

Unsupervised OMICs Integration

OMICs Integration and Systems Biology course

Nikolay Oskolkov, NBIS SciLifeLab

Lund, 6.09.2021

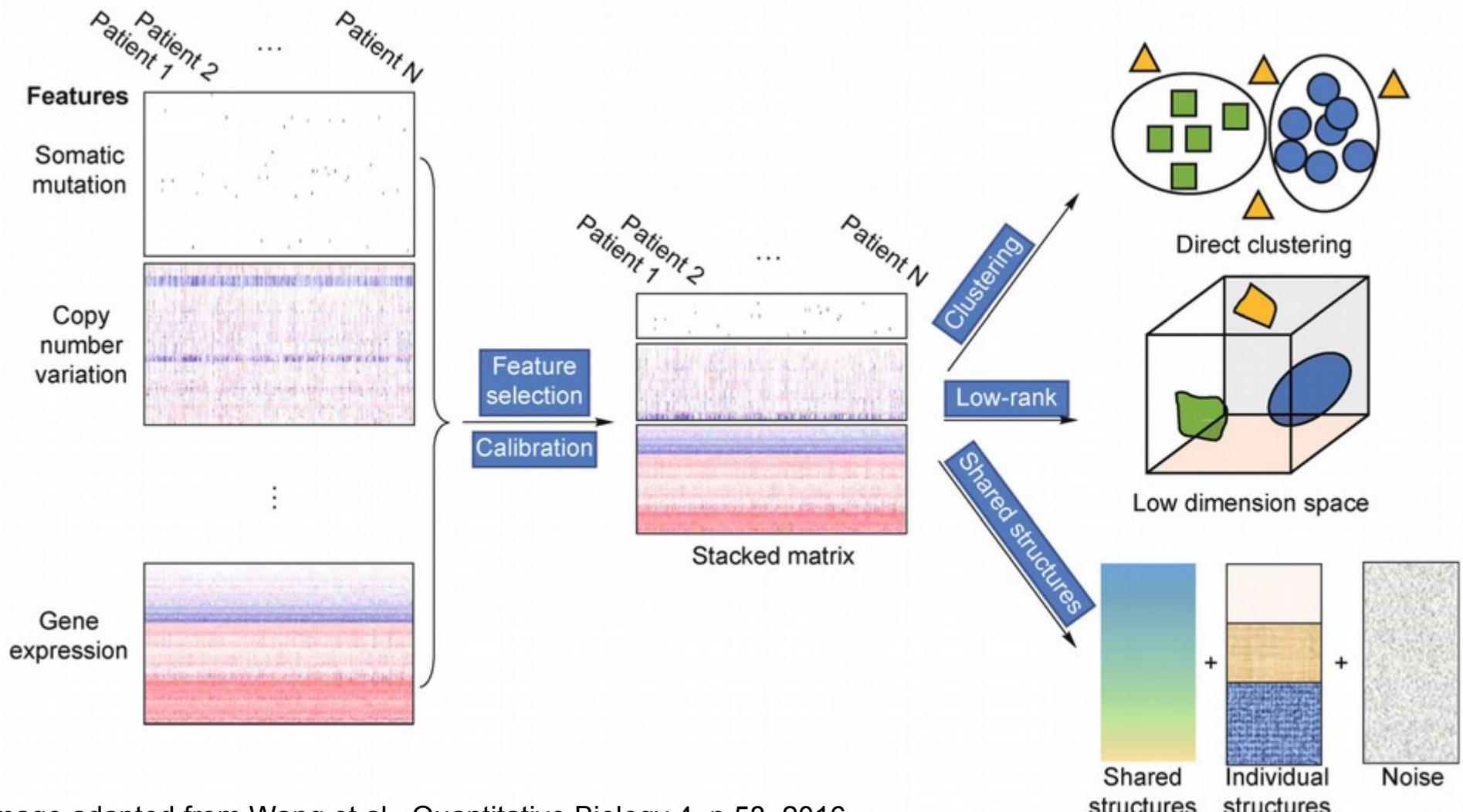
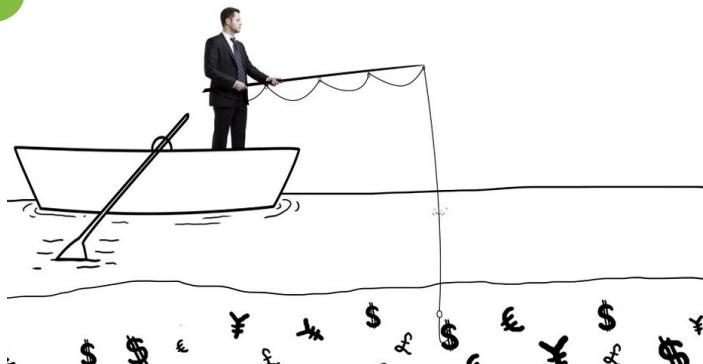


Image adapted from Wang et al., Quantitative Biology 4, p.58, 2016

Fishing expedition



BMC Part of Springer Nature

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

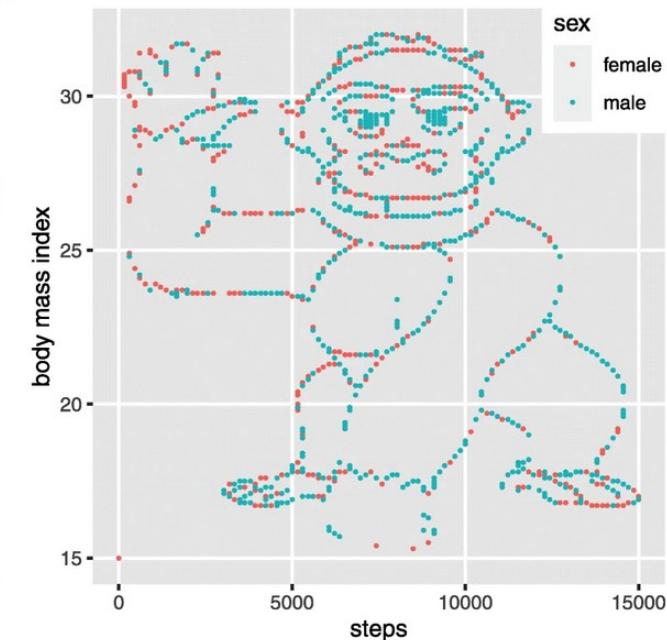
There is a hidden cost to having a hypothesis. It arises from the relationship between night science and day science, the two very distinct modes of activity in which scientific ideas are generated and tested, respectively [1, 2]. With a hypothesis in hand, the impressive strengths of day science are 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

- I do not understand your biological hypothesis
- I do not have any

a

ID	steps	bmi
3	15000	17.8
4	14861	17.2
5		
14		
15	15000	16.9
16	2	16.9
21	6	14861
23	7	16.8
26	8	14699
28	10	17.3
31	11	14560
33	13	20.5
34	17	14560
35	18	20.4
36	19	14560
38	20	19.8
39	22	14560
41	24	19.7
44	25	14560
45	27	19.6
46	29	14560
48	30	17.4
50	32	14560
52	37	20.9
53	40	14398
54	42	17.5
55	43	14398
56	43	17.1
57	43	14259
58	43	21.1
59	43	14259
60	43	21.1
61	43	14259
62	43	15.9

