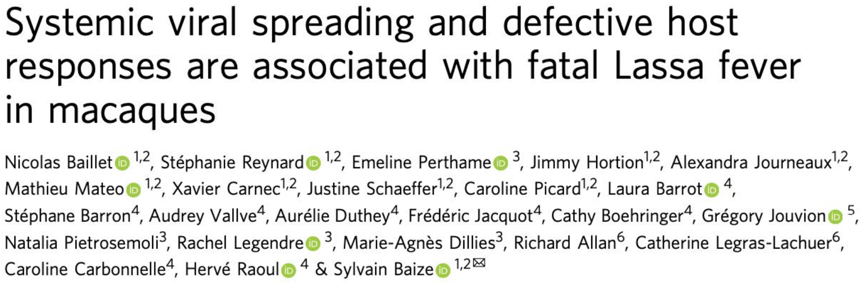
Application of deconvolution tools on RNA-Seq data

Emeline, Hélène, Natalia

Sys Bio group meeting
03/06/2021

Context and data



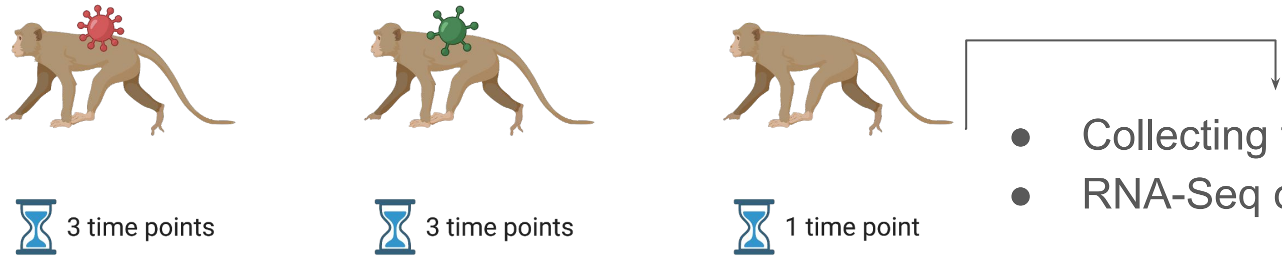
Systemic viral spreading and defective host responses are associated with fatal Lassa fever in macaques

Nicolas Baillet[1,2], Stéphanie Reynard[1,2], Emeline Perthame[3], Jimmy Hortion[1,2], Alexandra Journeaux[1,2], Mathieu Mateo[1,2], Xavier Carnec[1,2], Justine Schaeffer[1,2], Caroline Picard[1,2], Laura Barrot[4], Stéphane Barron[4], Audrey Vallve[4], Aurélie Duthey[4], Frédéric Jacquot[4], Cathy Boehringer[4], Grégory Jouvion[5], Natalia Pietrosemoli[3], Rachel Legendre[3], Marie-Agnès Dillies[3], Richard Allan[6], Catherine Legras-Lachuer[6], Caroline Carbonnelle[4], Hervé Raoul[4] & Sylvain Baize[1,2]✉

Sylvain Baize unit, Biology of Viral Emerging Infections (IP, Lyon)

Lassa fever = Hemorrhagic fever

Identify markers of early infection by Lassa fever

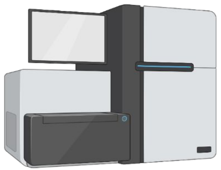


3 time points

3 time points

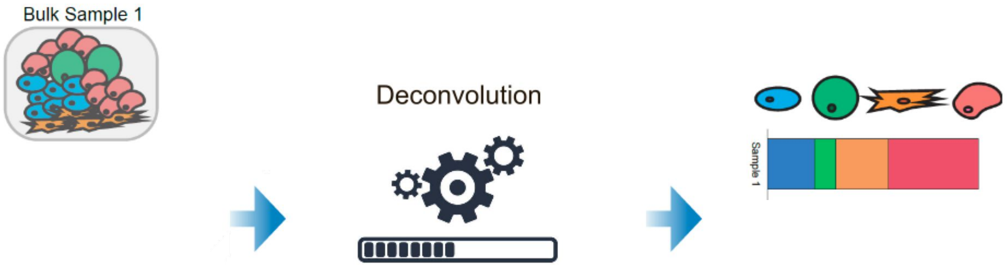
1 time point

- Collecting tissues
- RNA-Seq on PBMC



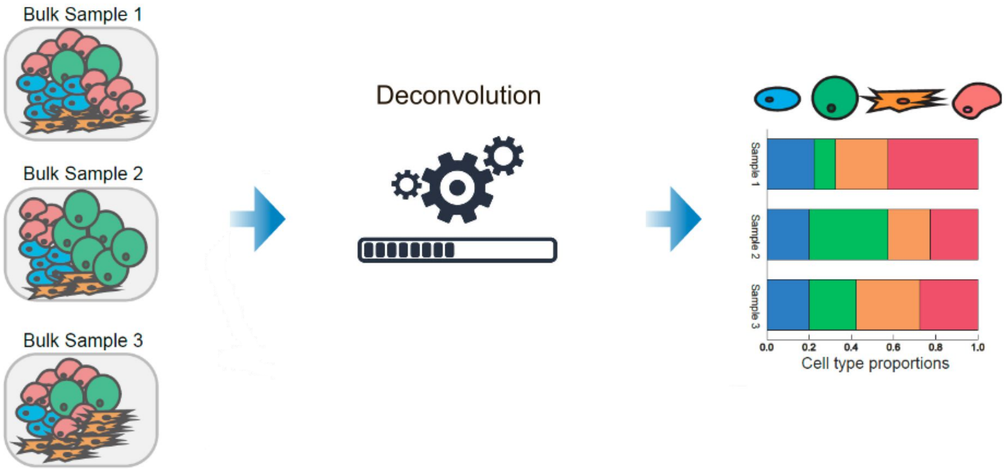
Differential analysis + functional analysis → publication

