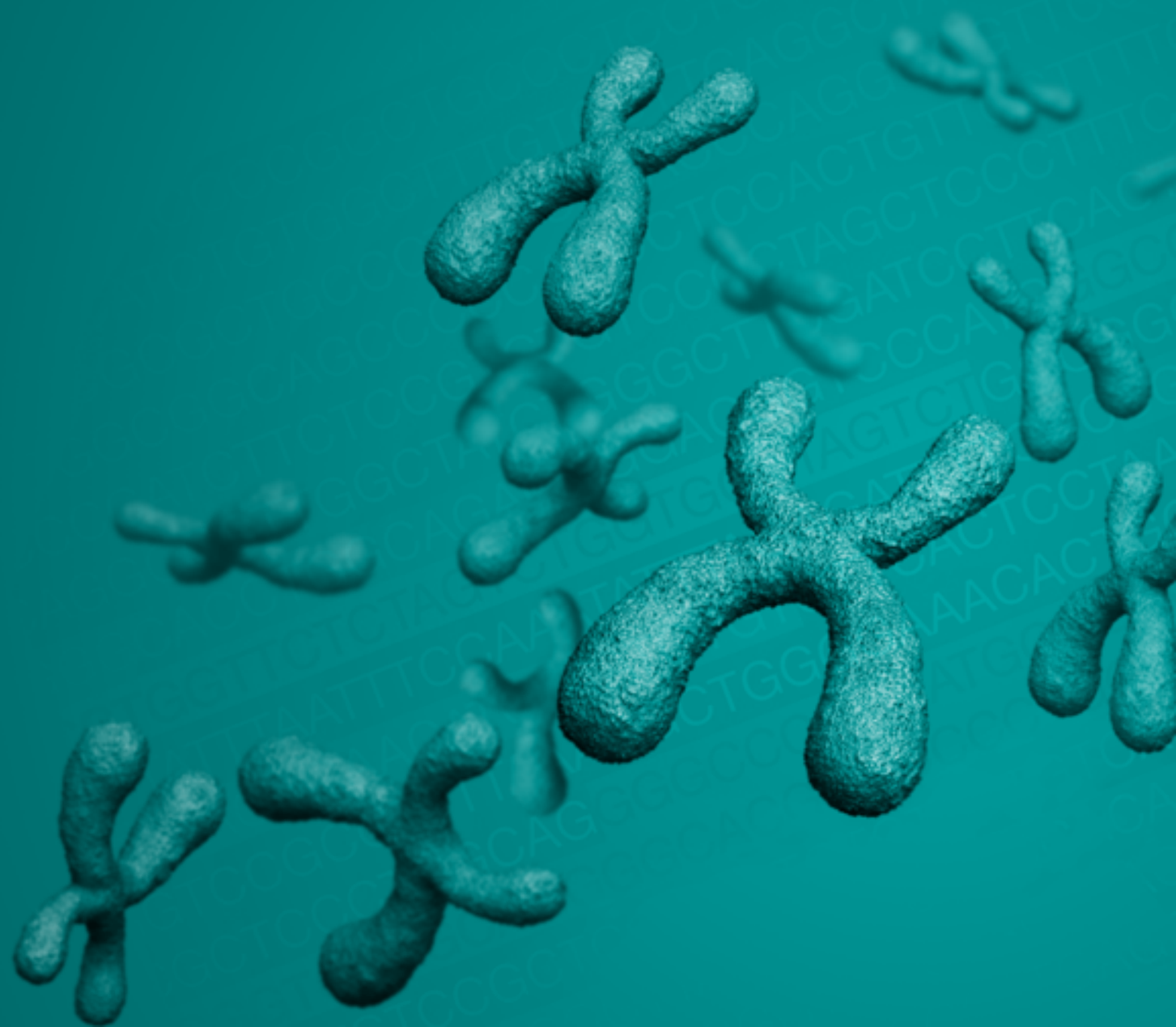
Statistical methods for multi-omics data integration: Multi-Omics Factor Analysis (MOFA)

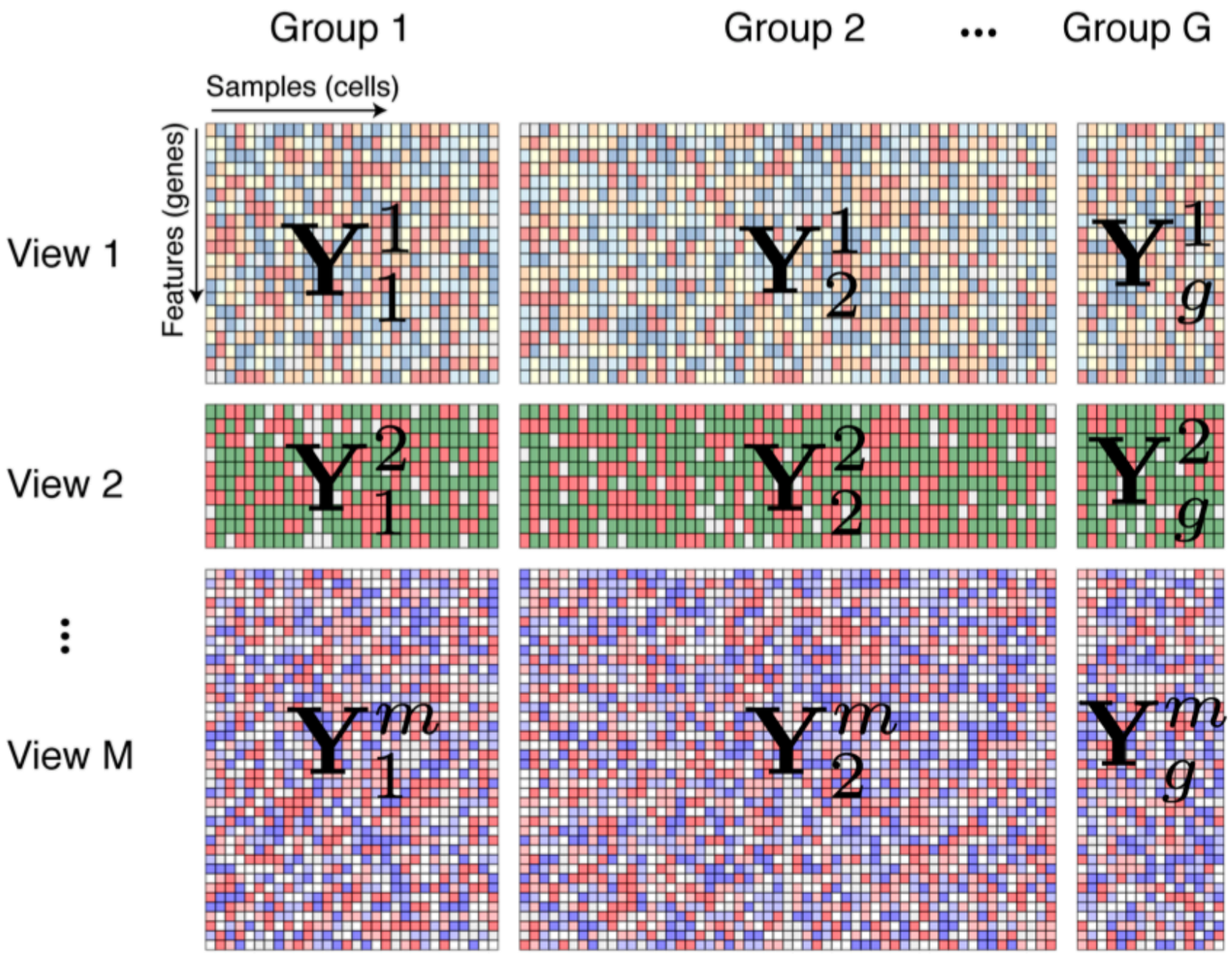
Ricard Argelaguet
ricard@ebi.ac.uk

PhD at European Bioinformatics Institute
John Marioni & Oliver Stegle groups

Postdoctoral scientist at Babraham Institute
Wolf Reik group

RArgelaguet



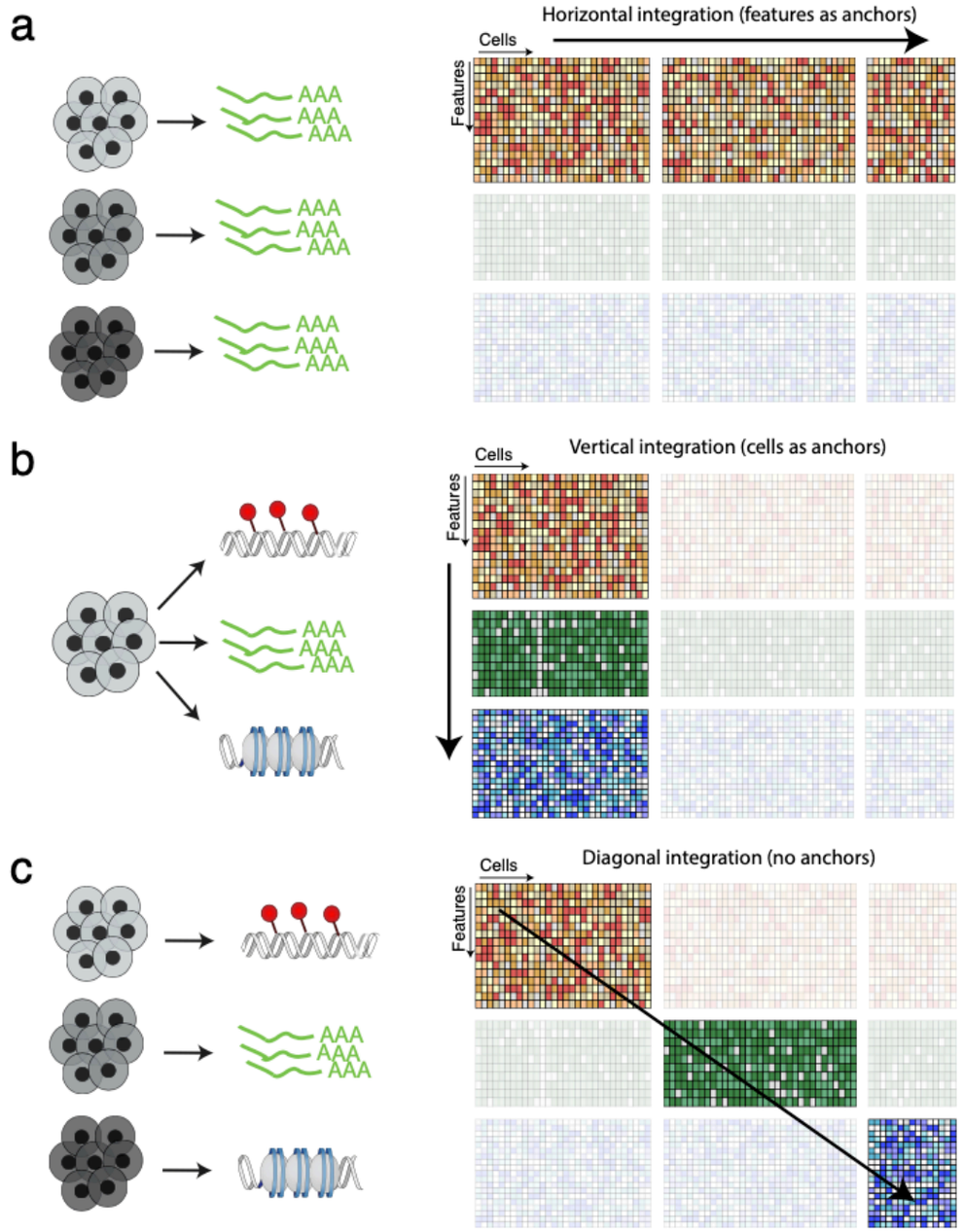

Group 1
Group 2
...
Group G
Samples (cells)
Features (genes)
View 1
View 2
View M
$\mathbf{Y}_1^1$
$\mathbf{Y}_2^1$
$\mathbf{Y}_g^1$
$\mathbf{Y}_1^2$
$\mathbf{Y}_2^2$
$\mathbf{Y}_g^2$
$\mathbf{Y}_1^m$
$\mathbf{Y}_2^m$
$\mathbf{Y}_g^m$

Choosing the integration anchor

(a) Genomic features as the anchor (horizontal integration): same data modality profiled from independent groups of samples

(b) Samples as the anchor (vertical integration): multiple data simultaneously profiled from the same sample

(c) No anchor in the high-dimensional space (diagonal integration): both samples and data modalities are different between experiments.



