Multivariate analysis and integration of omics data

Prof. Kim-Anh Lê Cao

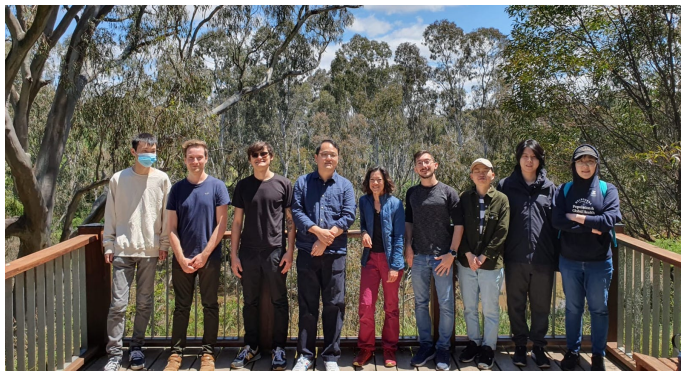
NHMRC L2 investigator
Melbourne Integrative Genomics
School of Mathematics and Statistics

THE UNIVERSITY OF MELBOURNE

@mixOmics_team | www.lecao-lab.science.unimelb.edu.au

Lê Cao lab

- Expertise in statistics and computational biology
- Team of statisticians, bioinformaticians, data analysts and software developpers
- Strengths: multi-disciplinary research, accessible software for the community, methods that are (often) technology agnostic
- mixOmics team finalist for the Australian Eureka prize 2023



We develop methods for microbiome and omics data integration.

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council

The Australian Research Council Centre of Excellence in Quantum Biotechnology

Outline

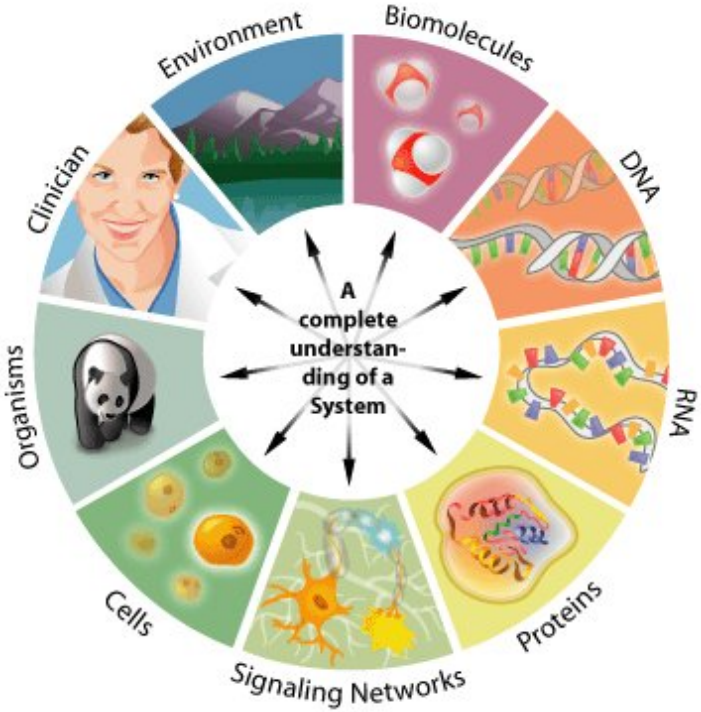
A holistic view of a biological system

From reductionism ...

1 gene = 1 hypothesis = 1 statistical test

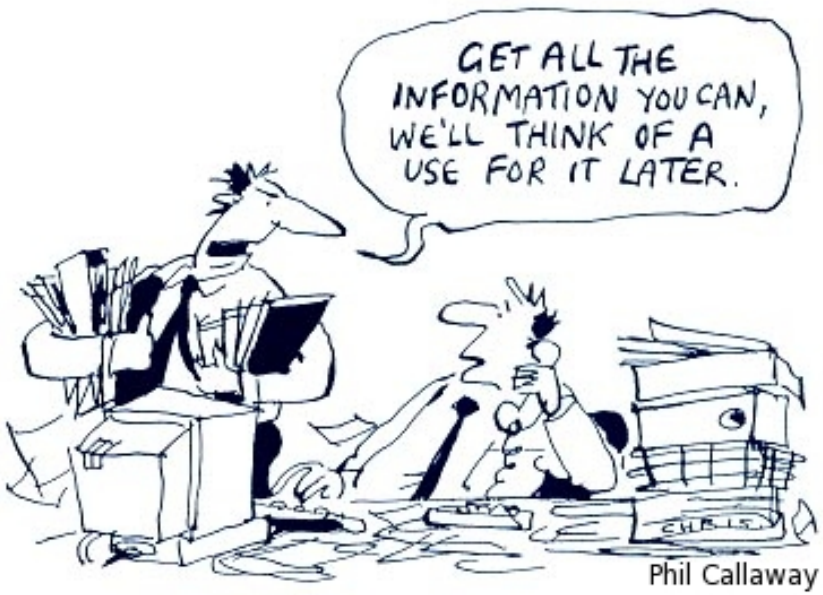
... to holism:

Thousands of molecules = ??



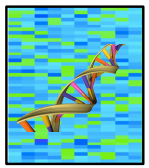
When biology and statistics meet

A close interaction between statisticians, bioinformaticians and molecular biologists is essential to provide meaningful results

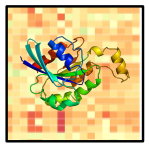


- Unlimited quantity of data from multiple and heterogeneous sources
- Computational issues to foresee
- Biological interpretation for validation
- Keep pace with new technologies

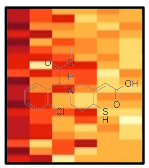
The 'omes and the biological dogma? not that straighforward



Transcriptomics
$\sim 100,000$ transcripts



$\downarrow$
Proteomics
$\sim 1,000,000$ proteins



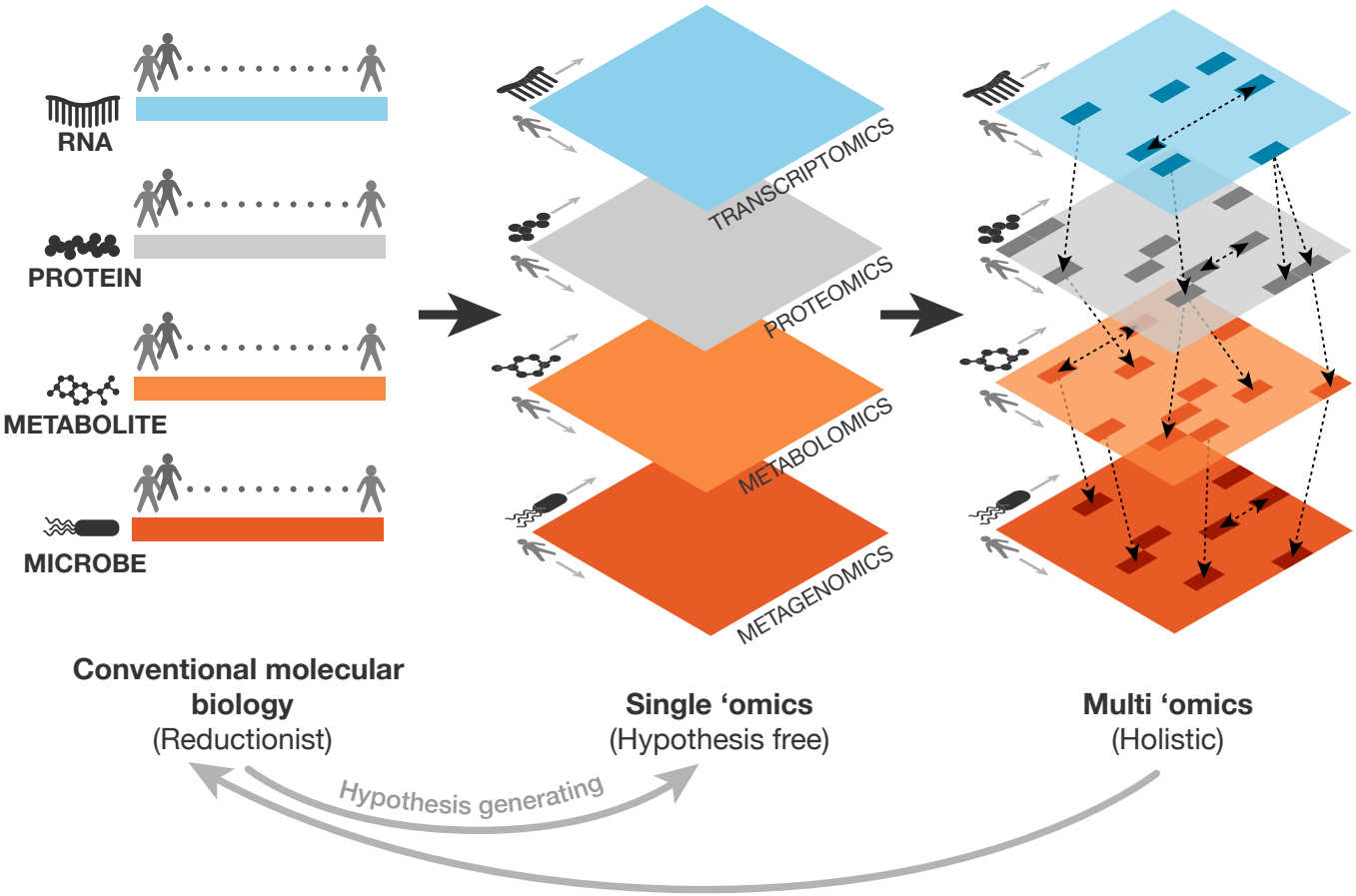
$\downarrow$
Metabolomics
$\sim 2,500$ compounds

Microbiome
$\sim 1,500$ trillion microbes

$\vdots$

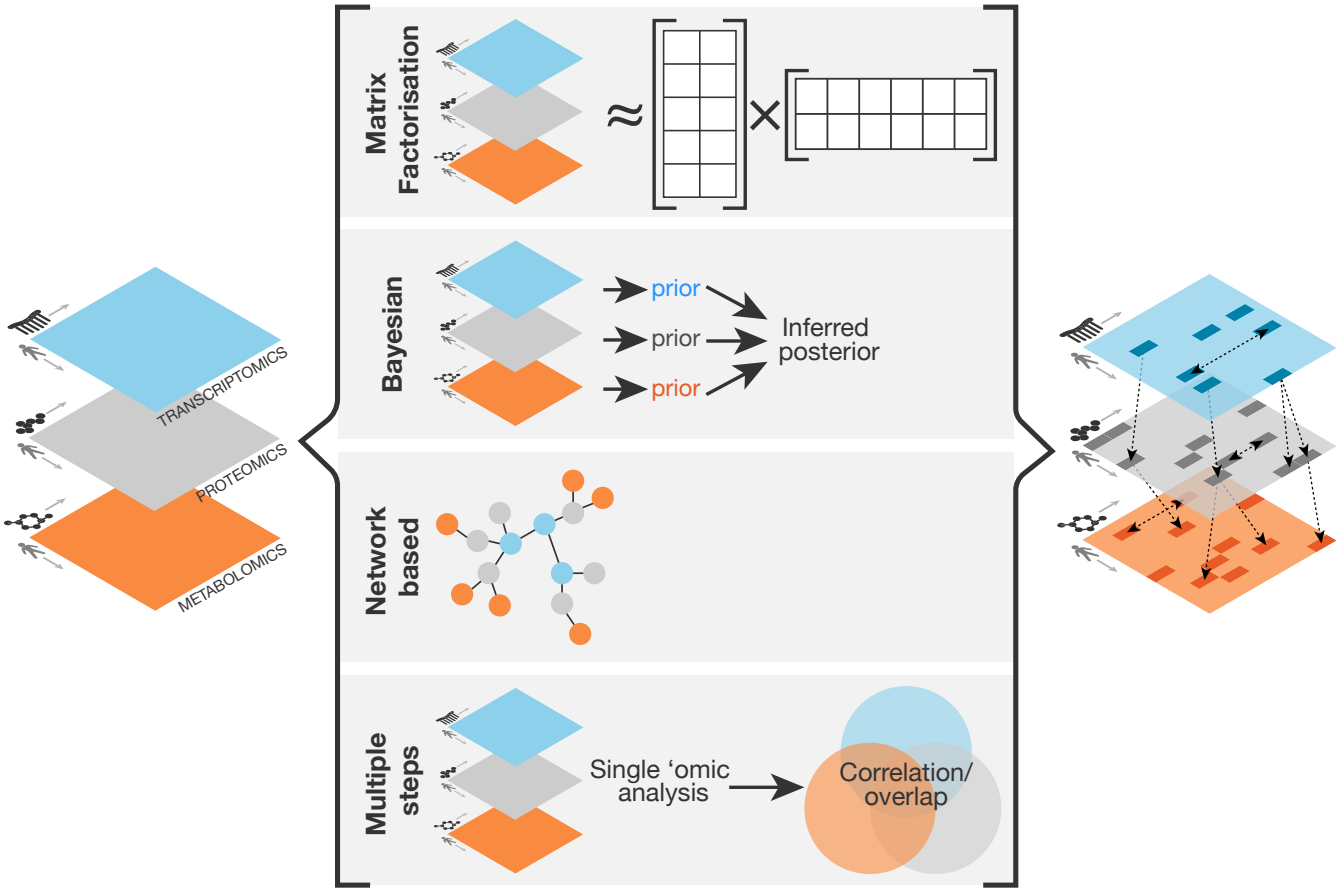
$\vdots$
Clinical data
many

A new research field



Start with a holistic view, rather than a traditional, reductionist, hypothesis-driven view, then generate new hypotheses.

Data integration from an analytical point of view



→ Our methods are based on matrix factorisation techniques to reduce data dimension

Multivariate analysis

Molecular entities act together to trigger cells' responses. We need to shift the 'one-gene hypothesis' paradigm to obtain deeper insight into biological systems.

- Identify a combination rather than univariate biomarkers
- Reduce the dimension of the data for a better understanding of complex biological systems
- Integrate multiple sources of biological data
- Supervised analysis for sample prediction

Example of multivariate methods: Principal Component Analysis (PCA), Projection to Latent Structures (PLS) models for data integration

Outline

Principal Component Analysis

PCA: the workhorse for linear multivariate statistical analysis is an (almost) compulsory first step in exploratory data analysis to:

- Understand the underlying data structure
- Identify bias, experimental errors, batch effects.

Original variables are replaced by artificial variables (principal components) which explain as much information as possible from the original data.

In PCA, the variance == information contained in the data

Linear combination of the variables Height and Weight

We assign two coefficients $w_1 = 0.5$ and $w_2 = 2$ to the variables Height and Weight respectively:

	Height		Weight		Linear combination
$0.5 \times$	174.0	$+ \quad 2 \times$	65.6	$=$	218.20
	175.3		71.8		231.25
	193.5		80.7		258.15
	186.5		72.6		238.45
	187.2		78.8		251.20
	181.5		74.8		240.35
	184.0		86.4		264.80
	184.5		78.4		249.05
	175.0		62.0		211.50
	184.0		81.6		255.20

Two variables are summarised into a single variable: a linear combination also called component.

$\rightarrow$ efficient algorithms for large datasets

$\rightarrow$ challenge: identify the coefficients assigned to each variable.

Now a 'bigger' data set

Data:

	Var1	Var2	Var3	Var4	Var5
Indiv 1	3.97	3.16	3.54	3.89	2.11
Indiv 2	2.05	2.36	3.89	3.50	1.58
Indiv 3	3.36	3.11	4.20	2.26	-0.09
Indiv 4	4.15	4.79	5.43	3.30	2.00
Indiv 5	2.91	3.05	3.87	4.99	1.69
Indiv 6	3.44	3.96	3.54	3.89	2.11
Indiv 7	3.65	2.22	3.89	3.50	1.58
Indiv 8	4.40	3.60	4.20	2.26	-0.09
Indiv 9	2.68	2.29	5.43	3.30	2.00
Indiv 10	3.85	3.89	3.87	4.99	1.69

$\Longrightarrow$

Components:

	PC1	PC2	PC3	PC4	PC5
Indiv 1	0.62	0.31	-0.45	0.71	-0.15
Indiv 2	0.41	-1.53	0.05	-0.34	-0.47
Indiv 3	-1.91	-0.53	-0.35	-0.39	-0.07
Indiv 4	-0.33	1.72	1.31	-0.08	-0.11
Indiv 5	1.36	-0.24	-0.38	-0.52	0.35
Indiv 6	0.69	0.64	-0.29	0.02	-0.69
Indiv 7	0.07	-0.71	-0.29	0.73	0.26
Indiv 8	-2.16	0.46	-0.53	0.04	0.18
Indiv 9	0.14	-1.10	1.45	0.13	0.27
Indiv 10	1.11	0.97	-0.52	-0.31	0.42

$\rightsquigarrow$ Replace the original variables by components that explain as much information as possible from the original data.

Not all components are needed to summarise the information.

Variance covariance matrix

COV(data) =

	Var1	Var2	Var3	Var4	Var5
Var1	0.52	0.39	0.00	-0.12	-0.14
Var2	0.39	0.69	0.08	0.06	0.03
Var3	0.00	0.08	0.48	-0.22	0.03
Var4	-0.12	0.06 -	0.22	0.87	0.54
Var5	-0.14	0.03	0.03	0.54	0.71

COV(PCs) =

	PC1	PC2	PC3	PC4	PC5
PC1	1.39	0.00	0.00	0.00	0.00
PC2	0.00	1.00	0.00	0.00	0.00
PC3	0.00	0.00	0.56	0.00	0.00
PC4	0.00	0.00	0.00	0.19	0.00
PC5	0.00	0.00	0.00	0.00	0.13

Sum of variances:

0.52+0.69+0.48+0.87+0.71= 3.27

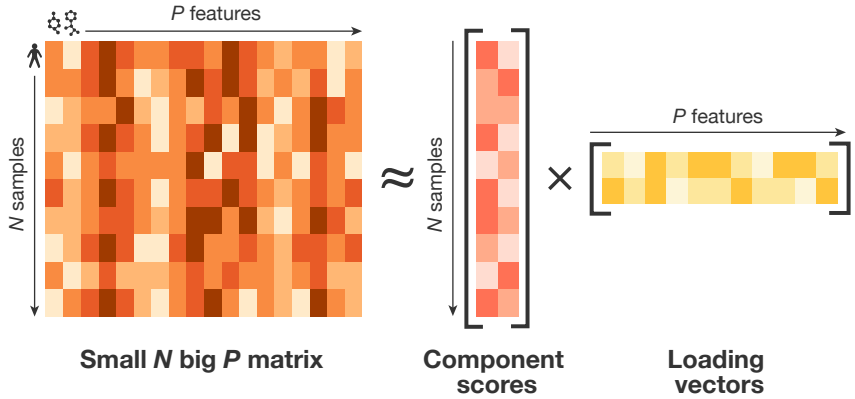
Sum of variances:

1.39+1+0.56+0.19+0.13=3.27

The top principal components explain as much information (variance) as possible from the original data.
PCs are orthogonal to each other (covariance = 0).

