization strategy that optimizes our initial parameters. Many strategies have been published, and a major area of future work will be to test and evaluate these published initialization strategies. We have had some early success using the so-called "greedy bandit" approach. This approach starts with 1000 random initializations and selects the one that has the lowest AJD when the model is used to embed the data. This adds a small computational cost, but that cost is a fraction of the cost of the algorithm and is theoretically justified by the increased accuracy. Second, we can refine our training algorithm. One method that has shown early promise is a variation on conformational space annealing[72]. By searching the landscape of model parameters more efficiently, we hope to obtain a global optimal result with each run of the algorithm.

To conclude, using deep neural networks as a means of dimensionality reductions shows great promise, and there are a wide variety of strategies and techniques to be drawn from in the literature. This technique has the potential to create lower distortion visualizations of scRNA-seq data and hence clearer pictures than have ever been seen before. It also has the potential to aid in the elucidation of the high dimensional manifold from which the data is originally sampled. This manifold, if properly characterized, will give us a clearer and more complete picture of gene expression than has ever been possible in history of science, which in turn would give us a greater understanding than ever before of how life changes and adapts to the Universe around it.

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. Syst.* **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. *et al.* 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. Approximate nearest neighbors: Towards removing the curse of dimensionality. in *Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC '98* (1998). doi:10.1145/276698.276876.

8. Friedman, J. H. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.* **1**, 55–77 (1997).

9. Pearson, K. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).

10. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* (1933) doi:10.1037/h0071325.

11. Cichocki, A. & Phan, A.-H. Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *Ieice Trans. Fundam.* **E92-A**, 708–721 (2009).

12. DeMers, D. & Cottrell, G. Non-linear dimensionality reduction. in *Advances in neural information processing systems* 580–587 (1993).

13. Moon, K. R. *et al.* 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**, 2319–23 (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 local linear embedding. *Science (80-.).* **290**, 2323–2326 (2000).

18. Maaten, L. Van Der & Hinton, G. Visualizing Data using t-SNE. **1**, 1–25 (2008).

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

20. Trapnell, C. *et al.* 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 Univ.* **8**, 406–424 (2004).

24. Siebert, S. *et al.* Stem cell differentiation trajectories in Hydra resolved at single-cell

resolution. *Science (80-.).* **365**, (2019).

25. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
26. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).
27. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science (80-.).* **360**, (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 (2018).
29. Kim, T. *et al.* 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. Anal.* **40**, 68–96 (2016).
32. Santos, J. M. & Embrechts, M. On the use of the adjusted Rand index as a metric for evaluating supervised classification. *Artif. Neural Networks–ICANN 2009* **5769**, 175–184 (2009).
33. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (80-.).* **360**, 176–182 (2018).
34. Schiebinger, G. *et al.* Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *Cell* **176**,

928-943.e22 (2017).

35. He, X. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

36. Zhang, W. *et al.* Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. *Transl. Lung Cancer Res.* **7**, 32–49 (2018).

37. Shue, Y. T., Lim, J. S. & Sage, J. Tumor heterogeneity in small cell lung cancer defined and investigated in pre-clinical mouse models. *Transl. Lung Cancer Res.* **7**, 21–31 (2018).

38. Böttger, F. *et al.* Tumor Heterogeneity Underlies Differential Cisplatin Sensitivity in Mouse Models of Small-Cell Lung Cancer. *Cell Rep.* **27**, 3345-3358.e4 (2019).

39. Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat. Cancer* **1**, 423–436 (2020).

40. Carney, D. N. *et al.* Establishment and Identification of Small Cell Lung Cancer Cell Lines Having Classic and Variant Features. *Cancer Res.* **45**, 2913–2923 (1985).

41. Fargion, S. *et al.* Heterogeneity of Cell Surface Antigen Expression of Human Small Cell Lung Cancer Detected by Monoclonal Antibodies. *Cancer Res.* **46**, 2633–2638 (1986).

42. Poirier, J. T. *et al.* Selective tropism of seneca valley virus for variant subtype small cell lung cancer. *J. Natl. Cancer Inst.* **105**, 1059–1065 (2013).