Challenges in multi-omics data integration

- Data collected using different techniques (i.e. data modalities) generally exhibit heterogeneous statistical properties
- Large amounts (and different patterns) of missing values
- Undesired sources of heterogeneity
- Overfitting
- Complexity of the data requires unsupervised interpretable approaches

Latent variable models

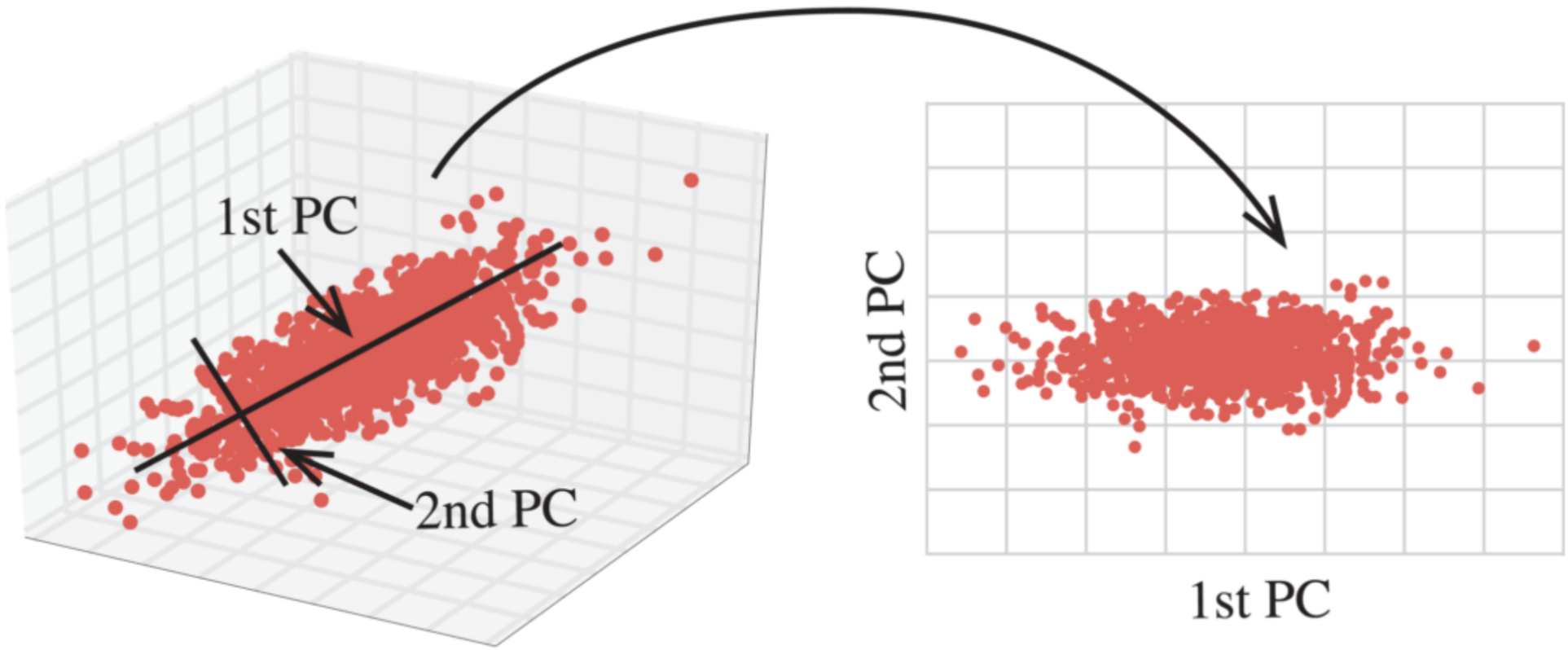
Y

Observed variables

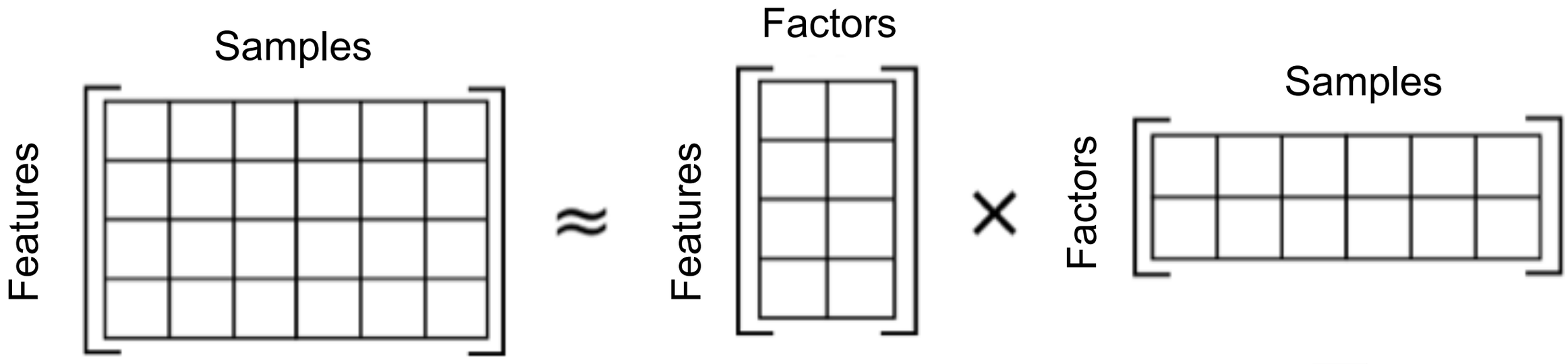
Z

Latent variables

f(x)



Matrix factorisation



Y **W** **Z**

Y is the observed measurements

W is the inferred feature weights

Z is the inferred latent factors

Principal Component Analysis is an instance of matrix factorisation!

Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) is a simple extension of PCA to find linear components that capture correlations between **two** datasets[3].

Given two data matrices $\mathbf{Y}_1 \in \mathbb{R}^{N \times D_1}$ and $\mathbf{Y}_2 \in \mathbb{R}^{N \times D_2}$ CCA finds a set of linear combinations $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ with maximal cross-correlation.

For the first pair of canonical variables, the optimisation problem is:

$$(\hat{\mathbf{u}_1}, \hat{\mathbf{v}_1}) = \underset{\|\mathbf{u}_1\|=1, \|\mathbf{v}_1\|=1}{\arg\max} \; corr(\mathbf{u}_1^T \mathbf{Y}_1, \mathbf{v}_1^T \mathbf{Y}_2)$$

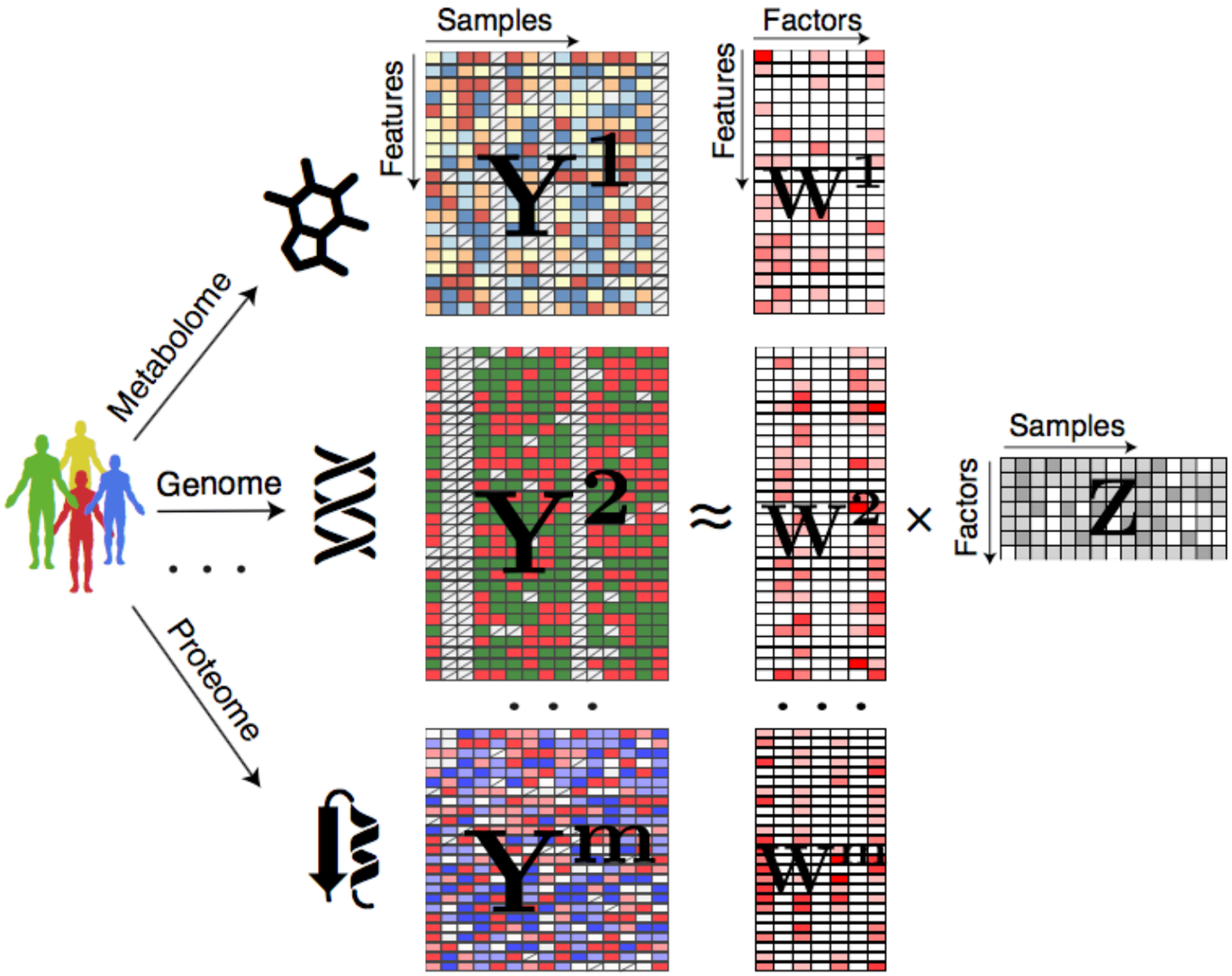
[3] Hotelling, H. "Relations between two sets of variates" *Biometrika* 1936.

