A Statistical Approach to Dimensionality Reduction Reveals Scale and Structure in scRNA-seq Data

Eric Johnson[1,2], William Kath[1,2] and Madhav Mani[1,2,3]

[1]Department of Engineering Sciences and Applied Mathematics, Northwestern University
[2]NSF-Simons Center for Quantitative Biology at Northwestern University
[3]Department of Molecular Biosciences, Northwestern University

Submitted: November 18, 2020

Abstract

Single-cell RNA sequencing (scRNA-seq) experiments often measure thousands of genes, making them high-dimensional data sets. As a result, dimensionality reduction (DR) algorithms such as t-SNE and UMAP are necessary for data visualization. However, the use of DR methods in other tasks, such as for cell-type detection or developmental trajectory reconstruction, is stymied by unquantified non-linear and stochastic deformations in the mapping from the high- to low-dimensional space. In this work, we present a statistical framework for the quantification of embedding quality so that DR algorithms can be used with confidence in unsupervised applications. Specifically, this framework generates a local assessment of embedding quality by statistically integrating information across embeddings. Furthermore, the approach separates biological signal from noise via the construction of an empirical null hypothesis. Using this approach on scRNA-seq data reveals biologically relevant structure and suggests a novel "spectral" decomposition of data. We apply the framework to several data sets and DR methods, illustrating its robustness and flexibility as well as its widespread utility in several quantitative applications.

Introduction

Recent advances in high-throughput measurement techniques have revolutionized cellular and molecular biology. In particular, the advent of single-cell RNA sequencing (scRNA-seq) has made it possible to ask detailed questions about cellular differentiation, patterning, signaling, and variation at a single-cell resolution [1–13]. However, transcriptional sequencing's attempt to characterize the entire cellular transcriptome at once means that thousands of genes must be measured simultaneously, making the data inherently high-dimensional and subject to the "curse of dimensionality" [14]. As a result, sophisticated methods must be employed in order to make statistical inferences from the data (e.g., the DESeq2 algorithm to infer simple differences in fold-change expression [15]).

Traditionally, a researcher would seek to find a reduced set of features (genes) or combination of features on which statistical methods can be applied with more power; this is known as **dimensionality reduction** (DR), and when done correctly, can be used to find a more "natural" description of a system [16]. Significant effort has been put into the development and application of DR algorithms such as PCA [17], t-SNE [18], UMAP [19], and others [20–34], which attempt to find lower-dimensional (usually two- or three-dimensional) representations of the data that preserve some aspect of the original structure (for a review, see [35–37]; in application to -omics data, see [38]). However, regardless of the choice of algorithm, DR will always incur a loss of information [39, 40], which manifests itself as distortions of high-dimensional structure in the lower-dimensional embedding [27, 36, 41] (See Figure S1 for an example). Furthermore, it is impossible to detect "by eye" which parts of an embedding are signal and which are noise, since these methods are often non-linear or stochastic, and therefore do not homogeneously distort the data [42, 43].

As a result, the use of DR algorithms in scRNA-seq analysis may be treated skeptically [9, 12, 44] or require additional guidance [45]. To underscore these concerns, consider Figure 1, where scRNA-seq data from over 5000 bone-marrow cells that were collected by the Tabula Muris Consortium [8] have been embedded in two dimensions using several DR algorithms. As noted earlier, the detection of different cell types in a heterogeneous tissue is a biologically interesting task, so the annotated cell types from [8] have been used to color the embeddings (the legend can be found in the S2). However, a quick examination of the embeddings in panels **B** and **E** reveals that the arrangement and shape of the clusters are different between two different runs of the t-SNE algorithm. As a result, even though algorithms such as t-SNE are *provably good* at clustering [46], they cannot reliably be used by themselves for unsupervised clustering due to these non-linear and stochastic effects. The rest of Figure 1 underscores that the addition of algorithmic hyperparameters and the choice of algorithms only serve to complicate this process.

To address these issues, much work has been done to provide guidelines on how to use these algorithms [10, 44, 45, 47] and to make improvements to the algorithms themselves [46, 48–54] that aim to correct or account for these distortions. At the same time, an entire set of methods for

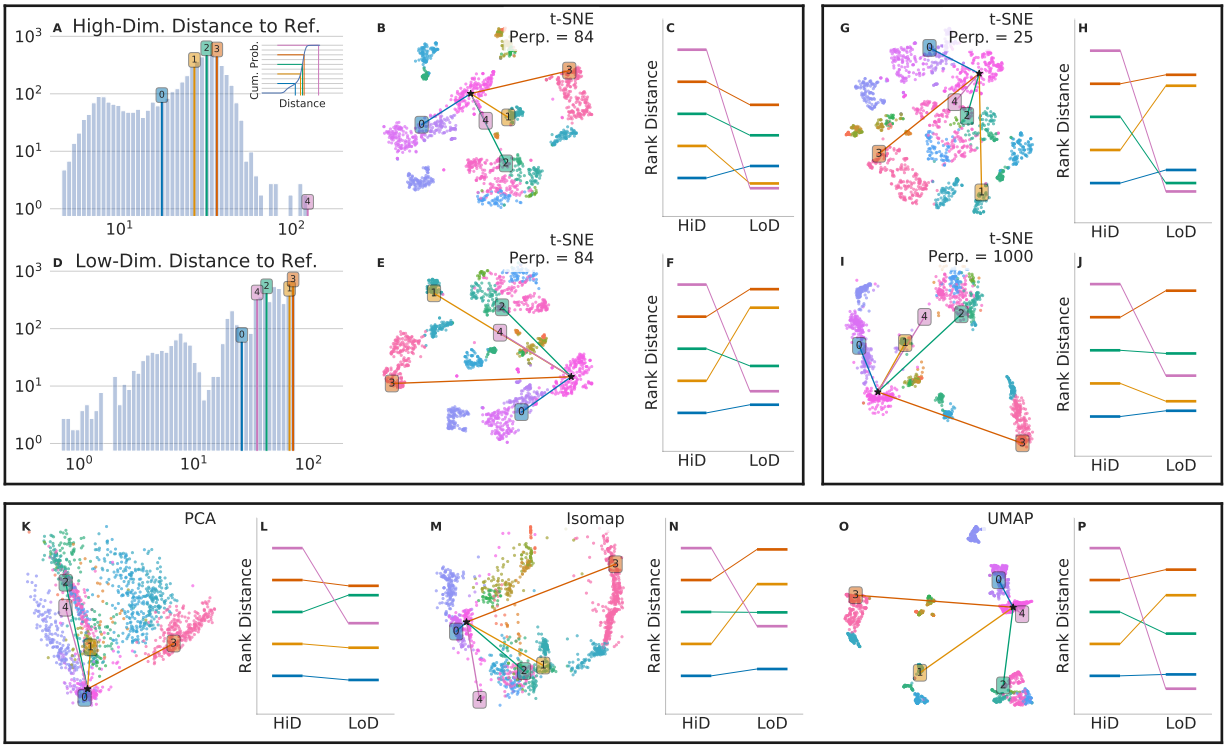

Figure 1: **A Zoology of Embeddings Demonstrates the Non-Linear and Stochastic Effects of Dimension Reduction Algorithms:** 5037 FACS-sorted bone marrow cells from several mice were sequenced by the Tabula Muris project [8] and have been embedded by several DR algorithms. (A) The pair-wise Euclidean distances between a reference cell and all other cells in the data set are shown as a histogram, with five "marker" cells highlighted by colored lines and numbered boxes. (A, Inset) The cumulative distribution of pair-wise distances shows that the five marker cells correspond to the 20th, 40th, 60th, 80th, and 100th percentile distances. (B) The cells are embedded in 2D by the standard t-SNE algorithm with perplexity set to 84 (as chosen in [8]). The coloring of the cells corresponds to the annotated cell type from [8], the legend can be found in Figure S2. The five marker cells from panel A are indicated with lines and numbered boxes. (C) The rank distance of the five marker cells in the high- and low-dimensional spaces (HiD and LoD, respectively) are compared. (D) The distance distribution from all points in the embedding in (E) to the reference point. The five marker cells from panel A are indicated by colored lines and numbered boxes. (E & F) The same as in panels B & C except with a different random initialization to the t-SNE algorithm. Notice the reordering of the originally marker points differs from that in panel C. (G & H) Same as in panels B & C except with the perplexity parameter set to 25. (I & J) Same as in panels B & C except with the perplexity parameter set to 1000. (K & L) Same as in panels B & C except the data are now projected onto their first 2 principal components. (M & N) Same as in panels B & C except the data have now been embedded with Isomap (default scikit-learn parameters) [22]. (O & P) Same as in panels B & C except the data have now been embedded with UMAP (default umap-learn parameters) [19].

quantitatively assessing the *quality* of DR methods has been developed [37, 43, 55]. These metrics can roughly be categorized as being global [27, 49, 56–64] or local [26, 42, 43, 65, 66] in scope, and either based on preserving distances [64], neighborhoods [27, 43, 55–57, 59, 67, 68], or topology [63, 69, 70], but in all cases, they attempt to summarize the extent to which a given DR algorithm preserves some aspect of the original data's structure. These metrics have been successfully used to compare DR methods and optimize hyperparameters, and a recent comprehensive benchmarking of these algorithms noted that t-SNE and UMAP were, in fact, consistently high-quality methods across many data set and metrics [37].

