PRACTICE 1. Understanding the role of cell-type specific factors

The aim of the experiment was to understand the role of several transcription factors. To achieve this goal, the experiment performed consisted in studied the following transcriptome profiles: wild type (J0571), mutant short root (shrJ0571) and mutant short root complemented with several transcriptional factors (shrJ0571 + TFs).

Question 1. Calculate PCA among ground tissue samples for shr mutant, the wild type and the complemented lines with the transcription factors BLUEJAY (BLJ), JACKDAW (JKD), MAGPIE (MGP), NUTCRACKER(NUC), IMPERIAL EAGEL (IME) and SCARECROW (SCR).

Principal Component Analysis (PCA) is a technique to reduce data dimensionality. It consists in the transformation of the set of original variables into another set of variables called principal components obtained as a linear combination of the originals. The principal components retain the variability of the original variables and only a few components are necessary to reasonably represent and explain the original data without a significant loss of information.

Transcriptome profiling data has a large dimensionality and applying PCA technique helps to manage, visualize and explain it. PCA variance captured for each component is shown in **Figure 1.1**. Most of the variance is explained by component 1, which captures the 93.36% of it.

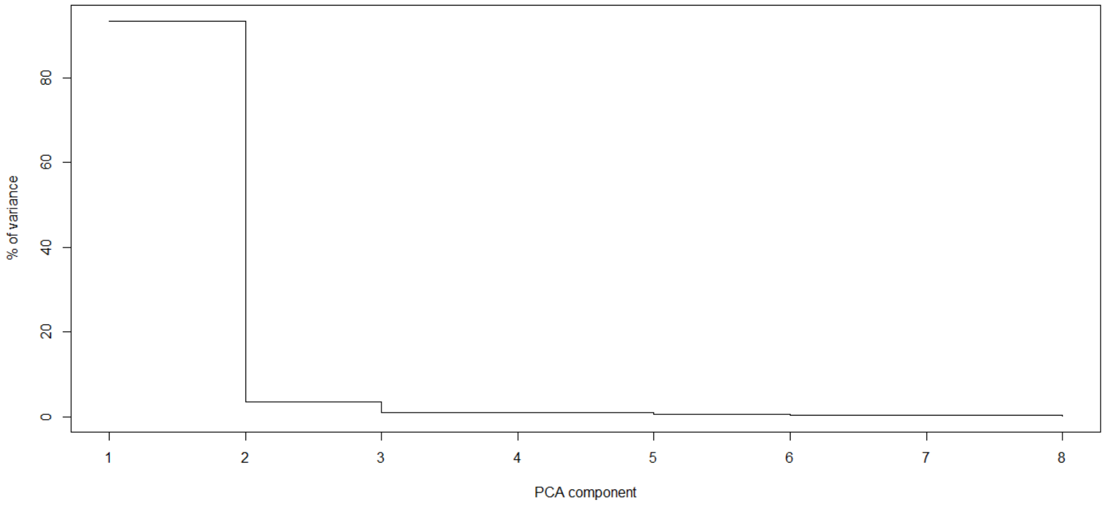

Figure 1.1. Variance captured for each PCA component.

PCA loadings for most significant components (component 1 and component 2) are displayed in **Figure 1.2**. The dots represent the loadings for each variable, or in other words, the coefficients of the linear combination of the initial variables from which the principal components, in this case component 1 and component 2, are constructed.

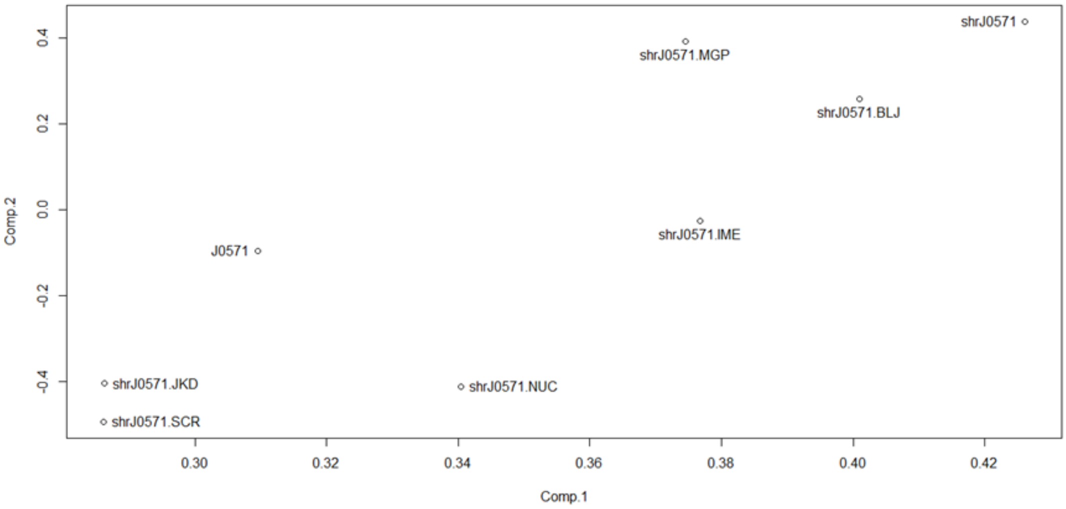

Figure 1.2. PCA loadings for most significant components (component 1 and component 2).

From the distribution of loadings, we can infer that the samples can be clearly distinguished by their variability. However, the samples which correspond to the mutant complemented with JKD and SCR are not being separated by the component 1, so these transcriptional factors may play similar effects on the transcriptome and their function is related. Nonetheless, this hypothesis should be further investigated. In addition, SCR and JKD complemented lines are those whose transcriptome variability are the closest to the wild type (J0571).

Question 2. Create intermediate transcriptomes between shr mutant and the wild type (J0571) which represent 25%, 50% and 75% of recovery (complementation) in gene expression. Recalculate PCA. Plot component variance and loadings. Do these transcription factors recover identity of mutant as compared to the wild type? What role may you establish for these transcription factors?

The intermediate transcriptomes between mutant and wild type were calculated and the PCA was performed again. The variance is shown in **Figure 1.3**. As it can be expected, most of the variance is explained by component 1, which in this case captures 94.52 % of it.

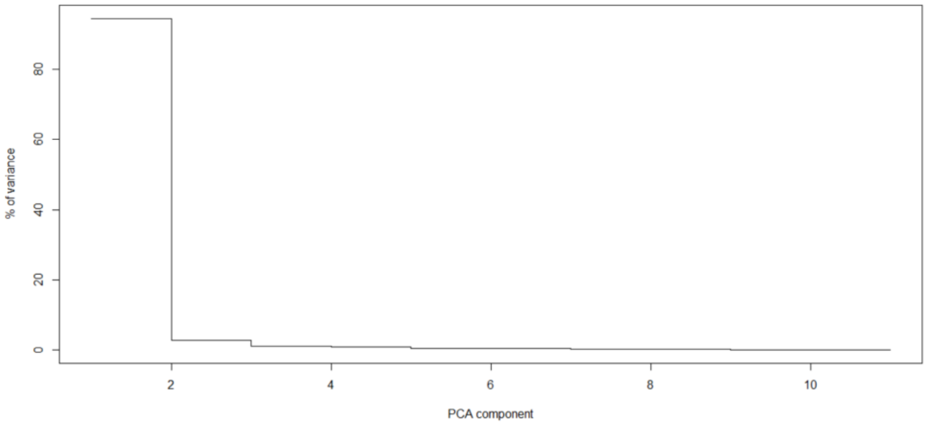

Figure 1.3. PCA variance including intermediate transcriptomes

PCA loadings are shown in **Figure 1.4.** The position for the loadings of the intermediate transcriptomes get closer to the wild type when the percentage of complementation is higher.

