Supervised Non Overlapping Single Cell Gene Set Clustering

Hadrien Lorenzo, Samson Koelle, Boris Hejblum, Raphaël Gottardo, Rodolphe Thiebaut

2017-02-15

Contents

1 Context

1.1 Idea

The main idea of that work is to define gene sets which would describe given cell types according. This is a supervised approach. We work on different types of genes such as **b-cells**, **cd14-monocytes**, **cd34**, **cd4-t-helper**, **cd56-nk** or **cytotoxic-t**.

Those different cell types have common and different pathways of expression, we think it might be interesting to find the pathways which are characteristic to one type of cells. Which means that we want to find genes which are activated for one type of cells and only for that type of cells.

1.2 Technology & data

Currently, we have found the **Single cell RNA-Seq** to be able to show those pathways, our wish has been to find enough well designed human single cell RNA-Seq datasets as to perform coherent analyses.

The **10X** technology permits to sequence a large amount of cells and has built **public dataset**.

Indeed, through (Zheng et al. 2016), the authors depict the quality of the generated data. We have had access to 29 dataset, the table 1 shows some of the properties of the data avalable on the data, the ones used by (Zheng et al. 2016). Those information are also disponible on the website, other useful information are also avalable in the table avalable **here**, this is a `.xls` file which recaps most of the information on the different datasets. We recall here that we are only concerned by human dataset and this is why **ercc** does not show any relevant inforamtion, this is also why some of the information that you can find in the `.xls` file (sheet 1 especially) will not be the same as in 1 : some cells came from mouse also.

	Number of Cells	Number of Genes	Proportion of non null genes (%)
293t_3t3	482	32738	51
293t	2885	32738	57
aml027_post_transplant	3965	32738	51
aml027_pre_transplant	3933	32738	48
aml035_post_transplant	909	32738	41
aml035_pre_transplant	3592	32738	49
b_cells	10085	32738	48
cd14_monocytes	2612	32738	43
cd34	9232	32738	59
cd4_t_helper	11213	32738	49
cd56_nk	8385	32738	50
cytotoxic_t	10209	32738	49
ercc			
fresh_68k_pbmc_donor_a	68579	32738	62
frozen_bmmc_healthy_donor1	1985	32738	50
frozen_bmmc_healthy_donor2	2472	32738	50
frozen_pbmc_b_c_50_50	8186	32738	54
frozen_pbmc_b_c_90_10	7046	32738	52
frozen_pbmc_b_c_99_1	8340	32738	53
frozen_pbmc_donor_a	2900	32738	48
frozen_pbmc_donor_b	7783	32738	52
frozen_pbmc_donor_c	9519	32738	55
jurkat_293t_50_50	3388	32738	60
jurkat_293t_99_1	4185	32738	57
jurkat	3258	32738	54
memory_t	10224	32738	49
naive_cytotoxic	11953	32738	47
naive_t	10479	32738	48
regulatory_t	10263	32738	49

Table 1: Structure of the Single Cell data for humans avalaible on the 10X website

2 Import the data

The data have a structure that might be opened with a tool called `cellranger`[1], we were not able to use it. But actually the R Package `Matrix` has permitted to open such datasets. We are talking about `.mtx` files which use sparse way of compressing the data matrix, very important in the context of Single Cell Data.

2.1 Which datasets ?

As we want relevant results... we have chosen to work on the datasets detailed in table 2. Which means that we have $K = 5$ cell types.

2.2 How many cells to be taken into account ?

As we want to carry quick and flexible analyses, we will not work on the all cells. We have decided to use an amount of $nb_{cells} = 1000$.

[1] See [**here**](https://support.10xgenomics.com/single-cell/software/pipelines/latest/output/matrices) to get further information over the `cellranger` tool.