It is in this context that we propose a *statistical* framework for characterizing the stability and variability of embedding quality by posing a point-wise metric as an **Empirical Embedding Statistic**. We propose this approach to address several aspects of scRNA-seq data that have limited the direct application of many of the tools in the quality assessment literature. Specifically, we note that any assessment methodology for scRNA-seq should **(1)** measure quality *locally*, not globally across an embedding, **(2)** estimate the *variation* in embeddings that is introduced by the DR algorithms themselves, and **(3)** estimate where embeddings show structure consistent with actual high-dimensional structure and not noise. In the rest of this section, we explain why these criteria are necessary for a useful DR assessment framework. We then outline the approach in the next section before demonstrating its application and utility on several data sets.

First, we note that while global assessment of DR methods for hyperparameter optimization is important, the direct use of DR output for clustering or lineage reconstruction is limited by concerns about the *local* quality of the samples within an embedding. (Heuristics for maximizing global quality have been proposed [45], and human-optimized parameters are often close to those selected algorithmically [49].) That is, it is much more important to know the answer to "*Is cell A close to cell B in the embedding because they are similar in gene-space?*" than it is to know whether the average cell is well-embedded. Therefore, we want to make use of the extensive work on local quality metrics [41, 42, 47, 66–69, 71] in developing our approach.

In addition, the non-linearity and stochasticity of popular DR methods is well known to cause

large variation in the arrangement and shape of embedded structures [45], as can be seen in Figure 1 for t-SNE. (Similar results can be shown for other common algorithms, such as UMAP [19] and PHATE [34].) As a result, we should expect that this will introduce variation into any quality metrics, and this variation should be incorporated into any downstream analysis. That is, we don't just want to know whether a cell is well-embedded one time, but whether it is *consistently* well-embedded. While there is a large body of work on ensemble visualization [72, 73], only recently [74] has there been an attempt to apply this theory to assess the variability of DR embeddings. Our approach differs in that it proposes a *statistical* framework in which to consider quality metric variability. Specifically, we can consider each cell's local quality score to be a measurement of a quality *statistic* and we now want to assess the distribution of this statistic across embeddings.

This approach then allows us to incorporate concerns about signal and noise as a statistical *hypothesis test*, where we can use consistently elevated embedding quality as evidence of real biological structure. To perform this test, we propose the use of resampling to generate "null" data sets that contain no biological structure. These null data are then embedded to provide a null distribution for local quality scores. Combining this null distribution with the actual quality metrics from the data, the output of our method is a p-value assigned to each sample, requiring no further corrections, indicating the likelihood that it was embedded better than noise. This assessment of the presence of biological structure is especially useful in the context of scRNA-seq data, which is notoriously noisy and sparse [13, 44].

In this way, the statistical approach to dimensionality reduction provides a biologically relevant quantification of DR quality. The approach, outlined in the next section, addresses several unique concerns that arise with scRNA-seq data including the local fidelity of embeddings, and the variability in embedding that is due to both biological noise and the DR algorithm. In our results, we show that the application of this approach indicates that heterogeneity in embedding quality is generic across data sets and DR algorithms. We then show that examining this cell-wise embedding variability across scale parameters reveals a spectral view of the data. We demonstrate that the method can be used to rigorously compare DR methods and data sets, allowing the user to untangle

analysis choices like those presented in Figure 1. Finally, we show that the approach may have utility in downstream analyses such as unsupervised clustering that can incorporate uncertainty in embedding.

The Statistical Approach

The statistical approach to dimensionality reduction consists of three components: **(1)** the embedding of the data, **(2)** the construction and embedding of the null data, and **(3)** the calculation of the embedding statistic and performance of a hypothesis test. These are illustrated heuristically in Figure 2 and more technically in S3. These steps are centered on the calculation of a local quality statistic, the Empirical Embedding Statistic (EES), for each sample (cell) in the data set. To clarify the notation throughout the rest of this paper: consider a data set X to be a collection of N_{Cells} vectors, where each cell contains measurements for each of D genes. Recalling that the data can be embedded multiple times to yield different embeddings, we denote the position of each ith cell in the nth embedded space by $\vec{y}_{i,n}$, where the number of embeddings is N_{Embed}. For each cell, in each embedding, we will calculate the embedding statistic, which we denote $EES_{i,n}$. We use an $*$ to indicate null data generated by resampling, so that a resampled high-dimensional data vector is $\vec{x}_i^*$ and it's position in the embedded space would be $\vec{y}_i^*$. The final step of the hypothesis test process involves calculating the p-value: $p_{i,n} = P(EES^* \leq EES_{i,n})$ using the empirically generated distribution of EES^*. We elaborate on each of these three steps below.

1. **Data Embedding:** The first step in our approach is simply to embed the data one or more times, and to calculate a sample-wise quality metric on each sample (single cell) for each embedding, $EES_{i,n}$. Depending on what aspects of data structure the researcher wants to assess, several of the quality metrics mentioned earlier may be adapted for this purpose. In this work, we calculate a similarity score between each pair of samples in the high-dimensional space and again in the low-dimensional space. For similarity in the high-dimensional space, we calculate a Gaussian probability for two points being a certain Euclidean distance apart,

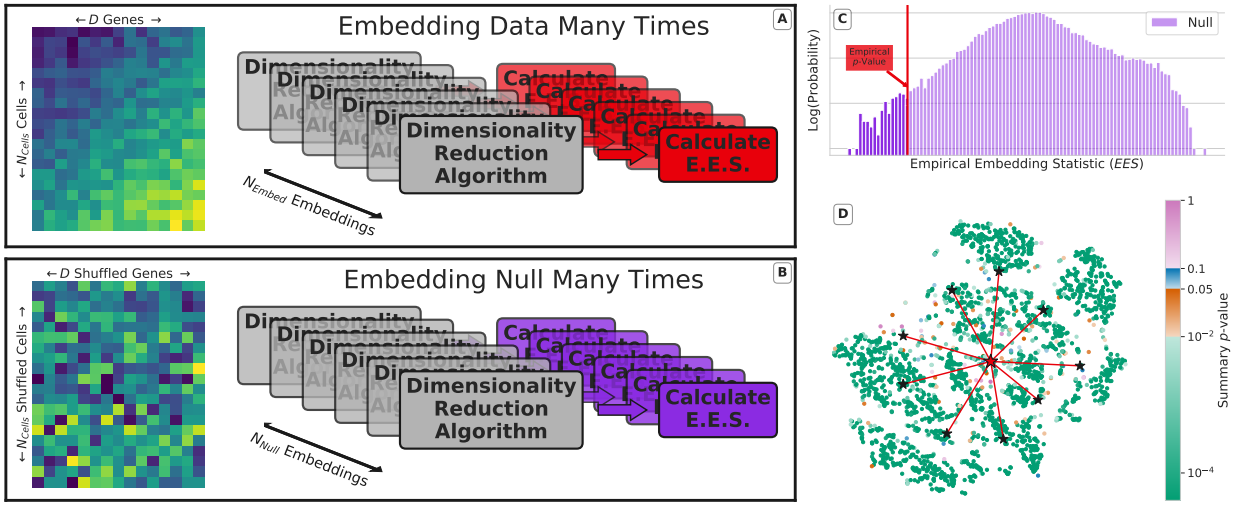

Figure 2: **The Statistical Approach:** a heuristic illustration of the proposed method. (A) A data set of N_{cell} cells and D genes is embedded N_{Embed} times. For each embedding and each cell, $EES_{i,n}$ is calculated. (B) A no-covariance resampling of the data provides a "null" data set. This is embedded N_{Null} times. For each embedding and cell, $EES^*_{i,n}$ is calculated. (C) The $EES^*_{i,n}$ are combined into an empirical null distribution. Each $EES_{i,n}$ is compared to the null distribution in turn, giving a p-value, $p_{i,n}$. (D) The p-values are synthesized for each cell across embeddings as in [75] to provide a point-wise quality metric, p_i. One of the data embeddings (MNIST digits [76], t-SNE, perplexity=30) is visualized with p_i as a color, allowing for the visual assessment of where the embedding best preserves high-dimensional structure. Cells are well-embedded when their neighbors, such as those indicated by red lines, are similarly situated in the 2D space and in the original, high-dimensional space.

where the scale is set by fixing the overall entropy of a sample's similarities, as in [25]. The low-dimensional similarity is given by the likelihood of a certain distance under a Student's t-distribution, with $\nu = 1$ degree of freedom. We then compare these similarity distributions using the Kullback-Leibler Divergence D_{KL} [77], and we use $D_{KL} \equiv EES$ as our embedding statistic. This choice of statistic is justified in that it directly relates the quality statistic to the similarities upon which t-SNE operates, and because it can be seen as a continuous generalization of *recall*, which is the likelihood that a sample's neighbors in the embedded space are also neighbors in the original data [27]. (See Figure S4 for example distributions of the EES.) Choosing a neighborhood-preservation metric is also consistent with our goal of assessing local quality in the embeddings. As shown in the results section, this also gives our quality metric the same scale parameter as t-SNE, making the results easily interpretable. However, the choice of local quality metric is flexible, and other metrics may be preferable in different contexts.

