Multi-Omics Factor Analysis (MOFA)

A statistical framework for the unsupervised integration of multi-omics data

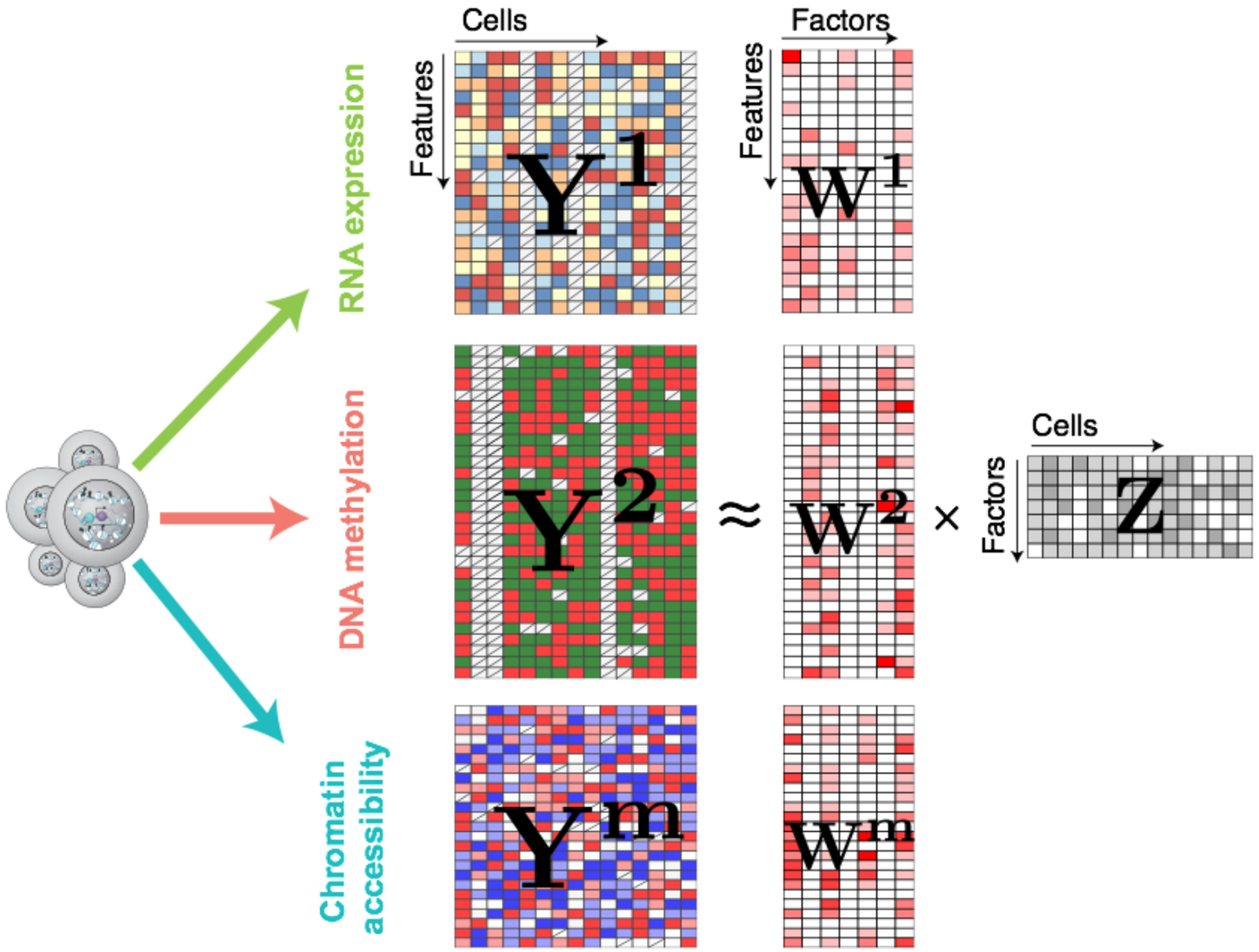
What are the problems of CCA for multi-omics data integration?

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

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

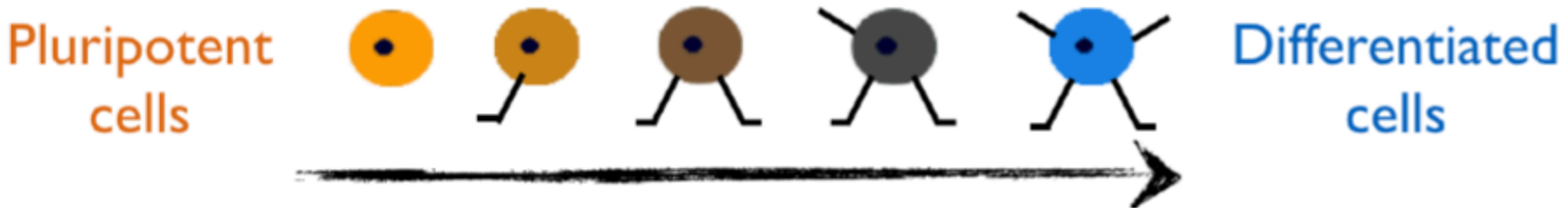
- It only works with *M=2* data sets
- It only finds sources of variation that are present in both data sets. CCA is not able to find the sources of variation (i.e. factors) that are present on the individual data sets

Multi-omics Factor Analysis (MOFA)

