MSc Data Science & Information Technologies

Bioinformatics - Biomedical Data Science Specialization

Course: Machine Learning in Computational Biology

Professor: Elias Manolakos

Enhanced Analysis of Gastric Cancer Metastasis Using Clustering Techniques and Single-Cell RNA Sequencing

Final Project Report

Authors	*Student ID*
Konstantinos Giatras	7115152300005
Olympia Tsiomou	7115152200035

DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens
EST. 1837

giatraskonstantinos99@gmail.com

o.tsiomou@gmail.com

2023-2024

Abstract

Inspired by a notable study on single-cell RNA sequencing (scRNA-seq) for gastric cancer, our project applies a comprehensive computational framework to delve deeper into the cellular complexities of the disease. We utilized dimensionality reduction techniques such as PCA, t-SNE, and UMAP, combined with various clustering algorithms, to explore cellular heterogeneity across tumor samples. Our methodologies facilitated the identification of marker genes and allowed us to conduct Gene Ontology (GO) analysis, providing new insights into the biological functions and pathways prevalent in distinct cell populations. These efforts underscore the robust capabilities of scRNA-seq to enhance our understanding of gastric cancer's cellular landscape, offering a nuanced perspective that aligns with and expands upon previous findings.

Introduction

Gastric cancer (GC) is the fifth most common cancer globally and ranks as one of the leading causes of cancer-related deaths, characterized by a notably poor five-year survival rate despite aggressive treatments. The inherent complexity of gastric cancer, particularly its propensity for lymphatic metastasis, underscores the urgent need for advanced diagnostic and therapeutic strategies. Single-cell RNA sequencing (scRNA-seq) has emerged as a crucial tool, offering unprecedented insights into the cellular heterogeneity of tumors that are often obscured in bulk analyses [1]. The study by Wang et al., which serves as the foundation for our project, leverages scRNA-seq to delve into the intratumoral heterogeneity of gastric cancer and identify cellular subpopulations linked to metastatic potential.

Wang et al.'s study provides a comprehensive exploration of primary tumors and paired metastatic lymph node tissues from GC patients. The authors effectively highlight potential therapeutic targets by mapping the scRNA-seq data to known cancer pathways. Their significant findings include the identification of marker genes such as ERBB2, CLDN11, and CDK12, and evolutionary drivers like FOS and JUN, which offer valuable insights into the mechanisms driving lymph node metastasis in GC [1]. This work not only emphasizes the distinct carcinoma profiles and diverse microenvironmental subsets but also underscores the transitional states of cancer during metastasis, providing a new perspective on potential targets for precision therapy.

The importance of scRNA-seq in revealing tumor heterogeneity and immune profiling has been further corroborated by related works in the field. For instance, Zhao et al. (2021) [2] utilized multiplexed drug perturbation combined with scRNA-seq to identify cell type-specific drug responses in glioblastoma, enhancing personalized medicine approaches in oncology. Similarly, the integration of scRNA-seq with machine learning, as discussed by Qi and Zou (2023) [3], has significantly improved our capacity to predict cellular responses to immune therapies and drug treatments, thus refining targeted treatment strategies.

Challenges and Workflow

Our motivation to replicate and extend the study by Wang et al. stems from the critical clinical challenges posed by gastric cancer and the potential to uncover deeper insights into its cellular diversity using advanced analytical techniques.

The process of replicating and extending the study by Bin Wang et al. on gastric cancer using single-cell

RNA sequencing presented several challenges. A primary difficulty was the lack of code availability from the original authors, necessitating that we develop our analysis procedures from scratch. This situation required a careful interpretation and implementation of the described methodologies without a reference point, significantly complicating our task but also providing an opportunity for creative problem-solving.

Our initial proposal aimed to replicate the original study's methodologies, including quality control (QC), filtering, and analysis, using the dataset provided by the authors. Instead of R, which was utilized in the original study, we decided to conduct our analyses in Python. We used packages such as Scanpy [4], which allowed us to adapt and expand the computational tools more suited to our expertise.

In practice, we not only replicated the key aspects of the original study, but also introduced novel analytical improvements. Our contributions include a comprehensive methodological framework for scRNA-seq data analysis in gastric cancer, validated through robust comparisons of several dimensionality reduction techniques, including PCA [5], t-SNE [6], and UMAP [7] for data visualization and clustering algorithms, such as Gaussian Mixture Models (GMM), Average Link, Ward, Spectral, Louvain, and Leiden, which provided a broader view of the data structure and helped in identifying marker genes effectively. This comprehensive approach helped us address another major challenge: ensuring the explainability and validity of our results. By comparing different methodologies, we could validate our findings robustly and provide a reasoned justification for our analytical choices. An overview of our pipeline can be seen in Figure 11.

Authors' Pipeline and Results

The authors of the original study employed a pipeline to analyze single-cell RNA sequencing (scRNA-seq) data from gastric cancer (GC) samples, including both primary tumors and paired metastatic lymph nodes. Their workflow is illustrated in Figure 12), which depicts the authors' pipeline as presented in their paper, outlining each stage of the process.

Their computational methods began with data preprocessing, where low-quality reads were removed using Trimmomatic [8], and high-quality reads were mapped to the human genome (UCSC hg19) with HiSat2 [9]. FeatureCounts [10] was then used to calculate gene expression levels, and scater (an R package) [11] was utilized to normalize read counts to transcripts per million (TPM) and perform log2 transformation.

For dimensionality reduction, the authors applied Principal Component Analysis (PCA) to capture the most significant features of the dataset, followed by t-Distributed Stochastic Neighbor Embedding (t-SNE) for visualization, which highlighted the separation between primary and metastatic tumor cell populations (Figure 13). In the clustering stage, hierarchical clustering was employed to identify distinct cell clusters based on gene expression profiles. Seurat [12] analysis further identified four main clusters within the tumor tissues, each representing distinct cell subpopulations (Figure 14a).

The results of this workflow included the identification of significant intratumoral heterogeneity, with distinct carcinoma profiles and varied gene expression patterns across the identified clusters. Differentially expressed marker genes, such as ERBB2, CLDN11, and CDK12, were found in metastatic cancer, and trajectory analysis revealed evolutionary paths of cancer cells, emphasizing key regulatory genes like FOS and JUN (Figure 14b). Functional annotation through Gene Ontology (GO) analysis (Figure 14c-f) and pseudotime trajectory analysis with TSCAN [13] and Monocle [14] provided deeper insights into the biological processes and pathways involved in gastric cancer metastasis.

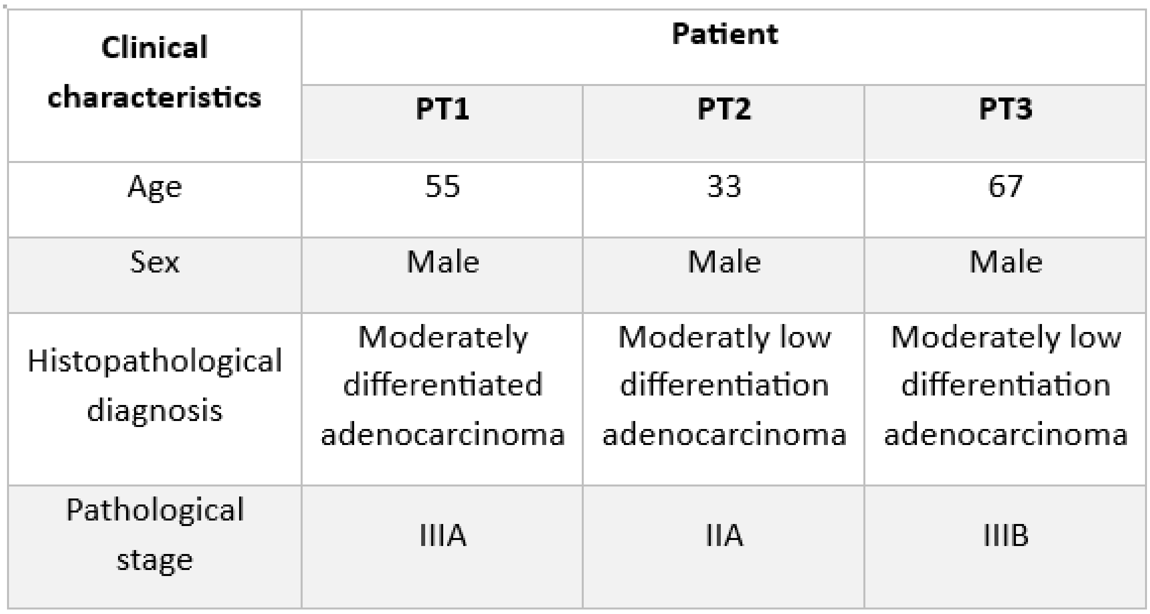
The rest of this report is organized as follows: In the next section, we provide a detailed explanation of our methodology, including preprocessing, normalization, and advanced clustering techniques. We then present our results, comparing them with the findings of Wang et al., and discuss the implications of our enhanced analytical approach. Finally, we conclude with a summary of our contributions and potential directions for future research in the field of gastric cancer.

Materials and Methods

Dataset Description

Our study utilized a dataset consisting of single-cell RNA sequencing (scRNA-seq) data derived from primary gastric cancer tissues and paired metastatic lymph node tissues from three patients. The dataset comprises a counts matrix of 21,196 genes (rows) and 94 single-cell samples (columns), which includes 65 samples from primary tumor (TT) sites and 29 from metastatic lymph node (LN) sites. Each column in the dataset represents the expression levels of each gene in individual single-cell samples, uniquely identified, such as GC1-TT1, GC1-LN1, GC2-TT1, and GC3-LN12. The raw data of this scRNA-seq study have been deposited and are accessible via the GEO database under the accession code GSE158631. This comprehensive dataset, which is available in CSV format, serves as the basis for our analysis, providing a detailed view of the gene expression patterns across different stages and conditions of gastric cancer.