2. **Null Construction and Embedding:** The most crucial step in our process is to generate a biologically-realistic synthetic data set that has no biological structure, which we define as having zero inter-gene correlation. This is achieved via **marginal resampling**, where genes in the null data are independently drawn from the original data's gene distribution (See Figure S5 for an illustration of this process). In this way, the null data contains biologically realistic distributions of individual genes, but no grouping of the samples as a function of these genes. This provides a basis on which to empirically generate a distribution of quality scores by embedding the null data multiple times. Figure S6 shows that the null distribution is generally stable across embeddings, so that only 5-20 are needed to generate a sufficient distribution.

3. **Empirical Hypothesis Test:** Once the null data have been created and the embedding statistic EES^* has been calculated for every point over several embeddings, each of the data statistics, $EES_{i,n}$ can be compared to the aggregated distribution of null statistics, as in the panel in Figure 2. This yields an empirical p-value, which can be summarized across the N_D

embeddings [75, 78, 79] to give a single quality metric, p_i, for each cell.

We should note that while others have performed calculations for random rank orderings to adjust quality statistics [26, 57, 60], it is not immediately clear that a generic DR method will produce completely randomized neighborhoods when applied to noise. Furthermore, these calculations do not describe the *spread* with which we expect to observe neighborhood preservation from noise, so that unlike our proposed method, they cannot evaluate the likelihood of extreme values.

Results

Embedding Quality is Heterogeneous Across an Embedding

As noted earlier, there is no good reason to assume *a priori* that a generic data set has any uniformity (in terms of density, continuity, topology, etc.) in the high-dimensional space. This lack of uniformity, is one of the central difficulties of dimensionality reduction and the analysis of high-dimensional data. There have been many methods proposed to address this heterogeneity, from t-SNE's scale-sensitive kernel [25] to Isomap's method of making local graph approximations [22], and yet even the most advanced algorithms necessarily end up setting global hyperparameters. As a result, we expect that even for parameters that are *globally optimal* there will be regions that are poorly embedded compared to the rest of the data (we know that global quality is bounded [40], but not the variance in quality within an embedding).

This local heterogeneity in embedding quality has previously been shown [42, 43, 66, 68, 71, 80], and we also recover this phenomenon, as shown in Figures 2, 3, and 4. Using the bottom panels of Figure 3 as an example, we note that embedding quality varies considerably across an embedding, and not necessarily with any generic pattern. Especially in comparison to an unmarked embedding it is impossible to deduce what parts of an embedding are interpretable features - and what parts are noise - without the use of a local quality metric.

Sweeping Across Scale Reveals Spectral Structures in the Data

The previous result highlights one of the difficulties of working with real data: that we don't know whether the data are spread throughout the high-dimensional space with a uniform spatial scale. To deal with this, t-SNE applies a similarity measure between data points where the scale of that measure is tuned to the size of each individual point's local neighborhood. Similarly, UMAP builds a similarity graph by only considering the distances to a sample's k nearest neighbors. The effect of these choices is to maintain an equal weighting of local scales across an embedding, which is ideal because it allows for both densely-packed regions of the data space to be examined with equal standing to more diffuse regions. However, this balancing of scales comes at the cost of specifying a neighborhood size in the form of an algorithmic hyperparameter: either `perplexity` for t-SNE or `n_neighbors` for UMAP. Considerable effort has been dedicated to disentangling the effect of this hyperparameter or eliminating it altogether [29, 30, 45, 48, 49, 81], but we demonstrate in Figure 3 that examining the embedding quality over many scales reveals important structure in the data.

First, we note that sweeping across t-SNE's perplexity parameter can be used to find a scale at which most of the samples in a data set are well embedded. In the case of the Tabula Muris FACS Marrow data, (3A) shows that this occurs at perplexity$\approx$1200, which is considerably higher than most recommendations for the hyperparameter [25, 45]. Considering neighborhood sizes considerably smaller or larger than this results in embeddings that have neighborhoods that are indistinguishable from noise (3A left and right insets).

While 3A shows how we can summarize the effect of choosing a neighborhood size, our local statistical approach also allows for the examination of different portions of the data independently. Examining cells according to their annotated labels as in 3B-E, shows that different cell types might have different characteristic structures that are better represented at certain neighborhood sizes than others (See Figure S8 for all annotated cell types). This suggests that examining embedding "power" as a function of scale - a sort of spatial power spectrum - might be a useful way to explore scRNA-seq data even if cell type annotations are not available.

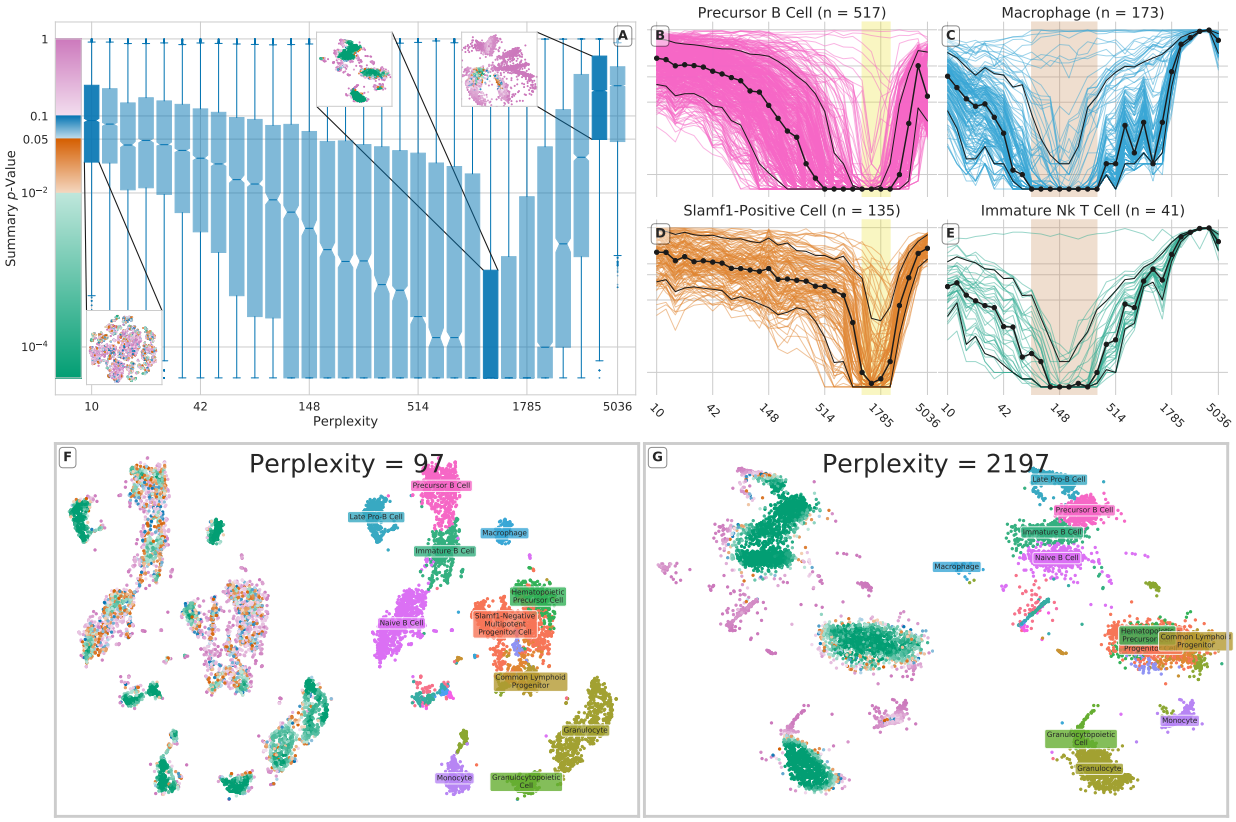

Figure 3: **A Spectral View of the Data Reveals Biological Structure:** Sweeping across t-SNE's scale parameter ("perplexity") reveals significantly different embedding quality at different scales (Panel A). Setting the scale parameter too large or small results in embeddings that are not much better structured than noise (Left and Right insets). Setting the perplexity parameter to that of the smallest average p-value shows indicates a potential consensus scale at which to examine the embedding (Middle inset). Examining previously annotated cell types' embedding quality as a function of perplexity suggests that cell types might be able to be characterized by their scale spectra (Panels B-E). PCA analysis of the scale spectra in Figure S7 indicate two spectral modes, with peaks near ~100 and ~2200. Applying the statistical approach at each mode (left sides of Panels F and G, respectively) reveals new biologically relevant structure, especially when compared to the expert annotations from the Tabula Muris project [8] (right side of Panels F and G).

