Unsupervised Clustering of Human Tissue Samples Based on Gene Expression Profiles

1. Introduction

Bulk RNA sequencing (RNA-Seq) is a powerful technique used to quantify gene expression across multiple tissue types or experimental conditions. In this study, transcriptomic data were retrieved from the Genotype-Tissue Expression (GTEx) Project, a comprehensive resource that offers high-quality RNA-Seq data across various human tissues. Proper preprocessing and normalization of this dataset are essential to ensure the reliability of downstream analyses, including differential expression and clustering.

2. Data Source and Overview

The bulk RNA-Seq dataset was downloaded from the GTEx portal (https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression). The selected dataset is already TPM-normalized (Transcripts Per Million), which is a common normalization approach that accounts for sequencing depth and gene length, facilitating comparisons across samples and genes.

The original matrix contains expression values for approximately:

- 59,035 genes (in rows)
- 19,616 tissue samples (in columns)

This comprehensive dataset spans a wide range of adult human tissues, making it a valuable resource for tissue-specific expression profiling and comparative analyses.

3. Preprocessing Steps

To refine the dataset for more focused and reliable analysis, we employed a systematic filtering strategy based on gene expression thresholds and data quality checks. The major preprocessing steps are outlined below:

3.1 Mean Expression Filtering

For each gene, the row-wise mean (across all tissue samples) was calculated. Genes exhibiting low average expression across tissues are often considered biologically uninformative or prone to noise. To mitigate this, We excluded genes with mean expression values less than 1 TPM.

This step helps reduce dimensionality, focus on consistently expressed genes, and enhance the signal-to-noise ratio in the data.

3.2 Post-Filtering Gene Count

After applying the above filtering criterion, the dataset was reduced to a more manageable and informative set of: 20,545 genes

This reduced gene set provides a balance between inclusivity and quality, retaining genes with sufficient expression levels for robust downstream analyses.

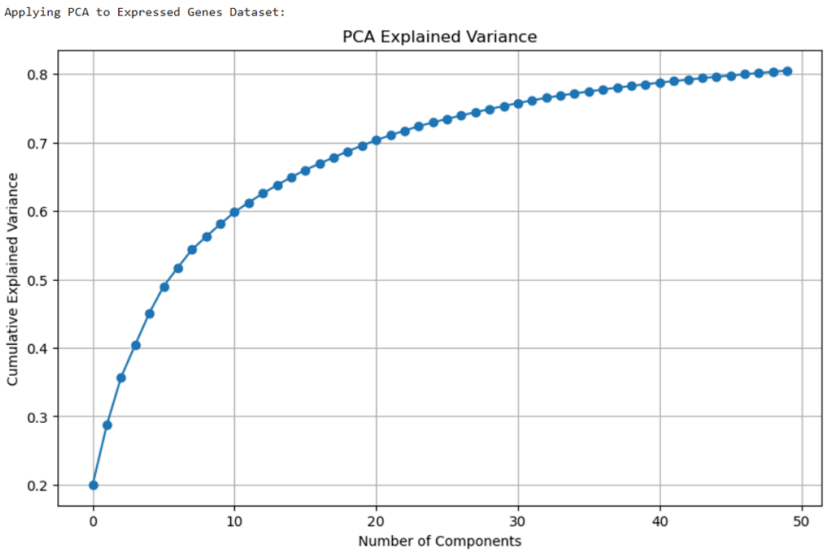
3.3 Handling Missing Values

Although no missing values were identified in the expression matrix, a conservative approach was adopted. If any missing (NA) values had been present, the corresponding genes would have been removed entirely. This ensures that the final dataset is free of technical artifacts or imputation biases.

4. Clustering and Modeling of GTEx Gene Expression Data

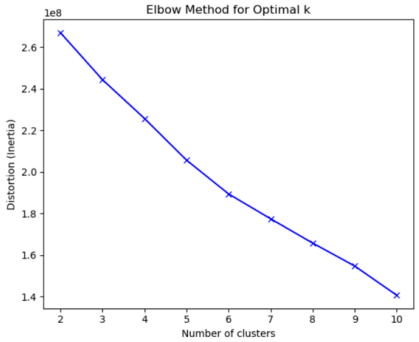
In this study, we focused on designing and executing the clustering and modeling pipeline to uncover biologically meaningful tissue clusters in high-dimensional bulk RNA-Seq data. The overarching objective was to implement unsupervised machine learning algorithms, reduce dimensionality, tune model parameters, and validate the clusters using both internal and external evaluation metrics. This analysis was performed without using any tissue label information during model training, relying solely on expression profiles.

