Bayesian Integration of Biological Data for Life Sciences

Bayes Lund 2023, Lund, Sweden, 23.01.2023
Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden

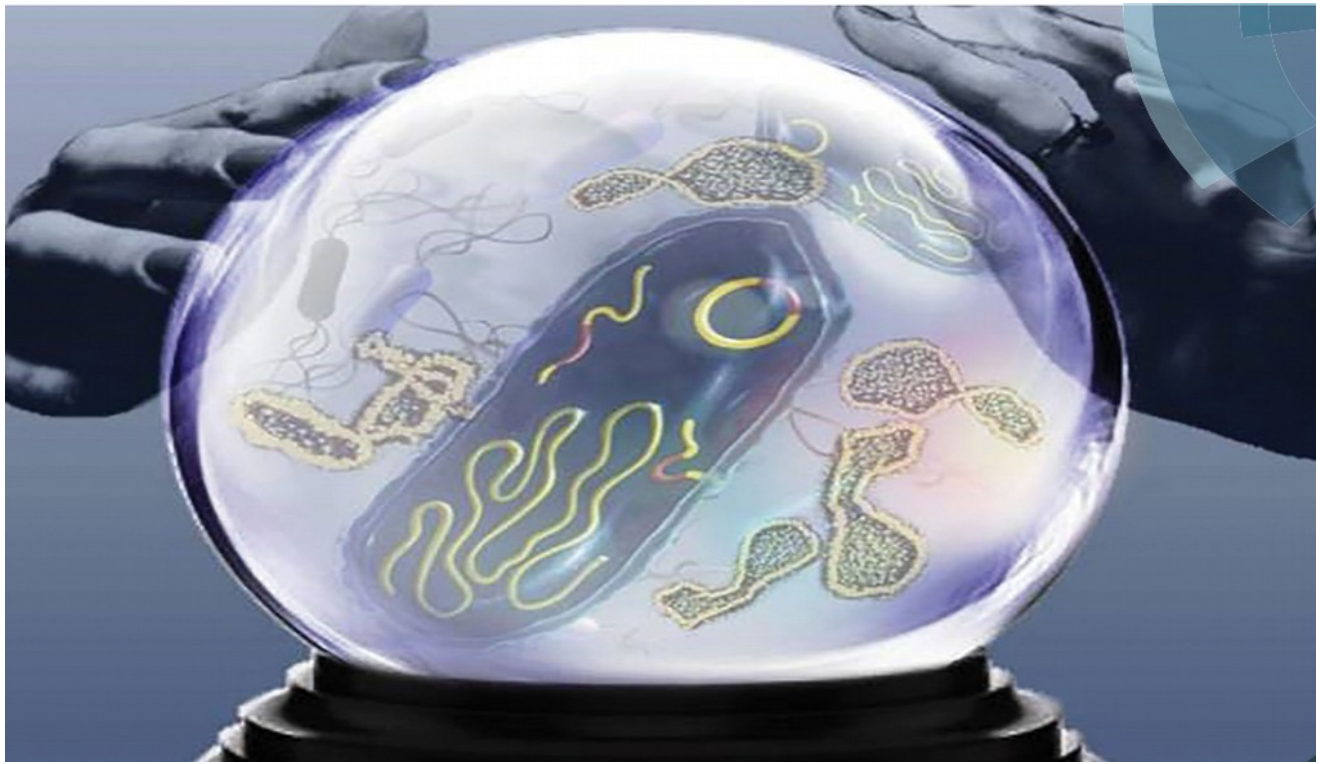
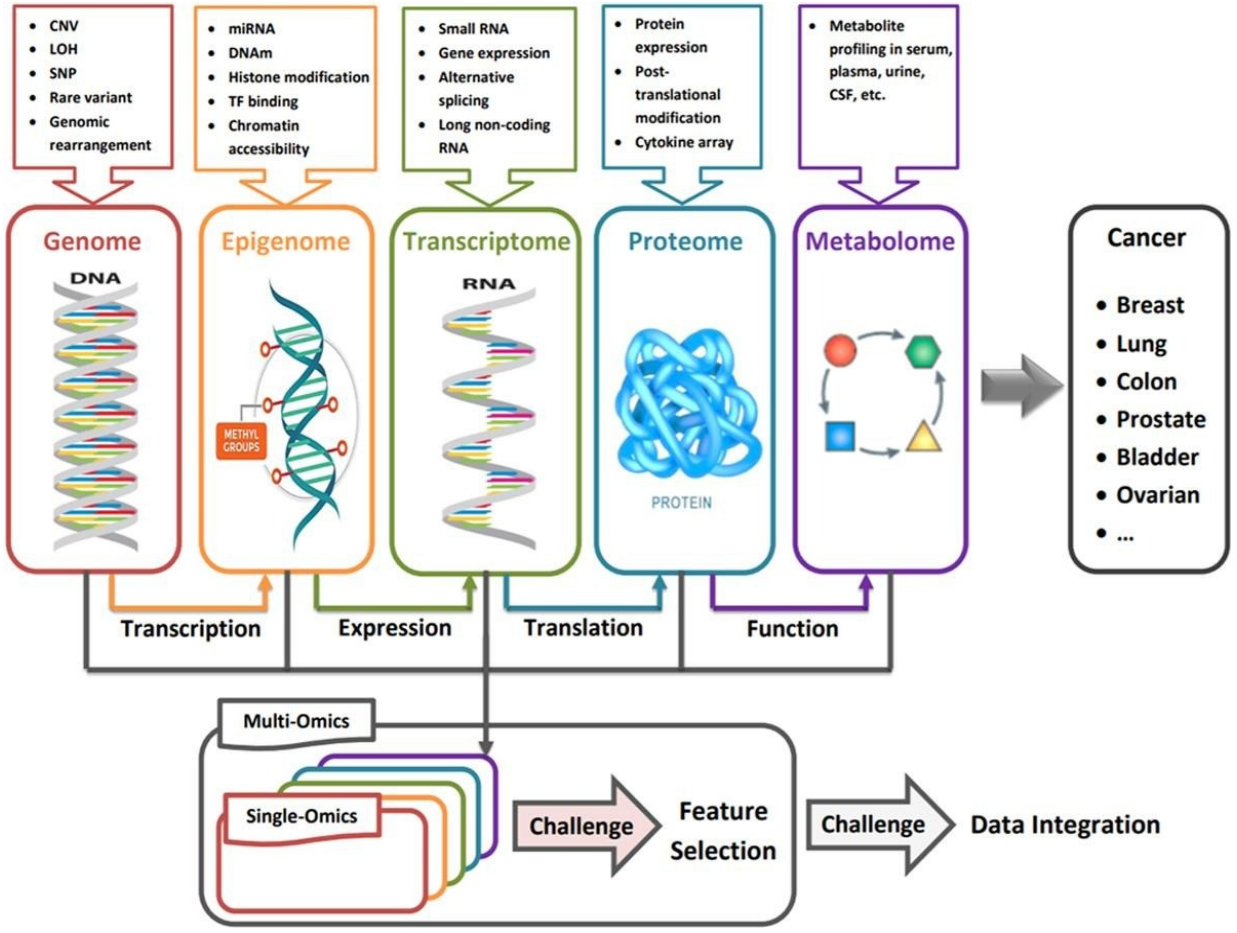
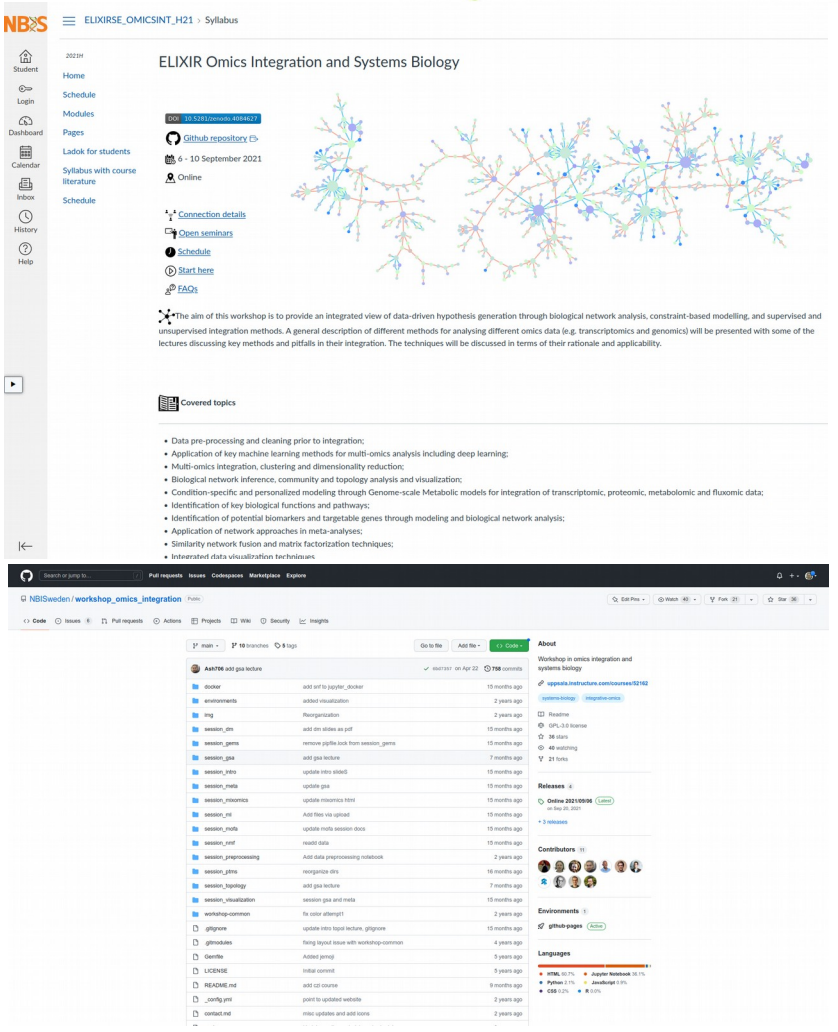