What are the problems of CCA for multi-omics data integration?

In CCA the canonical components are defined as linear combinations of features that maximise the cross-correlation between the two data sets. This implies that:

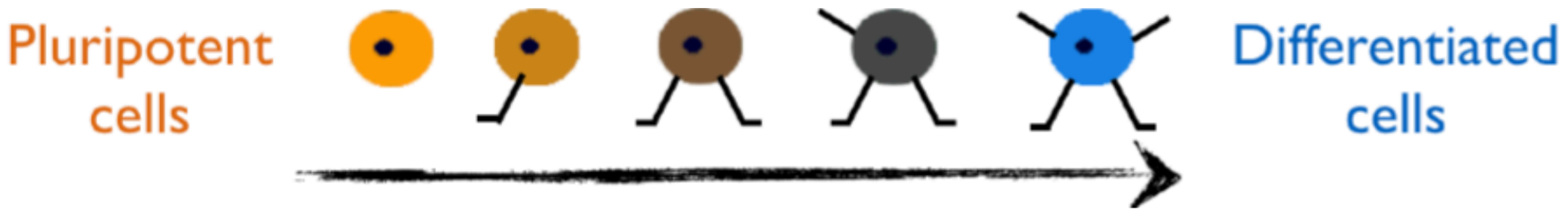
- It only works for the integration of *2* data sets
- It only finds sources of covariation between the two data sets. CCA is not able to find the sources of variation that are present within individual data sets

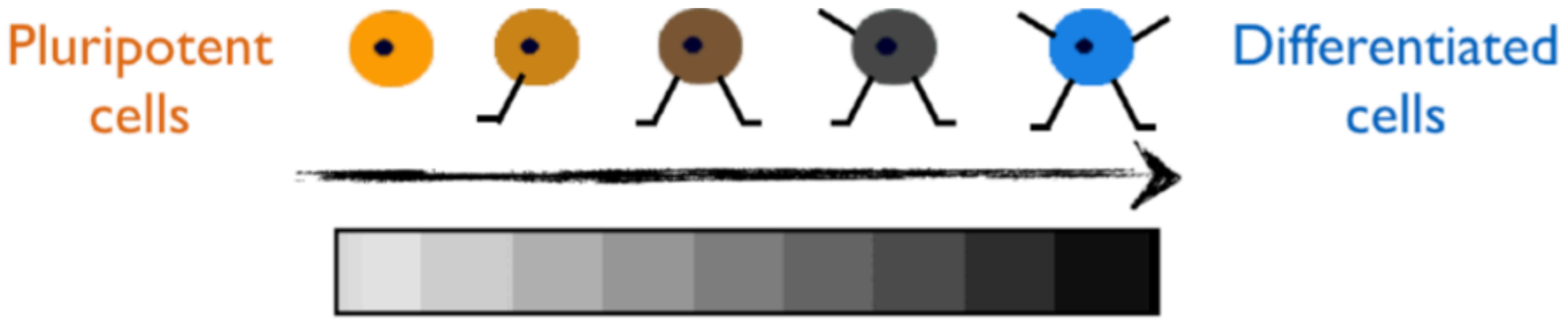
Multi-Omics Factor Analysis (MOFA)



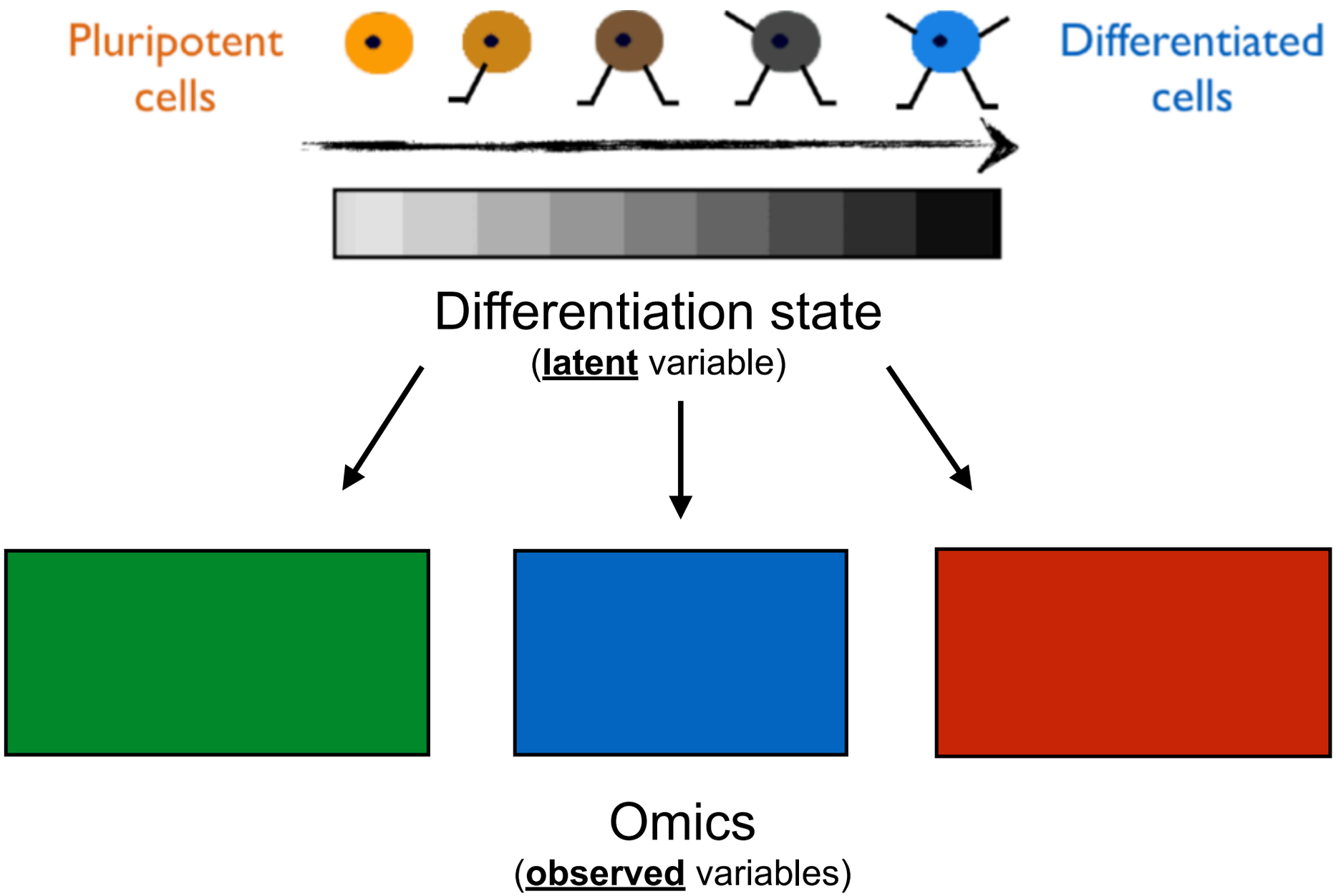
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

- The structure of the data is specified in the prior distributions of the Bayesian model
- The critical part of the model is the use sparsity priors, which enable automatic relevance determination of the factors