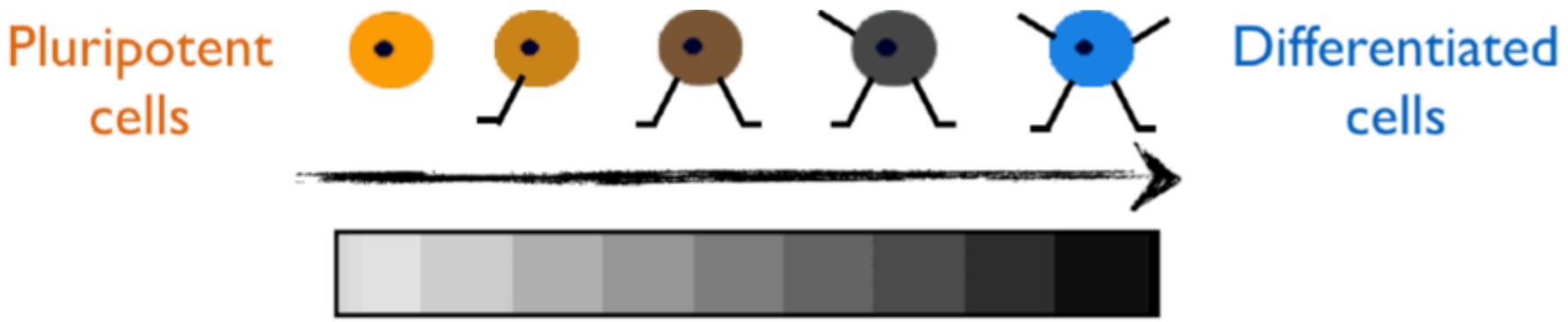


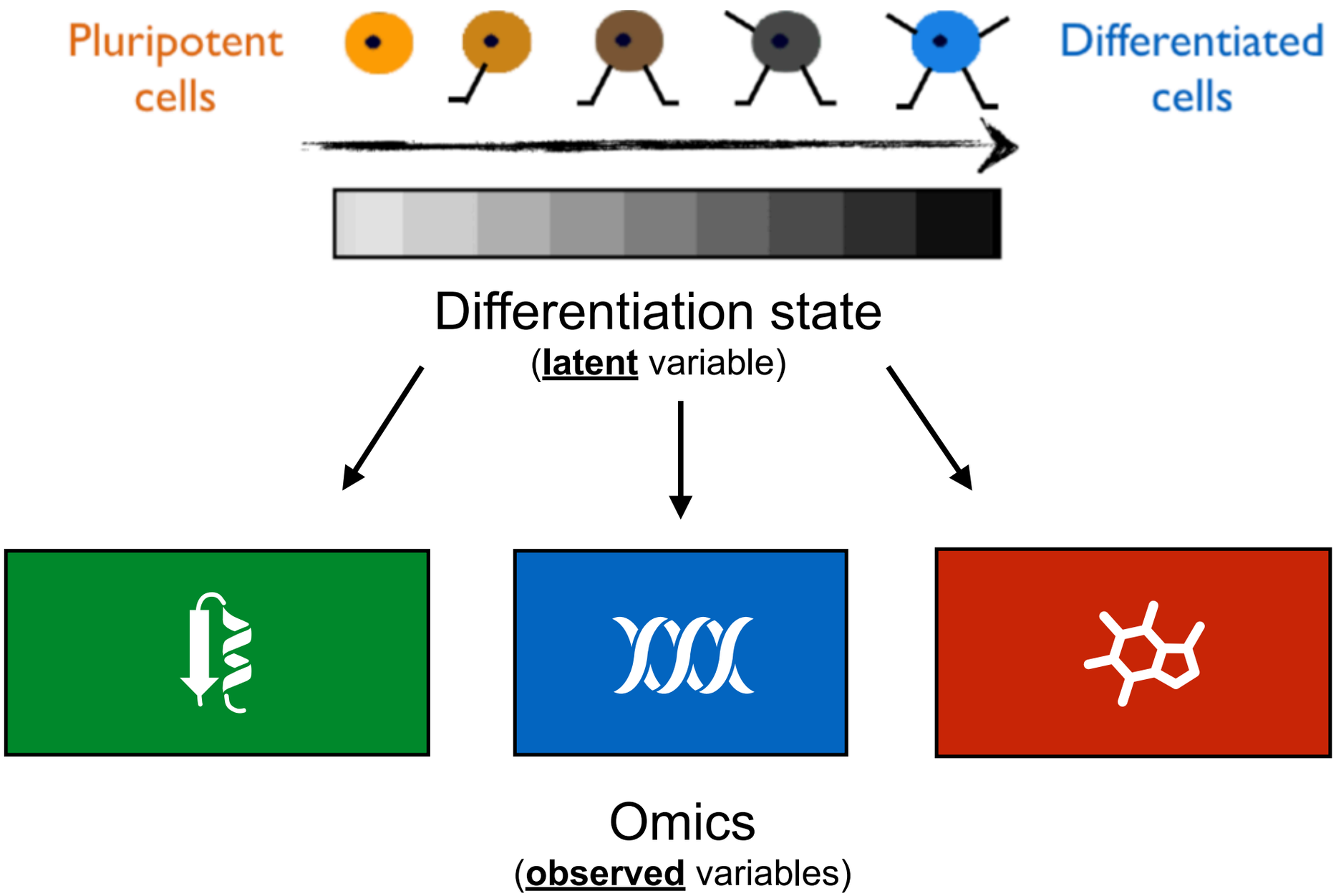
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

- MOFA is a probabilistic generalisation of PCA to multi-omics data
- 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

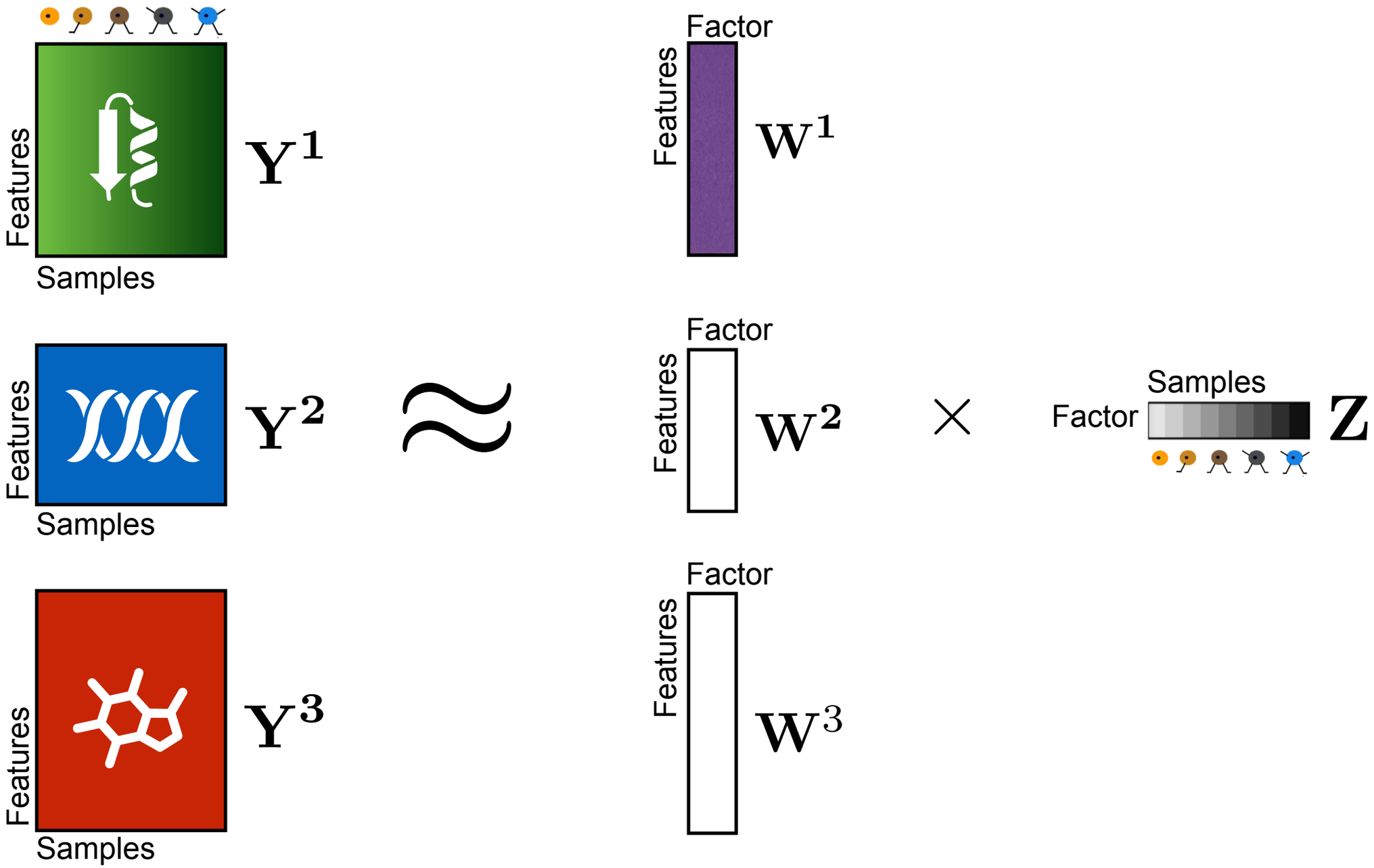

Pluripotent cells
Differentiated cells

Illustrative example

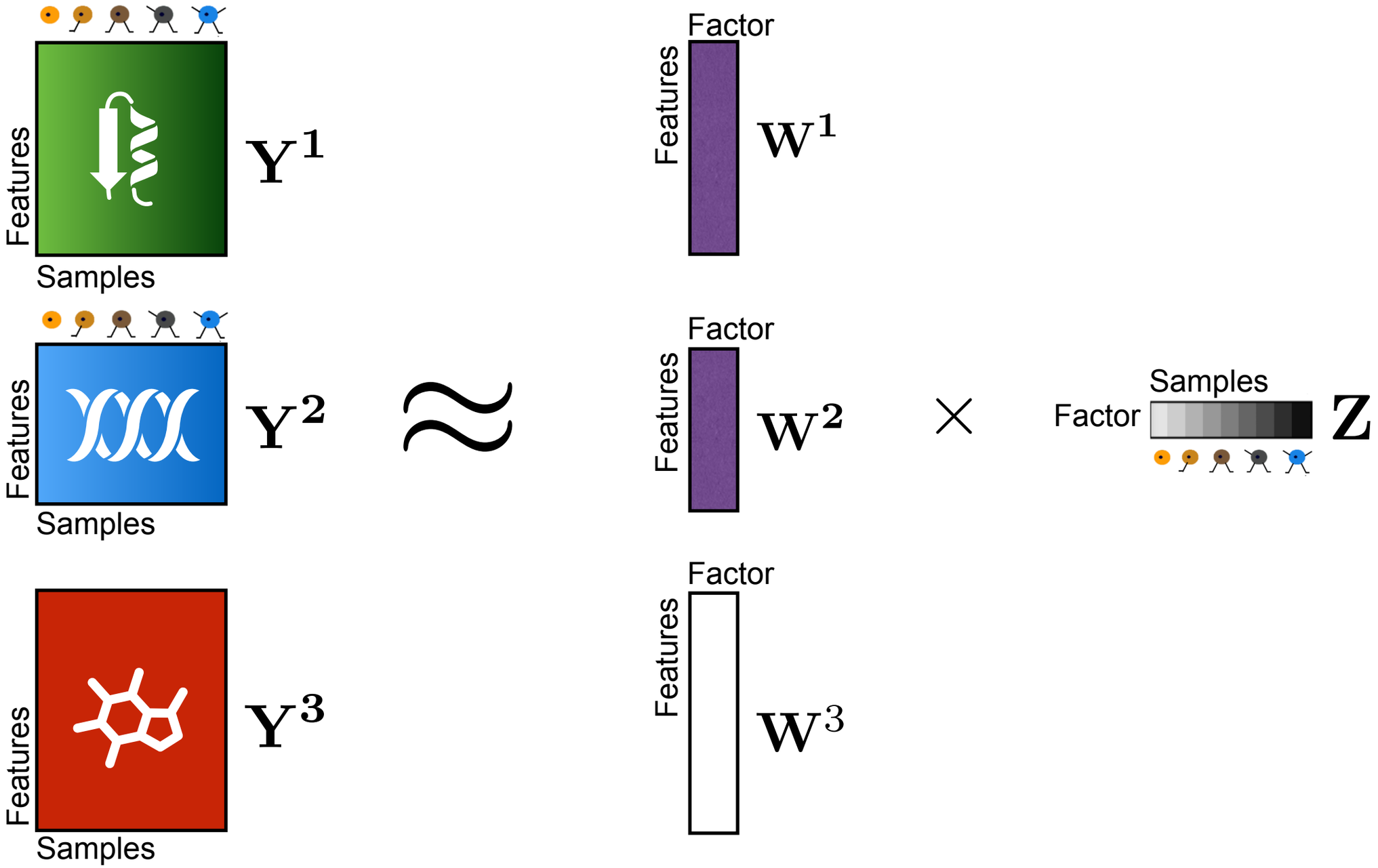



Pluripotent cells
Differentiated cells
Differentiation state
(latent variable)
Omics
(observed variables)

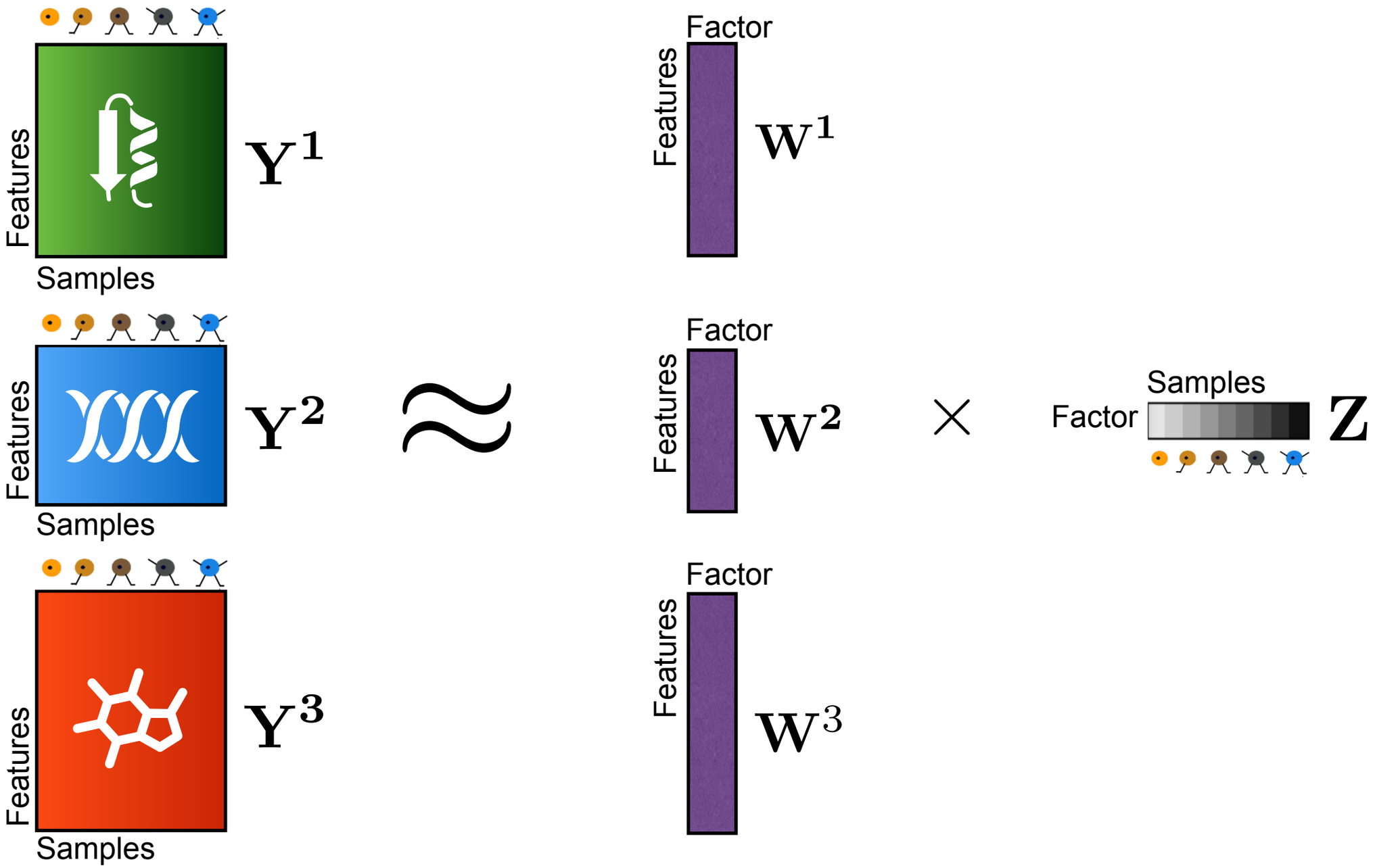
The differentiation state is the only driver of variation in **transcriptomics**



