Published in final edited form as:

Nat Biotechnol. 2021 June ; 39(6): 765–774. doi:10.1038/s41587-020-00801-7.

Assessing Single-Cell Transcriptomic Variability through Density-Preserving Data Visualization

Ashwin Narayan[1,2,3], **Bonnie Berger**[1,2,3,*], **Hyunghoon Cho**[2,3,*]

[1]Department of Mathematics, MIT, Cambridge, MA 02139, USA

[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

[3]Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA 02139, USA

Abstract

Nonlinear data-visualization methods, such as t-SNE and UMAP, summarize the complex transcriptomic landscape of single cells in 2D or 3D, but they neglect the local density of data points in the original space, often resulting in misleading visualizations where densely populated subsets of cells are given more visual space than warranted by their transcriptional diversity in the dataset. We present den-SNE and densMAP, density-preserving visualization tools based on t-SNE and UMAP, respectively, and demonstrate their ability to accurately incorporate information about transcriptomic variability into the visual interpretation of single-cell RNA-seq data. Applied to recently published datasets, our methods reveal significant changes in transcriptomic variability in a range of biological processes, including heterogeneity in transcriptomic variability of immune cells in blood and tumor, human immune cell specialization, and the developmental trajectory of *C. elegans*. Our methods are readily applicable to visualizing high-dimensional data in other scientific domains.

Introduction

Exploratory analyses of large-scale biological datasets typically begin with visualizing the data in low dimensions, in the hopes of revealing high-level structural insights to be probed in downstream analyses. This approach has been especially critical in the rapidly emerging field of single-cell transcriptomics, where high-throughput single-cell RNA sequencing (scRNA-seq) technologies are empowering researchers to study gene expression at an unprecedented resolution across diverse tissues, organisms, and biological conditions. Driven by the high-dimensionality of scRNA-seq datasets (thousands of different transcripts per cell) and their increasingly large-scale (hundreds of thousands of cells), many researchers rely on 2D or 3D data visualizations for quickly and intuitively finding structural

[*]**Correspondence** Correspondence and requests for materials should be addressed to Hyunghoon Cho (hhcho@broadinstitute.org) and Bonnie Berger (bab@mit.edu).

Author Contributions All authors conceived the method, evaluated results, and wrote the manuscript. A.N. and H.C. implemented the software and conducted the experiments. B.B. and H.C. guided the research.

Competing Interests The authors declare that they have no competing financial interests.

patterns (e.g. clusters or trajectories) and communicating biological insights with the scientific community[1,2].

Two of the most popular techniques for high-dimensional data visualization are t-stochastic neighborhood embedding[3] (t-SNE) and uniform manifold approximation and projection[4] (UMAP), both of which have been widely adopted in scRNA-seq analysis[5,6]. In contrast to traditional methods for dimensionality reduction, e.g. principal component analysis (PCA), both t-SNE and UMAP learn a *nonlinear* embedding of the original space by optimizing the embedding coordinates of individual data points using iterative algorithms. Both methods aim to accurately preserve the original local neighborhood of each data point in the visualization, while being more permissive of distortions in long-range distances. Because of the expressiveness of nonlinear embeddings, t-SNE and UMAP are well-regarded for their empirical performance at elucidating sophisticated manifold structures and clustering patterns in high-dimensional data[1,2].

Despite their strengths, t-SNE and UMAP suffer from a major, often-overlooked pitfall: they neglect information about the local density of data points in the source dataset. In other words, data points whose neighbors are close-by in the original data are not distinguished in the visualization from those whose neighbors are far away. This limitation leads to misleading visualizations where the apparent size of a cluster largely reflects the number of points in the cluster rather than its underlying heterogeneity, as our results demonstrate. In scRNA-seq data, this omitted information about heterogeneity corresponds to the *variability* of gene expression within a subpopulation of cells. Thus, accurately portraying differences in local density in the visualization could provide another “dimension” of information, reflecting heretofore hidden insights into the transcriptomic landscape of single cells.

Here, we introduce *density-preserving* data visualization methods den-SNE and densMAP that build upon t-SNE and UMAP, respectively, to enable researchers to more accurately visualize and extract deeper biological insights from the growing compendium of single-cell transcriptomic experiments. Our methods leverage the insight that, since both t-SNE and UMAP construct their embeddings by iteratively optimizing an objective function, we can augment that objective function with an auxiliary term that measures the distortion of local density at each data point in the visualization. To this end, we develop a general, differentiable measure of local density, called the *local radius*, which intuitively represents the average distance to the nearest neighbors of a given point. Our design of this measure enables efficient optimization of the density-augmented visualization objective. The algorithmic techniques we introduce could be used to augment other visualization tools based on iterative optimization and thus are of general interest.

To demonstrate the utility of density-preserving visualization, we applied den-SNE and densMAP to a diverse range of published scRNA-seq datasets from lung cancer patients[7], human peripheral blood cells[8] and embryonic roundworm *Caenorhabditis elegans*[9], as well as the UK Biobank human genotype profiles and the canonical MNIST hand-written digit images. These methods not only capture additional information beyond existing visualization techniques but also biological insights others miss, including immune cell transcriptomic variability in tumors; specialization of monocytes and dendritic cells; and temporally modulated transcriptomic variability across developmental lineages of *C. elegans*. Our work shows that density-preserving data visualization can detect unforeseen patterns in single-cell transcriptomic landscapes and enrich our understanding of biology.

Results

Overview of density-preserving data visualization.

Our density-preserving visualization methods, den-SNE and densMAP, augment t-SNE and UMAP respectively, generating embeddings that preserve both local structure and variability in the original data (Figure 1 and Methods). To capture the local structure of the data, t-SNE and UMAP both create a nearest-neighbors graph and preserve only the distances between neighboring points in this graph. We use the same nearest-neighbors graphs underpinning each of the original methods to calculate a *local radius* around each point, which represents the average distance from the point to its nearest neighbors; this conveys the density of that point's neighborhood. The two original algorithms have an objective function that quantifies the agreement between a given embedding and the original nearest-neighbors graph, and they rearrange the embedding to maximize this agreement. We augment these objective functions with an additional term that measures the agreement between the local radii in the original dataset and in the embedding, ensuring that local structure is still preserved in the embedding while also conveying information about local variability. Our techniques have strong theoretical foundations, enable efficient optimization, and are easily generalized to other data visualization algorithms that similarly use gradient-based optimization (Methods, Supplementary Notes 1–3).

Applying our methods to simulated datasets featuring heterogeneous density landscapes revealed the misleading visual conclusions that could be made without density preservation (Figure 2). Visualizing a mixture-of-Gaussian point clouds with different variances, t-SNE and UMAP generate clusters that are all similarly sized, while den-SNE and densMAP accurately depict the different variances (Figure 2a). When the point clouds are translated linearly with overlap, reflecting a trajectory, lack of density preservation in t-SNE and UMAP obscures dynamic changes in variability over the trajectory (Figure 2c). Conversely, when size is constant but a region is oversampled, t-SNE and UMAP overrepresent this oversampled region, giving the impression of increased variability and downplaying undersampled regions (Figures 2b and d). Our following results show that these considerations are critical in biological analyses.

Visualizing the heterogeneity of immune cells in tumor.

To illustrate the value of density-preserving visualization for biological studies, we first applied our methods to a scRNA-seq dataset of 41,861 immune cells in matched tumor and peripheral blood samples from seven non-small-cell lung cancer (NSCLC) patients[7]. The original study identified distinct transcriptomic states spanned by tumor-infiltrating myeloid cells that were reproducibly observed across different individuals, suggesting their potential relevance for cancer immunotherapies. We asked whether our methods could more accurately capture the transcriptomic landscape of tumor-infiltrating immune cells than existing tools.

Comparison of den-SNE and t-SNE embeddings revealed several immune cell types with noticeable differences between the visualizations (Figure 3): tumor-infiltrating neutrophils and plasma cells occupy considerably more space in the den-SNE visualization than their t-SNE counterparts, while tumor-infiltrating T cells are relatively smaller in den-SNE. These discrepancies arise because visual size of a cluster in t-SNE corresponds more closely to the number of cells in the cluster than to underlying variability. Thus, in t-SNE, tumor-infiltrating neutrophils ($n = 2861$) occupy much less space than circulating neutrophils ($n = 9217$) despite den-SNE indicating they have comparable variability. The rich transcriptomic diversity of tumor-infiltrating plasma cells is also lost in t-SNE. Conversely, T cells, the most populous cell type in tumors ($n = 10701$) are visually overrepresented in t-SNE relative to their actual variability.

To quantify the improvement in density preservation that our algorithms offer, we calculate two complementary measures of local density in the visualization—(i) local radius and (ii) neighborhood count (see Methods)—and assess their correlation with the local radii in the original data space, which represent underlying variability in the dataset. Both measures quantify our perception of density in the visualizations (inversely-related for local radius); intuitively, the local radius captures the size of a neighborhood that contains a fixed number of nearest neighbors, and the neighborhood count captures the number of points within a fixed radius around each point. The former is consistent with how our algorithms model density for efficient optimization, while the latter is arguably a more direct notion of density previously used in the literature on visual perception[10].

The accuracy of den-SNE's visualization of local density is confirmed by the high correlation based on both measures ($R^2 = 0.650$ for local radius; average $R^2 = 0.657$ for neighborhood count across different length-scales) compared to t-SNE ($R^2 = 0.004$; $R^2 = 0.023$) (Figure 3c; Supplementary Figure 1). Results with densMAP ($R^2 = 0.590$; $R^2 = 0.632$) and UMAP ($R^2 = 0.045$; $R^2 = 0.008$) are analogous (Supplementary Figures 1 and 2). Different parameter choices for UMAP and t-SNE did not improve their density-preservation performance ($R^2 < 0.05$ in all cases; Supplementary Figure 3), as is expected based on our theoretical analysis (see Methods). We also observed that even on previously proposed metrics of visualization quality based on clustering accuracy and pairwise distance preservation[6], our density-preserving tools largely preserve or improve upon the performance of the original methods (Supplementary Note 4; Supplementary Figures 4–8). Traditional dimensionality reduction approaches, including principal component analysis[11] (PCA), multidimensional scaling[12] (MDS), and Isomap[13], were ineffective *both* at preserving density and at visualizing clustering structure (Supplementary Figure 9). Our improved visualizations of simulated datasets in Figure 2 are similarly supported by our quantitative measures (Supplementary Figure 10).

Our visualizations motivate *transcriptomic variability* as a key distinguishing factor among cell types and biological conditions. To illustrate, we examined tumor-infiltrating lymphocytes (TILs) compared to those in blood. While essential in the anti-tumor immune response[14], these cells' molecular mechanisms in cancer remain poorly understood. Density-preserving visualization highlighted the increased transcriptomic variability of T and B cells compared to their counterparts in blood (Figure 3d). Despite an apparent size-difference between the tumor and blood TILs in t-SNE, lack of density-preservation means this pattern could only imply a difference in cell counts, not in variability of expression.

Ranking genes by their contribution to the increase in transcriptomic variability in tumor implicated several biological processes as potential driving factors of TIL diversity (Methods; Supplementary Tables 1–10). Top genes for CD8 T cells and CD4 memory T cells were significantly enriched in negative regulation of IL2 production, transcription, and metabolic processes, suggesting that T cells in tumor are subjected to variable degrees of proliferation control, likely in response to biochemical signals in the tumor microenvironments (Supplementary Tables 6 and 7). Notably, RGS1 and DUSP4 showed the largest difference in variability for both T cell types. RGS1 encodes a regulator for the G-protein signaling pathway known to be involved in chemokine-induced lymphocyte migration[15], and DUSP4 encodes a phosphatase that modulates a T cell receptor signaling pathway with known association with immunological disorders[16]. We validated the variability difference of these two genes in CD8 T cells between tumor and blood based on another scRNA-seq dataset of TILs from NSCLC patients[17], along with 7 other genes in our list of genes ranked by contribution to variability (9 out of 19 genes were found to have significant increase in variance in tumor in the validation dataset; Supplementary Table 11). On the other hand, top genes for naïve CD4 T cells are enriched in proteins targeting membranes and in those that ensure the decay of mis-transcribed mRNA (Supplementary Table 8). For B cells, key biological processes underlying the variability difference included leukocyte activation and protein complex assembly for memory B cells, and response to cyclic AMP (a known modulator of cell proliferation) and biotic stimulus for naïve B cells, along with transcriptional and metabolic regulation processes similar to those implicated for T cells (Supplementary Tables 9 and 10).

While many genes implicated here are lowly-expressed in blood and activated in tumor, we also found a substantial portion (42% among top 20 genes across all cell types) that show statistically significant *overdispersion* in tumor, whereby the increase in variance cannot be explained by an increase in mean expression (Methods, Supplementary Tables 1–5). In fact, some genes, e.g. RPS27 in naïve B cells, which encodes the MPS-1 protein that modulates the activity of tumor-suppressor p53[18], show a significant increase in variance *without* a significant change in mean. These genes are especially common in the top genes for naïve CD4 T cells. Their stability in mean expression implies that these key distinguishing genes cannot be identified by conventional differential expression analysis. Moreover, since standard visualization algorithms separate clusters largely based on difference in mean expression, the effects of these genes are lost in their visualizations. Our findings demonstrate that the transcriptomic variability landscape uncovered by our visualizations helps open new analytic directions for the study of anti-tumor immune response.

Visualizing immune cell specialization and diversification in peripheral blood.

