Unsupervised OMICs Integration

ISMB / ECCB 2021

Tutorial: a practical introduction to multi-omics integration and network analysis

Nikolay Oskolkov, NBIS SciLifeLab, 22.07.2021

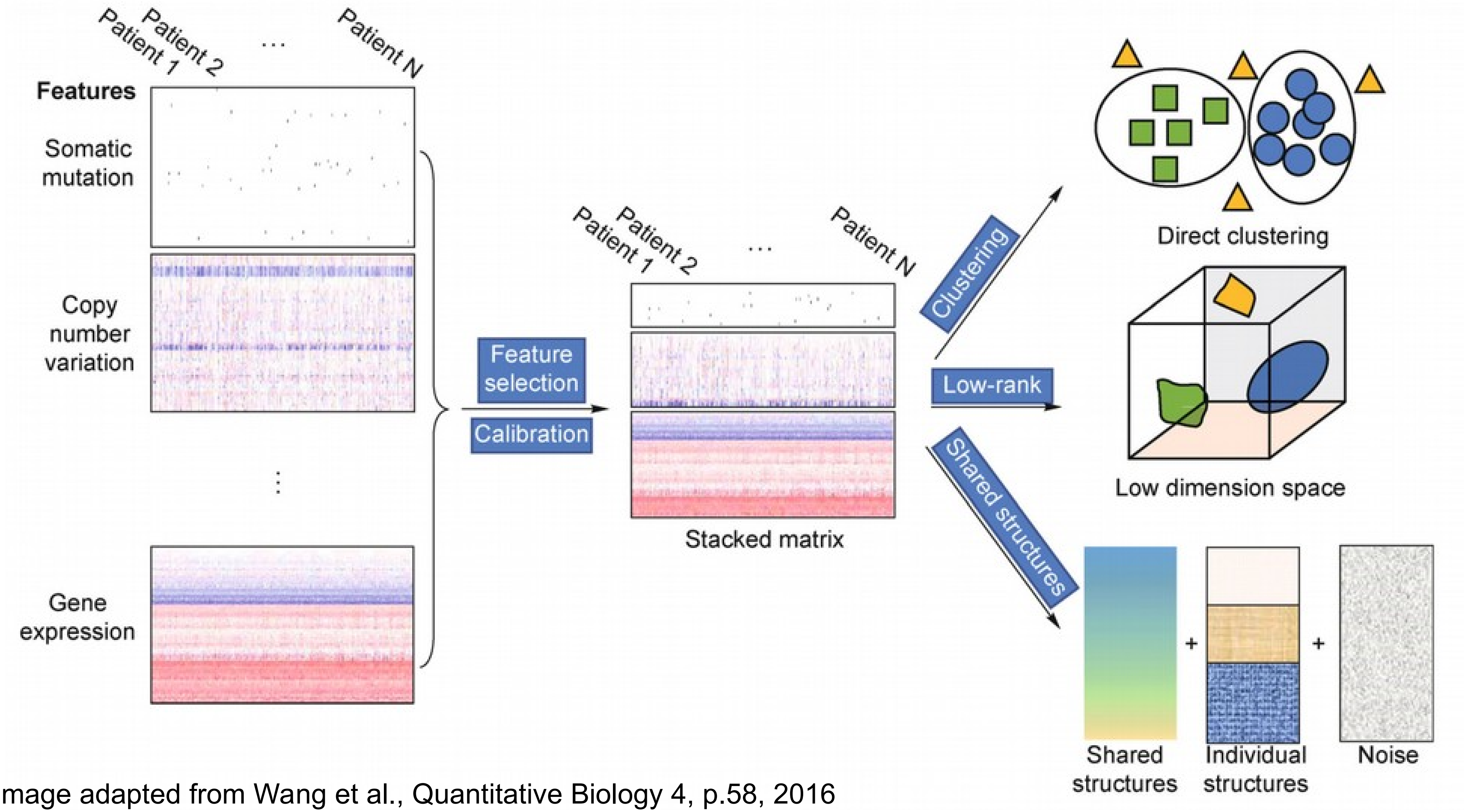
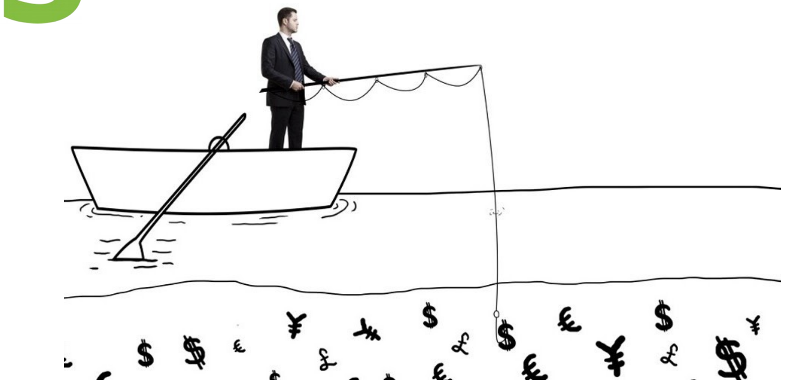


Image adapted from Wang et al., Quantitative Biology 4, p.58, 2016

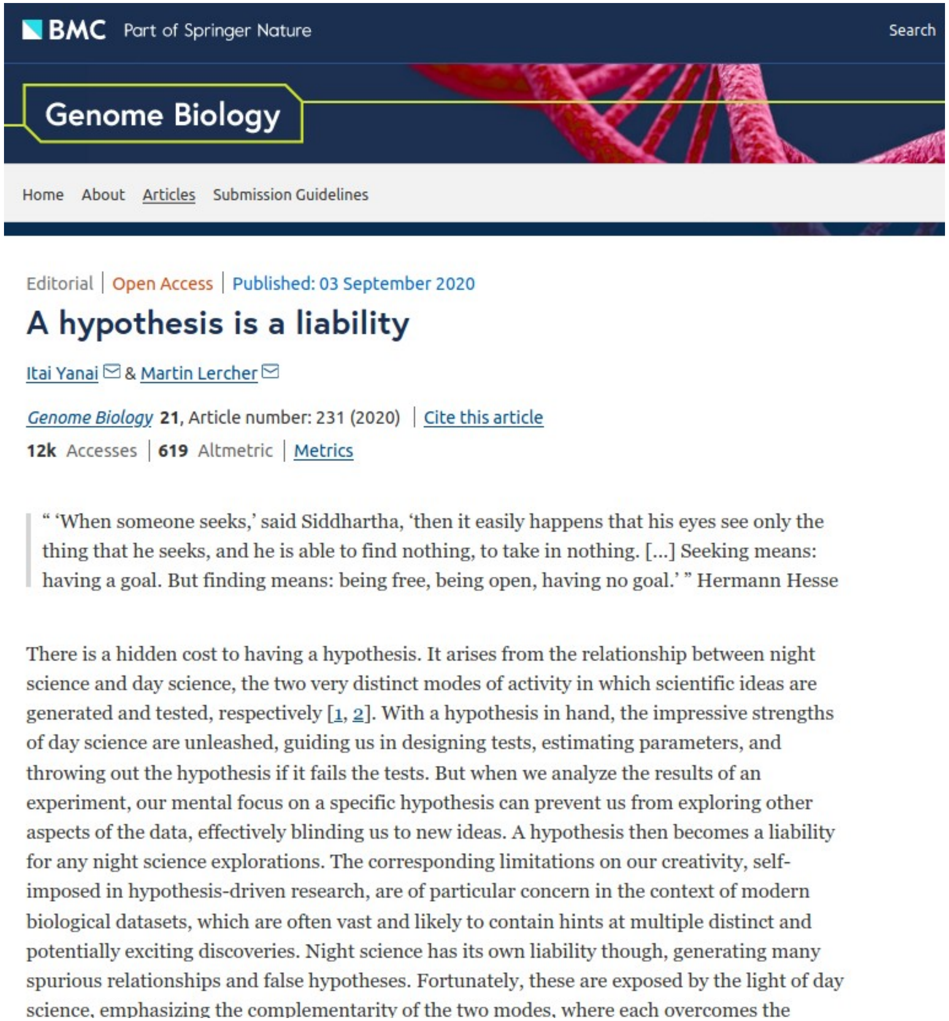
Find Something in My Data

Fishing expedition



- I do not understand your biological hypothesis

- I do not have any



BMC Part of Springer Nature

Search

Genome Biology

Home About Articles Submission Guidelines

Editorial | Open Access | Published: 03 September 2020

A hypothesis is a liability

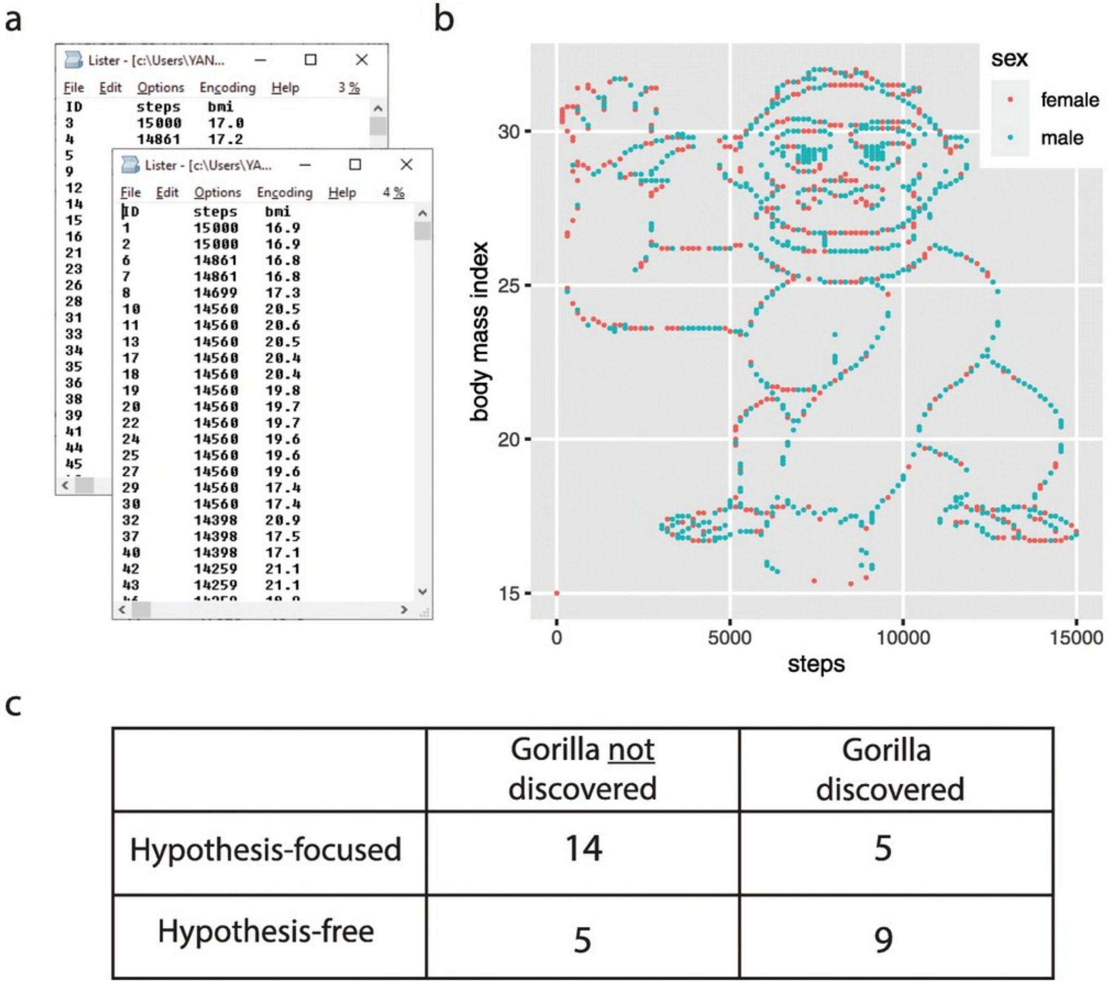
Itai Yanai & Martin Lercher

Genome Biology **21**, Article number: 231 (2020) | Cite this article

12k Accesses | **619** Altmetric | Metrics

> " 'When someone seeks,' said Siddhartha, 'then it easily happens that his eyes see only the thing that he seeks, and he is able to find nothing, to take in nothing. [...] Seeking means: having a goal. But finding means: being free, being open, having no goal.' " Hermann Hesse

