A PCA-Enhanced *t*-SNE Plot and Its Application in Biological and Medical Research

Meng Guo[1], Haoyu Liu[1] and Jiajuan Liang[1,2,*]

[1]*Department of Statistics and Data Science, Beijing Normal–Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China*

[2]*Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal–Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China*

Abstract: In this paper, we apply a two-step dimension reduction method, PCA-*t*-SNE to a real gene expression dataset as case study. It turns out that the PCA-*t*-SNE can signigicantly improve the visualization and cluster separation of high-dimensional biological data. While *t*-SNE alone often fails to reveal clear cluster structures in complex datasets, our approach first applies Principal Component Analysis (PCA) to reduce noise and dimensionality, followed by *t*-SNE to condense the data into a two-dimensional space and then apply the *k*-means to clustering the two-dimensional data. We demonstrate that PCA-*t*-SNE produces more distinct and interpretable clusters compared to the standard *t*-SNE. Statistical validation via a projected *F*-test for MANOVA confirms that clusters derived from PCA-*t*-SNE exhibit significantly greater mean separation, with lower *p*-values, underscoring the enhanced discriminative power of the method. The proposed PCA-*t*-SNE plot proves particularly effective for nonlinear data where conventional *t*-SNE performs poorly, offering a robust visualization tool and supporting the utility of sequential dimension reduction in exploratory data analysis for biological and medical research.

Purpose: This study aims to evaluate the effect from a combination of the classical PCA and the modern *t*-SNE technique for dimension reduction in clustering of high-dimensional gene expression data from the aspects of both visualization and MANOVA.

Methods: This paper presents a combined approach to dimension reduction for high-dimensional gene expression data. The effect of the visual approach is re-enhanced by the classical MANOVA method for large sample sizes ($n > p$) and the newly developed MANOVA method for small sample sizes ($n \leq p$).

Results: The proposed PCA *t*-SNE approach significantly improves the pure *t*-SNE approach for the selected gene expression dataset in the sense of clearer classification of the data from both visual observation and statistical significance tests. This provides a pre-processing of high-dimensional gene expression data before implementing the nonlinear dimension reduction, making the *t*-SNE approach more effective.

Contribution: We carry out a successful application of the two-step dimension reduction method PCA-*t*-SNE to a real gene expression dataset as case study. The idea of the PCA-*t*-SNE approach to visualizing high-dimensional gene expression data, enhanced by the projection-type MANOVA tests, opens a new way to discrimination of complex high-dimensional with statistical significance in the case of high dimension with a small sample size ($n \leq p$). It enhances the clustering of those nonlinear-type of data where the pure *t*-SNE almost fails to discriminate the clusters, and provides insight into a two-step dimension reduction approach.

Keywords: Clustering, Gene expression data, *k*-means algorithm, Principal component analysis, projected *F*-test, *t*-SNE plot.

1. INTRODUCTION

With the rapid development of high-throughput sequencing technology (e.g., RNA-seq), high-dimensional gene expression data have become increasingly important for studying tumor mechanisms, characterizing cellular heterogeneity, and identifying functional genes. However, such data are typically characterized by extremely high dimensionality, a limited number of samples, and complex noise structures, forming a typical (small n, large p) analysis scenario [20, 21]. In such cases, traditional statistical methods are often difficult to apply directly.Therefore, how to effectively reduce dimensionality, reveal latent structures, and achieve reliable interpretation while retaining essential information has become a central challenge in modern biostatistics and data science [21, 22].

Principal Component Analysis (PCA) is one of the most widely used linear dimensionality reduction techniques. By identifying orthogonal directions that maximize global variance, PCA compresses high-dimensional data and is frequently employed in genomics and transcriptomics for noise reduction, identification of major sources of variation, and data visualization [2, 3, 10]. Due to its computational efficiency and geometrically interpretable results, PCA

*Address correspondence to this author at the Guangdong Provincial/Zhuhai Key Laboratory of Interdisciplinary Research and Application for Data Science, Beijing Normal-Hong Kong Baptist University, 2000 Jintong Road, Tangjiawan, Zhuhai 519087, China; E-mail: jiajuanliang@bnbu.edu.cn

is often used as a preliminary step in analyzing high-dimensional gene expression data. However, as a linear method, PCA primarily captures global structures and often performs poorly when facing with complex nonlinear-manifold-type of data, local heterogeneity, or fine-grained cluster boundaries [2, 10, 12]. Although several PCA variants have been proposed in recent years such as FeatPCA [4], which relies on feature subspaces or improved projection strategies they fundamentally inherit the limitations of linear dimensionality reduction.

In contrast, the t-distributed Stochastic Neighbor Embedding (t-SNE, [23, 24]) is a nonlinear technique that emphasizes the preservation of local neighborhood relationships. It has shown considerable promise in visualizing high-dimensional biological data, such as single-cell RNA-seq data [5-8, 19]. By constructing probability distributions in both high- and low-dimensional spaces and minimizing their divergence, the t-SNE ensures that similar samples in the original space remain proximate in the embedded space, thereby revealing potential cell subpopulations [5, 7]. Nevertheless, studies have also highlighted t-SNE's tendency to distort global geometry, its high sensitivity to hyperparameters and initialization, and the difficulty of quantitatively comparing embeddings across different runs [15, 17]. Moreover, the t-SNE lacks a built-in statistical inference framework to assess whether clusters observed in low-dimensional visualizations correspond to statistically significant differences in the original high-dimensional space [8, 9].

