Multivariate projection-based methods for data exploration and data integration

Kim-Anh Lê Cao

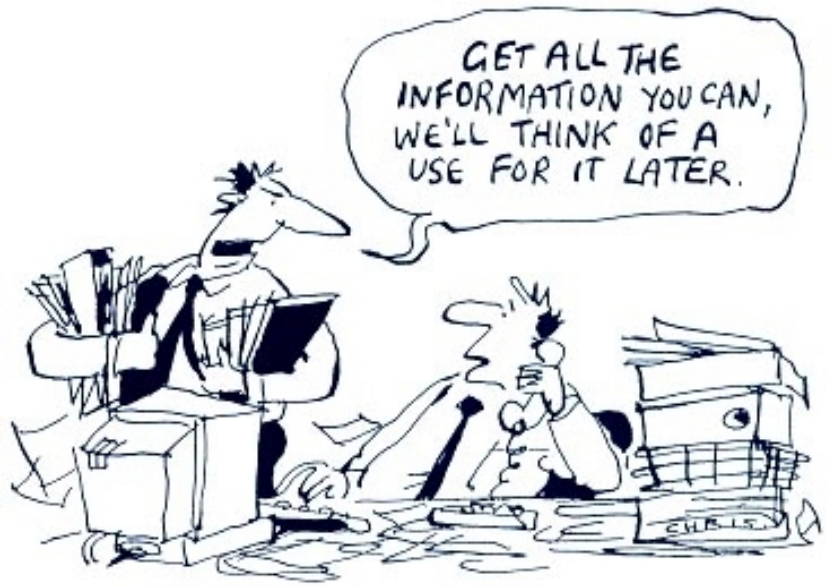
NHMRC Career Development Fellow
Centre for Systems Genomics
School of Mathematics and Statistics

THE UNIVERSITY OF MELBOURNE

Slides: www.mixOmics.org/presentations

Outline

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



- Large quantity of data from multiple and heterogeneous sources
- Computational issues
- Biological interpretation for validation
- Keep pace with new technologies
- Multi-disciplinary field

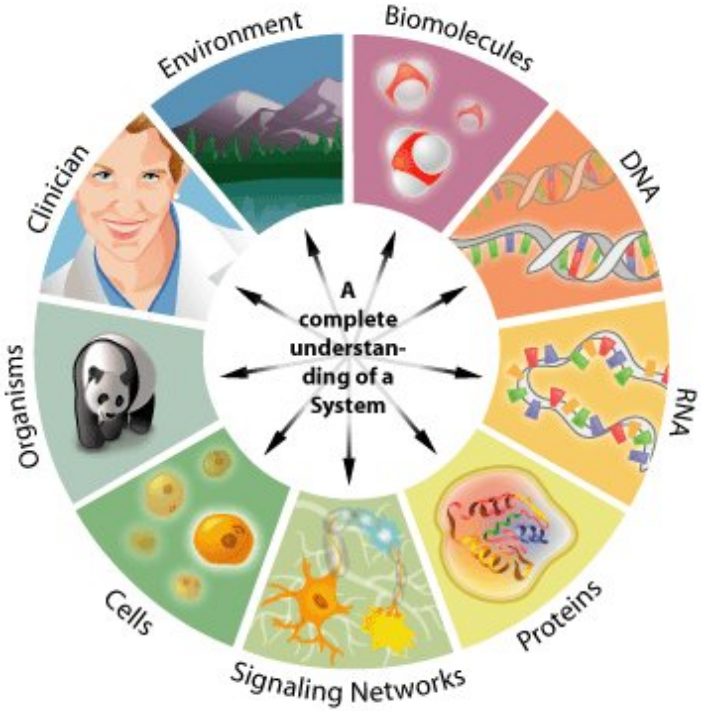
Systems biology and challenges

Reductionism:
1 gene = 1 hypothesis = 1 statistical test

$\Downarrow$

Holism:
Thousands of molecules = ??

- New ways of thinking
- Well stated biological question is crucial
- Need cutting-edge methods to harness the potential of biological data



$\rightsquigarrow$ postulate novel biological hypotheses to be validated in the lab

Research aim and hypothesis

Shift the univariate statistics paradigm to obtain deeper insight into biological systems

> Molecular entities act together to trigger cells' responses and need to be appropriately identified with novel multivariate statistical methods.

Research aim and hypothesis

Shift the univariate statistics paradigm to obtain deeper insight into biological systems

Molecular entities act together to trigger cells' responses and need to be appropriately identified with novel multivariate statistical methods.

"Essentially, all models are wrong, but some are useful."

George EP. Box, statistician
1919 - 2013

Outline

Linear multivariate methods

Linear multivariate methods use components that aggregate observable *variables* (e.g. genes, transcripts, proteins) in a model to summarise sources of variation in the data.

- Reduce data dimension
- Handle highly correlated, noisy, missing variables
- Capture experimental and biological variation

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

A multivariate model

Assign a **score** to each individual, based on a linear combination of variables (e.g. genes):

$$\text{score of Bob} = a_1 * \text{gene}_1^{Bob} + a_2 * \text{gene}_2^{Bob} + \ldots + a_P * \text{gene}_P^{Bob}$$

A multivariate model

Assign a **score** to each individual, based on a linear combination of variables (e.g. genes):

$$\begin{aligned}
\text{score of Bob} &= a_1 * \text{gene}_1^{Bob} + a_2 * \text{gene}_2^{Bob} + \ldots + a_P * \text{gene}_P^{Bob} \\
\text{score of Jane} &= a_1 * \text{gene}_1^{Jane} + a_2 * \text{gene}_2^{Jane} + \ldots + a_P * \text{gene}_P^{Jane} \\
\vdots &= \vdots
\end{aligned}$$

A multivariate model

Assign a **score** to each individual, based on a linear combination of variables (e.g. genes):

$$\begin{aligned} \text{score of Bob} &= a_1 * \text{gene}_1^{Bob} + a_2 * \text{gene}_2^{Bob} + \ldots + a_P * \text{gene}_P^{Bob} \\ \text{score of Jane} &= a_1 * \text{gene}_1^{Jane} + a_2 * \text{gene}_2^{Jane} + \ldots + a_P * \text{gene}_P^{Jane} \\ \vdots &= \vdots \end{aligned}$$

- **What we know**: the gene expression levels value for each individual
- **What we don't know**: the weights $a_1, a_2, \ldots, a_P$ assigned the genes (the same weight a_j is assigned to the same gene j across the whole cohort)

$\rightarrow$ statistical optimisation procedure

Data dimension reduction via components

The expression levels of P genes for each individual are summarised into one score value:

$$\text{score of Bob} = a_1 * \text{gene}_1^{Bob} + a_2 * \text{gene}_2^{Bob} + \ldots + a_P * \text{gene}_P^{Bob}$$

Data dimension reduction via components

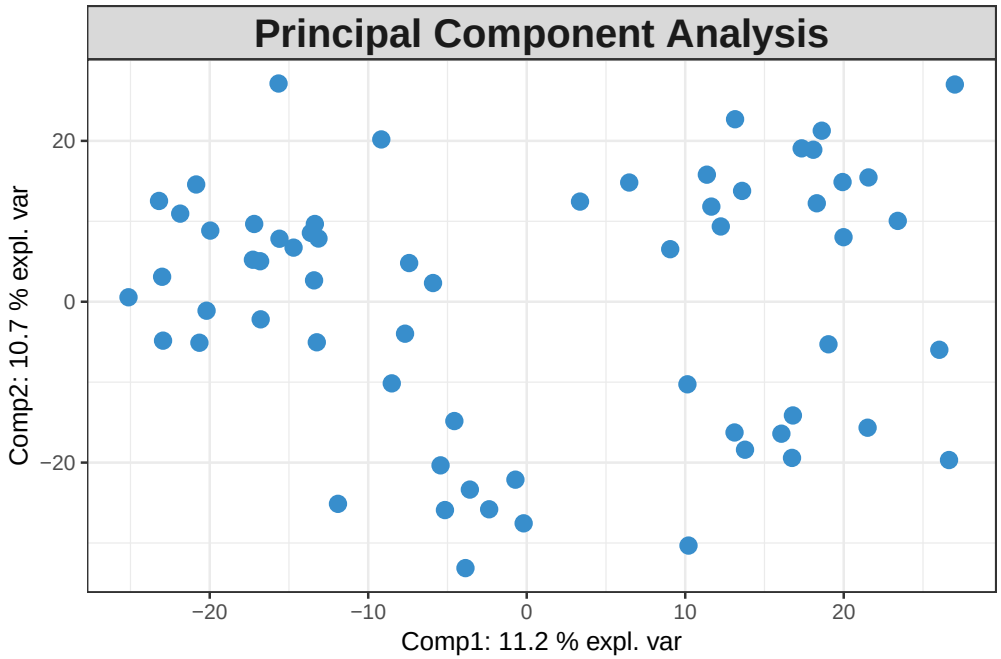
The expression levels of P genes for each individual are summarised into one score value:

$$\text{score of Bob} = a_1 * \text{gene}_1^{Bob} + a_2 * \text{gene}_2^{Bob} + \ldots + a_P * \text{gene}_P^{Bob}$$

- Each score value corresponds to a component score value
- As one component may not be enough to summarise the data we sometimes use several components (several linear combinations)
- As we summarise P data points into a few component scores, we project the data into a smaller subspace ('multivariate projection-based' methods)

Components for data visualisation

Principal Component Analysis: visualisation of 63 samples x 2,300 genes

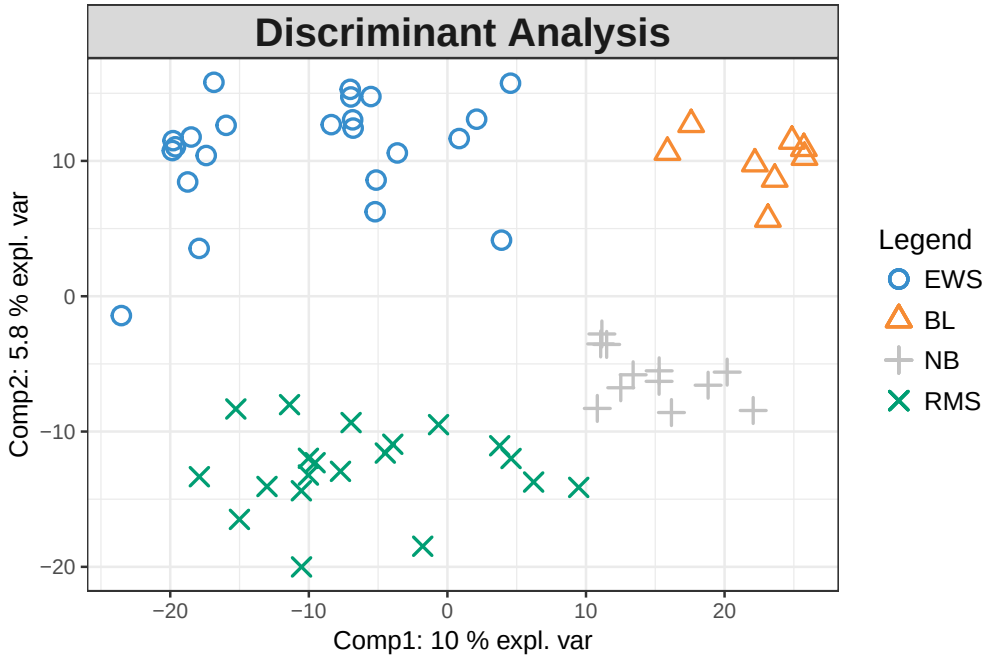


Unsupervised (exploratory) analysis: samples with similar gene expression values cluster together

→ no apriori on the samples group membership

Components for data visualisation

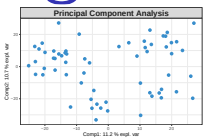
Discriminant Analysis: visualisation of 63 samples x 2,300 genes according to their tumour subtype



Supervised analysis: samples cluster according to their group

$\rightarrow$ the aim is to separate sample groups

Different types of methods for different biological questions

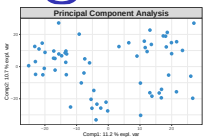


Unsupervised analysis

individuals' scores $\simeq$ $a_1 * \text{gene}_1 + a_2 * \text{gene}_2 + \cdots + a_P * \text{gene}_P$

PCA maximises the variance of each component

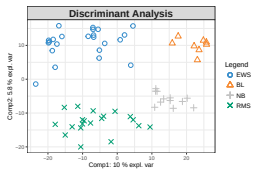
Different types of methods for different biological questions



Unsupervised analysis

individuals' scores $\simeq a_1 * \text{gene}_1 + a_2 * \text{gene}_2 + \cdots + a_P * \text{gene}_P$

PCA maximises the variance of each component

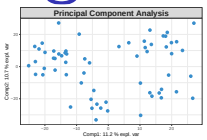


Supervised analysis

phenotype category $\simeq a_1 * \text{gene}_1 + a_2 * \text{gene}_2 + \cdots + a_P * \text{gene}_P$

PLS Discriminant analysis (PLS-DA) maximises the covariance between each component and the phenotype

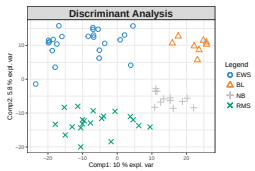
Different types of methods for different biological questions



Unsupervised analysis

individuals' scores $\simeq a_1 * \text{gene}_1 + a_2 * \text{gene}_2 + \cdots + a_P * \text{gene}_P$

PCA maximises the variance of each component



Supervised analysis

phenotype category $\simeq a_1 * \text{gene}_1 + a_2 * \text{gene}_2 + \cdots + a_P * \text{gene}_P$

PLS Discriminant analysis (PLS-DA) maximises the covariance between each component and the phenotype

Sparse method

phenotype category $\simeq 0 * \cancel{\text{gene}}_1 + 0 * \cancel{\text{gene}}_2 + \cdots + a_P * \text{gene}_P$

$\rightsquigarrow$ only a few genes are included in model and are selected

Multivariate methods are highly flexible

	One 'omics data set e.g. transcriptomics only	**Multiple 'omics data sets** e.g. integrate transcriptomics, proteomics, miRNA
	Data exploration, data mining, & visualisation	
Unsupervised (No response provided)	Select a subset of correlated variables	
Supervised classification (Categorical response)	Select a subset of discriminative variables	

→ My main research focus is to mine data and to identify a subset of diagnostic or prognostic biomarkers that are assessed in combination.