There is a hidden cost to having a hypothesis. It arises from the relationship between night science and day science, the two very distinct modes of activity in which scientific ideas are generated and tested, respectively [1, 2]. With a hypothesis in hand, the impressive strengths of day science are unleashed, guiding us in designing tests, estimating parameters, and throwing out the hypothesis if it fails the tests. But when we analyze the results of an experiment, our mental focus on a specific hypothesis can prevent us from exploring other aspects of the data, effectively blinding us to new ideas. A hypothesis then becomes a liability for any night science explorations. The corresponding limitations on our creativity, self-imposed in hypothesis-driven research, are of particular concern in the context of modern biological datasets, which are often vast and likely to contain hints at multiple distinct and potentially exciting discoveries. Night science has its own liability though, generating many spurious relationships and false hypotheses. Fortunately, these are exposed by the light of day science, emphasizing the complementarity of the two modes, where each overcomes the



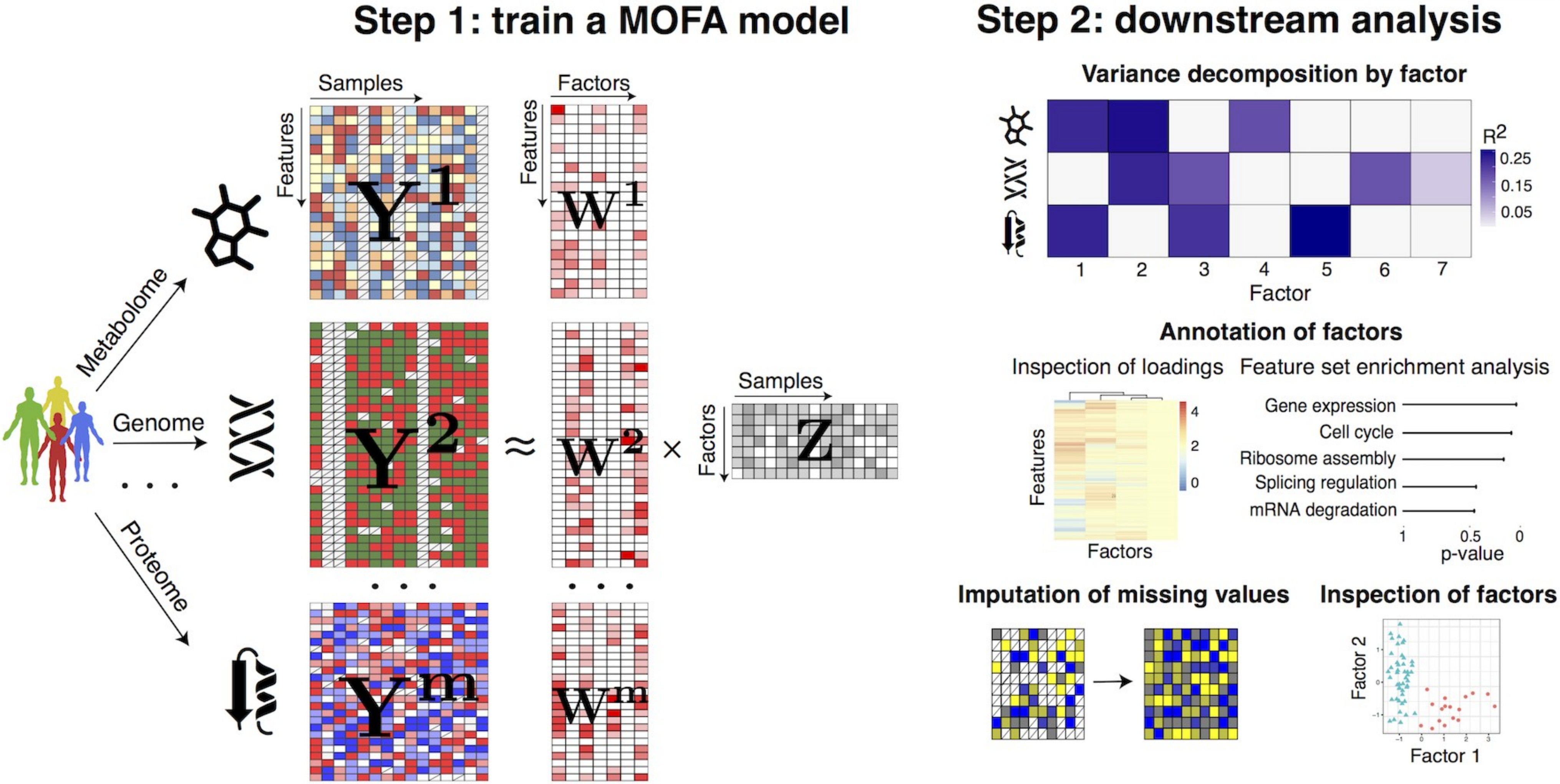
```
ID       steps   bmi
3        15000   17.0
4        14861   17.2
5
9
12
14
15
16
21
23
26
28
31
33
34
35
36
38
39
41
44
45
```

```
ID       steps   bmi
1        15000   16.9
2        15000   16.9
6        14861   16.8
7        14861   16.8
8        14699   17.3
10       14560   20.5
11       14560   20.6
13       14560   20.5
17       14560   20.4
18       14560   20.4
19       14560   19.8
20       14560   19.7
22       14560   19.7
24       14560   19.6
25       14560   19.6
27       14560   19.6
29       14560   17.4
30       14560   17.4
32       14398   20.9
37       14398   17.5
40       14398   17.1
42       14259   21.1
43       14259   21.1
```

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

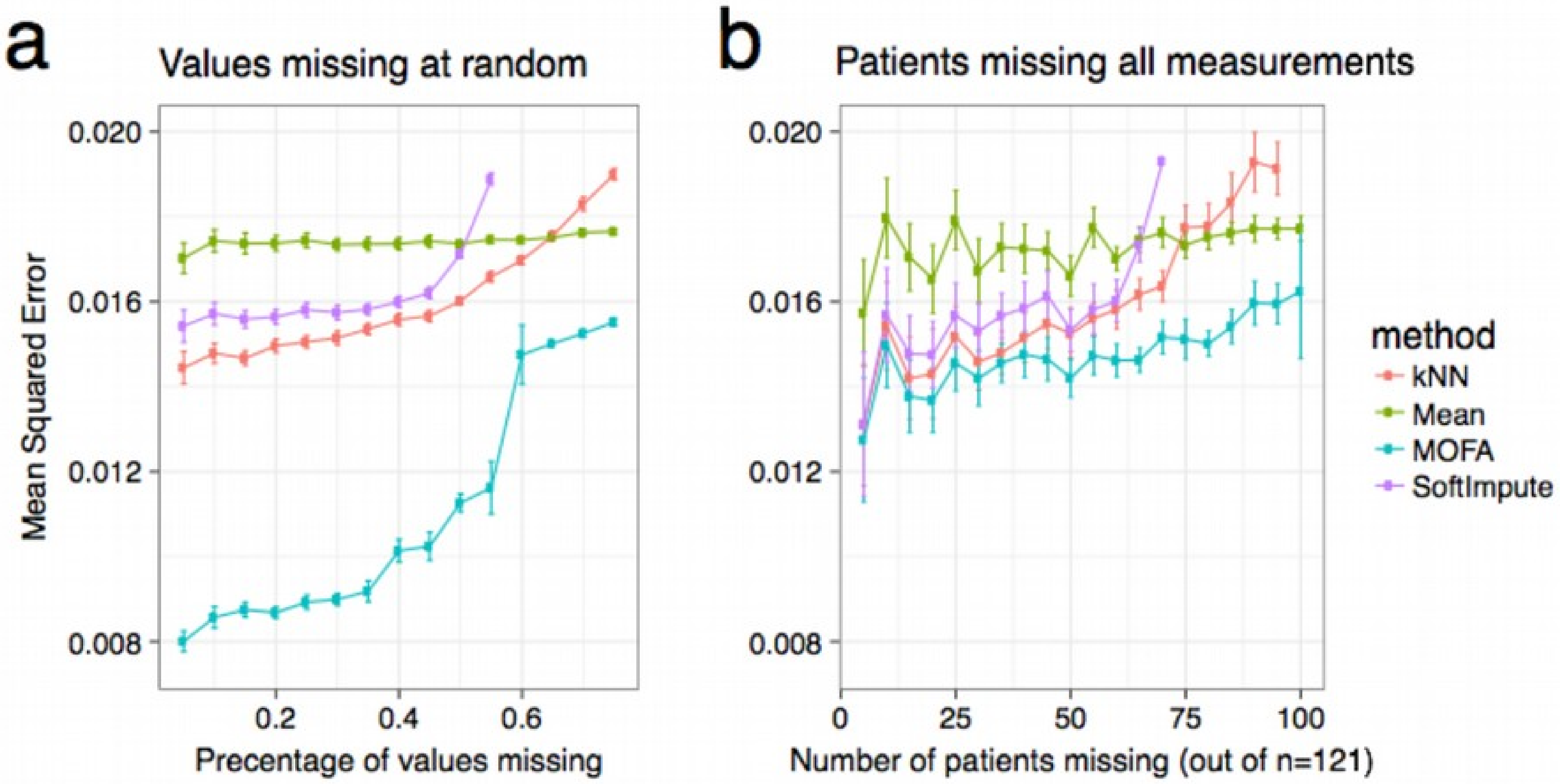
a An artificial dataset given to students with and without explicit hypotheses on the relationship between BMI and the steps taken on a particular day, for men and women. **b** A plot of the dataset. **c** The contingency table for students in the two groups ("hypothesis-focused," "hypothesis-free") that discovered the gorilla or not [6]

MOFA: Hypothesis-Free Exploratory Tool



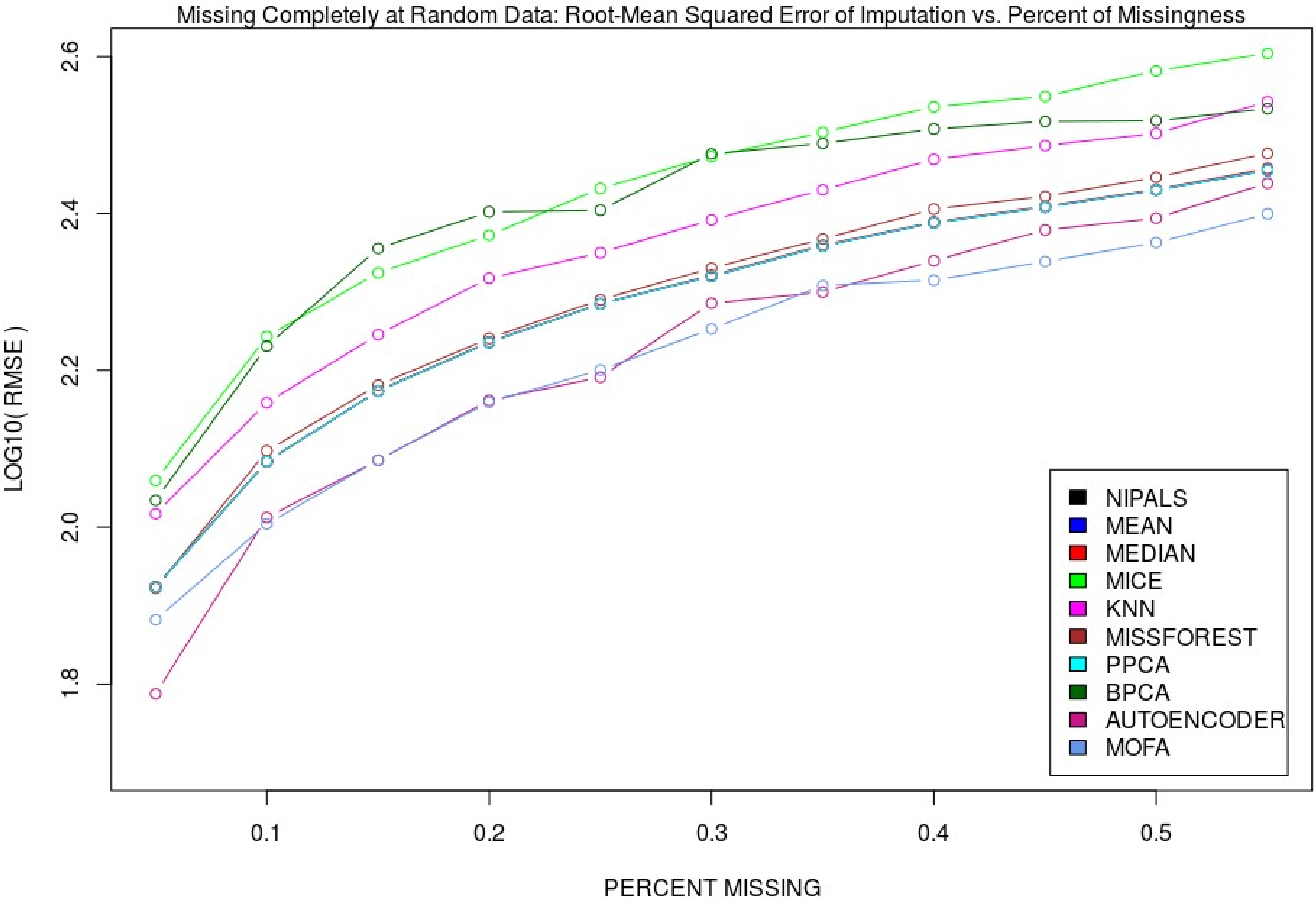
- Visualisation of samples in factor space
- Annotation of factors using (gene set) enrichment analysis
- Imputation of missing values
- Support of OMICs with non-Gaussian distribution including binary and count data