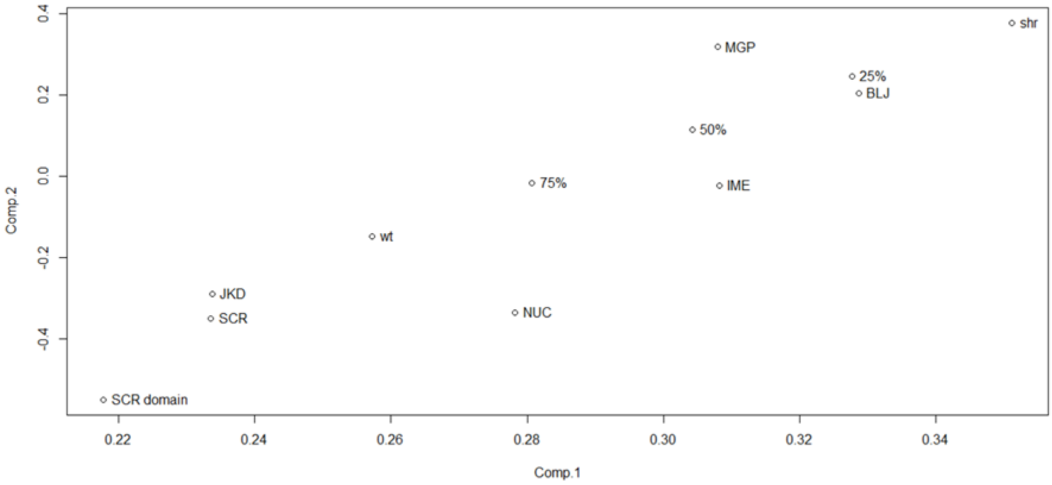

Figure 1.4. PCA loadings for intermediate transcriptomes, wild type and mutant.

Comparing all samples and intermediate transcriptome loadings, it is observed that IME, MGP and BLJ are found between the wild type and mutant, while NUC as well if only the first component is considered. Therefore, these transcription factors are involved in the recovery of identity of mutant up to a certain degree. However, as mentioned before, JKD and SCR are the ones that achieve the highest recovery of the mutant.

Question 3. Add the transcriptome of cells corresponding to SCR domain and recalculate PCA. Plot variance for components and loadings. What might you conclude for several of these transcription factors?

The variance is shown in **Figures 1.5**. As it can be expected, most of the variance is explained by component 1, which in this case captures 93.13 % of it.

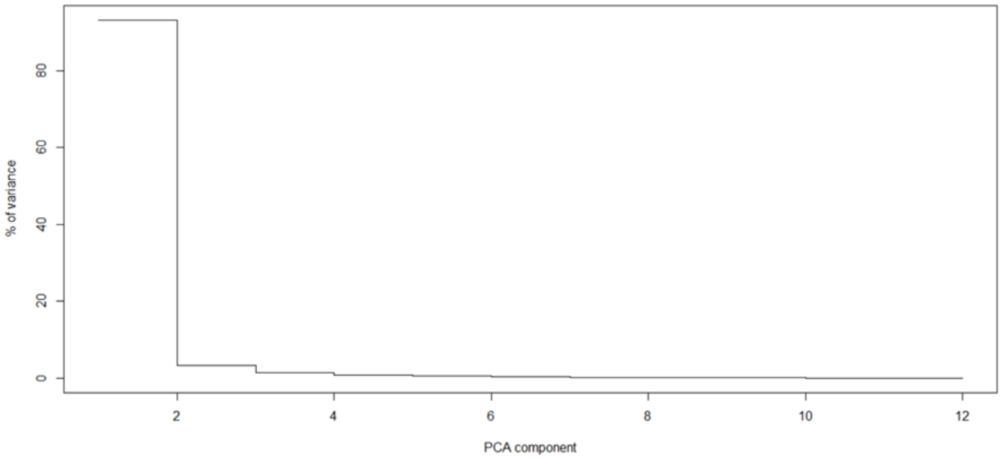

Figure 1.5. PCA components including intermediate transcriptomes and SCR domain.

PCA loadings are shown in **Figure 1.6**. The complemented lines with SCR and JKD transcriptional factors are the ones closer to SCR domain. Therefore, it is possible that these factors are involved in functions specific from this domain. SCR and JKD transcription factors would induce the transcription of specific genes from SCR domain.

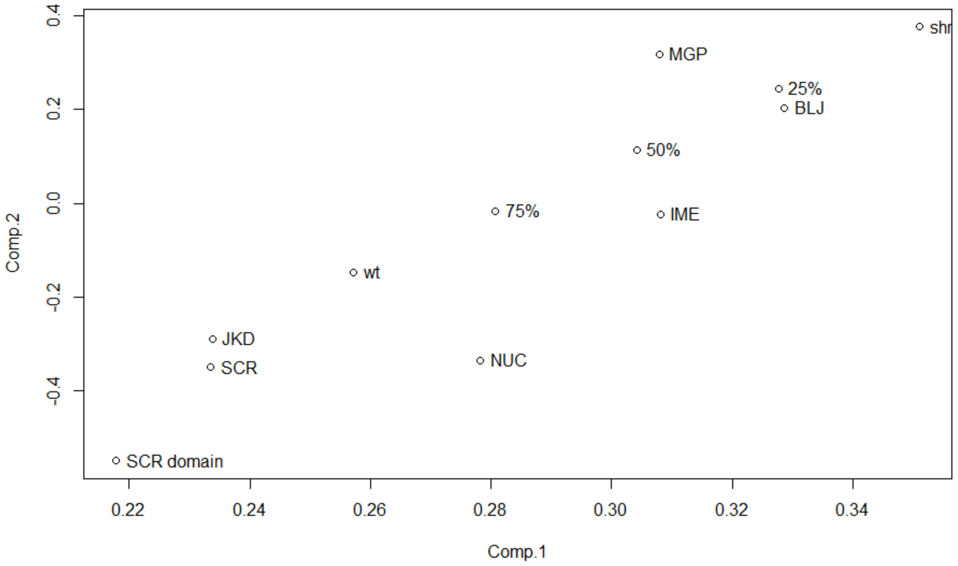

Figure 1.6. PCA loadings for intermediate transcriptomes, wild type, mutant, complemented lines and SCR domain.

