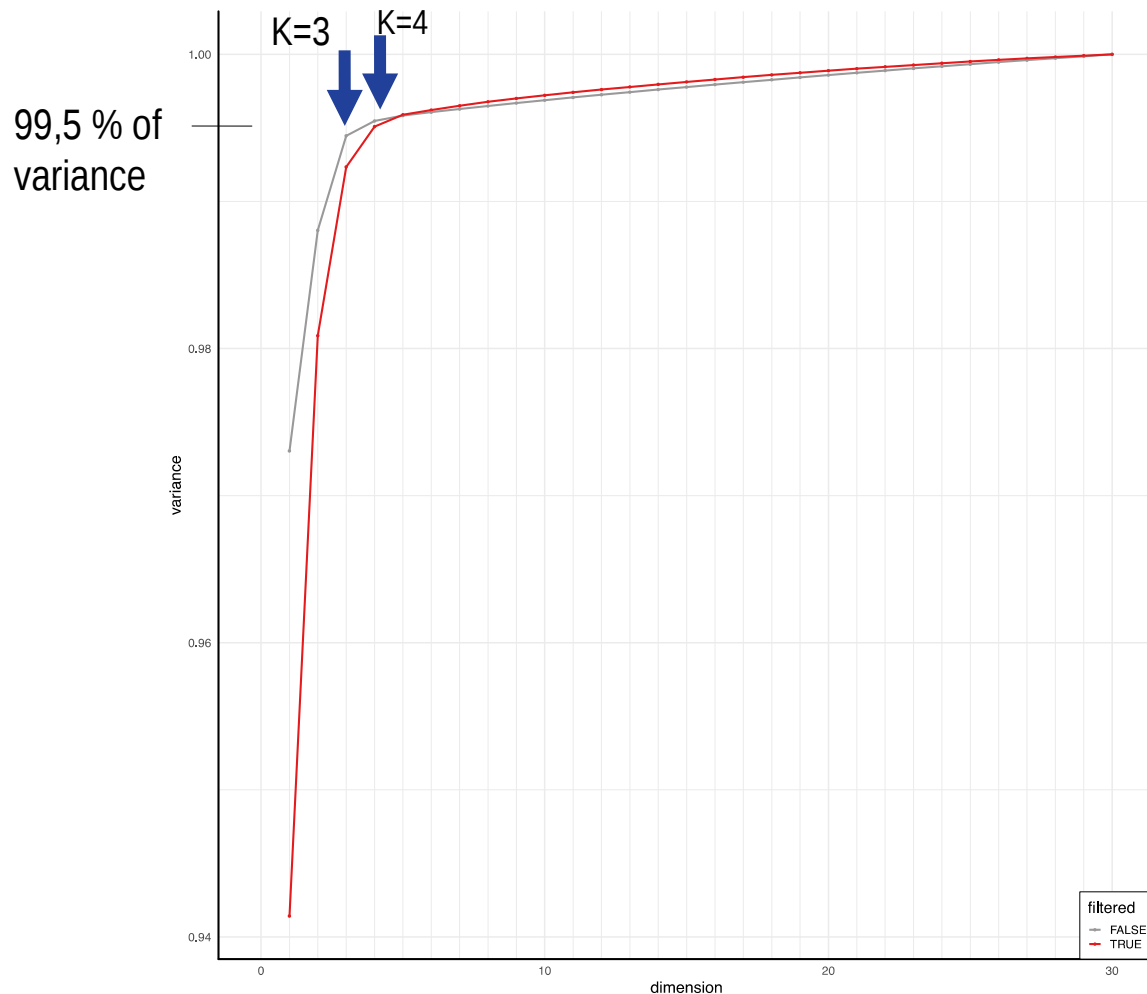




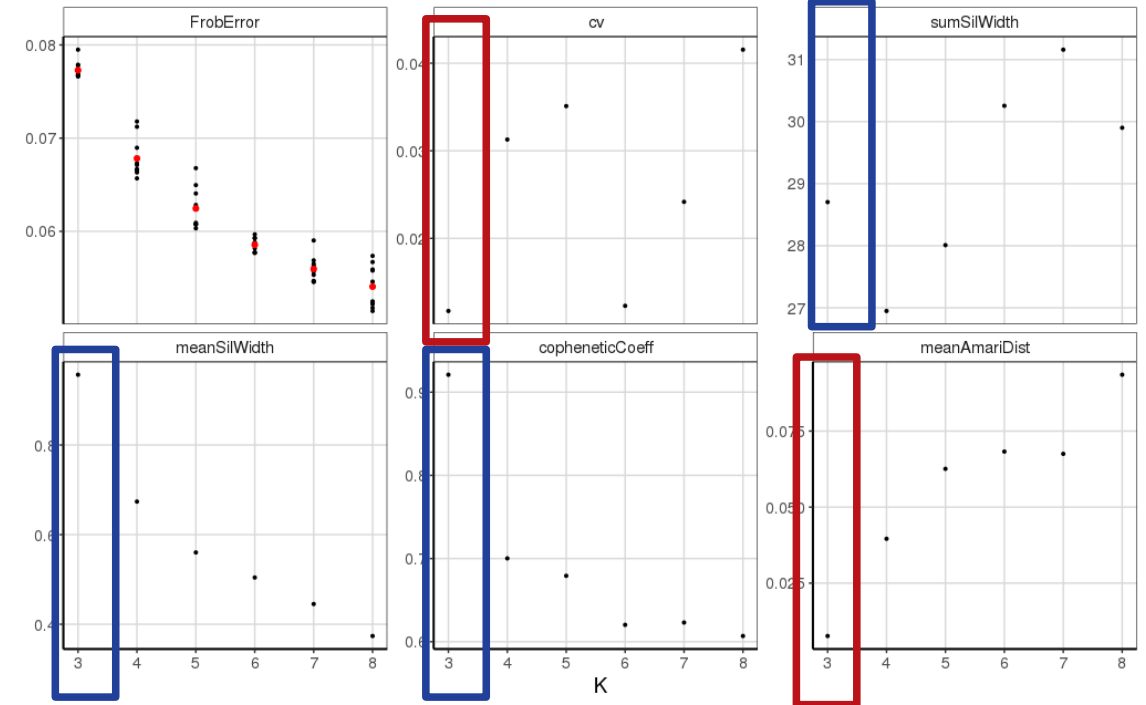
Team #1  
DA.FRA.MA

# Choice of K

## A. Linseed Algorithm (complete Deconvolution)



## B. Bradwurst library measures of control



Minimize :

- the Frobenius error,
- the coefficient of variation & Mean Amari distance,

Maximize :

- the Sum and Mean silhouette Width & the cophenetic coefficient.

# Process and Deconvolution

---

## 1.- Pre-processing

- Anti-logarithmic functions
- ( $D' = 2^D$ , or  $D' = \exp(D)$ )
- vs
- keeping data in log-scale &
- TMM normalization of linear data( $D'$ )

## 2.- Feature selection

- Variance of expression values  
⇒ Threshold : 85 % highest expression

## 3.- Deconvolution method

- 1) NMF (method = Lee | Brunet)
- 2) Bratwurst (Tensorflow implementation of NMF)

# Interpretation

---

- Given Pancreatic Cancer dataset we can suppose that  $K=3$  indicates :
  - Immune cells
  - Tumor cells
  - Fibroblasts
- PROS
  - Applied 2 methods for Unsupervised Deconvolution
  - No confounding factors added to data → No need for normalisation
  - Interesting platform of Codalab, to evaluate our results
  - Creative time for brainstorming and fruitful collaboration :-)
- CONS
  - Not much time for biological interpretation of data.
  - Restriction of tools to use in Unsupervised method-More familiar with (semi-)supervised
  - One of our methods couldn't be fully implemented (Tensorflow dependencies)



## Team 2: Methylome data

---

Nicolas Alcala<sup>1</sup>, Ghislain Durif<sup>2</sup>, Tiago Maié<sup>3</sup>

November 26, 2019

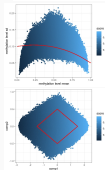
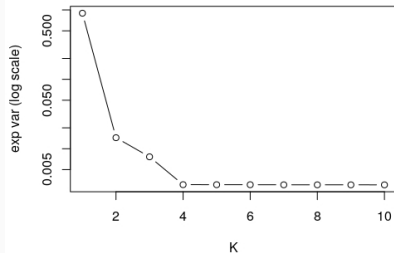
<sup>1</sup>IARC Lyon

<sup>2</sup>CNRS Montpellier

<sup>3</sup>RWTH University Hospital Aachen

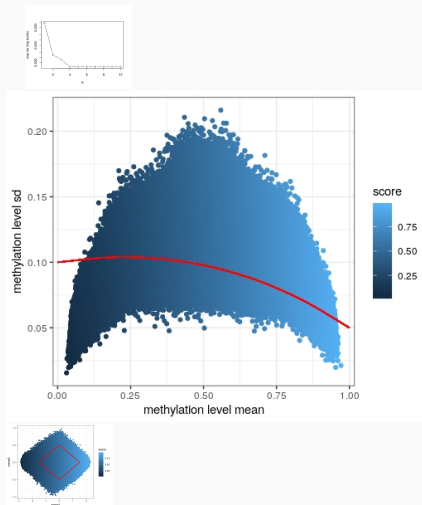
# Unsupervised approaches

1. K-choice: **PCA explained variance**
2. Prefiltering
  - Variance-based
  - PCA-based or NMF-based
  - probes selection (sex, CpG Island)
3. Learning of the matrix **A** with NMF-based approaches (RefFreeCellMix, NMF)



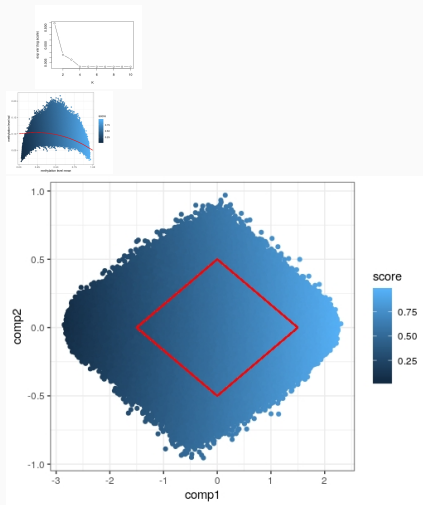
# Unsupervised approaches

1. K-choice: PCA explained variance
2. Prefiltering
  - Variance-based
  - PCA-based or NMF-based
  - probes selection (sex, CpG Island)
3. Learning of the matrix **A** with NMF-based approaches (RefFreeCellMix, NMF)



# Unsupervised approaches

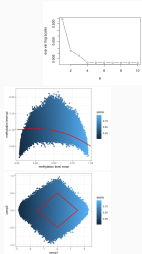
1. K-choice: PCA explained variance
2. Prefiltering
  - Variance-based
  - PCA-based or NMF-based
  - probes selection (sex, CpG Island)
3. Learning of the matrix **A** with NMF-based approaches (RefFreeCellMix, NMF)





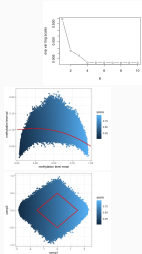
# Unsupervised approaches

1. K-choice: PCA explained variance
2. Prefiltering
  - Variance-based
  - PCA-based or NMF-based
  - probes selection (sex, CpG Island)
3. Learning of the matrix **A** with NMF-based approaches (RefFreeCellMix, NMF)



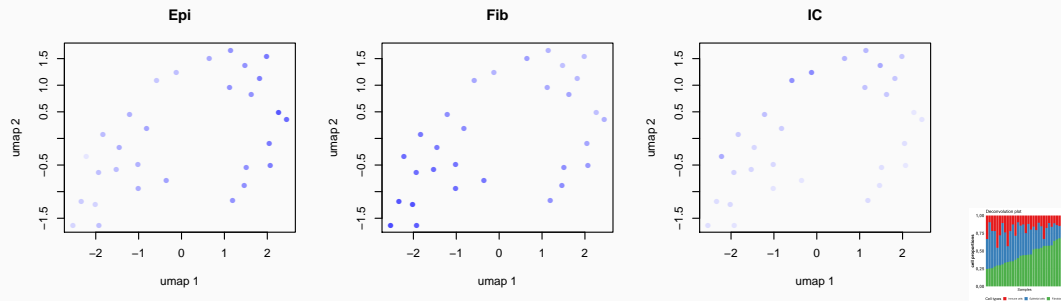
# Unsupervised approaches

1. K-choice: PCA explained variance
2. Prefiltering
  - Variance-based
  - PCA-based or NMF-based
  - probes selection (sex, CpG Island)
3. Learning of the matrix **A** with NMF-based approaches (RefFreeCellMix, NMF)



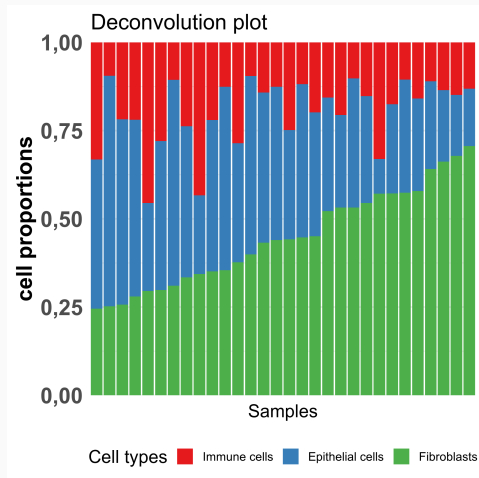
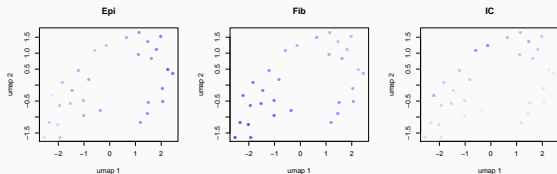
# Supervised approach

EpiDISH (<https://github.com/sjczheng/EpiDISH>)



# Supervised approach

EpiDISH (<https://github.com/sjczheng/EpiDISH>)



# Advantages/Drawbacks

EpiDISH (<https://github.com/sjczheng/EpiDISH>)

## Advantages

- Easy to use (a single function `EpiDISH::epidish`)
- Pre-selection of the probes is already done
- Supervised approach with known cell types

## Drawbacks

- Pre-selection of the probes is already done
- Supervised approach with known cell types (we got lucky it was the good ones)