	Number of Cells	Number of Genes	Proportion of non null genes (%)
b_cells	10085	32738	48
cd14_monocytes	2612	32738	43
cd34	9232	32738	59
cd4_t_helper	11213	32738	49
cd56_nk	8385	32738	50

Table 2: Datasets used in the work

2.3 Which genes to be watched ?

As we work on RNA-Seq, and even more with Single Cell, it is important to not put all the genes in th to be studied dataset. The common way is to use genes with a mean value higher than a particular level, called here $count_{min} = 1$.

The mean is taken over the current dataset for each of the **K** datasets previously chosen.

According to the fact that not all the genes will have a min count higher that $count_{min}$ for the all K datasets, it would be interesting to check which genes are selected for which datasets. We can look at this through a Venn Diagram on the sets of selected genes for each cell type :

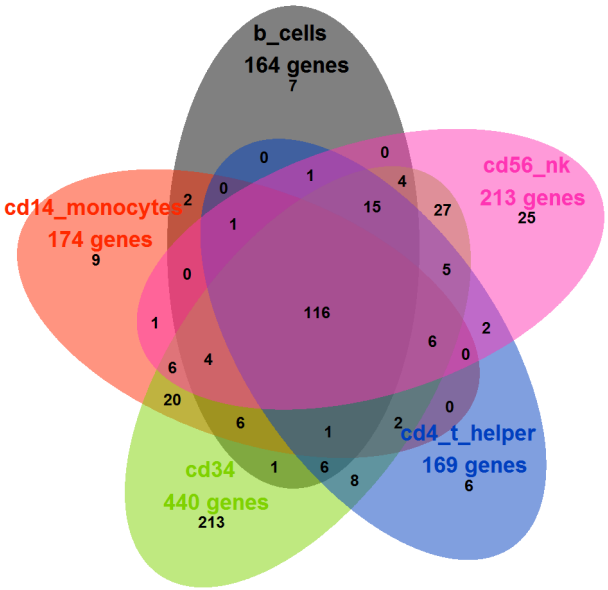

Figure 1: Venn Diagramm of the genes with a mean count higher than $count_{min} = 1$ in each of the **K=5** classes

That diagram shows that **CD34** have more very well explained genes than others. If we take the reunion of the **K** list of genes with a mean count higher than $count_{min}$, we get a total of $nb_{genes} = 494$.

Table 3 shows that **CD34** is definitly more represented through the different levels of gene expression.

2.4 Summary

Once those choices has been applied we get a matrix of dimensions $(n, p) = 5000, 494$ as

	Total number	Total proportion (%)	No Overlap number	No Overlap proportion (%)
b_cells	164	33	7	1
cd14_monocytes	174	35	9	2
cd34	440	89	213	43
cd4_t_helper	169	34	6	1
cd56_nk	213	43	25	5

Table 3: Main details of the dataset for the chosen parameters

Figure 2 shows the dispersion for each gene selected in the **K** different situations. We have normalised the counts with a `log` transformation [2], more precisely

$$count \to log(1 + count),$$

It seems that **CD56** shows high variability for the high counts while **CD14** and **B-cells** seem to shrink their variability in that high counts region. In the medium counts region, the one in which counts are the most variable, it seems that **CD56** show huge variability, but also **CD34** and **CD4**.

```
## pdf
##   2
```