Existing research indicates that PCA and the t-SNE each possess distinct strengths in analyzing high-dimensional gene expression data, yet each exhibits clear limitations when used alone: PCA reliably captures global variation but often fails to reflect nonlinear local structures; the t-SNE can reveal local neighborhood patterns but depends heavily on high-quality input representations and may sacrifice global structural consistency [7, 15]. Both theoretical and empirical studies emphasize that the t-SNE's performance is highly influenced by preprocessing and denoising steps, and inadequate input can lead to unstable or even misleading embeddings [7, 16, 17]. Consequently, designing a dimensionality reduction pipeline that balances global stability with local expressive power has emerged as a key methodological challenge in high-dimensional biological data analysis.

Motivated by these considerations, this paper investigates an integrated dimensionality reduction and clustering framework that combines PCA and the t-SNE. The approach aims to stabilize the global geometric structure through linear dimensionality reduction (PCA) and enhance local neighborhood representation via nonlinear embedding (t-SNE), thereby producing more robust and interpretable low-dimensional visualizations. Furthermore, we introduce a projection-based multivariate statistical testing procedure inspired by classical and projection-type MANOVA (multivariate analysis of variance) to statistically validate whether cluster structures observed in the low-dimensional embedding reflect a significant separation in the original high-dimensional space [9, 13, 14]. The remainder of this paper is organized as follows. Section 2 describes the gene expression dataset and preprocessing steps. Section 3 gives details about the methodology, including the selection of the number of clusters and the comparative clustering analysis based on standard t-SNE and PCA-enhanced t-SNE. Section 4 compares the clustering effects from the standard t-SNE and the PCA-t-SNE and evaluates their statistical significance using a projected F-test [9]. Some concluding remarks are given in the last section.

2. MATERIALS AND METHODS

2.1. The t-SNE Plot

The t-distributed Stochastic Neighborhood Embedding (t-SNE) is a non-linear dimensionality reduction technique designed to visualize high-dimensional data in a low-dimensional space while preserving local neighborhood relationships between observations. This study employs the t-SNE as an exploratory visualization tool to analyze latent clustering structures within high-dimensional gene expression data. It is known [23, 24] that the t-SNE plot possesses the following characteristics in visualizing high-dimensional data: [label=)]

1. Preservation of local neighborhood structures: the t-SNE emphasizes pairwise local similarity, enabling the visualization of fine-scale structures and latent clusters that may not be captured by linear dimensionality reduction techniques.

2. Non-linear representational capability: by permitting non-linear mappings from the original data space to the embedded space, the t-SNE adapts to the complex data geometries

commonly encountered in high-dimensional environments.

Despite these advantages, the direct use of t-SNE in high-dimensional settings is accompanied by several well-recognized limitations [7]: [label=)]

1. Sensitivity to noise and parameter adjustment parameters: When the dimension of the data is too high relative to the available sample size, the t-SNE embedding may be significantly affected by noise and parameter modulation parameters such as confusion and initialization. Therefore, multiple runs may produce significantly different embedding, which makes the evaluation of cluster stability complicated.

2. The interpretability of geometric structure is limited: The distance, relative cluster size and separation observed in low-dimensional embedding are not directly quantitatively explained in the original characteristic space. Therefore, the visualization results of t-SNE usually show a large central cluster area or a smooth transition pattern between different groups, which makes it difficult to objectively quantify the boundaries of the cluster.

3. Lack of formal inference framework: the t-SNE is mainly designed as a visualization tool, not a method for statistical inference. Therefore, the obvious cluster pattern revealed by embedding needs to be independently verified by appropriate statistical procedures formulated in the original high-dimensional space.

These considerations show that when the t-SNE is directly applied to high-dimensional data, appropriate preprocessing steps are usually required to achieve better results. Specifically, retaining the main global source of change while reducing the dimension helps to reduce the impact of noise and redundant characteristics, and can generate more stable and easy-to-understand low-dimensional representations.

2.2. PCA for Initial Dimension Reduction

Principal component analysis (PCA) is a classic linear dimension reduction technique that projects high-dimensional observations onto a low-dimensional subspace composed of mutually orthogonal directions. The construction of these directions aims to capture the largest possible variation in the data, thus providing a compact representation that can not only retain the main global structure, but also discard subtle changes and noise. In a high-dimensional environment, PCA provides many practical advantages for subsequent data analysis [2]: [label=)]

1. Extract the main global changes: by sorting the components according to the differences in interpretation, PCA concentrates the main sources of global change in a few major components. This feature enables PCA to effectively summarize the overall data structure when the original dimension is large.

2. Noise reduction and de-redundancy processing: high-dimensional data usually contains related variables and noise characteristics, which contribute little to system changes. Retaining only the main components can weaken such effects, thus providing a more stable representation for subsequent analysis.

3. Projection with determinism and high computational efficiency: for a given dataset, PCA can generate a unique projection, which is not affected by random initialization. This certainty and computational efficiency make PCA a reliable preprocessing tool in high-dimensional analysis.

At the same time, PCA as an independent data visualization tool also has its own limitations: [label=)]

1. Restrictions on linear structure: Since this is a linear method, PCA cannot clearly simulate the possible nonlinear relationships in complex high-dimensional data.

2. Limited sensitivity to local regional patterns: Because PCA focuses on overall variance, fine local structure or fine cluster separation may not be clearly reflected in low-dimensional PCA representation.

Overall, these characteristics show that the principal component analysis method is very suitable for capturing the main structure of the global situation and reducing the dimension, but when used alone, it may not be enough to reveal the detailed local pattern in the high-dimensional data.

2.3. The PCA-Enhanced t-SNE

The PCA t-SNE is motivated by the guidelines for appropriate use of t-SNE in high-dimensional data

analysis elaborated by Kobak and Berens [7]. These guidelines emphasize that the t-SNE should be regarded as an exploratory visualization tool, and proper pretreatment can significantly improve the quality and interpretability of the obtained embedding. Based on this principle, a unified framework is constructed, integrating PCA and t-SNE to solve the complementarity problem of high-dimensional structures.

