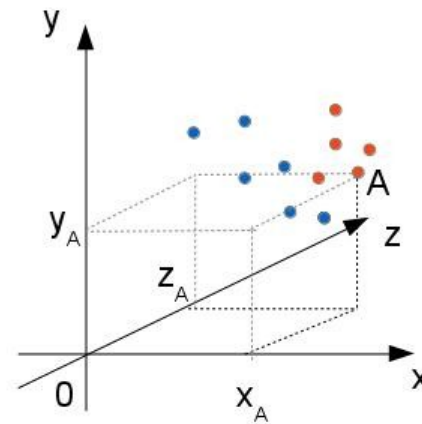
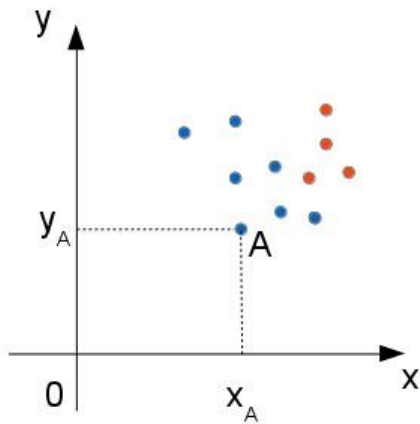


Multivariate analysis for omics data

Introduction

- individuals
- quantitative variables (values on axes)
- qualitative variables (colors)



Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

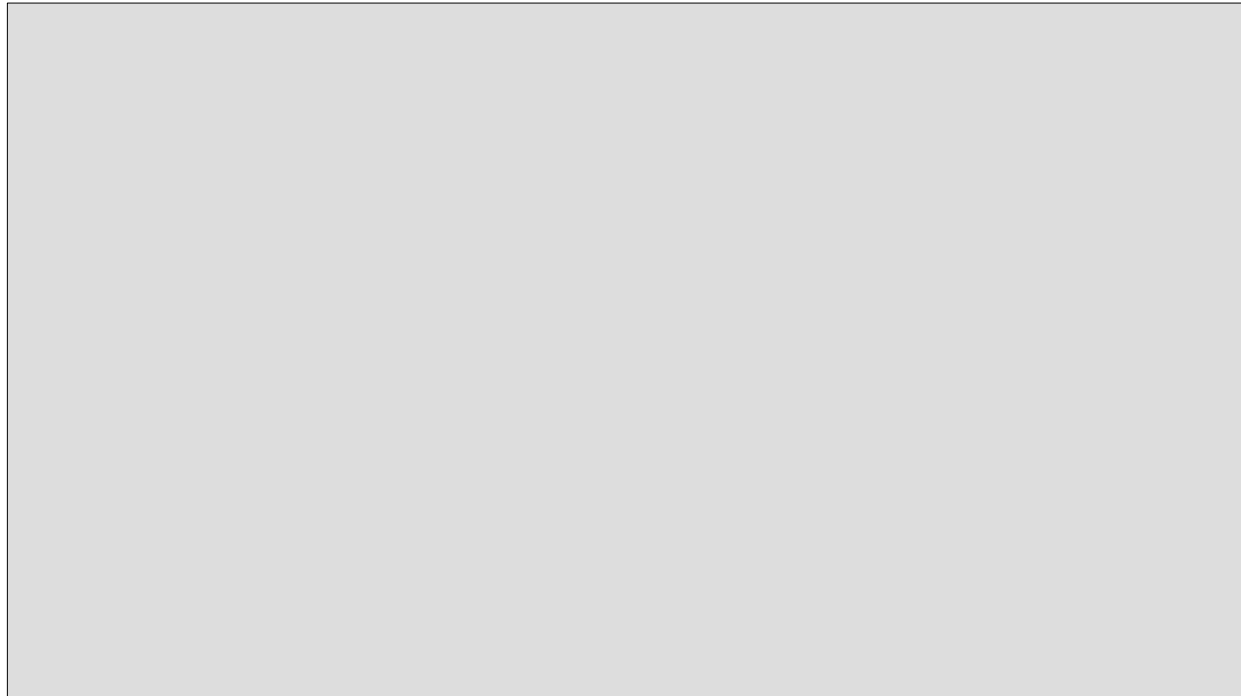
Omics datasets

Variables (p) on (n) individuals

$n \lll p$

p variables

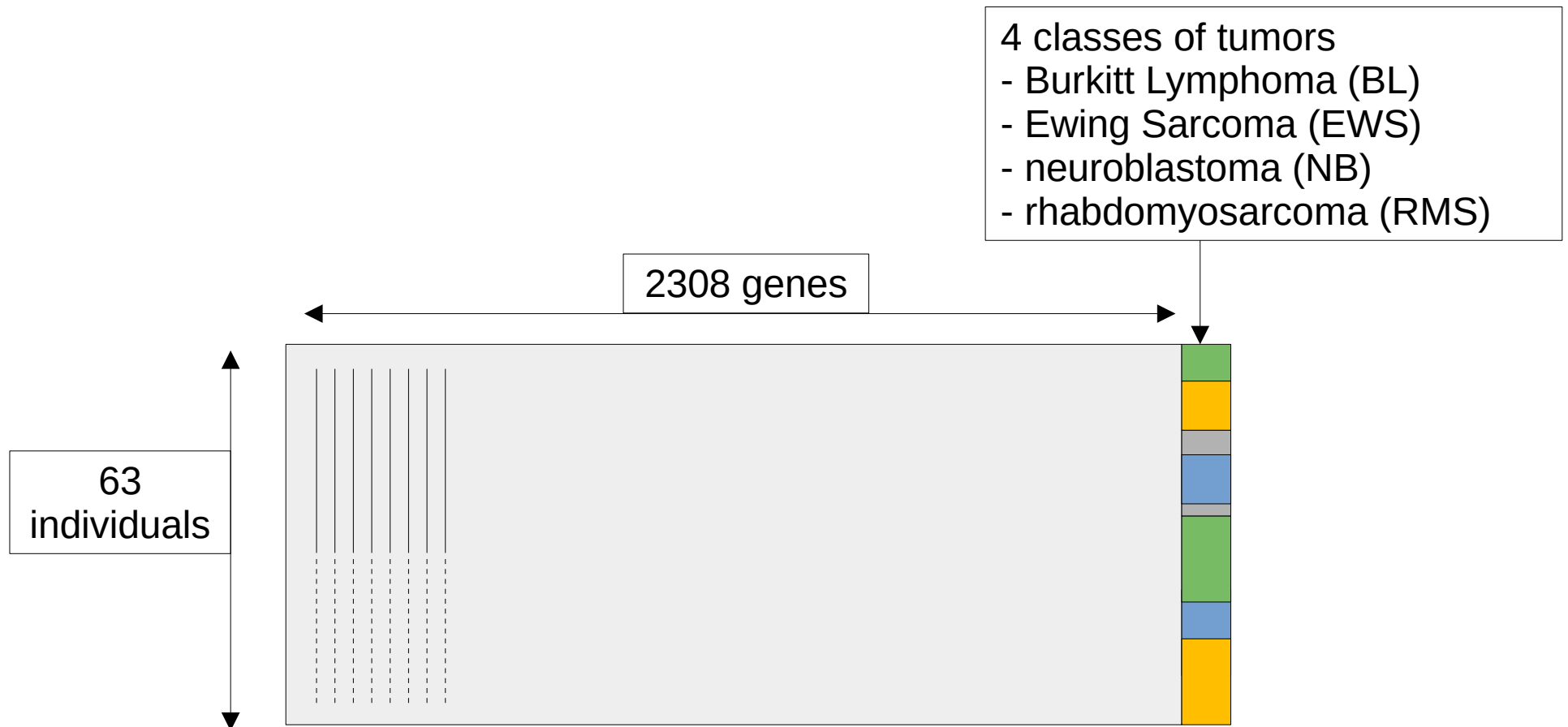
n individuals



Dataset “Small Round Blue Cell Tumours”

Expression of 2308 genes on 63 individuals

from Khan et al. 2001



Package R “mixOmics”



Exploration and
Integration of
Omics datasets

Kim-An Lê Cao, Florian Rohart et Sébastien Déjean

Multivariate analysis methods

$n \ll p$

Heterogeneous omics data

- Methods
- Graphical output
- Example datasets

sparsePCA

Shen, H. et al. (2008). *Sparse principal component analysis via regularized low rank matrix approximation*
Journal of Multivariate Analysis

sPCA

- Sparse PCA component are constructed with a limited number of variables

To select variables bearing the more information of the dataset

→ p is reduced

Several algorithms

(Zou et al. (2018). "A Selective Overview of Sparse Principal Component Analysis". Proceedings of the IEEE)

sPCA – how many variables ?

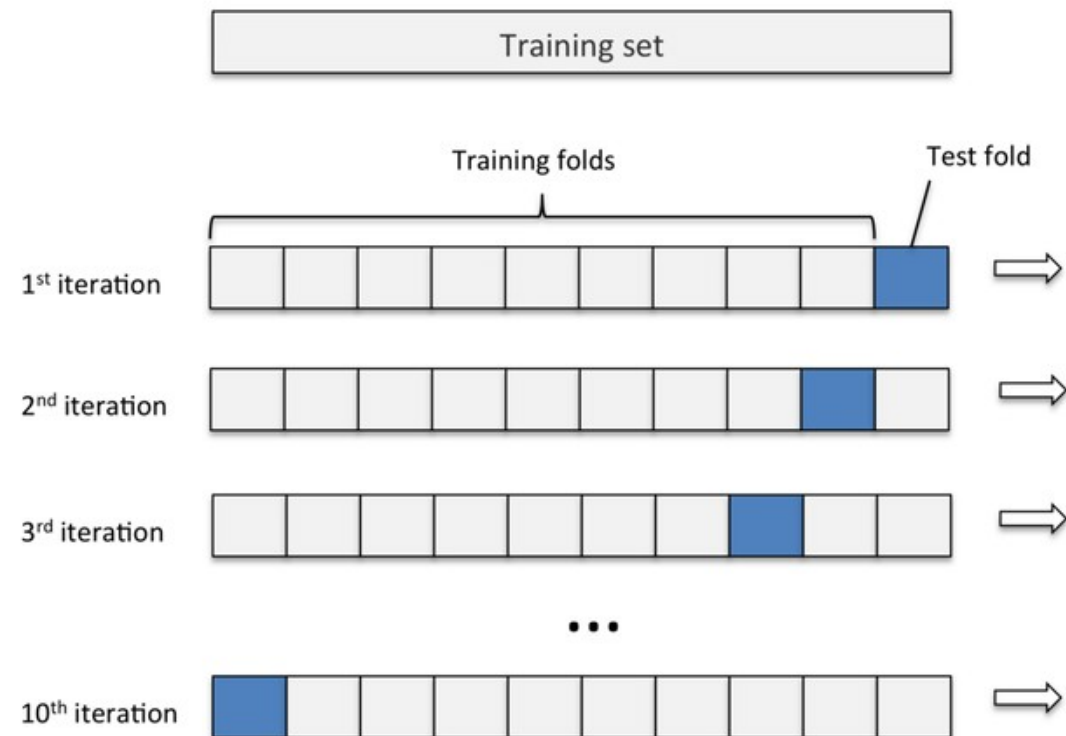
Cross-validation allows to identify the optimal number of variables (change between components)

- Dataset is splitted in x sub groups (*folders*) :

- a subset “test” (to predict)
- a training set

- Correlation between prédiction of “test” individuals and observed values

- cross validation repeated (*repeats*)



Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

PLS-DA

Partial Least Squares – Discriminant Analysis

- Omics quantitative values + 1 qualitative variable (groups)

We want to identify variable allowing to discriminate the groups

Quantitatives variables → that explain

Qualitative variables → that we want to explain

PLS-DA

- Intra-group dispersion is characterized by a variance-covariance matrix named W
- Inter-group dispersion is characterized by a variance-covariance matrix named B

Total dispersion is characterized by variance-covariance matrix V : $V=W+B$.

B (*between*) is maximised W (*within*) is minimised

=> can allow to predict some groups of new individuals !!

Dataset “Small Round Blue Cell Tumours”

Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

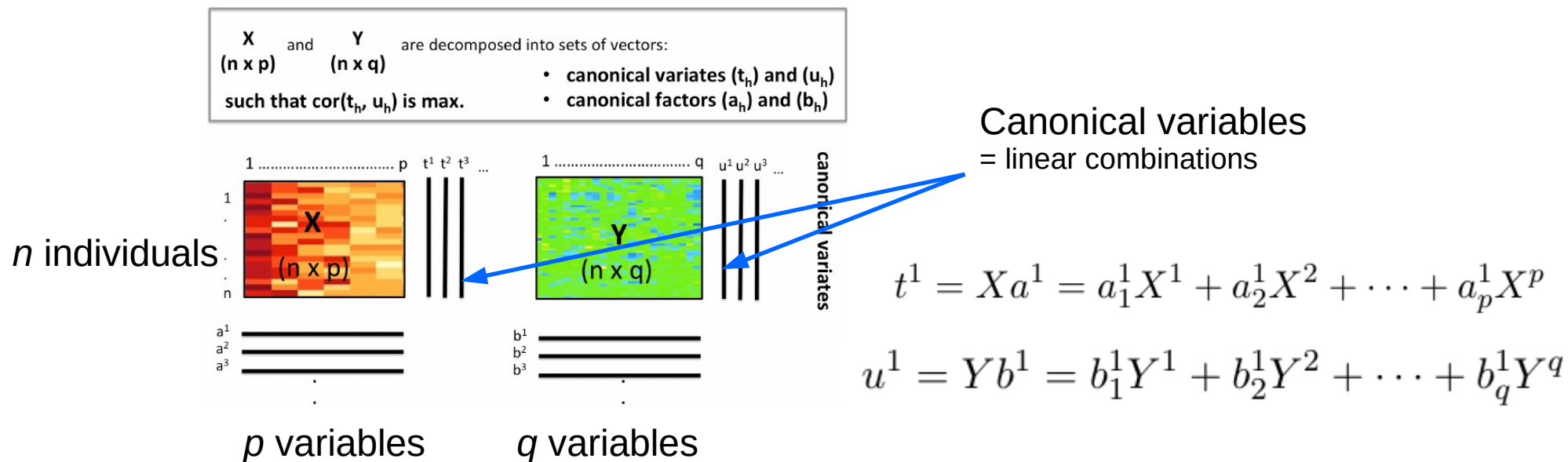
CCA : Canonical Correlation Analysis

On omics datasets collected on similar samples

- identify correlations between datasets
- extract common information of datasets

Components of the CCA are linear combinations of variables from both of the datasets

Correlation between pairs of components is maximised



CCA : correlation $\text{cor}(t^h, u^h)$ is maximized
 PLS : covariance $\text{cov}(t^h, u^h)$ is maximized

rCCA : Canonical Correlation Analysis

CCA can be applied only when $P+Q < N$

→ not the case of omics data

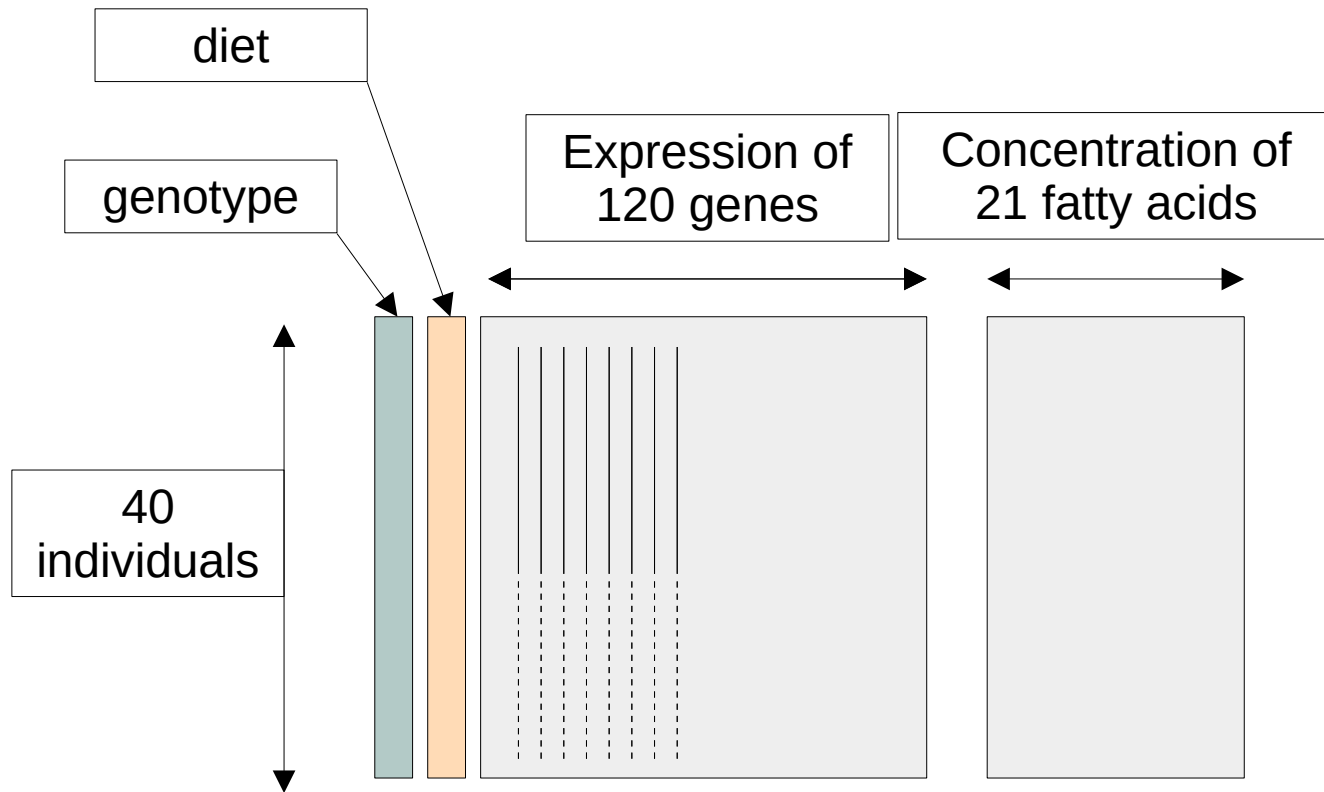
When $P+Q > N$, a regularized version of CCA can be applied (a penalization is applied)

→ cross validation (need calculation time for big datasets)

→ shrinkage (but it does not guarantee the best correlation between both of the datasets)

“nutri-mouse” dataset

Expression of 120 genes and 21 fatty acid concentrations were measured on 40 mice (Martin et al. 2007)



Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

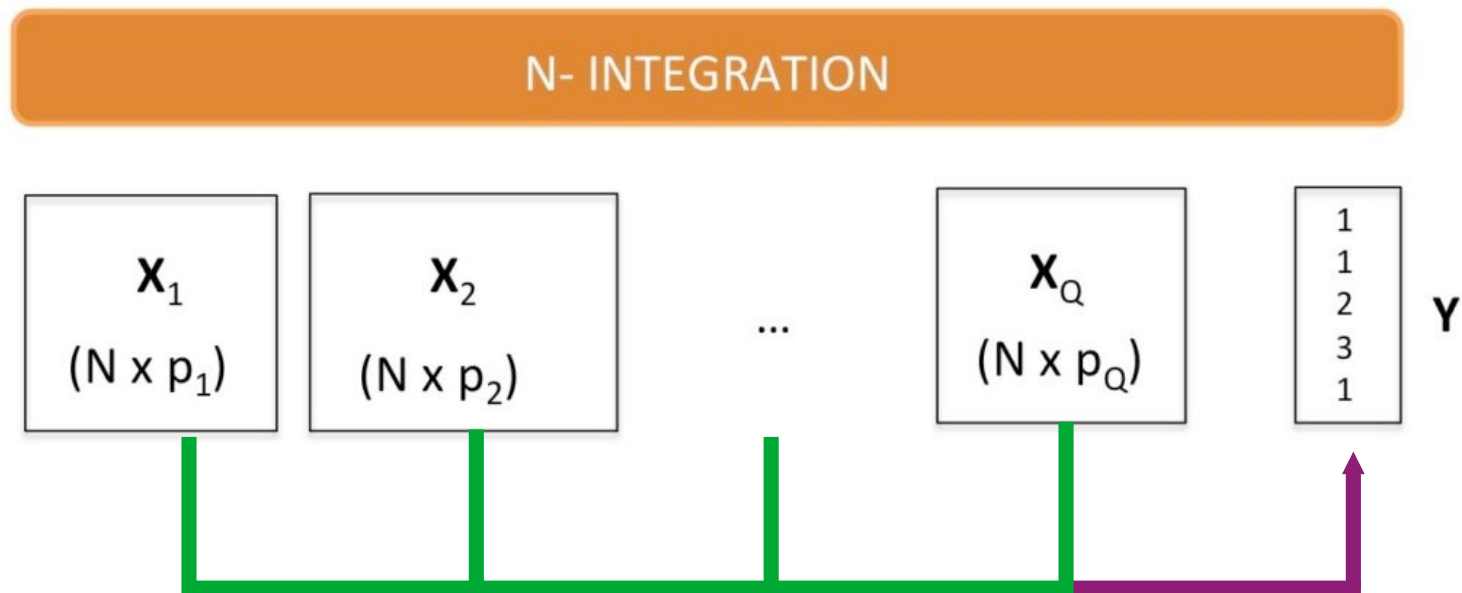
II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

Identification of biomarkers from multiple omics data

DIABLO : Data Integration Analysis for Biomarker discovery using Latent variable approaches for Omics studies

- integration of several datasets
- explain a qualitative variable from these datasets



Maximize covariance between datasets
design matrix values >0.5

Maximize prediction
design matrix values <0.5

Apply DIABLO method

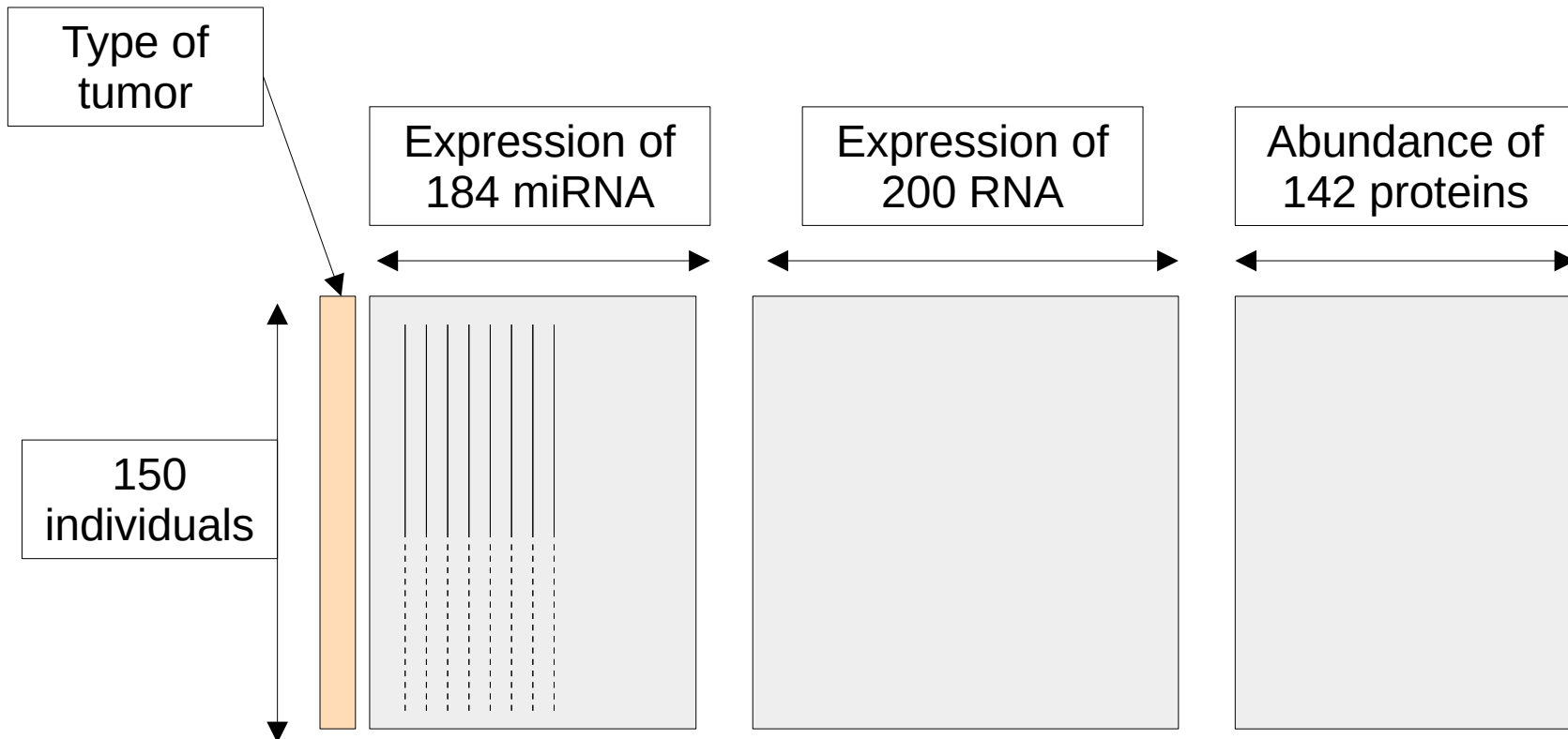
1. Perform pairwise PLS with omics datasets to evaluate the correlation between datasets
2. Choose `design` matrix values depending on the question we ask
3. Apply DIABLO model (`block.splsda`) with a given number of components (just to create the model)
4. Use `perf` function to choose the optimal number of components (cross validations)
5. Use `block.splsda`, using the number of variables to choose for each dataset. These optimal values are determined using the function `tune.block.splsda`

“human breast cancer” dataset

Three types of tumors : Basal, Her2 et LumA.

Expression of miRNA, of mRNA et protein abundances are collected

We aim to identify which variables of these datasets are related and allow to discriminate the different types of tumors.



Outline

I – Explore a dataset

- Principal component analysis (PCA and sPCA)
- Discriminant analysis (PLS-DA)
- Clustering methods (other session)

II - Integration of heterogeneous omics data (mixOmics)

- Identify links between different types of variables : Canonical correlation analysis (CCA, rCCA)
- Discriminate samples from links between different types of variables : DIABLO
- Extract information from different studies : meta-analyses

Meta-analyses – MINT tools

- Same variables collected with different “labs”, “times”, “protocols” that can lead to biased results because of strong datasets effect
 - Origin of the dataset is taken into account before to apply the methods in order to correct variations between dataset
- Multivariate INTEGRative method MINT
- `mint.pcs`
 - `mint.pls`
 - `mint.plsda`
 - `mint.splsda`
 -

Allow to mix data generated in different projects, labs or coming from databases...

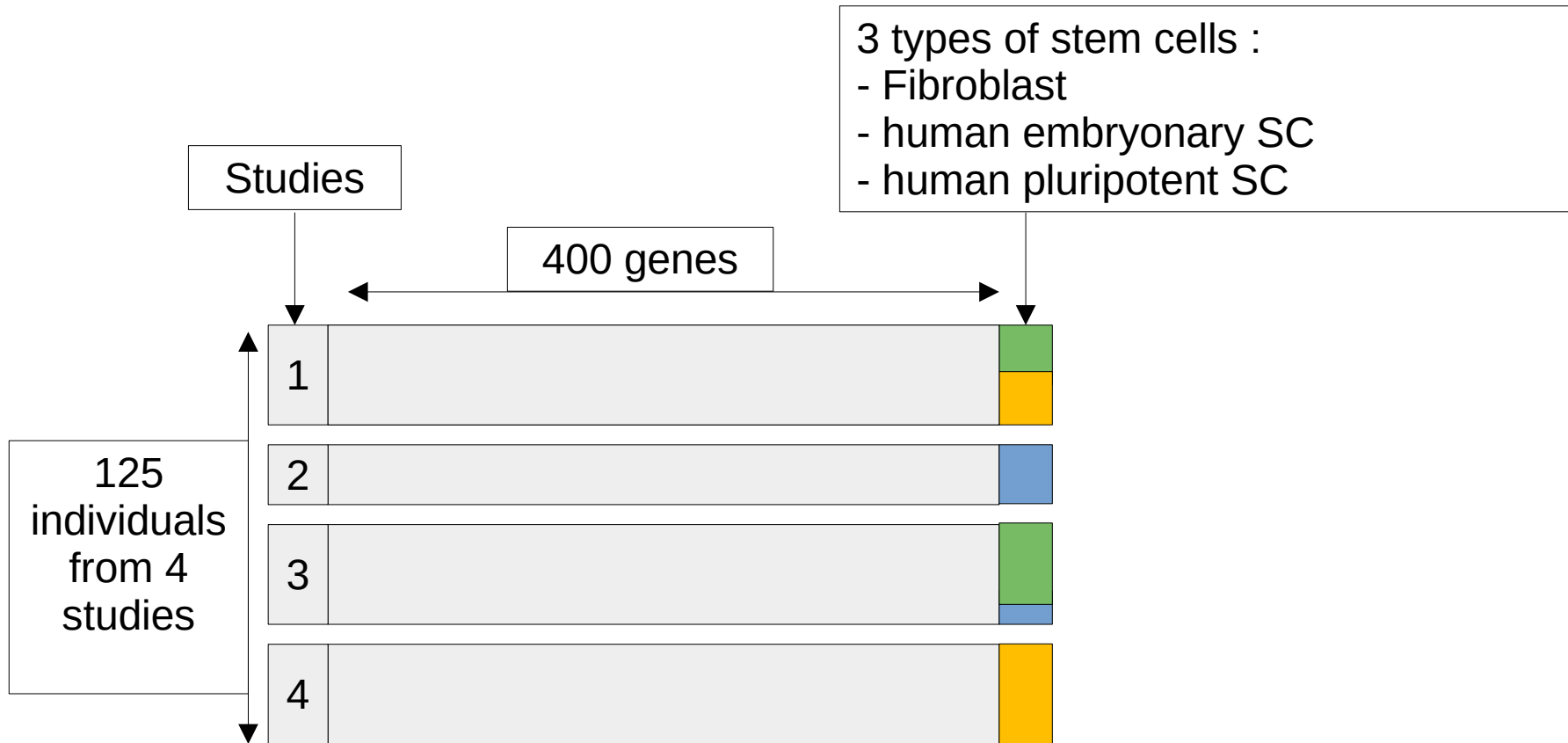
To consider in order to save :



“stem cells” dataset


Transcriptomes of stem cells coming from 4 studies (125 samples in total)

Expression of 400 genes in common to the 4 studies were collected on the 125 samples



To summarize...

	Explore	Discriminate / Classify
Only one dataset	<ul style="list-style-type: none">• PCA• sPCA (variables selection) (srbct)	<ul style="list-style-type: none">• PLS-DA (supervised) (srbct)• Clustering (not supervised)
Several datasets on the same individuals	<ul style="list-style-type: none">• CCA / rCCA (nutrimouse)	<ul style="list-style-type: none">• DIABLO (breast cancer data)
Meta-analyses	<ul style="list-style-type: none">• MINT (stem cells)	



<https://mixomics.org>