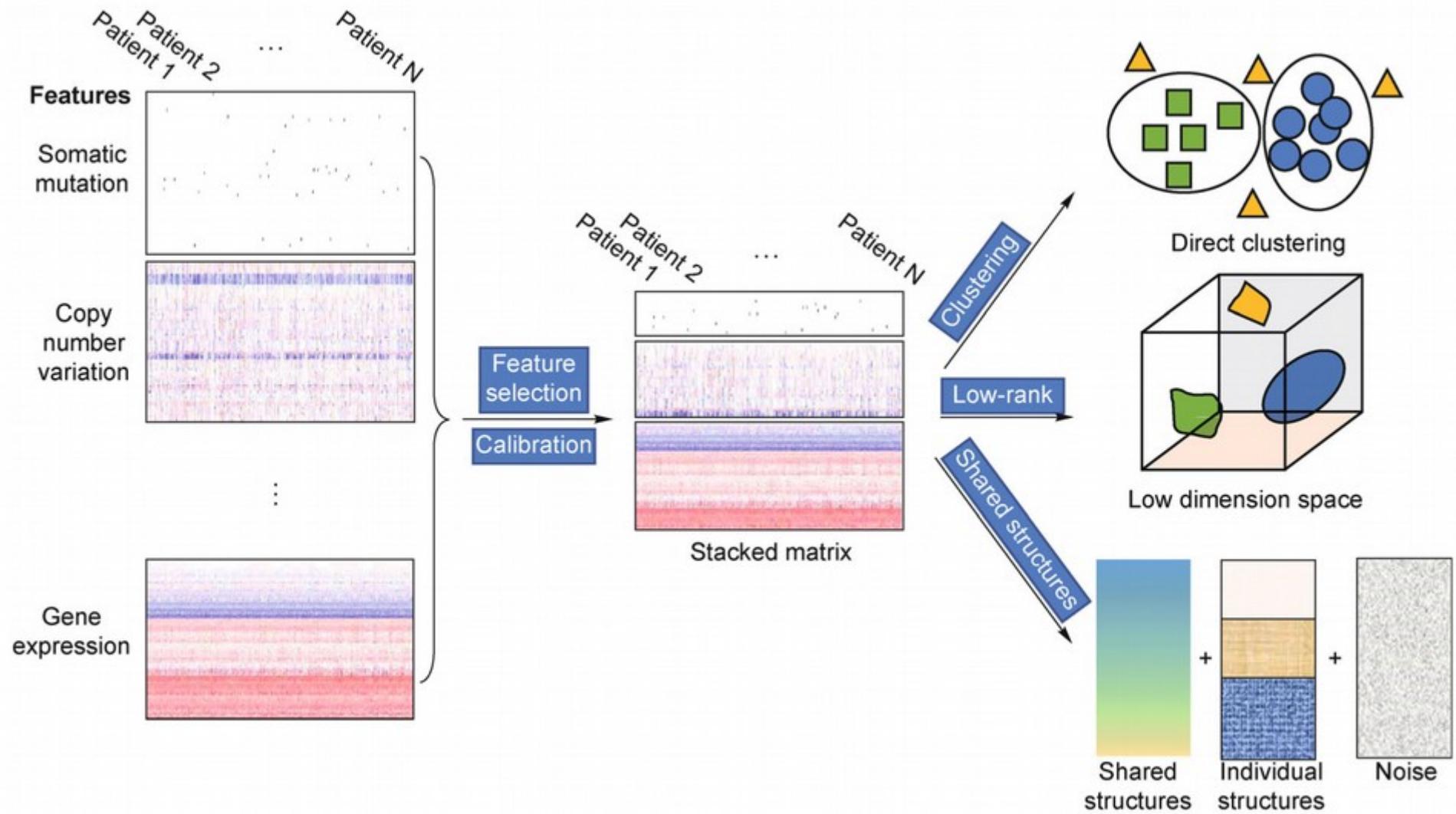


# Unsupervised OMICs Integration

OMICs Integration and Systems Biology course

Nikolay Oskolkov, NBIS SciLifeLab

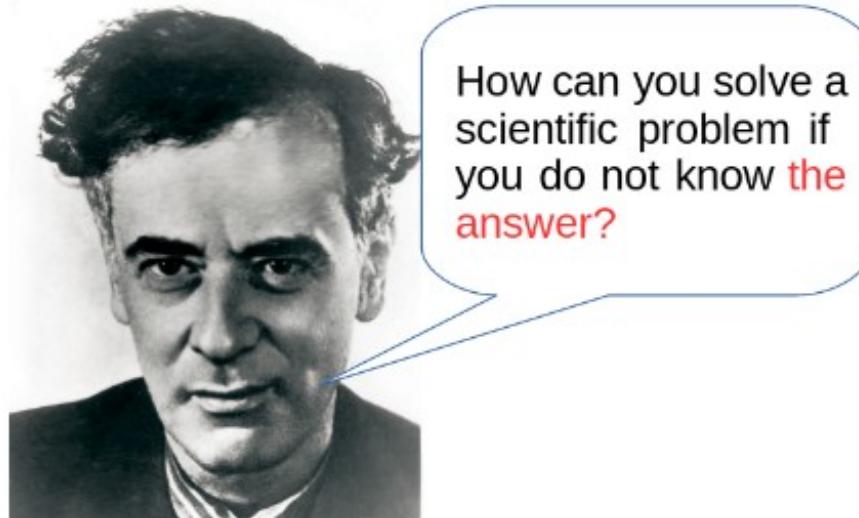
Lund, 5.10.2020



## Supervised Analysis: Commercial



Hypothesis-driven approach  
Mainstream research  
Guaranteed pubs and funds  
**Small leap in development**



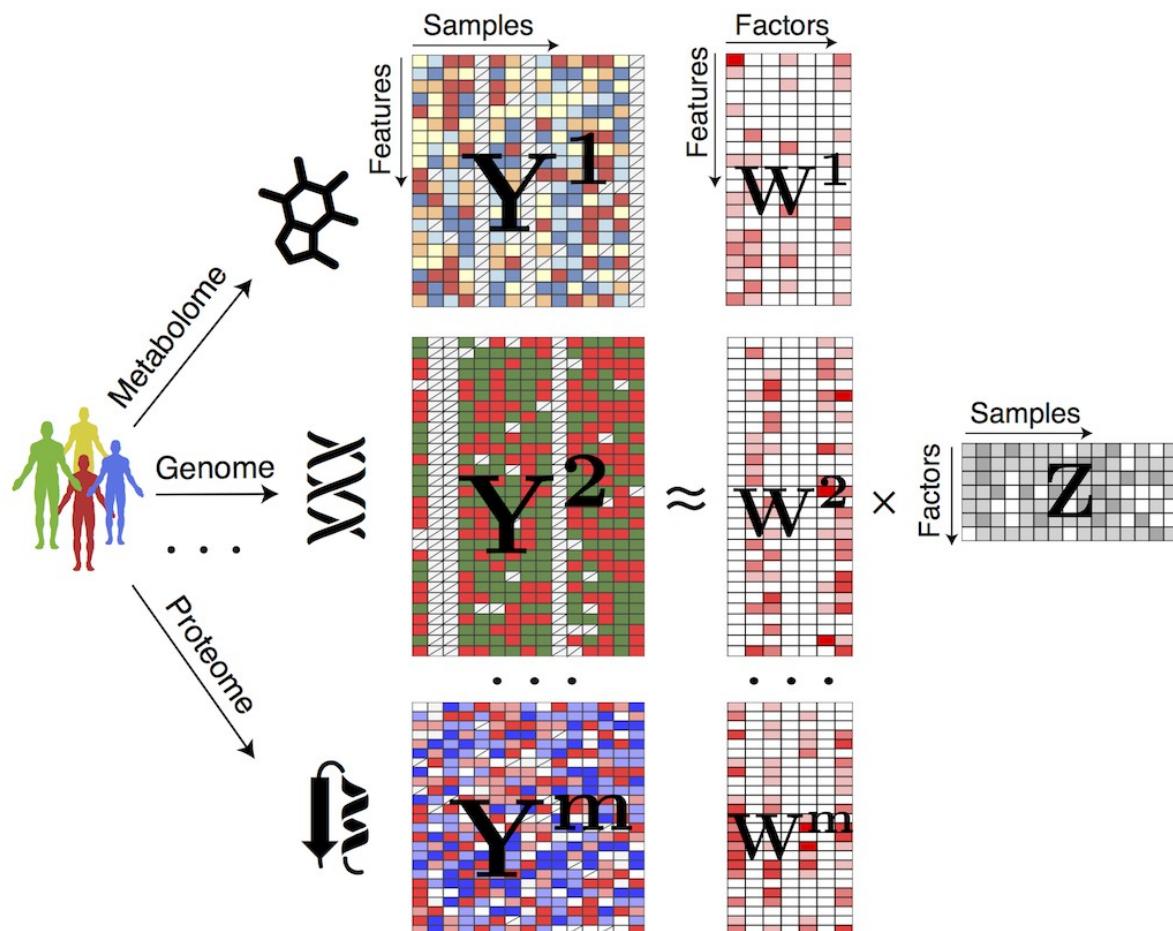
## Unsupervised Analysis: Underground



Data-driven approach  
"Fishing Expedition", "Very descriptive paper",  
"You do not know what you are doing"  
Hard to publish, no funding  
**Can be a revolution in research**