Figure 1 below provides additional clinical characteristics of the patients from whom the samples were derived:



Clinical characteristics	Patient		
	PT1	PT2	PT3
Age	55	33	67
Sex	Male	Male	Male
Histopathological diagnosis	Moderately differentiated adenocarcinoma	Moderatly low differentiation adenocarcinoma	Moderately low differentiation adenocarcinoma
Pathological stage	IIIA	IIA	IIIB

Figure 1: Clinical characteristics of patients [1].

Data Preprocessing

Data Loading and Quality Control

The initial step in our computational analysis was loading the raw counts data from a CSV file into a structured format using our custom ‘GCSingleCellAnalysis‘ class, which provides a comprehensive workflow for preprocessing, dimensionality reduction, clustering, and functional annotation of single-cell RNA-seq data. This process involved converting the raw matrix into an AnnData object, a format particularly suited for handling complex annotated data matrices in bioinformatics. This structured data format integrates seamlessly with the Scanpy library [4], which was used extensively in our analysis.

Our preprocessing approach included a thorough examination of the data to identify mitochondrial, ribosomal, and hemoglobin genes. Mitochondrial genes are often used as indicators of cell health, as high mitochondrial activity can suggest cell stress or damage. As such, cells exhibiting high mitochondrial gene expression are typically filtered out to enhance data quality [15]. Ribosomal genes, if expressed highly, can indicate either technical artifacts or highly active cells; thus, monitoring and potentially filtering out cells with excessive ribosomal gene expression helps focus analyses on biologically relevant cell subpopulations. This is crucial to avoid confounding factors in downstream analyses such as clustering or differential gene expression [16]. Hemoglobin genes, on the other hand, may serve as prognostic markers in gastric cancer, potentially identifying subpopulations relevant to anemia-related clinical outcomes; therefore identifying these latter genes is crucial for our analysis [17] [18]. We categorized these genes to compute specific quality control metrics, which are crucial for assessing the integrity of our sequencing data. We also applied logarithmic transformations to stabilize variance and normalize distributions, making the data more suitable for sensitive analytical methods that followed. As shown in Figure 2, the top 20 most highly expressed genes after preprocessing highlight the significant presence of hemoglobin genes (HBA2, HBB, and HBA1).

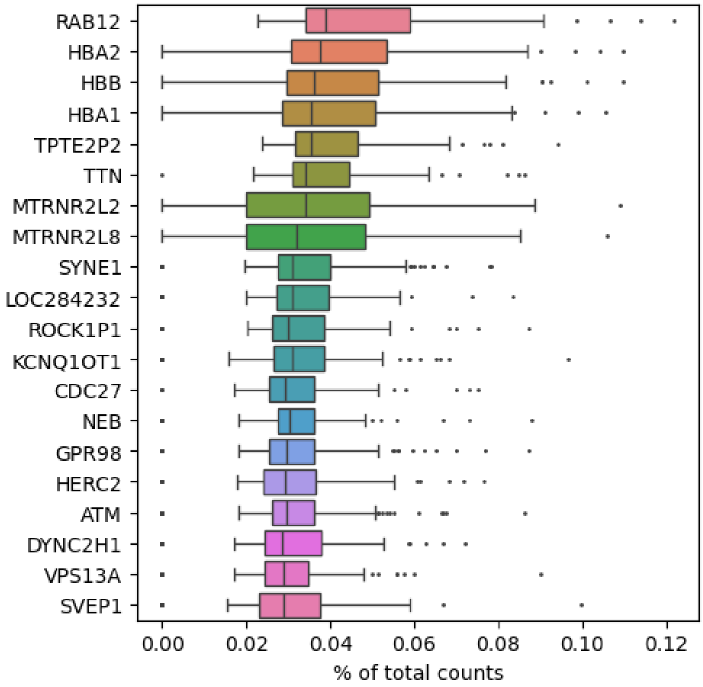


Figure 2: Top 20 most highly expressed genes of the preprocessed dataset.

Normalization and Filtering

Normalization was crucial to make the expression levels across cells comparable, thereby reducing biases due to differences in sequencing depth among samples. We followed the normalization procedure recommended by the original study's authors, transforming raw count data into Transcripts Per Million (TPM) and subsequently applying a log2 transformation.

After normalization, we applied stringent filtering criteria to ensure the retention of biologically significant signals. We removed genes that did not meet a minimum expression threshold of more than one average read count per gene across all cells. This filtering reduced the number of genes from 21,196 to 11,500 that showed significant expression levels. Additionally, we implemented filters based on the proportion of mitochondrial and ribosomal genes, but the number of cells remained consistent at 94 after these filters were applied.

Final Data Preparation

As the final step in preparing our dataset for in-depth analysis, we organized the data by extracting and annotating significant metadata from the cell identifiers. We extracted patient IDs and tissue types and combined these into a new column that detailed each cell's origin and type. This organization was crucial for subsequent analyses, allowing for precise stratification and comparative studies across different cell types and conditions within gastric cancer.

These detailed preprocessing steps established a robust foundation for the intricate explorations into cellular heterogeneity and the molecular dynamics of gastric cancer, discussed in further sections of this report.

Dimentionality Reduction and Visualization Approach

This section details the computational methods employed to reduce the dimensionality of the high-dimensional scRNA-seq dataset, which aids in the visualization and further analysis of the data. Our analytical approach leverages Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP).

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [5] was utilized to reduce the dimensionality of our preprocessed scRNA-seq dataset, focusing on the preservation of significant sources of variance across the cells. Using the Scanpy library's ‘sc.tl.pca‘ function, we specified the ARPACK solver for efficient computation. This method transforms high-dimensional data into a set of linearly uncorrelated variables known as principal components. The number of principal components retained was based on the dataset's characteristics, with a typical initial consideration of up to 50 components. However, this number was adjusted according to the dataset's constraints (the minimum of observations or variables minus one) to 85 components. The variance explained by each component was calculated to determine their significance.

t-distributed Stochastic Neighbor Embedding (t-SNE)

t-SNE [6], implemented via the Scikit-learn library, was applied to visualize high-dimensional data by reducing it to two-dimensional space. This non-linear dimensionality reduction technique is particularly well-suited for the visualization of large datasets like ours. We provided a standard parameter set for initial runs (perplexity: 30, learning rate: 200, iterations: 5000) and included an option to optimize these parameters based on the trustworthiness metric, which assesses how well the local structure is preserved in the reduced dimensionality.

Uniform Manifold Approximation and Projection (UMAP)

Similar to t-SNE, UMAP [7] was employed to project the dataset into a two-dimensional space, facilitating an alternative visualization that might preserve more of the global structure compared to t-SNE. UMAP's implementation from the 'umap-learn' library was utilized, with standard parameters initially set (number of neighbors: 15, minimum distance: 0.1, metric: 'euclidean'). The technique allows for optional optimization of these parameters to potentially enhance the trustworthiness of the embeddings.

The specific results and insights derived from the application of these methods are detailed in the subsequent "Results and Discussion" section of our report.

Cell Categorization through Clustering

Our clustering analysis aimed to categorize single cells from gastric cancer tissues into distinct groups based on their gene expression profiles, facilitating the understanding of cell heterogeneity and potential subtypes. We employed several clustering methods, each suited to different assumptions about data structure and distribution.

Gaussian Mixture Model (GMM)

GMM assumes data points are generated from a mixture of several Gaussian distributions. We used the 'GaussianMixture' class from the sklearn.mixture module, which provides soft-clustering capabilities, assigning probabilities of each point belonging to particular clusters. This approach is beneficial for handling uncertainties in cluster assignments effectively.

Hierarchical Clustering

We implemented two hierarchical clustering methods:

- **Average Linkage**: Uses the average of the distances between all observations of two sets. Implemented via 'AgglomerativeClustering' with linkage='average' from sklearn.cluster.
- **Ward's Method**: Aims to minimize the total within-cluster variance. It is executed using 'AgglomerativeClustering' with linkage='ward', ensuring minimal increase in total within-cluster variance after merging.

Spectral Clustering

Spectral Clustering leverages the eigenvalues of the similarity matrix of the data to perform dimensionality reduction before clustering. The method constructs an affinity matrix from the nearest neighbor graph and maps points onto a lower-dimensional space. We used the 'SpectralClustering' function from sklearn.cluster, specifying the number of clusters and using the nearest neighbors' affinity.

Louvain and Leiden Algorithms

These community detection algorithms are designed to find high-resolution clusters in large datasets by optimizing modularity. They are particularly useful for large datasets like ours. We used these methods via the Scanpy library, initiating with ‘sc.pp.neighbors‘ to create the neighbor graph followed by ‘sc.tl.louvain‘ or ‘sc.tl.leiden‘ for the clustering.

Optimization and Evaluation

Our methodology included an optional optimization feature to determine the ideal number of clusters, guided by the silhouette score. This metric evaluates the separation between clusters, helping identify the optimal cluster count that maximizes this separation. We tested cluster counts ranging from 2 to 10 for each method and recorded silhouette scores to determine the best configuration for our data.

We assessed the clustering quality using several metrics:

- **Silhouette Score**: Assesses how similar an object is to its own cluster compared to other clusters.
- **Calinski-Harabasz Index**: Measures the ratio of the sum of between-clusters dispersion to within-cluster dispersion.
- **Davies-Bouldin Index**: Indicates the average ’similarity’ between each cluster and its most similar cluster, where lower values signify better clustering.

Implementation Details

Our clustering algorithms were applied to data transformed via PCA, t-SNE, and UMAP embeddings, allowing us to leverage noise-reduced and dimensionality-reduced datasets for more reliable clustering outcomes. This approach ensured flexibility and robustness in our analysis framework by allowing choices in optimizing clustering parameters and selecting embeddings based on specific research needs.

This detailed methodological framework underpins our approach to understanding the intricate biological variations in gastric cancer tissues, preparing the ground for further biological insights in the "Results and Discussion" section.

Post-Clustering Analysis

Marker Gene Identification

For the identification of marker genes from the clustered data, we employed the Scanpy Python library, specifically utilizing its ‘rank_genes_groups‘ function. Our methodological approach was structured to identify genes that exhibited statistically significant differences in expression across the clusters defined by our optimal clustering results.

