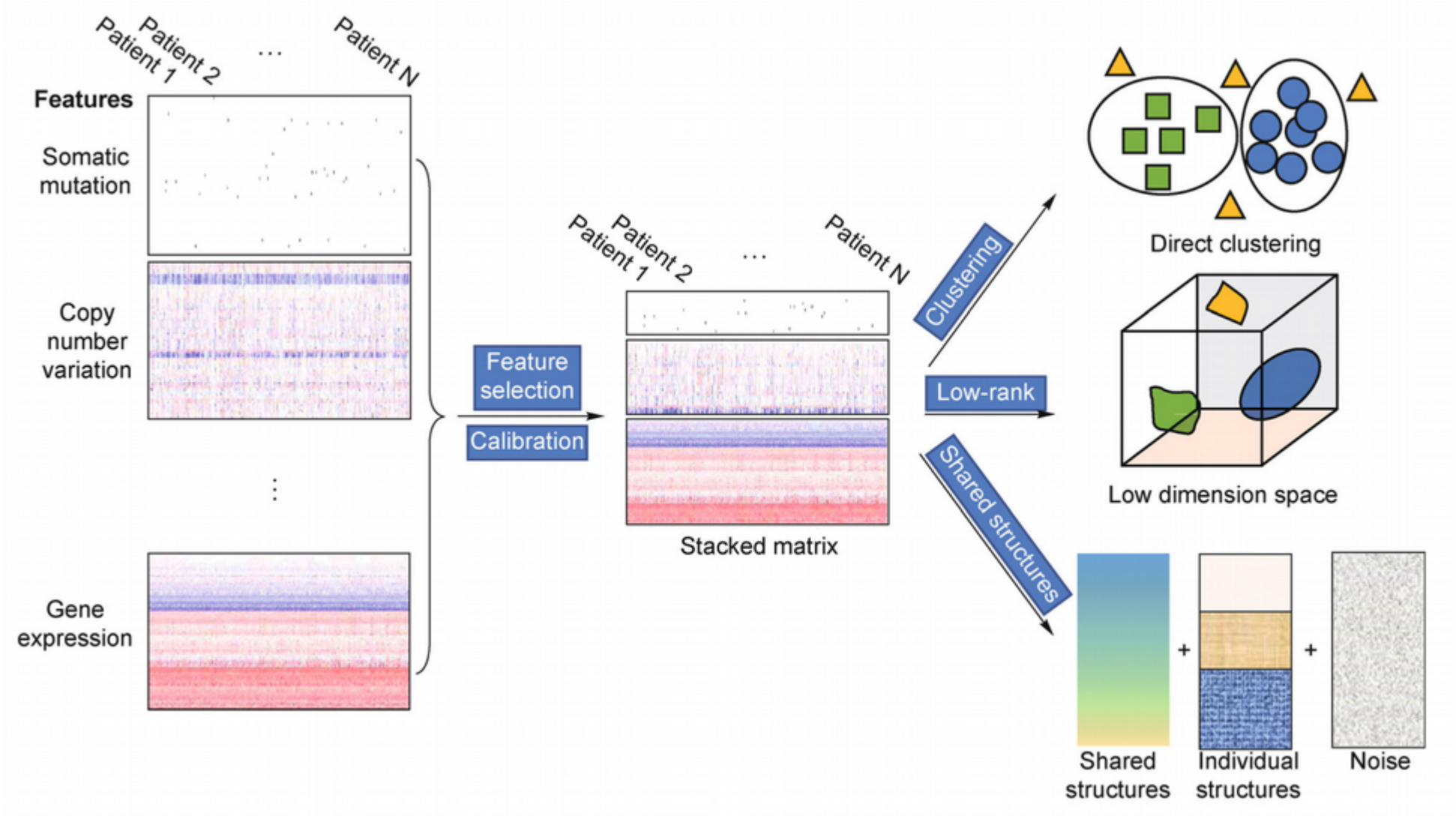
Unsupervised OMICs Integration

OMICs Integration and Systems Biology course

Nikolay Oskolkov, NBIS SciLifeLab

Lund, 5.10.2020



Supervised vs. Unsupervised Analysis

Supervised Analysis: Commercial

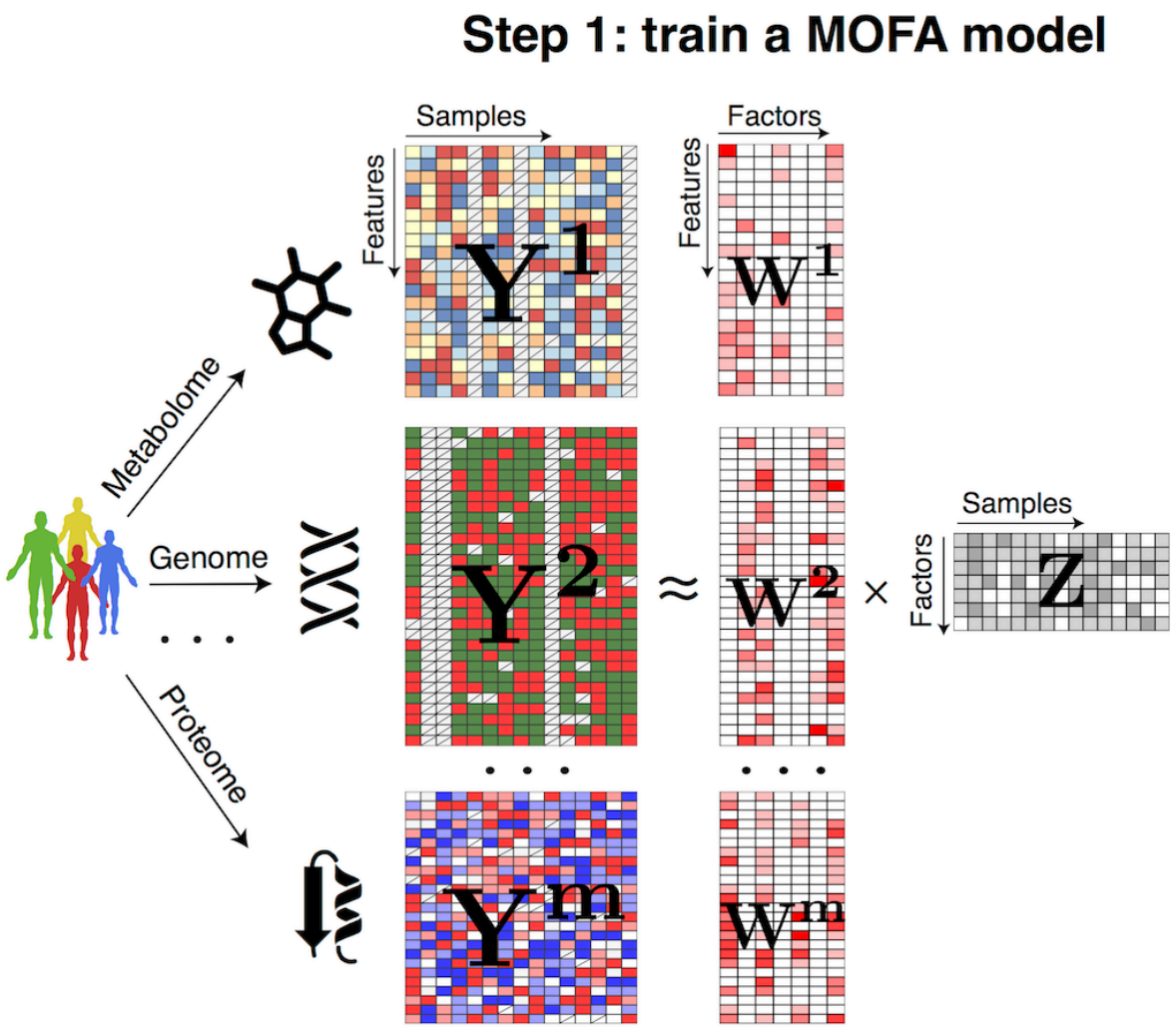
Hypothesis-driven approach
Mainstream research
Guaranteed pubs and funds
Small leap in development

