Advanced Systems Physiology - Guided Project

Océane Li

Master of Bioinformatics & Modelling, Sorbonne University, Paris, France

November 23, 2024

Abstract

Understanding cellular diversity and differentiation is crucial for unraveling the complexity of tissue development and disease mechanisms. In this study, we perform a comprehensive single-cell RNA sequencing (scRNA-seq) analysis to investigate cellular heterogeneity in a biological model (*Drosophila melanogaster*, focusing on the identification and characterization of distinct cell types. Using Principal Component Analysis (PCA) with varying numbers of principal components (PC), we explored the impact of dimensionality reduction on clustering results, finding that 30 PC provided the optimal balance between cluster resolution and biological coherence. UMAP visualization revealed eight distinct clusters corresponding to specific cell types, including muscle cells, muscle precursors, tendon cells, epithelial cells, and precursor cells of external sensory organs. Manual annotation based on marker genes confirmed the consistency of cluster identities, with notable gene expression patterns such as high log2 fold changes and low adjusted p-values for some markers (*eyg, Him, cpo*). Although some clusters showed overlap in marker gene expression, the clusters remained robust across different PCA dimensions. Our findings highlight the importance of dimensionality selection in clustering accuracy and provide insights into the molecular signatures of key cell types.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has become a revolutionary tool for studying cellular heterogeneity, enabling gene expression profiling at the individual cell level. This technology provides unprecedented insights into tissue complexity by revealing distinct cellular populations, their functional roles, and their interactions within heterogeneous environments. Unlike bulk RNA-seq, which measures the average gene expression across multiple cells, scRNA-seq uncovers the molecular variability between individual cells, making it an indispensable approach for studying developmental processes, disease mechanisms, and cellular responses to various stimuli.

In this study, we apply scRNA-seq to explore cellular diversity in *Drosophila melanogaster*, a model organism widely used in genetic and developmental research. The goal of this analysis is to identify and characterize distinct cellular populations within Drosophila tissues, based on their gene expression profiles. Using advanced computational methods, we aim to classify cells into biologically meaningful groups and highlight their associated molecular signatures. This study provides an in-depth analysis of the cellular landscape in Drosophila, contributing to a better understanding of its developmental biology and offering new insights into cellular diversity, which could inform broader studies on gene regulation and tissue complexity.

2 Materials & Methods

2.1 Data filtering

The unpublished raw data is provided in the form of a sparse matrix, in the dgCMatrix format. This structure, particularly suited for datasets with numerous zero values, contains 17,753 genes and 2,331,657 barcodes. To prepare the data for analysis, a Seurat object was created using the `CreateSeuratObject` function. This step included an initial filtering to exclude cells expressing fewer than 200 genes (to remove debris or under-sequenced cells) and to eliminate genes expressed in fewer than 5 cells (to discard uninformative genes). After this filtering, the Seurat object contains 11,752 genes and 43,688 cells, ensuring that only biologically relevant entities are retained.

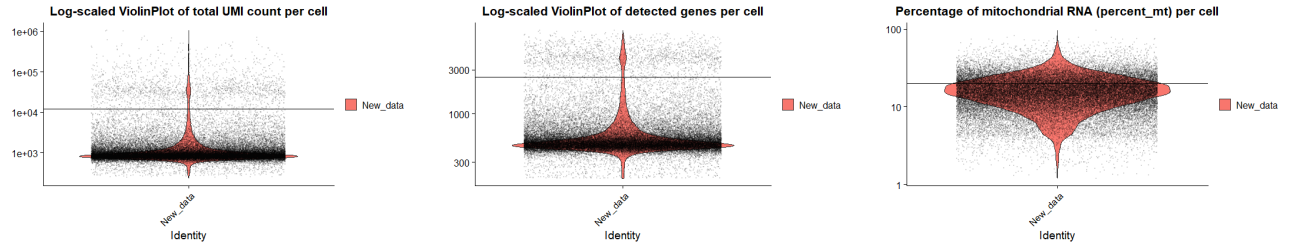
The genes in the raw data were identified by standardized but biologically unintuitive identifiers. To facilitate functional analysis and integration with external databases, these identifiers were converted into gene symbols using the `biomaRt` package. Gene in-

