Team A

Francisco Avila Cobos,

Silvia Yahel Bahena Hernandes,

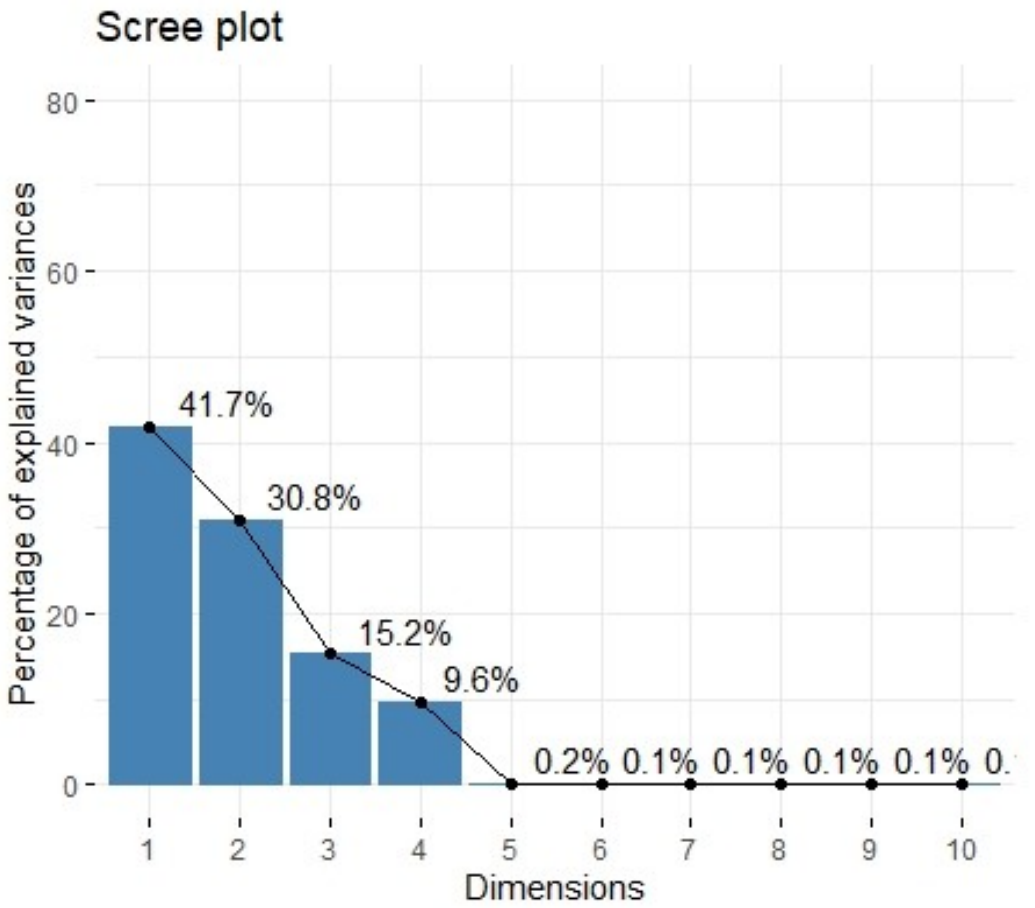
Petr Nazarov,

Florent Chuffart

Determining K with PCA scree plot

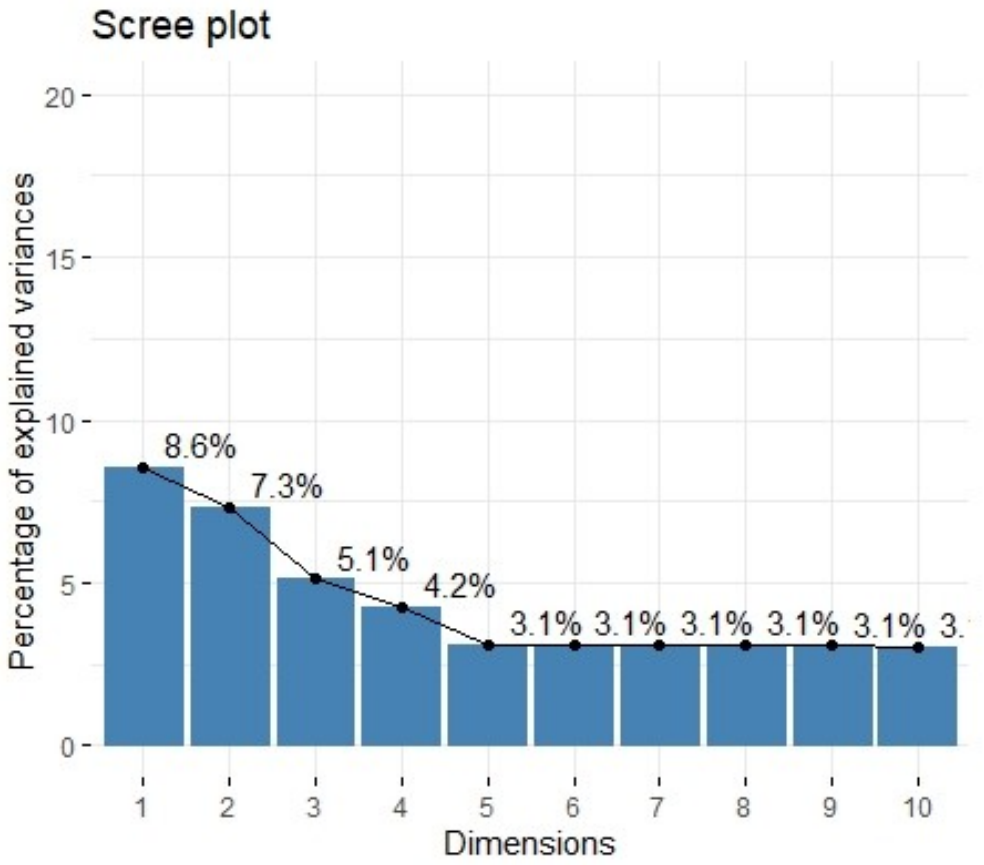
Transcriptome

K = 5

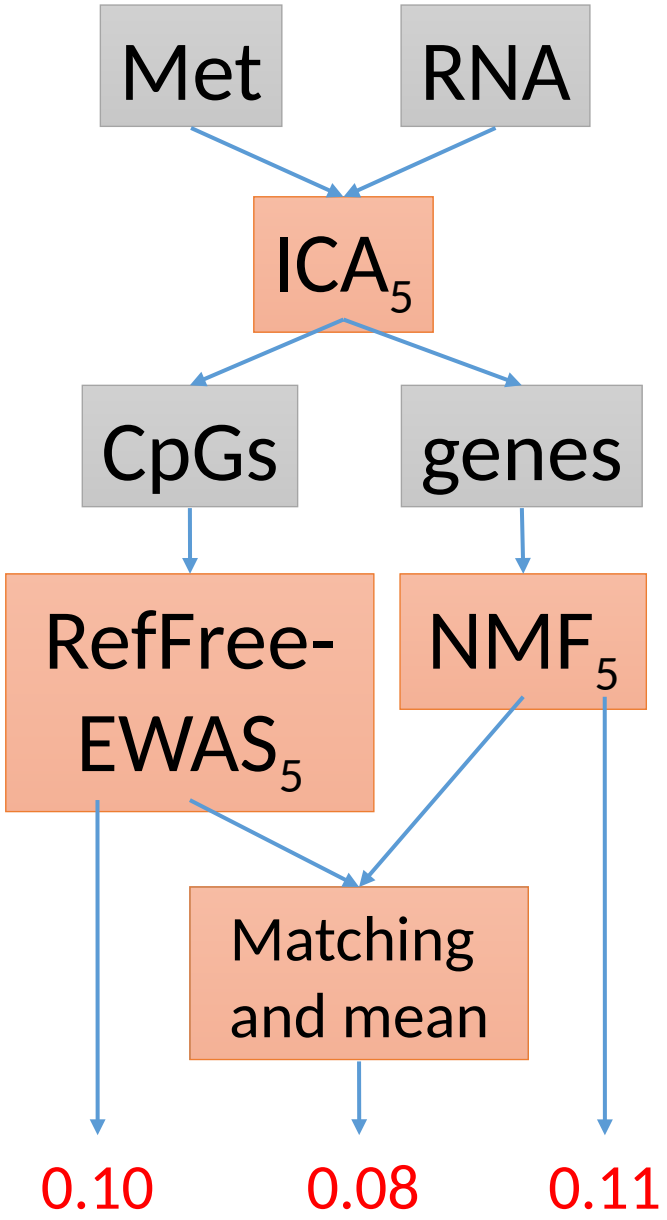


Methylome

K = 5



Cattell's rule : K = PCs + 1



ICA results

Component # 1 (stability = 0.980)