We can use these spectra to select interesting scales at which to examine the data. For example, (3B, D) suggest that these cell types are best embedded at large perplexity (∼1500), while (3C, E) suggest that perplexity ∼100 may be more appropriate. A more rigorous approach that applies PCA to the scale spectra is shown in Figure S7, and suggests that perplexity ∼100 and ∼2200 may be more "natural" scales for the data. Applying perplexity = 97 and 2197 yields Figures 3F and 3G, respectively.

Interestingly, the large perplexity 3G indicates that three large clusters of cells are well-embedded, and the labels in the right half confirm that these are biologically consistent clusters corresponding to B Cells, Progenitor Cells, and Granulocytes. However, examination of these clusters at a smaller neighborhood scale in 3F shows that much of the apparent structure in the B cells and progenitor cells is not as well resolved as at the higher scale, except for the Late Pro-B cells, which have broken off into their own well-embedded cluster. On the other hand, the granulocytes also have a different apparent structure at the smaller perplexity, with the granulopoeitic cells breaking off into their own cluster and the granulocytes separating into a well-embedded hourglass shape. In this way, we can examine how meaningful biological differences show up at different scales - the B cells seem to be similar at a wider scale, but are less distinguishable at a narrower resolution. Meanwhile, granulocytes are also well-grouped at a large scale, but might actually be composed of several distinct sub-types, as suggested by examination at a smaller scale.

A Statistical Approach Allows for Comparisons of Data and Algorithms

The statistical approach to dimensionality reduction also provides a rigorous method to evaluate and compare data analysis protocols or the performances of dimensionality reduction algorithms on specific data sets. These topics have been the subject of significant debate [10, 11, 13, 37, 44, 45, 82] because, as Figure 1 shows, these choices can significantly impact analysis and interpretation.

In Figure 4, a comparison between embeddings of the Tabula Muris marrow tissue generated by PCA, t-SNE, and UMAP at their default parameters are all shown. Rather than making heuristic arguments, it can now be seen quantitatively that the structures in PCA and t-SNE are better

representations of the original data's structure than UMAP, when using default parameters. Applying the procedure in Figure 3A suggests a methodology for choosing algorithmic parameters, regardless of the specific algorithm being used.

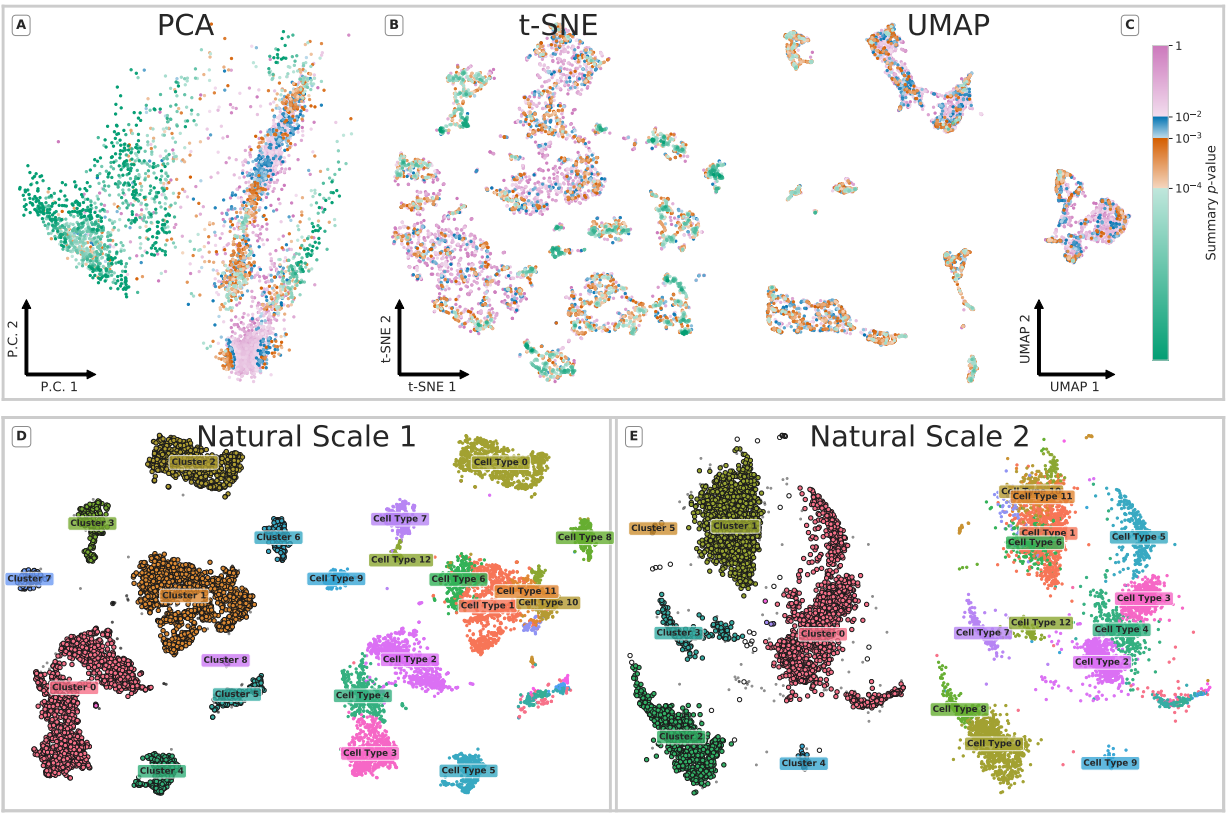

Figure 4: **The Statistical Approach Allows for Algorithmic Comparison and Facilitates Cell-Clustering:** Applying the statistical approach to assessing the quality of embeddings generated by PCA (A), t-SNE (B), and UMAP (C), reveal significantly different results between algorithms. All embeddings were generated with default parameters [19, 25], where applicable. Clustering with DBSCAN [83] at the scales suggested by Figure S7 (Panels D and E) using the empirical p-values as weights suggests that the statistical approach also has use in quantitative downstream applications.

A Statistical Approach Can Be Used Upstream of Unsupervised Clustering

Since the statistical approach is able to summarize each cell's neighborhood-fidelity with a p-value, we can leverage this quantity in downstream applications that depend on the geometry of the low-dimensional representations, such as unsupervised clustering. In the bottom of Figure 4 the results of applying DBSCAN [83] to embeddings at the two scales from Figure 3 are compared

to the cell ontology annotations generated by the Tabula Muris consortium. In generating these clusterings, the p-values, p_i, for each cell were leveraged in two ways: first, cells that are never embedded well at any scale are omitted from the clustering process, then the inverse of the p-values, $1/p_i$ are used as *weights* in the DBSCAN algorithm, allowing for more high fidelity regions of the embedding to take priority in the clustering process. Comparing the unsupervised clustering to the expert annotations reiterates some of the observations made about Figure 3 concerning the annotated cell types that appear to be part of larger, coherent structures detected by the statistical method applied to t-SNE. However, on a more fundamental level, Figure 4 indicates the immediate and obvious utility of our statistical approach in a variety of contexts.

Discussion

Dimensionality reduction is a complicated procedure, even in the best of circumstances. Single-cell RNA sequencing offers a path towards untold biological discovery, but its high-dimensional nature and relatively noisy measurements require the careful application of dimensionality reduction algorithms in order to make progress. Unfortunately, the state of the art in dimensionality reduction currently rests on ever-changing heuristics to a degree that limits data analysis.

The statistical approach presented in this work provides a rigorous approach to the evaluation of these heuristics, and at the same time unearths information about data sets that is of immediate utility to biological researchers. The statistical approach is relatively simple (Figure 2), and can be applied and utilized in a variety of contexts (Figure 4). Perhaps more importantly, statistically analyzing the EES promises to reveal previously hidden structures and scales in data sets (Figure 3).

This paper presents a broad view of the approach and its applications, but there are a few limitations that will require further consideration. Most practically, the code as written rests on the speed of current implementations of DR algorithms that can be chained together to generate many (null) embeddings of the same data. This is somewhat slow, requiring several hours to run a full scale-parameter sweep, however the structure of the method is obviously parallelizable, so there is

some optimism that this can be improved.

This efficiency concern directly relates to the fact that because the statistics are performed empirically, there is a finite resolution to the calculated p-values. This is not a huge practical concern except that it also provides a lower-bound on the p-values, as there will often be cells whose embedding statistics are completely outside the support of the null distribution. Other than improved computational efficiency, remedies may include theoretical work to describe the tails of these null distributions or a principled method for parameterizing the null distribution. Because of this effect, in this work, we have refrained from making more precise interpretations of these p-values (we do not make any significance assessments or cutoffs, nor do we do any ordered analysis of the cells by p-value), instead leveraging the fact that the p-values definitely convey strong *relative* information within the context of a data set and DR algorithm.

Moving forward, it is clear that this information can be leveraged in a variety of ways not presented in this work. Several of these directions are suggested in Figure 4, where more comprehensive efforts could be undertaken to assess the quality of DR algorithms generically, such as in [36, 37], or to incorporate the statistical approach into an unsupervised clustering algorithm more directly. Non-computationally, Figure 3 suggests that this approach may be of widespread utility in the analysis of high-dimensional biological data sets in order to detect and to assess the stability of biologically relevant structures. The ability of the method to form model-free, non-parametric scale spectra presents a new way to look at these data sets that may reveal heretofore unseen phenomena.