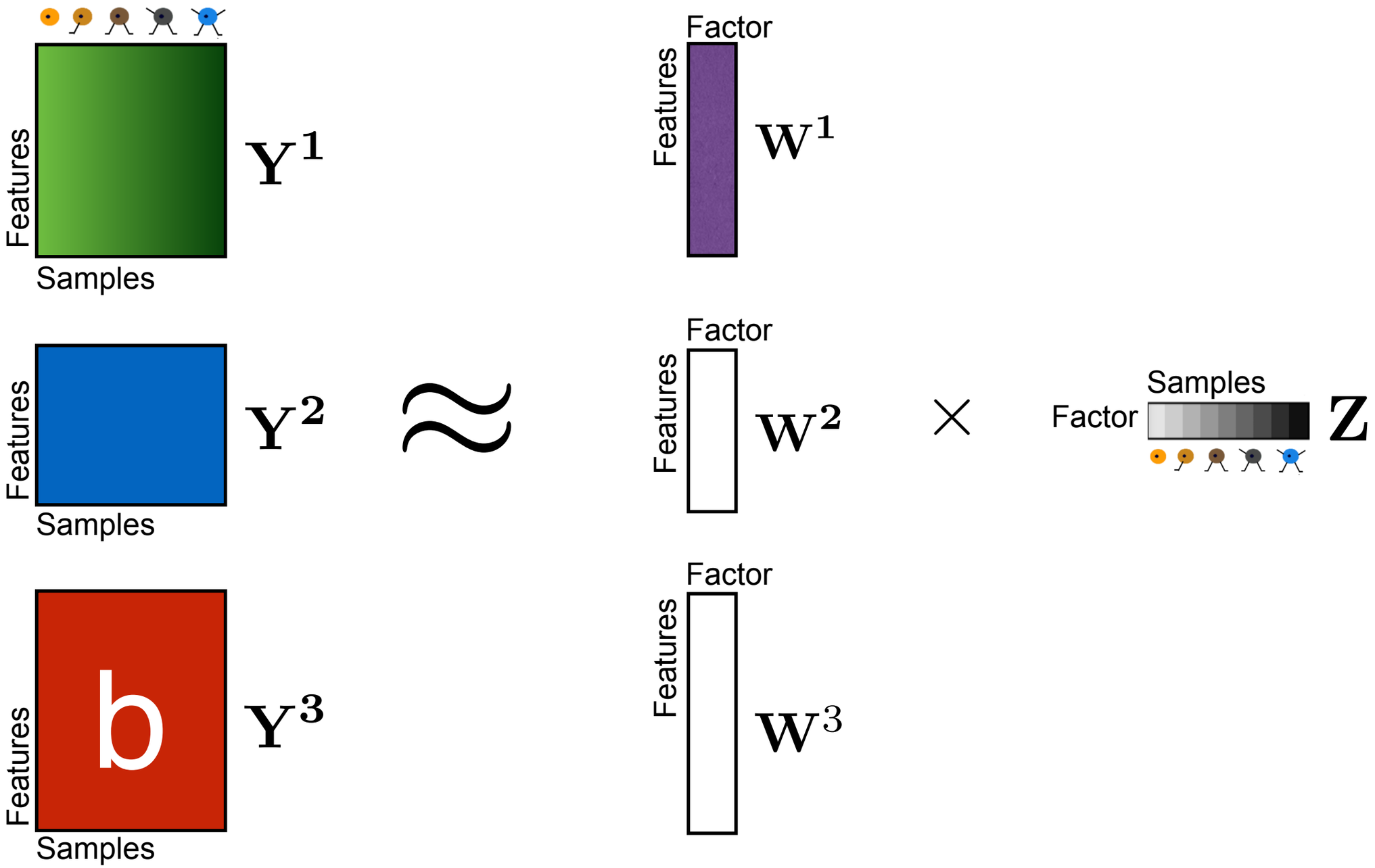

Pluripotent cells
Differentiated cells



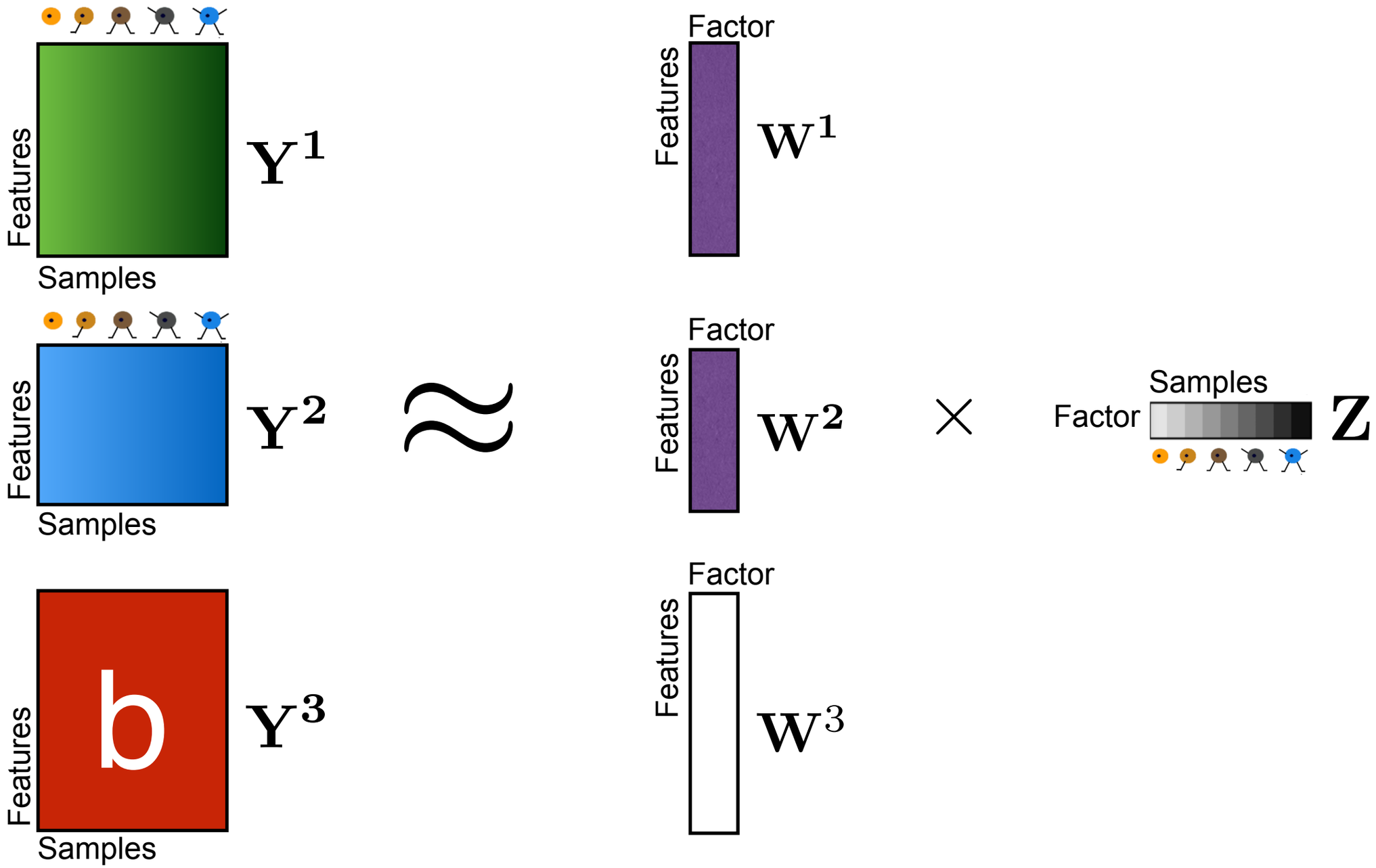
Differentiation state
(latent variable)


Pluripotent cells
Differentiated cells
Differentiation state
(**latent** variable)
Omics
(**observed** variables)

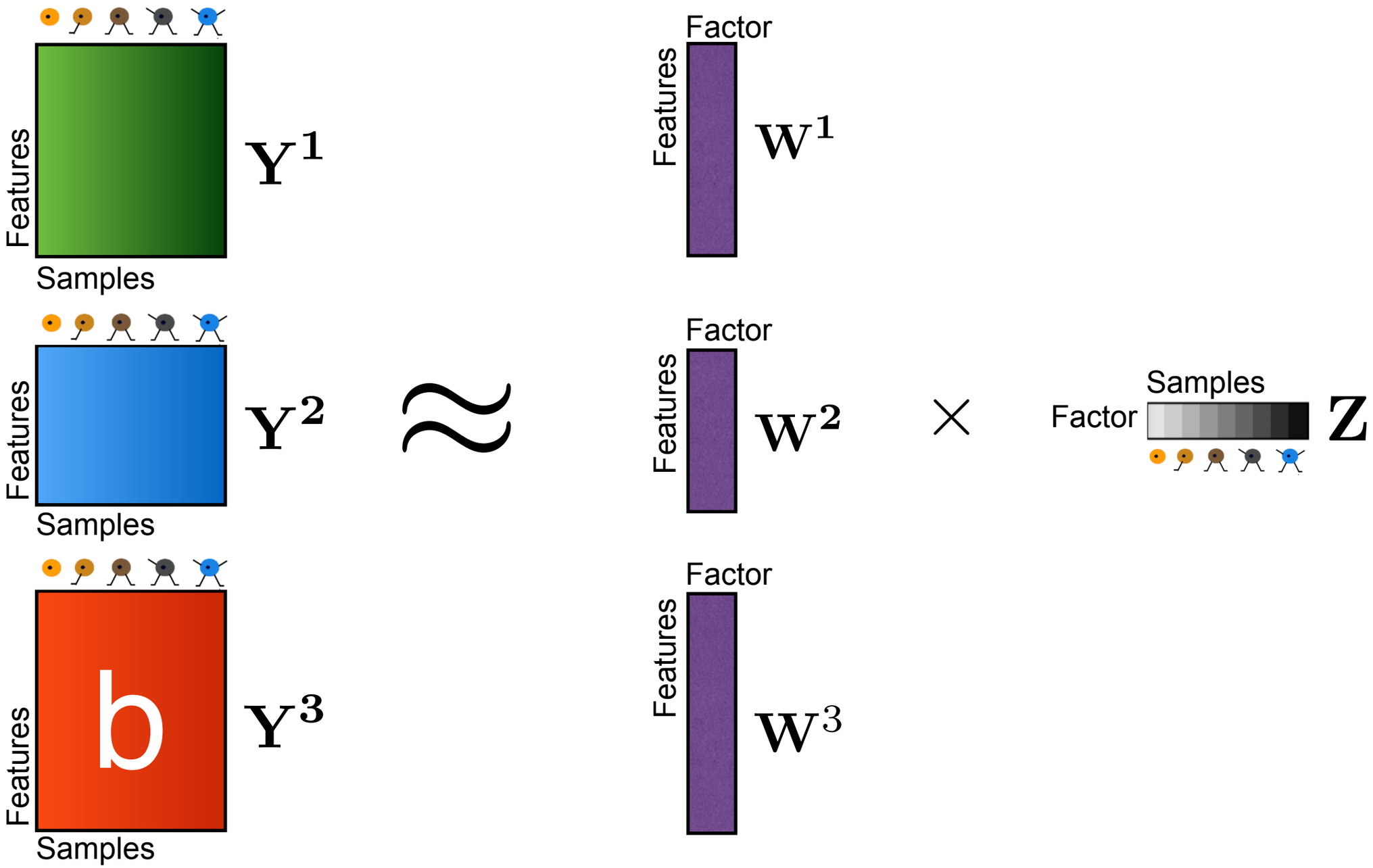
The differentiation state is the only driver of variation in **transcriptomics**



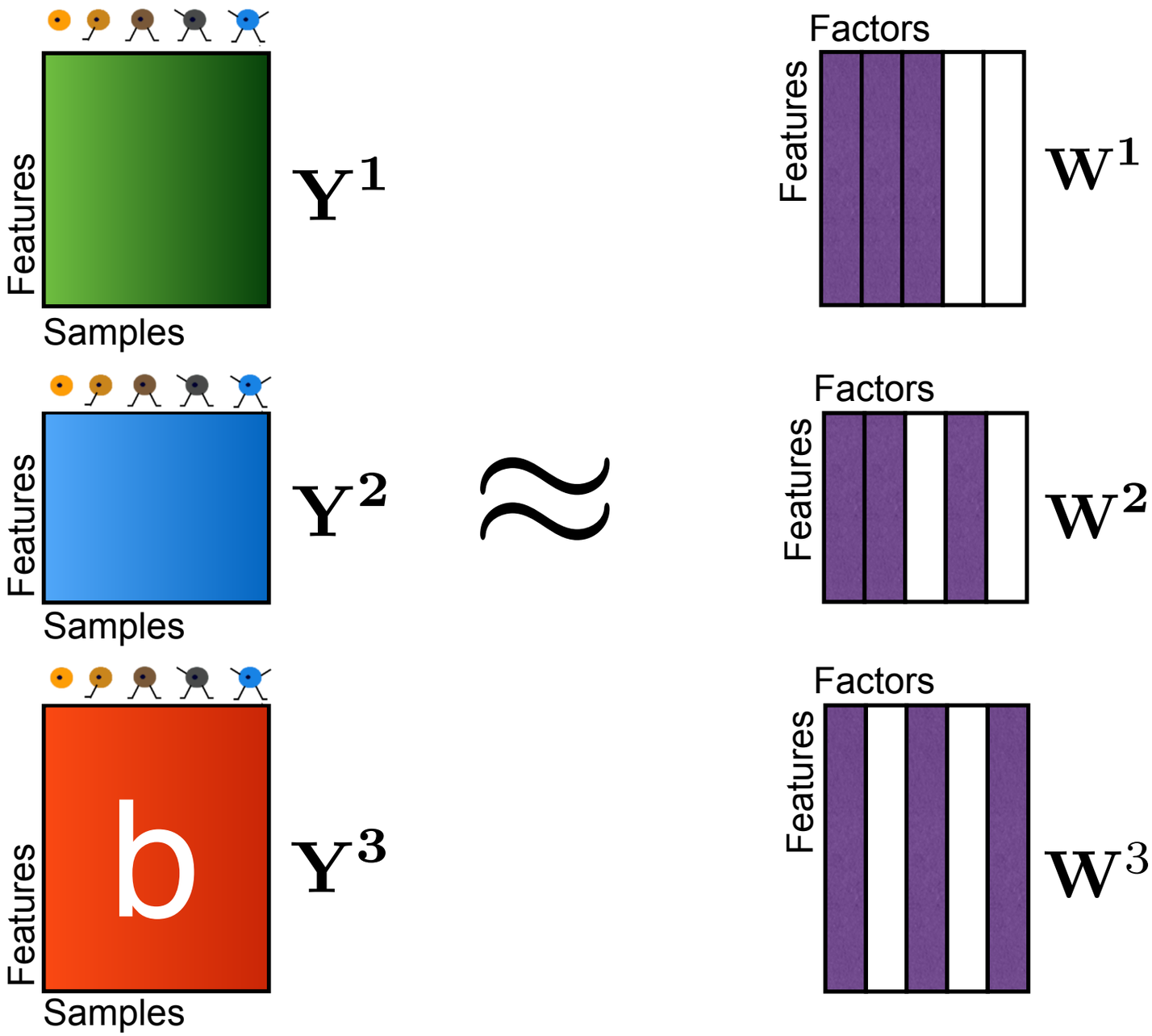
The differentiation state is the only driver of variation in **transcriptomics and genetics**



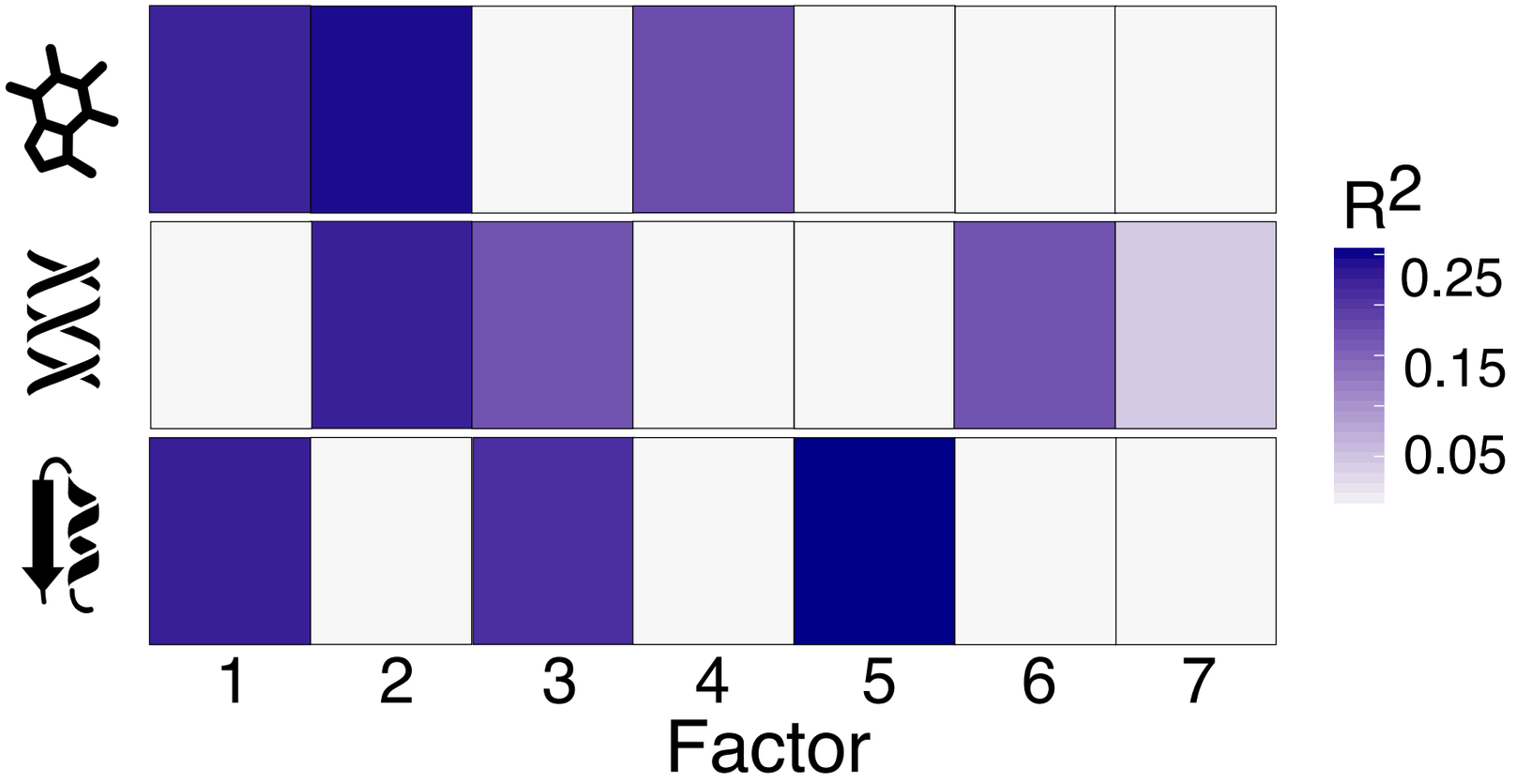
The differentiation state is the only driver of variation in **all omics**



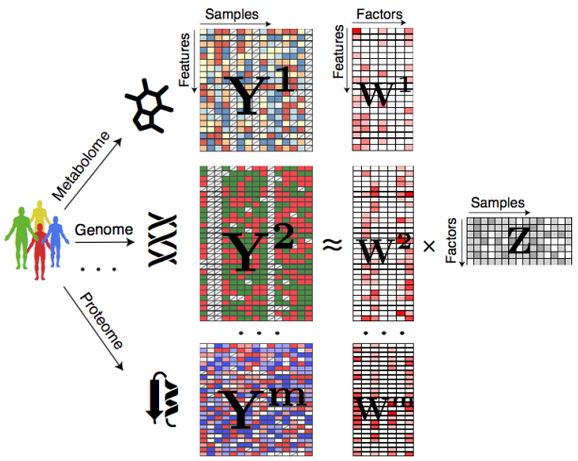
The differentiation state is the only driver of variation in **all omics**