MOFA: Imputation of Missing Values

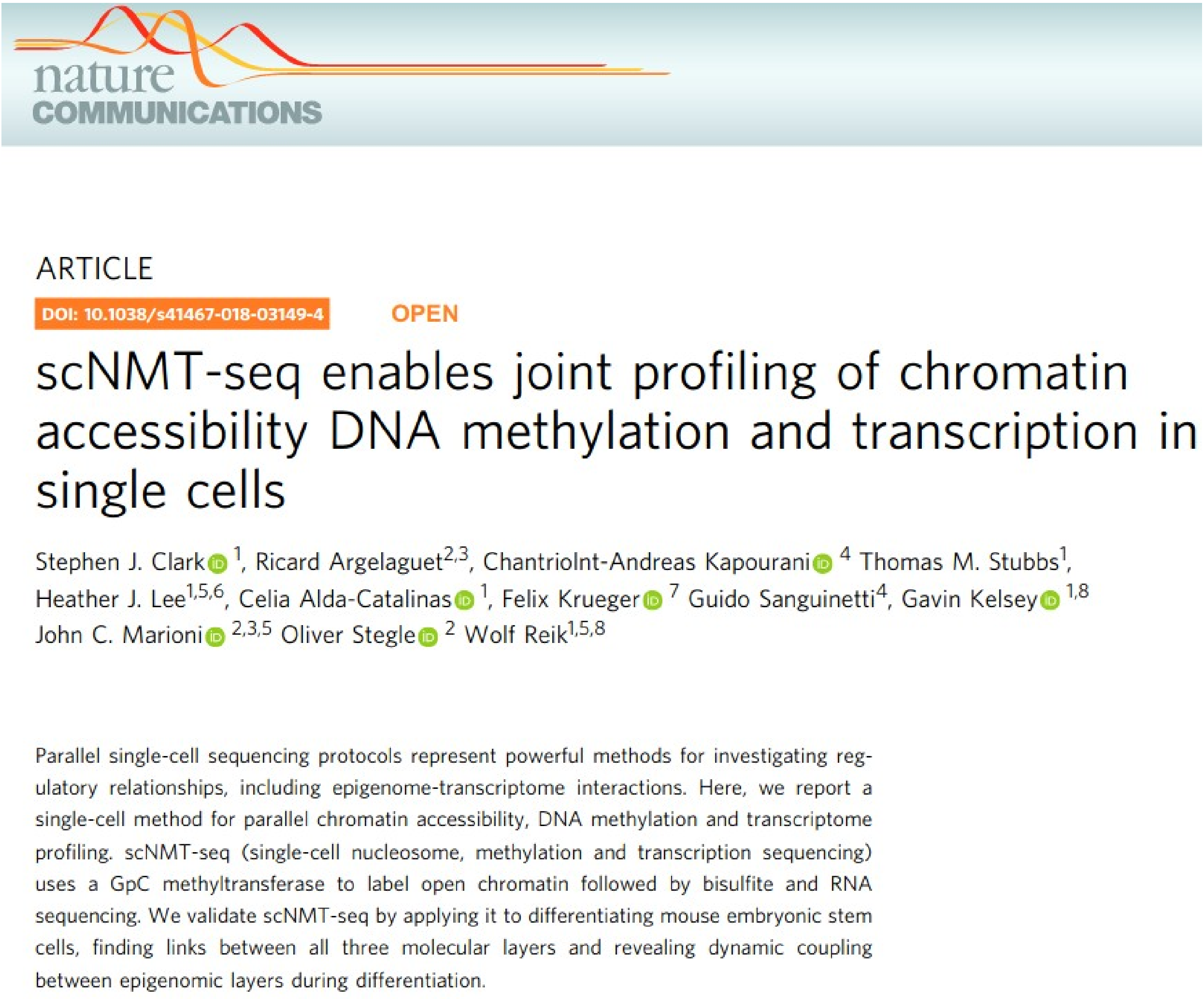


Bayesian framework is insensitive to missing data, priors compensate for the lack of data

MOFA: Imputation of Missing Values



scNMT Data Set: scRNAseq + scBSseq + scATACseq



nature COMMUNICATIONS

ARTICLE

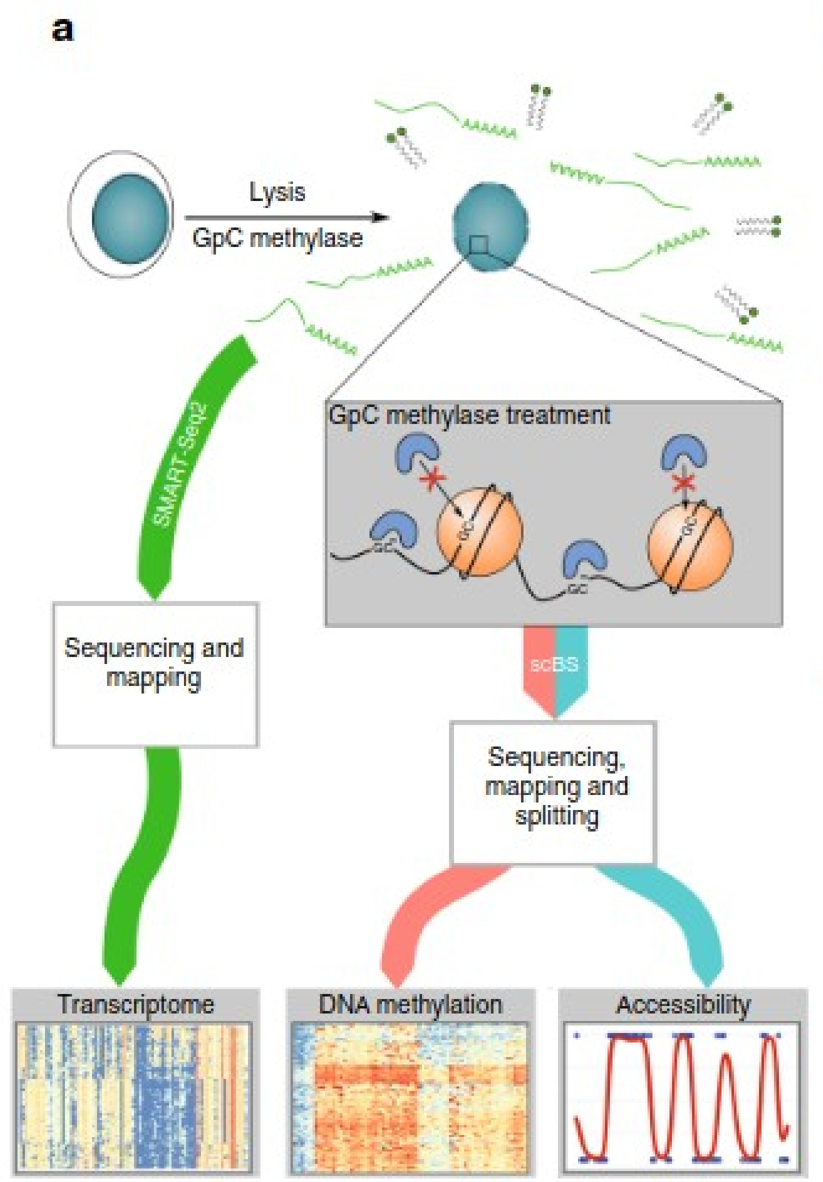
DOI: 10.1038/s41467-018-03149-4 OPEN

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

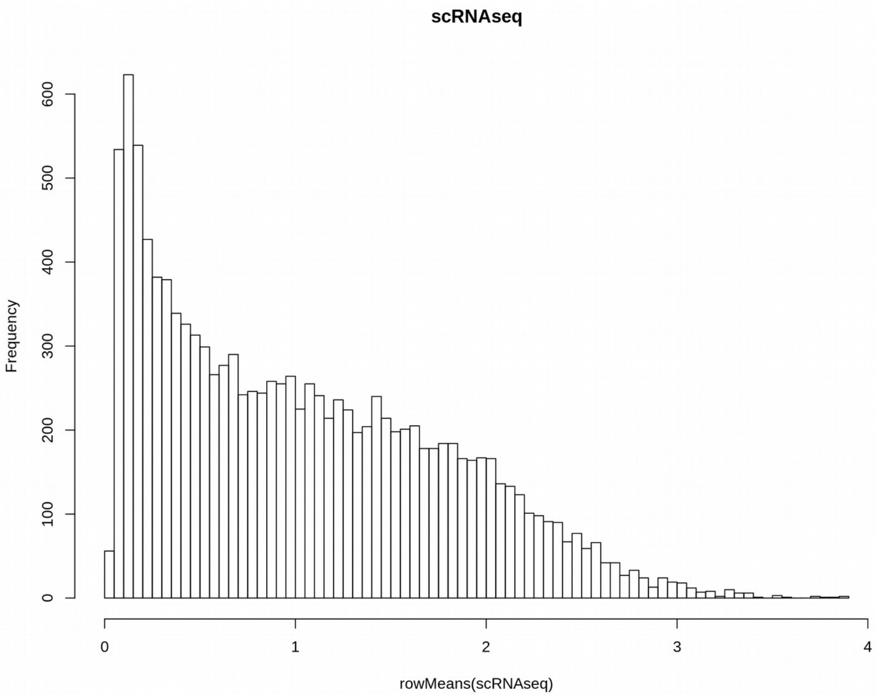
Stephen J. Clark [1], Ricard Argelaguet[2,3], ChantrioInt-Andreas Kapourani [4] Thomas M. Stubbs[1], Heather J. Lee[1,5,6], Celia Alda-Catalinas [1], Felix Krueger [7] Guido Sanguinetti[4], Gavin Kelsey [1,8] John C. Marioni [2,3,5] Oliver Stegle [2] Wolf Reik[1,5,8]

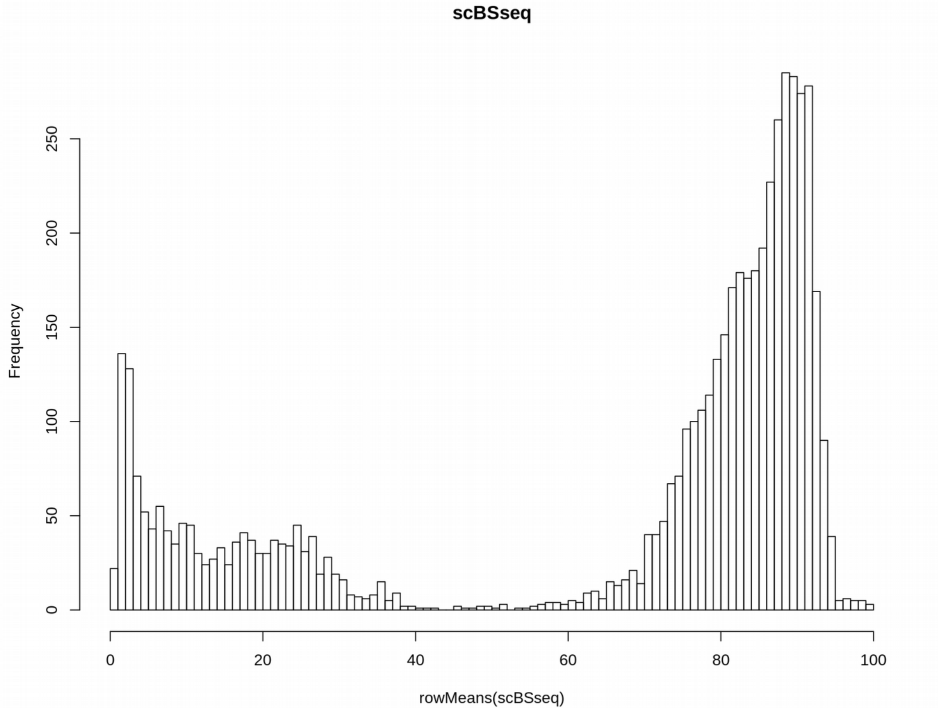
Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell 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.