GO:BP neg : 74 terms(FDR<0.01)

Term	FDR
cell division	2.07e−16
microtubule cytoskeleton organization in...	5.98e−12
cytokine−mediated signaling pathway	1.49e−09
immune system process	3.23e−09
flavonoid glucuronidation	5.75e−09
kinetochore organization	2.13e−08
mitotic nuclear division	1.00e−07

GO:BP pos : 45 terms(FDR<0.01)

Term	FDR
peptide hormone secretion	5.11e−08
cardiac muscle cell action potential inv...	2.23e−05
pancreas development	3.40e−05
extracellular structure organization	2.11e−04
ERK1 and ERK2 cascade	7.34e−04
regulation of heart rate by cardiac cond...	1.54e−03
second−messenger−mediated signaling	2.51e−03

Component # 2 (stability = 0.873)

GO:BP neg : 106 terms(FDR<0.01)

Term	FDR
extracellular matrix organization	1.60e−27
blood vessel development	1.92e−23
skeletal system development	3.40e−18
cell adhesion	1.76e−16
regulation of cell migration	3.19e−14
animal organ morphogenesis	2.21e−13

GO:BP pos : 18 terms(FDR<0.01)

Term	FDR
cornification	1.52e−09
homophilic cell adhesion via plasma memb...	2.39e−09
flavonoid glucuronidation	3.75e−07
O−glycan processing	3.75e−07
xenobiotic glucuronidation	1.15e−05
regulation of microvillus organization	1.06e−04

Component # 3 (stability = 0.909)

GO:BP neg : 12 terms(FDR<0.01)

Term	FDR
cell adhesion	8.46e−10
cornification	1.84e−09
extracellular matrix organization	7.45e−08
cardiovascular system development	8.78e−07
regulation of cell migration	5.11e−03
SRP−dependent cotranslational protein ta...	6.65e−03

GO:BP pos : 13 terms(FDR<0.01)

Term	FDR
xenobiotic metabolic process	3.19e−12
flavonoid glucuronidation	2.00e−07
O−glycan processing	2.23e−05
flavone metabolic process	3.47e−04
digestion	3.83e−04
regulation of microvillus organization	9.58e−04

Component # 4 (stability = 0.956)

GO:BP neg : 49 terms(FDR<0.01)

Term	FDR
extracellular matrix organization	1.60e−27
cell adhesion	1.52e−08
cartilage development	3.35e−07
tube development	6.70e−05
cell motility	6.70e−05

GO:BP pos : 151 terms(FDR<0.01)

Term	FDR
immune response	2.66e−28
defense response	2.66e−28
immune response−activating cell surface ...	2.66e−28
cell surface receptor signaling pathway	2.66e−28
positive regulation of leukocyte cell−ce	2.66e−28

Transcriptome

Linseed + ICA + PCA ✉ K = 5

- Linseed + 5000 most variable genes
- Markers from Linseed
  - NMF with those features
  - ssKL with CT – marker relationship
  - supervised ✉ cell-type enrichment:
    - activated stellate
    - immune (NK / eosinophil)
    - ductal
    - endothelial
- sd > 0.05 , 0.1 … Q3 + NMF(5, brunet/lee)
  - CV, IQR…

Methylome

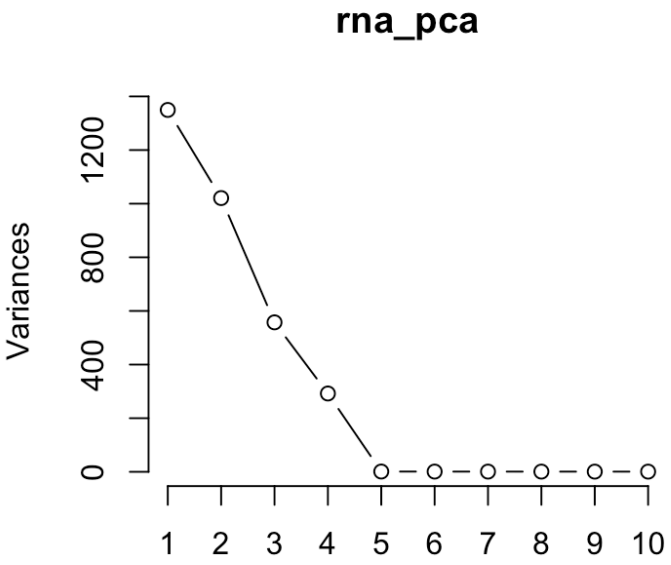
ICA + PCA ✉ K = 5

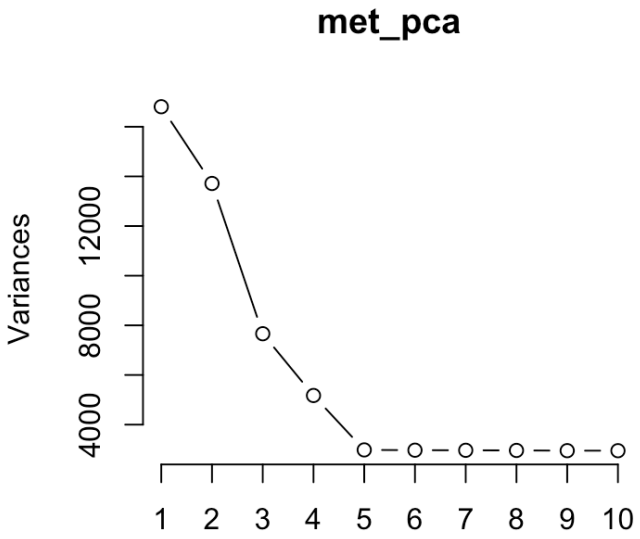
- mean >= 0.1…0.2 & mean <= 0.8…0.9 (avoid SNPs, focus on biology)
- sd <= Q2, Q3…
- removal of chrX , chrY – probes
- NMF(5, brunet/lee)
- MeDeCom(D, 5, c(0,10^(-3:1)), NINIT = 30, NFOLDS = 5, ITERMAX = 20)

Challenge #2

Team B

Estimation of k and preprocessing





- Variance filtering for both datasets separately

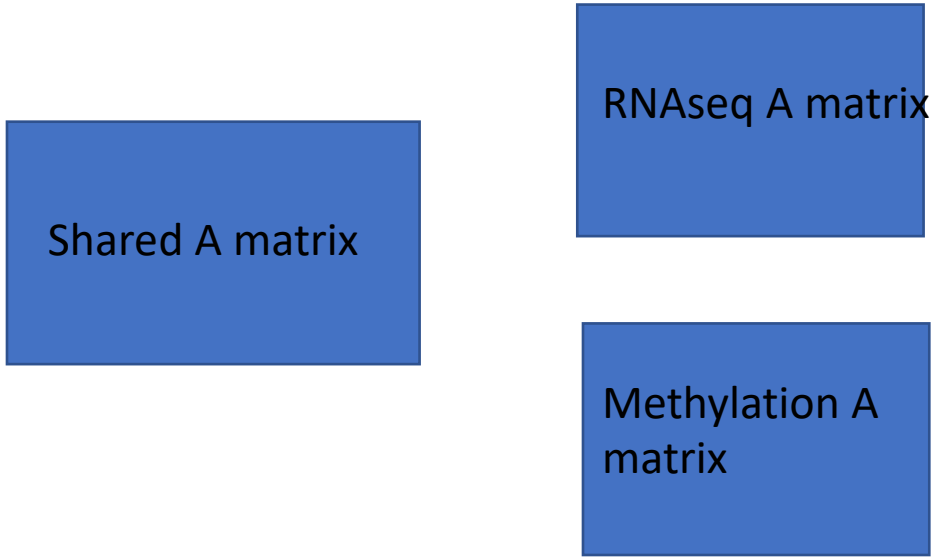
(85% and 95% quantile) for methods deconICA and integrative NMF

- No filtering for method EpiDish

Methods

Integrative NMF

Output



Best achieved MAE: 0.082
-> maybe due to bad feature selection
-> maybe method does not work well on methylation data

deconICA

- Separately for RNAseq and methylation matrix
- MAE: 0.07

MOFA

- Integrative approach
- We could not get it to work properly
- Got only two cell types

Method used: EpiDish

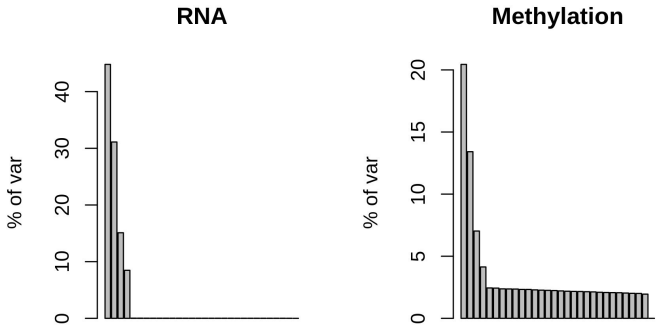
- Best result: MAE of 0.065 in first round
- Used hepiDish with five cell types:
    - Epi
    - Fat
    - Fibroblasts
    - NK cells
    - CD4T cells