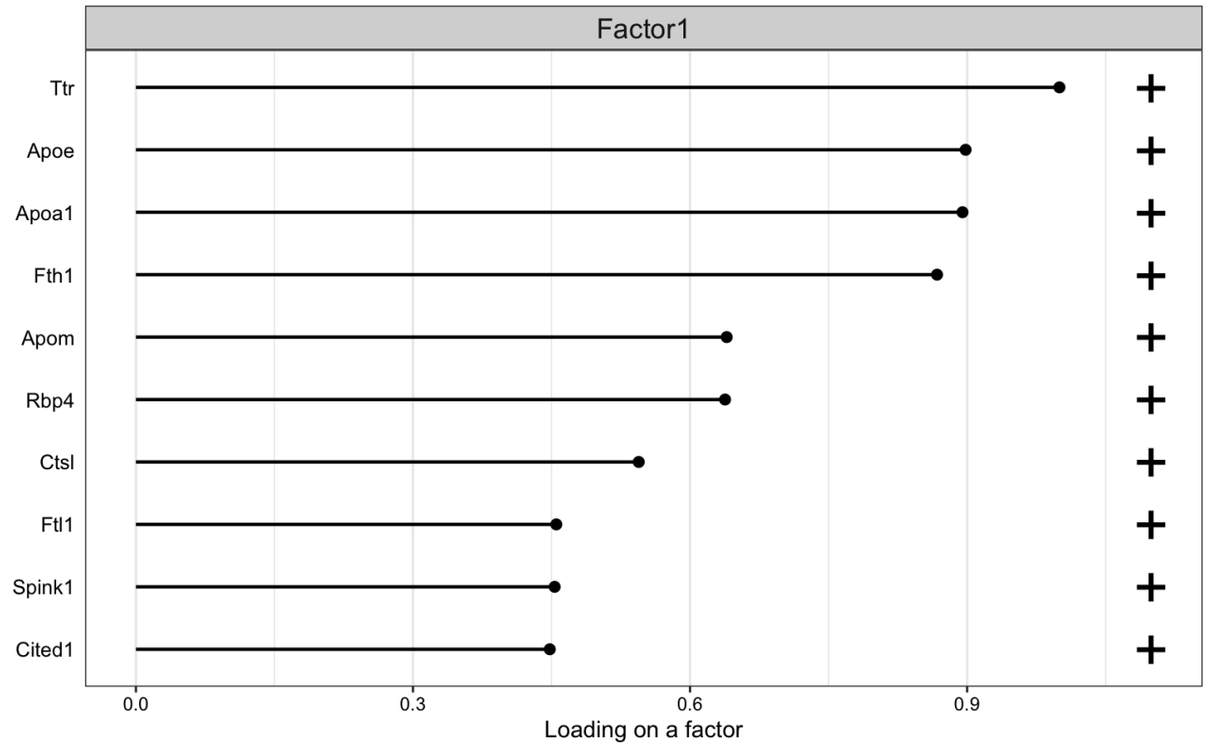
Variance decomposition by factor



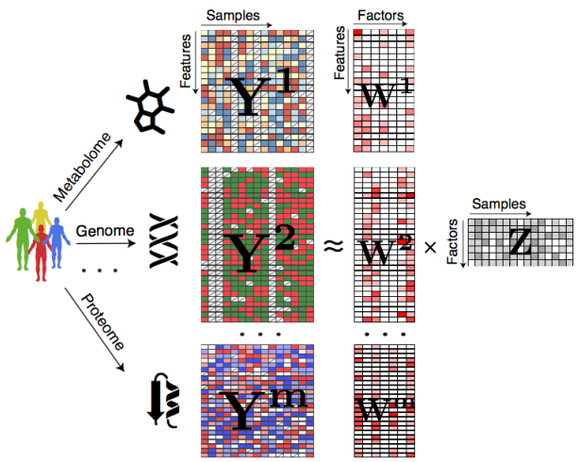
Downstream analysis



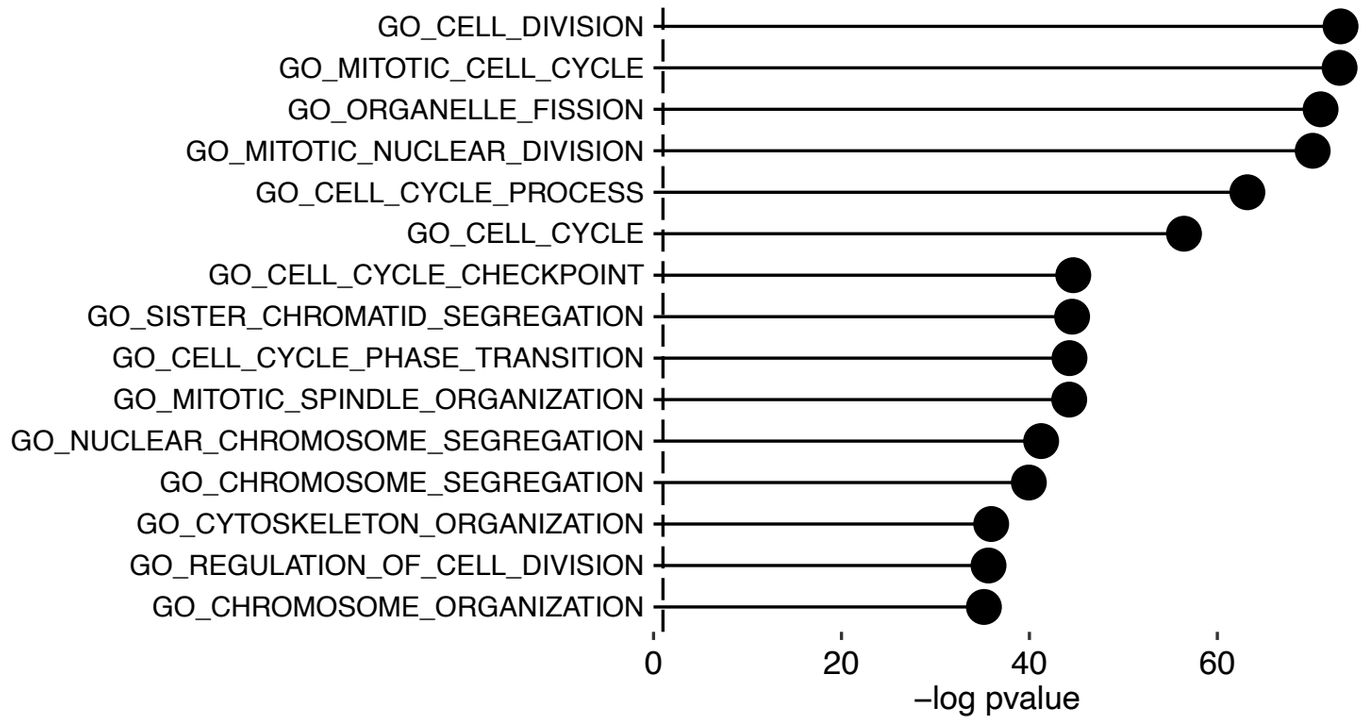
Inspection of feature weights



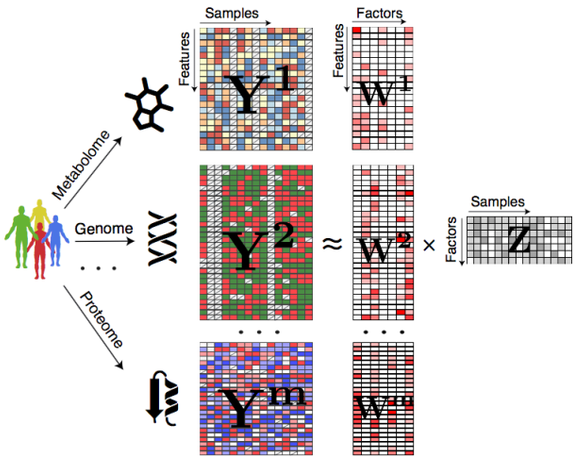
Downstream analysis

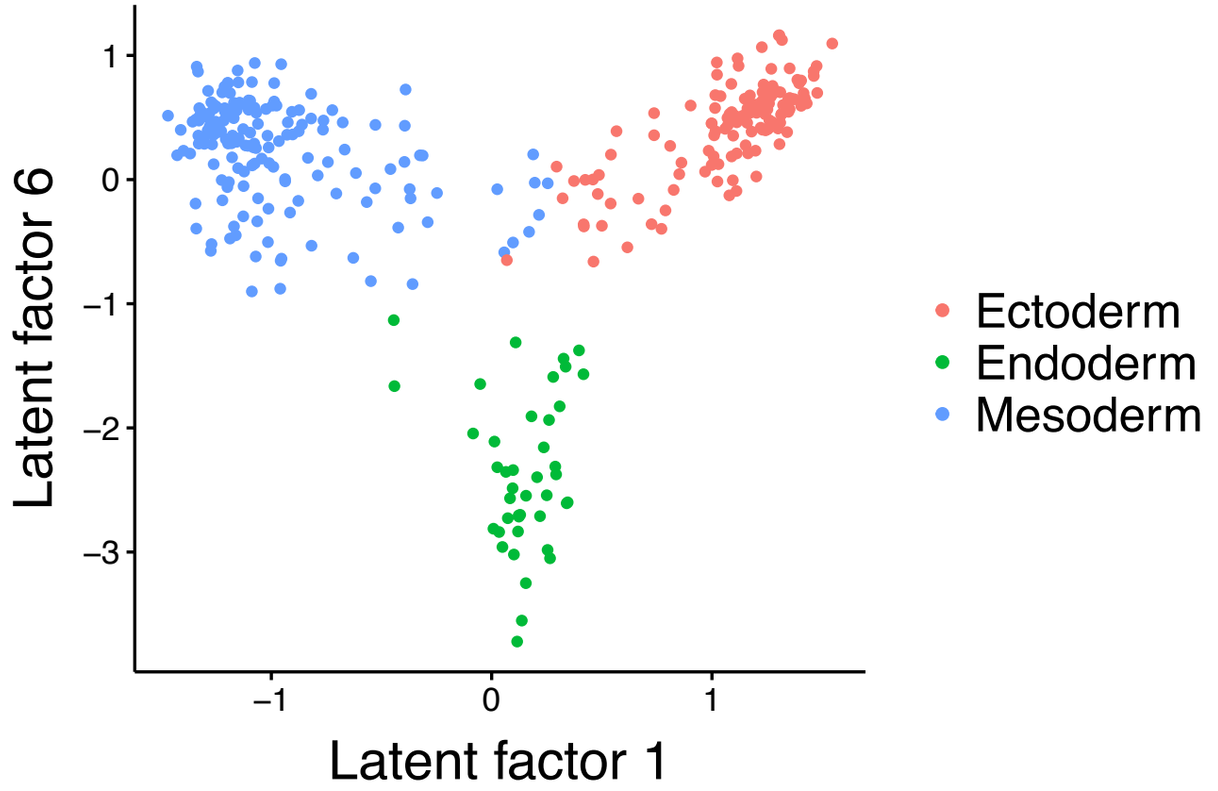


Gene set enrichment analysis