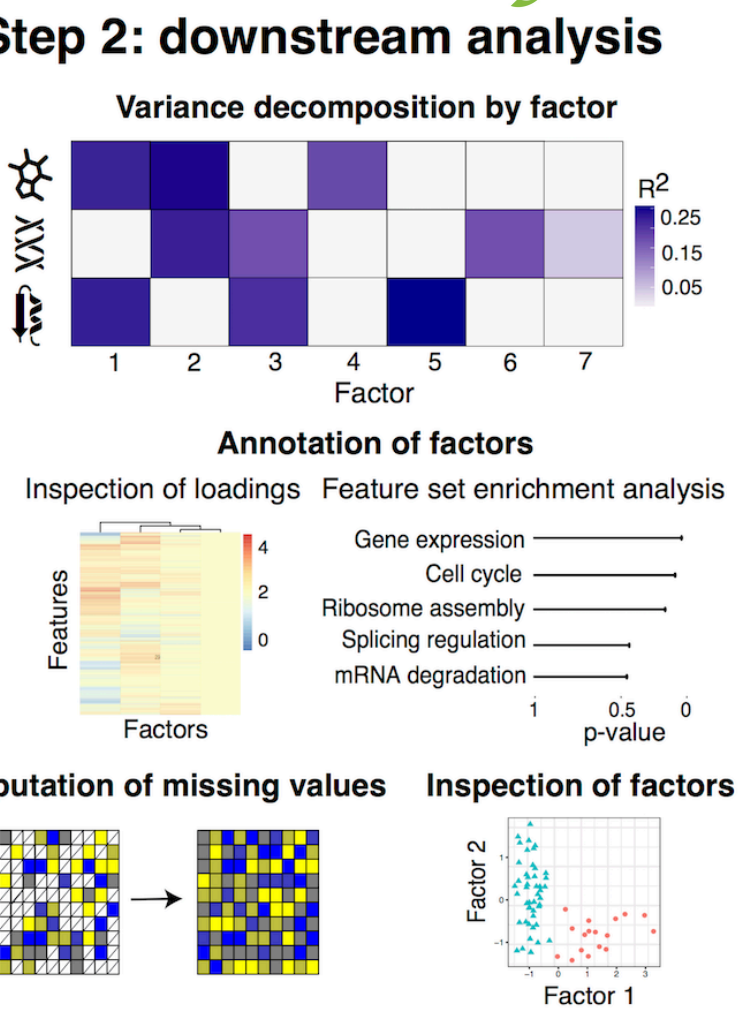
How can you solve a scientific problem if you do not know the answer?

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

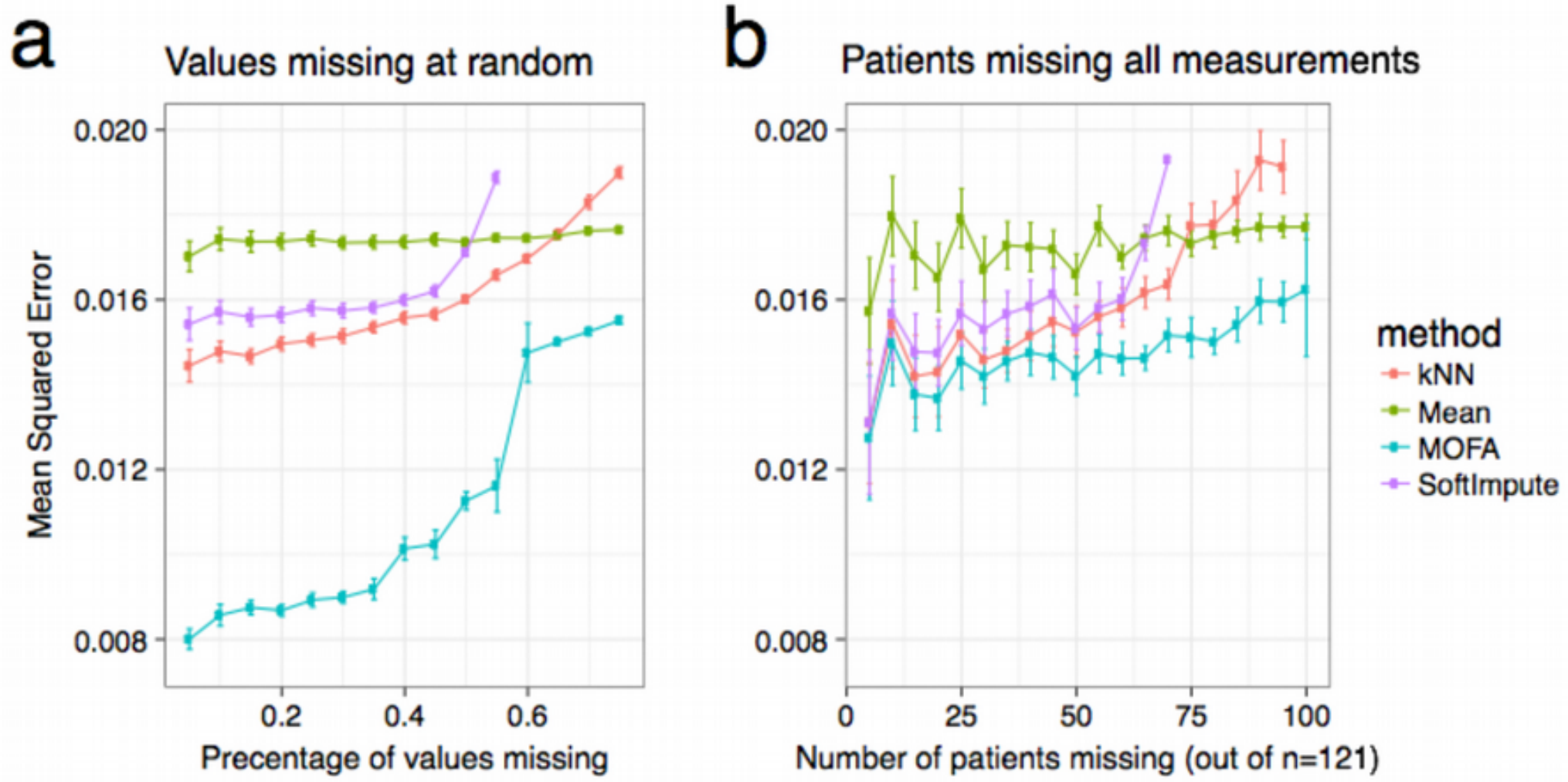
MOFA: Hypothesis-Free Exploratory Tool





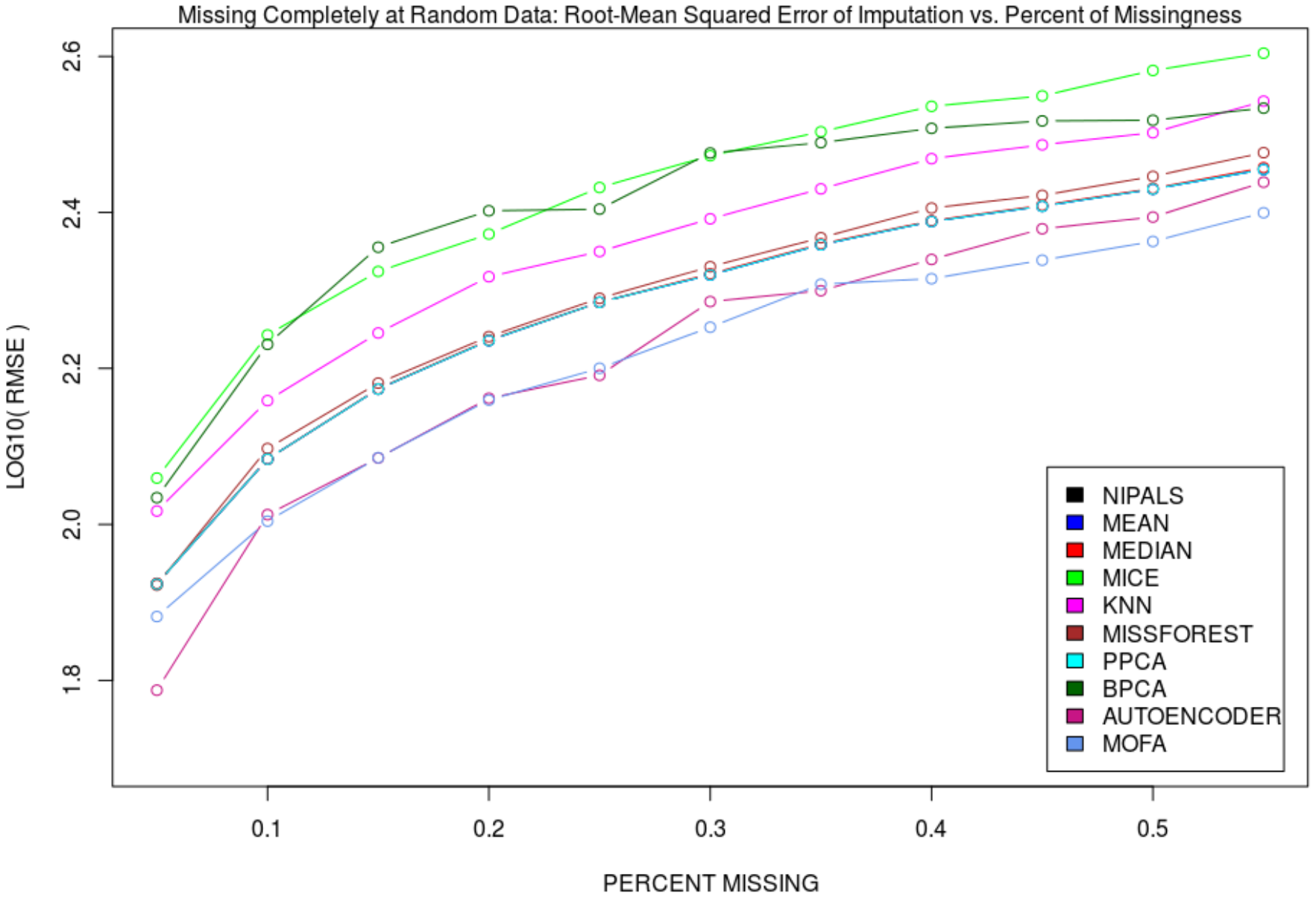
- Visualisation of samples in factor space
- Annotation of factors using (gene set) enrichment analysis
- Imputation of missing values
- Support of OMICs with non-Gaussian distribution including binary and count data