formation for *Drosophila melanogaster* was retrieved from the `dmelanogaster_gene_ensembl` dataset, and a mapping table (`gene_mapping`) was generated, linking each Ensembl identifier (`ensembl_gene_id`) to its corresponding gene symbol (`external_gene_name`). The gene names in the Seurat object were then updated: when a corresponding gene symbol was available, it replaced the original identifier; otherwise, the Ensembl identifier was retained (ex: `FBgn0031081` becomes `Nep3`).

To gain an initial understanding of the dataset, a preliminary descriptive analysis was performed to assess data quality and identify potential anomalies prior to specific processing steps. The metadata associated with the Seurat object provide global information for each cell, including the total number of RNA molecules detected (UMI) and the number of genes expressed.

The total number of UMI (Unique Molecular Identifiers) per cell reflects the overall transcriptional activity of each cell. The distribution of UMI was examined using a histogram with 200 intervals to identify potential technical or biological biases, such as abnormally high or low transcription levels. Additionally, a logarithmic version of the histogram was generated to better visualize variations in the data. The choice of 200 intervals was made to ensure sufficient granularity for detecting subtle distribution patterns.

Similarly, the distribution of the number of genes expressed per cell was analyzed. This metric reflects the transcriptional diversity of each cell and is critical for identifying low-complexity cells (potential debris) or highly complex cells (potentially biased). These analyses provide a comprehensive overview of the global properties of the data, helping to detect potential technical or biological artifacts. The observations will inform subsequent steps of data normalization and processing.



(a) Distribution de `nCount_RNA` (b) Distribution de `nFeature_RNA` (c) Distribution de `percent_mt`

Figure 1: **Quality assessment of scRNA-seq data using key metrics:** (a) Distribution of the total number of UMIs per cell, with a minimum threshold of 12,000 UMIs to exclude empty droplets and low-quality cells. (b) Distribution of the number of detected genes, with a minimum threshold of 2,500 genes to remove uninformative cells. (c) Percentage of mitochondrial RNA per cell, with a maximum threshold of 20% to exclude dead cells. These criteria ensure the removal of irrelevant cells prior to the main analysis.

Subsequently, we performed cell filtering to retain only those likely to contain reliable information for the analysis. This process is essential in single-cell RNA sequencing (scRNA-seq) analysis to eliminate low-quality cells, such as empty droplets, dead cells, duplicates, or damaged cells. Several criteria were used for this filtering. First, the total number of UMI per cell was used as a key indicator of sequencing data quality. A low UMI count may indicate an empty droplet (containing ambient RNA), whereas an excessively high UMI count can suggest the presence of doublets or multiplets (droplets containing two or more cells). To visualize this distribution, a violin plot was generated for the UMI counts per cell (Figure 1a). Using a logarithmic scale, we visually determined a minimum threshold of 12,000 UMIs to exclude cells suspected to be empty droplets. Additionally, a knee plot was used to validate this threshold. This plot identifies the point at which the distribution curve flattens, providing an objective basis for setting filtering thresholds. Regarding doublets and multiplets, although they are also to be excluded due to their excessive UMI counts, we did not set a maximum threshold at this stage. Instead, an automated doublet identification method will be implemented later in the analysis. The number of genes detected per cell was used as another important quality criterion. A low number of expressed genes often indicates dead cells or empty droplets, whereas an abnormally high number may suggest doublets. Similar to the UMI distribution, a violin plot was used to visualize this metric (Figure 1b). Based on this analysis, a minimum threshold of 2,500 genes was set to eliminate cells with anomalously low values.

The percentage of mitochondrial genes is another quality indicator. Dead or damaged cells often exhibit a high percentage of mitochondrial genes, as they release mitochondrial RNA into the medium. We calculated this percentage for each cell using the `PercentageFeatureSet` function, which determines the proportion of mitochondrial genes among all expressed genes in the cell. A maximum threshold of 20% was set to exclude cells with excessively high mitochondrial expression (Figure 1c). By strictly ap-

plying these thresholds, a new Seurat object was created containing only the cells deemed "high quality". The results of this filtering were visualized using violin plots and histograms, confirming the exclusion of low-quality cells. This preprocessing step is crucial to ensure meaningful analysis results and to avoid biases introduced by poor-quality cells. The aim is to filter out abnormal droplets while preserving as many cells as possible.

To refine our doublet filtering, we used the automated method provided by the `scDblFinder` package, which allows for more precise removal of doublets, sources of artifacts in single-cell analyses. The `scDblFinder` package operates on `SingleCellExperiment` objects, so we first converted the pre-filtered Seurat object into a `SingleCellExperiment object`. Once converted, we used the `scDblFinder` function to predict doublets. This function generates artificial doublets based on random combinations of cells, then calculates a proximity measure in feature space between these simulated doublets and the real cells. Cells that are close to these artificial doublets are classified as doublets. The results of this classification are visualized using a count table to obtain the proportion of cells classified as "doublets" versus "singlets." We found 1795 singlets and 103 doublets using this method. To eliminate these doublets, we filtered the cells, retaining only those classified as "singlets." The distribution of UMI counts and detected genes in the remaining cells was visualized to assess the impact of the filtering. It is important to note that some cells in the G2M phase of division, which exhibit a similar expression profile to doublets due to their high transcriptional activity, may be incorrectly classified as doublets. Normally, the `CellCycleScoring` function is used to assign a cell cycle phase and correct for this issue. However, due to the lack of cell cycle-specific gene lists for Drosophila, this approach could not be implemented in our analysis.

As observed in the previous graphs, the cells do not have the same total number of UMI. This could be due to biological differences (some cells express less RNA than others), but is likely a result of cell-specific sequencing biases (some cells have been sequenced less than others). If these biases are not corrected, they may distort the analysis by attributing higher expression levels to cells that were more heavily sequenced, regardless of actual biological differences. The normalization step aims to correct this bias. To do this, we used the `LogNormalize` method, which assumes that all cells have similar RNA content. This method involves dividing the expression of each gene by the total UMI count of the cell, multiplying by a scaling factor (in this case, the median total UMI per cell), and then applying a log-transformation to the results. This approach standardizes the data and corrects technical biases related to sequencing depth differences across cells, ensuring that the data are comparable between cells.

2.2 Identification of cell populations

Once the data were normalized, the next step involved identifying subpopulations of cells by clustering them based on their gene expression profiles. To ensure biological relevance while optimizing computations, the analysis was restricted to a subset of highly variable genes (HVG). These genes were selected using the `FindVariableFeatures` function in Seurat, applying the variance-stabilizing transformation (VST) method. This approach, specifically suited for single-cell transcriptomics data, stabilizes gene variance by accounting for their mean expression levels, thereby mitigating the influence of genes with very low or very high expression. A total of 2000 genes with the highest adjusted variance were identified for downstream analyses. To validate the relevance of the selected genes in the studied biological context, a targeted verification was performed to confirm the presence of known markers associated with specific cell types (e.g. muscle cells, hemocyte, precursor cells of external sensory organs, etc). All of these markers were detected in the dataset except for "zip," a gene primarily expressed in epithelial cells, whose minimal expression may reflect a low abundance of these cells in the analyzed sample.

Although gene selection already significantly reduces the data size, dimensionality reduction is necessary to capture the most important variabilities. Principal Component Analysis (PCA) is a common method for this purpose. However, before performing PCA, it is important to scale the data. This involves centering each gene around its mean and normalizing its standard deviation to 1. This step ensures that all genes have the same weight during dimensionality reduction and prevents highly expressed genes from dominating the analysis. The `RunPCA` function is used to perform PCA: it is applied to the most variable features, i.e. genes with the largest expression variation between cells, in order to maximize the information captured by PCA. The PCA output shows the genes most associated with each principal component (PC). For example, for the first principal component (PC1), genes such as `sing`, `twi` and `Him` show positive loadings, while genes like `mbl`, `CG1850`, and `CG33978` are associated with negative loadings. For each principal axis, it is useful to analyze the genes contributing the most. This helps identify genes whose expression increases or decreases together. `DimPlot` generates a representation of cells in the PCA space, and the `ElbowPlot` function allows examination of the variance explained by each principal component, aiding in the decision of how many PCs to use for the next analysis step.

After PCA, we need to determine the optimal number of principal components (PCs) to retain for UMAP and clustering. In this case, we test different values of PCs ranging from 10 to 50 in steps of 10. Once this choice is made, cells are grouped into clusters through two steps: the construction of a k-

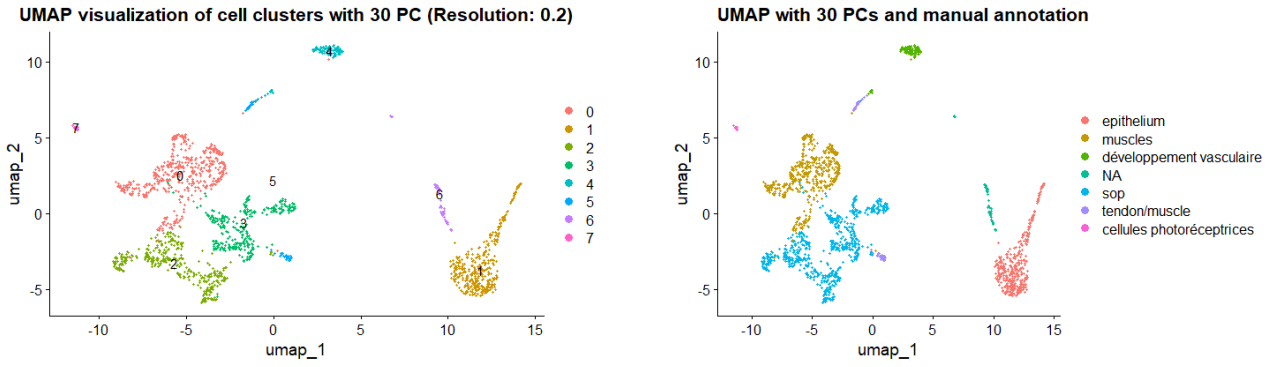
nearest neighbors (KNN) graph, and cluster identification using the Louvain algorithm. The KNN graph connects each cell to its 20 nearest neighbors (default parameter), with distances calculated from the PCA coordinates. Then, the Louvain algorithm generates clusters by optimizing modularity, a measure of the link density within clusters compared to outside them. The resolution of this process determines the number of clusters, which increases with the chosen resolution value. The maximum modularity obtained (e.g., 0.9269 for 10 PCs at a resolution of 0.2) allows for selecting the optimal resolution to use for subsequent steps, including the construction of the UMAP.

UMAP is a visualization tool for the clusters previously formed. To identify cell types or subpopulations, a manual annotation of the clusters was performed based on the expression of marker genes specific to certain cell types. Lists of marker genes associated with various cell types (muscles, tendons, epithelial cells, etc.) were defined to assign expression scores to each cell. To do this, the `AddModuleScore` function from Seurat was used to calculate the average expression score for each group of marker genes. This function requires the following parameters: `features` (list of marker genes), `nbin` (number of bins for normalization, here 5), and `ctrl` (number of control genes, equivalent to the number of markers). Before adding the scores, a check was performed to ensure that all genes from each list were present in the data, and absent genes were removed. Lists with only one or no genes present in the data were excluded from the analysis. The resulting expression scores were then added to the Seurat object, and visualizations were generated using the `VlnPlot` and `FeaturePlot` functions. These plots allow the distribution of marker gene expression within the clusters to be observed, thereby facilitating the identification of cell types based on their genetic signature.

In addition to the approach based on predefined marker gene lists (`marker_genes`), we sought to identify differentially expressed genes for each cluster in the dataset. This analysis was performed using the `FindMarkers` function from the Seurat package, which detects gene expression differences between a target cluster and the remaining cells. The goal was to identify statistically significant marker genes for each cluster, characterized by high expression specific to the target cluster, to associate clusters with cell types or biological functions. For each cluster, `FindMarkers` was used to compare gene expression within a given cluster (`ident.1 = cluster`) against all other clusters. A minimum expression proportion threshold (`min.pct = 0.25`) was applied to include only genes expressed in at least 25% of the cells within the target cluster. This threshold ensures that the identified genes are biologically relevant and not artifacts of stochastic expression.

Differentially expressed genes were subsequently filtered based on two criteria: only genes with a log2FC > 2, corresponding to at least a fourfold higher expression in the target cluster, were retained. Additionally, only genes with an adjusted p-value < 0.05, corrected for the false discovery rate (FDR), were considered statistically significant. The retained genes were organized into cluster-specific dataframes, including information such as log2FC, adjusted p-value, and the associated cluster. These analyses enabled the molecular profiling of clusters and the identification of robust markers, essential for annotating cell types and validating clustering results.

In a second analysis, this approach was repeated, restricting the study to genes included in predefined lists of known cell type markers (`marker_genes`). This list directly links the identified genes to specific cell types. The same criteria for log2FC and adjusted p-value were applied.



(a) UMAP of clusters built with 30 PC

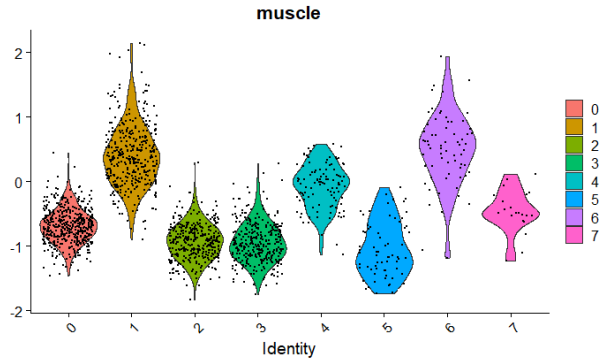
(b) UMAP of annotated clusters with 30 PC

Figure 2: UMAP representations showing the clustering results. (a) UMAP of clusters constructed using 30 principal components (PC). (b) UMAP with annotated clusters based on gene marker expression.

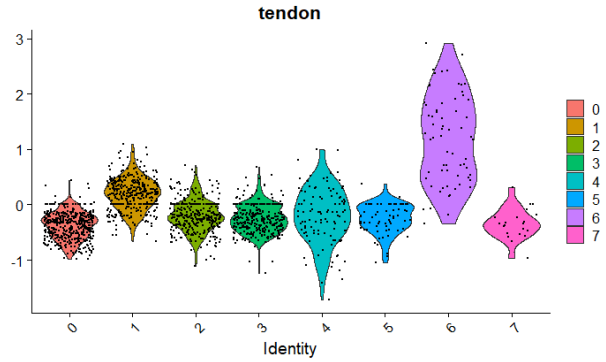
3 Results

The results of the descriptive analysis of the scRNA-seq data show significant variability in the number of UMI (RNA molecules) and the number of genes detected per cell. For the number of UMI (`nCount_RNA`), the minimum observed value is 234, while the maximum reaches 1,055,142. The median is 902, and the mean is 3,900, suggesting a skewed distribution with cells exhibiting a high number of UMIs. Regarding the number of detected genes (`nFeature_RNA`), values range from 199 to 8,029, with a median of 486 and a mean of 732. This broad range of detected genes reflects the diversity of cells in the sample, with some cells being less informative with a limited number of detected genes, while others contain a greater genetic diversity. These results highlight the need to filter cells based on these criteria to remove those that are uninformative or of poor quality.

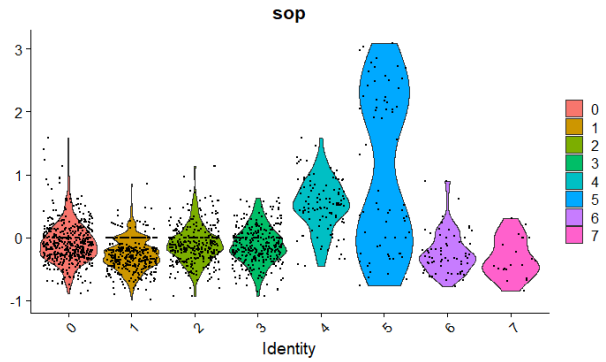
After comparing the results obtained using different numbers of principal components (PC), ranging from 10 to 50, we determined that using 30 PCs provides the most optimal representation for our clustering analysis. While the clusters identified with 10, 20, and 30 PC were similar in their general structure and the cell types represented, the use of 30 PC yielded sharper cluster boundaries, enabling better visual separation. Notably, the boundaries between clusters became more distinct, suggesting improved resolution in cell organization. The UMAP results revealed eight major clusters corresponding to specific cell types, including muscle cells, muscle precursors, tendon cells, epithelial cells, and sensory organ precursors, with a slightly adjusted distribution compared to other PC values (Figure 2a).



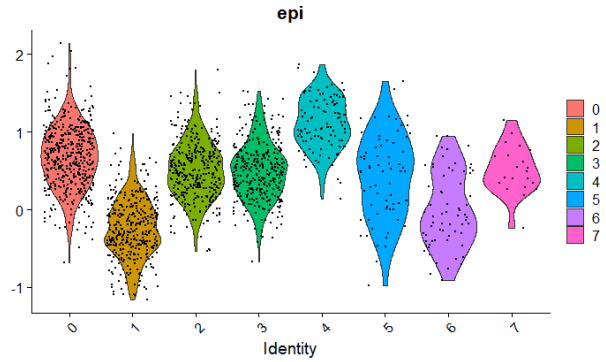
(a) Violin plot of muscle genes expression by cluster



(b) Violin plot of tendon genes expression by cluster



(c) Violin plot of sop genes expression by cluster



(d) Violin plot of epithelium genes expression by cluster

Figure 3: **Violin plots showing the expression of specific genes in different clusters.** (a) Expression of muscle genes across clusters, illustrating the number of cells expressing muscle genes per cluster. (b) Expression of tendon genes across clusters, showing the number of cells expressing tendon genes per cluster. (c) Expression of external sensory organs genes across clusters, illustrating the number of cells expressing sensorial genes per cluster. (d) Expression of epithelium genes across clusters, showing the number of cells expressing epithelium genes per cluster.

The manual annotation of clusters using marker gene lists revealed consistent gene expression profiles for each cluster, regardless of the number of PCs used (whether assessed through visual analysis of violin plots (Figure 3) or via `FindMarkers`). For instance, cluster 0 is consistently associated with epithelial genes (Figure 3d), notably *eyg*, whose expression is significantly differentiated with a high log2FC (2.06) and a very low adjusted p-value (1.59e-97). Muscle clusters, primarily observed in clusters 1 and

	p_val <dbl>	avg_log2FC <dbl>	pct.1 <dbl>	pct.2 <dbl>	p_val_adj <dbl>	cluster <chr>
vvl	1.560035e-140	3.151531	0.927	0.416	1.833353e-136	2
Cpr78E	5.545746e-139	5.560066	0.720	0.142	6.517361e-135	2
CG13053	1.502545e-101	2.622016	0.986	0.649	1.765790e-97	2
Dr	1.420341e-93	3.141058	0.709	0.204	1.669185e-89	2
nw	1.074316e-80	2.163286	0.912	0.486	1.262537e-76	2

Figure 4: Dataframe showing the top marker genes for Cluster 2 identified using the `FindMarkers` function. Genes with an average log2 fold change (`avg_log2FC`) greater than 2 and an adjusted p-value (`p_val_adj`) less than 0.05 were retained. This figure displays the most significant marker genes for Cluster 2 that meet these thresholds.

6 Figure 3a, are characterized by the expression of genes such as *Him, htl, Pdp1, sls*, and *Grip*, which exhibit high log2FC values and very low adjusted p-values. These observations remain consistent across clusterings performed with other PC values, yielding highly similar results.

An inversion of cluster identifications was observed between analyses using 10/20 PCs and those with 30/40/50 PCs: clusters 1 and 6 in the PC30 analysis correspond to clusters 2 and 5 in the PC10/PC20 analyses, respectively. However, this inversion did not affect the gene composition of the clusters, which remains consistent, suggesting that this variation in cluster assignment is primarily due to the UMAP projection. Analyses with 40 and 50 PCs yield nearly identical results to those obtained with 30 PCs. Additionally, cluster 4 consistently remains associated with precursor genes of external sensory organs (Figure 3c), notably *cpo*. Tendon-associated genes, such as *vkg, Hand, kon*, and *Alk*, are shared within cluster 6 alongside muscle-related genes (Figure 3b). Other clusters maintain specific gene expression profiles, and the distribution of cell types within each cluster remains stable regardless of the PC value used.

It is worth noting that among the subpopulation of cells retained after the various selection steps, almost no cells express characteristic hemocyte genes (*eater, et, NimC4, srp*), likely due to their absence or very low abundance in the initial population, or potentially due to suboptimal data partitioning. Furthermore, although precursor muscle genes were not identified under the applied logFC and adjusted p-value thresholds, some cells (notably in cluster 1) exhibit expression of these genes.

The analysis based on predefined marker gene lists did not allow for a clear assignment of cell types to clusters 2, 3, and 7. Therefore, we focused on the differentially expressed genes within each cluster. In cluster 2, genes such as *vvl, Cpr78E, CG13053, Dr*, and *nw* (Figure 4) show strong differential regulation (log2FC from 2.1 to 5.5, adjusted p-value $< 10^{-76}$). For example, *vvl* is expressed in 92.7% of the cells in cluster 2 (pct.1), compared to 41.6% in other clusters, suggesting involvement in biological processes partially shared with other cell types. In cluster 3, marker genes show moderate log2FC (2.6 to 3.6) and are expressed in less than 75% of the cells in the cluster, which may indicate an intermediate or heterogeneous cell population. In contrast, cluster 7 is characterized by highly specific genes (*CG5653, lz, gl*) expressed in nearly 100% of the cells in the cluster but rarely elsewhere (pct.2 $< 2\%$). The high log2FC values (up to 8 for some genes) indicate marked regulation and a unique cell identity.

Thus, after comparing analyses using between 10 and 50 PCs, the choice of 30 PCs appears optimal, providing better separation of the clusters while maintaining coherent and biologically meaningful gene expression profiles.

4 Discussion

The principal component-based clustering analysis revealed distinct cellular structures across a wide range of PC values (10 to 50). Among these configurations, the use of 30 PCs was identified as the most optimal, providing the best separation of clusters while maintaining consistency in gene expression profiles. Although analyses with 10 and 20 PCs yielded similar results in terms of the overall cluster structure, increasing the number of PCs allowed for better resolution of cluster boundaries, particularly for clusters 1 and 6, suggesting an improved underlying cellular organization. This observation aligns with previous work, which demonstrates that using an appropriate number of PCs enhances the discriminative capacity of UMAP for separating cell populations while reducing noise (McInnes et al., 2018 [1]).