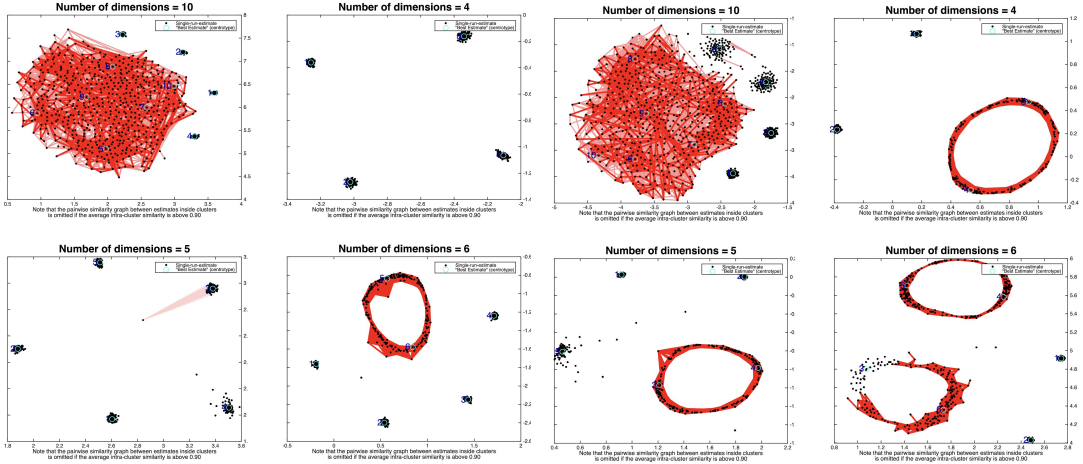
-> No time for more biological interpretation

Challenge #2

Team C - *Nicolas Sompairac, Lara Dirian, Jane Merlevede, Jules Marécaille*

Component selection





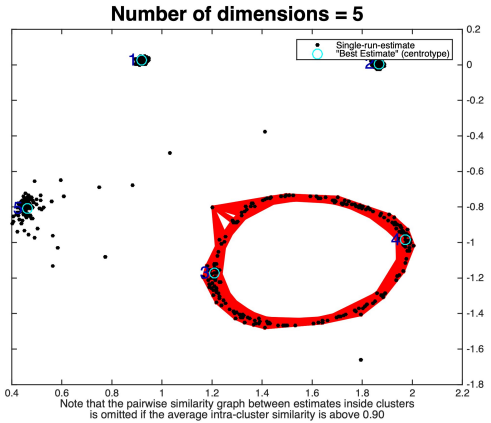
RNA decomposition

MET decomposition

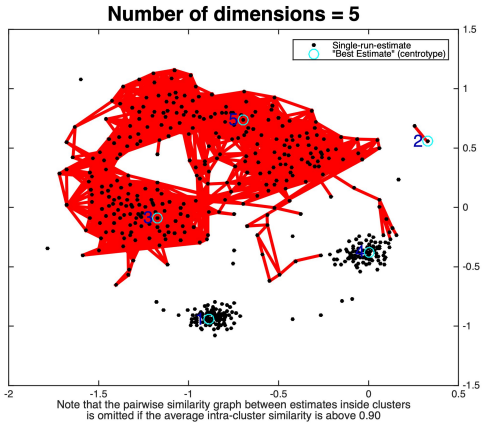
Filtering

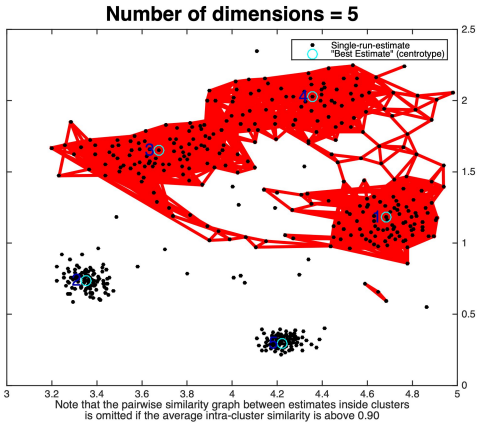
- Filtering the CpGs using the literature
- Filtering by variance (threshold : 0.95)
- Removing the information contained on the sexual chromosome and SNP
- None

Without filtering



With filtering 95%





With filtering 80%

Merging

- Merging the datasets before the deconvolution (append)
- Deconvoluting the datasets separately and merging them afterwards blindly (mistake)
- **Deconvoluting the datasets separately and merging them afterwards by permuting the components and checking the correlation between matrices.**

Deconvolution methods

- EDec
- NMF
- RefFreeEWAS
- **ICA (with deconICA + consICA)**

Scores

Prefiltering	Method name	Score (MAE)
None	Starting Kit	0.116
Variance + literature-based	NMF	0.082
Variance	NMF + Post merging	0.10
gender+SNP+variance+M values	consICA (both)	0.118
none	ICA (deconICA)	0.048
gender+SNP+variance	*consICA (rna) + RefFreeEwas*	*0.057*
None	**Submitted**	**0.0774**

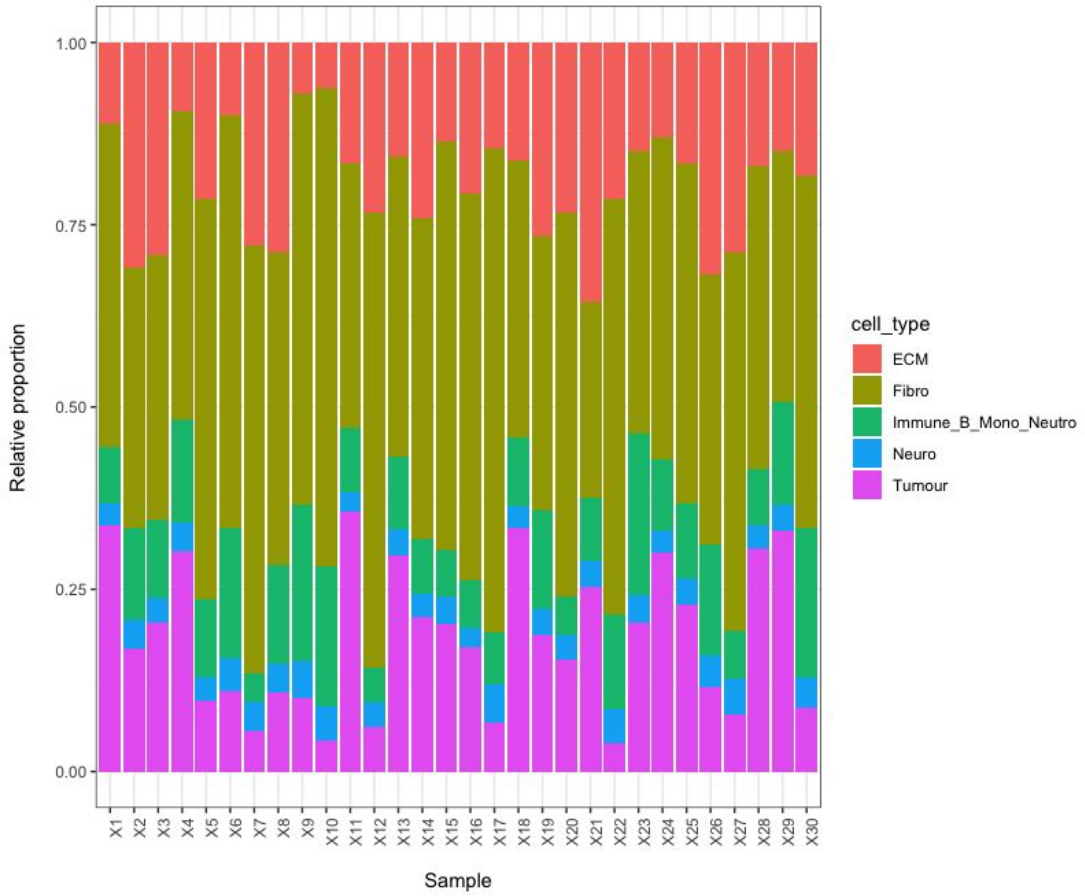
Conclusions

PROs

- ICA: Unsupervised approach
- ICA: Gives the K number

CONs

- Hard to explain the components
- Isn't really robust on Methylome data
- Could be improved with some filtering
- Additional step to merge components (unsupervised approaches)