Deconvolution methods for transcriptomic data



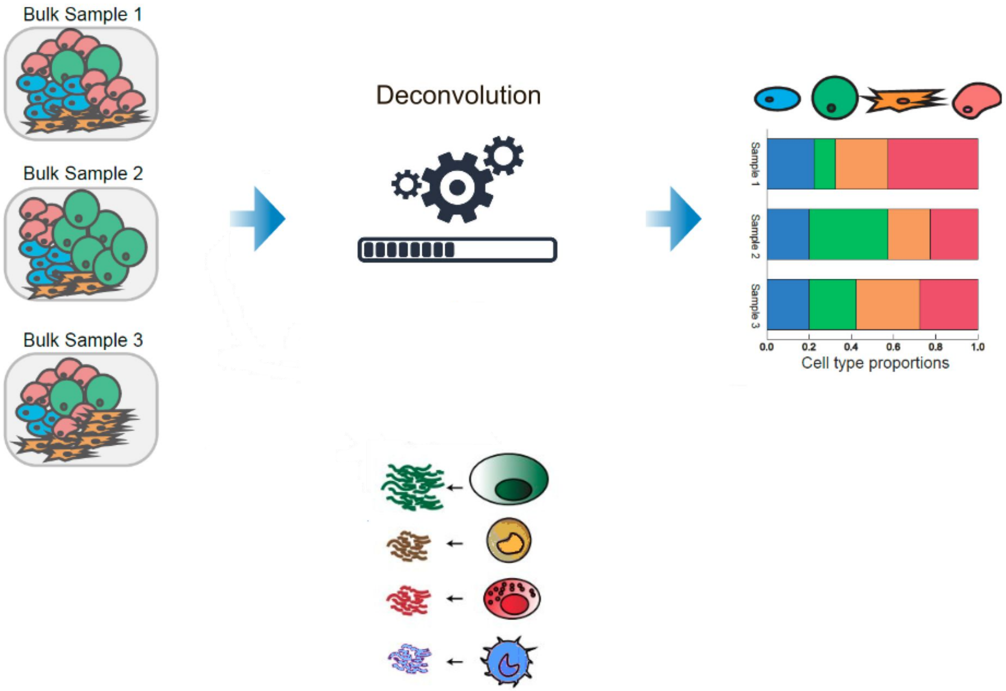
Computationally **inferring cell type proportions** from bulk heterogeneous mixtures

Deconvolution methods for transcriptomic data



Computationally **inferring cell type proportions** from bulk heterogeneous mixtures

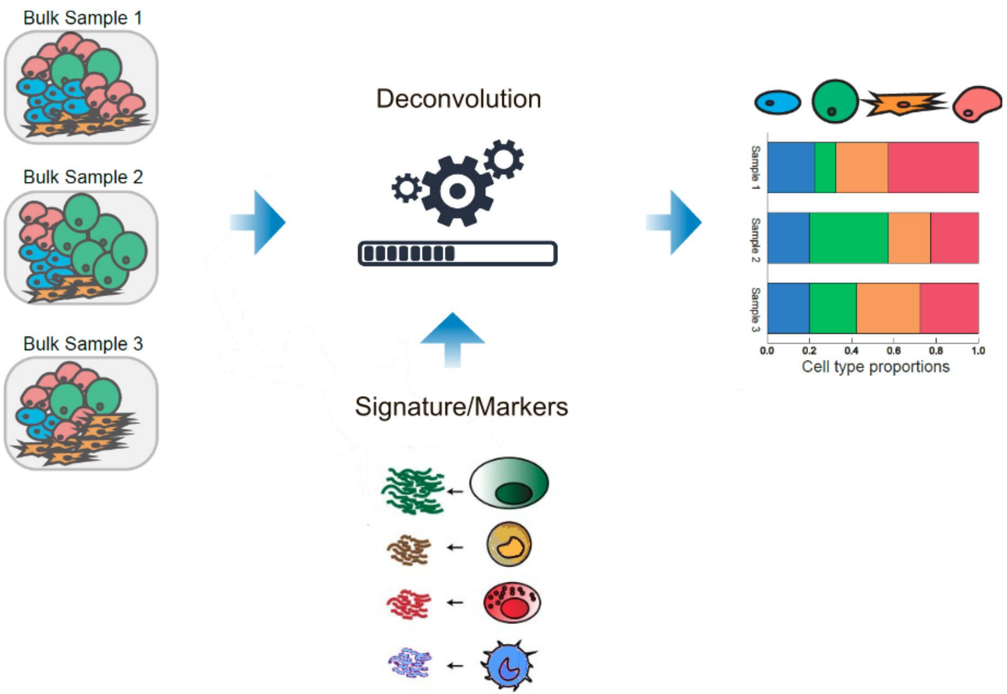
Deconvolution methods for transcriptomic data



Computationally **inferring cell type proportions** from bulk heterogeneous mixtures

Averaged expression levels of indiv. genes ≠ individual measures for each gene across the different cell types

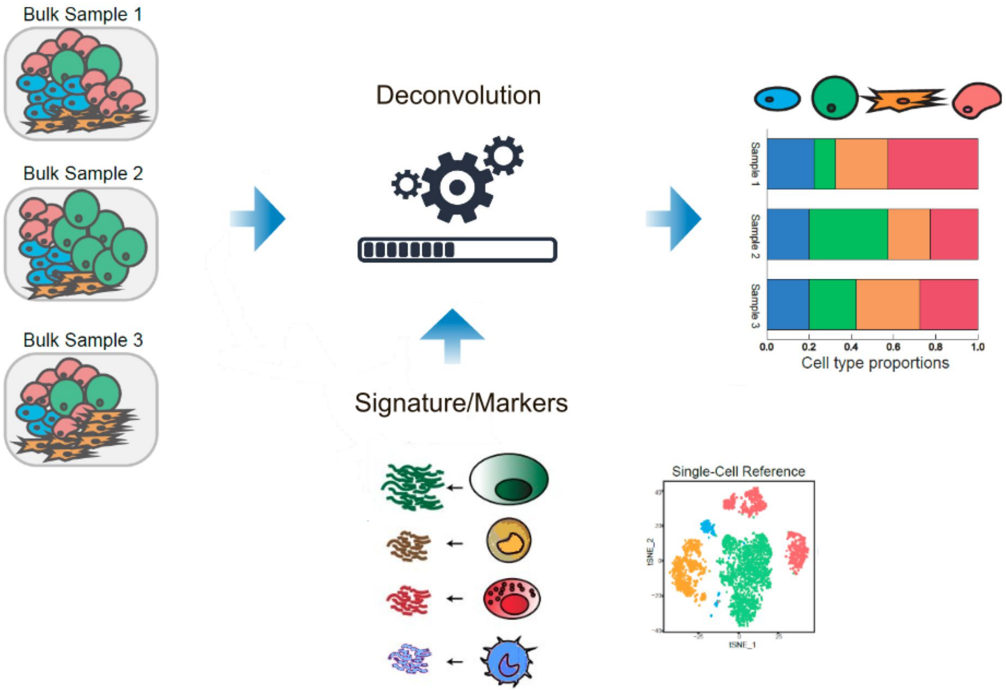
Deconvolution methods for transcriptomic data



Computationally **inferring cell type proportions** from bulk heterogeneous mixtures

Averaged expression levels of indiv. genes ≠ individual measures for each gene across the different cell types

