Project Almaso Report

A pipeline of flow cytometry data analysis with a single cell approach

5BIM Project 2022 – 2023

Albane Flocon, Marianne Richaud, Sonia Canjura

CRCL tutors : Laurie Tonon, Cyril Dégletagne

INSA tutor : Nicolas Parisot

The goal of the project is to build a pipeline to analyse flow cytometry data in the manner of scRNA-Seq data analysis. This present document will list all choices made when developing this pipeline (libraries used, pre-processing methods…) based on the differences or similarities to scRNA seq data analysis. The presentation of the final pipeline and the explanation for each function created is to be found in the vignette attached.

Table des matières

The Data

The datasets on which we work are from mice with and without tumors. We had flow cytometry datasets and scRNA-seq datasets. The aim is to conduct an analysis on both cytometry data and scRNA-Seq data and highlight similarities and differences. As cytometry data are usually analyzed through gating techniques, our work was more focused on cytometry data analysis.

As a reminder, flow cytometry is a technique that analyzes the expression of cell surface and intracellular molecules. The fluorescent intensity produced by fluorescent-labeled antibodies detecting protein are measured as cells flow one by one in front of lasers. Different fluorescent markers can be used simultaneously to identify different specific proteins. 2 types of parameters are measured in flow cytometry:

- "structure" parameters: forward scatter (FSC), which detects cell size, and side scatter, that provides information about the internal complexity of a cell. This information is usually used in the first steps of the gating process.
- Fluorescence markers: the intensity measurement gives the expression level of the specific protein the marker stain for.

Our dataset is composed of 19 parameters (6 structurals and 13 markers), stored in columns, and thousands of cells (stored in lines). We decided, as structural parameters and fluorescent markers do not convey the same type of information, that we will not use them both for clustering. We will focus mostly on the fluorescent markers.

Pre-processing

The first step of our pipeline is to preprocess the data in order to clean them for further analysis. Our preprocessing pipeline contains 3 steps :

- Compensation : this is a necessary processing step for flow cytometry data, to control the spillover phenomenon happening during the fluorescence detection.

- Quality control : this step is as necessary in scRNA-Seq data analysis as in flow cytometry data analysis in order to suppress outliers. Since the sources of outliers are very different for cytometry and RNA-Seq, we could not use scRNA-Seq QC packages. After comparing 3 different packages (*flowAI*, *flowClean* and *peacoQC*), we first decided to use flowAI because it is one of the most used in flow cytometry and it seemed to give more satisfactory results for the clustering. However, it was not always working on all computers because of version conflicts. We then choose the *peacoQC* package. The amount of removed cells seems reasonable.

Table 1 - Quality control summary

Sample	Cells counts before preprocessing	Cells counts after preprocessing	% of cells removed
VP4_Tumor_CD45+ cells.fcs	205,726	178,726	0.87%
Vp6_Control_CD45+ cells.fcs	64,213	58,963	0.91%
MixC_tumeur_CD45+ cells.fcs	66,816	62,566	0.93%
CD45pos2_control_CD45+.fcs	69,421	67,171	0.96%

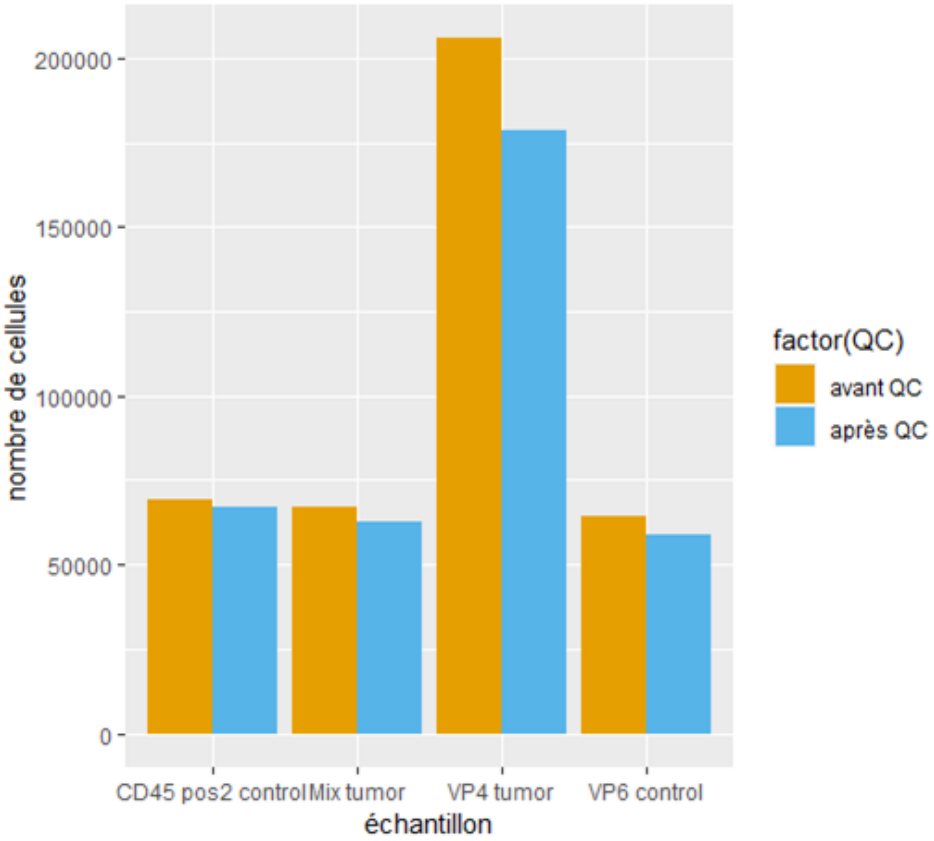

Figure 1 - Barplot of QC summary

- **Normalization** : For this last step of pre-processing, we decided to treat the flow cytometry data as scRNA-Seq data. So we decided to use a logCPM normalization method on the expression matrix. We also started by using a TMM normalization which was abandoned for reasons explained later on.

We did not integrate the log or arcsinh transformation step that is usual in flow cytometry pipelines because a log transformation is included in our normalization step.

Dimensionality reduction

Since our datasets contained only approximately 20 parameters, dimensionality reduction was not evidence. We therefore decided to run the pipeline with and without PCA and to compare the result. We observed that when using PCA, our clusters were better separated than without it.

We concluded that using PCA tends to increase the amount of clusters (especially when using only 2 dimensions) but that the separation between clusters was increased.

Our recommendation would therefore be to do a PCA, increase the number of PC dimensions used for clustering to reduce the number of clusters, and lower the Louvain clustering resolution in order to increase even more the separation between clusters. The clustering method will be explained later on.

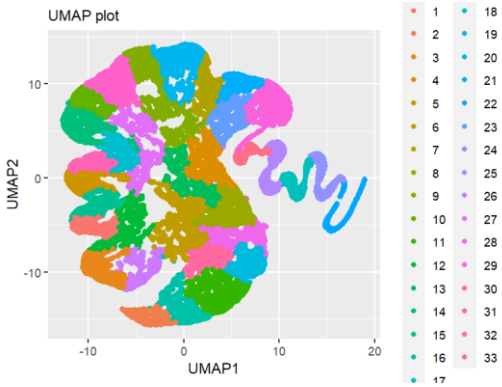

Figure 2 - UMAP of CD45pos2 control, 2 PC axes, resolution 0.5

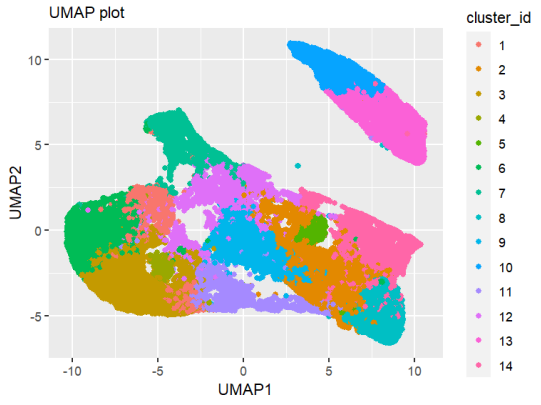

Figure 3 - UMAP of CD45pos2 control without PCA resolution 0.5

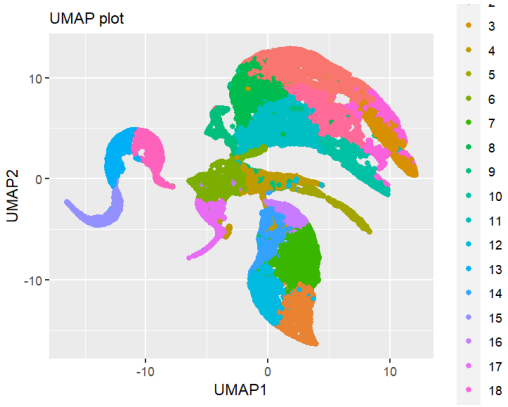

Figure 4 - UMAP of CD45pos2 control, 5 PC axes, resolution 0.5

Clustering

In order to have a similar approach to what is usually done on scRNA-Seq data, we decided to do a UMAP representation. Then, we tested several different clustering methods :

- SNN + Louvain clustering, which is the clustering method used in Seurat for scRNA-Seq data ;
- Cytosplore, which is an app using HSNE-based Gaussian Mean Shift clustering to automatically compute flow cytometry clustering ;
- FlowSOM clustering, which is a flow cytometry clustering package that seemed very popular in the literature.

We first tested Cytosplore which is an app that will calculate H-SNE clustering representation from cytometry data and output FCS files containing the clustering, that you can then visualize in RStudio. The advantage of this method is that it automatically calculates the clusters, but the issue is that this app can not be used on a computing cluster. Moreover, Cytosplore is based on a R package called *cytofast*, which is no longer updated in the last version of Bioconductor. We therefore decided to drop Cytosplore.

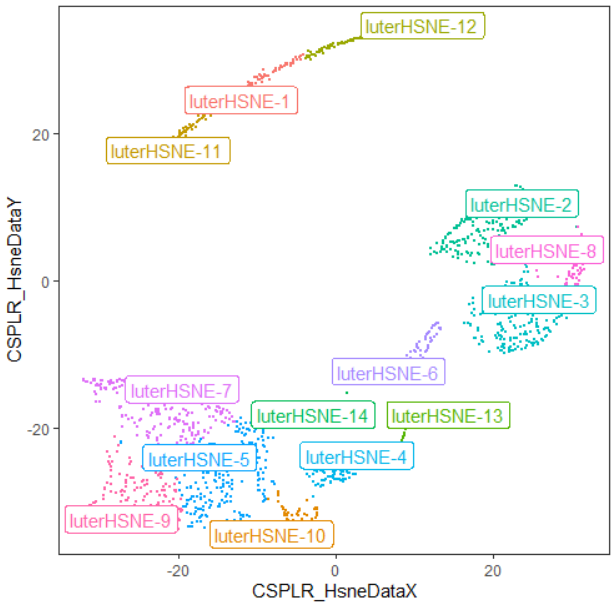

Figure 5 - Clustering with Cytosplore on CD45pos2

We then compared the clustering results from SNN + Louvain and FlowSOM clusterings. We observed that with FlowSOM, most of the cells were always contained in one very big cluster, no matter the pre-processing or the dimensionality reduction. On the other hand, we had a reasonable amount of clusters with a good separation. Based on those observations we decided to only include the SNN + Louvain clustering in our pipeline. This method also has the advantage of being used for scRNA-Seq data so it also met the goal of treating cytometry data as scRNA-Seq data.

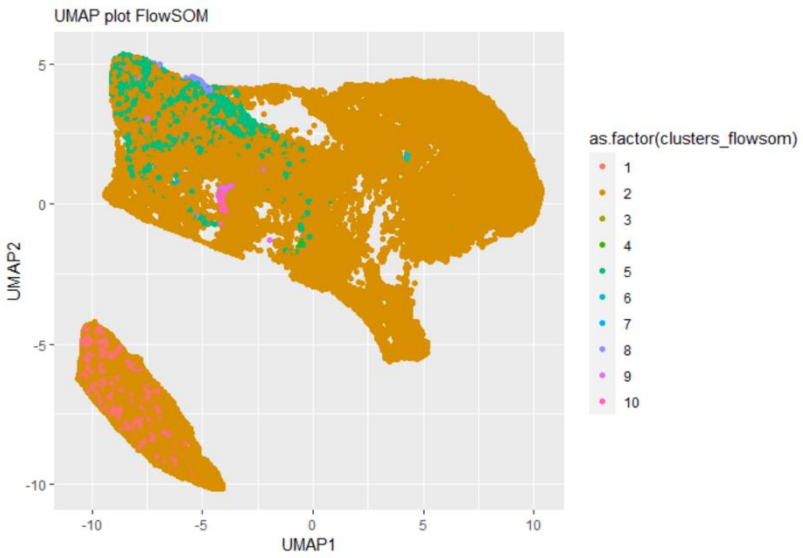

Figure 6 - UMAP of CD45_pos2 with FlowSOM clustering

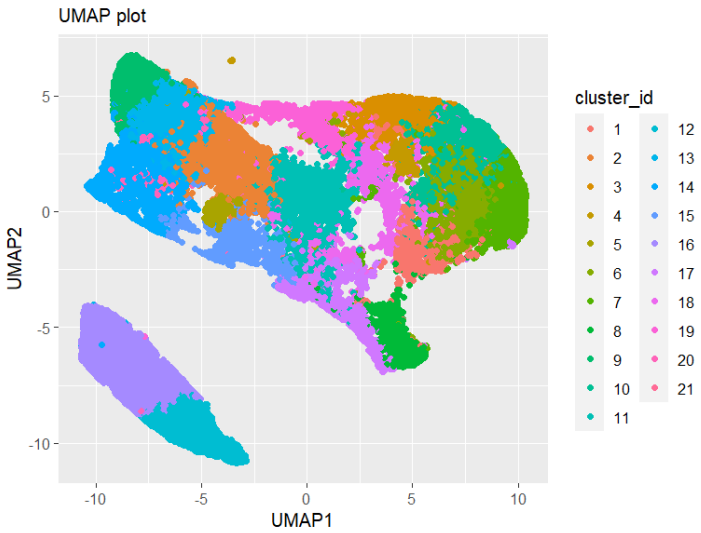

Figure 7 - UMAP of CD45pos2 control with SNN+Louvain clustering

Adjustments after first run of the pipeline

After running all the precedent steps on the “Vp6_Control_CD45+ cells.fcs" file, we observe an odd mechanism. When plotting the PCA, we saw that the first axis accounts for 91% of the variance. Then when taking 2 axes for PCA, we observe an UMAP with a potato shape.

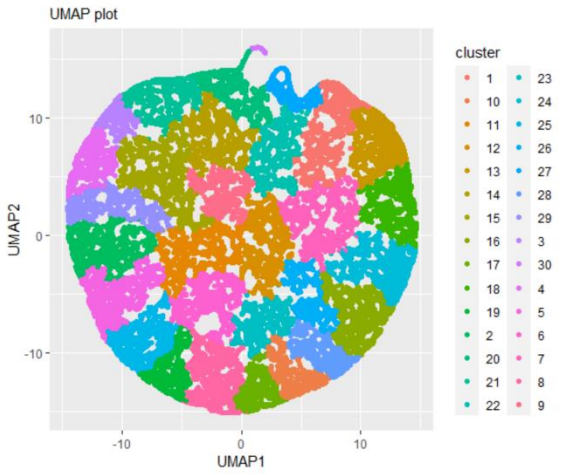

Figure 8 - UMAP of Vp6_Control, 2 PC axes, resolution 0.5

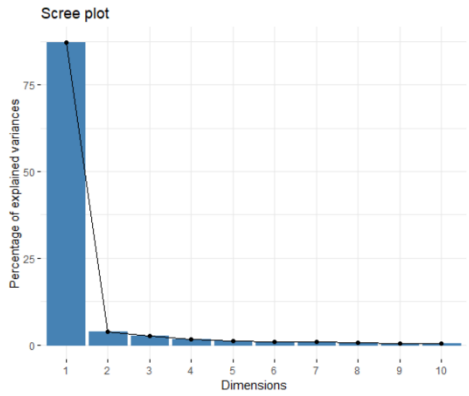

Figure 9 - Percentage of variance explained by PCA on Vp6 control

What could cause this issue? To have a better understanding of what is happening, we plot the contribution of each marker in the PCA. The following graph indicates that the APC marker, that is linked to the PDC cell population, contributes a lot. The plot of the two firsts PCA axis also shows that only a few points are separated by the first axis (91% of variance). In fact, PDC cells are a rare type of immune cells that secrete a lot of one specific protein. So this is why the UMAP was of such a shape.

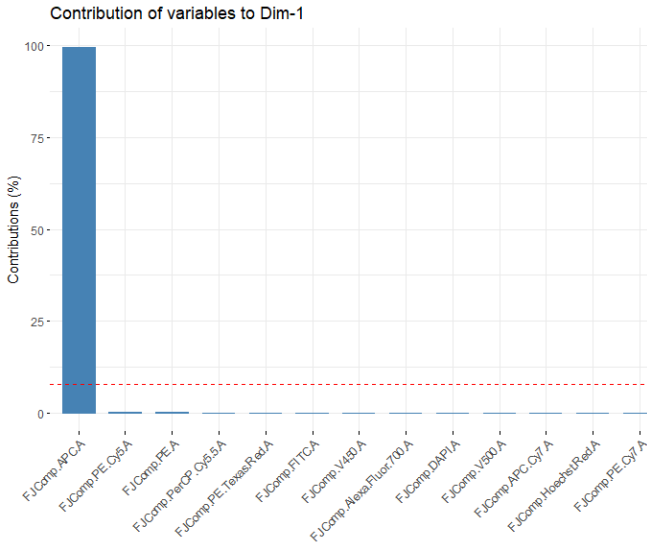

Figure 10 - Contribution of markers to first axis of PCA on Vp6 control

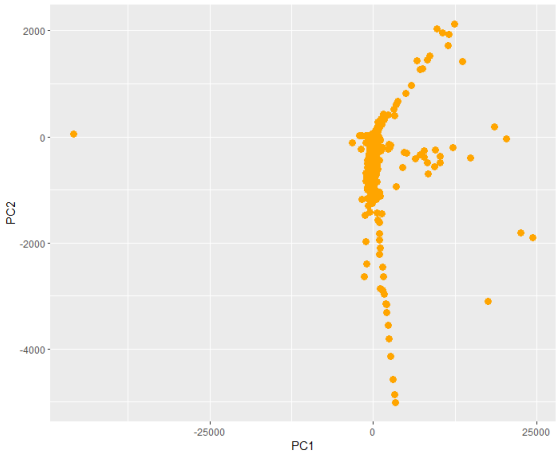

Figure 11 - Visualisation of PCA in 2 dimensions of Vp6 control

- **Negative expression**: we observe in the matrix expression that some values were negative. Yet negative fluorescent expression is not biologically acceptable. So, we set all negative values to 0.

- **Normalization**: At the beginning, we used a TMM (trimmed mean of M-values, M-values are the log fold change between each sample and a reference) from *EdgeR* package with the logCPM method. It actually does not normalize but instead calculates normalization factors. It trims off the most highly variable genes and then calculates a normalization factor that is then used to adjust the logCPM values. Here we are applying the logCPM on our whole data low and high value. When applying TMM, the value related to the APC marker expression bring a high normalisation factor in its favor and disturb the PCA. So, we remove the TMM, keep the logCPM normalization and choose to scale within the PCA, with the argument SCALE = TRUE.

With those adjustments, we find better and more coherent results.

Visualization

Heatmap

In order to visualize the expression of each marker across all clusters, we built a heatmap that illustrates the average expression value of each marker across the cells in each cluster. To accentuate the gradient, we scaled the values across each marker. The colours of the heatmap allow us to observe how much a marker is expressed in a cluster.

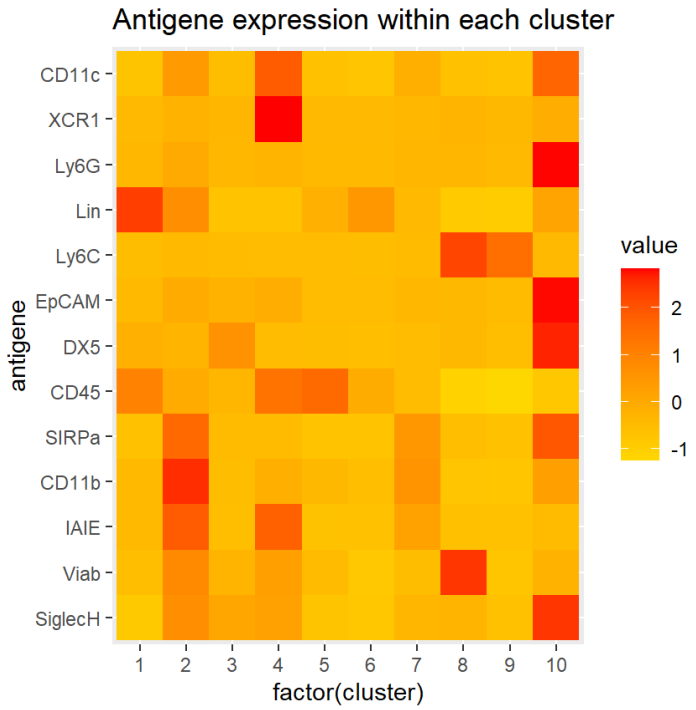

Figure 12 - Heatmap of markers expression in each cluster of Vp6 control with final pipeline

Marker's expression across clusters.

To visualize how each marker was expressed across each cluster, we plotted the UMAP of clustered cells and then illustrated each marker's expression level with a color scale. This allows us to identify markers that are expressed in a specific marker and that can later define the biological nature of the cluster.

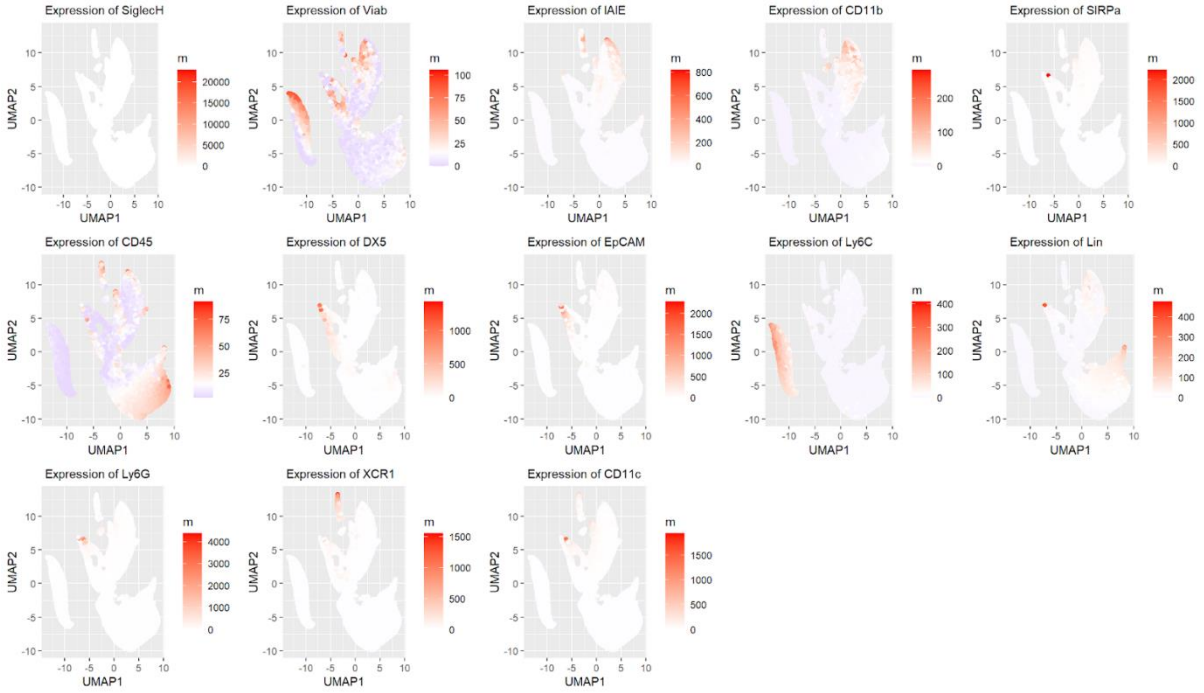

Figure 13 - UMAP with expression of each markerof Vp6 control with final pipeline

Differential Expression

In order to identify the cell types based on our clustering, we needed to identify the markers that had a significant differential expression among the clusters.

To do that, we explored the existing packages and the existing statistical tests that we could apply to our data. We only found two packages achieving this goal : *diffcyt* and *cytoTree*, both being flow cytometry data analysis pipelines containing preprocessing, clustering and differential expression methods. However, we realized that *diffcyt* uses a FlowSOM clustering object for its differential expression method, and since we had decided not to use FlowSOM we were not able to use *diffcyt*. As for *cytoTree* we realized that it performed differential expression analysis on the branches of the tree created by the pipeline but not on UMAP clusters. We therefore decided to drop *CytoTree* and to perform our own differential analysis method.

For that, we decided to perform a Wilcoxon test, which is a non-parametric Student test to compare two samples. We then identified top expressed markers for each cluster based on adjusted p-values and log fold changes, as is usually done to identify DEGs for RNA-Seq data. As many fold change values were negative, they were removed in the process.

We have recapitulated in the table below the markers that seem highly expressed in the heatmap and markers visualization and those that are differentially expressed for each cluster.

Table 2 - Significant markers identified by heatmap, expression plot and wilcox test on Vp6 control

Cluster	Heatmap	Expression graph	LogFC
1	Lin + CD45	Lin + CD45	Lin + CD45
2	CD11b + IAIE + SIRPa	Siglec-H	Ly6G + EpCAM + DX5
3	DX5	XCR1	IAIE + DX5
4	CD11c + CD45 + IAIE + XCR1	XCR1	DX5
5	CD45	CD45	XCR1 + CD45 + CD11c + IAIE
6	Lin + CD45	CD45	CD45
7	SIRPa + CD11b + IAIE	CD45	Lin
8	Viab + Ly6C	Ly6C	CD11b + SIRPa
9	Ly6C	Ly6C	Ly6C + Viab
10	DX5 + EpCAM + Ly6G + S	CD49b + EpCAM+ Lin + CD11c + Ly6G	Ly6C

Biological Interpretation

The goal is now to interpret biologically the results of the clustering. To do this, we have 3 different tools. The most robust is the statistical test performed to detect differentially expressed genes in certain clusters. We also have the heatmap and the umap of each marker expression across the clusters that give us an idea of certain markers that are highly expressed in certain clusters. By combining these 3 methods we can identify markers that are specific to certain clusters and that define the biological nature of the cluster. It is important to underline that this is possible because this type of data has a low number of markers and a combined analysis is still feasible.

Concerning our data *d45_pos2_control*, the results of the three methods are not coherent for all clusters. We mainly based our analysis in the statistical result that seemed more robust to analyze. Some clusters don't have specific markers (with a significant p value and a positive and high fold change), some others have a logFC that is negative for some markers. This leads us to believe that it is possible that we could have "over-clustered" our data. To address this issue, we tried to modify the clustering's resolution to try to find big population types (like B and T cells) and then try to narrow the analysis by increasing the resolution and identify clusters that correspond to cell types that only have a specific marker (like PDC for siglecX or neutrophils for Y6G).

Another way to overcome "overclustering" would be to merge some clusters that present important similarities /homogeneities (same markers, in different amounts). We could apply hierarchical clustering

or do that by hand. Finally, we believe that trying other clustering methods could lead to better results. However, the other methods (e.g. FlowSom) that we tried did not seem to separate the clusters in a way that could be interpretable and given the time constraints we decided to leave in our pipeline the SNN + Louvain clustering method.

Since our three methods for results interpretation (heatmap, marker expression graph and logFC) do not give the same results, it might also be possible that there is a coding issue in one of these functions. We did not have time to do interpretation for all the samples and to solve the overclustering issue. Therefore, if after doing that it turns out that results for all methods are always inconsistent, it might be interesting to check for coding issues in the heatmap or findMarkers_cyto functions. We already did some verifications and did not find any code problems but a mistake is always a possibility.

Discussion

- Verification of our interpretation with scRNA-Seq

The only way to check if our pipeline is correct is to do biological interpretations of the results and see if they seem coherent. Another possible way might be to compare our results with scRNA-Seq results. We did the analysis of our scRNA-Seq data on the same samples using Seurat. However, since our interpretations for the cytometry data were very limited, we could not do a real comparison of the results. Moreover, the real issue here is that we do not know yet if there actually is a correlation between genes and surface protein expression. Therefore we can not be sure if it is really possible to compare scRNA-Seq and flow cytometry results. We think that this might be an investigation perspective for the future in order to evaluate our pipeline.

- Adaptation of the pipeline for scRNA-Seq data

In order to achieve the goal of a joint pipeline for the analysis of flow cytometry and scRNA-Seq data, and to compare the results obtained for both types of data, we attempted to run our pipeline on scRNA-Seq data. For this, we performed pre-processing of the data in Seurat, and extracted the expression matrix as a data frame after QC and scaling. We wanted to input this data in the PCA function of the pipeline, but unfortunately the size of the data frame made the execution time of the PCA too long. So we had to abandon this axis of study due to lack of time. However, we think that this can be a way to improve the pipeline in the future.

- Integration of dimensional data

The FSC and SSC columns of the flow cytometry expression array represent a different type of data than the data from fluorescence markers. It was therefore not possible to integrate them into the pipeline with the same format. We therefore researched multi-omics clustering techniques to integrate them. In particular, we identified the COCA (and MOFA2 packages as potential multi-omics clustering methods. COCA is a solution to integrate multi-omics data with hierarchical clustering. Unfortunately, once again, the size of the data frame did not allow us to use the COCA package because of memory size issues.
We then tried the MOFA2 package, which allows the integration of multi-omics data by training a model in order to identify the main axes of variation. To use this package, we tried using markers as features and cells as samples to create a MOFA object. We then considered the FSC/SSC data and the markers as two Views.

We obtained the following two graphs, however, with so many cells it is very difficult to interpret the results.

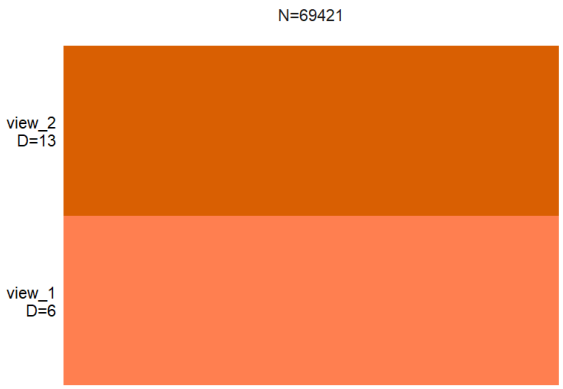

Figure 14 – Visualization of MOFA object

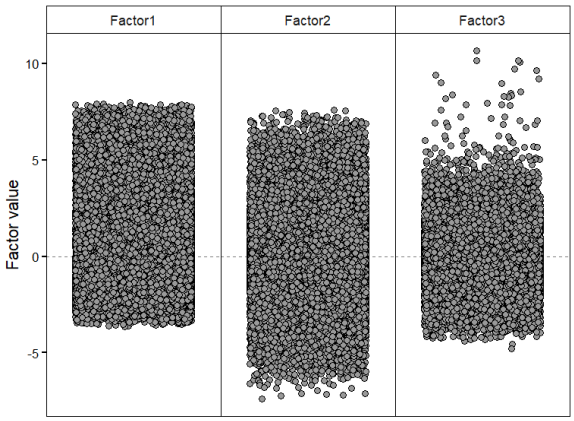

Figure 15 - Plot of MOFA model on 3 factors

We then tried to represent SSC/FSC and markers as two groups, but we are not sure how to interpret the data in this case. Indeed, we did not fully understand the difference between Views and groups. We tested this with 5000 cells since it was not possible to do so with all cells, however, again the results are difficult to exploit.