Image adapted from Molecular Omics, Issue 1, 2018

Multi-Omics Data in Life Sciences





Momeni et al., J Biomed Inform. 2020

https://github.com/NBISweden/workshop_omics_integration

Biological Data are High Dimensional

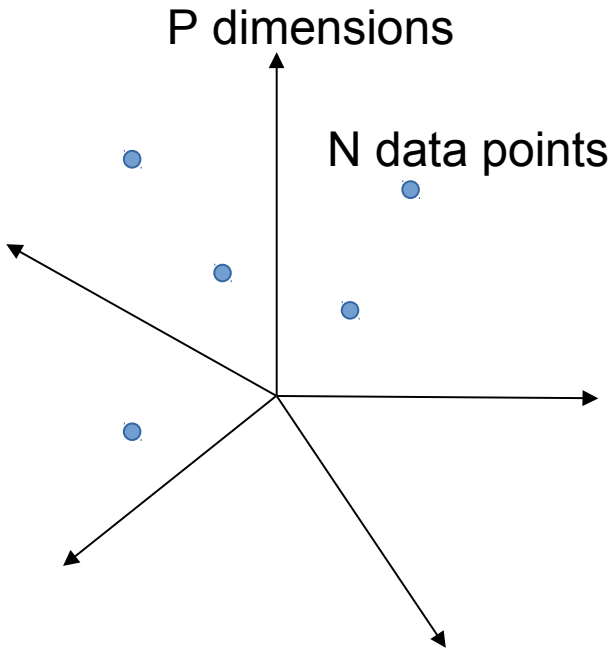
Statistical observations:
e.g. samples, cells etc.