MOFA: Imputation of Missing Values



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

MOFA: Imputation of Missing Values



scNMT Data Set: scRNAseq + scBSseq + scATACseq

nature COMMUNICATIONS

ATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-031

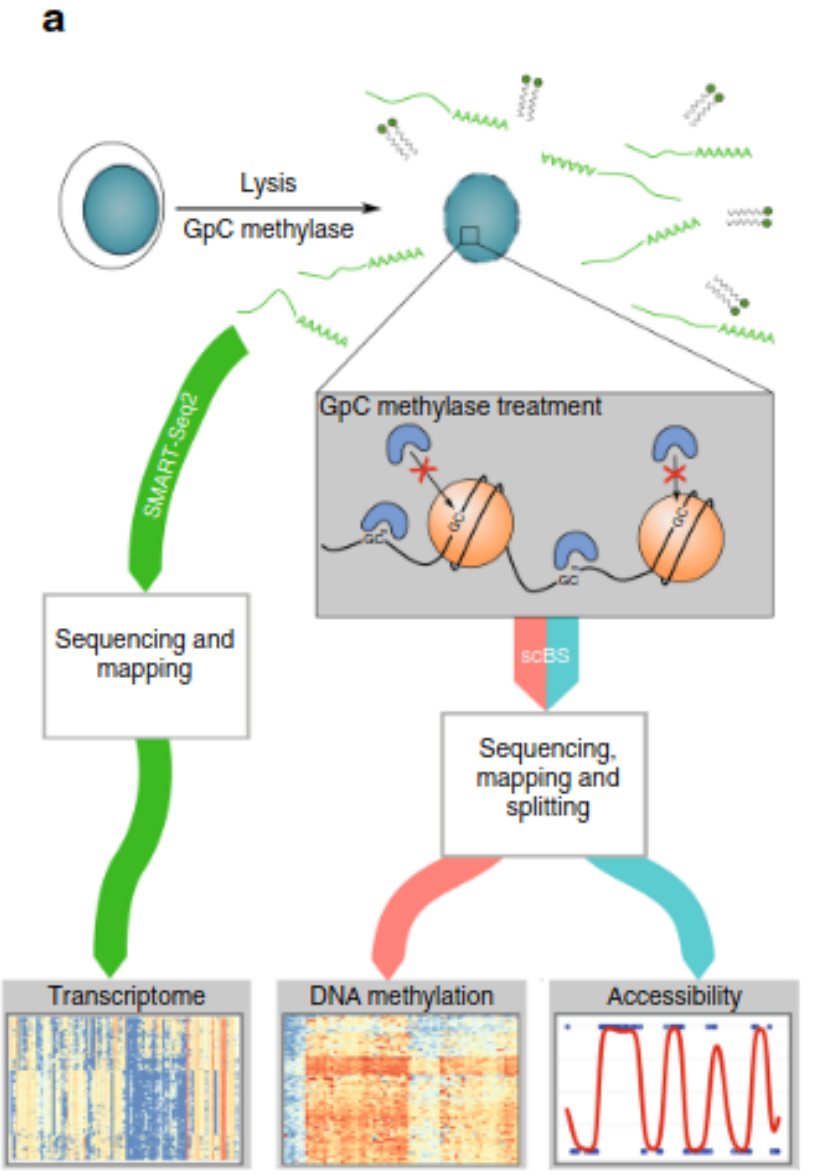
ARTICLE

DOI: 10.1038/s41467-018-03149-4 OPEN

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

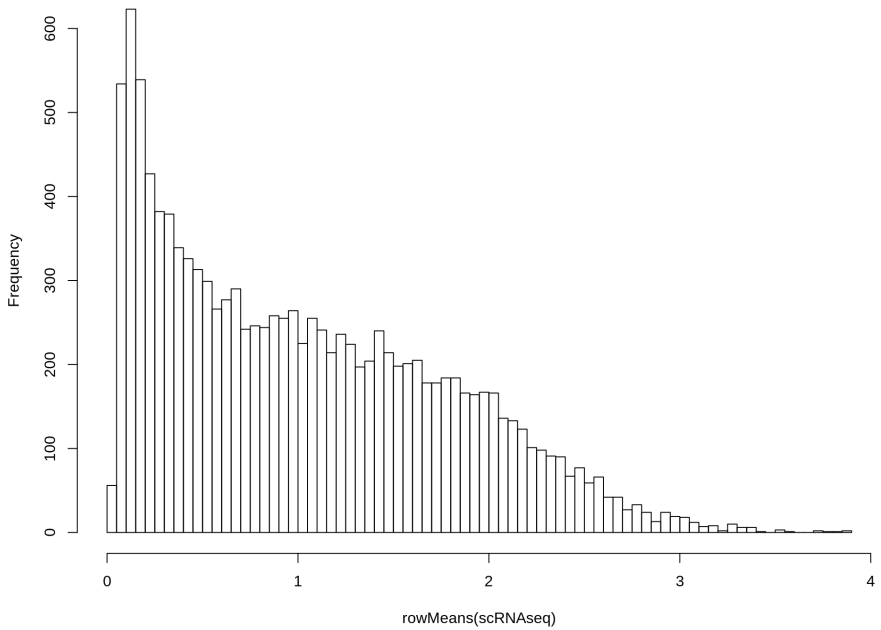
Stephen J. Clark[1], Ricard Argelaguet[2,3], ChantriolInt-Andreas Kapourani[4] Thomas M. Stubbs[1], Heather J. Lee[1,5,6], Celia Alda-Catalinas[1], Felix Krueger[7] Guido Sanguinetti[4], Gavin Kelsey[1,8] John C. Marioni[2,3,5] Oliver Stegle[2] Wolf Reik[1,5,8]

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.

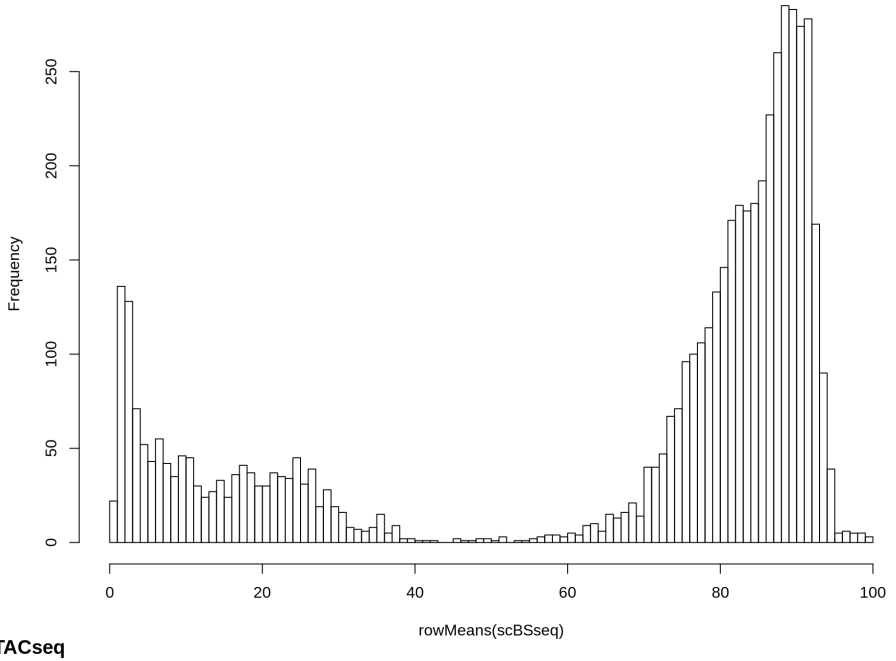


scNMT Data Set: Distributions

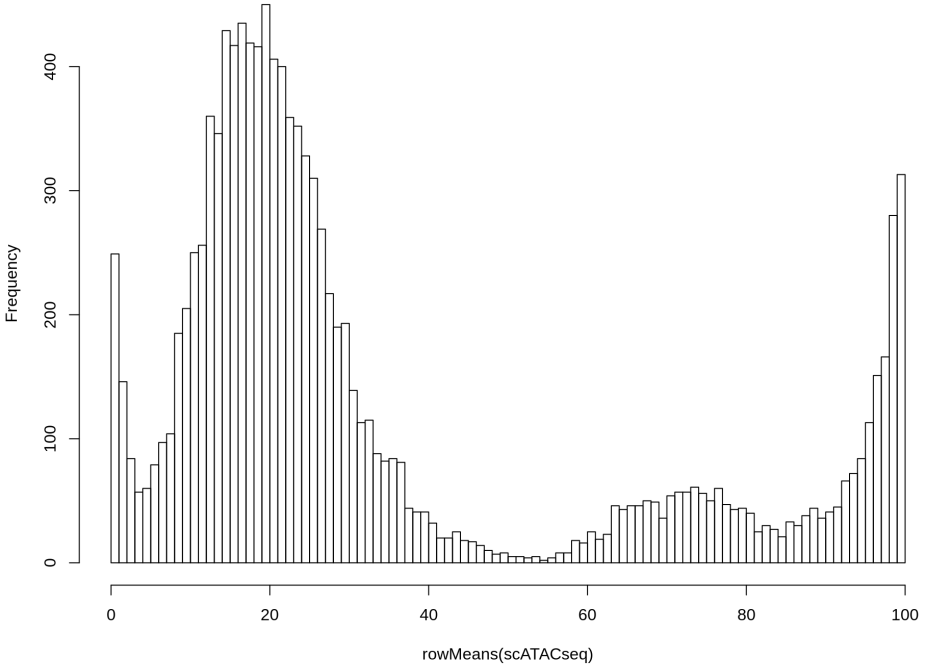
scRNAseq



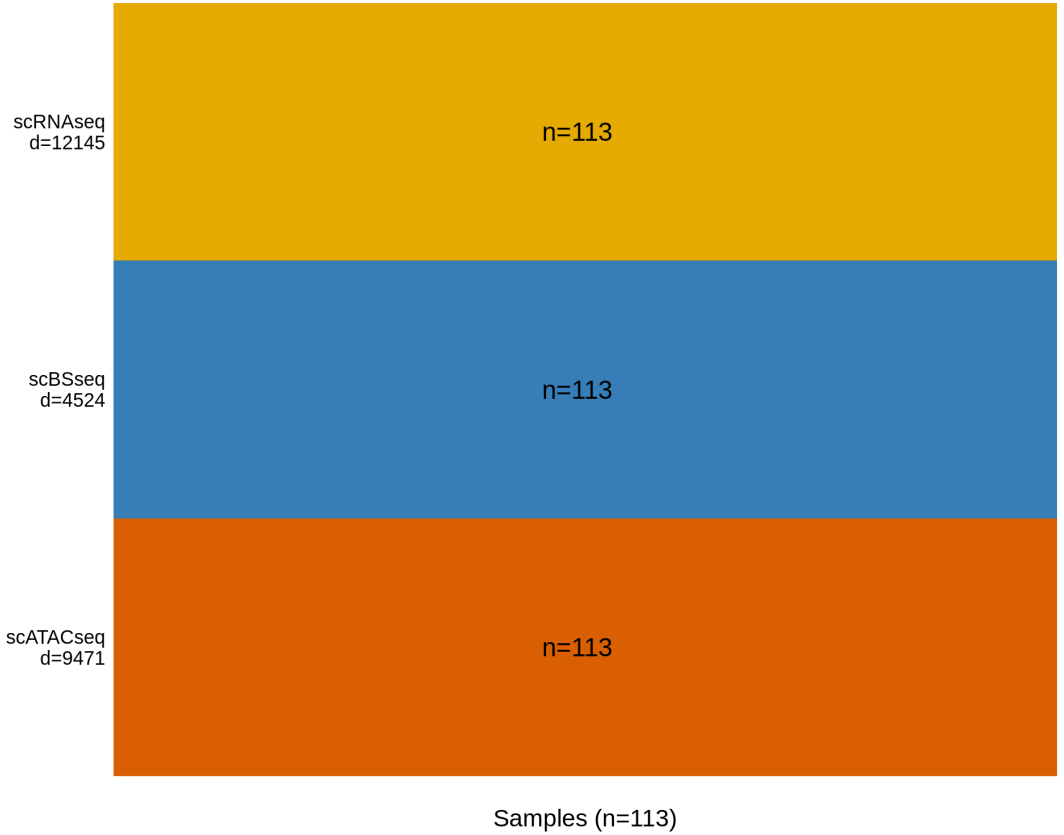
scBSseq



scATACseq



scNMT Data Set: Summary Stats

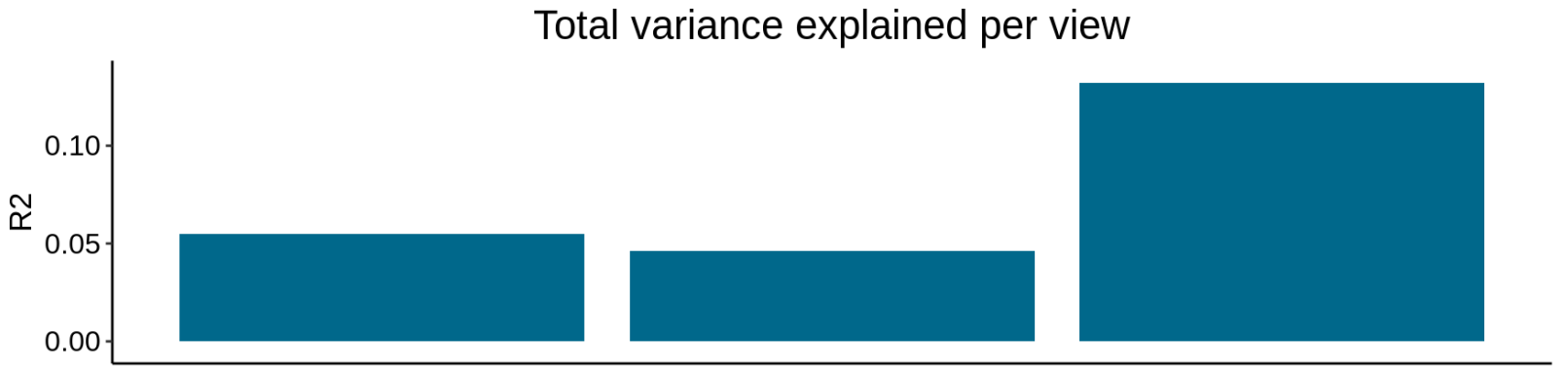


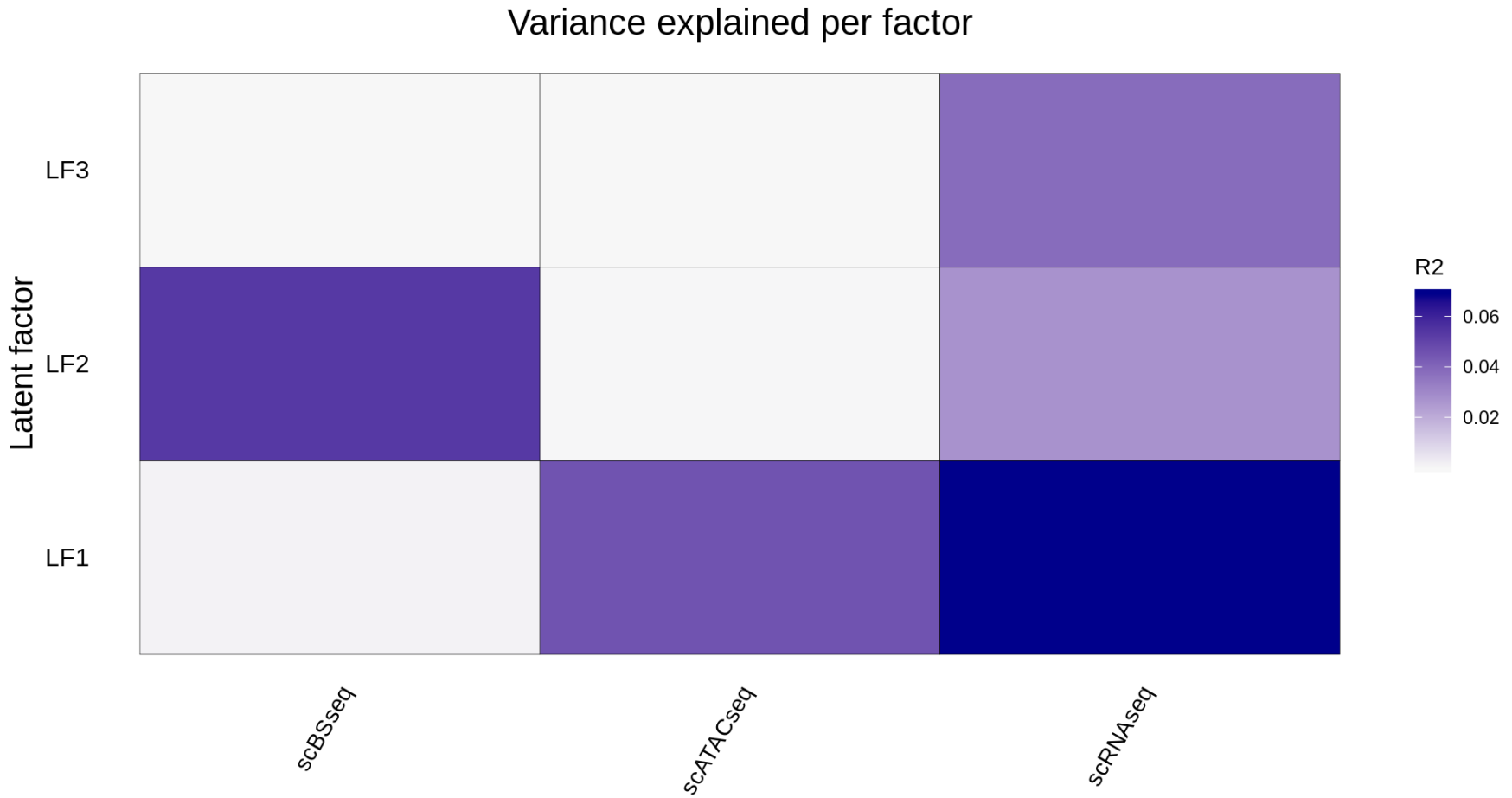

```r
library("MOFA")
omics<-list(scRNAseq = scRNAseq, scBSseq = scBSseq, scATACseq = scATACseq)
MOFAobject <- createMOFAobject(omics)
plotDataOverview(MOFAobject)
DataOptions <- getDefaultDataOptions()
ModelOptions <- getDefaultModelOptions(MOFAobject)
mydistr <- c("gaussian", "bernoulli", "bernoulli")
names(mydistr) <- c("scRNAseq", "scBSseq", "scATACseq")
ModelOptions$likelihood <- mydistr
ModelOptions$numFactors <- 20
TrainOptions <- getDefaultTrainOptions()
TrainOptions$seed <- 2018
# Automatically drop factors that explain less than 3% of variance in all omics
TrainOptions$DropFactorThreshold <- 0.03
TrainOptions$tolerance <- 0.1; TrainOptions$maxiter <- 1000
```

Prepare_MOFA.R hosted with ❤ by GitHub view raw

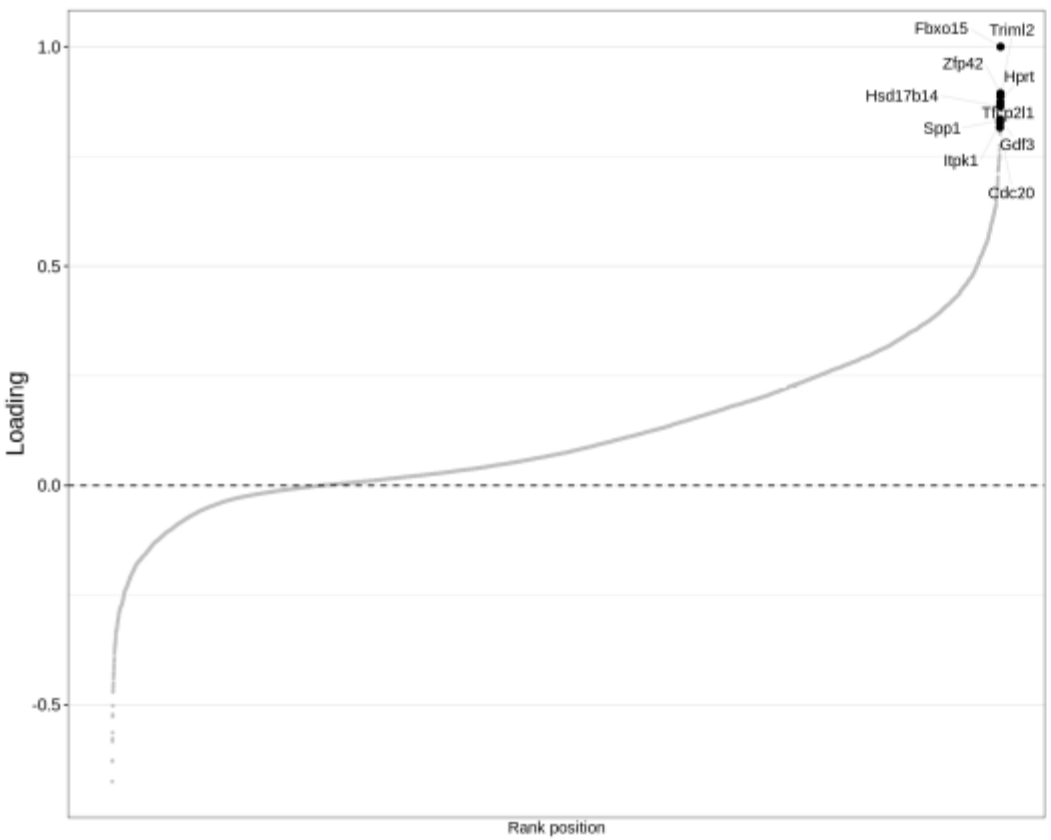
Bayesian framework of MOFA allows to explicitly model non-Gaussian distributions via Bayes rule