While the above illustrates changing patterns of variability that come about due to disease, we show here that variability of expression *within* cellular subtypes also reveals underlying biology. We used densMAP to visualize a benchmark scRNA-seq experiment that profiled 68,551 peripheral blood mononuclear cells (PBMC) from 10X Genomics[8]. While both

UMAP and densMAP separate the various clusters corresponding to different cell types, the densMAP embedding considerably expands the sizes of natural killer (NK) cells, cytotoxic T cells, CD14+ monocytes and dendritic cells (DCs), and shrinks naïve cytotoxic T cells (Figure 4a). Similar to the cancer dataset, the sizes of these clusters in UMAP correspond to the number of cells belonging to them and do not accurately reflect their variability of expression. By quantifying the agreement between the local radius in the original dataset and the local density measures in each visualization (Methods), we confirmed that densMAP more accurately preserves density ($R^2 = 0.712$ for local radius; average $R^2 = 0.727$ for neighborhood count), compared to UMAP ($R^2 = 0.000$; $R^2 = 0.000$) (Figure 4c; Supplementary Figure 11). The same pattern is observed when comparing den-SNE to t-SNE, with the density correlations in den-SNE much higher ($R^2 = 0.704$; $R^2 = 0.696$) than in t-SNE ($R^2 = 0.052$; $R^2 = 0.037$) (Supplementary Figures 11 and 12).

We focus here on the monocyte and DC clusters, which are strikingly different between the two visualizations (Figure 4b). While both reveal two subtypes of monocytes, densMAP separates them by density, with a dense subcluster adjacent to a much sparser one. Clustering these cells in the original gene expression space indeed identifies the two subtypes as separate clusters (Supplementary Figure 13). These cells begin life as *classical* monocytes, characterized by expression of CD14 and a lack of CD16 (also called FCGR3A); these can then differentiate into CD16 monocytes, macrophages, or dendritic cells (DCs)[19] (Figure 4d). Marker gene expression associated the sparse cluster with classical monocytes and the dense cluster with CD16 monocytes (Figure 4f), suggesting that classical monocytes exhibit a high level of variability before developing into more homogeneous CD16 monocytes. This trajectory has intriguing biological significance. Recent work has revealed that monocytes are an extremely heterogeneous cell type with complex intermediate states[20] and high transcriptional diversity[21]. However, non-classical monocytes are more specialized: they are thought to emerge from a small population of intermediate (CD14+CD16+) monocytes and spike rapidly during infections[20]; since their progenitor cell is rare and accounts for a small portion of transcriptional diversity represented by CD14 monocytes (Figure 4f), this supports the notion of a bottleneck in the development of non-classical monocytes.

We validated this difference in variability between classical and non-classical monocytes in two other scRNA-seq datasets of immune cells, one that profiled 1,078 monocytes, DCs and their subtypes[22] (PBMC2) and the other that profiled 13k PBMCs from two healthy donors[23] (PBMC3). In both, classical monocytes were sparser than non-classical ones: classical monocytes had larger local radii in the gene expression space than non-classical monocytes (one-sided Mann-Whitney U test, $p = 6.61\times10^{-7}$ for the PBMC2 dataset, $p = 2.89\times10^{-4}$ for the PBMC3 dataset, see Methods and Supplementary Figure 14).

A similar analysis can be performed on the DC subset: this cell type shows (i) a dense cluster of cells adjacent to the CD14 monocytes, (ii) a dense cluster overlapping the CD16 monocytes, and (iii) a sparser cluster near the CD14 monocytes (Figure 4b). While the classification of dendritic cells is still actively researched[24], the colocalization of the DCs (i) and (ii) and the monocytes in the densMAP visualization indicates that these DCs originate from monocytes. By analyzing the expression of the marker genes of DC subtypes identified by the PBMC2 study[22] in these DC subsets, we hypothesize that (i) corresponds to classical monocyte-derived DCs (cDCs; DC3 in PBMC2); (ii) corresponds to the poorly understood CD141– CD1C– DCs (DC4 in PBMC2); and (iii) corresponds to plasmacytoid DCs (pDCs; DC6 in PBMC2) (see Supplementary Figure 15). Despite the apparent closeness of the DC6 and DC4 clusters, we did not find any evidence that either subtype is derived from the other.

Our visualizations reveal that the DC3 cluster is far denser than the CD14 monocytes colocated with it, hinting that, as with CD16 monocytes, these cells specialize as they develop from CD14 monocytes. Similarly, in PBMC2, the DC3 cluster is significantly denser than the classical monocyte cluster (one-sided Mann-Whitney U test, $p = 5.43\times10^{-14}$; Methods and Supplementary Figure 14). In addition, the pDC cluster expands drastically in the density-preserving visualization compared to the standard visualization, revealing previously hidden variability (Figure 4b). The PBMC3 dataset was omitted from this analysis as it contained too few DCs to draw conclusions about subtypes.

We also note the DCs dispersed throughout the CD14 monocytes (Figure 4b). When we classify the DC3 subset into dense and sparse categories based on their original local radius (with a log-scale threshold of 3.9), we find that the sparse subset has *intermediate* expression of the marker genes of DC3 and those of CD14 monocytes (Supplementary Figure 15). While this could be due to misclassification (the original study assigned cell types based on similarity to purified samples), it could also indicate a bridging state between the two cell types, offering insights into the dynamics of cell state transition. These results suggest that there are key differences in transcriptomic variability among immune cell subtypes that are obscured by existing visualization tools.

Visualizing time-dependent transcriptomic variability in *C. elegans* development.

To explore embryo development at high-resolution, Packer et al. (2019) performed scRNA-seq profiling of *C. elegans* to create an atlas of gene expression at almost every cell division of the embryo[9]. We asked whether density-preserving visualization could better capture the diversification (or lack thereof) of different developmental lineages, complementing investigations into time-dependent patterns of gene expression in organism development and cellular differentiation[25–27].

For most of the cell types profiled, the lineage distance between cells correlates strongly with transcriptomic dissimilarity, and many cells from the same progenitor diverge after gastrulation[9]. Thus, an accurate visualization should show that the density of cells for most cell types decreases over time (reflecting increasing diversity), as the cells adopt their terminal fates. While both densMAP and UMAP show a central “progenitor” region that branches into the different major tissues, densMAP more clearly highlights the increase in variability in the outer branches of the lineages (Figures 5a and b). Evaluating the agreement between the local radius in the original dataset and both measures of local density in the visualization show that densMAP ($R^2 = 0.590$ for local radius; average $R^2 = 0.585$ for neighborhood count) more accurately preserves density than UMAP ($R^2 = 0.045$; $R^2 = 0.052$) (Figure 5c and Supplementary Figure 16). Results are analogous when comparing den-SNE ($R^2 = 0.619$; $R^2 = 0.596$) to t-SNE ($R^2 = 0.000$; $R^2 = 0.063$) (Supplementary Figures 16 and 17).

While transcriptomic variability generally increases over the course of differentiation, notable exceptions are also made apparent by densMAP. Specifically, of the cell types well-represented (greater than 1000 cells), the intestinal, body-wall muscle (BWM), and hypodermis cells show relative homogeneity in density (measured by the average local radius in the original dataset across time) throughout embryo development when compared to other cell types, e.g. both non-amphid and amphid neurons and seam cells; densMAP more accurately preserves these temporal changes in local density than UMAP (Figures 5d and e).

The underlying biology supports these visual patterns since intestinal, BWM, and hypodermis cells are so-called *semi-clonal lineage clades*[9]. A semi-clonal lineage model is intermediate between *clonal* development, which closely adheres to the lineage structure whereby branching patterns in cell proliferation leads to increasingly more divergent cells, and *non-clonal* development, where daughter cells are only loosely associated with their progenitors and different lineage branches share commonalities through horizontal transitions[28]. Semi-clonal cell types are thus expected to remain more compact in expression space than clonal lineages. Indeed, when we compare the average change in density over embryo time for semi-clonal cells, this change is considerably lower than the average change for the other cell types (Supplementary Figure 17). The difference in density between these semi-clonal cell types and the rest is made clear in our density-preserving visualization but completely hidden by UMAP. In fact, the UMAP plots tend to show a *decrease* in density in many lineages because fewer cells were profiled at the late time-points (Supplementary Figure 17). Our methods can thus accurately portray continuous changes in transcriptomic variability in developmental trajectories, which are not captured by existing visualization tools.

General applicability of density-preserving data visualization.

Visualizing high-dimensional data is broadly useful both within and outside biology. Like t-SNE and UMAP, our density-preserving methods require only a distance metric defined between data points. To illustrate the performance of our methods on other data domains, we analyzed a genotype dataset from the UK Biobank and the MNIST image dataset widely used by the machine learning community (Methods).

The UK Biobank[29] (UKBB) project collects extensive genotypic and phenotypic data from British individuals for use in health-related research. Due to the skew in ethnicity of the British population, most of the individuals in the dataset self-identify as white (94% of the 534k individuals). This lack of diversity has raised concerns about ethnic biases in downstream scientific analyses[30]. When visualizing the individuals in the dataset based on their genotype profiles, an analytic approach that is increasingly being explored[31], t-SNE and UMAP show the cluster corresponding to whites disproportionately large, while the clusters corresponding to Asian and black people can scarcely be seen (Extended Data Figure 1). Visualizing this data using den-SNE and densMAP results in a more balanced representation of ethnicities, considerably expanding the people-of-color clusters and shrinking the white cluster. Existing visualization tools thus grossly under-represent the genetic diversity of minority populations due to their limited sample sizes. Even among the white population, density-preserving visualizations obtain a more balanced representation of subpopulations (computationally identified; Methods). In the UMAP and t-SNE visualizations, only the two most populous subgroups take up significant space, whereas densMAP and den-SNE show five subgroups with comparable diversity.

A complementary situation occurs in the MNIST dataset, a dataset of handwritten digit images (Methods). Here, t-SNE and UMAP generate ten evenly sized clusters; den-SNE and densMAP visualizations, however, reveal that the cluster corresponding to the digit **1** is strikingly less variable than the other digits (Extended Data Figure 2). This is as expected, since **1** is drawn with considerably limited degrees of freedom. Analyzing the local radii in the original data reveals that, indeed, **1** has a higher density than the other digits. The improved accuracy of our visualizations for UKBB and MNIST datasets are supported by both density-preservation metrics based on local radius and neighborhood count (Extended Data Figures 1 and 2 and Supplementary Figures 18 and 19). Taken together, these results show that density-preserving visualization reveals insights about the data not captured by the existing methods on diverse types of datasets.

Density-preserving visualization is almost as computationally efficient as existing approaches.

As experimental methods continue to generate larger datasets, computational tools to analyze them need to scale as well. By leveraging computations already done by t-SNE and UMAP, our density-preserving methods incur only $O(n)$ additional computation (in dataset size) and achieve the same asymptotic scaling as those methods. Although density preservation increases the overall runtime of den-SNE and densMAP (overhead of about 30% for den-SNE and 20% for densMAP on our largest dataset with 250k points; Extended Data Figure 3), we believe that this additional cost is not onerous, when weighed against additional information conveyed by accurately depicting density. While t-SNE, even without density preservation, has limited scalability to datasets approaching many hundreds of thousands of cells, recent computational improvements to t-SNE for massive datasets[32,33] could be augmented with our density-preservation technique. The memory requirements of den-SNE and densMAP are nearly identical to those of t-SNE and UMAP, respectively (Extended Data Figure 3).

Discussion

Effective tools for visualizing the single-cell landscapes captured by ever-larger single-cell experiments are pivotal for accelerating and disseminating discoveries. den-SNE and densMAP overcome a major limitation of the state-of-the-art tools t-SNE and UMAP that they neglect differences in the *local variability* of gene expression across the transcriptomic landscape. While t-SNE and UMAP remain useful for revealing clustering or trajectory patterns, we demonstrated on a range of datasets that the local density information we incorporate into our visualizations harbors insights that can enrich our understanding of biology beyond what existing visualization tools offer. Our density-preservation techniques are broadly applicable to other visualization algorithms, including recent extensions of t-SNE[33,34] and force-directed layout embedding[35,36] (FDLE), and also to other types of biological data where visualization has been useful, such as scATAC-seq[37] and metagenomics[38].

In theory, targeted analyses could also capture the changes in transcriptomic variability made apparent by our visualizations (e.g. by comparing the variance of gene expression between cell types[39]). However, by visualizing this information over the entire dataset, our approach allows easier interpretation and understanding. This methodological shift is akin to how t-SNE and UMAP have streamlined cell-type identification workflows by visually revealing clustering patterns in the dataset, despite the fact that clustering algorithms could be applied independently of visualization. Similarly, our methods can help researchers to easily grasp variability changes in their data and, consequently, to generate biological hypotheses.

Its analytical benefits aside, density-preserving visualization, as our results illustrate, more faithfully represents the underlying structure of the dataset. Even as the community becomes increasingly aware of the intricate limitations of existing visualization tools, inaccurate visualizations will continue to expose researchers to potential biases in data interpretation. A large body of work in the social sciences highlights the problematic nature of inaccurate visualizations: for example, even though distortions in Mercator projections of the world map are well-known, they still suggest biased conclusions to viewers[40,41]. Our density-preserving visualization tools will reduce such distortions and can help prevent unintentional biases and misdirection when researchers interpret and share insights from these data.