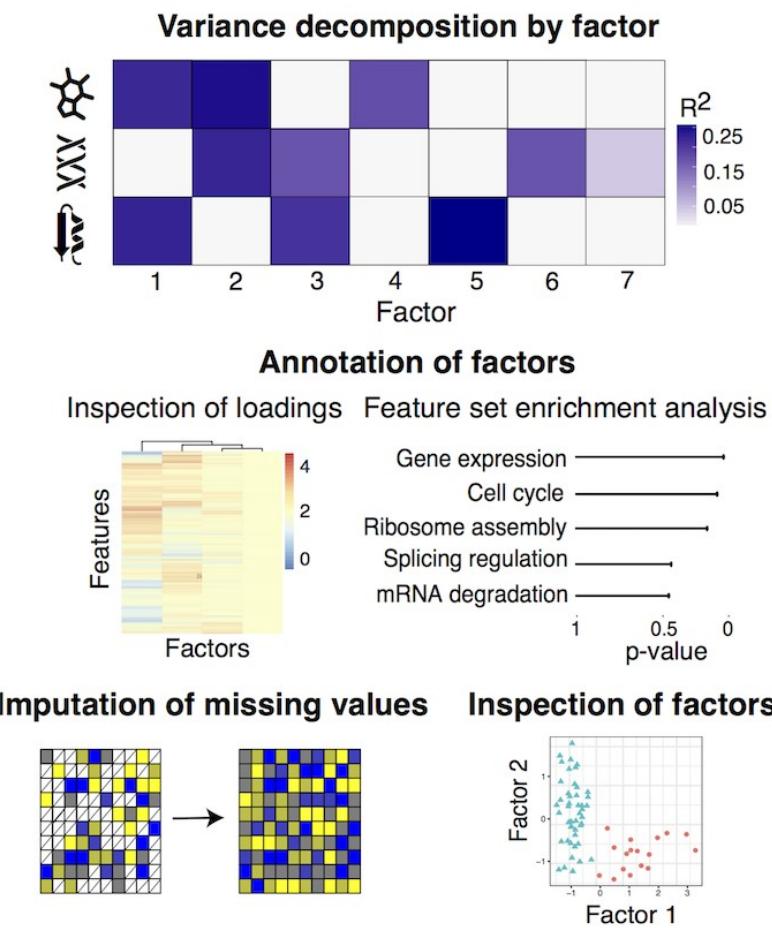


## Step 1: train a MOFA model



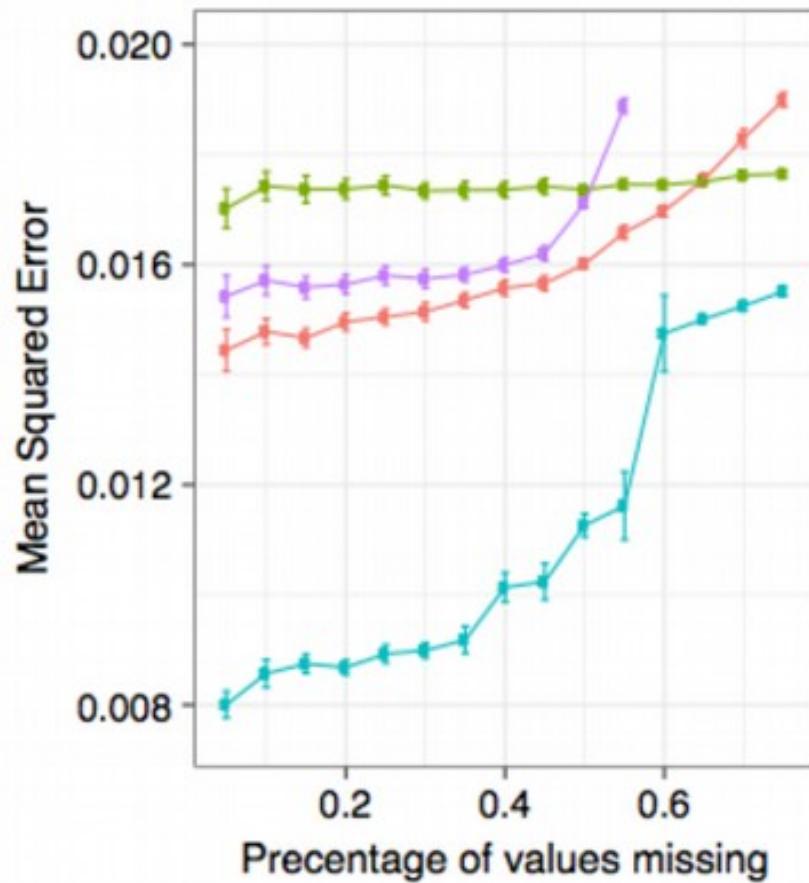
- Visualisation of samples in factor space
- Annotation of factors using (gene set) enrichment analysis
- Imputation of missing values
- Support of OMICs with non-Gaussian distribution including binary and count data