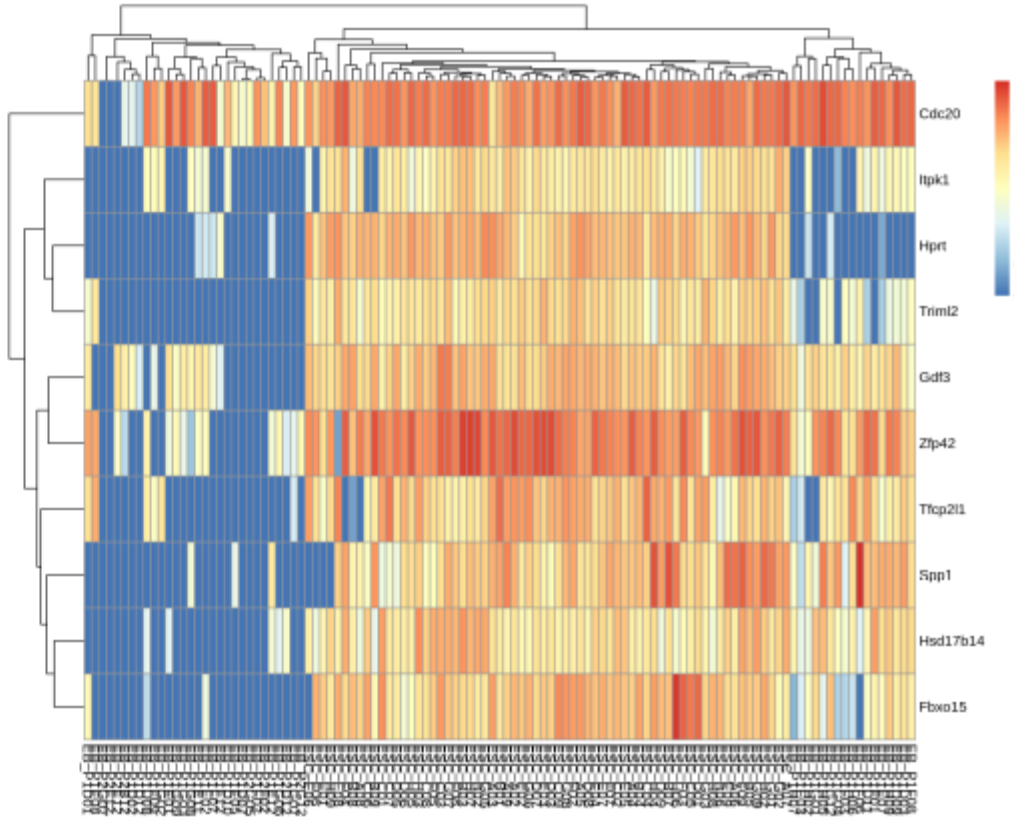
scNMT Data Set: Variance Explained





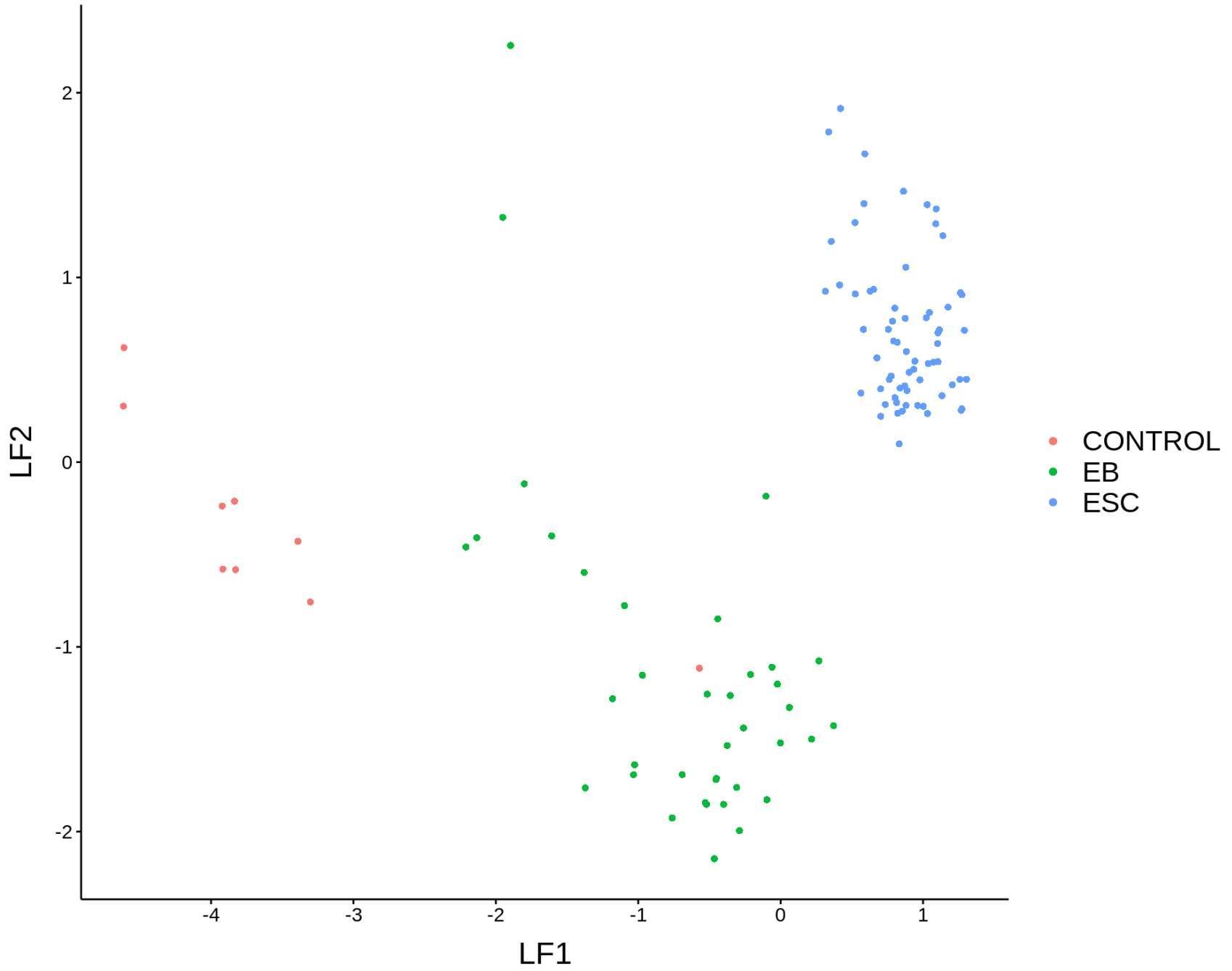
MOFA: Interpretation of Factors





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

MOFA: Visualizing Low-Dimensional Space



Clustering of Clusters