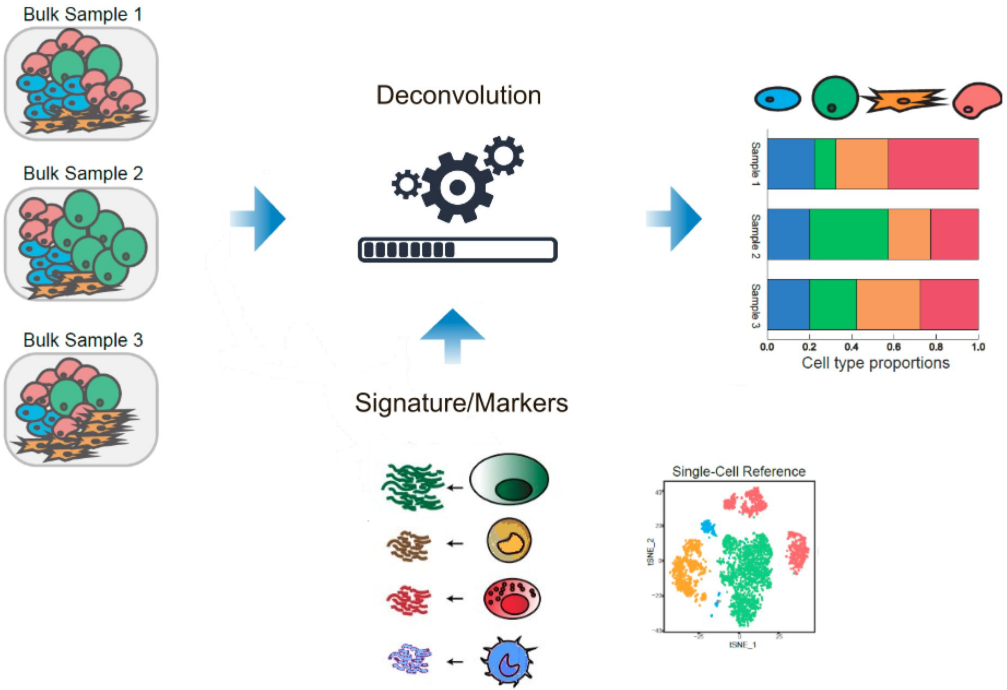
Deconvolution methods for transcriptomic data



Computationally **inferring cell type proportions** from bulk heterogeneous mixtures

Averaged expression levels of indiv. genes ≠ individual measures for each gene across the different cell types

Deconvolution methods for transcriptomic data



What does it add:

- A more specific analysis than the differential analysis, which can be confounded by differences in cell type proportions

- Being able to infer specific cell type behaviour so they can be targeted

- Alternative to single cell transcriptomics sequencing, fluorescence-activated cell sorting (FACS), immunohistochemistry (IHC) (cost-effective, time-effective)

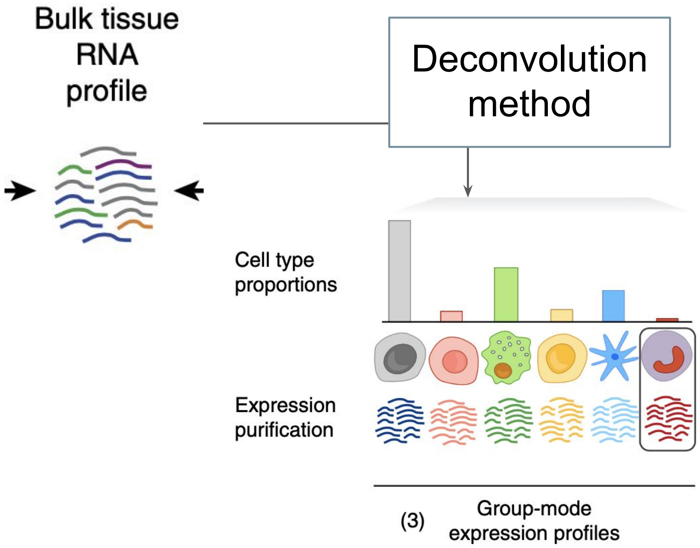
Deconvolution methods for transcriptomic data

Computationally **infering cell type proportions** from bulk heterogeneous mixtures

Averaged expression levels of indiv. genes ≠ individual measures for each gene across the different cell types

What does it add:

- A more specific analysis than the differential analysis, which can be confounded by differences in cell type proportions
- Being able to infer specific cell type behaviour so they can be targeted
- Alternative to single cell transcriptomics sequencing, fluorescence-activated cell sorting (FACS), immunohistochemistry (IHC) (cost-effective, time-effective)



Newman, A.M., Steen, C.B., Liu, C.L. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 37, 773–782 (2019). https://doi.org/10.1038/s41587-019-0114-2

Overview of deconvolution methods

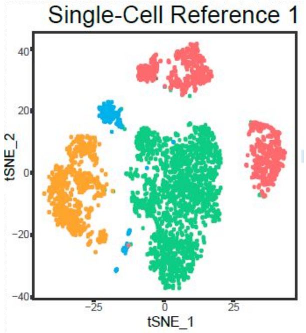
Two families of methods:

- Methods using a signature "file" to identify cellular types and quantify them
  - The famous lm22 matrix that defines genes & cellular types providing an expression level for each
  - Another simpler signature that indicates only specific markers for each cell type