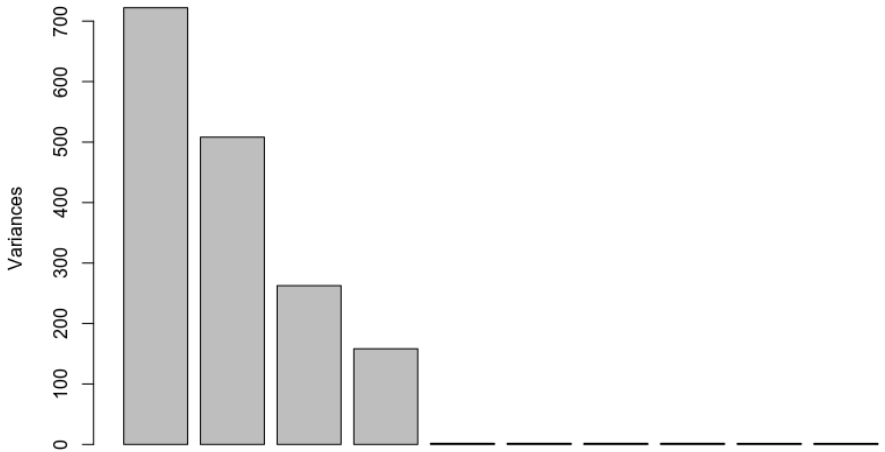
Team D Challenge 2

Maria Kondili,
Novella Rausell Claudio,

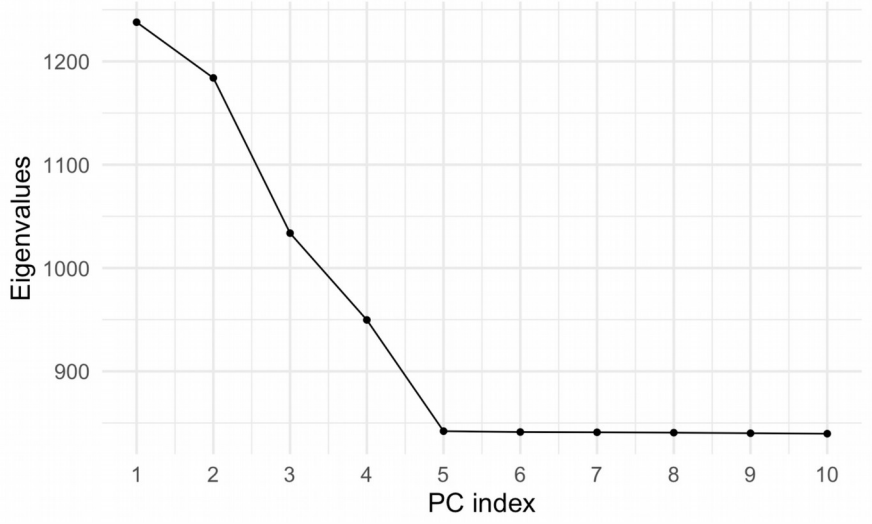
Zacharouli Markella-Achilleia

Choice of K

RNA



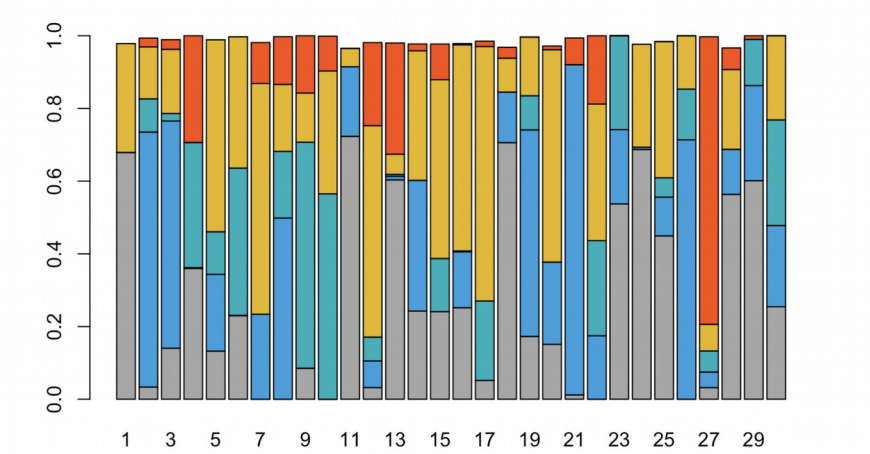
MET



Our Deconvolution Methods

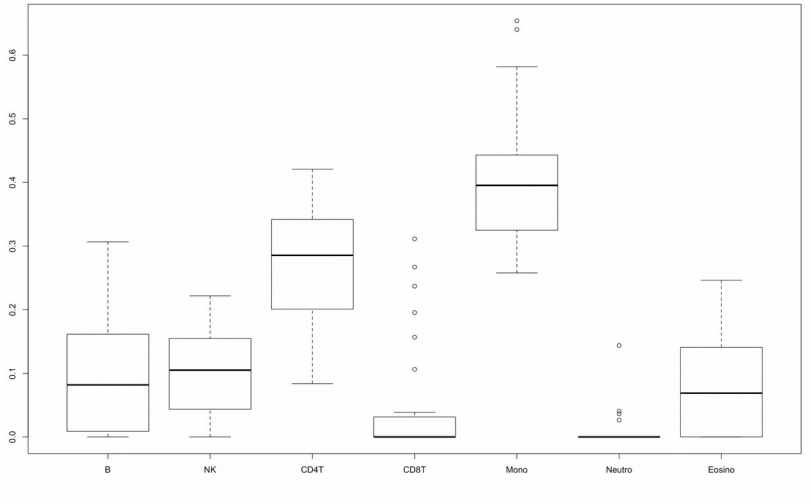
Non-supervised

- NMF
- RFE-SVD



Supervised

- EpiDISH, RPC =robust partial correlation



Integration Potentials

Ideas we would like to apply:

- Integrate initial datasets (MOFA) or,
- After independent deconvolution > correlate components (ICA)
- Associate methylation Annotation (promoter site | ProbeID) to the gene expression

Choice of K / Preliminary analyses

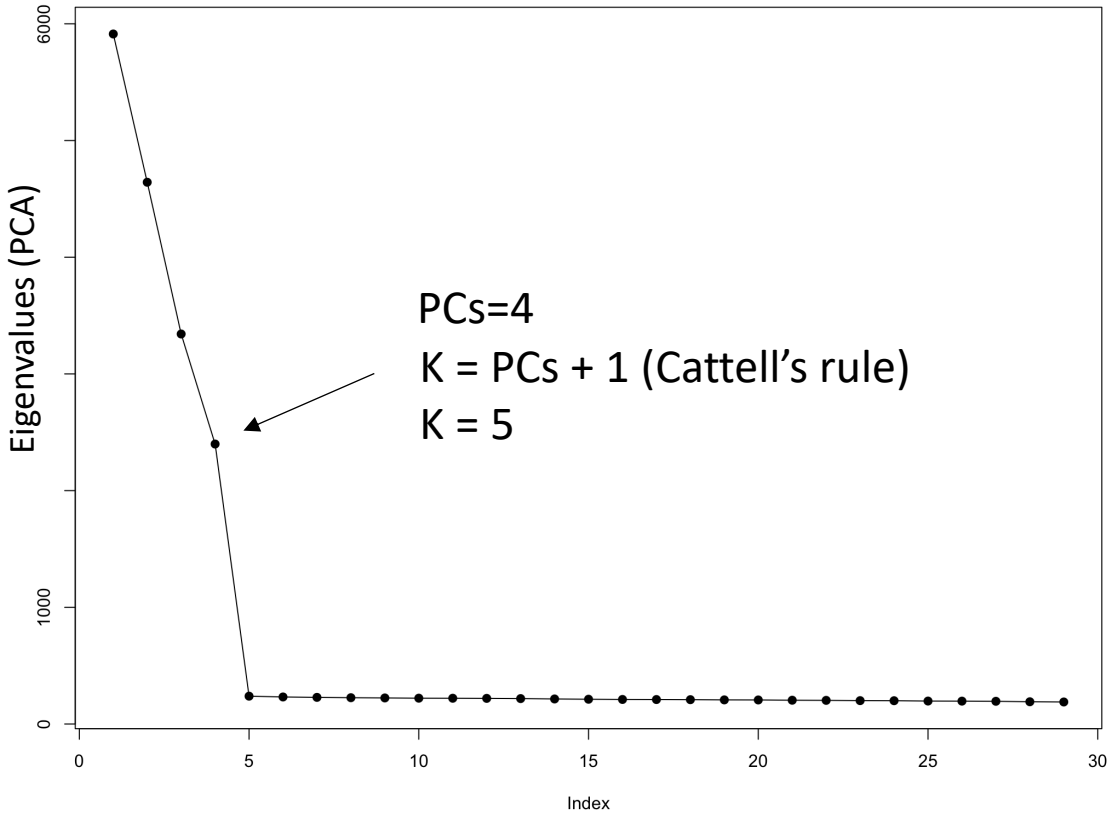
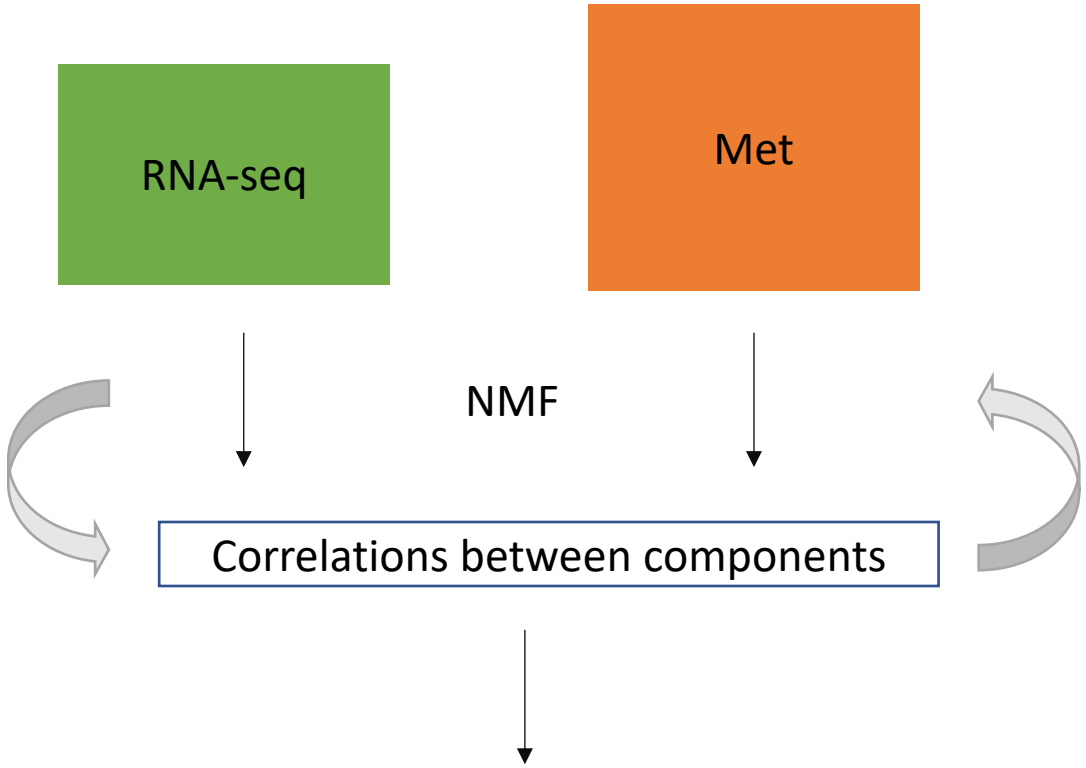


Figure: Scree plot

Optimization of the NMF (K = 5)



Two best NMF methods

- snmf/r: regularization
- nsNMF: includes an intermediate smoothing matrix

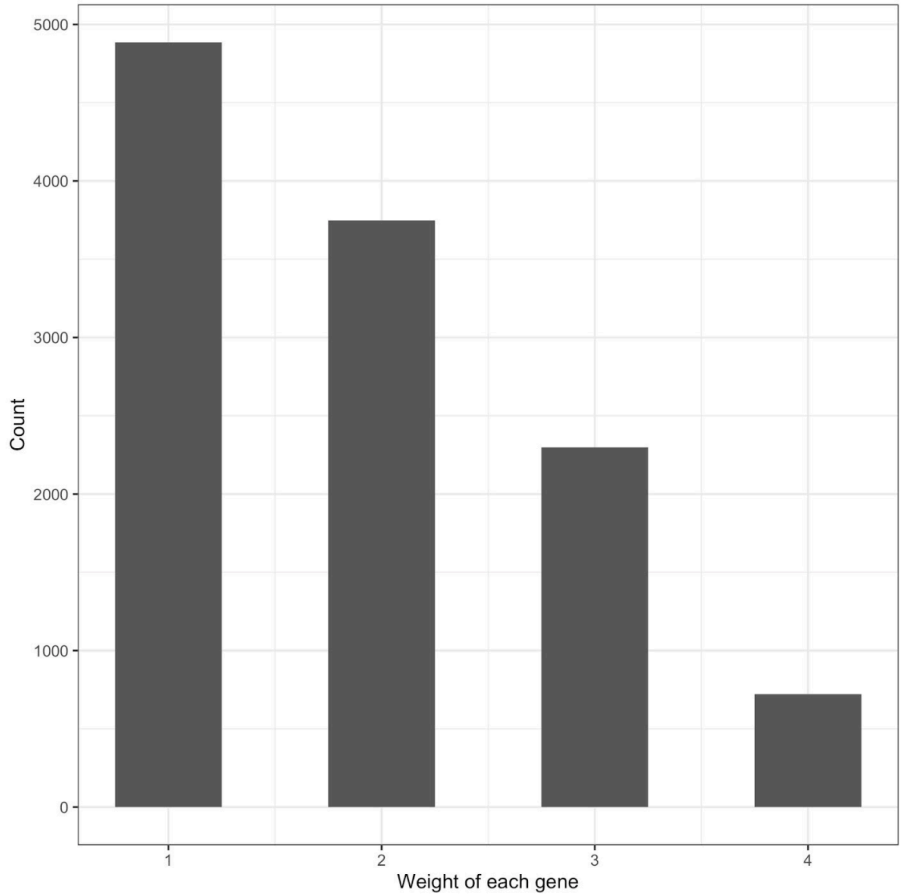
Deconvolution method – Tween: Two-step weighted NMF

Step 1 - Preselection of features

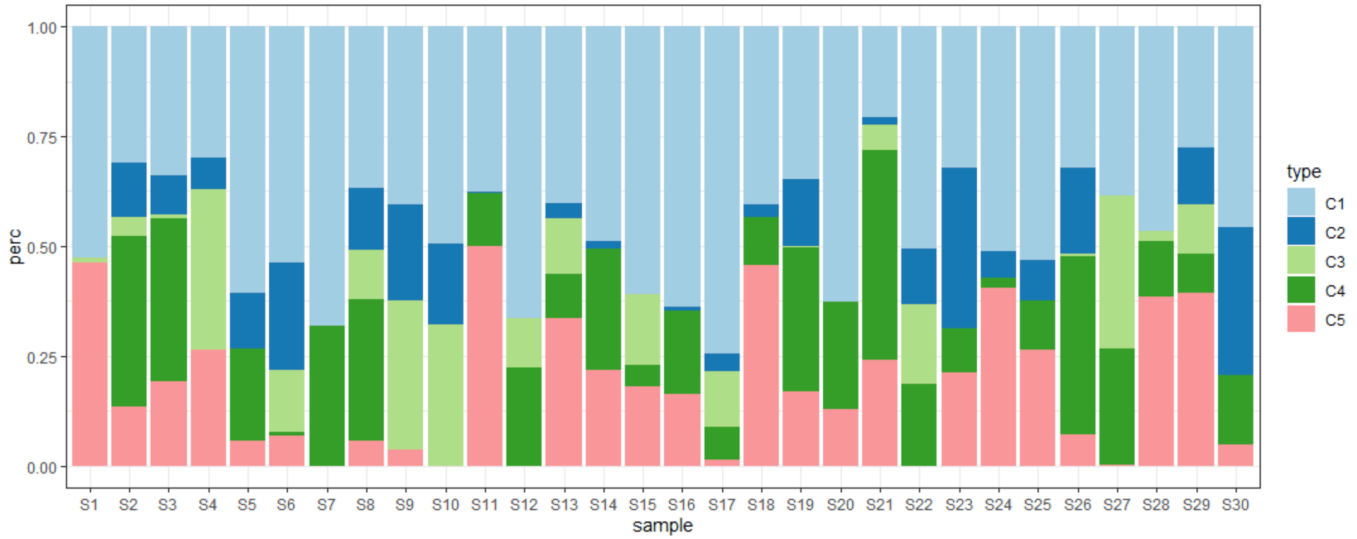
- Consensus ICA
- Select features significantly associated with ICs (loose FDR cutoff: 0.2)