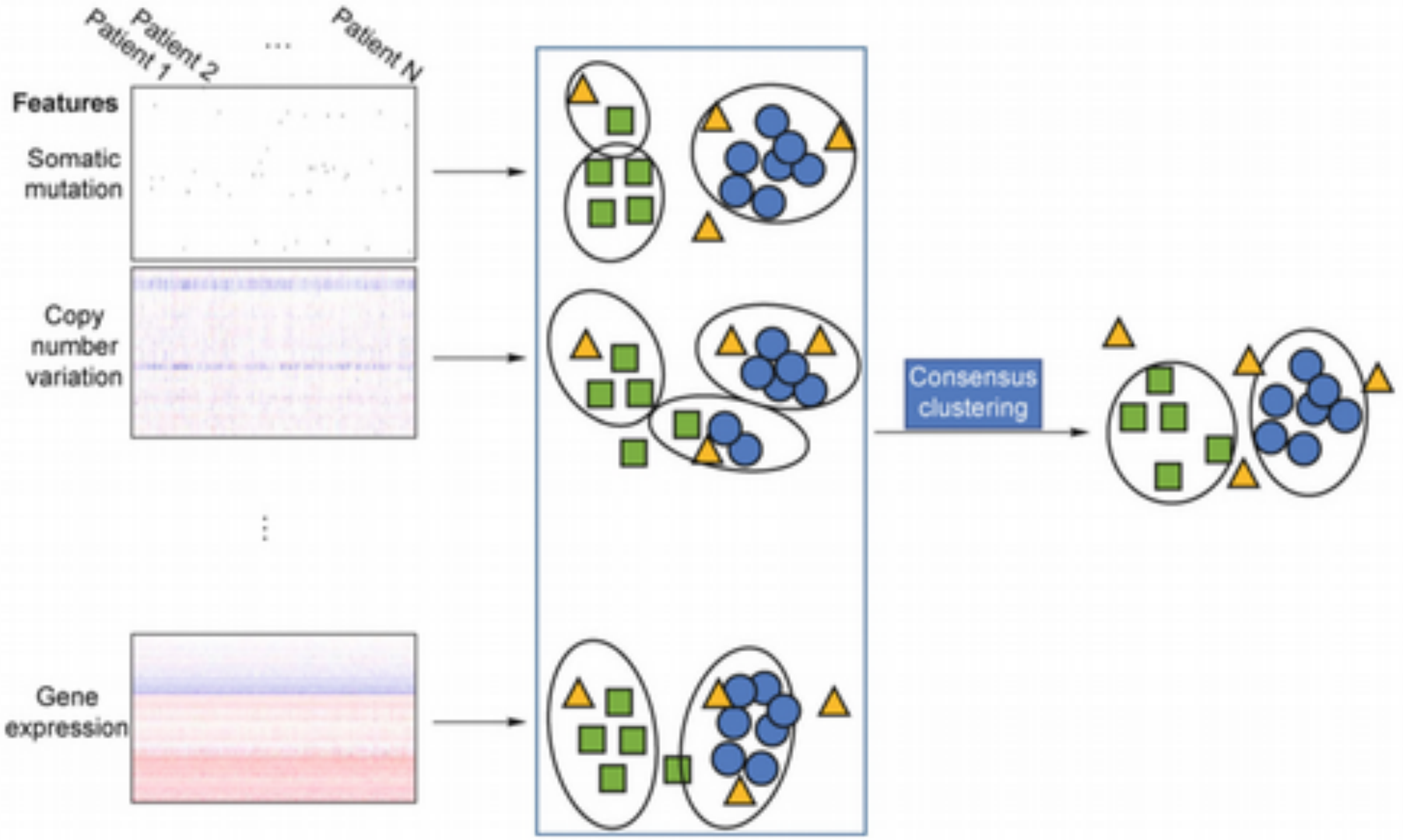
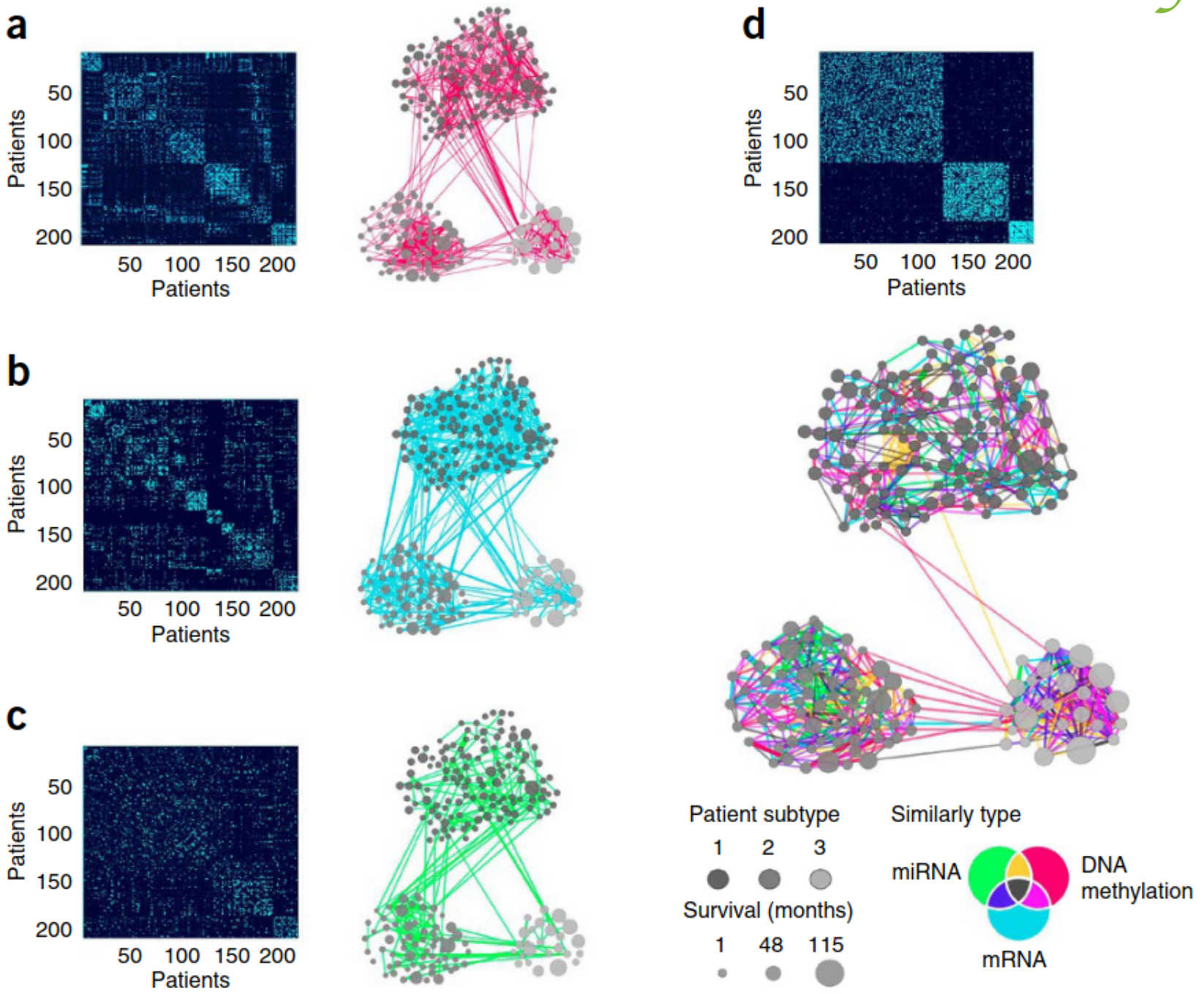
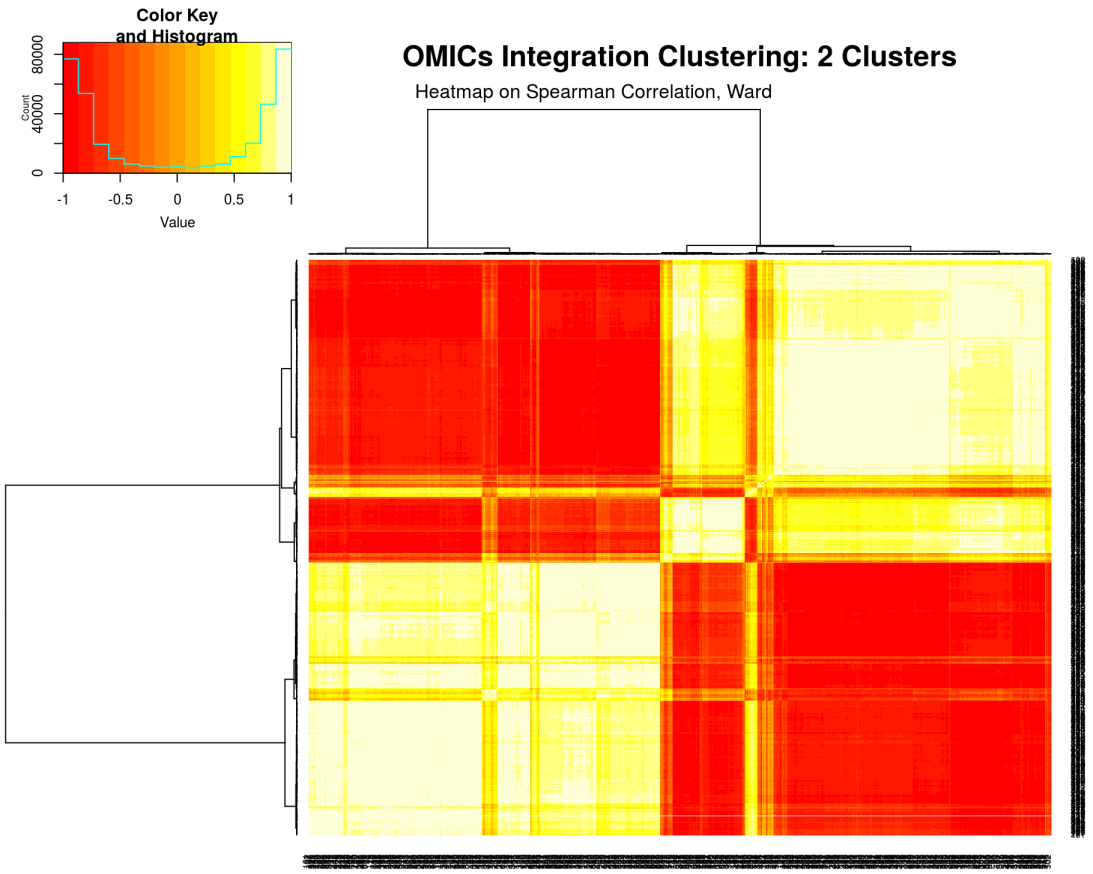


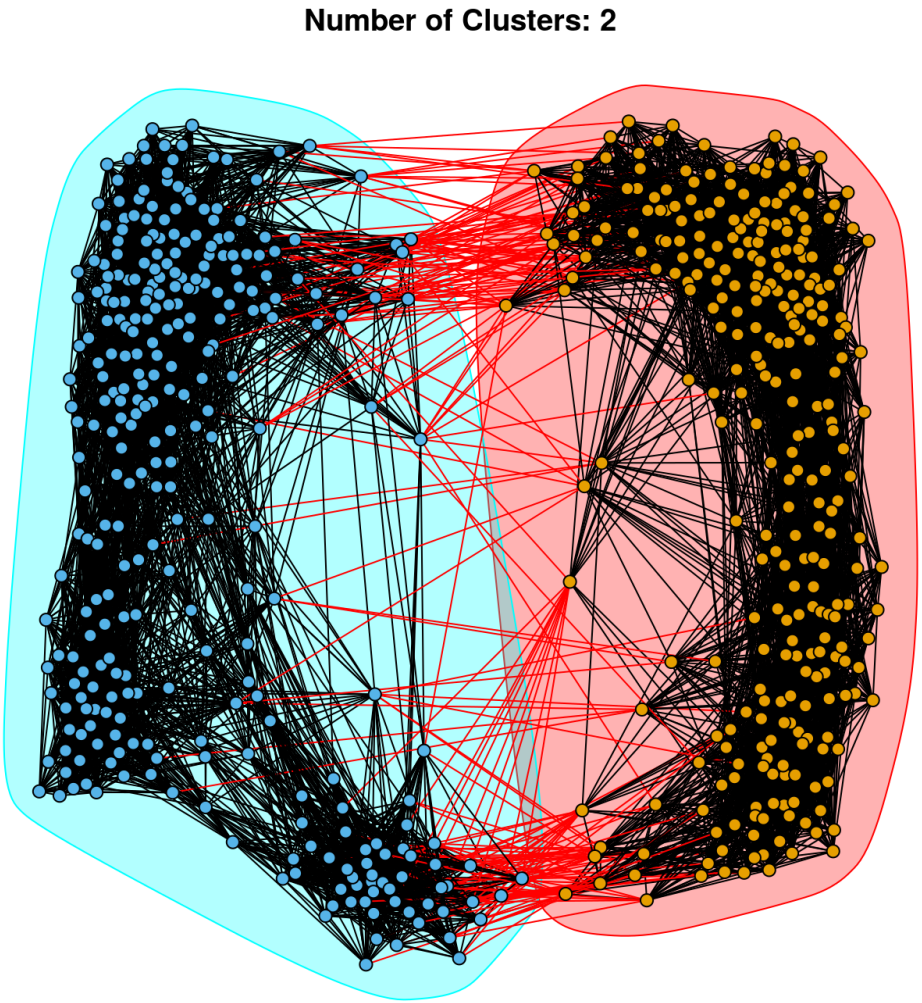
Figure 2. Clustering of clusters. This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.