Our work motivates a number of directions for further research. First, the changes in transcriptomic variability we discovered in tumor-infiltrating immune cells suggest *differential variability* as a general tool for characterizing different cell states. A change in variability likely reflects underlying alterations of gene regulatory programs, and identifying the key drivers of this pattern and their roles merits further exploration. Our visualizations also motivate local density measures for noise reduction, as they often reveal fine-grain structure within a cell type, typically a dense “core” surrounded by a sparse cloud of cells with more divergent expression patterns. By focusing on only this core, one could obtain crisper canonical representations of cell states and developmental trajectories. Lastly, other popular tools for scRNA-seq analysis based on the nearest-neighbors representation of the transcriptomic landscape may also benefit from information about local variability, motivating density-augmented algorithms for tasks such as clustering[42], trajectory analysis[43], and data integration[44]. Our work represents a key step forward in understanding the dynamic structure of complex single-cell transcriptomic landscapes.

Methods

Review of t-SNE and UMAP.

The most widely-used nonlinear visualization algorithms in single-cell transcriptomic analysis are t-SNE[3] and UMAP[4], and both follow a similar methodology. They first compute a nearest-neighbor graph of the high-dimensional data and introduce a type of probability distribution on the edges of this graph that assigns larger weights on smaller distances. (For t-SNE, this distribution is over *all* edges, and for UMAP, it is over *each* edge.) They then choose an embedding that minimizes the distance between this original probability distribution and a similar distribution computed on the embedding. The key differences between the two algorithms lie in their choices of these distributions and the objective function quantifying the difference between the two distributions.

Let $X = \{x_i\}_{i=1}^{n}$ be our input dataset with *n* data points, where each $x_i \in \mathbb{R}^d$ (e.g. gene expression profile of a cell). Let *E* be the set of edges (*i, j*) in the (directed) *k*-nearest neighbor graph constructed on this dataset, where *j* is one of the *k* points closest to *i*. For t-SNE, the probability distribution on the original data, $P_{ij}^{t-\text{SNE}}$, is given by normalizing and symmetrizing Gaussian kernel distances:

$$\begin{aligned}\widetilde{P}_{j \mid i} &= \exp\left(- \parallel x_i - x_j \parallel^2 / \sigma_i^2\right)\\ Z_i &= \sum_{j:(i,j) \in E} \widetilde{P}_{j \mid i}\\ P_{ij}^{\text{t}-\text{SNE}} &= \frac{1}{2n}\left(\frac{\widetilde{P}_{j \mid i}}{Z_i} + \frac{\widetilde{P}_{i \mid j}}{Z_j}\right)\end{aligned} \tag{1}$$

where σ_i is chosen adaptively for each *i* and corresponds the length-scale at x_i.

UMAP uses a slightly different kernel, representing a rescaled exponential distribution:

$$\begin{aligned}\widetilde{P}_{j \mid i} &= \exp\left(-\left(\parallel x_i - x_j \parallel - \text{dist}_i\right)/\gamma_i\right)\\ P_{ij}^{\text{UMAP}} &= \widetilde{P}_{j \mid i} + \widetilde{P}_{i \mid j} - \widetilde{P}_{j \mid i}\widetilde{P}_{i \mid j}\end{aligned} \tag{2}$$

where γ_i is chosen adaptively and also corresponds to the length-scale, and dist_i is the distance from x_i to its nearest neighbor. We expand on the role of σ_i and γ_i in the next section.

For the probability distributions computed on the embedding, both t-SNE and UMAP use a heavy-tailed distribution (e.g. Student's *t*-distribution for t-SNE), which emphasizes preserving local structure in the original dataset while being more lenient towards longer distances (see the original papers[3,4] for a thorough explanation). Formally, the probability distributions $Q_{ij}^{t-\text{SNE}}$ and Q_{ij}^{UMAP} in the embedding are defined as

$$\widetilde{Q}_{ij}(a,b) = \left(1 + ad_{ij}^{2b}\right)^{-1} \tag{3}$$

$$\mathscr{Z}_i(a,b) = \sum_{j \neq i} \widetilde{Q}_{ij}(a,b) \tag{4}$$

$$Q_{ij}^{\text{t}-\text{SNE}} = \widetilde{Q}_{ij}(1,1)\left(\sum_k \mathscr{Z}_k(1,1)\right)^{-1} \tag{5}$$

$$Q_{ij}^{\text{UMAP}} = \widetilde{Q}_{ij}(a,b) \tag{6}$$

where d_{ij} represents the distance between points i and j in the embedding (Euclidean for both methods), and a and b are additional shape parameters UMAP introduces to control the spread of the distribution according to a user parameter. In the following, we omit the superscripts of P and Q when they are clear from the context.

The goal of both algorithms is to generate an embedding that minimizes the difference between P and Q. The loss function used by t-SNE to quantify this difference is the Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(P \parallel Q) = -\sum_{ij} P_{ij}\left(\log P_{ij} - \log Q_{ij}\right).$$

UMAP instead uses the cross-entropy (CE) loss summed over all the edges:

$$\mathrm{CE}(P \parallel Q) = -\sum_{ij} P_{ij}\log Q_{ij} + \left(1 - P_{ij}\right)\log\left(1 - Q_{ij}\right).$$

Both methods optimize the embedding coordinates to minimize the respective loss functions using standard gradient descent optimization techniques (see Supplementary Note 2 for details). Notably, the fact that UMAP does not require Q to be renormalized over all edges allows UMAP to use *stochastic* gradient descent (whereby the embedding coordinates are updated for one data point at a time), making it more computationally efficient than t-SNE in general.

Adaptive length-scale selection in t-SNE and UMAP erases density information.

The length-scale parameters σ_i and γ_i play an important role. The exponentially-decaying tails of the P distribution in both t-SNE and UMAP mean that the points a few multiples of the length-scale away from x_i are effectively omitted from the conditional distribution $P_{\cdot|i}$. Thus, the choice of the length-scale at point x_i determines the radius of the local structure around x_i that the embedding aims to preserve. Since different points in the dataset can have vastly different distribution of distances to their respective nearest neighbors, it is desirable to use a different σ_i or γ_i for each point x_i in order to evenly capture the local structure across all parts of the data.

In t-SNE, the σ_i's are chosen by setting the *perplexity* of each conditional distribution $P_{\cdot|i}$ constant. Perplexity can be thought of as a "smooth" analog of the number of nearest neighbors and is formally defined as $\mathrm{Perp}_i = 2^{H_i}$, where H_i denotes the entropy of the conditional distribution $P_{\cdot|i}$:

$$H_i = -\sum_j P_{j\,|\,i}\log_2 P_{j\,|\,i}. \tag{7}$$

Since perplexity monotonically increases in σ_i (more points are significantly represented in $P_{\cdot|i}$ as σ_i increases), t-SNE performs a binary search on each σ_i to obtain a constant perplexity for all i. UMAP's length-scale selection is analogous, but instead of fixing the value of perplexity, it fixes the marginal sum of probabilities at each point *i*, $\sum_j P_{ij}$, by choosing an appropriate γ_i.

Although it is effective for capturing local structure, adaptive choice of length-scale has the undesirable consequence of canceling out differences in density around each point in the original data, as t-SNE (implicitly) and UMAP (explicitly) both assume the data points are distributed uniformly on an underlying manifold. Note that, in both t-SNE and UMAP, a sparse neighborhood of x_i leads to a large length-scale, whereas a dense neighborhood leads to a small length-scale. Since the distance between points is divided by the length-scale parameter in the computation of *P*, we can intuitively see that this normalization removes density information from the data.

More formally, consider a dataset of points $X = \{x_i\}_{i=1}^{n}$ with Euclidean pairwise distances $\{d_{ij}\}_{i,j=1}^{n}$. Suppose we dilate the data space by a factor of $\alpha > 1$ to generate a sparser dataset $Z = \{z_i\}_{i=1}^{n}$ with the same underlying structure, where the new pairwise distances are scaled by α, i.e. $\|z_i - z_j\| = \alpha d_{ij}$. A key observation is that the distribution *P* computed on *X* by t-SNE or UMAP will be *identical* to *P* computed on *Z*, even though *Z* represents a more heterogeneous set of points than *X*. Intuitively, this is because obtaining the same perplexity/marginal sum of probabilities on *Z* requires that the respective length-scales be scaled by α, which cancels out the increase in distances and leaves the resulting *P unchanged*. Since *P* is the only information about the dataset provided as input to the embedding step of each algorithm, the original differences in density in different regions of the data space are entirely lost in the embedding. We provide a more detailed description of this property and its generalization to a broader class of generative models for the underlying data in Supplementary Note 3.

Our approach: capturing density information using the local radius.

To generate embeddings that retain information about the density at each point, we introduce the notion of a *local radius* to make concrete our intuition of spatial density. Intuitively, a point is in a *dense* region if its nearest neighbors are very close to it, and in a *sparse* region if its nearest neighbors are far away. Thus, we use average distance to nearest neighbors as a measure of density for a given point.

To formalize this notion, for a point x_i, we require two components: (i) a pairwise distance function $d(x_i, x_j)$, and (ii) a probability distribution $\rho_{j|i}$ that weighs each x_j based on its distance from x_i, with faraway points having lower weights. We define the local radius at x_i, denoted $R_\rho(x_i)$, as the expectation of the distance function over x_j with respect to $\rho_{j|i}$, thus capturing the average distance from x_i to nearby points:

$$R_\rho(x_i) := \mathbb{E}_{j \sim \rho_{j \mid i}}\left[d\left(x_i, x_j\right)\right]. \qquad (8)$$

In the following, we let the distance function be the squared Euclidean distance, i.e. $d(x_i, x_j) = \|x_i - x_j\|^2$, which we found to have better empirical performance than standard Euclidean distance. Other choices of distance function can be easily incorporated into our framework.

In den-SNE and densMAP, we take advantage of the probability distributions $P^{\text{t-SNE}}$ and P^{UMAP} which already capture local relationships; for the local radius in the original embedding, we renormalize the edge probabilities P_{ij} to obtain a conditional distribution $\rho_{j|i} = P_{ij}/\sum_j P_{ij}$ and calculate the local radius as

$$R_P(x_i) = \frac{1}{\sum_j P_{ij}} \sum_j P_{ij} \parallel x_i - x_j \parallel^2$$

for both methods. Note that P vanishes rapidly outside the neighborhood of each x_i and is thus well-suited for density estimation. We can show in fact that this representation of density (inversely-related) has the desirable property that it scales with the *variance* of a range of data-generating distributions and increases when the length-scale term σ_i increases (Supplementary Note 3).

Next, we define the local radius in the embedding. Let y_i be the embedding coordinates of the point x_i given by the algorithm of choice. We need a distribution analogous to P for calculating the expected distance between y_i and its neighbors in the embedding. It would still be desirable for this distribution to have adaptive length-scales like P in order to ensure that a comparable number of nearest neighbors are taken into consideration for calculating the local radius at different points in the dataset. However, this would present a major hurdle for optimization because the binary search used to determine σ_i and γ_i is not differentiable. Instead, we leverage the embedding distribution Q computed by t-SNE and UMAP as an approximation for the adaptive scheme. It is worth noting that, in the case of t-SNE, Q is based on a Cauchy distribution, which can be interpreted as the marginalization of a Gaussian distribution over an unknown variance[45]. Thus, Q intuitively reflects an average over all length-scales. Letting $\rho_{j|i} = Q_{ij}/\sum_j Q_{ij}$ and $d(y_i, y_j) = \|y_i - y_j\|^2$, the local radius in the embedding is given as

$$R_Q(y_i) = \frac{1}{\sum_j Q_{ij}} \sum_j Q_{ij} \parallel y_i - y_j \parallel^2 . \tag{9}$$

Note that we adopt the squared Euclidean distance for consistency with local radius computation in the original space.

For ease of notation, we denote the local radius in the original data as R_o and the local radius in the embedding as R_e in the following sections.

Augmenting the visualization objective to induce density preservation.

To preserve density, we aim for a *power-law* relationship between the local radius in the original dataset and in the embedding, i.e. $R_e(y_i) \approx \alpha[R_o(x_i)]^\beta$ for some α and β, inspired by the exponential scaling of density with respect to dimensionality (see Supplementary Note 1). This can be reframed as an *affine* relationship between the logarithms of the local radii, i.e.,

$$r_e(y_i) \approx \beta r_o(x_i) + \alpha,$$

where we define $r_o(x_i) := \log R_o(x_i)$ and $r_e(y_i) := \log R_e(y_i)$. The goodness-of-fit of this relationship can be measured by the *correlation coefficient*

$$\mathrm{Corr}(r_e, r_o) = \frac{\mathrm{Cov}(r_e, r_o)}{(\mathrm{Var}(r_e)\mathrm{Var}(r_o))^{1/2}}, \tag{10}$$

which is invariant to the parameters α and β. $\mathrm{Cov}(\cdot, \cdot)$ denotes the covariance function, and $\mathrm{Var}(\cdot)$ denotes the variance function. Note that these quantities are estimated by considering the tuples $\{(x_i, y_i)\}_{i=1}^n$ as n independent samples from the same distribution; e.g., the mean of r_e is estimated as $\frac{1}{n}\sum_{i=1}^{n} r_e(y_i)$.

Our density-preservation objective is to choose the embedding $\{y_i\}_{i=1}^n$ such that correlation between the log local radii of the original dataset and the embedding is maximized. This approach is closely related to canonical correlation analysis[46] (CCA), which finds a linear transformation of a dataset that maximizes its correlation with another. We are further motivated by recent work that extends CCA to nonlinear transformations[47].

