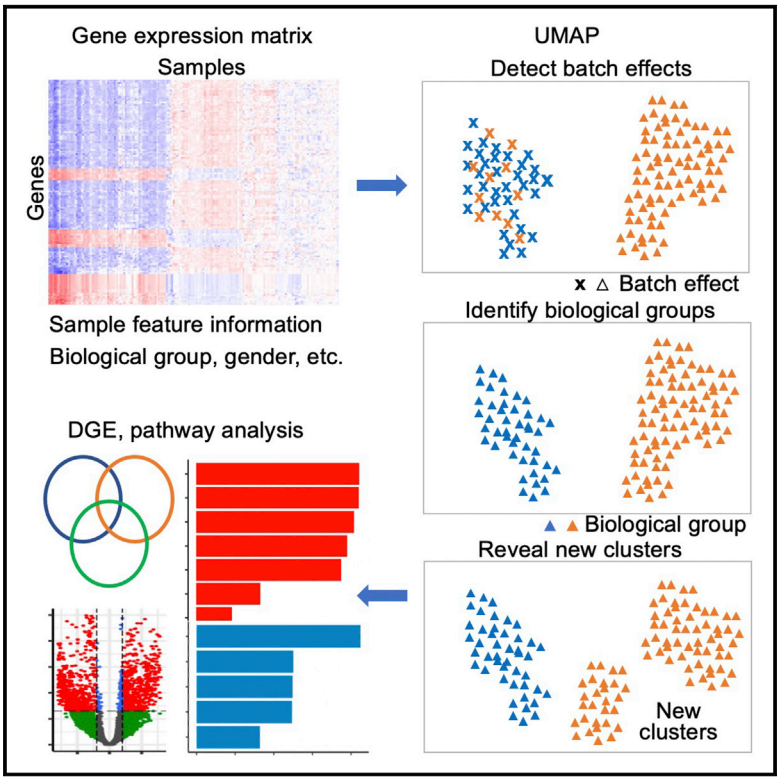
Cell Reports

Resource

Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data

Graphical abstract



Highlights

- Four methods, PCA, MDS, t-SNE, and UMAP, are evaluated on 71 bulk transcriptomic datasets
- UMAP is overall superior to PCA and MDS and shows some advantages over t-SNE
- UMAP can efficiently and effectively reveal clusters in two-dimensional space
- Clusters revealed by UMAP are associated with biological features and clinical traits

Authors

Yang Yang, Hongjian Sun, Yu Zhang, ..., Yuchen Yang, Di Wu, Di Yu

Correspondence

dwu@unc.edu (D.W.),
di.yu@uq.edu.au (D.Y.)

In brief

Yang et al. compare four major dimensionality reduction methods (PCA, MDS, t-SNE, and UMAP) in analyzing large bulk transcriptomic datasets. UMAP is overall superior to PCA and MDS and shows some advantages over t-SNE in differentiating batch effects, identifying pre-defined biological groups, and revealing in-depth clusters in two-dimensional space.

Yang et al., 2021, Cell Reports *36*, 109442
July 27, 2021
https://doi.org/10.1016/j.celrep.2021.109442

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data

Yang Yang,[1,2] Hongjian Sun,[2,3] Yu Zhang,[4] Tiefu Zhang,[5] Jialei Gong,[6] Yunbo Wei,[4] Yong-Gang Duan,[6] Minglei Shu,[2] Yuchen Yang,[7,8] Di Wu,[9,10,11,*] and Di Yu[1,2,4,12,*]

[1]The University of Queensland Diamantina Institute, Faculty of Medicine, The University of Queensland, Translational Research Institute, Brisbane, QLD, Australia
[2]Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[3]School of Microelectronics, Shandong University, Jinan, China
[4]Laboratory of Immunology for Environment and Health, School of Pharmaceutical Sciences, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[5]University of Electronic Science and Technology of China, Chengdu, China
[6]Shenzhen Key Laboratory of Fertility Regulation, Center of Assisted Reproduction and Embryology, University of Hong Kong, Shenzhen Hospital, Shenzhen, China
[7]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[8]McAllister Heart Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[9]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[10]Division of Oral and Craniofacial Health Science, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[11]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA
[12]Lead contact
*Correspondence: dwu@unc.edu (D.W.), di.yu@uq.edu.au (D.Y.)
https://doi.org/10.1016/j.celrep.2021.109442

SUMMARY

Transcriptomic analysis plays a key role in biomedical research. Linear dimensionality reduction methods, especially principal-component analysis (PCA), are widely used in detecting sample-to-sample heterogeneity, while recently developed non-linear methods, such as t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), can efficiently cluster heterogeneous samples in single-cell RNA sequencing analysis. Yet, the application of t-SNE and UMAP in bulk transcriptomic analysis and comparison with conventional methods have not been achieved. We compare four major dimensionality reduction methods (PCA, multidimensional scaling [MDS], t-SNE, and UMAP) in analyzing 71 large bulk transcriptomic datasets. UMAP is superior to PCA and MDS but shows some advantages over t-SNE in differentiating batch effects, identifying pre-defined biological groups, and revealing in-depth clusters in two-dimensional space. Importantly, UMAP generates sample clusters uncovering biological features and clinical meaning. We recommend deploying UMAP in visualizing and analyzing sizable bulk transcriptomic datasets to reinforce sample heterogeneity analysis.

INTRODUCTION

Bulk transcriptomic profiling quantifies the transcripts in a given biological sample, achieved by technologies including microarray (Heller, 2002; Lim et al., 2009) and RNA sequencing (RNA-seq) (Wang et al., 2009; Wu et al., 2013). This tool is ubiquitously adopted in modern biomedical research and application to reveal features of gene expression for specific cell or tissue type and biological process. The principal task of bulk transcriptomic profiling is to analyze differential gene expression (DGE) of samples between biological groups. When statistically modeling DGE, an implicit assumption is that data of individual samples within a given group are relatively homogeneous. For instance, to investigate the biomarker for a certain disease, the group comparison between patient and healthy control cohorts presumes that the biological characteristics of individual patients are largely indistinguishable when compared with healthy controls and vice versa. However, there exists heterogeneity within a group, which can lie in samples' distinct biological states. For example, patients with systemic lupus erythematosus (SLE) show distinct disease activities and can be classified based on the levels of disease activity index (Bombardier et al., 1992). Other heterogeneity can result from different sample preparation or processing conditions, often referred to as batch effects (Leek et al., 2010; University of Texas MD Anderson Cancer Center, 2020). Therefore, it is crucial to scrutinize sample-to-sample heterogeneity within groups so that subgroups or outliers can be identified. Only with such information can appropriate analytic methods be used to correct

batch effects, remove outliers, and distinguish subgroups. In contrast, DGE analysis simply in given groups without the knowledge of sample-to-sample heterogeneity within groups can often lead to biased or even wrong conclusions.

To detect among-sample heterogeneity in bulk transcriptomic profiling, individual samples are usually visualized in two-dimensional embedded space by dimensionality reduction methods; we adopted the convention to analyze the transcriptomic datasets in two-dimensional space. Principal-component analysis (PCA) (Wold et al., 1987) and multidimensional scaling (MDS) (Torgerson, 1952) have been thoroughly exploited to obtain an overview of sample relationship in a low-dimensional space (Law et al., 2016; Love et al., 2014; Ritchie et al., 2015; Robinson et al., 2010). Both methods succeeded in visualizing biological or technical variation among samples by uncovering the overall structure of the sample-to-sample relationship, which represents the key information of among-sample heterogeneity.

Since 2009 (Tang et al., 2009), the era of characterizing transcriptome at single-cell level has arrived. Numerous single-cell RNA-seq (scRNA-seq) technologies enable simultaneous profiling of thousands of cells' transcriptomes in a given sample so that the analysis of population heterogeneity can identify complex compositions, reveal rare cell populations, detect differentially expressed genes between multiple cell populations or between samples for cell types, uncover cell differentiation trajectories, and so forth (Hwang et al., 2018; Van Buren et al., 2021). However, PCA and MDS show inefficient performance for dimensionality reduction of scRNA-seq data, while two non-linear methods, *t*-distributed stochastic neighbor embedding (t-SNE) (Kobak and Berens, 2019; Maaten and Hinton, 2008) and uniform manifold approximation and projection (UMAP) (Becht et al., 2018; McInnes et al., 2018), exhibit better capability because of the advantage of maintaining cell-to-cell neighbor information and visualizing local structure. Studies suggested that UMAP is better than t-SNE to retain the global structure in scRNA-seq data analysis because of Laplacian Eigenmaps initialization and cross-entropy object function of the former (Becht et al., 2018; McInnes et al., 2018). Nevertheless, by tuning parameters, especially initialization, t-SNE was shown to achieve comparable performance (Kobak and Berens, 2019; Kobak and Linderman, 2021).

The continuous improvement and invention of sequencing platforms has hugely improved the efficiency and throughput of DNA sequencing and resulted in a dramatic reduction in costs, which enable generation of a large number of samples and datasets of bulk transcriptomic profiling. For example, the landmark cancer genomics program, The Cancer Genome Atlas (TCGA), has profiled and integrated over 20,000 primary cancer and matched normal samples spanning 33 cancer types and generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data (GTex Consortium, 2015; Lou et al., 2021; Tomczak et al., 2015). Although PCA remains as the mainstream tool recommended for detecting among-sample heterogeneity in bulk transcriptomic profiling, such as by TCGA Batch Effects (Leek et al., 2010; University of Texas MD Anderson Cancer Center, 2020), we hypothesize that, for datasets with large sample sizes, local structure of sample-to-sample relationship becomes more prominent for sample heterogeneity analysis. Therefore, non-linear methods t-SNE and UMAP might outperform PCA and MDS.

In this study, we visually and quantitatively compared the capabilities of PCA, MDS, t-SNE, and UMAP in heterogeneity exploration of bulk transcriptomic profiling, when parameters were set as default. By visualizing and interpreting 71 sizable datasets of bulk transcriptomic profiling in two-dimensional space, we found that UMAP was superior in preserving sample-level neighborhood information and maintaining clustering accuracy, thus conspicuously differentiating batch effects, identifying pre-defined biological groups, and identifying clustering structures associated with biological features and clinical meaning.

RESULTS

Overview of the evaluation

The bulk transcriptomic profiling datasets were collected from the Gene Expression Omnibus (GEO) database within the past 5 years (available online 10.17632/nddfs37dkh.2). To minimize the cell-type effects interacting with our results, which are usually strong and very easy to be identified, we chose only the datasets of human samples from peripheral blood mononuclear cells (PBMCs) or whole blood for bulk transcriptomic analysis, which are among the most frequent cell populations. Datasets with a size of fewer than 100 samples were excluded in order to generate observable and meaningful clusters. The collection covered a diverse range of biomedical research, including the investigations on disease features such as SLE (Banchereau et al., 2016; Davenport et al., 2018; Figgett et al., 2019; Guthridge et al., 2020; Hong et al., 2019; Oon et al., 2019; Toro-Domínguez et al., 2018) and influenza infection (Dunning et al., 2018; Hoang et al., 2014; Zhai et al., 2015), and the evaluation of interventions such as therapies and vaccination (Le-Niculescu et al., 2020; Narang et al., 2018; Rawat et al., 2020; Tasaki et al., 2018; Thakar et al., 2015).