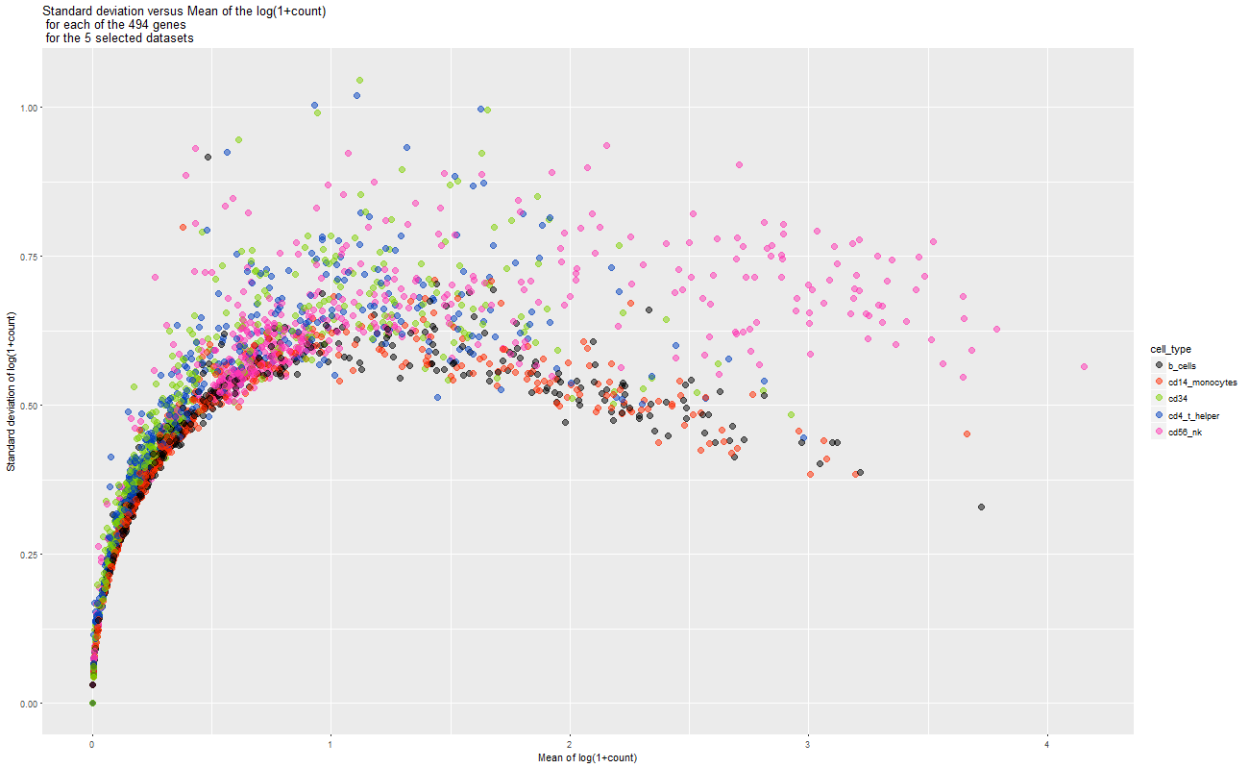

Figure 2: Dispersion profile

[2]Because all those cells are from the same very sample and we do not know anything about replicates.

3 Unsupervised Analysis

3.1 PCA for *Principal Component Analysis*

Indeed, PCA is of a great help to show colinear variables in the context of a large quantity of variables, which is the case here. We have decided to compute the **K+1** first components of that dataset and plotted the corresponding variates on figure 3. As we have **K** cell types to discriminante, it has been interesting to check the component **K+1**, at least, to check if one cell type is discriminate on that very last compoent.