43. Poirier, J. T. *et al.* DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. *Oncogene* **34**, 5869–5878 (2015).

44. Borromeo, M. D. *et al.* ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. *Cell Rep.* **16**, 1259–1272 (2016).

45. Mccoll, K. *et al.* <Oncotarget-08-73745.Pdf>. **8**, 73745–73756 (2017).

46. Huang, Y. H. *et al.* POU2F3 is a master regulator of a tuft cell-like variant of small cell

lung cancer. *Genes Dev.* **32**, 915–928 (2018).

47. Poirier, J. T. *et al.* New Approaches to SCLC Therapy: From the Laboratory to the Clinic. *J. Thorac. Oncol.* **15**, 520–540 (2020).

48. Wooten, D. J. *et al.* Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. *bioRxiv* 506402 (2019) doi:10.1101/506402.

49. Mørup, M. & Hansen, L. K. Archetypal analysis for machine learning and data mining. *Neurocomputing* **80**, 54–63 (2012).

50. Groves, S. M. *et al.* Cancer Hallmarks Define a Continuum of Plastic Cell States between Small Cell Lung Cancer Archetypes. *bioRxiv* 2021.01.22.427865 (2021).

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

52. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

53. Cooley, S. M., Hamilton, T., Deeds, E. J. & Ray, J. C. J. A novel metric reveals previously unrecognized distortion in dimensionality reduction of scRNA-Seq data. *bioRxiv* 689851 (2019) doi:10.1101/689851.

54. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science (80-.).* **367**, 405–411 (2020).

55. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, 1–12 (2008).

56. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

57. Wagner, A. H. *et al.* Recurrent WNT pathway alterations are frequent in relapsed small

cell lung cancer. *Nat. Commun.* **9**, (2018).

58. George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47–53 (2015).

59. Lim, J. S. *et al.* Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature* **545**, 360–364 (2017).

60. Ouadah, Y. *et al.* Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch. *Cell* **179**, 403-416.e23 (2019).

61. Zhao, L. *et al.* Flotillin1 promotes EMT of human small cell lung cancer via TGF-β signaling pathway. *Cancer Biol. Med.* **15**, 400–414 (2018).

62. Liu, Y. *et al.* Transcriptional landscape of the human cell cycle. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 3473–3478 (2017).

63. Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: The industrial revolution of DNA around histones. *Nat. Rev. Mol. Cell Biol.* **7**, 437–447 (2006).

64. Tee, W.-W. & Reinberg, D. Chromatin features and the epigenetic regulation of pluripotency states in ESCs. *Development* **141**, 2376–2390 (2014).

65. Waskom, M. Seaborn: Statistical Data Visualization. *J. Open Source Softw.* **6**, 3021 (2021).

66. Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 3642–3649 (2012) doi:10.1109/CVPR.2012.6248110.

67. Jia, P., Liu, Q. & Sun, Y. Detection and Classification of Astronomical Targets with Deep Neural Networks in Wide-field Small Aperture Telescopes. *Astron. J.* **159**, 212 (2020).

68. Goh, G. B., Hodas, N. O. & Vishnu, A. Deep learning for computational chemistry. *J.*

Comput. Chem. **38**, 1291–1307 (2017).

69. Harrison, D. & Rubinfeld, D. L. Hedonic housing prices and the demand for clean air. *Reveal. Prefer. Approaches to Environ. Valuat. Vol. I II* 99–120 (2019).

70. Kulczycki, P. & Charytanowicz, M. A complete gradient clustering algorithm formed with kernel estimators. *Int. J. Appl. Math. Comput. Sci.* **20**, 123–134 (2010).

71. Siebert, S. *et al.* Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science (80-.).* **365**, eaav9314 (2019).

72. Joung, I. S., Kim, J. Y., Gross, S. P., Joo, K. & Lee, J. Conformational Space Annealing explained: A general optimization algorithm, with diverse applications. *Comput. Phys. Commun.* **223**, 28–33 (2018).