b

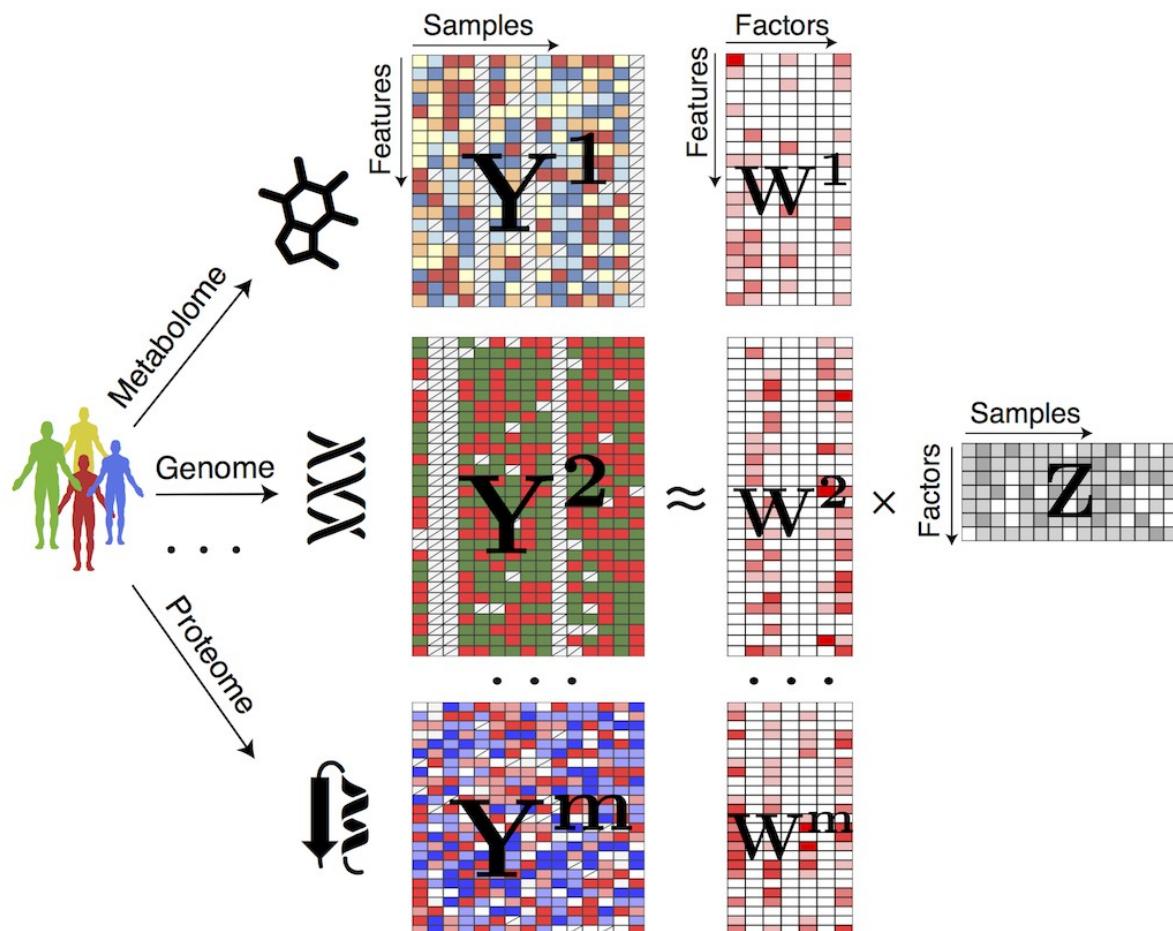


c

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

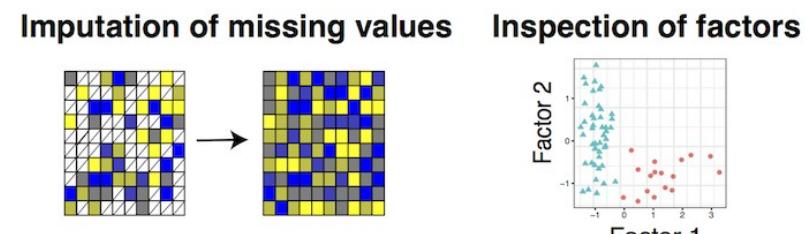
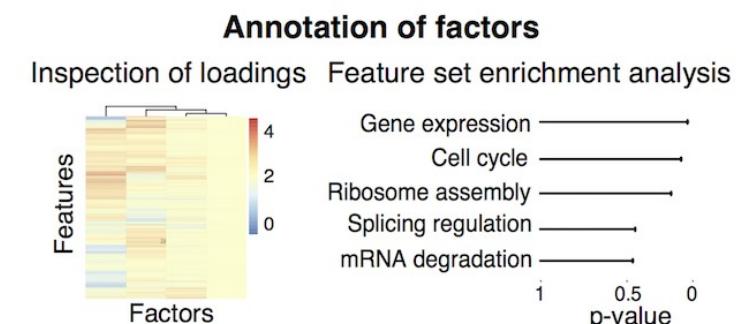
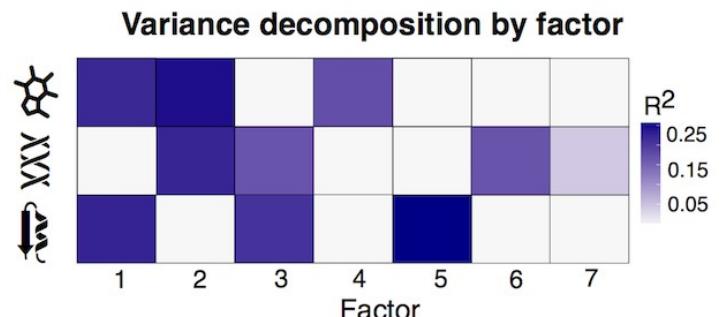
a An artificial dataset given to students with and without explicit hypotheses on the relationship between BMI and the steps taken on a particular day, for men and women. b A plot of the dataset. c The contingency table for students in the two groups ("hypothesis-focused," "hypothesis-free") that discovered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

Step 1: train a MOFA model



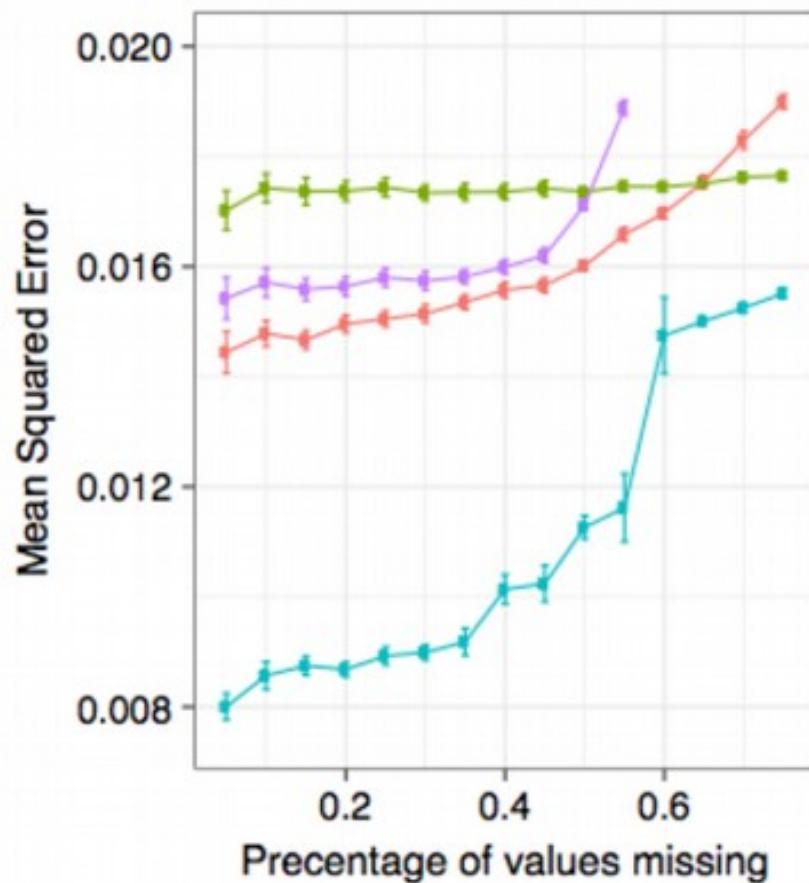
- Visualisation of samples in factor space
- Annotation of factors using (gene set) enrichment analysis
- Imputation of missing values
- Support of OMICs with non-Gaussian distribution including binary and count data