4.1 Dimensionality Reduction using PCA



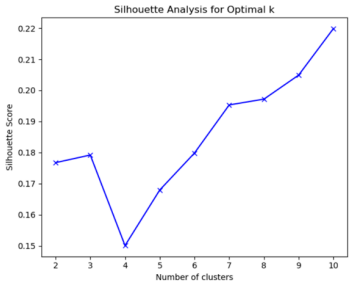
The original GTEx dataset comprised ~20,000 genes per sample, making direct clustering computationally inefficient and prone to noise. To address this, we applied Principal Component Analysis (PCA) to reduce the dataset's dimensionality. The first 50 principal components were selected, capturing approximately 80% of the total variance. This dimensionality reduction enabled faster model training and improved our ability to separate clusters by removing low-variance noise.

4.2 Clustering Models and Parameter Tuning: Elbow Plot and Silhouette Score Analysis



To determine the optimal number of clusters (k) for K-Means and GMM, we used two techniques: the Elbow Method and Silhouette Score Analysis. The Elbow plot displays how inertia, or the within-cluster sum of squares, decreases as k increases. Around k=10, the rate of decrease significantly slows, forming a visible 'elbow'—suggesting that adding more clusters beyond this point yields diminishing returns. This inflection point guided our initial cluster selection.

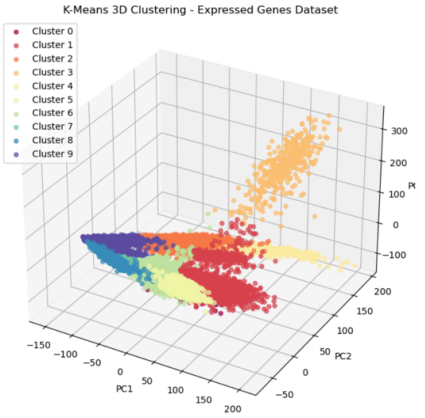
The Silhouette Score plot offers a complementary perspective by measuring how well each sample fits within its cluster versus the nearest alternative. Higher scores indicate better separation. In our results, silhouette scores steadily increased and peaked at k=10, reinforcing the elbow plot's suggestion. This dual-evidence approach ensured robust selection of the number of clusters for downstream modeling.



The Elbow plot showed a clear inflection around k=10, where further increases in k yielded diminishing returns in reducing intra-cluster inertia. The corresponding silhouette analysis supported this choice, with silhouette scores increasing steadily and peaking at k=10. This combination of methods helped us justify the use of 10 clusters in downstream modeling.

We implemented and tuned three clustering models: **K-Means**, **Gaussian Mixture Models (GMM)**, and **DBSCAN**.

5.1. K-Means Clustering



The performance of the K-Means clustering model is visualized using two types of figures. The first diagram shows the 2D PCA plot of K-Means clustering using the first two principal components (PC1 and PC2). Each point represents a tissue sample, colored by its assigned cluster. This visualization highlights how samples form well-separated groups along PC axes, reflecting the algorithm's ability to distinguish tissue-specific gene expression patterns. Clusters such as those representing testis, skeletal muscle, and thyroid are particularly distinct, with minimal overlap.