e.g **MCP counter**, **CIBERSORT**, **OLS**, **nnls**, **RLR** & **FARDEEP**

- (Newer) methods based on annotated single cell RNA-Seq datasets

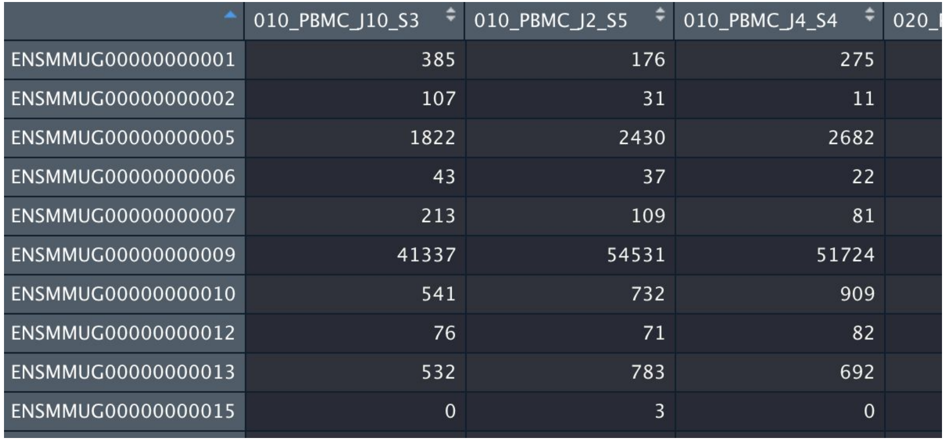
e.g. MUSIC, SCDC, CIBERSORTx, DWLS



Many of them are based in linear models

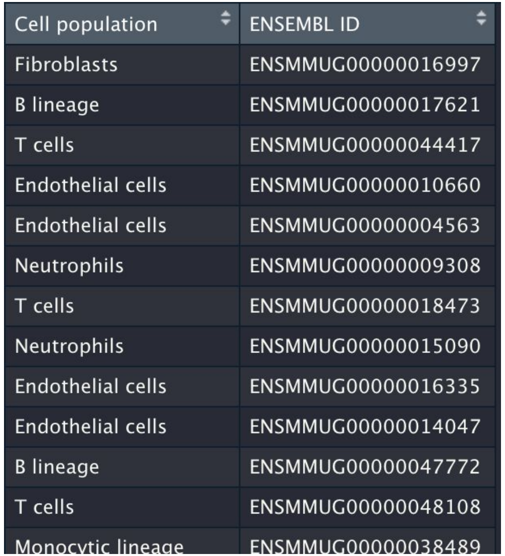
Input

1) Gene counts matrix



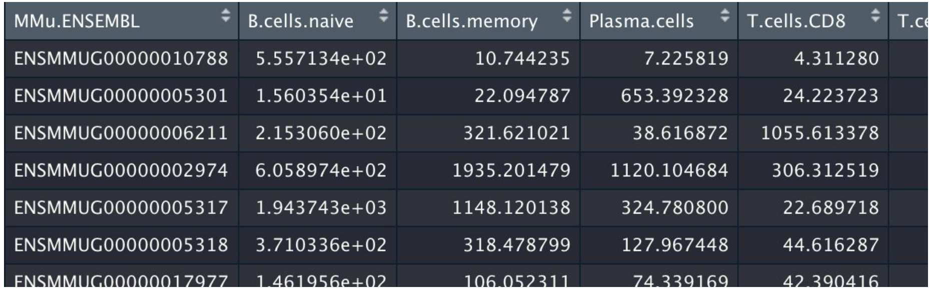
	010_PBMC_J10_S3	010_PBMC_J2_S5	010_PBMC_J4_S4	020_
ENSMMUG00000000001	385	176	275	
ENSMMUG00000000002	107	31	11	
ENSMMUG00000000005	1822	2430	2682	
ENSMMUG00000000006	43	37	22	
ENSMMUG00000000007	213	109	81	
ENSMMUG00000000009	41337	54531	51724	
ENSMMUG00000000010	541	732	909	
ENSMMUG00000000012	76	71	82	
ENSMMUG00000000013	532	783	692	
ENSMMUG00000000015	0	3	0	

2) Signature **vector**



Cell population	ENSEMBL ID
Fibroblasts	ENSMMUG00000016997
B lineage	ENSMMUG00000017621
T cells	ENSMMUG00000044417
Endothelial cells	ENSMMUG00000010660
Endothelial cells	ENSMMUG00000004563
Neutrophils	ENSMMUG00000009308
T cells	ENSMMUG00000018473
Neutrophils	ENSMMUG00000015090
Endothelial cells	ENSMMUG00000016335
Endothelial cells	ENSMMUG00000014047
B lineage	ENSMMUG00000047772
T cells	ENSMMUG00000048108
Monocytic lineage	ENSMMUG00000038489

2) or signature **matrix (LM22)**



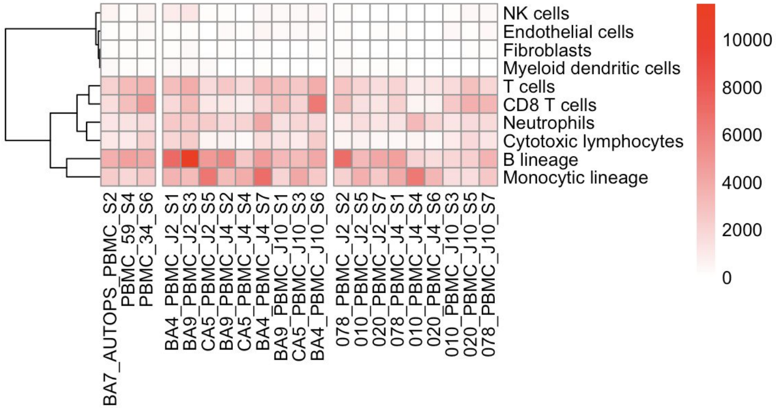
MMu.ENSEMBL	B.cells.naive	B.cells.memory	Plasma.cells	T.cells.CD8	T.ce
ENSMMUG00000010788	5.557134e+02	10.744235	7.225819	4.311280	
ENSMMUG00000005301	1.560354e+01	22.094787	653.392328	24.223723	
ENSMMUG00000006211	2.153060e+02	321.621021	38.616872	1055.613378	
ENSMMUG00000002974	6.058974e+02	1935.201479	1120.104684	306.312519	
ENSMMUG00000005317	1.943743e+03	1148.120138	324.780800	22.689718	
ENSMMUG00000005318	3.710336e+02	318.478799	127.967448	44.616287	
ENSMMUG00000017977	1.461956e+02	106.052311	74.339169	42.390416	

Signatures available for human and mice → Macaca?

Output

A score indicating the abundance level of cell types candidates

Basically, a matrix such as



Different variations around linear model

Several methods rely on different ways to estimate the following linear model

$$Y_i = X\beta_i + \varepsilon_i$$

Y_i is a p-vector of (transformed/normalized) gene counts for sample i.

X is a $p \times q$ signature matrix.

ε_i is a random error term.

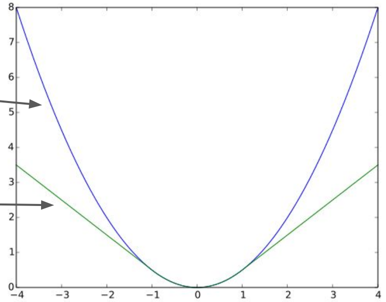
Deconvolution methods estimate the mixture coefficients β_i for each sample.

OLS minimizes the least squares (quadratic loss)

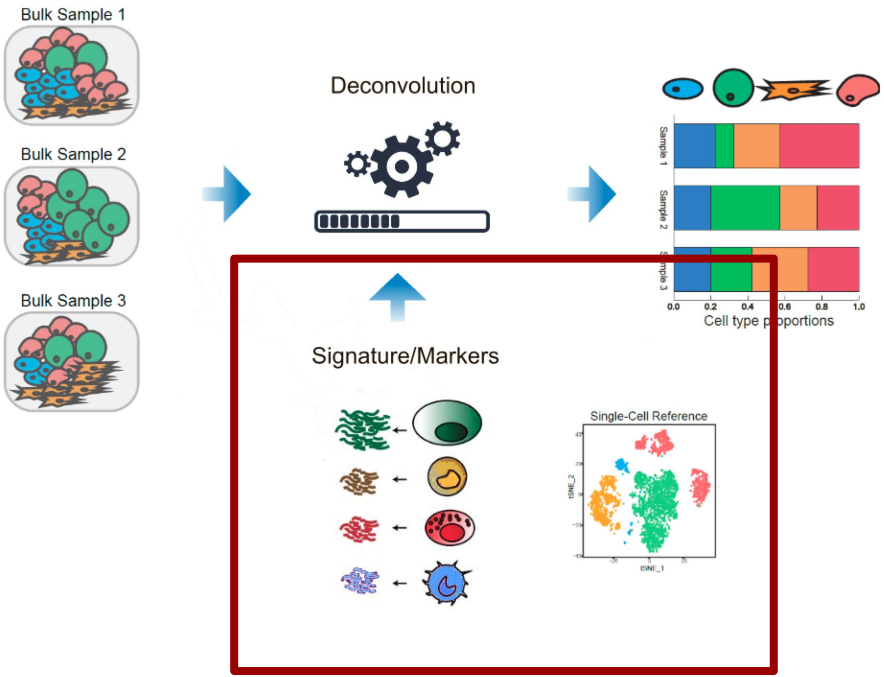
nnls minimizes the least squares with constraint $\beta_i \geq 0$