Univariate and multivariate statistics are complementary

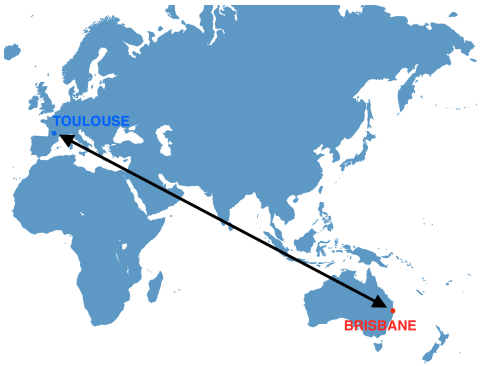
Univariate (e.g. T-test/linear model)	**Multivariate (e.g. PLS-DA)**
• Data distribution assumption	• No data distribution required
• Test one biomarker at a time	• Assess all biomarkers in combination
• Output a p-value per biomarker	• No p-value, unless permutations
• No visualisation	• Visualisation (samples, variables) based on the components
• Adjust for covariates	• Does not adjust for covariates (in progress)
• No prediction if $P > N$ (linear model)	• Prediction of a phenoype for a new sample

Sparse multivariate methods identify a small subset of biomarker candidates to suggest new biological hypotheses and make downstream univariate analysis amenable.

Open-source and user-friendly tools

The mixOmics R toolkit (since 2009)

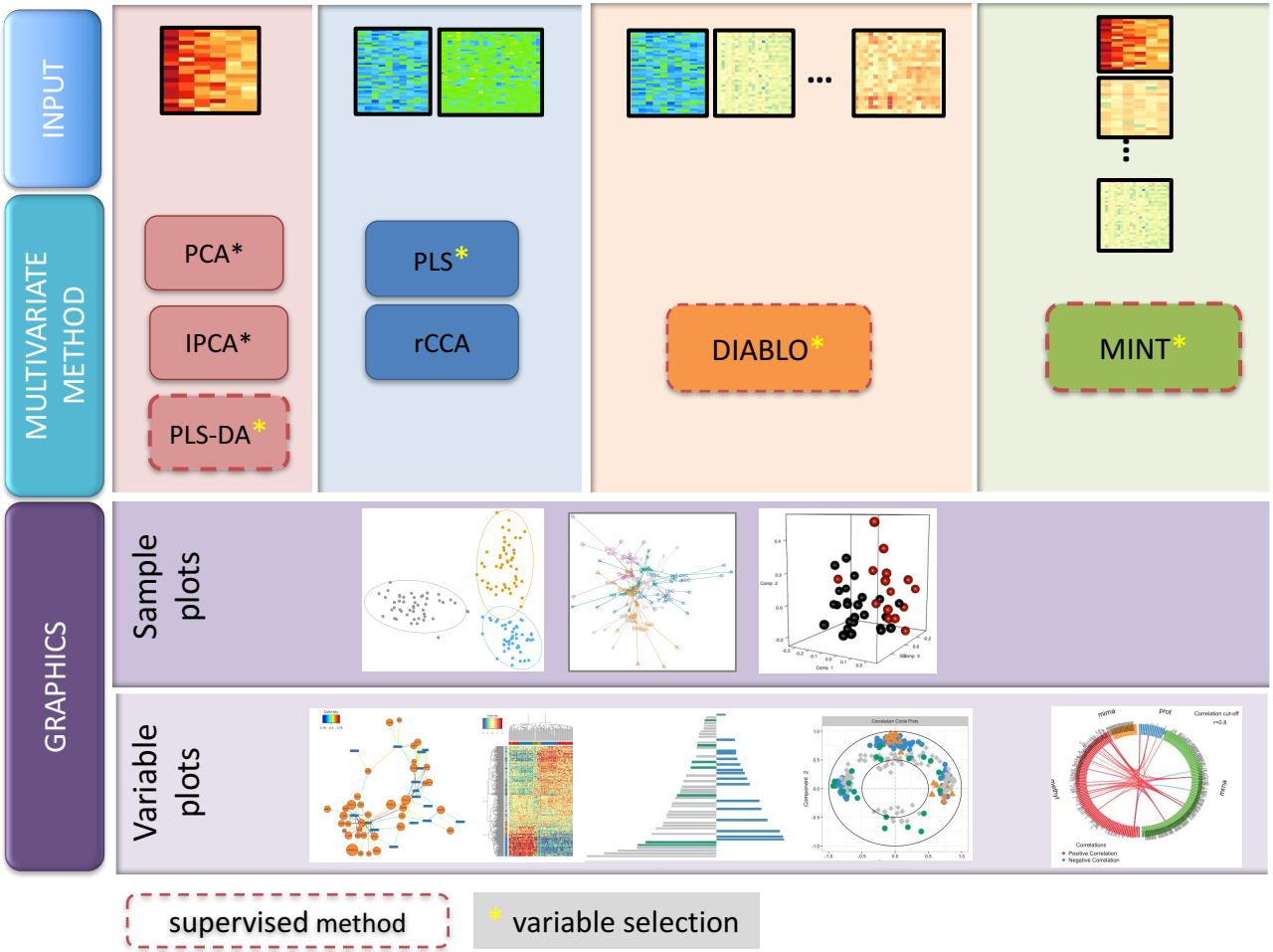
- French'Oz team:
  4 core, 1 developer, students, collaborators
- Today: 17 novel multivariate methods
- 21K downloads in 2016
- 14 multi-day workshops since 2014 (FR, AUS, NZ)



Our **research program** focuses on the development of multivariate statistical methodologies, their applications in areas informed by biology, and the training of the new generation of computational biologists.

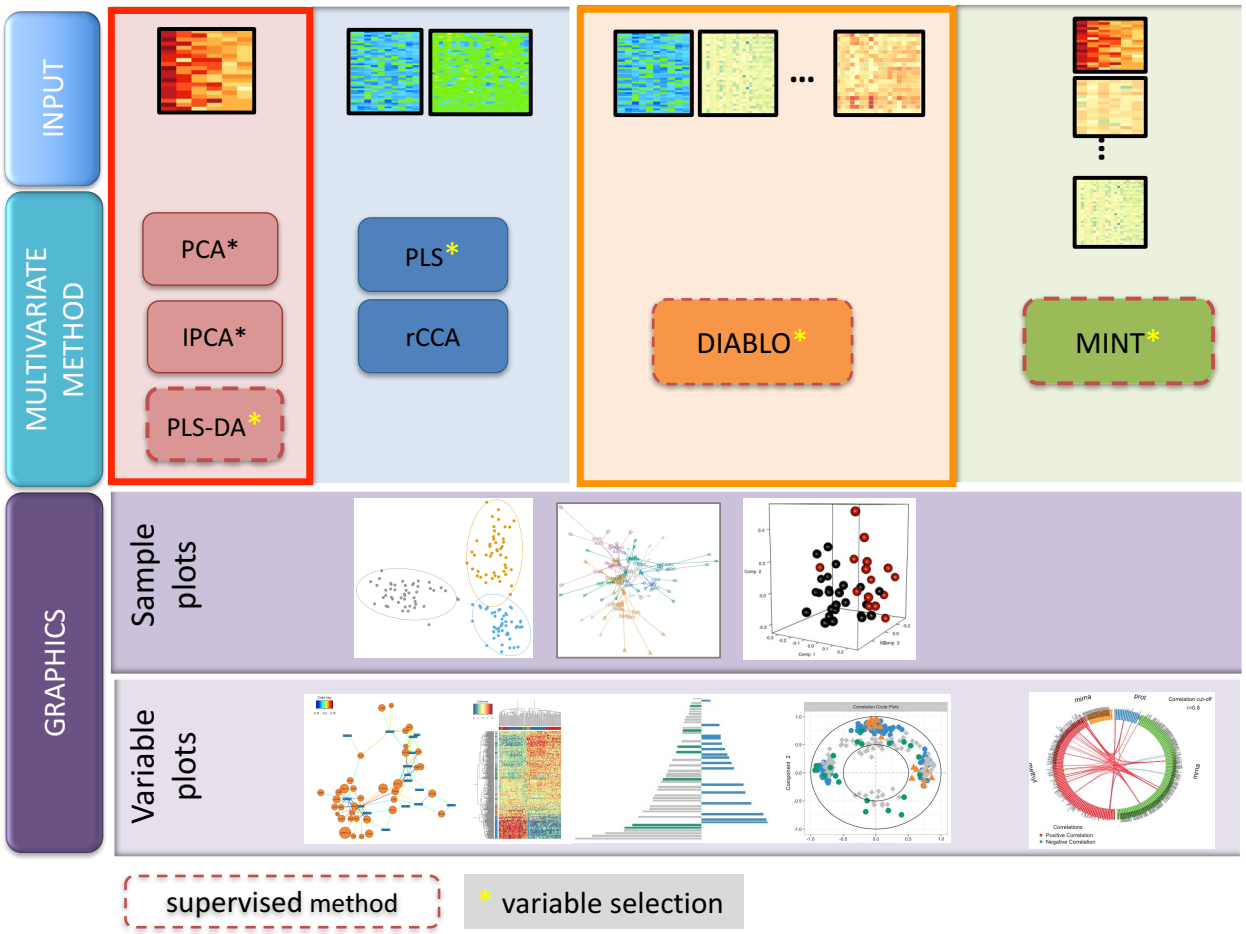
www.mixOmics.org

'omics data analysis with mixOmics



⇝ Visualisations based on components and selected biomarkers

'omics data analysis with mixOmics



single 'omics with

PCA:

better understanding of the major sources of variation in the data

sparse PLSDA:

identify a molecular signature to predict the response

multi 'omics integration with DIABLO:

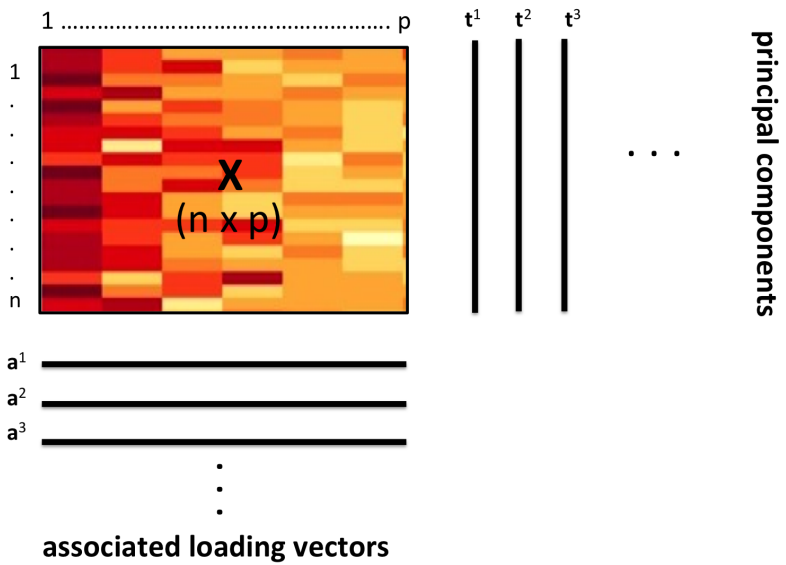
identify a multi 'omics signature

Multivariate projection-based methods (take-home message)

- Assess variables in combination
- Reduce data dimension via components (linear combinations of original variables)
- Data visualisation based on the components
- Flexible models, well-suited for 'difficult' data and first steps for biomarker discovery
- Complement univariate statistical models

Outline

PCA is a matrix decomposition technique



Solved with Singular Value Decomposition :

$$X = U\Delta A^T$$

Singular vectors:

- $T = U\Delta$, T contains the PCs $\boldsymbol{t}^h$
- A contains the loading vectors $\boldsymbol{a}^h$

Singular values:

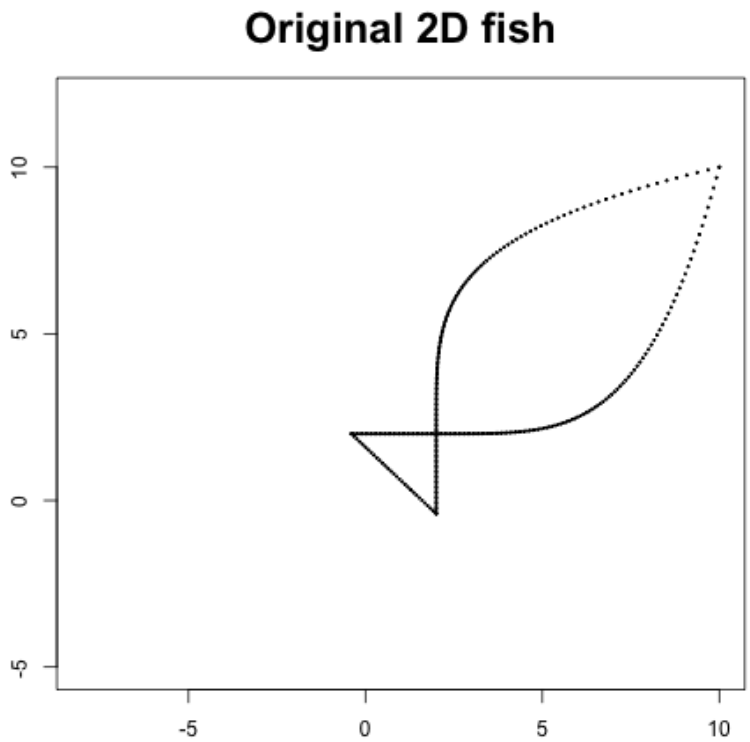
- Δ diagonal matrix with $\sqrt{\delta_h}$

$h = 1..H$ is the number of PCs

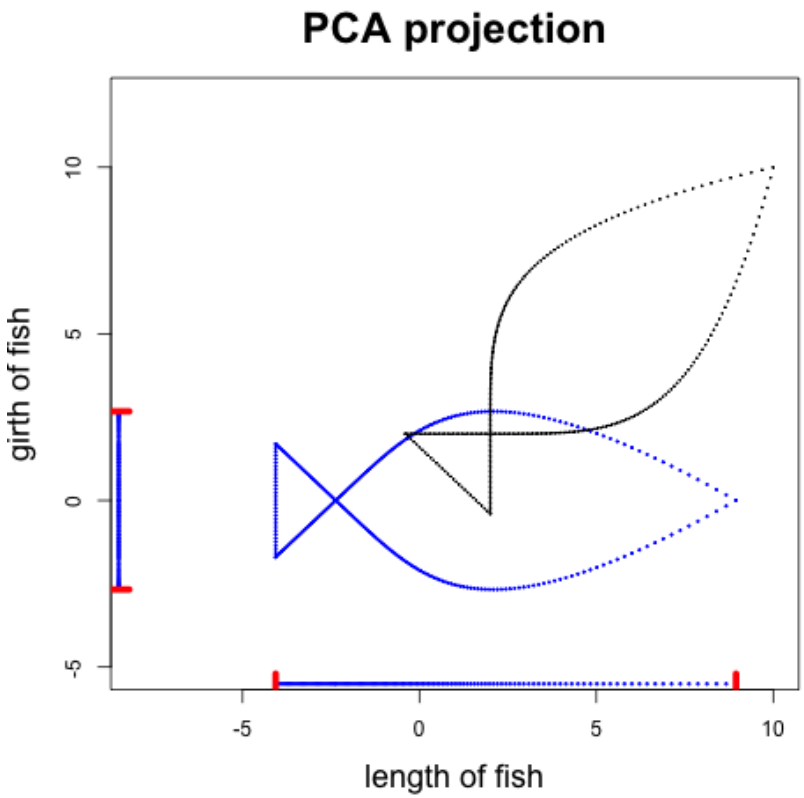
The variance of the first principal component $\boldsymbol{t}^1$ is the largest $(= \delta_1)$.

The eigenvalues δ_h decrease and correspond to the explained variance per component.

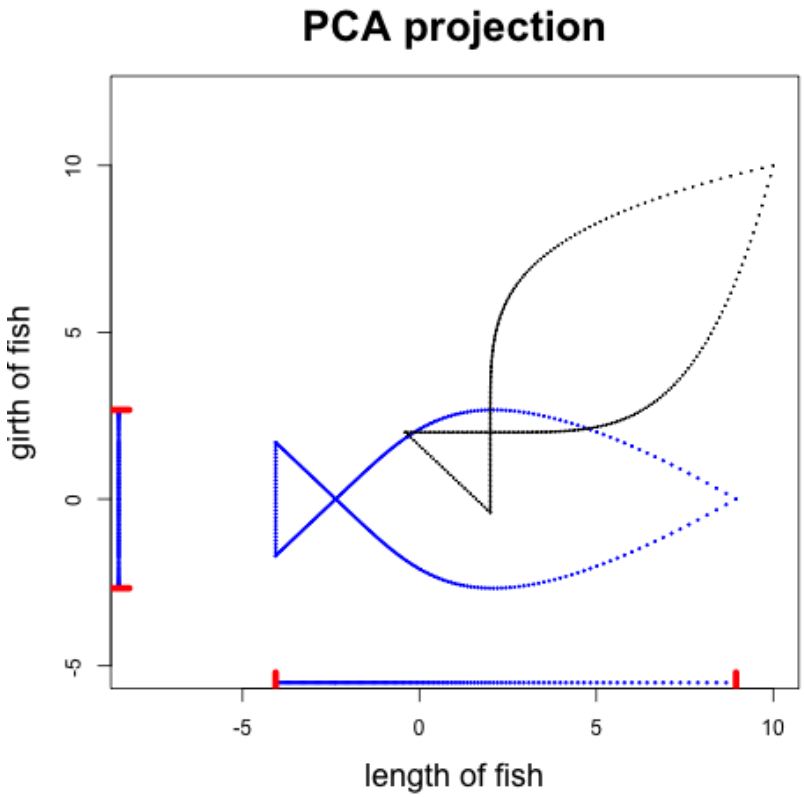
Summarize a fish



Summarize a fish



Summarize a fish



PCA:

- First component maximises the variance
- Second component maximises the remaining variance

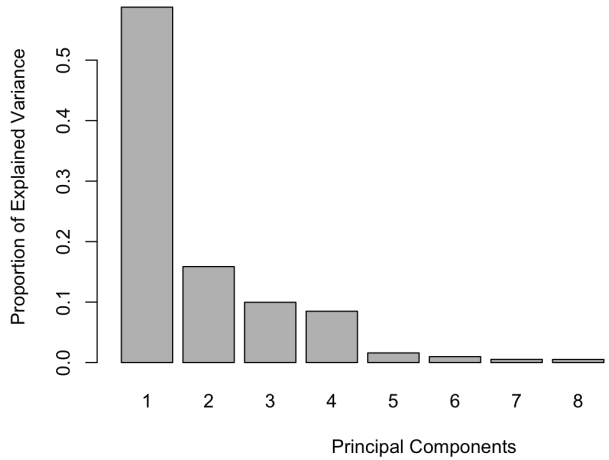
Acknowledgements: B Gautier for plotting a fish in R!

Parameters in PCA

How many principal components to summarize most of the information?

There are as many components as the rank of the matrix X

- Screeplot of eigenvalues: any elbow?
- Sample plot: makes sense?
- Cumulative proportion of explained variance
- Some stat tests to estimate the 'intrinsic' dimension, but limitations when $n << p$



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

PCA is a visualisation tool

- Sample plot plots two components to visualise similarities between samples
- Correlation circle plots* to visualise the cross-correlations between variables
- Biplot to relate samples and variables in the same plot.

*More details about CC plots: González, I., Lê Cao, K.-A., et al. (2012). Visualising associations between paired 'omics data sets. BioData mining, 5(1), 19.

Example: yeast metabolomics

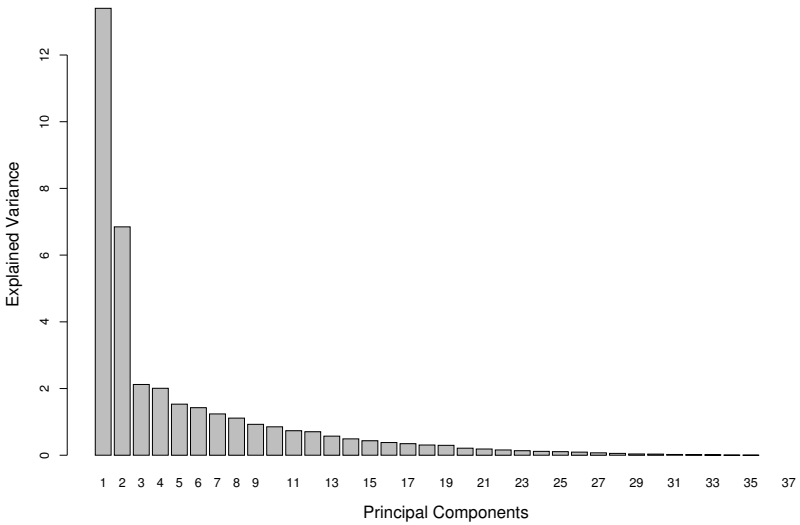
Yeast study from Villas-Boâs et al, 2005:

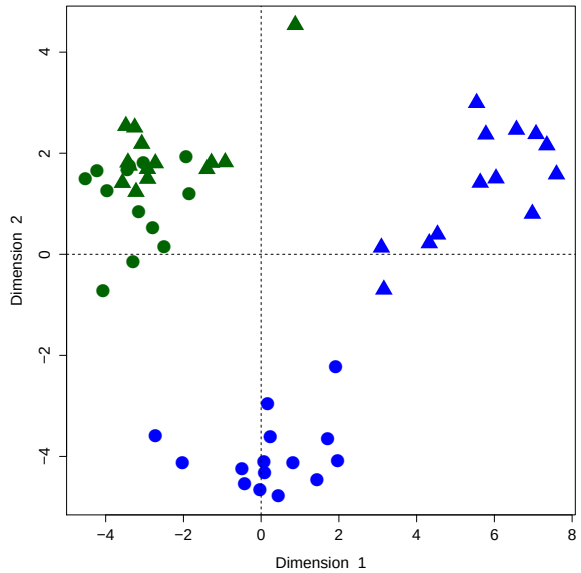
- Two Saccharomyces Cerevisiae strains: WT and MT
- Two environmental conditions: aerobic (AER), anaerobic (ANA)
- 37 metabolites and 55 samples
  (13 MT-AER, 14 MT-ANA, 15 WT-AER, 13 WT-ANA)

Question:

- What is the strongest source of variation in the metabolomics data: strain or environment? something else?

Number of components and sample plot



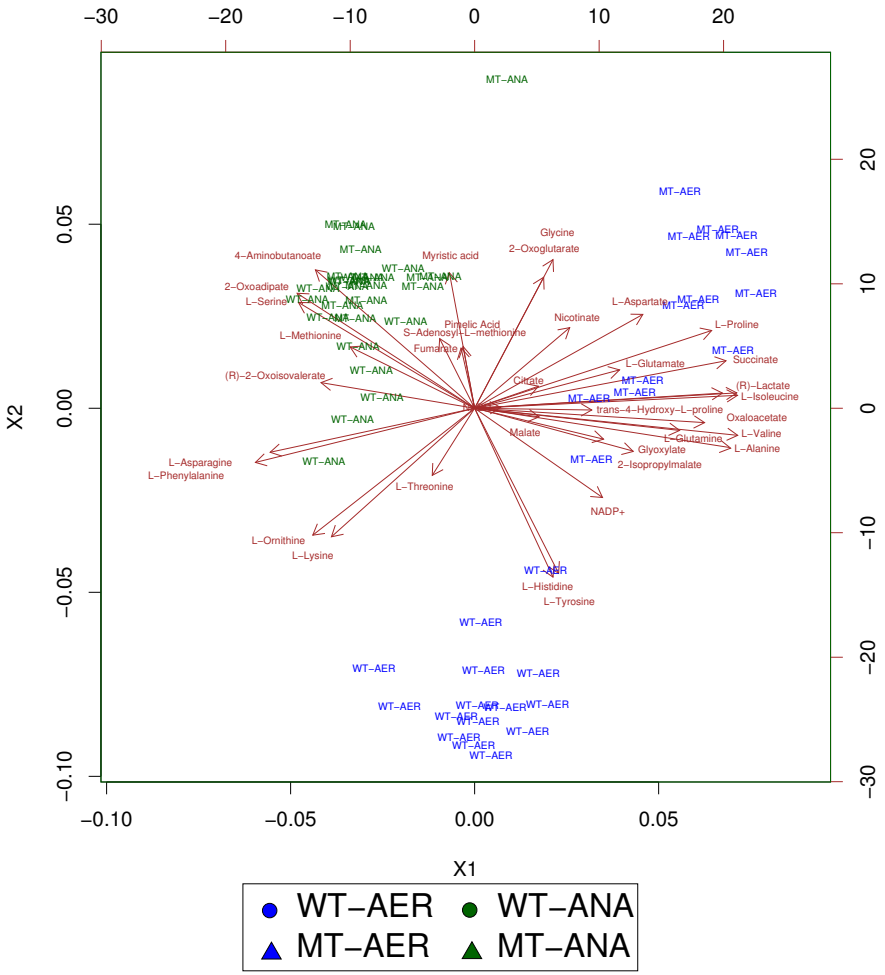


Cumulative % of explained variance:

- Two PCs: 54.72% of total var
- Three PCs: 60.45% of total var

Strong separation of environment conditions on the first 2 PCs

Biplot: variables and samples plot



- Correlations (pos and neg) between variables (cosine angle btw arrows)
- Contribution of each variable to each component / axis
- 'Relationship' btw samples and variables

PCA summary

- PCA is a matrix decomposition technique for dimension reduction.
- Perform a PCA first to understand the major sources of variation in your data.
- Always report the % explained variance per component.
- PCA can highlight 'batch effect' in the data and can be used to check that batch-effect removal techniques are efficient.

When PCA may fail

- When the biological question may not be related to the highest variance in the data
  → Independent Component Analysis (ICA) or variants.
- When there are too many noisy variables that contribute to the variance
  → sparse PCA but need to specify # variables to select.
- When there are too many missing values
  → Algorithm variant NIPALS
- When samples are **not** independent (e.g. time course data, repeated measures) do **not** use PCA as subject variation $>$ the time variation
  → multilevel approach for multivariate analysis

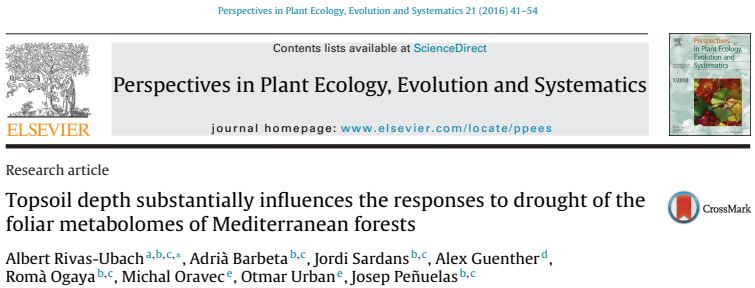
Outline

Single 'omics supervised analysis with sPLS-DA:
- identify biomarkers to discriminate phenotype groups
- predict the phenotype of an external sample

Input: an expression data and a phenotype group

Applications: any single 'omics data, in particular proteomics, metabolomics data, and microbiome

Some examples for 2 data sets integration:


Perspectives in Plant Ecology, Evolution and Systematics 21 (2016) 41–54

Contents lists available at ScienceDirect

Perspectives in Plant Ecology, Evolution and Systematics

journal homepage: www.elsevier.com/locate/ppees

ELSEVIER

Research article

Topsoil depth substantially influences the responses to drought of the foliar metabolomes of Mediterranean forests

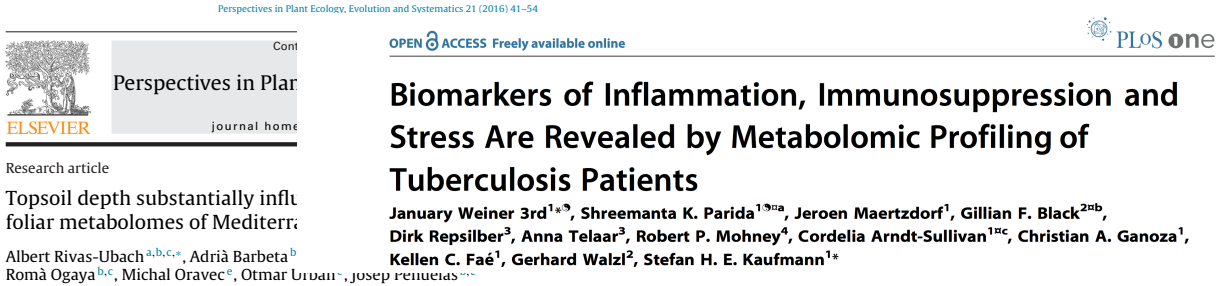
Albert Rivas-Ubach[a,b,c,*], Adrià Barbeta[b,c], Jordi Sardans[b,c], Alex Guenther[d], Romà Ogaya[b,c], Michal Oravec[e], Otmar Urban[e], Josep Peñuelas[b,c]

Single 'omics supervised analysis with sPLS-DA:
- identify biomarkers to discriminate phenotype groups
- predict the phenotype of an external sample

Input: an expression data and a phenotype group

Applications: any single 'omics data, in particular proteomics, metabolomics data, and microbiome

Some examples for 2 data sets integration:


Perspectives in Plant Ecology, Evolution and Systematics 21 (2016) 41–54

Perspectives in Plar

journal home

ELSEVIER

Research article

Topsoil depth substantially influ
foliar metabolomes of Mediterra

Albert Rivas-Ubach[a,b,c,*], Adrià Barbeta[b]
Romà Ogaya[b,c], Michal Oravec[e], Otmar Urban[~], Josep Peñuelas[~,~]

OPEN ACCESS Freely available online

PLoS one

Biomarkers of Inflammation, Immunosuppression and Stress Are Revealed by Metabolomic Profiling of Tuberculosis Patients

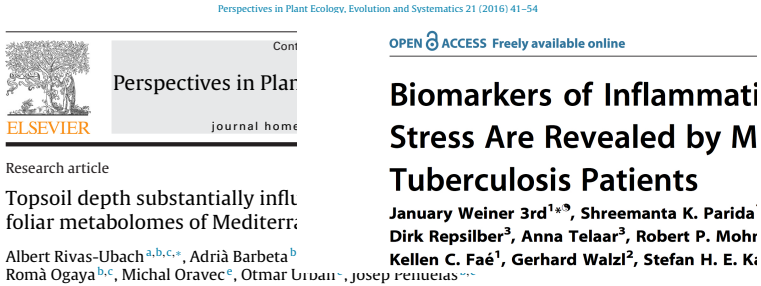
January Weiner 3rd[1*¶], Shreemanta K. Parida[1¶¤a], Jeroen Maertzdorf[1], Gillian F. Black[2¤b], Dirk Repsilber[3], Anna Telaar[3], Robert P. Mohney[4], Cordelia Arndt-Sullivan[1¤c], Christian A. Ganoza[1], Kellen C. Faé[1], Gerhard Walzl[2], Stefan H. E. Kaufmann[1*]

Single 'omics supervised analysis with sPLS-DA:
- identify biomarkers to discriminate phenotype groups
- predict the phenotype of an external sample

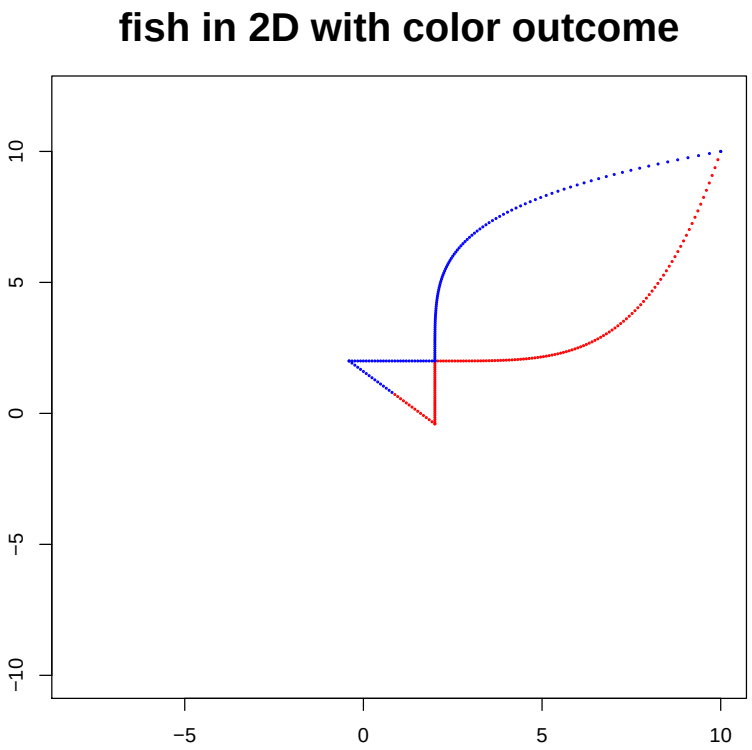
Input: an expression data and a phenotype group

Applications: any single 'omics data, in particular proteomics, metabolomics data, and microbiome

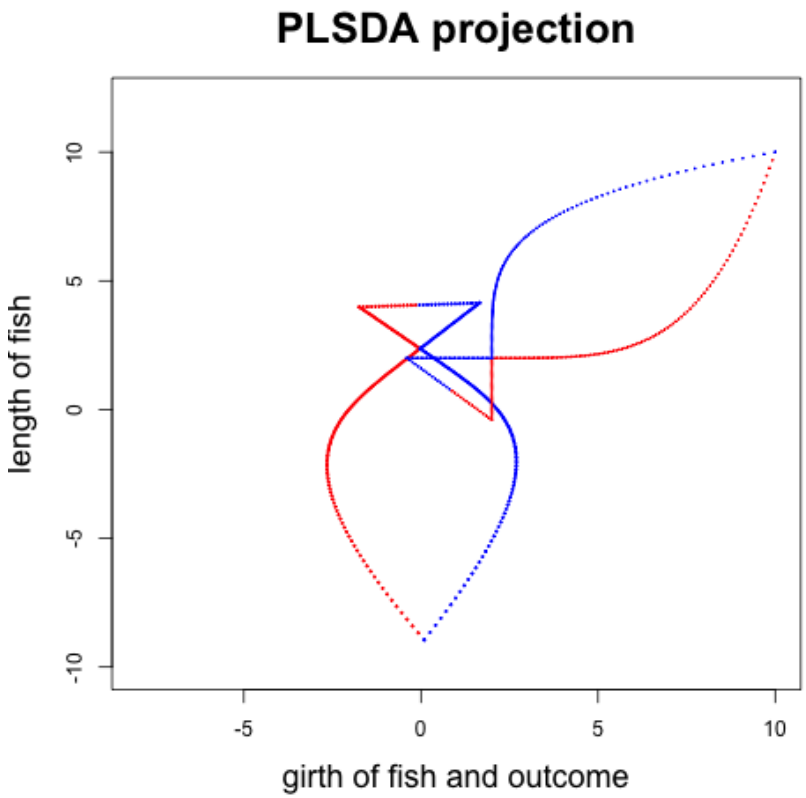
Some examples for 2 data sets integration:



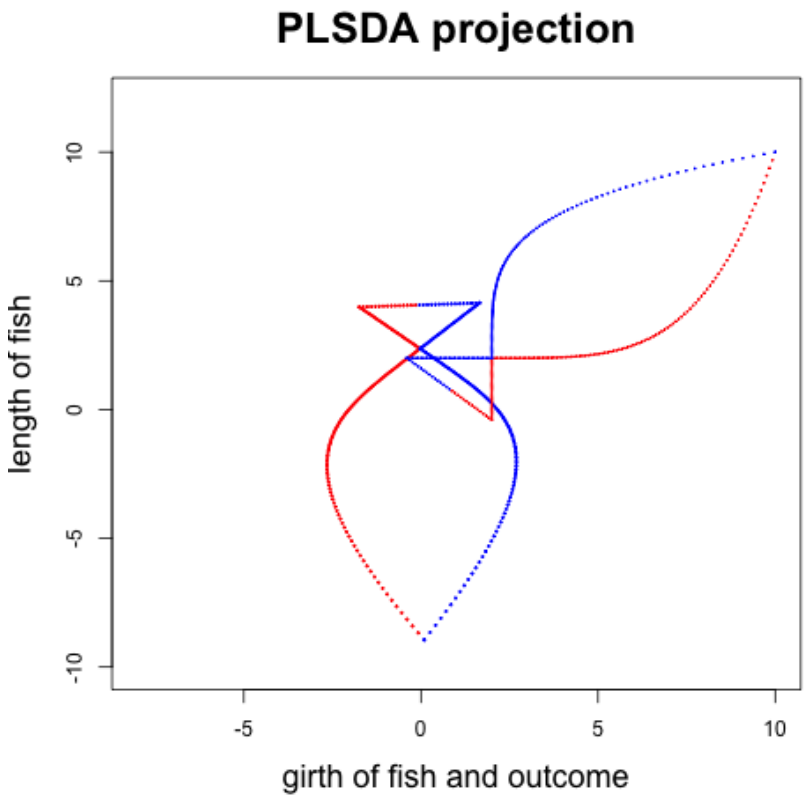
Summarize a fish



Summarize a fish



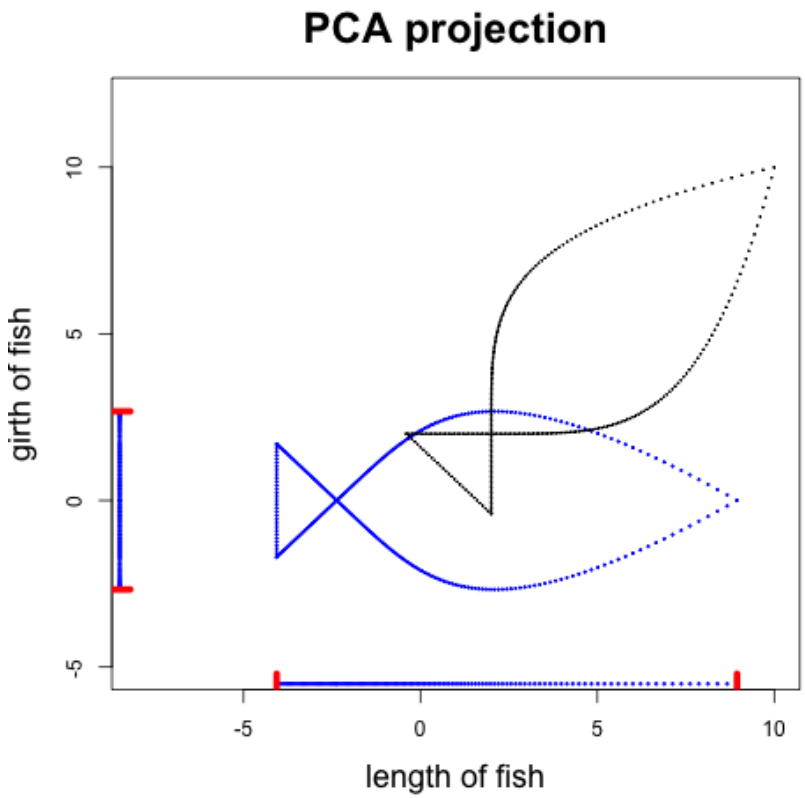
Summarize a fish

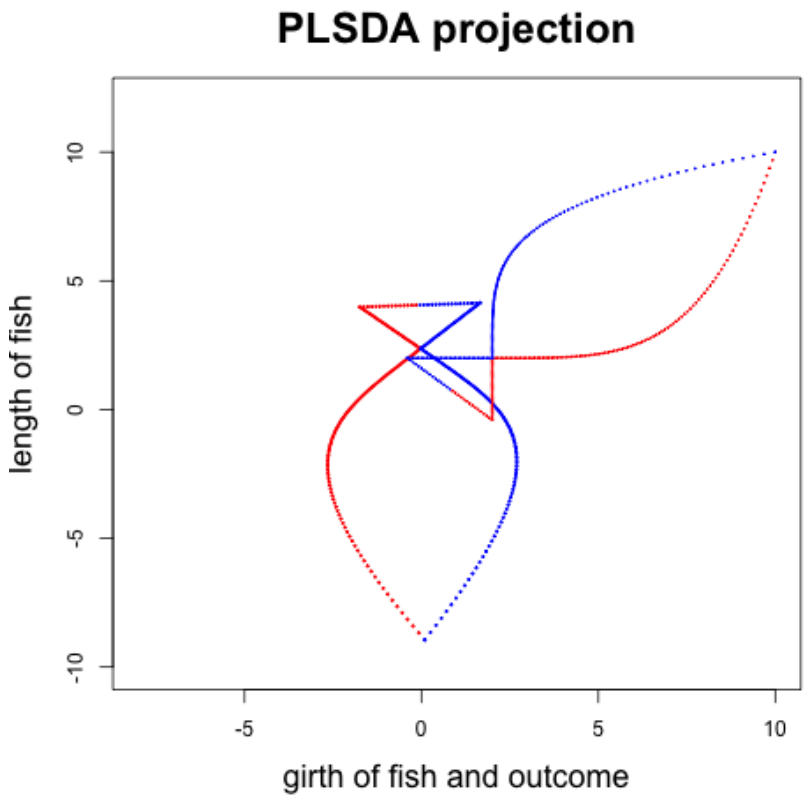


PLSDA:

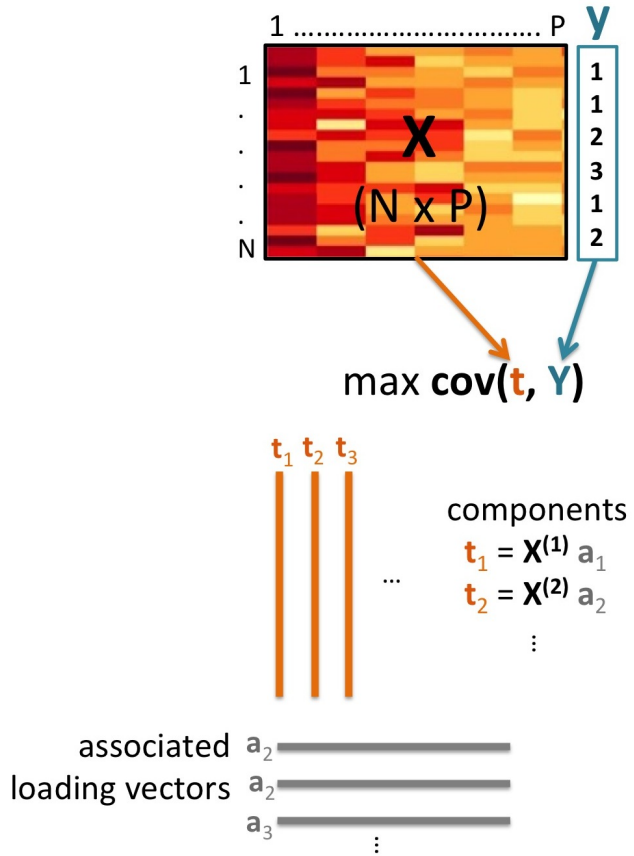
- First component maximises the differences between colors (outcome)
- Second component maximises the variance

Summarize a fish



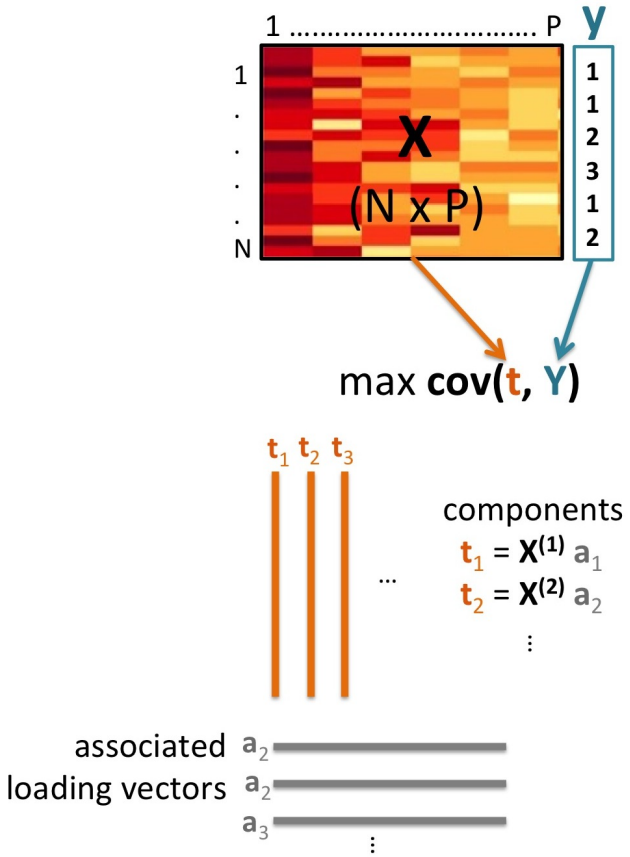


PLS Discriminant Analysis matrix decomposition



PLS-DA maximises the covariance between components and the outcome (phenotype groups)

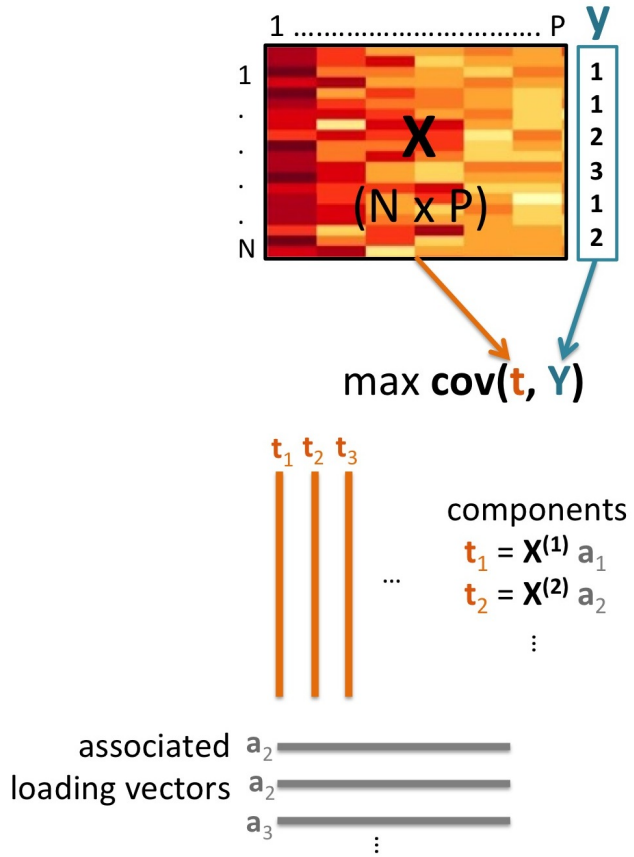
PLS Discriminant Analysis matrix decomposition



PLS-DA maximises the covariance between components and the outcome (phenotype groups)

sparse PLS-DA selects the most discriminative variables (with LASSO penalisations)

PLS Discriminant Analysis matrix decomposition



PLS-DA maximises the covariance between components and the outcome (phenotype groups)

sparse PLS-DA selects the most discriminative variables (with LASSO penalisations)

Predictive model: based on the linear combination of variables, predict the phenotype group of a new sample.

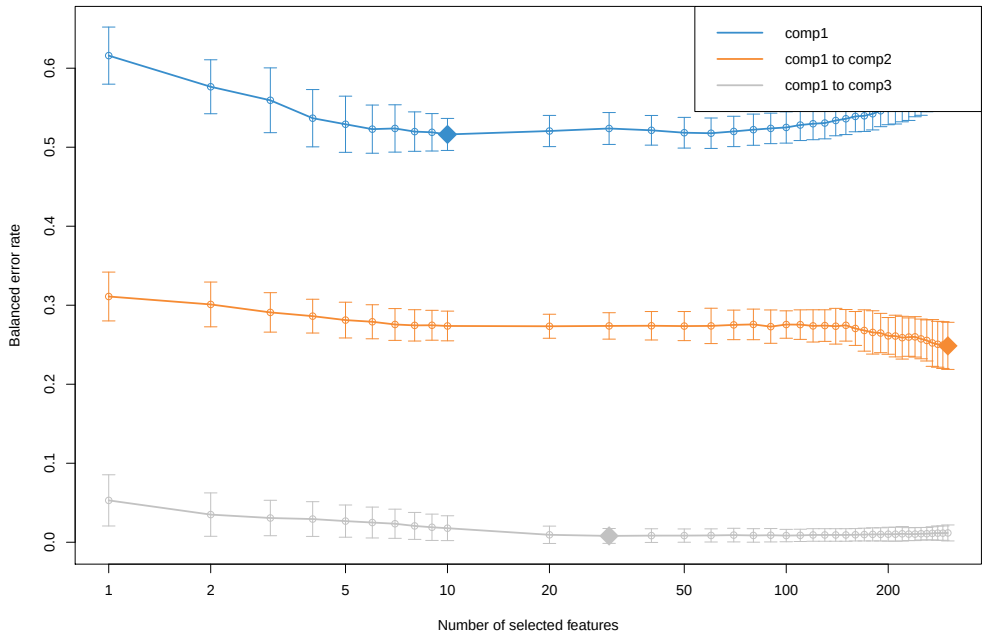
Lê Cao K-A., et al. (2011). Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems, *BMC Bioinformatics*, 12:253.

Parameters in sparse PLS-DA

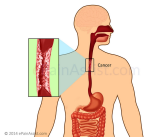
- The number of PLS components
- The number of variables to select on each component

Parameters in sparse PLS-DA

- The number of PLS components
- The number of variables to select on each component



$\rightarrow$ we use repeated cross-validation and choose the parameters that achieve the lowest classification error rate.



Oesophageal cancer proteomics study

with **Alok Shah**, Benoît Gautier & Michelle Hill (UQ)

Proteomics targeted assay (129 proteins) including 20 Barrett's oesophagus benign (BE) or 20 oesophageal (EAC) adenocarcinoma cancer samples.

Aim: develop blood tests for detection and personalised treatment

Statistical challenges:

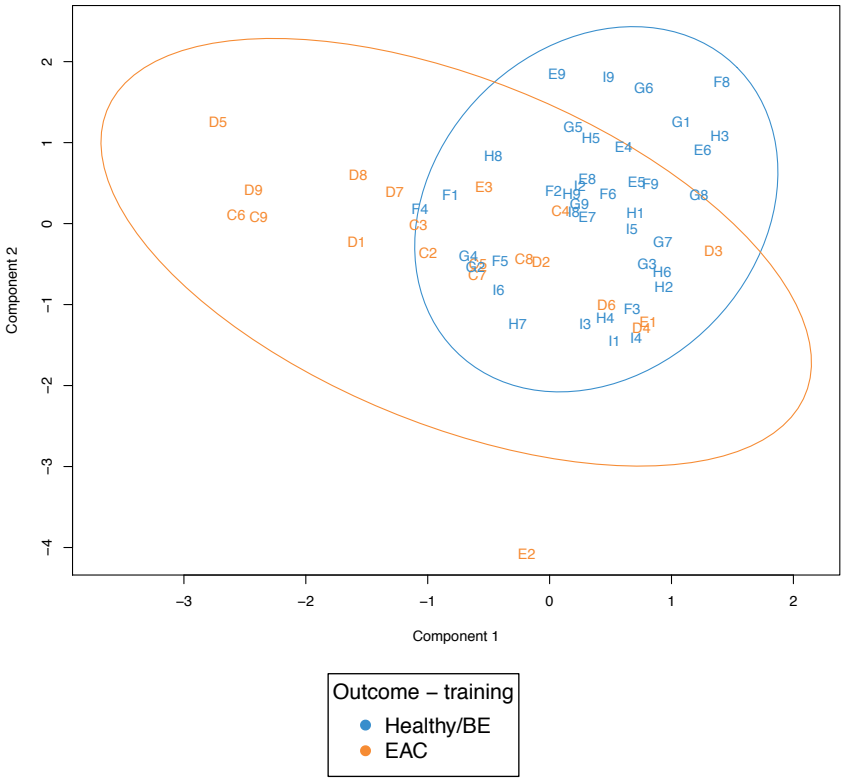
- Small cohort (40 samples)
- Data range and variability with proteomics data
- Classical univariate statistical methods failed

Hill M, Shah AK, Lê Cao K-A (2014). Blood Test for Throat Cancer. WO/2016/077881. Priority 17/11/2015

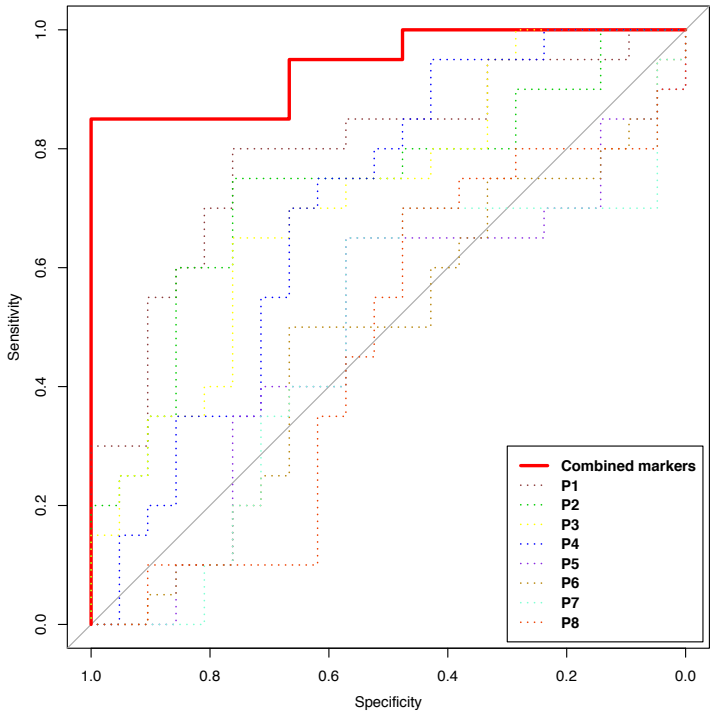
Shah AK, et al. (2015) Serum glycoprotein biomarker discovery and qualification pipeline reveals novel diagnostic biomarkers for oesophageal adenocarcinoma. *Mol Cell Prot 14(11)*.

sPLSDA: a signature of 11 proteins

Discovery cohort:



ROC comparison:

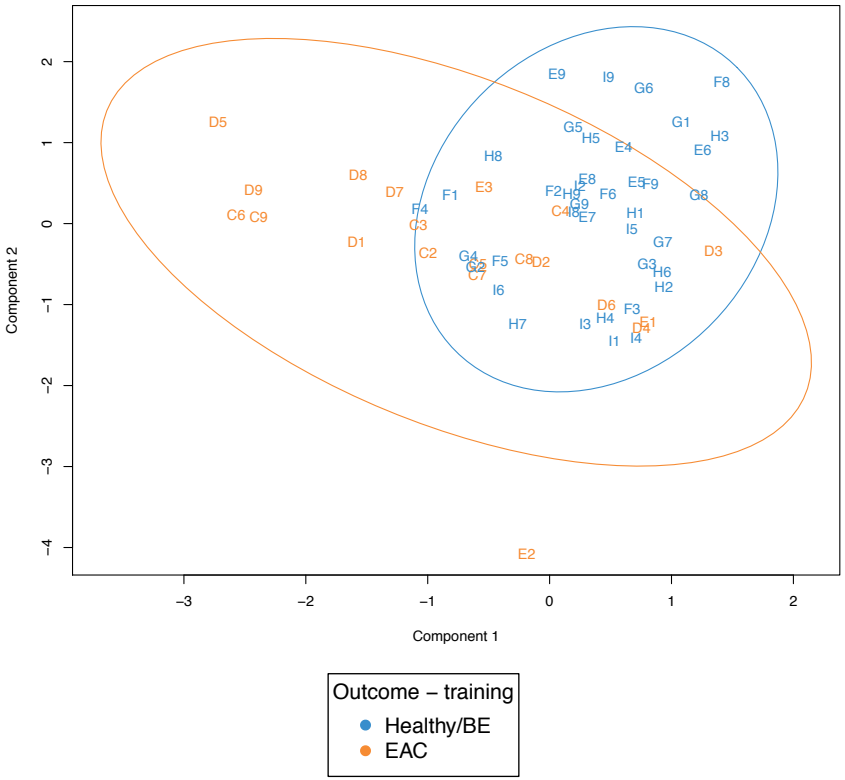


Sample representation based on the 11 selected proteins

AUC = 0.9675 of combined 11 proteins > each indiv. protein

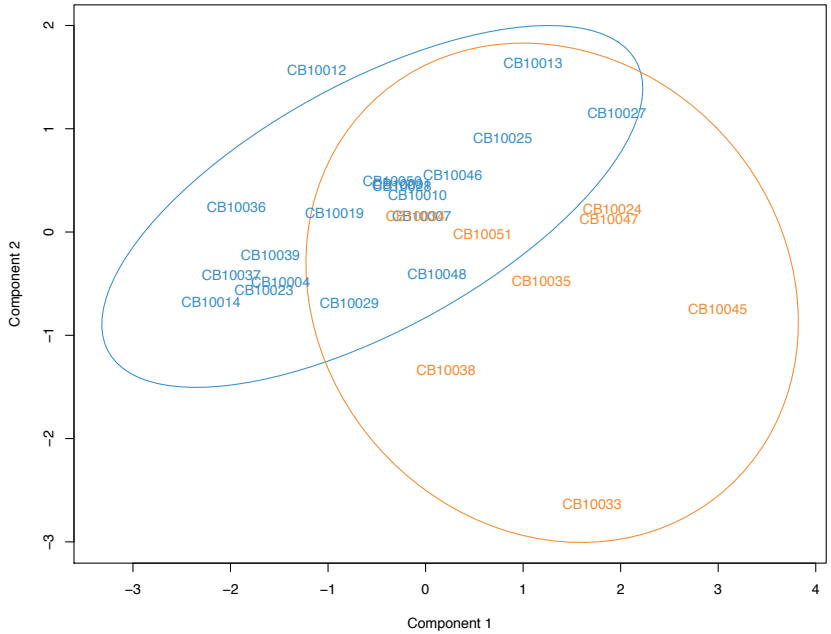
sPLSDA: a signature of 11 proteins

Discovery cohort:



Sample representation based on the 11 selected proteins

Validation cohort:



Sample representation based on the same 11 selected proteins

PLSDA summary

- PLSDA is a supervised method that aims to discriminate sample phenotype groups
- Suitable for biomarker discovery when seeking for a molecular signature
- Results validation in external cohorts or using repeating cross-validation on the training data set.

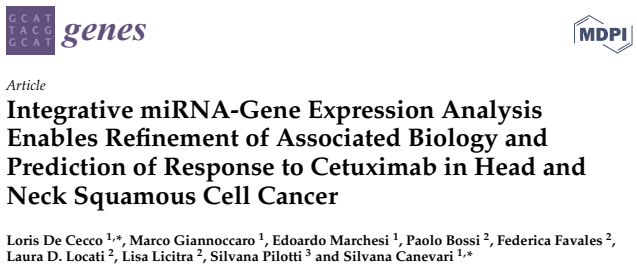
Outline

Multiple 'omics integration and supervised analysis:
- identify a multi 'omics signature that explain a phenotype
- seek for maximal correlation between molecular features of different types to obtain greater biological insights

Input: several 'omics expression datasets matching the same samples, and phenotype/disease status information

Applications: 2 'omics, microbiome-environmental variables integration

Some examples:


genes

MDPI

Article

Integrative miRNA-Gene Expression Analysis Enables Refinement of Associated Biology and Prediction of Response to Cetuximab in Head and Neck Squamous Cell Cancer

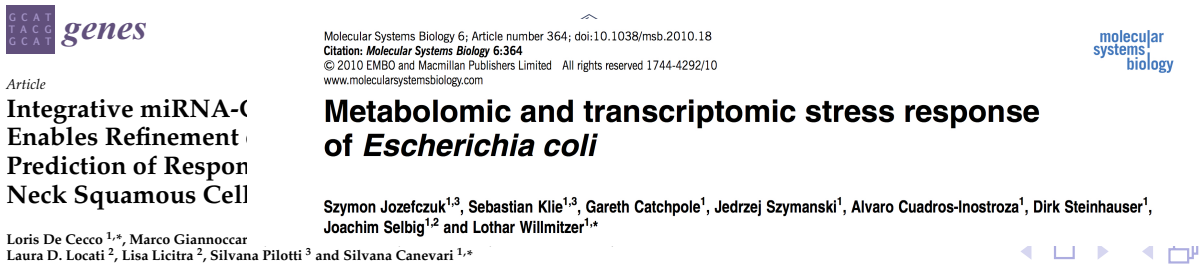
Loris De Cecco [1,*], Marco Giannoccaro [1], Edoardo Marchesi [1], Paolo Bossi [2], Federica Favales [2], Laura D. Locati [2], Lisa Licitra [2], Silvana Pilotti [3] and Silvana Canevari [1,*]

Multiple 'omics integration and supervised analysis:
- identify a multi 'omics signature that explain a phenotype
- seek for maximal correlation between molecular features of different types to obtain greater biological insights

Input: several 'omics expression datasets matching the same samples, and phenotype/disease status information

Applications: 2 'omics, microbiome-environmental variables integration

Some examples:


genes

Article

Integrative miRNA-(
Enables Refinement
Prediction of Respon
Neck Squamous Cell

Loris De Cecco [1,*], Marco Giannoccar
Laura D. Locati [2], Lisa Licitra [2], Silvana Pilotti [3] and Silvana Canevari [1,*]

Molecular Systems Biology 6; Article number 364; doi:10.1038/msb.2010.18
Citation: *Molecular Systems Biology* 6:364
© 2010 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/10
www.molecularsystemsbiology.com

molecular systems biology

Metabolomic and transcriptomic stress response of *Escherichia coli*

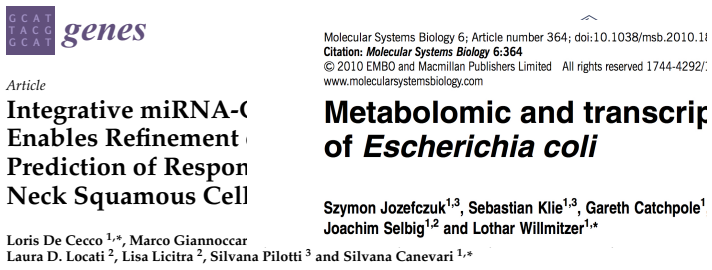
Szymon Jozefczuk[1,3], Sebastian Klie[1,3], Gareth Catchpole[1], Jedrzej Szymanski[1], Alvaro Cuadros-Inostroza[1], Dirk Steinhauser[1], Joachim Selbig[1,2] and Lothar Willmitzer[1,*]

Multiple 'omics integration and supervised analysis:
- identify a multi 'omics signature that explain a phenotype
- seek for maximal correlation between molecular features of different types to obtain greater biological insights

Input: several 'omics expression datasets matching the same samples, and phenotype/disease status information

Applications: 2 'omics, microbiome-environmental variables integration

Some examples:


genes

Article

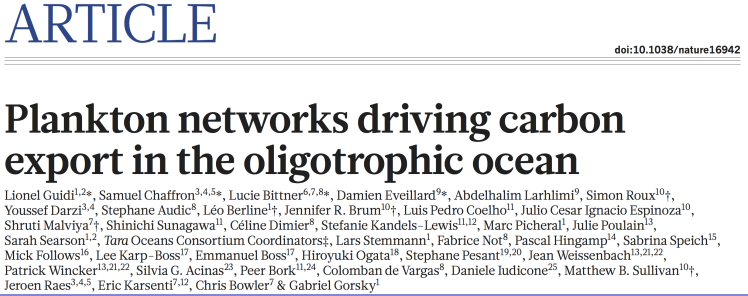
Integrative miRNA-(
Enables Refinement
Prediction of Respon
Neck Squamous Cell

Loris De Cecco [1,*], Marco Giannoccar
Laura D. Locati [2], Lisa Licitra [2], Silvana Pilotti [3] and Silvana Canevari [1,*]

Molecular Systems Biology 6; Article number 364; doi:10.1038/msb.2010.1
Citation: Molecular Systems Biology 6:364
© 2010 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/
www.molecularsystemsbiology.com

Metabolomic and transcrip
of *Escherichia coli*

Szymon Jozefczuk[1,3], Sebastian Klie[1,3], Gareth Catchpole[1]
Joachim Selbig[1,2] and Lothar Willmitzer[1,*]


ARTICLE

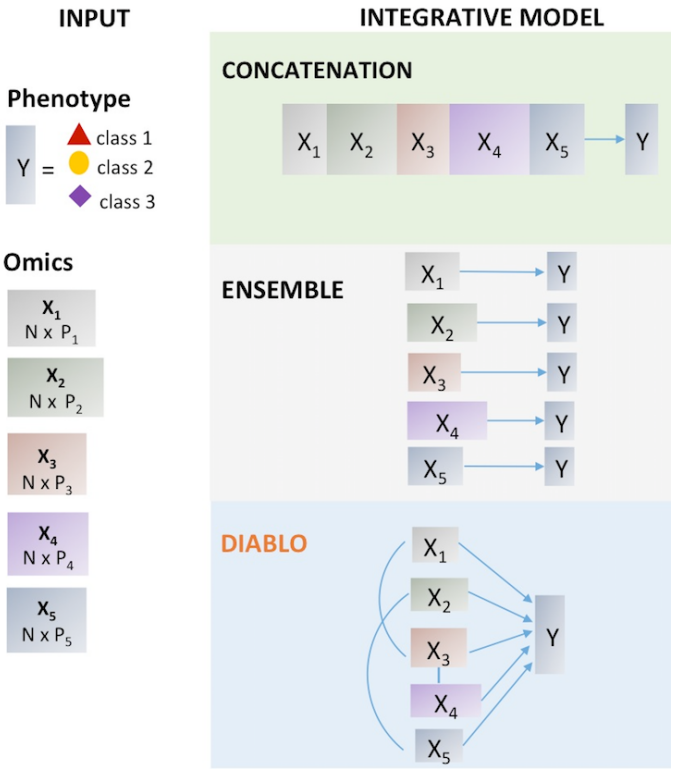
doi:10.1038/nature16942

Plankton networks driving carbon export in the oligotrophic ocean

Lionel Guidi[1,2*], Samuel Chaffron[3,4,5*], Lucie Bittner[6,7,8*], Damien Eveillard[9*], Abdelhalim Larhlimi[9], Simon Roux[10†], Youssef Darzi[3,4], Stephane Audic[8], Léo Berline[1†], Jennifer R. Brum[10†], Luis Pedro Coelho[11], Julio Cesar Ignacio Espinoza[10], Shruti Malviya[7†], Shinichi Sunagawa[11], Céline Dimier[8], Stefanie Kandels-Lewis[11,12], Marc Picheral[1], Julie Poulain[13], Sarah Searson[1,2], *Tara* Oceans Consortium Coordinators‡, Lars Stemmann[1], Fabrice Not[8], Pascal Hingamp[14], Sabrina Speich[15], Mick Follows[16], Lee Karp-Boss[17], Emmanuel Boss[17], Hiroyuki Ogata[18], Stephane Pesant[19,20], Jean Weissenbach[13,21,22], Patrick Wincker[13,21,22], Silvia G. Acinas[23], Peer Bork[11,24], Colomban de Vargas[8], Daniele Iudicone[25], Matthew B. Sullivan[10†], Jeroen Raes[3,4,5], Eric Karsenti[7,12], Chris Bowler[7] & Gabriel Gorsky[1]

Multi-'omics data integration frameworks for supervised analysis

Amrit Singh (UBC), Florian Rohart (UQ)



Concatenation: fit a supervised model on concatenated data

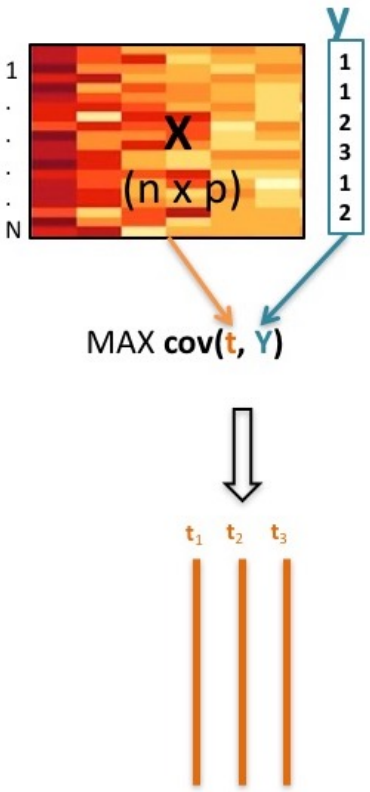
Ensemble: fit a supervised model on each 'omics dataset, then aggregate the results

DIABLO: maximise the covariance between specific 'omics datasets and the phenotype group

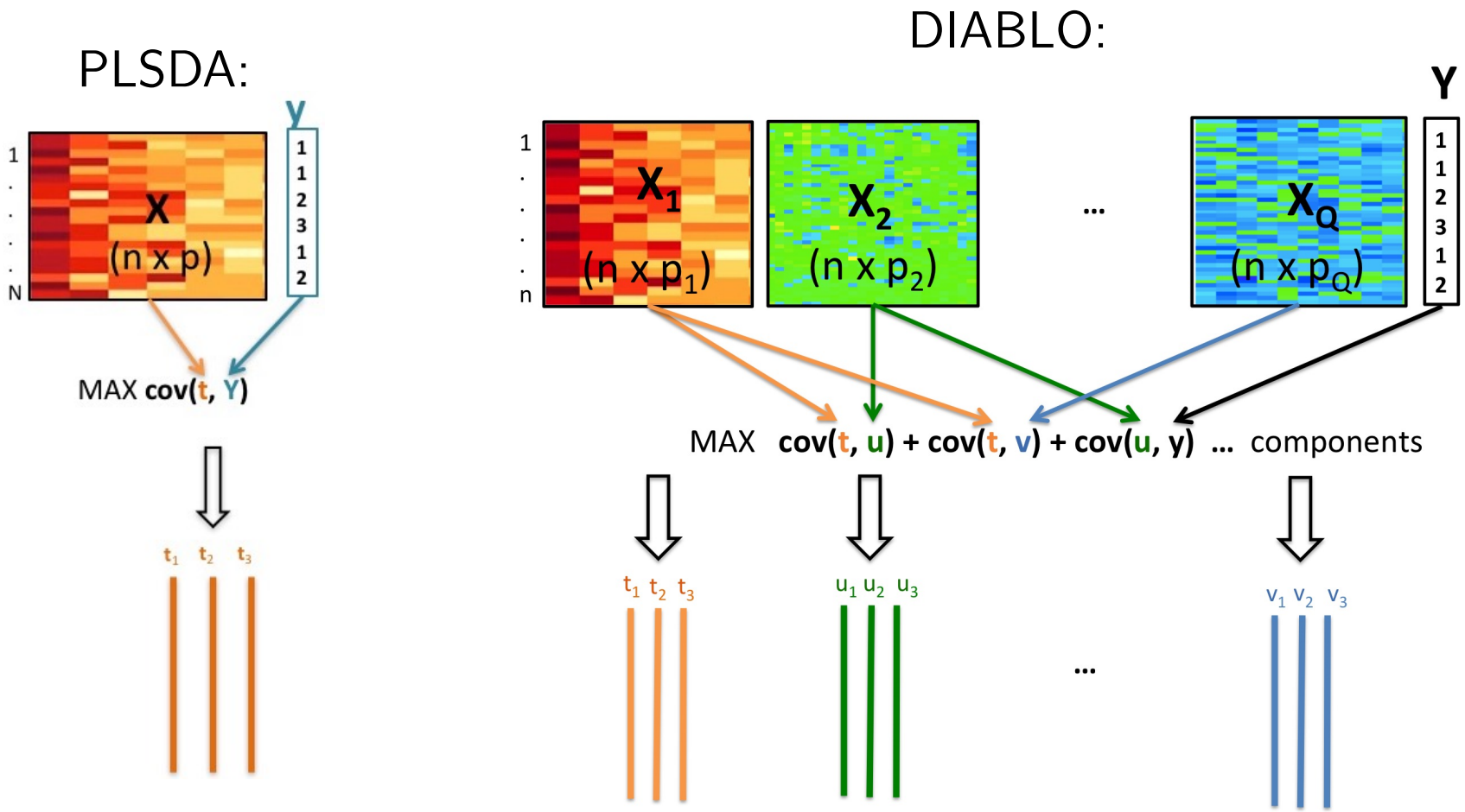
Singh A, Gautier B, Shannon C, Vacher M, Rohart F, Tebbutt S, Lê Cao K-A. **DIABLO - an integrative, multi-omics, multivariate method for multi-group classification.** *bioRxiv* 067611

Maximize the sum of covariances between two components at a time

PLSDA:

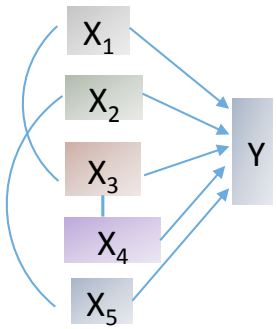


Maximize the sum of covariances between two components at a time



Maximise the covariance between specific 'omics datasets and the phenotype group

C matrix design 'links' datasets (link == maximise the covariance btw 2 data sets).



is coded in R as

```
> design
   X1 X2 X3 X4 X5
X1  0  0  1  0  0
X2  0  0  0  0  1
X3  1  0  0  1  0
X4  0  0  1  0  0
X5  0  1  0  0  0
```

Choose C based on prior knowledge or data-driven preliminary exploratory analyses.

Parameters in DIABLO

- The design C: what is the biological question?
- The number of components (`K-1`, with K number of classes)
- Number of variables to select on *each* component and for *each* data set.
  - intensive cross-validation or
  - arbitrary values for pragmatic biological validation of the identified biomarkers (GO analysis, experimental validation).

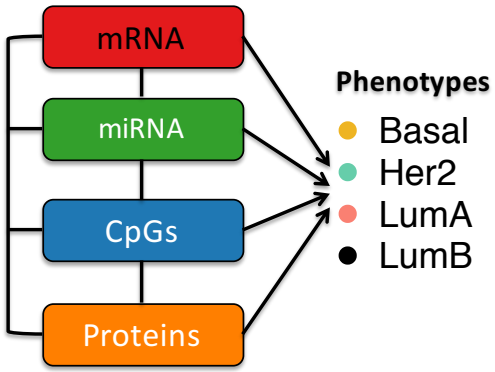
The multi-'omics module is still in active development
→ follow us: http://mixomics.org/category/news/

Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *bioRxiv 108597*

Multi-'omics breast cancer study

with **Amrit Singh** (UBC), Benoît Gautier, Florian Rohart (UQ)

- Outcome $\boldsymbol{Y}$: lumA, lumB, HER2, basal cancer subtypes ($n = 379$)
- 'Omics: mRNA (2,000), miRNA (184), CpGs (2,000), Proteins (142)



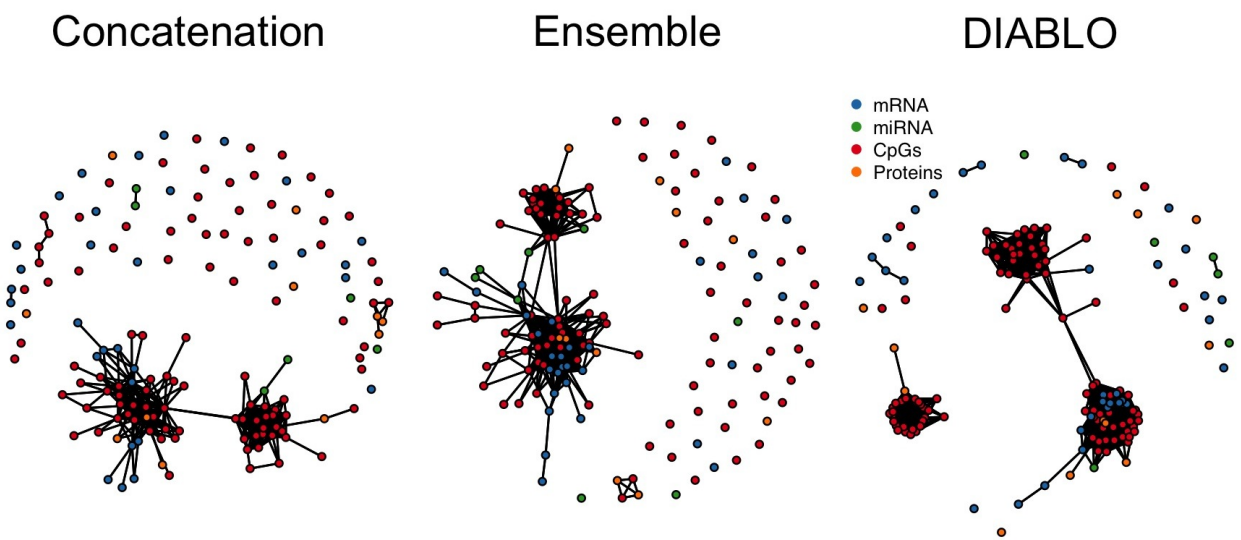
DIABLO design

$\rightsquigarrow$ Identify a correlated multi-'omics signature to explain/predict $\mathbf{Y}$.

* Prefiltered based on sd

powered by mixOmics

Multi-'omics signature connectivity

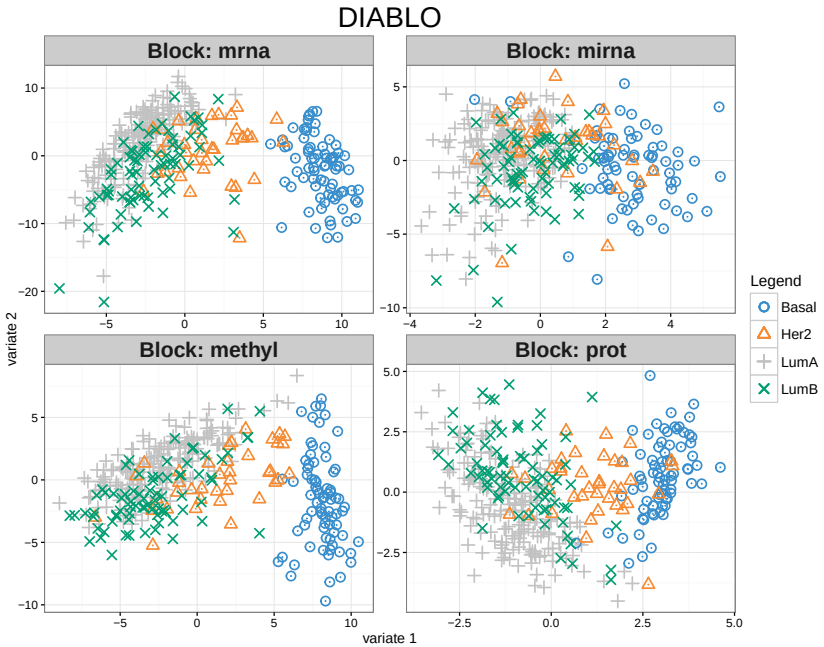


Relevance network of the signature identified by the integrative methods (link indicate $|r| > 0.6$).

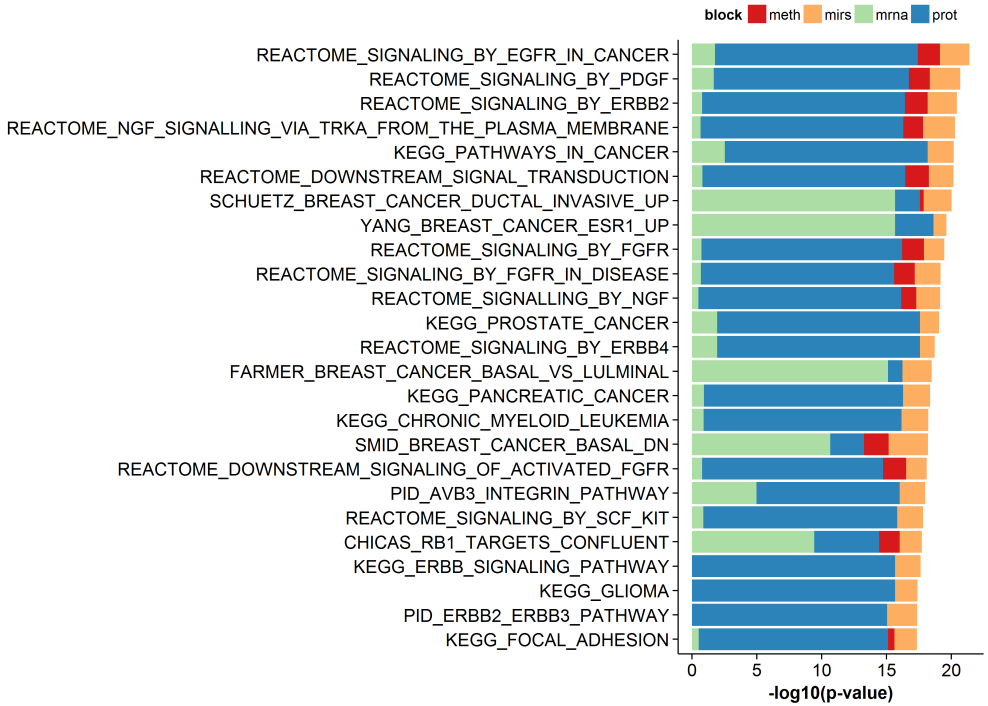
DIABLO seeks for subsets of 'omics variables maximally correlated.

Signature between the 'omics is more balanced than concatenation.

Multi-'omics signature



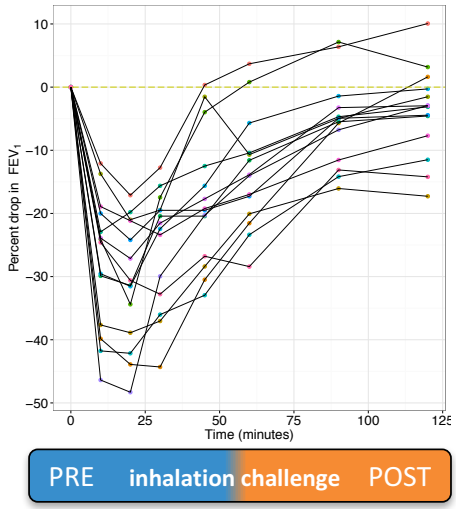
Sample plot
(agreement btw data sets)

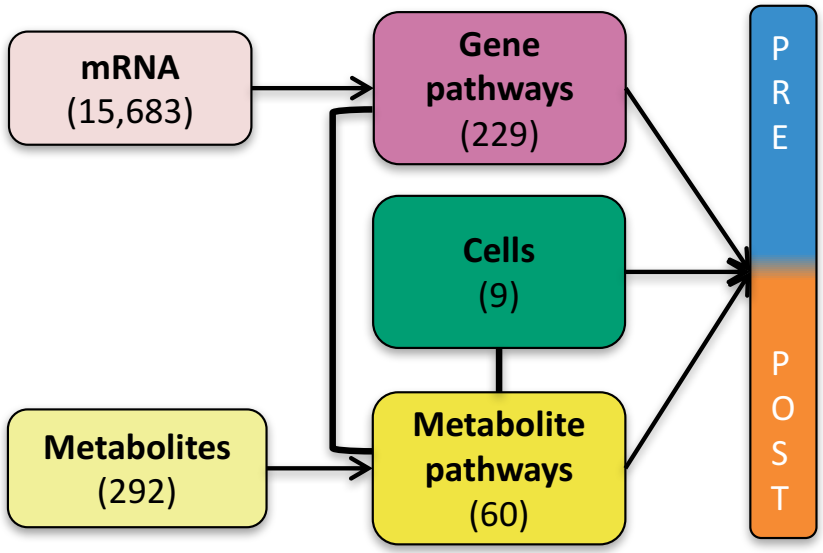


Pathway analysis, not in mixOmics yet
(complementarity of multi-'omics signature)

Multi-'omics asthma study

14 asthmatic individuals undergoing allergen inhalation challenge



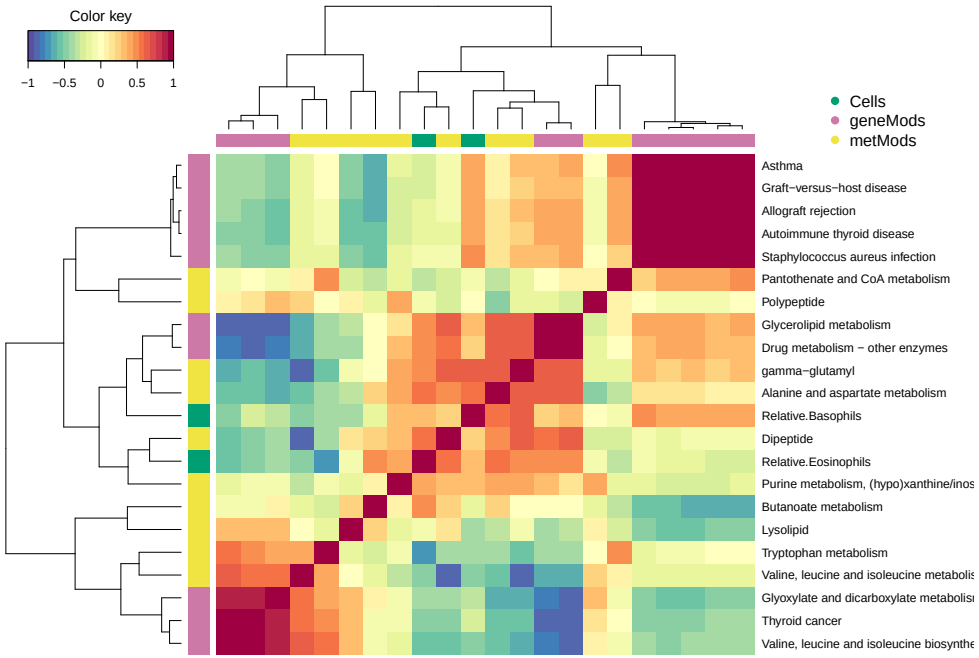


Leukocyte gene expression and plasma metabolite abundance reduced to pathway scores using component scores ('eigengene summarisation',).

Singh A, Gautier B, Shannon C, Vacher M, Rohart M, Tebbutt S, Lê Cao K-A. DIABLO - an integrative, multi-omics, multivariate method for multi-group classification, bioRxiv 067611

Multi-'omics asthma study

Our multi-'omics signature suggests mechanistic link with response to allergen challenge across different biological layers



- Cell types: eosinophils and basophils (hallmarks of allergic asthma)
- Gene pathways: Asthma KEGG pathway, Valine, leucine and isoleucine biosynthesis module (↑ post challenge)
- Metabolite pathway: Valine, leucine and isoleucine metabolism module (↑ post challenge)

Correlation btw cells, genes/metab pathways

Singh A, Gautier B, Shannon C, Vacher M, Rohart M, Tebbutt S, Lê Cao K-A. DIABLO - an integrative, multi-omics, multivariate method for multi-group classification, bioRxiv 067611

Statistics in the 'omics era

- Strong potential of multivariate methods, especially for challenging, unconventional, high variability data
  - data exploration; classification; integration of multiple data sets; biomarker identification
  - may provide a deeper understanding of a biological system
- Multiple 'omics integration is a complex statistical problem that requires well-stated biological questions and strong multi disciplinary collaborations

www.mixOmics.org @mixOmics_team

mixomics@math.univ-toulouse.fr kimanh.lecao@unimelb.edu.au

Acknowledgements

mixOmics team

France: Sébastien Déjean, François Bartolo

Australia: K-A Lê Cao, Florian Rohart, Benoît Gautier

Many mixOmics users

Lê Cao past lab at UQ

Staff: Florian Rohart, Nicholas Matigian, Anne Bernard, Benoît Gautier

PhD students: Ralph Patrick, Chao Liu, Amrit Singh, Jasmin Straube, Aimee Hanson

Masters/Hons: Eva Wang, Vanessa Lakis, Zoe Welham, Nick d'Arcy, Thom Cuddihy

Case studies collaborators

Hill lab: Alok Shah, Michelle Hill (UQ/QIMR)

Thomas lab: Helen Benham, Ranjeny Thomas (UQ)

PROOF Centre: Oliver Günther, Amrit Singh, Scott Tebutt (UBC)

a senior postdoc, computational biostatistics, University of Melbourne

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council