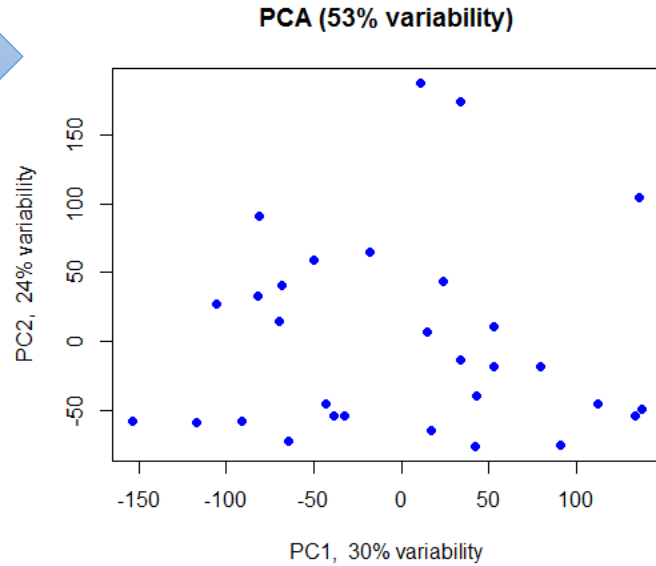
# Team 3

**Paulina Jedynak, Milan Jakobi, Petr Nazarov**

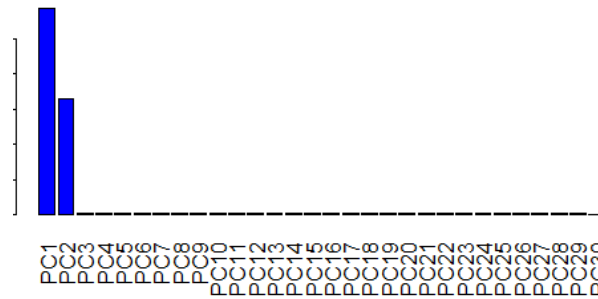
RNA-seq

# Exploration & Processing

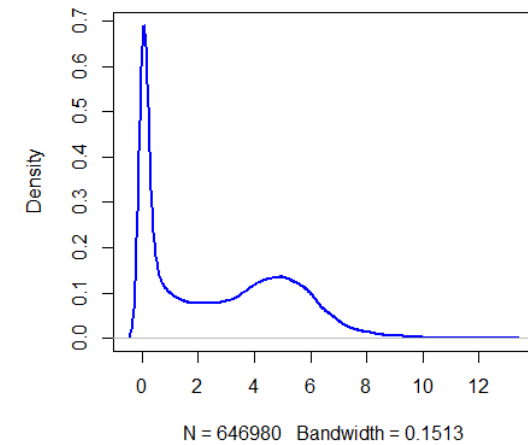
PCA



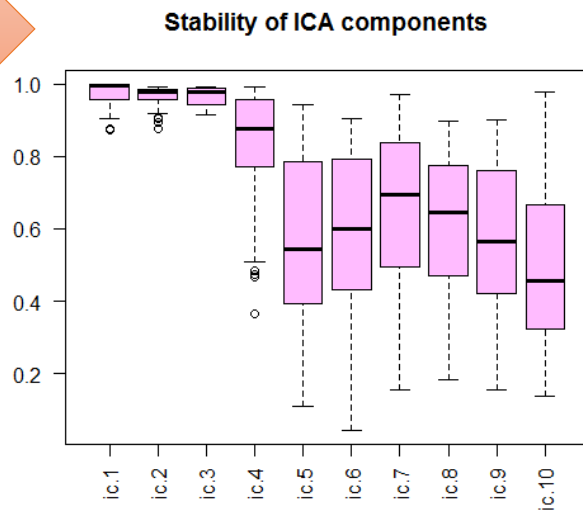
Screeplot



P.D.F.: log-transformed counts



ICA<sub>10</sub>



ICA<sub>3</sub>

Select features:  
5791 genes

IC1: *no*  
IC2: blood vessel dev <-> cornification  
IC3: immune response <-> cell division  
Other: *no*

Log scale

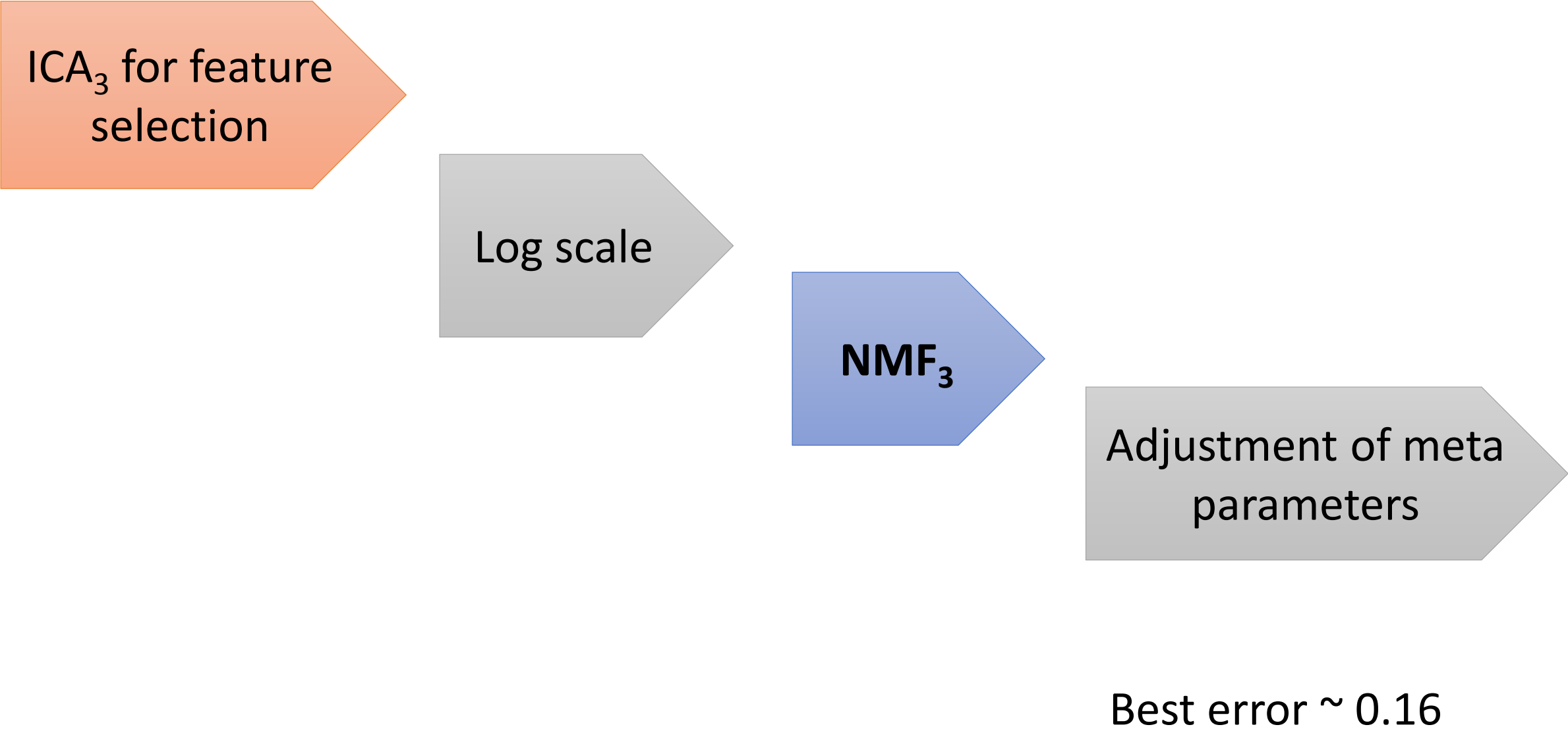
ok

Linear scale

X

# Deconvolution

ICA<sub>3</sub> for feature  
selection



```
graph LR; A[ICA3 for feature selection] --> B[Log scale]; B --> C[NMF3]; C --> D[Adjustment of meta parameters]; D --- E[Best error ~ 0.16]
```

Log scale

**NMF<sub>3</sub>**

Adjustment of meta  
parameters

Best error ~ 0.16



# Interpretation

1. The data were quite simple – 2 PCs only
2. ICA successfully worked as feature selection tool. But only two components were annotated by biological functions
3. We get better results with log-transformed data
4. Basic NMF works not bad, though it showed some stochasticity

⇒ *Multiple runs are recommended*

⇒ *ICA, perhaps, can be used as an initial estimation for NMF*



# Results for challenge #1

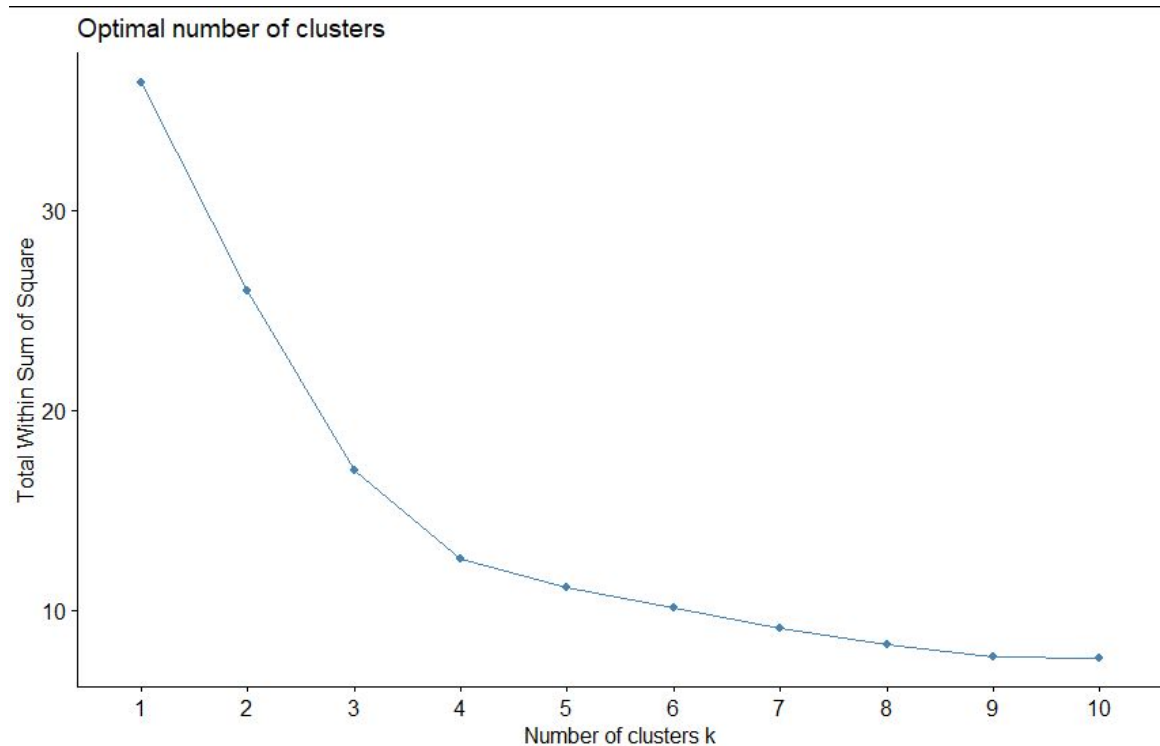
HADAC 2019

**Team 4** - Rémy Jardillier - Lara Dirian - Jules Marécaille



# Preprocessing

- We filtered the initial dataset using a subset of pancreatic cancer hyper and hypo methylated CpGs we got from the literature
- We used **k-means** and analysed the elbow curve to determine the number of LMCs (4)





# Deconvolution

- We used the **EDec** algorithm for deconvolution



# Conclusion

- We may have restrained the number of features to much, maybe we should have look up subsets coming from different studies.
- We found 4 methylation patterns even though it might not reflect perfectly on the number of cell types

# team5

## Transcriptome deconvolution

Florent Chuffart  
Jane Merlevede  
Nicolas Sompairac

# Variable selection

- Method 1:

```
sds = apply(D, sd) ; D = D[sds > 0.2, ]
```

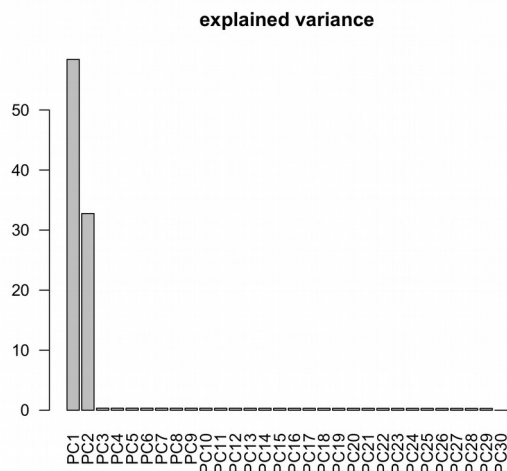
- Method 2 :

none

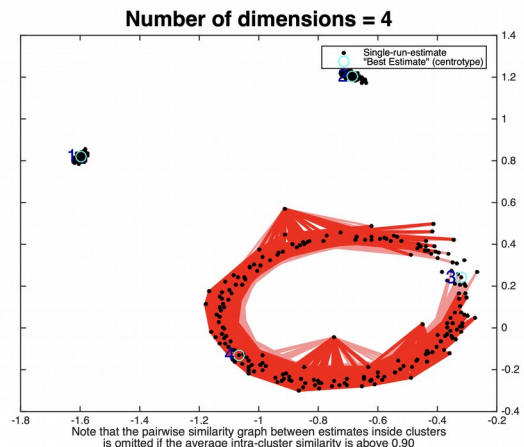
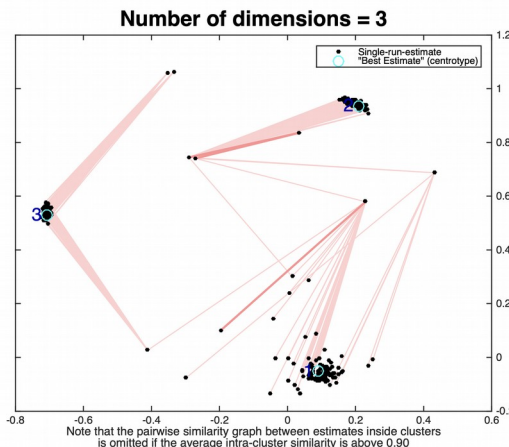
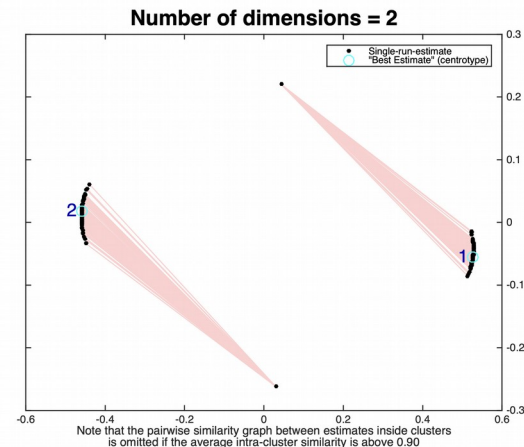
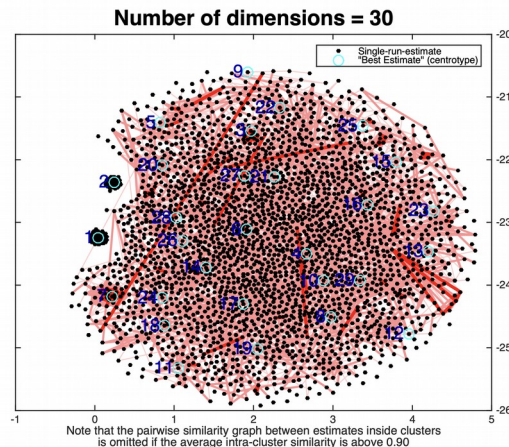


# Deconvolution methods

- Method 1: NMF with default parameters and  $k=3$  according to PCA/ICA



- Method 2: ICA



# Pros and cons

- Pros
  - Fast and simple (sd based + NMF)
  - ICA related to biological interpretation
- Cons
  - Local minimum with  $sd > 0.2$  (over fitting)
  - NMF depend on random initialization (nrun did not work)

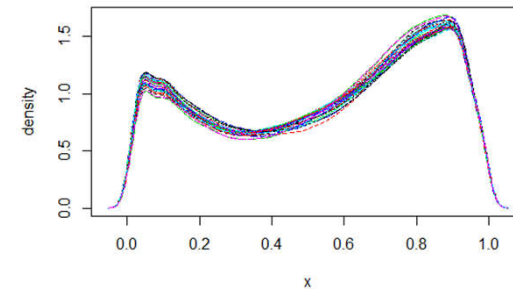
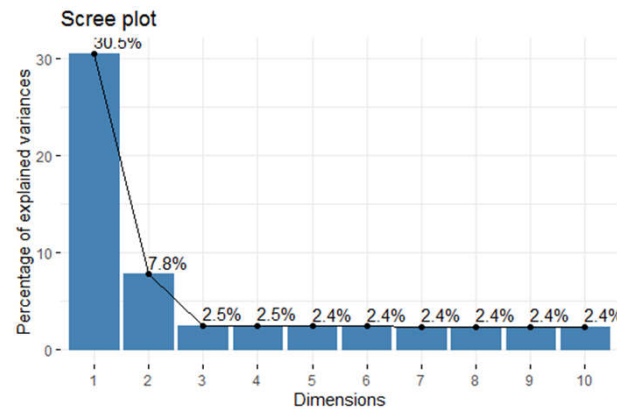
	none	Sd > 0.2
NMF	0.18	0.15
ICA	0.11	0.13

# PRE-PROCESSING

## 1) Choice of k

**K=3**

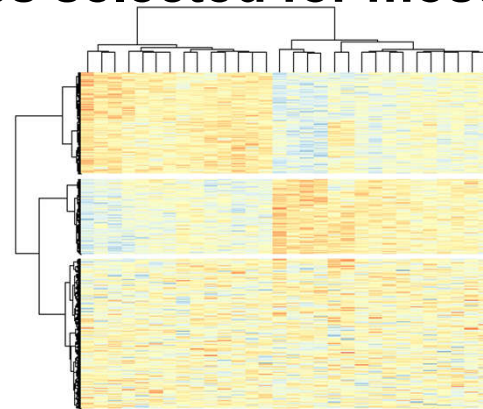
`medepir::plot_k`



## 2) Feature selection

**5000 or 10000 most variable features selected for most tools**

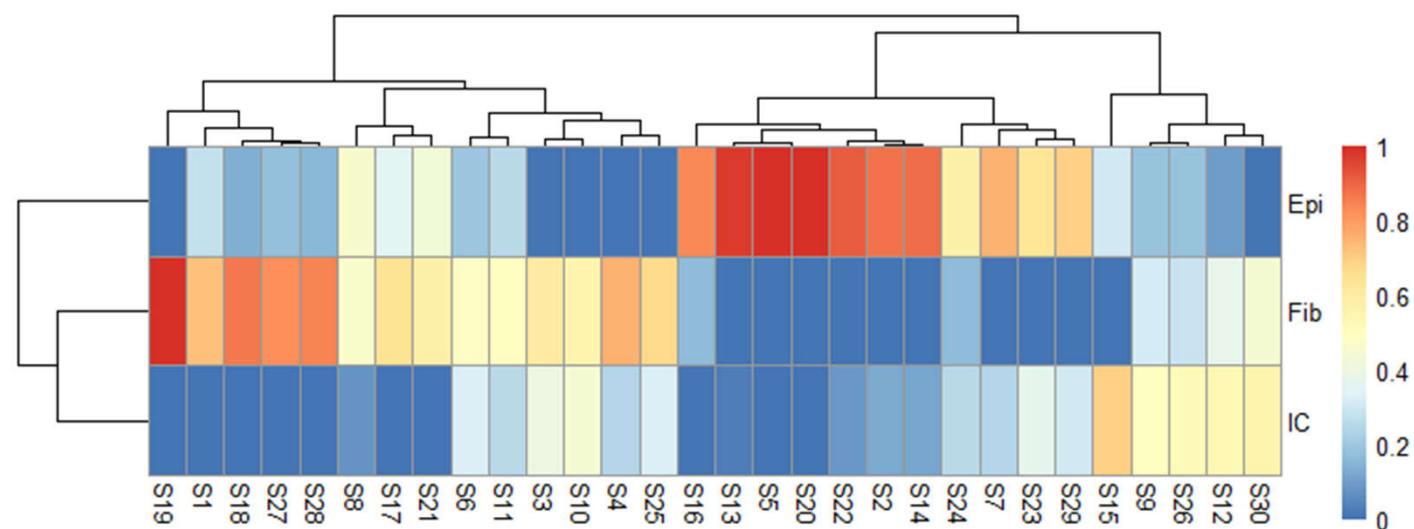
`medepir::feature_selection`



# TOOLS

- 1) **NMF**
- 2) **RefFreeEWAS** / medepir::RFE(D\_FS, nbcell = k)
  - 1) Initialize euclidean distance and manhattan
- 3) **EDec** / medepir::Edec(D\_FS, nbcell = k, infloci = infloci)
  - 1) RefFactor score to select features (500)
  - 2) Reference examples data from EDec
  - 3) CpG matrix from EpiDISH
- 4) **EpiDISH**
  - 1) Selection of features (variables + epidish ref)
  - 2) All features
  - 3) Methods "RPC", "CBS", "CP":  
Robust Partial Correlations-RPC(Teschendorff et al. 2017),  
Cibersort-CBS(Newman et al. 2015),  
Constrained Projection-CP(Houseman et al. 2012))

# RESULTS



# Team “007”

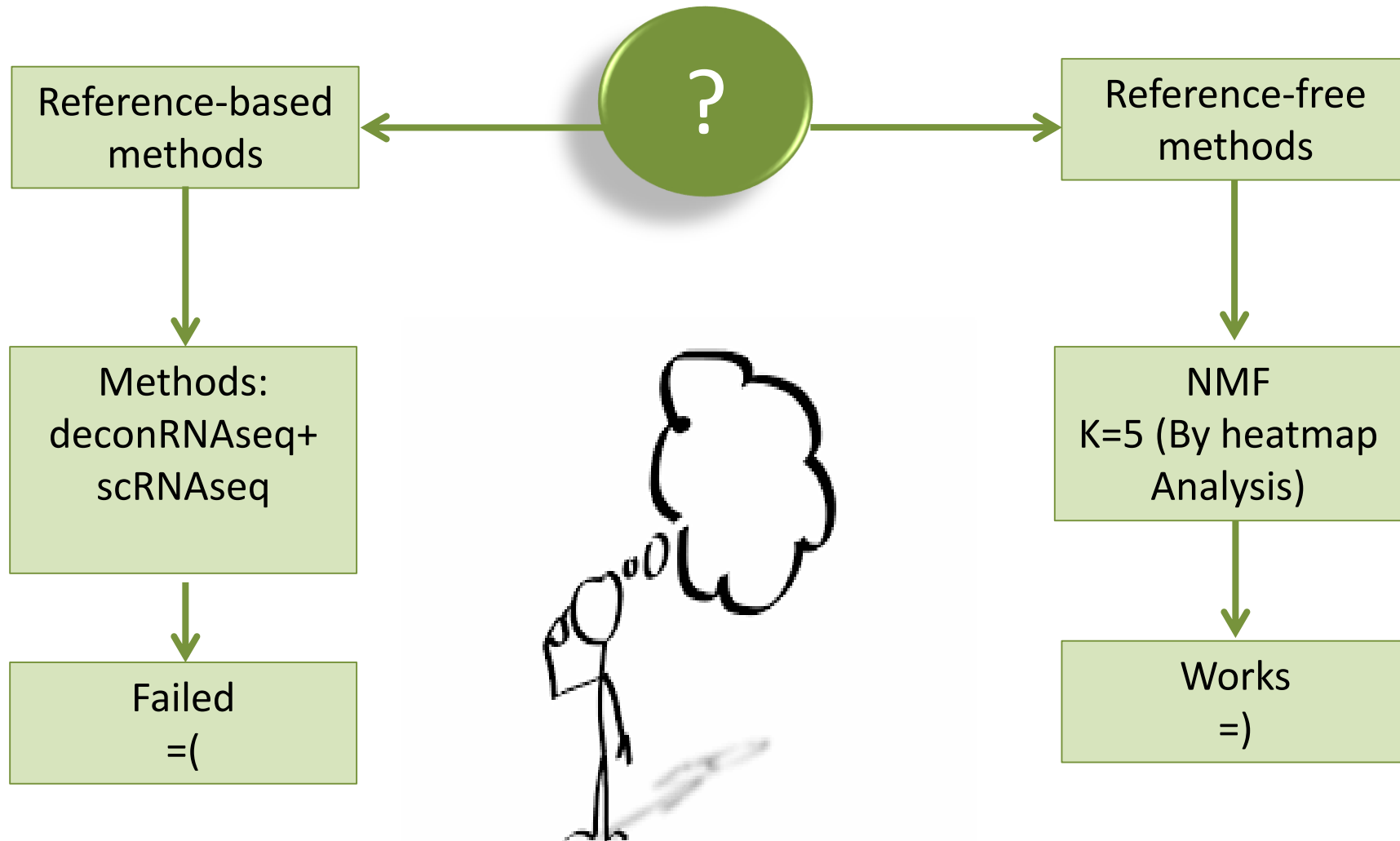
Michael Scherer

Aleksandra Kakoichenkova

Kapil Newar



# Approach



## Deconvolution of the RNA-Seq data by using NMF

### **1). Instalation of NMF package**

```
if ( !{ "NMF" %in% installed.packages( ) } ) {  
  install.packages(pkgs = "NMF", repos = "https://cloud.r-project.org")  
}
```

### **2). Input data**

```
dat <- input$rna  
sort.var <- apply(dat,1,sd,na.rm=T)  
sel.dat <- dat[order(sort.var,decreasing = T)[1:5000],]
```

### **3). NMF analysis**

```
nmf.mod <- nmf(sel.dat,rank = 5)  
A.estimate <- nmf.mod@fit@H  
col.sums <- 1/apply(A.estimate,2,sum)  
for(i in 1:ncol(A.estimate)){  
  A.estimate[,i] <- A.estimate[,i]*col.sums[i]  
}  
return(A.estimate)
```



- Choosing the right reference profiles is crucial and hard
- NMF for RNAseq technically works, but results are not really interpretable
- Determining the number of cell types itself is not trivial from RNAseq data
- Further things to be considered:
  - Feature selection
  - Rescaling of the  $A$  estimate



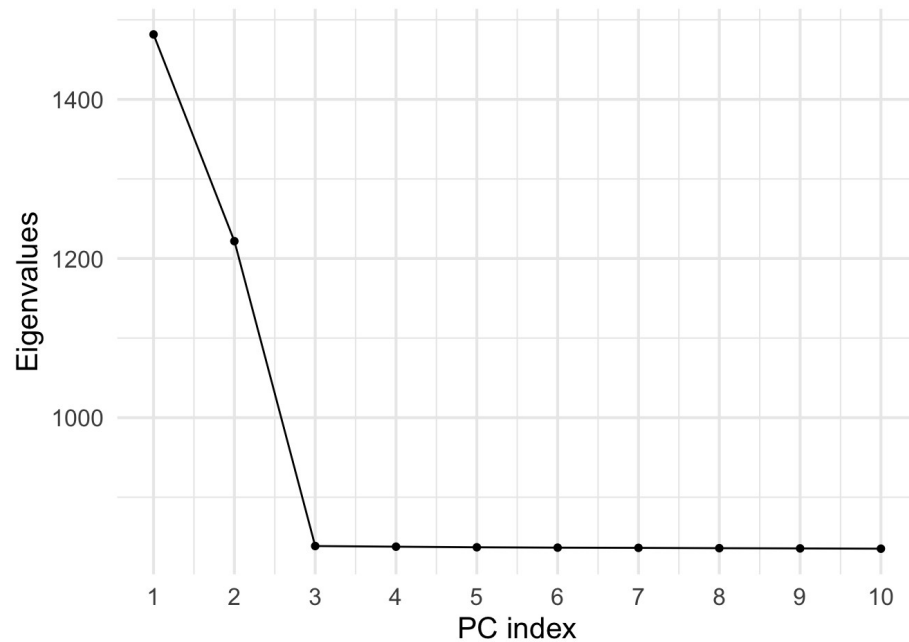
# Challenge 1

Team 8



# Choice of K

$$K = \text{nrPC} + 1$$



# Deconvolution Method

- **RefFreeEWAS**

Permits reference-free deconvolution. RefFreeEWAS offers a method for evaluating the extent to which the underlying reflects specific types of cells.

Solution to a convolution equation of the form  $D = A * T$

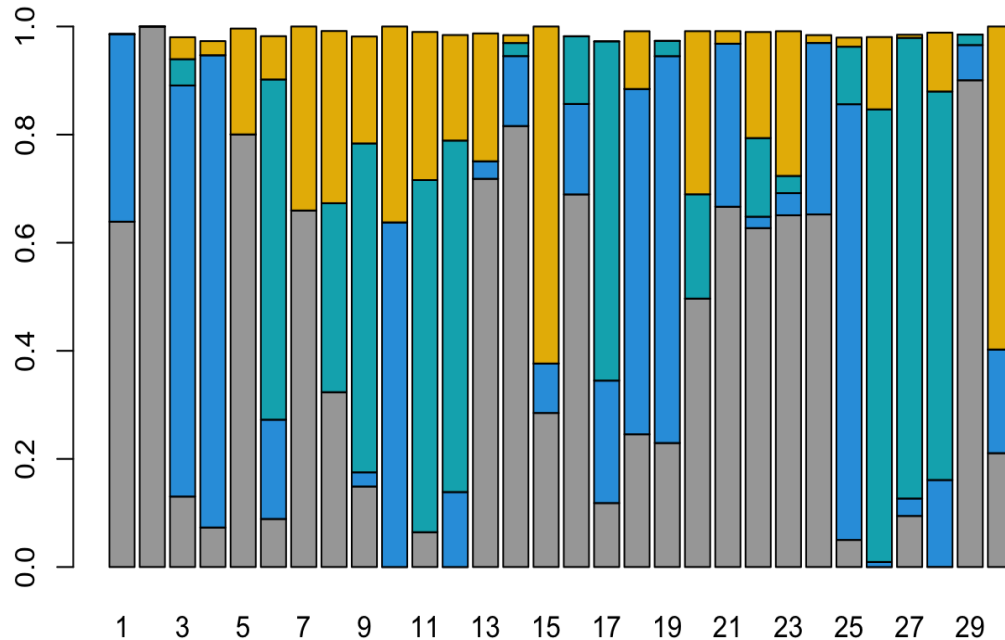
Feature selection of the 5000 most variable genes in D

- Regression based methods
- Probabilistic methods
- Enrichment methods
- **Matrix factorization methods**

# Interpretation

Reference-free based approach

Pros and cons



# Pre-treatment / Choice of K

Input: normalized/log-transformed RNA-seq data

## Data transformation

- Log-transformed data vs. Linear data

## Feature selection

- Variance-based feature selection (10 to 40%) vs. none

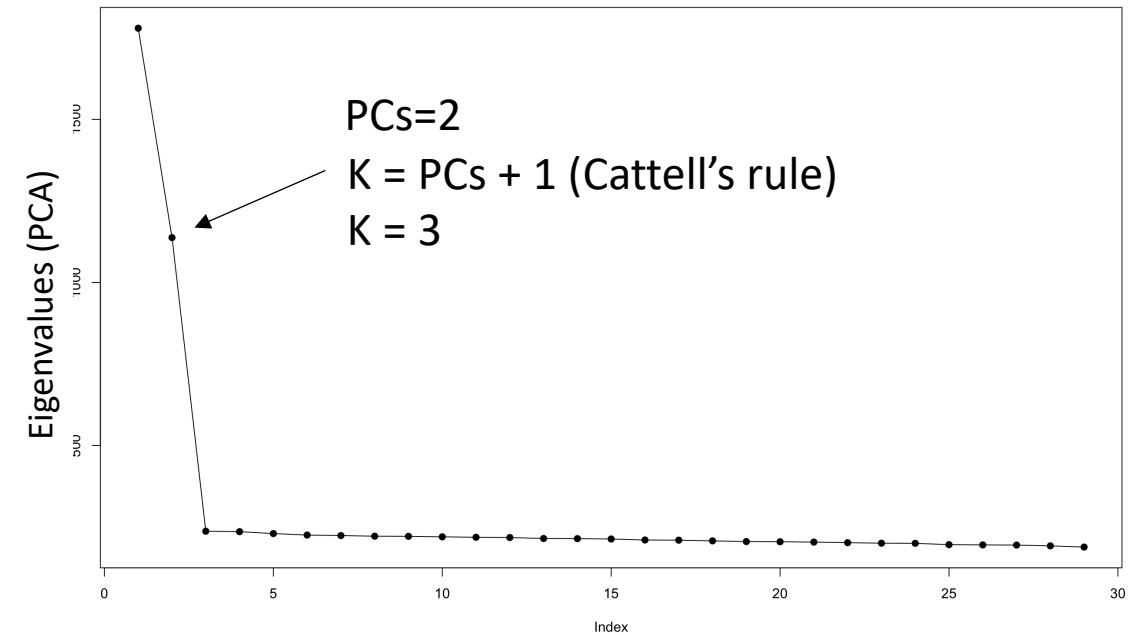


Figure: Scree plot

# Deconvolution method

## Unsupervised approaches

### NMF-based approaches

- Basic NMF
- Consensus NMF:
  - > compute a consensus A matrix averaging different NMF clusterings

## Supervised approaches

### Pre-requirement

- Fibroblast estimation

### Method: MCP-counter

- Marker-based approach
- Produces an abundance score for 8 immune cell populations and 2 stromal cell pops.
- Alternative strategies: focus on the 3/4 most abundant cell pop, include an additional 'consensus' component

### Estimation of A:

- Derive proportions from abundance scores by dividing  $\sum s_c$  for each patient

# Interpretation: Pros & Cons

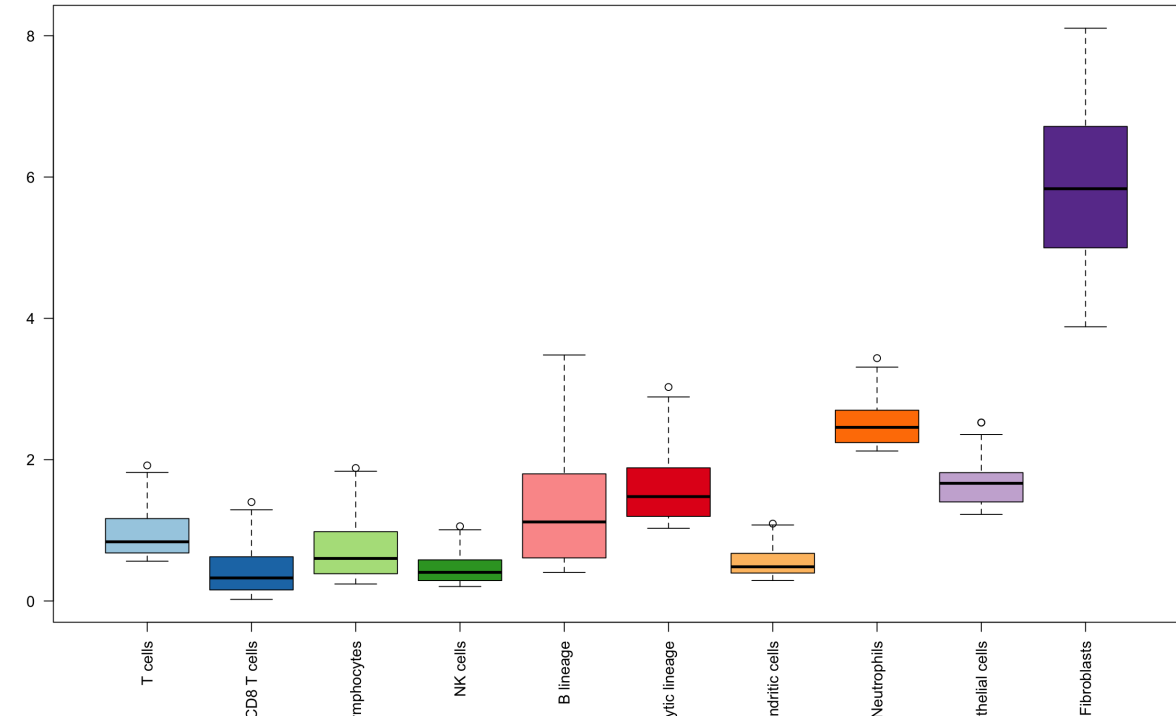
MCPcounter: promising !

- Pros: easy to run & interpret, fast
  - Cons:
    - gives abundance scores and not proportions
- > The approach to estimate proportions could be refined (?)
- could allow some cell pop to be discarded (semi-sup)

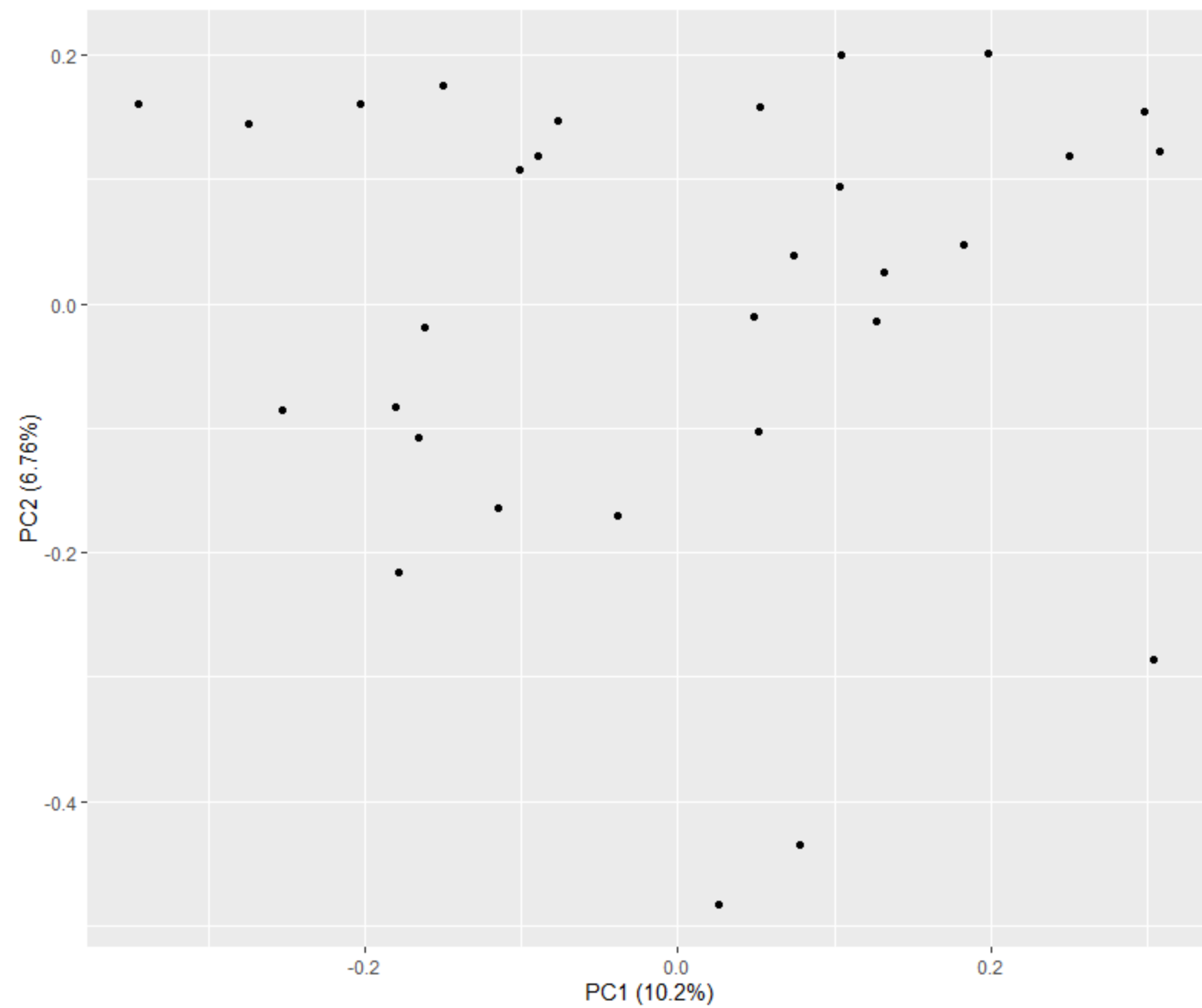
Best result (MAE\_D1=0.1/MAE\_D2=0.08):

NMF with no feature selection // 3 components // log-transformed data

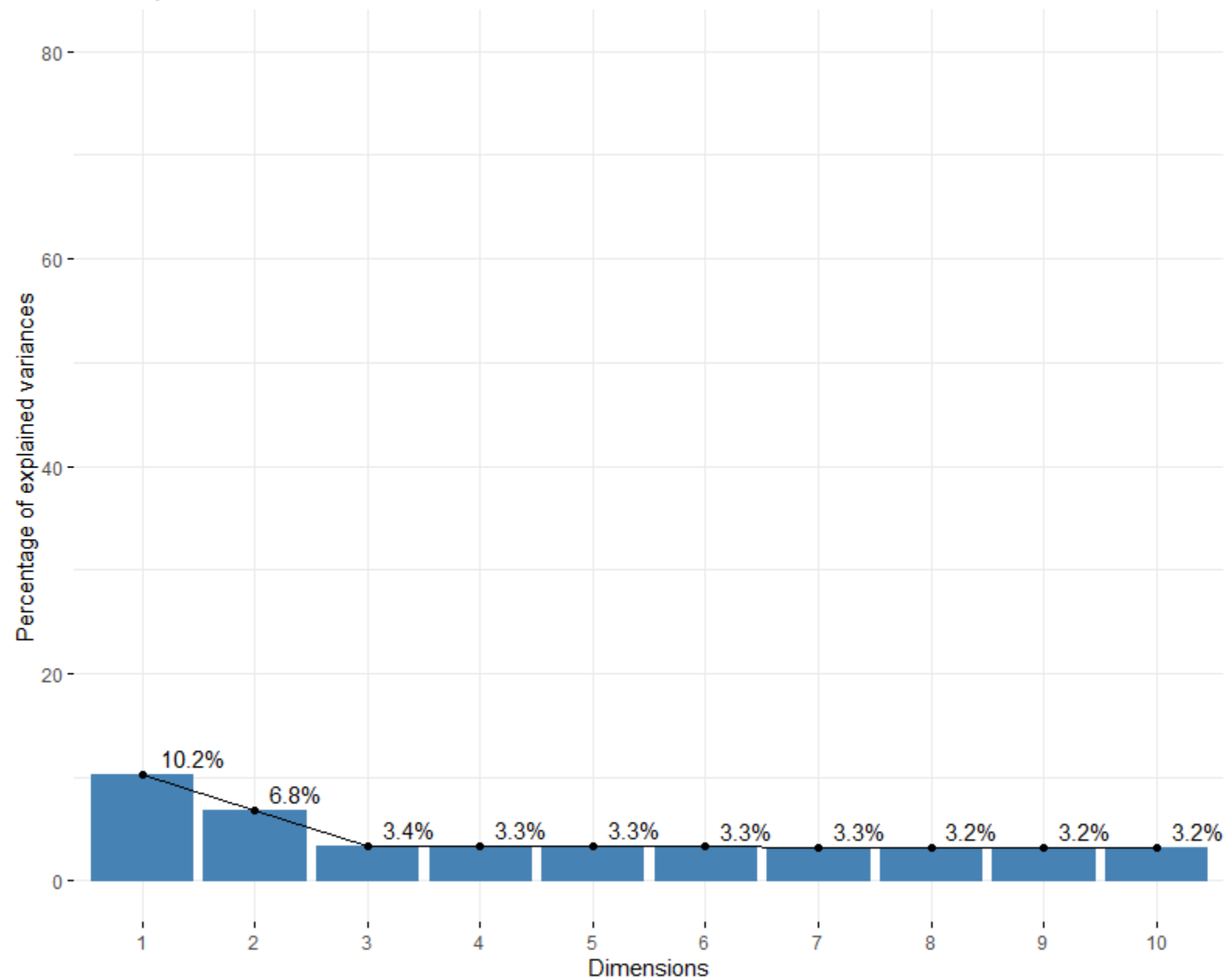
- Pros: easy to run, fast
- Cons:
  - interpretation of the components needs further analyses
  - can be trapped in suboptimal local minima



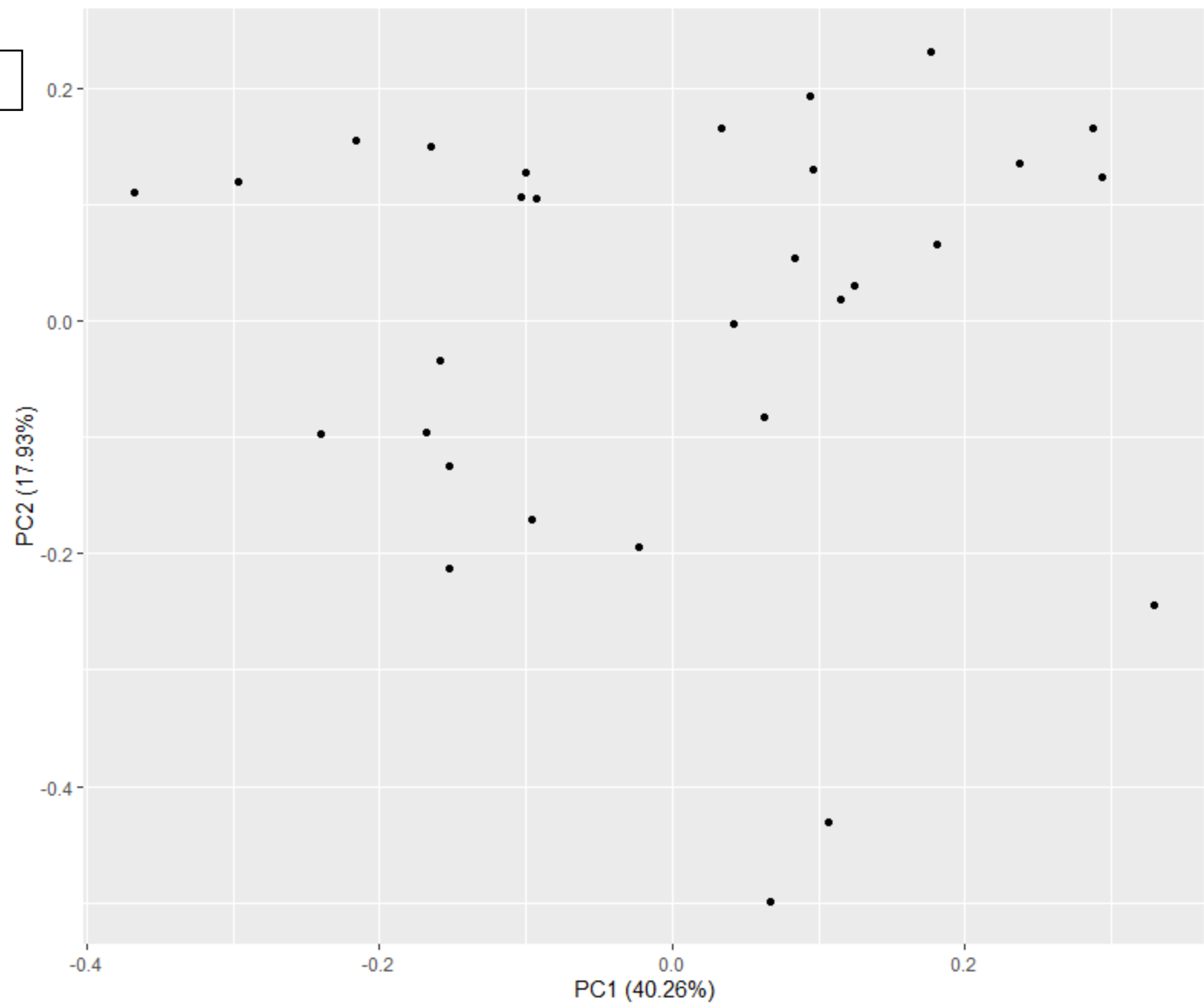




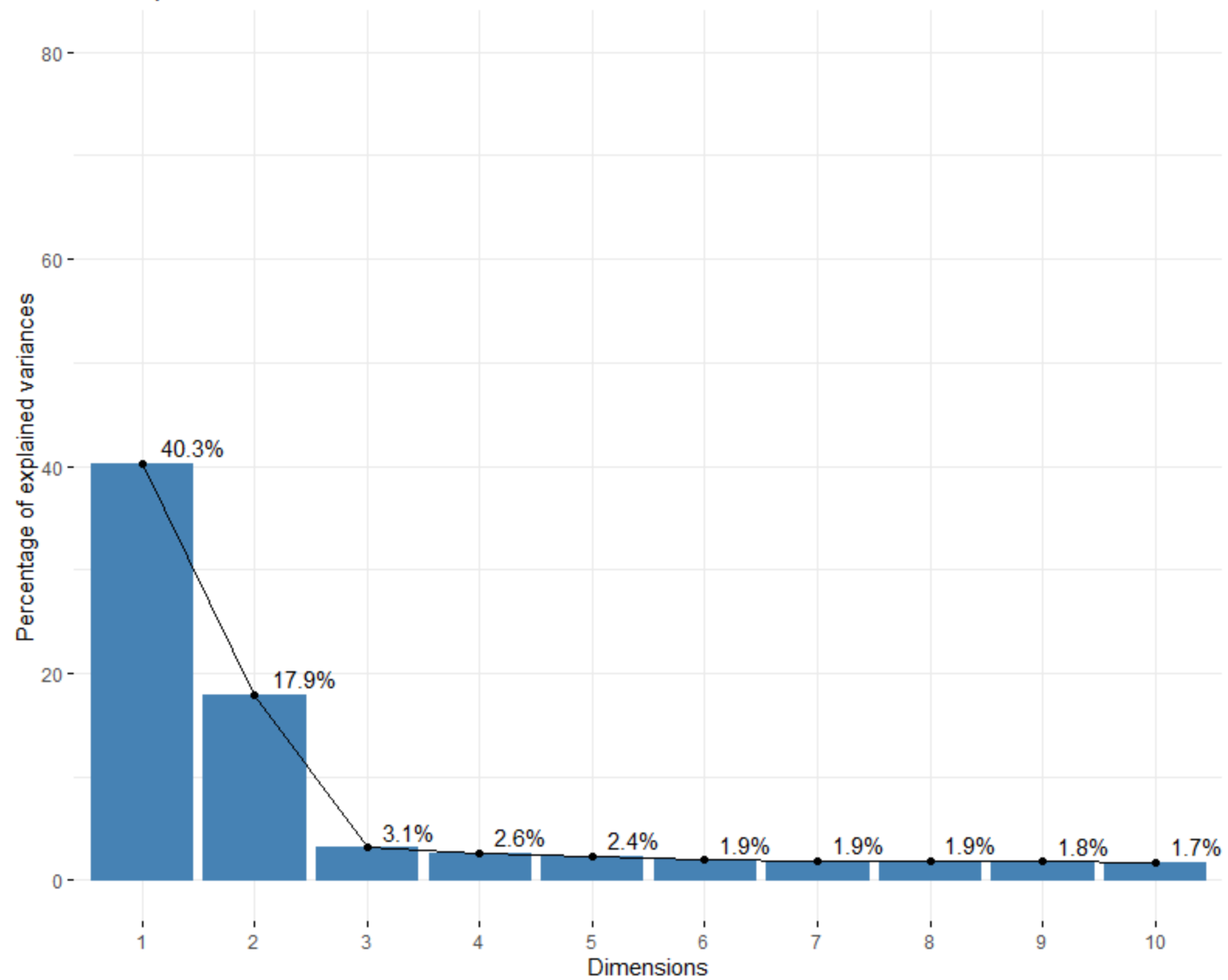
Scree plot



500 Higher Rank



Scree plot



```
RefFreeCellMix(factors,mu0=NULL,K=3,itors=9,Yfinal=NULL,verbose=TRUE)
```

Default

1	0.2967613432
---	--------------

MAE ▲	MAE 1 ▲	MAE 2 ▲
0.2864 (8)	0.0952 (6)	0.1912 (9)