RLR minimizes Huber loss

FARDEEP uses adaptive least trimmed squares



Signatures - gene markers

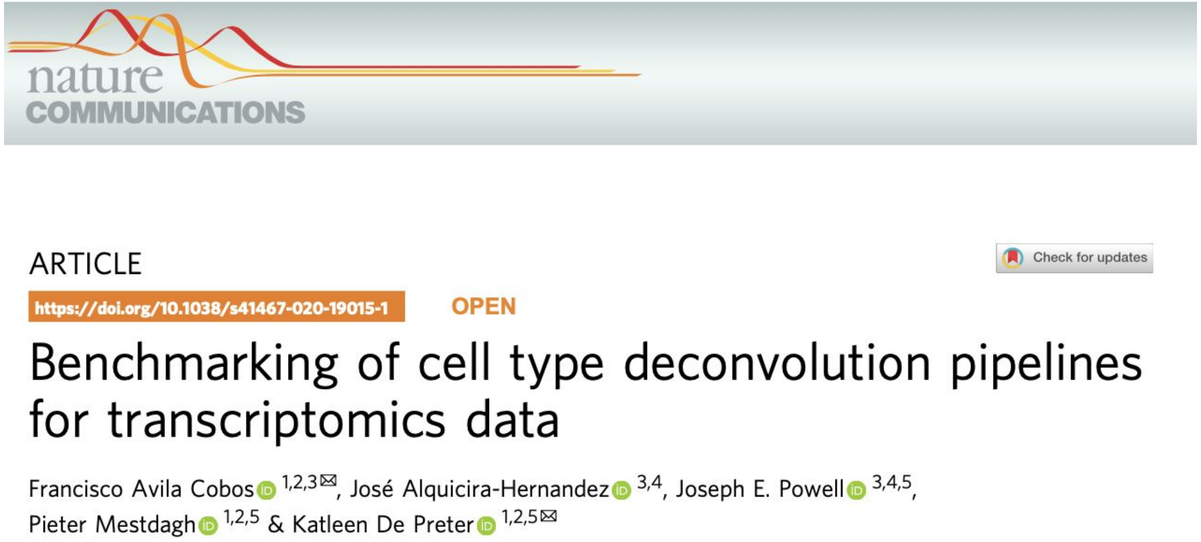


Existing Signatures : immune cells (Tumor env.)

Custom Signatures is a critical step for both bulk and single cell tools

- Datasets available for tissues/species
- All cell types must be present
- Close cell types sharing similar signatures

Reference and script



nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-19015-1 OPEN

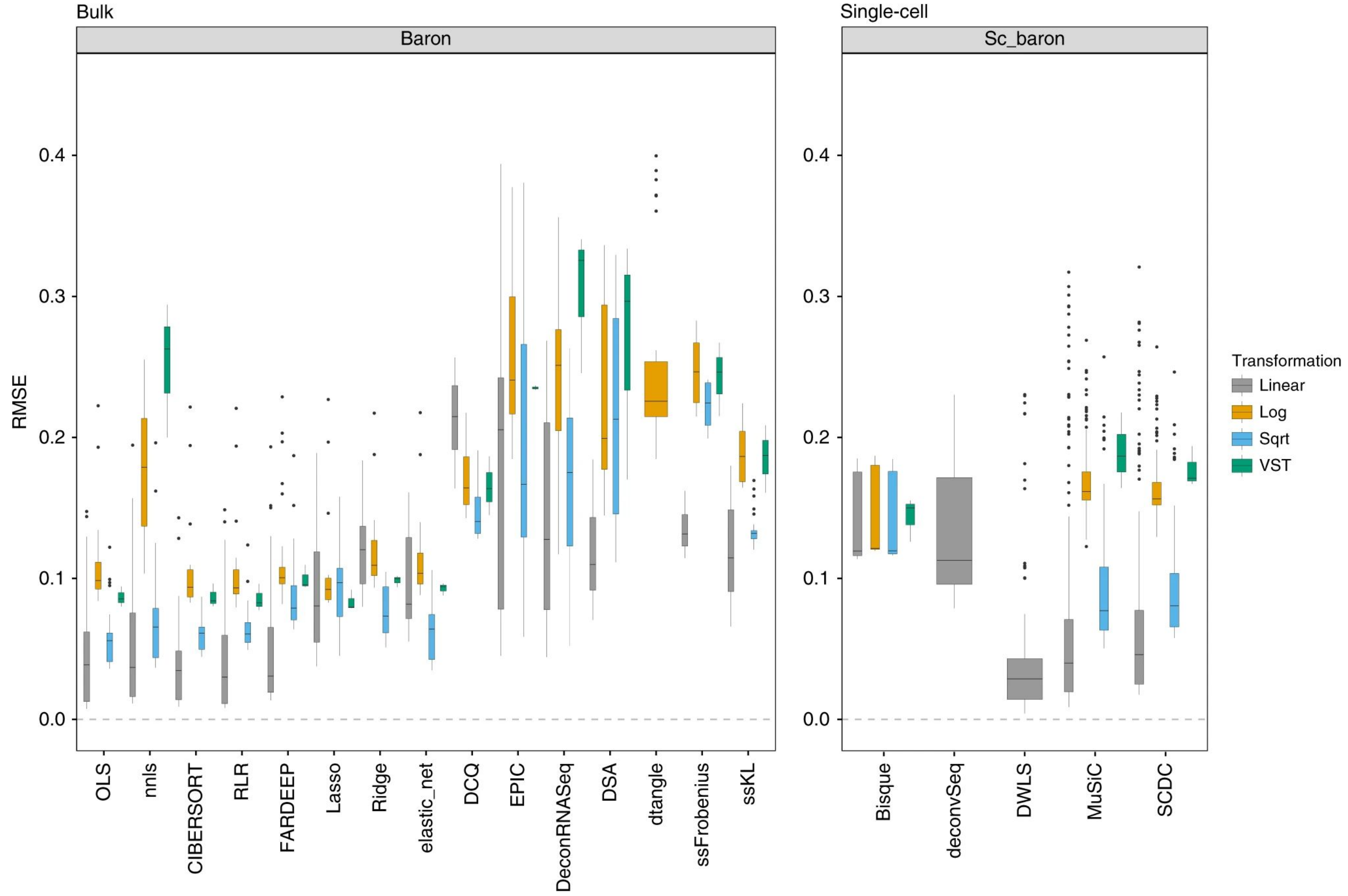
Benchmarking of cell type deconvolution pipelines for transcriptomics data