Step 2: downstream analysis

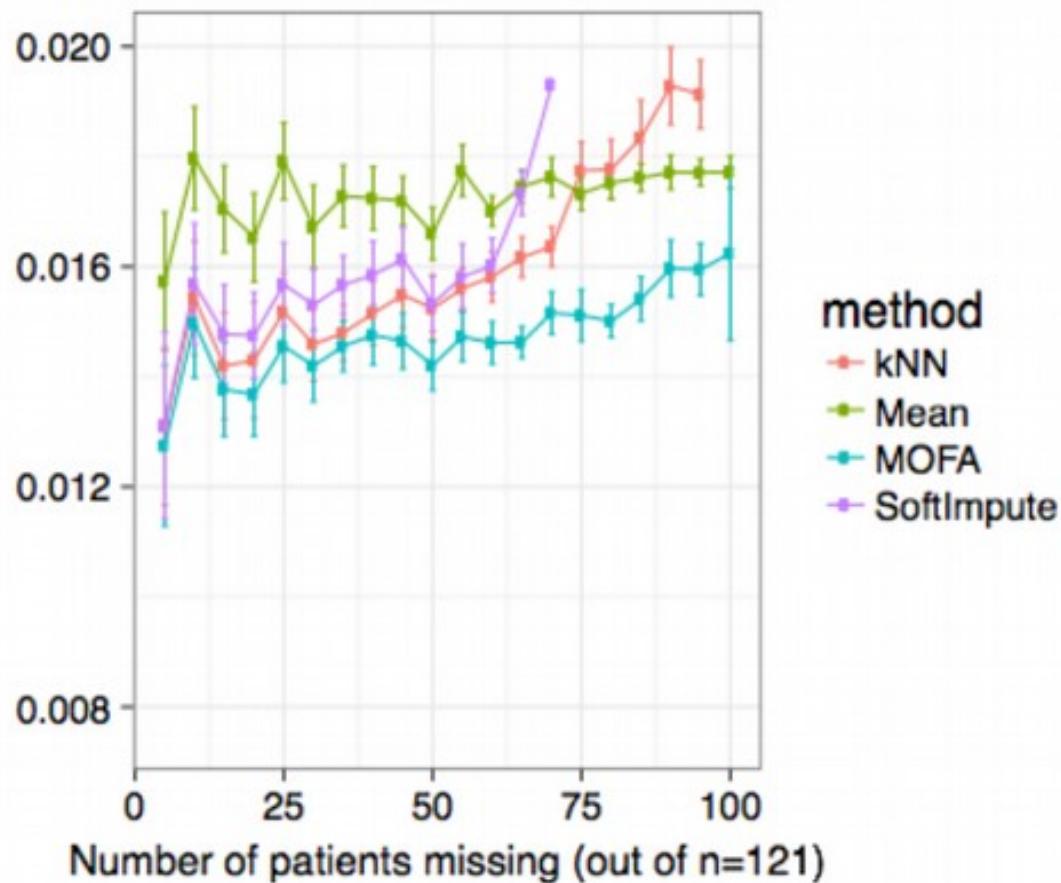


a

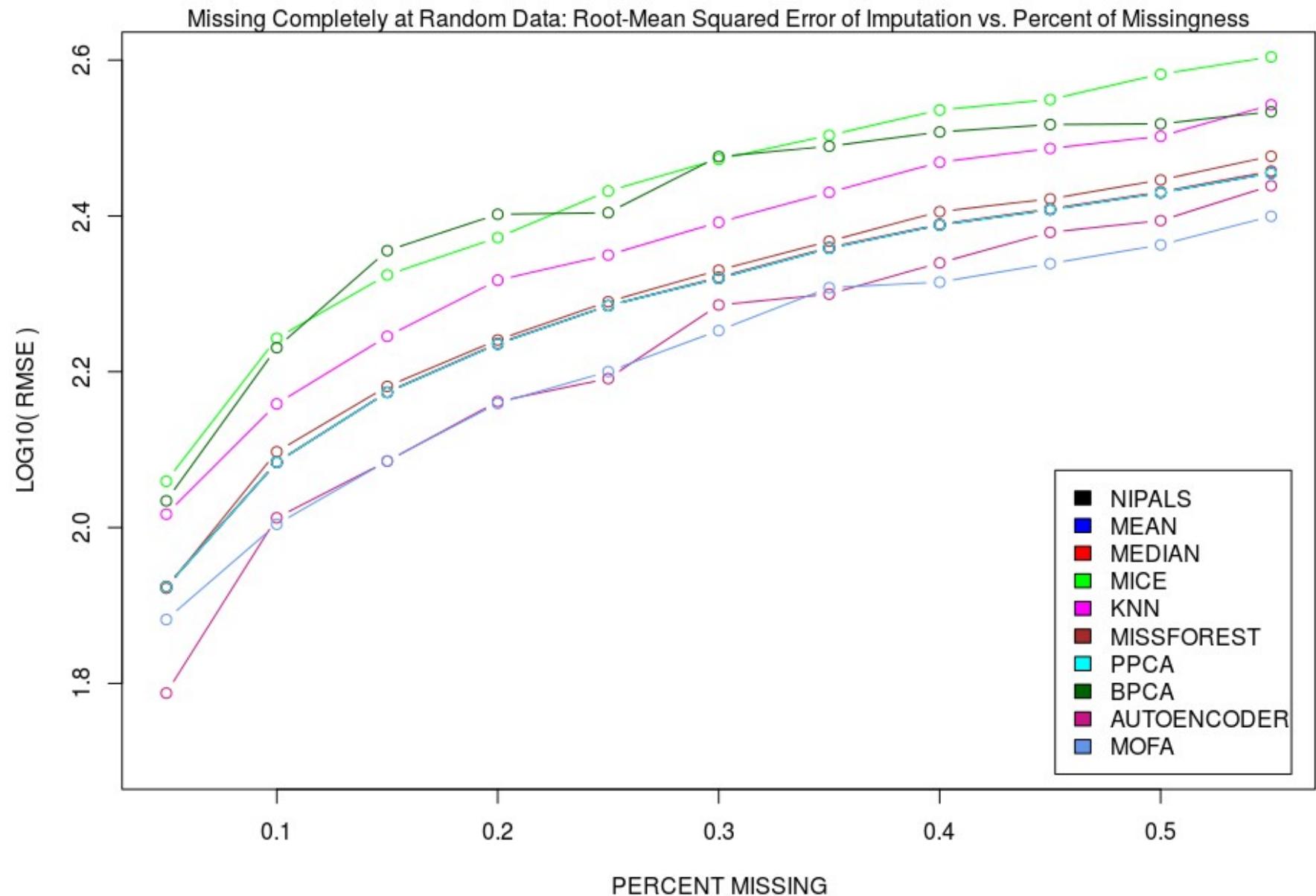
Values missing at random

**b**

Patients missing all measurements



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



ARTICLE

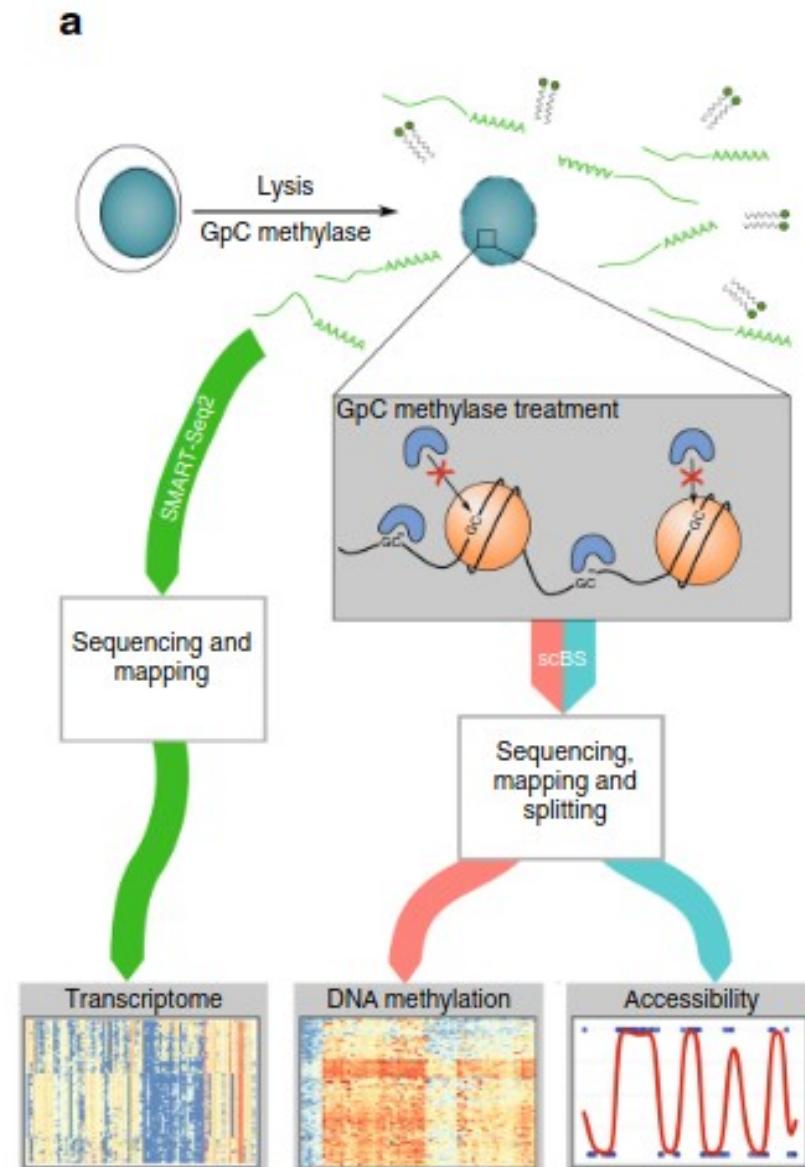
DOI: 10.1038/s41467-018-03149-4

OPEN

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

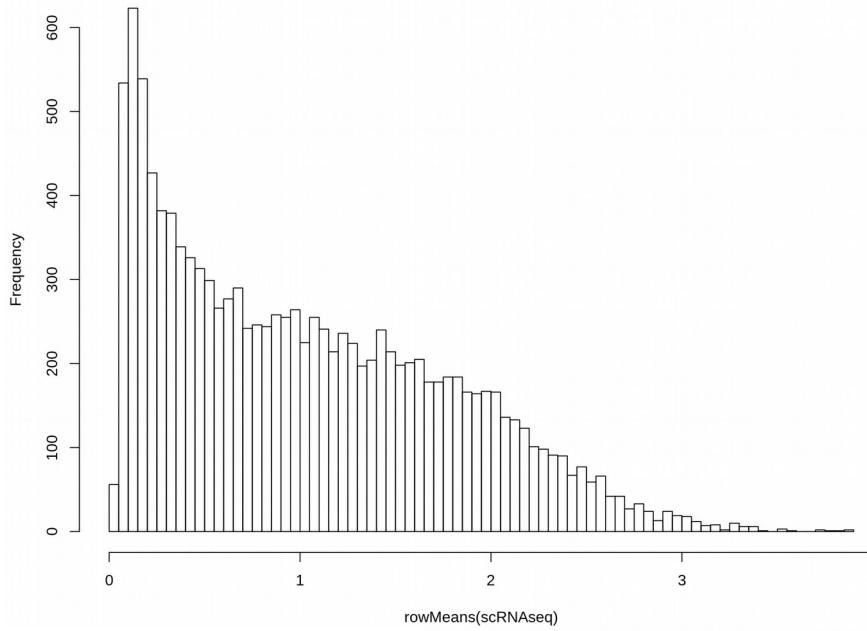
Stephen J. Clark¹, Ricard Argelaguet^{2,3}, Chantriolnt-Andreas Kapourani¹, Thomas M. Stubbs¹, Heather J. Lee^{1,5,6}, Celia Alda-Catalinas¹, Felix Krueger¹, Guido Sanguinetti⁴, Gavin Kelsey^{1,8}, John C. Marioni¹, Oliver Stegle¹, Wolf Reik^{1,5,8}

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.