N

P

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

Features: genes, proteins, microbes, metabolites etc.



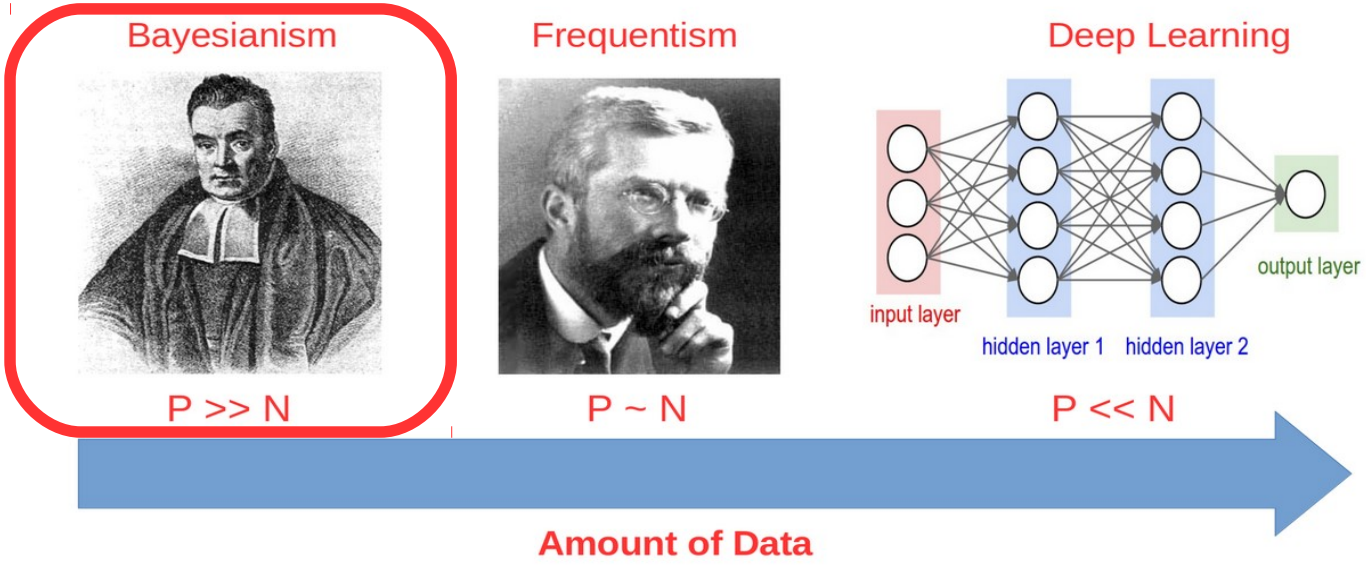
High Dimensional Data:
P >> N

For a robust statistical analysis, one should properly “sample” the P-dimensional space, hence large sample size is required, N >> P

Data Dimensionality and Choice of Analysis

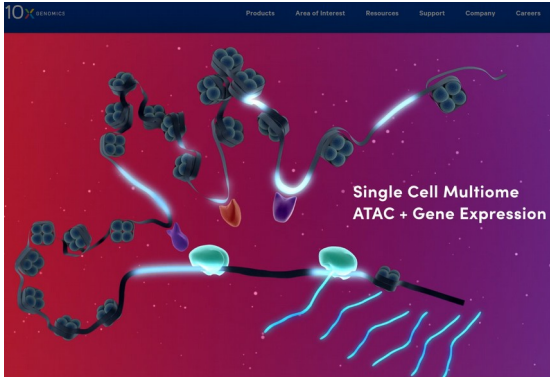
P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biology / Biomedicine



Big Data in Single Cell Genomics:

It becomes common to work with millions of statistical observations



nature methods

Explore content | About the journal | Publish with us

nature › nature methods › editorials › article

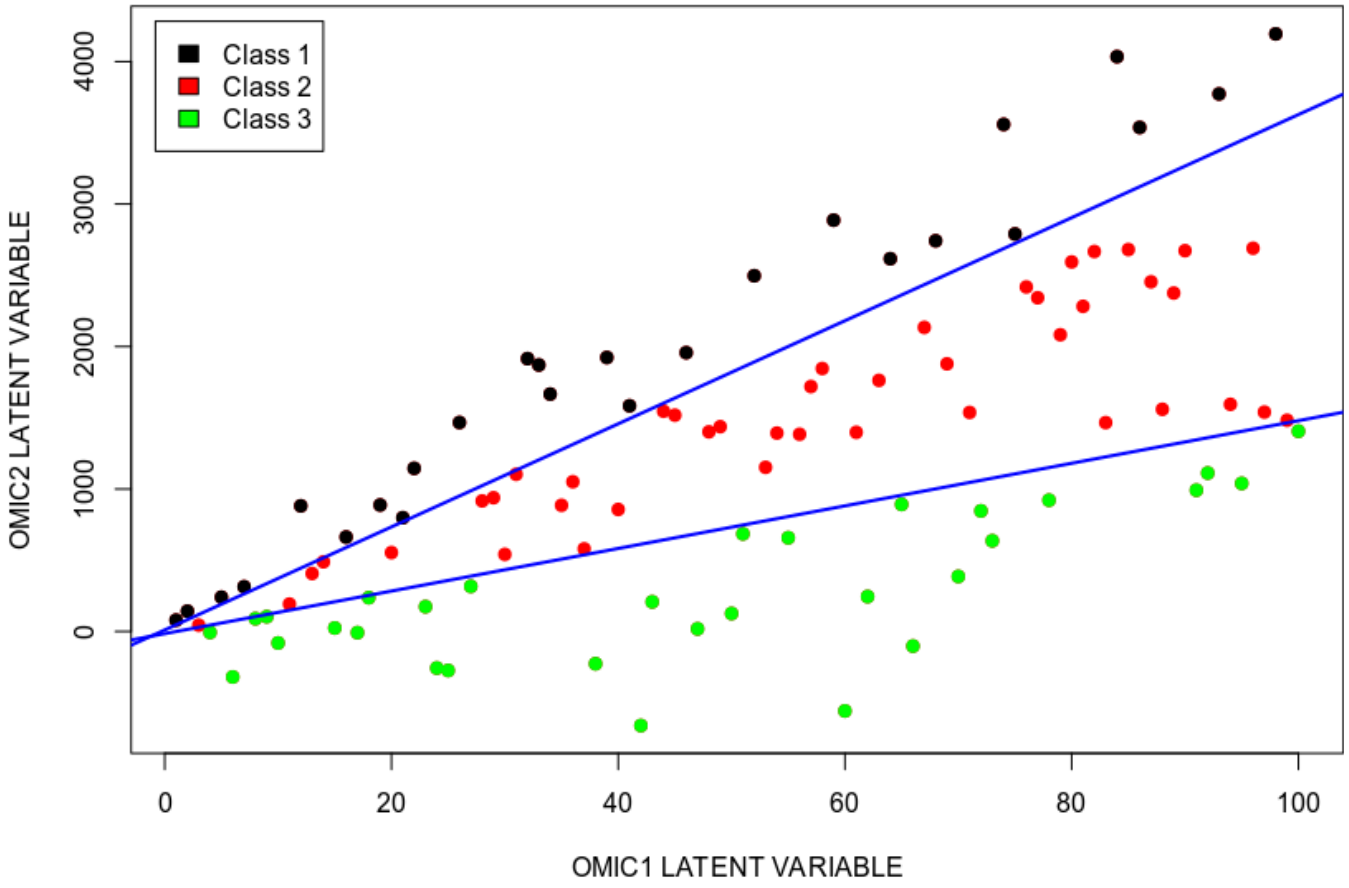
Editorial | Published: 06 January 2020

Method of the Year 2019: Single-cell multimodal omics

Nature Methods 17, 1 (2020) | Cite this article

40k Accesses | 54 Citations | 133 Altmetric | Metrics

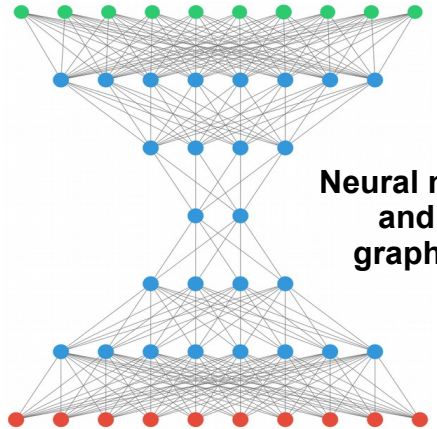
What we mean by data integration



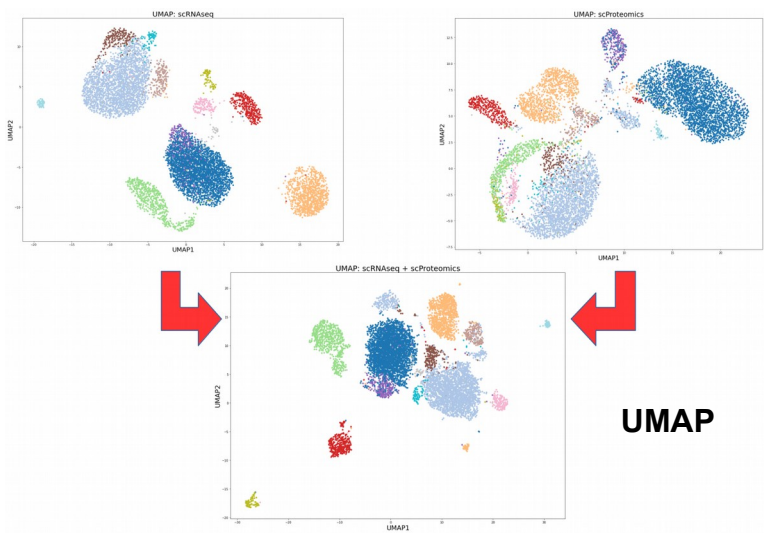
The idea behind Omics integration: see patterns hidden in individual Omics