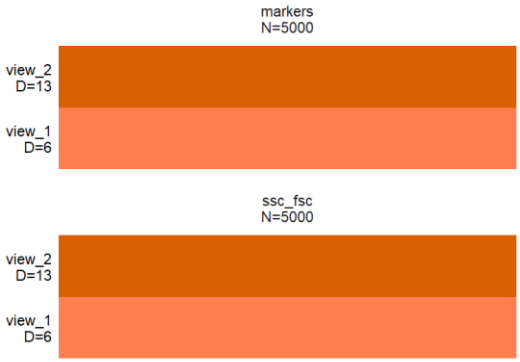

Figure 16 - Representation of MOFA object with groups

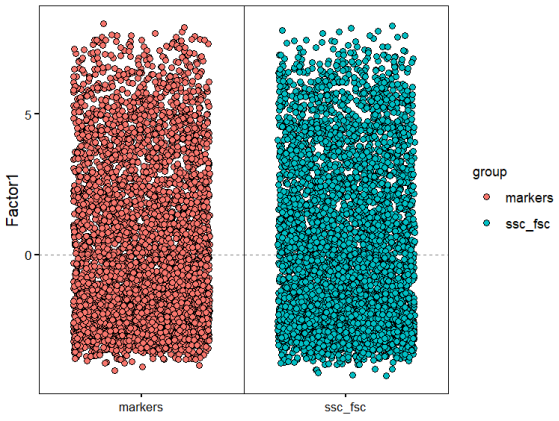

Figure 17 - Representation of 1rst factor of MOFA object coloured by groups

We are therefore not sure that MOFA is usable in the case we are interested in. Indeed, MOFA is usually used on many more features, and on samples and not cells. We did not have the time to investigate this method further and to interpret the results. However, it may be interesting to continue testing codes using MOFA, or to test other multi-omics clustering methods.

Bibliography

Flow Cytometry

McKinnon, K. M. (2018). Flow Cytometry: An Overview. *Current Protocols in Immunology*, *120*(1). https://doi.org/10.1002/cpim.40

Pre - processing

Melsen, J. E., van Ostaijen-ten Dam, M. M., Lankester, A. C., Schilham, M. W., & van den Akker, E. B. (2020). A Comprehensive Workflow for Applying Single-Cell Clustering and Pseudotime Analysis to Flow Cytometry Data. *The Journal of Immunology*, *205*(3), 864–871. https://doi.org/10.4049/jimmunol.1901530

den Braanker, H., Bongenaar, M., & Lubberts, E. (2021). How to Prepare Spectral Flow Cytometry Datasets for High Dimensional Data Analysis: A Practical Workflow. *Frontiers in Immunology*, *12*, 768113. https://doi.org/10.3389/fimmu.2021.768113

Monaco, G. et al (2016). flowAI: Automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics*, *32*(16), 2473–2480. https://doi.org/10.1093/bioinformatics/btw191

Fletez-Brant, K., Špidlen, J., Brinkman, R. R., Roederer, M., & Chattopadhyay, P. K. (2016). flowClean: Automated identification and removal of fluorescence anomalies in flow cytometry data: flowClean for Quality Control of Flow Cytometry Data. *Cytometry Part A*, *89*(5), 461–471. https://doi.org/10.1002/cyto.a.22837

Hahne, F., LeMeur, N., Brinkman, R. R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., & Gentleman, R. (2009). flowCore: A Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, *10*(1), 106. https://doi.org/10.1186/1471-2105-10-106

Law, C. W., Alhamdoosh, M., Su, S., Dong, X., Tian, L., Smyth, G. K., & Ritchie, M. E. (2018). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research*, *5*, 1408. https://doi.org/10.12688/f1000research.9005.3

Clustering

Beyrend, G., Stam, K., Höllt, T., Ossendorp, F., & Arens, R. (2018). Cytofast: A workflow for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations. *Computational and Structural Biotechnology Journal*, *16*, 435–442. https://doi.org/10.1016/j.csbj.2018.10.004

Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., & Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM. *Cytometry Part A*, *87*(7), 636–645. https://doi.org/10.1002/cyto.a.22625

Thomas Höllt, Nicola Pezzotti, Vincent van Unen, Frits Koning, Elmar Eisemann, Boudewijn Lelieveldt, and Anna Vilanova. (2016) Cytosplore: Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. *Computer Graphics Forum (Proceedings of EuroVis 2016)*. https://www.lcbc.nl/publications/2016_eurovis_cytosplore/

Zhu X, Zhang J, Xu Y, Wang J, Peng X, Li HD. Single-Cell Clustering Based on Shared Nearest Neighbor and Graph Partitioning. (2020) *Interdiscip Sci*. doi: 10.1007/s12539-019-00357-4.

https://pubmed.ncbi.nlm.nih.gov/32086753/

Differential expression

Weber, L. M., Nowicka, M., Soneson, C., & Robinson, M. D. (2019). diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, *2*(1), 183. https://doi.org/10.1038/s42003-019-0415-5

scRNA-Seq

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Multi-omics clustering

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M. D. M., Niu, B., McLellan, M. D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A. A., van't Veer, L. J., Lopez-Bigas, N., … Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, *158*(4), 929–944. https://doi.org/10.1016/j.cell.2014.06.049

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., & Stegle, O. (2018). Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, *14*(6). https://doi.org/10.15252/msb.20178124

Acknowledgements

We would like to thank Cyril and Laurie for all their patient and kind explanations. We really appreciated working on this very interesting project (especially during our matcha tea breaks) and we learned a lot from you. We also want to thank Nicolas Parisot for his help throughout the whole project.