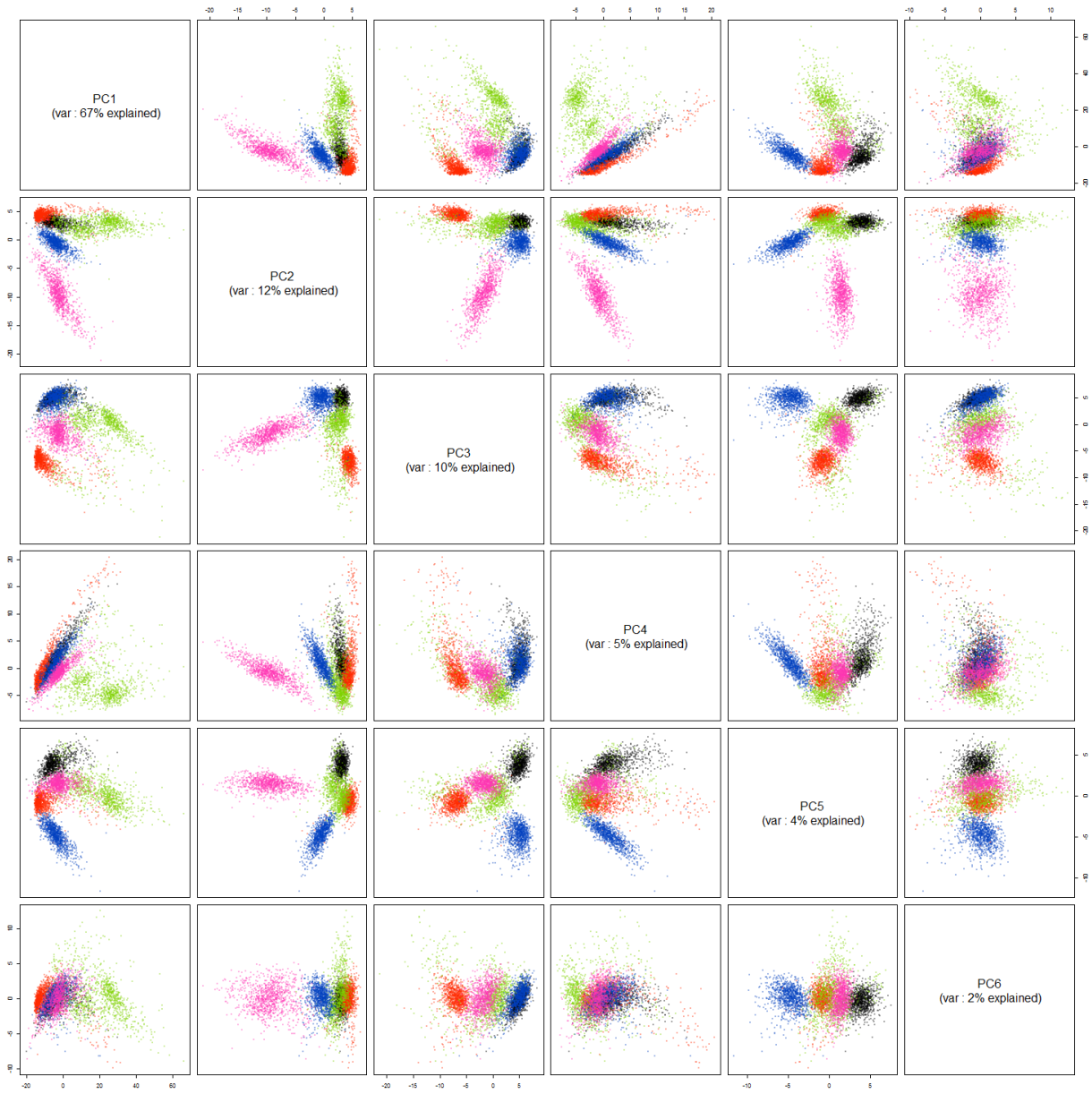

Figure 3: PCA biplots

The variance explained are actually computed on the **K** first components, this is why this is so huge. We can see also taht the first component describes more the original data than does the second component.