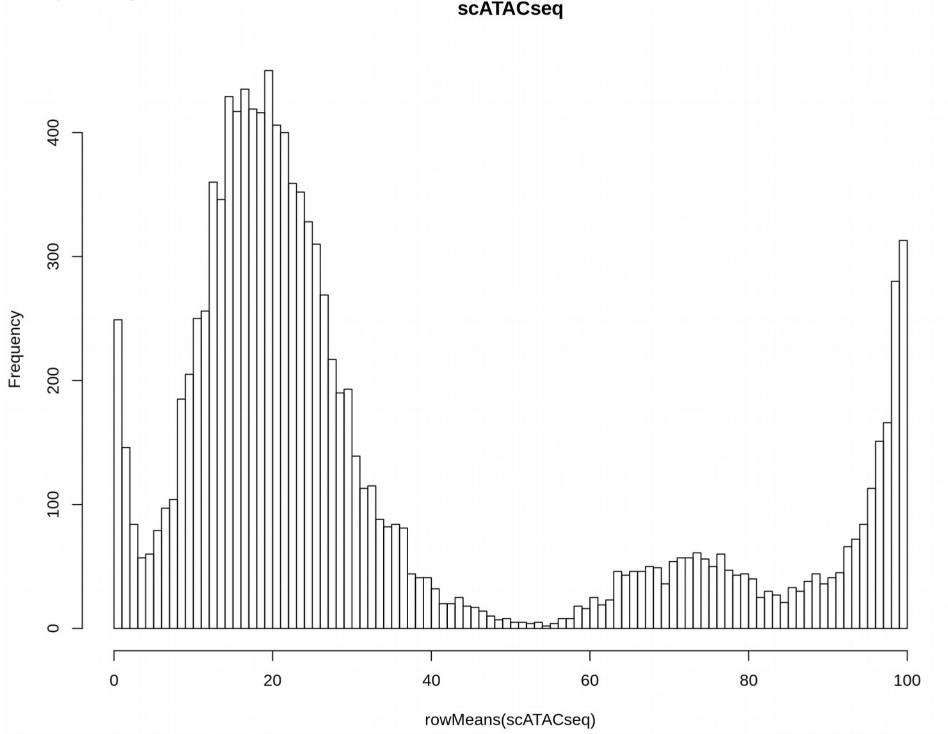
ATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-031