References

1. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell* **18,** 675–685. ISSN: 15345807 (2010).
2. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology* **29,** 1120–1127. ISSN: 10870156 (2011).
3. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201. ISSN: 10974172 (May 2015).
4. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161,** 1202–1214. ISSN: 10974172 (2015).
5. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360,** eaar3131. ISSN: 0036-8075 (June 2018).
6. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature* **555,** 457–462. ISSN: 14764687 (2018).
7. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360,** 1–39. ISSN: 10959203 (June 2018).
8. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372. ISSN: 14764687 (Oct. 2018).
9. Kester, L. & van Oudenaarden, A. Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23,** 166–179. ISSN: 19345909 (Aug. 2018).
10. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental and Molecular Medicine* **50.** ISSN: 20926413 (2018).
11. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360,** 981–987. ISSN: 0036-8075 (June 2018).
12. Dasgupta, S., Bader, G. D. & Goyal, S. Single-Cell RNA Sequencing: A New Window into Cell Scale Dynamics. *Biophysical Journal* **115,** 429–435. ISSN: 00063495 (Aug. 2018).
13. Grün, D. Revealing routes of cellular differentiation by single-cell RNA-seq. *Current Opinion in Systems Biology* **11,** 9–17. ISSN: 24523100 (Oct. 2018).
14. Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nature Methods* **15,** 399–400. ISSN: 1548-7091 (June 2018).
15. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 1–21. ISSN: 1474760X (2014).
16. Transtrum, M. K. *et al.* Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *Journal of Chemical Physics* **143.** ISSN: 00219606 (2015).
17. Jollife, I. T. & Cadima, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374.** ISSN: 1364503X (2016).
18. Van Der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605. ISSN: 15729338 (2008).

19. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (Feb. 2018).

20. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43,** 59–69. ISSN: 0340-1200 (1982).

21. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation* **10,** 1299–1319. ISSN: 0899-7667 (July 1998).

22. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **290,** 2319–2323. ISSN: 00368075 (Dec. 2000).

23. Roweis, S. T. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290,** 2323–2326. ISSN: 00368075 (Dec. 2000).

24. Belkin, M. & Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* **15,** 1373–1396. ISSN: 0899-7667 (June 2003).

25. Van Der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2625. ISSN: 15324435 (2008).

26. Chen, M. *et al.* The Bayesian Elastic Net: Classifying Multi-Task Gene-Expression Data (2009).

27. Venna, J., Kaski, S., Aidos, H., Nybo, K. & Peltonen, J. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research* **11,** 451–490. ISSN: 15324435 (2010).

28. Joia, P., Paulovich, F. V., Coimbra, D., Cuminato, J. A. & Nonato, L. G. Local Affine Multidimensional Projection. *IEEE Transactions on Visualization and Computer Graphics* **17,** 2563–2571. ISSN: 1077-2626 (Dec. 2011).

29. Najim, S. A. & Lim, I. S. Trustworthy dimension reduction for visualization different data sets. *Information Sciences* **278,** 206–220. ISSN: 00200255 (Sept. 2014).

30. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell rna-seq data by kernel-based similarity learning. *Nature Methods* **14,** 414–416. ISSN: 15487105 (2017).

31. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J. P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9.** ISSN: 20411723 (2018).

32. Wu, Y., Tamayo, P. & Zhang, K. Visualizing and Interpreting Single-Cell Gene Expression Datasets with Similarity Weighted Nonnegative Embedding. *Cell Systems* **7,** 656–666. ISSN: 24054712 (Dec. 2018).

33. Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. & Wang, B. Self-assembling manifolds in single-cell RNA sequencing data. *eLife* **8,** 1–29. ISSN: 2050084X (Sept. 2019).

34. Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* **37,** 1482–1492. ISSN: 15461696 (2019).

35. Van Der Maaten, L. J. P., Postma, E. O. & Van Den Herik, H. J. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research* **10,** 1–41. ISSN: 0169328X (2009).

36. Gracia, A., González, S., Robles, V. & Menasalvas, E. A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. *Information Sciences* **270,** 1–27. ISSN: 00200255 (2014).

37. Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. T. & Telea, A. C. Towards a Quantitative Survey of Dimension Reduction Techniques. *IEEE Transactions on Visualization and Computer Graphics* **X,** 1–1. ISSN: 1077-2626 (2019).

38. Fanaee-T, H. & Thoresen, M. Performance evaluation of methods for integrative dimension reduction. *Information Sciences* **493,** 105–119. ISSN: 00200255 (Aug. 2019).

39. Gracia, A., González, S., Robles, V., Menasalvas, E. & Von Landesberger, T. New insights into the suitability of the third dimension for visualizing multivariate/multidimensional data: A study based on loss of quality quantification. *Information Visualization* **15,** 3–30. ISSN: 14738724 (Jan. 2016).

40. Lui, K. Y. C., Ding, G. W., Huang, R. & McCann, R. J. Dimensionality Reduction has Quantifiable Imperfections: Two Geometric Bounds. *Advances in Neural Information Processing Systems* **2018-Decem,** 8453–8463. ISSN: 10495258 (Oct. 2018).

41. Colange, B., Vuillon, L., Lespinats, S. & Dutykh, D. *Interpreting Distortions in Dimensionality Reduction by Superimposing Neighbourhood Graphs* in *2019 IEEE Visualization Conference (VIS)* (IEEE, Oct. 2019), 211–215. ISBN: 978-1-7281-4941-7.

42. Aupetit, M. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing* **70,** 1304–1330. ISSN: 09252312 (Mar. 2007).

43. Mokbel, B., Lueks, W., Gisbrecht, A. & Hammer, B. Visualizing the quality of dimensionality reduction. *Neurocomputing* **112,** 109–123. ISSN: 0925-2312 (July 2013).

44. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *Cmgh* **5,** 539–548. ISSN: 2352345X (2018).

45. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nature Communications* **10,** 5416. ISSN: 2041-1723 (Dec. 2019).

46. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods* **16,** 243–245. ISSN: 1548-7091 (Mar. 2019).

47. France, S. L. & Akkucuk, U. A Review, Framework and R toolkit for Exploring, Evaluating, and Comparing Visualizations (Feb. 2019).

48. Lee, J. A., Peluffo-Ordóñez, D. H. & Verleysen, M. *Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction* in *22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2014 - Proceedings* (2014), 177–182. ISBN: 9782874190957.

49. Cao, Y. & Wang, L. Automatic Selection of t-SNE Perplexity. *arXiv* (Aug. 2017).

50. Bodt, C. D., Mulders, D., Verleysen, M. & Lee, J. A. *Perplexity-free t -SNE and twice Student tt -SNE* in *European Symposium on Artificial Neural Networks* (Bruges, Belgium, 2018). ISBN: 978-287587047-6.

51. Poličar, P., Stražar, M. & Zupan, B. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *bioRxiv,* 1–2 (2019).

52. Aliverti, E. *et al.* Projected t-SNE for batch correction. *Bioinformatics* **36** (ed Wren, J.) 3522–3527. ISSN: 1367-4803 (June 2020).

53. Häkkinen, A. *et al.* qSNE: Quadratic rate t-SNE optimizer with automatic parameter tuning for large data sets. *Bioinformatics,* 1–7. ISSN: 1367-4803 (2020).

54. Belkina, A. C. *et al.* Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications* **10,** 5415. ISSN: 2041-1723 (Dec. 2019).

55. Lee, J. A. & Verleysen, M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing* **72,** 1431–1443. ISSN: 09252312 (Mar. 2009).

56. Venna, J. & Kaski, S. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* September, 485–491 (2001). ISBN: 3540424865.

57. France, S. & Carroll, D. in *Machine Learning and Data Mining in Pattern Recognition* 499–517 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007).

58. Lee, J. A. & Verleysen, M. *Quality assessment of nonlinear dimensionality reduction based on {K}-ary neighborhoods* in *JMLR: Workshop and conference proceedings* **4** (2008), 21–35.

59. Goldberg, Y. & Ritov, Y. Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. *Machine Learning* **77,** 1–25. ISSN: 0885-6125 (Oct. 2009).

60. Lee, A. Circular data. *Wiley Interdisciplinary Reviews: Computational Statistics* **2,** 477–486. ISSN: 19395108 (2010).

61. Meng, D., Leung, Y. & Xu, Z. A new quality assessment criterion for nonlinear dimensionality reduction. *Neurocomputing* **74,** 941–948. ISSN: 09252312 (Feb. 2011).

62. Zhang, P., Ren, Y. & Zhang, B. A new embedding quality assessment method for manifold learning. *Neurocomputing* **97,** 251–266. ISSN: 09252312 (Nov. 2012).

63. Paul, R. & Chalup, S. K. A study on validating non-linear dimensionality reduction using persistent homology. *Pattern Recognition Letters* **100,** 160–166 (Dec. 2017).

64. Heiser, C. N. & Lau, K. S. A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Reports* **31,** 107576. ISSN: 22111247 (2020).

65. Kaski, S. *et al.* Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics* **4.** ISSN: 14712105 (2003).

66. Lespinats, S. & Aupetit, M. CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings. *Computer Graphics Forum* **30,** 113–125. ISSN: 01677055 (Mar. 2011).

67. Schreck, T., von Landesberger, T. & Bremm, S. *Techniques for precision-based visual analysis of projected data* in *Visualization and Data Analysis 2010* (eds Park, J., Hao, M. C., Wong, P. C. & Chen, C.) **7530** (Jan. 2010), 75300E. ISBN: 9780819479235.

68. Martins, R. M., Minghim, R. & Telea, A. C. Explaining neighborhood preservation for multidimensional projections. *Computer Graphics and Visual Computing, CGVC 2015,* 7–14 (2015).

69. Rieck, B. & Leitte, H. Persistent Homology for the Evaluation of Dimensionality Reduction Schemes. *Computer Graphics Forum* **34,** 431–440. ISSN: 01677055 (June 2015).

70. Rieck, B. & Leitte, H. in *Topological Methods in Data Analysis and Visualization IV* (eds Carr, H., Garth, C. & Weinkauf, T.) 103–117 (Springer International Publishing, Cham, 2017). ISBN: 978-3-319-44684-4.

71. Martins, R. M., Coimbra, D. B., Minghim, R. & Telea, A. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graphics* **41,** 26–42. ISSN: 00978493 (June 2014).

72. Wang, J., Hazarika, S., Li, C. & Shen, H.-W. Visualization and Visual Analysis of Ensemble Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics* **25,** 2853–2872. ISSN: 1077-2626 (Sept. 2019).

73. Kehrer, J. & Hauser, H. Visualization and Visual Analysis of Multifaceted Scientific Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics* **19,** 495–513. ISSN: 1077-2626 (Mar. 2013).

74. Reinbold, C., Kumpf, A. & Westermann, R. Visualizing the Stability of 2D Point Sets from Dimensionality Reduction Techniques. *Computer Graphics Forum* **39,** 333–346. ISSN: 14678659 (2019).

75. Heard, N. & Rubin-Delanchy, P. Choosing Between Methods of Combining p-values (July 2017).

76. Bottou, L., Haffner, P. & LeCun, Y. *Efficient conversion of digital documents to multilayer raster formats* in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* (2001). ISBN: 0769512631.

77. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22,** 79–86. ISSN: 0003-4851 (Mar. 1951).

78. Loughin, T. M. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics and Data Analysis* **47,** 467–485. ISSN: 01679473 (2004).

79. Cousins, R. D. Annotated Bibliography of Some Papers on Combining Significances or p-values. *arXiv* (May 2007).

80. Chen, L. & Buja, A. Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Drawing, and Proximity Analysis. *Journal of the American Statistical Association* **104,** 209–219. ISSN: 0162-1459 (Mar. 2009).

81. Lee, J. A. & Verleysen, M. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters* **31,** 2248–2257. ISSN: 01678655 (Oct. 2010).

82. Gisbrecht, A. & Hammer, B. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **5,** 51–73. ISSN: 19424787 (Mar. 2015).

83. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (Portland, OR, USA, 1996), 226–231.

84. Simes, R. J. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* **73,** 751–754 (1986).

Supplemental Materials

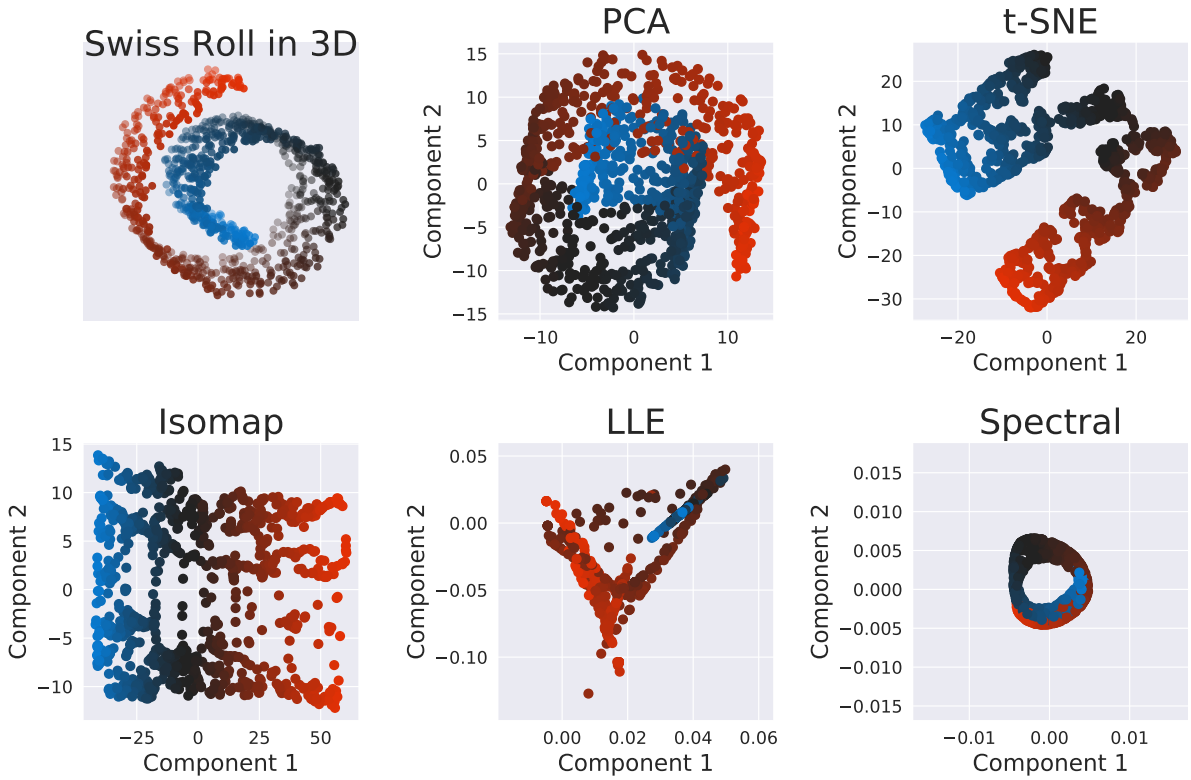

Figure S1: **Dimensionality Reduction Algorithms Generate Heterogeneous and Unpredictable Distortions in Lower-Dimensional Embeddings:** The classic "Swiss Roll" data set is embedded in two-dimensions by several DR algorithms. The variability and quality of the results varies significantly even for this simple data set. All methods shown here were run at default parameters using the scikit-learn Python package.

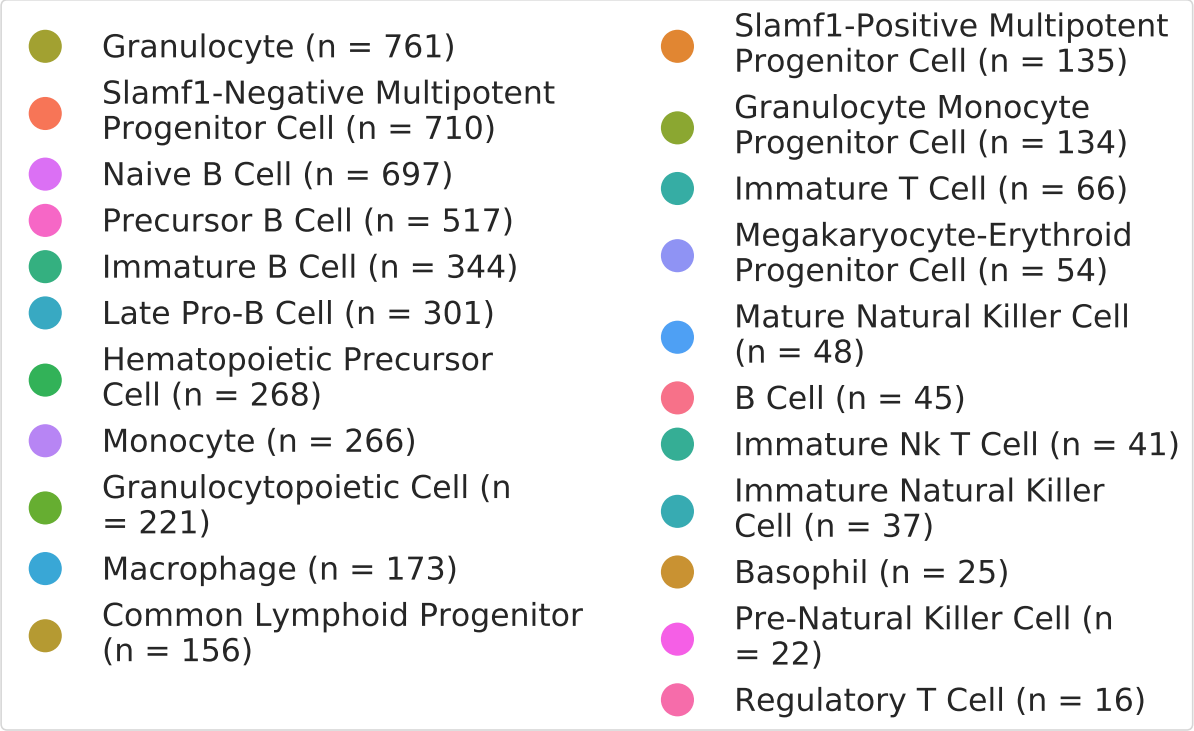

Figure S2: **Tabula Muris Marrow Tissue Cell Ontology Labels:** Legend for cell ontology classes in Figures 1, 3, and 4. Cell types were identified by the Tabula Muris project [8].

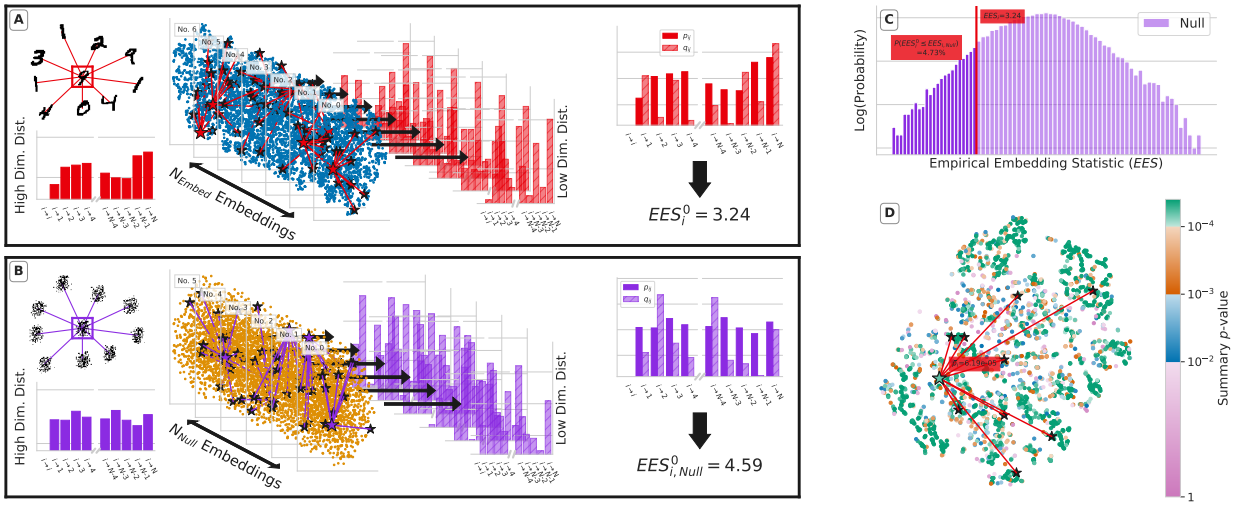

Figure S3: **The Statistical Approach with More Details:** an illustration of the method using the MNIST digits [76]. (A, Left) The neighboring samples of a digit in the high-dimensional space is identified and quantified via an array of "affinities", p_{ij}, which are calculated as in [25], but are *normalized per sample*, rather than over all samples. (A, Middle) The data set is embedded N_{Embed} times and the affinities in the low-dimensional space between each sample and its neighbors are again calculated as in [25]. Again, the normalization of these affinities is performed sample-wise, rather than across the entire data set. (A, Right) The high- and low-dimensional affinities for each cell are compared using the Kullback-Leibler divergence, which measures differences between discrete distributions. This value, $EES_{i,n}$ is stored for each sample in the data set for each embedding of the data. (B) The process in panel A is repeated for *null* data sets generated via marginal resampling (see S5 for an illustration), resulting in the calculation of $EES^*_{i,n}$. Examples of resampled MNIST digits are shown (B, Left). (C) The EES values calculated from the null embeddings are aggregated into a distribution, here shown in purple, and the $EES_{i,n}$ for each ith sample in the nth embedding of the data (shown in red) is compared to this null distribution to generate an empirical p-value. (D) The p-values for each sample can be summarized across embeddings using Simes' method [84]. These summary p-values can be indicated on any particular embedding using a custom colormap.

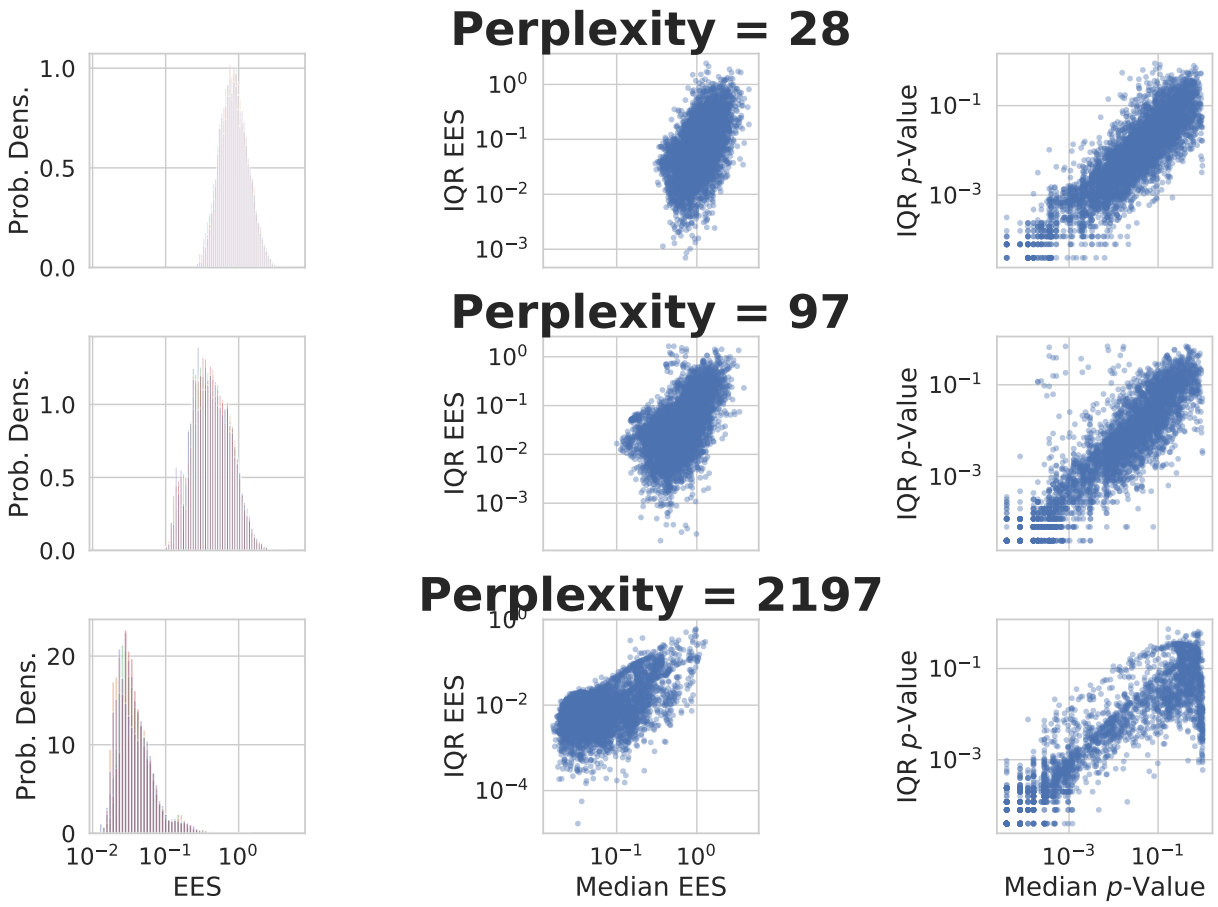

Figure S4: **The Empirical Embedding Statistic Varies Between Embeddings:** Examples of the EES distribution for several t-SNE embeddings of the Tabula Muris marrow data set at three perplexity values. The left column shows the EES distributions, with each of 5 embeddings denoted by a different color. The distributions appear similar, but the middle and right columns show the inter-quartile range (IQR) of each sample's EES (middle column) or p-value (right column) as a function of the median EES or p-value. These plots show that for some samples, the EES and p-value can vary significantly between embeddings.

Figure S5: **Illustration of Marginal Resampling for Generating Null Data Sets:** The null data sets are generated via marginal resampling, which is a process by which "null" cells are created by randomly selecting from the distribution for each measurement (gene) independently. This results in data sets that have the same marginal distributions for each gene as the original data, but have no correlative structure (compare the left and right panels).