We selected the Wilcoxon rank-sum test (’wilcoxon’ method) to rank genes due to its effectiveness in handling non-normally distributed data, which is typical in single-cell RNA sequencing analyses. For each cluster, we determined the top 10 highly expressed marker genes based on their differential expression levels. These genes were then saved to a CSV file for further analysis and visualization.

The heatmap of the top 10 highly expressed marker genes for each cluster was generated using the ‘rank_genes_groups_heatmap‘ function of Scanpy. This visualization helped us understand the expression pattern across different clusters, emphasizing genes that were distinctly expressed in one cluster compared to others.

Functional annotation of clusters using Gene Ontology (GO)

For the GO analysis, we integrated the g:Profiler web tool to fetch and analyze GO annotations associated with the top marker genes identified from the optimal clustering. The g:Profiler tool was chosen due to its comprehensive database and the ability to handle high-throughput gene lists for functional enrichment analysis.

Each list of top marker genes was submitted to g:Profiler to retrieve relevant GO terms that describe biological processes, cellular components, and molecular functions associated with these genes [19]. We focused on annotations that reached statistical significance, which helps in understanding the biological implications of our clustering results. The annotations were filtered by significance and relevance to the study's context, with a particular focus on pathways and processes that are potentially implicated in gastric cancer.

Results and Discussion

This section outlines the key findings from our analysis of the scRNA-seq dataset derived from gastric cancer cells. We explore the inherent heterogeneity within the dataset, crucial for a deeper understanding of the biological variations among single cells in gastric cancer. The results from our dimensionality reduction and clustering analyses are presented to illustrate the structural and phenotypic diversity captured within the data.

Dimensionality Reduction

Principal Component Analysis (PCA)

The pairwise PCA plots, visualized in Figure 3, illustrate a broad distribution of data points across the first three principal components, with no obvious grouping into distinct clusters. The plots display a mix of samples from three patients, differentiated by tissue type; tumor tissue (TT) and lymph node (LN). The colors, representing six different categories (Patient 1 TT, Patient 1 LN, Patient 2 TT, Patient 2 LN, Patient 3 TT, Patient 3 LN), indicate that the samples are interspersed, reflecting significant variability within and between these groups. This overlap suggests that the gene expression profiles are highly heterogeneous, not clearly separating by patient or tissue type. The PCA results, therefore, do not reveal distinct subgroups, implying complex biological dynamics that might require alternative analytical approaches.

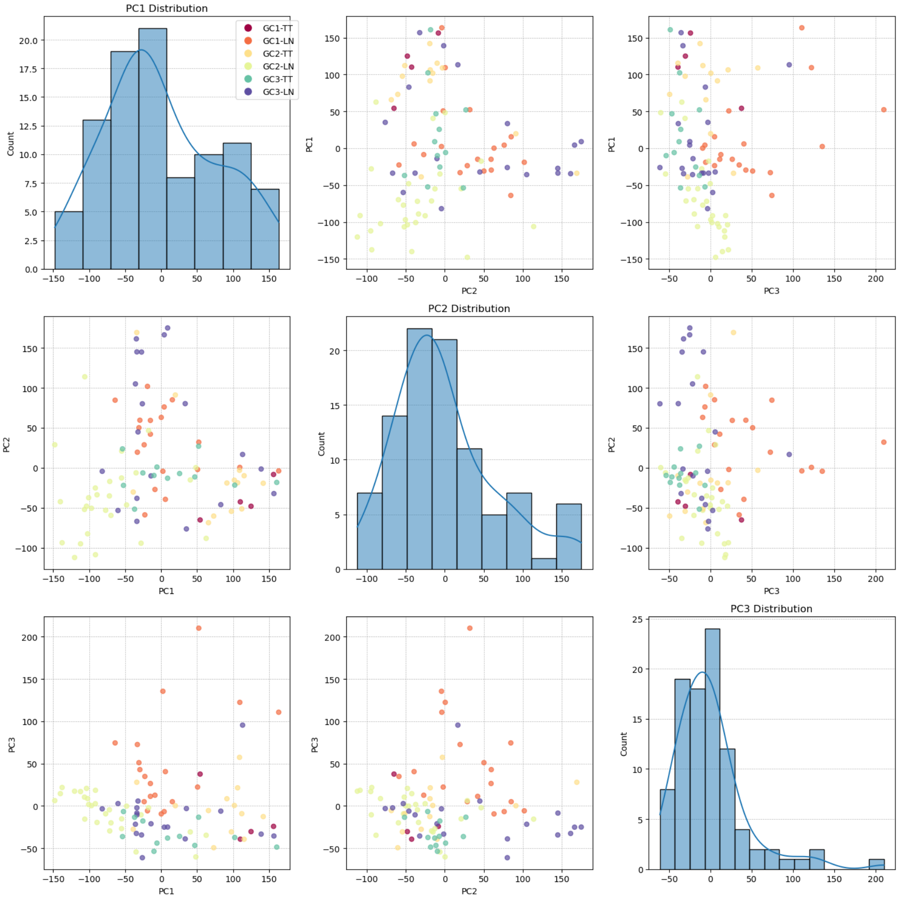


Figure 3: Pairwise PCA distribution of gastric cancer single-cell RNA-seq data across different patients and tissue types.

Moreover, the cumulative variance analysis indicated that a substantial number of principal components, specifically 81, were required to capture at least 90% of the variance within the data, as shown in Figure 4. This suggests the complexity and high dimensionality inherent to the dataset, justifying the use of PCA to reduce noise and focus on the most informative features for subsequent analysis.

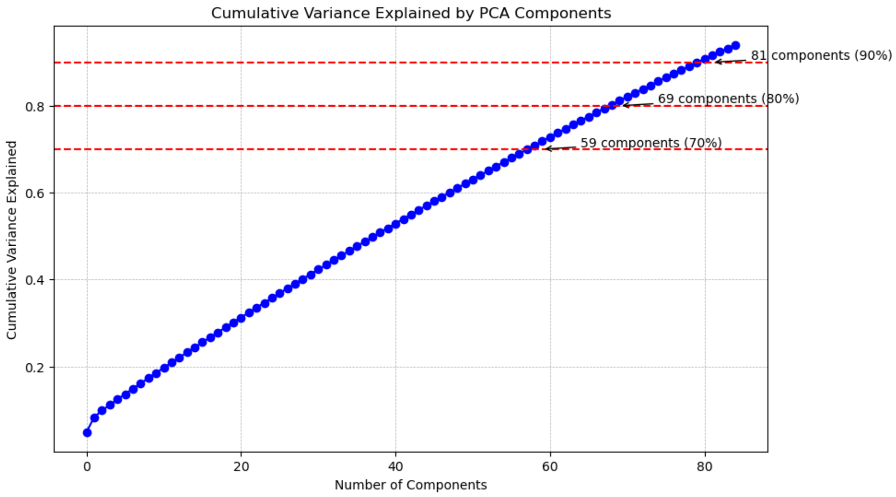


Figure 4: Cumulative Variance explained by PCA components.

t-distributed Stochastic Neighbor Embedding (t-SNE)

The t-SNE plot (Figure 5) provides a two-dimensional visualization of the high-dimensional scRNA-seq dataset, aiming to capture the inherent structure and variability within the data. The results indicate some level of spatial separation among the samples, particularly noticeable for the GC2-LN (second patient's lymph node) samples, suggesting distinct gene expression profiles that might be influenced by the biological and pathological conditions specific to that sample. This partial segregation could reflect underlying differences in cellular composition or state, influenced by the pathological stage of the cancer which ranges from stage IIA to IIIB among the patients.

However, the overall distribution of points does not show a clear or consistent separation across all groups, indicating a complex interplay of cellular environments that might not be fully resolved by t-SNE alone. This observation aligns with insights from the referenced paper, which noted that primary and metastatic tumor subgroups were partly merged, suggesting an overlap in cellular characteristics between these conditions [1].

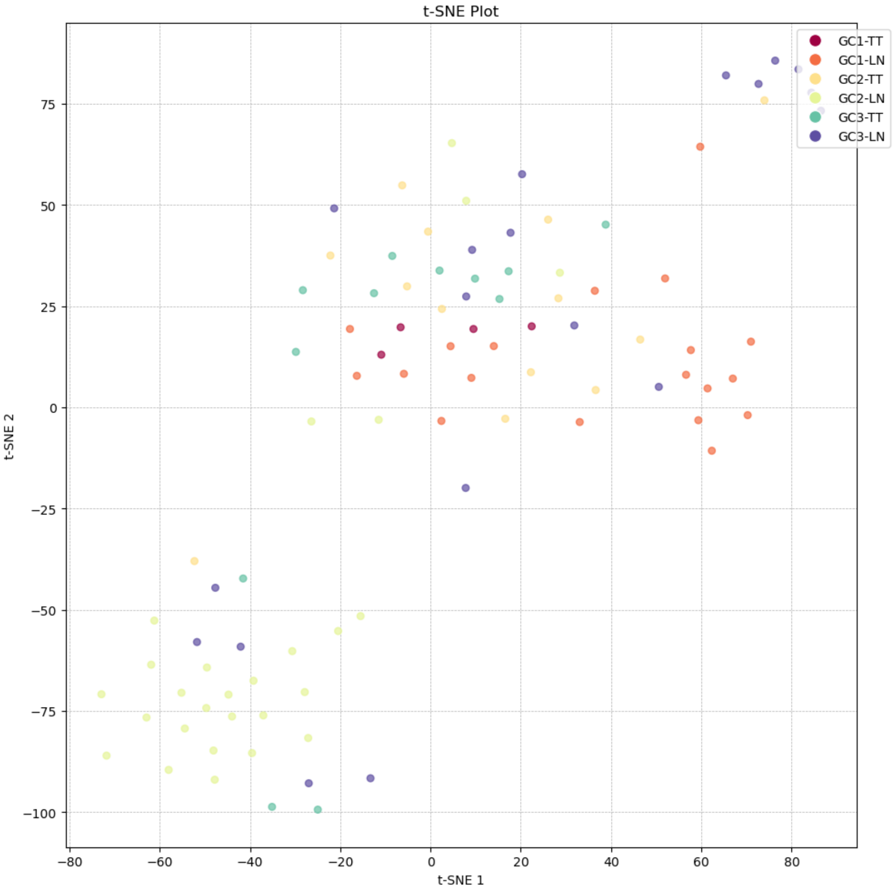


Figure 5: t-SNE visualization of single-cell RNA-seq data.

Uniform Manifold Approximation and Projection (UMAP).

The UMAP plot presents an alternative visualization of the scRNA-seq data, leveraging UMAP's strengths in maintaining both local and more global structures compared to t-SNE. The plot displayed in Figure 6 shows a more distinct separation of samples, particularly notable for samples GC2-LN and GC1-LN, indicating clearer delineation of these groups based on their gene expression profiles.

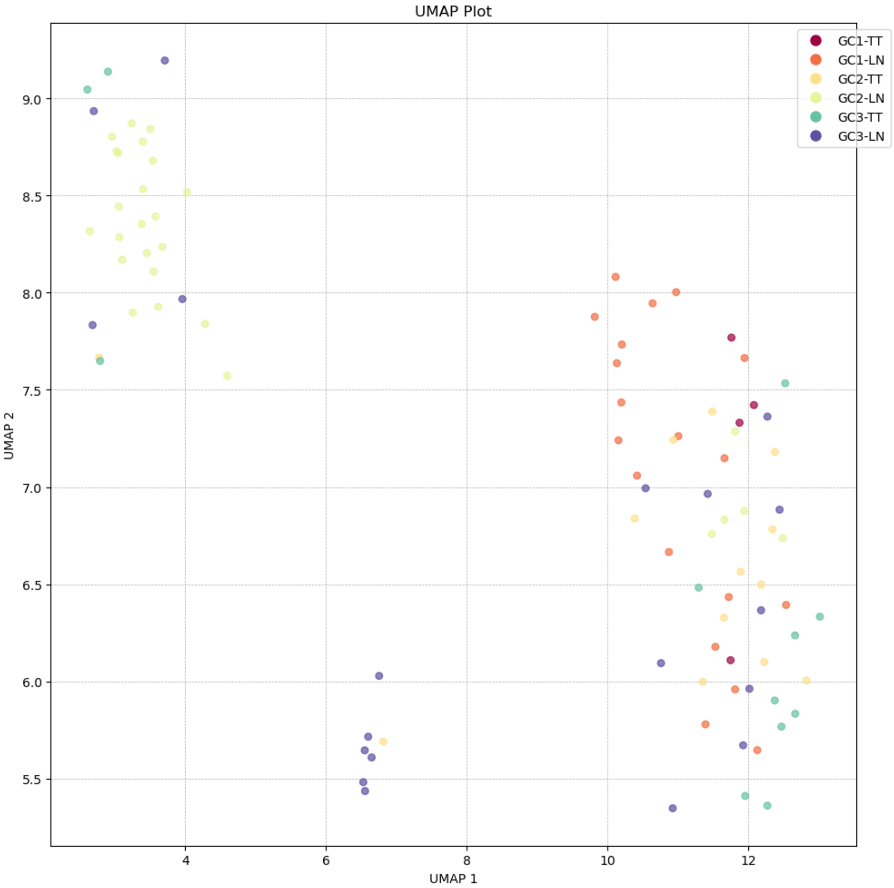


Figure 6: UMAP visualization of single-cell RNA-seq data.

The separation in the GC2-LN samples, which were not as distinct in the PCA and t-SNE plots, suggests that UMAP may be better at capturing the subtleties and nuances of the data, potentially due to its ability to preserve broader data structures. Similarly, the GC1-LN samples show a more consolidated grouping in the UMAP plot, hinting at consistent patterns or states within these samples that are distinguishable from other groups.

This improved separation could be reflective of underlying biological differences linked to the disease stages of the patients, as previously noted. For instance, the patient stages range from II to III, and the distinctions in the clustering may correlate with these stages, potentially revealing how gene expression varies with cancer progression.

Overall, the UMAP results offer a promising direction for further analysis, suggesting that this method may provide a more refined understanding of the cellular heterogeneity and the microenvironment interactions in gastric cancer. These results underscore the utility of advanced dimensionality reduction techniques in uncovering the intricate architecture of single-cell datasets, offering valuable insights into cellular behavior and interactions in gastric cancer. The distinct patterns observed across different techniques emphasize the importance of method selection based on specific analysis goals and data characteristics.

Clustering with Optimal Number of Clusters

In our clustering experiments, we applied different clustering methods with an optimization approach to determine the ideal number of clusters. Despite targeting the identification of four clusters, as specified in the reference study [1], our automated approach often distinguished only two clear clusters. This suggests a variance in cluster optimization effectiveness across different methods and datasets. Notably, methods like Louvain and Leiden applied to PCA and UMAP embeddings identified four to five clusters. However, these clusters were not well-separated, exhibiting significant overlap and thus leading to unsatisfactory separation.

This preliminary attempt using the optimal number of clusters indicates that while the method can potentially reveal interesting insights, it may not always align perfectly with expected outcomes based on prior studies. The algorithms that produced four clusters—specifically, Louvain on PCA, Leiden on PCA, and Louvain on UMAP—demonstrate a nuanced detection of substructures within the data but lack compactness and clear separation.

To better illustrate the performance of these clustering techniques, comparative barplots of evaluation metrics such as the Silhouette score, Calinski-Harabasz Index, and Davies-Bouldin Index are presented in Figure 7.

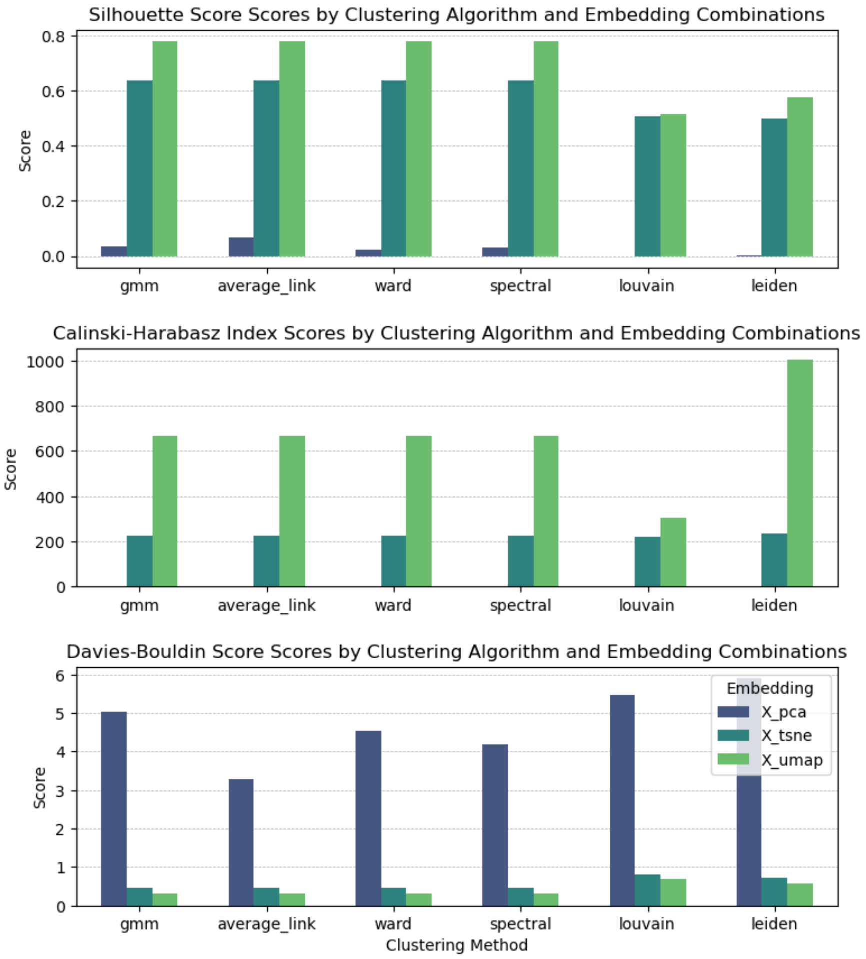


Figure 7: Comparative analysis of clustering metrics for optimized cluster numbers across different algorithms and embeddings.

The barplot visualizations offer a comprehensive comparison of clustering performance across various algorithms and embeddings, as quantified by Silhouette scores, Calinski-Harabasz Index, and Davies-Bouldin scores.

- Silhouette Score: The Silhouette Score is significantly higher for embeddings derived from t-SNE and UMAP, indicating better cluster separation and cohesion. Particularly, the UMAP embedding achieves exceptionally high scores across all algorithms, suggesting a superior structuring of the data which enhances the distinctiveness of clusters. Conversely, the PCA embeddings yield lower Silhouette scores, reflecting a relatively poor separation of clusters. This observation is consistent with the known characteristics of PCA, which, while useful for dimensionality reduction, may not always result in the most cluster-friendly transformation of the data.
- Calinski-Harabasz Index: The Calinski-Harabasz Index, which measures cluster dispersion, shows markedly higher values for UMAP embeddings. Notably, the Leiden algorithm on UMAP reaches the highest score, surpassing 1000, which underscores its effectiveness in maximizing the variance between clusters compared to within-cluster variance. The low scores observed with PCA indicate less effective variance enhancement between clusters, which aligns with the observed overlap and less distinct clustering.
- Davies-Bouldin Score: For Davies-Bouldin scores, which ideally should be low as they denote average similarity between clusters, the UMAP embeddings again outperform others, particularly in conjunction with GMM, Average Link, Ward, and Spectral clustering, all scoring around 0.3. This indicates minimal intra-cluster similarity, which is desirable in clustering contexts. The scores for PCA are considerably higher, suggesting more significant intra-cluster similarity and less effective separation, which could be problematic for clear cluster delineation.

These results collectively suggest that while PCA provides a foundational understanding of data structure, advanced embeddings like t-SNE and especially UMAP, combined with suitable clustering algorithms, are crucial for achieving clear and meaningful cluster separation in high-dimensional data scenarios such as single-cell RNA sequencing analyses. Overall, the best clustering result is achieved using UMAP with GMM, Average Linkage, Ward, or Spectral Clustering. This combination yields the highest Silhouette Score (0.781) and the lowest Davies-Bouldin Score (0.314), indicating the clusters are well-defined, compact, and well-separated. The high Calinski-Harabasz Index (667.831) further supports the quality of these clusters.

Clustering with a Predefined Number of Clusters

In addition to the optimization of cluster numbers, we further explored clustering with a predefined number of clusters set to four, across various combinations of dimensionality reduction techniques and clustering algorithms. This approach aimed to evaluate the performance of each method when constrained to produce a specific number of clusters, reflecting a controlled experimental setup similar to the original study [1].

PCA Results

Clustering on PCA embeddings demonstrated mixed results. For Gaussian Mixture Models (GMM) and Average Linkage, while clusters appeared distinct, there was a notable imbalance in the distribution of data points across clusters. Clusters 2 and 3 contained very few points compared to cluster 1, which may suggest potential misclassifications or the influence of outlier points disproportionately affecting the clustering results. This skew in cluster sizes might be attributed to PCA's linear nature, which may not effectively capture the nonlinear relationships present in high-dimensional biological data.

For Spectral, Leiden, and Louvain methods applied to PCA-reduced data, the clusters were less distinct, with considerable overlap among points. This suggests that PCA may not have provided the most conducive transformation of the data for these clustering algorithms.

t-SNE Results

t-SNE embeddings showed a significant improvement in cluster definition. GMM, Average Link, Ward, and Spectral all displayed clear separation into four well-defined clusters. However, Average Link exhibited slightly different classifications for data points within clusters 2 and 3, though the clusters remained compact. In contrast, Leiden and Louvain algorithms did not perform optimally with a predefined four-cluster setup; Louvain yielded three clusters, and Leiden produced five, indicating their sensitivity to the density and distribution of data points within the t-SNE space.

UMAP Results

UMAP provided the most satisfactory clustering outcomes among the tested dimensionality reduction methods. GMM, Average Link, Ward, and Spectral all produced four distinct and aligned clusters. Notably, even Louvain managed to capture four clusters under UMAP, although the allocation of data points in clusters 1 and 2 differed from other algorithms. Leiden, however, was unable to conform to the four-cluster constraint and resulted in five clusters.

Quantitative Metrics and Best Performance

The clustering performance was further quantified using Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Score and by plotting the comparative barplots as shown in Figure 8. Given these metrics, the best clustering result is achieved using UMAP with Average Linkage Clustering because it has the highest Silhouette Score (0.607), the second highest Calinski-Harabasz Index (954.279), and the lowest Davies-Bouldin Score (0.448). These indicators suggest well-separated and compact clusters, highlighting the robustness of UMAP in maintaining global and local data structures conducive to effective clustering. UMAP with Average Linkage Clustering graph is presented in Figure 9.

Given these metrics, the best clustering result is achieved using UMAP with Average Linkage Clustering because it has the highest Silhouette Score (0.607), the second highest Calinski-Harabasz Index (954.279, second only to Leiden on UMAP), and the lowest Davies-Bouldin Score (0.448), indicating well-separated and compact clusters.

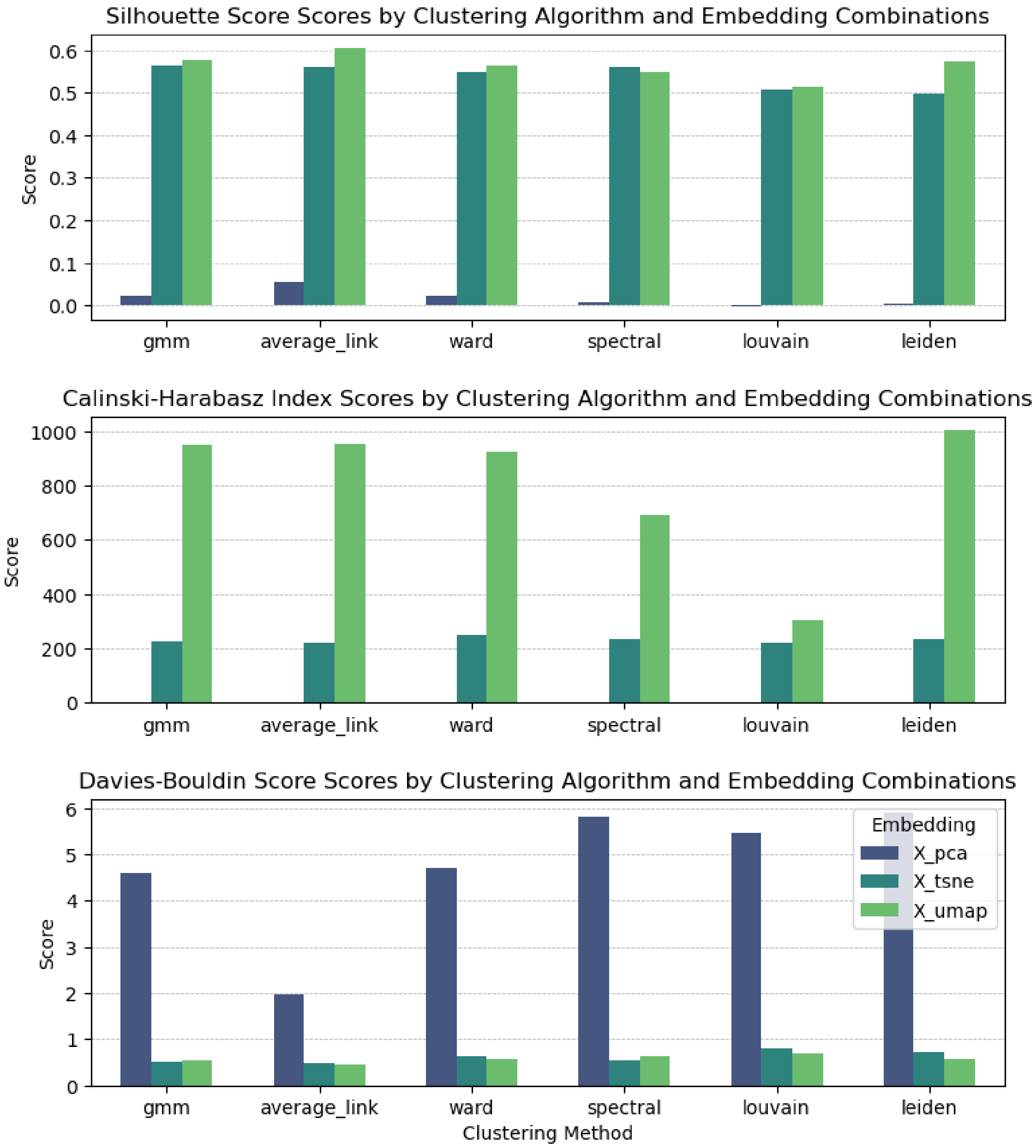


Figure 8: Comparative analysis of clustering metrics for 4 Clusters across different algorithms and embeddings.

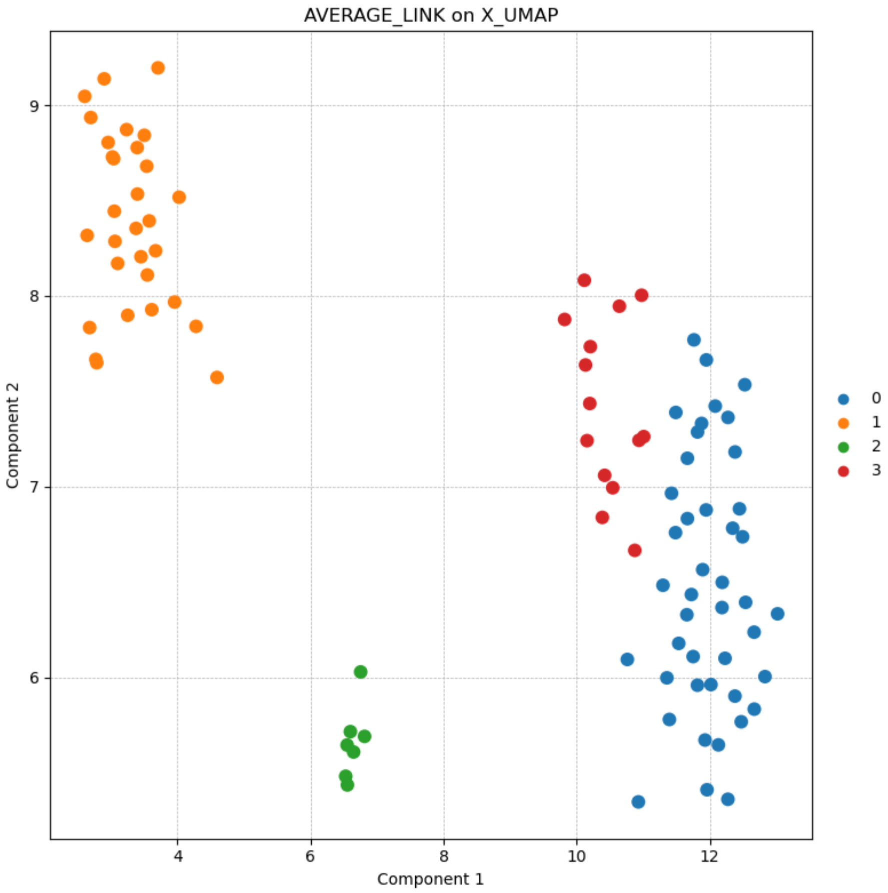


Figure 9: Clustering Visualization of UMAP Combined with Average Linkage Demonstrating Four Principal Clusters Across Single-Cell Populations.

These results affirm that while PCA is valuable for initial data exploration, advanced embeddings like t-SNE and especially UMAP, combined with suitable clustering algorithms, are crucial for achieving clear and meaningful cluster separations in complex datasets such as single-cell RNA sequencing.

Post Clustering Analysis

Marker Gene Identification

Following the clustering analysis, we focused on identifying and characterizing marker genes for each cluster derived from the best-performing clustering result (UMAP with Average Linkage). Marker genes are pivotal for understanding the distinct biological signatures and functional roles associated with each identified cluster.

Marker genes, distinct from merely highly expressed genes, are characterized by their differential expression across different conditions or clusters. These genes show significant expression differences that help in distinguishing between clusters, offering insights into the unique biological states or phenotypes represented by each cluster. The heatmap displayed in Figure 10 illustrates the expression patterns of the top 10 highly expressed marker genes in each cluster. This visualization helps in comparing the expression levels across clusters, highlighting the distinct and unique profiles that characterize each group.

In more detail:

- Cluster 0 is characterized by genes predominantly associated with hemoglobin complex and mitochondrial RNA processing, such as HBA2, HBA1, and various members of the MTRNR2L family. This suggests that this cluster might be representing cells with high metabolic activity or specific red blood cell precursors.
- Cluster 1 features genes involved in signaling pathways and cellular structure, such as FZD3 and ARL4A, indicating a cluster potentially rich in signaling molecules influencing cell communication and structure.
- Cluster 2 includes genes like GDPD5 and SLC22A13, which are linked to metabolic processes and transport functions, suggesting these cells may be involved in active metabolite transport and processing.
- Cluster 3 is marked by genes associated with cellular stress responses and nuclear elements such as MALAT1 and FOSB, indicating a potential involvement in rapid response mechanisms to external stimuli or stress.

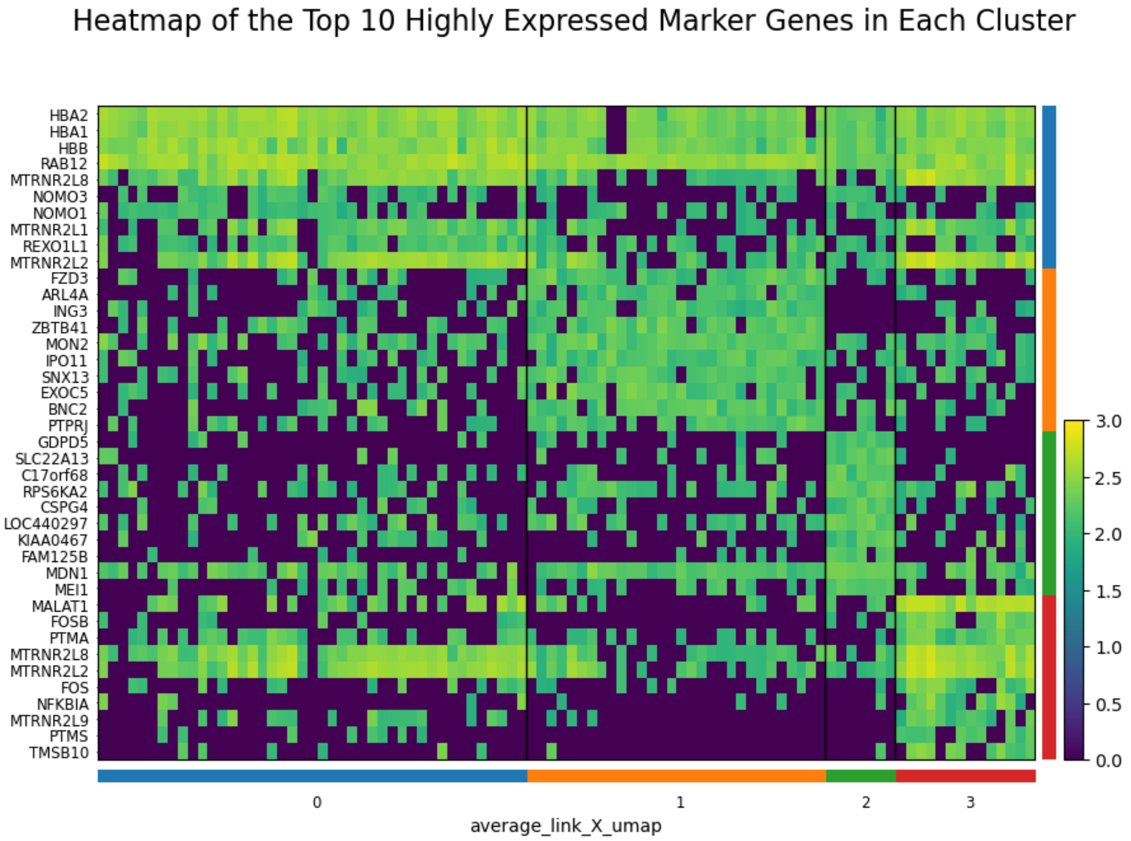


Figure 10: Heatmap of the top 10 highly expressed marker genes in each cluster, identified from UMAP with Average Linkage clustering.

The differential expression of these marker genes across clusters not only validates the biological relevance of our clustering results but also provides a foundation for deeper insights into the cellular composition and functionality within the sampled tissue.

Comparative Analysis of Marker Gene Identification in Gastric Cancer

Our study identified several marker genes that are significant in the context of gastric cancer, albeit different from those highlighted in the original study. Key genes such as HBA2, HBA1, and HBB, found within Cluster 0, are primarily associated with hemoglobin complexes. While these are not directly related to the genes identified in the original paper, their expression in cancer could reflect alterations in oxygen transport and cellular environments within tumors, which could be linked to hypoxic conditions often observed in rapidly growing tumors.

In Cluster 1, genes like ARL4A and ZBTB41 were notable. ARL4A belongs to the ADP-ribosylation factor family of GTP-binding proteins, involved in actin cytoskeleton remodeling and therefore could play a role in cancer cell migration and metastasis, somewhat echoing the functions of the CLDN11 gene highlighted in the original study concerning tumor migration.

Furthermore, Cluster 2 featured genes such as GDPD5 and CSPG4, with CSPG4 being involved in cellular proliferation and migration, similar to the roles proposed for KIF5B and ERBB4 in the original paper. This suggests that while the exact genes differ, the functional pathways they influence—particularly those driving cancer cell proliferation and migration—remain a consistent theme.

MALAT1, identified in Cluster 3, is a well-known long non-coding RNA (lncRNA) associated with various cancers, including gastric cancer. It is implicated in the regulation of metastasis and cellular growth, linking well with the discussion on ERBB2 and CDK12 in the original study, which are also associated with metastatic behavior and cancer progression.

Our study's focus on differentially expressed genes across a spectrum of clusters rather than isolating markers for primary versus metastatic cancer provided a broader view of the genetic landscape but also introduced a point of divergence from the original study. This broader approach allowed us to uncover a variety of functional genes that, while different, still relate to critical pathways involved in cancer biology.

Overall, while our identified genes may not exactly mirror those found in the original paper, they belong to similar biological pathways or gene families, underscoring their potential roles in gastric cancer pathology. This reinforces the relevance of our findings within the established framework of gastric cancer research and highlights the complexity of cancer genomics where different studies may reveal complementary aspects of the disease

Functional annotation of clusters using Gene Ontology (GO)

Functional annotation of the gene clusters, utilizing Gene Ontology (GO), highlights specific biological processes, cellular components, and molecular functions that are associated with each cluster identified by the best clustering approach, Average Linkage on UMAP. These annotations provide a deeper understanding of the roles that differentially expressed genes may play within distinct clusters.

- Cluster 0 Annotations: Cluster 0 exhibits significant GO annotations associated with protein transport and ER membrane integration, featuring specific complexes like the "multi-pass translocon complex" and "ER membrane insertion complex." These annotations indicate a strong involvement in the mechanisms of protein translocation across cellular membranes, a critical process for cellular function and homeostasis. The presence of specific transcription factors, such as ZNF606 and ZNF699, suggests a regulatory layer that may be essential for these protein transport processes. Although the recall values for these annotations are modest, they indicate that these processes involve a significant subset of genes within the cluster, suggesting specialized functions that could be critical in cellular responses to environmental or pathological conditions.

Source	**Native**	**Name**	**P-value**
GO:CC	GO:0160064	multi-pass translocon complex	0.000081
GO:CC	GO:0072379	ER membrane insertion complex	0.000205
TF	TF:M06584	Factor: ZNF606; motif: GSCCTCTAGAAK	0.000490
GO:BP	GO:0160063	multi-pass transmembrane protein insertion into ER	0.000651
TF	TF:M05758	Factor: ZNF699; motif: NGKATYAAAATA	0.002122
GO:BP	GO:0045048	protein insertion into ER membrane	0.009532
GO:BP	GO:0051205	protein insertion into membrane	0.018669
TF	TF:M05617	Factor: ZNF500; motif: NGTTCCGGGGRW	0.045386
GO:MF	GO:0043022	ribosome binding	0.046598
HP	HP:0034280	Target cells	0.049924

Table 1: Gene Ontology annotations for cluster 0.

- Cluster 1 Annotations: For Cluster 1, no significant GO annotations were found, which could be indicative of several factors including a smaller size of distinct marker genes list, or these genes not meeting the statistical significance thresholds for GO annotation. This lack of significant annotations could suggest that the genes in this cluster may not be involved in well-defined or specific biological processes or might represent a more diverse set of functions that are not captured by existing GO terms.
- Cluster 2 Annotations: Cluster 2 reveals a specific functional annotation related to "nicotinate transmembrane transporter activity." This annotation suggests that cells within this cluster may play a role in the transport of nicotinate across cellular membranes, which is crucial for maintaining NAD+ levels in cells. This function could be pertinent to cellular energy metabolism and possibly the cellular response to stress or damage, which are vital aspects in cancer physiology.

Source	**Native**	**Name**	**P-value**
GO:MF	GO:0090416	nicotinate transmembrane transporter activity	0.049875

Table 2: Gene Ontology annotations for cluster 2.