One of the core design options of the framework is to apply PCA before t-SNE embedding. This sorting reflects the different goals of these two methods. PCA acts directly on the original data space and generates a linear and deterministic projection, which can summarize the main global changes. In contrast, t-SNE constructs a nonlinear embedding, and its geometric structure is optimized for the retention and visualization of local neighborhoods, not for representing variance. Therefore, the application of PCA after applying t-SNE is equivalent to linear transformation of embedding that has undergone nonlinear deformation, which cannot restore the meaningful global structure. The selected sorting ensures that the dimension reduction operation is carried out in an environment that can explain global changes. Under this framework, the functions of PCA and the t-SNE are complementary but independent of each other. As a preprocessing step, PCA projects high-dimensional data into a medium-dimensional subspace, which can retain major global changes while suppressing noise and redundancy. This intermediate representation provides a structured input for subsequent embedding and reduces the impact of high-dimensional features that may cover up local neighborhood relationships. Then, the representation after the reduction of the PCA dimension is applied to t-SNE to build a low-dimensional embedding that emphasizes the local structure, so as to facilitate visualization and clustering.

By combining PCA and t-SNE in this gradual way, the framework makes full use of the advantages of these two methods while overcoming their respective limitations. Compared with the direct application of the t-SNE, the method of enhancing PCA is to operate on a more stable and informative representation, so the generated embedding is less sensitive to false changes. Compared with using only PCA, the framework retains the ability to reveal nonlinear local patterns that are not well captured by linear projection. Therefore, the proposed PCA t-SNE provides a coherent and principle-based method for exploratory analysis, clustering and subsequent statistical verification in a high-dimensional environment.

3. A PRACTICAL COMPARISON BETWEEN THE t-SNE AND THE PCA t-SNE

3.1. Data Description and Preprocessing

This study is based on a publicly available high-dimensional gene expression dataset. Let

$$\in \mathbb{R}^{n\times p}$$

denote the gene expression data matrix, where n is the sample size, which is the number of biological samples , p denotes dimension, which is the number of measured gene-expression variables. For the dataset studied in this paper, the data matrix has the dimension:

$$\in \mathbb{R}^{n\times p} = \mathbb{R}^{29\times 22447},$$

corresponding to 29 samples with expression measurements for approximately 22,447 genes per sample. Therefore, each line of the matrix stands for a gene expression pattern. The data matrix demonstrates the characteristics of high dimension with a small sample size, which is commonly encountered in medical and biological research when the number of patients is limited or the limited number of genes with large-scale expression by modern DNA-RNA sequencing technology. The dataset was downloaded from the ArrayExpress repository maintained by the European Bioinformatics Institute (EBI), available at https://www.ebi.ac.uk/biostudies/ArrayExpress/studies/E-MTAB-9428?query=carrer%20RNA-Seq with the data file provided as MTAB-9428.zip. This dataset serves as the basis for all subsequent clustering and visualization analyses.

Before dimension reduction and clustering, the standard preprocessing steps were applied. All non-human genes are eliminated to avoid cross-species contamination, and genes with missing values are also excluded. The original RNA-sequencing counting data is standardized using the M-value mean (TMM) method in the edgeR framework and converted to count per million (CPM) to ensure comparability between samples with different sequencing depths. Then, the processed data is sorted into a sample matrix of genes for subsequent analysis. Since the original data set contains more than 20,000 gene expression variables, a feature selection step was carried out to reduce high-dimensional noise and improve the stability of nonlinear

embedding and clustering. Specifically, only the first 2,000 genes with the highest expression variability were retained as input characteristics for all subsequent analyses. This step significantly reduces the computational complexity while retaining the main biological signals, thus providing a stable data basis for the selection of the number of clusters through the elbow diagram and the comparative analysis of the t-SNE and PCA t-SNE.

3.2. Determination of the Number of Clusters

Prior to applying dimension reduction and clustering methods, the number of clusters was assessed using the elbow method based on within-cluster dispersion. Specifically, k-means clustering was performed for a range of candidate cluster numbers, and the corresponding within-cluster sum of squares was examined as a function of the number of clusters. The elbow method [6] provides a visual observation for determining a suitable cluster number by locating a point at which the marginal decrease in within-cluster dispersion begins to diminish, indicating a trade-off between model complexity and goodness of fit.

As illustrated in the elbow plot shown in Figure **1** the curve exhibits a clear change in slope in the region corresponding to three to five clusters. In particular, the reduction in within-cluster dispersion is substantial when increasing the number of clusters up to this range, while further increases lead to comparatively smaller gains. Based on this observed pattern, the values k = 3, 4, 5 are selected as candidate cluster numbers for subsequent analyses. These choices allow for a visualization-based comparison of clustering behavior across different levels of granularity and provide a practical basis for evaluating the stability and interpretability of the resulting low-dimensional representations.

3.3. Clustering Based on Pure t-SNE

Figures **2**-**4** present the t-SNE-based clustering results obtained using k-means for three candidate numbers of clusters, k = 3, 4, 5. For each value of k, the t-SNE embeddings were computed using perplexity values of 10, 20, 30, and 40, and the overall clustering patterns were compared across these settings. While changes in perplexity lead to moderate variations in the geometric layout of the embedding, the main clustering characteristics remain similar and are therefore discussed with respect to the number of clusters. Here we employ the R package Rtsne (https://cran.r-project.org/web/packages/Rtsne/index.html to do all t-SNE plots and follow the idea in [5, 15, 17] to adjust the perplexity parameter to obtain a desirable t-SNE plot for each case.