Francisco Avila Cobos[1,2,3]✉, José Alquicira-Hernandez[3,4], Joseph E. Powell[3,4,5], Pieter Mestdagh[1,2,5] & Katleen De Preter[1,2,5]✉

Code to run the methods compared in this paper

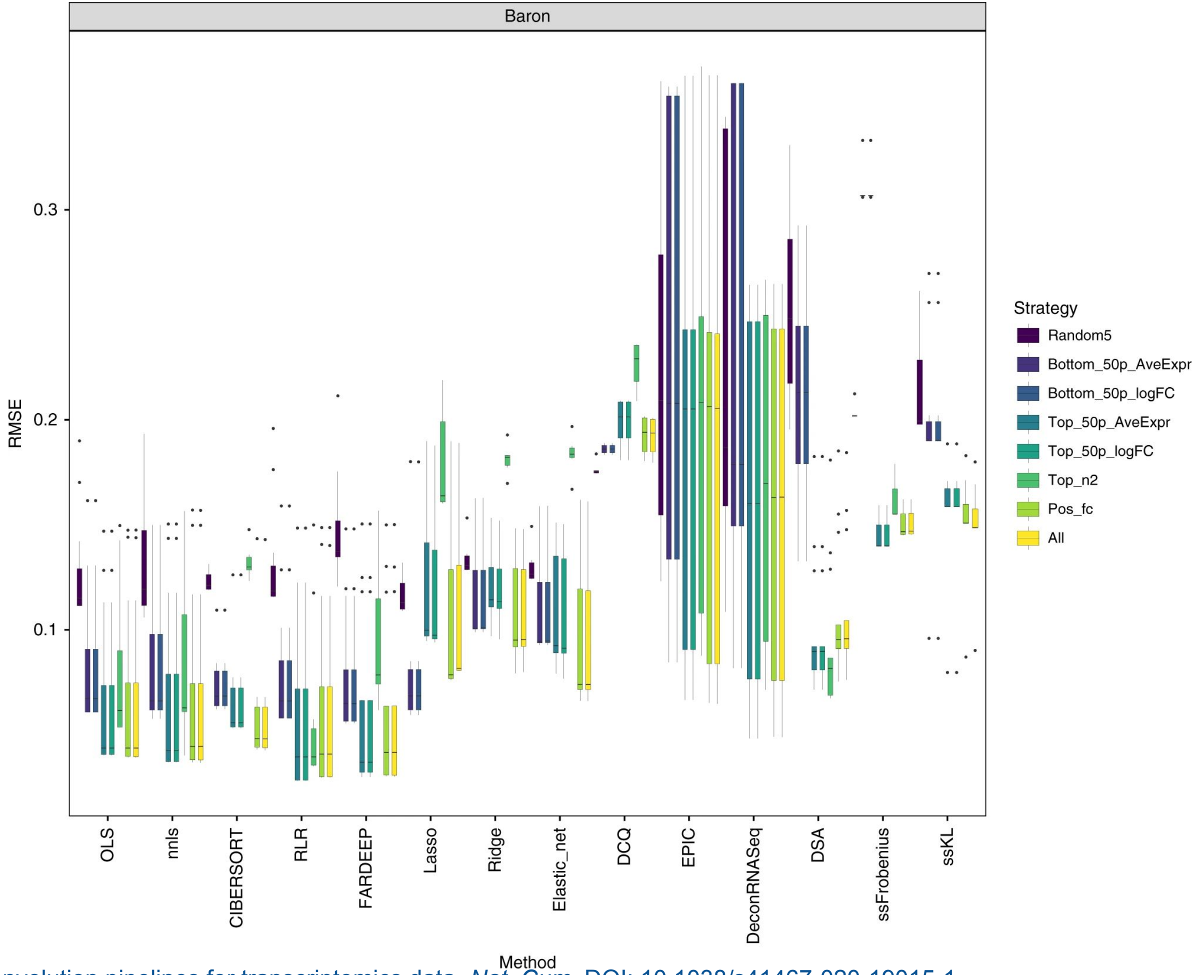
https://github.com/favilaco/deconv_benchmark

Impact of the data transformation on the deconvolution results



From: Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Cum*. DOI: 10.1038/s41467-020-19015-1

Impact of the marker selection on the deconvolution results



From: Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Cum*. DOI: 10.1038/s41467-020-19015-1

Deconvolution tools comparison

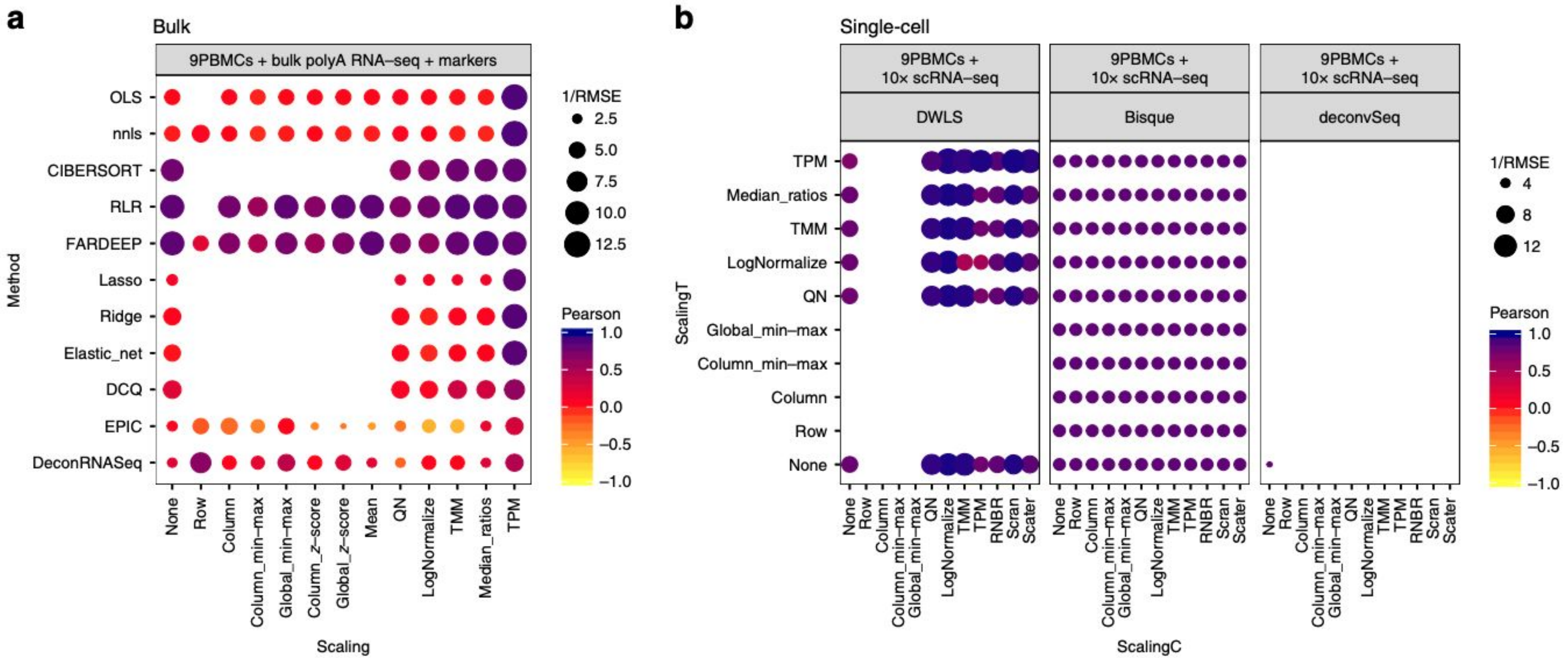
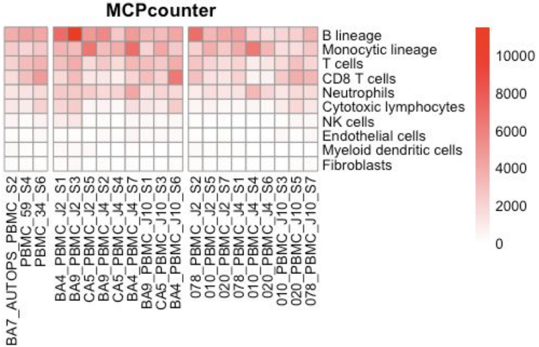
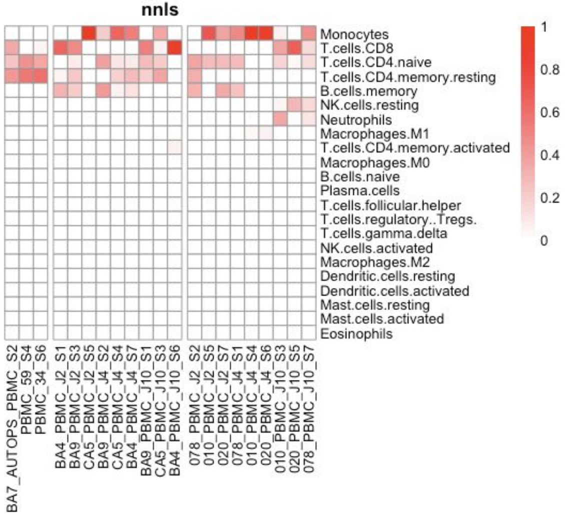


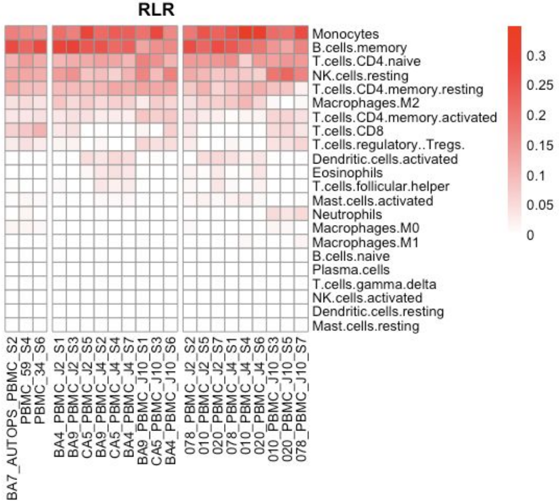
Fig. 7 Deconvolution performance on nine human PBMC bulk samples. With **a** bulk deconvolution methods; **b** deconvolution methods using scRNA-seq as reference.

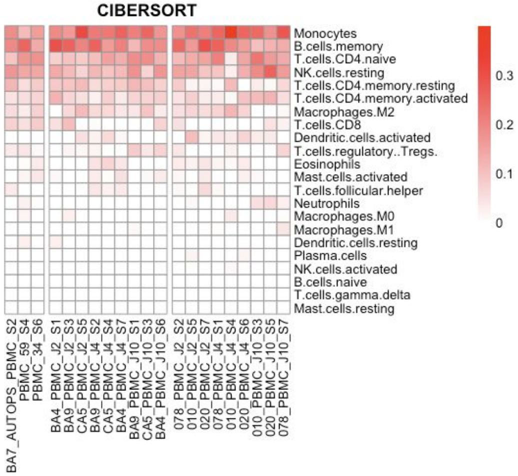
From: Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Cum*. DOI: 10.1038/s41467-020-19015-1

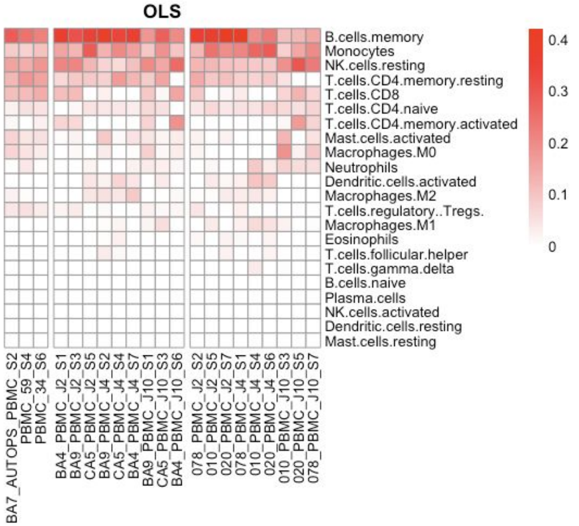
Results on our Macacas

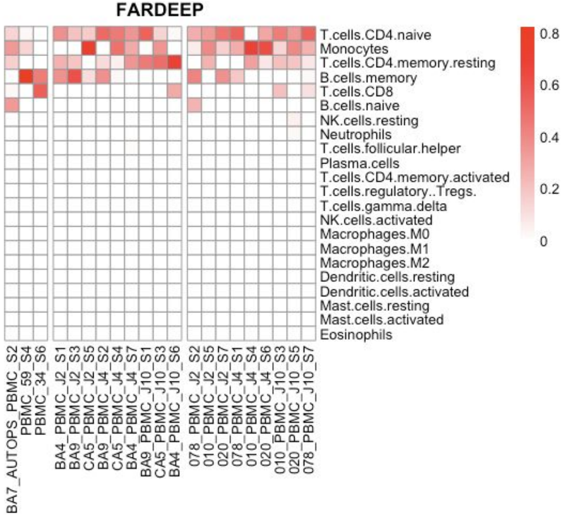

MCPcounter


nnls


RLR


CIBERSORT


OLS

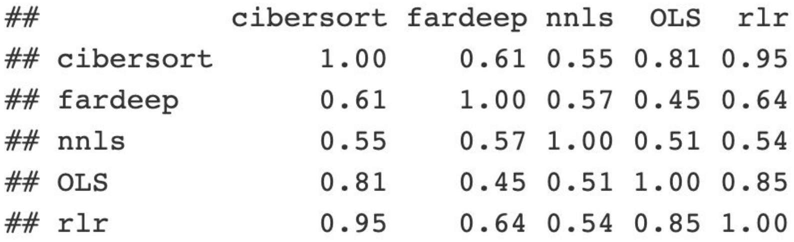

FARDEEP

Results on our Macacas - Similarity between methods

Correlation between scores (all cell types + all samples)