- Cluster 3 Annotations: Cluster 3 is enriched in pathways related to immune response and signaling with pathways like the "IL-17 signaling pathway" and "Osteoclast differentiation." These pathways are pivotal for immune modulation and inflammation, processes often co-opted in cancer to promote tumor growth and metastasis. The high precision values in these annotations emphasize a strong and specific correlation between the cluster's genes and these immune-related functions, suggesting their potential roles in the pathology of gastric cancer, particularly in its interaction with the host's immune system.

Source	Native	Name	P-value
KEGG	KEGG:04657	IL-17 signaling pathway	0.000061
KEGG	KEGG:04380	Osteoclast differentiation	0.000182
WP	WP:WP2355	Corticotropin releasing hormone signaling pathway	0.001089
WP	WP:WP4919	Neuroinflammation	0.001957
WP	WP:WP2435	Quercetin and Nf kB AP 1 induced apoptosis	0.003009
REAC	REAC:R-HSA-9031628	NGF-stimulated transcription	0.006812
KEGG	KEGG:05031	Amphetamine addiction	0.009631
KEGG	KEGG:05140	Leishmaniasis	0.010491
GO:BP	GO:0051412	response to corticosterone	0.011370
WP	WP:WP4880	Host pathogen interaction of human coronavirus	0.013185

Table 3: Gene Ontology annotations for cluster 3.

Discussion on Functional Annotations

The functional annotations derived from our analysis underscore the biological diversity among the clusters and align with distinct transcriptional profiles previously identified. These enriched biological themes offer insights into the roles of various cell populations in gastric cancer. For example:

Cluster 0's focus on protein transport and ER-related processes suggests these cells may be involved in heightened synthetic and secretory activities, potentially related to the production of factors that influence tumor growth or response to therapy. The absence of annotations in Cluster 1 raises interesting questions about the heterogeneity and complexity of the tumor microenvironment, possibly reflecting a transitional or less characterized cell state within the tumor. Cluster 3's enrichment in immune-related pathways highlights the interplay between cancer cells and the immune system, potentially indicating cells that are either responding to immune activity or manipulating immune responses to facilitate cancer progression. These insights not only validate some of the observations from the original study but also provide new avenues for understanding the cellular mechanisms in gastric cancer.

Conclusions - Further Research

This study aimed to replicate and extend previous scRNA-seq analyses of gastric cancer, utilizing a range of dimensionality reduction and clustering techniques to probe the complex cellular heterogeneity characteristic of this disease. Our approach underscored the utility of UMAP and Average Linkage Clustering for revealing distinct cellular subpopulations, although our results suggest room for further refinement in cluster definition and separation. The integration of Gene Ontology annotations enriched our understanding of the functional implications of our clustering results, albeit revealing gaps that warrant deeper investigation.

Future work could benefit from exploring additional clustering algorithms (e.g. DBSCAN) and cluster evaluation metrics (e.g. BIC criterion), improving our 'GCSingleCellAnalysis' class so it can handle raw data instead of starting from a counts matrix and integrating pseudotime trajectory analysis, similarly to the authors, to better understand the temporal dynamics of cellular states within gastric cancer, potentially aligning closer with clinical outcomes and therapeutic strategies.

References

[1] B. Wang, H. Lyu, J. Pei, Z. Huang, D. Wang, and W. Sun, "Comprehensive analysis of metastatic gastric cancer tumour cells using single-cell rna-seq," *Scientific Reports*, vol. 11, no. 1, Jan. 2021.

[2] W. Zhao *et al.*, "Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell rna-seq," *Genome Medicine*, vol. 13, no. 1, May 2021.

[3] R. Qi and Q. Zou, "Editorial: Machine learning methods in single-cell immune and drug response prediction," *Frontiers in Genetics*, vol. 14, Jun 2023.

[4] F. Wolf, P. Angerer, and F. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, p. 15, 2018.

[5] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[6] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[7] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018. [Online]. Available: https://doi.org/10.21105/joss.00861

[8] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: A flexible trimmer for illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug 2014.

[9] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, "Graph-based genome alignment and genotyping with hisat2 and hisat-genotype," *Nature Biotechnology*, vol. 37, no. 8, pp. 907–915, 2019.

[10] Y. Liao, G. K. Smyth, and W. Shi, "featurecounts: An efficient general purpose program for assigning sequence reads to genomic features," *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr 2014.

[11] D. J. McCarthy, K. R. Campbell, A. T. Lun, and Q. F. Wills, "Scater: Pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r," *Bioinformatics*, vol. 33, no. 8, pp. 1179–1186, 2017.

[12] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, "Integrating single-cell transcriptomic data across different conditions, technologies, and species," *Nature Biotechnology*, vol. 36, no. 5, pp. 411–420, 2018.

[13] Z. Ji and H. Ji, "Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis," *Nucleic Acids Research*, vol. 44, no. 13, p. e117, Jul 2016.

[14] C. Trapnell, D. Cacchiarelli, J. Grimsby *et al.*, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nat Biotechnol*, vol. 32, pp. 381–386, 2014.

[15] M. Su, T. Pan, Q.-Z. Chen, Z. Weiwei, Y. Gong, G. Xu, H.-Y. Yan, S. Li, Q.-Z. Shi, Y. Zhang, X. He, C. Jiang, S.-C. Fan, X. Li, M. Cairns, X. Wang, and Y. Li, "Data analysis guidelines for single-cell rna-seq in biomedical studies and clinical applications," *Military Medical Research*, 12 2022.

[16] A. Subramanian, M. Alperovich, B. Li, and Y. Yang, "Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics," 10 2021.

[17] Q. Wei, X. Yuan, G. Huang, J. Li, L. Chen, and J. Ying, "Correlation between hemoglobin levels and the prognosis of first-line chemotherapy in patients with advanced gastric cancer," *Cancer Management and Research*, vol. 12, pp. 7009–7019, 08 2020.

[18] S. Wang, S. Tao, Y. Liu *et al.*, "Identification of significant genes associated with prognosis of gastric cancer by bioinformatics analysis," *Journal of the Egyptian National Cancer Institute*, vol. 34, p. 55, 2022.

[19] National Center for Biotechnology Information, "Human gene nc_000023 (provided in gene id 9606)," Genetic sequence, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/gene/?term=%229606%22%5BTaxonomy+ID%5D+AND+alive%5Bproperty%5D+AND+genetype+protein+coding%5BProperties%5D

Code availability

To access the code and the results, please refer to our GitHub repository at
`https://github.com/GiatrasKon/MLCB-Final-Project`.

Data availability

The CSV file of the raw counts matrix used for our scRNA-seq analysis can be found in GEO (GSE158631) as originally deposited by the authors.

LLM Usage

During this project, ChatGPT was used for a variety of purposes:

- Explanation of concepts to assist with further understanding and insight.
- Assistance in code readability (e.g. help with adding informative comments).
- Suggestions for code error handling and bug fixing.
- Task management and organization (e.g. breaking a larger task into more, smaller ones).

Supplementary Material

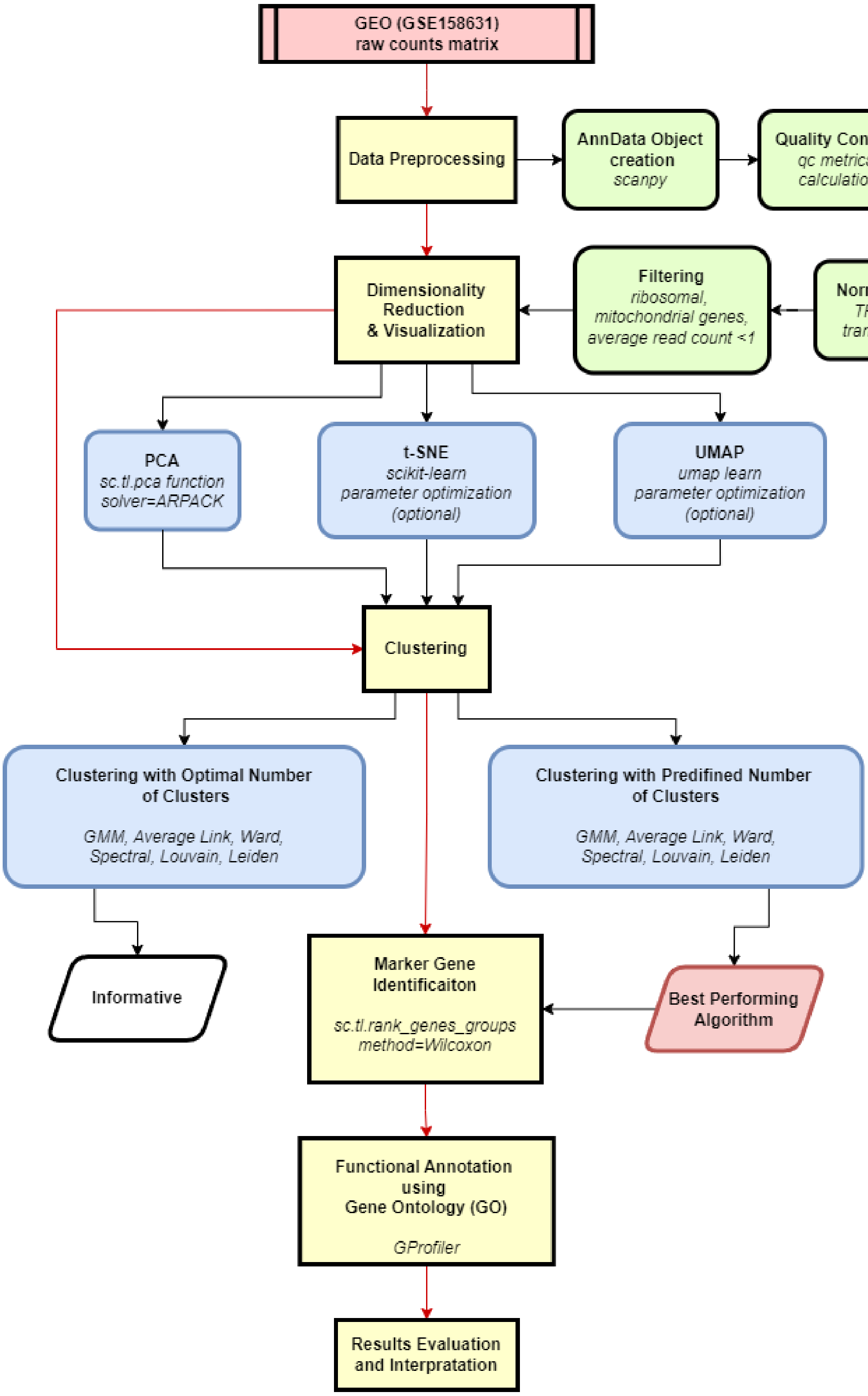

GEO (GSE158631)
raw counts matrix
Data Preprocessing
AnnData Object creation
scanpy
Quality Control
qc metrics calculation
Normalization
TPM, log2 tranformation
Filtering
ribosomal, mitochondrial genes, average read count <1
Dimensionality Reduction & Visualization
PCA
sc.tl.pca function solver=ARPACK
t-SNE
scikit-learn parameter optimization (optional)
UMAP
umap learn parameter optimization (optional)
Clustering
Clustering with Optimal Number of Clusters
GMM, Average Link, Ward, Spectral, Louvain, Leiden
Clustering with Predifined Number of Clusters
GMM, Average Link, Ward, Spectral, Louvain, Leiden
Informative
Marker Gene Identificaiton
sc.tl.rank_genes_groups method=Wilcoxon
Best Performing Algorithm
Functional Annotation using Gene Ontology (GO)
GProfiler
Results Evaluation and Interpratation

Figure 11: Single-cell RNA-seq analysis pipeline in our study.

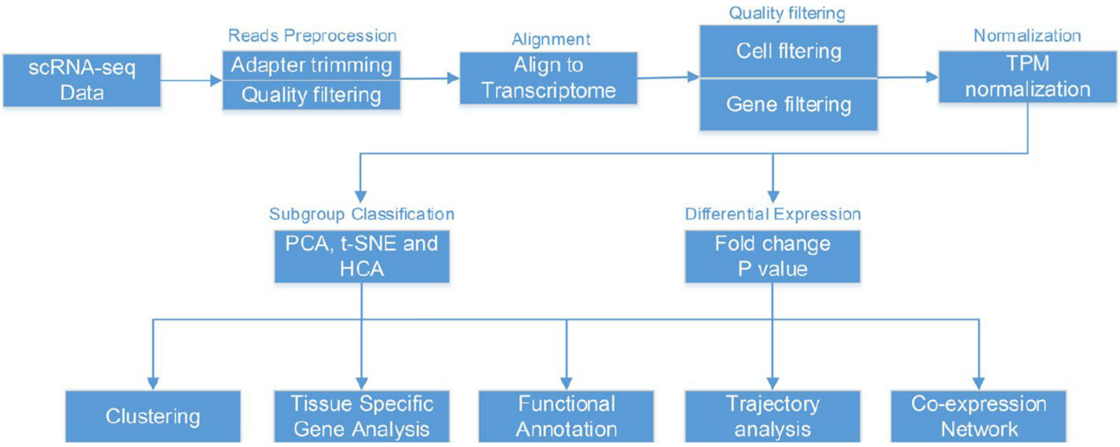


Figure 12: Single-cell RNA-seq analysis pipeline in the original study.

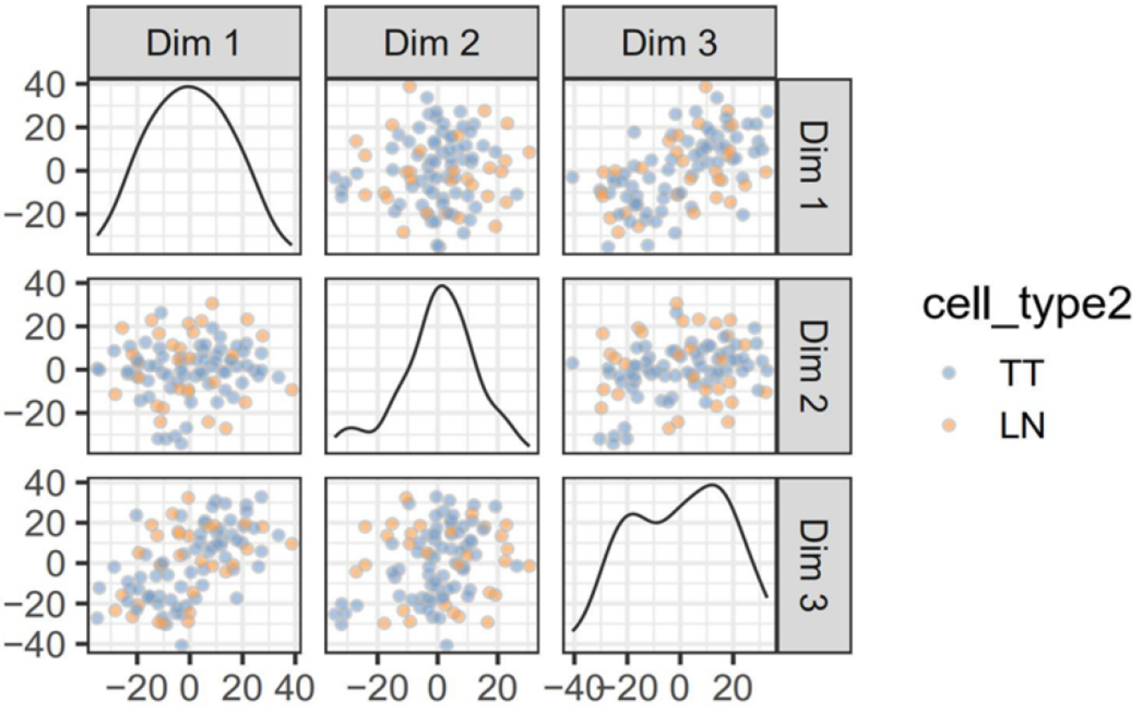


Figure 13: Pair plot of the PCA-reduced data visualized using t-SNE in the original study.

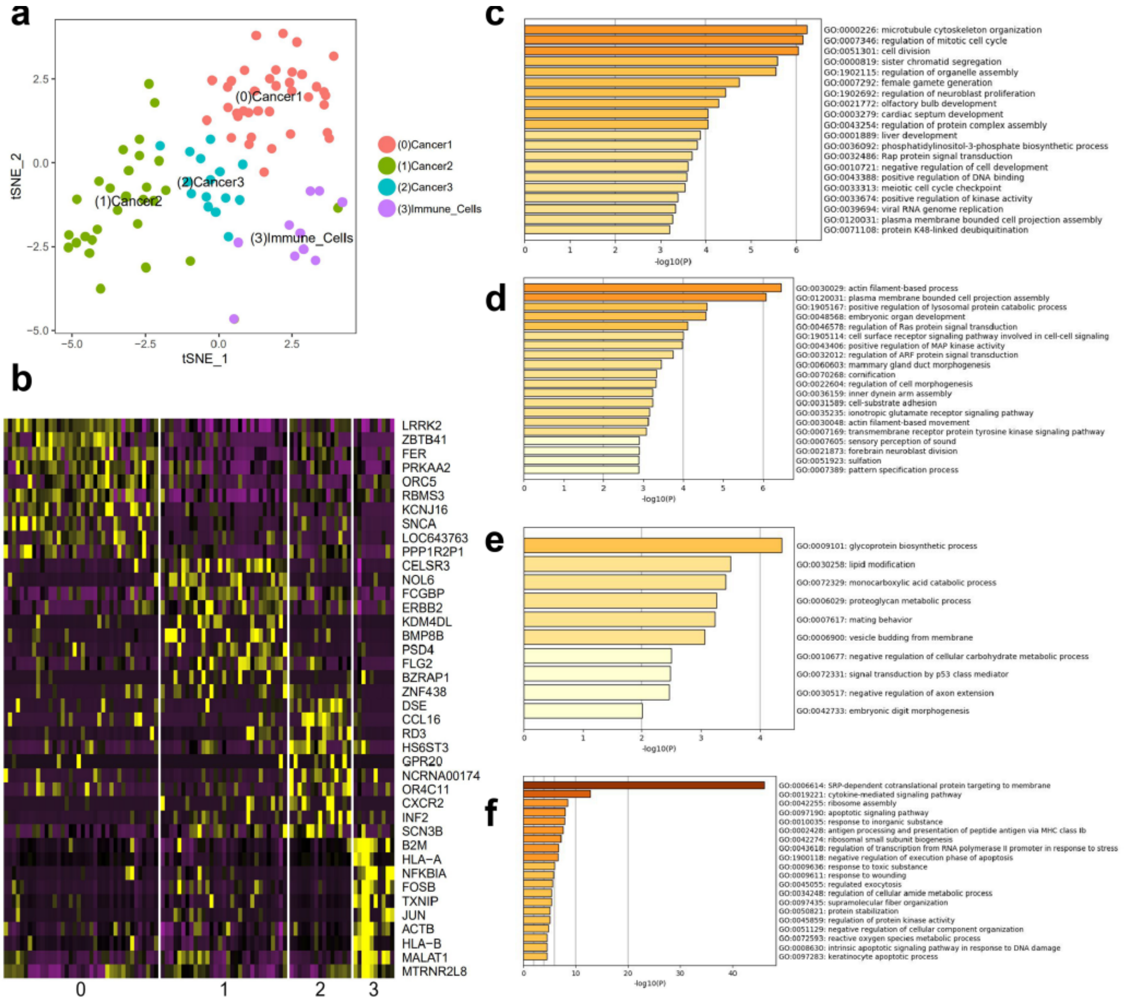


Figure 14: Clustering results and post-clustering analysis in the original study. (**a**) 4 main clusters identified in the tumor tissues after applying hierarchical clustering on the PCA-reduced data and visualizing the results with t-SNE. (**b**) Heatmap indicating the top 10 highly expressed marker genes in each of the 4 clusters. (**c-f**) GO funtional annotations for each of the 4 clusters based on their top marker genes.