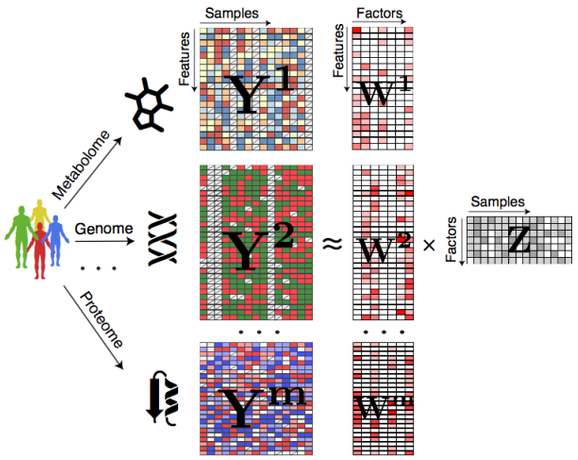


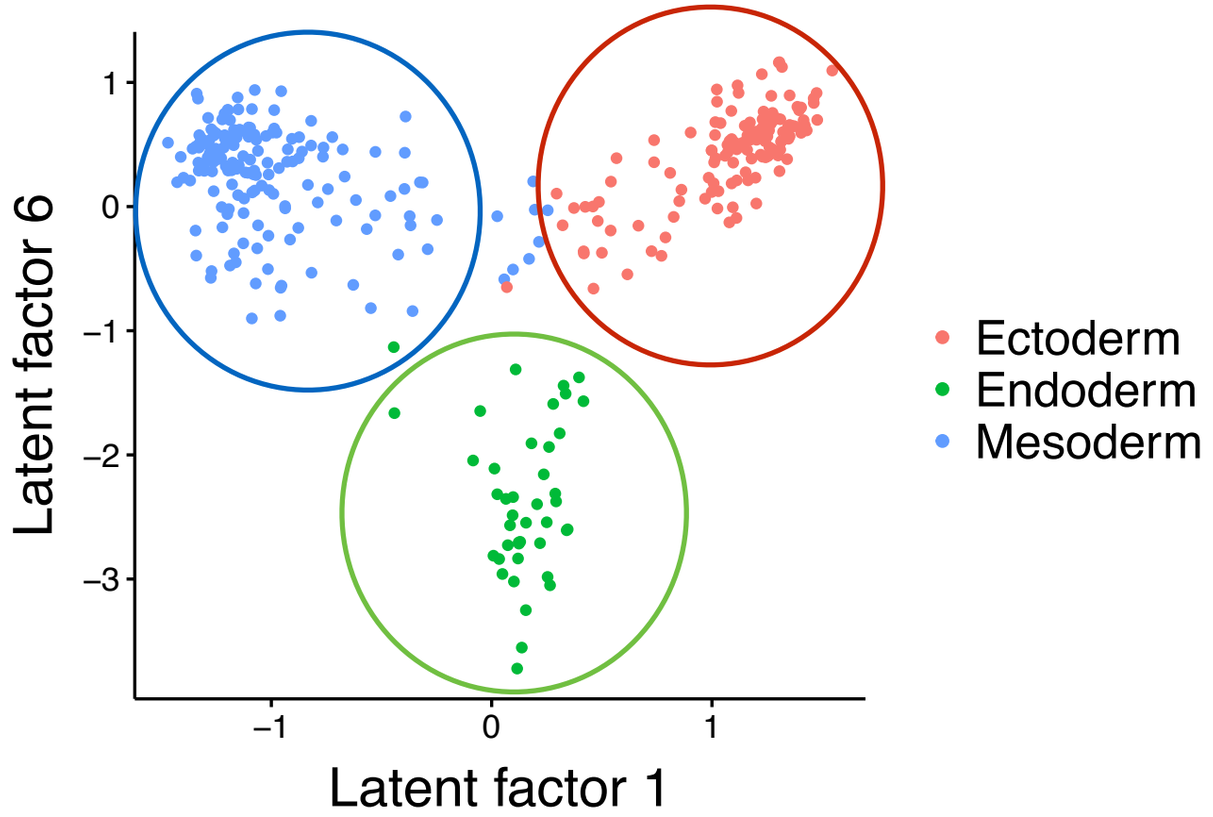
Downstream analysis



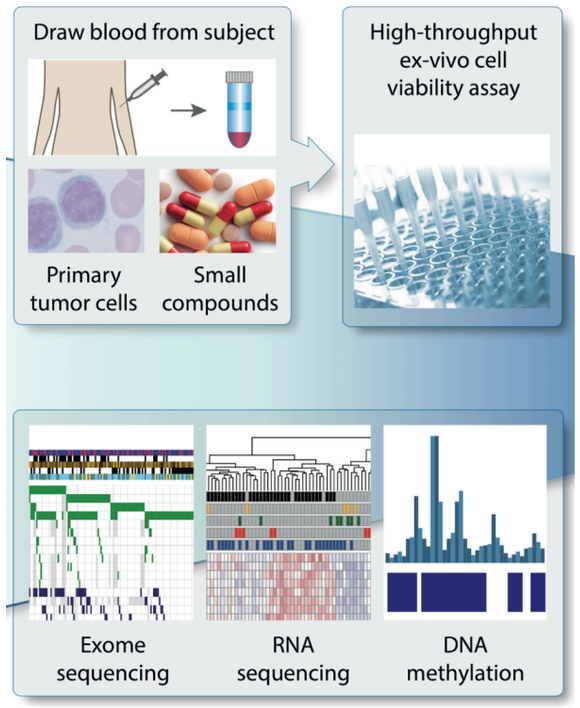


Downstream analysis

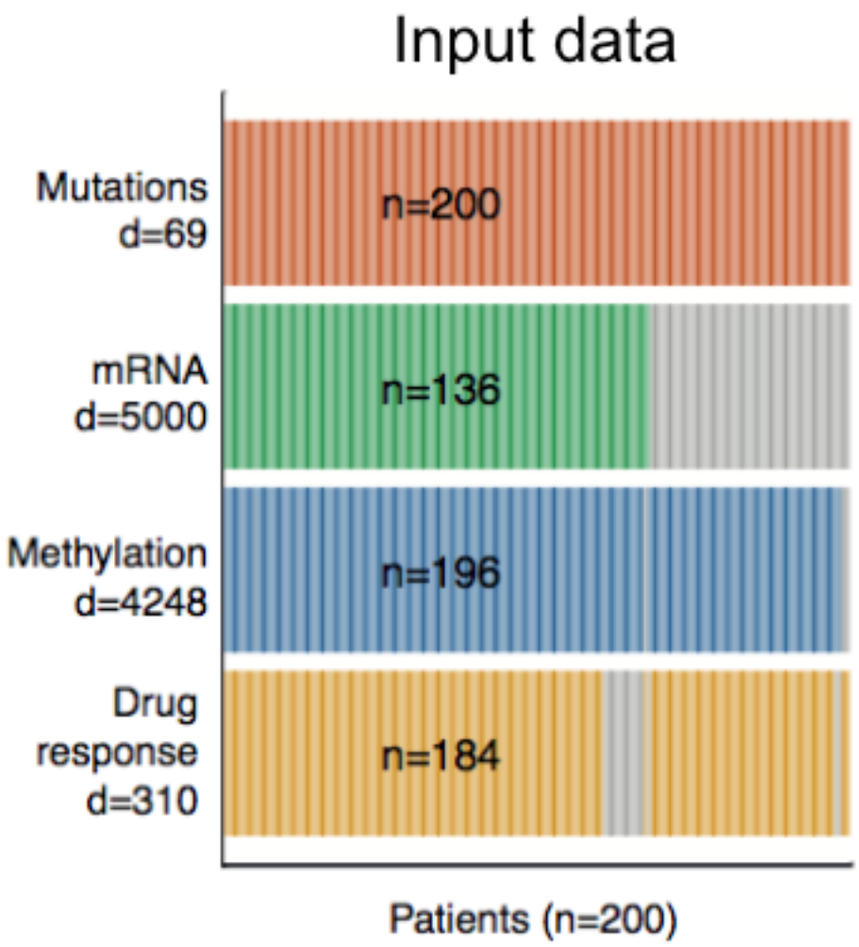




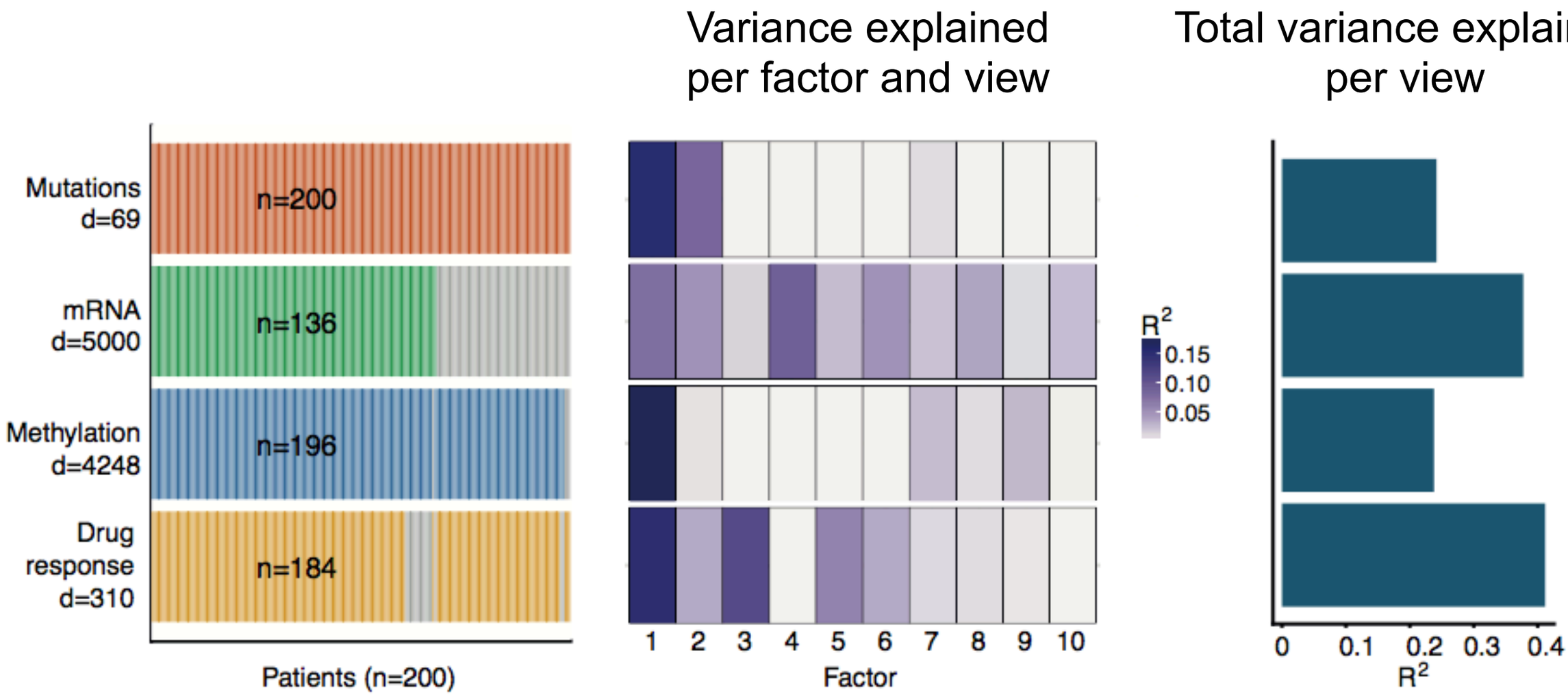
Application to Chronic Lymphocitic Leukaemia



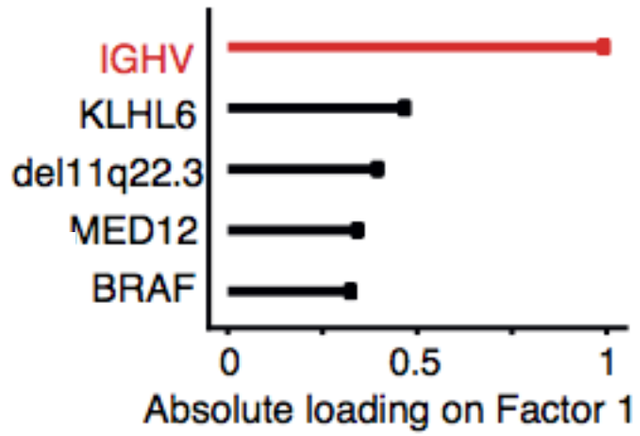
Thorsten Zenz group (Heidelberg)