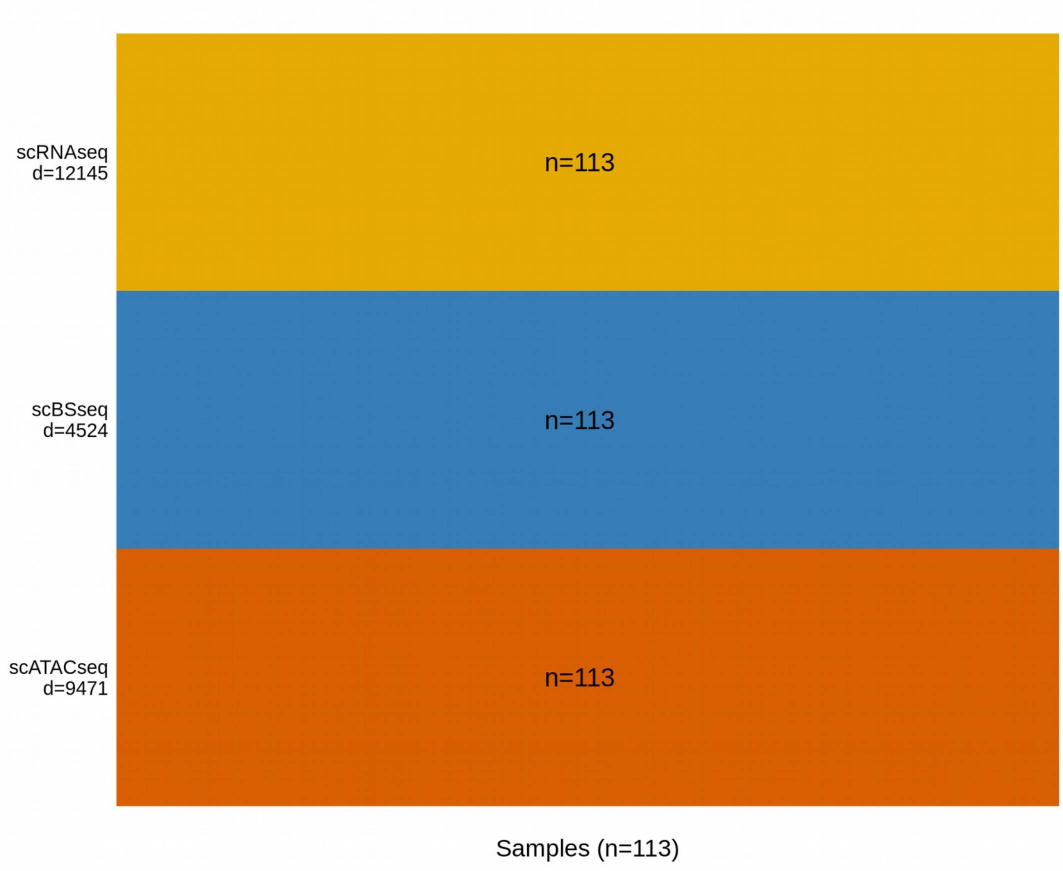
scNMT Data Set: Distributions

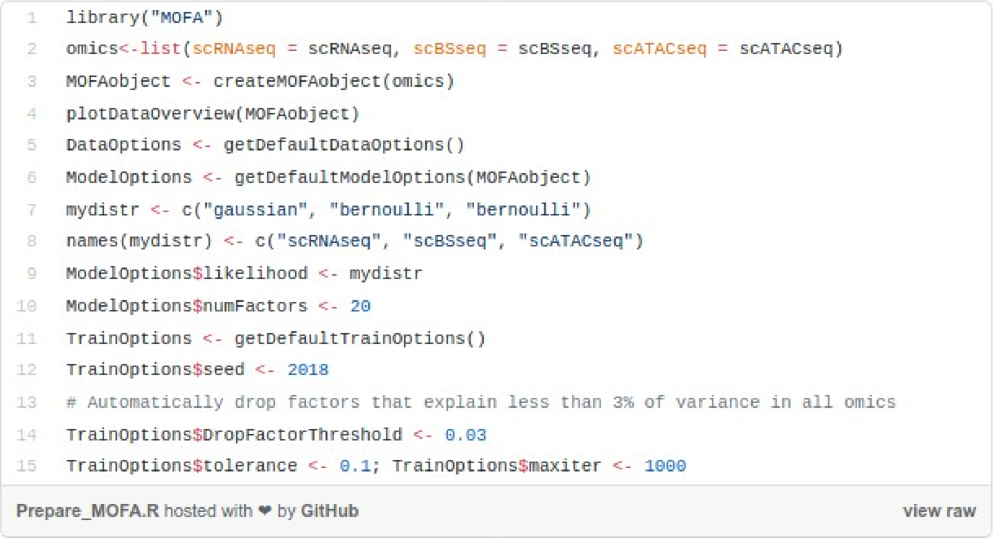






scNMT Data Set: Summary Stats

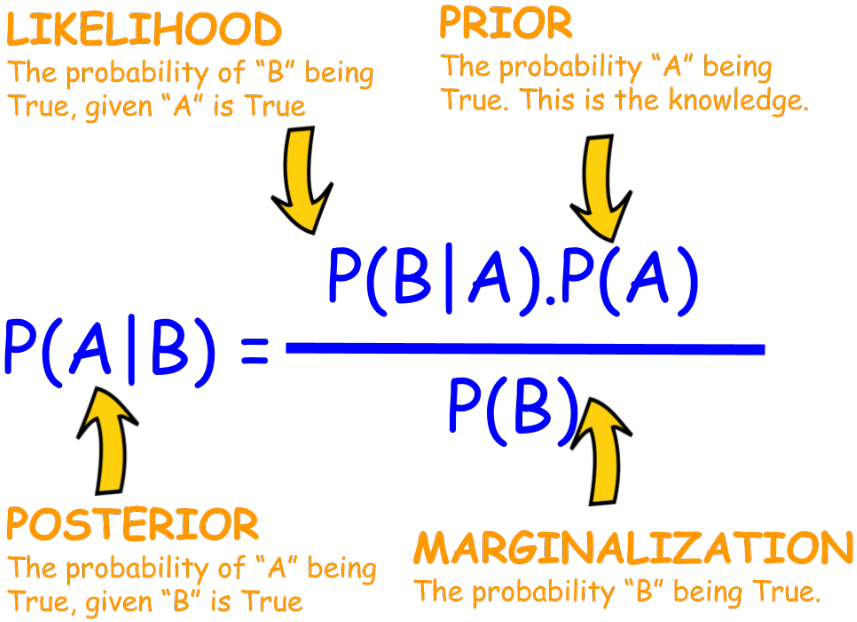




```r
library("MOFA")
omics<-list(scRNAseq = scRNAseq, scBSseq = scBSseq, scATACseq = scATACseq)
MOFAobject <- createMOFAobject(omics)
plotDataOverview(MOFAobject)
DataOptions <- getDefaultDataOptions()
ModelOptions <- getDefaultModelOptions(MOFAobject)
mydistr <- c("gaussian", "bernoulli", "bernoulli")
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



LIKELIHOOD
The probability of "B" being True, given "A" is True

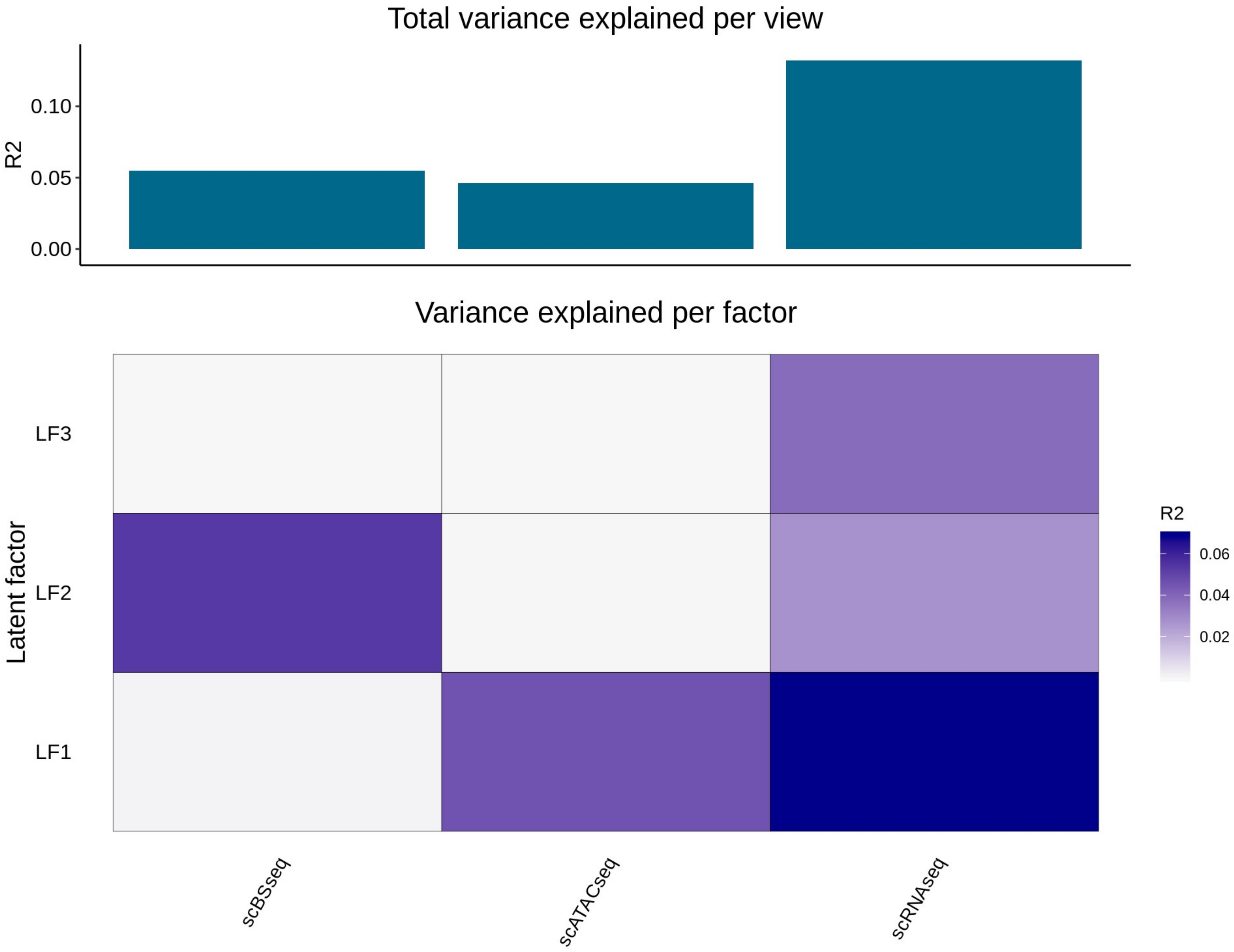
PRIOR
The probability "A" being True. This is the knowledge.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

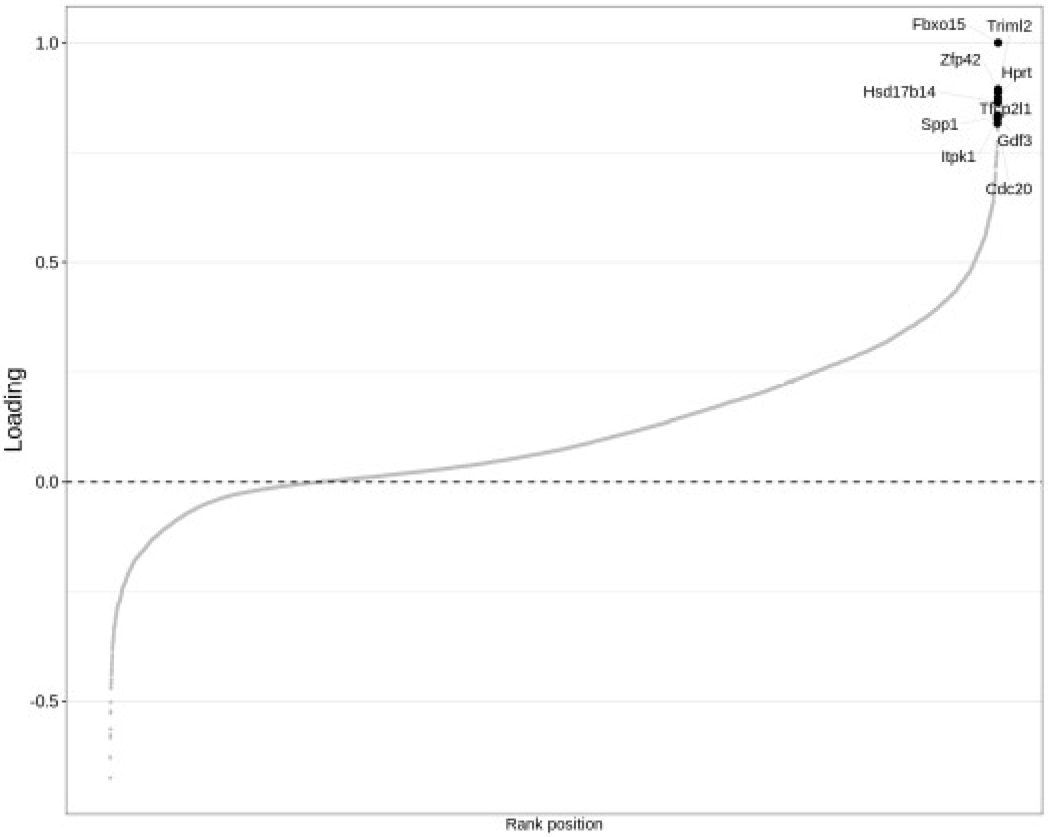
POSTERIOR
The probability of "A" being True, given "B" is True

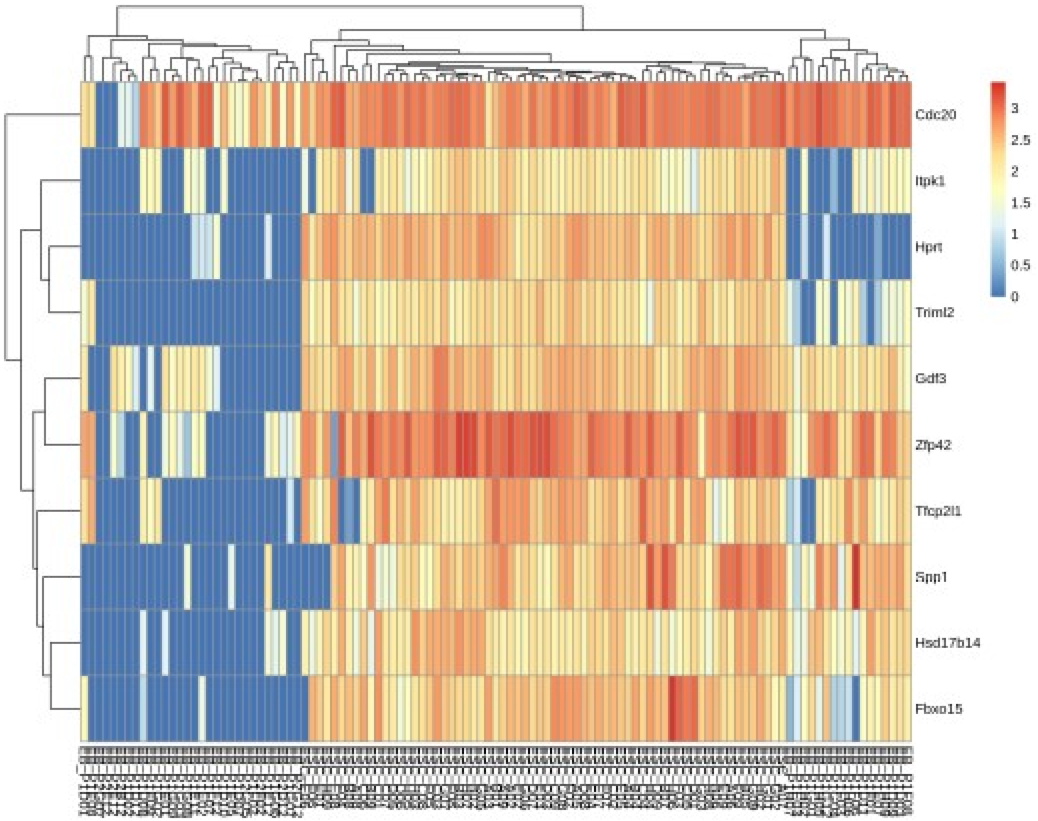
MARGINALIZATION
The probability "B" being True.

scNMT Data Set: Variance Explained



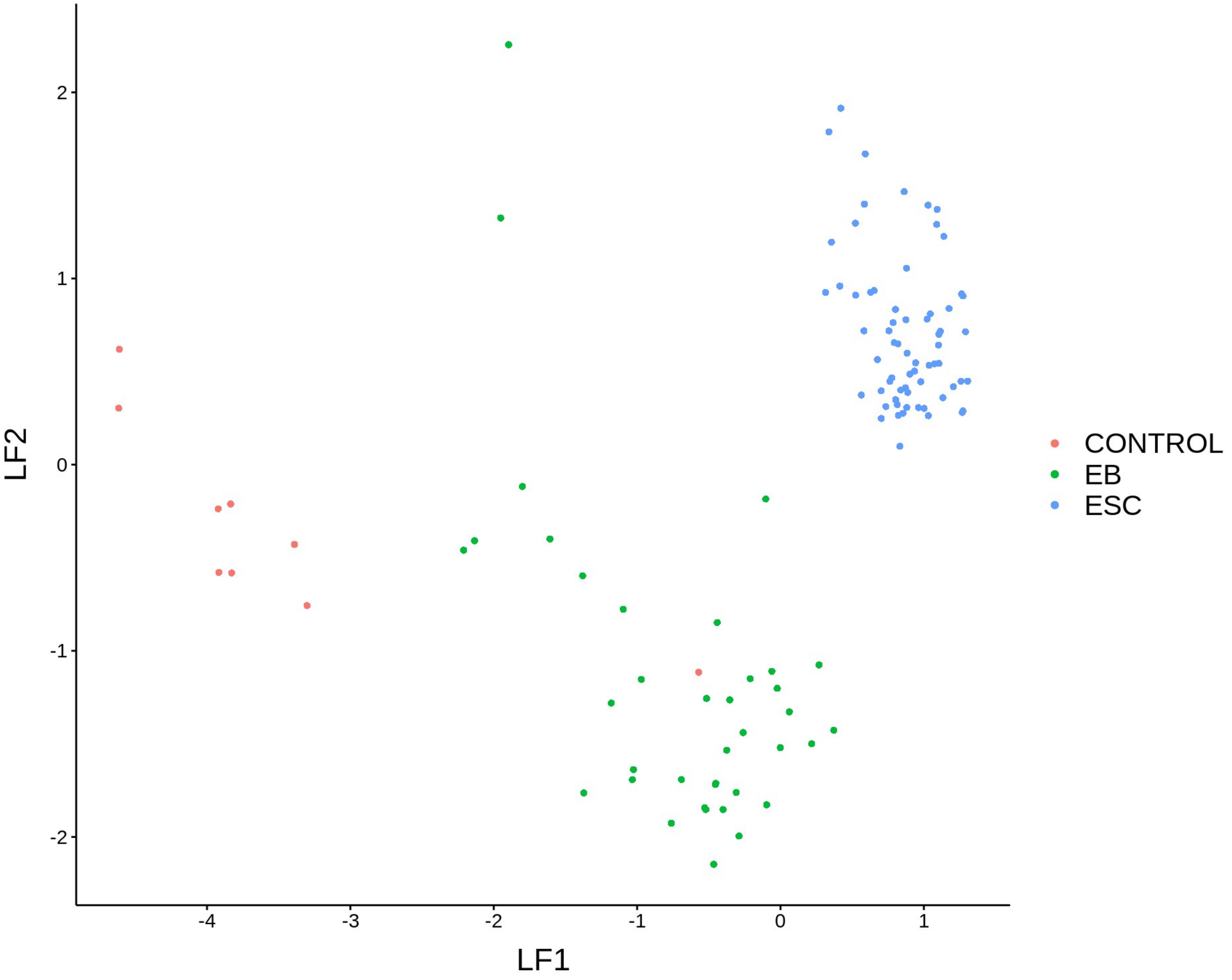
MOFA: Interpretation of Factors





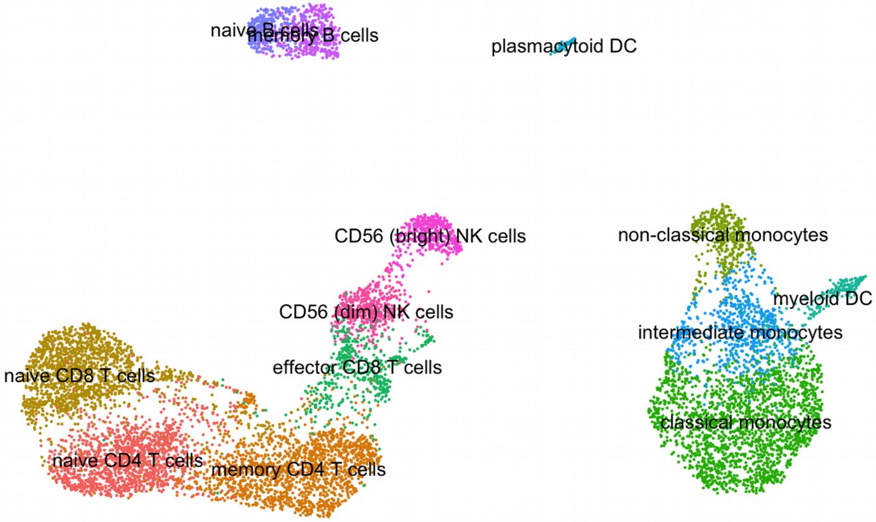
ESC and EB cells are separable on the heatmap built on loadings of the MOFA latent factors

MOFA: Visualizing Low-Dimensional Space

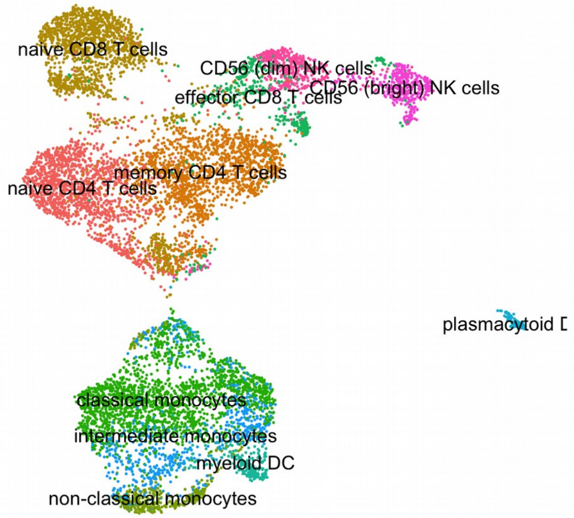


MOFA for scOmics Integration: 10X PBMC

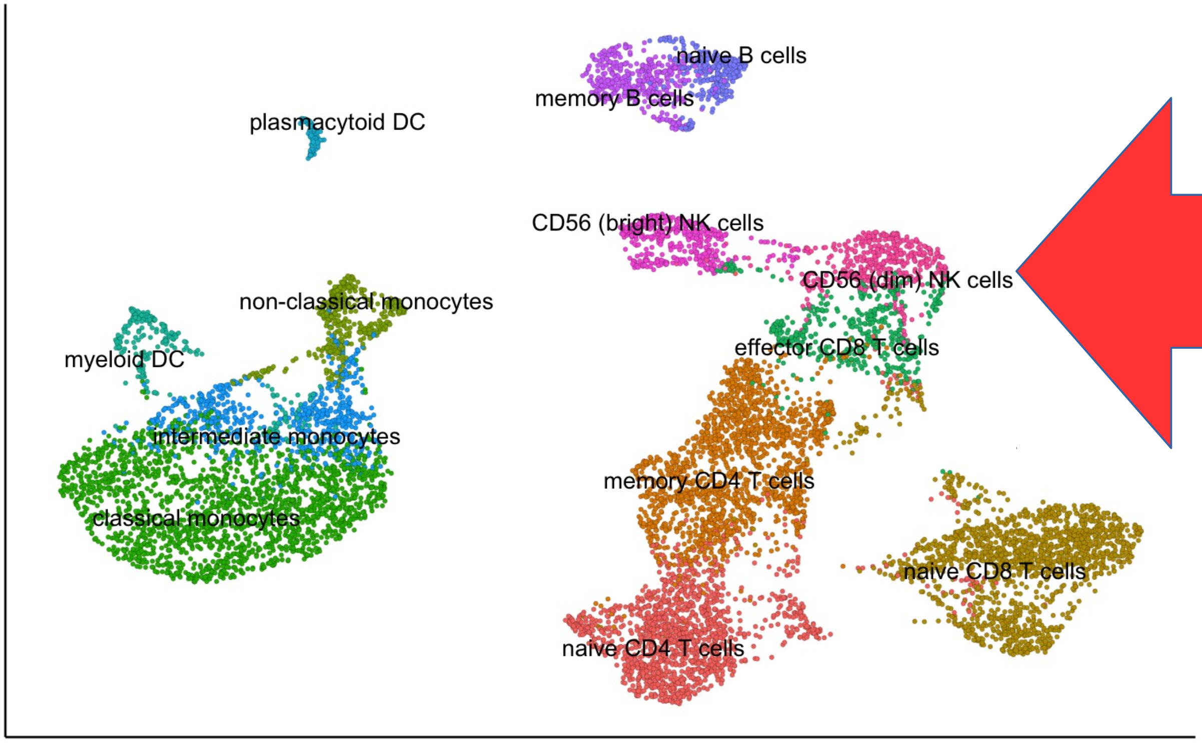
scRNAseq



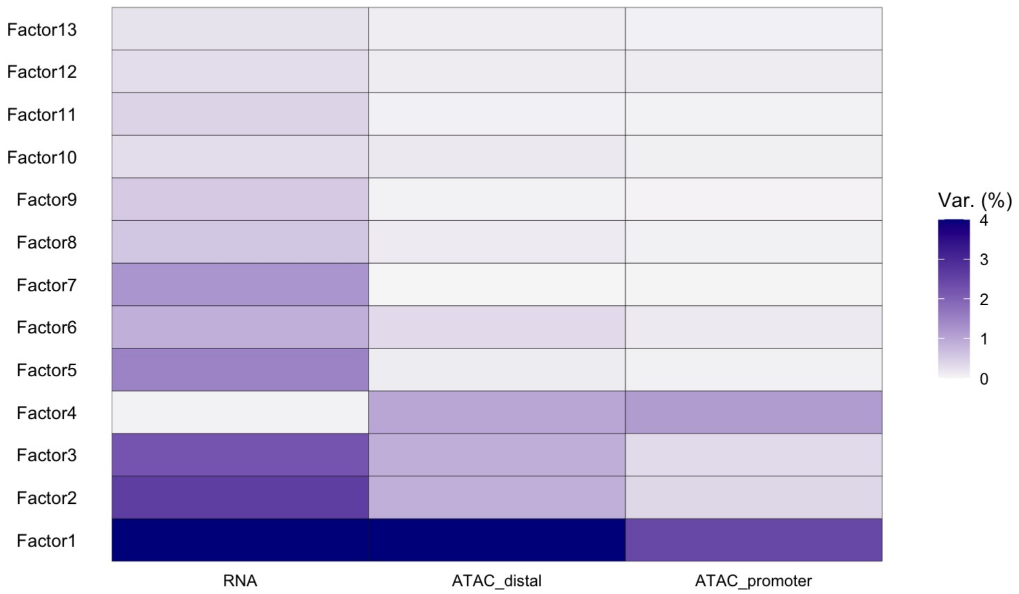