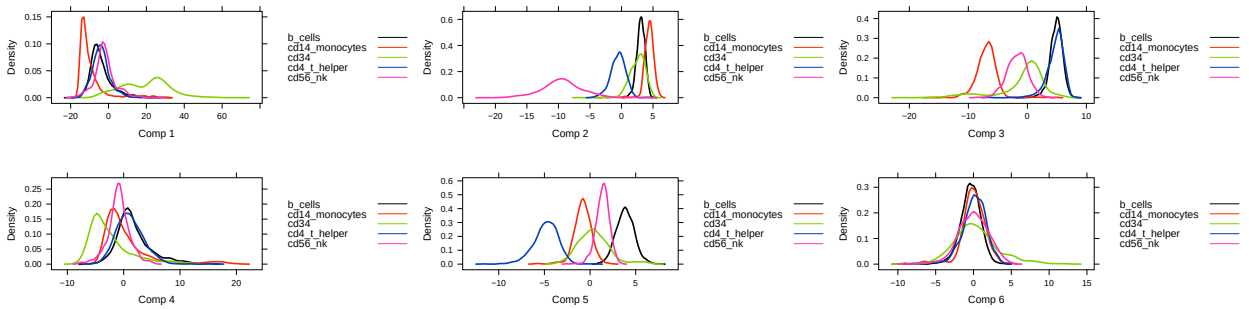

Figure 4: PCA density plots

Whatever, we can see than the 1^{st} and the 2^{nd} components are particularly capable of describing two different populations **CD34** and **CD56**. We recall, according to 3 that a large proportion of genes show large expressions for **CD34**. Consequently, **CD56** show genes with large expressions.

It also seems that the 3^{rd} component could be an expression od **CD14** but that group seems to have an expression close to the expression of the **CD34**.

The 4^{th} component does not discriminate a special type of cells. Maybe that component shows cellular variability due to the method of measure or the normalization and/or thresholding method that we used.

Whatever, the 5^{th} component is able to show a discrimination of **CD4** and **B-cells**.

Finally, the last component computed does not discriminate any cluster information, and this is why the las line of the plot is so appealling to check the univariate discrimination of each component.

A question could be, would an efficient unsupervised clustering algorithm find clusters ? We could ask Chariff!

Actually, this cannot be an answer to the question here. We want to find groups of genes which discriminate the different cell types. In that sense this is a lack of power of considering most of the variance along the first component and an useless 4^{th} component. This is why we have decided to try **sparse** methods with components, such as `SPLS-DA` ((Chung, Keles, and others 2010) and (Lê Cao, Boitard, and Besse 2011)) for *Sparse Partial Least Square Discriminant Analysis* or SDA ((Clemmensen et al. 2011)) for *Sparse Discriminant Analysis*.

3.2 A link with the Cellular Discovery Rate

As there is no way to normalize along the library size, it might be interesting to look at the CDR. As a first shot we have decided to check at the following formula

$$pseudoCDR_i = \frac{1}{N}\sum_{g=1}^{N} z_{ig}$$

Where

- i is the index of a cell,

- g is the index of a gene,
- N is the total number of genes,
- z_{ig} is an indicator if gene g in cell i was expressed above *background*.

We have used the previous dataset, for which $(n, p) = 5000, 494$, in our case $N = p = 494$.

We have fixed the background to $background = 2$ and so considered a gene expressed if its expression for cell i is above the threshold which is here equal to $background = 2$.

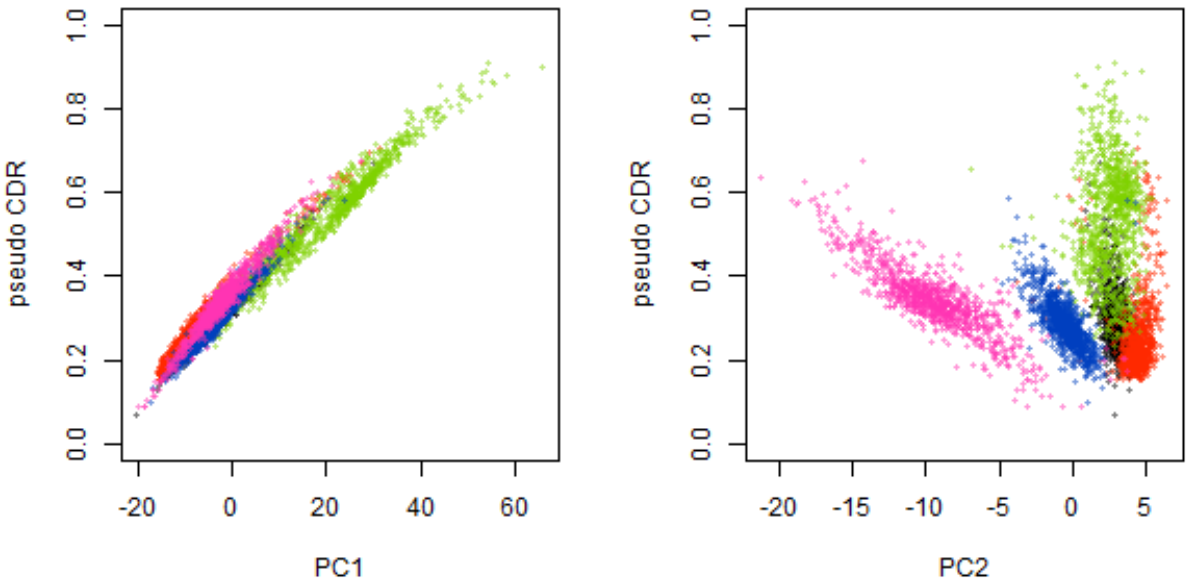

Figure 5: Pseudo CDR according to PC1 and PC2

4 Supervised Analysis

4.1 SPLS-DA

As a proof of concept we will use the `mixOmics` package, we have used a common $keep_X = 50$ to the **K** components and we have constructed $ncomp = 5$ components.

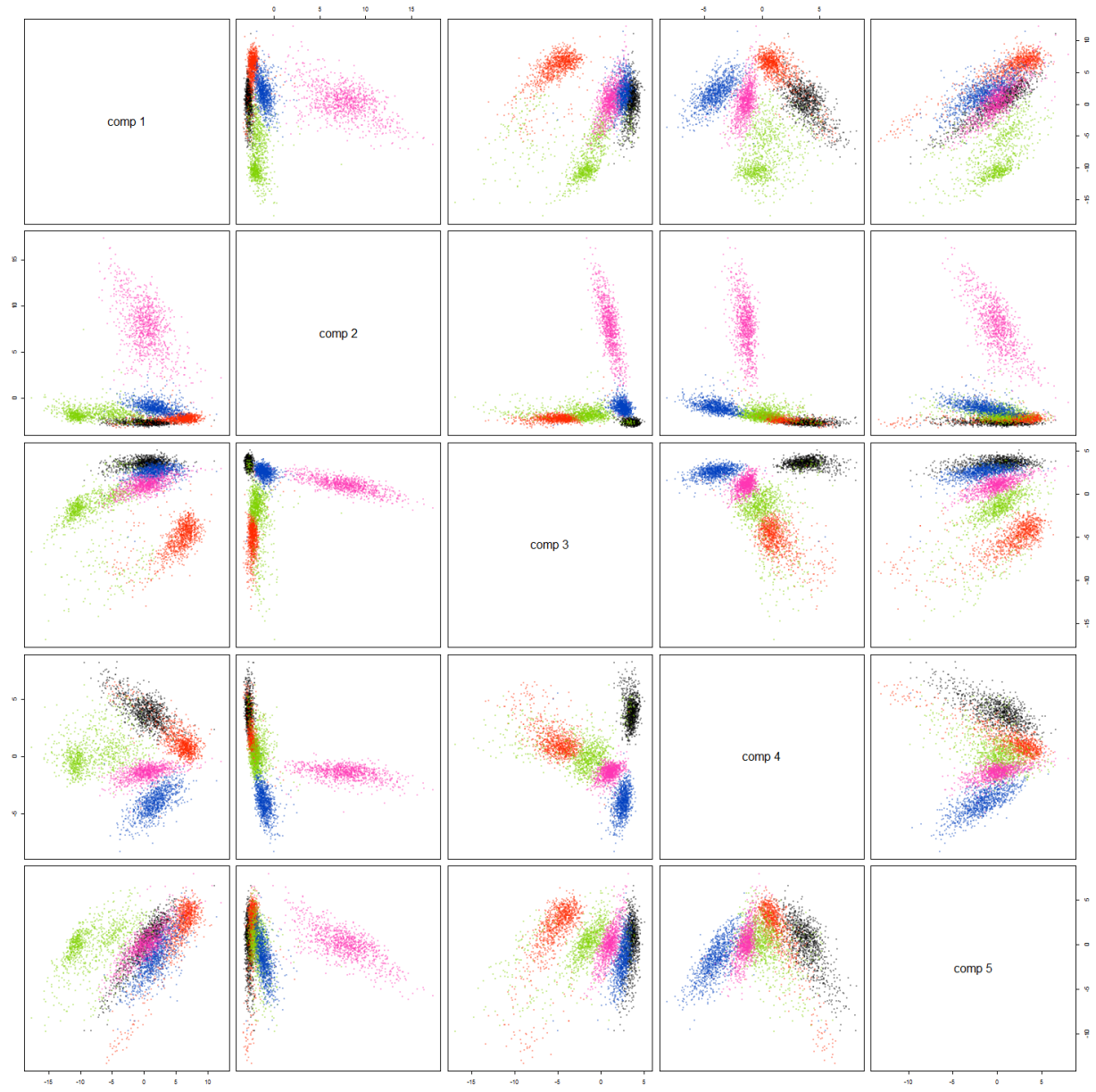

Figure 6: SPLSDA biplots

We can make a few comments on those results :

- 1^{st} component discriminates **CD34** before others,

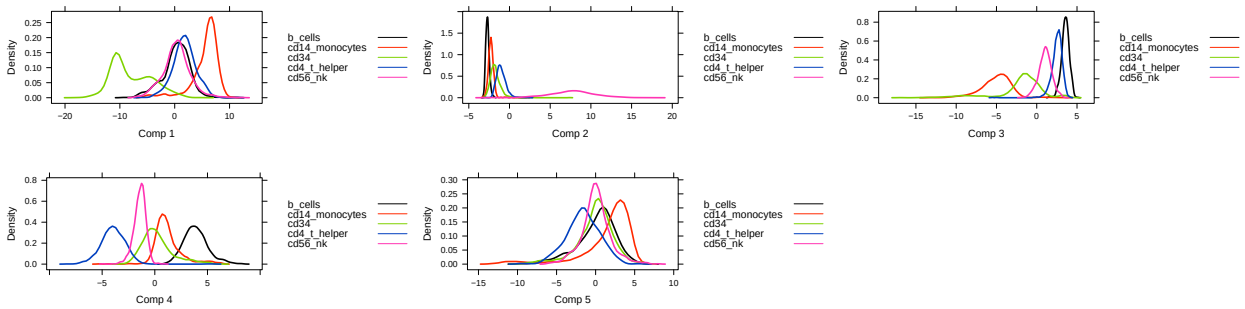

Figure 7: SPLSDA density plots

- 2^{nd} component discriminates **CD56** before others,
- 3^{rd} component discriminates **CD14** before others,
- 4^{th} component discriminates **CD4** and **B-cells** before others,
- 5^{th} component does not seem to descriminate a particular group of cells.

The following Venn Diagramm gives the behaviors of the different component selected genes.

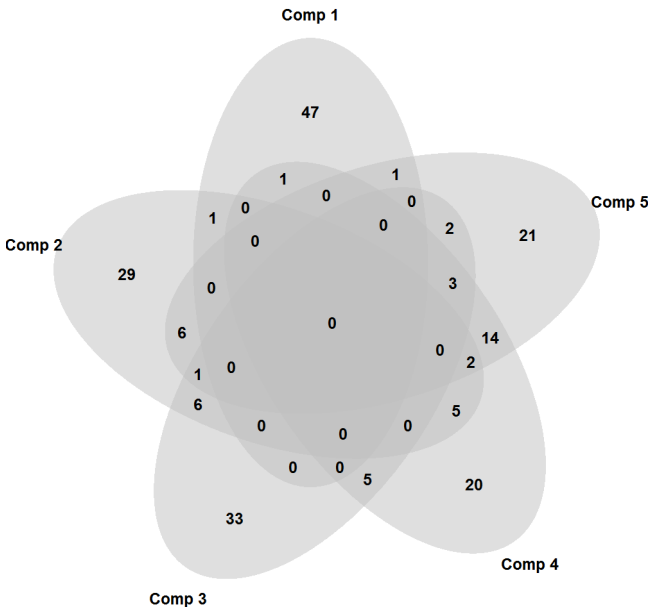

Figure 8: Venn Diagramm of the genes selected by the SPLSDA on the **K** first components

4.1.1 A link with the Cellular Discovery Rate

As there is no way to normalize along the library size, it might be interesting to look at the CDR. We have fixed the parameters as we did for unsupervised part.

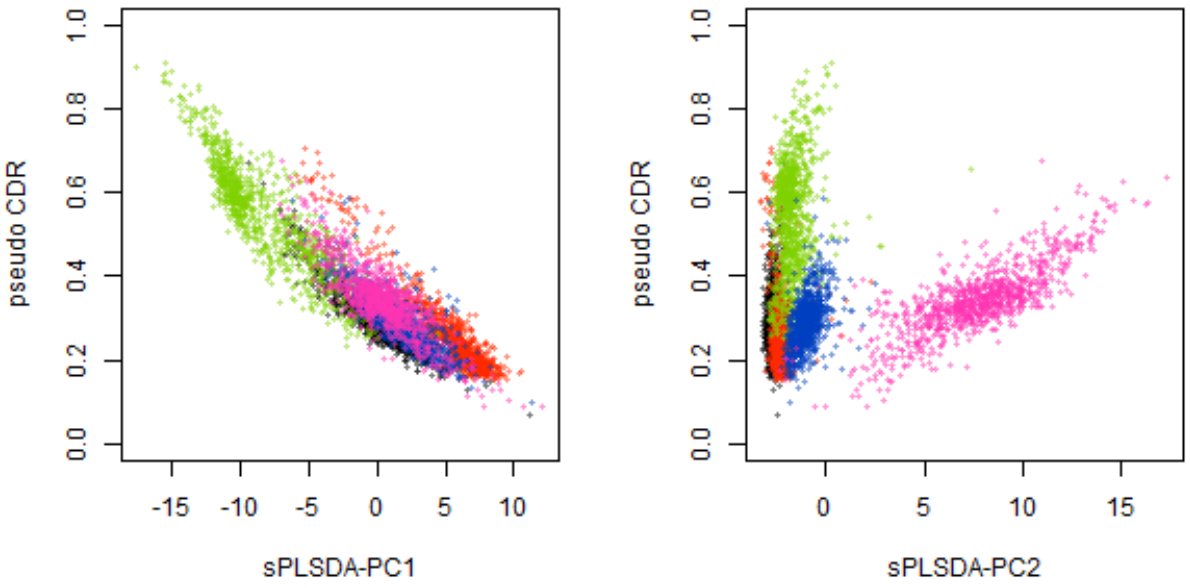

Figure 9: Pseudo CDR according to sPLSDA-PC1 and sPLSDA-PC2

References

Chung, Dongjun, Sunduz Keles, and others. 2010. "Sparse Partial Least Squares Classification for High Dimensional Data." *Statistical Applications in Genetics and Molecular Biology* 9 (1). bepress: 17.

Clemmensen, Line, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. 2011. "Sparse Discriminant Analysis." *Technometrics* 53 (4). Taylor & Francis: 406–13.

Lê Cao, Kim-Anh, Simon Boitard, and Philippe Besse. 2011. "Sparse PLS Discriminant Analysis: Biologically Relevant Feature Selection and Graphical Displays for Multiclass Problems." *BMC Bioinformatics* 12 (1). BioMed Central: 253.

Zheng, Grace XY, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, et al. 2016. "Massively Parallel Digital Transcriptional Profiling of Single Cells." *BioRxiv*. Cold Spring Harbor Labs Journals, 065912.