Augmenting the loss functions of t-SNE and UMAP with this density-preservation objective yields the den-SNE and densMAP objectives, respectively:

$$\mathscr{L}^{\text{den-SNE}} = \mathrm{KL}\left(P^{\mathrm{t-SNE}} \parallel Q^{\mathrm{t-SNE}}\right) - \lambda \mathrm{Corr}\left(r_o^{\mathrm{t-SNE}}, r_e^{\mathrm{t-SNE}}\right), \tag{11}$$

$$\mathscr{L}^{\mathrm{densMAP}} = \mathrm{CE}\left(P^{\mathrm{UMAP}} \parallel Q^{\mathrm{UMAP}}\right) - \lambda \mathrm{Corr}\left(r_o^{\mathrm{UMAP}}, r_e^{\mathrm{UMAP}}\right), \tag{12}$$

where λ is a user-chosen parameter that determines the relative importance of the density-preservation term compared to the original objective.

Optimizing the embedding with respect to density-augmented objectives.

Our differentiable formulation of the local radius enables us to optimize the density-augmented objective functions (11) and (12) using standard gradient descent techniques. Since both t-SNE and UMAP are also based on gradient descent, it suffices for us to calculate the contribution of the density-preservation objective to the overall gradient and add it to the existing t-SNE and UMAP gradients.

The gradient of the density-preservation objective with respect to the embedding coordinates y_i is given by

$$\nabla_{y_i}\mathrm{Corr}(r_e, r_o) = \sum_{j \neq i}\left[\frac{\partial}{\partial d_{ij}^2}\mathrm{Corr}(r_e, r_o)\right]\left(y_i - y_j\right),$$

where $d_{ij} = \|y_i - y_j\|$. To simplify the notation, let $\mu_e = \mathbb{E}[r_e], r_i^e = r_e(y_i)$, and $r_i^o := \left(r_o(x_i) - \frac{1}{n}\sum_i r_o(x_i)\right)/\mathrm{Var}^{1/2}(r_o)$. Note that the centering of r_i^o and normalizing by standard deviation does not depend on the embedding and thus can be precomputed. Now, the inner gradient term with respect to d_{ij}^2 can be calculated as

$$\frac{\partial}{\partial d_{ij}^2}\mathrm{Corr}(r_e, r_o) = \frac{\mathrm{Var}(r_e)\left(r_i^o\frac{\partial r_i^e}{\partial d_{ij}^2} + r_j^o\frac{\partial r_j^e}{\partial d_{ij}^2}\right) - \mathrm{Cov}(r_e, r_o)\left((r_i^e - \mu_e)\frac{\partial r_i^e}{\partial d_{ij}^2} + (r_j^o - \mu_e)\frac{\partial r_j^e}{\partial d_{ij}^2}\right)}{(n-1)\mathrm{Var}(r_e)^{\frac{3}{2}}}$$

where

$$\frac{\partial r_i^e}{\partial d_{ij}^2} = \frac{\widetilde{Q}_{ij}^2(a,b)}{\mathscr{Z}_i(a,b)}\left[abd_{ij}^{2(b-1)} + e^{-r_i^e}\left(1 + a(1-b)d_{ij}^2\right)\right].$$

The terms $\widetilde{Q}_{ij}(a,b)$ and $Z_i(a,b)$, defined in (3) and (4), respectively, are quantities computed by t-SNE and UMAP to capture the local structure of the embedding. ($Z_i(a,b)$ is required only in t-SNE.) Setting the parameters $a = b = 1$ results in the t-SNE formulation, whereas UMAP sets these two parameters as a function of a user parameter. A detailed derivation of our gradients above is provided in Supplementary Note 2.

Optimizing the densMAP objective requires special consideration because UMAP uses stochastic gradient descent (SGD), whereby edges are sampled according to P_{ij} and the gradient update is performed for one edge at a time. Since the gradient formula (10) involves a sum over its neighbors with equal weights, edges sampled from P must be re-weighted to obtain unbiased estimates of our gradient. To this end, we multiply the density term in the gradient for an edge $\{i, j\}$ by Z/nP_{ij} where $Z = \sum_{\{k,\ell\}\in E} P_{k\ell}$ to correct for sampling bias. In addition, there are a number of global terms that are computationally burdensome to update for every edge, which include $\mathrm{Var}(r_e)$, $\mathrm{Cov}(r_e, r_o)$, and μ_e. We compute these terms in the beginning of each epoch (a round of edge-wise updates for the entire dataset) and consider them as fixed during the epoch. This can be viewed as a form of coordinate descent, where the objective is optimized with respect to a subset of variables at a time while conditioning on the rest. We describe these techniques in detail in Supplementary Note 2.

Implementation details.

To ensure that our methods find good local optima of (11) and (12) that are as effective as t-SNE and UMAP in separating clusters, we take a two-step approach where we run the original algorithms *without* the density-preserving objective for the first q fraction of iterations, then optimize the full objective for the remaining $1-q$ fraction of iterations. This approach is akin to t-SNE's "early exaggeration", whereby the first several iterations of the optimization emphasize attractive forces to help guide the direction of the optimization. We note that an alternative approach is to smoothly activate the density-preserving objective, but because any non-zero weight on this term incurs all of the associated computational overhead with little benefit, we opted for the two-step approach instead.

For computational efficiency, we approximate the embedding distribution Q used in our local radius computation (9) by allowing Q_{ij} to be non-zero only when P_{ij} is non-zero (i.e. i and j are k-nearest neighbors in the original space), thus inducing sparsity in Q. This technique is especially well-suited for the aforementioned two-step approach, since the embedding already closely follows the nearest-neighbor structure in P when this approximation takes effect.

There are several parameters of den-SNE and densMAP that the users can modify to tailor the behavior of these algorithms. We inherit all of the parameters from t-SNE and UMAP, including perplexity (t-SNE) or number of neighbors (UMAP), number of iterations/epochs, and the “min-dist” parameter for UMAP (which controls the a and b parameters in Q_{ij}; see (6)). We refer the readers to the original publications for a detailed discussion of these parameters. There are two additional parameters we introduce in den-SNE and densMAP: the weight $\lambda \geq 0$ given to the density-preserving objective, and the fraction $q \in [0, 1]$ of iterations that take the density term into account. All of our experimental results are based on the following default parameter settings that we recommend. For den-SNE, we use perplexity of 50 and 1000 iterations (same as the default setting of t-SNE), along with $q = 0.3$ and $\lambda = 0.1$. For densMAP, we use 30 neighbors, 750 epochs, $q = 0.3$, and $\lambda = 2$. We note that changing the value of λ leads to qualitatively different embeddings that achieve different trade-offs between the original visualization objective and the density-preservation term (Supplementary Figure 20). For MNIST, we took advice from the scientific community and Kobak et al. (2019) to increase the early exaggeration parameter for t-SNE and den-SNE to 1,000, which resulted in better clustering of the digits[48].

Quantitative evaluation of density preservation.

To assess the performance of visualization algorithms at preserving density, we compute the correlation between the log local radii in the original dataset and two measures of visual density in the embedding generated by the algorithm.

The first measure of visual density is the local radius computed in the same manner as in the original space. Recall that during the optimization, we compute the local radius in the embedding *approximately* using the heavy-tailed distribution Q computed by t-SNE or UMAP and consider only the edges present in the nearest-neighbors graph of the original data. For accurate evaluation, here we compute the local radius more directly as follows. Given the embedding points $\{y_i\}_{i=1}^{n}$, we compute the analog of the P matrix on the original data on these embedding points, denoted P'. For t-SNE and den-SNE, we define P' as:

$$\begin{aligned} \tilde{P}'_{j \mid i} &= \exp\left(- \| y_i - y_j \|^2 / \sigma_i\right) \\ Z'_i &= \sum_j P'_{j \mid i} \\ P'_{j \mid i} &= P'_{j \mid i} / Z'_i \end{aligned}$$

Where σ'_i, the length-scale parameter, is chosen to achieve the same perplexity as in the original P matrix.

For UMAP and densMAP, we define P' as:

$$\tilde{P}'_{j\mid i} = \exp\left(-\left(\| y_i - y_j \| - \text{dist}_i\right)/\gamma'_i\right)$$
$$P'_{ij} = \left(\tilde{P}'_{j\mid i} + \tilde{P}'_{i\mid j} - \tilde{P}'_{j\mid i}\tilde{P}'_{i\mid j}\right)$$
$$P'_{j\mid i} = P'_{ij} / \sum_{j\neq i} P'_{ij}$$

where dist_i is the distance to the nearest neighbor of y_i, and γ'_i is chosen to achieve the same constant marginal $\sum_j P'_{j\mid i}$ as the original P matrix.

Since P' more explicitly focuses on the local neighborhoods of points in the embedding than Q by adaptively choosing the length-scale, calculating the local radius using this distribution more accurately reflects the actual density of each point in the embedding:

$$R_{P'}(y_i) = \sum_{j\neq i} P'_{j\mid i} \| y_i - y_j \|^2$$

Note that the adaptive length-scale ensures that a similar number of neighbors are considered when computing the local radius for both dense and sparse neighborhoods in the embedding. Our quantitative metric of density preservation is then the Pearson correlation coefficient (R^2) between log $R_{P'}(y_i)$ and $r_o(x_i)$ = log $R_P(x_i)$, where the latter is the log local radius in the original data space.

The second measure of visual density in the embedding is the *neighborhood count*, which is motivated by the visual perception of density as the number of points in a given area. For a point y_i in the embedding and a radius ℓ, the ℓ-neighborhood count of y_i is the number of points y_j that are within a distance of ℓ from y_i in the embedding. Thus, dense regions will have large neighborhood counts and sparse regions, small counts. This natural notion of local density has been extensively used in the psychology of vision[10,49].

To systematically choose ℓ for each dataset, we first compute the area A of the smallest bounding box of the embedded points, then calculate an average length-scale $\ell_{ave} = \sqrt{A/n}$, where n is the number of points in the dataset. To assess density preservation across different length-scales, we tested different multiples of ℓ_{ave}; for den-SNE and t-SNE, we chose ℓ from $\{\ell_{ave}, 2\ell_{ave}, 4\ell_{ave}\}$, and for densMAP and UMAP, from $\left\{\frac{1}{2}\ell_{ave}, \ell_{ave}, 2\ell_{ave}\right\}$. We chose smaller values for densMAP and UMAP because those embeddings are more compact in general than those of den-SNE and t-SNE for our parameter choices. For each chosen ℓ, we calculate the ℓ-neighborhood count for each point in the embedding and calculate the correlation (in log space) with the local radii in the original space as a quantitative metric of density preservation. A strong negative correlation is desirable, which indicates that points with a higher neighborhood count (higher visual density) tend to have a smaller local radius in the original dataset (smaller underlying variability).

Additional metrics for evaluating visualization quality.

We additionally evaluated the performance of our methods on three previously proposed metrics of visualization quality on scRNA-seq data[6]: classification score (CS), mutual information score (MIS), and pairwise distance score (PDS). Intuitively, CS and MIS measure clustering accuracy based on the visualization, and PDS measures the preservation of pairwise distances among the datapoints.

More specifically, CS evaluates the accuracy of classifiers that assign each datapoint to one of the known clusters based on the visualization coordinates. Following prior work[6], we trained a random forest classifier on the visualization (60% of the data) to predict the cluster labels from the original dataset using the RandomForestClassifier class in Python scikit-learn package with default parameters. We then calculated the CS as the accuracy of the trained classifier on a held-out test set (40% of the data). We averaged the results across three trials of cross-validation to produce the final score.

MIS measures the agreement between the output of a clustering algorithm in the original and the embedding space. As previously proposed[6], we used agglomerative clustering with $k =$ 100 clusters to generate a high resolution clustering of the original dataset, then applied the same procedure to obtain a clustering based on the visualization. We performed the clustering using scikit-learn's AgglomerativeClustering class with the default Ward linkage. MIS is calculated as the *mutual information* between the two cluster assignments, which measures their agreement. To produce a robust estimate of the score, we computed MIS on three 60% subsamples of the original dataset and averaged the results.

Lastly, for PDS, we sampled 1,000 points at random from the dataset and calculated the score as the squared correlation coefficient (R^2) between the pairwise distances among the chosen points in the original space and those in the visualization, again following the previously proposed approach[6]. Note that this score equally considers all pairs of points regardless of their distance, even though the nonlinear data visualization algorithms like t-SNE and UMAP are designed to focus on preserving distances within local neighborhoods. To more comprehensively assess the preservation of pairwise distances at different scales in the original dataset, we calculated PDS for different subsets of pairwise distances with an increasing upper limit on their original distance in the dataset. More precisely, we calculated the PDS for the bottom x% of pairwise distances in the original space for x ranging from 0 to 100.

Data preprocessing.

We obtained three publicly available scRNA-seq datasets for the main analyses: a dataset of immune cells in lung cancer and blood[7], a dataset of peripheral blood mononuclear cells (PBMCs) in healthy individuals[8], and a dataset that profiled the developmental trajectory of *C. elegans*[9]. We used three additional scRNA-seq datasets for validation experiments, including another lung cancer dataset[17] and two blood immune cell datasets[22,23]. For each dataset, we applied the same cell and gene filtering schemes used by the original publications, then normalized the data so that each cell has the same total number of counts (10k). Following the standard in scRNA-seq analysis, we then log-transformed the normalized counts, i.e. $x \to \log(1 + x)$. Principal component analysis (PCA) was then used to produce lower-dimensional representations of individual cells, which are provided as input to the visualization algorithms. We used the number of principal components (PCs) prescribed by the original publications if present, or used 50 dimensions otherwise. The resulting datasets for the main experiments included 48,969 cells and 306 PCs (34.7% of total variance) for lung cancer, 68,551 cells and 50 PCs (9.7%) for PBMCs, and 86,024 cells and 100 PCs (25.2%) for *C. elegans*. We used the cell type labels provided by the original datasets for visualization.