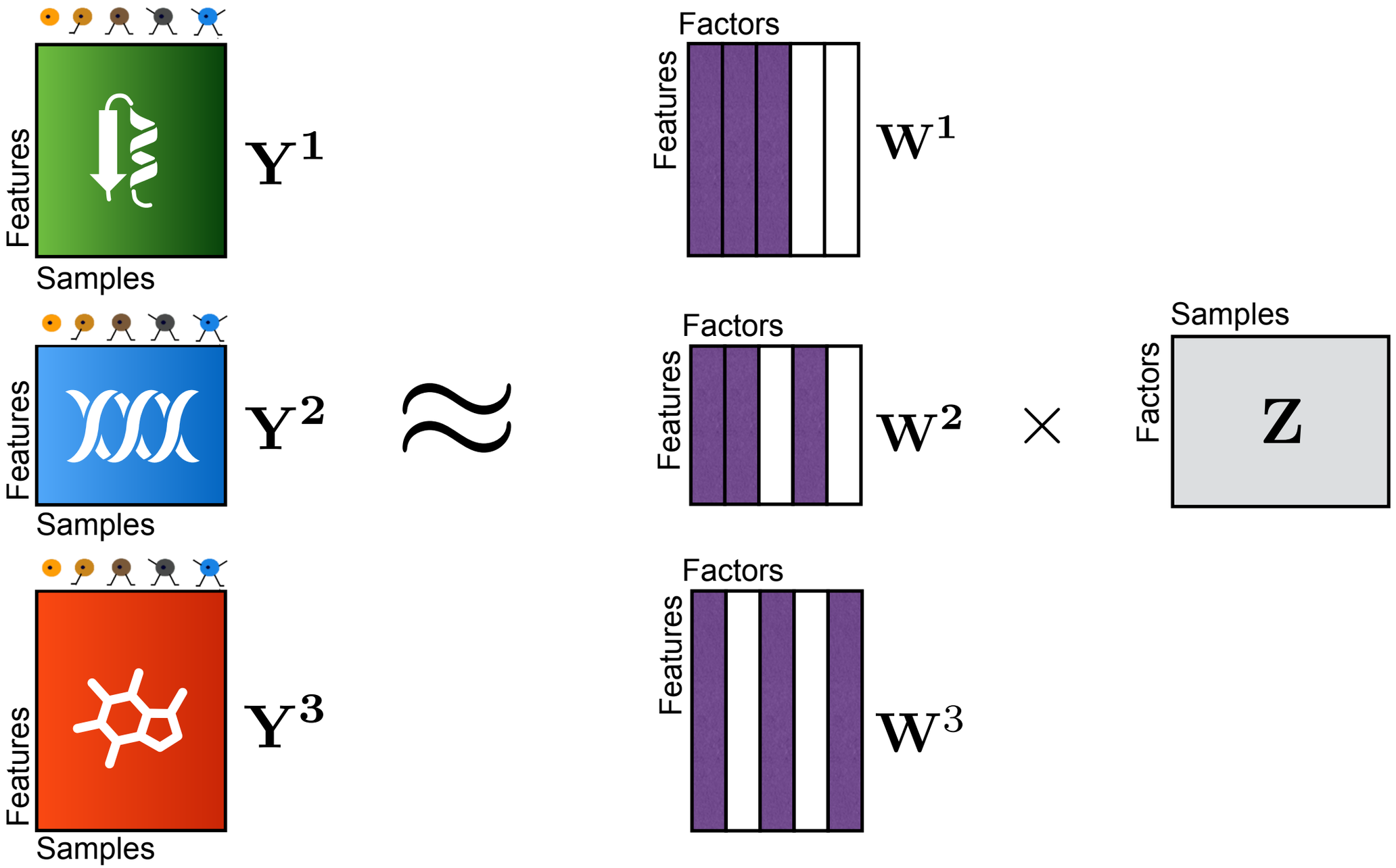
The differentiation state is the only driver of variation in **transcriptomics and genetics**



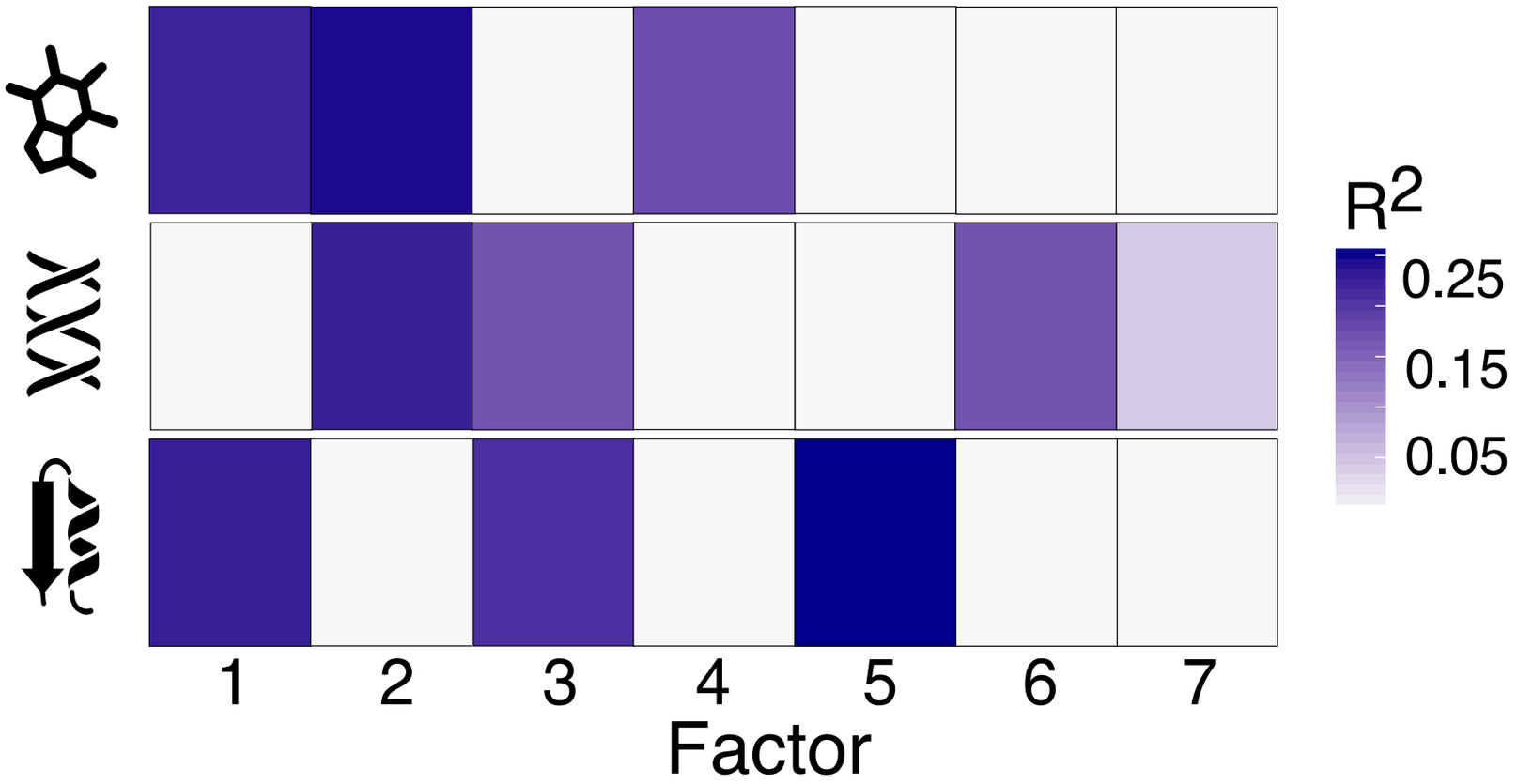
The differentiation state is the only driver of variation in **all omics**



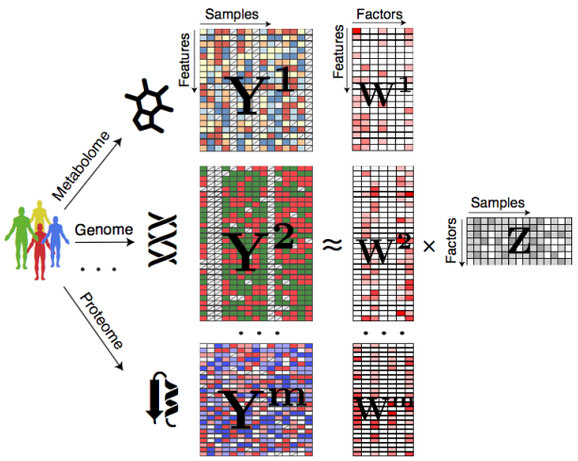
A more realistic solution…



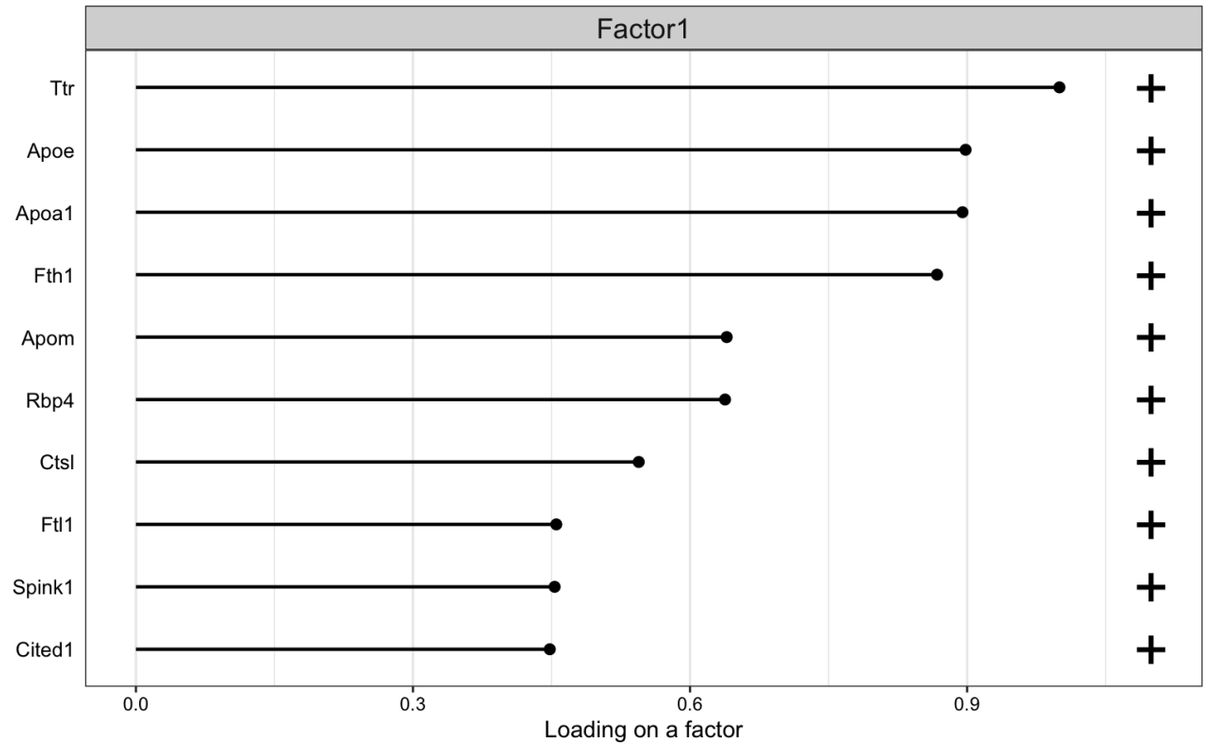
Variance decomposition by factor



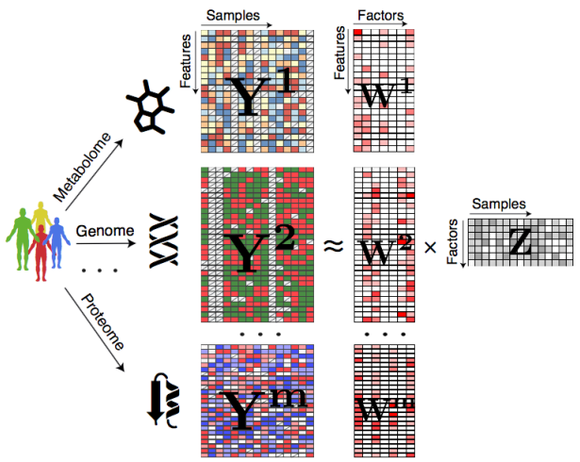
Downstream analysis

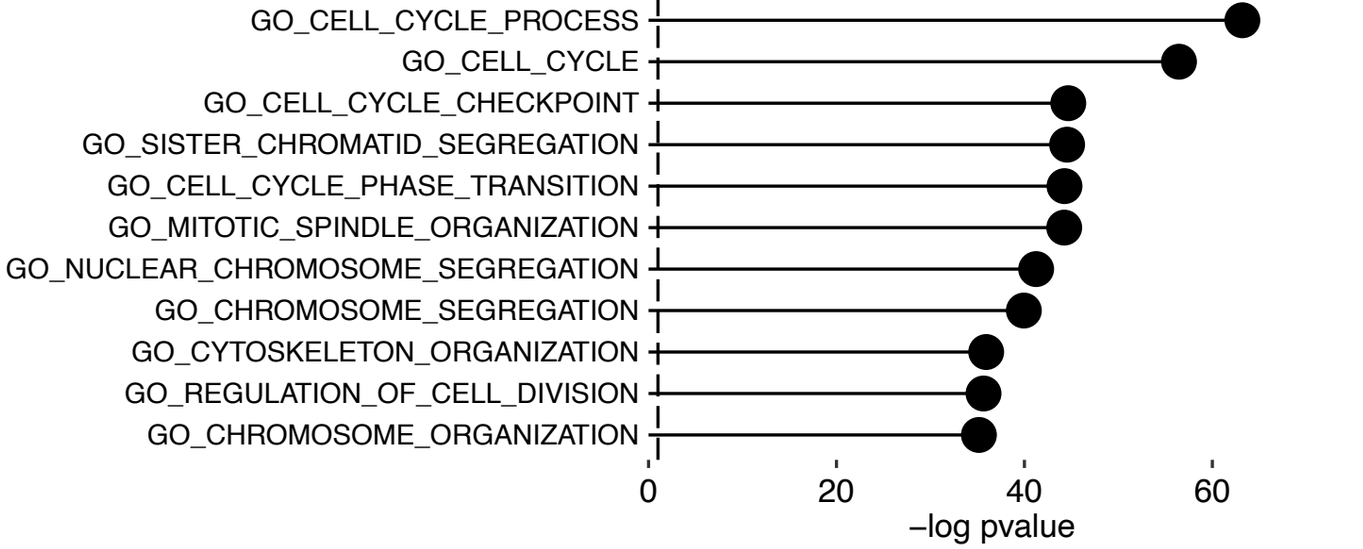


Inspection of feature weights

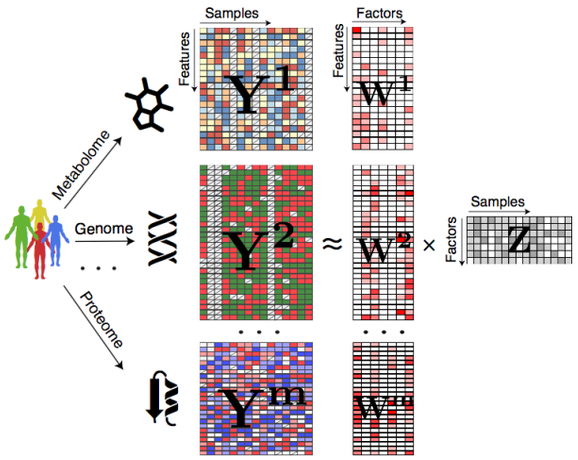


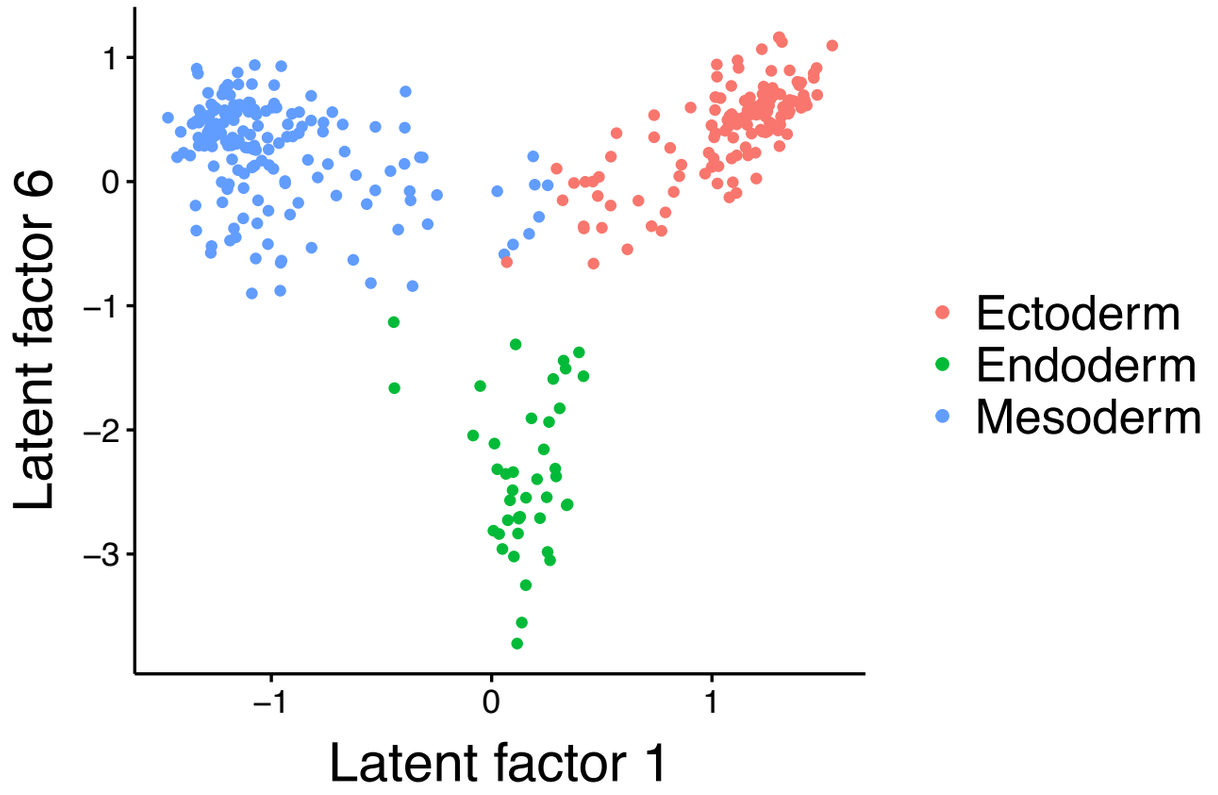
Downstream analysis



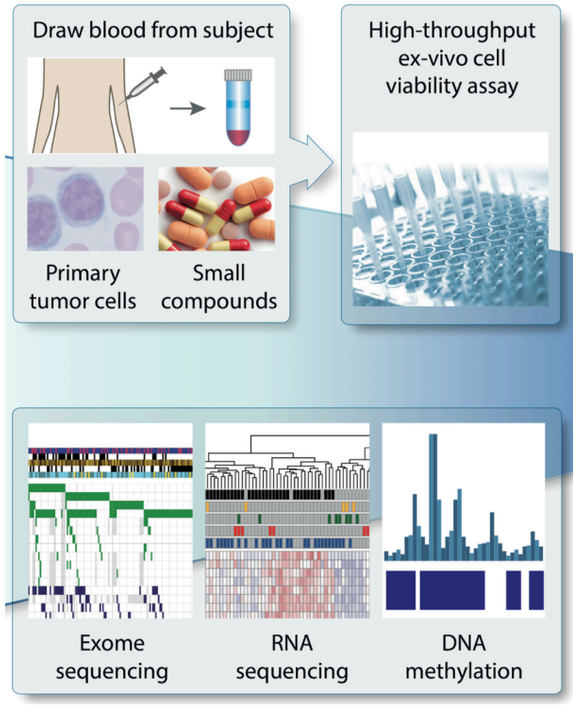


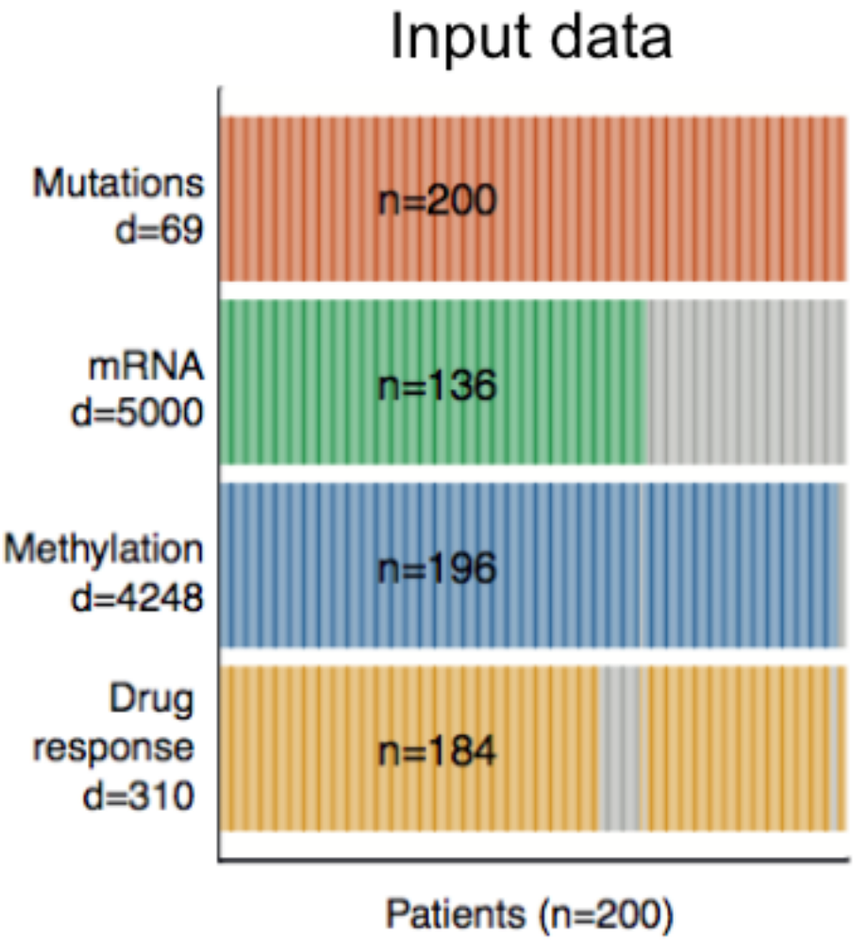
Downstream analysis



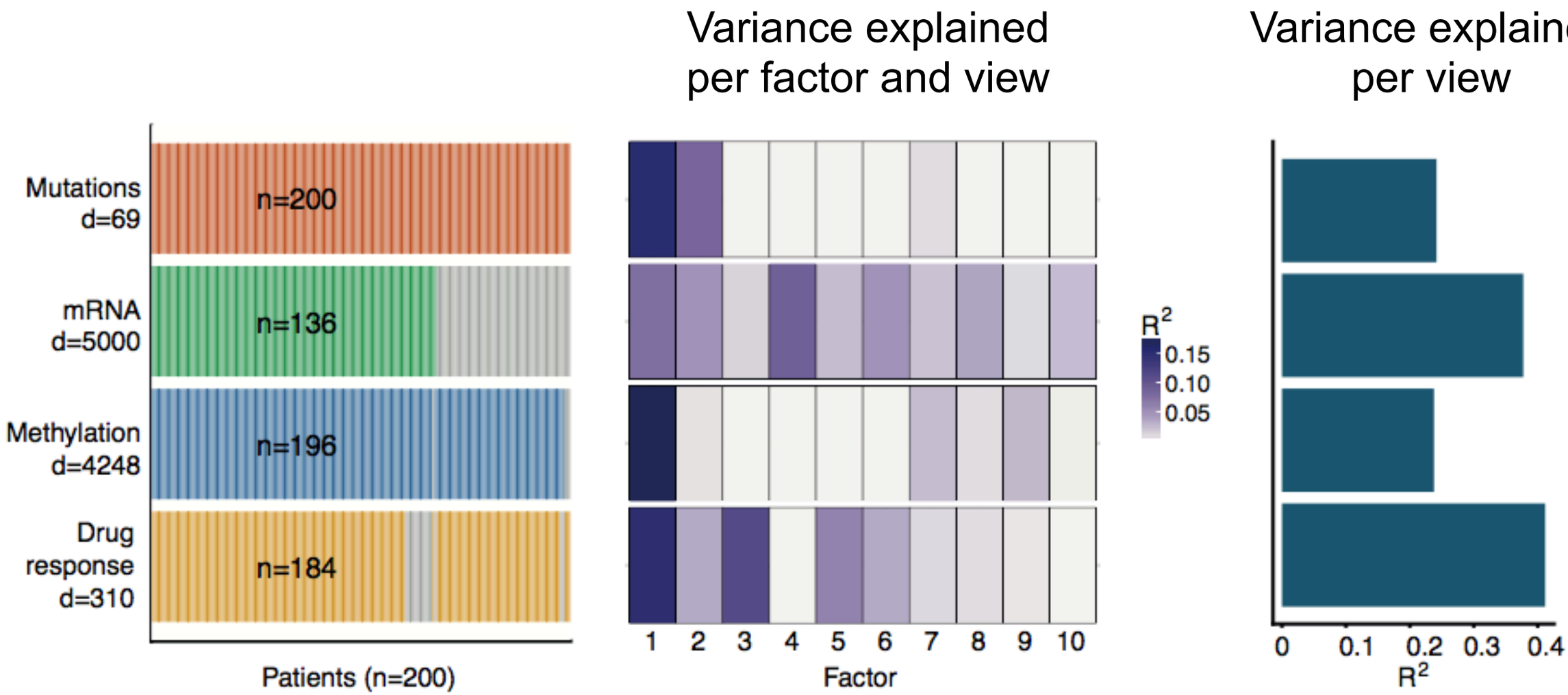


Personalised medicine application, a cohort of Chronic Lymphocitic Leukemia patients

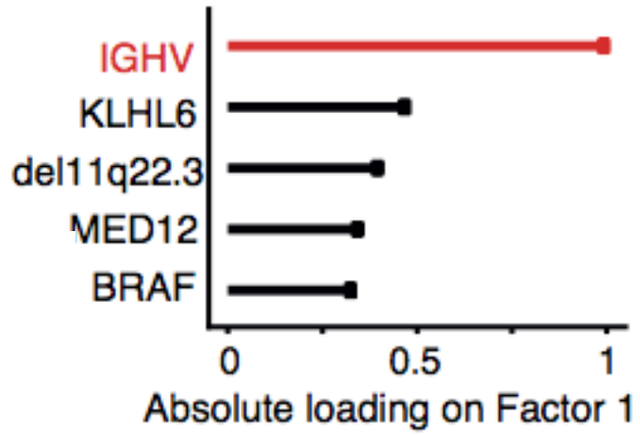


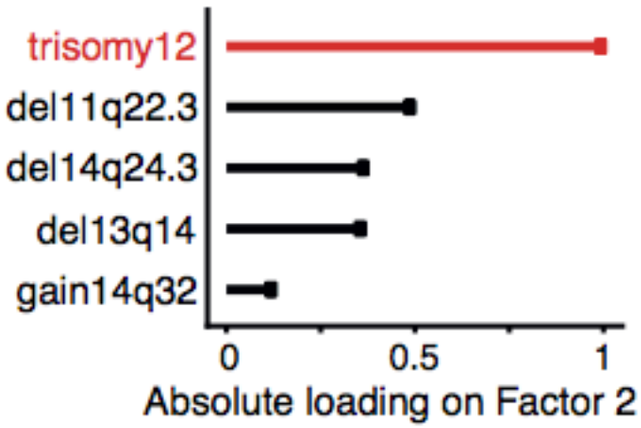


Personalised medicine application, a cohort of Chronic Lymphocitic Leukemia patients



Inspection of Somatic mutation 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

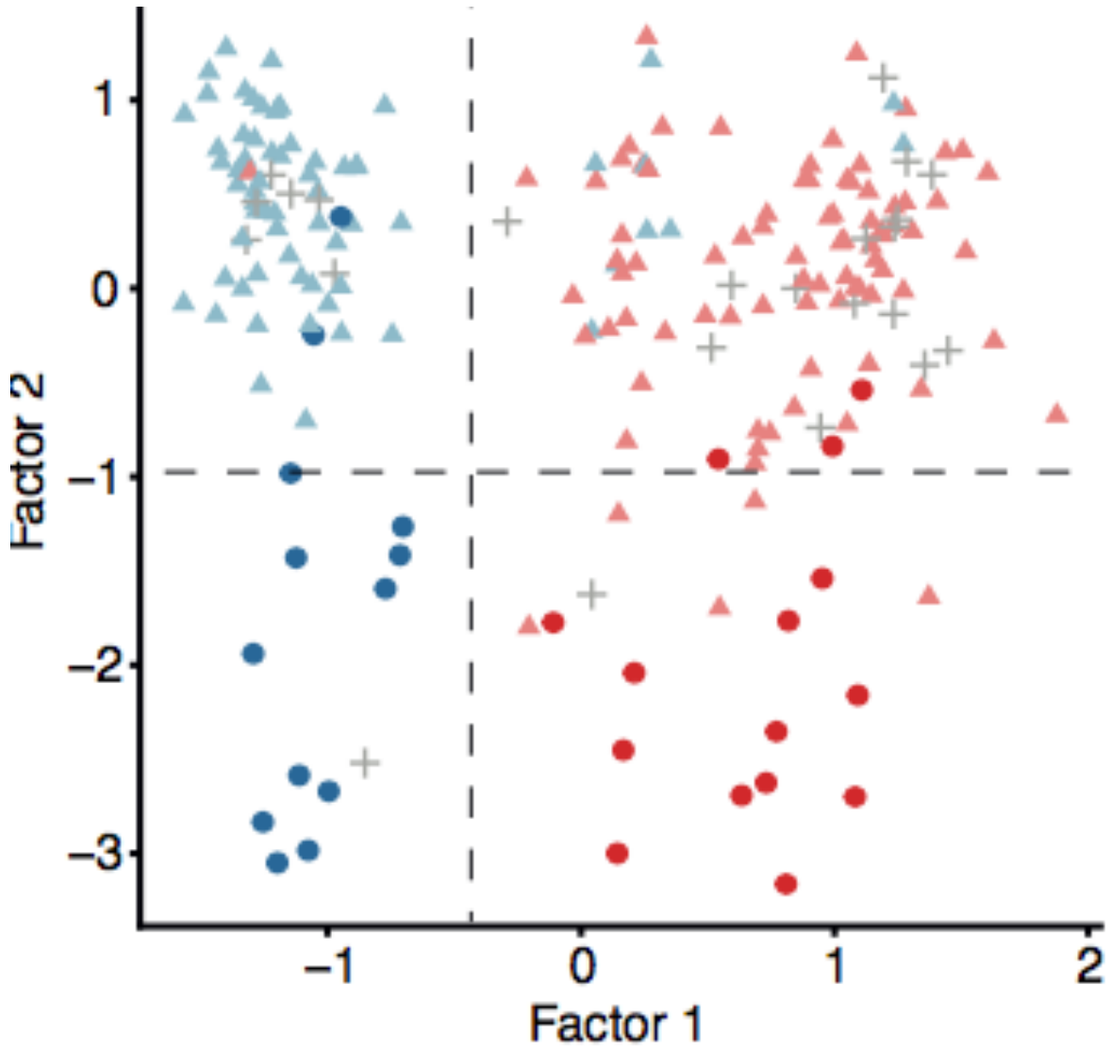
Trisomy 12 chronic lymphocytic leukemia cells

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

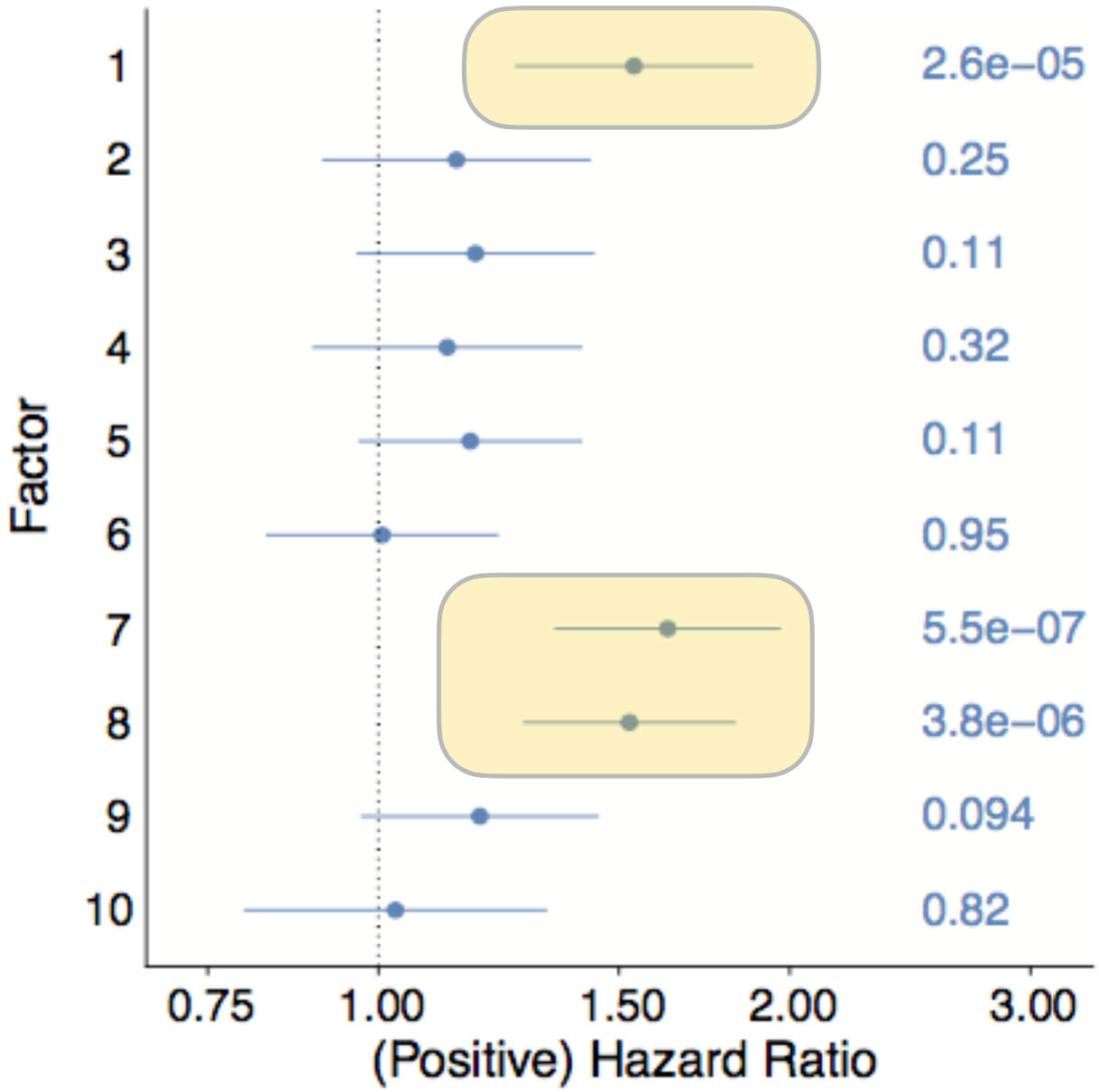
Blood 2014 123:4101-4110; doi: https://doi.org/10.1182/blood-2014-01-552307

Visualisation of samples in the latent space

Factor 1: IGHV+ vs IGHV-
Factor 2: tr12+ (circle) vs tr12- (triangle)



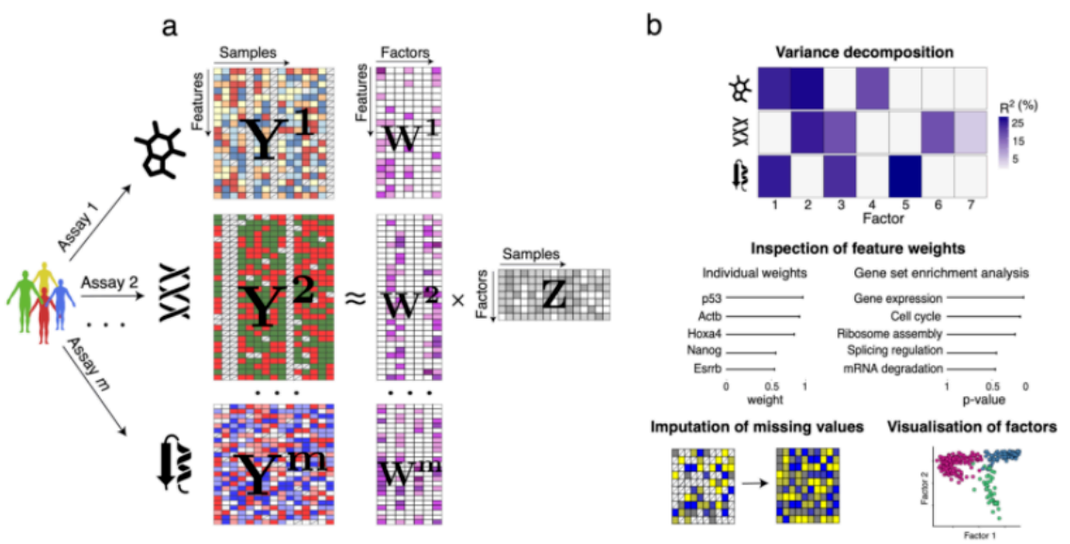
Association of Factors with clinical covariates



MOFA

Multi-Omics Factor Analysis V2 (MOFA+)

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.



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