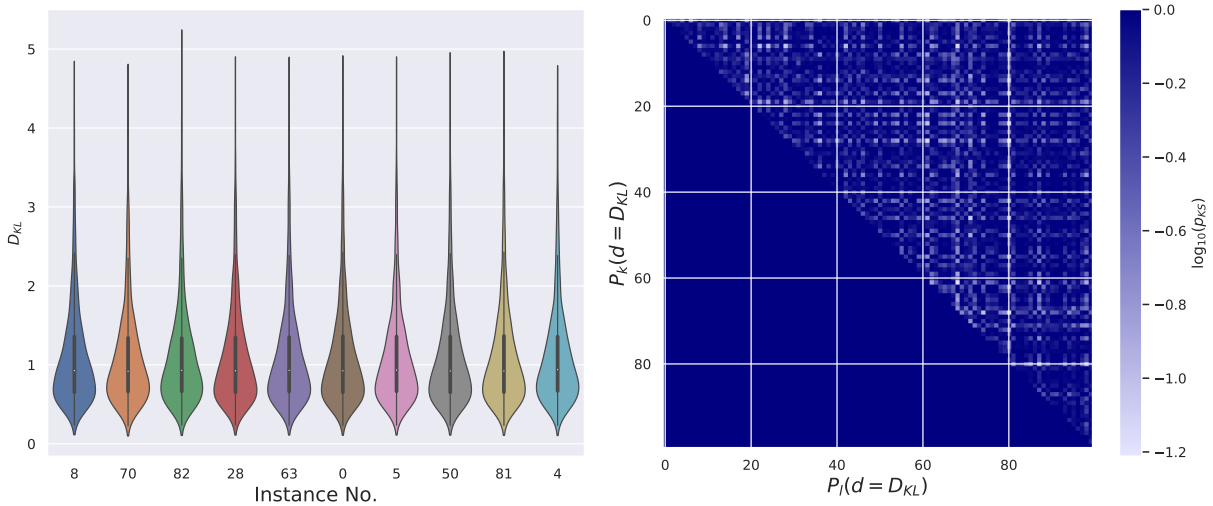

Figure S6: **Comparing Single Null Embeddings Shows Relative Stability of the EES Null Distribution:** 100 null distributions are generated for the MNIST digits data at perplexity of 50. The left panel shows a random selection of these distributions, while the right panel shows the p-value from a Kolmogorov-Smirnov test between each pair of null distributions. The right panel shows that most pairs display a high p-value, suggesting close agreement between these distributions. Based on this and other tests (not shown), we recommend combining between 5-20 null data sets in order to form a stable null distribution for the EES.

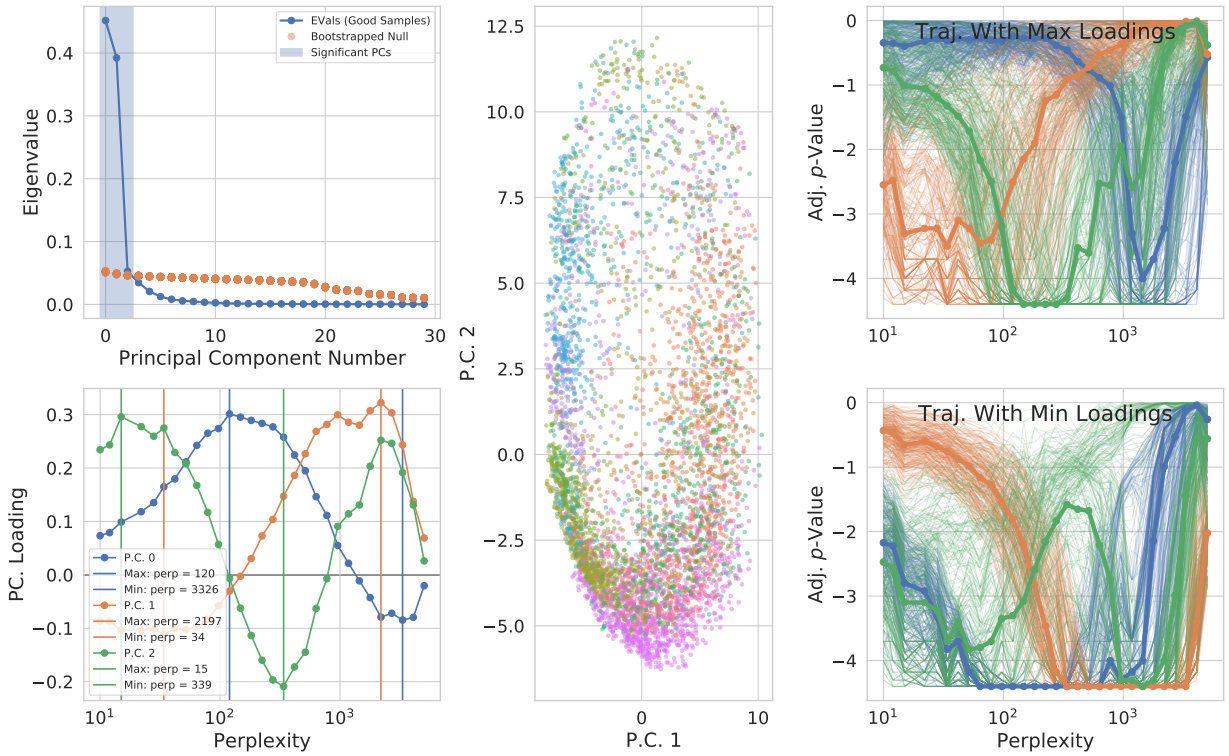

Figure S7: **PCA Analysis of Scale Spectra Reveals "Natural" Data Scales:** Performing PCA on the scale-spectra from Figure 3 reveals three significant principal components, of which two are extremely significant (Top Left). Examining these components (Bottom Left) indicate several natural scales at which to examine the data. Plotting the spectra in PC space (Middle) show a ring-like structure. Colors indicate cell ontology annotations from Figure S2. Examining the spectra with the top loadings in each principal component (Positive in Top Right and Negative in Bottom Right; PC1 in blue, PC2 in orange, and PC3 in green) show that perplexities of 100 and 2000 may be natural and interesting scales at which to examine the data.

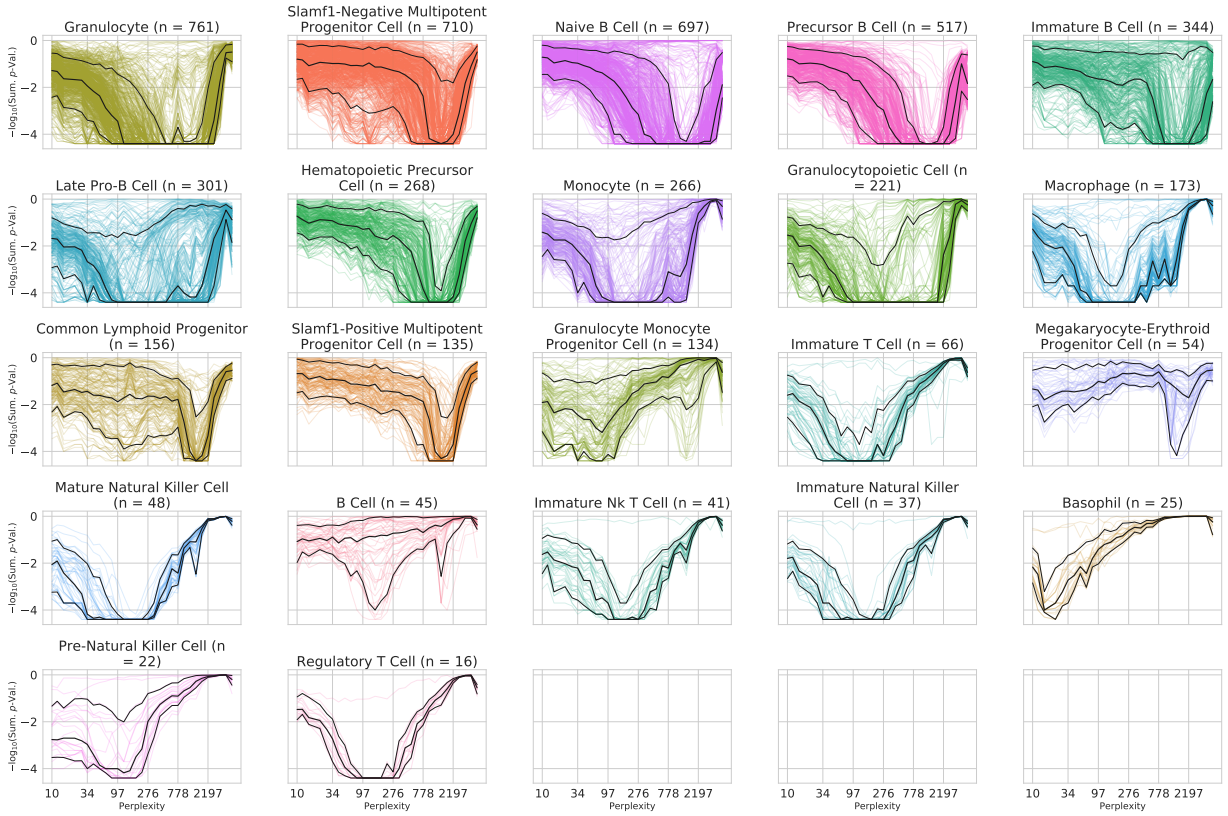

Figure S8: **Sweeping Across Perplexity Reveals Different Natural Scales for Different Cell Types:** Examining the scale-spectra of each cell type as in Figure 3 suggests that some cell types may have characteristic spectra or characteristic scales at which they are well resolved in a 2D embedding.