Question 4. Find the most important genes which contributing to observed transcriptomic changes. What do you observe?

The most important genes contributing to observed transcriptomic changes were extracted according to the scores of the component 1 (because it explains most of the variance) for the first PCA performed. The heatmap is shown in **Figure 1.7**, displaying the genes with the lowest expression in the first half and the ones with the highest expression in the second half. The genes which are highly expressed in the wild type (J0571) have a lower expression in the mutant lines, even when complemented with transcriptional factors. The genes which have a low expression in the wild type (J0571), have a higher expression in the mutant line (shrJ0571) and supplemented with BLJ and MGP transcriptional factors.

The mutant complemented lines with SCR and JKD have the most similar expression pattern to the wild type. This observation is in accordance with which has been observed before analyzing PCA loadings for these transcription factors.

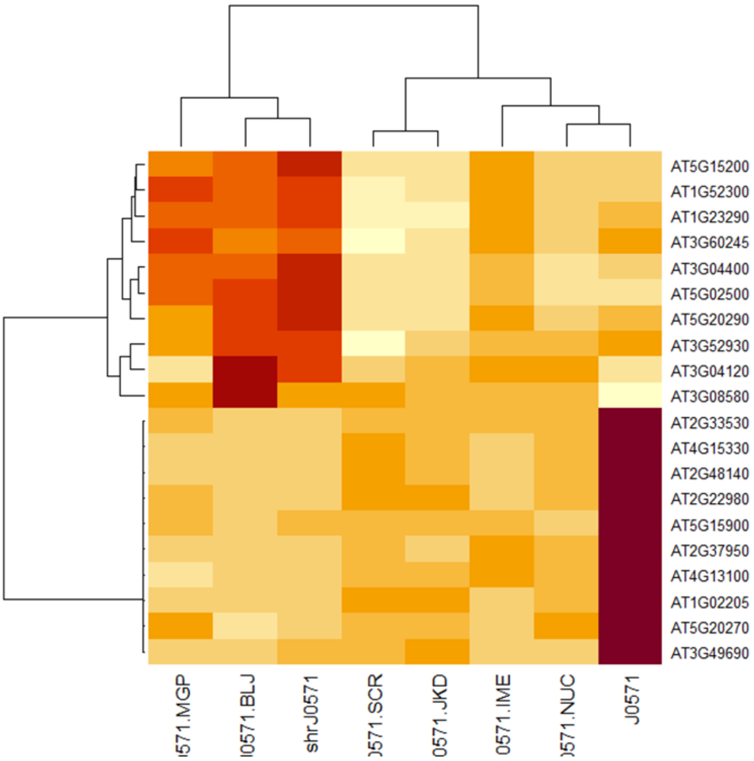

Figure 1.7. Heatmap representation of the most changing genes in the transcriptome.

PRACTICE 2. Finding biomarkers

The aim of the experiment was to study the transcriptomic profile of several cell types to find biomarkers. According to the World Health's Organization, a biomarker is "almost any measurement reflecting an interaction between a biological system and a potential hazard, which may be chemical, physical, or biological."

Question 1. Visualize distribution of gene expression values across samples. What are the samples with more variability?

The distribution of gene expression values across samples is visualized through a boxplot, shown in the **Figure 2.1.** In the boxplot, for each cell type the expression of each gene is represented by a dot. In general terms, the samples analyzed have gene expression variables within a range between 0 and 5,000. Cell types S18 and E30 have the most variability as they have the greatest number of genes outside this range and the most remarkable outliers.

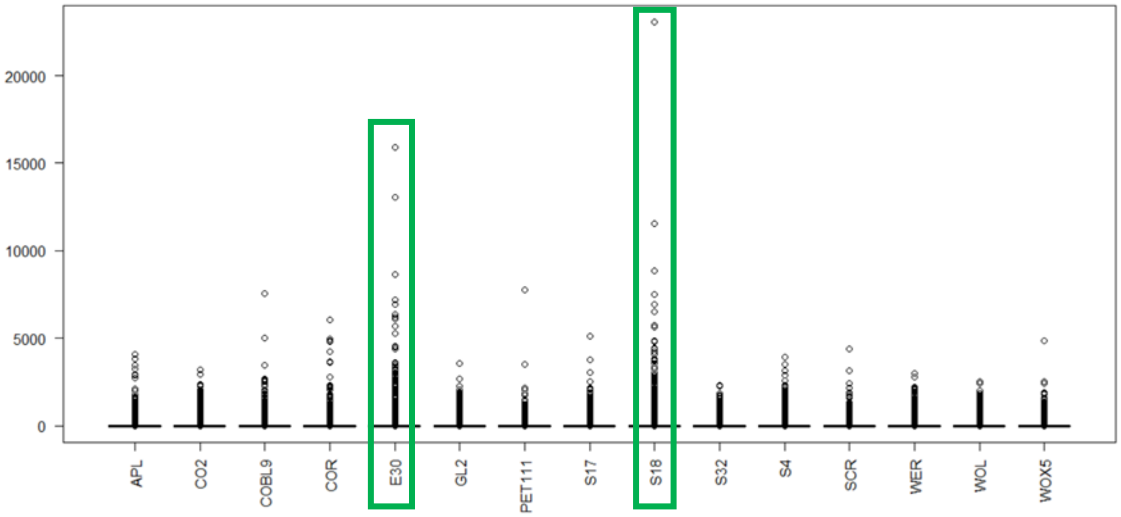

Figure 2.1. Gene expression values distribution across samples. E30 and S18 (green rectangle) are the samples with the most variability.

Question 2. Analyze transcriptome differences for all cell types using PCA. Is there any possible relationship between expression value distribution in boxplots (step1) and PCA representation of loadings for most important components?

To study the differences between the transcriptomes for cell types analyzed, PCA calculated to reduce the dimensionality of our data and visualize the differences (explained in Practice 1). The variance captured for each PCA component is displayed in **Figure 2.2**. The first 4 components which account for most of the variance: Component 1: 43.64% ; Component 2: 21.51%; Component 3: 18.27% ; Component 4: 6.66%. Considering these components, we can explain the variability and the differences between our cell lines without a significant loss of information.

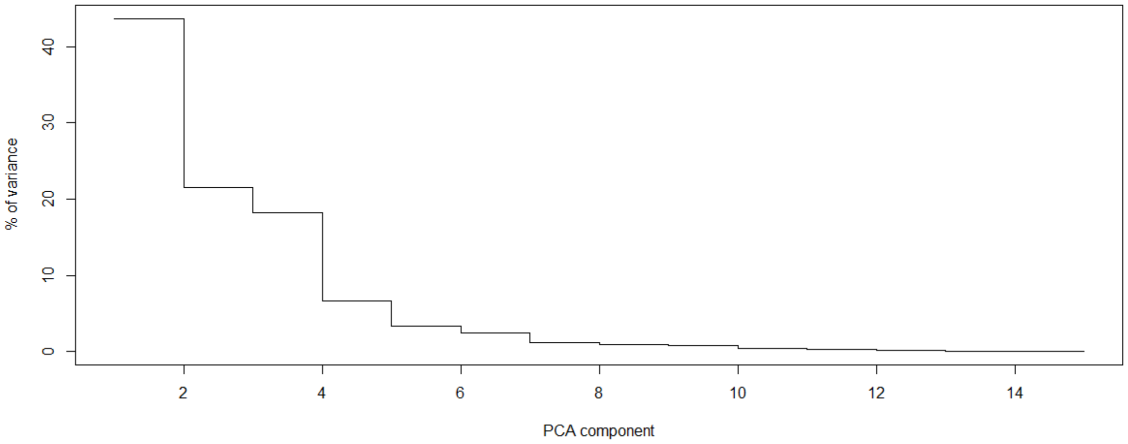

Figure 2.2. PCA variance captured by each component.

PCA loadings for component 1 vs component 2 are displayed in **Figure 2.3**. There is a clear relationship between the expression value distribution (**Figure 2.1.**) and the separation of the samples according to the PCA loadings for component 1 and 2. Only the most variable cell types, E30 and S18, can be clearly distinguished from the other cell types according to the variability of their transcriptome, captured by the principal components of the PCA. Except S18 and E30, the samples have very low variability transcriptome, so they are located closer in the plot.

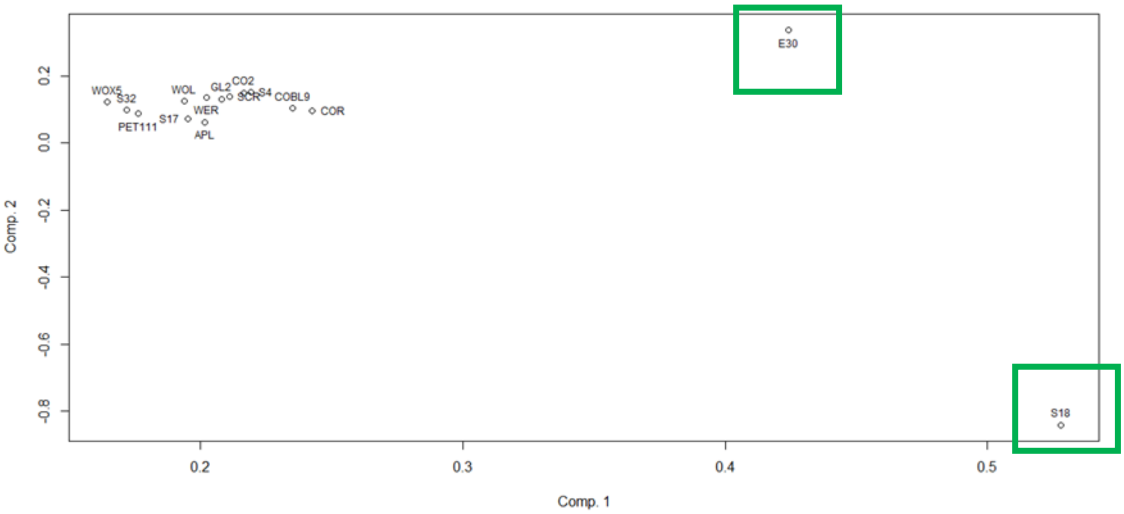

Figure 2.3. PCA loadings for component 1 and component 2. The samples with the highest variability, E30 and S18 (green box) are the only ones which can be clearly separated from the other samples.

PCA loadings for component 1 vs component 3 are displayed in **Figure 2.4**. The conclusions are the same than for the previous plot.

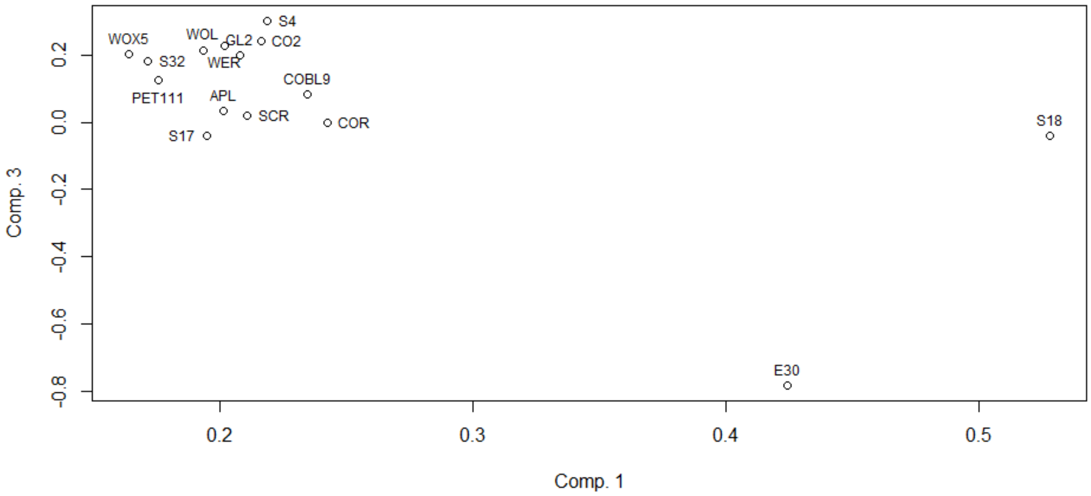

Figure 2.4. PCA loadings for component 1 and component 3.

Question 3. Find biomarkers for Stem Cells as those genes only expressed in WOX5 domain (threshold for expression=1) and not expressed in rest of cell types. Obtain biomarker names and visualize their expression patterns in a heatmap.

In order to find biomarkers for Stem Cells, the dataset was filtered considering a threshold for expression of 1, and thus, genes which had an expression higher than 1 in WOX5 cell type and genes in all the other cell types which had an expression lower than 1 were selected. In total, 417 genes were obtained with these conditions. Out of these 417 genes, there were 2 with a remarkable high expression (above 8) in WOX5: AT1G63240.1 and AT2G15590.2.

The expression of some of these genes in the different cell lines is shown in **Figure 2.4** (not all of them fit in the figure). As it can be expected because of the restrictions used to filter the dataset, genes represented have a higher expression in WOX5 (red) than in the other cell types (yellow and orange).

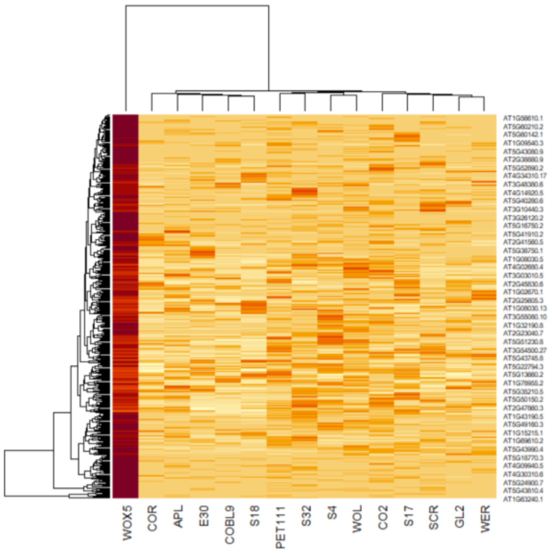

Figure 2.5. Representation of stem cells potential biomarkers expression across samples.

Question 4. Separate now samples by performing PCA only in genes previously identified as Stem Cell Biomarkers (in step 3). What do you observe? Why?

The potential stem cells biomarkers were separated and PCA was calculated for all the samples only considering these genes.

Variance captured by each component is shown in **Figure 2.6**. Components 1 and 2 account for the most variance: 58.09% and 12.57% respectively.

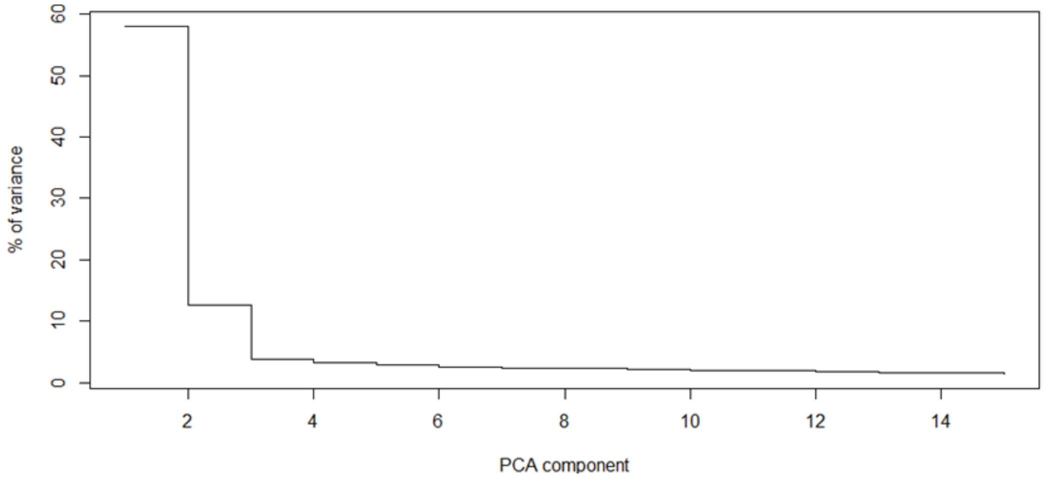

Figure 2.6. Variance captured by each component of the PCA of potential stem cells biomarkers.

Loadings are shown in **Figure 2.7**. WOX5 is separated from the other samples while S18 and E30 are integrated among the other samples. This result makes sense considering that the genes used for the PCA are those with a differential expression for WOX5.

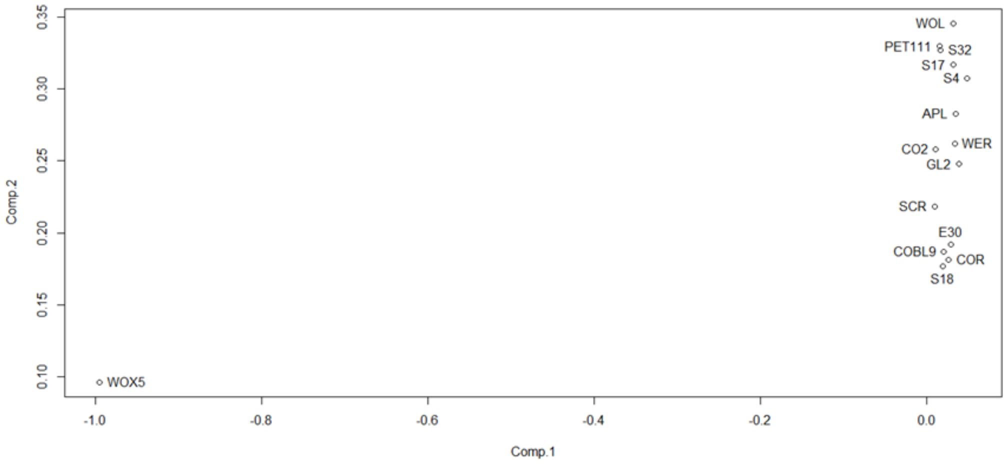

Figure 2.7. Loadings of PCA for the potential stem cells biomarkers.

PCA scores for stem cells potential biomarkers are displayed in **Figure 2.8**. In the figure we can see that almost are the genes are clustered while only a few of them are separated according to their score value for component 1 and 2. The most separated genes are the ones with the remarkably high expression for WOX5.

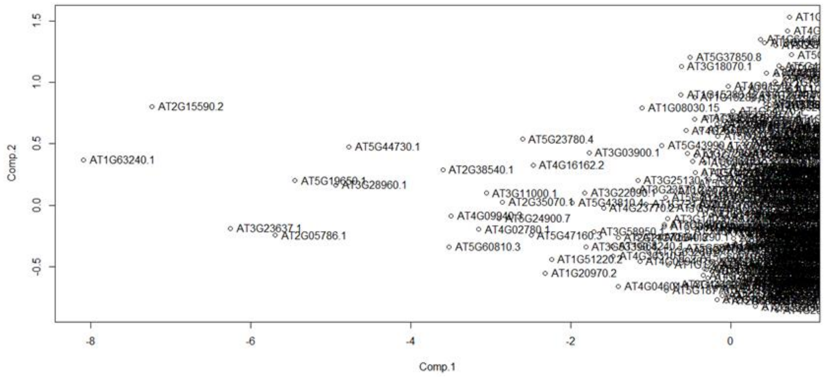

Figure 2.8. PCA scores for potential stem cells biomarkers.

Question 5. Plot PCA scores of Stem Cell biomarkers in PCA analysis of step 2. Can biomarkers be easily identified from PCA scores? Speculate about for what cell type genes with highest/lowest scores might be biomarkers.

PCA scores for all genes are shown in **Figure 2.9.** The 2 genes which have the highest expression for WOX5 cell type (AT1G63240.1 and AT2G15590.2.) cannot be identified as biomarkers. This is because WOX5 cell type is not significantly different from the other cell lines regarding variability and their biomarkers can just be detected when processing and re-calculating PCA. Nonetheless, it is possible to observe genes separated from the main cluster of genes. These separated genes will be biomarkers for the cell types which are significantly different from the other cell types (the ones with the highest and lowest loadings). In this case, the genes with the most different scores (green circles) may be biomarkers of S18 and/or E30.

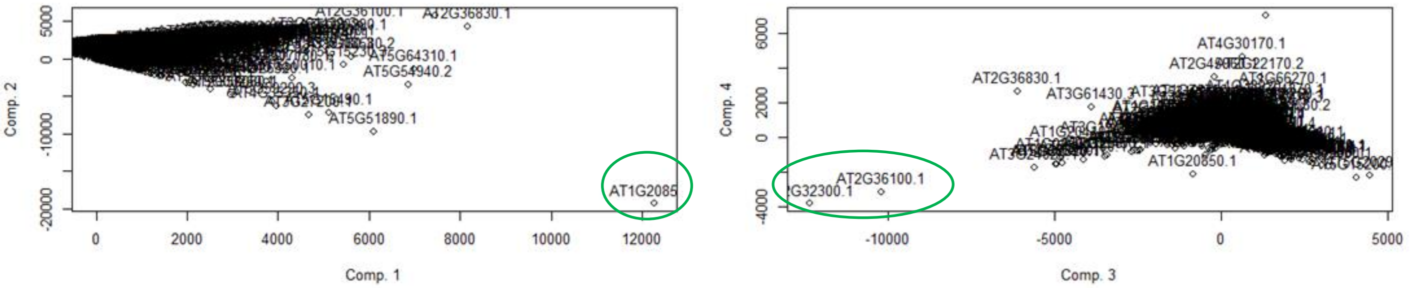

Figure 2.9. PCA scores of all genes for components 1 vs 2 and 3 vs 4. Most differentiated genes are shown within a green circle. Left plot: potential biomarker of S18 (highest component 1, **Figure 2.3**). Right plot: potential biomarkers of E30 (lowest component 3, **Figure 2.4**).

PRACTICE 3. SingleCellOmics planaria

Question 1. Initialize a Seurat Object. Perform quality control, normalization, variable features selection and scaling of the data. Perform a PCA analysis and obtain a t-SNE clustering plot.

Seurat is an R package designed for analyzing of single-cell RNA-seq data. Seurat allows to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data [1].

In this practice we are dealing with single cell omics RNA seq data. First, data is loaded, transformed into a matrix and subsequently to a sparse matrix (these steps are necessary to be able to create a Seurat object). Then, a Seurat object is created with planaria data processed. Note that no min. cells nor features are selected.

Plotting number of features vs number of counts we can check the distribution of our data according features and counts (**Figure 3.1**). This will be considered to avoid considering damaged cells, such as cells with very low quality or with more than a sample per cell. Considering this, data was subset selecting a number of features higher than 200 and a number of counts lower than 1500, removing outliers.

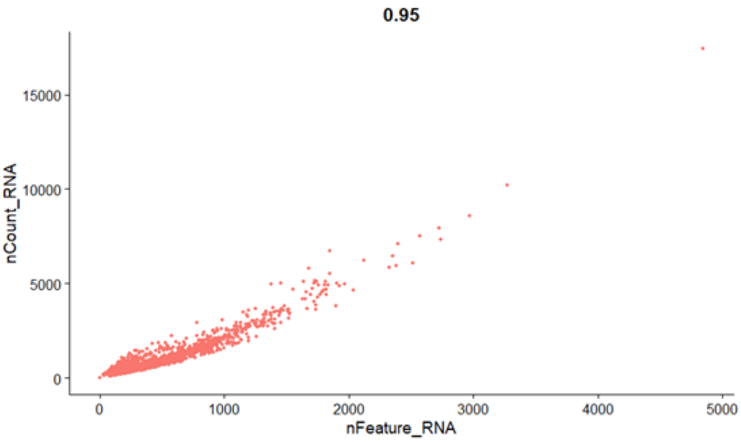

Figure 3.1. Number of RNA features vs counts for planaria single cell omics data.

Data was normalized through a logarithmic normalization. The effects of the normalization of our data can be seen in **Figure 3.2**.

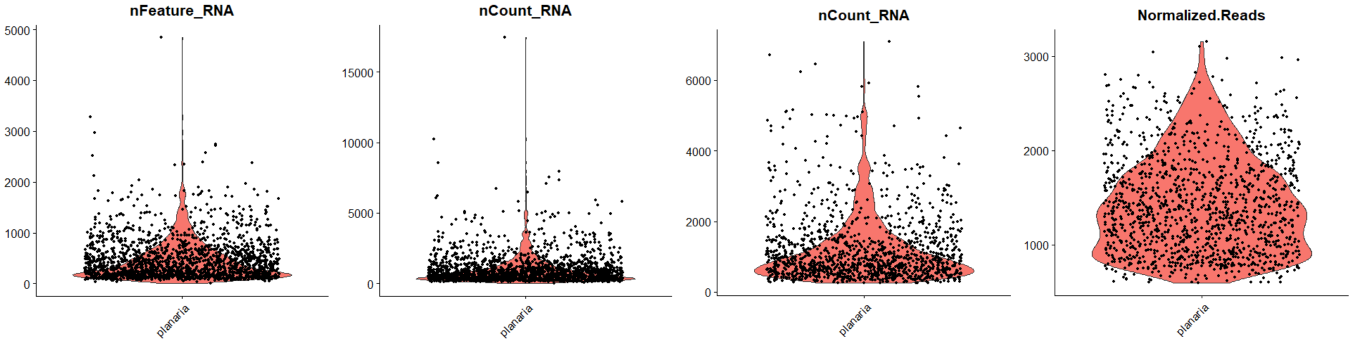

Figure 3.2. RNA features and Counts before and after normalizing data.

Then, feature selection was performed, data was centered and scaled.

PCA analysis was performed to decrease the dimensionality of our data. Neighbors and clusters were found and the resulting t-SNE clustering plot is displayed in **Figure 3.3**.

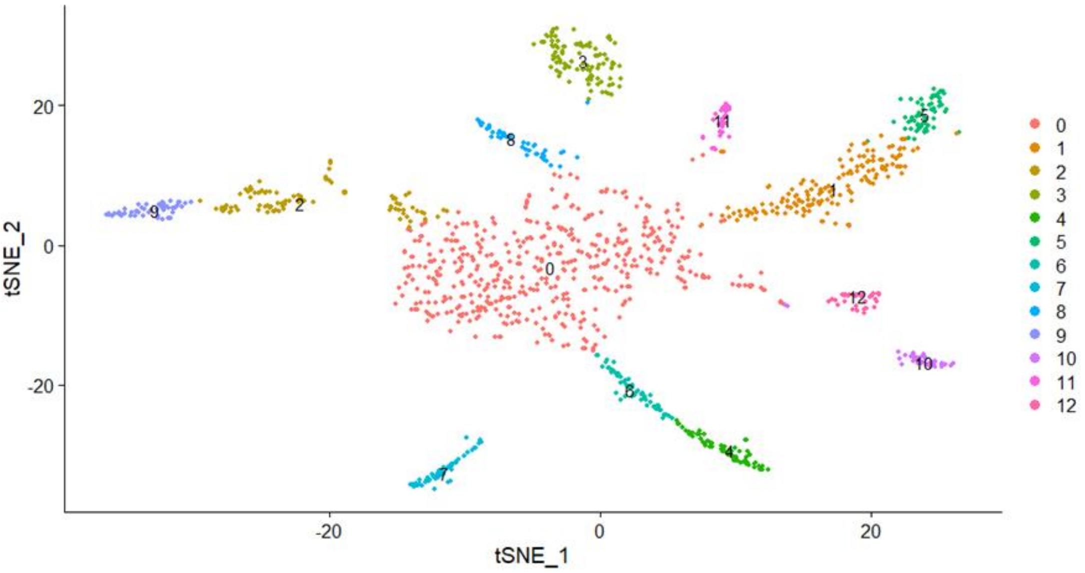

Figure 3.3. tSNE clustering plot for planaria SingleCellOmics data.

Question 2. Repeat previous exercise with the following parameters: min. expressed genes = 3 and min. detected genes = 200; cells with 200-2500 detected genes; LogNormalize, scaling factor = 10.000; 300 variable features; regress out variability for a number of genes in each cell; PCs 1-5 and resolution value of 0.6.

After performing the analysis with the parameters indicated, the new tSNE clustering plot is shown in **Figure 3.4**. Less clusters are observed (10 vs 12 in the previous case). This may be explained because the parameters used in this case are more restrictive.

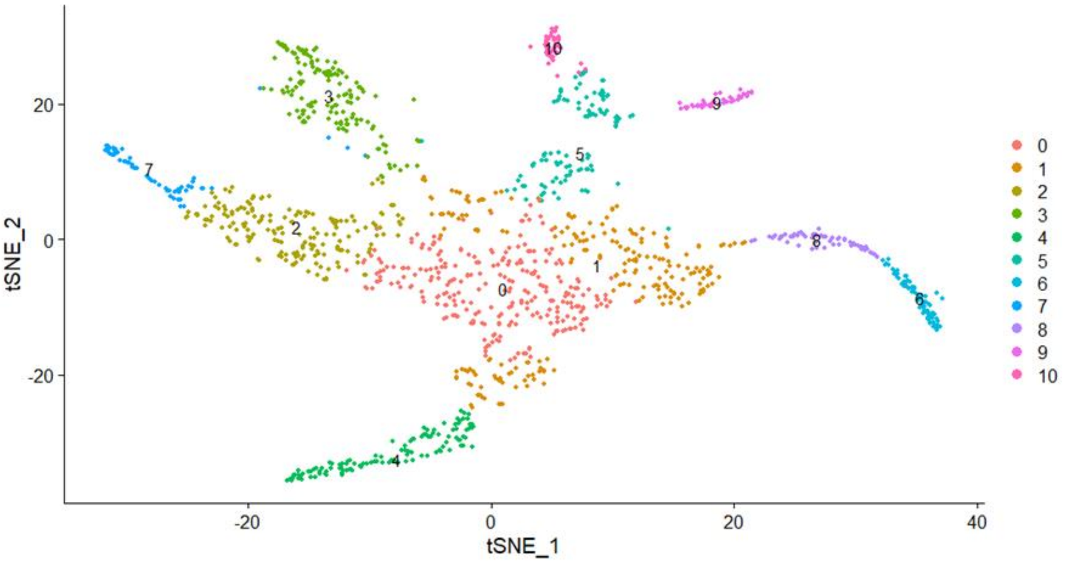

Figure 3.4. tSNE clustering plot for planaria SingleCellOmics data with new parameters.

Question 3. Extract a table of the 5 top biomarkers for each of your clusters. Comment the results.

The top 5 biomarkers for each cluster were extracted from planaria data filtered, normalized, scaled and clustered using more restrictive parameters (question 2). The resulting heatmap is shown in **Figure 3.5**. For each cluster, top 5 biomarkers appear in yellow. There are several clusters which show a high expression of the biomarkers from other clusters: cluster 6: biomarkers from 8 and vice versa; cluster 7: biomarkers from 2; clusters 9 and 10: biomarkers from 5.

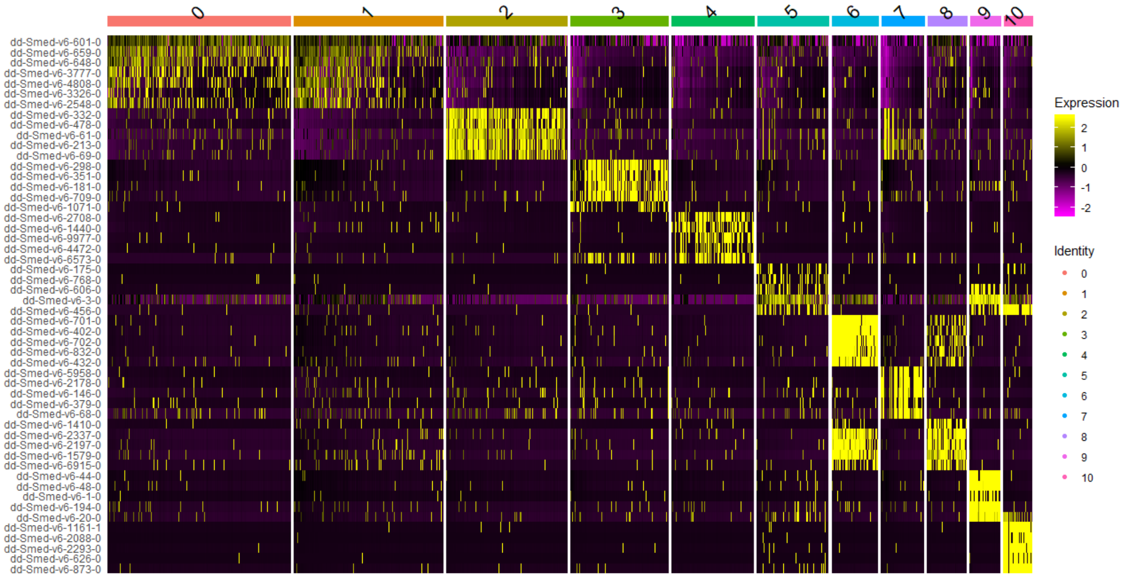

Figure 3.5. Heatmap representation of the top 5 biomarkers for each cluster.

Question 4. Considering biomarkers for several cell identities, identify and rename your clusters.

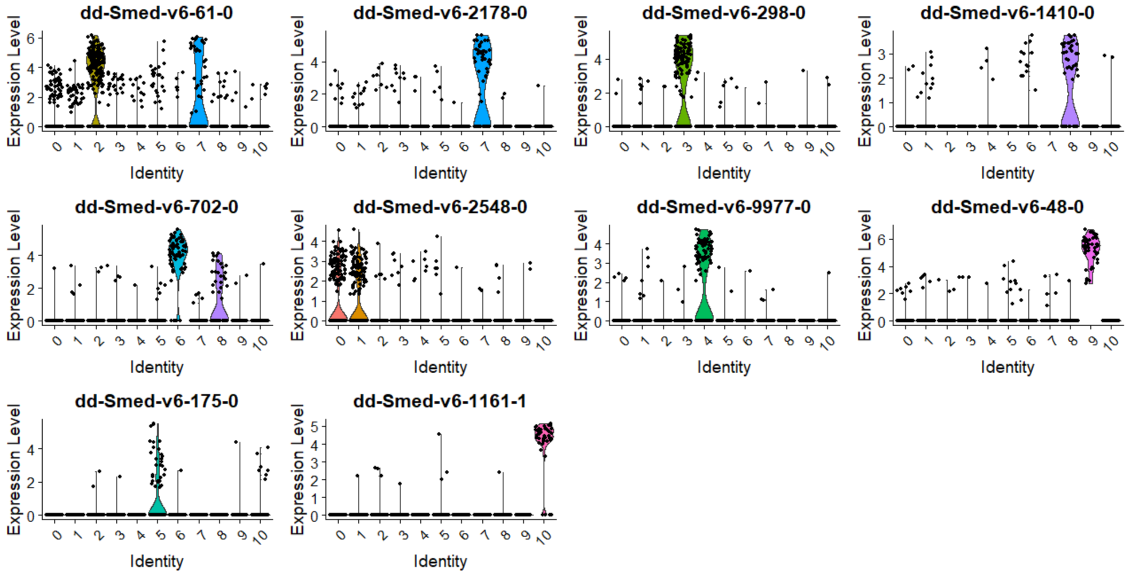

Figure 3.6. Violin plots for biomarkers belonging to several cell identities.

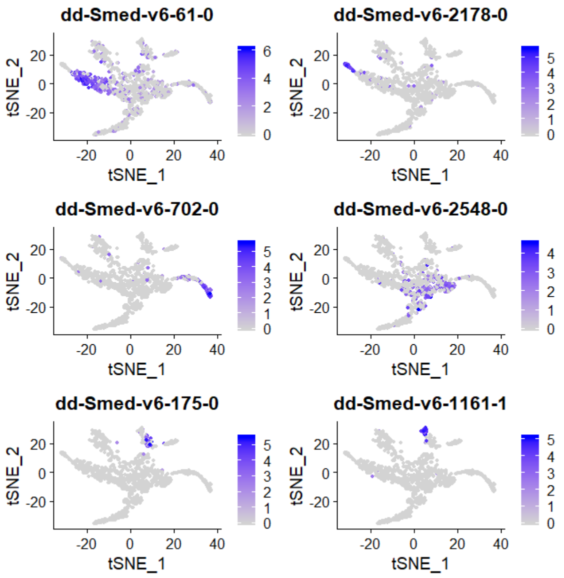

Figure 3.7. Feature plots for biomarkers belonging to several cell identities.

Based on the information provided by cell type biomarkers, the identity of each cluster is the following: (0) Neural progenitors - 0; (1) Neural progenitors - 1 ; (2) Early epidermal progenitors; (3) Epidermis; (4) GABA neurons; (5) Parenchymal cells; (6) Muscle body; (7) Late epidermal progenitors; (8) Muscle progenitors; (9) Phagocytes; (10) Pigment. For the cases in which more than one cell type biomarker was being expressed, the one with the highest expression was chosen as the identity of the cluster.

Question 5. Compare your results with this lineage tree reconstruction of planarian cell types (Plass et al., 2018). What cell type do you think cluster 0 (the central cluster) is? Can all of the planarian cell types be found in your plot? Why?

The central cluster may be neoblast 1. This result is in accordance with our results as the biomarker with the highest expression for the central cluster was a biomarker of neural progenitors' cell line.

Not all the planarian cell types can be found in the plot. This is explained because a limited number of biomarkers have been analyzed. To find all planarian cell types in the plot, it would be necessary to use biomarkers which were used in the article [2] such as pharynx, gut, etc. and we have not analyzed. However, most important planarian cell types are found in the plot.

Question 6. dd-Smed-v6-1999-0 is a neoblast (stem) marker gene. Show its expression distribution and explain the result.

The expression of a neoblast (stem) marker gene was analyzed. The distribution can be observed in the violin plot (**Figure 3.8**) and the features plot (**Figure 3.9**). The expression of this marker is remarkably high in early epidermal progenitors, late epidermal progenitors and muscle progenitors. However, the distribution of this marker is found across most of the cell types. This may be explained because neoblasts are the stem cells of planarians and these cells are distributed through the body and are activated to regenerate tissues which have been removed [3]. The regeneration may be more important in the epidermal and muscle tissues (more expression), but these cells are found through the body (wide expression of this gene).

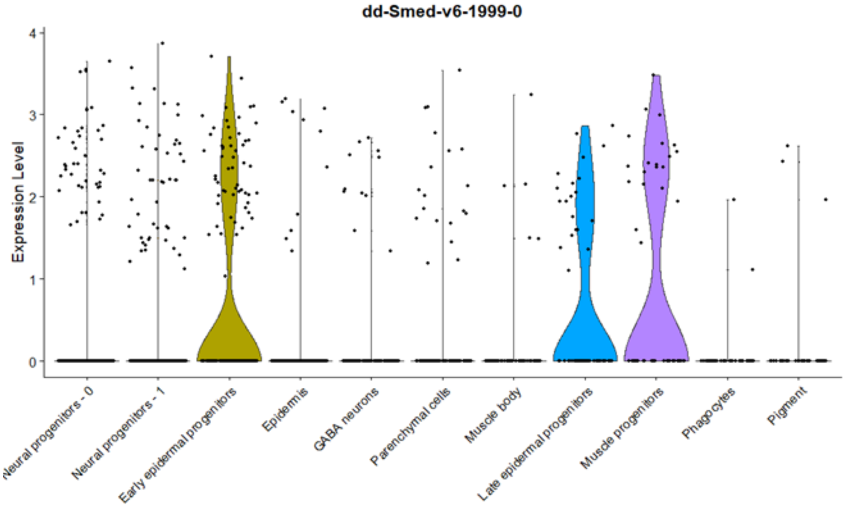

Figure 3.8. Expression of a neoblast (stem) marker gene shown as a violin plot.

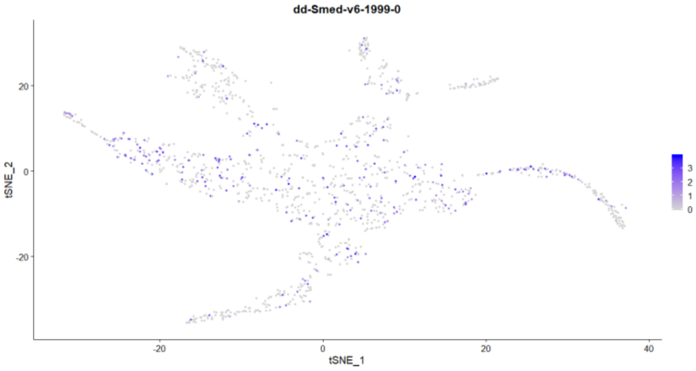

Figure 3.9. Expression of a neoblast (stem) marker gene shown as a features plot.

References

[1] Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature biotechnology, 36(5), 411.

[2] Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., ... & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science, 360(6391), eaaq1723.

[3] Rossant, J. (2014). Planaria: Genes for regeneration. Elife, 3, e02517.