For the UK Biobank dataset[29], we used the 40 PC loadings provided as part of the genetic data for visualization. We analyzed a 20% subsample of the dataset including 97,676 individuals, for computational efficiency. Ethnicity labels for the individuals were obtained from Data Field 21000, which was collected from the participants via a touchscreen questionnaire. To visualize subpopulation structure within the white British individuals, we performed spectral clustering using the 40 PCs as input to identify five subclusters.

For the MNIST dataset, we flattened each of the 60,000 28×28 pixel images to a 784-dimensional vector and used the top 50 PCs (82.4% of total variance) as our input to the visualization algorithms. Labels classifying the handwritten digits were provided in the dataset.

Differential analysis of gene expression variability in the lung cancer data.

For each cell type with visible expansion of transcriptomic variability in tumor in our visualizations — CD8 T cells (1,621 cells in blood, 443 cells in tumor), CD4 memory resting T cells (1,036 cells in blood, 9,019 cells in tumor), CD4 naïve T cells (437 cells in blood, 61 cells in tumor), memory B cells (67 cells in blood, 4,811 cells in tumor), and naïve B cells (83 cells in blood, 396 cells in tumor) — we identified twenty genes with the largest increase in variance in tumor compared to blood for further analysis. For each gene and cell type, we calculated the differences in the mean and variance of expression between tumor and blood. The statistical significance of the observed differences is assessed using a permutation test, whereby the assignment of cells to tumor or blood is randomly permuted, and the statistic computed on the permuted dataset is viewed as samples from the null distribution where there is no difference between tumor and blood. For comparing the variance, we centered the expression levels for each group (tumor or blood) before the permutation procedure to control for the shift in mean. The p-value is calculated as the fraction of permutations that result in a statistic whose magnitude is larger than the statistic computed on the original dataset. We used 100k permutations to estimate the p-values and applied Bonferroni correction within each cell type to account for multiple hypothesis testing.

When considering changes in the variance of gene expression, it is important to note that an increase in variance can often be explained by an increase in mean. For example, under the Poisson process model of underlying count distributions, variance of the observed counts naturally scales with the mean[39]. Thus, we additionally calculated the difference in *dispersion index* to assess the extent to which the change in variance is unexplained by a corresponding change in mean. The dispersion index (DI) is given by σ^2/μ, where μ and σ^2

are the mean and variance of expression. We assessed the statistical significance of the difference in DI also using a permutation test. For the null distribution, we assume that in the absence of excess difference in dispersion, the variance of expression has a linear dependence on the mean (as suggested by the dispersion index). A permutation scheme that correctly reflects this null distribution is one where the expression levels within each group (tumor or blood) are transformed as $x \mapsto \mu^{-1/2}(x-\mu)+1$ before the permutation, where μ is the sample mean of the group. This transformation maps both groups to the same mean (μ = 1) while preserving the DI, so that permuting the labels leads to a valid sample from the null distribution. Similar to the mean and variance tests, we used 100k permutations to estimate the p-values and applied Bonferroni correction.

Assessing significance of density differences in monocytes and dendritic cells.

To verify our claims that classical (CD14+) monocytes have more variability of expression than both CD16+ monocytes and DC3 dendritic cells (as characterized by the PBMC2 dataset), we compared the distribution of the log local radius in the original data for each of these cell types in the PBMC2 and PBMC3 datasets. To assess significance, we used the one-sided Mann-Whitney U (MWU) test[50], which tests the hypothesis that values drawn from one distribution are larger than those drawn from another. We calculated the MWU test statistic for: CD14+ monocytes and CD16+ monocytes in the PBMC2 and PBMC3 datasets; and for CD14+ monocytes and DC3 dendritic cells in PBMC2. In PBMC2, there are 163 CD14+ monocytes, 122 CD16+ monocytes, and 107 DC3 cells; in PBMC3, there are 1,264 CD14+ monocytes, 398 CD16+ monocytes, and 142 DCs.

Runtime and memory benchmarking.

To evaluate runtime and memory usage of our density-preserving visualization methods, we used each of the five datasets (three scRNA-seq datasets, UK Biobank, and MNIST) along with logarithmically downsampled subsets of each (i.e. subsamples of size $N/2$, $N/4$, down to 1,000 datapoints for a dataset of size N). dataset from Packer et al. (2019) with 86,024 cells, which is the largest scRNA-seq dataset used in this paper. In addition to the full dataset, we subsampled it into smaller datasets, including 43,012 cells, 21,506 cells, 10,753 cells, and 5,376 cells. We measured the runtimes of denSNE, densMAP, t-SNE, and UMAP on each of the datasets with the default parameter settings and profiled memory usage using the psrecord package (https://github.com/astrofrog/psrecord). All experiments were run on an Intel Xeon Gold 6130 (2.30 GHz) processor and used a single core.

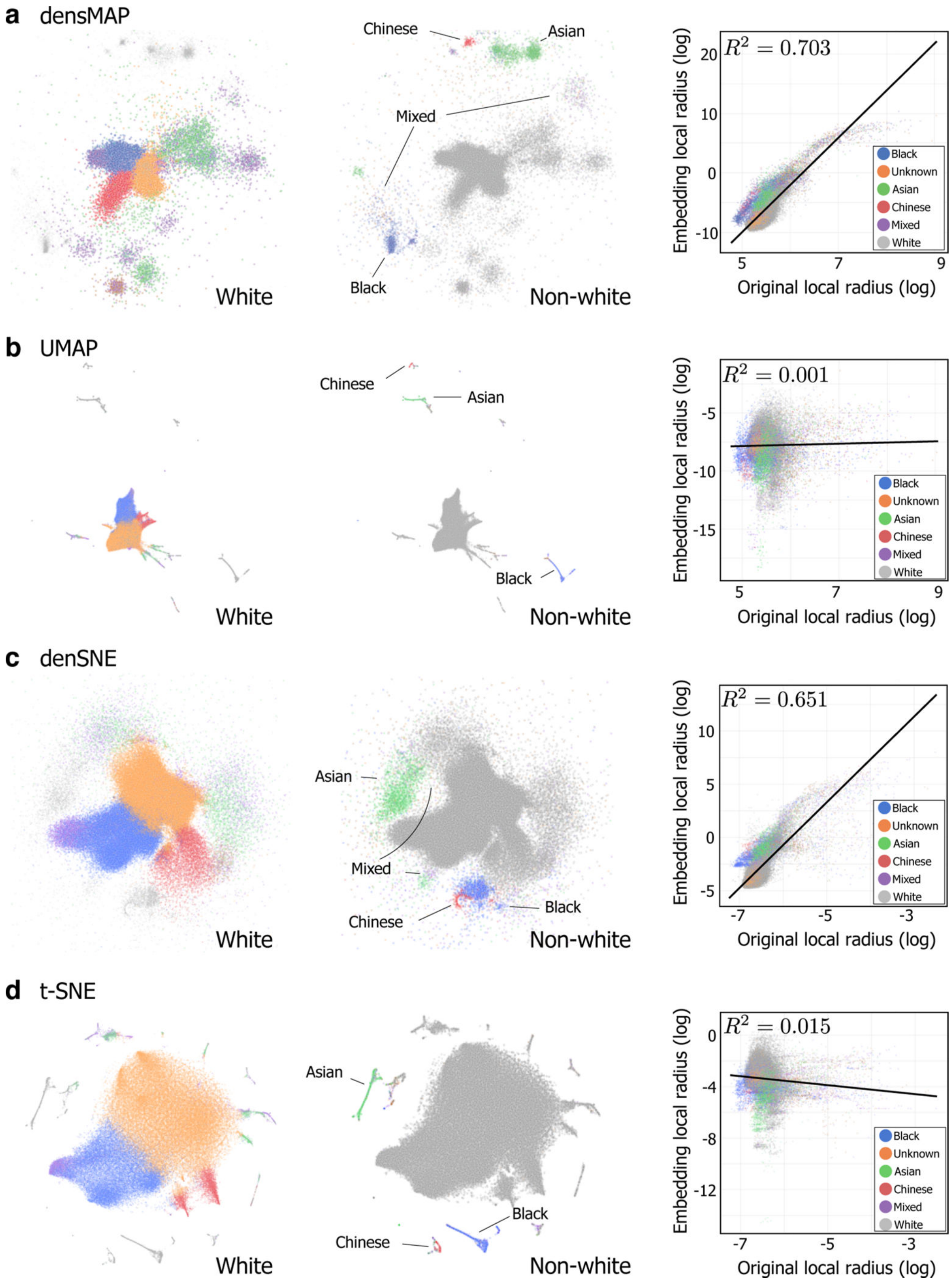
Data availability.

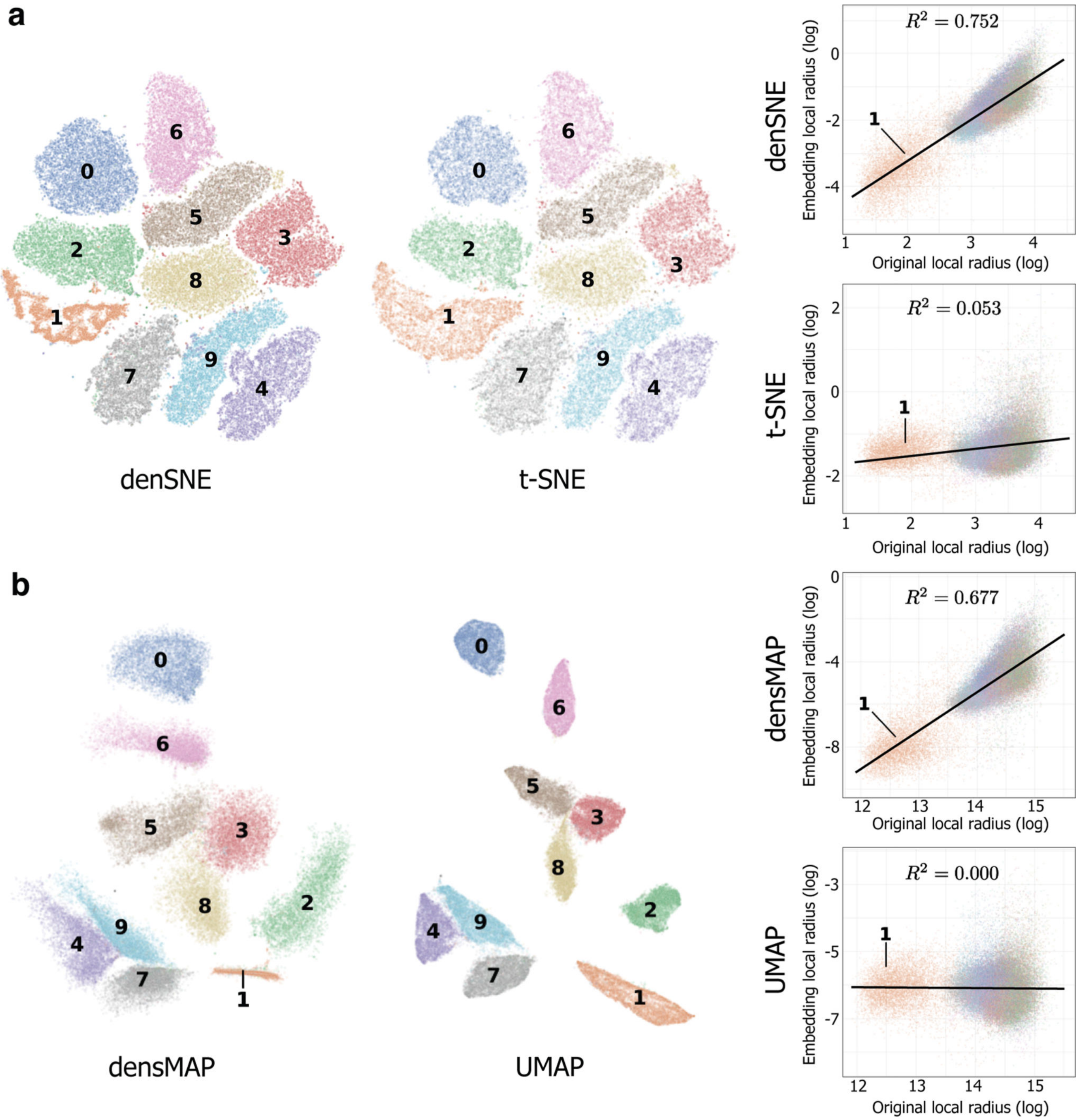
The lung cancer[7] and *C. elegans*[9] datasets are available from the Gene Expression Omnibus (GEO) database with accession numbers GSE127465 and GSE126954, respectively. The PBMC dataset[8] is available from 10x Genomics at: https://support.10xgenomics.com/single-cell-gene-expression/datasets. For our validation datasets, the secondary lung cancer dataset[17] is available from GEO (GSE99254), and the PBMC2[22] and PBMC3[23] datasets can be accessed through the Broad Institute's Single Cell Portal (https://singlecell.broadinstitute.org/) with dataset IDs SCP43 and SCP345, respectively. Data access applications for the UK Biobank data can be submitted at: https://www.ukbiobank.ac.uk/. The MNIST dataset is available at: http://yann.lecun.com/exdb/mnist/. We also provide our preprocessed data for the main datasets (lung cancer, PBMC, and *C. elegans*) at: http://densvis.csail.mit.edu/datasets.

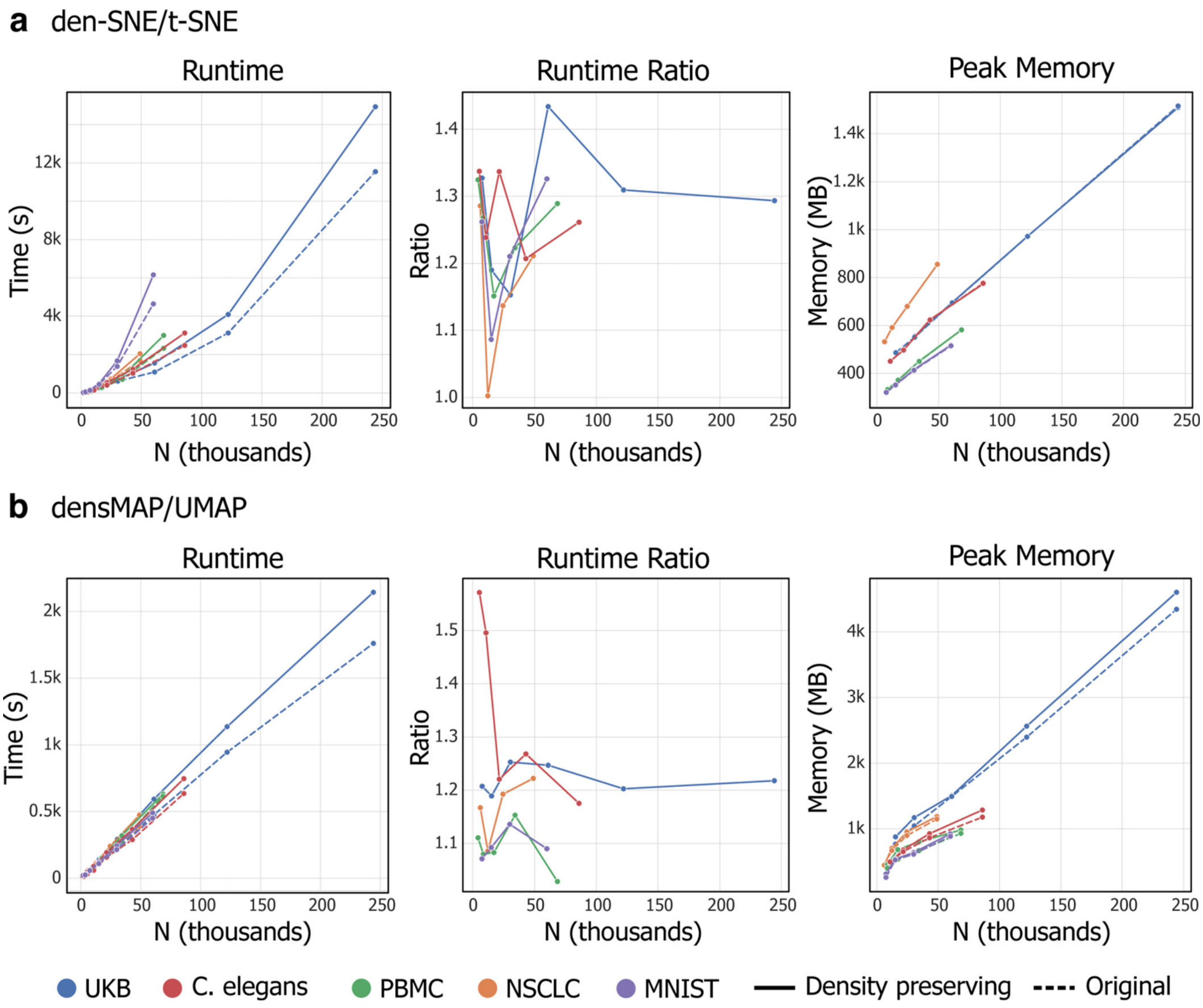
Code availability.

We provide the software for den-SNE and densMAP in the densVis package available at: http://densvis.csail.mit.edu/ and https://github.com/hhcho/densvis. Our densMAP implementation is also available as part of the Python umap package (https://github.com/lmcinnes/umap).

Extended Data




a
0
6
5
2
3
8
1
7
9
4
denSNE
0
6
5
2
3
8
1
7
9
4
t-SNE
denSNE
$R^2 = 0.752$
Embedding local radius (log)
0
-2
-4
1
1
2
3
4
Original local radius (log)
t-SNE
$R^2 = 0.053$
Embedding local radius (log)
2
0
-2
1
1
2
3
4
Original local radius (log)
b
0
6
5
3
8
2
4
9
7
1
densMAP
0
6
5
3
8
2
4
9
7
1
UMAP
densMAP
$R^2 = 0.677$
Embedding local radius (log)
0
-4
-8
1
12
13
14
15
Original local radius (log)
UMAP
$R^2 = 0.000$
Embedding local radius (log)
-3
-5
-7
1
12
13
14
15
Original local radius (log)



Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is in part supported by NIH U01 CA250554 (to B.B). H.C. is partially supported by Eric and Wendy Schmidt through the Schmidt Fellows Program at the Broad Institute. The authors thank Brian Hie, Benjamin DeMeo, Ellen Zhong, and Joshua Peters for helpful discussions. Our visualization of genotype data was conducted using the UK Biobank Resource under Application Number 46341 in keeping with the informed consent given by its participants. BioRender.com was used to generate Figure 4d.

References

1. Hie B. et al. Computational Methods for Single-Cell RNA Sequencing. Annual Review of Biomedical Data Science 3, 339–64 (2020).

2. Chen G, Ning B. & Shi T. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Frontiers in Genetics 10, 317 (2019). URL https://www.frontiersin.org/article/10.3389/fgene.2019.00317. [PubMed: 31024627]

3. van der Maaten L. & Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008). URL http://www.jmlr.org/papers/v9/vandermaaten08a.html.

4. McInnes L. & Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv (2018).

5. Amir E-a. D. et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nature Biotechnology 31, 545–552 (2013). URL 10.1038/nbt.2594.

6. Becht E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology 37, 38 (2019).

7. Zilionis R. et al. Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity 50, 1317–1334 (2019). [PubMed: 30979687]

8. Zheng GX et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications 8 (2017).

9. Packer JS et al. A lineage-resolved molecular atlas of C. Elegans embryogenesis at single-cell resolution. Science 365 (2019).

10. Healey CG & Enns JT Large datasets at a glance: combining textures and colors in scientific visualization. IEEE Transactions on Visualization and Computer Graphics 5, 145–167 (1999).

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

12. Cox T. & Cox M. Multidimensional scaling, second editionth edn (2001).

13. Tenenbaum JB, De Silva V. & Langford JC A global geometric framework for nonlinear dimensionality reduction. science 290, 2319–2323 (2000). [PubMed: 11125149]

14. Whiteside TL & Parmiani G. Tumor-infiltrating lymphocytes: their phenotype, functions and clinical use. Cancer Immunology, Immunotherapy 39, 15–21 (1994). URL 10.1007/BF01517175. [PubMed: 8044822]

15. Bignon A. et al. DUSP4-mediated accelerated T-cell senescence in idiopathic CD4 lymphopenia. Blood, The Journal of the American Society of Hematology 125, 2507–2518 (2015).

16. Agenes F, Bosco N, Mascarell L, Fritah S. & Ceredig R. Differential expression of regulator of G-protein signalling transcripts and in vivo migration of CD4+ naive and regulatory T cells. Immunology 115, 179–188 (2005). [PubMed: 15885123]

17. Guo X. et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nature medicine 24, 978–985 (2018).

18. Xiong X, Zhao Y, He H. & Sun Y. Ribosomal protein S27-like and S27 interplay with p53-MDM2 axis as a target, a substrate and a regulator. Oncogene 30, 1798–1811 (2011). [PubMed: 21170087]

19. Palucka KA, Taquet N, Sanchez-Chapuis F. & Gluckman JC Dendritic Cells as the Terminal Stage of Monocyte Differentiation. Journal of Immunology 160, 4587–4595 (1998).

20. Stansfield BK & Ingram DA Clinical significance of monocyte heterogeneity. Clinical and Translational Medicine 4, 5 (2015). URL 10.1186/s40169-014-0040-3. [PubMed: 25852821]

21. Wells CA et al. Alternate transcription of the Toll-like receptor signaling cascade. Genome Biology 7, R10 (2006). URL 10.1186/gb-2006-7-2-r10. [PubMed: 16507160]

22. Villani A-C et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, eaah4573 (2017). URL http://science.sciencemag.org/content/356/6335/eaah4573.abstract.

23. Slyper M, Waldman J, Dionne D. & Li B. Study: ICA: Blood Mononuclear Cells (2 donors, 2 sites). URL https://singlecell.broadinstitute.org/single_cell/study/SCP345/ica-blood-mononuclear-cells-2-donors-2-sites.

24. Guilliams M. et al. Dendritic cells, monocytes and macrophages: A unified nomenclature based on ontogeny (2014).

25. Hutchison LAD, Berger B. & Kohane IS Meta-analysis of Caenorhabditis elegans single-cell developmental data reveals multi-frequency oscillation in gene activation. Bioinformatics (2019). URL 10.1093/bioinformatics/btz864.

26. Freytag V. et al. Genome-wide temporal expression profiling in Caenorhabditis elegans identifies a core gene set related to long-term memory. Journal of Neuroscience 37, 6661–6672 (2017). [PubMed: 28592692]

27. Minkina O. & Hunter CP Intergenerational transmission of gene regulatory information in Caenorhabditis elegans. Trends in Genetics 34, 54–64 (2018). [PubMed: 29103876]

28. Maiden MCJ Multilocus Sequence Typing of Bacteria. Annual Review of Microbiology 60, 561–588 (2006). URL 10.1146/annurev.micro.59.030804.121325.

29. Sudlow C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine 12 (2015).

30. Nicol TL Detecting racial bias in algorithms and machine learning. Journal of Information, Communication and Ethics in Society 16, 252–260 (2018). URL 10.1108/JICES-06-2018-0056.

31. Diaz-Papkovich A, Anderson-Trocme L, Ben-Eghan C. & Gravel S. UMAP reveals cryptić population structure and phenotype heterogeneity in large genomic cohorts. PLOS Genetics 15, 1–24 (2019). URL 10.1371/journal.pgen.1008432.

32. Linderman GC, Rachh M, Hoskins JG, Steinerberger S. & Kluger Y. Fast interpolationbased t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods 16, 243–245 (2019). [PubMed: 30742040]

33. Cho H, Berger B. & Peng J. Generalizable and scalable visualization of single-cell data using neural networks. Cell systems 7, 185–191 (2018). [PubMed: 29936184]

34. Linderman GC, Rachh M, Hoskins JG, Steinerberger S. & Kluger Y. Fast interpolationbased t-SNE for improved visualization of single-cell RNA-seq data. Nature Methods 16, 243–245 (2019). URL 10.1038/s41592-018-0308-4. [PubMed: 30742040]

35. Eades P. A Heuristic for Graph Drawing. Congressus Numerantium 42, 149–160 (1984).

36. Harel D. & Koren Y. A fast multi-scale method for drawing large graphs. In International Symposium on Graph Drawing, 183–196 (Springer, Heidelberg, 2000).

37. Jansen C. et al. Building gene regulatory networks from scatac-seq and scrna-seq using linked self organizing maps. PLoS computational biology 15, e1006555 (2019).

38. Dai H. & Guan Y. The nubeam reference-free approach to analyze metagenomic sequencing reads. Genome Research 30, 1364–1375 (2020). [PubMed: 32883749]

39. Eling N, Richard AC, Richardson S, Marioni JC & Vallejos CA Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. Cell Systems 7, 284–294 (2018). [PubMed: 30172840]

40. Castex GM Frames of Reference: The Effects of Ethnocentric Map Projections on Professional Practice. Social Work 38, 685–93 (1993).

41. Haemer KW Area Bias in Map Presentation. The American Statistician 3, 19 (1949).

42. Kiselev VY, Andrews TS & Hemberg M. Challenges in unsupervised clustering of single-cell rna-seq data. Nature Reviews Genetics 20, 273–282 (2019).

43. Schiebinger G. et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 176, 928–943 (2019). [PubMed: 30712874]

44. Hie B, Bryson B. & Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. Nature Biotechnology 37, 685–691 (2019).

45. Gelman A. et al. Bayesian Data Analysis (CRC press, Boca Raton, 2013), 3 edn.

46. Hotelling H. Relations Between Two Sets of Variates. Biometrika 28, 321–377 (1936). URL http://www.jstor.org/stable/2333955.

47. Andrew G, Arora R, Bilmes J. & Livescu K. Deep Canonical Correlation Analysis. In International Conference on Machine Learning, vol. 28, 1247–1255 (2013).

48. Kobak D, Linderman G, Steinerberger S, Kluger Y. & Berens P. Heavy-tailed kernels reveal a finer cluster structure in t-sne visualisations. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 124–139 (Springer, 2019).

49. Healey CG & Enns JT Building perceptual textures to visualize multidimensional datasets. In Proceedings Visualization ‘98 (Cat. No.98CB36276), 111–118 (1998).

50. Mann HB & Whitney DR On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics 50–60 (1947).

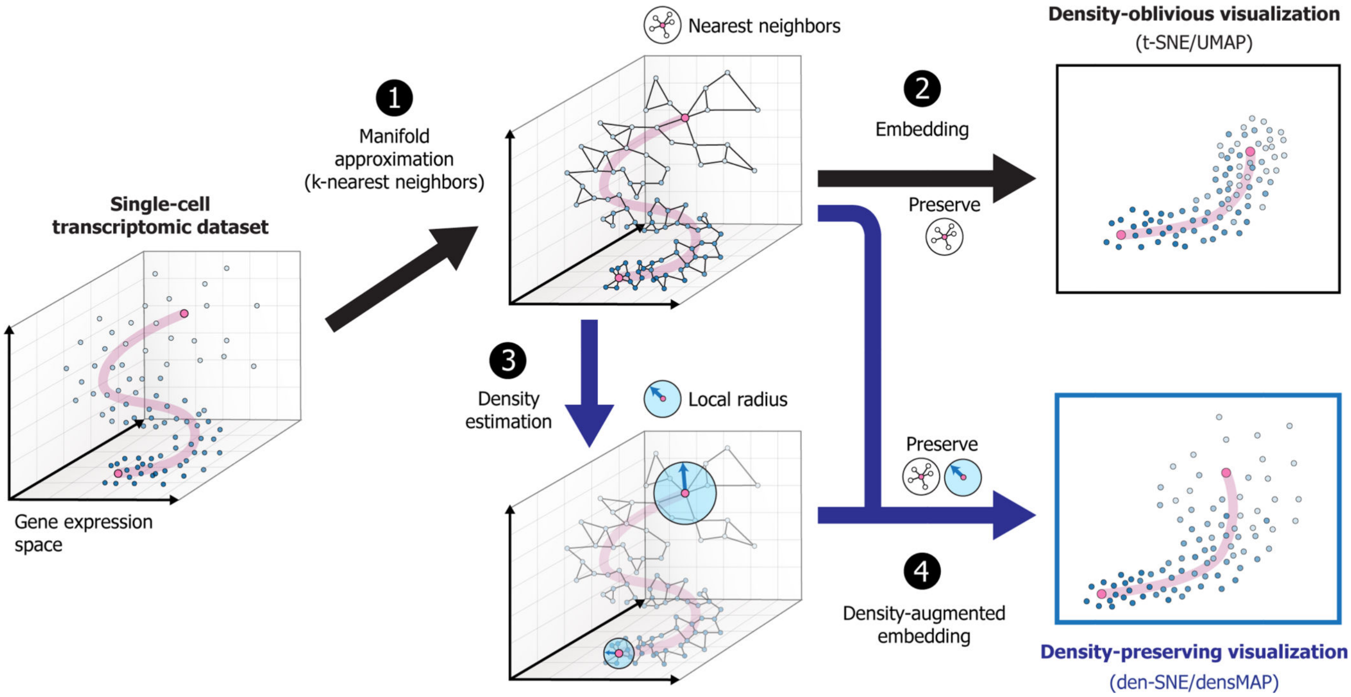

Single-cell transcriptomic dataset
Gene expression space
1
Manifold approximation (k-nearest neighbors)
Nearest neighbors
2
Embedding
Density-oblivious visualization
(t-SNE/UMAP)
Preserve
3
Density estimation
Local radius
Preserve
4
Density-augmented embedding
Density-preserving visualization
(den-SNE/densMAP)

Figure 1: Overview of density-preserving data visualization.

Given a set of points in a high-dimensional space as input (e.g. gene expression profiles from single-cell RNA-seq experiments), the goal of data visualization is to embed these points in 2D or 3D while preserving the structure of the original data. To this end, standard visualization tools t-SNE and UMAP first construct the *k*-nearest neighbor (KNN) graph as a compact summary of the data manifold (**1**). These methods then optimize the visualization coordinates of the points to maximally preserve local distances between neighbors in the graph (**2**). However, because t-SNE and UMAP adaptively choose length-scale to normalize local distances within each neighborhood, they produce visualizations that neglect information about density in the original space, thus omitting a key structural feature of the data. To enhance data visualization by incorporating density information, we introduce a general, differentiable measure of density called the local radius (Methods), which is efficiently calculated on the same KNN graphs that t-SNE and UMAP leverage (**3**). By augmenting the original visualization objective with an additional term that encourages local radii to be consistent between the original space and the visualization, we transform both t-SNE and UMAP into density-preserving counterparts, den-SNE and densMAP, which more accurately portray the structure of the underlying data (**4**).

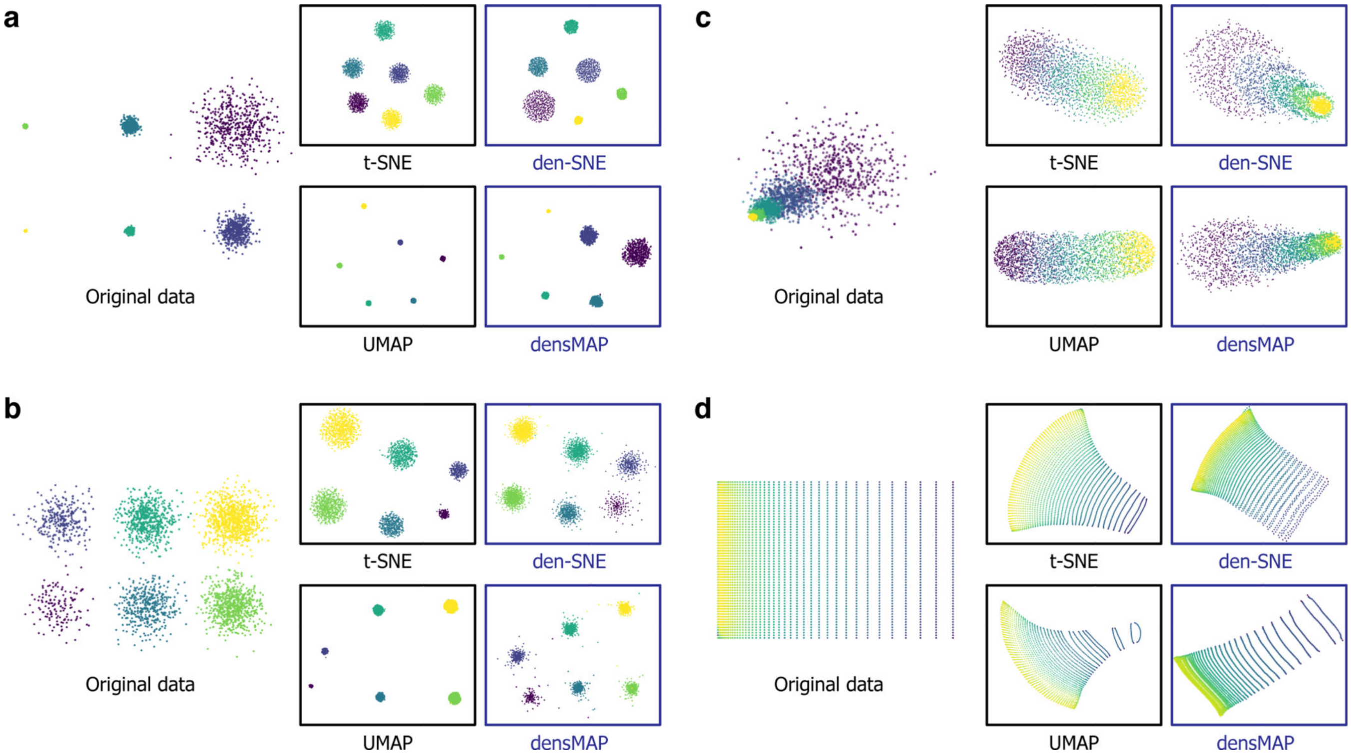


Figure 2: Density-preserving visualization more accurately captures the true underlying shape of synthetic datasets than existing tools.

We compared the visualizations of our density-preserving methods den-SNE and densMAP to those of t-SNE and UMAP on different synthetic datasets: mixture-of-Gaussian point clouds with (**a**) increasing variances with the same sampling rate; (**b**) same variance, but with increasing sampling rates; (**c**) increasing variances in a linear translational motion with overlap, representing a temporal trajectory; and (**d**) a grid of points, whereby the density grows linearly in one direction. The synthetic datasets are generated in twenty dimensions for the point clouds and two dimensions for the grid, and the depictions of the original data in the figure represent two-dimensional linear projections for the former. While t-SNE and UMAP produce misleading visualizations where the apparent size of a cluster of points (marked by different colors) is unrelated to the amount of space it occupies in the original space and is biased by sampling rate, den-SNE and densMAP more accurately portray the shape of the original data by preserving density information.

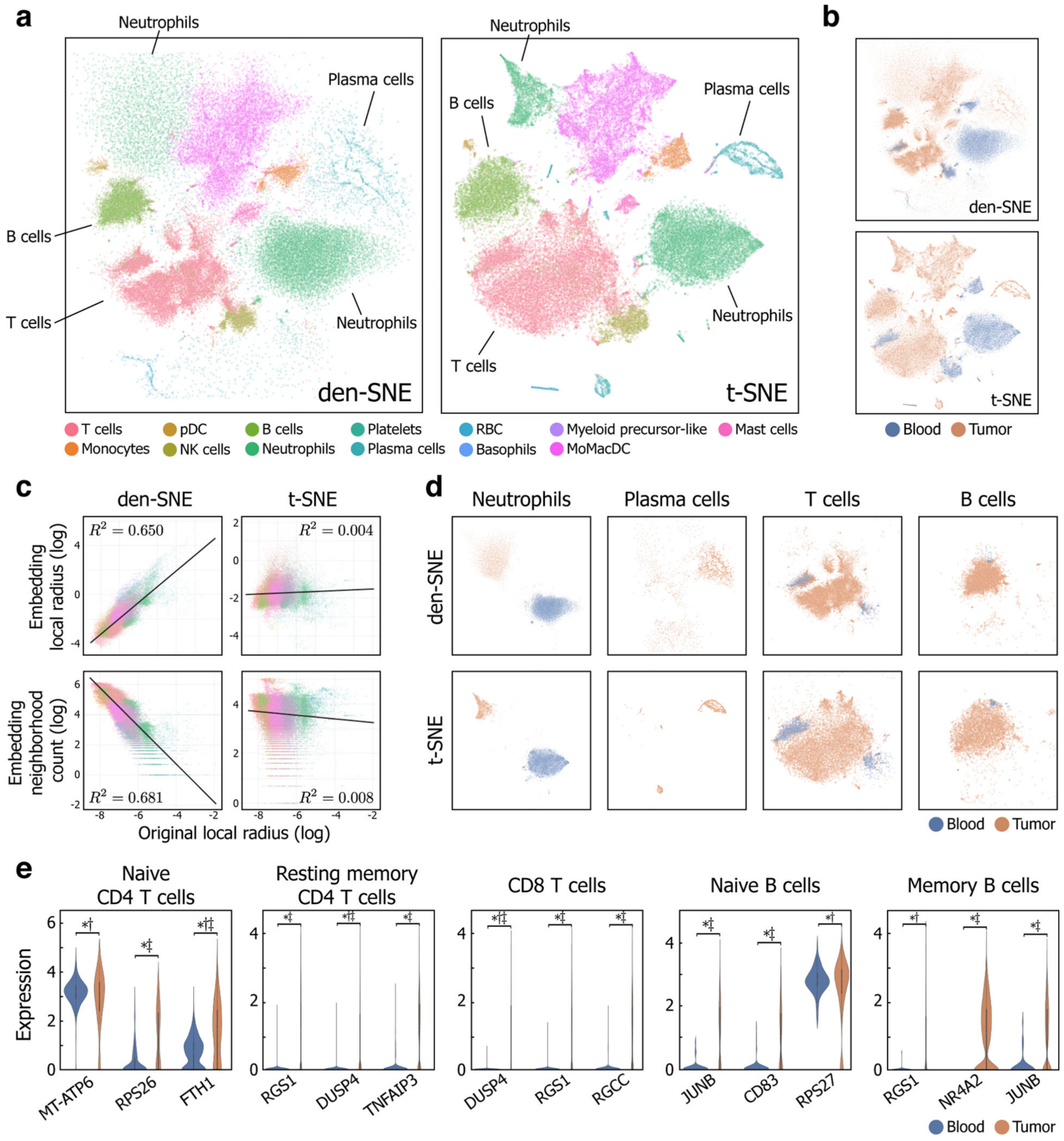


Figure 3: Density-preserving visualization reveals heterogeneity in transcriptomic variability of immune cells in blood and tumor.

We visualized a dataset of tumor and blood immune cells from lung cancer patients[7] using den-SNE and t-SNE, colored by (**a**) cell type and (**b**) tissue type (tumor or blood); den-SNE exposes striking density differences between immune cell types and between blood and tumor, which cannot be discerned from the t-SNE visualization due to its theoretical lack of density-preservation (Methods). Note that the relative heterogeneity of neutrophils, plasma cells, and T cells are misleadingly portrayed in the t-SNE visualization. **c.** Scatter plots comparing the local radii, our measure of local density (Methods), in the original space and two measures of visual density (local radius and neighborhood count; see Methods) in the visualization (embedding) for den-SNE and t-SNE. Points are colored by cell type, and the R^2 value of the correlation is shown for each plot. Higher correlations of den-SNE (inverse correlation for neighborhood count) show that den-SNE more accurately conveys the density landscape of the original data than t-SNE. The radius for neighborhood count is set to two times the average length-scale of each visualization (Methods); other choices of length-scale show a similar improvement for den-SNE (Supplementary Figure 1). **d.** For detailed comparison, we plot the same visualizations for den-SNE (top) and t-SNE (bottom), restricted to each of four notable cell types (neutrophils, plasma cells, T cells, and B cells) and colored by tissue type (tumor or blood). Neutrophils and plasma cells in tumor considerably expand in size in den-SNE, reflecting transcriptomic variability previously hidden in t-SNE. T and B cells show a large increase in heterogeneity in tumor compared to blood in den-SNE. Although t-SNE shows a similar pattern, its lack of a density-preservation property precludes reasoning about differences in heterogeneity. **e.** Violin plots showing the distributions of gene expression in tumor and blood for the top three genes with the highest increase in variance in tumor for each subtype of T and B cells. A more comprehensive list of genes for each cell type is included in Supplementary Tables 1–5. These genes indicate potential biological mechanisms underlying the increased heterogeneity (revealed by den-SNE) of T and B cells in tumor. The markers *, †, and ‡ denote a statistically-significant difference in variance, dispersion, and mean, respectively, between blood and tumor (Bonferroni-corrected $p < 0.01$; Methods). All genes shown have significant variance difference, and several of them are not accompanied by a shift in mean expression (e.g. RPS27 in naive B cells), suggesting biological insights about tumor not captured by conventional differential expression analysis. We provide the same plots for densMAP and UMAP in Supplementary Figure 2.

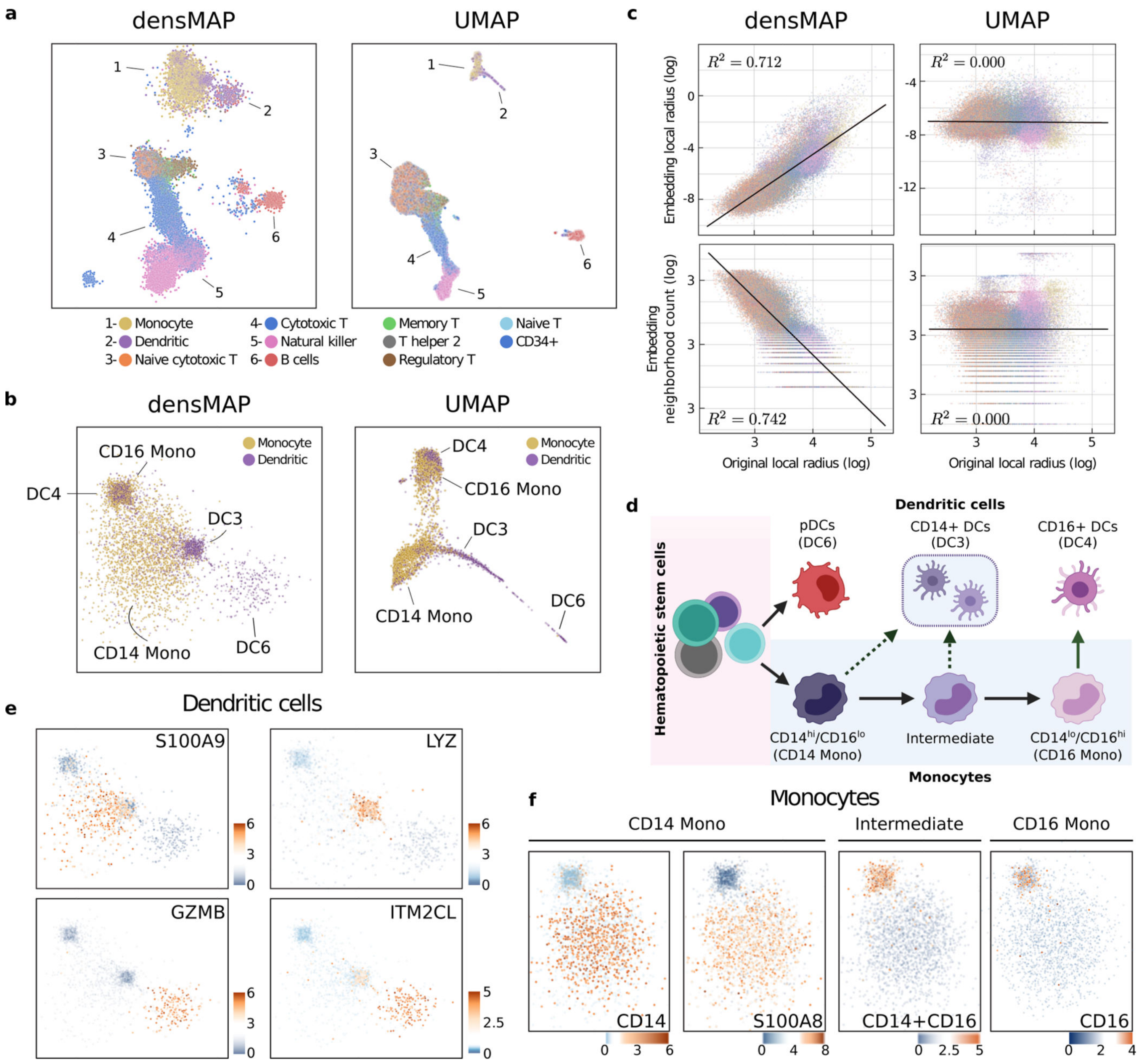


Figure 4: Density-preserving visualization of peripheral blood mononuclear cells reveals monocyte and dendritic cell subsets that differ in transcriptomic variability.

a. We visualized the PBMC dataset[8] using densMAP (left) and UMAP (right), colored by cell type. The group of clusters corresponding to monocytes (cluster 1) and dendritic cells (DCs; cluster 2) showed the most pronounced difference between the two visualizations. **b.** For a detailed comparison, we plotted the same visualizations restricted to the monocyte-DC subset, which revealed distinct subtypes of monocytes (CD16 Mono and CD14 Mono) and DCs (DC3, DC4, and DC6) with clear density differences in densMAP. Each subtype is annotated using the classification from the PBMC2 study[22] based on marker gene expression. Although the same subtypes are visible in UMAP, their relative density differences are lost. **c.** Scatter plots comparing the local radii, our measure of local density

(Methods), in the original space and two measures of visual density (local radius and neighborhood count; see Methods) in the visualization (embedding) for densMAP and UMAP. Points are colored by cell type, and the R^2 value of the correlation is shown for each plot. Higher correlations in densMAP (inverse correlation for neighborhood count) support the validity of the observed density differences between the monocyte and DC subtypes in the densMAP visualization. The radius for neighborhood count is set to the average length-scale of each visualization (Methods); other choices of length-scale show a similar improvement for densMAP (Supplementary Figure 11). **d.** Graphical illustration showing the biological relationships among the five monocyte and DC subtypes we found in the monocyte-DC subset. Under inflammatory conditions, CD14 Mono (classical monocytes) differentiate into CD16 Mono (non-classical monocytes) for immune response. Both CD14 Mono and CD16 Mono can differentiate into DCs (classified as DC3 and DC4, respectively). DC6 represents plasmacytoid DCs (pDCs), which come from a different differentiation trajectory than the rest. densMAP visualization suggests that the differentiation paths from CD14 Mono to CD16 Mono and DC3 both represent specialization with considerable decrease in transcriptomic variability. densMAP also reveals rich variability of DC6 previously hidden in UMAP. **e.** Gene expression heatmaps of DC marker genes from the PBMC2 study[22] for DC3 (top) and DC6 (bottom) in the densMAP visualization restricted to DCs. These support our assignment of DC clusters to DC3 and DC6. A comprehensive set of heatmaps as well as violin plots of all marker genes for DC3, DC4, and DC6 are provided in Supplementary Figure 15. **f.** Gene expression heatmaps of monocyte marker genes CD14, S100A8, and CD16 in the densMAP visualization restricted to monocytes. CD14+CD16 indicate joint expression of the two genes, which is set to their mean if both are expressed, and zero otherwise. The patterns of expression support our classification of the dense cluster as CD16 Mono and the sparse cluster as CD14 Mono. We provide the same plots for den-SNE and t-SNE in Supplementary Figure 12. Validation of the observed variability differences among monocyte and DC subtypes on two additional datasets[22,23] is included in Supplementary Figure 14.

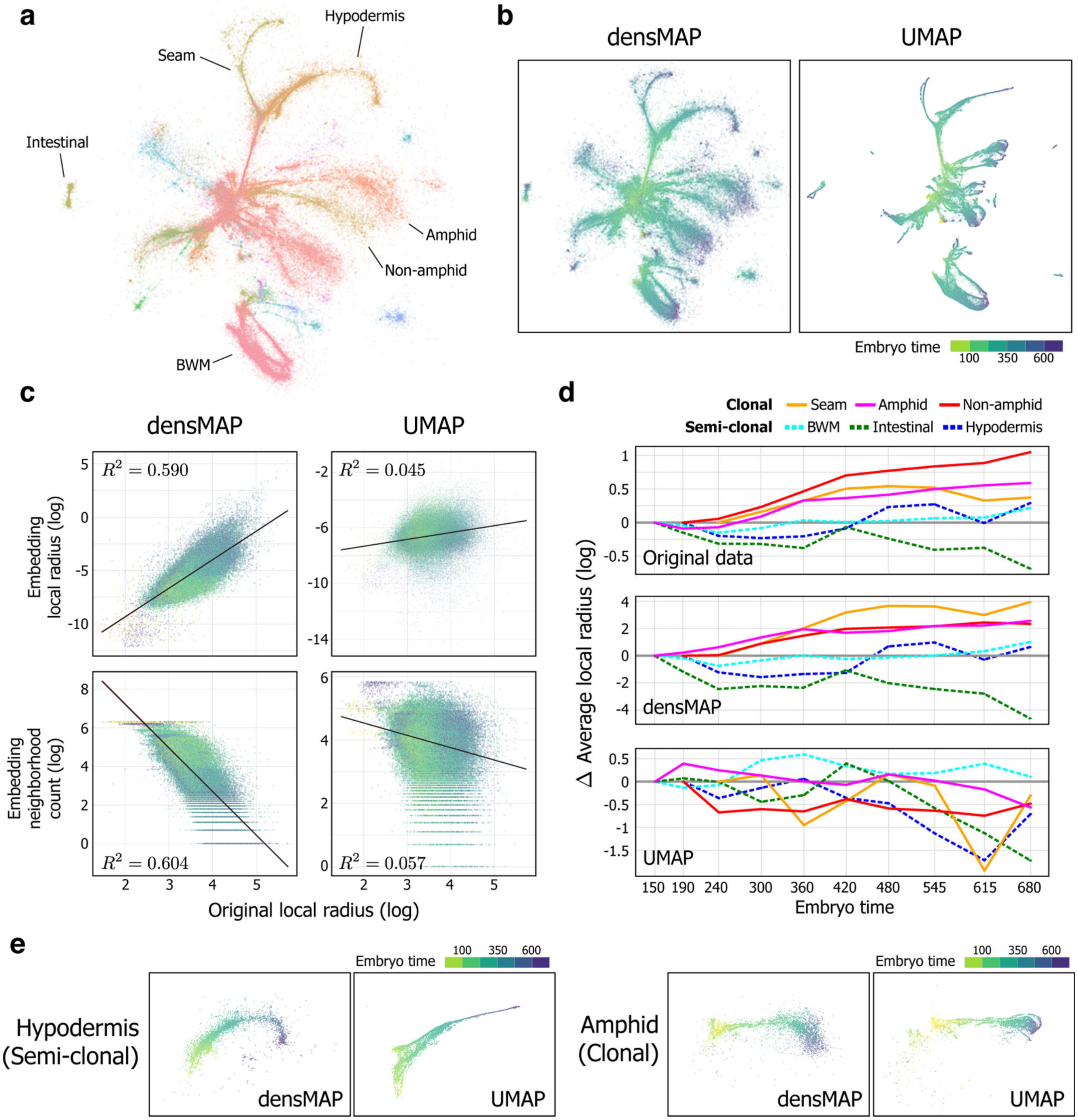


Figure 5: Density-preserving visualization of *C. elegans* development reveals temporal dynamics of transcriptomic variability in different developmental lineages.

We visualized the *C. elegans* dataset[9] using densMAP and UMAP, colored by (**a**) cell type (major cell types labeled) and (**b**) embryo time. UMAP visualization with cell type coloring is omitted for space. In contrast to UMAP, densMAP clearly conveys an overall increase in transcriptomic variability as the organism develops and realizes a wider range of biological functions. **c.** Scatter plots comparing the local radii, our measure of local density (Methods), in the original space and two measures of visual density (local radius and neighborhood count; see Methods) in the visualization (embedding) for densMAP and UMAP. Points are colored by embryo time, and the R^2 value of the correlation is shown for each plot. Higher correlations of densMAP (inverse correlation for neighborhood count) support the validity of the overall increase in transcriptomic variability over the course of development observed in densMAP. The radius for neighborhood count is set to the average length-scale of each visualization (Methods); other choices of length-scale show a similar improvement for densMAP (Supplementary Figure 16). **d.** To assess lineage-specific patterns of transcriptomic variability, we summarized the average local radius of each cell type (marked by different line style) within each embryo time interval for the original data (top), densMAP (middle), and UMAP (bottom). The plot for original data represents the temporal changes in the underlying transcriptomic variability of each cell type, and the plots for densMAP and UMAP represent apparent changes in variability based on the respective visualizations. We used the time intervals provided by the original study, and the y-axis shows the change in average local radius compared to the earliest time interval in log scale. densMAP closely follows the temporal patterns of each cell type in the original dataset, a structural insight that is lost in UMAP. These patterns uniquely captured by densMAP highlight the relatively constant variability of semi-clonal lineages (BWM, intestinal, and hypodermis) in contrast to the increasing variability of clonal lineages (seam, amphid and non-amphid neurons), which can be explained by the more intermixed nature of semi-clonal development. **e.** densMAP and UMAP visualizations restricted to hypodermis and amphid cells for comparison, colored by embryo time. densMAP captures constant variability of hypodermis cells during development, whereas UMAP vastly under-represents the variability of the terminal cell state. Similarly, for amphid cells, densMAP accurately portrays expanding variability, a pattern that is lost in UMAP. We provide the visualizations of other cell types and repeat the analyses for den-SNE and t-SNE in Supplementary Figure 17. BWM: body-wall muscle.