The analysis of marker genes for each cluster revealed consistency in gene expression, regardless of the number of PC used. For example, epithelial cell-specific genes, such as *eyg*, were consistently associated with cluster 0, with significantly differentiated expression compared to other clusters, confirming that gene expression can serve as a reliable indicator of cell type in clustering analyses. Similarly, muscle clusters, primarily represented by clusters 1 and 6, showed characteristic expression profiles for muscle-specific genes such as *Him, htl*, and *Pdp1*. This con-

sistency of results across different PC numbers suggests that gene expression plays a major role in the structuring of clusters.

The stability of gene expression profiles across different numbers of PCs also suggests that the optimal number of PCs (30) is large enough to capture essential biological variations while being sufficiently reduced to avoid dimensionality overload. The absence of cells expressing characteristic hemocyte genes likely results from their low abundance in the initial population or from inadequate data partitioning. In some cases, under-sampling of rare cell populations may lead to their exclusion during analysis, which is often the case in transcriptomic experiments where rare cells can be obscured by background noise. To better capture rare populations, it would be worthwhile to explore the use of oversampling techniques or specialized tools to detect less abundant cell types. Approaches such as t-SNE could allow for better separation and identification of these populations. Additionally, the integration of multi-omic data (such as transcriptomics and proteomics) may help overcome this limitation by providing complementary information.

Moreover, clusters 2, 3, and 7 were difficult to annotate using predefined marker genes, likely due to the heterogeneity of these cell populations. Such heterogeneity is common in single-cell transcriptomic analyses, where intermediate or transient populations may exhibit less distinct gene expression profiles, as observed in cluster 3 (Kharchenko et al., 2014 [2]). In contrast, cluster 7 is distinguished by highly specific genes, which are expressed in nearly all cells of the cluster and rarely in other clusters, indicating a clearly defined cellular identity. The gene *vvl* (ventral veinless), expressed in 92.7% of cluster 2, is essential for vascular morphogenesis during development in Drosophila. *vvl* is also involved in Wnt signaling pathways and can influence cell growth and endothelial cell proliferation in various tissues [3][4], suggesting that this cluster may contain endothelial cells or those regulating vascular development. Finally, the gene *gl* (Glass), expressed in cluster 7, represents a key transcription factor in the regulation of photoreceptor cell differentiation in the eyes of *Drosophila melanogaster*, further reinforcing the hypothesis of a unique functional identity for this cluster. This gene is known to play a pivotal role in the formation and differentiation of photoreceptor cells in Drosophila [5][6].

Thus, these results provide important insights into cellular structure and genetic organization in this biological model (Figure 2b), paving the way for future investigations into the function of specific genes in each cluster.

References

[1] E. Becht, L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, LG. Ng, F. Ginhoux, and E.W. Newell. Dimensionality reduction for visualizing single-cell data using umap. Nature Biotechnology, 36(1):41–48, 2018. https://doi.org/10.1038/nbt.4314.

[2] P.V. Kharchenko, L. Silberstein, and D.T. Scadden. Bayesian approach to single-cell differential expression analysis. Nature Methods, 11(7):740–742, 2014. https://doi.org/10.1038/nmeth.2967.

[3] National Center for Biotechnology Information. vvl ventral veins lacking [drosophila melanogaster (fruit fly)], 2024. Gene ID: 38752, updated on 2-Nov-2024.

[4] J. F. de Celis, M. Llimargas, and J. Casanova. Ventral veinless, the gene encoding the cf1a transcription factor, links positional information and cell differentiation during embryonic and imaginal development in drosophila melanogaster. Development, 121(10):3405–3416, 1995. https://doi.org/10.1242/dev.121.10.3405.

[5] National Center for Biotechnology Information. gl glass [drosophila melanogaster (fruit fly)], 2024. Gene ID: 42210, updated on 2-Nov-2024.

[6] K. Moses, M. Ellis, and G. Rubin. The glass gene encodes a zinc-finger protein required by drosophila photoreceptor cells. Nature, 340:531–536, 1989. https://doi.org/10.1038/340531a0.