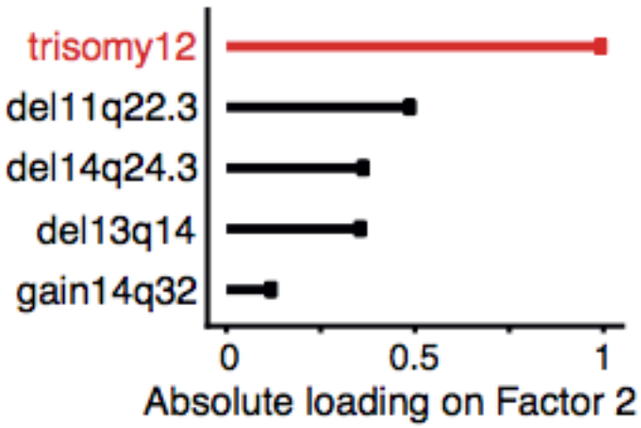


Application to Chronic Lymphocitic Leukaemia



Inspection of feature weights for Factors 1 and 2





IGHV: *Immunoglobulin heavy chain variable region*

American Journal of Hematology AJH
Explore this journal >

CLINICAL PEARLS IN BLOOD DISEASES

IGHV mutational status testing in chronic lymphocytic leukemia

Jennifer Crombie, Matthew S. Davids

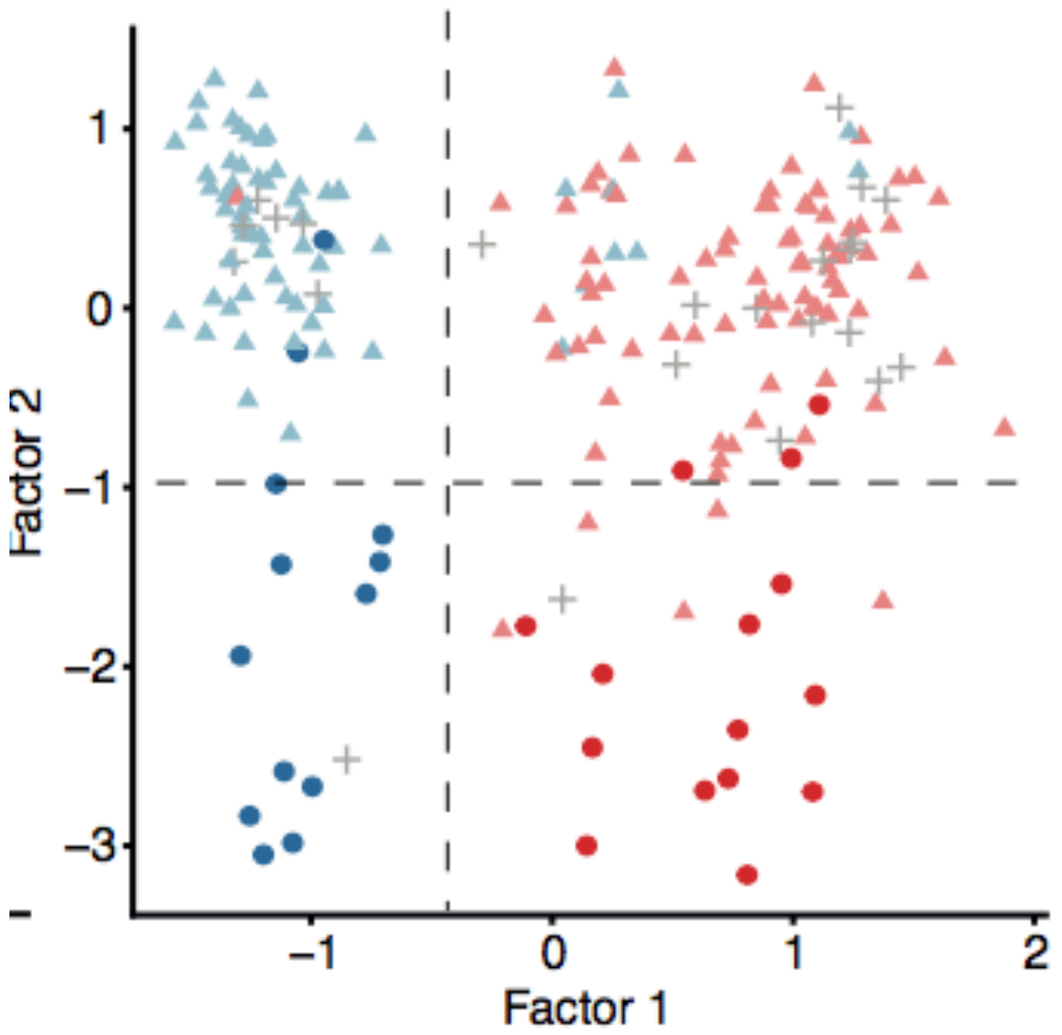
Trisomy 12 chronic lymphocytic leukemia cells

John C. Riches, Conor J. O'Donovan, Sarah J. Kingdon, Fabienne McClanahan, Andrew J. Clear, Laura Z. Rassenti, Thomas J. Kipps, and John G. Gribben

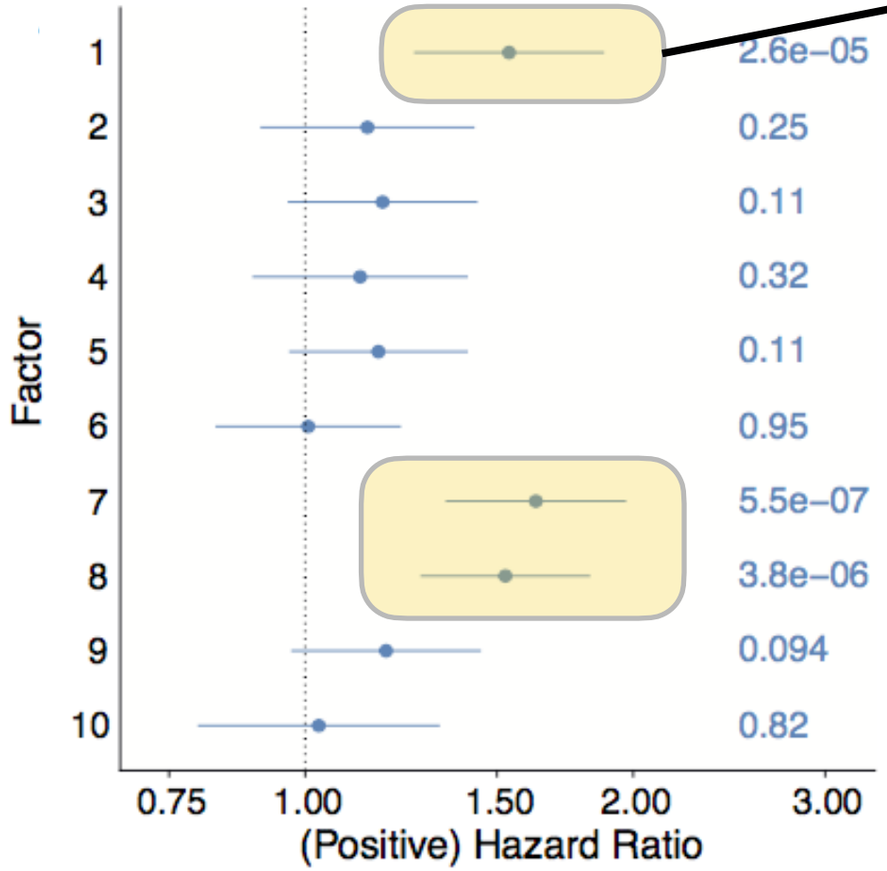
Blood 2014 123:4101-4110; doi: https://doi.org/10.1182/blood-2014-01-552307

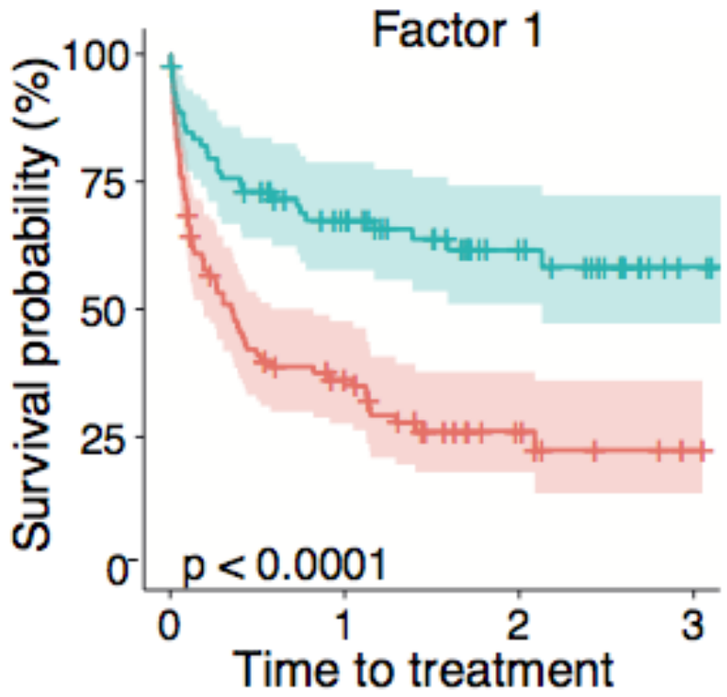
Visualisation of samples in the latent space

Factor 1: IGHV+ vs IGHV-
Factor 2: tr12+ (●) vs tr12- (▲)



Factors are associated with clinical response



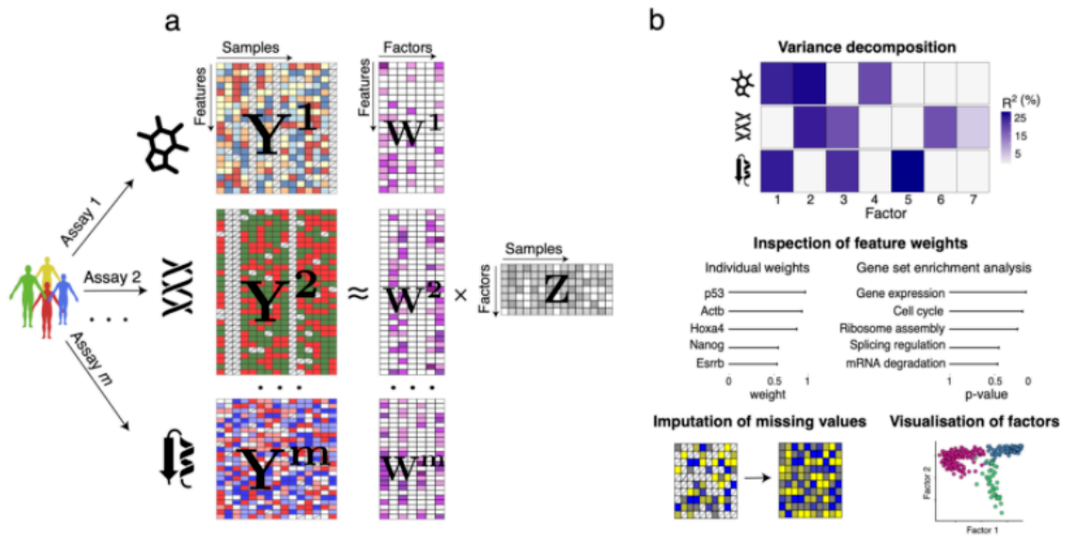


MOFA

Multi-Omics Factor Analysis V2 (MOFA+)

Home | Installation | Tutorials | Interactive web server | FAQ | Troubleshooting | MEFISTO | News | Contact | Citation | View on GitHub

MOFA is a factor analysis model that provides a **general framework for the integration of multi-omic data sets** in an unsupervised fashion.
Intuitively, MOFA can be viewed as a versatile and statistically rigorous generalization of principal component analysis to multi-omics data. Given several data matrices with measurements of multiple -omics data types on the same or on overlapping sets of samples, MOFA infers an **interpretable low-dimensional representation in terms of a few latent factors**. These learnt factors represent the driving sources of variation across data modalities, thus facilitating the identification of cellular states or disease subgroups.



Case examples using real data (in R)

- **(authors' favourite) Analysis of chronic lymphocytic leukaemia cohort for personalised medicine**: a bulk multi-omics data set. Figure 2 and 3 of the MOFA v1 paper.
- **(authors' favourite) Integrative analysis of the Chromium Single Cell Multiome ATAC + Gene Expression assay**: this is the result of a collaboration between the MOFA team and the 10x Genomics R&D team to provide a downstream analysis pipeline for the new RNA+ATAC multi-modal technology.
- **Analysis of a time course scRNA-seq data set using the multi-group framework**: Figure 2 of the MOFA+ paper. Demonstrates the multi-group functionality and how to train a MOFA model from a Seurat object.
- **Integration of scNMT-seq data (single-cell multi-omics)**: Figure 4 of the MOFA+ paper. Demonstrates the simultaneous multi-view and multi-group functionality using the scNMT-seq mouse gastrulation atlas.
- **Integration of SNARE-seq data (single-cell multi-omics)**. Demonstrates how MOFA can be used for the analysis of paired scRNA+scATAC data (from the same cell) using a multi-modal Seurat object. This data set is very noisy and the results are not fantastic, we suggest you have a look at the Chromium Single Cell Multiome ATAC + Gene Expression vignette instead.
- **Analysis of multi-modal microbiome data**: we demonstrate how to systematically integrate viral, fungal and bacterial sequence data.

Statistical methods for the integrative analysis of single-cell multi-omics data

Ricardo Argelaguet Calado

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

https://github.com/rargelaguet/thesis

EMBL-EBI