scATACseq

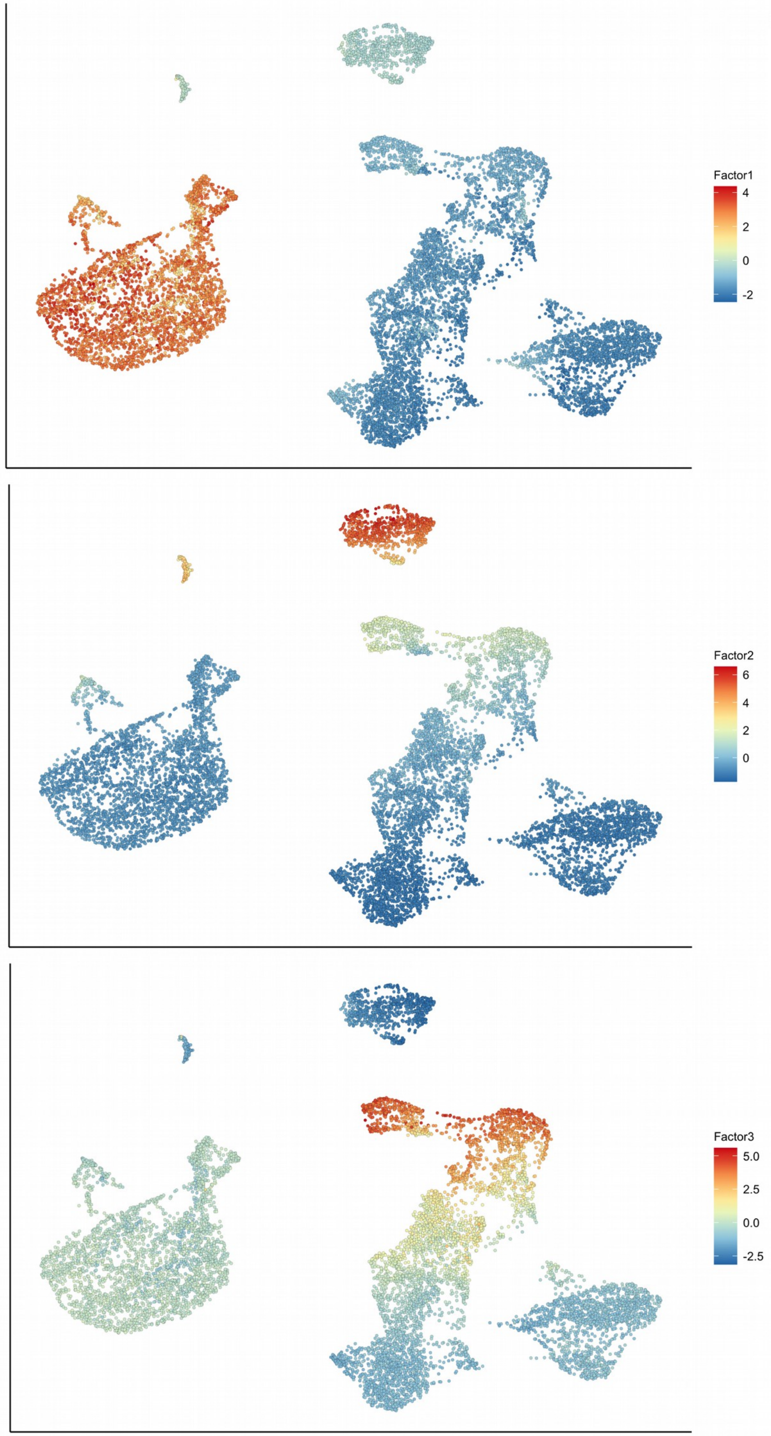


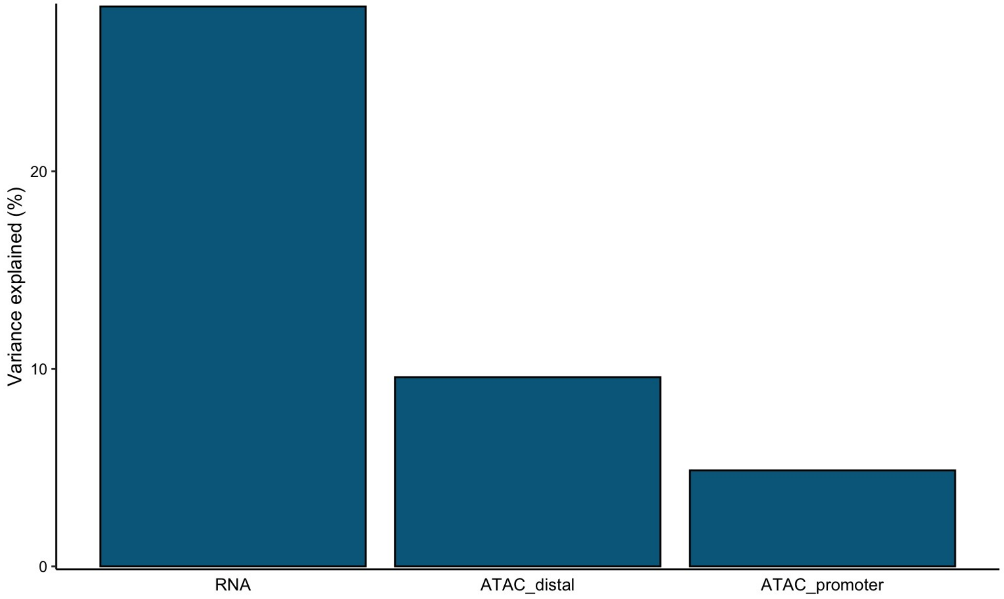
scRNAseq + scATACseq



MOFA for scOmics Integration: 10X PBMC







Other Unsupervised Integrative OMICs Methods

Clustering of Clusters

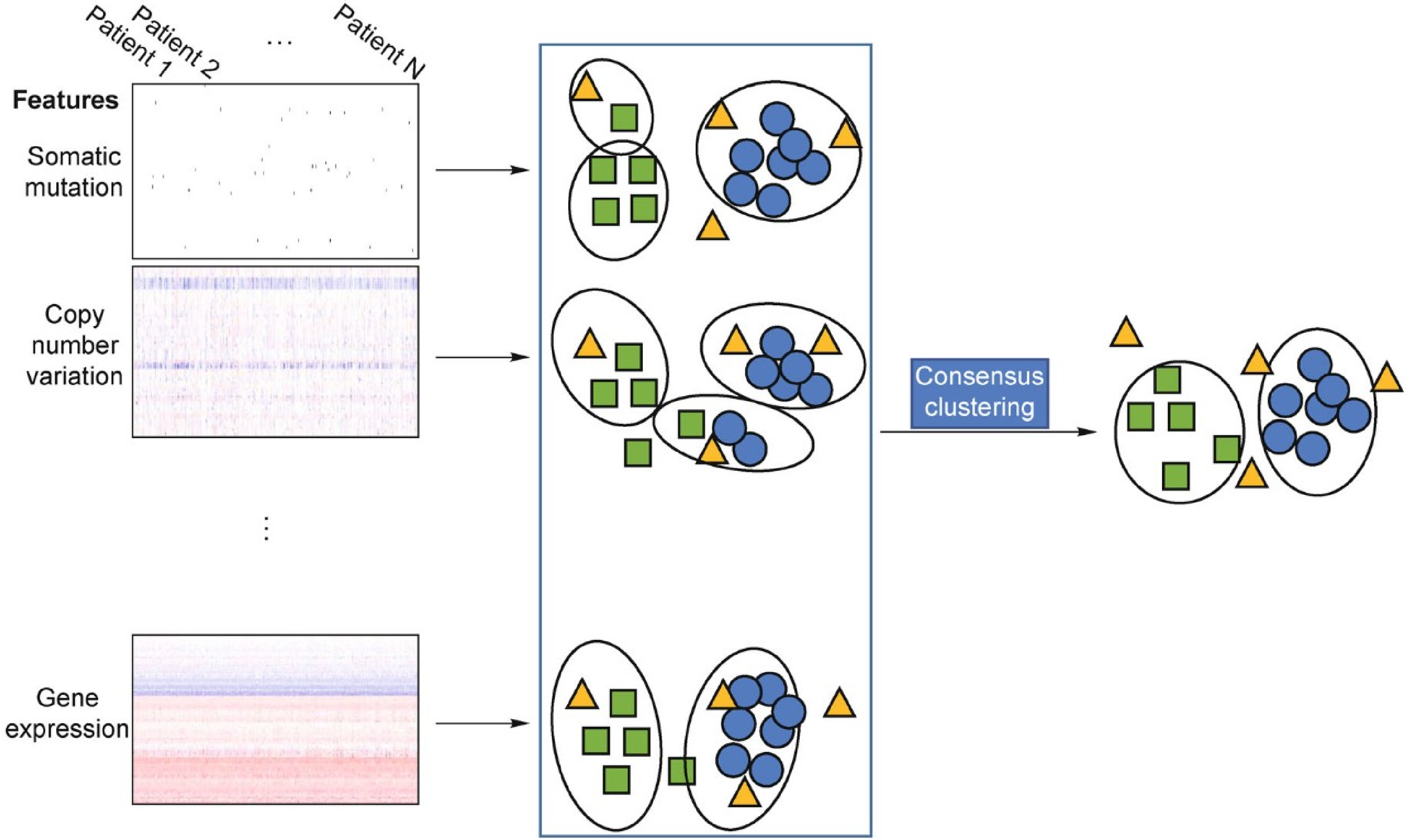
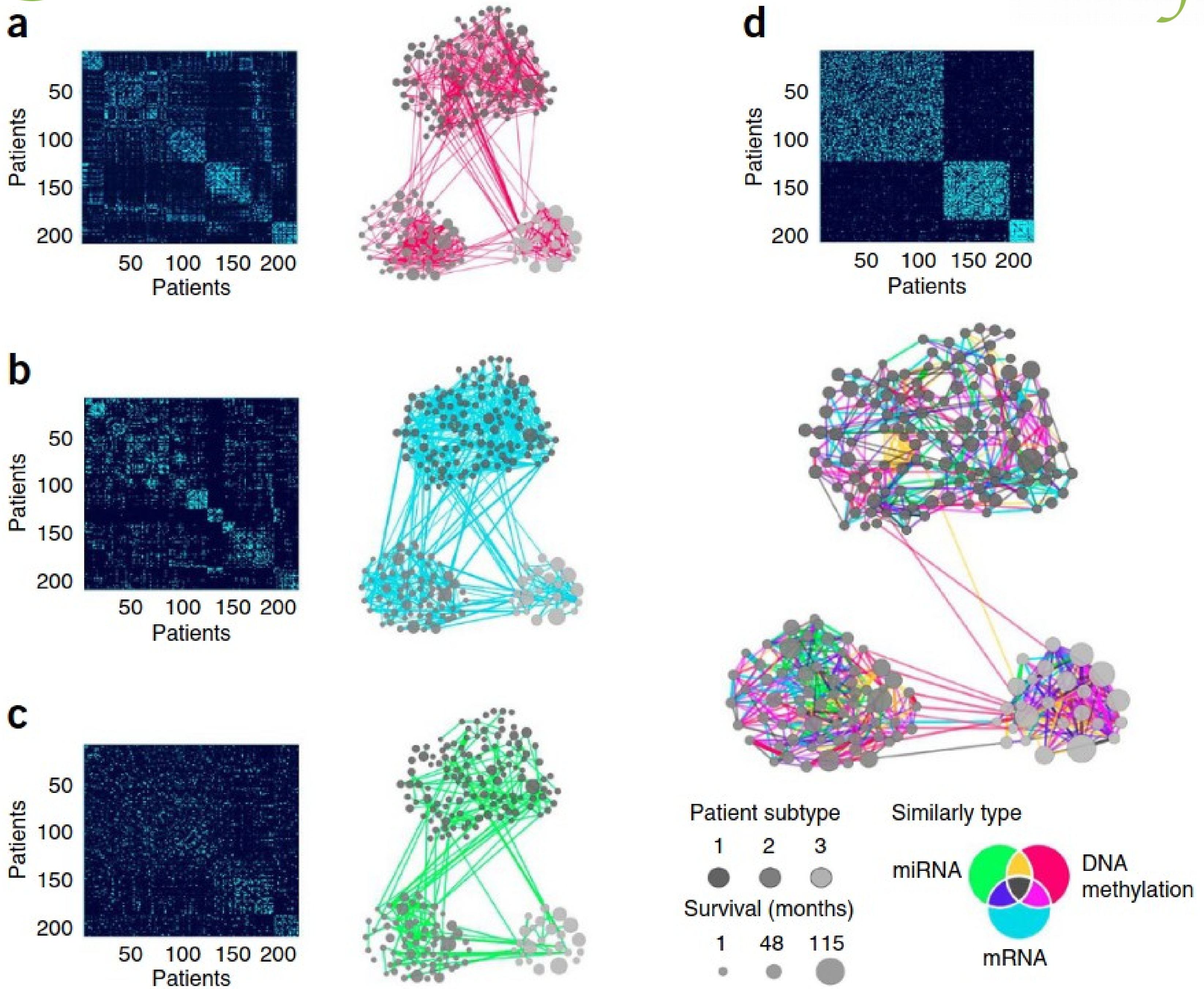


Figure 2. Clustering of clusters. This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.

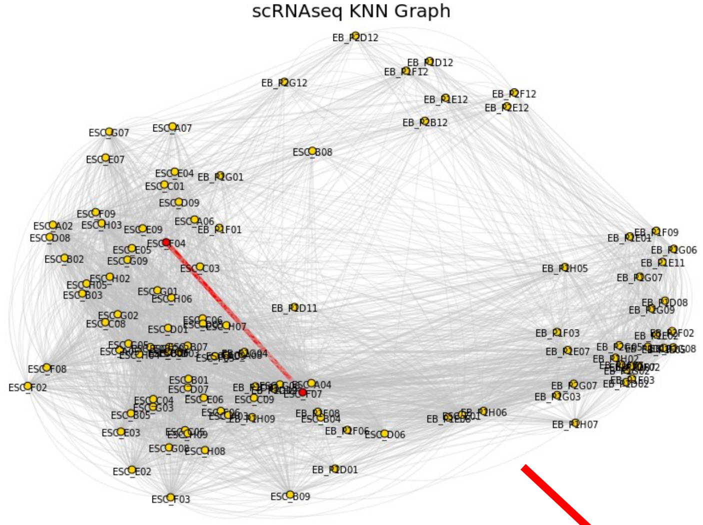
Wang et al., Quantitative Biology 2016, 4(1): 58–67

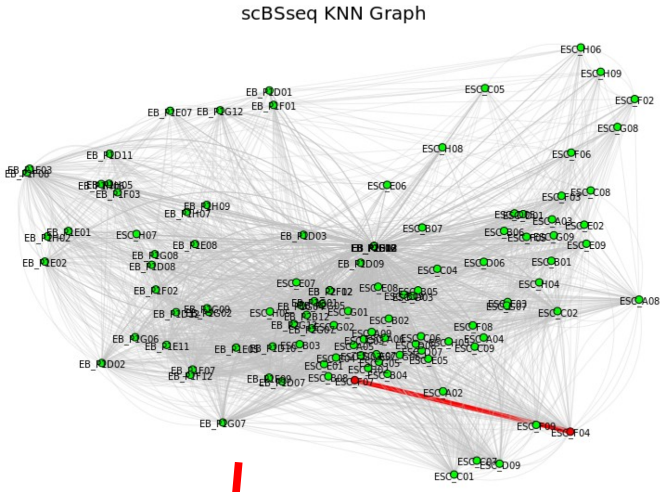
Similarity Network Fusion (SNF)

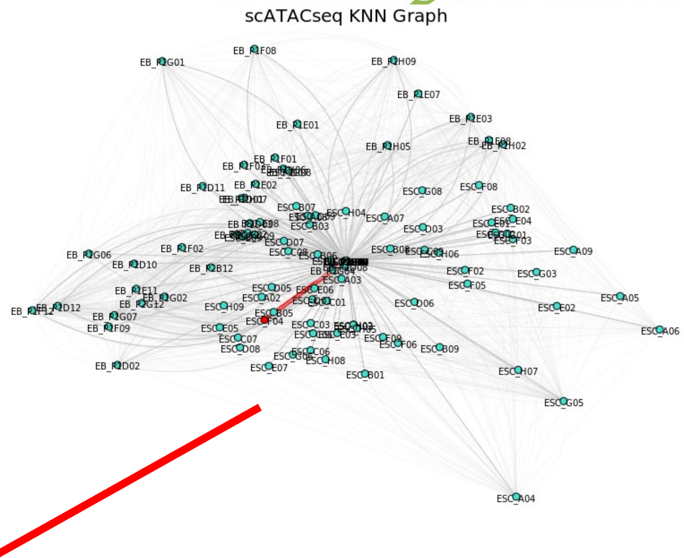


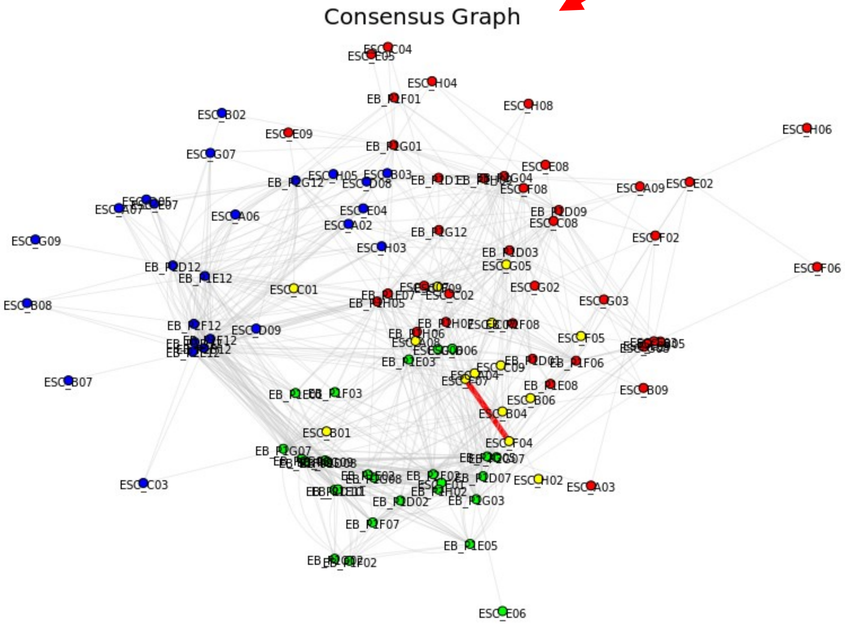
Wang, B., Mezlini, A., Demir, F. et al. . Nat Methods 11, 333–337 (2014)