Similarity Network Fusion (SNF)



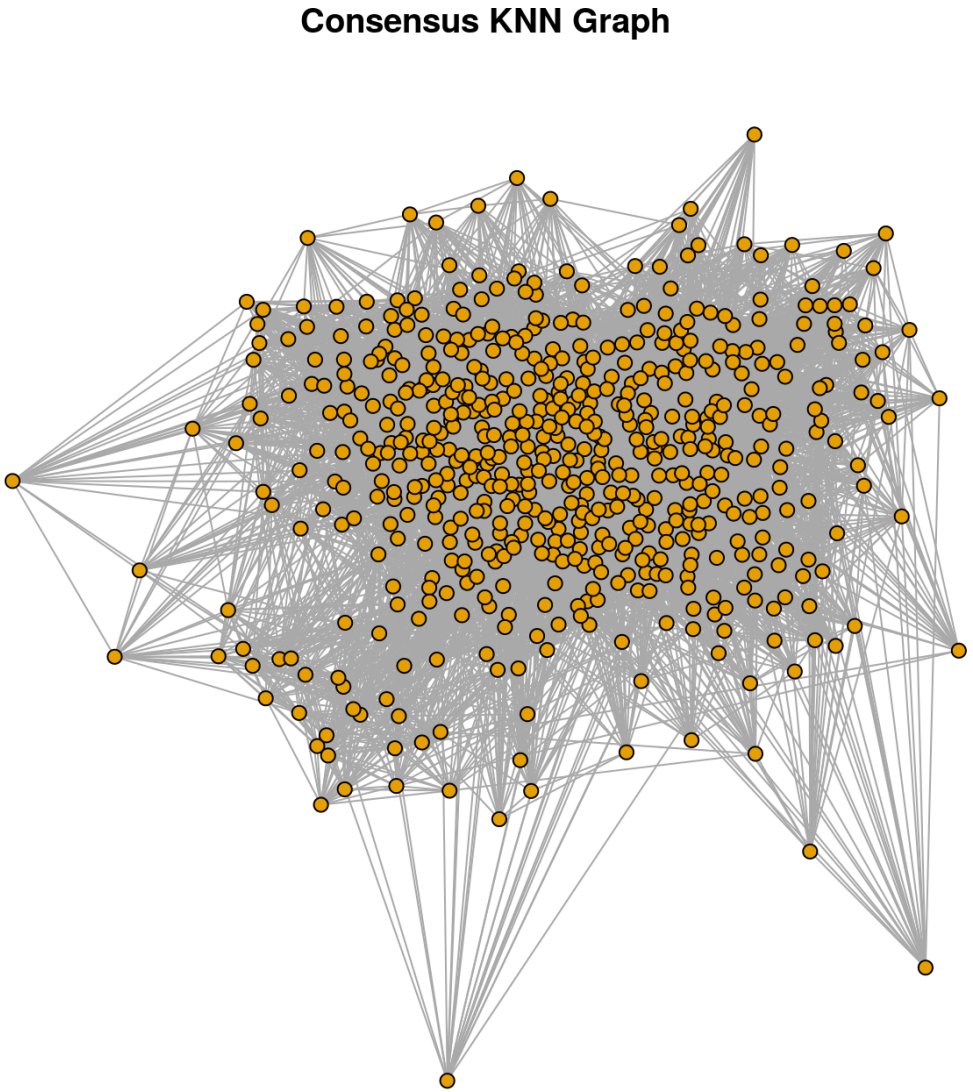
Clustering of Clusters: Bootstrap Consensus

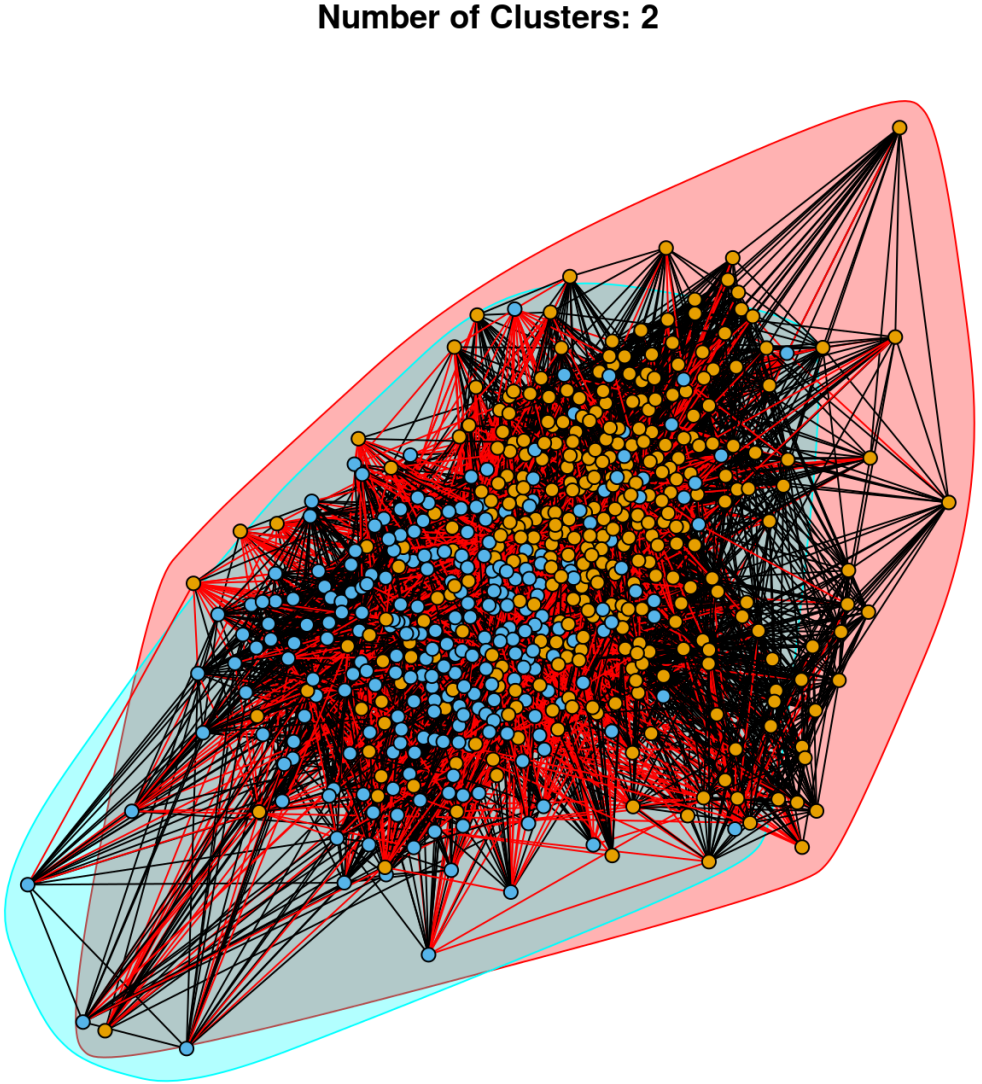




Instead of OMICs datasets construct pairwise adjucency matrices where elements are fractions each pair of samples belongs to the same cluster. Next we average the adjucency matrices, i.e. find samples that are consistently together across all OMICs

Graph-Based Consensus Clustering





Build fully-connected graph for each individual OMIC and average the weights of the edges of the graphs

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet