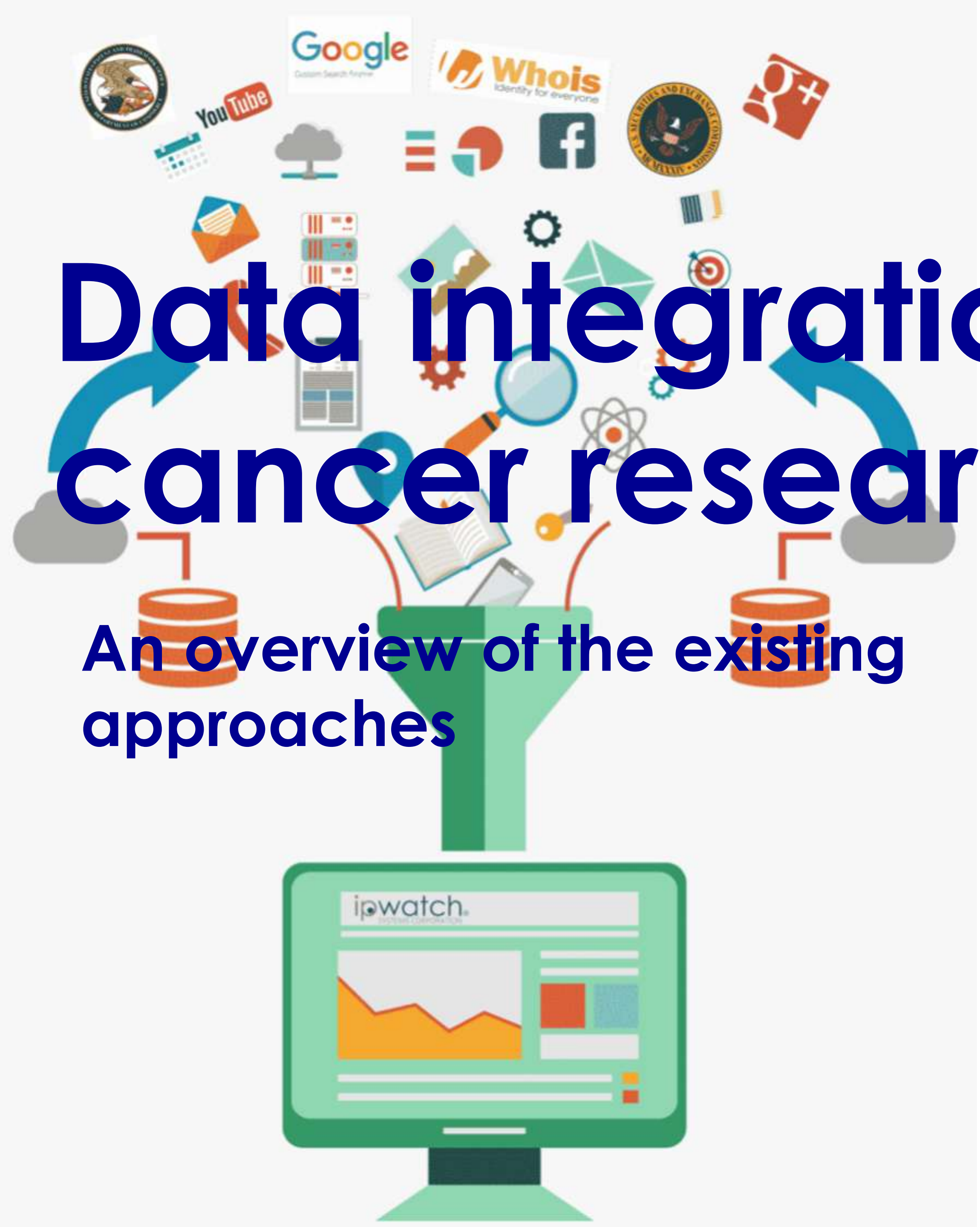


08-10-2020
EBAI2020, Roscoff

Data integration in cancer research

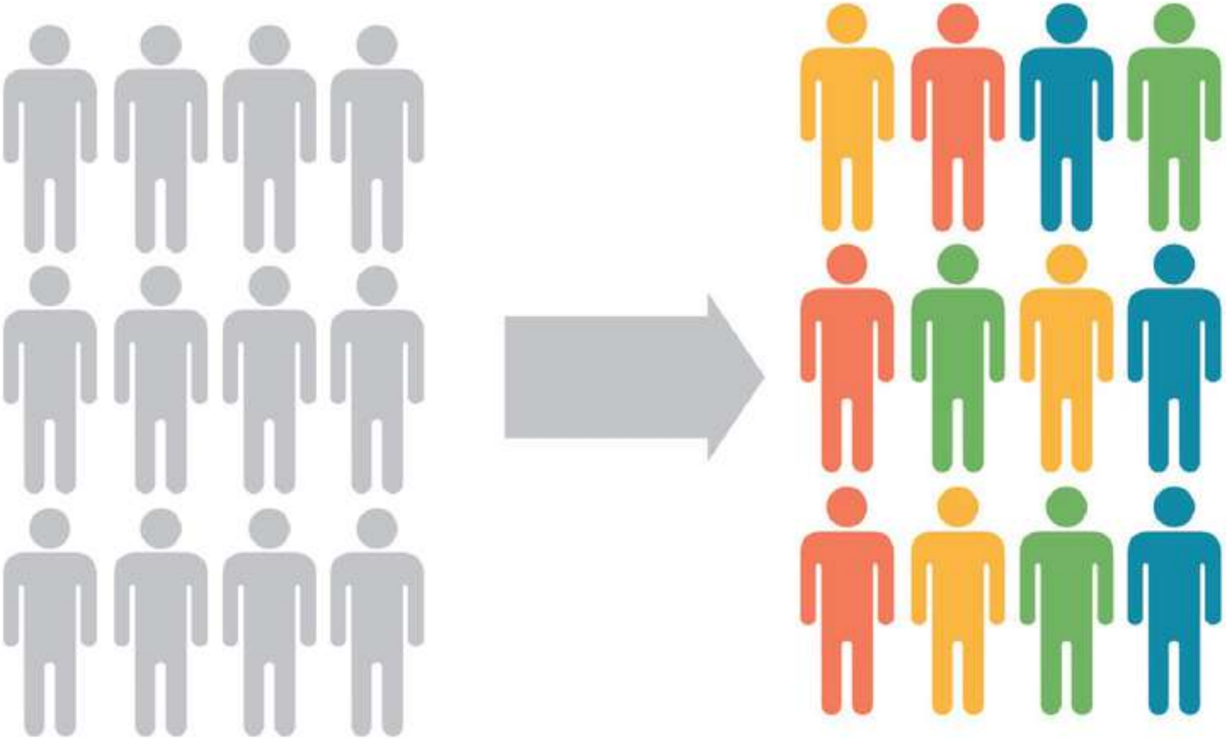
An overview of the existing approaches

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale
IBENS
cnrs
ENS
ÉCOLE NORMALE SUPÉRIEURE
PSL RESEARCH UNIVERSITY PARIS

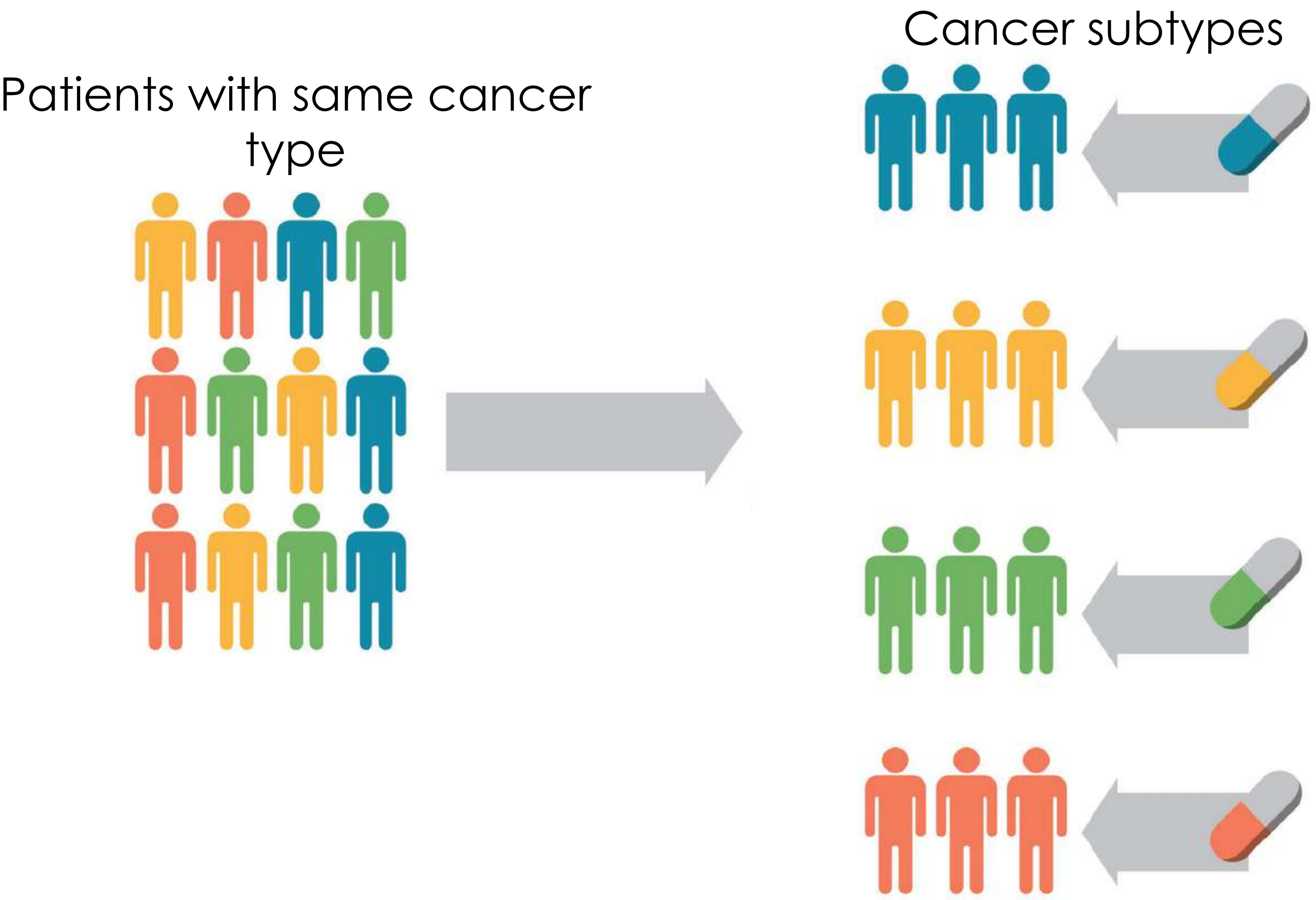
Laura Cantini
Computational Systems Biology Team
IBENS, Paris

Personalized cancer medicine

Patients with same cancer type don’t have the same survival, treatment response and molecular characteristics



Personalized cancer medicine

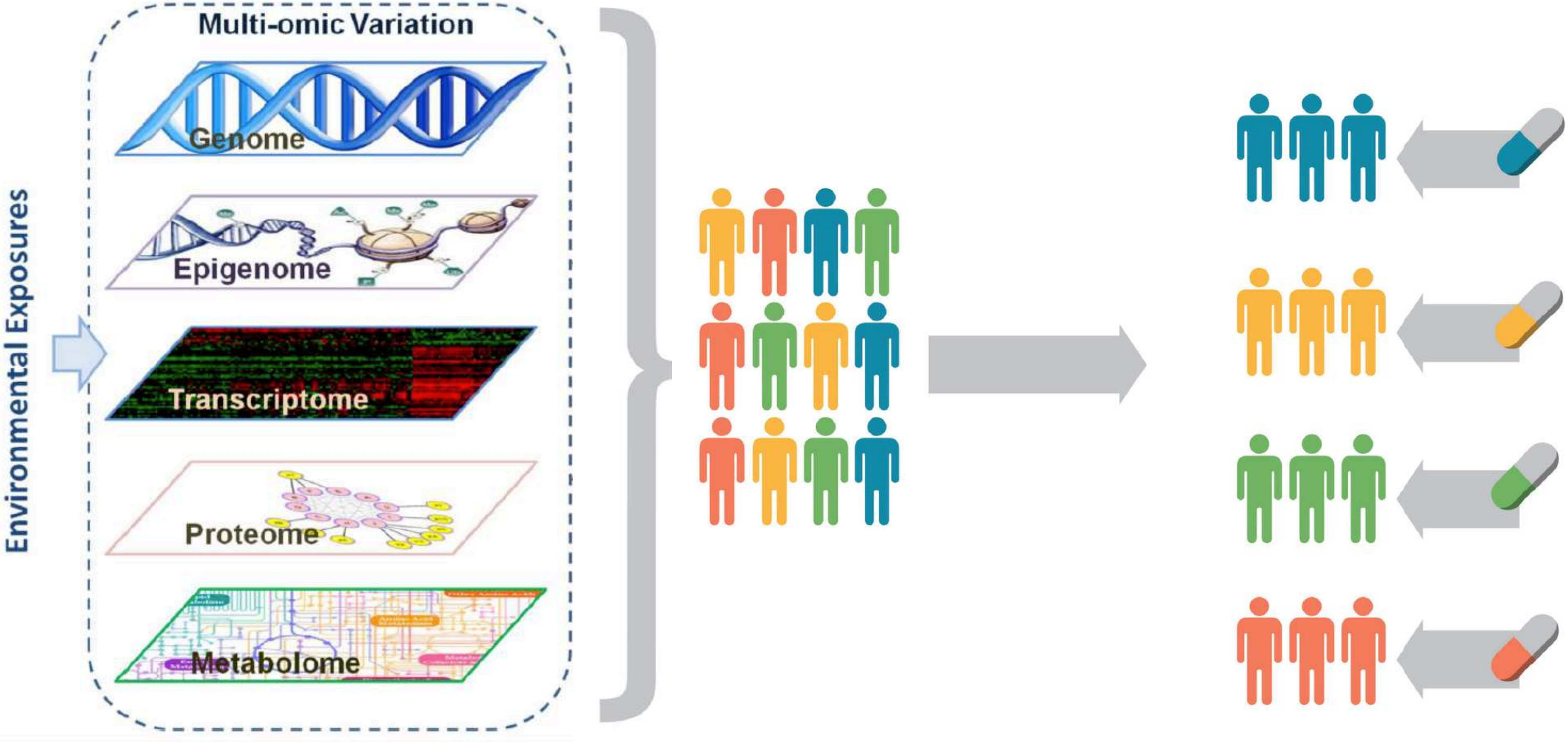


Patients with same cancer type

Cancer subtypes

Classify cancer patients into groups with similar prognosis, drug response or molecular features

Multi-omics data available

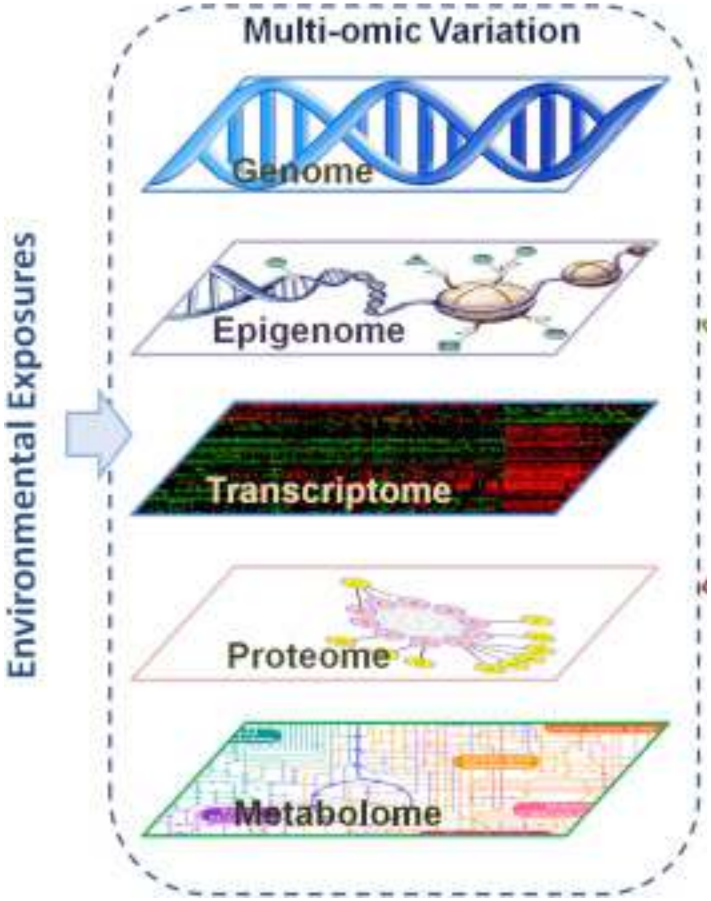


The Cancer Genome Atlas (TCGA) for example contains data from 10.000 patients, 33 cancer types, 6 omics, plus clinical data

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

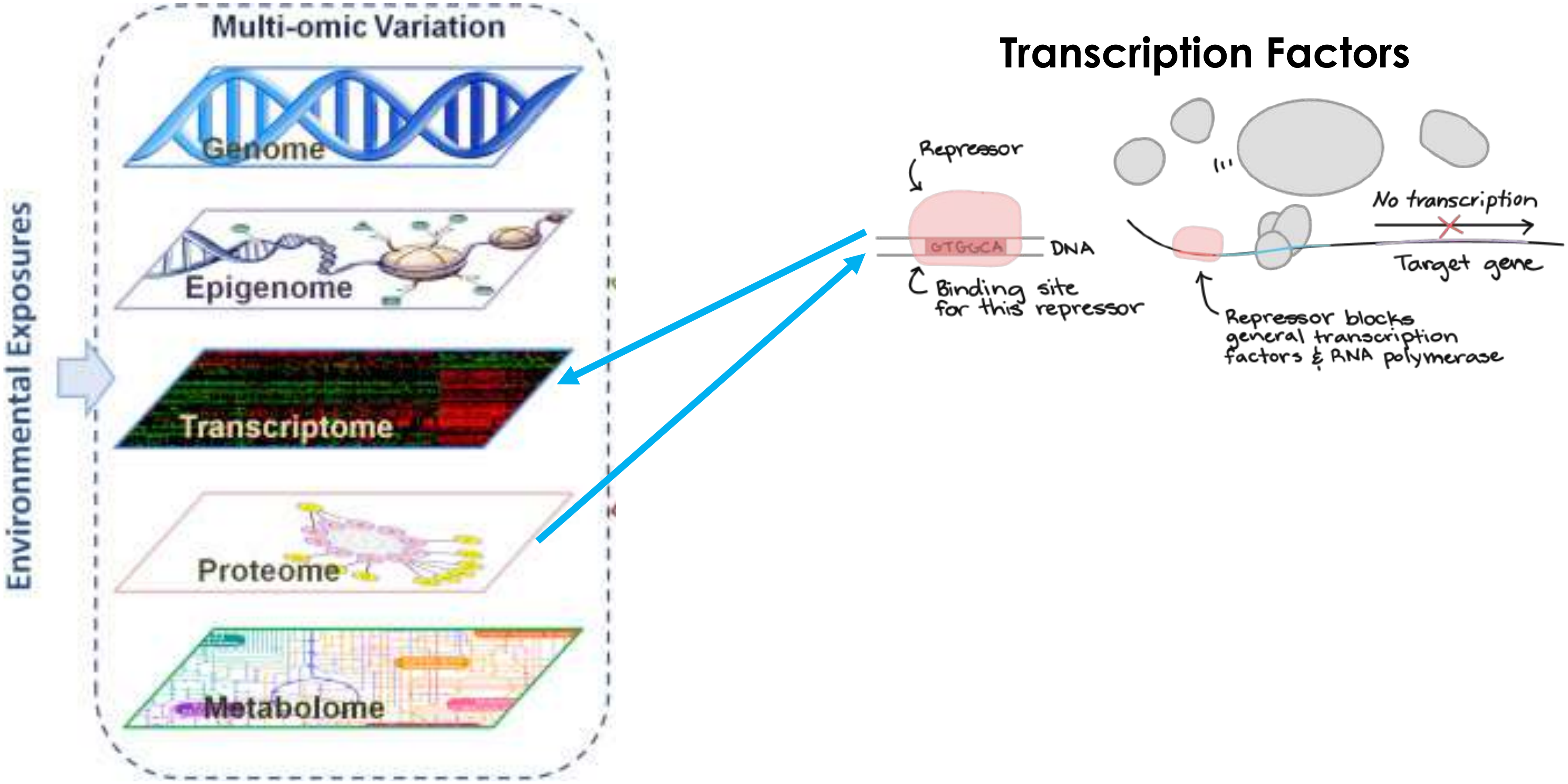
Multi-omics data are interconnected



The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

Multi-omics data are interconnected

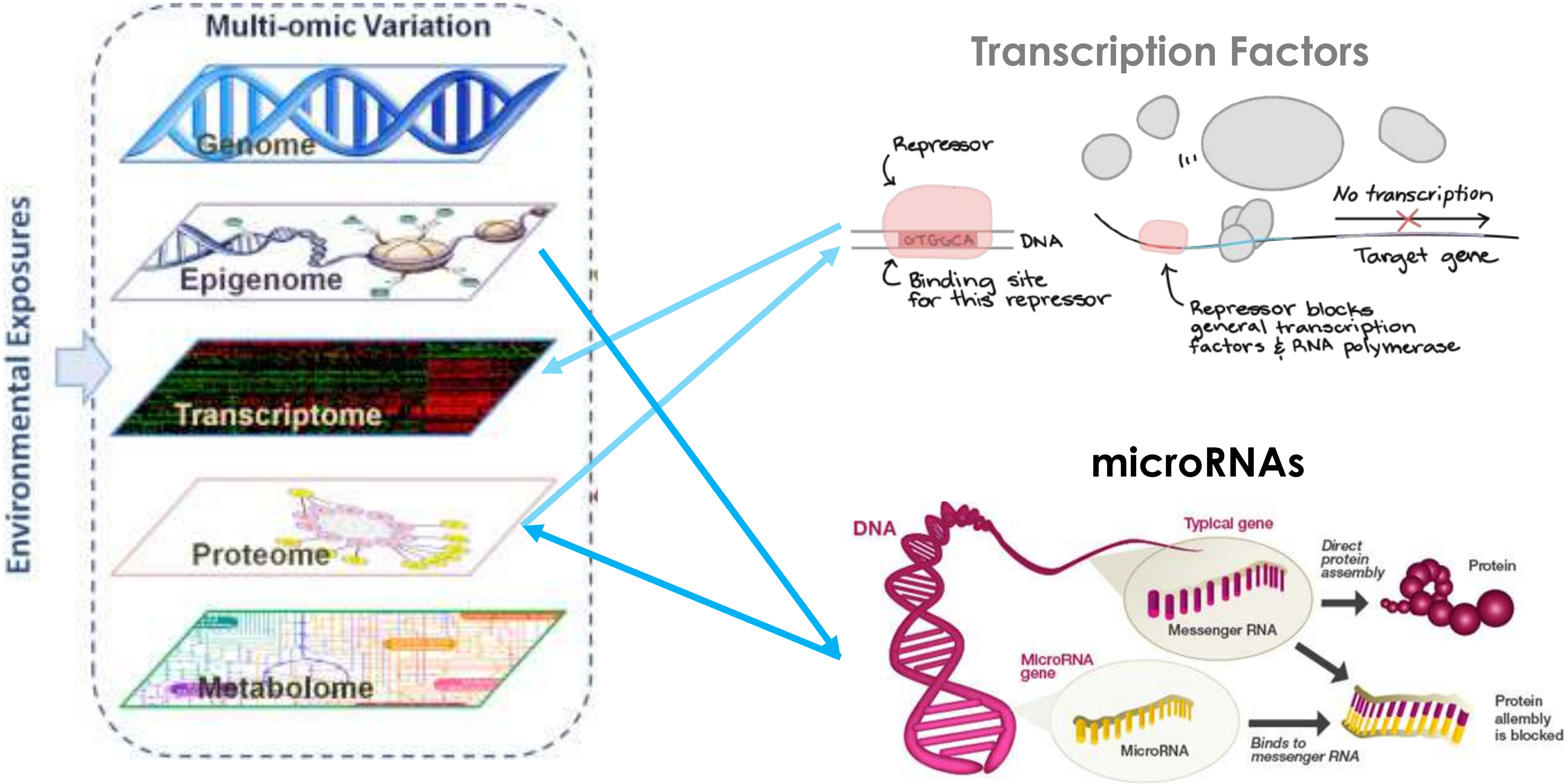


Transcription Factors

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

Multi-omics data are interconnected



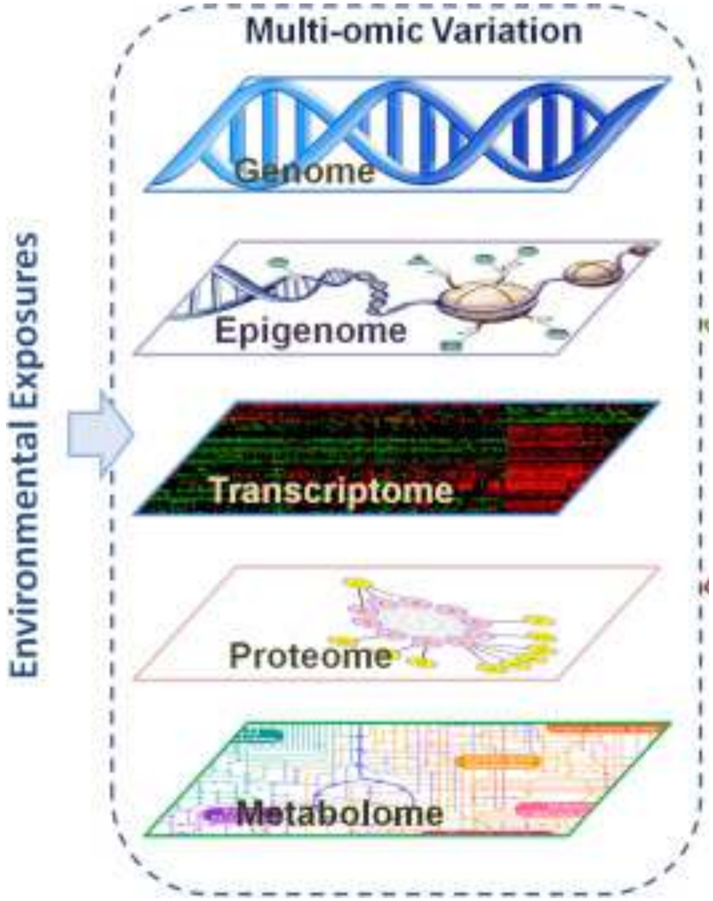
Transcription Factors

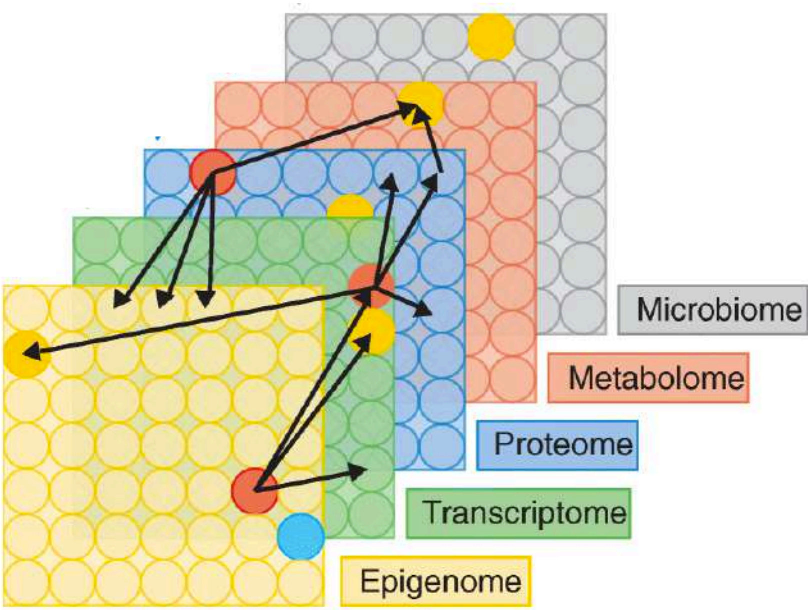
microRNAs

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

Multi-omics data are interconnected





The joint analysis of multiple omics is required

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. doi:10.1038/ng.2764

Sun, Yan V., and Yi-Juan Hu. "Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases." *Advances in genetics*. Vol. 93. Academic Press, 2016. 147-190.

Challenges of multi-omics integration

High-dimensionality -> Big-data

Heterogeneous variables

Different ranges of variation

Technical noise different for each omics

More omics is better, but how many more?

Is it always good to consider ALL the available omics?

Choosing which omics to integrate

Aim: predicting drug response

Available input data:

- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response

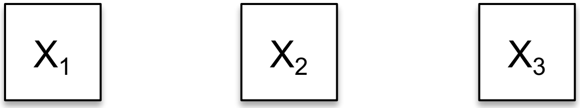
ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

Choosing which omics to integrate

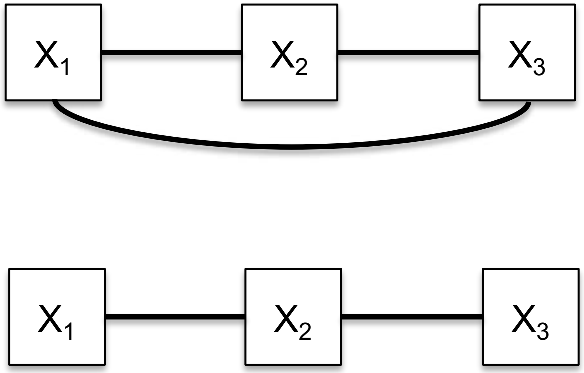
Aim: predicting drug response

Available input data:

- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response



Using correlation:



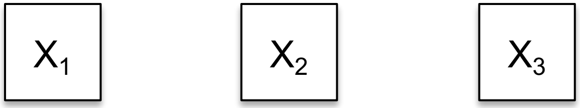
ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

Choosing which omics to integrate

Aim: predicting drug response

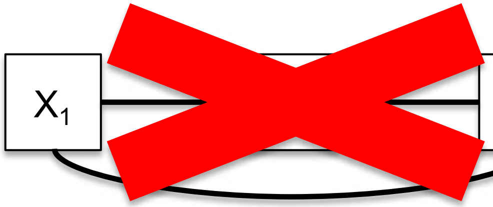
Available input data:

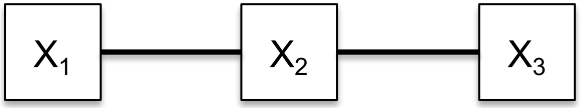
- Mutations
- Copy Number Alterations (CNA)
- Methylation
- Gene expression
- Proteomics
- Cancer types
- Drug response



Using partial correlation (iTOP):

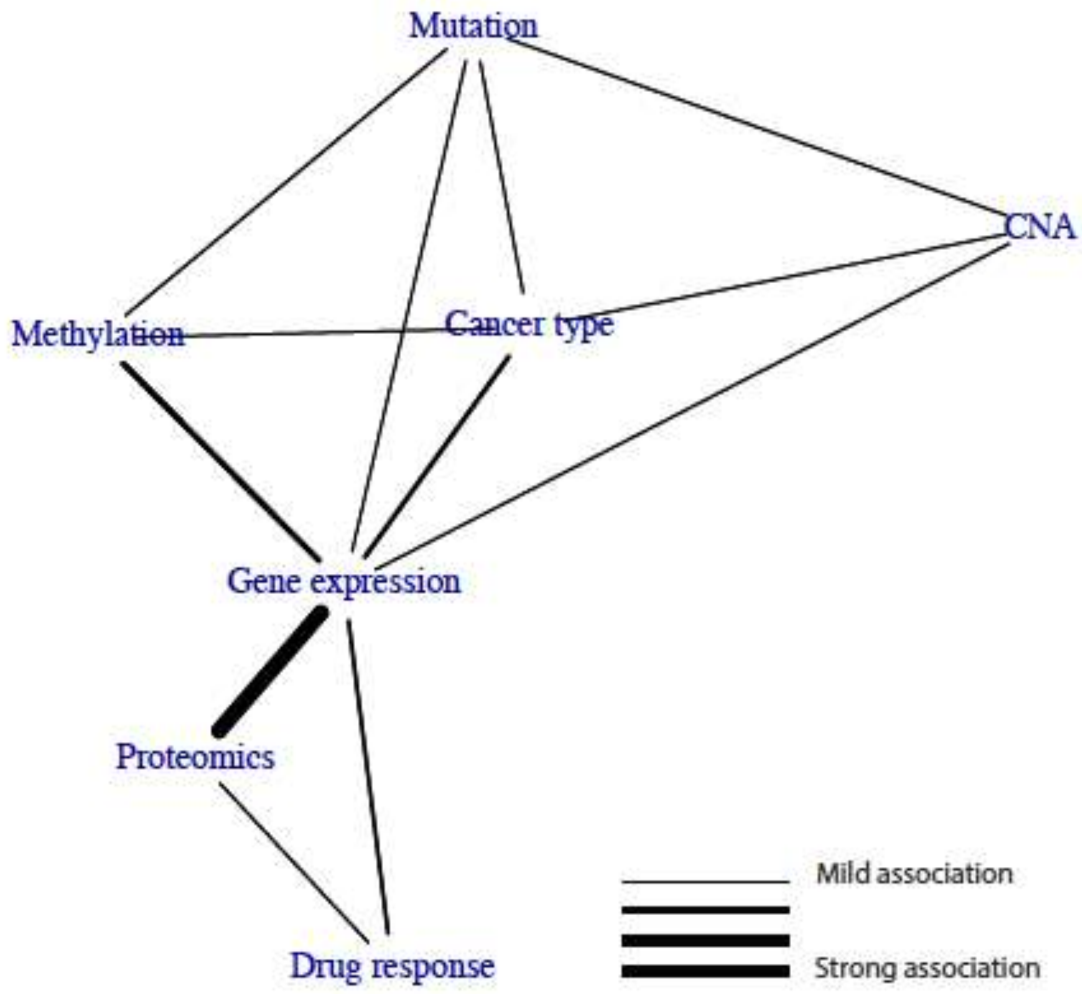
e.g. $cor(X_1,X_3 \,|\, X_2) \approx 0$





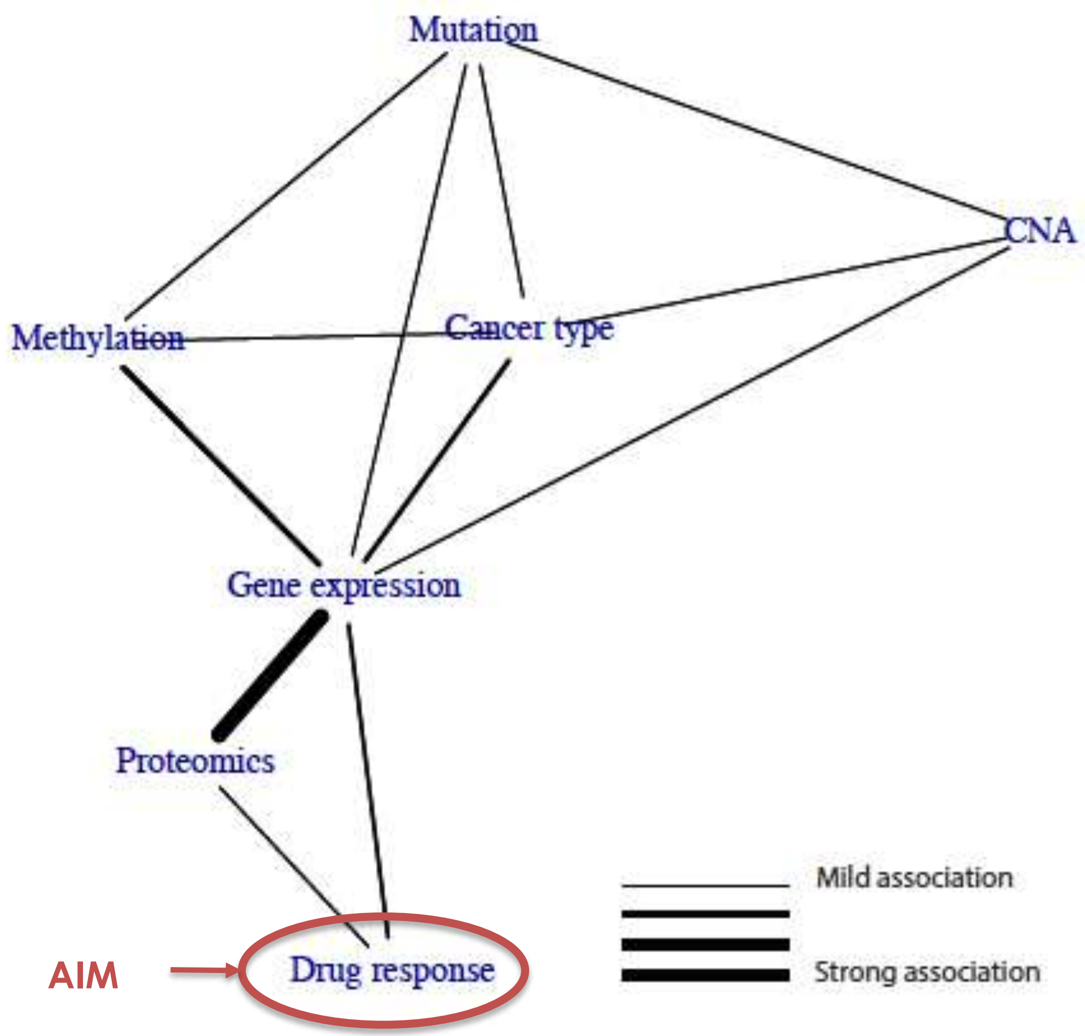
ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

Choosing which omics to integrate



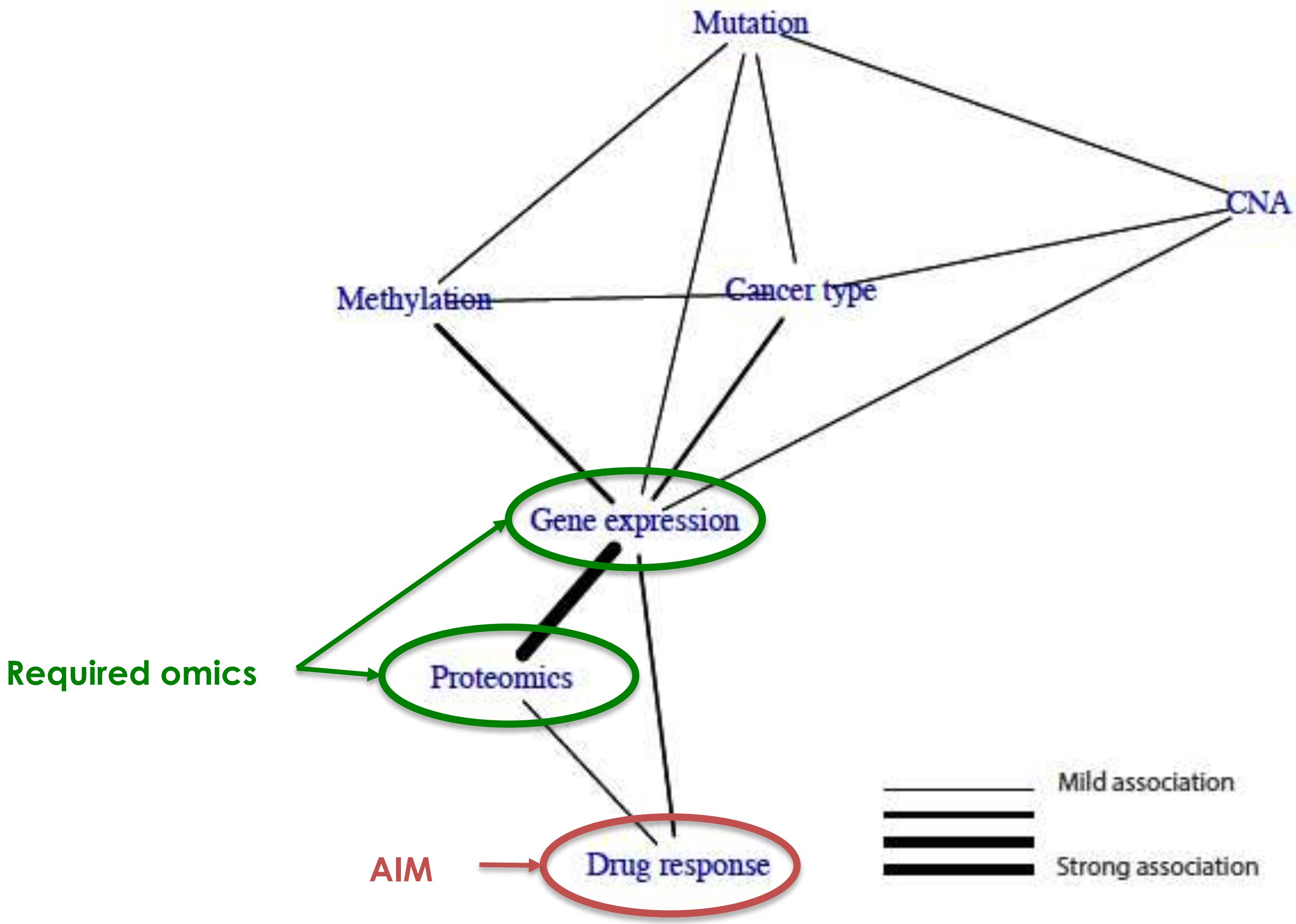
ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

Choosing which omics to integrate



ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

Choosing which omics to integrate



Required omics

AIM

ABEN, Nanne, et al. iTOP: inferring the topology of omics data. Bioinformatics, 2018, 34.17: i988-i996.

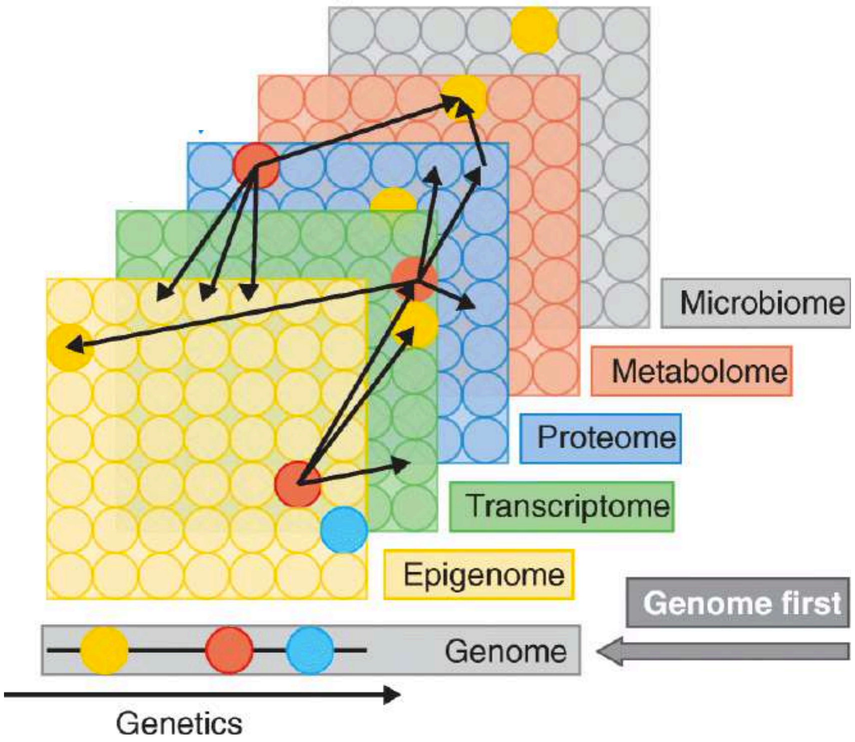
How the omics should be combined?

Integrating multi-omics data

Approach “Genome First”

Priority given to genome

Other omics are only used for interpretation



Hasin, Yehudit, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease." Genome biology 18.1 (2017): 83.

Integrating multi-omics data

Dataset 1

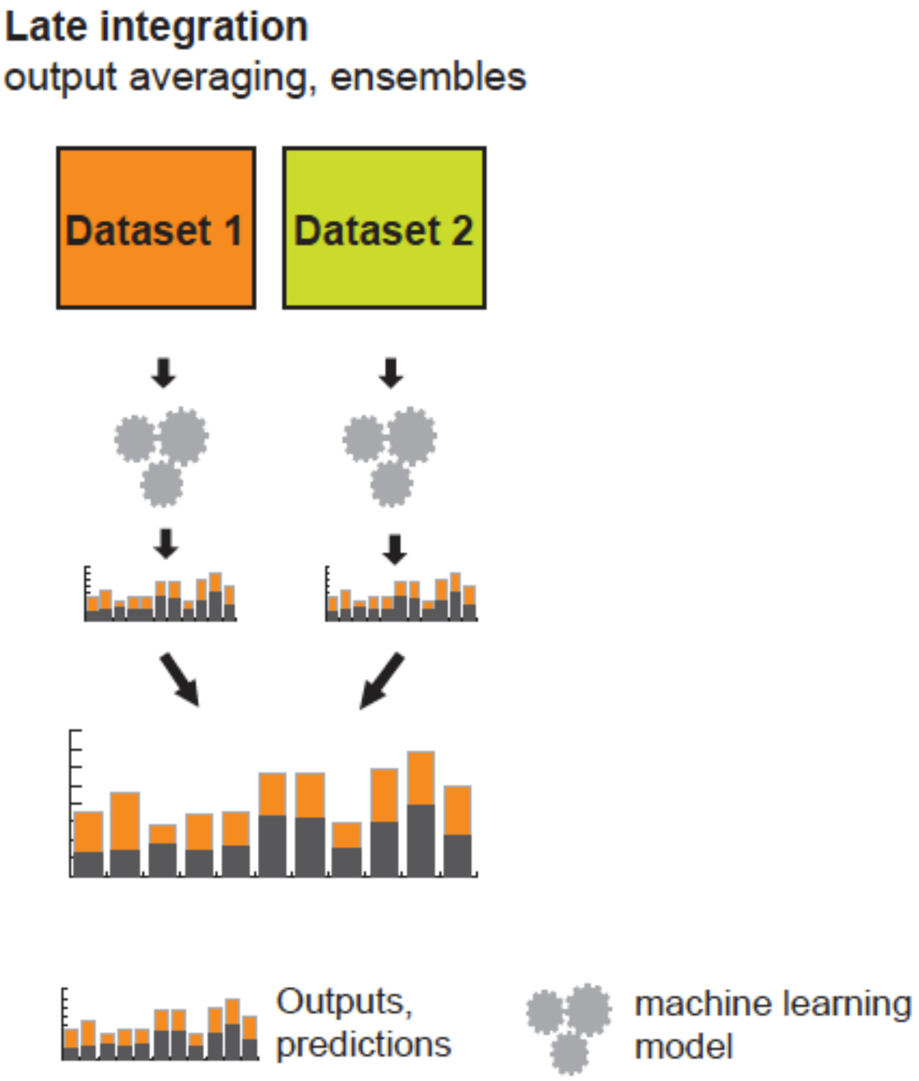
Dataset 2

Machine learning algorithm designed for a single dataset

How do I integrate them ???????????

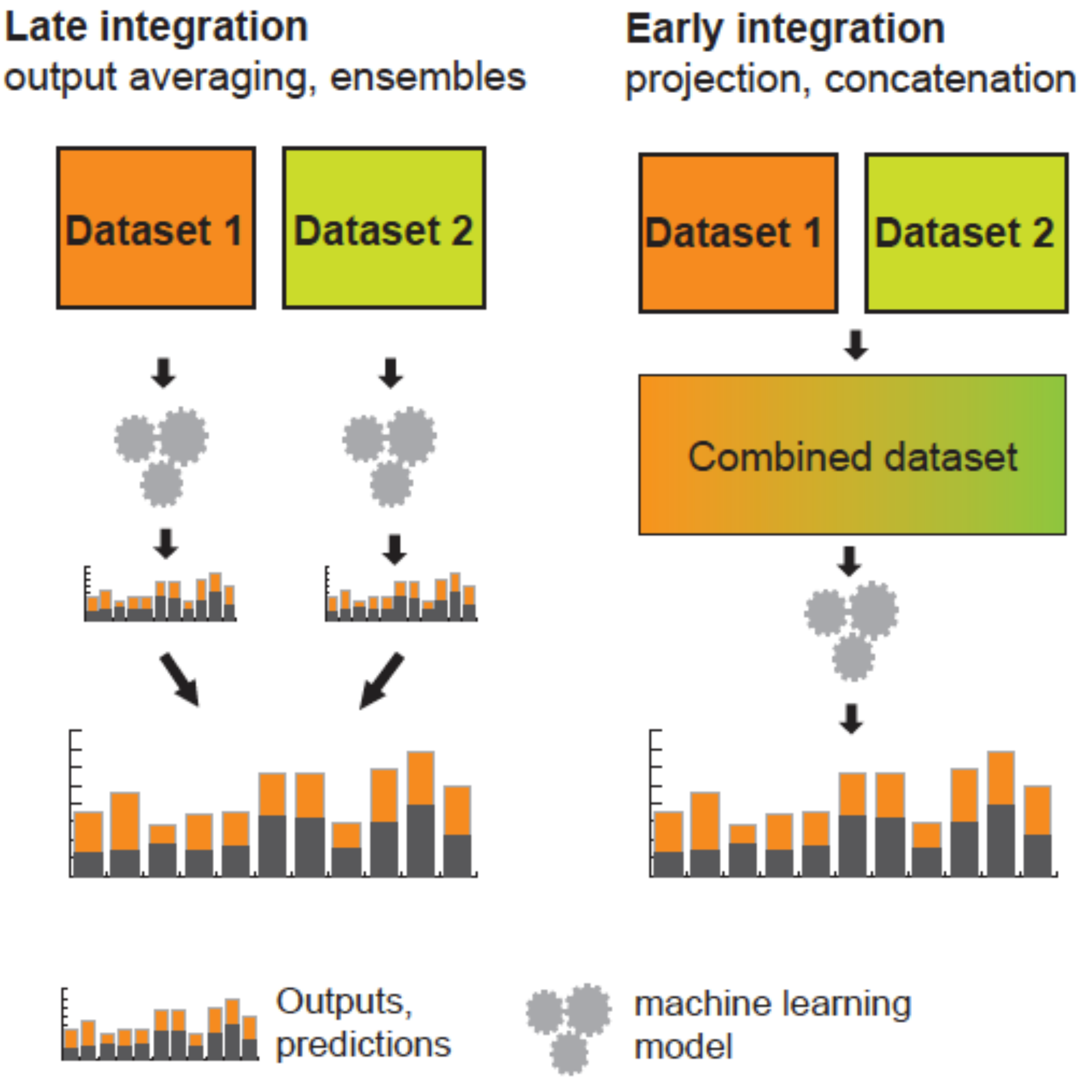
Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

Integrating multi-omics data



Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

Integrating multi-omics data



Late integration
output averaging, ensembles

Early integration
projection, concatenation

Dataset 1

Dataset 2

Dataset 1

Dataset 2

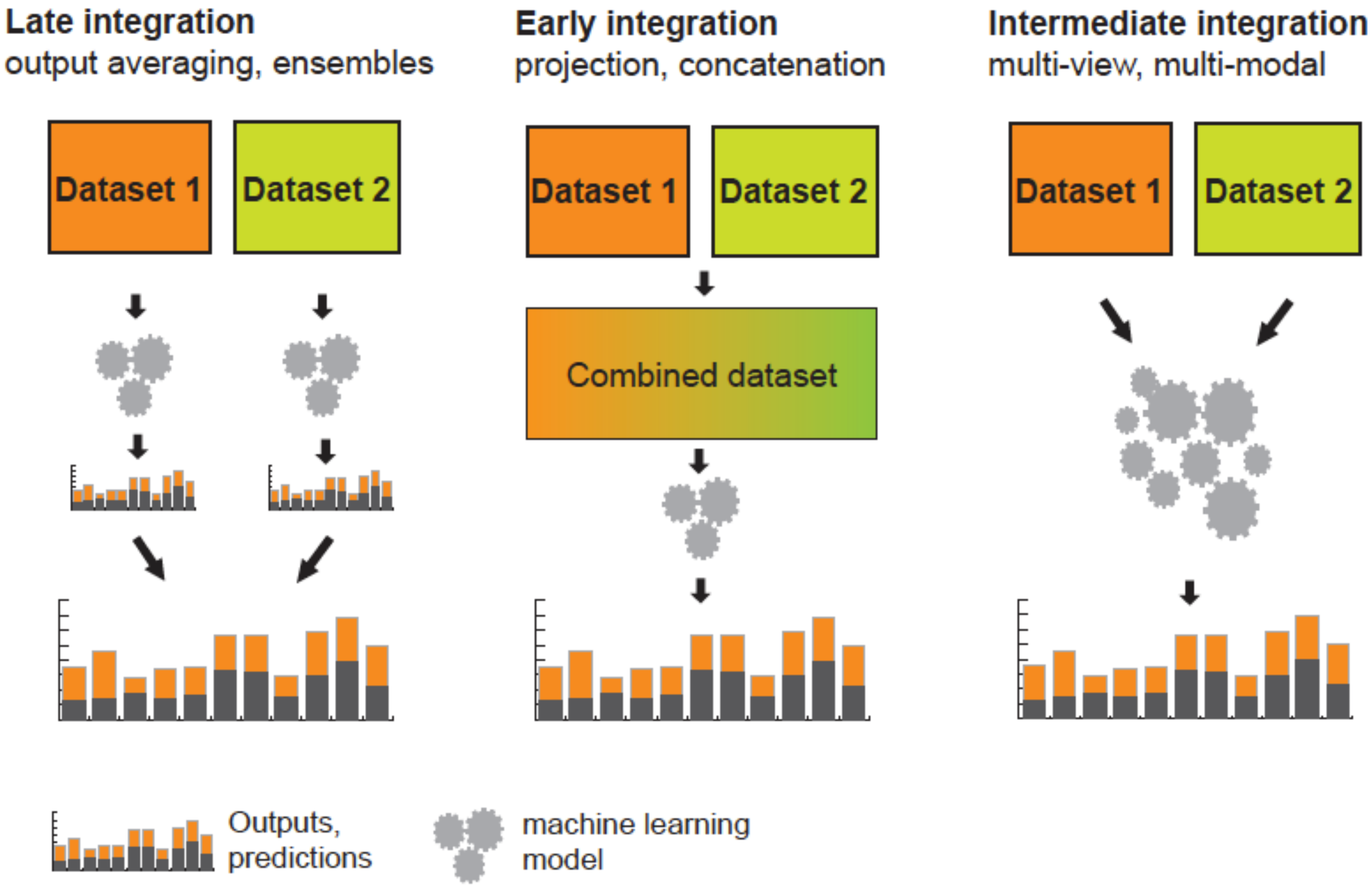
Combined dataset

Outputs, predictions

machine learning model

Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

Integrating multi-omics data



Late integration
output averaging, ensembles

Early integration
projection, concatenation

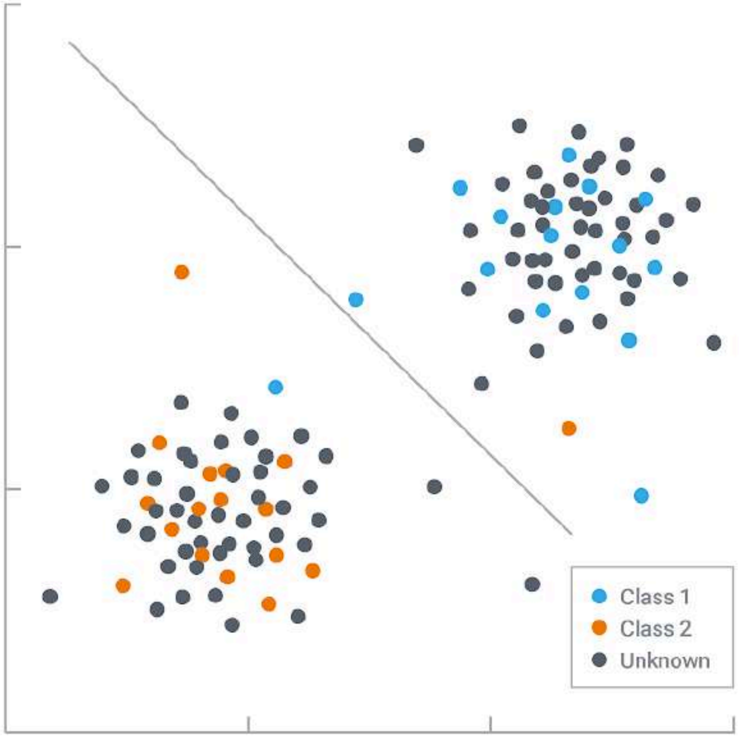
Intermediate integration
multi-view, multi-modal

Zitnik, Marinka, et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities." Information Fusion 50 (2019): 71-91.

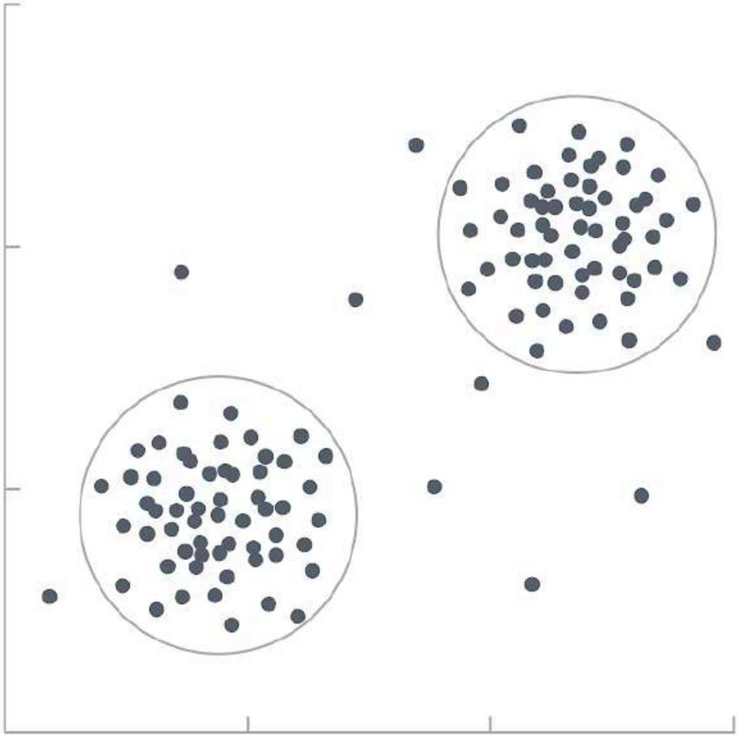
Main categories of existing multi-omics integrative approaches

Main categories of integrative approaches

Supervised methods

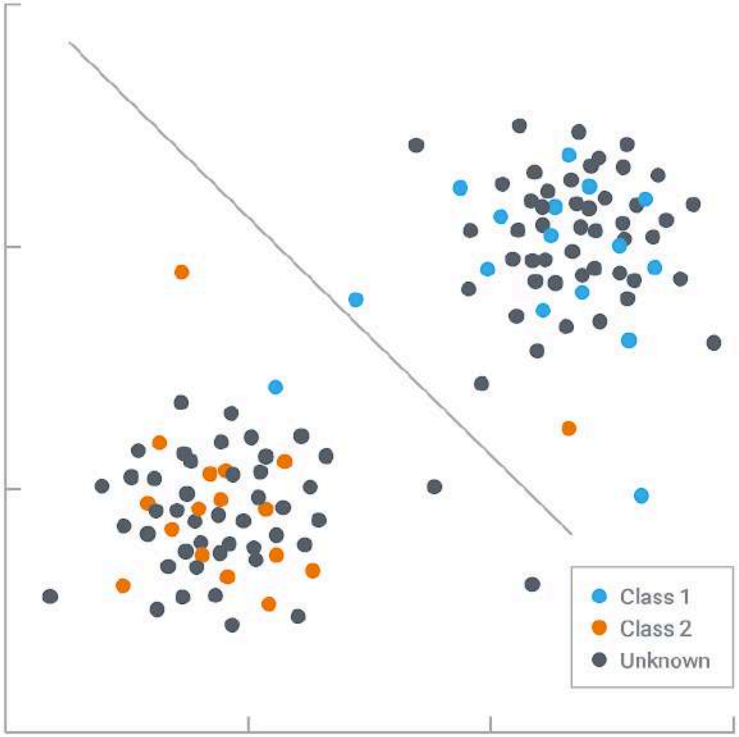


Unsupervised methods



Main categories of integrative approaches

Supervised methods

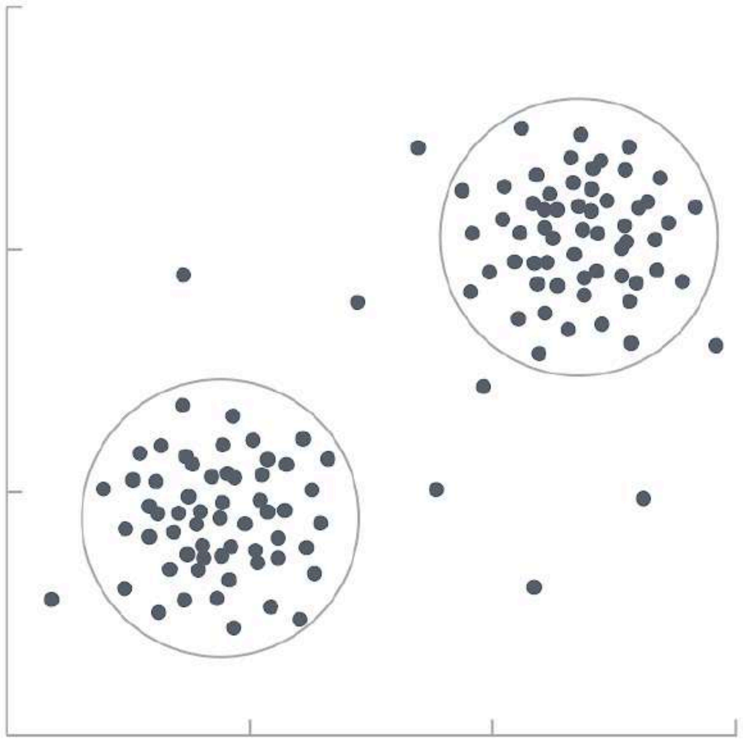


- They require 2 datasets in input: training and test datasets
- Labels must be avilable for the training dataset
- This information is used to infer labels on the test dataset

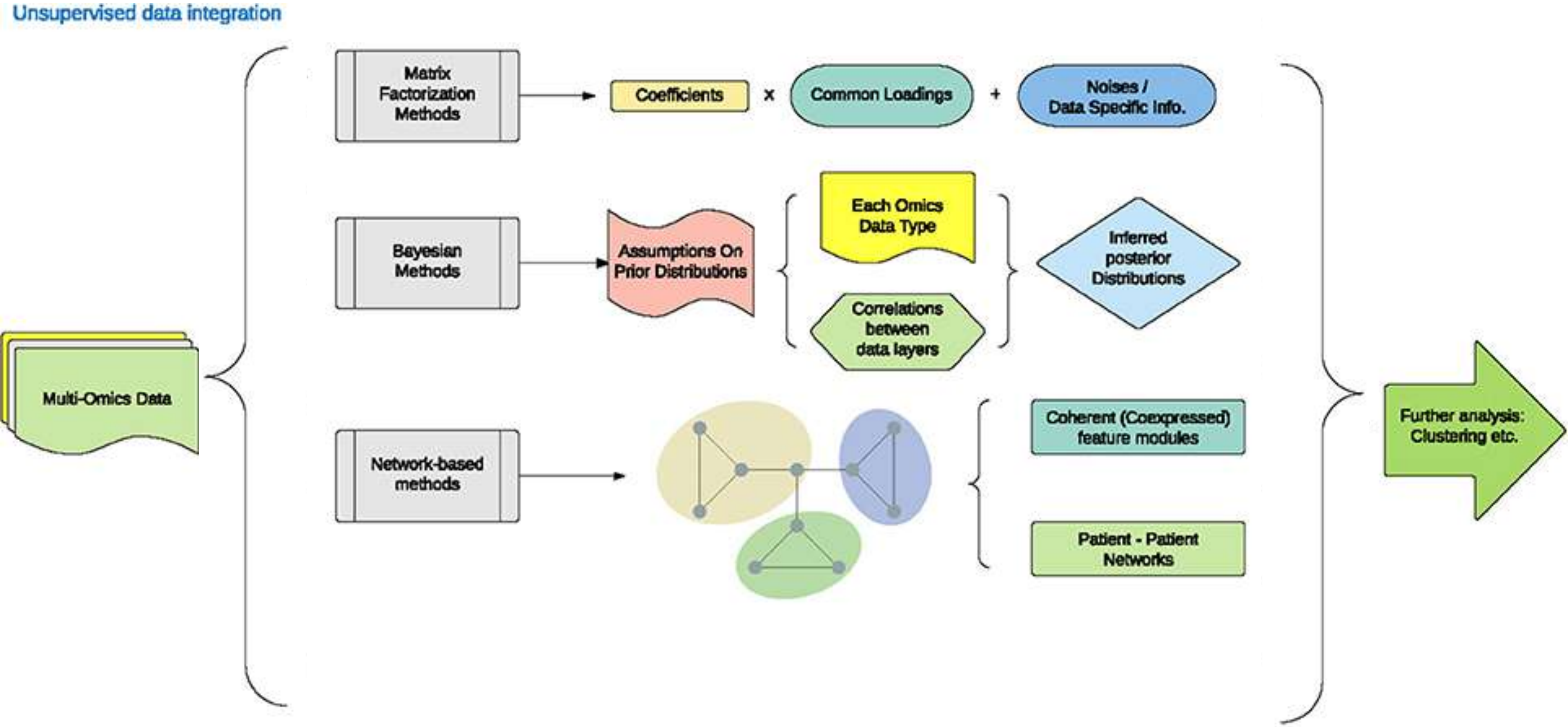
Main categories of integrative approaches

Unsupervised methods

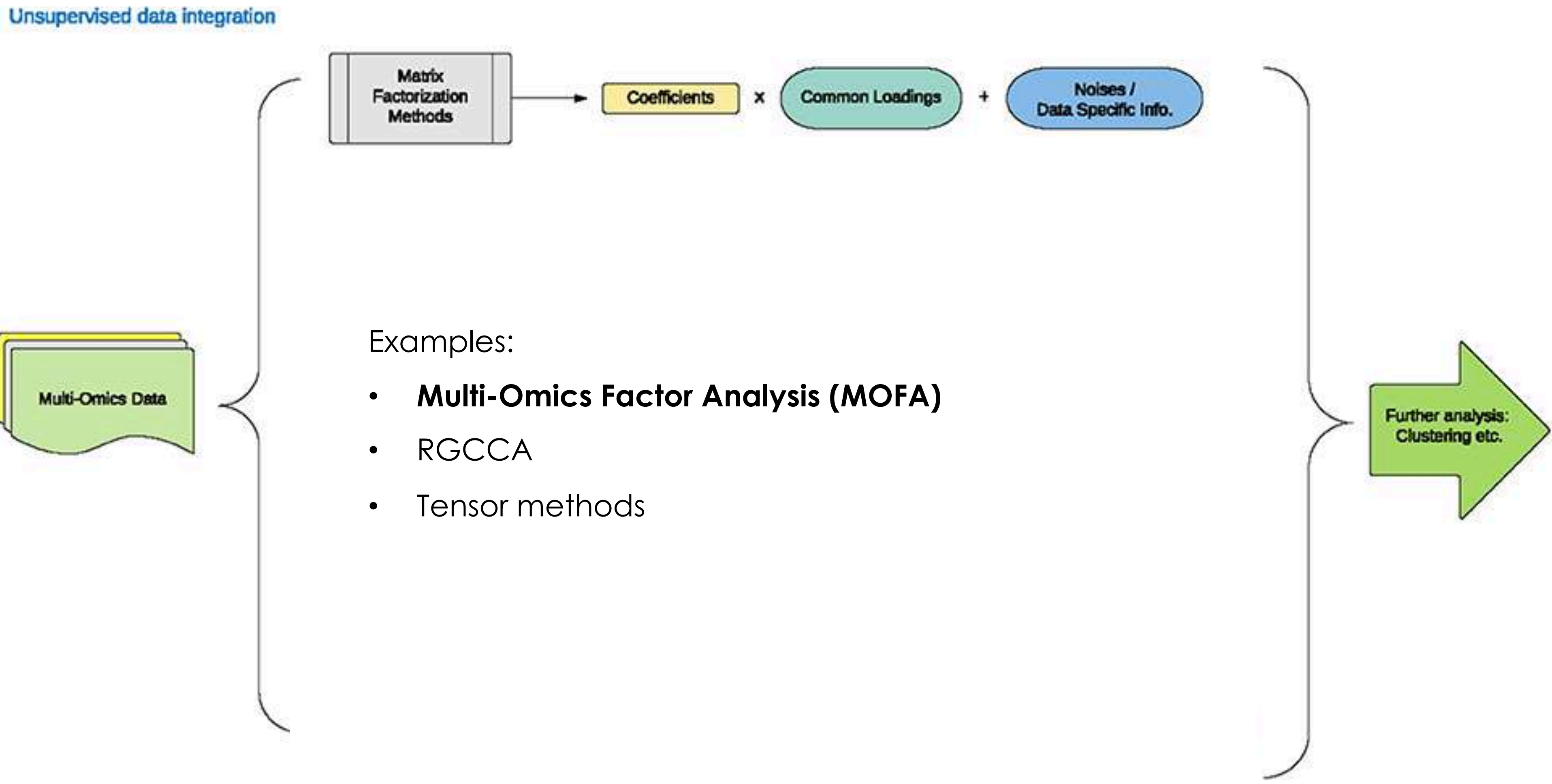
- The methodology is directly applied to one dataset
- They infer information from the structure of the data without any label information



Unsupervised integrative approaches


Unsupervised data integration
Multi-Omics Data
Matrix Factorization Methods
Coefficients
×
Common Loadings
+
Noises / Data Specific Info.
Bayesian Methods
Assumptions On Prior Distributions
Each Omics Data Type
Correlations between data layers
Inferred posterior Distributions
Network-based methods
Coherent (Coexpressed) feature modules
Patient - Patient Networks
Further analysis: Clustering etc.

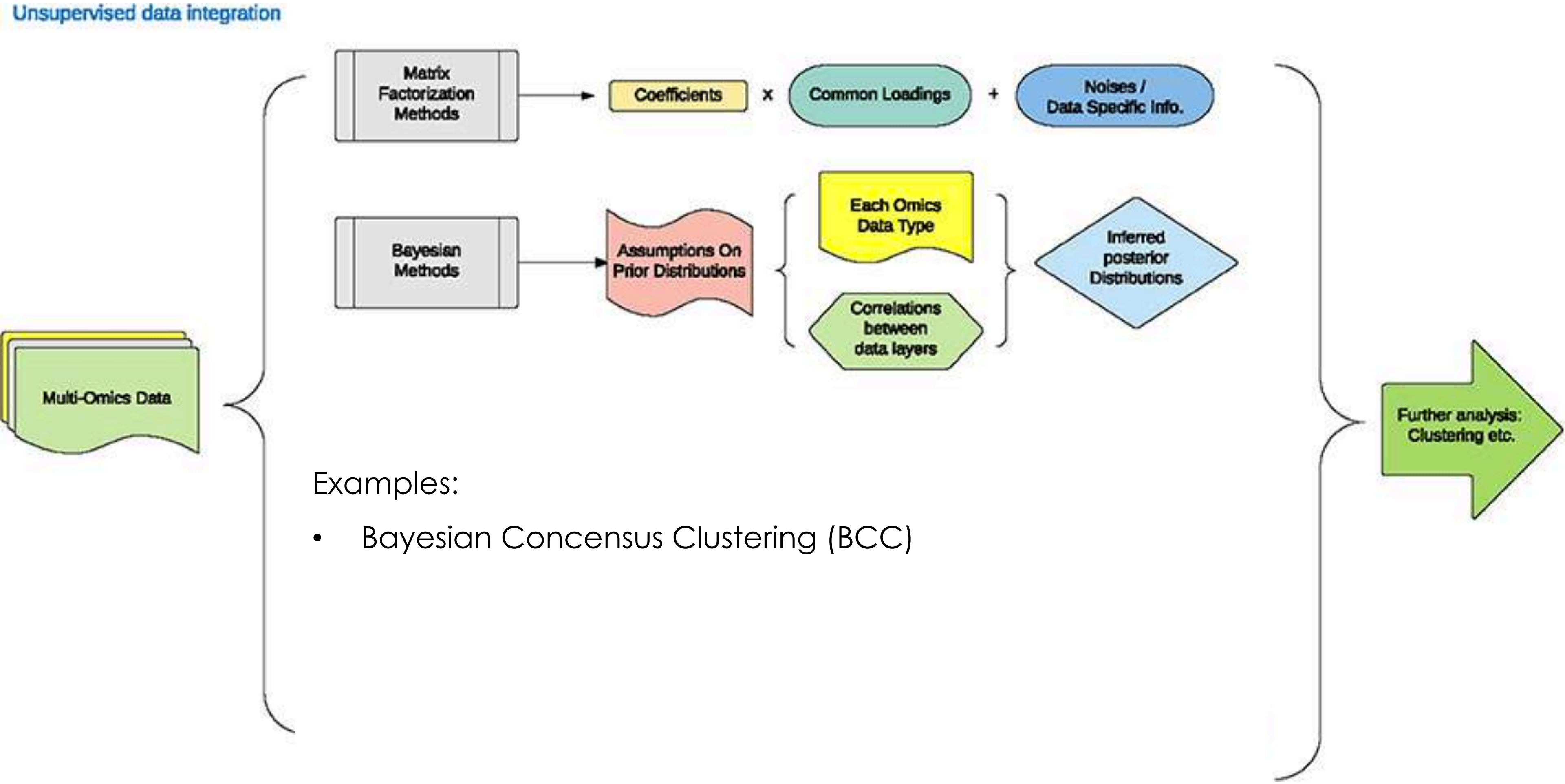
Unsupervised integrative approaches


Unsupervised data integration
Multi-Omics Data
Matrix Factorization Methods
Coefficients
×
Common Loadings
+
Noises / Data Specific Info.
Further analysis: Clustering etc.

Examples:

- **Multi-Omics Factor Analysis (MOFA)**
- RGCCA
- Tensor methods

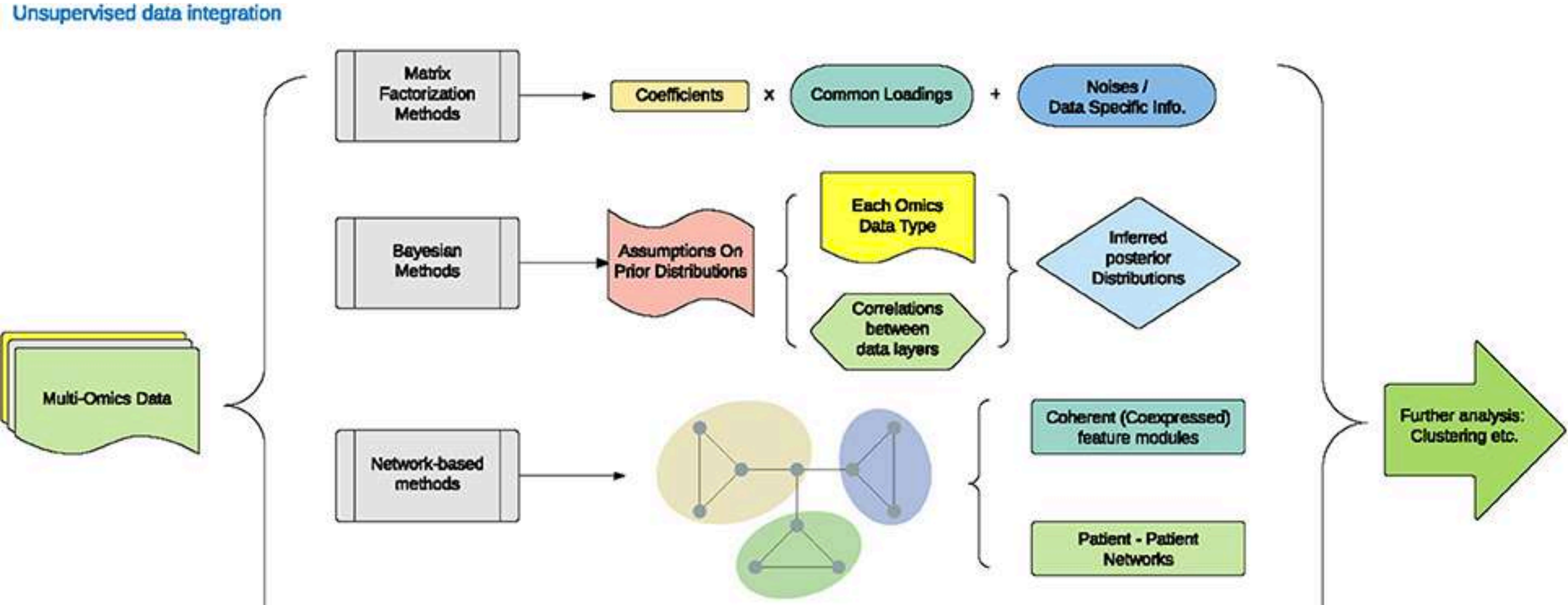
Unsupervised integrative approaches


Unsupervised data integration
Multi-Omics Data
Matrix Factorization Methods
Coefficients
×
Common Loadings
+
Noises / Data Specific Info.
Bayesian Methods
Assumptions On Prior Distributions
Each Omics Data Type
Correlations between data layers
Inferred posterior Distributions
Further analysis: Clustering etc.

Examples:

- Bayesian Concensus Clustering (BCC)

Unsupervised integrative approaches


Unsupervised data integration
Multi-Omics Data
Matrix Factorization Methods
Coefficients
×
Common Loadings
+
Noises / Data Specific Info.
Bayesian Methods
Assumptions On Prior Distributions
Each Omics Data Type
Correlations between data layers
Inferred posterior Distributions
Network-based methods
Coherent (Coexpressed) feature modules
Patient - Patient Networks
Further analysis: Clustering etc.

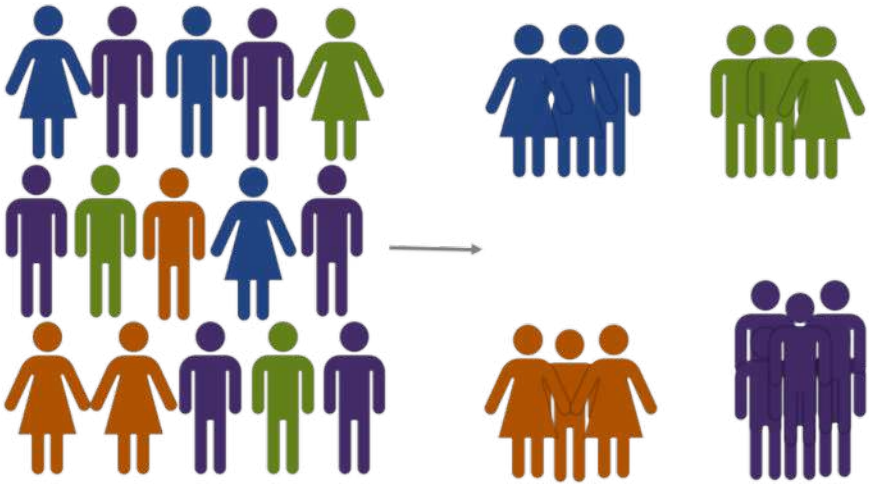
Examples:

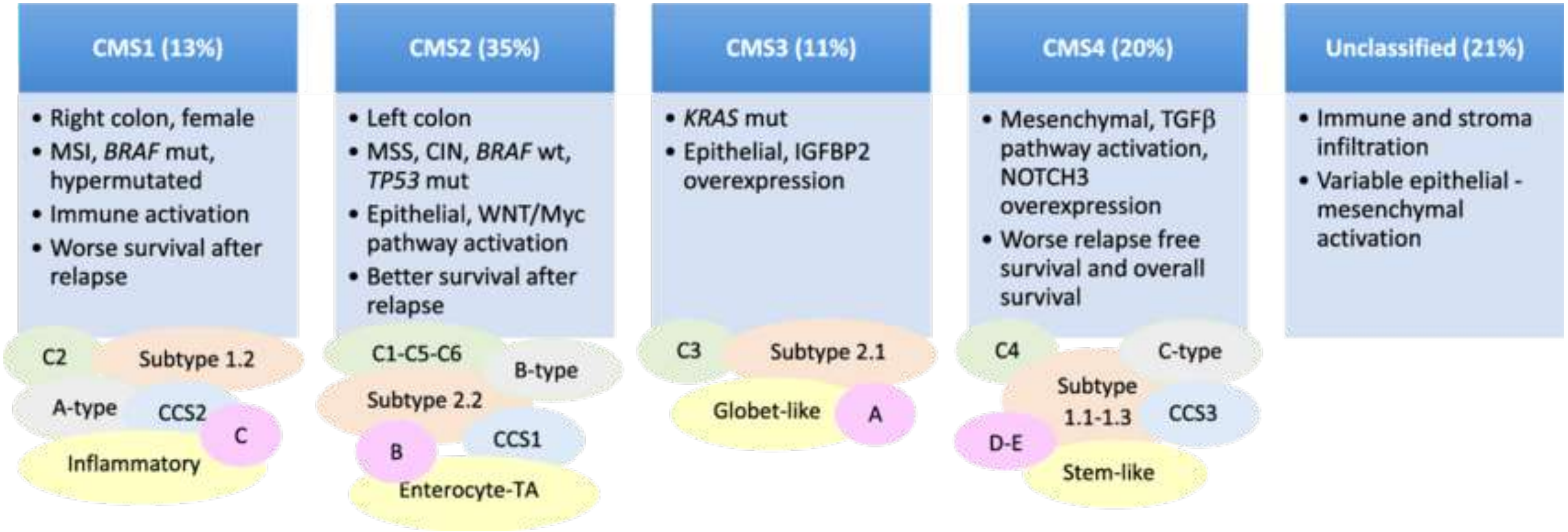
Network-based methods:

- Similarity between macromolecules: Multiplex networks
- Similarity between samples: Similarity Network Fusion SNF

Cancer insights from data integration methods

Cancer subtyping





CMS1 (13%)

- Right colon, female
- MSI, *BRAF* mut, hypermutated
- Immune activation
- Worse survival after relapse

CMS2 (35%)

- Left colon
- MSS, CIN, *BRAF* wt, *TP53* mut
- Epithelial, WNT/Myc pathway activation
- Better survival after relapse

CMS3 (11%)

- *KRAS* mut
- Epithelial, IGFBP2 overexpression

CMS4 (20%)

- Mesenchymal, TGFβ pathway activation, NOTCH3 overexpression
- Worse relapse free survival and overall survival

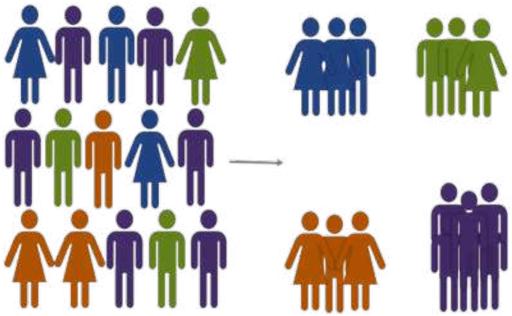
Unclassified (21%)

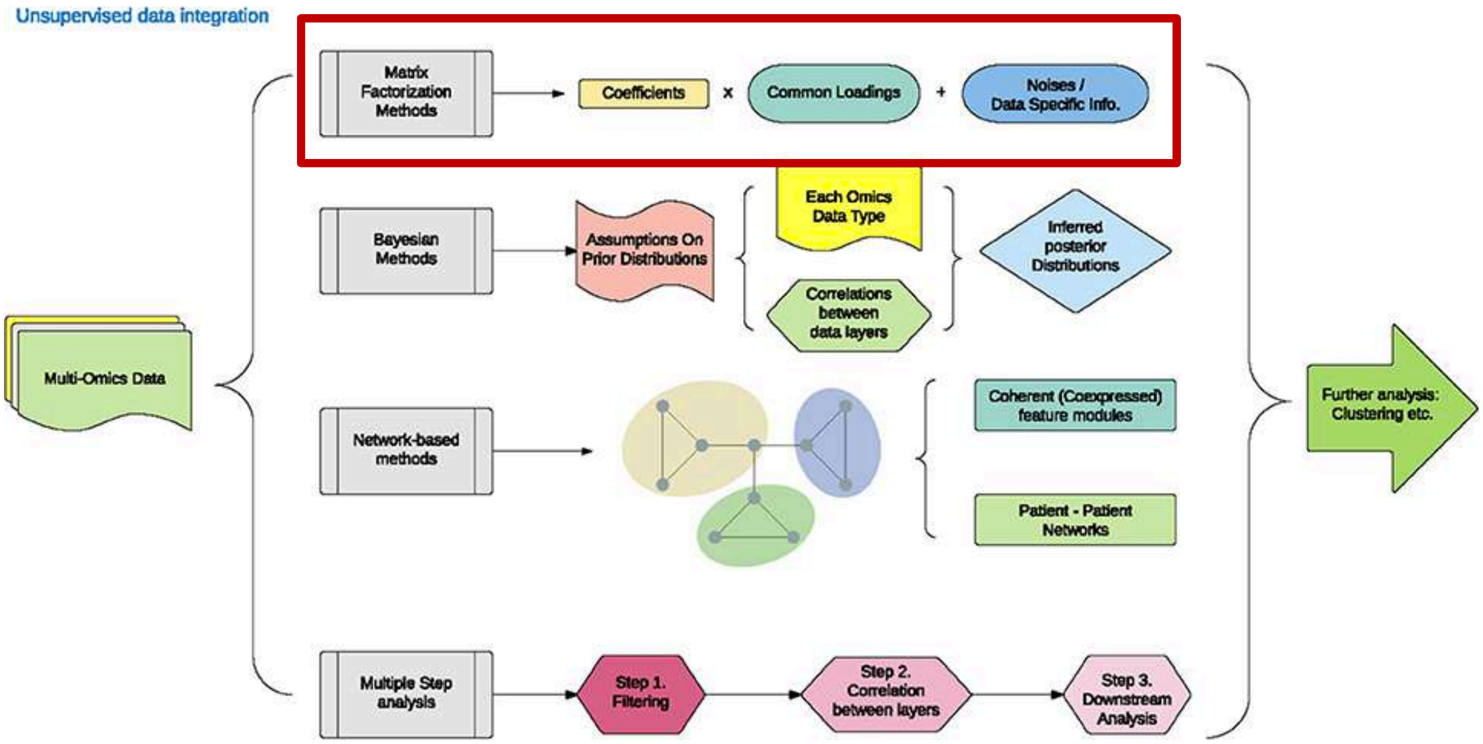
- Immune and stroma infiltration
- Variable epithelial - mesenchymal activation

Santos, Cristina, et al. "Intrinsic cancer subtypes-next steps into personalized medicine." Cellular oncology 38.1 (2015): 3-16.

Cancer subtyping

This problem is generally approached with unsupervised approaches.





Unsupervised data integration

Multi-Omics Data

Matrix Factorization Methods

Coefficients × Common Loadings + Noises / Data Specific Info.

Bayesian Methods

Assumptions On Prior Distributions

Each Omics Data Type

Correlations between data layers

Inferred posterior Distributions

Network-based methods

Coherent (Coexpressed) feature modules

Patient - Patient Networks

Multiple Step analysis

Step 1. Filtering

Step 2. Correlation between layers

Step 3. Downstream Analysis

Further analysis: Clustering etc.

Gene modules identification

Drug rensponding

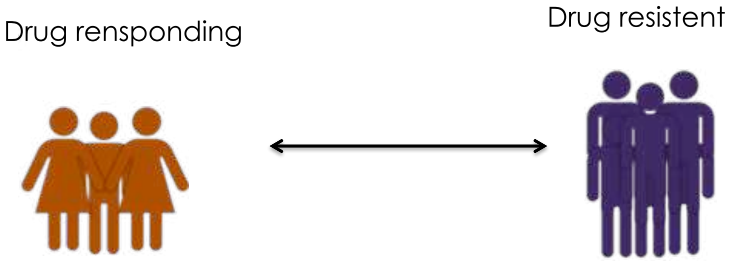
Drug resistent

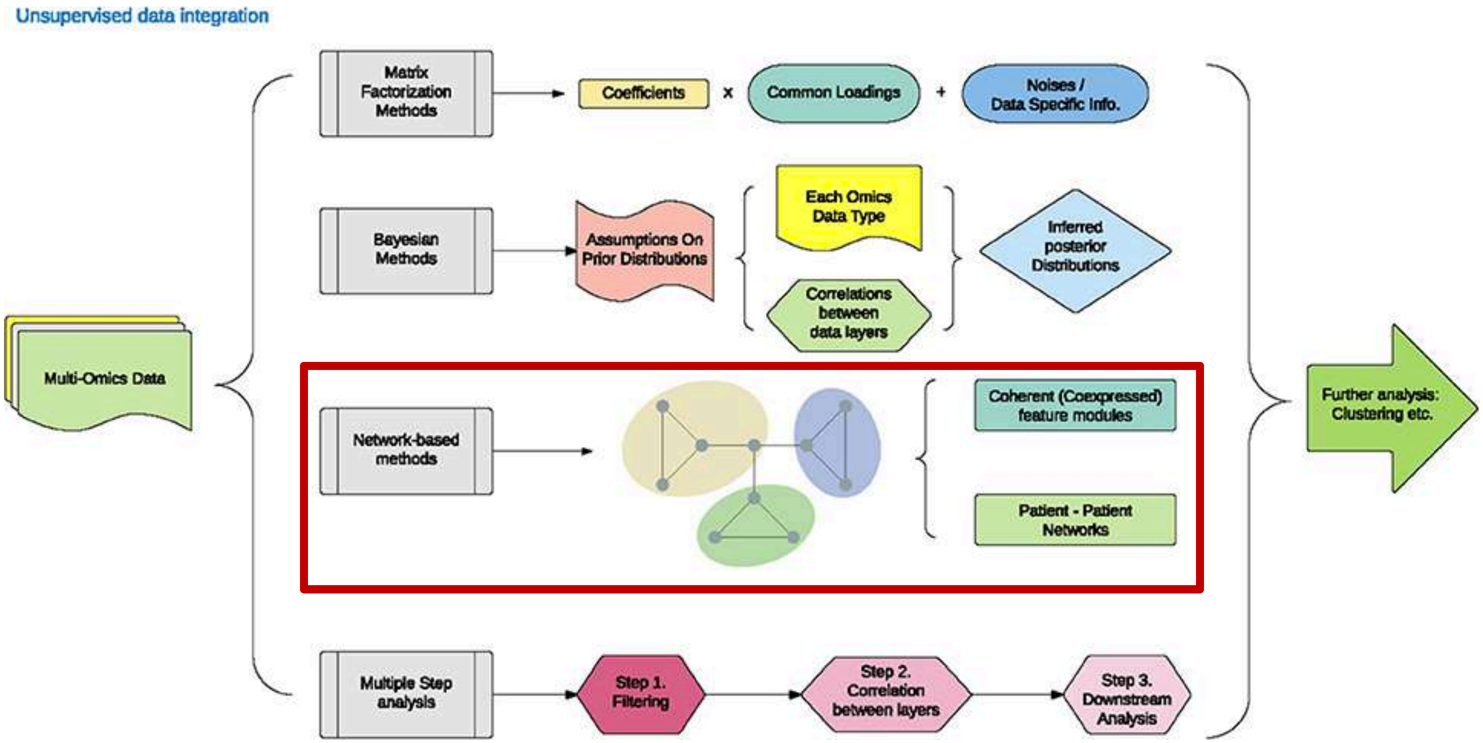
Which are the molecular mechanisms that make these two groups of patients having a different behaviour?

Can we identify a driver that can alter the behaviour of a set of patients?

Gene modules identification

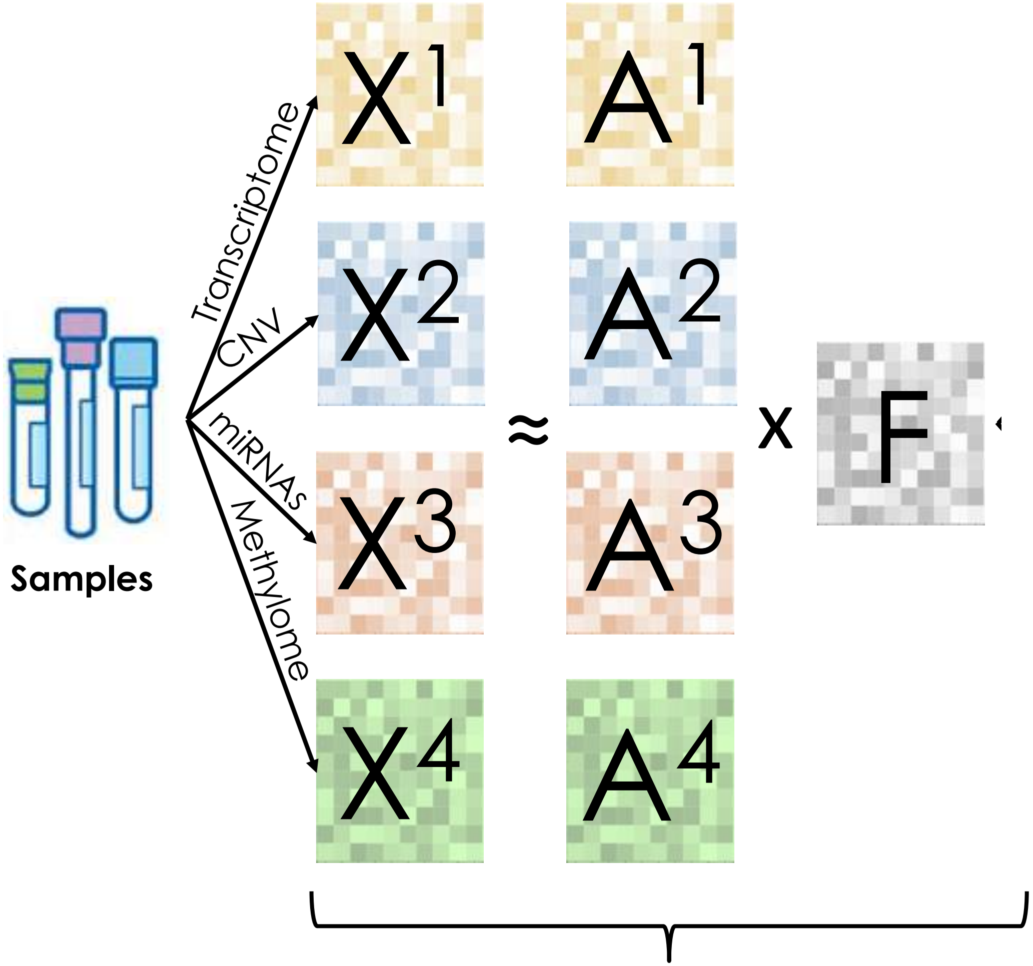
This problem is generally approached with unsupervised approaches.




Unsupervised data integration
Multi-Omics Data
Matrix Factorization Methods
Coefficients
×
Common Loadings
+
Noises / Data Specific Info.
Bayesian Methods
Assumptions On Prior Distributions
Each Omics Data Type
Correlations between data layers
Inferred posterior Distributions
Network-based methods
Coherent (Coexpressed) feature modules
Patient - Patient Networks
Multiple Step analysis
Step 1. Filtering
Step 2. Correlation between layers
Step 3. Downstream Analysis
Further analysis: Clustering etc.

Matrix Factorization

Joint Dimensionality Reduction (jDR)



$X^m = A^m F + e^m$

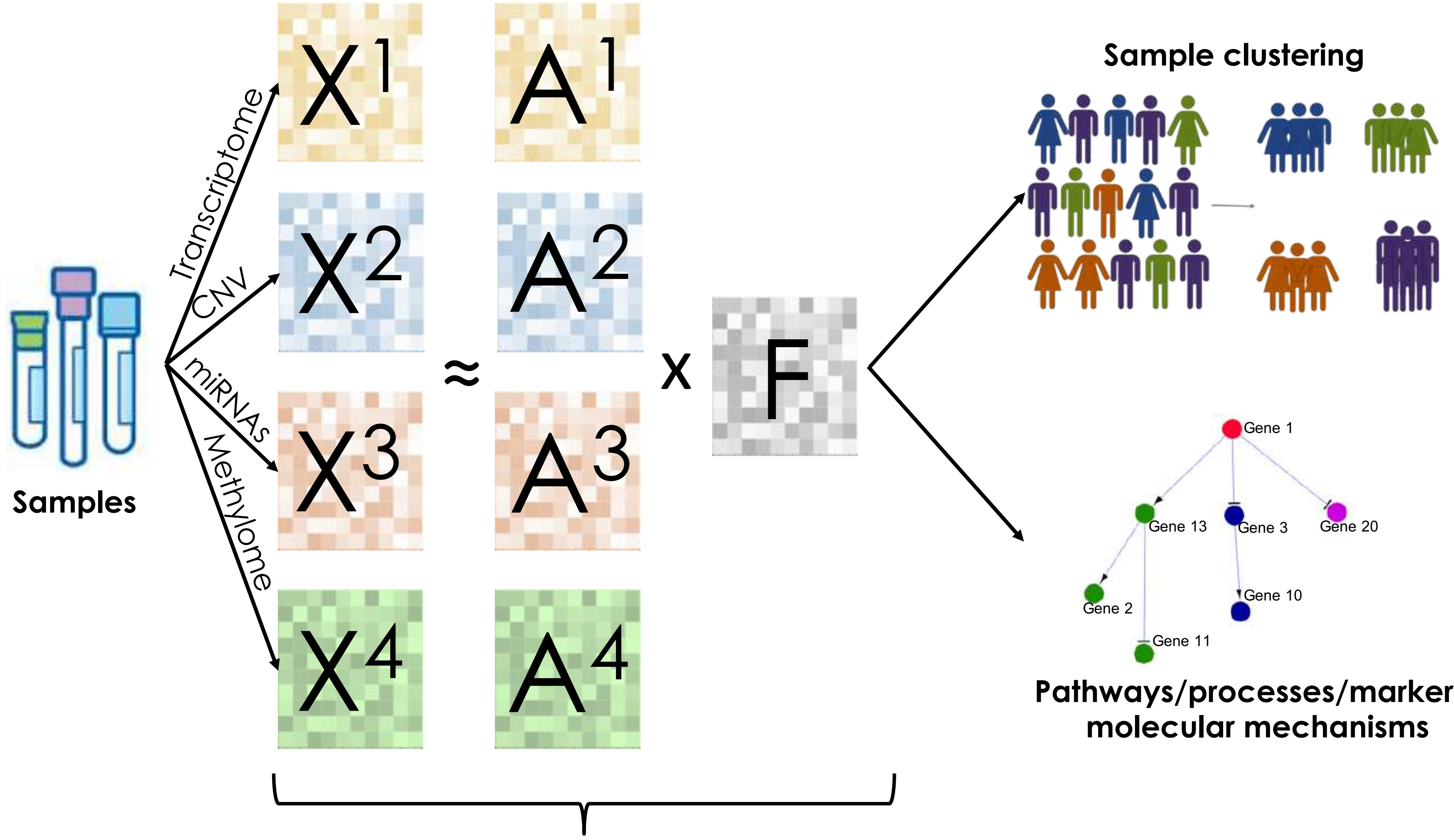
$e^m <<<<$

$m = 1, \ldots, P$

Meng, Chen, et al. "Dimension reduction techniques for the integrative analysis of multi-omics data." *Briefings in bioinformatics* 17.4 (2016): 628-641.

Joint Dimensionality Reduction (jDR)

Multi-omics joint Dimensionality Reduction (jDR)



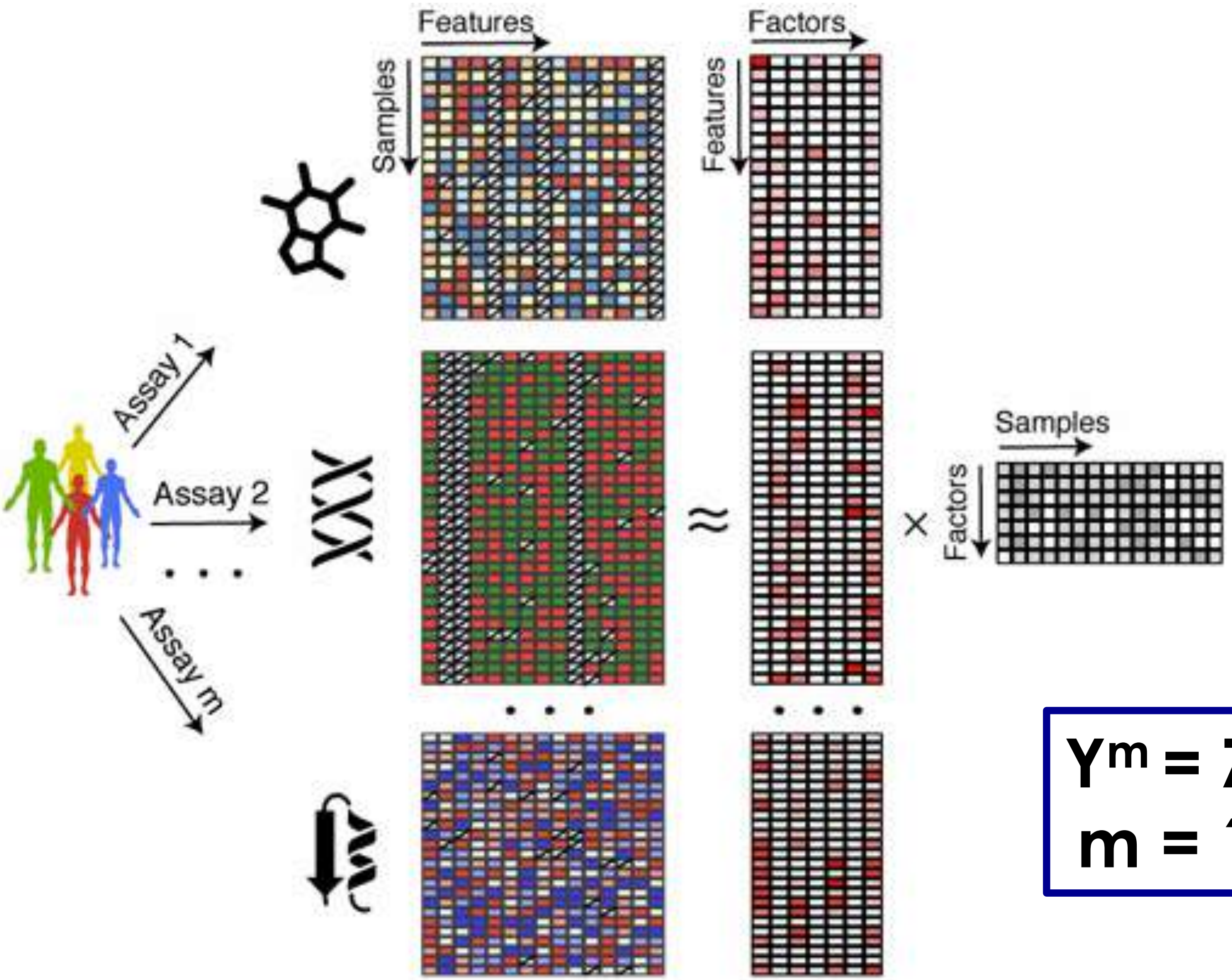
$X^m = A^m F + e^m$

$e^m <<<<$

$m = 1, \ldots, P$

Meng, Chen, et al. "Dimension reduction techniques for the integrative analysis of multi-omics data." *Briefings in bioinformatics* 17.4 (2016): 628-641.

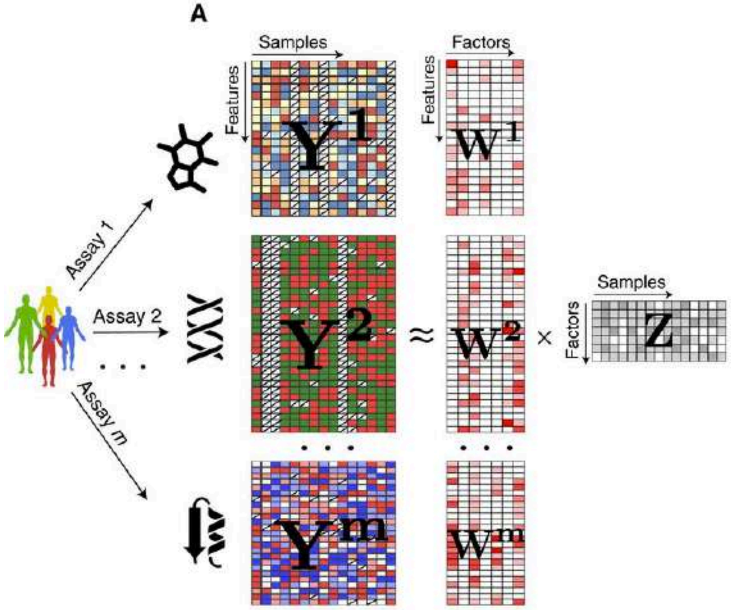
Multi-omics Factor Analysis (MOFA)

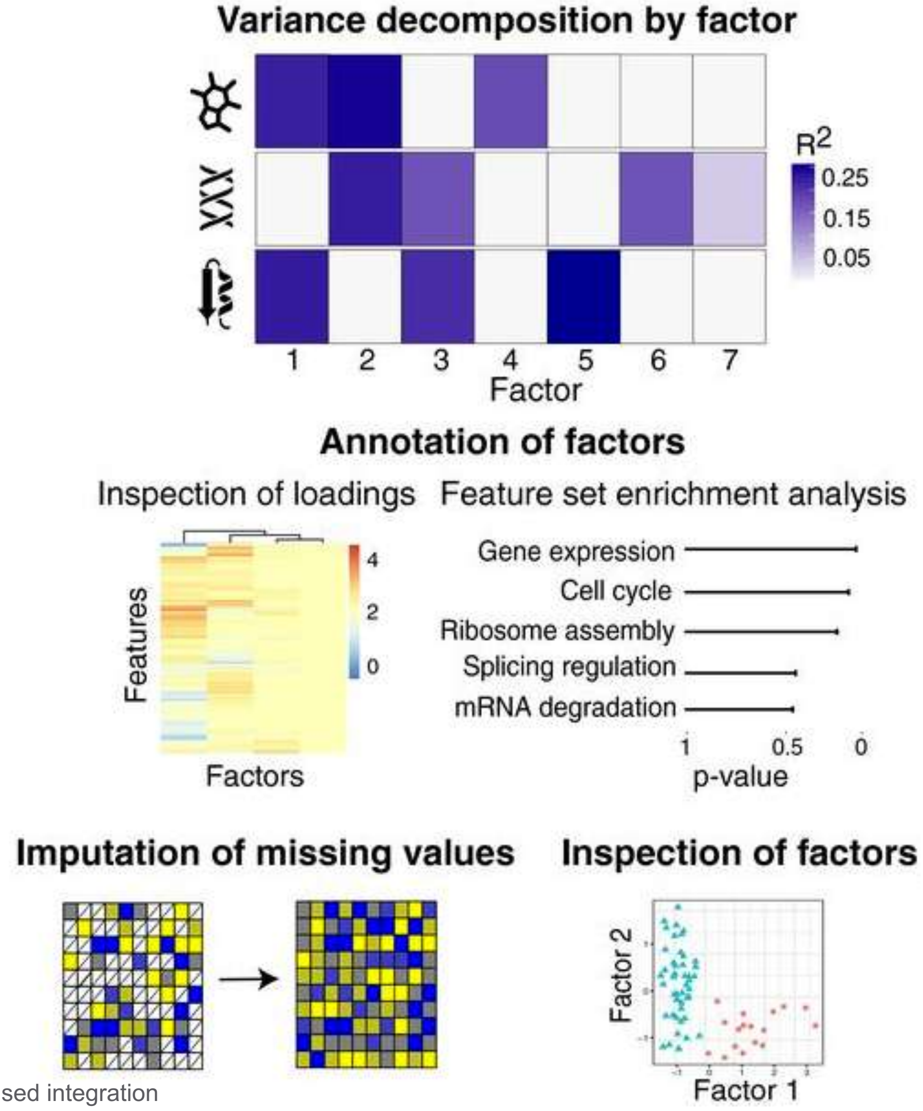


$$\boxed{\begin{array}{l} Y^m = ZW^m + e^m \\ m = 1, \ldots, M \end{array}}$$

Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

MOFA advantage: interpretability of factors

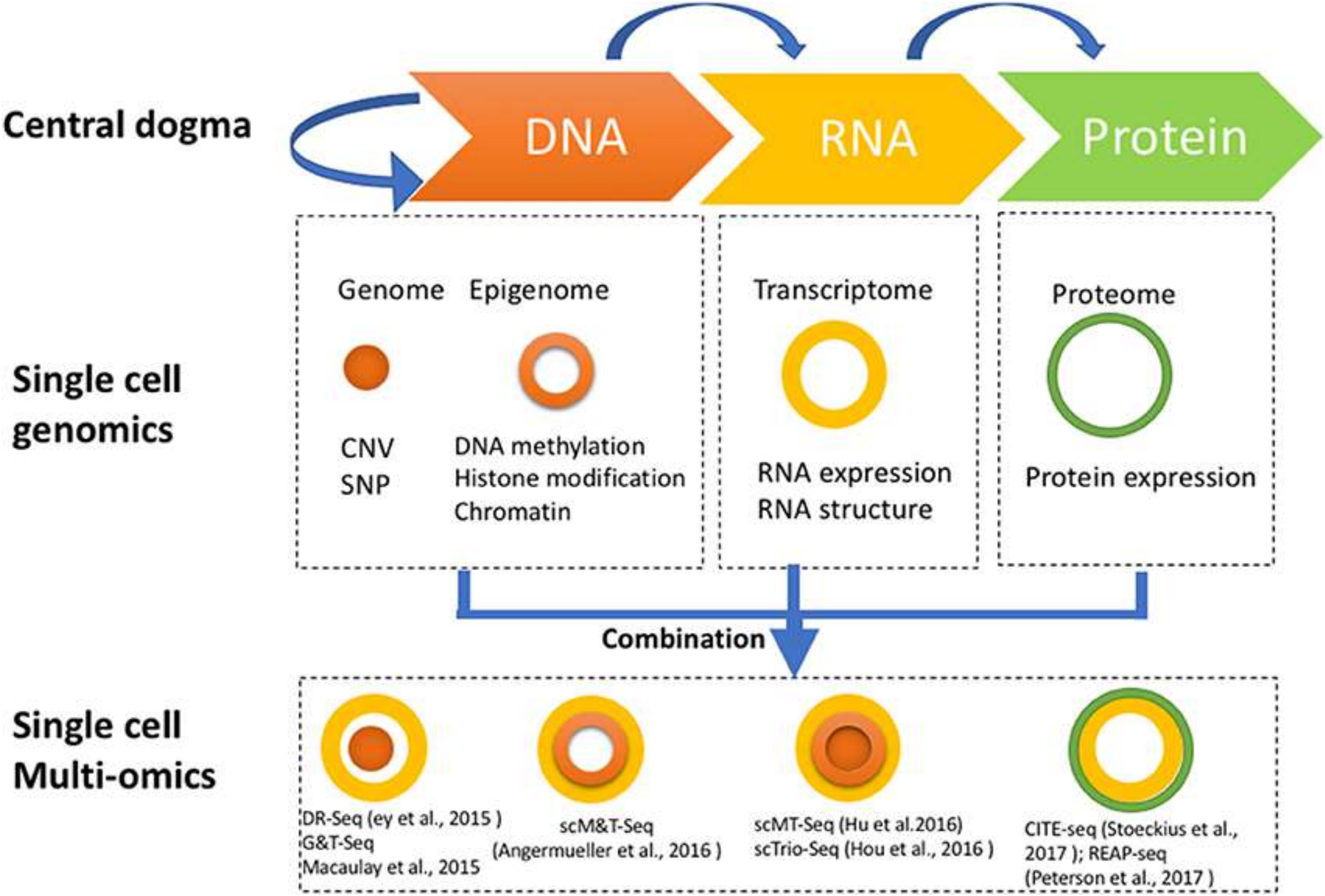




Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

Also single-cell multi-omics data can be integrated with matrix factorization

Multi-omics single-cell



Hu, Youjin, et al. " *Frontiers in cell and developmental biology* 6 (2018): 28.

Example MOFA application single-cell multi-omics

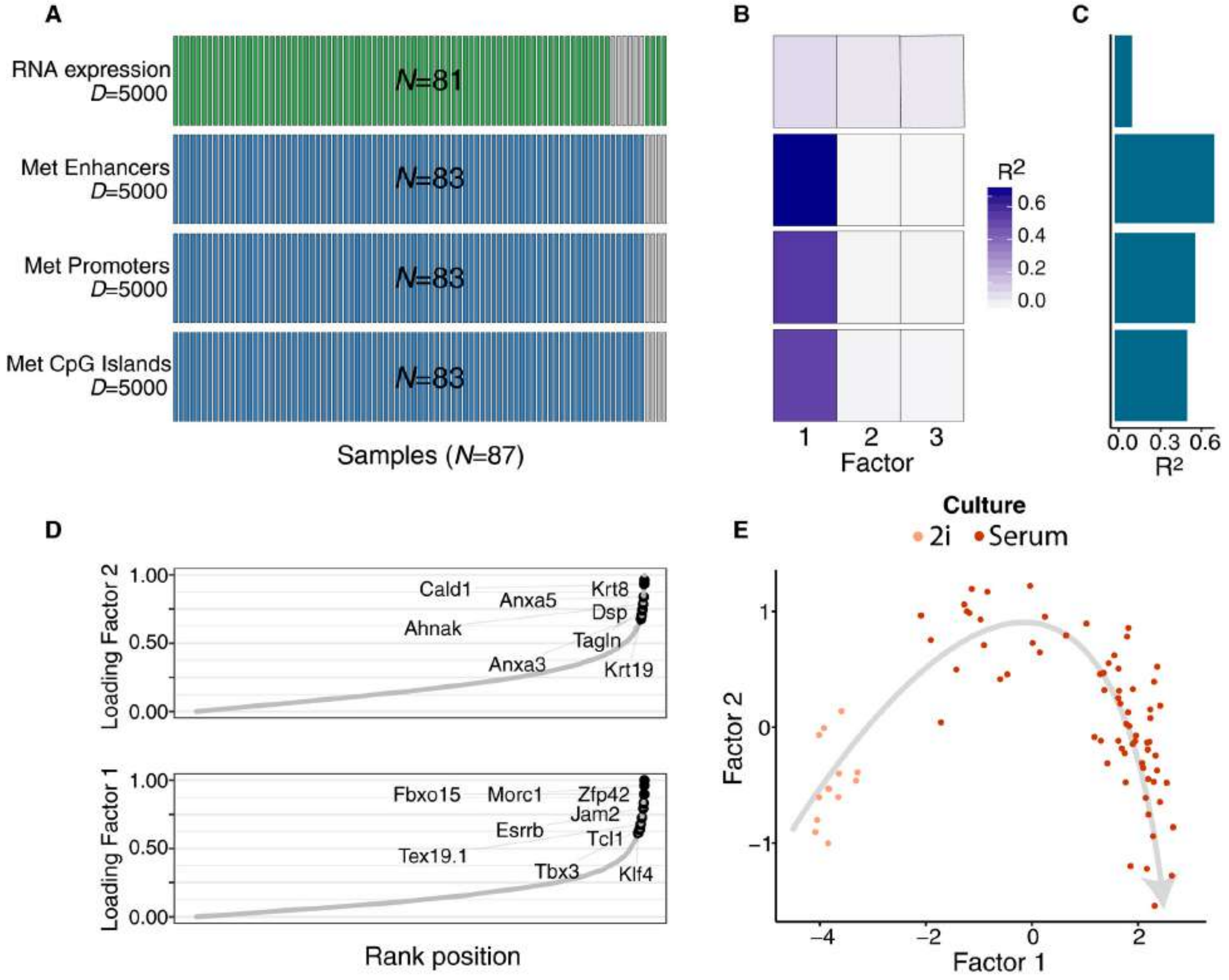
Dataset: 87 mouse embryonic stem cells (mESCs) comprising:

- 16 cells cultured in “2i” media, which induces a naive pluripotency state
- 71 serum-grown cells, which commits cells to a primed pluripotency state poised for cellular differentiation.

All cells were profiled using single-cell methylation and transcriptome sequencing

Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

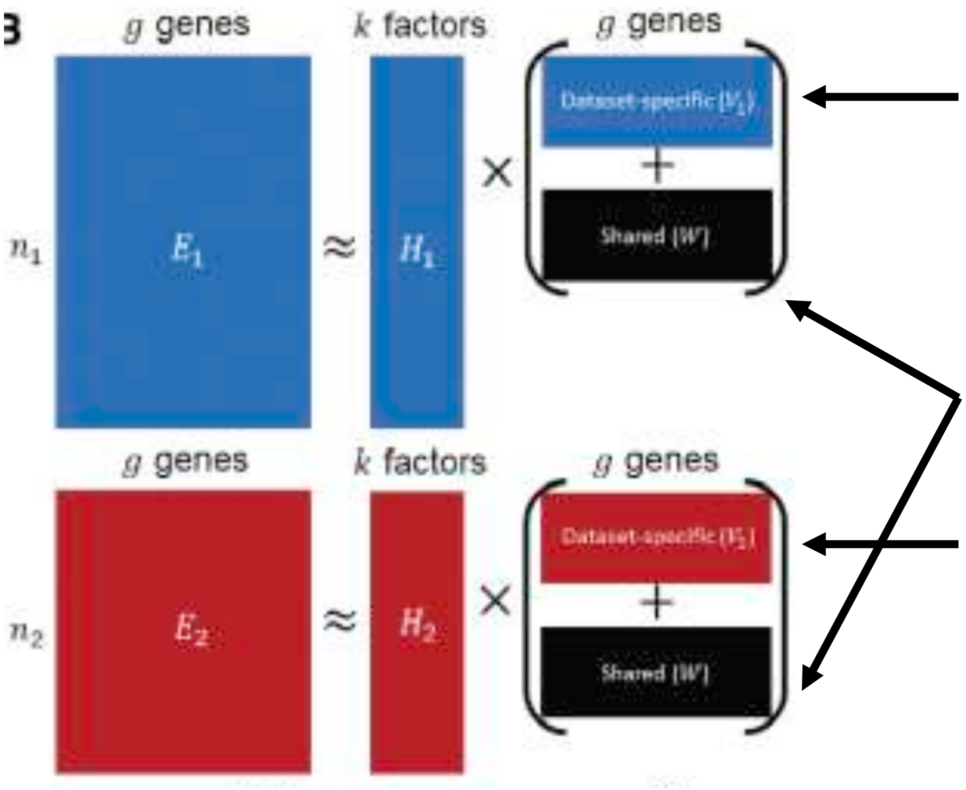
Example MOFA application single-cell multi-omics



Argelaguet, Ricard, et al. "Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets." Molecular systems biology 14.6 (2018): e8124.

Linked inference of genomic experimental relationships (LIGER)

Integrative Non Negative Matrix Factorization (iNMF)



Omic-specific factor

Shared Factor

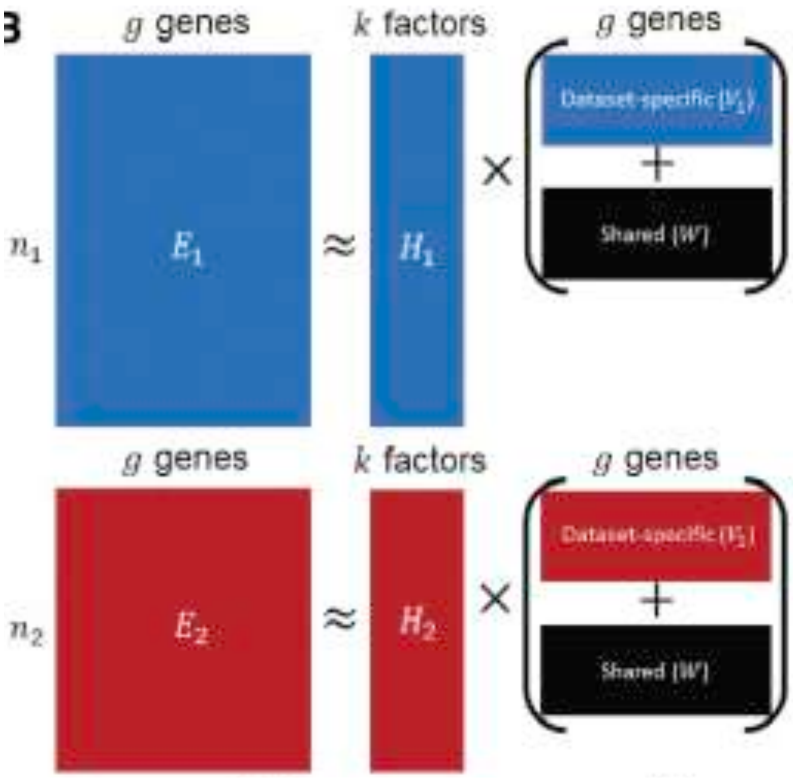
Omic-specific factor

$$E_i = H_i \, V_i + H_i \, W$$

Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.

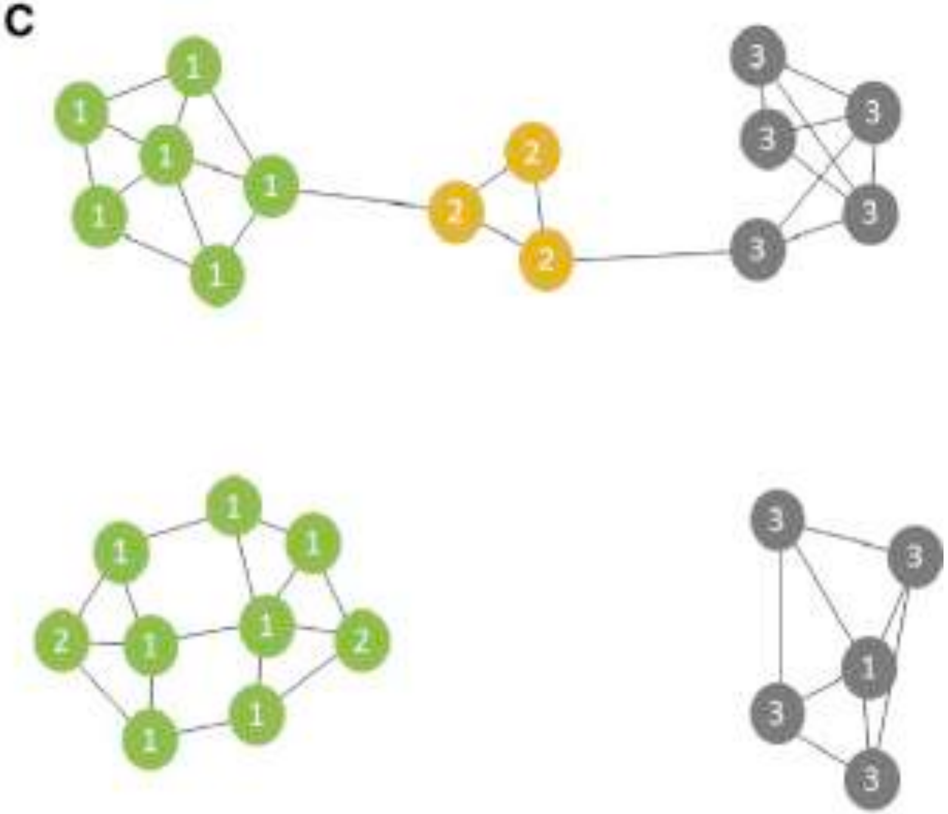
LIGER: multi-omics clustering

Integrative Non Negative Matrix Factorization (iNMF)



$$E_i = H_i\, V_i + H_i\, W$$

kNN graphs to derive clusters from factors



Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.

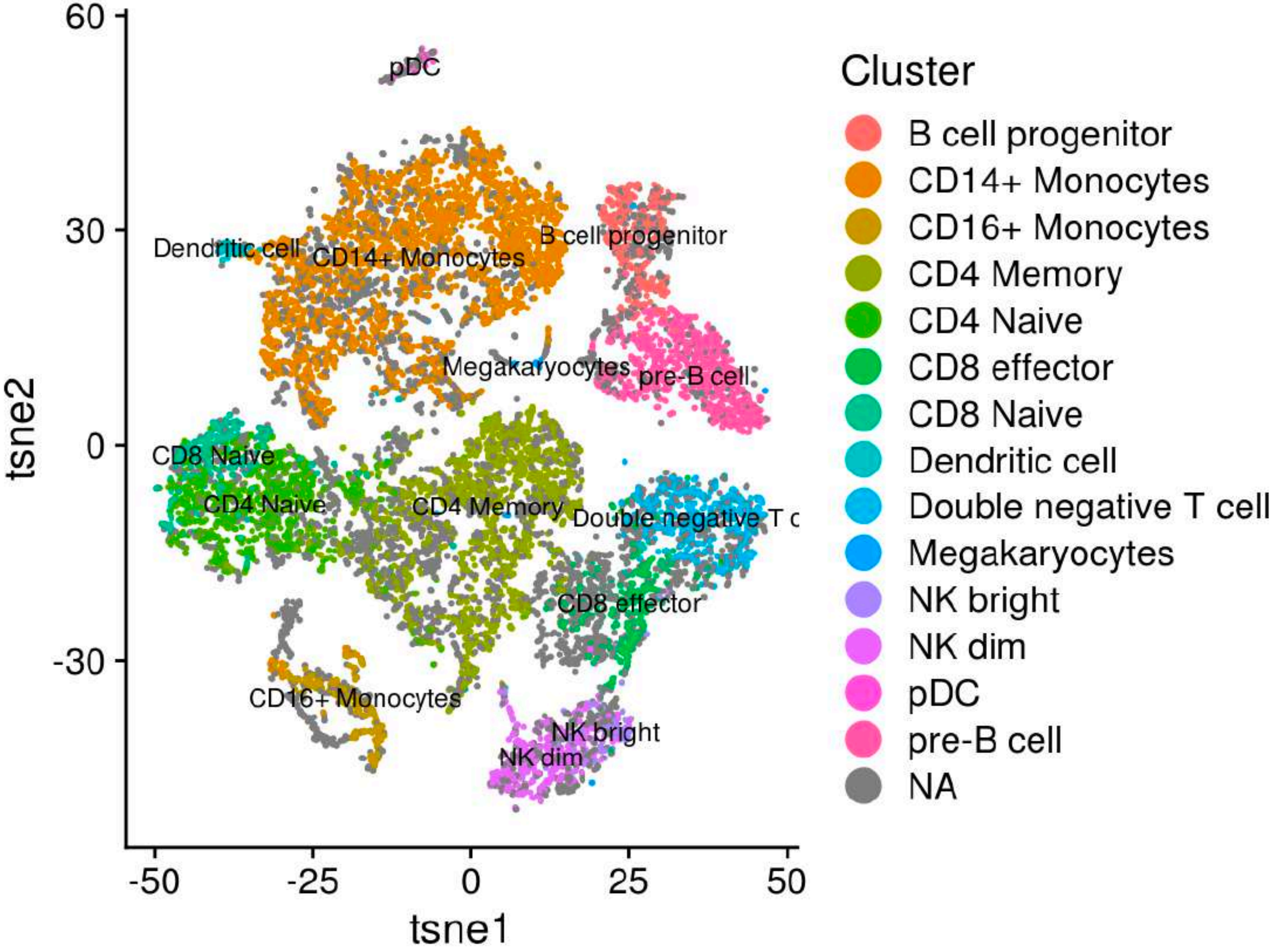
LIGER: peripheral blood mononuclear cell (PBMC)

scRNAseq and scATACseq data from approx. 10k cells PBMCs

We want to identify subtypes of cells based on the joint analysis of the two data types

Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.

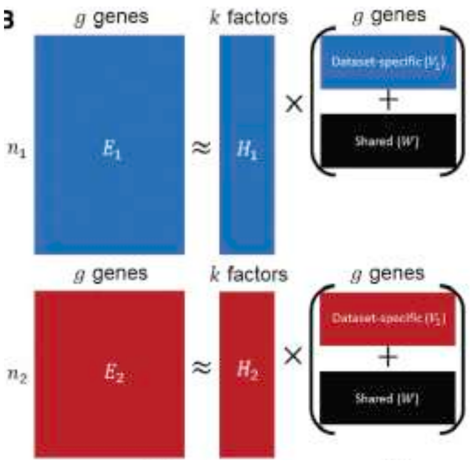
LIGER: peripheral blood mononuclear cell (PBMC)



Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.

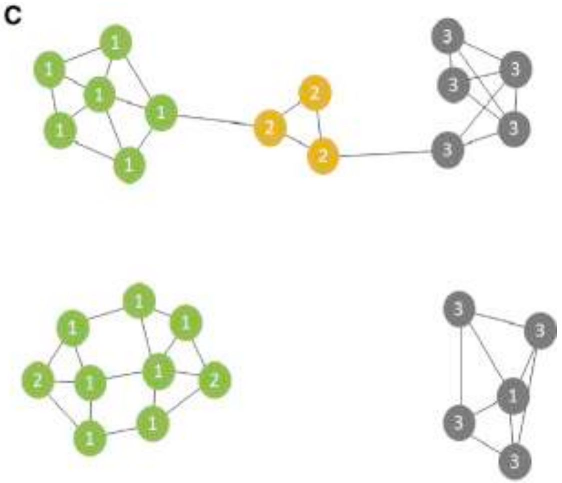
Pay attention this is not a TSNE plot of scRNAseq data

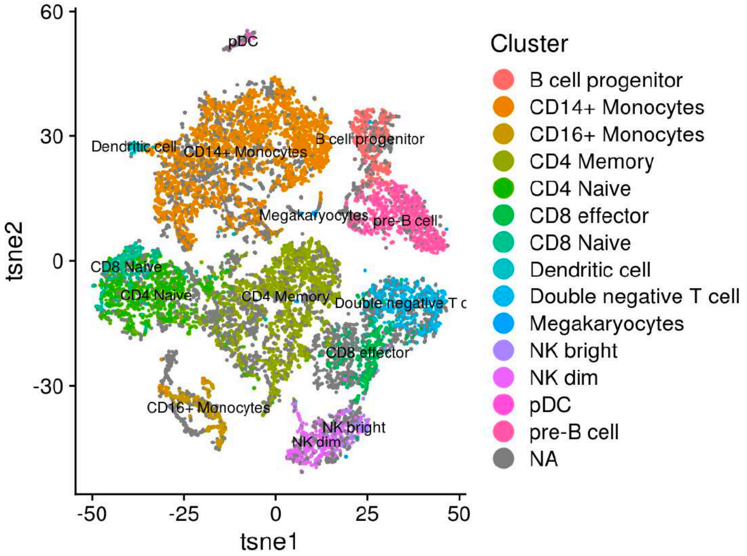
Integrative Non Negative Matrix Factorization (iNMF)



$E_i = H_i V_i + H_i W$

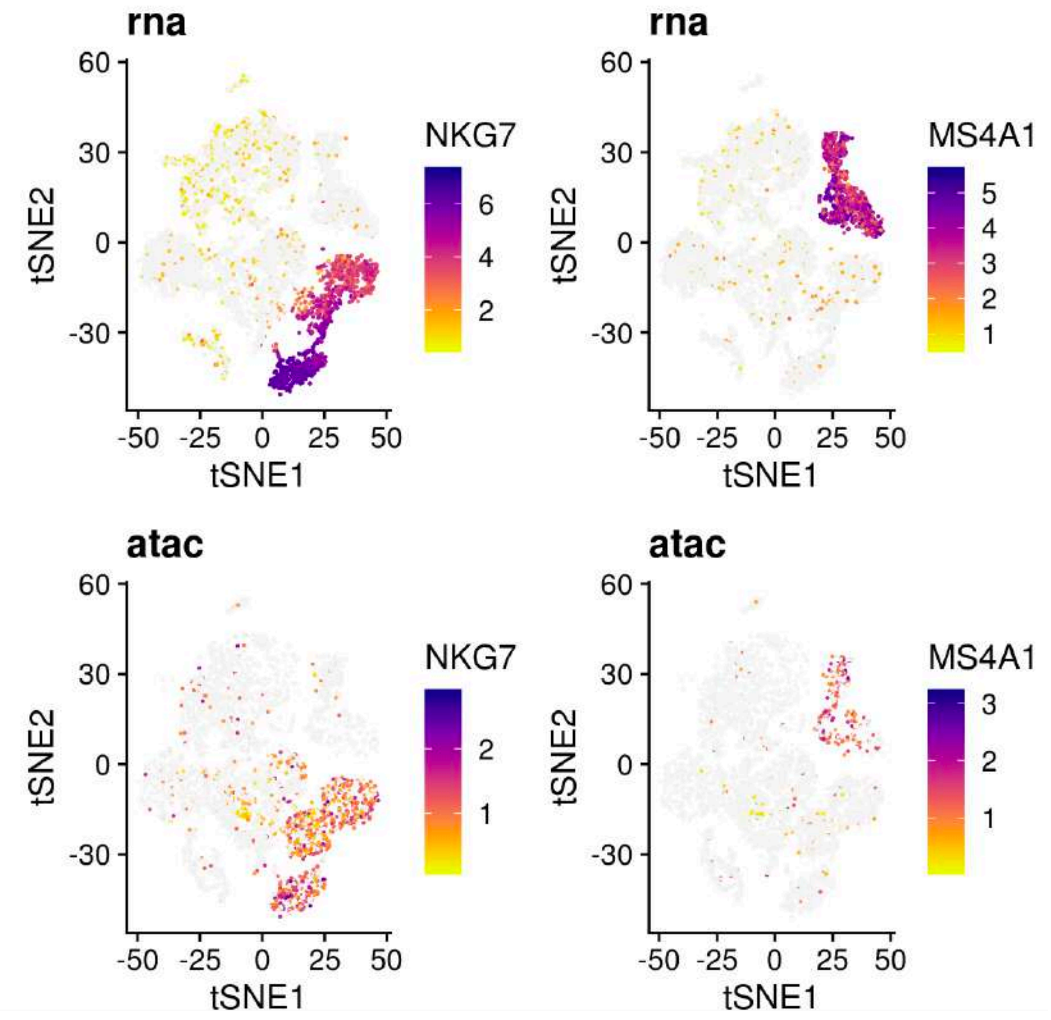
kNN graphs to derive clusters from factors

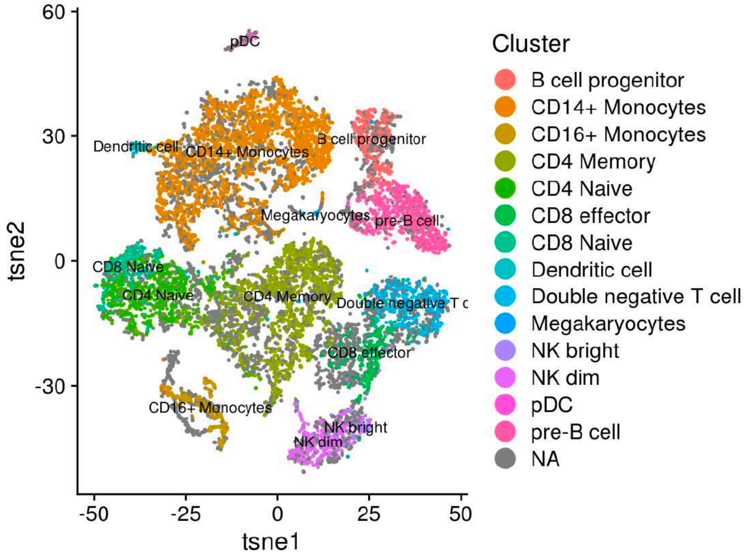




Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.

LIGER: peripheral blood mononuclear cell (PBMC)





Welch, Joshua D., et alCell 177.7 (2019): 1873-1887.