## Step 2: downstream analysis

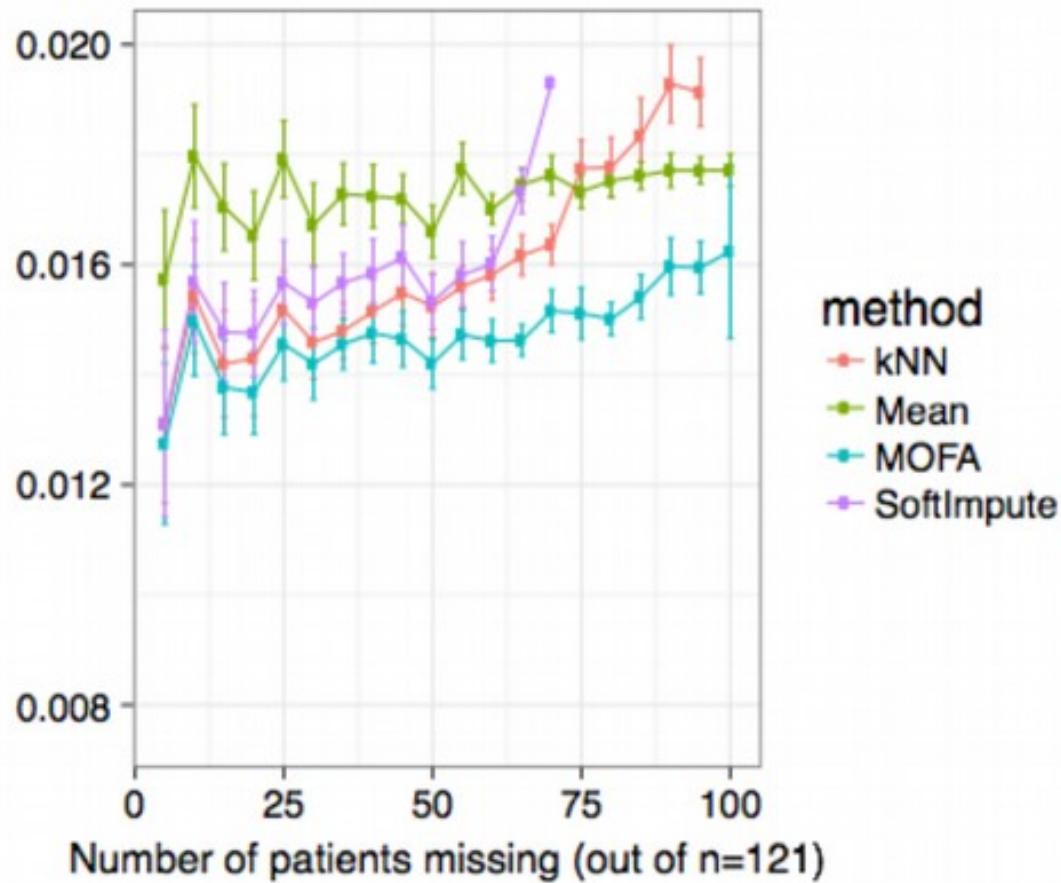


**a**

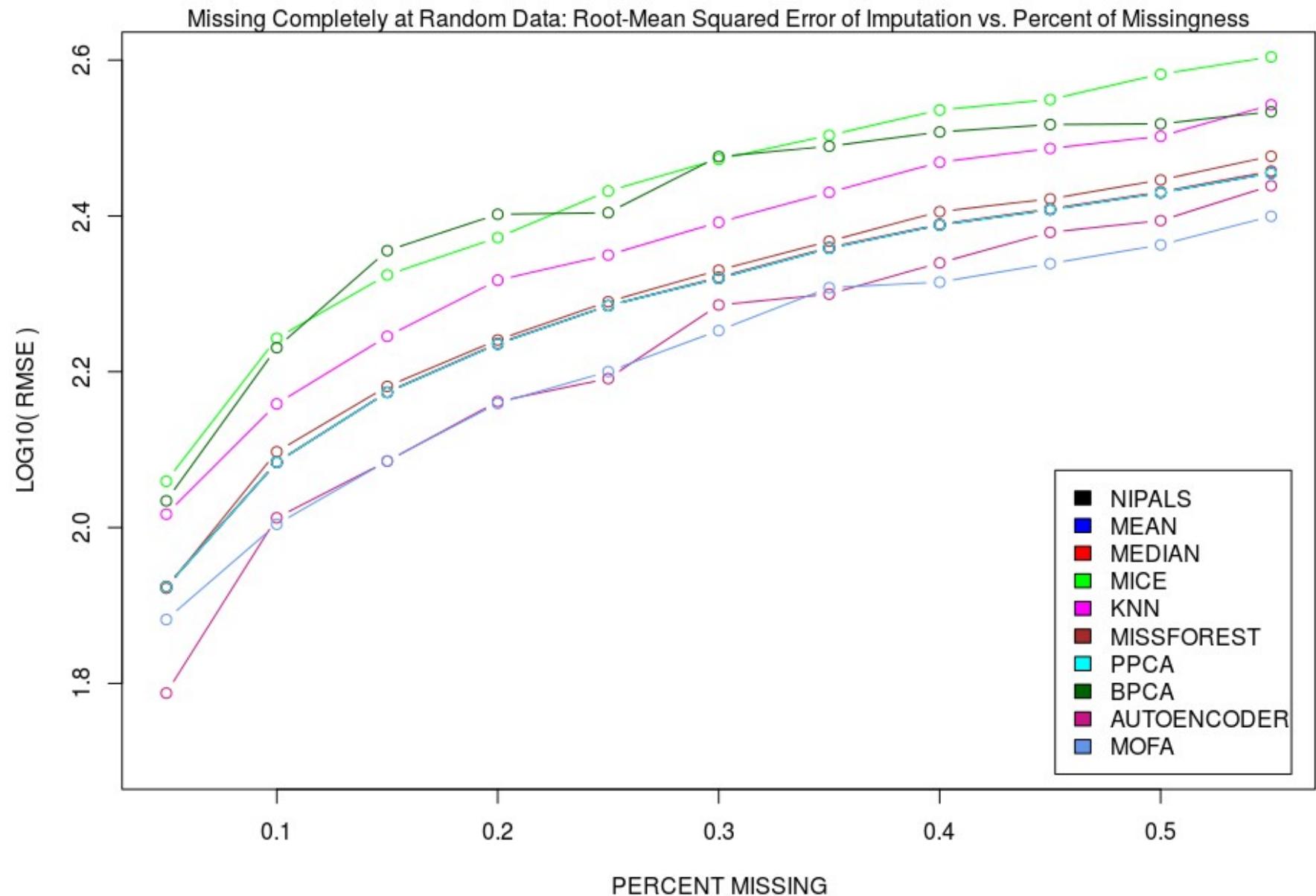
Values missing at random

**b**

Patients missing all measurements



Bayesian framework is insensitive to missing data, priors compensate for the lack of data





NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-03149-4

## ARTICLE

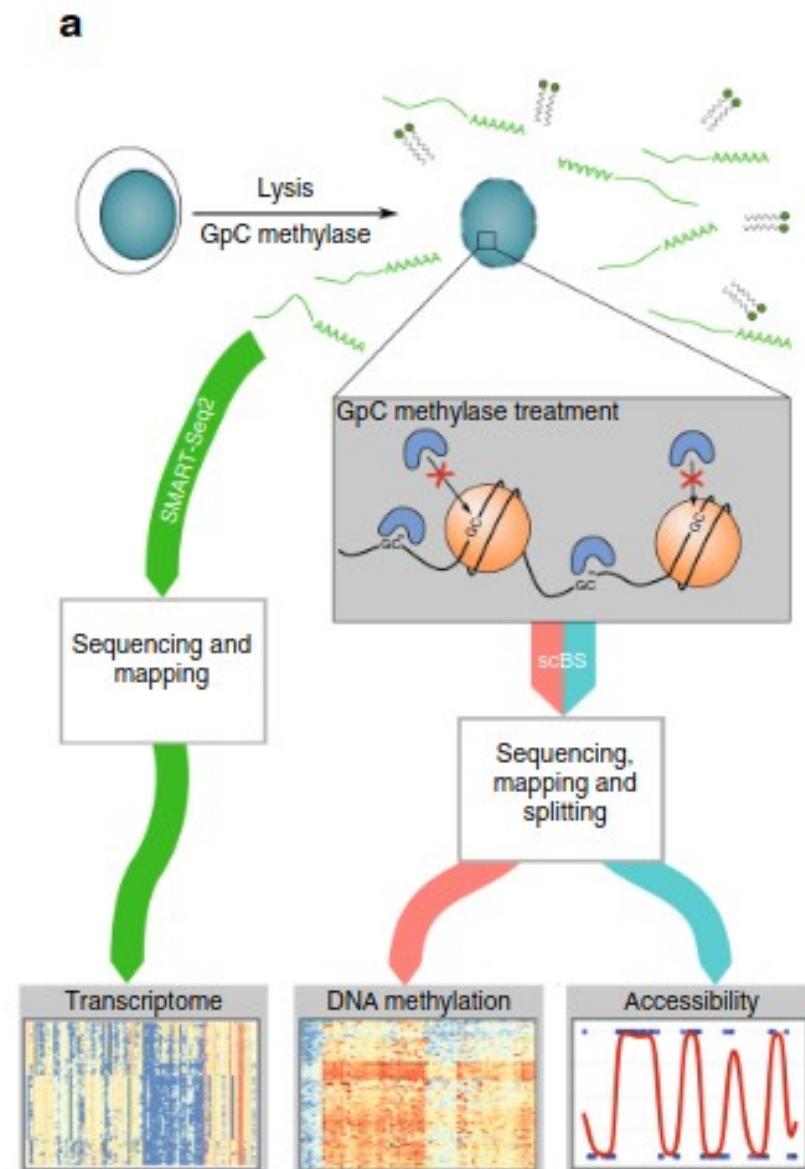
DOI: 10.1038/s41467-018-03149-4

OPEN

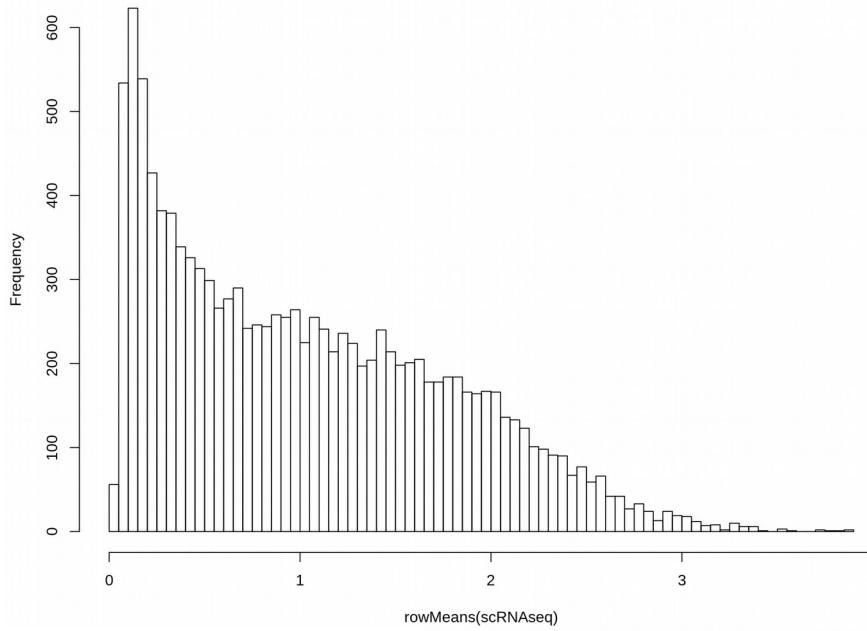
## scNMT-seq enables joint profiling of chromatin accessibility, DNA methylation and transcription in single cells

Stephen J. Clark<sup>1</sup>, Ricard Argelaguet<sup>2,3</sup>, Chantriolnt-Andreas Kapourani<sup>1</sup>, Thomas M. Stubbs<sup>1</sup>, Heather J. Lee<sup>1,5,6</sup>, Celia Alda-Catalinas<sup>1</sup>, Felix Krueger<sup>1</sup>, Guido Sanguinetti<sup>4</sup>, Gavin Kelsey<sup>1,8</sup>, John C. Marioni<sup>1,2,3,5</sup>, Oliver Stegle<sup>1</sup>, Wolf Reik<sup>1,5,8</sup>

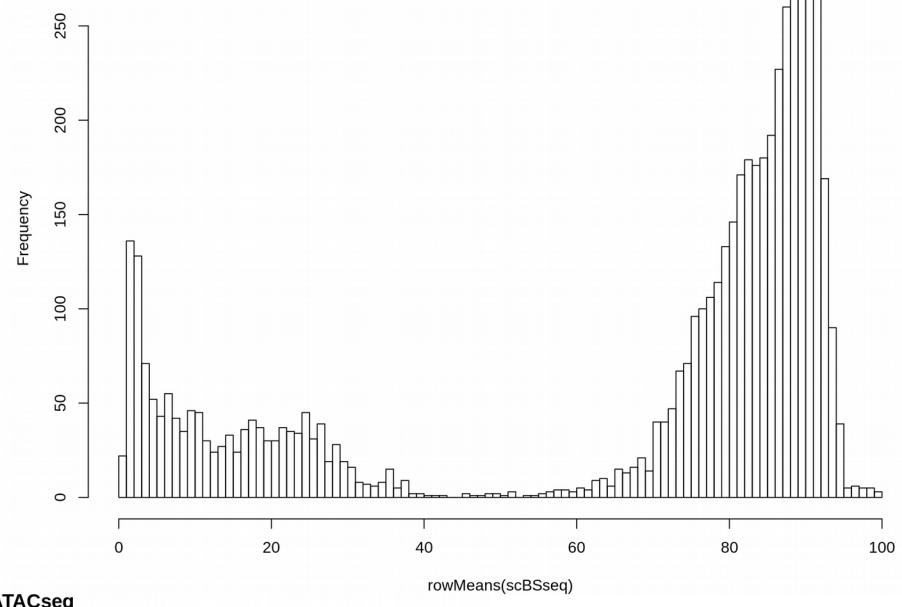
Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) uses a GpC methyltransferase to label open chromatin followed by bisulfite and RNA sequencing. We validate scNMT-seq by applying it to differentiating mouse embryonic stem cells, finding links between all three molecular layers and revealing dynamic coupling between epigenomic layers during differentiation.



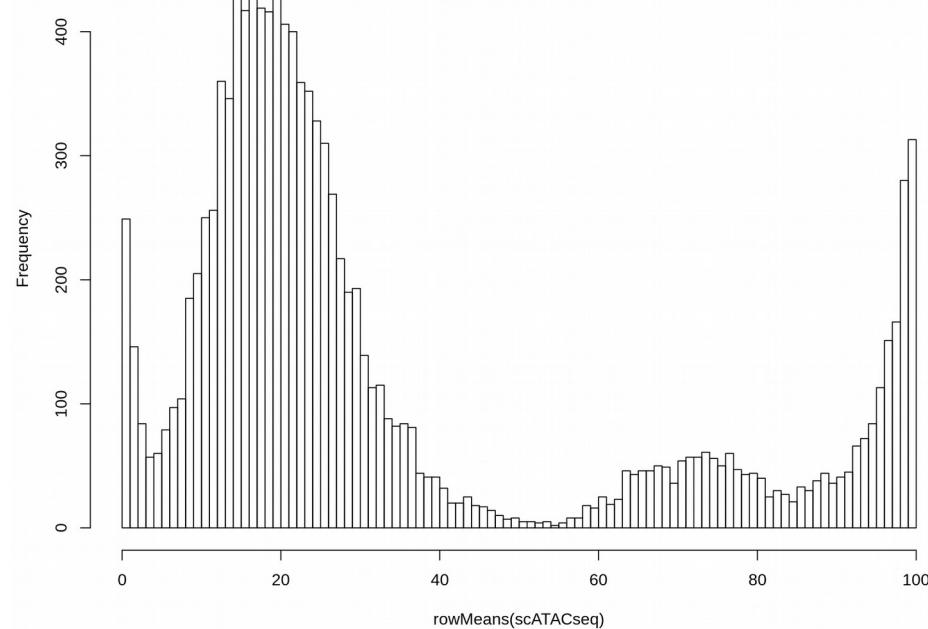
scRNAseq

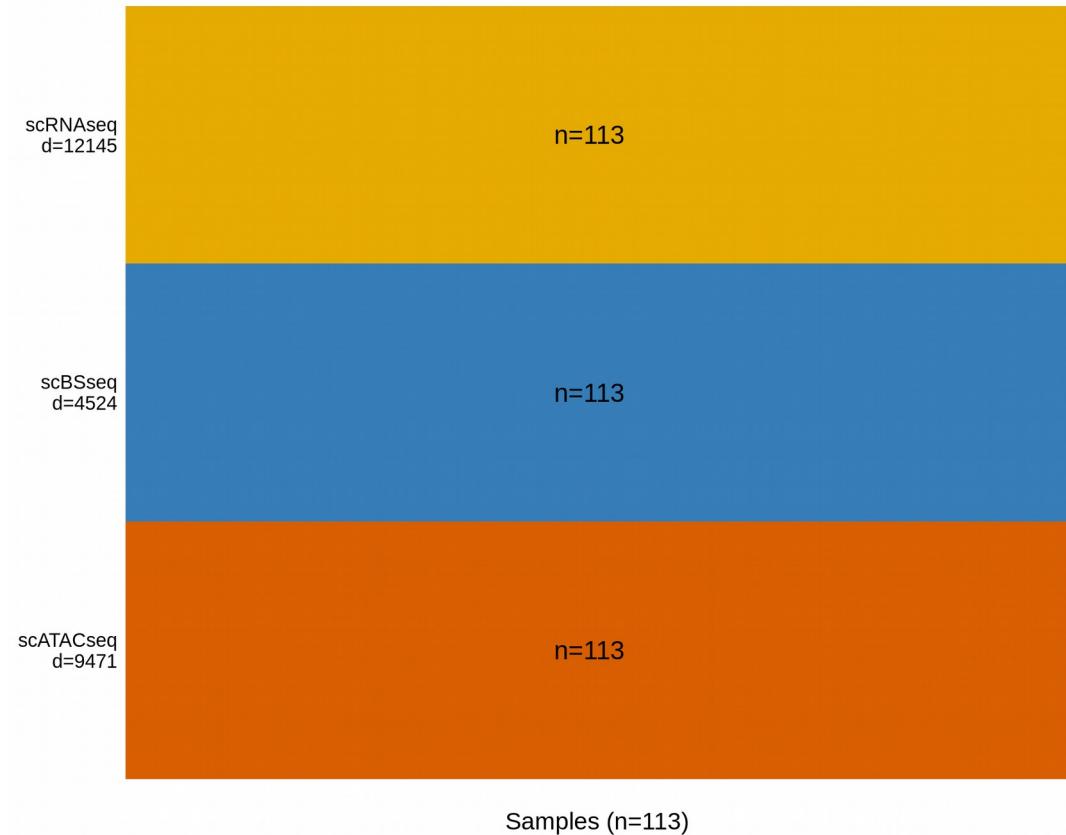


scBSseq



scATACseq



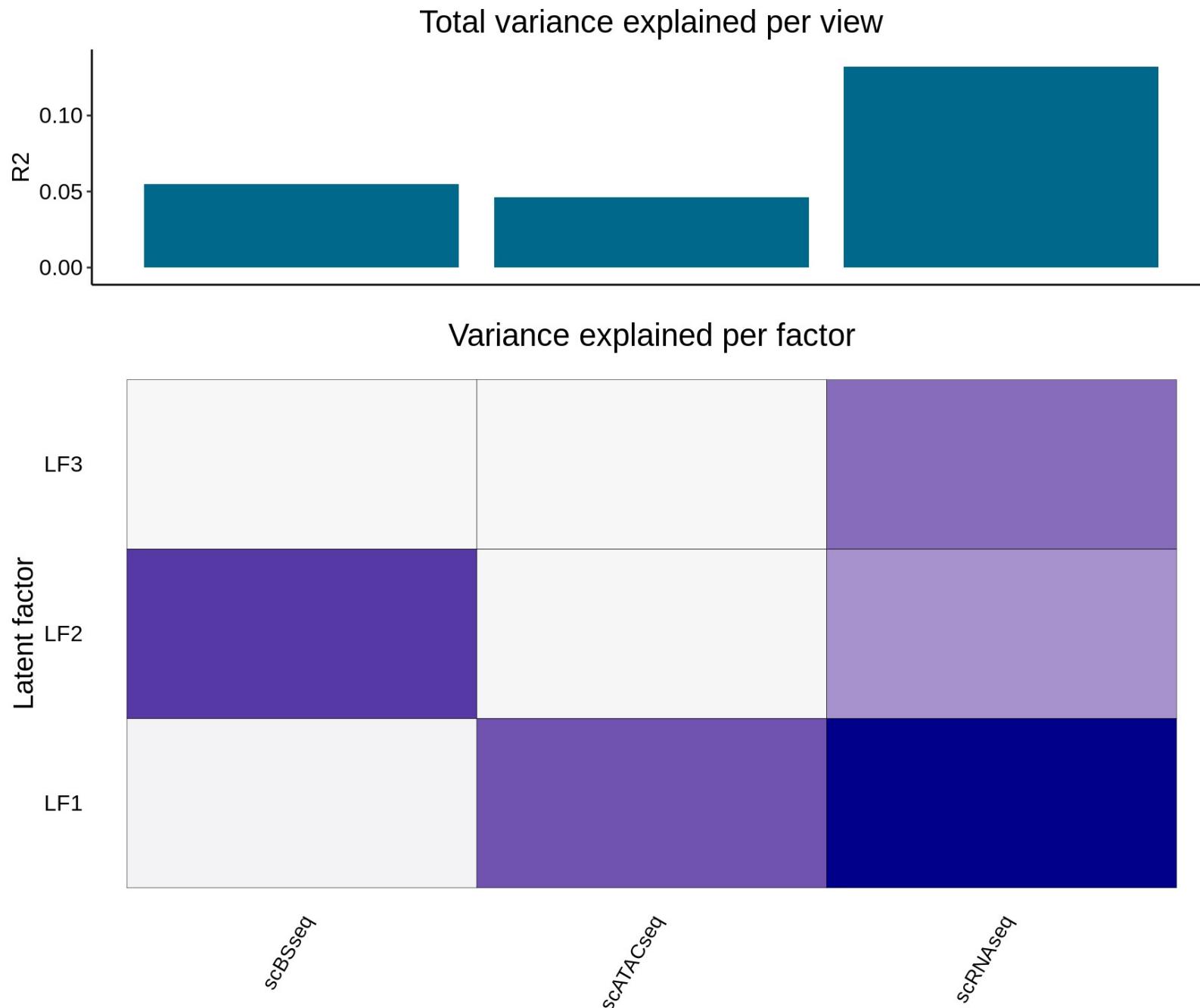


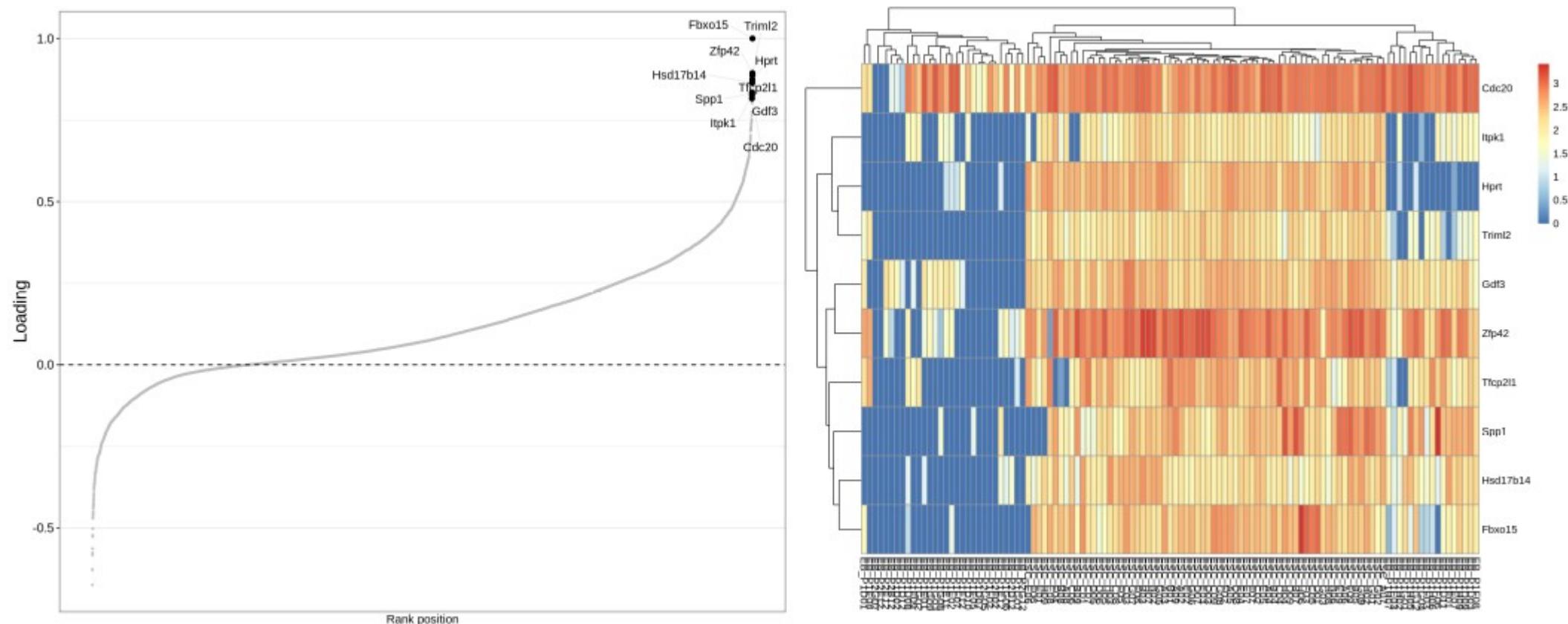
```
1 library("MOFA")
2 omics<-list(scRNaseq = scRNaseq, scBSSeq = scBSSeq, scATACseq = scATACseq)
3 MOFAobject <- createMOFAobject(omics)
4 plotDataOverview(MOFAobject)
5 DataOptions <- getDefaultDataOptions()
6 ModelOptions <- getDefaultModelOptions(MOFAobject)
7 mydistr <- c("gaussian", "bernoulli", "bernoulli")
8 names(mydistr) <- c("scRNaseq", "scBSSeq", "scATACseq")
9 ModelOptions$likelihood <- mydistr
10 ModelOptions$numFactors <- 20
11 TrainOptions <- getDefaultTrainOptions()
12 TrainOptions$seed <- 2018
13 # Automatically drop factors that explain less than 3% of variance in all omics
14 TrainOptions$DropFactorThreshold <- 0.03
15 TrainOptions$tolerance <- 0.1; TrainOptions$maxiter <- 1000
```

Prepare\_MOFA.R hosted with ❤ by GitHub

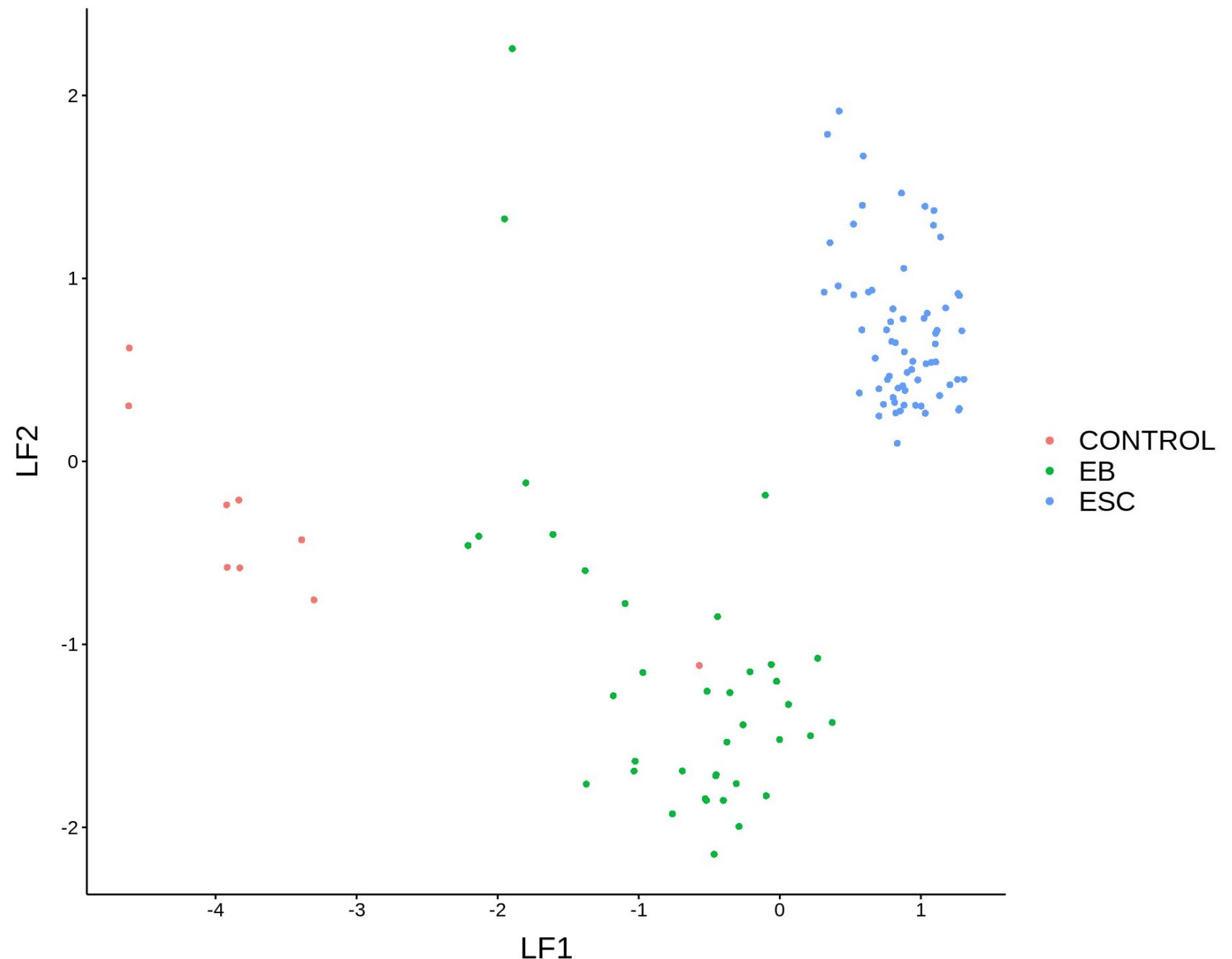
[view raw](#)

Bayesian framework of MOFA allows to explicitly model non-Gaussian distributions via Bayes rule

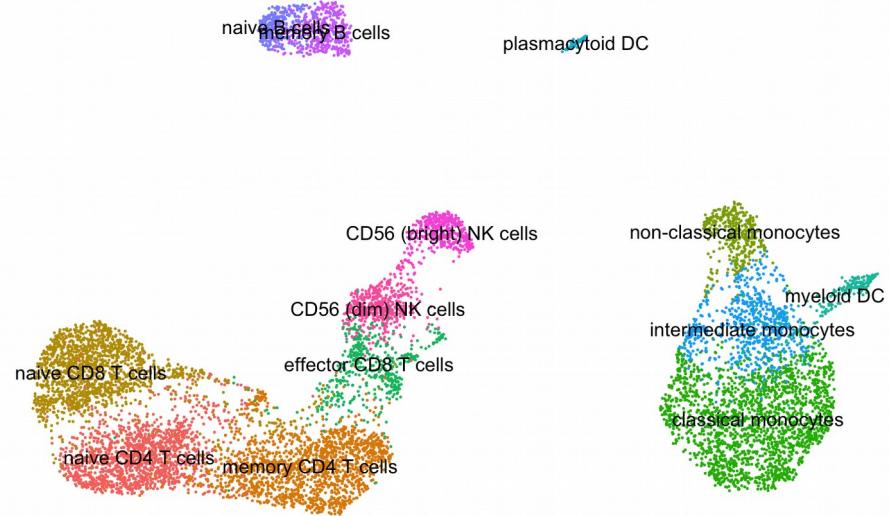




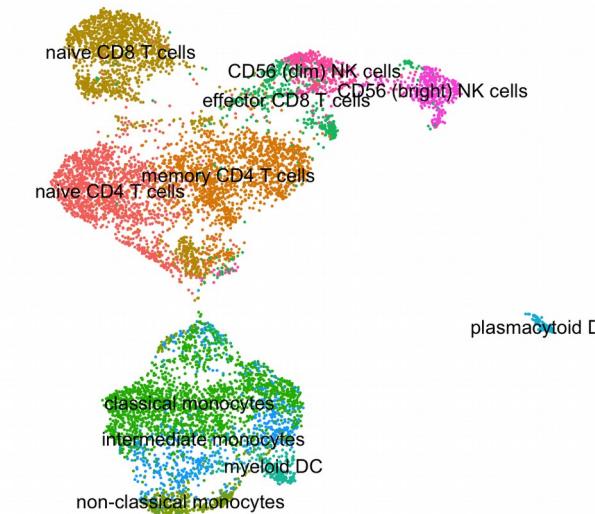
ESC and EB cells are separable on the heatmap built on loadings of the MOFA latent factors



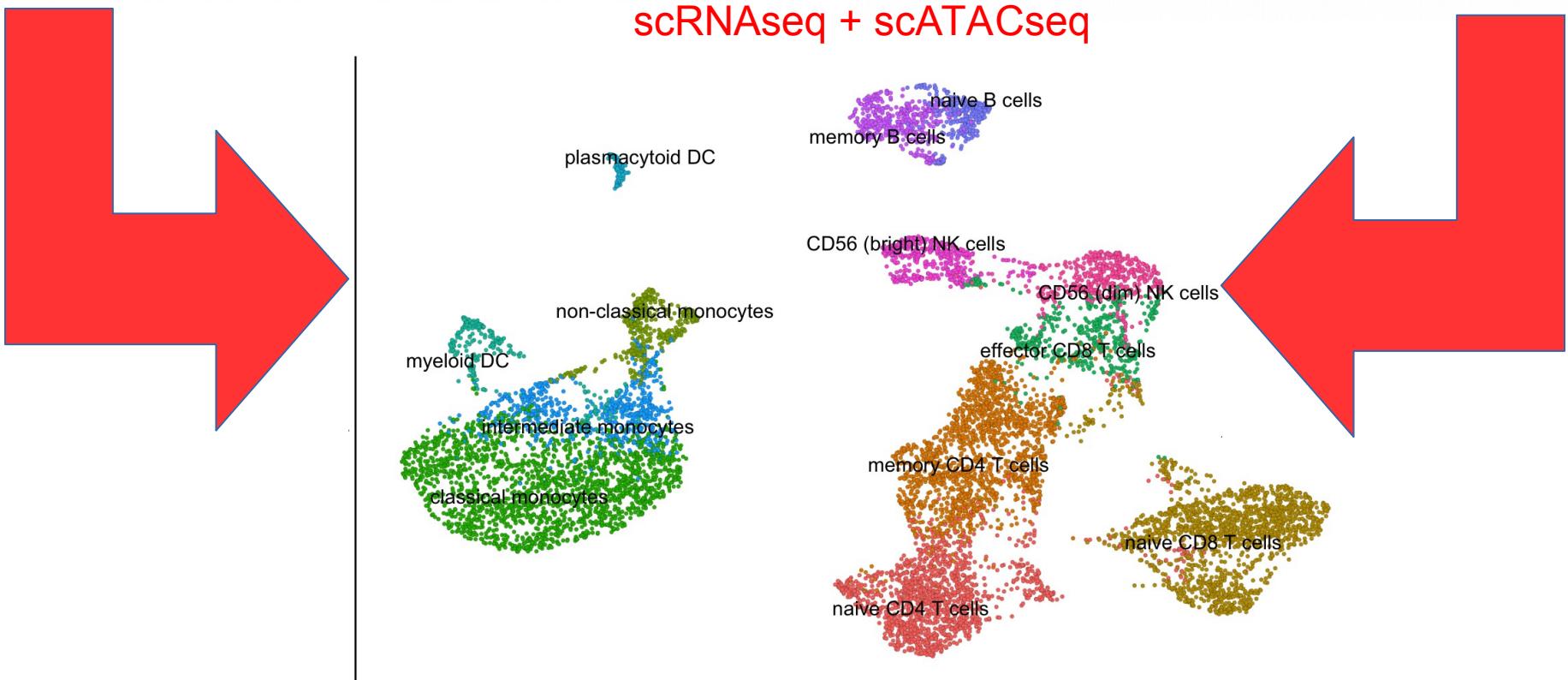
## scRNAseq

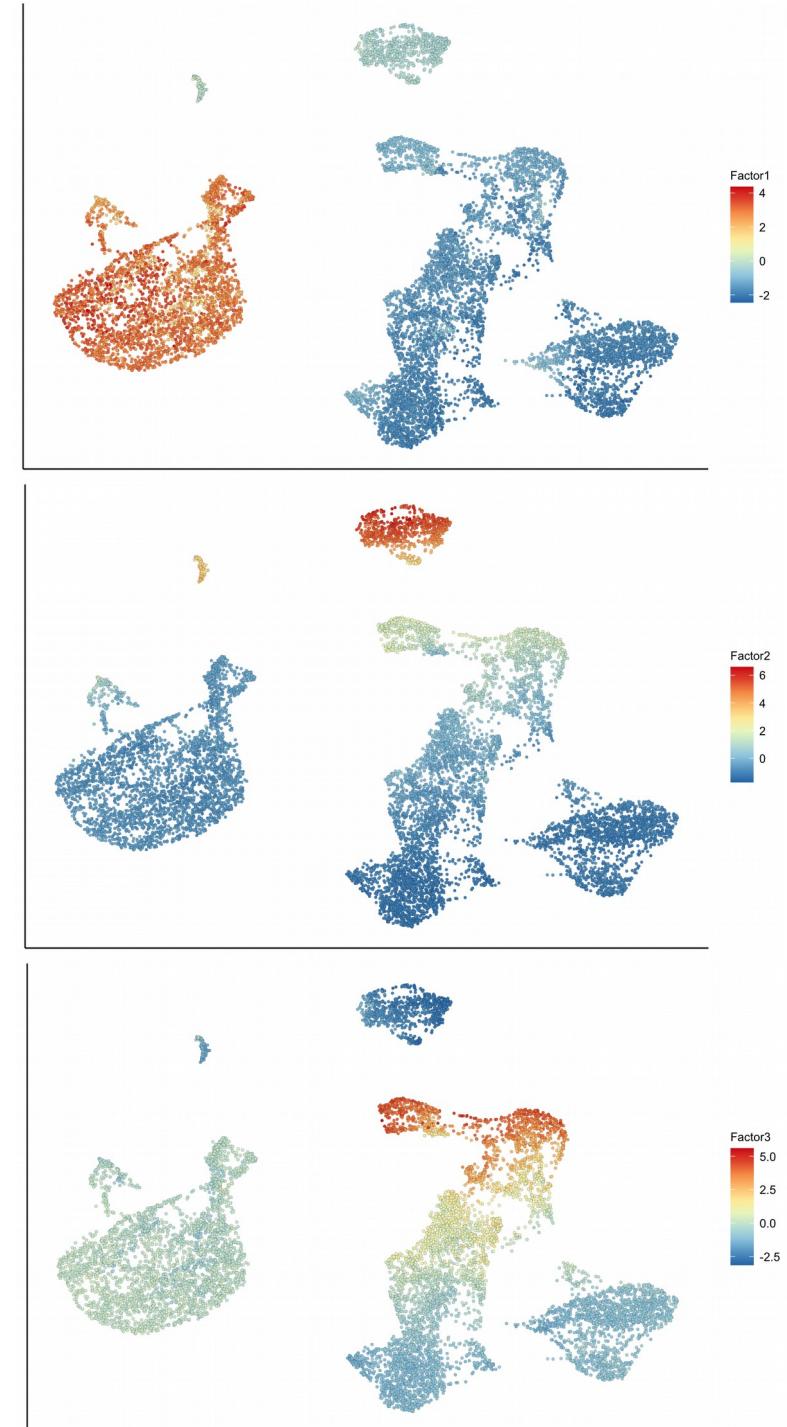
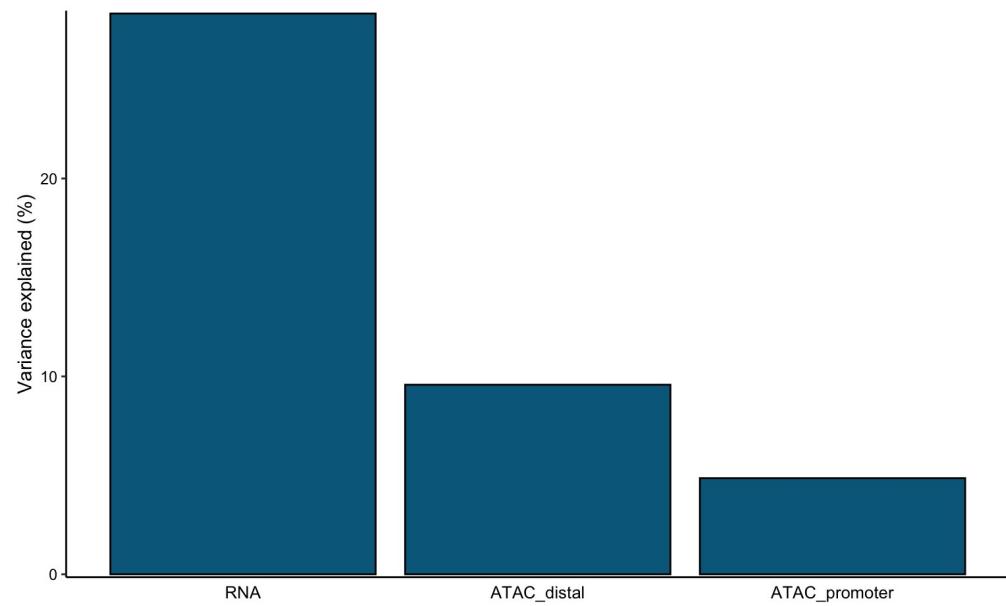


## scATACseq

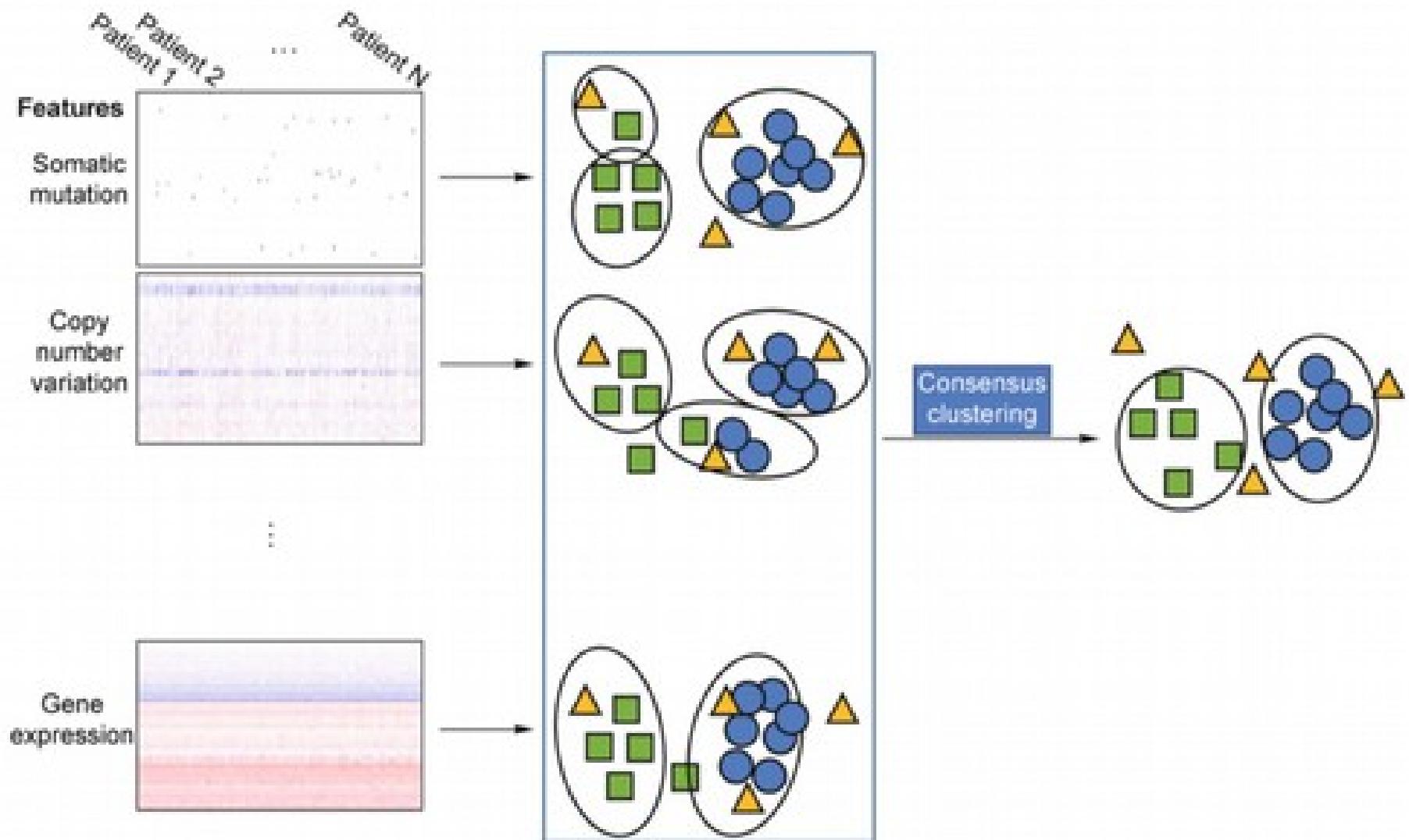


## scRNAseq + scATACseq

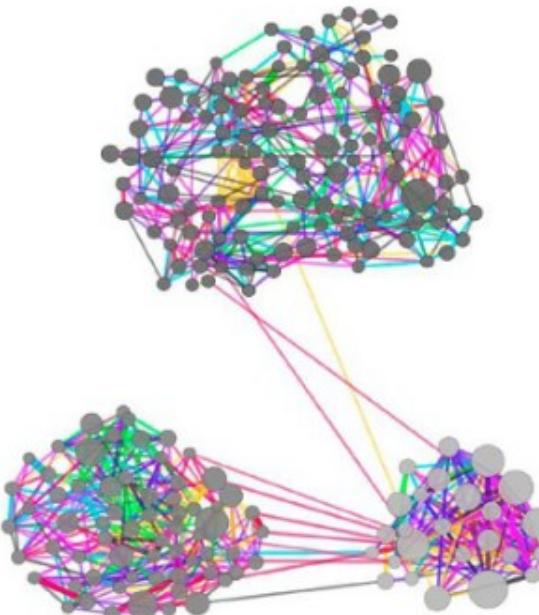
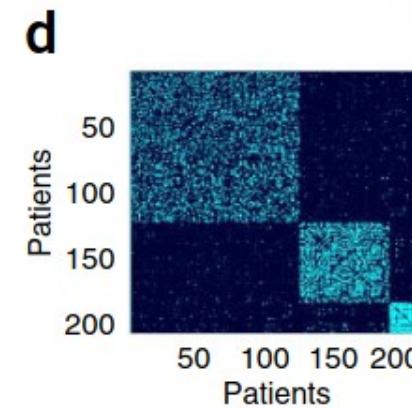
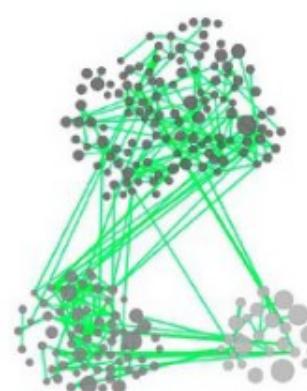
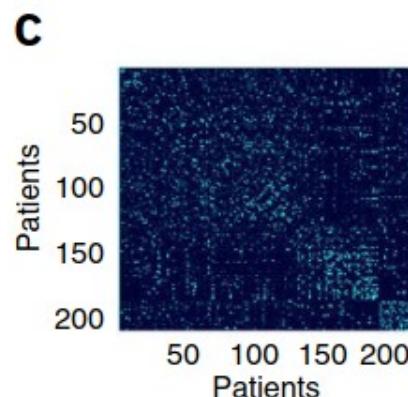
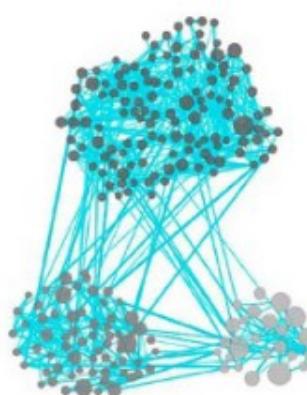
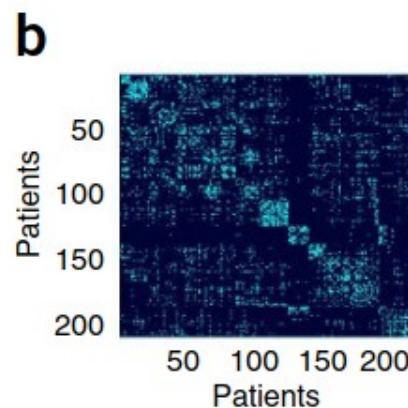
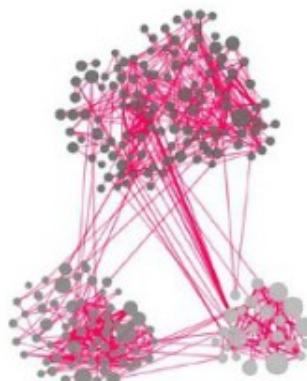
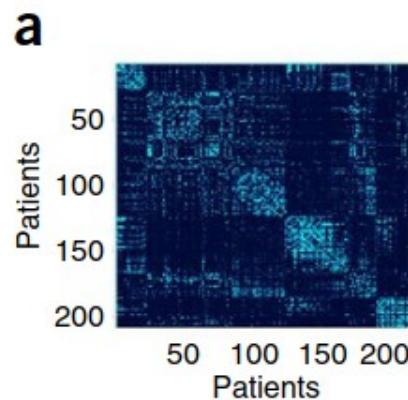


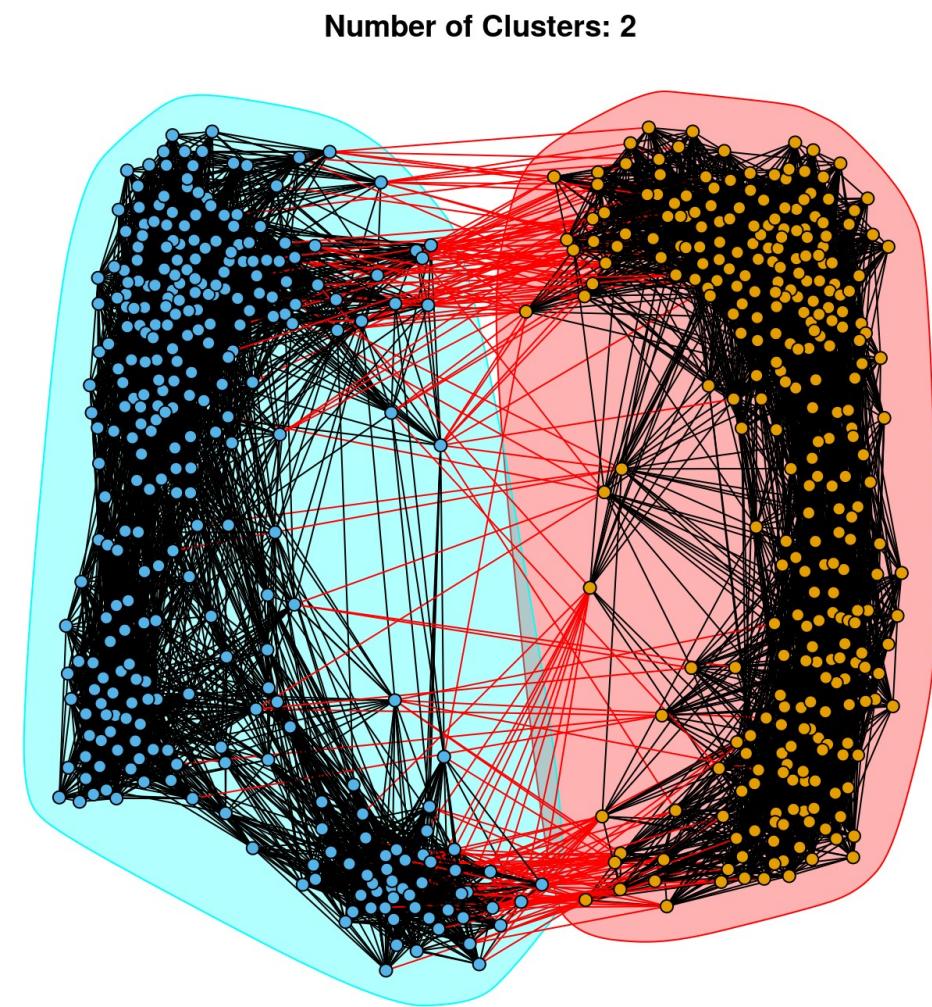
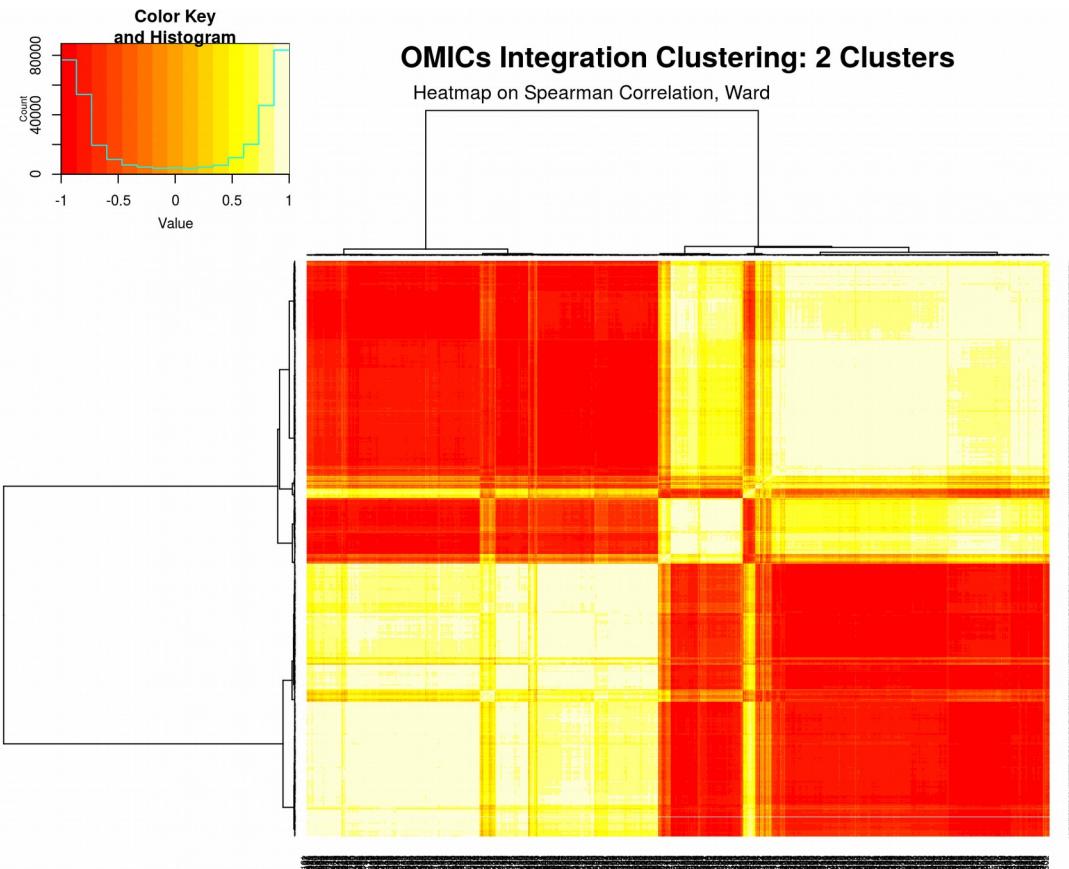


# Other Unsupervised Integrative OMICs Methods

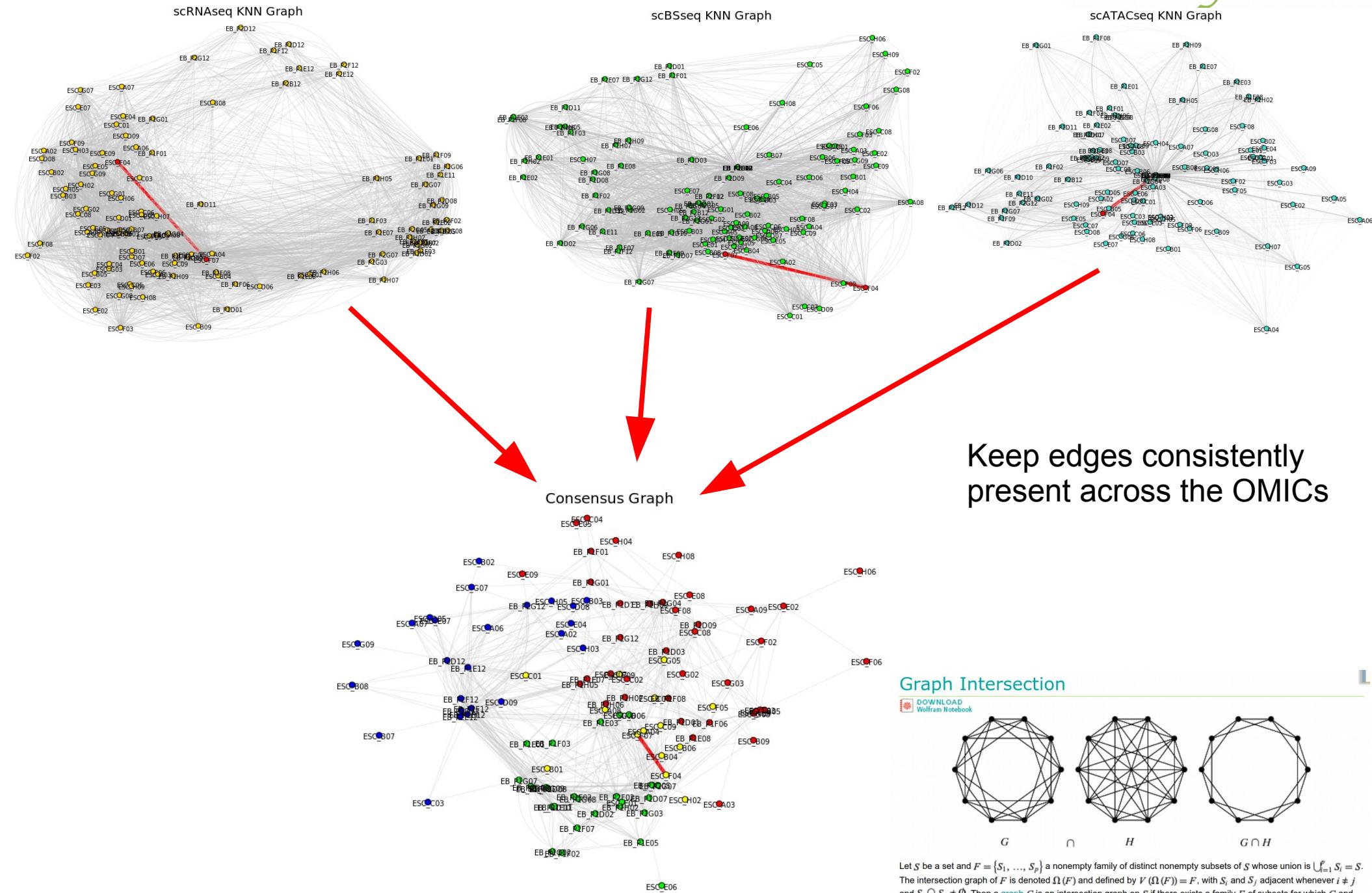


**Figure 2. Clustering of clusters.** This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.





Instead of OMICs datasets construct pairwise adjacency matrices where elements are fractions each pair of samples belongs to the same cluster. Next we average the adjacency matrices, i.e. find samples that are consistently together across all OMICs





# National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab



*Knut och Alice  
Wallenbergs  
Stiftelse*



**LUNDS**  
UNIVERSITET