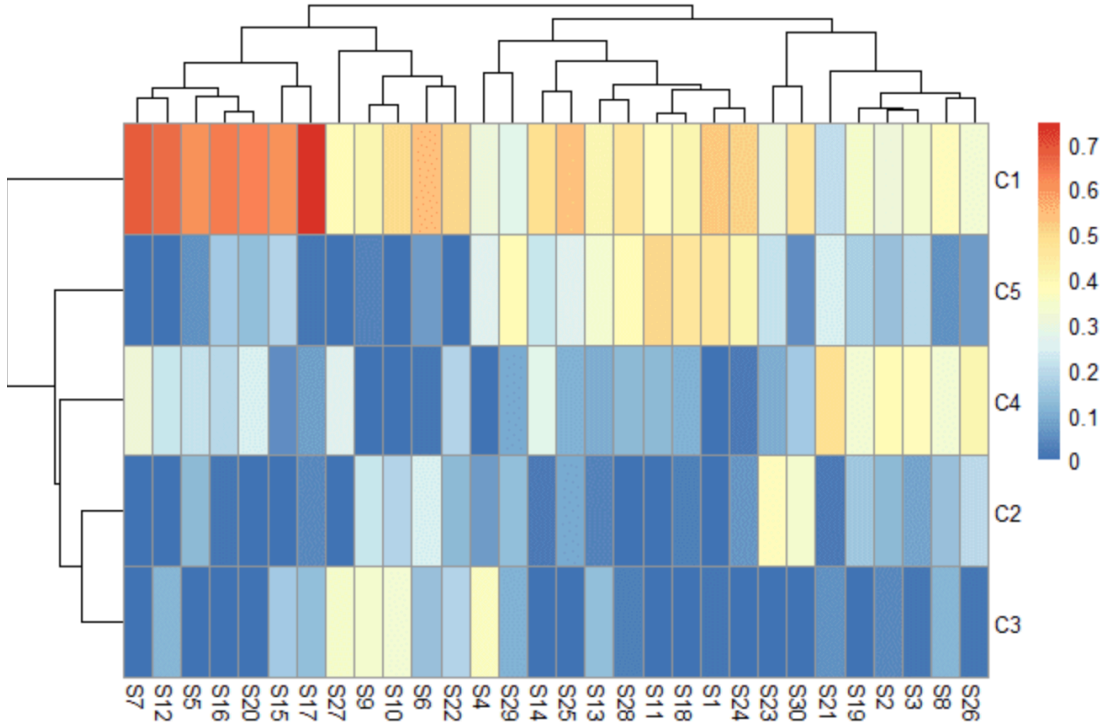
Step 2 - Regularized NMF (K = 5)

- Weighted features



Interpretation: Pros & Cons





Pros:

- Easy to implement/fast
- Good performances on the test and validation datasets

Cons:

- Unsupervised approach: needs further analyses to interpret the components

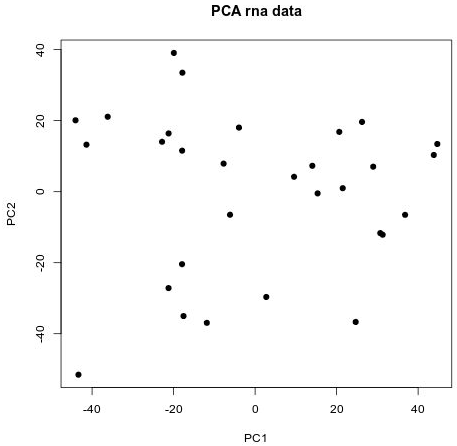
Challenge #2
#HADACA2019

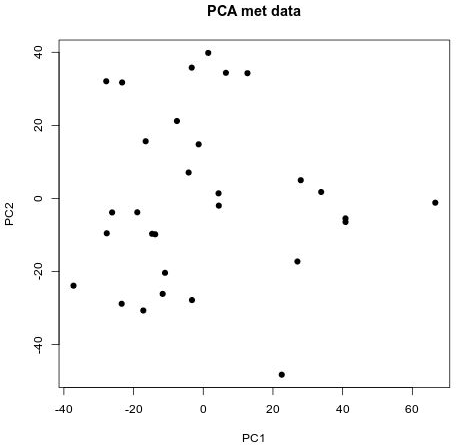
Group F

Anne-Françoise Batto
Katherine Waury
Tiago Maié

Find K!

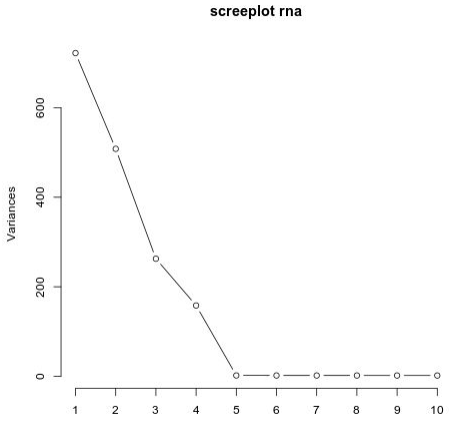
We started with **K = 5**, however preliminary results pushed us to use **K = 4** instead.

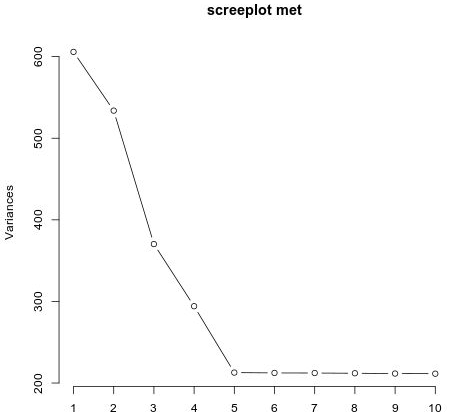




Pre-filtering

Given the results from challenge 1 we decided to only apply filtering to transcriptomic data (ICA)





Deconvolution methods

Methylation data

➔ EpiDISH

- Best method from challenge 1 (met)
- Supervised: pre-compiled list of CpGs for identification of fibroblasts, epithelial cells and immune cells
- Given the comments/results from challenge 1 we decided to stick for the most part with B cells
- Cibersort (CBS) method performed better than Robust Partial Correlations (RPC)
- Timed execution so that we could explore as many parameters (nu.v) as possible given the time frame

RNA data

➔ ICA + NMF

- Best method from challenge 1 (rna)
- Unsupervised: feature selection with ICA
- Promising results (at some point our best entry) but in general worse than EpiDISH. This led us to not explore this option as much.

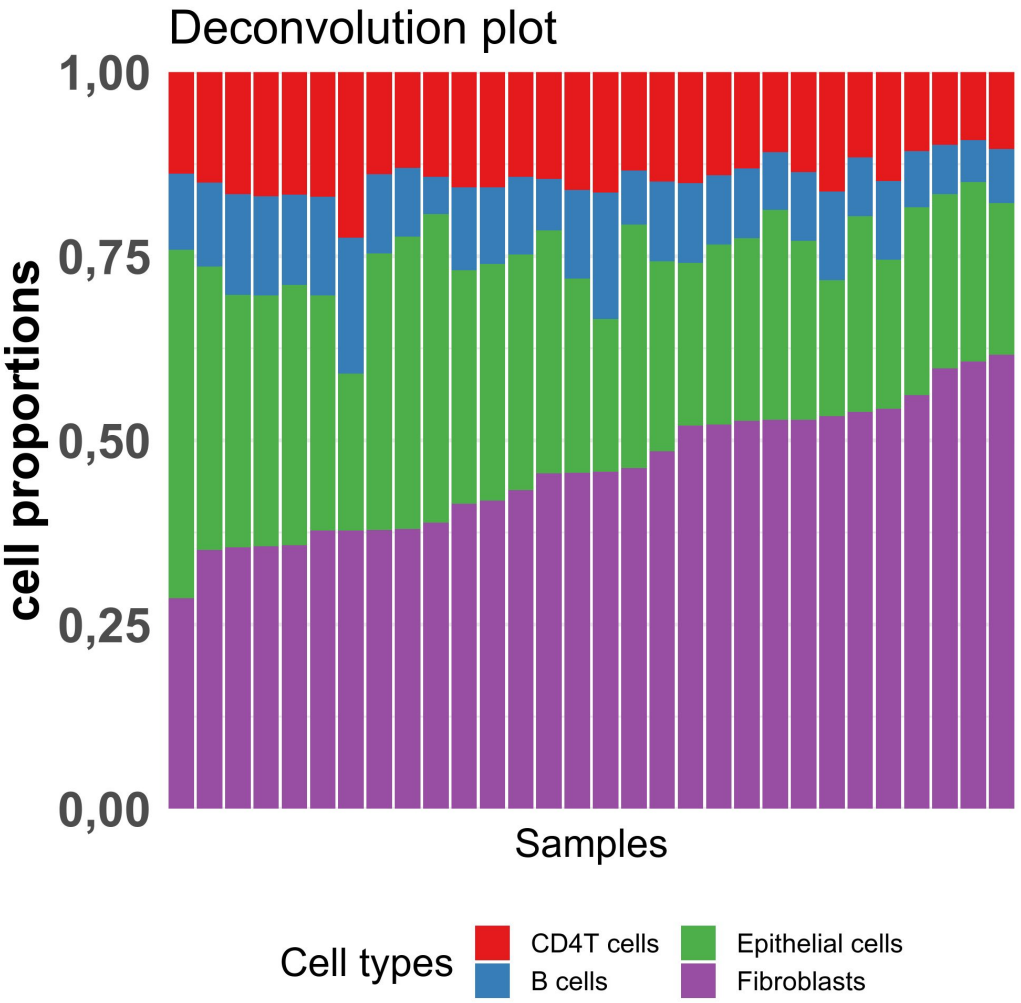
Integration (met+rna)