PCA is a matrix decomposition technique

Q: What are the major sources of variation in the data?

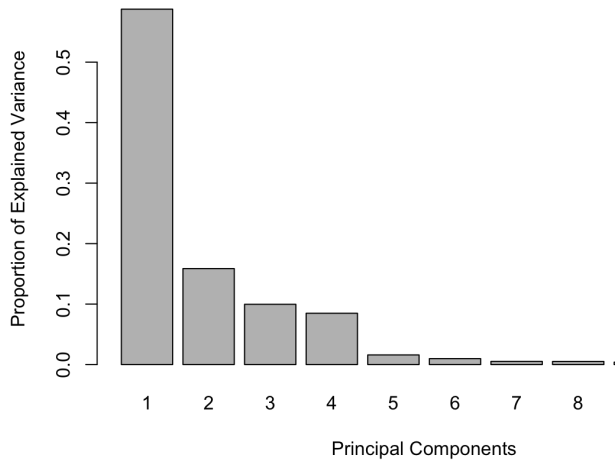


→ Maximise the variance of each component

- A component score is a linear combination of features
- Loading weights (variable coefficients) indicate the importance of each feature in the linear combination
- To each component is associated a loading vector

Choosing the number of components to summarise enough information

- Screeplot of prop of explained variance. Any elbow?
- Look at sample plot. Makes sense?
- Some stat tests exist to estimate the 'intrinsic' dimension, but limited

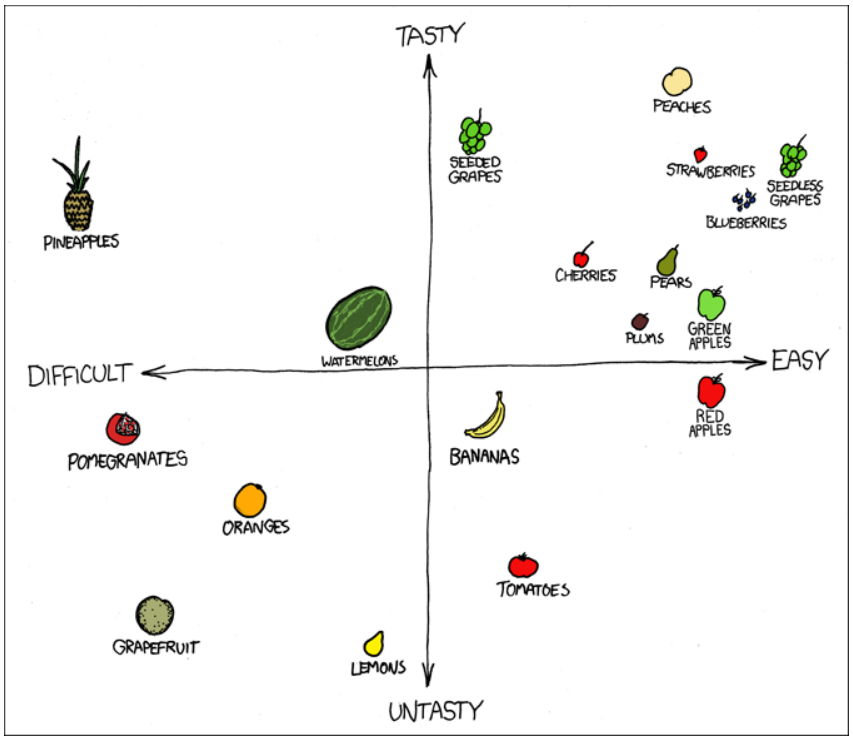


Cumulative proportion of explained variance for the first 8 principal components:

PC1	PC1 to 2	PC1 to 3	PC1 to 4	PC1 to 5	PC1 to 6	PC1 to 7	PC1 to 8
0.59	0.75	0.84	0.93	0.946	0.956	0.961	0.966

(We can have as many components as the rank of the matrix X)

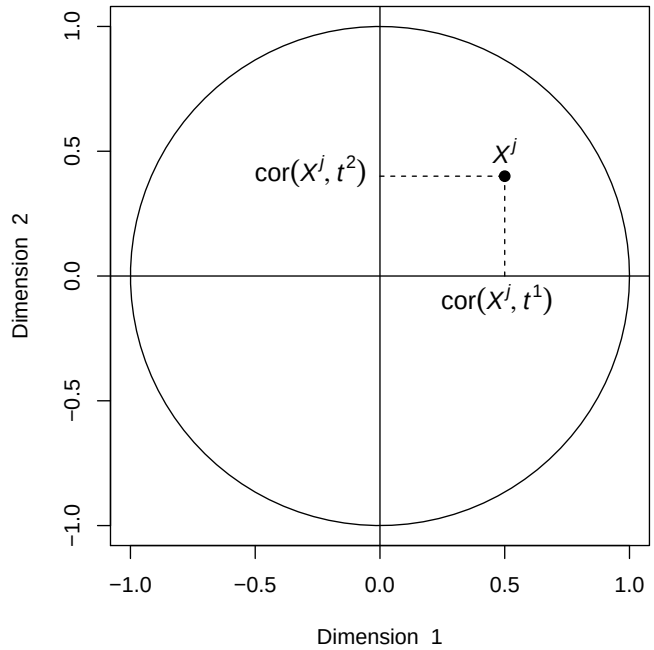
PCs are used for visualisation: sample plot



- 'Samples' here are fruit
- Variables are different characteristics (e.g texture, acidity)

• Summarise the data into 2 components (dimensions) then project the data in the space spanned by those 2 components.
• Each component has a 'meaning' as it tends to spread the samples according to their characteristics.

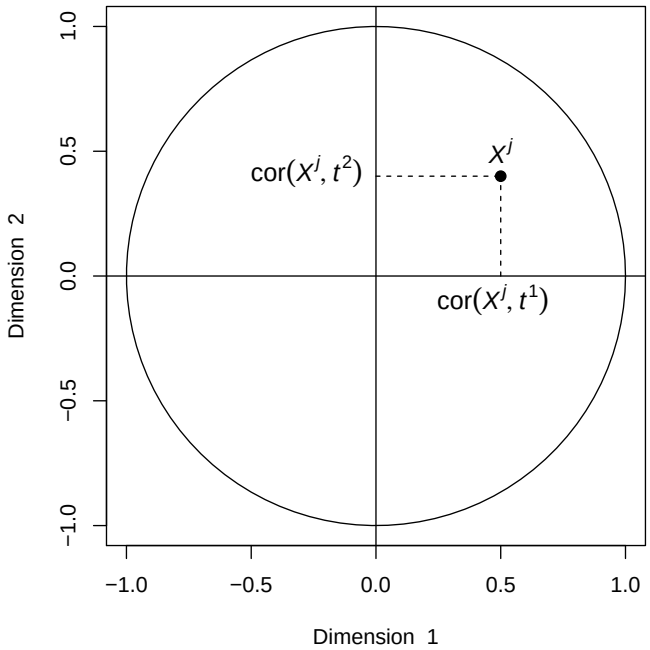
Variable representation



Coordinate of each variable: calculate the correlation between each variable and each PC: $\text{cor}(\boldsymbol{X}^j, \boldsymbol{t}^1), \text{cor}(\boldsymbol{X}^j, \boldsymbol{t}^2)$

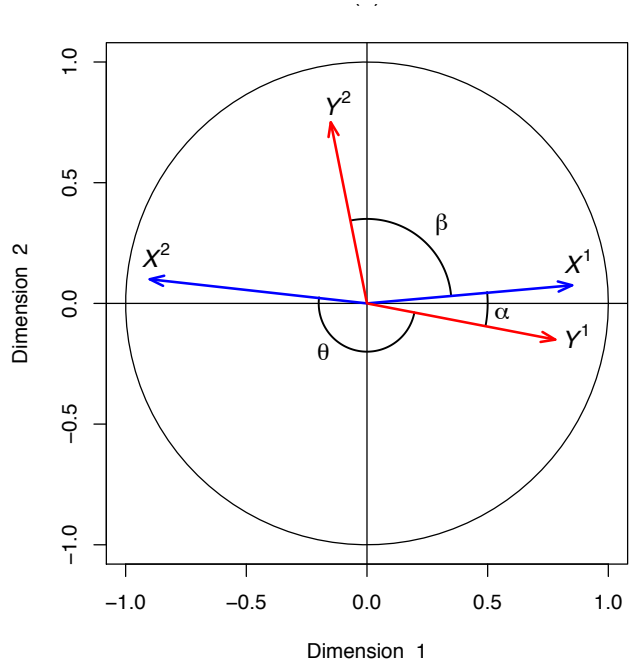
(data should be centered and scaled in PCA)

Variable representation



Coordinate of each variable: calculate the correlation between each variable and each PC: cor($\boldsymbol{X}^j$, $\boldsymbol{t}^1$), cor($\boldsymbol{X}^j$, $\boldsymbol{t}^2$)

(data should be centered and scaled in PCA)