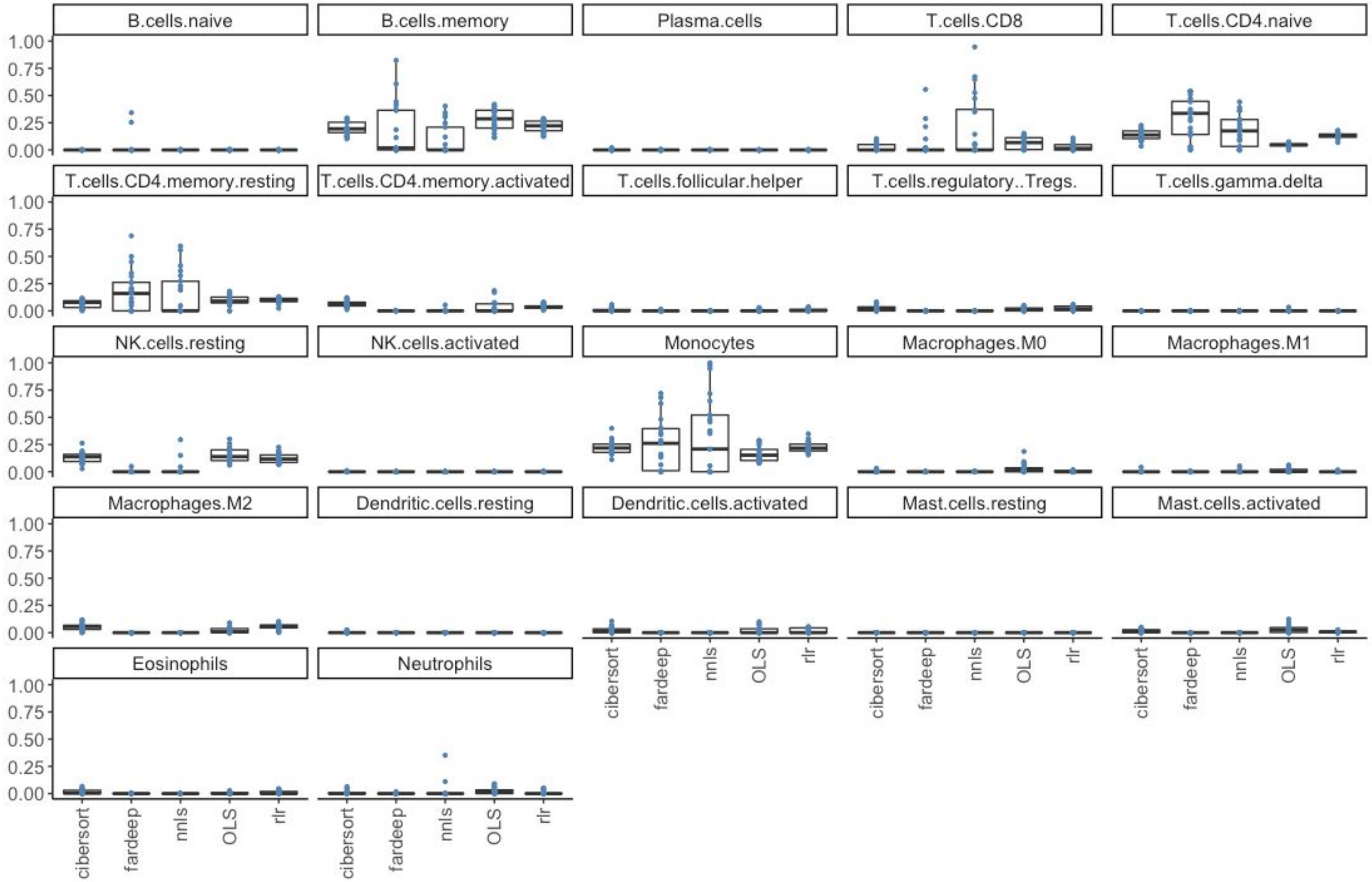
- CIBERSORT & RLR
- OLS & RLR

- OLS & FARDEEP
- OLS & NNLS



```
##              cibersort fardeep nnls  OLS  rlr
## cibersort         1.00    0.61 0.55 0.81 0.95
## fardeep           0.61    1.00 0.57 0.45 0.64
## nnls              0.55    0.57 1.00 0.51 0.54
## OLS               0.81    0.45 0.51 1.00 0.85
## rlr               0.95    0.64 0.54 0.85 1.00
```

Results on our Macacas



Main factors affecting deconvolution results

- Data transformation (e.g. **linear**, log, sqrt, VST)
- Scaling / normalization (column-wise, min-max, logNormalize, ...)
- Marker selection / reference matrix → Highly dependant of biology !
- Cell type composition
- Method

Limitations

Existence of datasets including the cells of interest (for signature) : reference markers should include **all cell types** being part of the mixture ("semi-supervised") / H. sapiens - missing cellular components in the reference

Cell types sharing similar signatures / marker genes not being sufficiently cell-type specific

Highly dependant of biology !

Summary of methods

Methods	**MCP counter** 😐	**CIBERSORT** 🙂	**OLS** 🙂	**nnls** 🙂	**RLR** 🙂	**FARDEEP** 🙂	**MUSIC** 🙂	**DWLS** ☹️
Main principle	(bulk)	(bulk) support-vector	(bulk) least-squares	(bulk) non-negative least squares	(bulk) robust linear regression	(bulk) robust linear regression	(sc) Multi-subject Single-cell Deconvolution	(sc) dapenned-weighted least-squares
Reference signature	Marker vector (absence/presence)	Human LM22 (matrix)	Human LM22 (matrix)	Human LM22 (matrix)	Human LM22 (matrix)	Human LM22 (matrix)	annotated single cell datasets	
R package usage / performance		fast and easy (cibersort function and web)	easy to implement (lm native function)	easy to implement (nnls function)	easy to implement (rlm function)		easy (music function)	time consuming + package hard to install + examples not executable
Overall appreciation						:(No ML, despite name !		