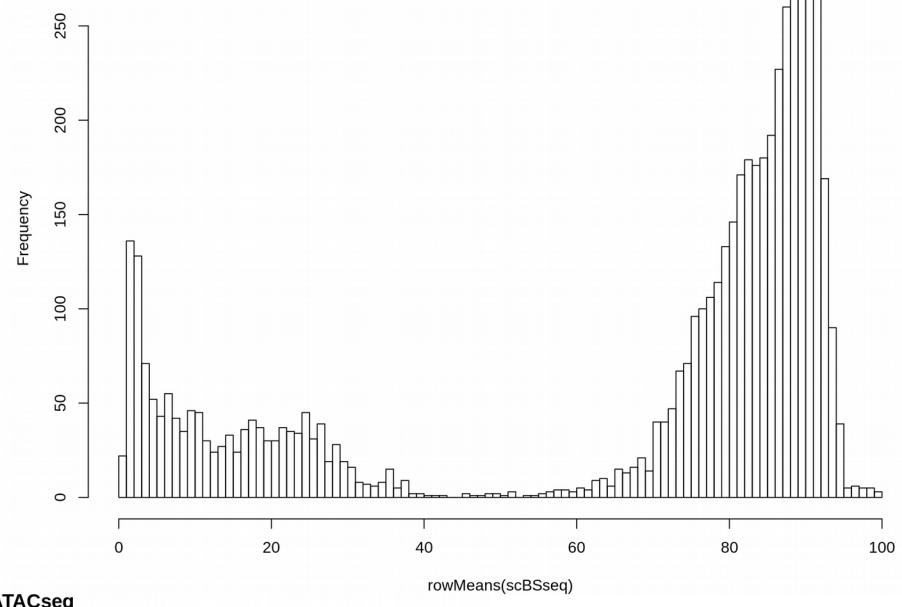


scNMT Data Set: Distributions

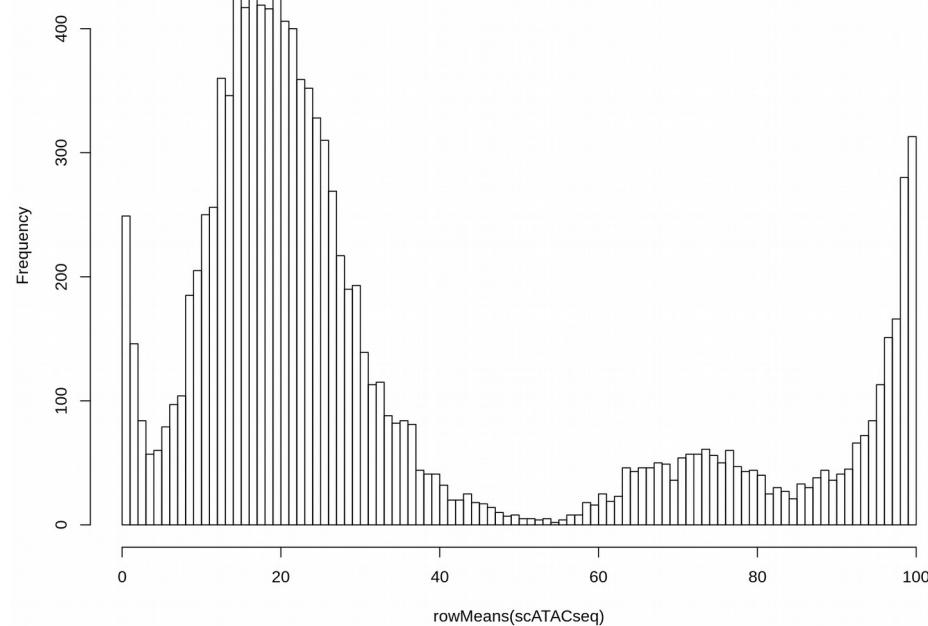
scRNAseq

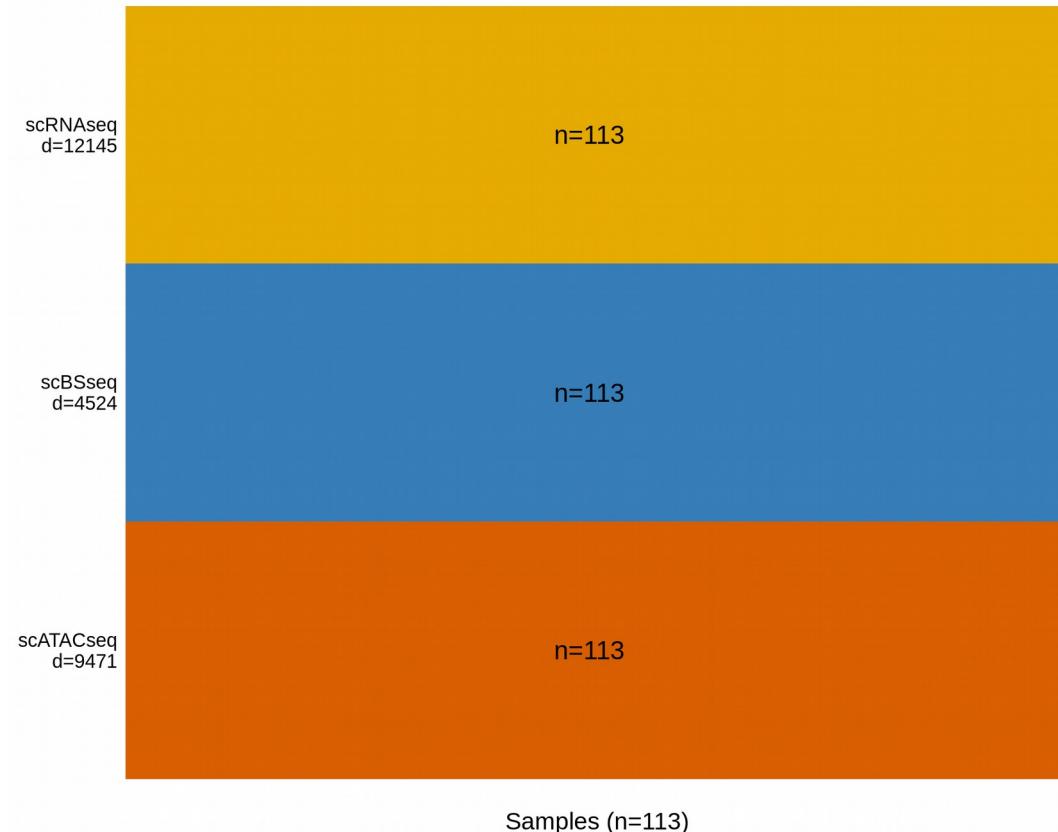


scBSseq



scATACseq





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

```

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](#)

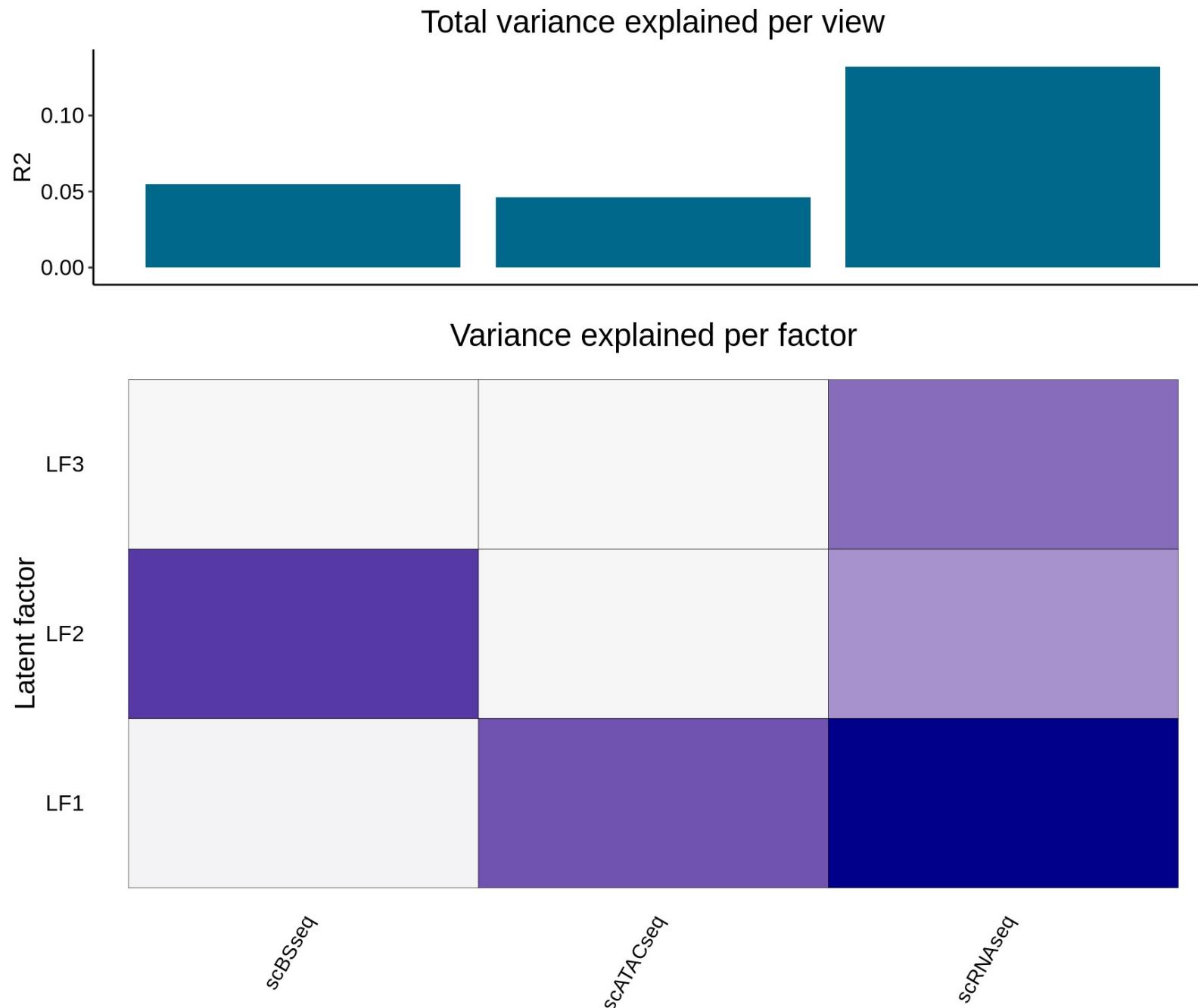
LIKELIHOOD
The probability of "B" being True, given "A" is True

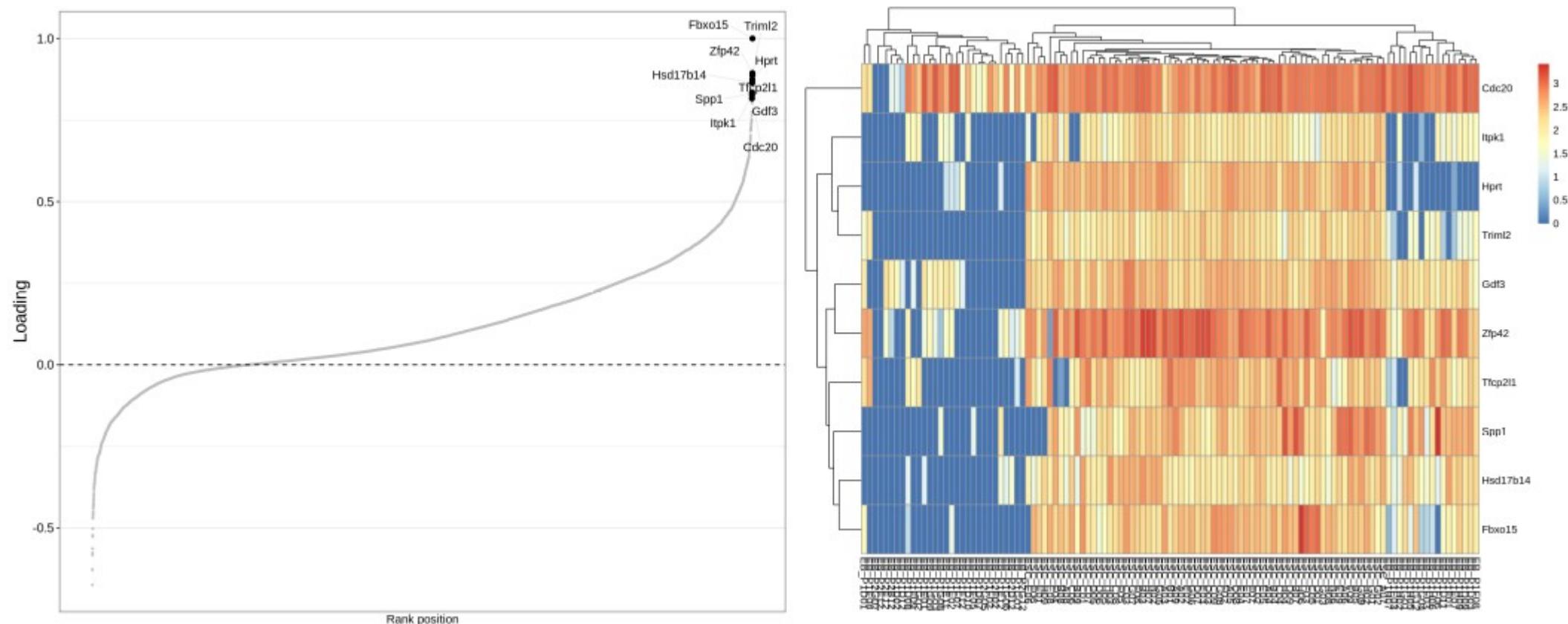
PRIOR
The probability "A" being True. This is the knowledge.

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

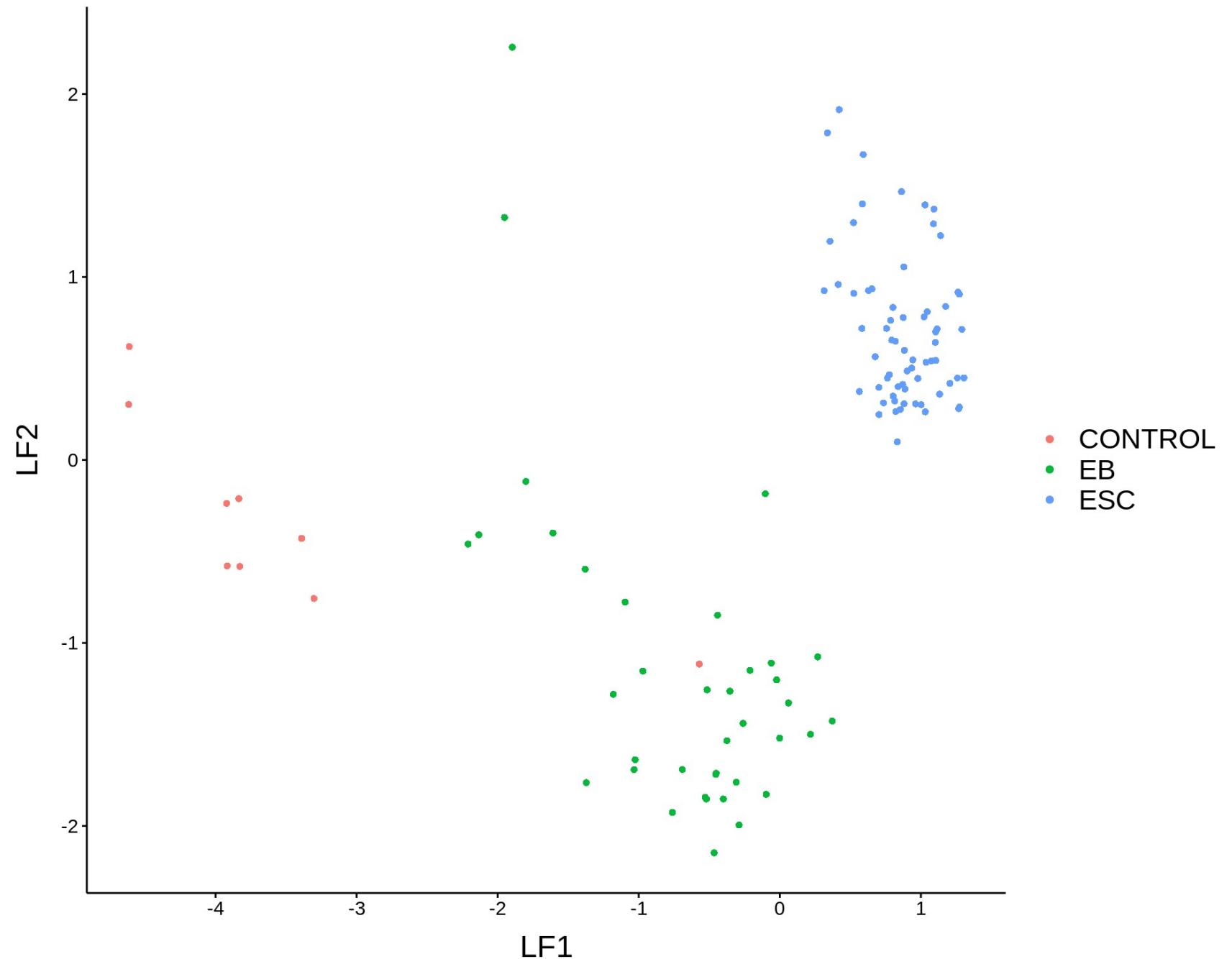
POSTERIOR
The probability of "A" being True, given "B" is True

MARGINALIZATION
The probability "B" being True.

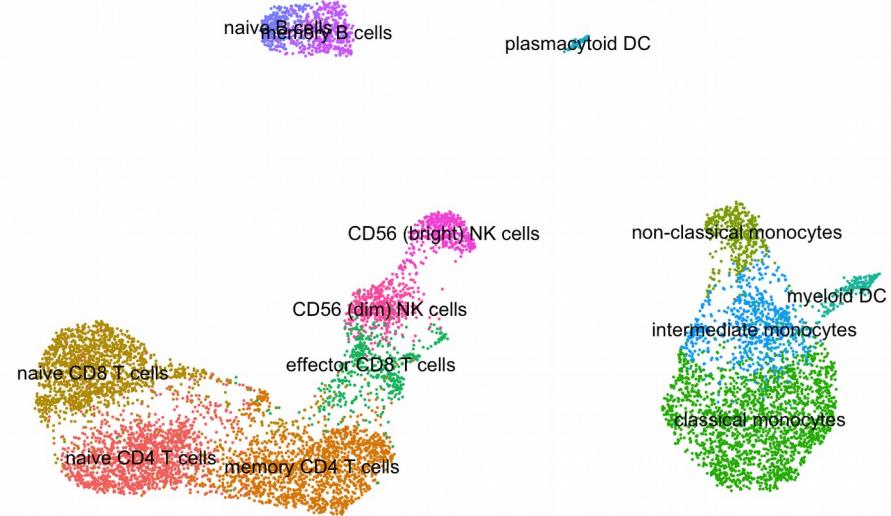




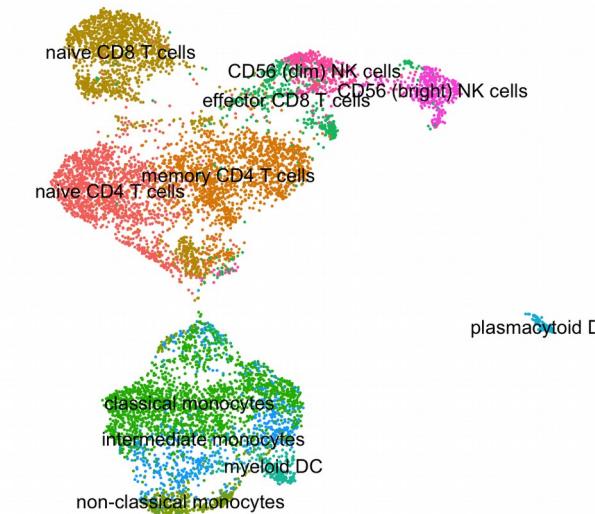
ESC and EB cells are separable on the heatmap built on loadings of the MOFA latent factors



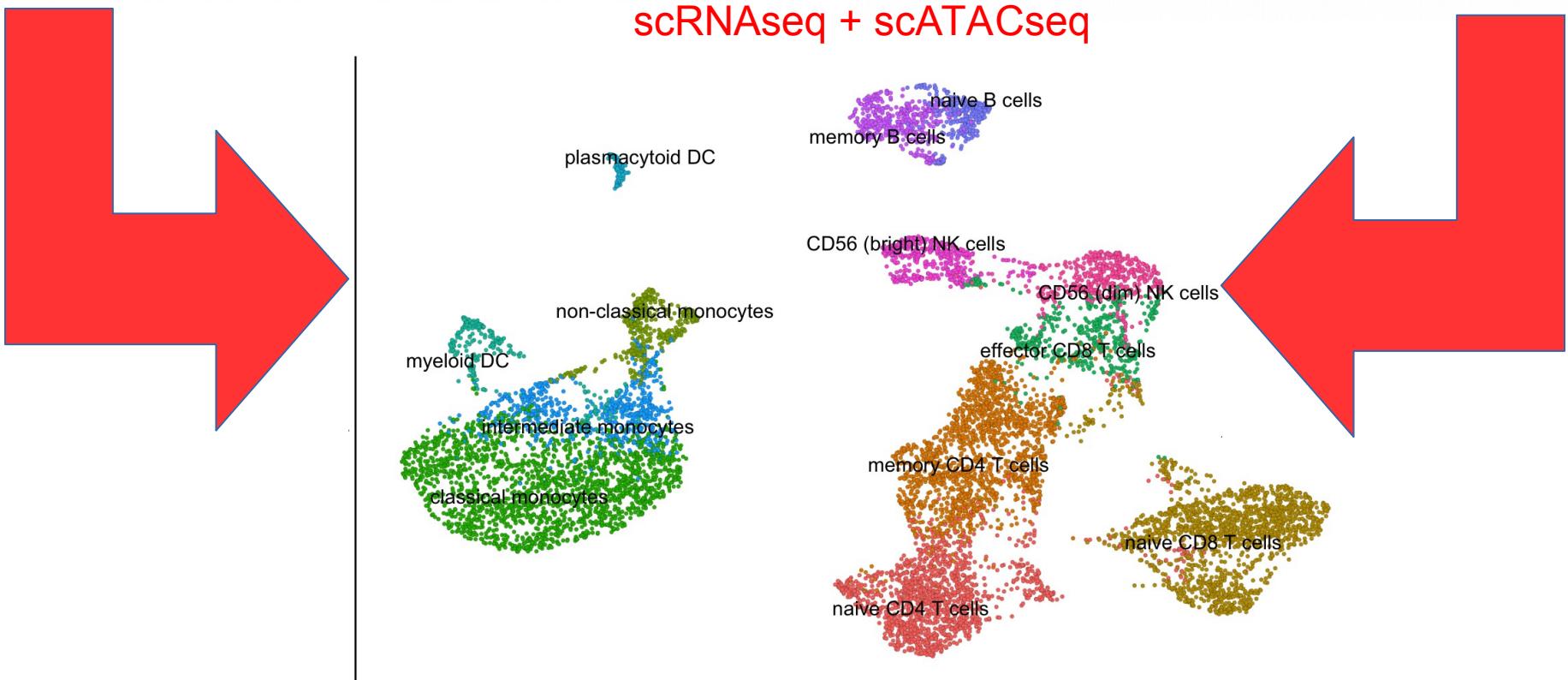
scRNAseq

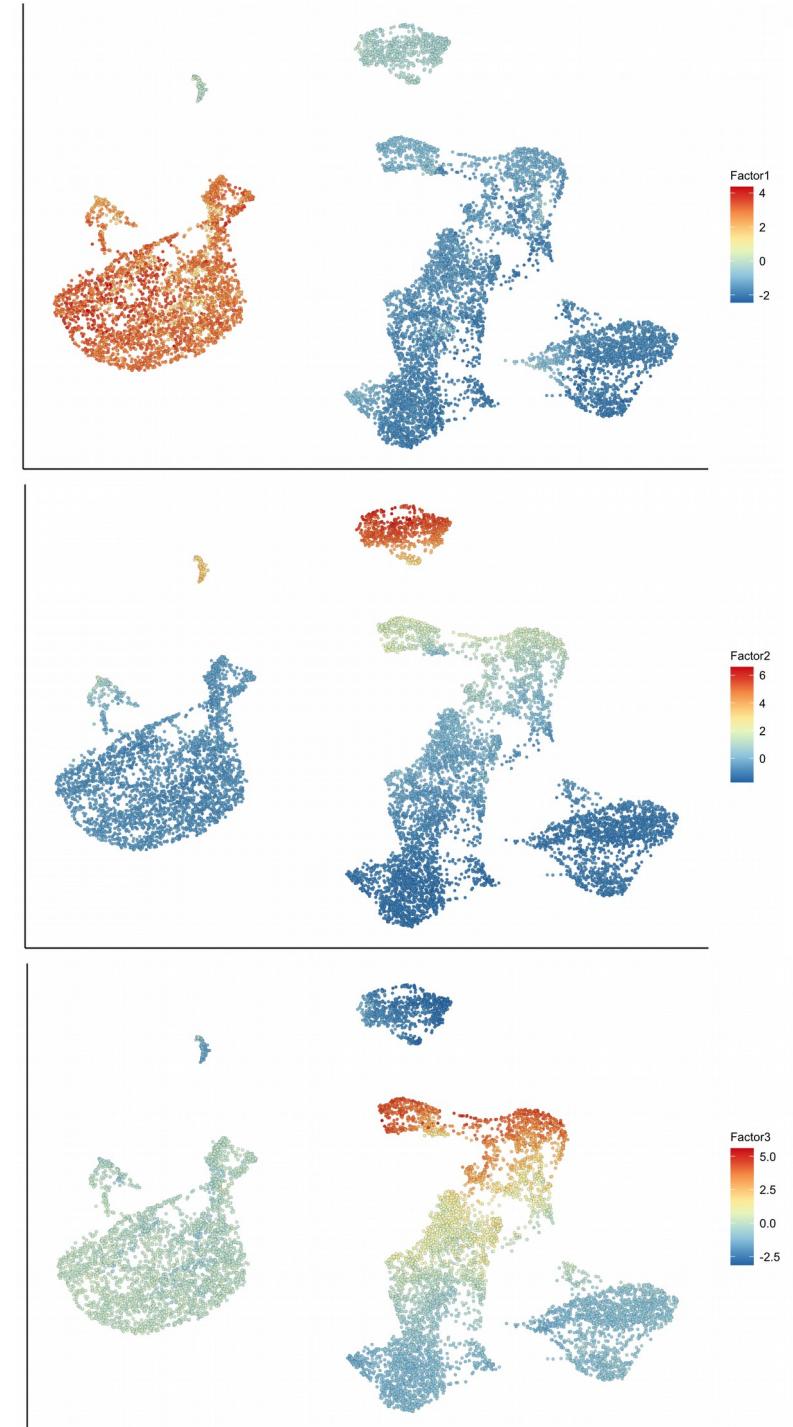
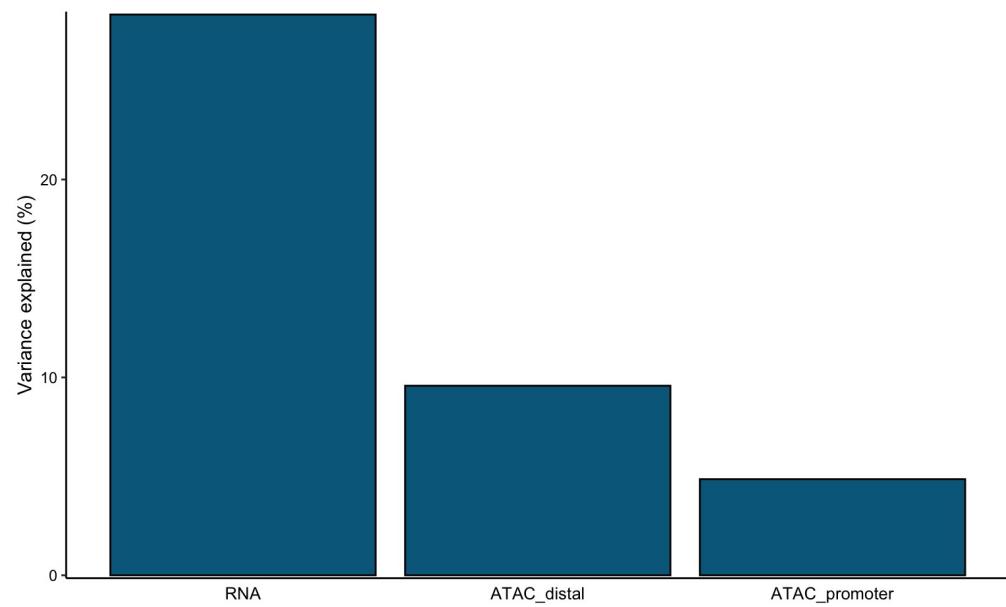


scATACseq



scRNAseq + scATACseq





Other Unsupervised Integrative OMICs Methods

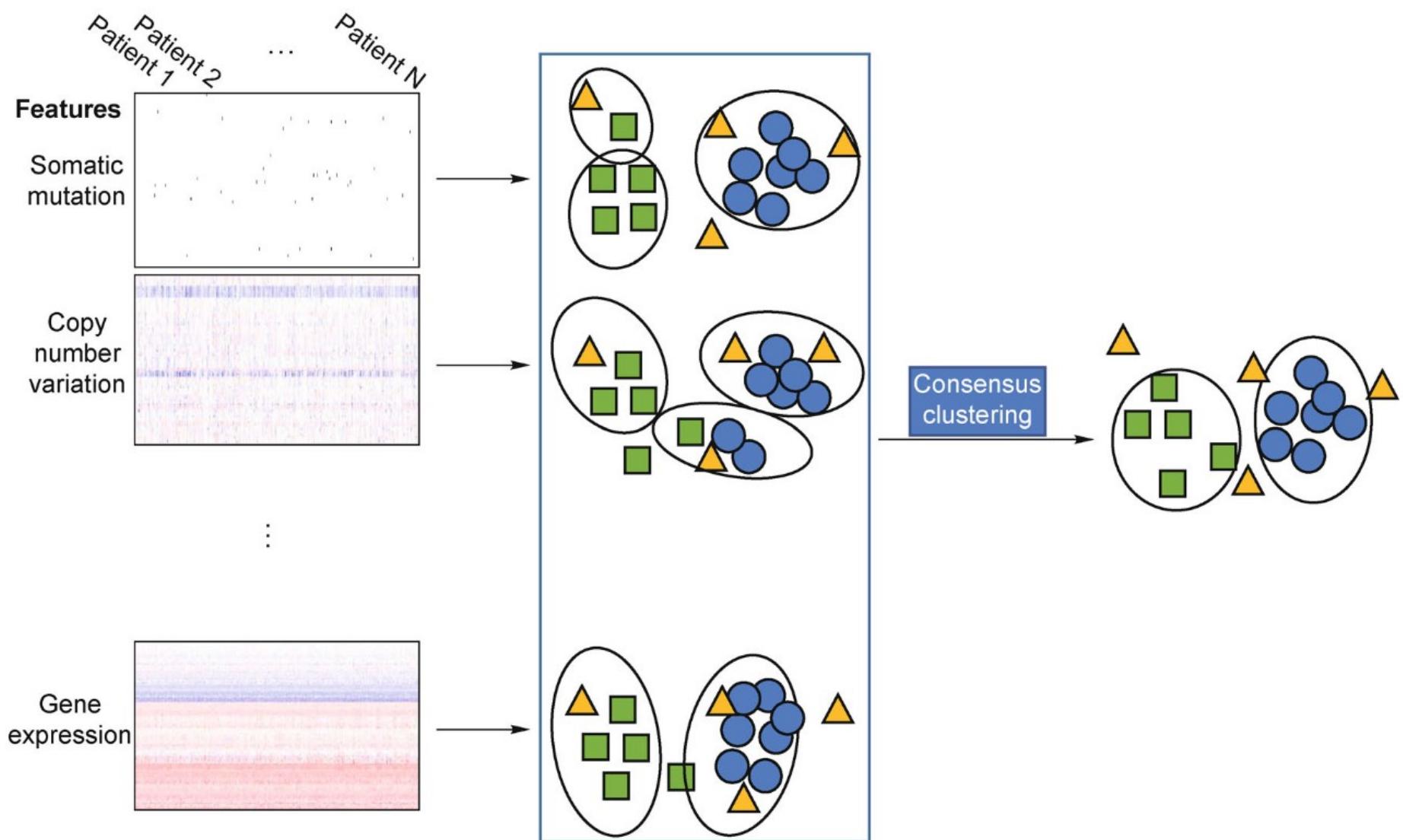
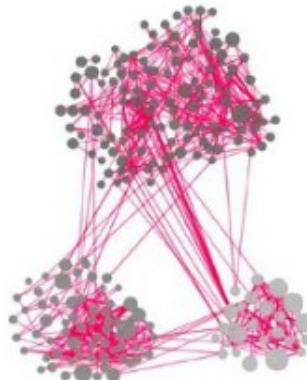
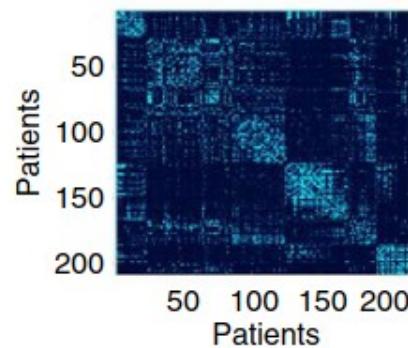
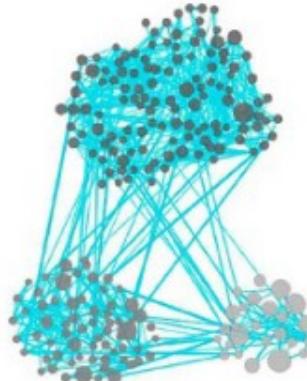
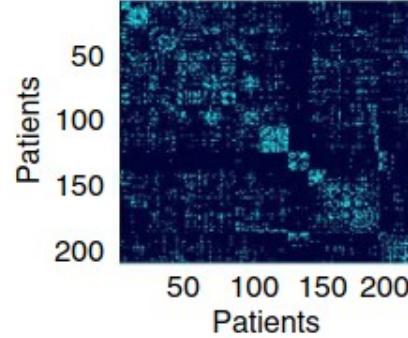
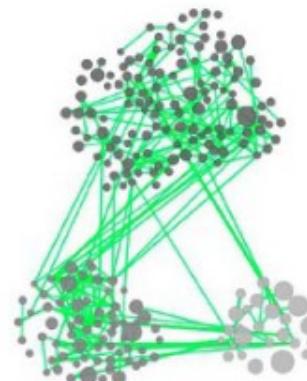
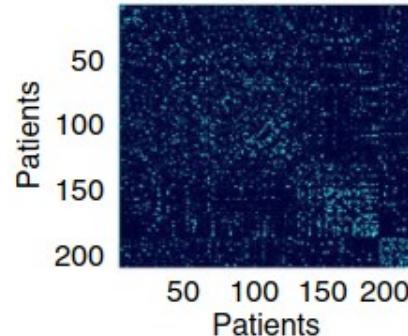
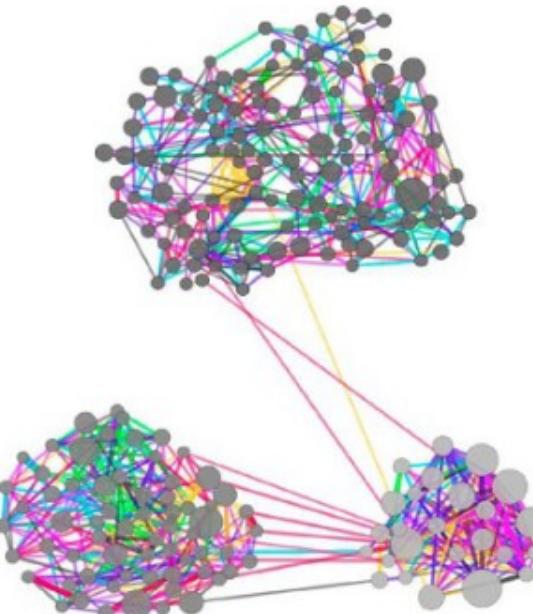
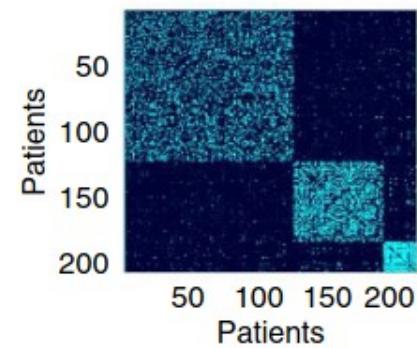


Figure 2. Clustering of clusters. This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.

a**b****c****d**

Patient subtype

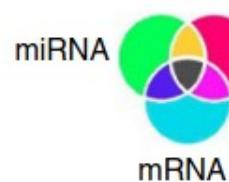
1 2 3

Survival (months)

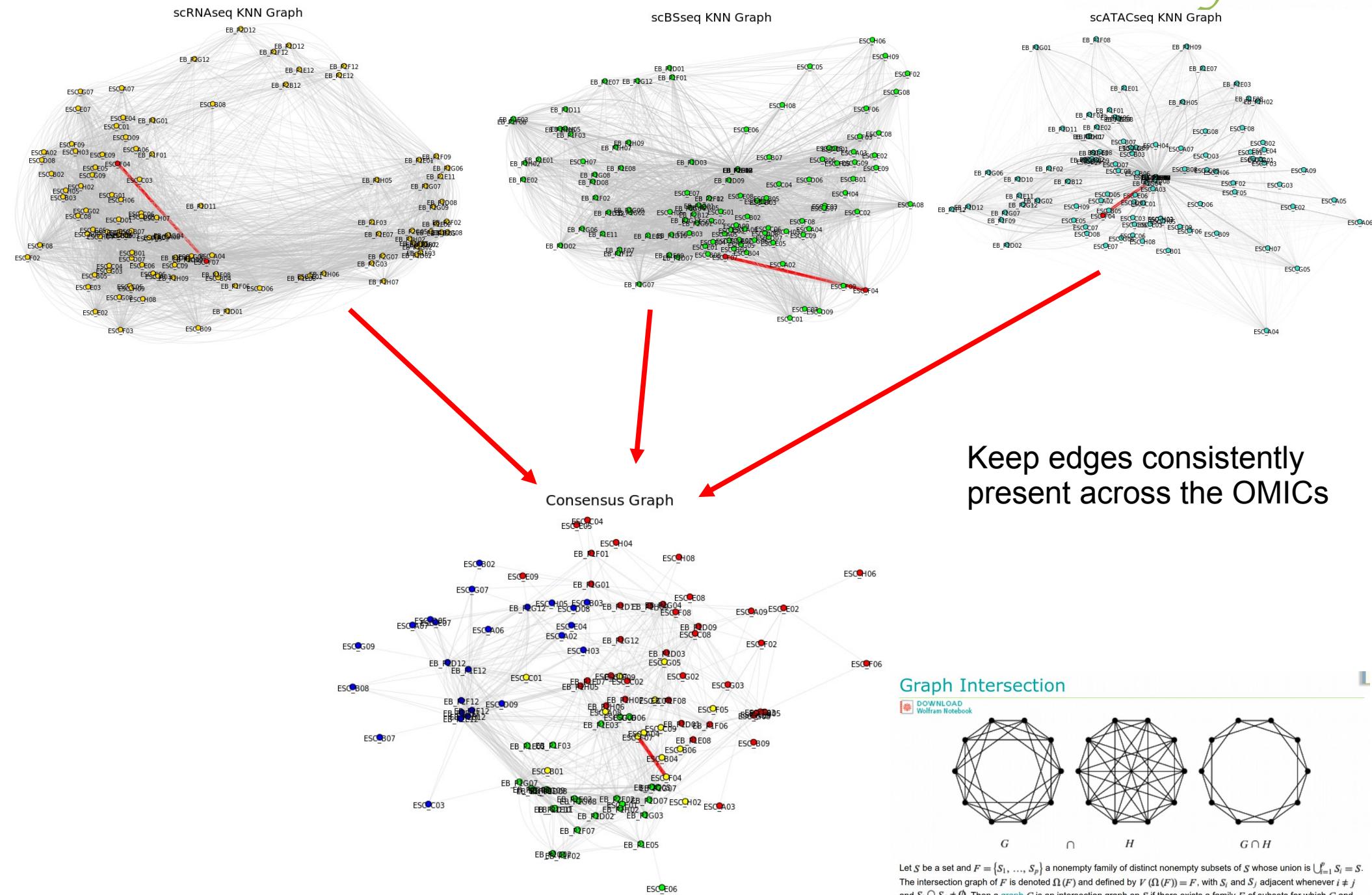
1 48 115

Similarly type

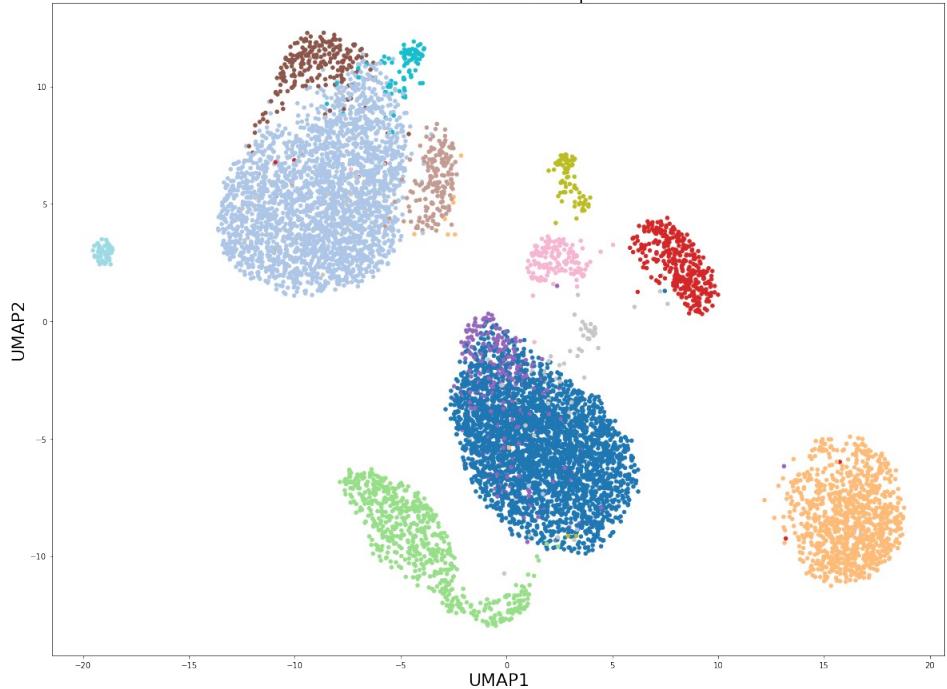
miRNA

DNA
methylation

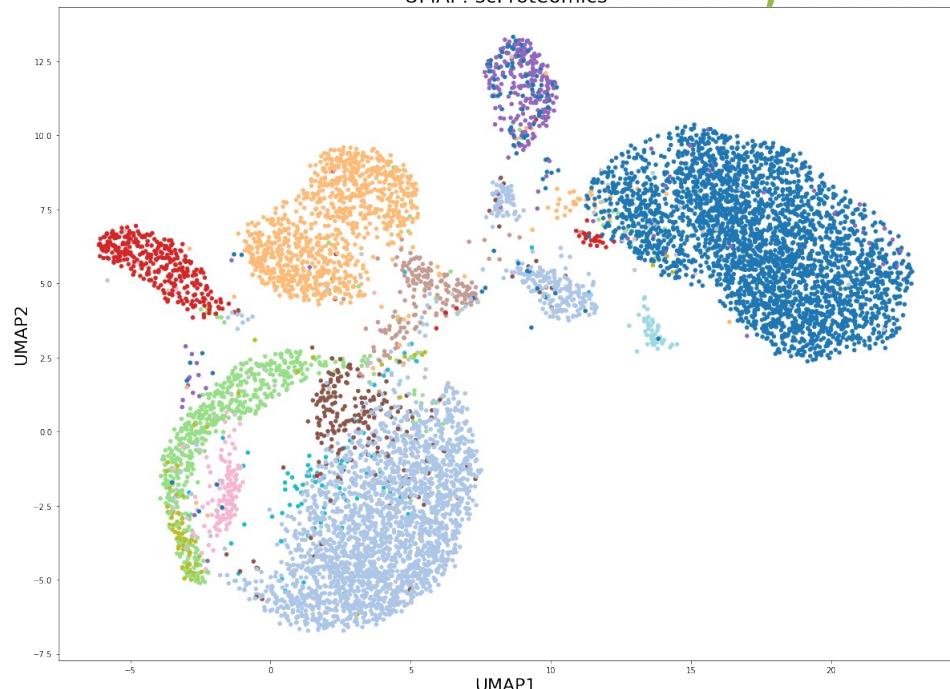
mRNA



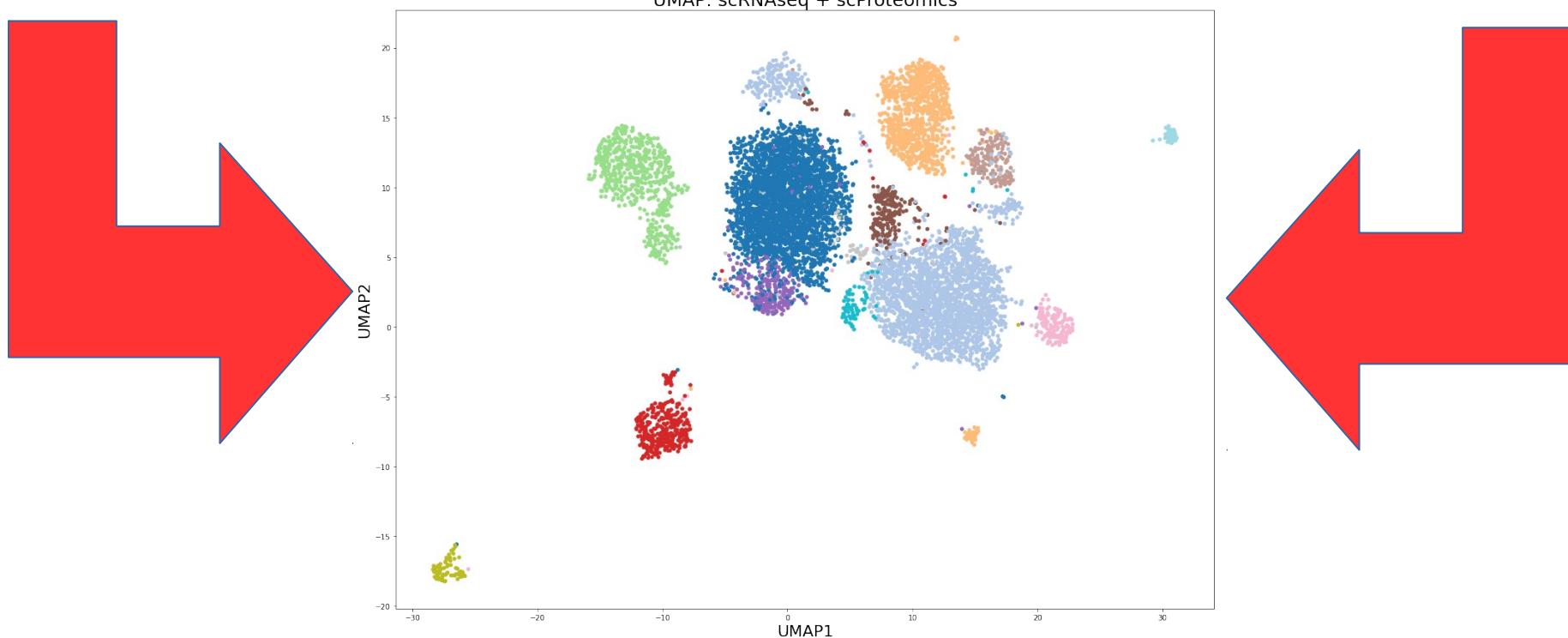
UMAP: scRNAseq

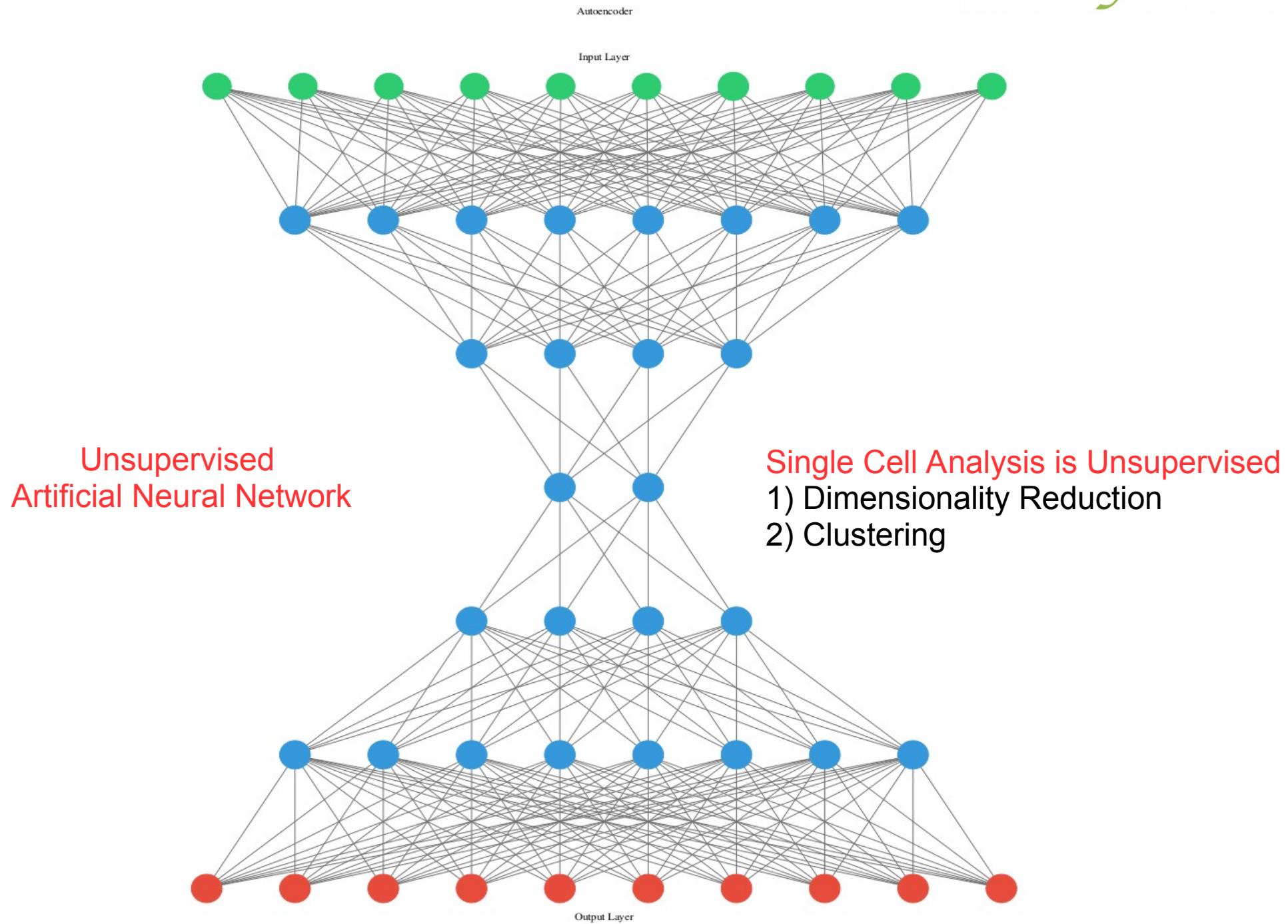


UMAP: scProteomics

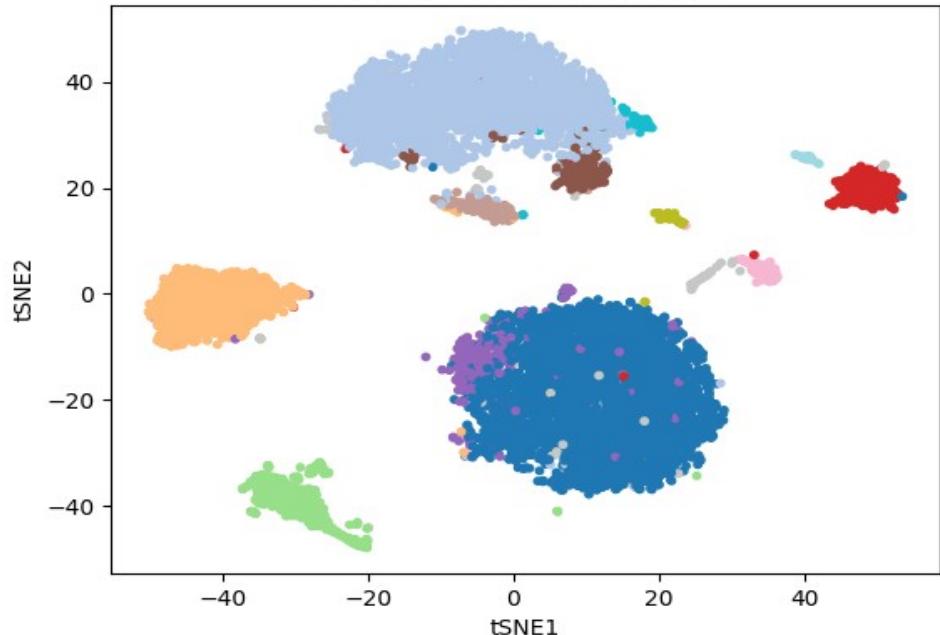


UMAP: scRNAseq + scProteomics

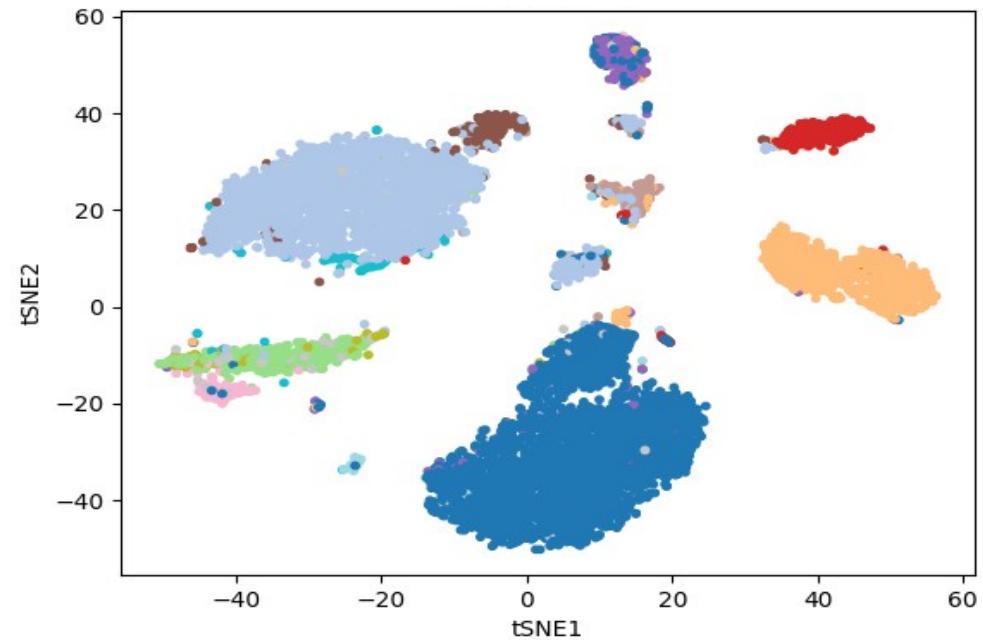




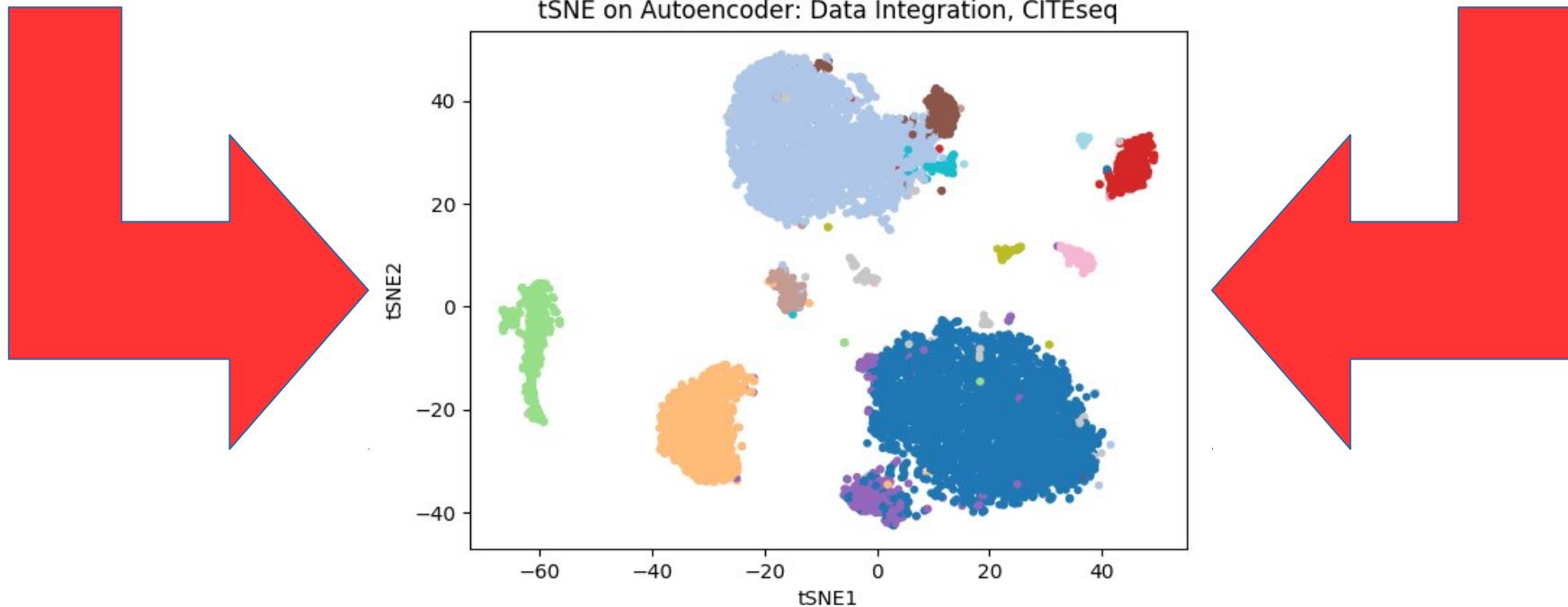
scRNAseq



scProteomics



tSNE on Autoencoder: Data Integration, CITEseq





National Bioinformatics Infrastructure Sweden (NBIS)

SciLifeLab



*Knut och Alice
Wallenbergs
Stiftelse*



LUNDS
UNIVERSITET