➔ Using a single method very good at a given task seems better than combining methods

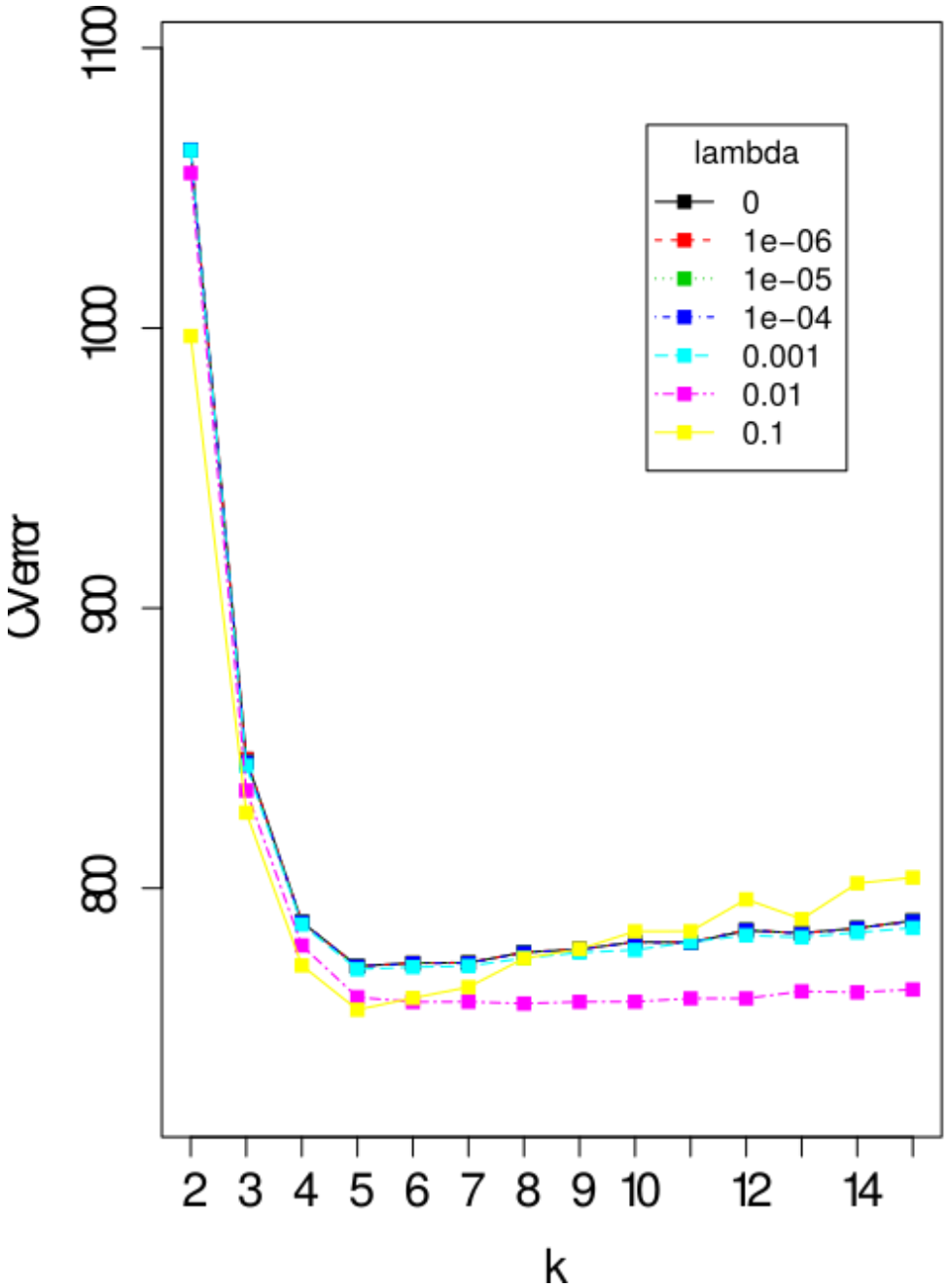
Interpretation

We kept K = 4 because despite trying with 5 different cell types for very many different parameter combinations, our best scores were always with 4 different cell types.

We chose as immune cells B and CD4T cells because on our tests these seemed to be the most relevant in the data



Team G



Quality filtering

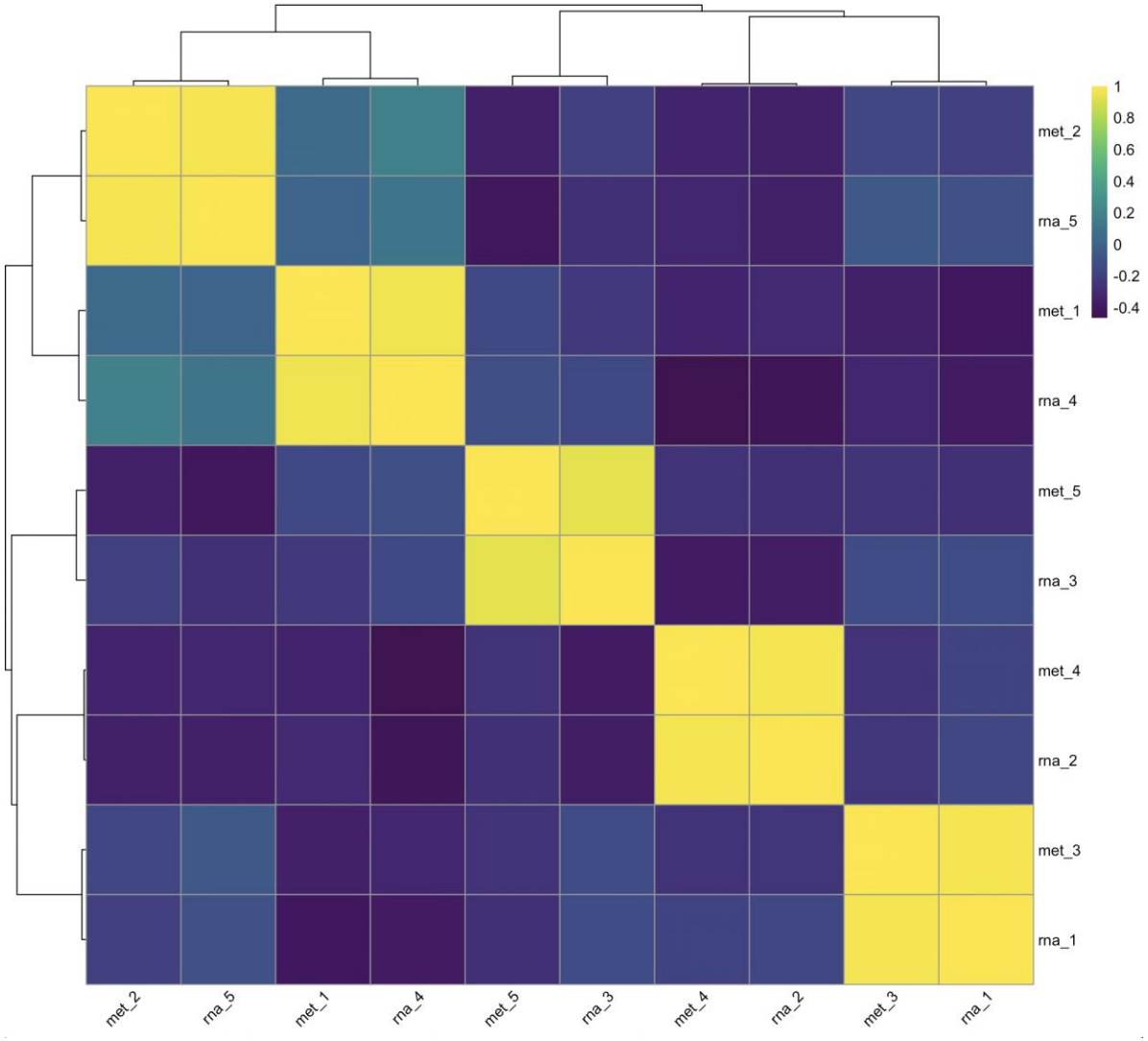
SNPs
Sex chromosomes
Cross-reactive sites
Outside of CpG context

CpG selection

4,000 (5,000) most variable

MeDeCom

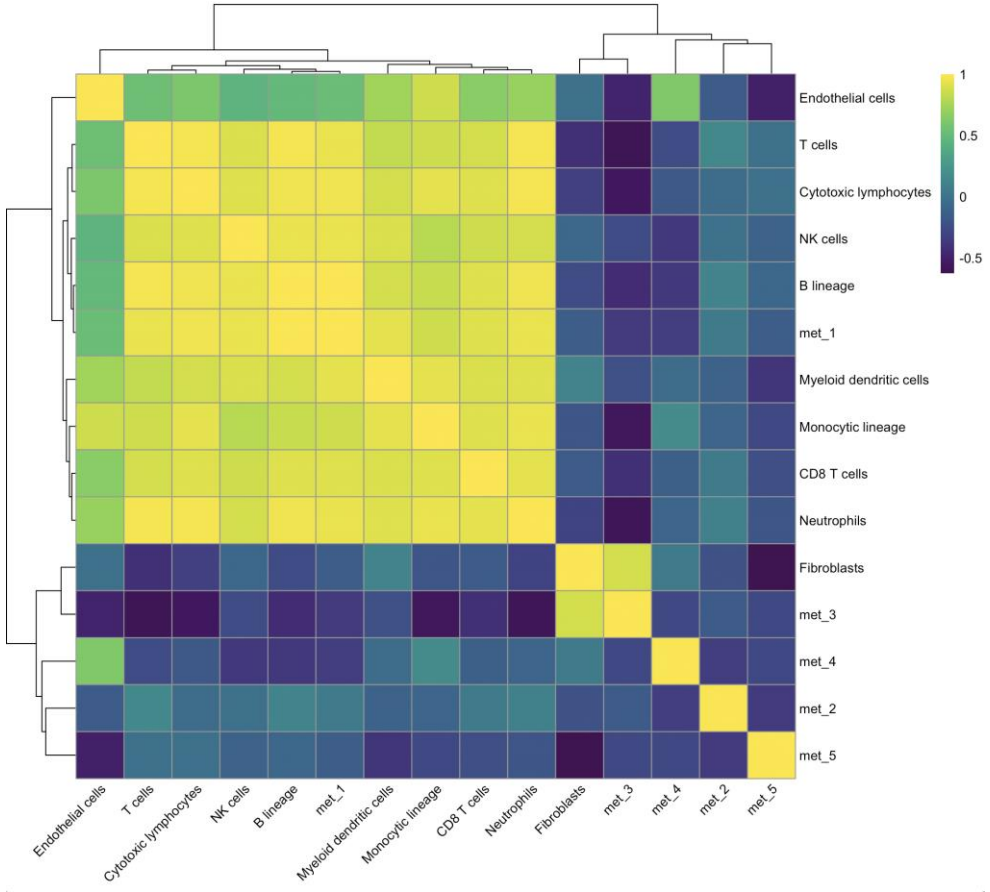
high correlation between RNA and methylation

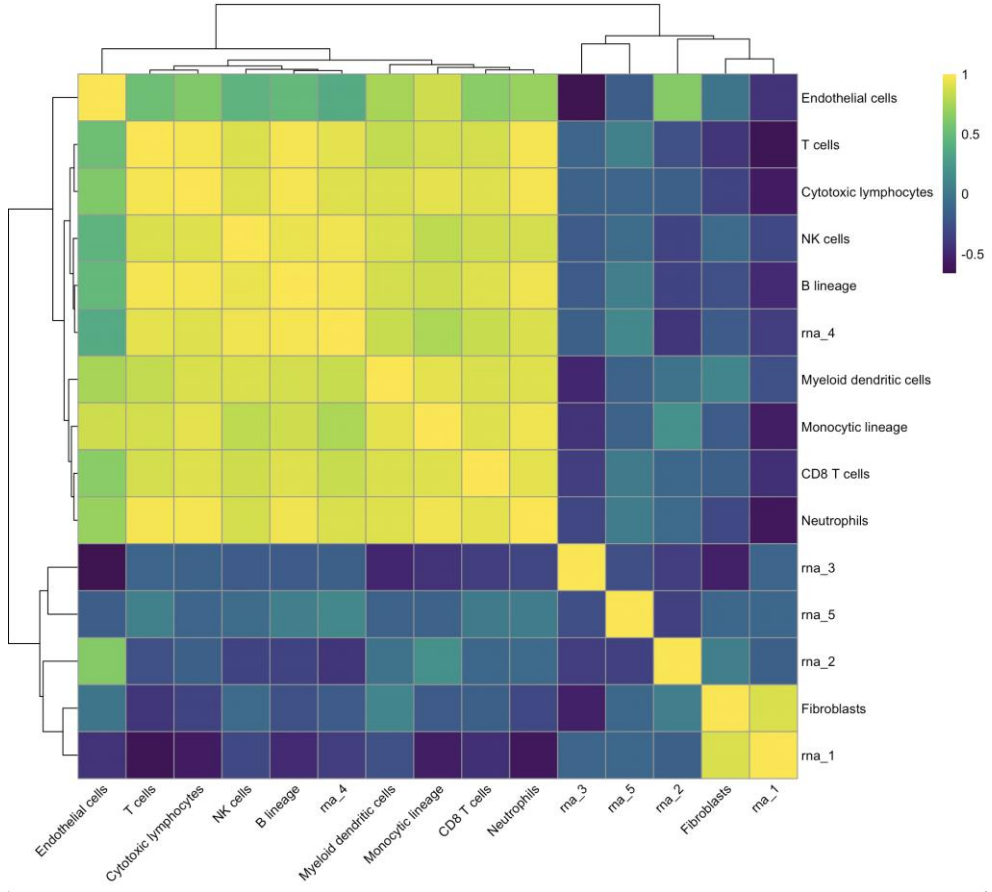


met_2 = rna_5
met_1 = rna_4
met_5 = rna_3
met_4 = rna_2
met_3 = rna_1

Annotation with MCP counter

- Correlation between mcp-counter scores and proportion matrices of RNA and methylation





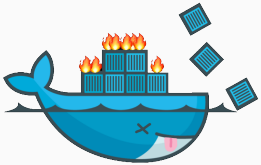
Team H(elloWorld)

Nicolas Alcala, Ghislain Durif, Milan Jakobi, Paulina Jedynak

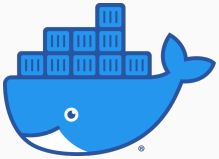
November 29, 2019

Disclamer

Docker works in mysterious ways...



...actually NOT, EpiDISH does!!!



Some explorations

Using NMF & improved NMF algorithm (e.g. pCMF) to estimate D with ”raw” data

- Poor performance on transcriptomic datas (> 0.10)
- Better ones on EPIC datas (< 0.10)

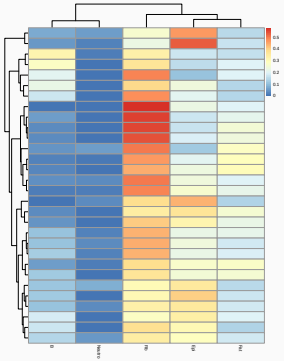
Trying to find some filters

- Removing probes on sexual chromosomes (no improvement)
- Using only 6 genes known as related to Pancreatic cancer led to improvement (~ 0.08)

In the end, we tried RefFreeEWAS with different parameters and those 2 filters but couldn’t reach a performance under 0.1

Tests: supervised approach (epiDISH)

epiDISH with various reference matrices



Issue: how to estimate what is not in the references?

Tests: unsupervised (MOFA)

Joint factorization of the Methylation and RNA matrices with MOFA

1. Filter sex probes
2. use most variable (75% genes from RNA, 5% CpGs from EPIC array)
3. Transform β-values into M-values
4. Run MOFA

Two strategies:

1. Hack the deconICA scoring method: get top genes/CpGs, compute their average level in each sample
2. Use weighted fuzzy clustering (C-means): weight by variance explained each axis,

Issue: does not take into account known types

Final: semi-supervised (epiDISH+NMF)

How to combine the supervised approach and unsupervised approach?

1. Compute estimate of some types using epiDISH
2. Filter sex probes
3. Regress the effect of the estimated cell type on RNA and methylation matrices
4. Compute NMF on the matrices