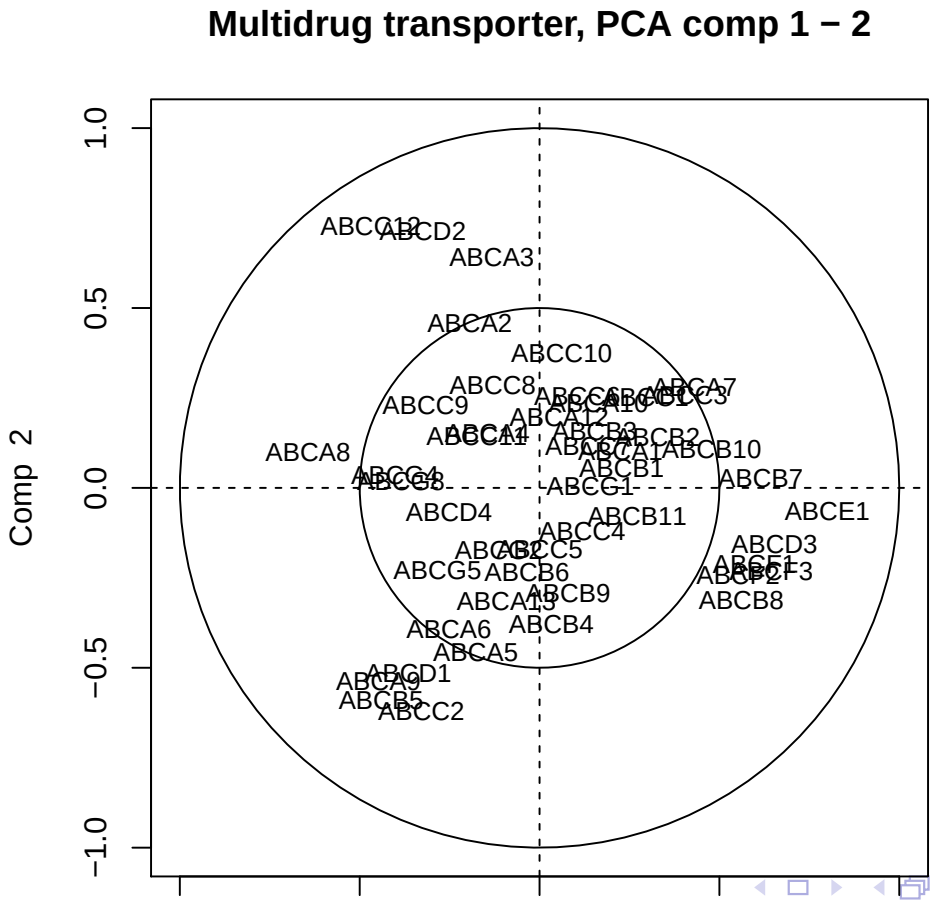


Correlation between two variables = cos(angle) between 2 variable vectors

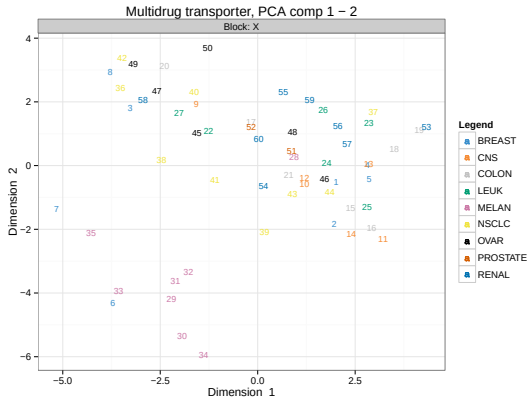
- $cos(\alpha)$ close to 1 $\rightarrow$ > 0 corr
- $cos(\beta)$ close to 0 $\rightarrow$ no corr
- $cos(\beta)$ close to -1 $\rightarrow$ < 0 corr

Variable representation

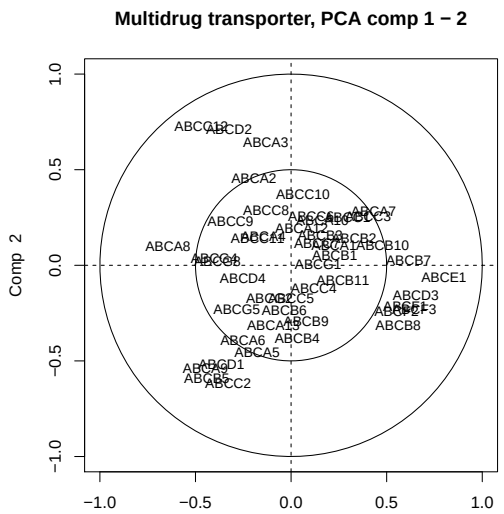
In `mixOmics` the variables are represented as a dot or a name (no arrow)



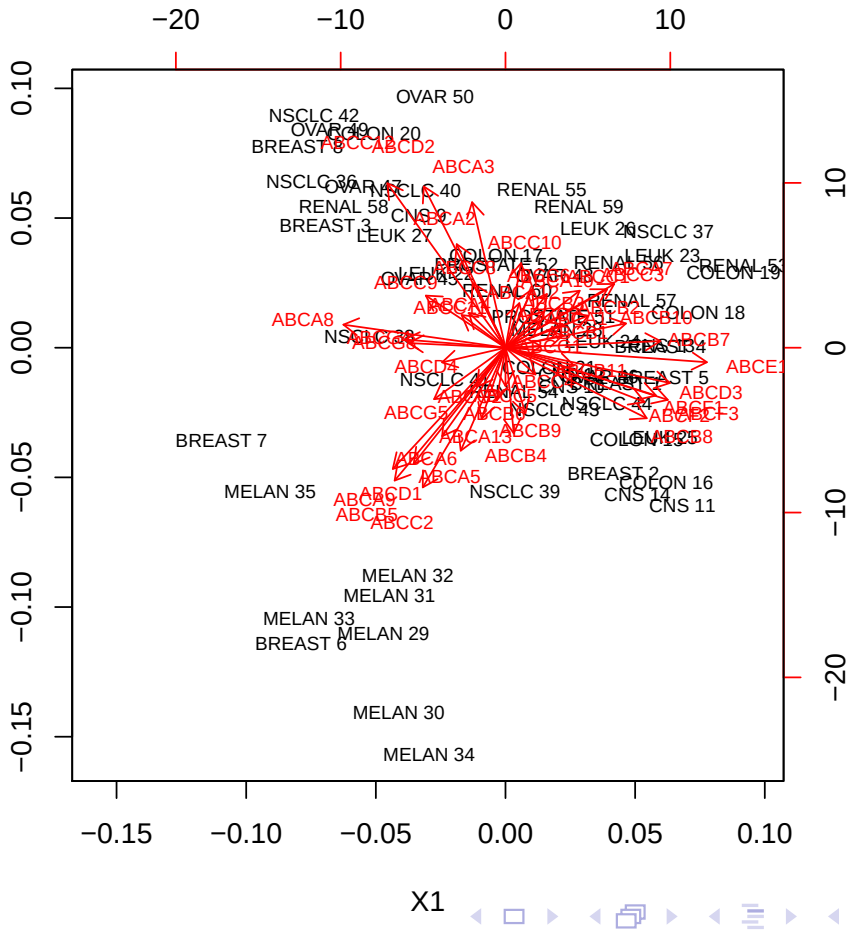
The `biplot` overlays sample and variable plots to understand their relationship



+



⟹



Outline

Classification analysis

Supervised methods model a relationship between the data and a measurable outcome ($\neq$ unsupervised analysis like PCA).
If the outcome is a discrete variable (e.g. type of treatment)
$\rightarrow$ classification.

Aims:

- **Descriptive**: weight the variables in an *optimal* manner so that their combination (component) best separates the k classes of samples,
  $\rightarrow$ according to a statistical criterion

- **Predictive**: predicting the class of a new samples
  $\rightarrow$ construction of a classifier (= set of rules)
  $\rightarrow$ diagnostic/prognostic measures w.r.t sensitivity and specificity (e.g. ROC, AUC)

Feature selection

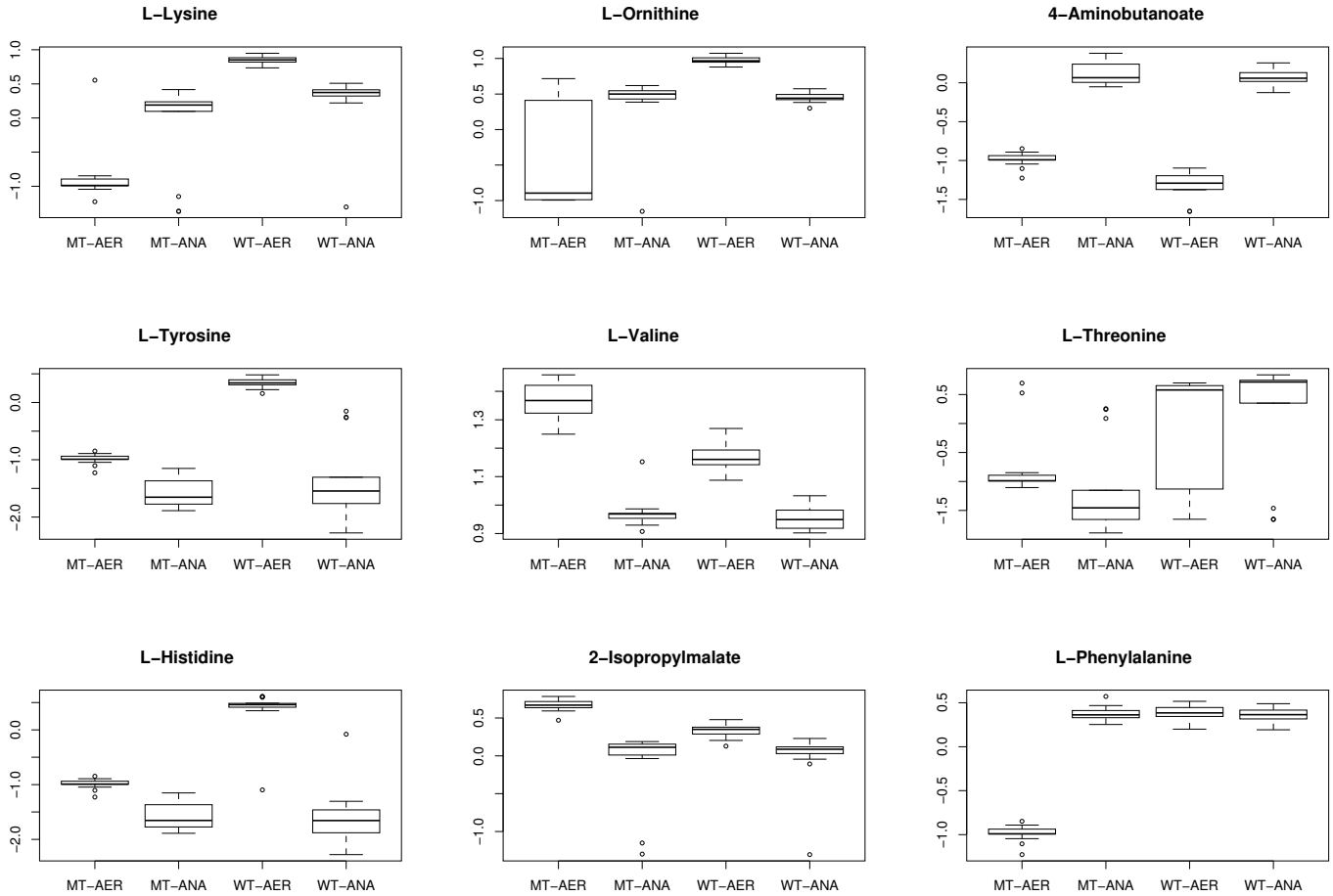
Classification rule built on

- All variables (e.g. genes) or
- A small subset of variables

$\rightsquigarrow$ In molecular biology, a *biomarker panel or molecular signatures* = subset of molecular features with high discriminative power.

$\rightsquigarrow$ Multivariate variable selection often represent a diverse biomarker signature that can not be obtained using univariate statistical methods.

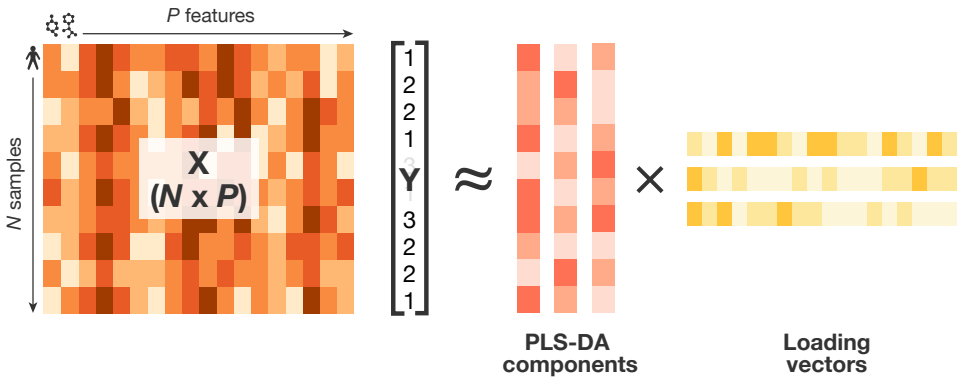
Example of molecular signature



Yeast metabolite data: multivariate biomarker signature w.r.t 4 groups

PLS - Discriminant Analysis (PLS-DA)

Q: Can I discriminate my samples according to their groups?



→ Maximise the covariance* between each component and the outcome

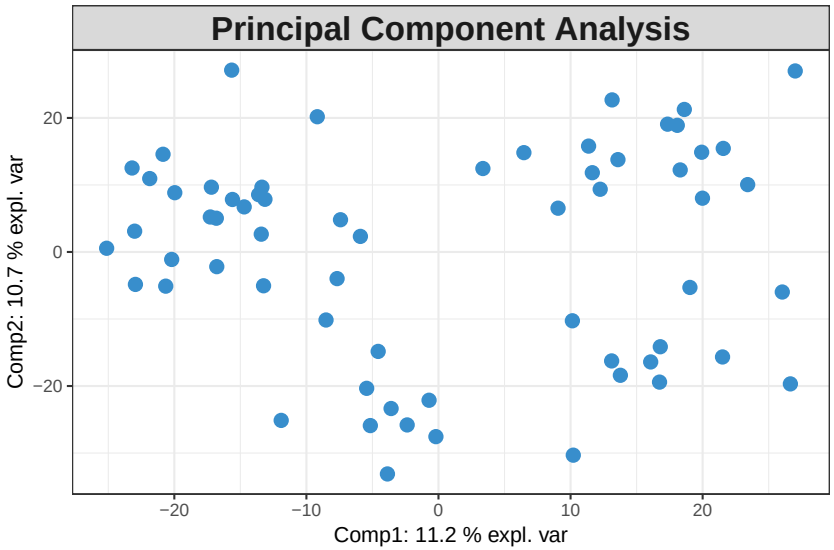
- Components discriminate sample groups
- Feature selection in the loading vectors: weights shrunk to zero using Lasso penalisation

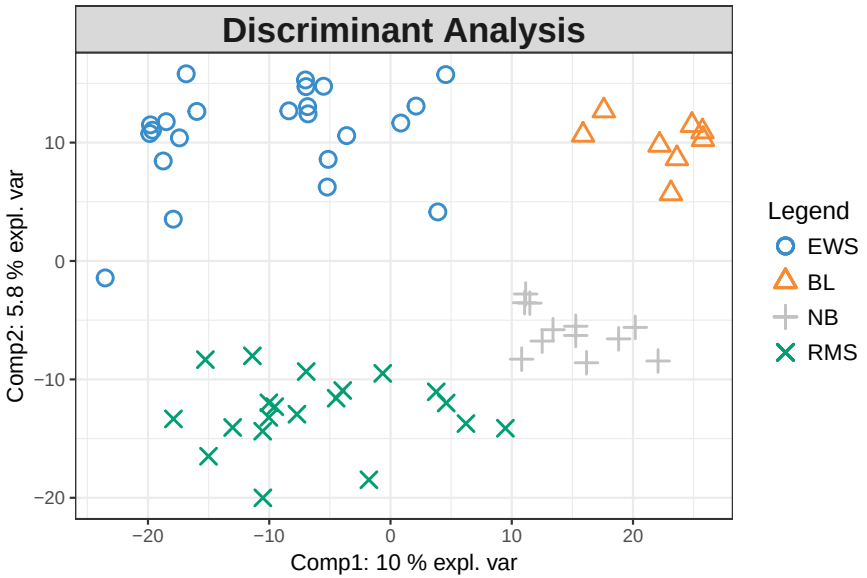
*covariance is akin to correlation

Lê Cao et al. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 12:253.

Sample visualisation in a reduced space using components

Example: SRBCT data with 63 samples (shown in plot) and 3,116 genes





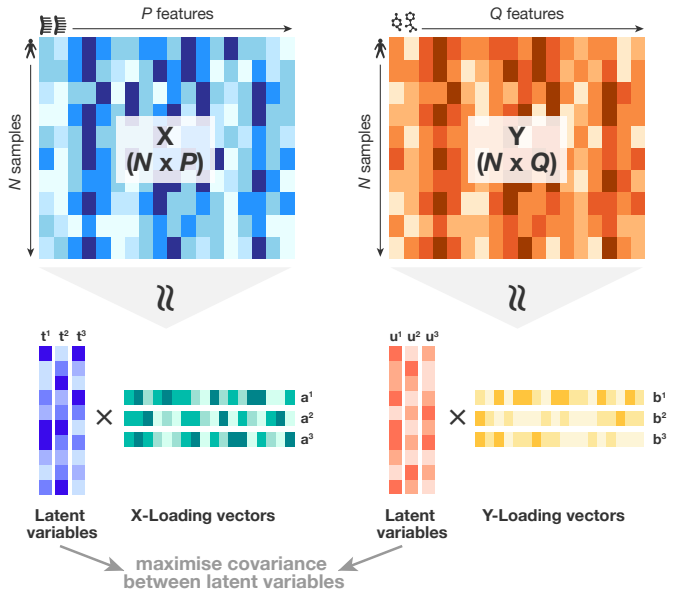
- Unsupervised exploratory analysis: 'similar' samples cluster but no information about sample groups is included in the analysis

- Supervised analysis: Samples cluster according to their respective group

Outline

Integration of two omics datasets with PLS

Q: Can I extract common information across 2 datasets?



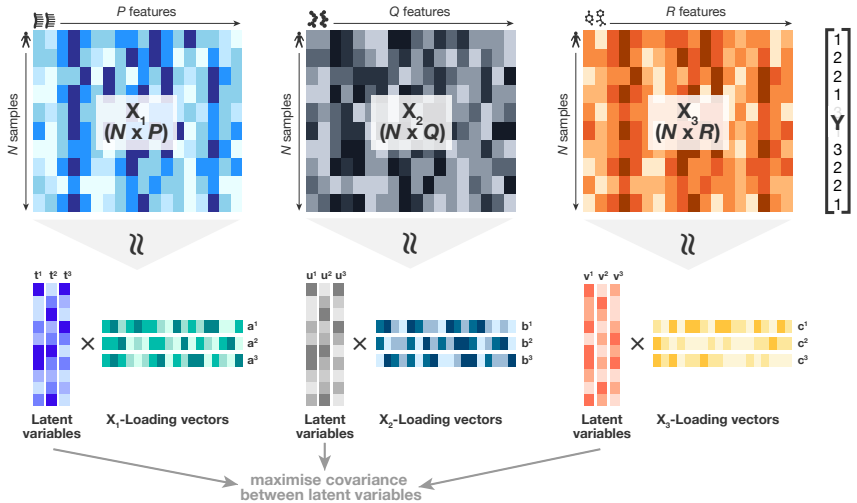
→ Maximise the covariance between pairs of components

- Components are maximally correlated across omics & are built on selected features that are correlated

Lê Cao KA et al. (2008). A sparse PLS for variable selection when integrating omics data *Stat App in Gen & Mol Biol*

Multi-omics data integration with DIABLO

Q: Can I extract common information across multiple datasets?

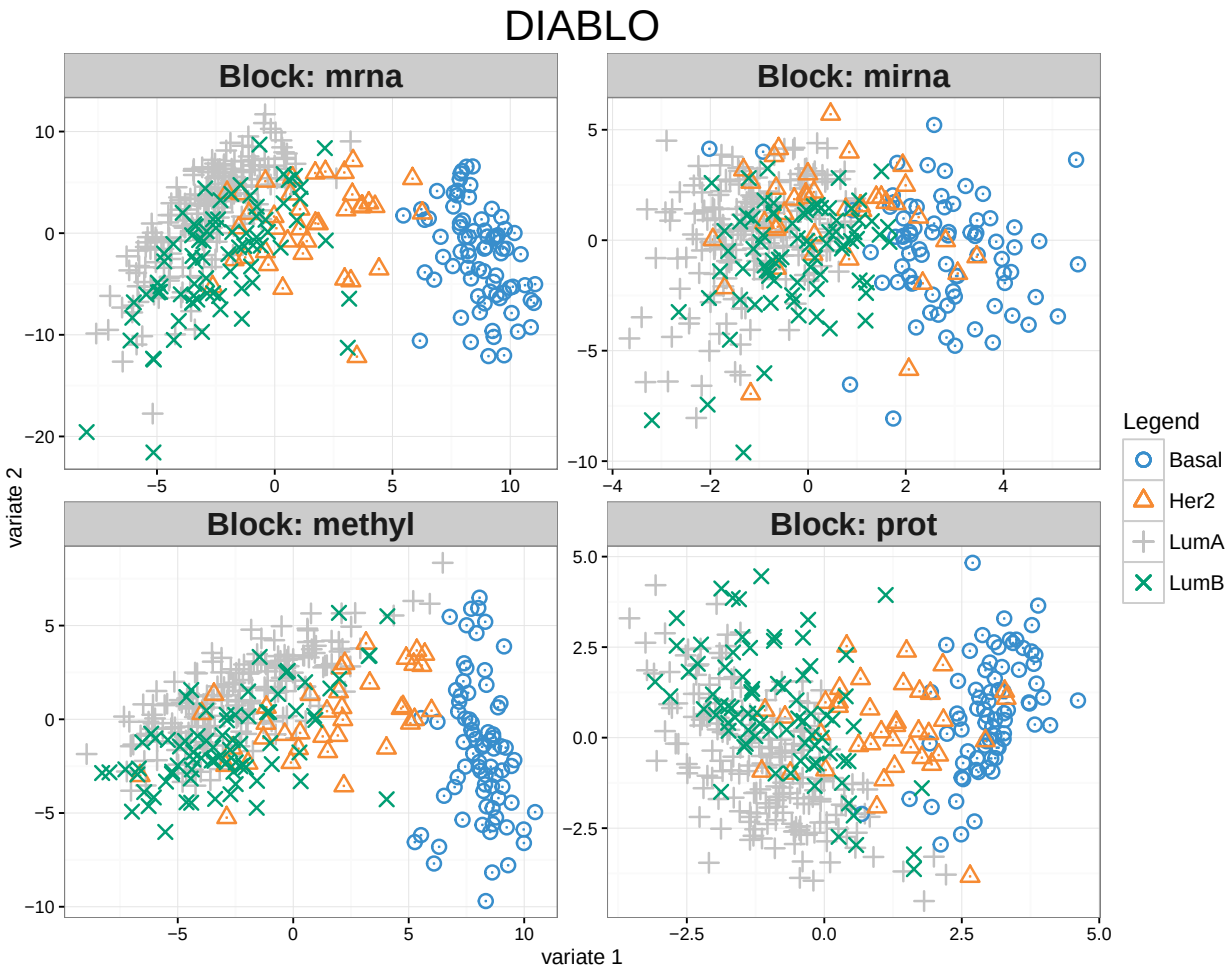


→ Maximise the covariance between pairs of components and the outcome

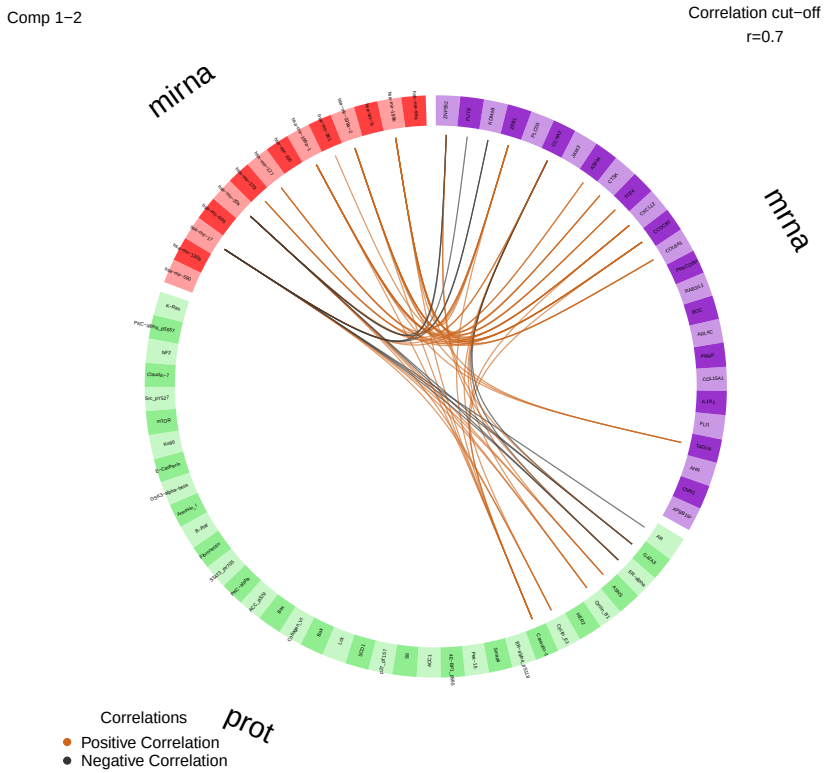
- Components are maximally correlated across omics, are built on selected features that are correlated and discriminate the outcome

Singh A, Gautier B, Shannon C, Vacher M, Rohart F, Tebbutt S, Lê Cao K-A (2019). DIABLO: identifying key molecular drivers from multi-omic assays, an integrative approach. *Bioinformatics*.

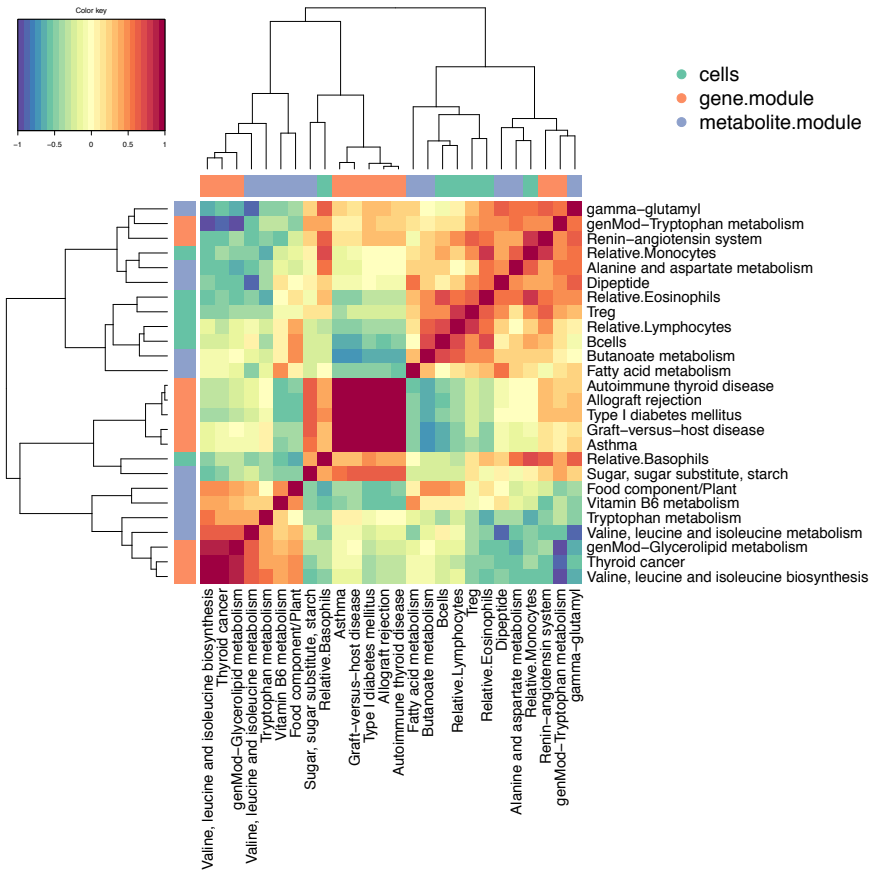
DIABLO: each sample is projected in each own reduced space through the components



DIABLO: multi-omics signatures are highly correlated



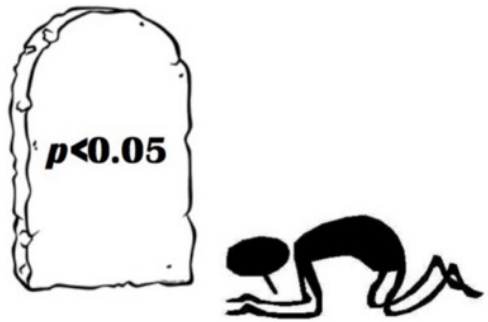
Circos plot



Clustered image map

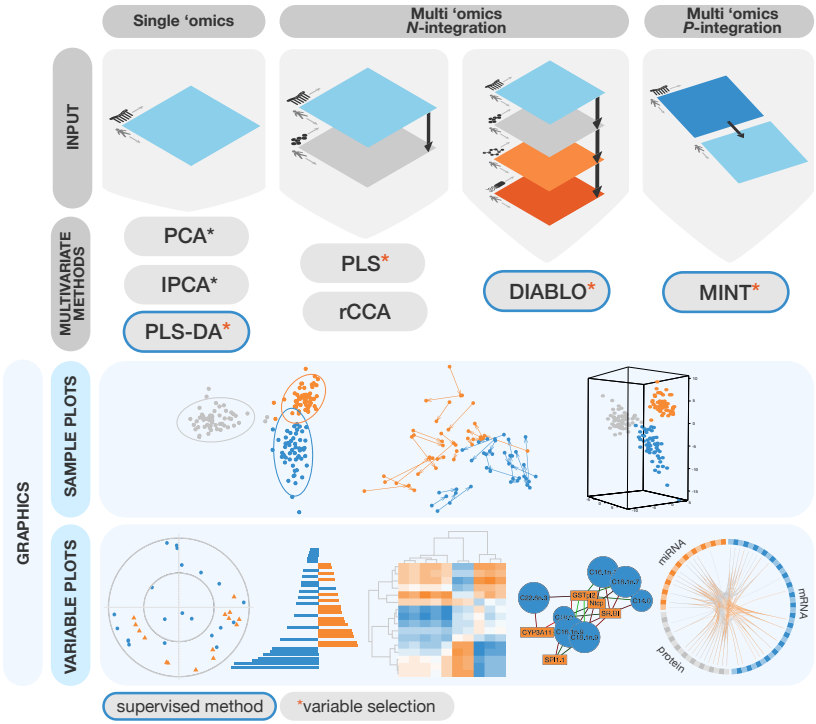
One note on p-values in multivariate data analysis

P-values are hardly relevant in this context



Outline

mixOmics: An R toolkit for multivariate data analysis and integration of 'omics data

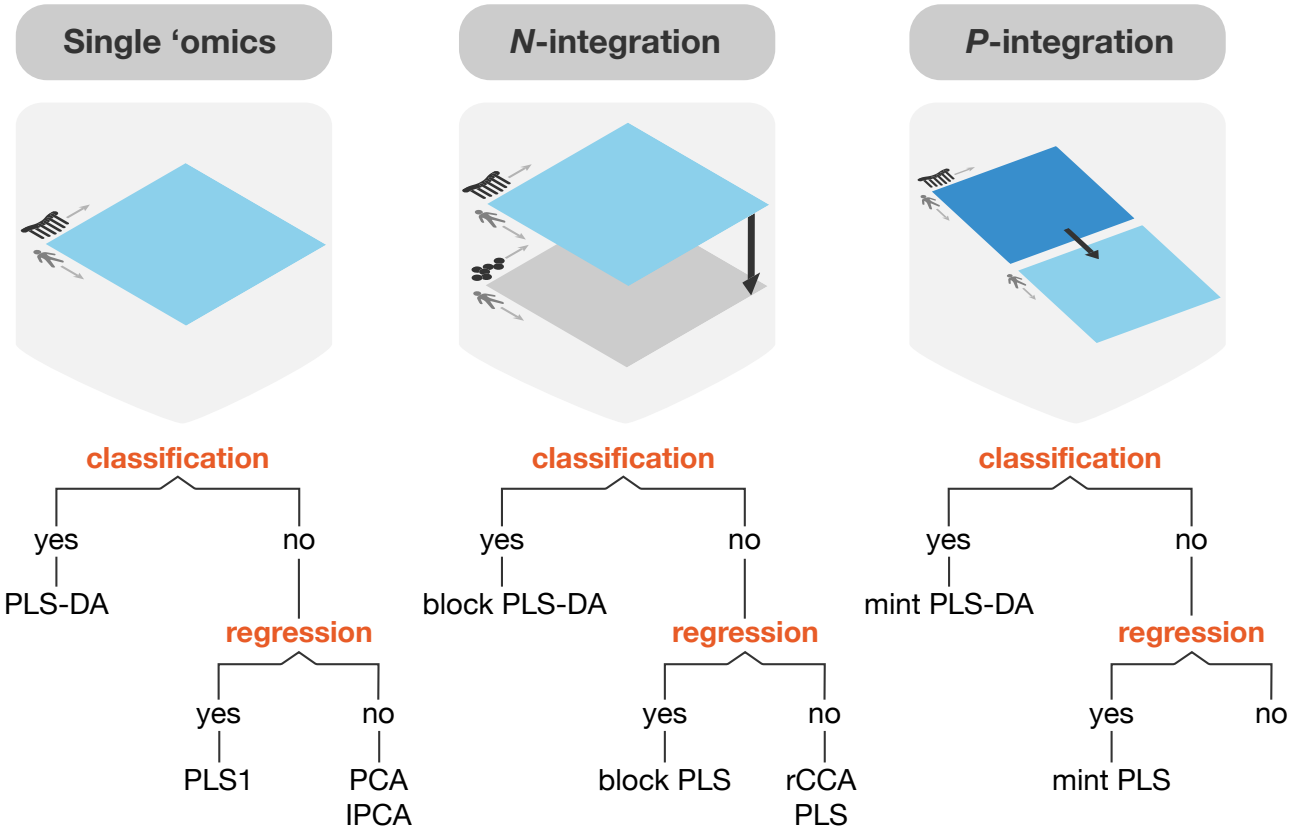


- 19 multivariate methods
- Top 5% Bioconductor downloads
- Tutorials on www.mixOmics.org, book and online courses
- Multi-day workshops to 750+ attendees since 2014.

  → Online course twice a year, see website

Rohart F, Gautier B, Singh, M, Lê Cao K-A. (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comp Biol* 13(11).

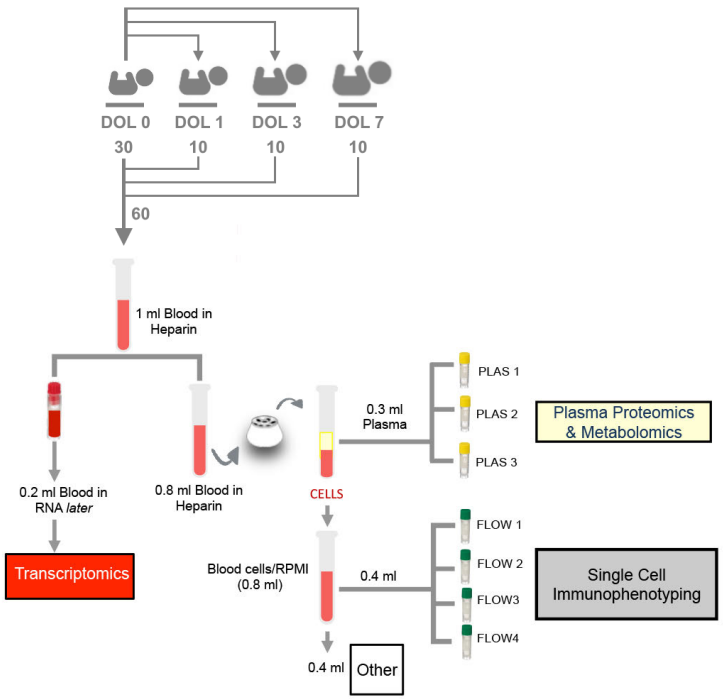
19 multivariate methods (13 novel)

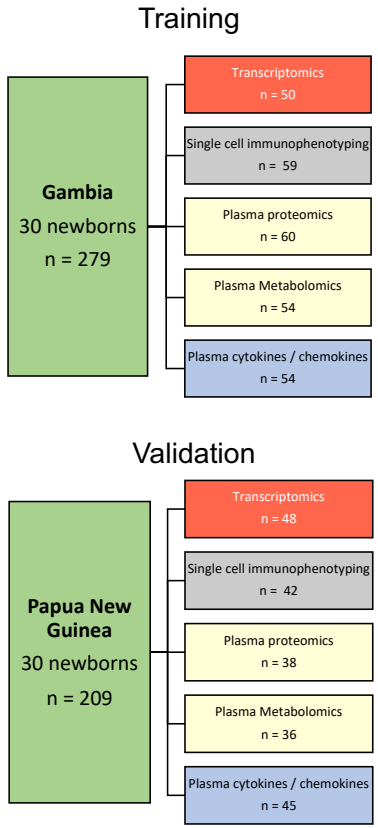


$\rightsquigarrow$ each method answers a specific biological question

Outline

#smallbig study: the first week of human life

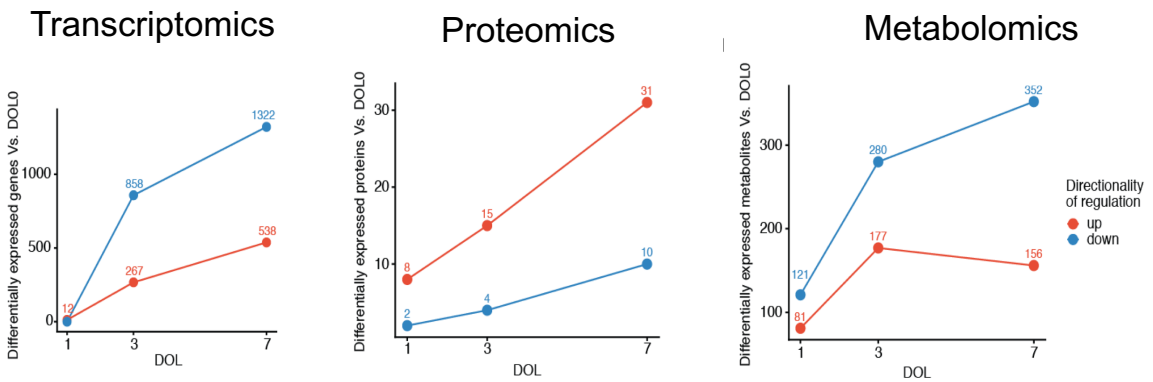


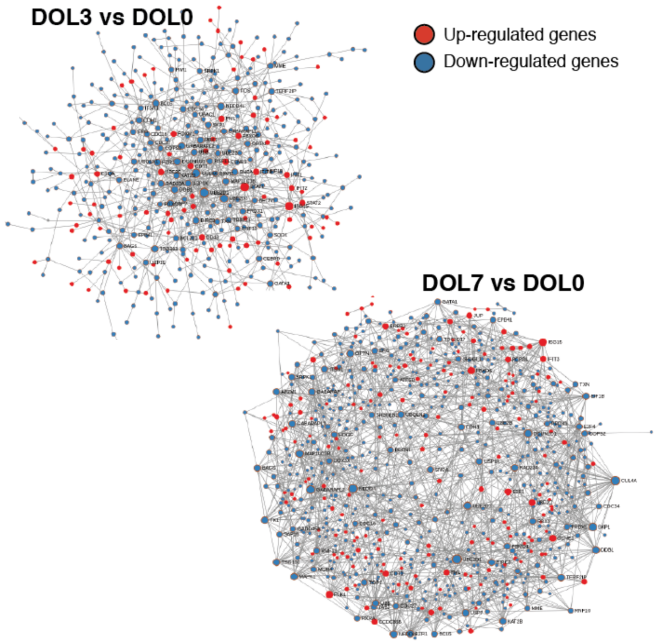


Small biosample amount: < 1ml of blood from newborns, five data types

Lee, Shannon, ..., Lê Cao, ... & Kollman (2019) Dynamic molecular changes during the first week of human life follow a robust developmental trajectory. *Nat Comm* 10:1092.

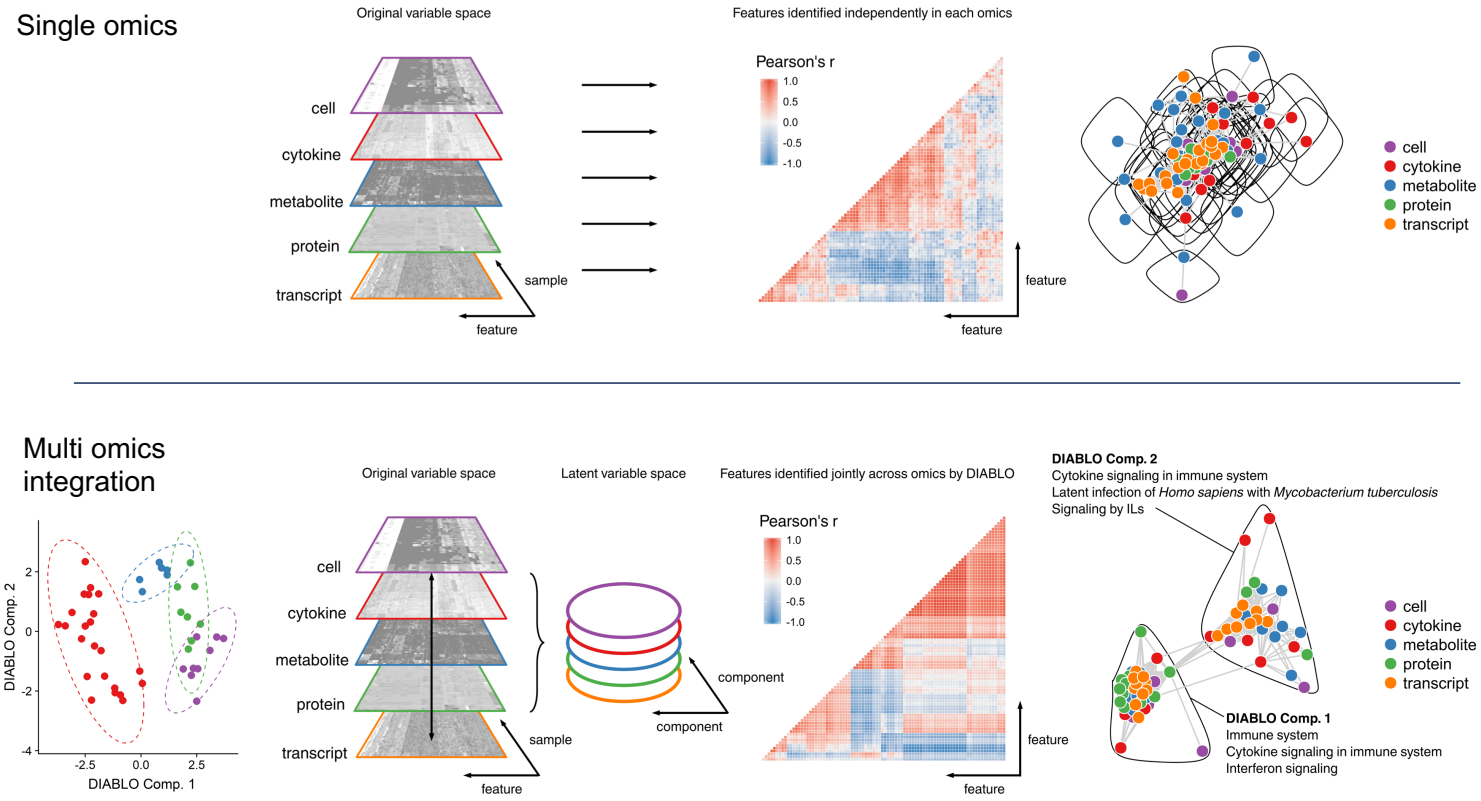
Single omics: too much information!





Dramatic developmental changes emerge when comparing later days of life to D0 but the correlation structure is a hairball mess!

Multi-omics integration: better correlation structure



New biological insights not revealed by single omics analysis (e.g. prostaglandin-endoperoxide synthase 2) and pathways common to all data types (interferon, neutrophil degranulation pathways, complement cascade).

Take-home messages

Multivariate matrix decomposition methods we presented

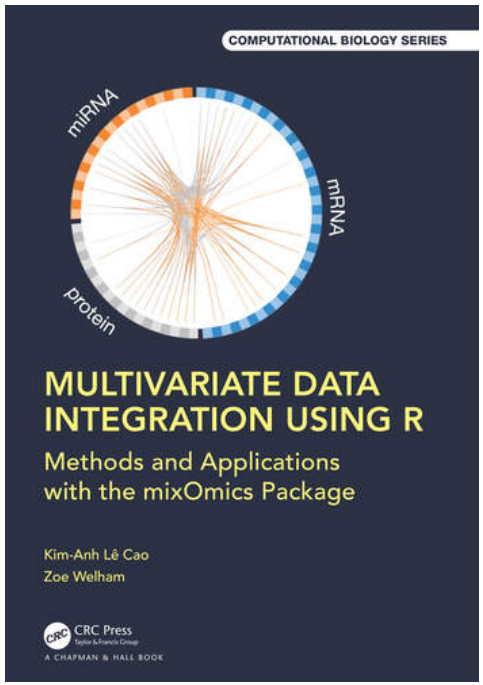
- provide insights into sources of variation in the data (PCA)
- Enable classification, prediction and variable selection (PLS-DA, DIABLO)
- Provide insighful visualisation via dimension reduction
- Help generate new biological hypotheses

Univariate and multivariate approaches are complementary!

→ www.mixomics.org | www.lecao-lab.science.unimelb.edu.au ←

How to go further with mixOmics after this session?

- Install `mixOmics` from Bioconductor
- Vignette + extended tutorials and videos on mixOmics.org
- Ask us questions on our discourse forum mixomics-users.discourse.group



Online full workshop (registration fees apply):
Oct - Nov 2024, mixOmics.org (advertised soon)