Graph-Based OMICs Integration

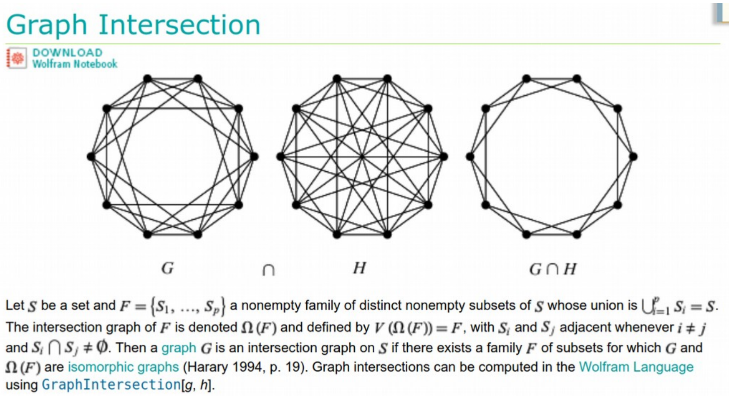








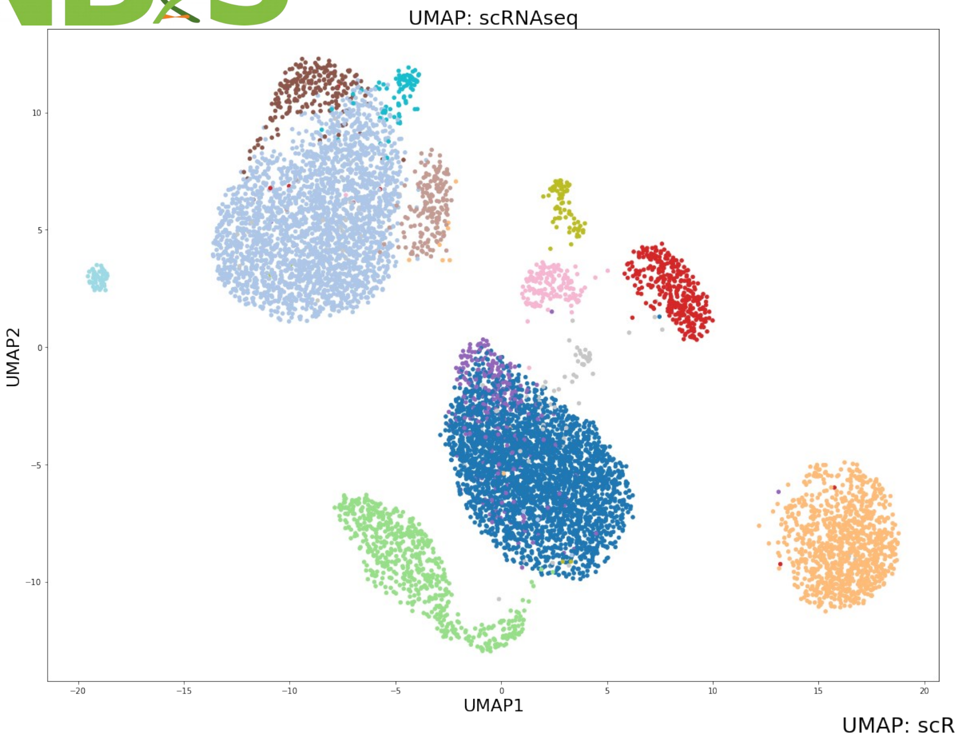
Keep edges consistently present across the OMICs

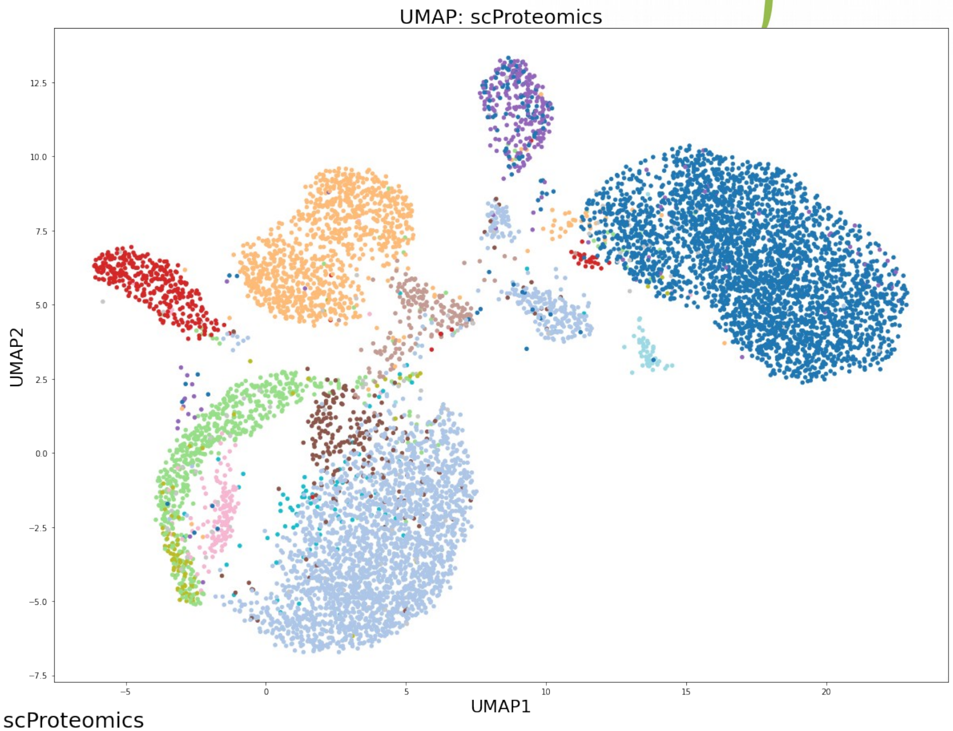


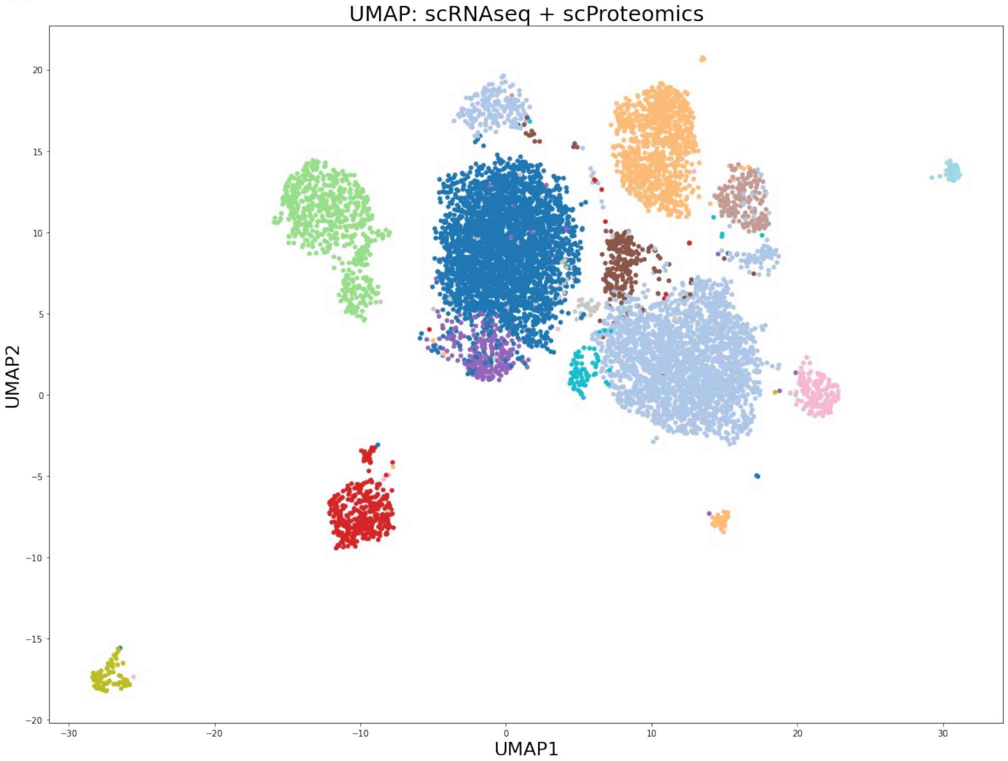
Graph Intersection

Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

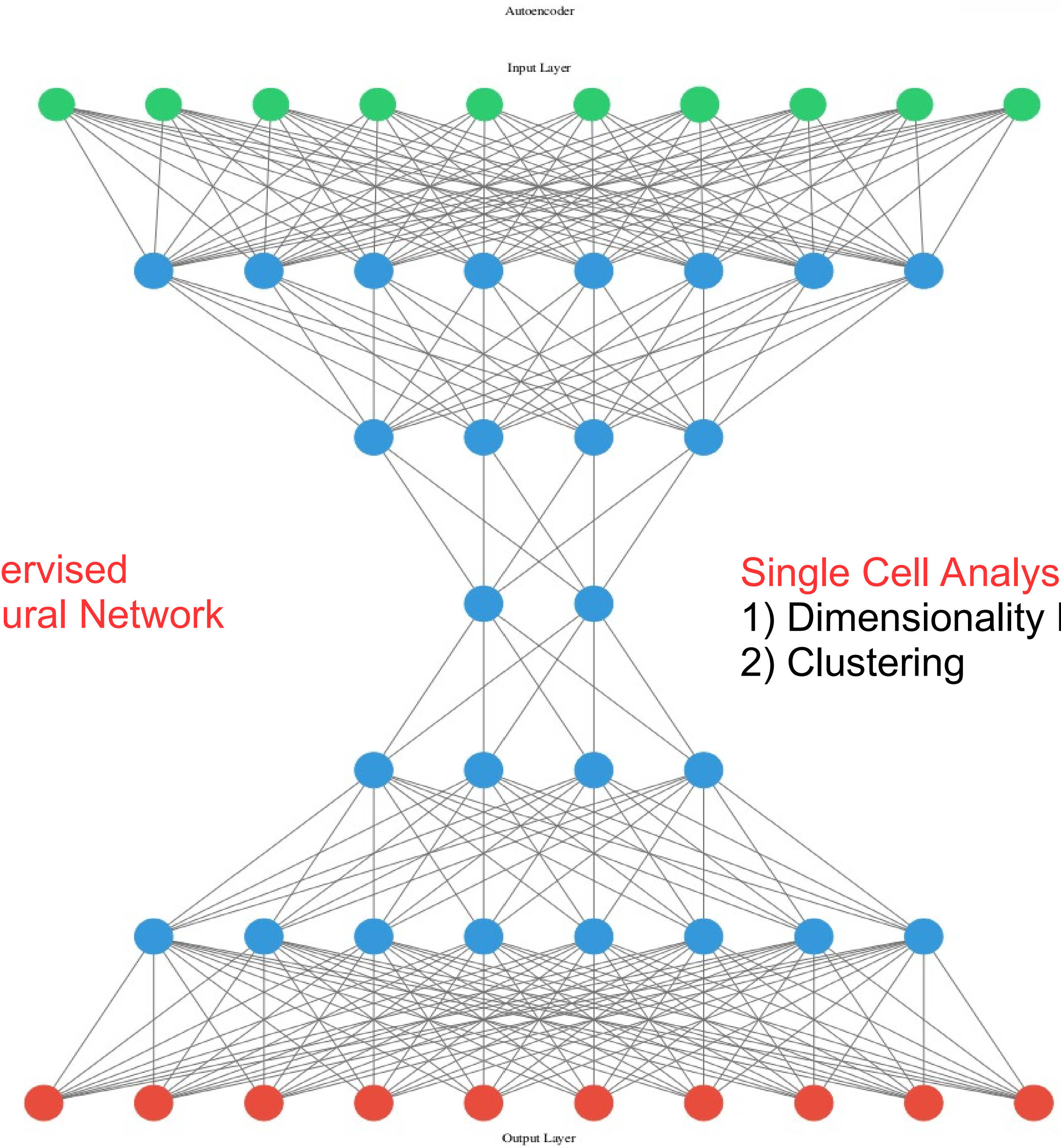
UMAP for Omics Integration







Autoencoder for Single Cell



Unsupervised
Artificial Neural Network

Single Cell Analysis is Unsupervised
1) Dimensionality Reduction
2) Clustering

NBIS

National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

SIGILLUM UNIVERSITATIS GOTHORUM CAROLINAE * AD VT RVMQVE 1666

LUNDS UNIVERSITET