The research design flowchart is shown in Figure 1A. All two-dimensional visualizations generated by four methods from 71 datasets were independently assessed by three assessors without the knowledge of datasets to report clustering structures. Clustering structures in two-dimensional visualization were manually identified in 41 datasets (details in STAR Methods), with the highest number of clustering structures by UMAP (n = 41) and the lowest number by PCA (n = 11). To validate the objectivity and accuracy of manual assessments, we applied the density-based clustering method hdbscan (McInnes et al., 2017) to the two-dimensional embedding coordinates and calculated the silhouette score (Rousseeuw, 1987) (details in STAR Methods) to measure the clustering level of data points (Data S1C available online 10.17632/gtfrr4j5cx.3). For each dimensionality reduction method, the silhouette scores of embedding coordinates with manually reported clustering structures were significantly higher than those without manually reported clustering structures (Data S1A available online 10.17632/gtfrr4j5cx.3). The silhouette scores for all two-dimensional embedding coordinates appeared as bimodal distribution, so two normal distributions were extracted by fitting a finite mixture model. More than 90% of the visualizations without manually reported clustering structures were within the range

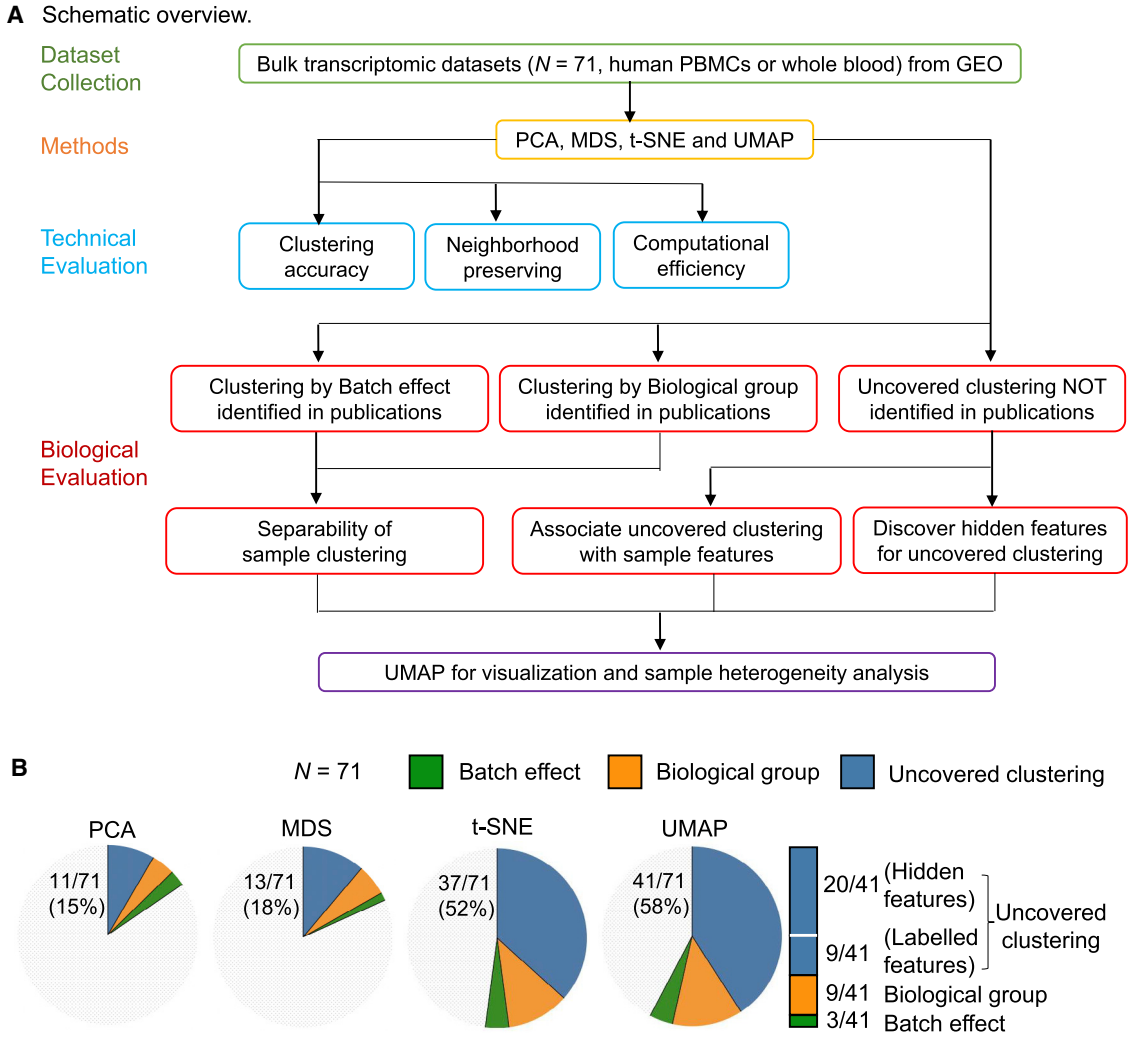


Figure 1. Evaluation overview for four dimensionality reduction methods

(A) Schematic overview of the evaluation. Bulk transcriptomic datasets (n = 71) were collected from the GEO database, followed by applying four methods to the datasets for visualization. The methods were evaluated in both technical and biological aspects. Finally, we presented the recommendation on UMAP for visualization.

(B) Pie chart showing the percentage of datasets by biological explanations for all revealed clustering structures. By associating features identified in publications, clustering structures were divided into three categories: batch effect (colored green), biological group (colored orange), and uncovered clustering (colored blue). Batch effect was the cluster associated with batch effects. Biological group was related to experimental design like control and treatment groups, while revealed clustering was the clusters related to other pre-defined features such as gender. Uncovered clusterings were further divided into clusters with sample features and clusters with hidden features by considering available feature information.

of mean ± 2 × SD in the normal distribution for low silhouette scores, and more than 85% of the visualizations with manually reported clustering structures were within the range of mean ± 2 × SD in the normal distribution for high silhouette scores (Data S1B available online 10.17632/gtfrr4j5cx.3). These results validated the manual assessments to report clustering structures.

Among the 41 datasets demonstrating clustering structures, UMAP reported all clustering (41/71) and, together with t-SNE (37/71), performed significantly better than PCA (11/71) and MDS (13/71) (Figure 1B). The datasets showing clustering structures by t-SNE, MDS, and PCA are a subset of those by UMAP.

The 41 datasets were classified into three categories by incorporating available features from metadata (details in STAR Methods) (Figure 1B; Data S1D available online 10.17632/gtfrr4j5cx.3). As in Figure 1B, three plots in the two-dimensional embedding space from the dimension reduction methods showed clusters related to batches (batch effect) described in studies for these datasets, while 9 plots showed clusters related to biological groups designated by study designs. In addition, 29 plots revealed clustering not related to batch information or biological group by study design, suggesting significant sample-to-sample heterogeneity in bulk transcriptomic analysis. We identified the relationship of clustering structures with known

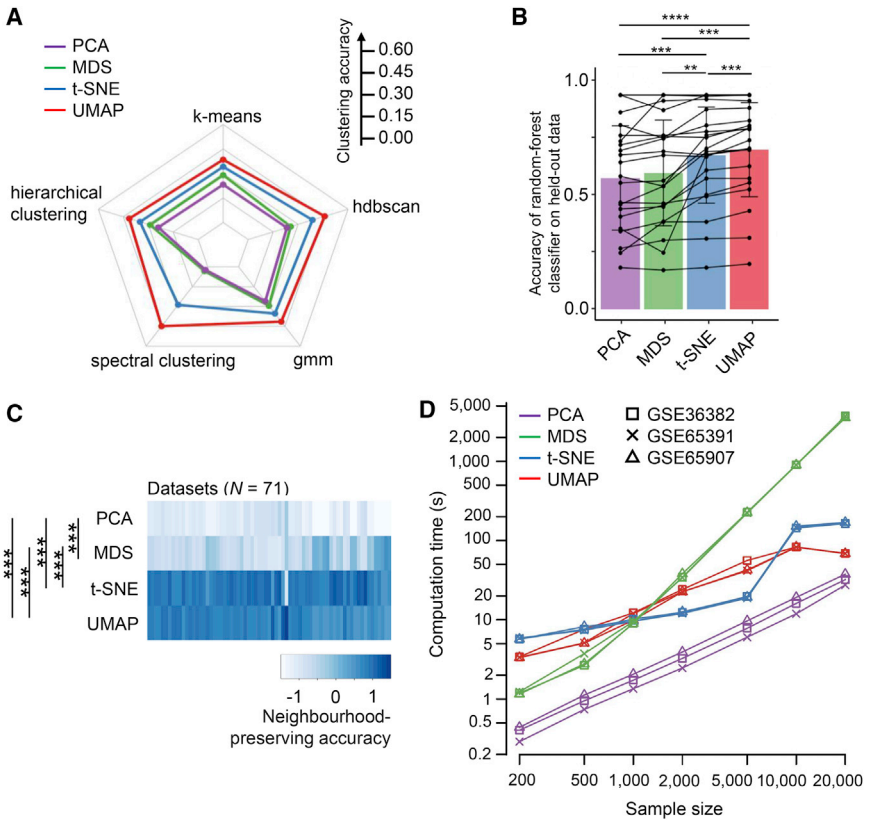


Figure 2. Quantitative analysis of four dimensionality reduction methods

(A) Radar plot of clustering accuracy (average NMI score) comparison using five clustering methods on 21 datasets with cluster labels. The input was the embedded two-dimensional coordinates of each dimensionality reduction method. A larger scale denotes better clustering accuracy.

(B) Classification accuracy on held-out data of random forest classifiers predicting cluster labels taking embedded coordinates as input. 21 datasets showing clustering with available features are included. The average score across datasets is shown, with vertical bars representing SD; paired t test was conducted on pairwise methods (*p < 0.05, **p < 0.01, ***p < 0.001).

(C) Heatmap for evaluating neighborhood preservation of each method on 71 datasets. The number of neighbors is set as 15. The darker the color is, the better the local information is retained. One-way ANOVA shows a significant difference among the four methods (***p < 0.001). R function *heatmap* in R package *stats* was used for (B).

(D) Running time evaluation of dimensionality reduction methods with varying sample sizes. Different sizes of data were generated by sampling with replacement from the three largest datasets respectively. t-SNE was implemented by openTSNE with parameter *neighbors* = 'exact'.

sample features for nine plots. The clustering structures of the rest of the 20 plots could result from hidden batch effect or biological features not reported by publications, thus referred to as clustering with hidden features (Figure 1B).

With clustering structures generated by PCA, MDS, t-SNE, and UMAP, we could evaluate individual methods' performance for clustering accuracy, local information preservation, and computational efficiency. For datasets with clustering structure by batch effect or biological group, we would then compare the separability of each method in detecting distinct groups. For revealed clustering structures, we would investigate the relationships of clustering structures with sample features. Based on these quantitative and qualitative assessments, we could provide the recommendation of the best performing method for dimensionality reduction in sizable bulk transcriptomic analysis (Figure 1A).

Comparison of dimensionality reduction methods by quantitative performance

Clustering accuracy

The foremost objective of dimensionality reduction for bulk transcriptomic analysis is to conspicuously distinguish clustering structures of samples that associate biological meaning. We applied five clustering algorithms—k-means (Lloyd, 1982; MacQueen, 1967), hierarchical clustering (Murtagh and Legendre, 2014), spectral clustering (Von Luxburg, 2007), Gaussian mixture model (Reynolds, 2009), and hdbscan (McInnes et al., 2017), with details of the five algorithms in the STAR Methods—to low-dimensional spaces projected by dimensionality reduction methods, and we compared the clustering accuracy.

The five clustering algorithms were performed on the embedding two-dimensional coordinates of 21 datasets that have available label information for groups (labeled in Table S1 available online 10.17632/nddfs37dkh.2). To assess clustering accuracy of dimensionality reduction methods, we then computed Normalized Mutual Information (NMI) (Danon et al., 2005) and Adjusted Rand Index (ARI) (Rand, 1971) for comparing the true group labels and inferred group labels obtained by clustering algorithms based on the low-dimensional components, and the lager score indicates better clustering accuracy. UMAP was scored the highest for both NMI and ARI, no matter what clustering algorithm used, achieving the best accuracy for clustering (Figures 2A and S1A– S1C); t-SNE was scored slightly lower than UMAP but well outperformed MDS and PCA.

Separating features

To evaluate the capability of each dimensionality reduction method in separating the features by the embeddings, we deployed random forests to train embedding data with features as labels and computed the prediction accuracy on held-out data (details in STAR Methods). All 21 datasets showing clustering structures with features associated were included. Paired t tests were performed between every two methods (Figures 2B, S2B, and S2C). PCA showed the lowest score (average, 0.61 ± 0.244) in separating features, followed by MDS (average, 0.64 ± 0.247). UMAP achieved the best performance (average, 0.75 ± 0.220), outperforming MDS ($p < 0.001$) and PCA ($p < 0.0001$);

UMAP was better than t-SNE (average, 0.72 ± 0.224; $p < 0.001$), although the difference was not considerable (Figure 2B).

Neighborhood preserving

We then evaluated the performance of different dimensionality reduction methods in retaining local information from original datasets, which was assessed by comparing the fidelity of local neighborhood structures between the reduced low-dimensional space and the original space using a Jaccard index (details in STAR Methods) (Levandowsky and Winter, 1971). The Jaccard indexes were computed for 15 (Figure 2C) and 30 neighbors (Figure S1D), respectively. PCA exhibited the worst performance in preserving neighborhood information (average, 0.19 ± 0.067), followed by MDS (average, 0.26 ± 0.114). The performance of UMAP (average, 0.35 ± 0.091) appeared comparable with that of t-SNE (average, 0.36 ± 0.095), and both were better than PCA and MDS. Paired t tests were performed between every two methods (Figure 2C), and statistically significant differences were detected between group means by one-way ANOVA ($F(3, 280) = 57.88$; $p < 0.001$). This was conceivable because UMAP and t-SNE are designed to utilize local information for dimensionality reduction.

Computational efficiency

We next measured the execution time of each dimensionality reduction method on data with sample sizes ranging from 200 to 20,000. The computer setting details are in the STAR Methods. The varied scales of data were generated by randomly sampling with replacement from the three largest datasets (GEO: GSE36382, GSE65391, and GSE65907). We utilized the heavily optimized Barnes-Hut and Fourier-interpolated t-SNE by openTSNE (Poličar et al., 2019) implementation. Notably, in the preprocessing stage, the scRNA-seq analyses have included PCA to first reduce the dimensionality before applying t-SNE or UMAP (Kobak and Berens, 2019; Luecken and Theis, 2019), and we adopted this pre-processing strategy for t-SNE and UMAP in running time comparison. All methods were set as default parameter values except for *n_jobs* = −1 (for MDS, t-SNE, and UMAP) and *neighbors* = 'exact' (for t-SNE). The detailed parameters settings were listed in STAR Methods. Consumed time was negligibly affected by different datasets (Figure 2C). PCA performed consistently faster than the other three methods, while MDS ran slowest. From 200 to 1,000 samples, consumed time was similar between t-SNE and UMAP; for 2,000 and 5,000 sample sizes, t-SNE performs better than UMAP, but UMAP gained an advantage for data with sample size larger than 10,000. PCA, t-SNE, and UMAP were more time efficient than MDS, in particular for sample sizes over 5,000 (Figure 2D).

Technically, UMAP not only identified more clustering structure in 71 datasets of bulk transcriptomic analysis (Figure 1B) but was also superior to the other three methods for the overall performance by assessing the four quantitative criteria. We next compared four dimensionality reduction methods for uncovering biological meaning.

Comparison of dimensionality reduction methods by associating feature information

Identification of batch effects

Batch effects are common in many types of high-throughput sequencing experiments, which are systematic technical variations introduced by processing samples in different batches (Leek et al., 2010, 2012). As for high-throughput sequencing experiments, it is essential to remove unwanted variations in the transcriptomic analysis by normalization (Gerstner et al., 2016; Risso et al., 2014) to avoid biased analysis and distorted results (Leek et al., 2010). The first step is to identify batch effects among samples. PCA is the most used tool, such as by TCGA project (Tomczak et al., 2015). It generates the clustering structure of samples in a two-dimensional embedding space to facilitate the visualization of batch information. Among the 41 datasets with explicit clustering structures, 3 datasets showed clustering structures related to batch effects reported by publications (Figure 1B). Each dimensionality reduction method was used to visualize batch effects for the three datasets (one in Figure 3A and two in Figure S2A). UMAP and t-SNE showed better segregation between samples from different batches.

Validation of biological groups by experimental design

One major purpose of bulk transcriptomic analysis is for the DGE analysis between biological groups defined by experimental design. Visualizing the segregation of samples from groups with distinct biological features by dimensionality reduction is often applied to the validation of group-to-group distinction. Among the 41 datasets with explicit clustering structures, 9 datasets showed clustering structures related to biological groups by experimental designs (Figure 1B). We compared four dimensionality reduction methods in visualizing biological groups and found that UMAP and t-SNE outperformed MDS or PCA in visually separating biological groups in nine datasets (one in Figure 3B and eight in Figure S3).

Uncovering associations between clustering structures and sample features

Only 12 of 41 datasets showed clustering structures explained by batch effects or biological groups (Figure 1B). The appearance of uncovered clustering structures in 29 plots demonstrated significant heterogeneity existing in bulk transcriptomic profiling, which could be efficiently revealed by UMAP. We next investigated the causes underlying uncovered clustering structures. The clustering structures in nine datasets were found to be associated with certain sample features reported by publications (Figure S4). These features were not used for the classification of sample groups in experimental designs, suggesting certain biological features with major impacts on sample heterogeneity were not included in experimental designs or data analyses. A good case was the dataset GEO: GSE71220, which was designed to determine the impact of cigarette smoking (former versus current smoker) on gene expression in peripheral blood of patients with chronic obstructive pulmonary diseases (COPDs) (Obeidat et al., 2016). Dimensionality reduction methods of UMAP and t-SNE generated plots showing clustering structures (right part in Figure 4A). However, such clustering was not associated with smoking status (Figure 4B). We applied other sample features, including age and disease status, to the two-dimensional plots. Surprisingly, the sample feature of gender demonstrated a clear association with clusters in the plots generated by UMAP and, to a lesser extent, t-SNE (Figure 4C). In the UMAP plot, one cluster was highly enriched by females (in orange), and another cluster was highly enriched by males (in blue), with the third cluster showing the pattern of a

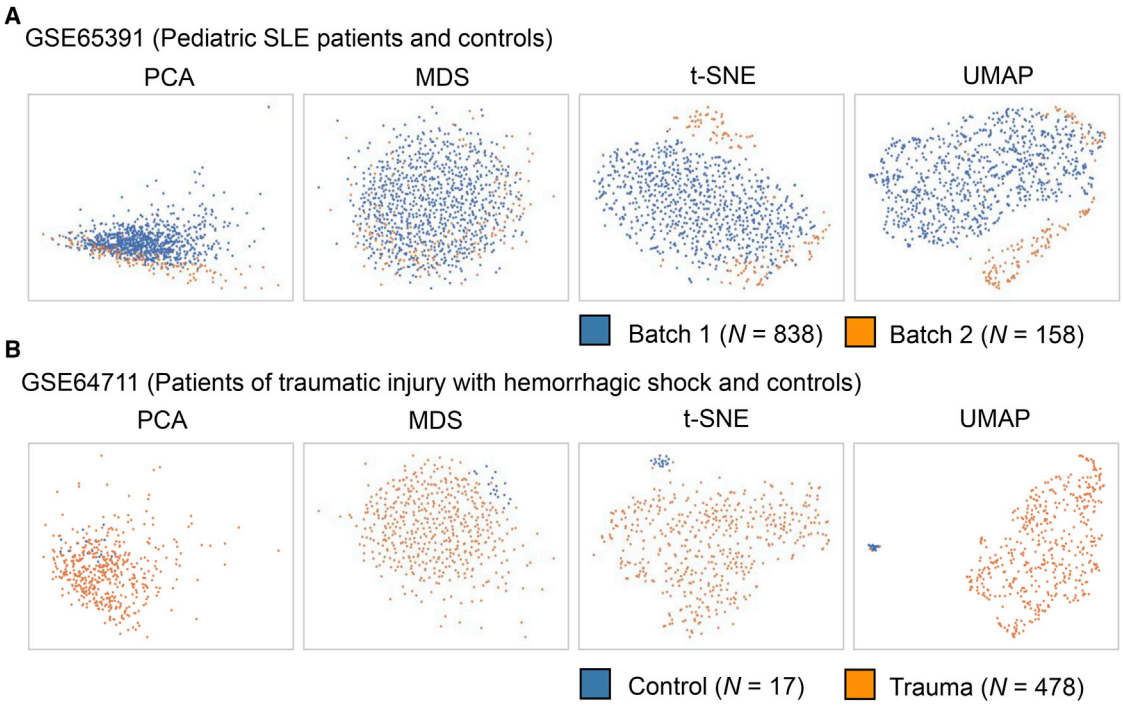


Figure 3. Biological explanation of clustering by batch effects and biological group

(A) Visualization of dataset GEO: GSE65391 showing the batch effects (colored by blue and orange) in two-dimensional space by dimensionality reduction methods.

(B) Visualization of dataset GEO: GSE55447 illustrating biological group by dimensionality reduction methods. The control group is labeled in blue, and the trauma group is in orange.

mixture (Figure 4C). By deploying k-means clustering (details in STAR Methods), samples were divided into three clusters with distinct gender composition: C1 %, 80% females; C2 %, 90% males; and C3, mixed (Figures 4D and 4E). This indicated that the transcriptomes of samples in this study were highly influenced by gender difference. Indeed, the heatmap of the top 100 differentially expressed genes demonstrated that the clustering of samples was strongly associated with gender (Figure 4F). Therefore, the heterogeneity uncovered by the dimensionality reduction using UMAP indicated that the gender difference should have been critically treated as a latent variable in the downstream transcriptomic analysis.

Discovering associations between clustering structures and hidden features

By dimensionality reduction using UMAP, 41 datasets showed clustering structures in two-dimensional embedding spaces in which the associations with batch effects, biological groups by experimental designs, or specific sample features reported by publications were identified in 21 datasets (Figure 1B). For the rest of the 20 datasets, clustering structures might derive from obscure heterogeneity of samples, biologically or technically (Data S1E available online 10.17632/gtfrr4j5cx.3). We made efforts to explore the biological meanings of clustering structures of these datasets and herein present the dataset GEO: GSE121239 as an example to support the notion that clustering structures generated by UMAP can reinforce sample heterogeneity analysis of bulk transcriptomic data to reveal important biological meaning.

Dataset GEO: GSE121239 originated from the study of SLE, which is the prototype of systemic autoimmune diseases with highly diverse manifestations in multiple tissues and organs, such as skin, kidney, and lung (National Institutes of Health, 2020). As a chronic disease, SLE patients often experience the unpredictable occurrence of disease flares (Buyon et al., 2005). In order to identify the heterogeneity of SLE patients and stratify patient groups of disease activity progression, the dataset GEO: GSE121239 collected longitudinal transcriptome profiles of 65 SLE patients with more than three clinical visits and 20 healthy individuals as controls (Toro-Domínguez et al., 2018). Data collected at each visit contributed to one sample in the dataset. Dimensionality reductions plot by UMAP and t-SNE, but not PCA or MDS, clearly demonstrated separated clusters for SLE patients (in orange) and healthy controls (in blue) (Figures 5A and 5B). In the UMAP plot, we noticed more than one cluster for patient samples (Figure 5C). To understand the biological meaning of clusters representing subgroups of SLE patients, we examined feature information of patients reported by the publication, including gender and patient ID, but found no direct association with the clustering structure of patient subgroups. Because the samples of patients were collected longitudinally from multiple clinical visits, we set samples collected at the first clinical visiting date as day 0, then labeled subsequently collected samples with the period from the first visiting date. The resulting contour plot showed samples in chronological order (Figure 5D). Importantly, the gradient from light to dark orange spreads from the middle of the plot to two sides,

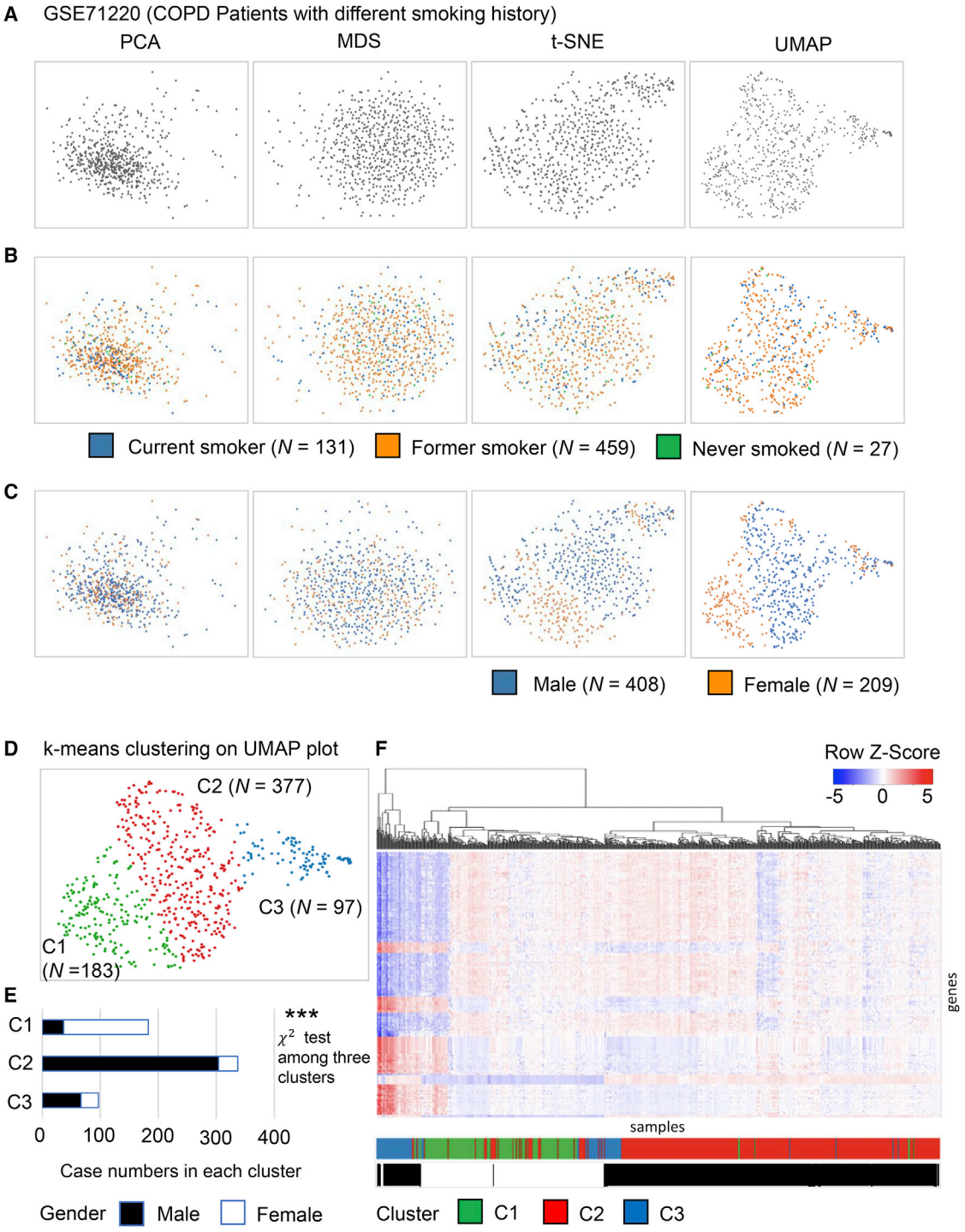


Figure 4. Uncovered clustering interpreted by available sample features

(A–C) Visualization of dataset GEO: GSE71220 in two-dimensional space by assigning no feature (A), group labels (B), and gender (C).

(D) k-Means clustering on two-dimensional embedded coordinates into three clusters: C1, C2, and C3.

(E) Gender proportion among three clusters by χ^2 test showing a significant difference (***$p < 0.001$). Male and female are colored by black and white, respectively.

(F) Heatmap of top 100 differentially expressed genes with three clusters, C1, C2, and C3, and two gender groups, male and female. R function heatmap.2 in R package gplots was used.

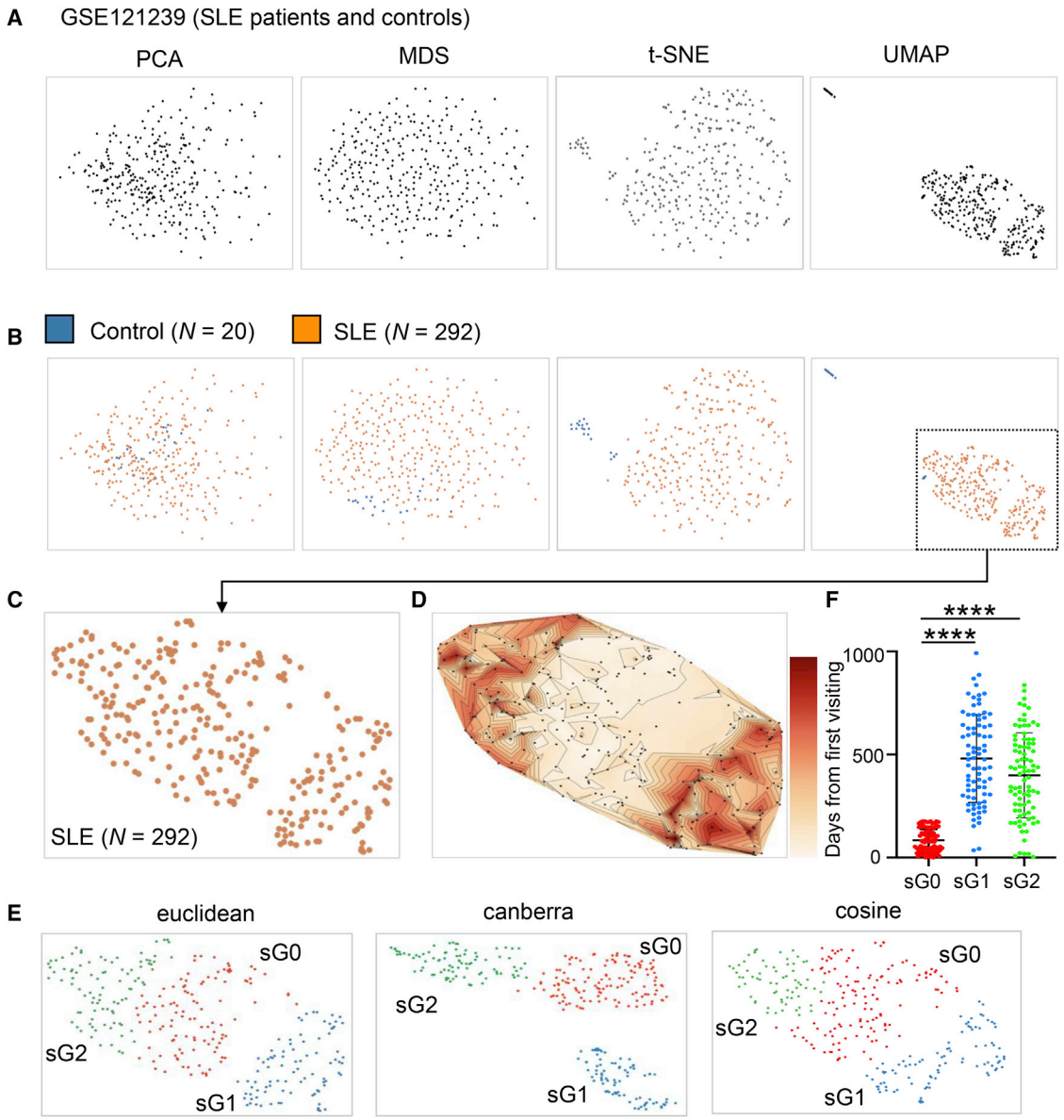


Figure 5. Discovering associations between clustering structures and hidden features

(A and B) Visualization of dataset GEO: GSE121239 in two-dimensional space by assigning no feature (A) and group labels (B).

(C) Patient (SLE) group (colored orange) showing clustering structure (sG1, lower right).

(D) Contour plot on patient groups by the order of visiting timestamp. Each data point is associated with one visiting timestamp. Data points are colored by the visiting time with light colors for early visits and dark color for late visits. The code to plot is in Data and code availability.

(E) Hierarchical clustering of patient group on two-dimensional embedded coordinates by UMAP with distance metrics as 'euclidean', 'canberra', and 'cosine', respectively. The patient group is divided into three subgroups: sG0, sG1, and sG2.

(F) Scatterplot of visiting days in three subgroups. The average visiting day is shown, with vertical bars representing SD; one-way ANOVA was conducted on pairwise subgroups (****$p < 0.0001$).

indicating the clustering structure generated by UMAP was associated with the timing evolution of clinical visits. For example, the bottom right cluster in Figure 5C represents samples collected from a subgroup of patients at their late clinical visits, indicated by dark orange in Figure 5D. This intriguing discovery suggested that clustering structures revealed by UMAP could facilitate the exploration of samples' hidden features.

To generate UMAP plots, there are several options for metric space, with '*euclidean*' distance as default (McInnes et al., 2018). We tested '*euclidean*' and another two representative metrics, '*canberra*' and '*cosine*' (details are in STAR Methods), and observed that the metric '*canberra*' led to more explicit clustering on UMAP projection, with patients' samples clustered into three subgroups: sG0, sG1, and sG2 (Figure 5E). By comparing

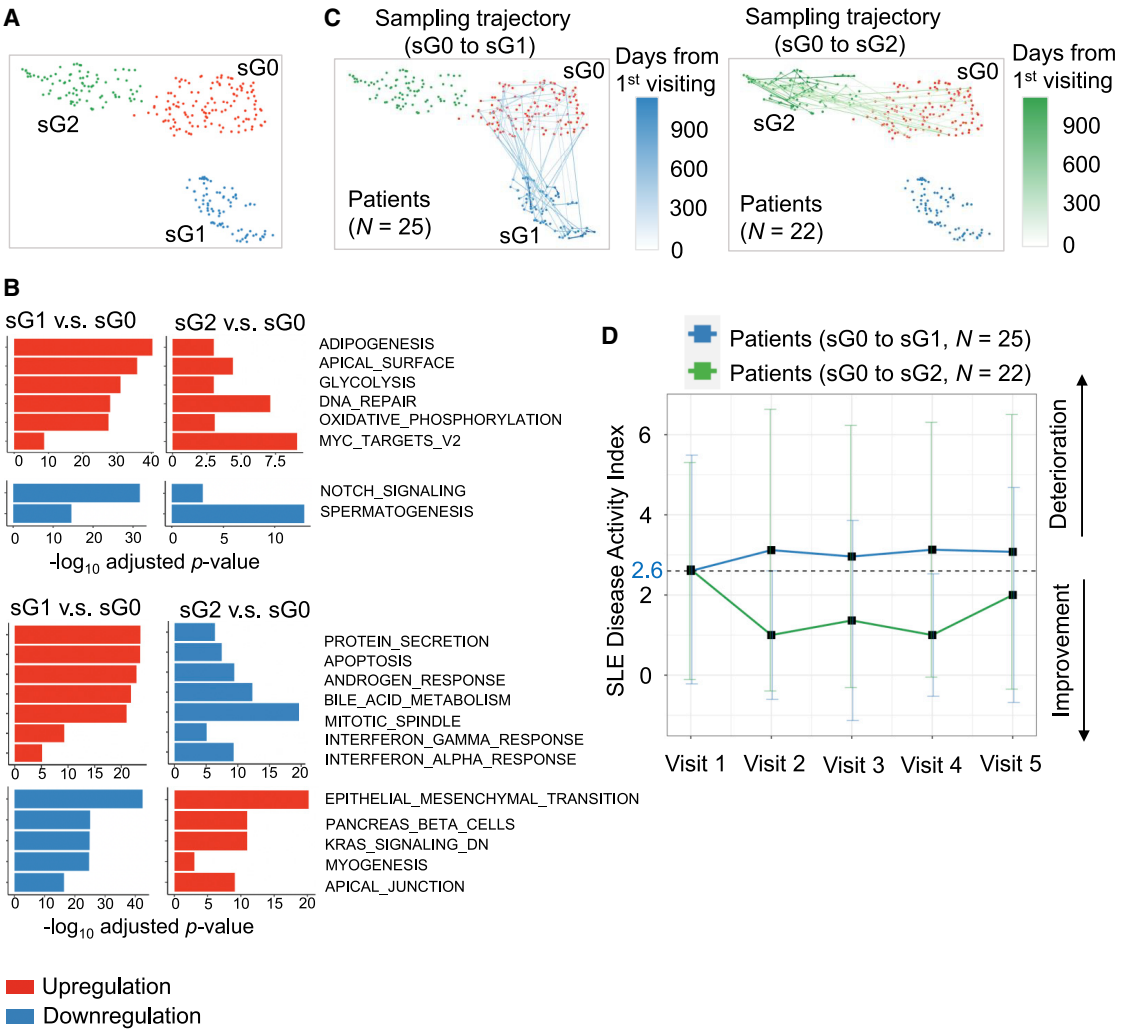


Figure 6. UMAP revealed clustering structures explained by clinical traits

(A) Hierarchical clustering of patient groups on two-dimensional embedded coordinates by UMAP with metric as 'canberra'.

(B) Histogram illustrating gene set enrichment analysis between sG1 versus sG0 and sG2 versus sG0 with top 20 differentially regulated molecular pathways (negative logarithm of the adjusted p value (base 10)). Red denotes upregulation, and blue for downregulation. The top two rows are in the same direction of regulation, and the bottom two rows are in the opposite direction.

(C) Visiting trajectories of each patient on UMAP plot with metric = 'canberra'. Each path connected data points corresponding to one patient with several visits. Data points in paths were connected by visiting timestamp. The light color denotes early visit and the dark color for late visits. The paths were mainly divided into two patterns: from sG0 to sG1 and from sG0 to sG2.

(D) Line chart of average SLEDAI changing along with visits between sG0 to sG1 and sG0 to sG1, with vertical bars representing SD. Both started with an average SLEDAI around 2.6; from sG0 to sG1 (colored by blue), the average SLEDAI increased, while from sG0 to sG2 (colored by green), the average SLEDAI decreased.

the timing of each sampling among these three subgroups, we found that samples in sG0 were collected significantly earlier than those collected in sG1 and sG2 ($p < 0.0001$), while the timing of sampling between sG1 and sG2 was comparable (Figure 5F).

According to the timing evolution (Figure 5D), samples of sG0 were collected earlier, while samples of sG1 or sG2 were collected later. The clear separation of late collected samples into two clusters of sG1 and sG2 suggested a biological divergence. To interpret the biological difference between sG1 and sG2, we applied gene set enrichment analysis (GSEA) using the R package EGSEA (Alhamdoosh et al., 2017), resulting in the top 20 differentially regulated molecular pathways between sG1 versus sG0 and sG2 versus sG0 (Figures 6B and S5). Comparing with sG0, sG1 and sG2 were common in six upregulated pathways (in red) and two downregulated pathways (in blue). However, seven upregulated and five downregulated pathways in sG1 showed opposite trends in sG2, suggesting the biological distinction between them.

Given longitudinal sampling of individual patients, we next investigated the visit trajectories of individual patients. Connection of samples from each patient demonstrated that most

patients (n = 47/65) showed one-directional trajectories from sG0 to sG1 or sG0 to sG2 (Figure 6C), in agreement with the timing evolution of patients' sample (Figure 5D). When initially admitted to the clinic to take samples (visit 1, Figure 6D), patients with distinct trajectories had comparable disease activities (SLE disease activity index [SLEDAI], mean ± SD, sG0 to sG1: 2.6 ± 2.71; sG0 to sG2: 2.6 ± 2.85). Widely used in clinical practice and research, SLEDAI is a global index that was developed as a clinical index for the assessment of lupus disease activity, and larger SLEDAI indicates worse disease conditions (Bombardier et al., 1992). Importantly, we noticed that the average SLEDAI at the following visits increased for patients with the trajectory from sG0 to sG1 (in blue, Figure 6D), indicating the disease deterioration of these patients, whereas the average SLEDAI at the following visits decreased for patients with the trajectory from sG0 to sG2 (in green, Figure 6D), indicating the disease improvement of these patients. The opposite disease progression between two trajectories was also supported by GSEA, which showed the key pathogenic pathways for SLE, including apoptosis (Muñoz et al., 2010), type I interferon (Banchereau and Pascual, 2006), and type II interferon (Ivashkiv, 2018) were increased in sG1 but decreased in sG2 (Figure 6B). Taken together, the deep exploration of the biological and clinical meaning of the clustering structure of dataset GEO: GSE121239 revealed by UMAP supports the future application of dimensionality reduction methods such as UMAP to reinforce sample heterogeneity analysis of bulk transcriptomic data.

Recommendation

Although PCA is often used in identifying sample-to-sample heterogeneity in the bulk transcriptomic analysis, our study demonstrated that the non-linear dimensionality reduction method UMAP improved the identification, visualization, and interpretation of clustering structures in sizable datasets. The analysis of the dataset GEO: GSE121239 suggested that the choice of the parameter *metric* in UMAP could affect the visualization of clustering structures of UMAP plots (Figure 6A). To understand how the parameter metrics of UMAP influence the visualization, we decided to compare different metrics in all 71 datasets. There are four major categories of metrics in UMAP, including Minkowski metrics, spatial metrics, angular metrics, and binary metrics, with binary metrics unsuitable for transcriptomic analysis. Therefore, we selected one representative metric from each category ('euclidean' for Minkowski metrics, 'canberra' for spatial metrics, and 'cosine' for angular metrics) and compared their performance. The two-dimensional UMAP visualization by metric 'euclidean', 'canberra', or 'cosine' revealed clustering structures in 41, 44, or 42 datasets, respectively, with 39 datasets in common (Figure 7A). Without any '*metric*' of the three showing a clear advantage, we recommend trying the three representative metrics for UMAP in visualizing the bulk transcriptomic data and being integrated into the pipeline for bulk transcriptomic analysis (Figure 7B). The analysis starts with transcript counts as the input, followed by data pre-processing, including log transformation and batch effect correction. UMAP will then be applied to visualize potential clustering structures. With identified clustering structures that may be caused by hidden batch effects, efforts will be made to remove hidden batch effects (Leek and Storey, 2007; Risso et al., 2014). The clustering structure should next be tested for the association with biological groups assigned by experimental design. The explicit association of the clustering structure with biological groups can ensure robust DGE analysis among different biological groups. If the clustering structure is related to specific sample features rather than biological groups, that feature should be treated as latent covariates in DGE analysis. In contrast, the clustering structure might reveal biological subgroups or a hidden factor to be analyzed separately for DGE analysis.

DISCUSSION

Transcriptomics, like other high-throughput omics technologies, faces the challenges to process large amounts of data with high dimensionality. Therefore, various informatics methods have been developed to generate meaningful results while reducing systemic errors in analysis. A systemic comparison of available analytic methods can help to illuminate the features of each method and advise the most suitable methods for specific applications. For example, the studies comparing t-SNE and UMAP in scRNA-seq analysis reported that UMAP, with Laplacian Eigenmaps initialization and cross-entropy object function (Becht et al., 2018; McInnes et al., 2018), usually outperformed t-SNE for dimensionality reduction, showing shorter running times, higher reproducibility, better neighborhood preservation, and more efficient organization of cell clusters (Becht et al., 2018; Sun et al., 2019). However, by tuning parameters, especially the initialization, t-SNE was shown to achieve comparable performance (Kobak and Berens, 2019; Kobak and Linderman, 2021). Since then, UMAP has become a mainstream method for scRNA-seq analysis and also has been repurposed in large-scale genotype datasets to explore fine structures and visualize genetic interactions (Dorrity et al., 2020; Sakaue et al., 2020).

In scRNA-seq analysis, tens or hundreds of thousands of cells are usually grouped into dozens of clusters with many clusters accounting for a small portion. Therefore, methods such as UMAP and t-SNE with the advantage of neighborhood preservation are necessary to distinguish fine structures at the local level. As for datasets of bulk RNA-seq, current methods for dimensionality reduction largely focus on the overall structure, so PCA is the mainstream method. It remains unknown whether the advantage of neighborhood preservation by UMAP and t-SNE results in a more reliable analysis for bulk transcriptomic analysis. If fine structures or clusters are revealed, is this information associated with any biological significance? To address this problem, we carried out the quantitative and qualitative comparison among four major dimensionality reduction methods in bulk transcriptomic analysis. Our results showed that UMAP was superior to other dimensionality reduction methods and could reveal biological information. Thus, we recommend UMAP as the visualization tool in the pipeline for bulk transcriptomic profiling and DGE analysis, particularly reinforcing sample heterogeneity analysis in datasets with large sample sizes.

Sample heterogeneity in bulk transcriptomic data reflects both biological and technical variation among samples. It is crucial to

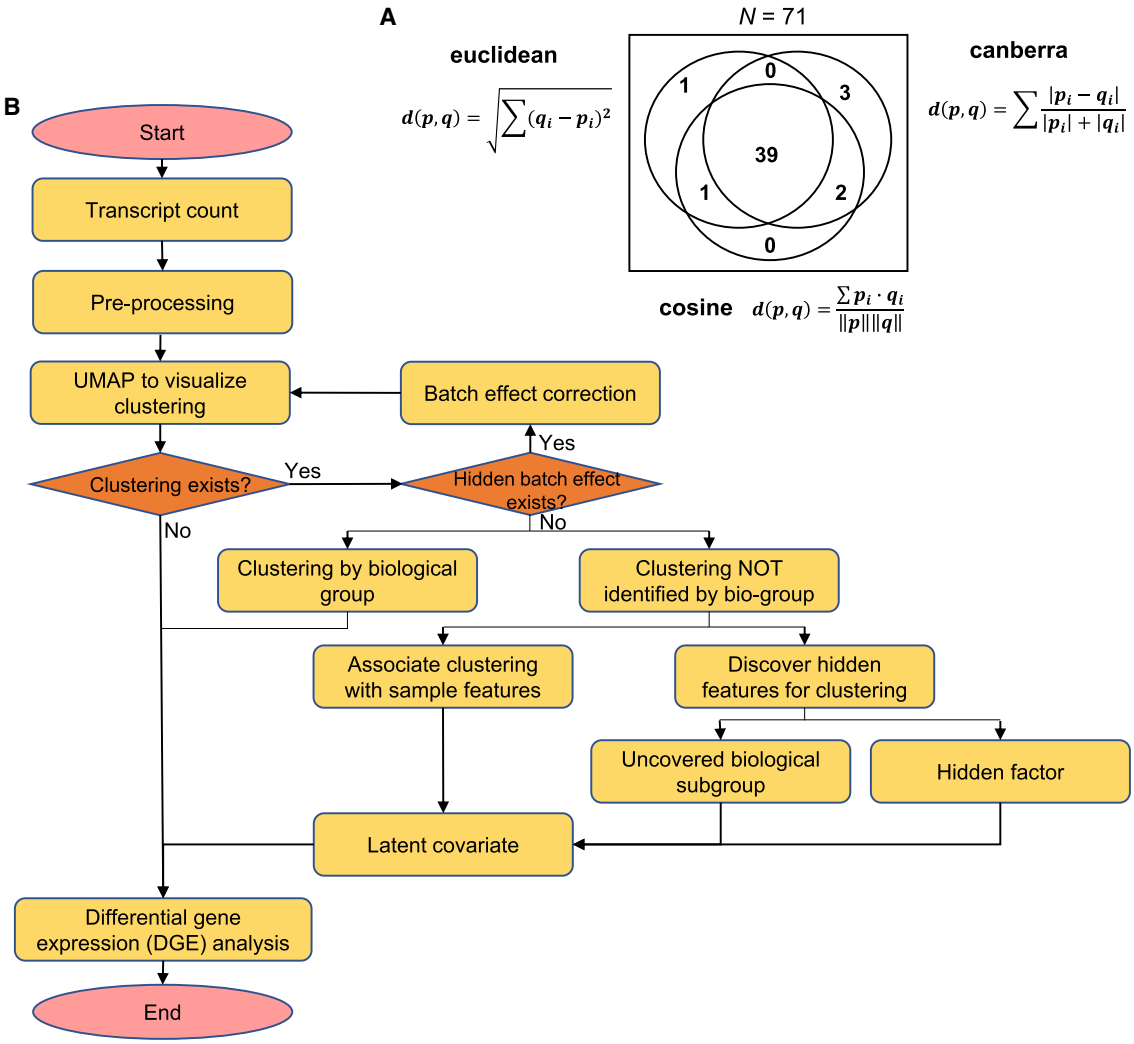


Figure 7. Recommendations for UMAP processing bulk transcriptomic datasets

(A) Venn diagram illustrating the overlap in the number of datasets having clustering structure by the UMAP plot under three different '*metric*' parameters: '*euclidean*', '*canberra*', and '*cosine*'.

(B) The recommendation pipeline for applying UMAP to bulk transcriptomic analysis.

detect among-sample heterogeneity before DGE analysis for bulk transcriptomic data so that appropriate analytic methods can be subsequently used to correct batch effects, remove outliers, and distinguish subgroups. Sample heterogeneity analysis by dimensionality reduction should consider both local and global information of datasets to congregate similar samples and distinguish different samples. PCA is the current mainstream tool of dimensionality reduction to visualize and detect among-sample heterogeneity, adopted by widely used analytic packages limma and edgeR (Ritchie et al., 2015; Robinson et al., 2010). PCA produces linear combinations of the original variables to generate the principal components (Holland, 2008), and visualization is generated by projecting the original data to the first two or more than two principal components; thus, PCA plot linearly shows global distance among data points. Similarly, the MDS method places each data point into two or higher dimensional space such that the between-point distances are preserved according to the pairwise distance of original data points (Borg and Groenen, 2005). Both PCA and MDS focus more on maintaining global information, which can fail to compactly cluster similar data points and face a major challenge with the rapid increase in sample sizes of bulk transcriptomic profiling datasets.

In contrast, t-SNE and UMAP model the pairwise distance by adopting the concept from the k-nearest neighbor (kNN) graph (Maaten and Hinton, 2008; McInnes et al., 2018) whereby two points are connected by an edge if their distance is among the k-th smallest distances compared with distances to other points (Preparata and Shamos, 2012). For dimensionality reduction by t-SNE or UMAP, all pairs of two points have edge weights indicating the probability for them being connected (connection probability). If the distance between two points is

among the k-th smallest distances compared with distances to other points, the connection probability between these two points is high. If the distance between two points is much greater than the k-th smallest distance, the connection probability between these two points is low (Maaten and Hinton, 2008; McInnes et al., 2018). Therefore, t-SNE and UMAP can efficiently preserve local distance information and cluster similar sample points. For large sample size in dataset resulting in the quadratic increase of pairwise comparisons, t-SNE and UMAP not only retain pairwise interaction but also focus on local information, thus outperforming PCA and MDS in detecting sample heterogeneity. Compared with t-SNE using random initialization and Kullback-Leibler divergence object function, UMAP utilizes Laplacian Eigenmaps initialization and cross-entropy object function (Kobak and Linderman, 2021; McInnes et al., 2018), which contribute to the global structure preservation. This might explain the overall better performance of UMAP than t-SNE. Notably, parameter tuning was found to significantly influence the performance of t-SNE, which demonstrated that t-SNE visualizations were improved to better preserve the global geometry of data by applying PCA initialization, a high learning rate, and multi-scale similarity kernels and considering exaggeration and downsampling-based initialization for very large datasets (Kobak and Berens, 2019). Moreover, both t-SNE and UMAP combine an attractive force between neighboring pairs of points with a repulsive force between all points; by changing the balance between the attractive and the repulsive forces, t-SNE yields the embeddings of UMAP with the exaggeration factor approximating to 4 (Böhm et al., 2020), indicating that UMAP has a stronger attraction that can better represent continuous manifold structures.

Among 71 bulk transcriptomic profiling datasets with >100 samples tested in this study, UMAP and t-SNE clearly outperformed PCA and MDS in identifying clusters associated with batch effects and biological groups pre-defined in study designs. It should be noted that, within 41 of 71 datasets that UMAP identified as clustering structures, fine-scale clustering structures were revealed and accounted for more than half (29 of 41) (Figure 1). The important question is whether the clustering structures discovered by UMAP represent biological significance. This question was then addressed in case studies of datasets with uncovered clustering structures. One case is the study that was initially designed to investigate how smoking influences blood gene expression of patients with COPD and utilized bulk transcriptomic profiling and DGE analysis (GEO: GSE71220) (Obeidat et al., 2016). Intriguingly, the PCA plot showed no clustering structure, while the UMAP plot revealed clustering structures, which was related to gender rather than smoking status (Figure 4). This information discovered by dimensionality reduction using UMAP suggests the gender feature should be treated as an important latent covariate in DGE analysis. Another example is the study that was designed to stratify patients with SLE, a highly complex autoimmune disease with heterogeneous clinical presentation, according to longitudinal disease activity and blood gene expression (GEO: GSE121239) (Toro-Domínguez et al., 2018). This study calculated a gene-by-patient correlation matrix computing a stringent Pearson correlation coefficient between gene expression data and SLEDAI scores across each patient's visits and then selected genes with the highest absolute correlation values by rank-sum method (Toro-Domínguez et al., 2018). Instead of this multiple-step process, dimension reduction by UMAP revealed the separation of samples by visit timestamp (Figure 5), which enabled the identification of two groups of patients with opposite changes of longitudinal disease activity (Figure 6). These results thus validate the application of UMAP in dimensionality reduction in stratifying SLE patients. Using several datasets as examples, we demonstrated that the clustering structures were associated with certain sample features and enabled to uncover unappreciated sample subgroups with specific biological and clinical features.

In analyzing 71 datasets, we demonstrated that UMAP was able to visualize the among-sample heterogeneity in two-dimensional space. Based on the two-dimensional embedding space of UMAP, clustering methods were deployed to define clusters of the data points (Figures 4D and 5E). The biological significance of the resulting clusters was validated by subsequent exploration and evaluation (Figures 4 and 6). For scRNA-seq data, a clustering algorithm is generally applied on low-dimensional space, for example, in the commonly used scRNA-seq package *Seurat* (Butler et al., 2018), a graph-based clustering algorithm to low dimensional space by PCA projection. The rationale of applying the clustering method to low-dimensional projected space mainly arises from the curse of dimensionality (Kriegel et al., 2009). When computing distance (e.g., Euclidean distance) in high-dimensional data, the difference in the distances between different pairs of samples becomes less precise, which hinders discriminating near and far points. Thus, applying clustering methods to low-dimensional embedding space is better to define clusters of data points. Therefore, we suggest that UMAP can be applied as a pre-processing step before generating clusters from bulk transcriptomic datasets.

Although UMAP has shown significant advantages in detecting among-sample heterogeneity, PCA has a property not present by other methods. PCA compresses the data by top-ranked principal components and computes the PCA score for each sample. Therefore, it can calculate the variable weight corresponding to the principal component coordinate system (PCA loadings), which explains the contribution of each variable to sample points. In contrast, the non-linear methods, including MDS, t-SNE, and UMAP, do not involve the variable weight such that dimensionality reduction embedding cannot be immediately explainable by variable weight. This might represent an area for the future improvement of UMAP or methods of a similar kind. Moreover, we compared dimensionality reduction methods for only two-dimensional visualization as the most common application; future work is required to reassess their performance when embedding space with dimensionality larger than two is conducted.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109442.

ACKNOWLEDGMENTS

We acknowledge the High-Performance Computing resources at the University of Queensland and Shandong Artificial Intelligence Institute. This work was supported by the Bellberry-Viertel Senior Medical Research Fellowship (to D.Y.) and Natural Science Foundation of Shandong Province (Major Basic Program, ZR2020ZD41 to M.S.).

AUTHOR CONTRIBUTIONS

Conceptualization: D.Y., Yang Yang, and D.W. Investigation: Yang Yang, H.S., Y.Z., and J.G. Formal analysis: Yang Yang and H.S. Supervision: D.Y., D.W., Yuchen Yang, Y.-G.D., and M.S. Writing – original draft preparation: Yang Yang and D.Y. Writing – review & editing: Yang Yang, T.Z., Y.W., D.W., and D.Y.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 13, 2021
Revised: June 1, 2021
Accepted: July 1, 2021
Published: July 27, 2021

REFERENCES

Alhamdoosh, M., Ng, M., Wilson, N.J., Sheridan, J.M., Huynh, H., Wilson, M.J., and Ritchie, M.E. (2017). Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics *33*, 414–424.

Banchereau, J., and Pascual, V. (2006). Type I interferon in systemic lupus erythematosus and other autoimmune diseases. Immunity *25*, 383–392.

Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.M., Acs, P., Turner, J., Anguiano, E., et al. (2016). Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. Cell *165*, 1548–1550.

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Böhm, J.N., Berens, P., and Kobak, D. (2020). A Unifying Perspective on Neighbor Embeddings along the Attraction-Repulsion Spectrum. arXiv *2007*, 08902.

Bombardier, C., Gladman, D.D., Urowitz, M.B., Caron, D., and Chang, C.H.; The Committee on Prognosis Studies in SLE (1992). Derivation of the SLEDAI. A disease activity index for lupus patients. Arthritis Rheum. *35*, 630–640.

Borg, I., and Groenen, P.J. (2005). Modern Multidimensional Scaling: Theory and Applications (Springer Science & Business Media).

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

Buyon, J.P., Petri, M.A., Kim, M.Y., Kalunian, K.C., Grossman, J., Hahn, B.H., Merrill, J.T., Sammaritano, L., Lockshin, M., Alarcón, G.S., et al. (2005). The effect of combined estrogen and progesterone hormone replacement therapy on disease activity in systemic lupus erythematosus: a randomized trial. Ann. Intern. Med. *142*, 953–962.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. (2005). Comparing community structure identification. J. Stat. Mech. *2005*, P09008.

Davenport, E.E., Amariuta, T., Gutierrez-Arcelus, M., Slowikowski, K., Westra, H.J., Luo, Y., Shen, C., Rao, D.A., Zhang, Y., Pearson, S., et al. (2018). Discovering in vivo cytokine-eQTL interactions from a lupus clinical trial. Genome Biol. *19*, 168.

Dorrity, M.W., Saunders, L.M., Queitsch, C., Fields, S., and Trapnell, C. (2020). Dimensionality reduction by UMAP to visualize physical and genetic interactions. Nat. Commun. *11*, 1537.

Dunning, J., Blankley, S., Hoang, L.T., Cox, M., Graham, C.M., James, P.L., Bloom, C.I., Chaussabel, D., Banchereau, J., Brett, S.J., et al.; MOSAIC Investigators (2018). Progression of whole-blood transcriptional signatures from interferon-induced to neutrophil-associated patterns in severe influenza. Nat. Immunol. *19*, 625–635.

Figgett, W.A., Monaghan, K., Ng, M., Alhamdoosh, M., Maraskovsky, E., Wilson, N.J., Hoi, A.Y., Morand, E.F., and Mackay, F. (2019). Machine learning applied to whole-blood RNA-sequencing data uncovers distinct subsets of patients with systemic lupus erythematosus. Clin. Transl. Immunology *8*, e01093.

Gerstner, J.R., Koberstein, J.N., Watson, A.J., Zapero, N., Risso, D., Speed, T.P., Frank, M.G., and Peixoto, L. (2016). Removal of unwanted variation reveals novel patterns of gene expression linked to sleep homeostasis in murine cortex. BMC Genomics *17* (*Suppl 8*), 727.

Guthridge, J.M., Lu, R., Tran, L.T., Arriens, C., Aberle, T., Kamp, S., Munroe, M.E., Dominguez, N., Gross, T., DeJager, W., et al. (2020). Adults with systemic lupus exhibit distinct molecular phenotypes in a cross-sectional study. EClinicalMedicine *20*, 100291.

Heller, M.J. (2002). DNA microarray technology: devices, systems, and applications. Annu. Rev. Biomed. Eng. *4*, 129–153.

Hoang, L.T., Tolfvenstam, T., Ooi, E.E., Khor, C.C., Naim, A.N.M., Ho, E.X.P., Ong, S.H., Wertheim, H.F., Fox, A., Van Vinh Nguyen, C., et al. (2014). Patient-based transcriptome-wide analysis identify interferon and ubiquination pathways as potential predictors of influenza A disease severity. PLoS ONE *9*, e111640.

Holland, S.M. (2008). Principal Components Analysis (PCA) (Department of Geology, University of Georgia), pp. 30602–32501.

Hong, S., Banchereau, R., Maslow, B.L., Guerra, M.M., Cardenas, J., Baisch, J., Branch, D.W., Porter, T.F., Sawitzke, A., Laskin, C.A., et al. (2019). Longitudinal profiling of human blood transcriptome in healthy and lupus pregnancy. J. Exp. Med. *216*, 1154–1169.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 1–14.

Ivashkiv, L.B. (2018). IFNγ: signalling, epigenetics and roles in immunity, metabolism, disease and cancer immunotherapy. Nat. Rev. Immunol. *18*, 545–558.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat. Commun. *10*, 5416.

Kobak, D., and Linderman, G.C. (2021). Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat. Biotechnol. *39*, 156–157.

Kriegel, H.-P., Kröger, P., and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Trans. Knowl. Discov. Data *3*, 1–58.

Law, C.W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G.K., and Ritchie, M.E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Res. *5*, 1408.

Le-Niculescu, H., Roseberry, K., Levey, D.F., Rogers, J., Kosary, K., Prabha, S., Jones, T., Judd, S., McCormick, M.A., Wessel, A.R., et al. (2020). Towards precision medicine for stress disorders: diagnostic biomarkers and targeted drugs. Mol. Psychiatry *25*, 918–938.

Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. *3*, 1724–1735.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. *11*, 733–739.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Levandowsky, M., and Winter, D. (1971). Distance between sets. Nature *234*, 34–35.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Lim, E., Vaillant, F., Wu, D., Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.L., Gyorki, D.E., Ward, T., Partanen, A., et al.; kConFab (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat. Med. *15*, 907–913.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theory *28*, 129–137.

Lou, J., Yang, Y., Gu, Q., Price, B.A., Qiu, Y., Fedoriw, Y., Desai, S., Mose, L.E., Chen, B., Tateishi, S., et al. (2021). Rad18 mediates specific mutational signatures and shapes the genomic landscape of carcinogen-induced tumors in vivo. NAR Cancer *3*, zcaa037.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Maaten, L.d., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, *1*, L.M. Le Cam and J. Neyman, eds., pp. 281–297.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. J. Open Source Softw. *2*, 205.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv *1802*, 03426.

University of Texas MDAnderson Cancer Center (2020). TCGA Batch Effects Viewer. https://bioinformatics.mdanderson.org/public-software/tcga-batch-effects/.

Muñoz, L.E., Lauber, K., Schiller, M., Manfredi, A.A., and Herrmann, M. (2010). The role of defective clearance of apoptotic cells in systemic autoimmunity. Nat. Rev. Rheumatol. *6*, 280–289.

Murtagh, F., and Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. *31*, 274–295.

Narang, V., Lu, Y., Tan, C., Camous, X.F.N., Nyunt, S.Z., Carre, C., Mok, E.W.H., Wong, G., Maurer-Stroh, S., Abel, B., et al. (2018). Influenza vaccine-induced antibody responses are not impaired by frailty in the community-dwelling elderly with natural influenza exposure. Front. Immunol. *9*, 2465.

National Institutes of Health (2020). Systemic Lupus Erythematosus (Lupus). https://www.niams.nih.gov/health-topics/lupus.

Obeidat, M., Ding, X., Fishbane, N., Hollander, Z., Ng, R.T., McManus, B., Tebbutt, S.J., Miller, B.E., Rennard, S., Paré, P.D., and Sin, D.D. (2016). The Effect of Different Case Definitions of Current Smoking on the Discovery of Smoking-Related Blood Gene Expression Signatures in Chronic Obstructive Pulmonary Disease. Nicotine Tob. Res. *18*, 1903–1909.

Oon, S., Monaghan, K., Ng, M., Hoi, A., Morand, E., Vairo, G., Maraskovsky, E., Nash, A.D., Wicks, I.P., and Wilson, N.J. (2019). A potential association between IL-3 and type I and III interferons in systemic lupus erythematosus. Clin. Transl. Immunology *8*, e01097.

Poličar, P.G., Stražar, M., and Zupan, B. (2019). openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. bioRxiv, 731877.

Preparata, F.P., and Shamos, M.I. (2012). Computational Geometry: An Introduction (Springer Science & Business Media).

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. *66*, 846–850.

Rawat, C., Kutum, R., Kukal, S., Srivastava, A., Dahiya, U.R., Kushwaha, S., Sharma, S., Dash, D., Saso, L., Srivastava, A.K., and Kukreti, R. (2020). Down-regulation of peripheral PTGS2/COX-2 in response to valproate treatment in patients with epilepsy. Sci. Rep. *10*, 2546.

Reynolds, D.A. (2009). Gaussian mixture models. In Encyclopedia of Biometrics, *741*, S.Z. Li and A.K. Jain, eds. (Springer), pp. 659–663.

Risso, D., Ngai, J., Speed, T.P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nat. Biotechnol. *32*, 896–902.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. *20*, 53–65.

Sakaue, S., Hirata, J., Kanai, M., Suzuki, K., Akiyama, M., Lai Too, C., Arayssi, T., Hammoudeh, M., Al Emadi, S., Masri, B.K., et al. (2020). Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat. Commun. *11*, 1569.

Sun, S., Zhu, J., Ma, Y., and Zhou, X. (2019). Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. Genome Biol. *20*, 269.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tasaki, S., Suzuki, K., Kassai, Y., Takeshita, M., Murota, A., Kondo, Y., Ando, T., Nakayama, Y., Okuzono, Y., Takiguchi, M., et al. (2018). Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. Nat. Commun. *9*, 2755.

Thakar, J., Mohanty, S., West, A.P., Joshi, S.R., Ueda, I., Wilson, J., Meng, H., Blevins, T.P., Tsang, S., Trentalange, M., et al. (2015). Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. Aging (Albany NY) *7*, 38–52.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. (Pozn.) *19* (*1A*), A68–A77.

Torgerson, W.S. (1952). Multidimensional scaling: I. Theory and method. Psychometrika *17*, 401–419.

Toro-Domínguez, D., Martorell-Marugán, J., Goldman, D., Petri, M., Carmona-Sáez, P., and Alarcón-Riquelme, M.E. (2018). Stratification of Systemic Lupus Erythematosus Patients Into Three Groups of Disease Activity Progression According to Longitudinal Gene Expression. Arthritis Rheumatol. *70*, 2025–2035.

Van Buren, E., Hu, M., Weng, C., Jin, F., Li, Y., Wu, D., and Li, Y. (2021). TWOSIGMA: A novel two-component single cell model-based association method for single-cell RNA-seq data. Genet. Epidemiol. *45*, 142–153.

Von Luxburg, U. (2007). A tutorial on spectral clustering. Stat. Comput. *17*, 395–416.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. Chemom. Intell. Lab. Syst. *2*, 37–52.

Wu, D., Pang, Y., Wilkerson, M.D., Wang, D., Hammerman, P.S., and Liu, J.S. (2013). Gene-expression data integration to squamous cell lung cancer subtypes reveals drug sensitivity. Br. J. Cancer *109*, 1599–1608.

Zhai, Y., Franco, L.M., Atmar, R.L., Quarles, J.M., Arden, N., Bucasas, K.L., Wells, J.M., Niño, D., Wang, X., Zapata, G.E., et al. (2015). Host Transcriptional Response to Influenza and Other Acute Respiratory Viral Infections—A Prospective Cohort Study. PLoS Pathog. *11*, e1004869.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Data S1	This paper	https://doi.org/10.17632/gtfrr4j5cx.3
Table S1	This paper	https://doi.org/10.17632/nddfs37dkh.2
Code	This paper	https://doi.org/10.17632/5gzyzffcyw.1
Software and algorithms		
PCA	(Wold et al., 1987)	https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
MDS	(Torgerson, 1952)	https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html
openTSNE	(Poličar et al., 2019)	https://github.com/pavlin-policar/openTSNE
UMAP	(Becht et al., 2018; McInnes et al., 2018)	https://github.com/lmcinnes/umap/
k-means	(Lloyd, 1982; MacQueen, 1967)	https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
Hierarchical clustering	(Murtagh and Legendre, 2014)	https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering
Spectral clustering	(Von Luxburg, 2007)	https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering
gmm	(Reynolds, 2009)	https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html
hdbscan	(McInnes et al., 2017)	https://github.com/scikit-learn-contrib/hdbscan
limma	(Ritchie et al., 2015)	https://www.bioconductor.org/packages/release/bioc/html/limma.html
EGSEA	(Alhamdoosh et al., 2017)	https://www.bioconductor.org/packages/release/bioc/html/EGSEA.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Di Yu (di.yu@uq.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyses existing, publicly available data from Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov). These GEO accession numbers for the datasets are listed in Table S1 (deposited at Mendeley) and is publicly available as of the date of publication. Data S1 has been deposited at Mendeley and are publicly available as of the date of publication. The DOIs are listed in the Key resources table.
- All original code has been deposited at https://github.com/yulmmuGroup/umap_on_bulk_transcriptomic_analysis and https://github.com/yulmmuGroup/transcriptomic_analysis_DGE_and_GSEA as well as Mendeley, and is publicly available as of the date of publication.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The total transcriptomic datasets were collected from the Gene Expression Omnibus (GEO) database with query conditions set as follows: the dataset type was expression profiling by array or by high throughput sequencing; the number of samples ranged from 100 to 10,000; the dimensionality of datasets ranged from 1443 to 70510 with median as 47878 (Figure S1E); organism was *Homo sapiens*; the publication date was from 2015/01/01 to 2020/03/01; sample source was PBMC or whole blood. Applying the query to the GEO database, we gained 214 results. We further manually removed the datasets in which each group owned less than 100 samples, resulting in 71 datasets. The datasets supporting the conclusions of this article are available in Gene Expression Omnibus repository (https://www.ncbi.nlm.nih.gov) with the GEO accession numbers in Table S1, including four columns (UMAP, t-SNE, MDS and PCA) showing which feature information explains the clustering structure of each dataset. Dimensionality reduction methods, clustering algorithms and bulk transcriptomic analysis R packages used in this paper are listed in the Key resources table.

METHOD DETAILS

Data pre-processing

Transcriptomic count data are pre-processed by log transformation (base 2) before applying dimensionality reduction methods. For PCA, we standardize the sample vectors by removing the mean and scaling to unit variance. For MDS, t-SNE (openTSNE implementation) and UMAP, we set the parameters n_jobs = −1: using multi-threads to run each method. For running time comparison, we first use PCA to reduce the dimensionality of data to 100 (Kobak and Berens, 2019; Luecken and Theis, 2019), before applying t-SNE and UMAP to the datasets. The initialization of openTSNE is 'pca' (default), 'spectral' or 'random' while UMAP is 'spectral' (default) or 'random'.

Dimensionality reduction methods

The four dimensionality reduction methods were all implemented in Python language. PCA (scikit-learn 0.23.1) was available at https://scikit-learn.org/stable/ with parameter settings as *n_components* = 2, *svd_solver* = 'auto', *iterated_power* = 'auto', *random_state* = None; MDS (scikit-learn 0.23.1) was available at https://scikit-learn.org/stable/ with parameter settings as *n_components* = 2, *n_init* = 4, *max_iter* = 300, *eps* = 0.001, *n_jobs* = −1, *dissimilarity* = 'euclidean' ; t-SNE (openTSNE 0.6.0) was available at https://github.com/pavlin-policar/openTSNE with parameter settings as *n_components = 2, perplexity = 30, initialization = 'pca', early_exaggeration = 12, n_iter = 500,metric = "euclidean," n_jobs = −1,neighbors = 'exact', random_state = 42*; UMAP (umap-learn 0.5.1) was available at https://github.com/lmcinnes/umap/ with parameter settings as *n_components = 2, n_neighbors = 15, init = 'spectral', min_dist = 0.1, metric = 'euclidean', n_jobs = −1, random_state = 42*.

Classifying clustering structures

Two-dimensional visualization of individual datasets was independent manually assessed by three assessors without the knowledge of datasets to report clustering structures. The majority rule was applied for disagreement. The density-based clustering method hdbscan (Bombardier et al.) was applied to the two-dimensional embedding coordinates to calculate the silhouette score (Rousseeuw, 1987) (Figure S1D), with a higher score value indicating a stronger clustering configuration. The silhouette scores were assumed to follow bimodal distribution, and two estimated normal distributions were generated by fitting a finite mixture model with the Expectation-Maximization algorithm (https://github.com/choisy/cutoff).

The datasets showing clustering structures are classified into three types (Batch effect, Biological group and Uncovered clustering) by associating the clustering structures with available feature information in the metadata. Since identifying clustering features or interpreting existing biological features are generally more important than detecting batch effects, a hierarchical classification is applied, which can also avoid double counting by assigning each dataset into only one classification. If the clustering structures show a feature that cannot be explained by batch effect or biological group, this dataset will be classified as "Uncovered clustering"; if there is no uncovered clustering but the feature is associated with any biological group, this dataset will be classified as "Biological group"; otherwise, the dataset will be assigned as "Batch effect" (Figure S1C).

Clustering algorithms

Five clustering methods were deployed to evaluate the clustering accuracy. All methods were implemented in Python. k-means, hierarchical clustering, spectral clustering and gmm were available at https://scikit-learn.org/stable with version scikit-learn 0.23.1. hdbscan was available at https://hdbscan.readthedocs.io/en/latest/ with version hdbscan 0.8.27.

Clustering accuracy (NMI, ARI)

For clustering accuracy analysis, we applied five clustering methods to the embedded low-dimensional space by dimensionality reduction methods. The clustering methods included k-means clustering (Python function *KMeans*), hierarchical clustering (Python function *AgglomerativeClustering*), spectral clustering (Python function *SpectralClustering*), hdbscan (Python function *hdbscan*) and Gaussian mixture model (Python function *GaussianMixture*). In these clustering methods, the number of clusters *k* was set to be the

known number of different groups in the data, except for hdbscan which is a density-based clustering algorithm (we set the *min_cluster_size* as 10). We applied the five clustering methods to the embedded space of 26 datasets with available features for groups. The retained partitions inferred using the low-dimensional components were compared to the true clusters. The level of agreement between the clustering partition and the true clusters was measured by two criteria: the Adjusted Rand Index (ARI) (Rand, 1971) and the Normalized Mutual Information (NMI) (Danon et al., 2005). Given two partitions $X = \{X_1, ..., X_r\}$ and $Y = \{Y_1, ..., Y_s\}$, the ARI and NMI are defined as:

$$\mathrm{ARI}(\mathrm{X},\ \mathrm{Y}) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}} \text{ and } NMI(X,\ Y) = \frac{2\mathrm{MI}(\mathrm{X},\mathrm{Y})}{\mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y})}$$

where $n_{ij} = \left|X_i \cap Y_j\right|$ is the number of common data points between X_i and Y_j, $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$, $\mathrm{MI}(\mathrm{X},\mathrm{Y})$ is the mutual information between cluster labels X and Y, $\mathrm{H}(\mathrm{X})$ and $\mathrm{H}(\mathrm{Y})$ are the entropy function for cluster labeling. We used Python function *adjusted_rand_score* and *normalized_mutual_info_score* to calculate ARI and NMI, respectively.

Silhouette score

The silhouette score is to compute the mean silhouette coefficient of all sample points. The silhouette score falls within the range [-1, 1]. The silhouette score of 1 means that the clusters are very dense and nicely separated; the score of 0 means that clusters are overlapping. A score of less than 0 means that data belonging to clusters may be incorrect. For each sample point $i \in C_i$(data point i in cluster C_i), the silhouette coefficient is calculated by $s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$, where a_i is the mean intra-cluster distance and $b_i$is the mean nearest-cluster distance with $a_i = \frac{1}{|C_i - 1|}\sum_{j \in C_i, i \neq j} d(i,j)$ and $b_i = min_{k \neq i}\frac{1}{|C_k|}\sum_{j \in C_k} d(i,j)$, respectively. $d(i,j)$ is the distance between data point i and j; C_k is another cluster different from C_i. The silhouette score is computed as $s = \frac{1}{\sum_i |C_i|}\sum_i s_i$.

Neighborhood preserving evaluation

The evaluation of neighborhood preserving is to assess how the reduced low-dimensional space retains the local information compared with the original high dimensional dataset. For the original space and embedded space, the k-nearest neighbors (kNNs) for each data point were computed respectively (denoted as sets X and Y). The Jaccard index (JI) (Levandowsky and Winter, 1971) was used to calculate the neighborhood similarity between original and embedded space: $JI = |\mathrm{X} \cap \mathrm{Y}|/|\mathrm{X} \cup \mathrm{Y}|$ where $|\cdot|$ means set cardinality, then the average Jaccard index (AJI) across all data points were computed to measure the neighborhood preservation.

Running time

We measured the running time of PCA, MDS, t-SNE and UMAP on Intel Xeon Gold 5218 @2.30 Ghz, 16 threads. The running time was determined in R using the ''elapsed'' (wall clock) time measurements, which allows for consistent timing across methods. For total-RNA datasets, the number of samples is moderately large with hundreds of data points. We generated datasets with sample size ranging from 200 to 20,000 by random sampling with replacement to evaluate the computation efficiency. The data were generated by randomly sampling with replacement from the three largest datasets GSE36382, GSE65391 and GSE65907, with dimensionality as 35900, 43787 and 28264, respectively. We evaluated the parameter '*neighbors*' of t-SNE in running time performance for processing high dimensionality (larger than 28000) datasets (Figure S1F). t-SNE with parameter '*neighbors*' set as 'exact' consumed less time than the others, especially the default 'auto'.

Separability of batch effects and biological groups

To evaluate the capability of each dimensionality reduction method in separating the features by the embeddings, we first assigned feature labels to 21 datasets. For each dataset, we used Python function *train_test_split* with parameter *test_size = 0.3* to divide the dataset into 70% training set and 30% test set. For each algorithm, a random-forest classifier by Python function *RandomForestClassifier* was trained using the group labels as target variable and the embedding's coordinates as training variables. We then utilized these classifiers to predict cluster identities on the test set and computed the accuracy of these predictions, thus assessing the ability of each method to separate groups.

Parameter 'metric' in dimensionality methods

To evaluate the parameter metric in the performance of UMAP, we chose the metric 'euclidean', 'canberra' and 'cosine' as the representatives for Minkowski metrics, spatial metrics and angular metrics, respectively. The three metrics are computed as follows:

$$d^{\mathrm{euclidean}}(p,q) = \sum_i (p_i - q_i)^2,\ \cdot d^{\mathrm{canberra}}(p,\ q) = \sum_i \frac{|p_i - q_i|}{|p_i| + |q_i|},\ \cdot d^{\mathrm{cosine}}(p,q) = \frac{\sum_i p_i \cdot q_i}{\|p\|\,\|q\|}$$

QUANTIFICATION AND STATISTICAL ANALYSIS

We applied one-way ANOVA to compare the performance of four methods in 71 datasets (Figure 2B). Two-tailed paired t tests were applied to compare the accuracy on 21 datasets (Figure 2B). The frequency difference of categorical variables was examined by χ^2 test on three clusters with size 183, 377 and 97, respectively (Figure 4E). To calculate the timing of sampling, the timing of individual patients' first clinic visit is set as day 0. The timing of sample collections during subsequent clinic visits is calculated accordingly. One-way ANOVA was applied to compare the visiting day difference (Figure 5F). The p value less than 0.05 was considered statistically significant. We used R (3.6.3) package *limma* (Law et al., 2016; Ritchie et al., 2015) for differential gene expression (DGE) analysis. Top 100 differential expressed genes were chosen to be included in the heatmap among control and experimental groups (Figure 4F). We applied gene set enrichment analysis (GSEA) by R package *EGSEA* (Figure 6B), where the Molecular Signatures Database (MSigDB) was set as H: hallmark gene sets (Liberzon et al., 2015).