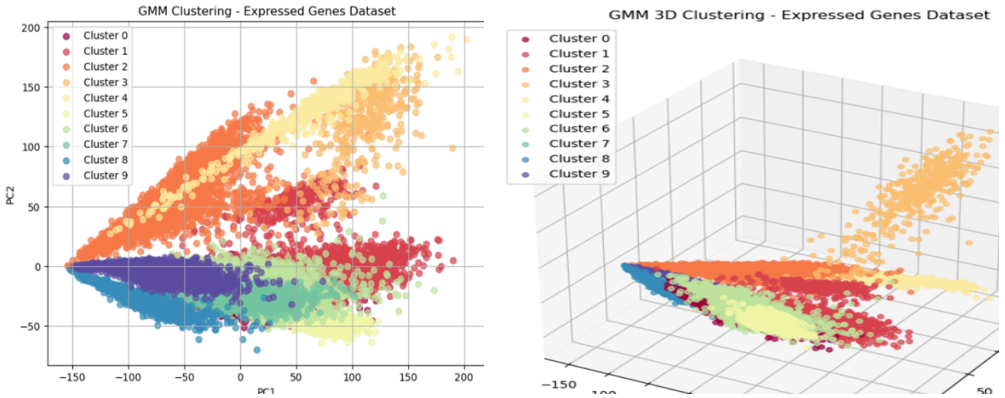
The second set of visuals includes two key projections:2D PCA plot showing sample separation across the first two components (PC1 vs. PC2), and a 3D PCA plot incorporating PC3 to offer a volumetric view of cluster relationships. In both cases, data points are color-coded by their assigned cluster, revealing clear group boundaries with minimal overlap. The 2D plot highlights how major clusters diverge horizontally and vertically in PCA space, while the 3D plot gives spatial insight into how tightly or loosely those clusters are distributed. These visualizations make the compactness and separability of K-Means clusters intuitively clear and support the quantitative findings from silhouette and index metrics. These figures illustrate that K-Means created compact and well-separated clusters. The supporting visuals make it clear how clustering performance metrics guide model selection and underscore the biological significance, especially for tissues with distinct profiles like testis and skeletal muscle.

5.2 Gaussian Mixture Models (GMM)

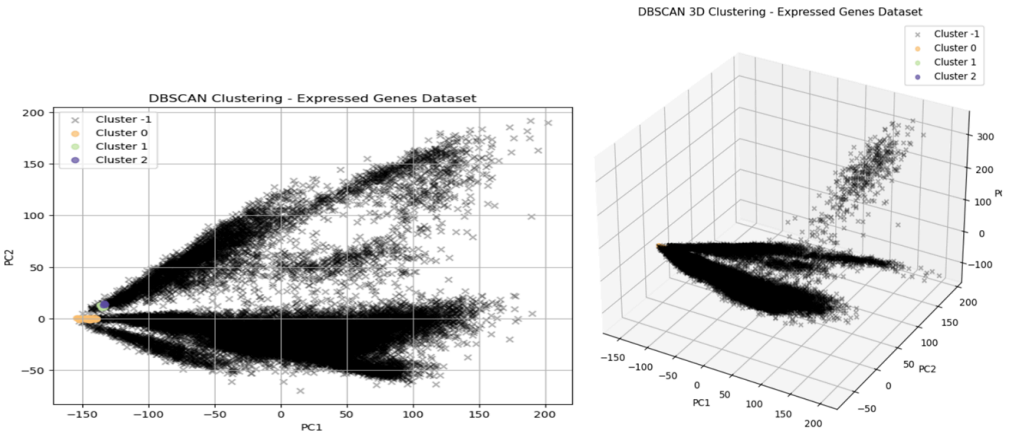
GMM performance is illustrated using both metric-based plots and dimensionality reduction-based projections. The GMM cluster evaluation visuals include a silhouette score trend plot across various k values, along with 2D and 3D PCA cluster projections. The silhouette trend plot peaks at k=10, showing moderate clustering quality. The PCA visualizations show overlapping cluster boundaries typical of GMM. The 2D and 3D PCA scatter plots display gradual transitions between clusters and confirm the model's ability to capture soft, probabilistic clustering structure.

This analysis is supported by metric plots showing the silhouette score peaking at k=10 with a value of 0.226, suggesting a moderate but valid clustering solution. Supplementary bar charts or curves

also reflect slightly higher Davies-Bouldin scores, indicating more overlap between clusters compared to K-Means. These diagrams provide context for the model's ability to accommodate tissues with transcriptomic similarity. GMM's flexibility in modeling soft boundaries becomes particularly apparent in these visuals, aligning with its mathematical foundations and practical performance.



5.3 DBSCAN



DBSCAN's results are visualized through both silhouette scoring and PCA cluster projections. DBSCAN's silhouette score diagram highlights the performance peak when using core points only, affirming the algorithm's precision in high-density areas. Complementary 2D and 3D PCA cluster plots show clearly defined cluster cores and scattered noise points. These visuals distinguish DBSCAN's capability to form well-separated clusters where density permits, while excluding ambiguous samples as noise.

In the PCA-based scatter plots, DBSCAN's clusters appear as dense regions with clearly separated boundaries. Samples labeled as noise are visible as scattered points that do not belong to any cluster. These visuals highlight the model's effectiveness in uncovering distinct, compact groupings like skeletal muscle and testis, while acknowledging its limitations in sparse or noisy regions of the dataset.

DBSCAN's evaluation includes a high silhouette score of 0.660 when considering only core points, which is visualized through silhouette plots demonstrating this peak. These figures help confirm the

model's success in isolating dense, confident clusters. In parallel, PCA projection visuals reveal how DBSCAN groups compact clusters while excluding low-density regions as noise. These combined diagrams make it clear that DBSCAN is highly effective for well-defined structures, even if it sacrifices overall data coverage in favor of cluster purity.

Summary of Performance Across Models:

Model	Best Params	Silhouette Score	Strengths
K-Means	k = 10	0.243	Consistent, interpretable, well-separated
GMM	k = 10	0.226	Handles overlaps, soft assignment
DBSCAN	eps = 2	0.660 (core only)	Finds dense clusters, sensitive to noise

6 External Validation with ARI and NMI:

To evaluate the biological relevance and accuracy of our unsupervised clustering results, we compared the derived cluster assignments with the known tissue labels available in the GTEx metadata. This metadata consists of SAMPID which is a Unique sample identifier, SMTS which is Broad tissue type and SMTSD which is Detailed tissue type. This comparison serves as external validation, offering a benchmark for how well the clustering algorithm captured real-world biological distinctions—despite not having access to any label information during training.

The Adjusted Rand Index (ARI) is a widely used metric for evaluating clustering performance when ground truth labels are available. It compares the similarity between two groupings by examining how consistently sample pairs are assigned to the same or different clusters in both the predicted and actual labels. Importantly, ARI adjusts for random chance, where a score of 0 indicates clustering no better than random, and a score of 1 signifies perfect agreement. In this study, the ARI score of 0.191 reflects a modest level of agreement between the predicted clusters and actual tissue types. While not indicative of high accuracy, this value is still meaningful, especially considering that the clustering was performed in an unsupervised manner without access to label information. It suggests that the algorithm was able to partially uncover real biological structures embedded within the data, even if the alignment with known labels was not exact.

The Normalized Mutual Information (NMI) score measures the degree of shared information between predicted clusters and true labels, providing insight into how well the clustering reflects known groupings. It ranges from 0, indicating no mutual information or relationship, to 1, representing perfect correlation between the cluster assignments and actual labels. In this analysis, an NMI score of 0.571 suggests a moderate to good alignment between the unsupervised clusters and the ground-truth tissue types. This indicates that the model was able to capture a significant portion of the underlying biological structure present in the gene expression data. The relatively high NMI value reinforces the idea that the clustering outcomes are not random but biologically meaningful, reflecting genuine patterns across tissue types.

Together, the ARI and NMI values indicate that the clustering process uncovered non-trivial, biologically valid groupings within the high-dimensional gene expression data. These results reinforce the hypothesis that unsupervised methods can reveal latent tissue-specific gene expression patterns, even when starting from unlabelled data. Though not perfect, the clustering results demonstrate a meaningful correlation with known biological categories, providing confidence in the utility of unsupervised learning for tissue classification and exploratory bioinformatics research.

6.1 Cluster Purity:

Cluster purity is an intuitive and visual method of evaluating how well the unsupervised clustering results align with known tissue types. In this project, we constructed confusion matrices mapping predicted cluster IDs to actual tissue labels from the GTEx metadata to evaluate how well the unsupervised clusters align with known tissue types. Each row represents a tissue type, and each column represents one of the 10 clusters. The values indicate the number of samples of a tissue assigned to each cluster.

Cluster	0	1	2	3	4	5	6	7	8	9
Tissue										
Adipose - Subcutaneous	3	3	0	0	0	0	704	0	0	4
Adipose - Visceral (Omentum)	2	9	0	0	0	0	520	0	0	56
Adrenal Gland	0	2	0	0	0	0	239	0	0	54
Artery - Aorta	466	1	0	0	0	0	5	0	0	0
Artery - Coronary	180	12	0	0	0	0	73	0	0	3
Artery - Tibial	675	1	0	0	0	0	15	0	0	0
Bladder	0	10	0	0	0	0	67	0	0	0
Brain - Amygdala	0	0	110	0	0	0	0	0	0	71
Brain - Anterior cingulate cortex (BA24)	0	0	185	0	0	0	0	0	0	48
Brain - Caudate (basal ganglia)	0	0	194	0	0	0	1	0	0	105
Brain - Cerebellar Hemisphere	0	0	16	0	240	0	0	0	0	21
Brain - Cerebellum	0	0	12	0	244	0	0	0	0	10
Brain - Cortex	0	0	237	0	0	0	0	0	0	33
Brain - Frontal Cortex (BA9)	0	1	241	0	0	0	1	0	0	26
Brain - Hippocampus	0	0	154	0	0	0	0	0	0	101
Brain - Hypothalamus	0	0	180	0	0	0	0	0	0	77
Brain - Nucleus accumbens (basal ganglia)	0	0	214	0	0	0	0	0	0	71
Brain - Putamen (basal ganglia)	0	0	142	0	0	0	0	0	0	112
Brain - Spinal cord (cervical c-1)	0	1	152	0	0	0	1	0	0	50
Brain - Substantia nigra	0	0	106	0	0	0	1	0	0	76
Breast - Mammary Tissue	0	124	0	0	0	0	372	0	0	18
Cells - Cultured fibroblasts	0	0	0	0	0	645	7	0	0	0
Cells - EBV-transformed lymphocytes	0	0	0	0	0	327	0	0	0	0
Cervix - Ectocervix	0	10	0	0	0	0	4	10	0	0
Cervix - Endocervix	0	23	0	0	0	0	0	0	0	0
Colon - Sigmoid	1	13	0	0	0	0	394	0	0	11
Colon - Transverse	0	5	0	0	0	0	353	0	0	121
Esophagus - Gastroesophageal Junction	3	11	0	0	0	0	364	0	0	25
Esophagus - Mucosa	0	4	0	0	0	0	15	578	0	17
Esophagus - Muscularis	2	5	0	0	0	0	524	0	0	30
Fallopian Tube	0	21	0	0	0	0	8	0	0	0
Heart - Atrial Appendage	0	0	0	0	0	0	11	0	0	450
Heart - Left Ventricle	0	0	0	0	0	0	7	0	0	445
Kidney - Cortex	0	6	0	0	0	0	23	0	0	75
Kidney - Medulla	0	1	0	0	0	0	6	0	0	4
Liver	0	0	0	0	0	0	3	0	0	259
Lung	0	94	0	0	0	0	510	0	0	0
Minor Salivary Gland	0	4	0	0	0	0	117	26	0	34
Muscle - Skeletal	0	0	0	0	0	0	26	0	0	792
Nerve - Tibial	1	618	0	0	0	0	51	0	0	0
Ovary	1	176	0	0	0	0	16	0	0	0
Pancreas	0	1	0	0	0	0	3	0	0	358
Pituitary	0	272	1	0	0	0	35	0	0	5
Prostate	2	140	0	0	0	0	134	0	0	6
Skin - Not Sun Exposed (Suprapubic)	0	1	0	0	0	0	8	642	0	0
Skin - Sun Exposed (Lower leg)	0	5	0	0	0	0	18	731	0	0
Small Intestine - Terminal Ileum	0	34	0	0	0	0	148	0	0	25
Spleen	0	58	0	0	0	0	218	0	0	1
Stomach	0	3	0	0	0	0	179	0	0	225
Testis	0	41	0	369	0	0	4	0	0	0
Thyroid	0	646	0	0	0	0	37	0	0	1
Uterus	12	122	0	0	0	0	19	0	0	0
Vagina	1	51	0	0	0	0	31	87	0	0
Whole Blood	0	0	0	0	0	0	1	0	690	112

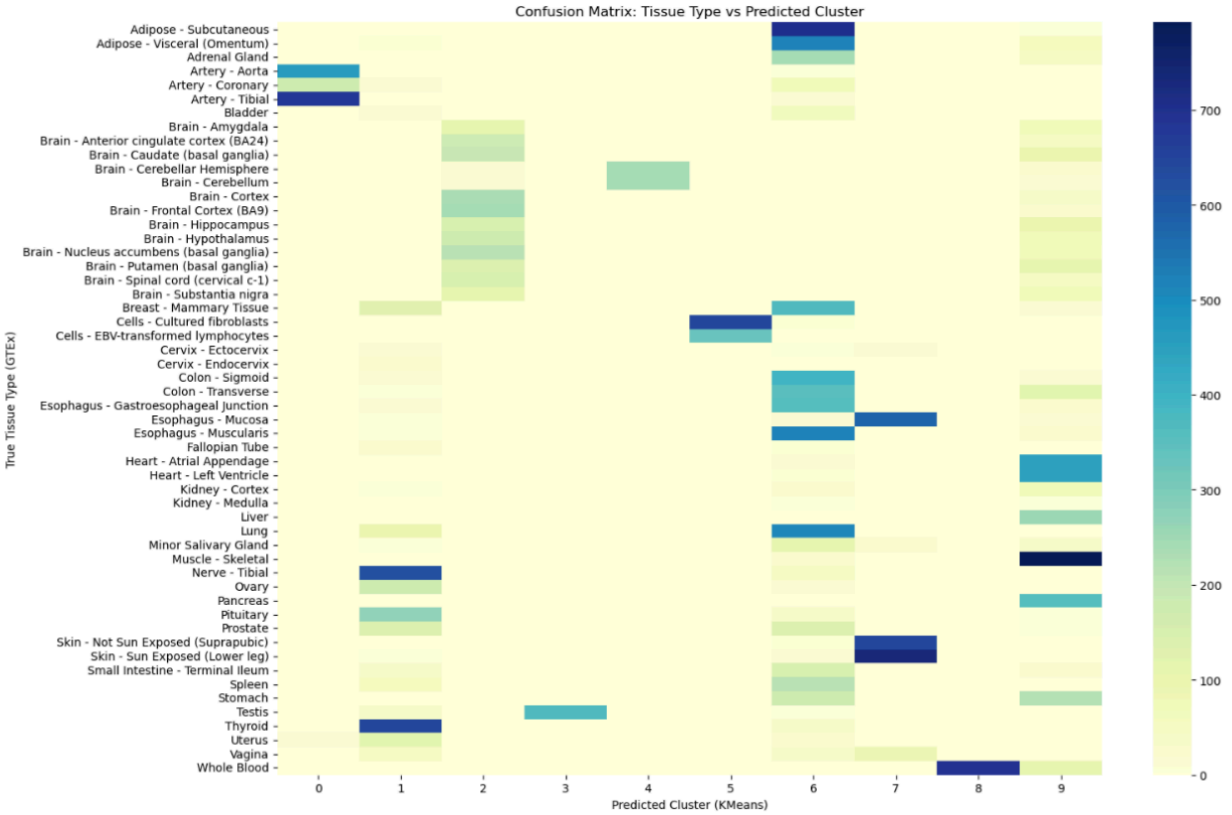
6.2 Confusion Matrix:

The confusion matrix visually maps the relationship between true tissue types (from GTEx metadata) and the predicted clusters assigned by the K-Means algorithm. Each row represents a tissue type, and each column corresponds to one of the ten clusters. The color intensity of each cell reflects the number of samples that fall into a particular tissue-cluster pair, darker cells indicate higher counts. This heatmap reveals how well the clustering preserved biological groupings: when most of a tissue's samples fall into a single cluster, it suggests high cluster purity and a strong match between the unsupervised model and the underlying biology.

The matrix highlights clear one-to-one relationships for certain tissues, such as skeletal muscle, testis, thyroid, and nerve, which appear as bright blocks along individual cluster columns, indicating strong internal consistency. Conversely, tissues like the esophagus, lung, and several brain regions show

more distributed patterns across multiple clusters, reflecting biological complexity or overlapping gene expression profiles. This visualization reinforces earlier metrics such as ARI and NMI, providing an intuitive view of clustering quality and confirming that the model, despite being unsupervised, was able to recover meaningful tissue-level structure in the gene expression data.

The below matrix confirms that distinct organs or tissues with strong functional identity (e.g., testis, thyroid, muscle) are easily separated. In contrast, functionally similar or histologically adjacent tissues (e.g., digestive tract regions or various brain areas) may present overlapping expression patterns, making separation more difficult without supervised fine-tuning. This aligns with known biology, some tissues have a unique transcriptomic identity, while others are transcriptomically complex or similar due to shared function or proximity.

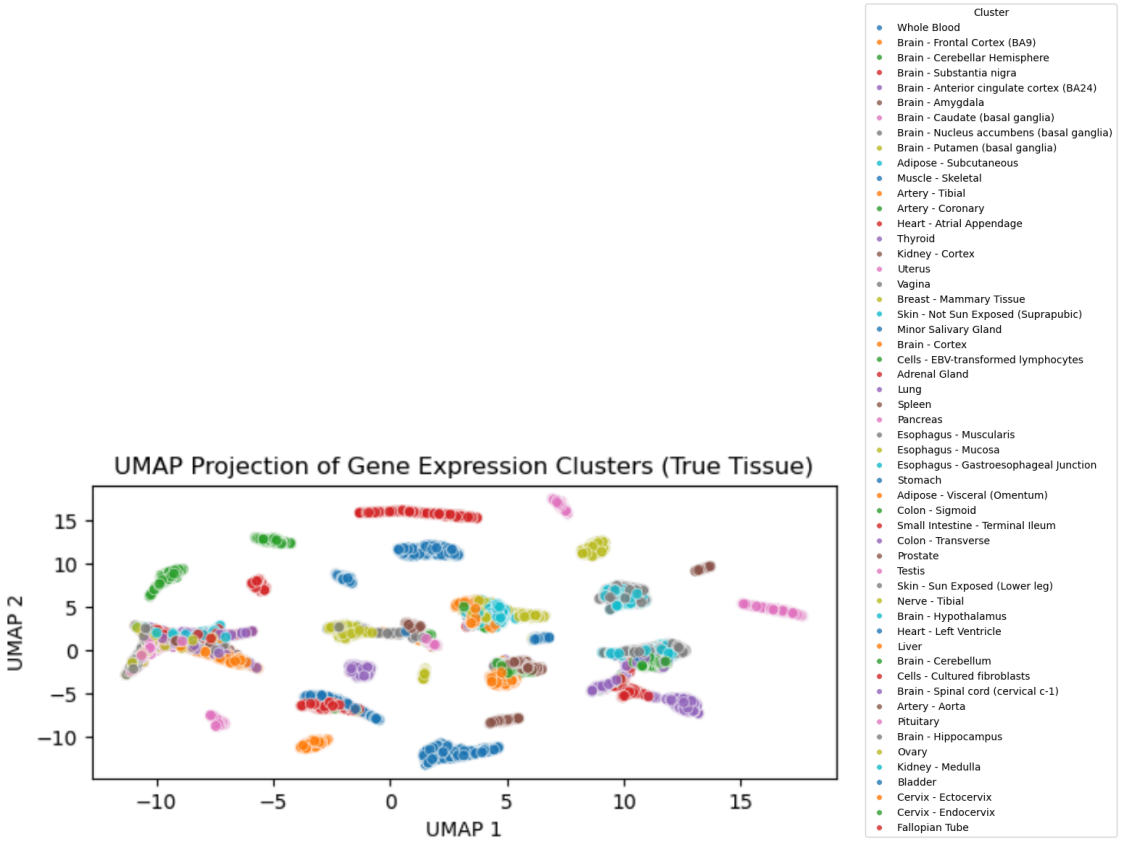


6.3 UMAP Visualization Analysis: Mapping Gene Expression Clusters

The UMAP (Uniform Manifold Approximation and Projection) visualization provides a compelling two-dimensional representation of gene expression patterns across GTEx tissue samples. Each point corresponds to a sample, color-coded by its true tissue label, and the plot reveals distinct, compact clusters for several tissue types, most notably skeletal muscle, testis, thyroid, nerve, and whole blood. These tightly grouped and spatially separated clusters suggest strong transcriptomic identity and low intra-tissue variation, supporting earlier metrics such as high cluster purity and strong NMI scores. The spatial distance between clusters further reflects biological dissimilarity, while proximity between certain tissue types (e.g., artery and heart tissues) highlights underlying similarities in their gene expression profiles.

In contrast, some tissues, particularly brain subregions, skin, and esophageal samples, appear partially overlapping or loosely clustered. This reflects either transcriptomic similarity due to shared function or structural complexity within those tissues. Despite these overlaps, the overall UMAP

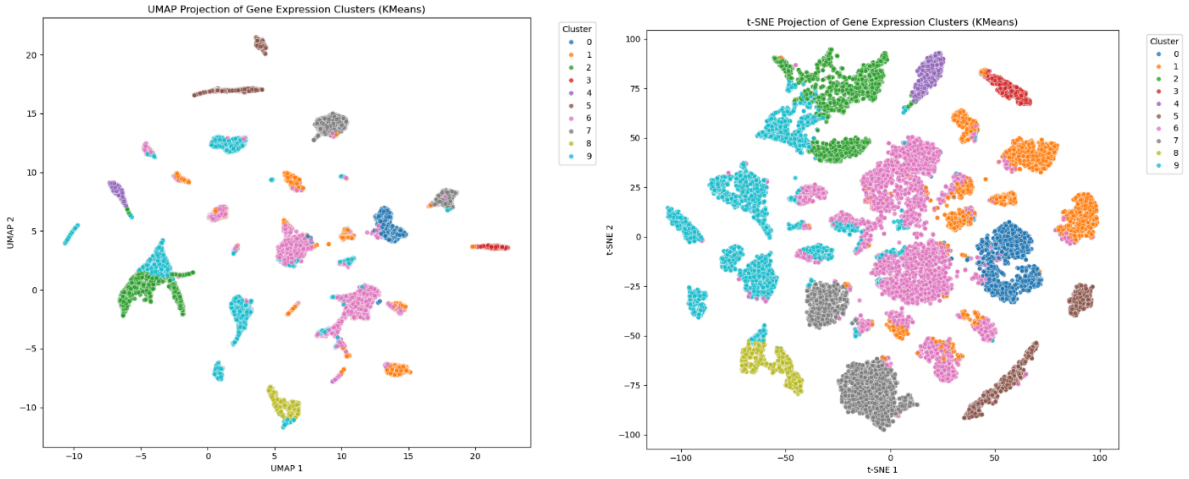
structure effectively captures both global and local relationships, confirming that meaningful biological patterns are preserved even in unsupervised, label-free settings. Thus, UMAP acts as a visual validation tool that reinforces the model's ability to uncover latent tissue-specific gene expression structure.



6.4 UMAP and t-SNE Projections

The UMAP and t-SNE plots shown demonstrate how nonlinear dimensionality reduction techniques can effectively reveal structure in high-dimensional gene expression data. In both visualizations, samples are projected into a two-dimensional space where distinct tissues form visibly compact clusters, particularly for skeletal muscle, testis, thyroid, nerve, and whole blood. This clear separation indicates that these tissues have unique and consistent transcriptomic profiles, which are successfully captured by the unsupervised clustering process. UMAP, in particular, maintains both global and local relationships, allowing biologically similar tissues to appear nearby while preserving distinct boundaries between unrelated ones.

t-SNE, on the other hand, is optimized for preserving local similarity, which makes it especially useful for highlighting small-scale differences among samples. While t-SNE does not retain the global structure as effectively as UMAP, it still reveals tightly knit groupings and subtle intra-tissue distinctions. Tissues with overlapping expression profiles, such as brain subregions or epithelial tissues, show some degree of cluster diffusion in both methods, reflecting biological complexity rather than modeling limitations. Combined, UMAP and t-SNE offer complementary views of the dataset, validating the clustering results and reinforcing the conclusion that meaningful tissue-specific patterns exist within the gene expression landscape.



7. Conclusion

In conclusion, this study demonstrates that unsupervised machine learning methods, when applied to high-dimensional RNA-seq gene expression data, can effectively uncover biologically meaningful clusters corresponding to human tissue types. Through the use of PCA for dimensionality reduction, followed by clustering algorithms like K-Means, GMM, and DBSCAN, we observed varying degrees of alignment with known tissue labels, supported by evaluation metrics such as ARI, NMI, and cluster purity. Visual tools like UMAP and t-SNE further reinforced the model's ability to reveal latent structure, showing clear separations for many tissue types and meaningful proximities among related ones. These findings highlight the potential of unsupervised learning to explore complex biological datasets without prior labeling, offering valuable insights for tissue classification and transcriptomic analysis.