For k = 3 in Figure **2**, the clustering results exhibit a pronounced imbalance in cluster sizes. One of the clusters occupies an absolute dominant position in the sample, containing about 93% of the observation points, and forming a large and continuous curved structure in the embedded space. In contrast, the smallest cluster accounts for less than 1% of the total, mainly composed of a small number of samples at the embedded edge, while the remaining clusters present a narrow band distribution adjacent to the dominant structure. The division essentially reflects the separation of a very small number of marginal samples from a large-scale and continuously changing subject group, rather than forming three relatively balanced clusters with sample support.

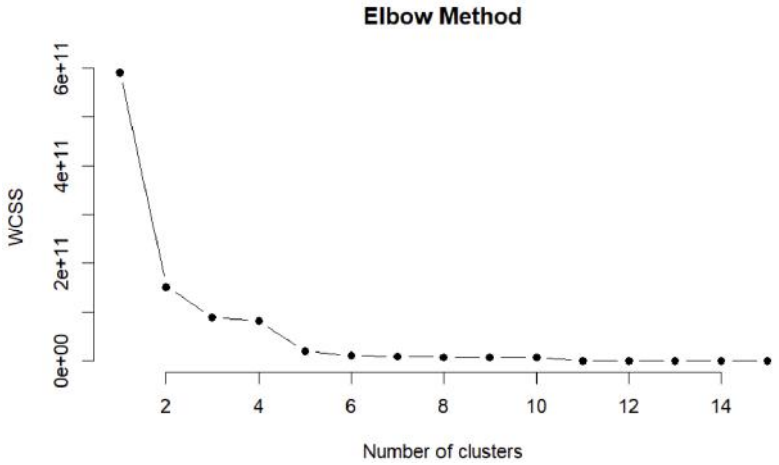

Figure 1: Number of clusters.

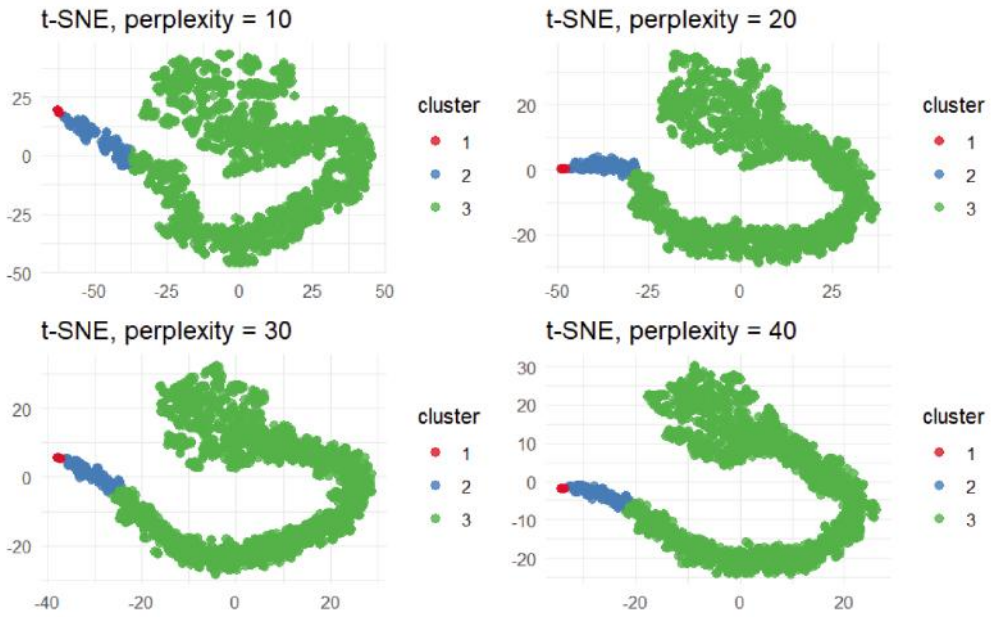

Figure 2: Traditional t -SNE plots under different perplexity values ($k = 3$).

When k = 4 in Figure **3**, the dominant structure observed for k = 3 is further subdivided. However, this subdivision still mainly occurs on the same continuous surface, and each cluster is more like a local cut along a continuous trajectory than an independent area that is clearly separated in space. Under this configuration, the largest cluster still contains more than 80% of the samples, while the sample ratio of the smallest cluster is still less than 1%. The rest of the samples are mainly distributed in one to two medium-sized clusters. On the whole, the cluster results still show the structural characteristics of a highly dominant cluster + several extremely small clusters, and there is an obvious continuous transition between different clusters.

A similar pattern is observed for k = 5 in Figure **4**. With the further increase in the number of clusters, the embedded structure is further divided, but this division still mainly affects the original continuous structure. At this time, the largest cluster accounts for about 77% of the total sample, while the smallest cluster still accounts for less than 1%. Although the cluster particle size has improved, the new cluster does not correspond to the obviously independent area in the

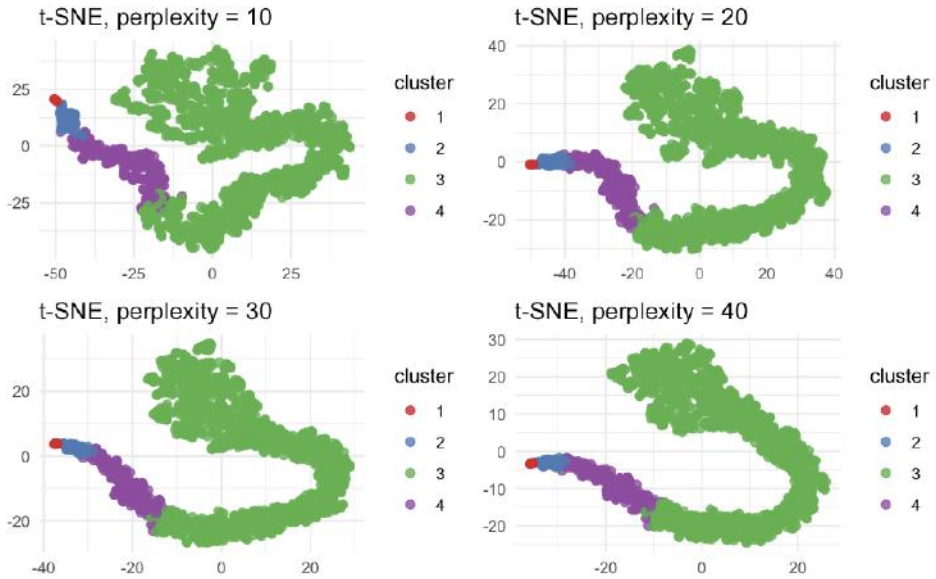

Figure 3: Traditional t -SNE plots under different perplexity values ($k = 4$).

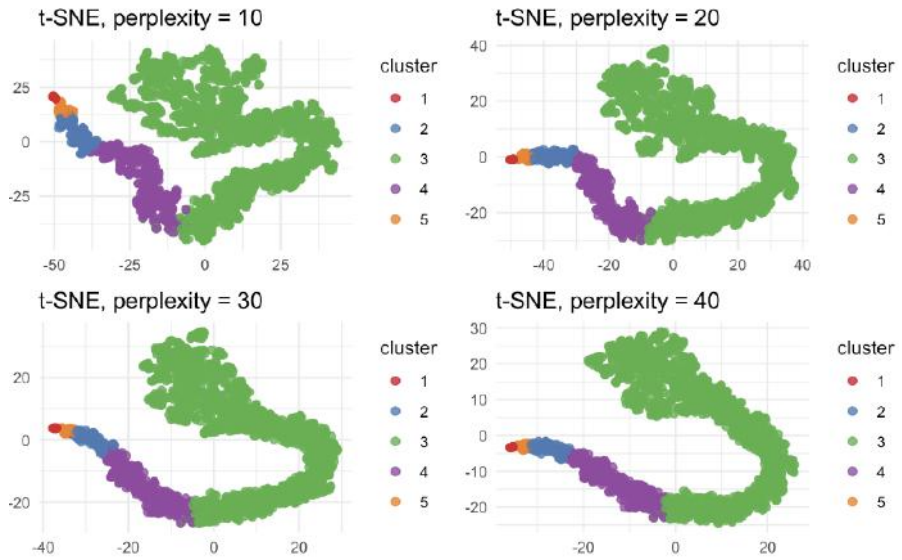

Figure 4: Traditional t-SNE plots under different perplexity values ($k = 5$).

embedded space, but is a further refinement of the dominant structure. As a result, the overall clustering results are still highly uneven in the distribution of samples, and the spatial differentiation of each cluster is limited.

Overall, across k = 3, 4, 5, the results obtained from clustering directly based on the t-SNE embedding generally show obvious cluster size imbalances, that is, there are clusters with extremely small sample proportions and large clusters that dominate at the same time. In addition, clusters are often divided along continuous structures instead of forming groups that are clearly separated in space. These phenomena show that in the current research context of high-dimensional and small samples, it is difficult to provide sufficient structured and stable input for the k-means clusters by relying only on the low-dimensional representation obtained by the t-SNE, thus limiting the interpretability of clustering results.

3.4. Clustering Based on PCA t-SNE

When PCA preprocessing is introduced prior to the t-SNE, the resulting embeddings and clustering outcomes exhibit noticeably different structures compared with those obtained from direct t-SNE embeddings. Figures **5**-**7** shows the clustering results based on PCA-enhanced the t-SNE representations for k = 3, 4, 5, where the k-means is applied to the

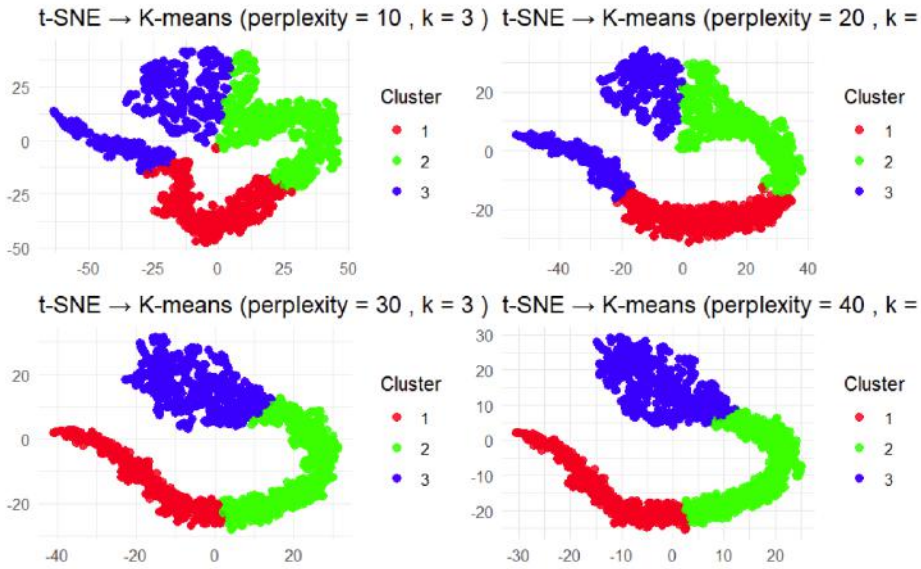

Figure 5: PCA-t-SNE plots under different perplexity values ($k = 3$).

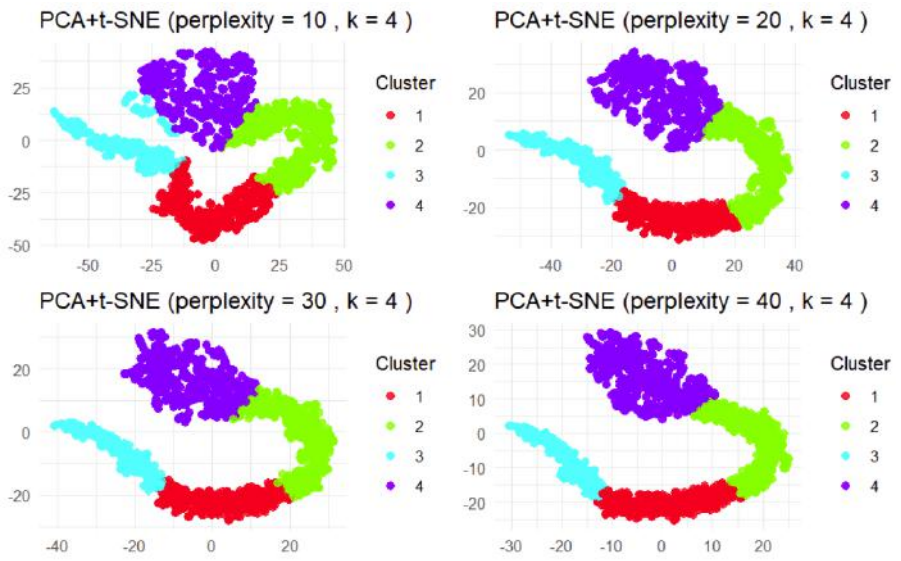

Figure 6: PCA-t-SNE plots under different perplexity values ($k=4$).

dimension-reduced data from PCA with reduced dimension $p_r = 50$ and coefficient of variation explanation [2] 100% from the selected principal components.

For k =3 in Figure **5**, the PCA t-SNE yields three clusters that are more evenly populated than those obtained from the t-SNE alone. Each cluster occupies a distinct region of the embedding, and none of the clusters is reduced to only a few isolated points. Although the overall structure still reflects a curved geometry, the partition no longer corresponds to separating a small set of peripheral samples from a single dominant group. Instead, the three clusters are supported by substantial numbers of observations, which improves the interpretability of the clustering result.

For k =4 in Figure **6**, the PCA t-SNE plots display clearer separation among clusters. The four clusters are distributed more uniformly across the embedding space, with reduced overlap at their boundaries. Compared with the t-SNE-alone clustering, the subdivision of the data is less driven by local distortions in the low-dimensional representation and more closely aligned with visually distinct regions. As a result, the clusters appear more coherent and balanced, facilitating a clearer interpretation of the four-group structure.

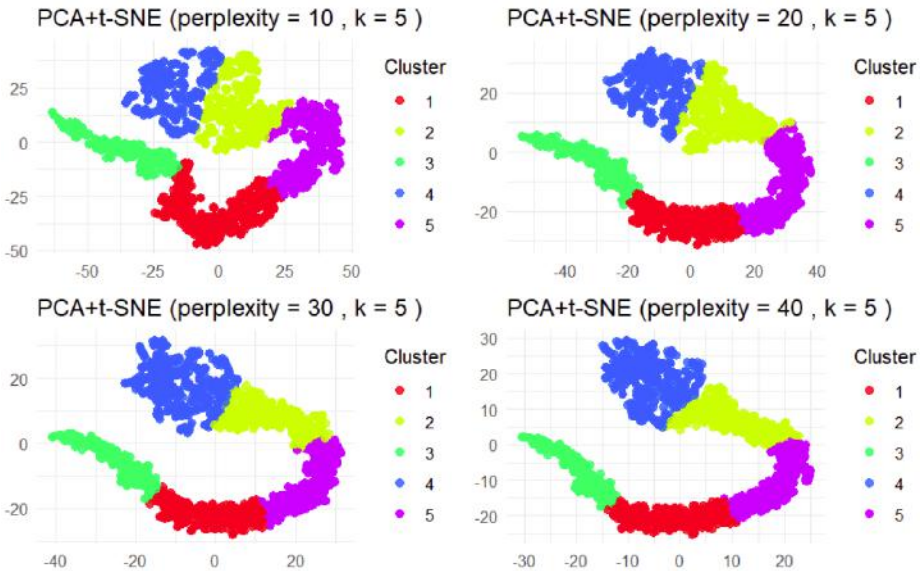

Figure 7: PCA-t-SNE plots under different perplexity values ($k=5$).

When k =5 in Figure **7**, the PCA t-SNE representation provides the most structured clustering outcome. The five clusters are well separated in the embedding, with relatively homogeneous cluster sizes and clearly defined spatial regions. Unlike the corresponding t-SNE-alone results, no cluster is dominated by only a handful of points, and the partition does not arise from arbitrary slicing of a continuous manifold. Instead, the PCA t-SNE plots display multiple distinct groups that are consistently identifiable in the two-dimensional space.

Overall, the results demonstrate that incorporating PCA as a preprocessing step leads to more balanced and interpretable clustering outcomes when combined with the t-SNE. Across the three candidate values of k, the PCA-enhanced approach consistently mitigates the severe cluster-size imbalance and ambiguous boundaries observed in the t-SNE-alone results. Among the configurations considered, the clustering with k =5 (Figure **7**) provides the clearest separation and the most coherent cluster structure, suggesting that this choice offers the most informative representation for the present dataset.

4. STATISTICAL VALIDATION OF CLUSTERING

4.1. The Projected F -Test

Let $\in \mathbb{R}^{n\times p}$ denote the centered gene expression matrix, where n is the number of samples and p is the number of genes. The objective is to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \quad (1)$$

against the alternative that at least one cluster mean differs. In high-dimensional settings where $p \gg n$, classical multivariate analysis of variance (MANOVA) is not applicable due to the singularity of the sample covariance matrix. To address this issue, we adopt the projected F-test [9], which conducts inference after projecting the data onto low-dimensional subspaces.

For any given projection dimension $r < \min\{n,p\}$, implementing the transformation in Theorem 1 of Cao and Liang (2025, [9]), we carry out the same projected F-tests as in Table **1** of Cao and Liang (2025, [9]). The projected F-statistic has an F-distribution $F(r, n-1-r)$ under the null hypothesis (1). Instead of relying on a single projection dimension, the projected F-test will be evaluated against a series of projection dimensions suitable for the scale of the problem. Let $r_{\max} = \min\{n-1, p\} - 1$ denote the maximum admissible projection dimension (Theorem 1 in [9]). In this study, four representative projection dimensions are defined as

$$r_1 = \left\lfloor \frac{r_{\max}}{4} \right\rfloor, \quad r_2 = \left\lfloor \frac{r_{\max}}{3} \right\rfloor, \quad r_3 = \left\lfloor \frac{r_{\max}}{2} \right\rfloor, \quad r_4 = \left\lfloor \frac{3r_{\max}}{4} \right\rfloor,$$

where $n = 29$, $p = 22{,}447$, $r_{\max} = \min\{n-1, p\} - 1 = 28$, the notation $[x]$ = the largest integer not exceeding the real number x, for example, $[2.1] = [2.9] = 2$. These values span low to moderately high projection dimensions while remaining well below the sample size. By examining statistical evidence across multiple projection levels, we assess whether cluster separation is stable with respect to the choice of projection dimension, rather than being driven by a particular value of r.

In practice, the projected F-test is applied separately for each value of r, and the resulting test statistics and p-values are summarized across projection dimensions. Consistent significance across multiple values of r is interpreted as stronger statistical evidence that the clusters identified through low-dimensional visualization correspond to distinct mean structures in the original high-dimensional gene expression space.

4.2. Testing Results under Different Cluster Configurations

The projected F-test results for clusters obtained from the traditional t-SNE pipeline under different cluster configurations are summarized in Table **1**. When the number of clusters varies from $k = 3$ to $k = 5$, the projected F-statistics remain relatively small for most projection dimensions, and the corresponding p-values are generally large or only marginal. This indicates that, under the direct t-SNE clustering, the inferred cluster structures provide limited statistical evidence for between-cluster mean differences in the original high-dimensional gene expression space. Although statistically significant results can be observed in more favorable configurations, the inferential support is not stable across different values of k.

The corresponding results for the PCA t-SNE pipeline are reported in Table **2**, where the projected F-test results for k =3, 4 5 are presented in a unified framework. In contrast to the traditional t-SNE approach, the PCA-enhanced pipeline consistently yields larger projected F-statistics and markedly smaller p-values across all cluster configurations and projection dimensions. Notably, even for smaller values of k, where the direct t-SNE fails to provide

Table 1: Projected F-Test for MANOVA of Clusters from the Traditional t-SNE

Projection dimension	$k=3$	$k=4$	$k=5$
$r_1=7,\ F(7,21)$	$F=240.28, p=.00$	$F=75.59, p=.00$	$F=35.21, p=.00$
$r_2=9, F(9,19)$	$F=197.35, p=.00$	$F=59.51, p=.00$	$F=27.90, p=.00$
$r_3=14, F(14,14)$	$F=150.80, p=.00$	$F=40.68, p=.00$	$F=21.04, p=.00$
$r_4=21, F(21,7)$	$F=105.26, p=.00$	$F=27.61, p=.00$	$F=15.33, p=.00$

Table 2: Projected F-Test for MANOVA of Clusters from PCA t-SNE

Projection dimension	$k=3$	$k=4$	$k=5$
$r_1=7,\ F(7,21)$	$F=110.63, p=.00$	$F=1.71, p=.10$	$F=2.70, p=.00$
$r_2=9, F(9,19)$	$F=92.80, p=.00$	$F=1.74, p=.07$	$F=2.63, p=.00$
$r_3=14, F(14,14)$	$F=72.83, p=.00$	$F=1.81, p=.03$	$F=2.44, p=.00$
$r_4=21, F(21,7)$	$F=50.28, p=.00$	$F=1.73, p=.02$	$F=2.25, p=.00$

convincing statistical evidence, the PCA-enhanced method produces strong and consistent rejection of the null hypothesis of equal cluster means. Overall, the comparison between Tables **1** and **2** demonstrates that, although both methods can produce statistically significant p-values under certain parameter choices, the inferential support provided by the PCA t-SNE framework is systematically stronger and more robust across different cluster configurations. By stabilizing the projected F-test results over a range of k values, the incorporation of principal component analysis provides a more reliable statistical basis for cluster analysis in high-dimensional gene expression data.

It should be pointed out that the projected F-tests in Tables **1-2** only provide the individual p-value for each selected projection dimension. Because the projected F-tests are not independent, the determination of the overall type I error rate and the effect sizes from all projected F-tests are challenging tasks involved in multiple comparison procedures. We are not able to further this research direction in this paper but would like to refer interesting readers to the books [26, 27].

5. CONCLUDING REMARKS

This paper investigated a PCA-enhanced t-SNE framework for the clustering and visualization of high-dimensional, low-sample-size gene expression data. While t-SNE is a powerful tool for exploratory analysis, its direct application to such data often yields unstable embeddings and ambiguous clusters. To address this, we proposed a preprocessing step using Principal Component Analysis (PCA) to capture dominant global structures and mitigate the influence of noise and redundancy prior to t-SNE. Our empirical analysis demonstrates that this two-step approach substantially refines the resulting visualizations. Compared to standard t-SNE, the PCA-enhanced method produces embeddings with clearer cluster separation, more balanced cluster sizes, and greater interpretability, particularly as the number of presumed clusters increases. This finding aligns with established guidance on preparing high-dimensional data for nonlinear dimensionality reduction. Beyond visual assessment, we introduced a formal statistical evaluation using a projection-based F-test to validate cluster separability directly in the original high-dimensional space. This test provided stable and meaningful statistical evidence for the identified clusters, whereas classical MANOVA based on Wilks-Λ proved less informative due to numerical instability in our high-dimensional setting.

In summary, our work supports the integration of PCA with t-SNE as a coherent and statistically grounded workflow for biomedical data exploration. This framework effectively bridges intuitive low-dimensional visualization with rigorous inference in the original feature space, offering a reliable strategy for pattern discovery in genomics and related fields.

Looking ahead, future research should expand this comparative analysis across a broader range of datasets and alternative methodologies. A systematic

comparison with other nonlinear techniques, such as UMAP [25], combined with diverse clustering algorithms, would further elucidate the relative strengths and optimal applications of the PCA-*t*-SNE pipeline. Additionally, extending the statistical validation framework to include resampling-based methods and other robust inference procedures would deepen our understanding of cluster stability and reproducibility in complex, high-dimensional biological data.

REFERENCES

[1] Yoshida K, Toyoizumi T. A biological model of nonlinear dimensionality reduction. Science Advances 2025; 11(6). https://www.science.org/doi/10.1126/sciadv.adp9048

[2] Jolliffe IT. Principal component analysis. Springer series in statistics. New York: Springer-Verlag; 2002. https://doi.org/10.1007/b98835

[3] Ringner M. What is principal component analysis? Nature Biotechnology 2008; 26(3): 303-304. https://doi.org/10.1038/nbt0308-303

[4] Islam MR, Shatabda S. FeatPCA: A feature subspace based principal component analysis technique for enhancing clustering of single-cell RNA-seq data 2025: https://arxiv.org/abs/2502.05647

[5] van der Maaten L, Hinton G. Visualizing data using *t*-SNE. Journal of Machine Learning Research 2008; 9: 2579-2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

[6] Ketchen DJ, Shook CL. The application of cluster analysis in strategic management research: an analysis and critique. Strategic Management Journal 1996; 17(6): 441-458.

[7] Kobak D, Berens P. The art of using *t*-SNE for single-cell transcriptomics. Nature Communications 2019 Nov 28. https://www.nature.com/articles/s41467-019-13056-x

[8] Yousuff M, Babu R, Anand Rathinam R. Nonlinear dimensionality reduction based visualization of single-cell RNA sequencing data. Journal of Analytical Science and Technology, 2024; 15(1). https://doi.org/10.1186/s40543-023-00414-0

[9] Cao Y, Liang J. Multiple mean comparison for clusters of gene expression data through the t-SNE plot and PCA dimension reduction. International Journal of Statistics in Medical Research 2025; 14: 1-14. https://doi.org/10.6000/1929-6029.2025.14.01

[10] Tsuyuzaki K, Sato H, Sato K, Nikaido I. Benchmarking principal component analysis for large-scale single-cell RNA-sequencing. Genome Biology 2020; 21(1): 9. https://doi.org/10.1186/s13059-019-1900-3

[11] GeeksforGeeks. Principal component analysis (PCA) [Internet] 2018 Jul 7. https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/

[12] Giraud C. Introduction to high-dimensional statistics. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2021. https://doi.org/10.1201/9781003158745

[13] Peter BM. A geometric relationship of F2, F3 and F4-statistics with principal component analysis. Philosophical Transactions of the Royal Society B: Biological Sciences 2022; 377(1852): 20200413. https://doi.org/10.1098/rstb.2020.0413

[14] Rychlik T. Projecting statistical functionals. Vol. 160. New York: Springer Science+Business Media; 2012.

[15] Wattenberg M, ViÃ©gas F, Johnson I. How to Use *t*-SNE Effectively. Distill [Internet] 2016; 1(10). https://distill.pub/2016/misread-tsne/

[16] Bibliography1.Arora S, Hu W, Kothari PK. An Analysis of the *t*-SNE Algorithm for Data Visualization. PMLR [Internet] 2018; 1455-62. https://proceedings.mlr.press/v75/arora18a.html

[17] Linderman GC, Steinerberger S. Clustering with *t*-SNE, provably. SIAM Journal on Mathematics of Data Science 2019; 1(2): 313-32. https://pubmed.ncbi.nlm.nih.gov/33073204/

[18] Gisbrecht A, Schulz A, Hammer B. Parametric nonlinear dimensionality reduction using kernel *t*-SNE. Neurocomputing 2015 ; 147: 71-82.

[19] Li W, Cerise JE, Yang Y, Han H. Application of *t*-SNE to human genetic data. Journal of Bioinformatics and Computational Biology 2017; 15(04): 1750017.

[20] Johnstone IM, Titterington DM. Statistical challenges of high-dimensional data. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2009; 367(1906): 4237-53.

[21] Assent I. Clustering high dimensional data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2012; 2(4): 340-50.

[22] Boulesteix AL ., Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics 2006; 8(1): 32-44.

[23] Tenenbaum JB, Silva VD, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000; 290: 2319-2323.

[24] Roweis ST. Nonlinear dimensionality reduction by locally linear embedding. Science 2000; 290(5500): 2323-6.

[25] McInnes L, Healy J, Saul N, Grobberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software 2018; 3(29): 861.

[26] Borenstein M (Ed.), Meta-analysis: A guide to calibrating and combining statistical evidence. Wiley 2024.

[27] Westfall PH, Young SS, Resampling-based multiple testing: Examples and methods for p-value adjustment. John Wiley & Sons 1993.

Received on 06-11-2025 Accepted on 10-12-2025 Published on 26-12-2025

https://doi.org/10.6000/1929-6029.2025.14.76