Overview of Integrative Methods

Convert to common space

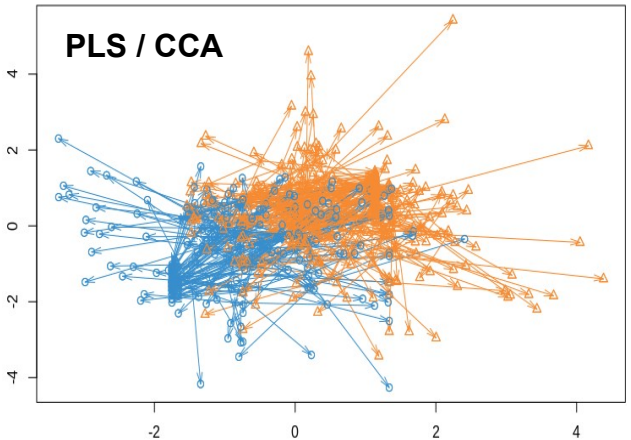


Neural nets and graphs

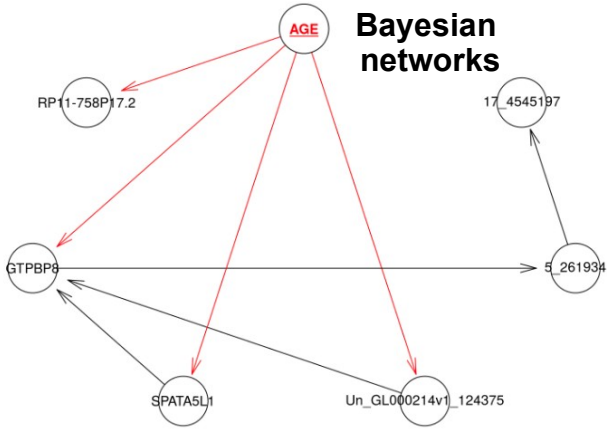


UMAP

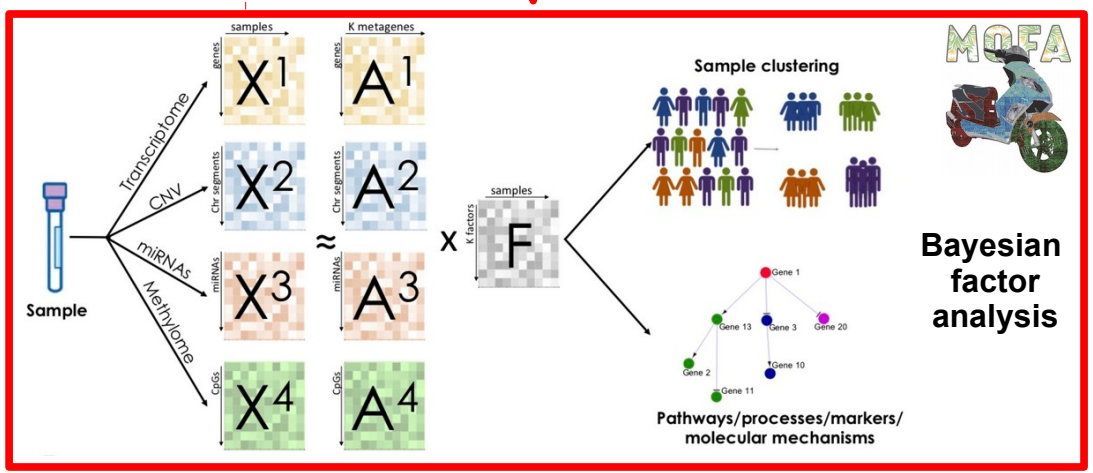
Extract common variation



Combine via Bayes rule

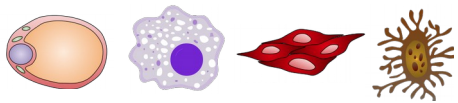


Bayesian networks

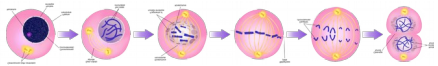


Bayesian factor analysis

Factor Analysis: Common Axes of Variation



CELL TYPE



CELL CYCLE

L_1

L_2

P_1

N

P_2

N

P_3

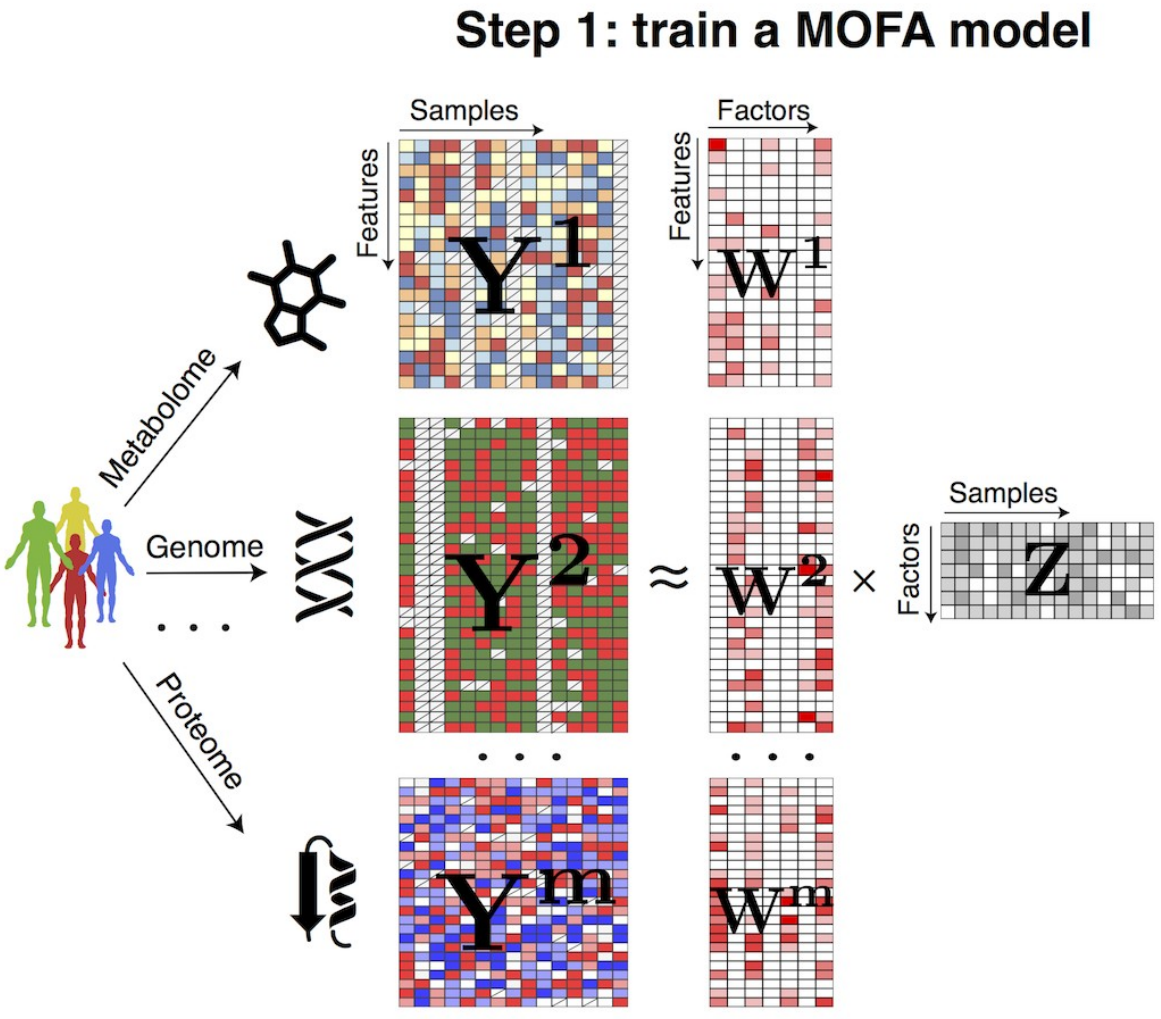
N

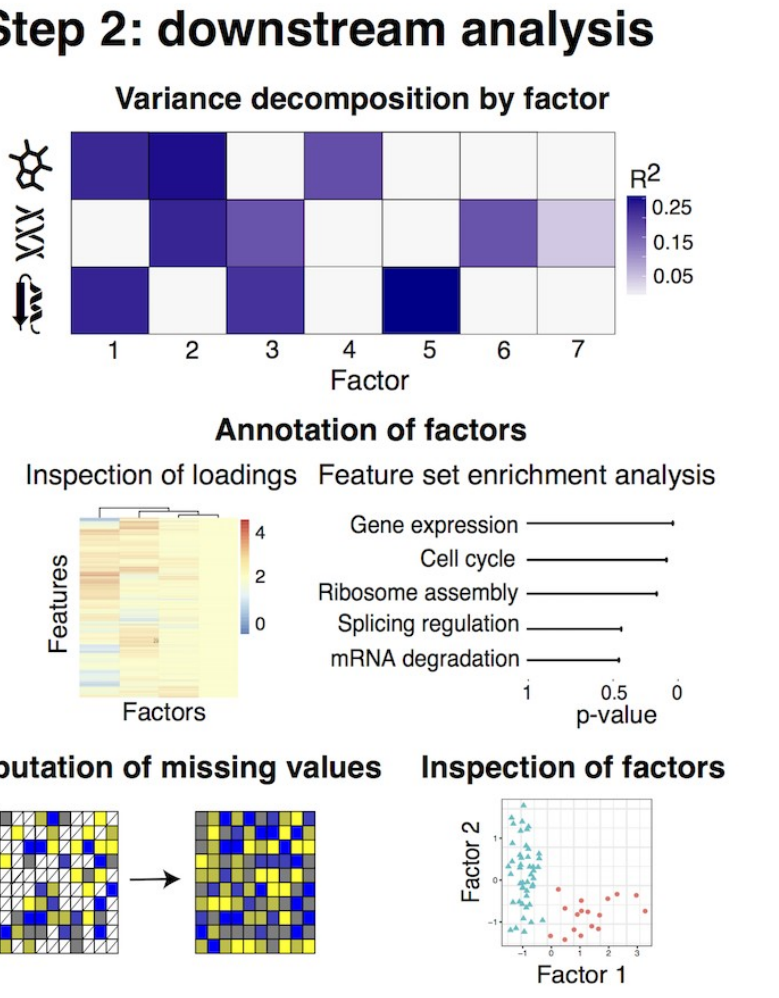
Gene expression

Methylation

Genetic mutation

Multi-Omics Factor Analysis (MOFA)





R. Argelaguet et al., Mol. Syst. Biol. 2018

MOFA: Bayesian Group Factor Analysis

Factor analysis models, also called latent variable models, are a probabilistic modelling approach which aim to reduce the dimensionality of a (big) dataset into a small set of variables which are easier to interpret and visualise. More formally, given a dataset $\mathbf{Y}$of N samples and D features, latent variable models attempt to explain dependencies between the features by means of a potentially smaller set of K unobserved (latent) factors. MOFA is a generalisation of traditional Factor Analysis where the input data consists of M matrices $\mathbf{Y}^m = [y^m_{nd}] \in \mathbb{R}^{N \times D_m}$ where each matrix m is called a view. Each view consists of non-overlapping features which usually, but not necessarily, represent different assays. The input data is then factorised as:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m, \tag{1}$$

where $\mathbf{Z} = [z_{nk}] \in \mathbb{R}^{N \times K}$ is a single matrix that contains the low-dimensional latent variables, $\mathbf{W}^m = [w^m_{dk}] \in \mathbb{R}^{D_m \times K}$ are loading matrices that relate the high-dimensional space to the low dimensional representation, and $\boldsymbol{\epsilon}^m = [\epsilon^m_d] \in \mathbb{R}^{D_m}$ denotes residual noise. We start by assuming Gaussian residuals $\boldsymbol{\epsilon}^m$, similar to standard (group) factor analysis models, while allowing for heteroscedasticity across features:

$$p(\epsilon^m_d) = \mathcal{N}\left(\epsilon^m_d \,|\, 0, 1/\tau^m_d\right). \tag{2}$$

This results in the following normal likelihood (for extensions to non-Gaussian settings see section 4):

$$p(y^m_{nd}) = \mathcal{N}\left(y^m_{nd} \,|\, \mathbf{z}_{n,:}\mathbf{w}^{mT}_{d,:}, 1/\tau^m_d\right), \tag{3}$$

where $\mathbf{w}^m_{d,:}$ denotes the d-th row of the loading matrix $\mathbf{W}^m$ and $\mathbf{z}_{n,:}$ the n-th row of the latent factor matrix $\mathbf{Z}$. For a fully probabilistic treatment we place prior distributions on the weights $\mathbf{W}^m$, the latent variables $\mathbf{Z}$ as well as on the precision of the noise $\boldsymbol{\tau}^m$. We use a standard Gaussian prior on the latent variables and a conjugate Gamma prior for the precision:

$$p(z_{n,k}) = \mathcal{N}\left(z_{n,k} \,|\, 0, 1\right), \tag{4}$$

$$p(\tau^m_d) = \mathcal{G}\left(\tau^m_d \,|\, a^\tau_0, b^\tau_0\right), \tag{5}$$

To ensure scalable inference we use a variational approach with a mean-field approximation [3]. Briefly, in variational inference the true intractable posterior distribution of the unobserved variables $p(\mathbf{X}|\mathbf{Y})$ is approximated by a simpler distribution of factorized form $q(\mathbf{X}) = \prod_i q(\mathbf{X}_i)$ that leads to an efficient inference scheme. Here, $\mathbf{X}$ denotes all the hidden variables (including parameters) and $\mathbf{Y}$ denotes all the observed variables.

