Batch effect in single cell data

Pierre Bost
pierre.bost@pasteur.fr

Causa latet, vis est notissima

Sources of batch effect

At least 3 possible sources of technical variations :

- Differences in **sequencing depth** / library size : number of UMI per cell will change **across batches**…

- Differences in **gene sensibility detection** : variation in primer/ capture process of the mRNA or in the mapping… Will change across **technology/platform**

- Difference in **global cell state** : cells can be stressed by many steps of the sequencing process.

Sources of batch effect

At least 3 possible sources of technical variations :

- Differences in **sequencing depth** / library size : number of UMI per cell will change **across batches**…

➡ Sampling strategy, adapted distance measure

- Differences in **gene sensibility detection** : variation in primer/capture process of the mRNA or in the mapping… Will change across **technology/platform**

- Difference in **global cell state** : cells can be stressed by many steps of the sequencing process.

Sources of batch effect

At least 3 possible sources of technical variations :

- Differences in **sequencing depth** / library size : number of UMI per cell will change **across batches**…

➡ Sampling strategy, adapted distance measure

- Differences in **gene sensibility detection** : variation in primer/capture process of the mRNA or in the mapping… Will change across **technology/platform**

- Difference in **global cell state** : cells can be stressed by many steps of the sequencing process.

➡ Remove/regress out the batch « effect » genes

Sources of batch effect

At least 3 possible sources of technical variations :

- Differences in **sequencing depth** / library size : number of UMI per cell will change **across batches**…

➡ Sampling strategy, adapted distance measure

- Differences in **gene sensibility detection** : variation in primer/capture process of the mRNA or in the mapping… Will change across **technology/platform**

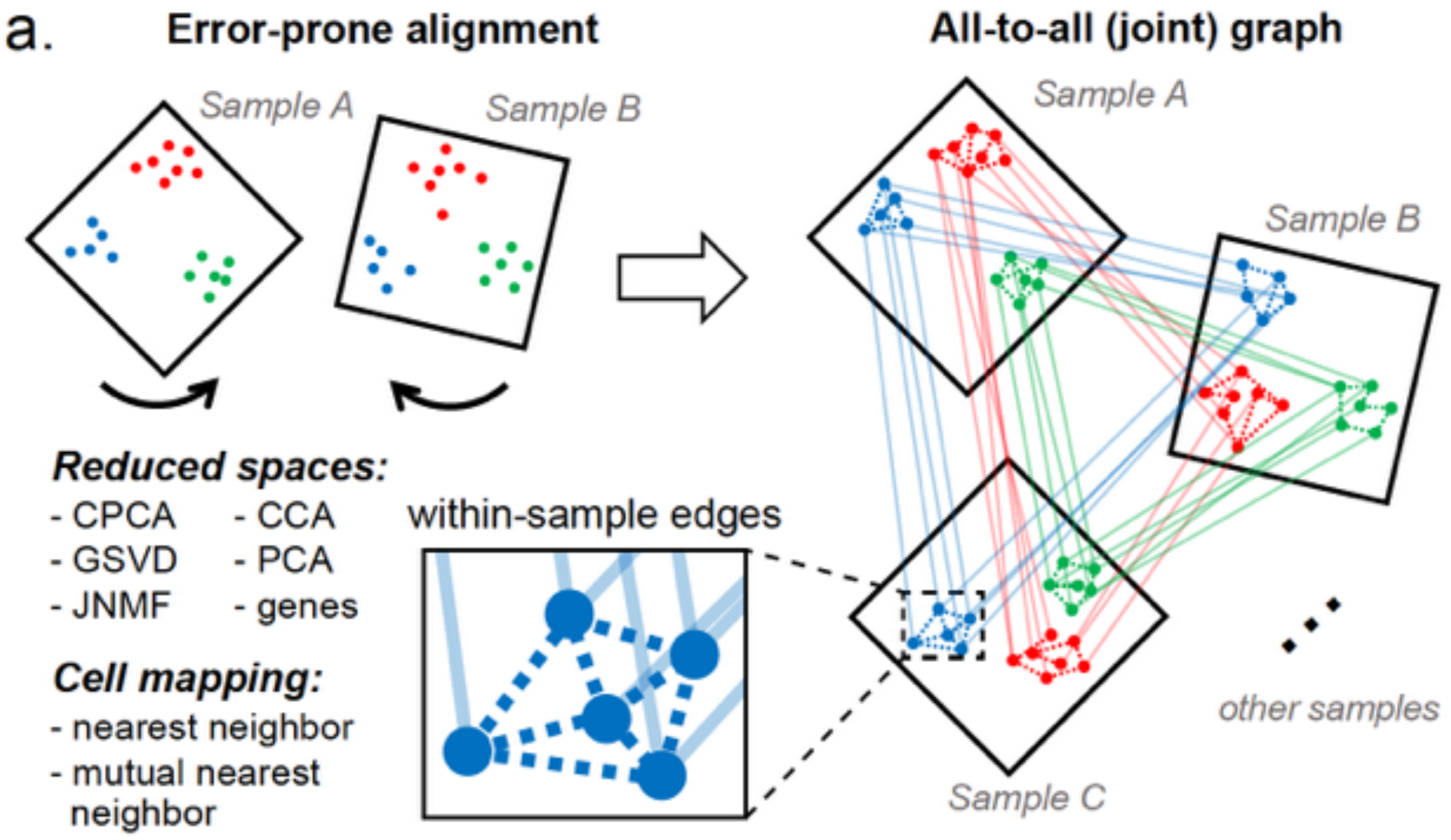
➡ Not clear… Multiple strategies are possible….

- Difference in **global cell state** : cells can be stressed by many steps of the sequencing process.

➡ Remove/regress out the « batch effect » genes »

Conos tool (1)

Wiring together large single-cell RNA-seq sample collection
Barkas et al



Published on biorxiv in November 2018 : aggregation of multiple batch using weighted graphs

Conos tool (2)

3 different steps of the pipeline :

- Pre-**processing/normalisation** of each dataset independently

- Removing low quality cells
- Normalising the expression/ variance adjustment
- Identification of highly variable genes
- Can be performed in Pagoda2 or Seurat pipeline

Conos tool (2)

3 different steps of the pipeline :

- Pre-**processing/normalisation** of each dataset independently
- Creation of a **joint space** and **mapping** of the different dataset

- Creation of a joint space that can be :
  - ➡ Merged gene space (union of variable genes)
  - ➡ PCA space of joined dataset
  - ➡ Result of joint factorisation (JNMF, CPCA but not CCA….)

Conos tool (2)

3 different steps of the pipeline :

- Pre-**processing/normalisation** of each dataset independently
- Creation of a **joint space** and **mapping** of the different dataset
- **Construction** of the graph + **imputation/smoothing** analysis

➡ Identification of inter-batch neighbour (mNN or simple kNN)

➡ Creation of a graph using strong inter-batch connection and adding intra-batch connection with low weight…

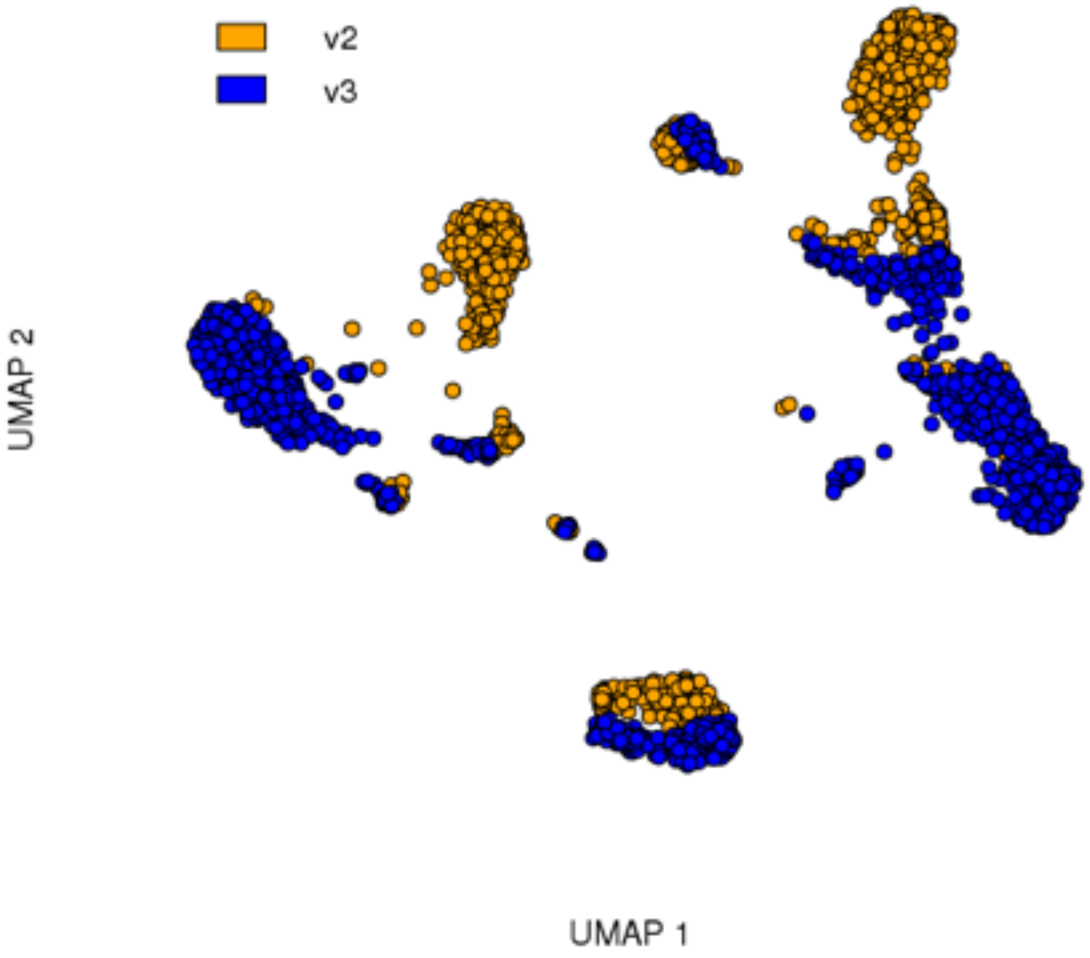
➡ Multiple analysis possible :

- Graph clustering (Louvain, Walktrap etc…)
- Variable imputation using local value smoothing…
- Label assignment…
- Graph visualisation

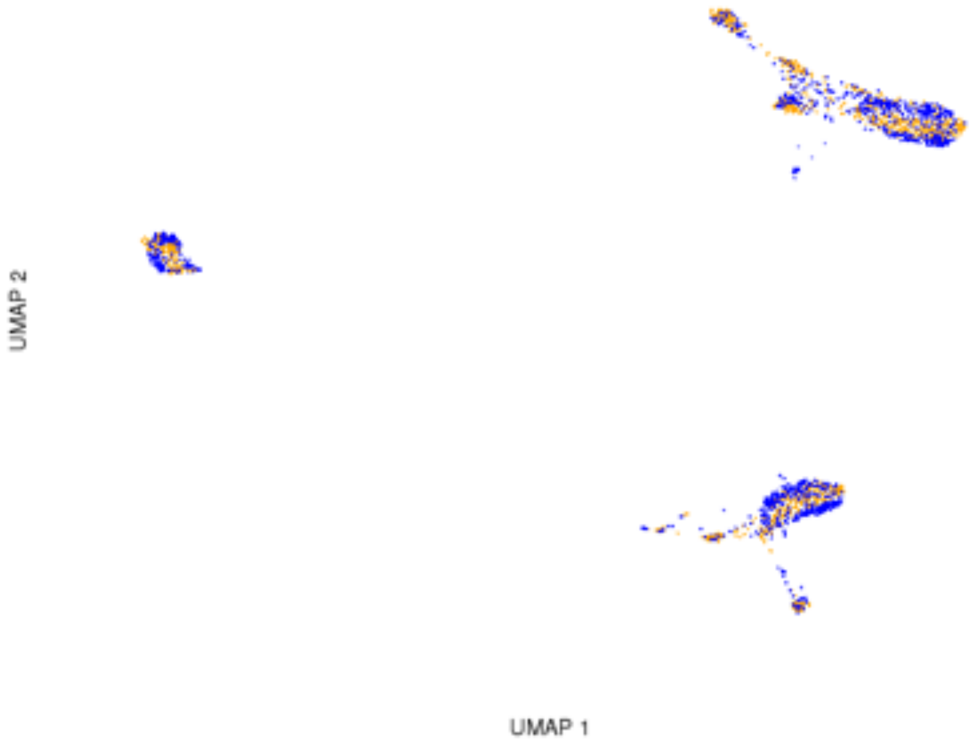
Conos tool : test

3 different datasets : PBMC from healthy donors, 10X sequencing.
V2, V3 and 5’ sequencing.

First test : merging v2 and v3 with CPCA and default parameters



UMAP plot of the
v2 and v3 datasets



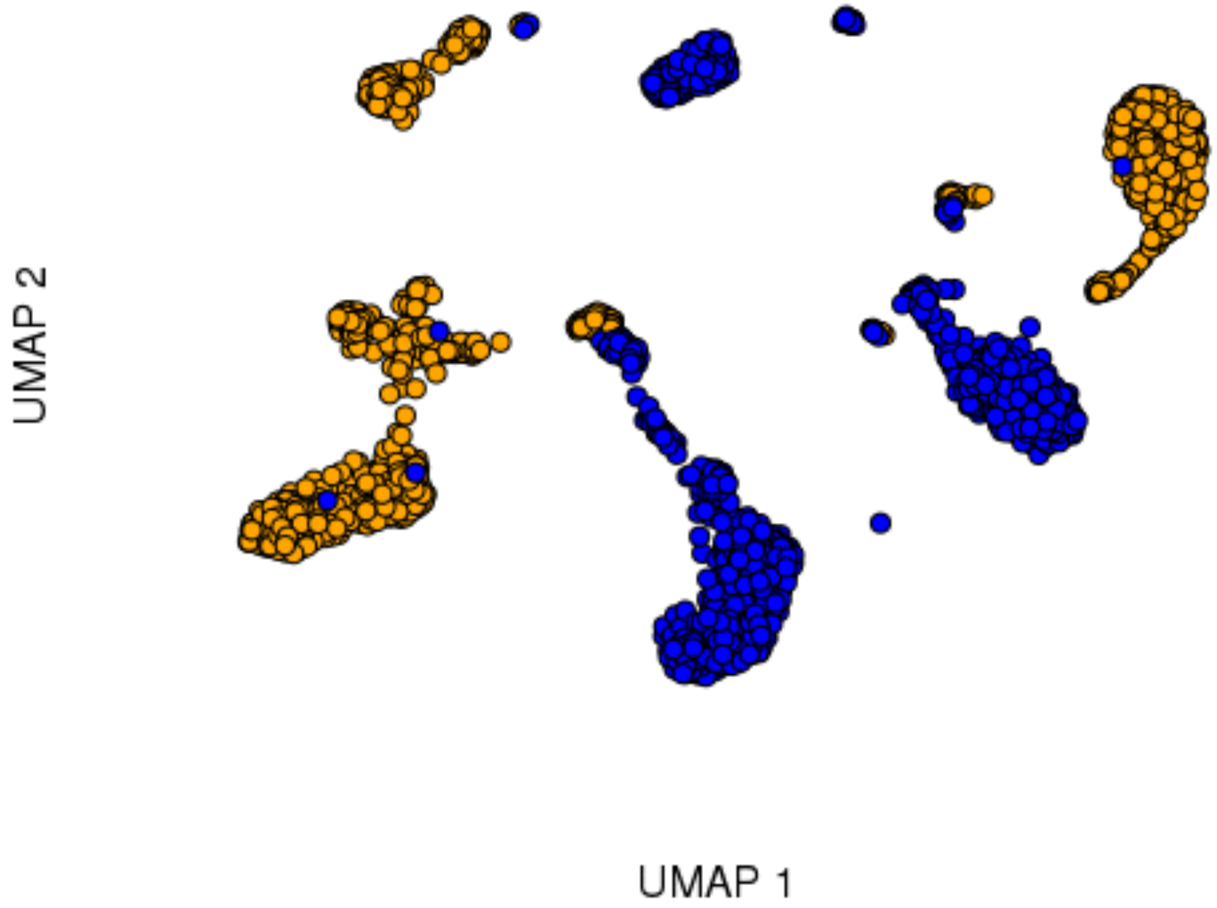
LargeVis plot of the merged graph

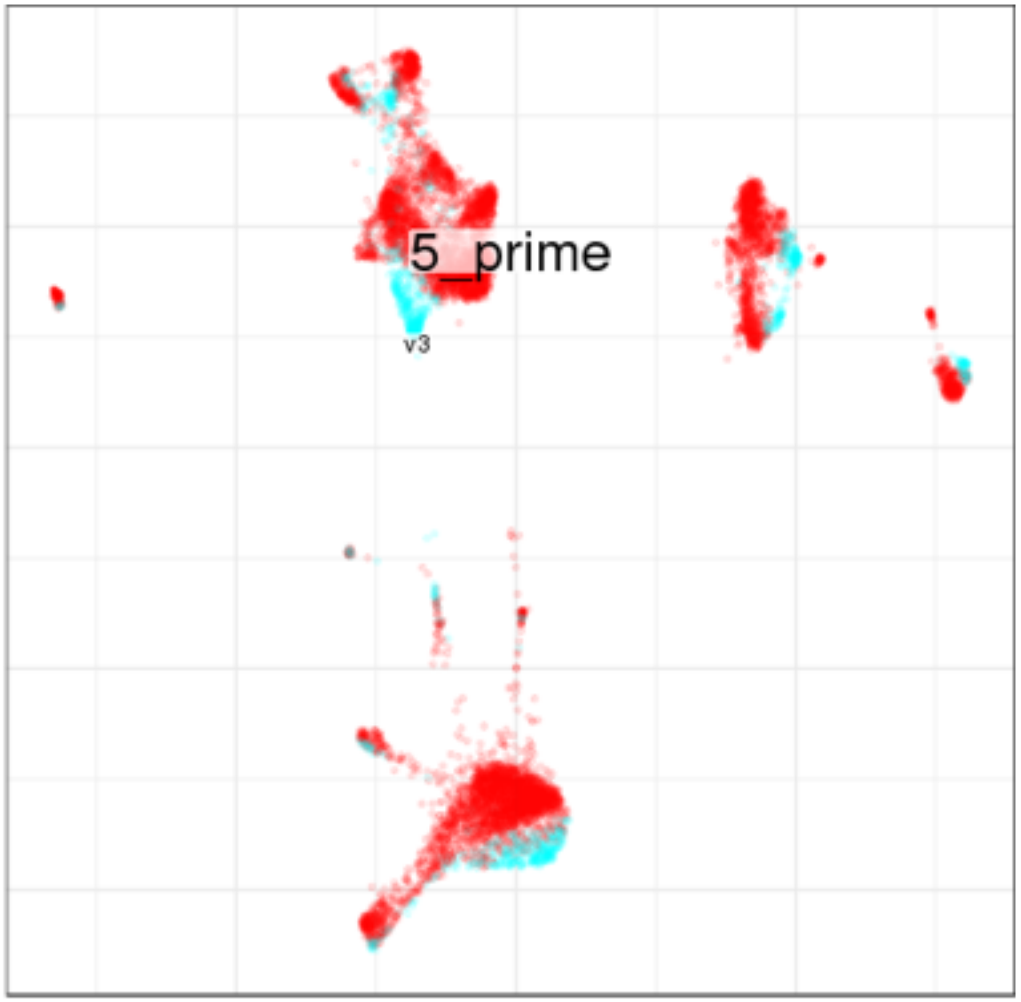
-> Not nice mixing….

Conos tool : test

3 different datasets : PBMC from healthy donors, 10X sequencing. V2, V3 and 5’ sequencing.

Second test : merging v3 and 5’ with CPCA and default parameters

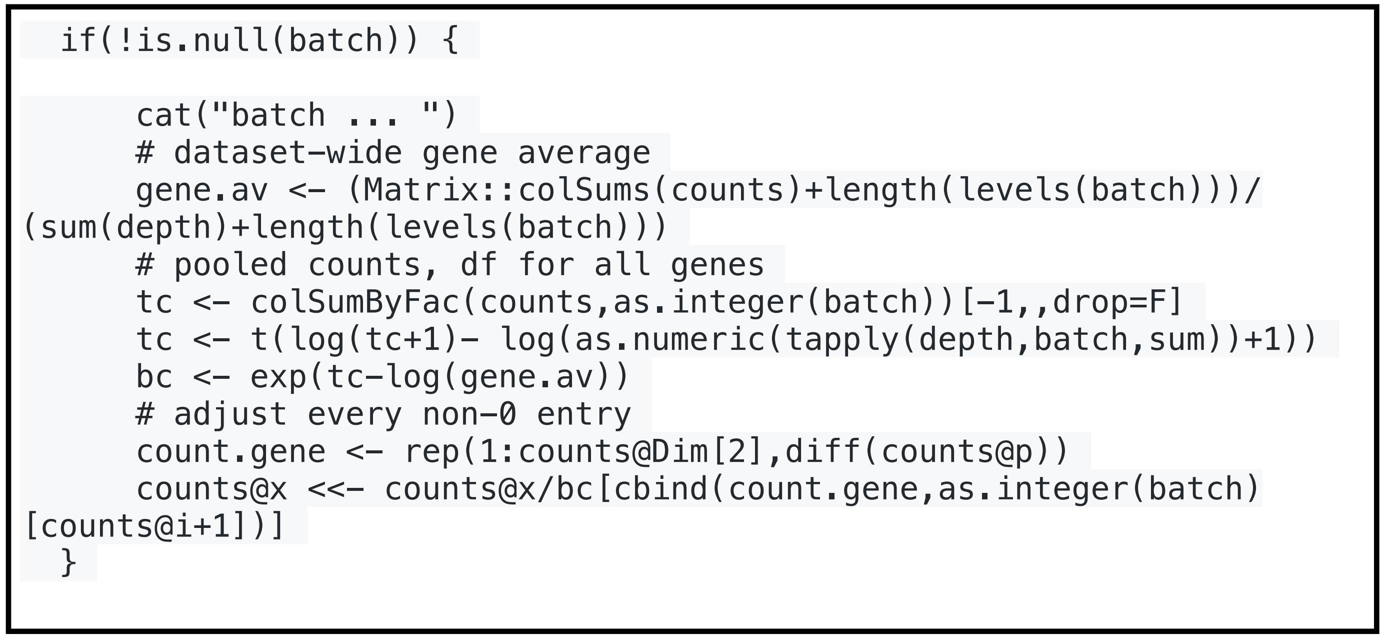




-> Batch correction does not work whatever the strategy is (JNMF, PCA, CPA…)

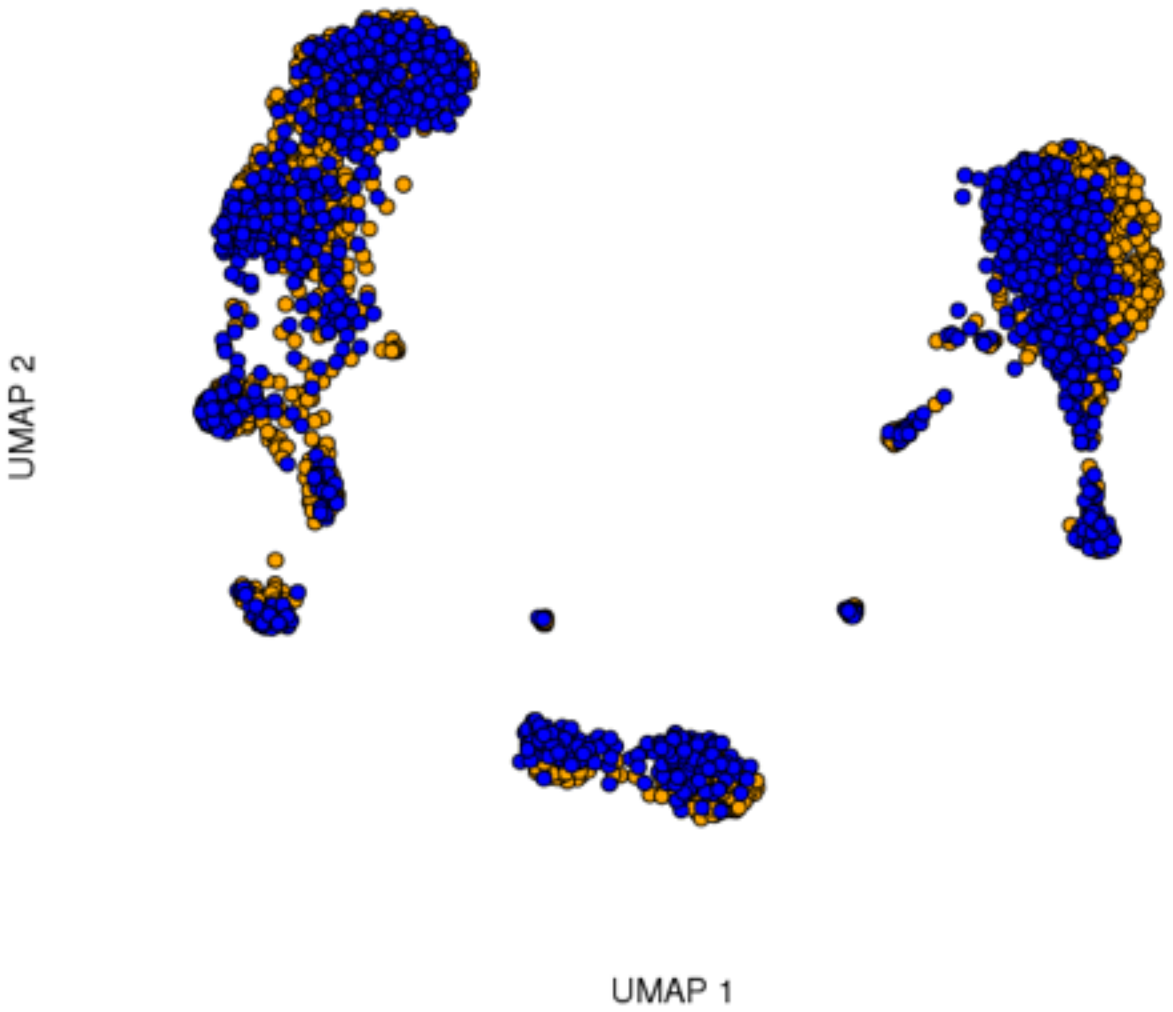
An « old » alternative in Pagoda2

Hidden function in the code : possibility to include batch correction during loading of the data….



```
  if(!is.null(batch)) {

      cat("batch ... ")
      # dataset-wide gene average
      gene.av <- (Matrix::colSums(counts)+length(levels(batch)))/
(sum(depth)+length(levels(batch)))
      # pooled counts, df for all genes
      tc <- colSumByFac(counts,as.integer(batch))[-1,,drop=F]
      tc <- t(log(tc+1)- log(as.numeric(tapply(depth,batch,sum))+1))
      bc <- exp(tc-log(gene.av))
      # adjust every non-0 entry
      count.gene <- rep(1:counts@Dim[2],diff(counts@p))
      counts@x <<- counts@x/bc[cbind(count.gene,as.integer(batch)
[counts@i+1])]
  }
```



Seems to work !

Seems to be a simple gene by gene scaling factor computation ….

Extremely fast to compute and as efficient as Conos but relies on a highly similar cell distribution….

CCA : Canonical Correlation Analysis

- **CCA** : generalisation of many factorial approches, including multiple linear regression, PCA etc…

- Developed by **Hotelling in 1936**.

- Usually considered as outdated and not practically useful, the progress in the single-cell omic field made it a key tool for the analysis !

Lets say we have two data matrix **X** and **Y** with the same number of rows (samples) ***n*** but with ***p*** and ***q*** columns (variables).

We are looking for two vectors ***a*** and ***b*** that are **maximizing the correlation** $r = cor(X * a, Y * b)$.

The random variables **U = X * a** and **V = Y * b** are the **first pair of canonical variables**

CCA : a bit of math

Lets say we have two data matrix **X** and **Y** with the same number of rows (samples) ***n*** but with ***p*** and ***q*** columns (variables).

- To make the calculus simpler : all variables have a mean equal to zero and a variance equal to 1.
  ➡ Hence **r** = cor(X * a, Y * b) = (X*a)' * (Y*b) = **a'X'Yb**

- We also want that the new vectors have the same L2 norms, hence :
  ➡ $\| X*a \|_2 = \| Y*b \|_2$ = **a'X'Xa = b'Y'Yb = 1**

We therefore want to solve

argmax(**a'X'Yb)**
with **a'X'Xa = b'Y'Yb = 1**

CCA : a bit of math (2)

$$\boxed{\begin{array}{c}\mathrm{argmax}(\mathbf{a'X'Yb}) \\ \text{with } a'X'Xa = b'Y'Yb = 1\end{array}}$$

-> Use of two Lagrange multiplier $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$

$$\boxed{L = \mathbf{a'X'Yb - \mu * a'X'Xa - \lambda * b'Y'Yb}}$$

If we derive in respect to **a** and **b** and that we look at a maximum point we have :

$$\boxed{\begin{array}{c}\mathbf{X'Yb - 2\mu * X'Xa = 0} \\ \mathbf{Y'Xa - 2\,\lambda * Y'Yb = 0}\end{array}} \;\mathbf{->}\; \boxed{\begin{array}{c}\mathbf{a'X'Yb = 2 * \mu * a'X'Xa = 2\mu} \\ \mathbf{b'Y'Xa = 2\,\lambda * b'Y'Yb = 2\lambda}\end{array}} \;\mathbf{->}\; \boldsymbol{\mu = \lambda}$$

We set $\mathbf{2\mu = 2\,\lambda = \beta},$ therefore

$$\boxed{\begin{array}{c}\mathbf{X'Yb = \beta * X'Xa} \\ \mathbf{Y'Xa = \beta * Y'Yb}\end{array}}$$

Lets suppose **X'X** and **Y'Y** are not singular hence

$$\boxed{\mathbf{a = (X'X)^{-1}X'Yb / \beta}}$$

$$\boxed{\begin{array}{c}\mathbf{Y'X(X'X)^{-1}X'Yb = \beta^2 * Y'Yb} \\ \mathbf{Mb = \beta^2\, b}\end{array}}$$

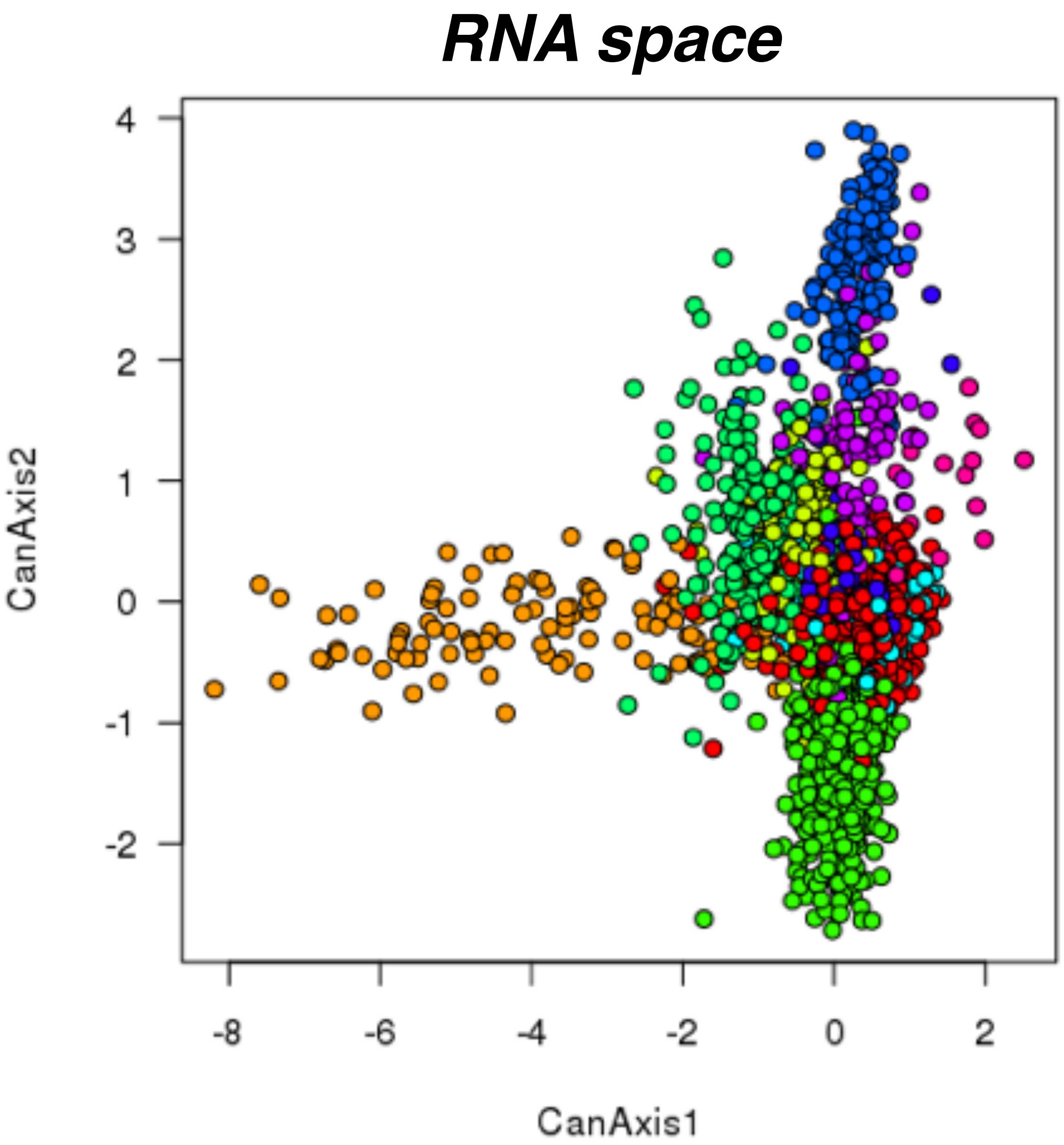
b is therefore the first eigenvector and $\boldsymbol{\beta}$ the largest eigenvalue of the matrix M with

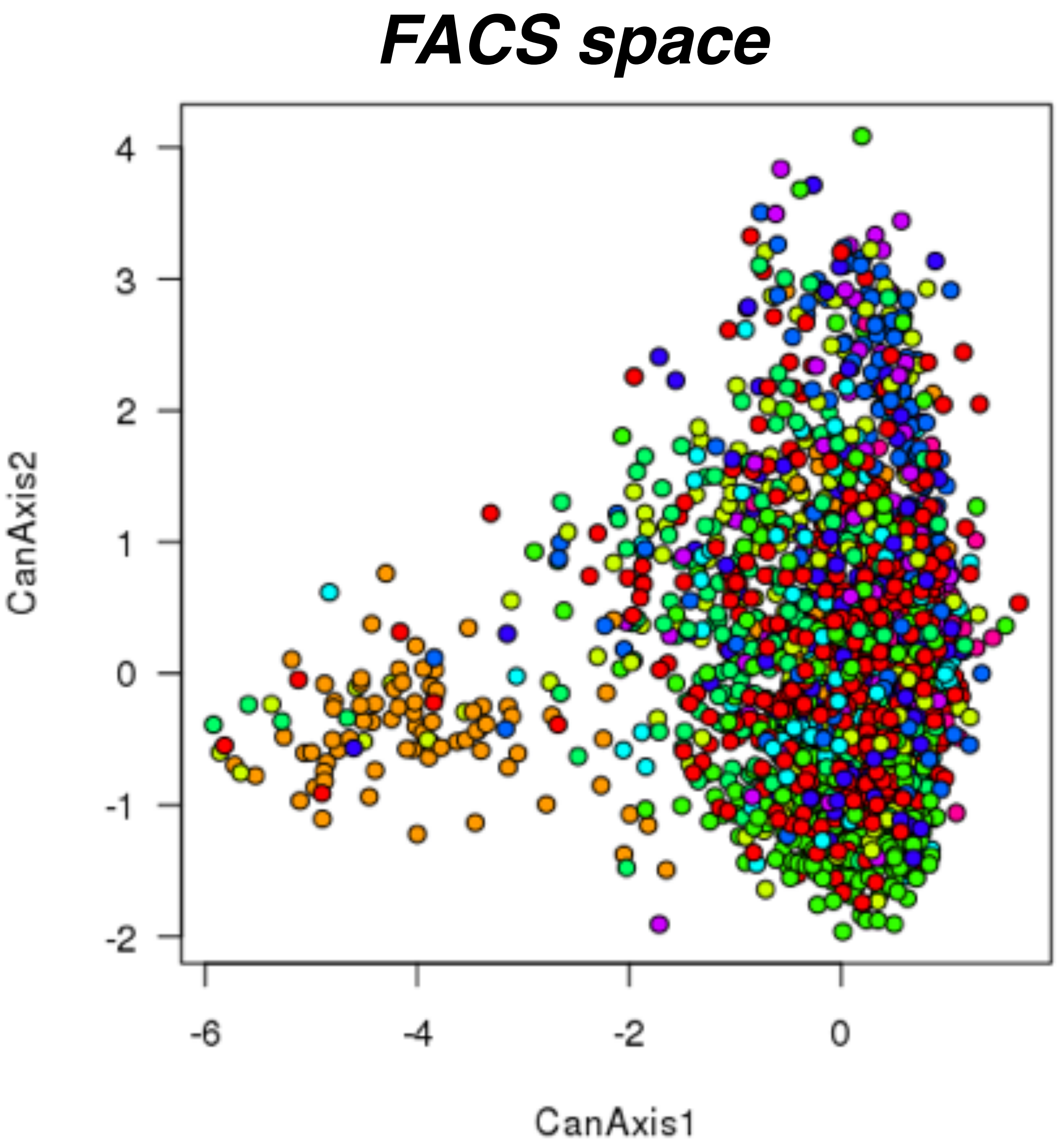
$$\boxed{\begin{array}{c}\mathbf{M = (Y'Y)^{-1}Y'X(X'X)^{-1}X'Yb} \\ \mathbf{M = (\Sigma_Y)^{-1}\Sigma_{YX}\,(\Sigma_X)^{-1}\Sigma_{XY}}\end{array}}$$

CCA : practical use

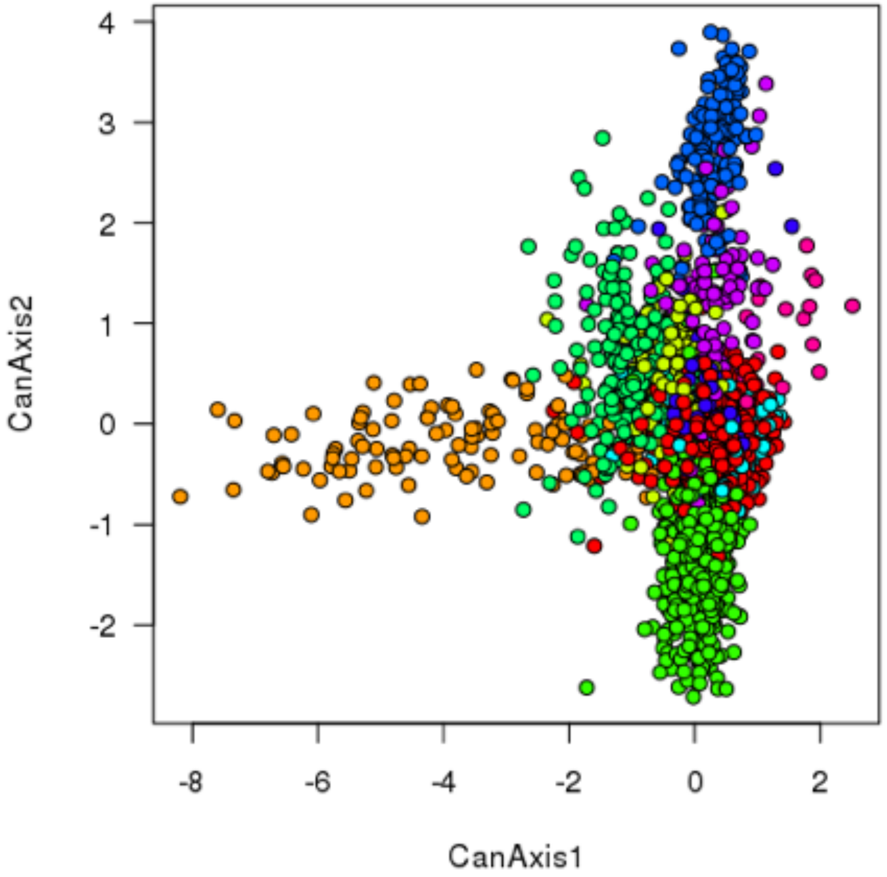
- 2 cases of use for single-cell datasets : Multi-omic analyse and batch merging.

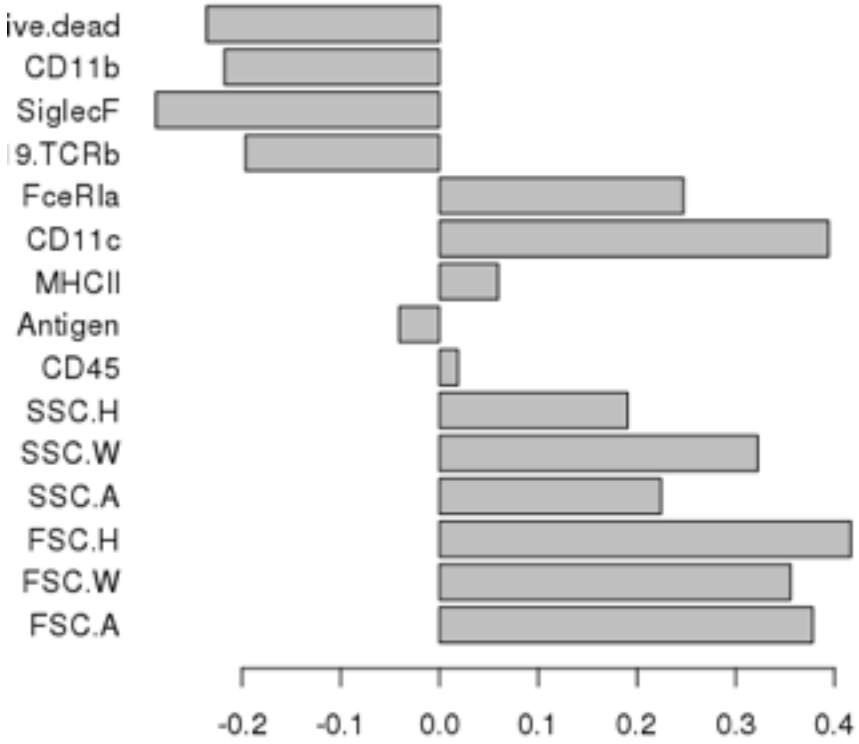
First case : Multi-omic. Here MARS-seq data + Index sorting data (~2800 cells)



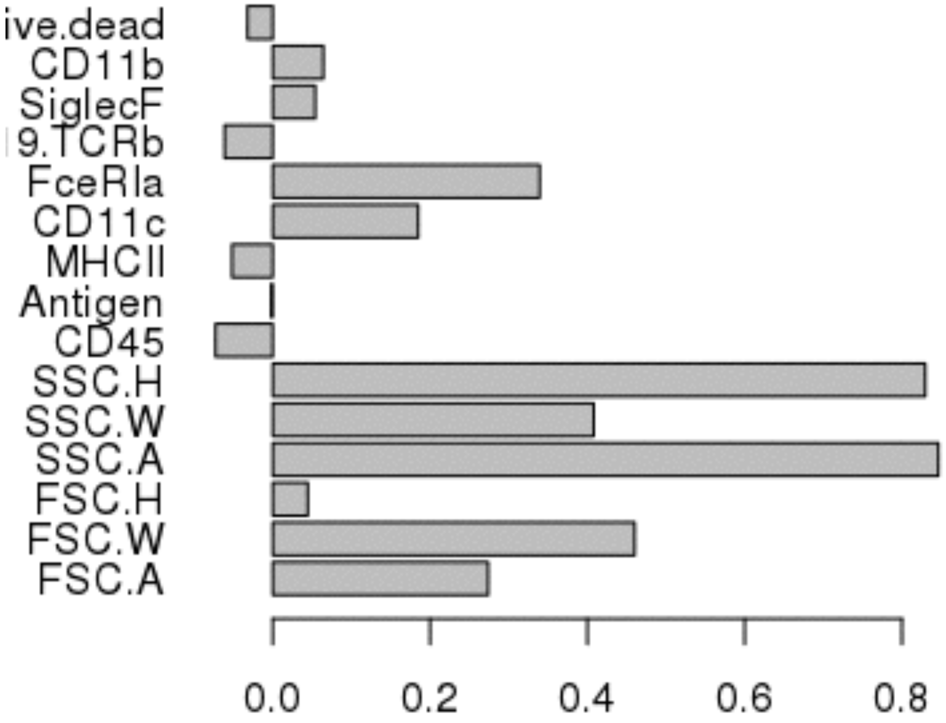


How to interpret ? Here dimension 1 corresponds to B cells vs Myeloid cells and dimension 2 correponds to Mono vs Macro…





Identify size and granularity as the key factor to distinguish B vs Myelo and Granularity as the only factor to distinguish Macro from mono….

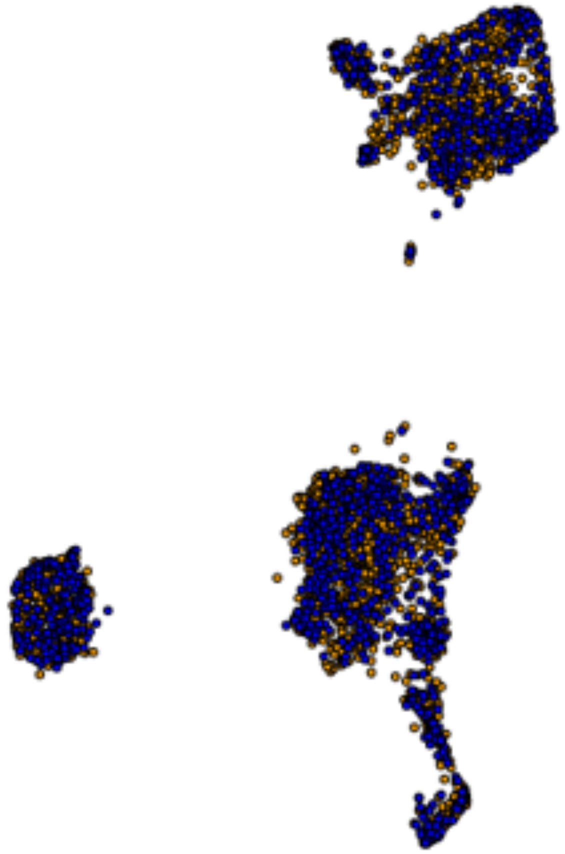


What about merging two different datasets ?

To merge 2 datasets : we take the transpose of both matrices…
X : [p genes ; n_x cells] and **Y** : [p genes ; n_y cells].

We want to identify **gene structures/modules** that are conserved across the two datasets. The two **projection vectors (a and b)** can be considered as the contribution of each cell to the **conserved gene modules.**

Use of a **diagonalised/penalized version of the CCA** developed by Hastie and Tibshirani (projecting vectors with the same norm?). Implementation in the PMA package that also supports multiple CCA….



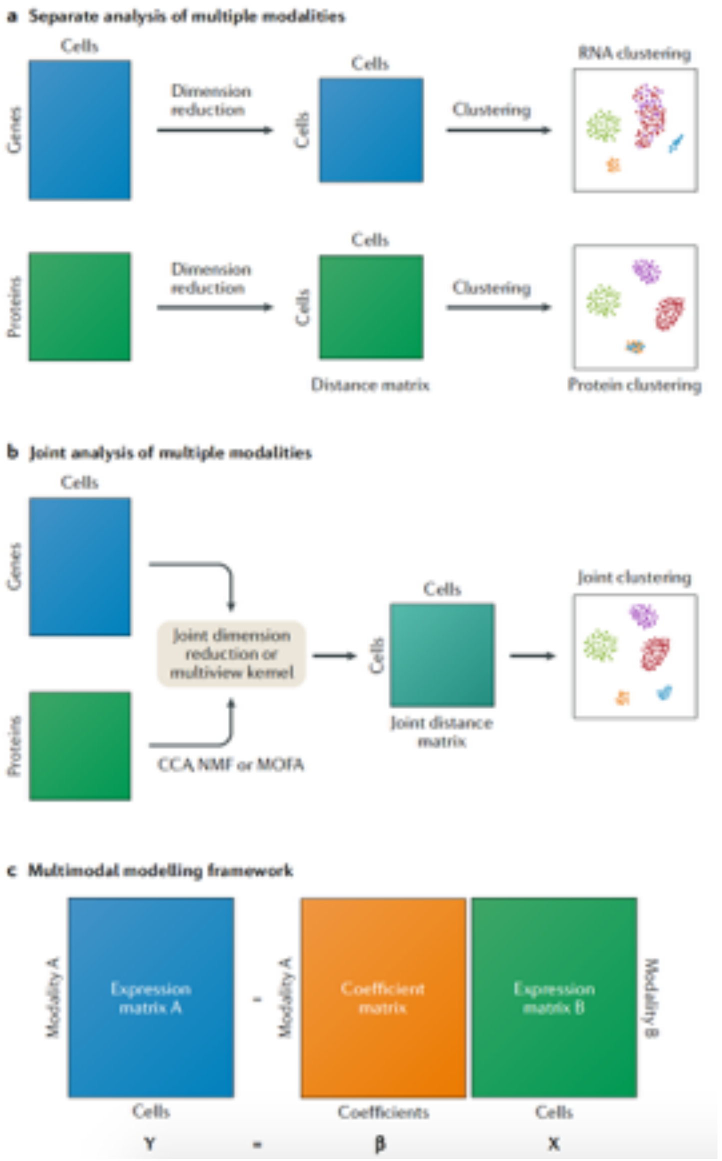
Really nice performance with the 5' and v3 chemistry…
Need to test it more deeply but could help for large scale studies of patients for instance…

What else to look at ?

Use of deep neural network : scVI (Nir Yosef Lab) or DCA (Fabian Theis Lab)…

Reported use of a specific metric : Maximum Mean Discrepancy (SAUCIE paper) that need to be tested…

Alternative : physical cell hashing using barcoded antibodies (Satija approach…)



Recent review by Satija in Nature Genet…

References

- **Seurat v3 paper** : https://www.biorxiv.org/content/10.1101/460147v1
- **Conos paper** : https://www.biorxiv.org/content/10.1101/460246v1
- **Cell hashing** : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6300015/
- **Review on multiomic/batch** : https://www.nature.com/articles/s41576-019-0093-7
- **Introduction to CCA** : Statistique exploratoires multidimensionnelles (Lebart, Piron et Morineau)
- **Diagonalised CCA paper** : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2697346/