Under this approximation, the true log marginal likelihood $\log p(\mathbf{Y})$ is lower bounded by:

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X})\Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y})\Big) d\mathbf{X} \\ &= \log p(\mathbf{Y}) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \\ &\leq \log p(\mathbf{Y})\end{aligned} \tag{11}$$

$\mathcal{L}(\mathbf{X})$ is called the Evidence Lower Bound (ELBO), which is equal to the sum of the model evidence and the negative KL-divergence between the true posterior and the variational distribution. The key observation here is that increasing the ELBO is equivalent to decreasing the KL-divergence between the two distributions.

Variational learning involves optimising the functional $\mathcal{L}(\mathbf{X})$ with respect to the distribution $q(\mathbf{X})$. If we allow any possible choice of $q(\mathbf{X})$, then the maximum of the lower bound $\mathcal{L}(\mathbf{X})$ will occur when the KL-divergence vanishes, which occurs when $q(\mathbf{X})$ equals the true posterior distribution $p(\mathbf{X}|\mathbf{Y})$. Nevertheless, since the true posterior is intractable, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of variational distributions that are tractable to compute and then seek the member of this family for which the KL divergence is minimised [2].

Mean-field approximation

The most common type of variational Bayes, known as mean-field approach, assumes that the variational distribution factorises over M disjoint groups of variables:

$$q(\mathbf{X}) = \prod_{i=1}^{M} q(\mathbf{x}_i)$$

Evidently, this family of distributions does not usually contain the true posterior because the unobserved variables have dependencies, but this assumption allows the derivation of an analytical inference scheme

R. Argelaguet et al., Mol. Syst. Biol. 2018

MOFA application for scNMTseq

nature COMMUNICATIONS

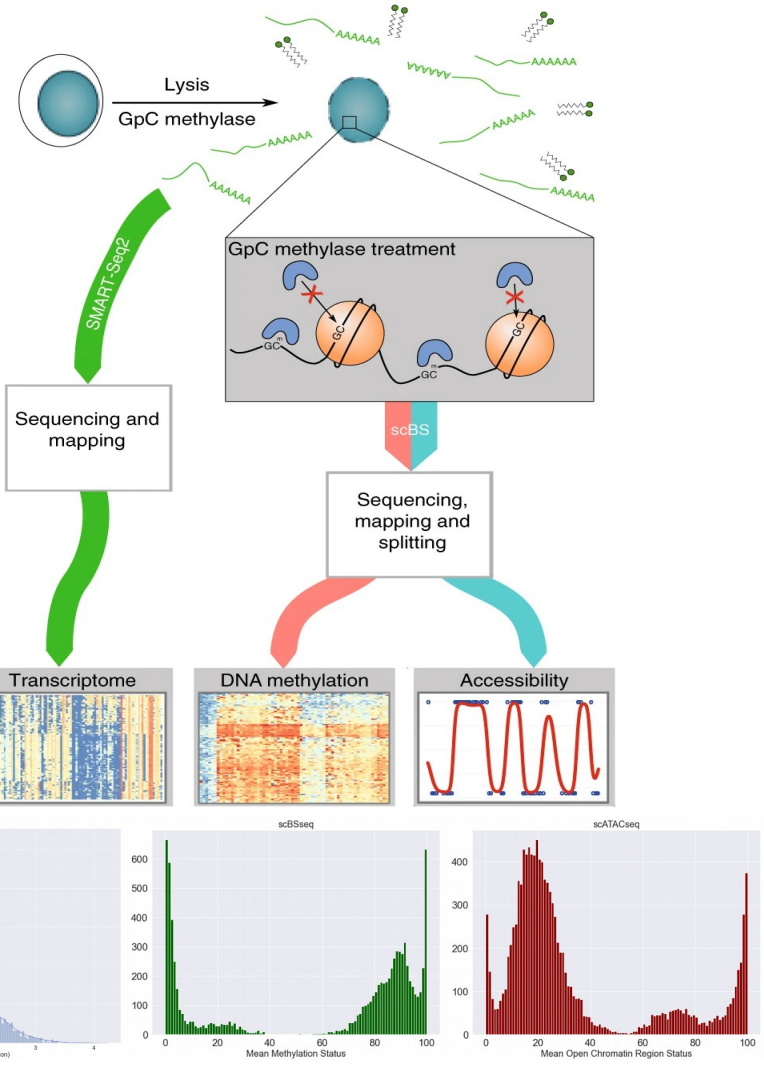
ARTICLE

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

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

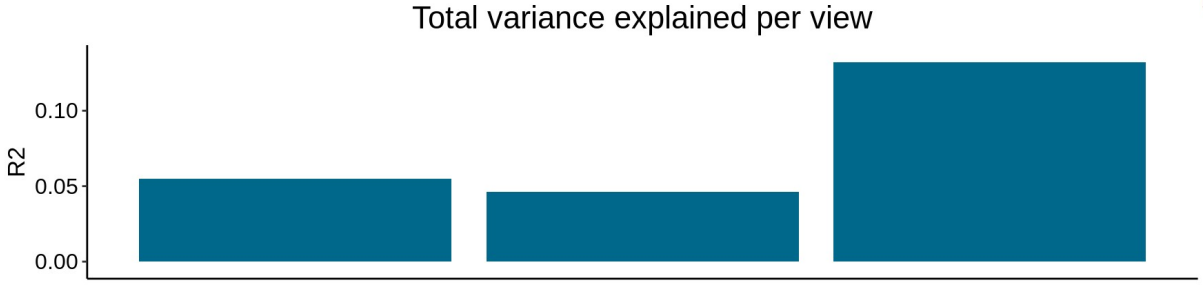
Stephen J. Clark[1], Ricard Argelaguet[2,3], ChantriolInt-Andreas Kapourani[4] Thomas M. Stubbs[1], Heather J. Lee[1,5,6], Celia Alda-Catalinas[1], Felix Krueger[7] Guido Sanguinetti[4], Gavin Kelsey[1,8] John C. Marioni[2,3,5] Oliver Stegle[2] Wolf Reik[1,5,8]

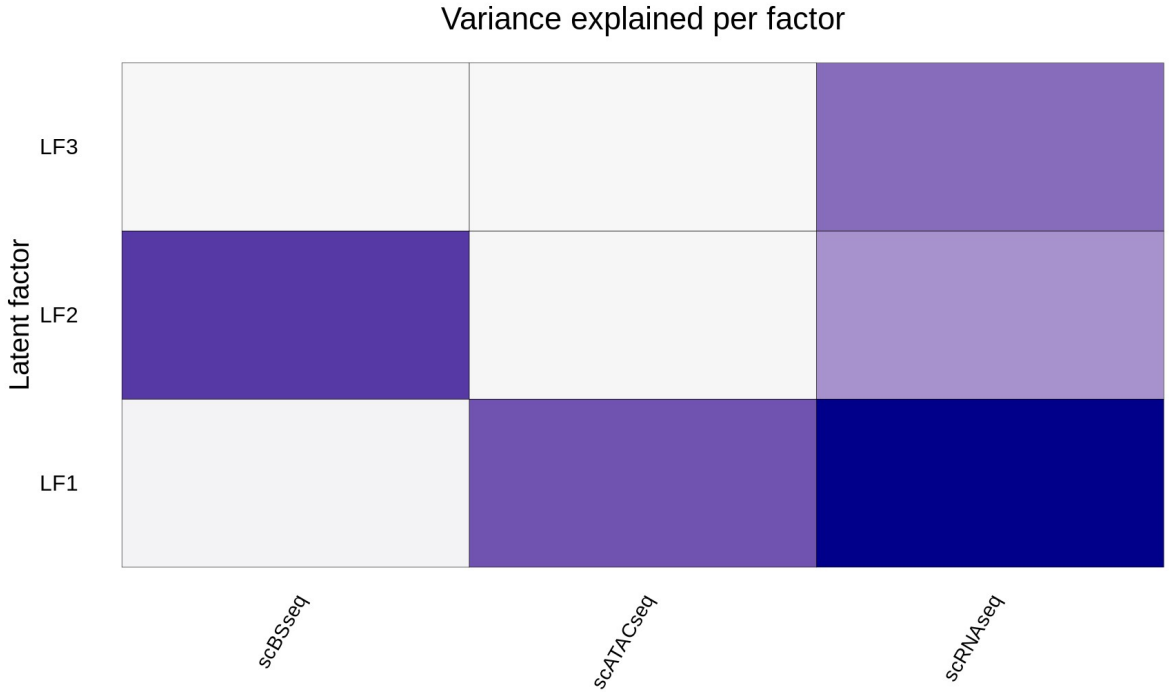
Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell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

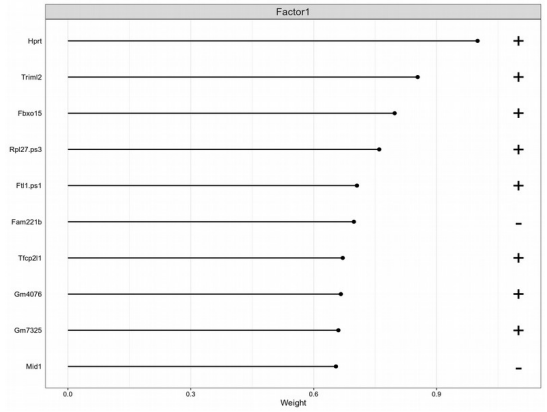


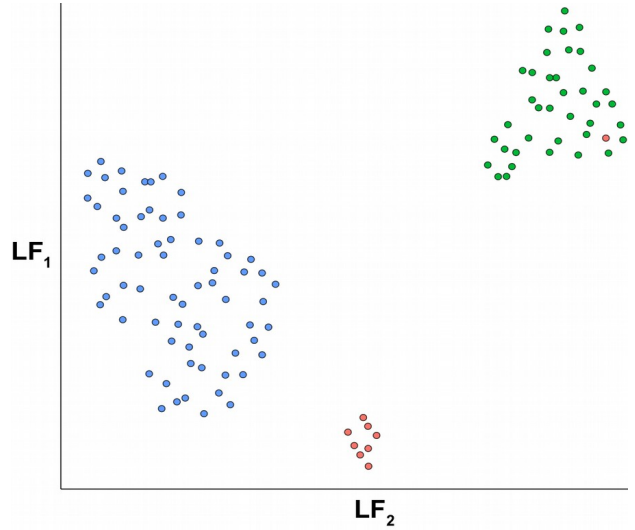
MOFA: Omics contributions to Latent Factors

$$R^2_{m,k} = 1 - \left(\sum_{n,d} y^m_{nd} - z_{nk} w^m_{kd} - \mu^m_d\right)^2 / \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$









National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET