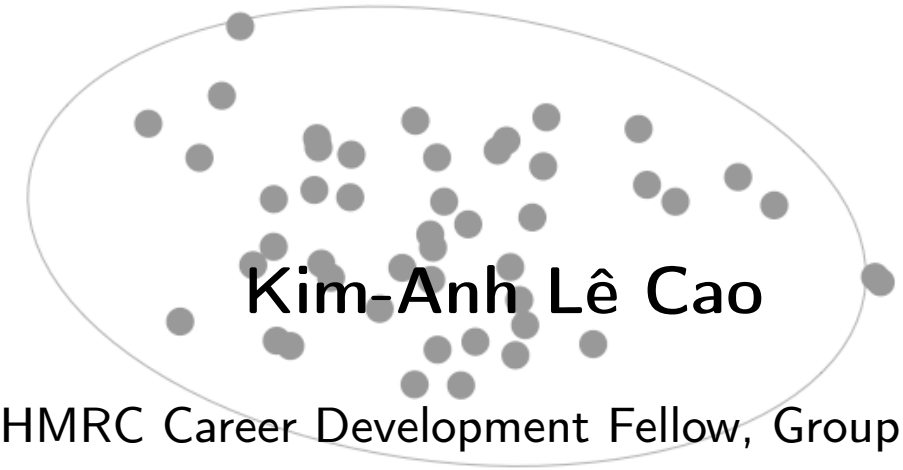
Multivariate integration of multiple biological datasets



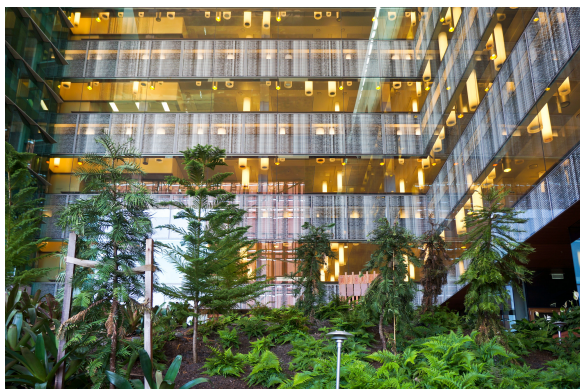
Kim-Anh Lê Cao

NHMRC Career Development Fellow, Group Leader

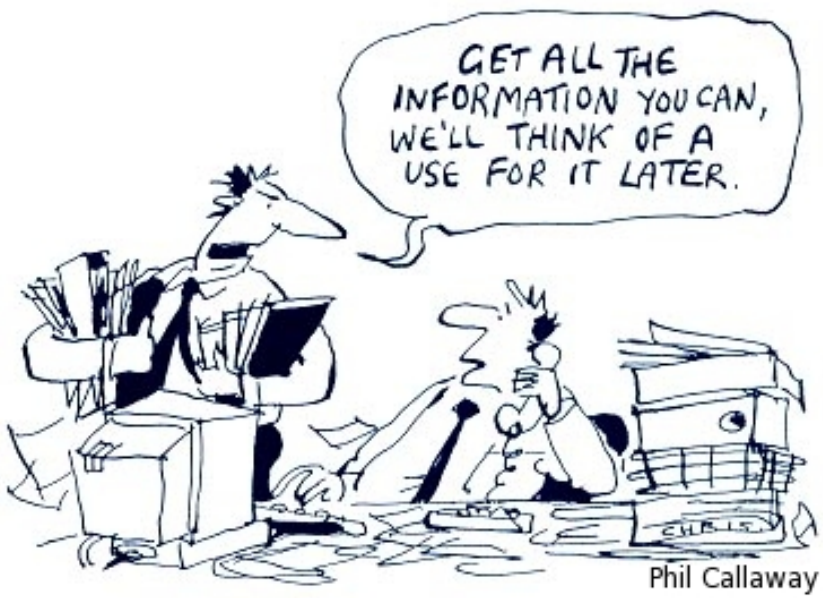
The University of Queensland Diamantina Institute
Brisbane, Australia

I moved from Toulouse, France to Brisbane, Australia in late 2008.

I lead a team of computational statisticians at the Translational Research Institute, a unique, Australian initiative of 'bench to bedside' medical research.



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

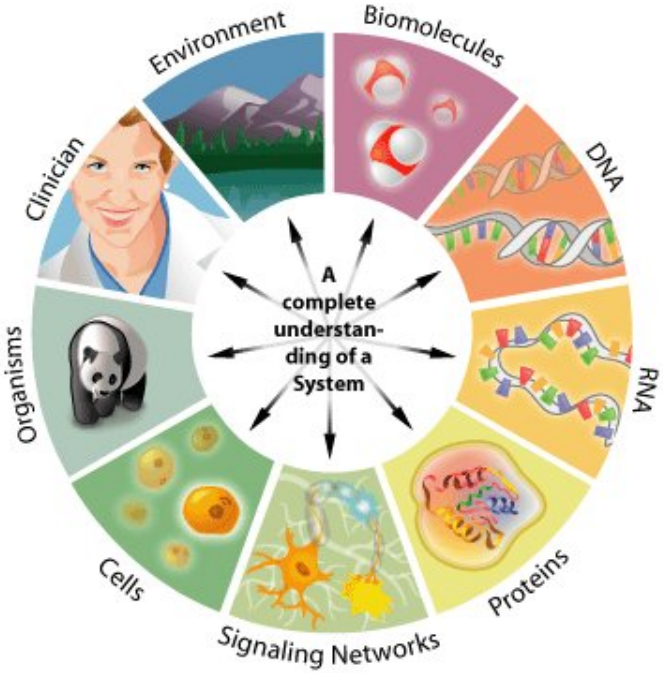
A holistic view of a biological system

Reductionism: 1 gene = 1 hypothesis

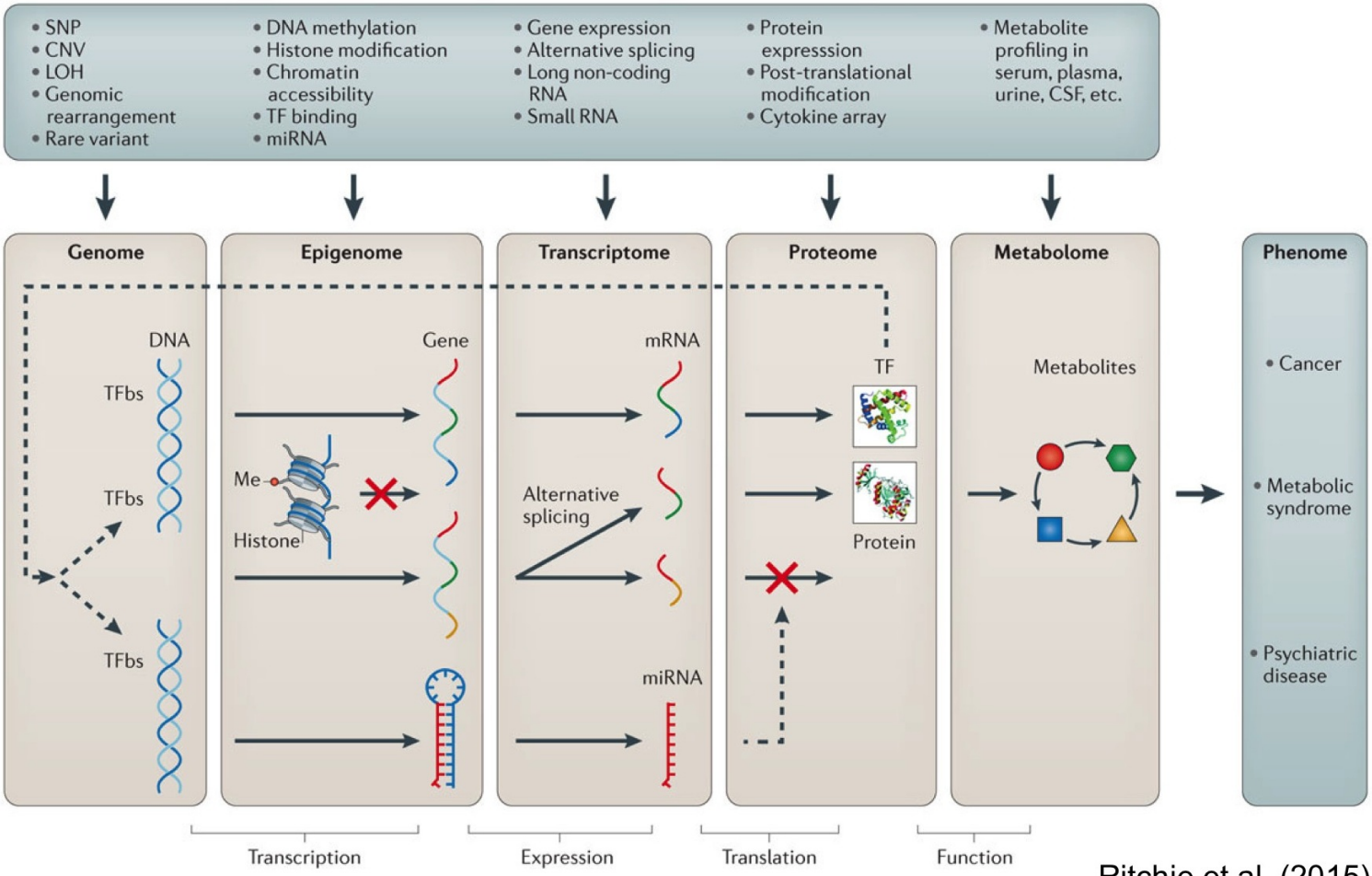
$\Downarrow$

Holism: thousands of molecules = ??

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



Heterogeneous 'omics data, from genome to phenome



Ritchie et al. (2015)
Nature Reviews | Genetics

Research aim and hypothesis

Shift the paradigm of univariate statistics to solve biological problems:

Molecular entities act together in biological pathways and need to be appropriately modelled using novel multivariate statistical methods.

Research aim and hypothesis

Shift the paradigm of univariate statistics to solve biological problems:

Molecular entities act together in biological pathways and need to be appropriately modelled using novel multivariate statistical methods.

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

George E. P. Box, statistician
1919 - 2013

Table of Contents

1 mixOmics R toolkit

2 n-integration

3 p-integration

mixOmics R package

- French'Oz team:
  4 core, 1 developer, students, collaborators
- 10K downloads in 2015, 16K downloads in 2016
- 12 multi-day workshops since 2014 (FR, AUS, NZ)
- 15 novel multivariate methods

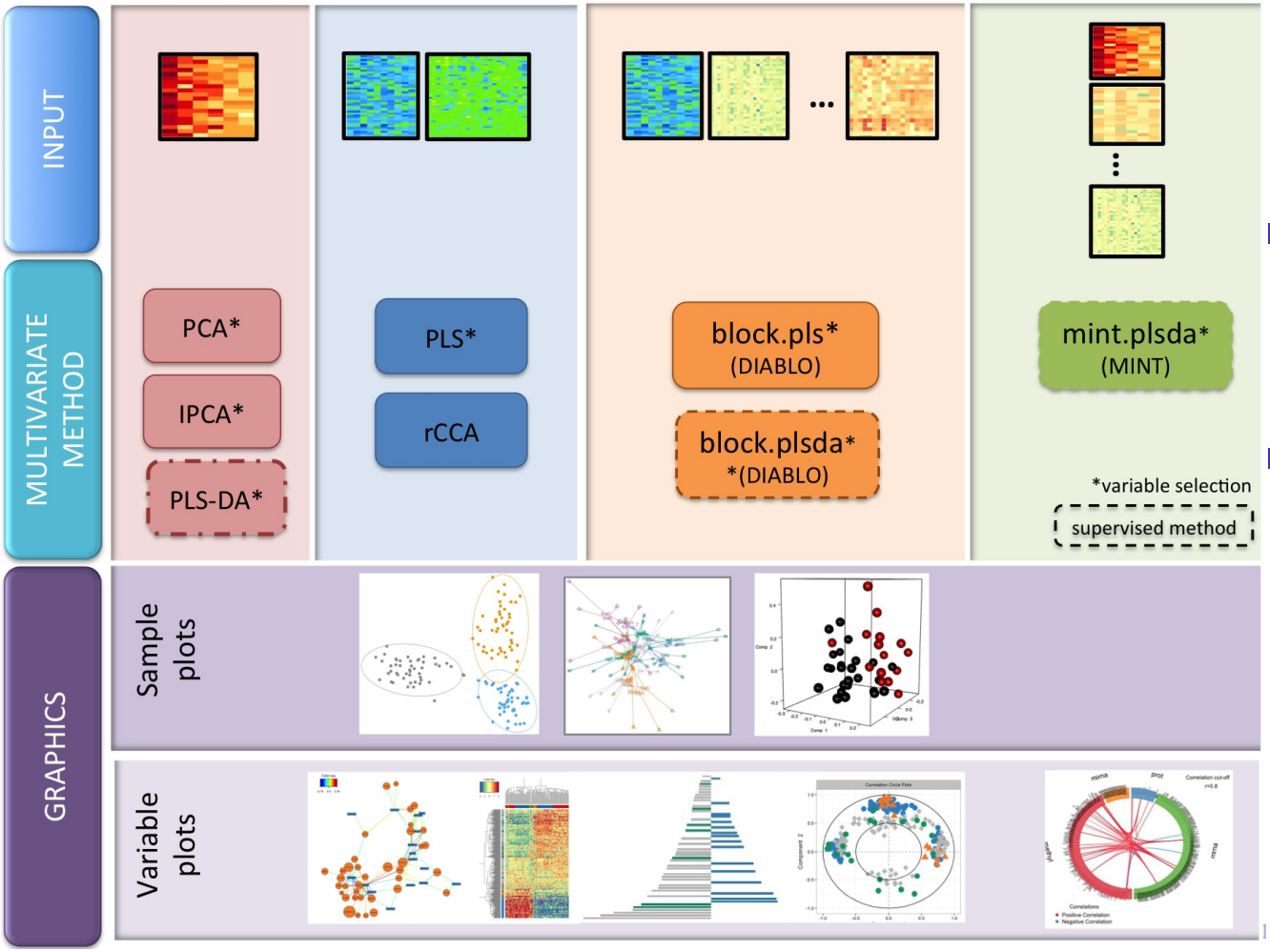
My research program focuses on the development of multivariate statistical methodologies, their applications in areas informed by biology, and the training of the new generation of computational biologists.

www.mixOmics.org

Flexibility of multivariate methods

	Single data set	**Multiple data sets**
Unsupervised (No response) **Supervised** (categorical response $\boldsymbol{Y}$)	Data exploration & visualisation Selection of a subset of correlated variables Selection of a subset of discriminative variables	

mixOmics framework for 'large p small n' biological data



- n-integration with **DIABLO**:
  $\neq$ 'omics variables
  same n samples
- p-integration with **MINT**:
  same p variables
  $\neq$ samples from independent studies

Multivariate methods

- Seek for components summarising the information from the data
- Each component is a linear combination of the original variables (e.g. gene expression levels, protein abundance)
- LASSO penalties to perform *variable selection* ('sparse' linear methods)
- Methods based on Projection to Latent Structures Discriminant Analysis (PLSDA) and sparse variants.

Lê Cao et al., 2008, 2009ab, 2011; Günther et al. 2014; Tenenhaus et al. 2014

Linear combinations of variables

$$\underset{(n \times p)}{\mathbf{X}} =$$

	Height	Weight
1	174.0	65.6
2	175.3	71.8
3	193.5	80.7
4	186.5	72.6
5	187.2	78.8
6	181.5	74.8
7	184.0	86.4
8	184.5	78.4
9	175.0	62.0
10	184.0	81.6

$$\mathbf{a} = \begin{pmatrix} 0.5 \\ 2 \end{pmatrix}$$

Assign to the variables Height and Weight the coefficients $a_1 = 0.5$ and $a_2 = 2$ respectively *(in matrix algebra: **Xa**)*.

Linear combination of variables for data reduction

		Height			Weight		Linear combination
		174.0			65.6		218.20
		175.3			71.8		231.25
		193.5			80.7		258.15
		186.5			72.6		238.45
a_1	$\times$	187.2	$+$	a_2	$\times$ 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

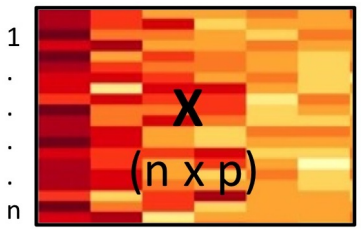
Component X**a** summarises the original variables.

$\rightarrow$ Efficient algorithms for large data sets (Singular Value Decomposition)

$\rightarrow$ Variable coefficient vector (loading) ***a*** defined according to some statistical criterion.

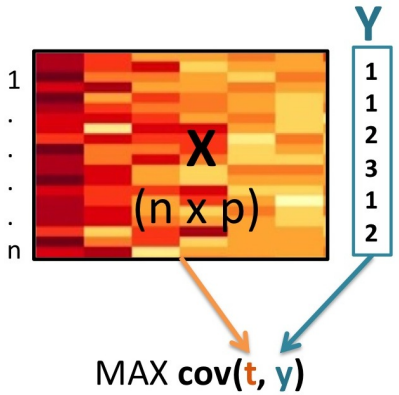
Matrix decomposition for dimension reduction

PCA maximises the **variance**

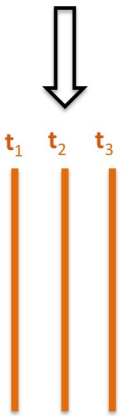


MAX **var(t)** components

PLSDA maximises the **covariance** with a categorical outcome

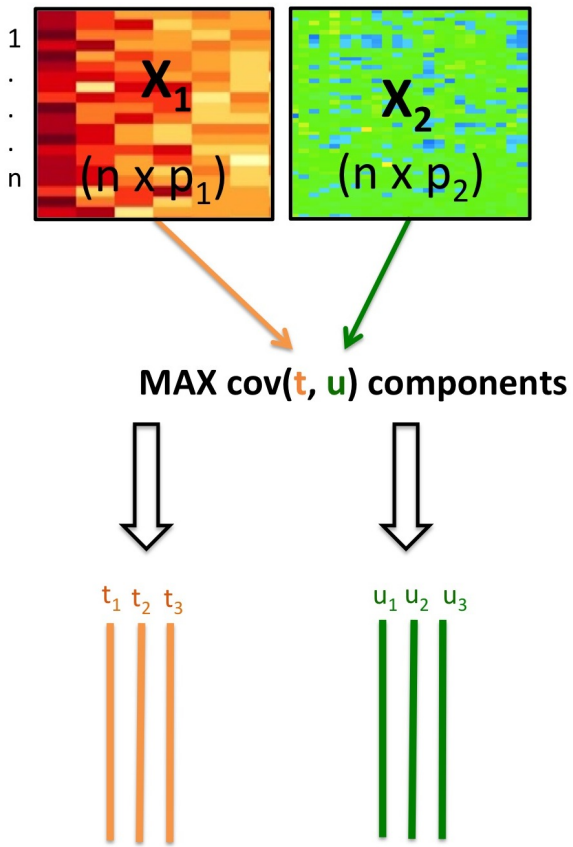


MAX **cov(t, y)**



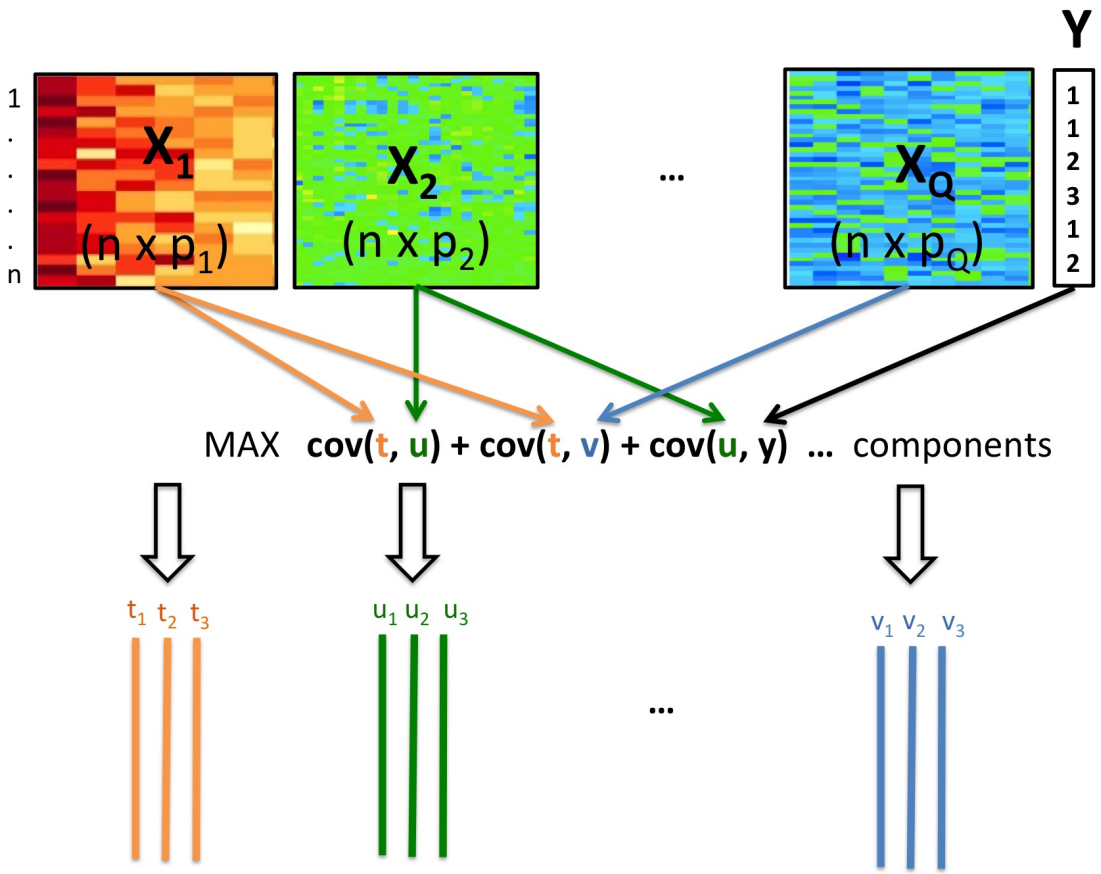
Matrix decomposition for dimension reduction

PLS maximises the **covariance** between two data sets



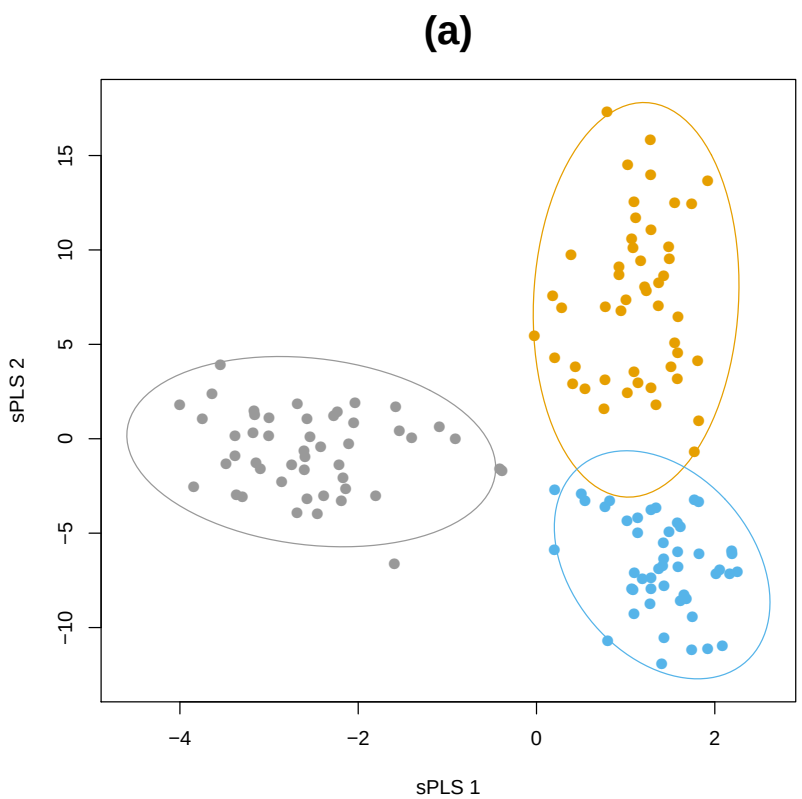
Matrix decomposition for dimension reduction

Multiblock-PLSDA maximises the **sum of covariances** between each pair of data sets and with the outcome

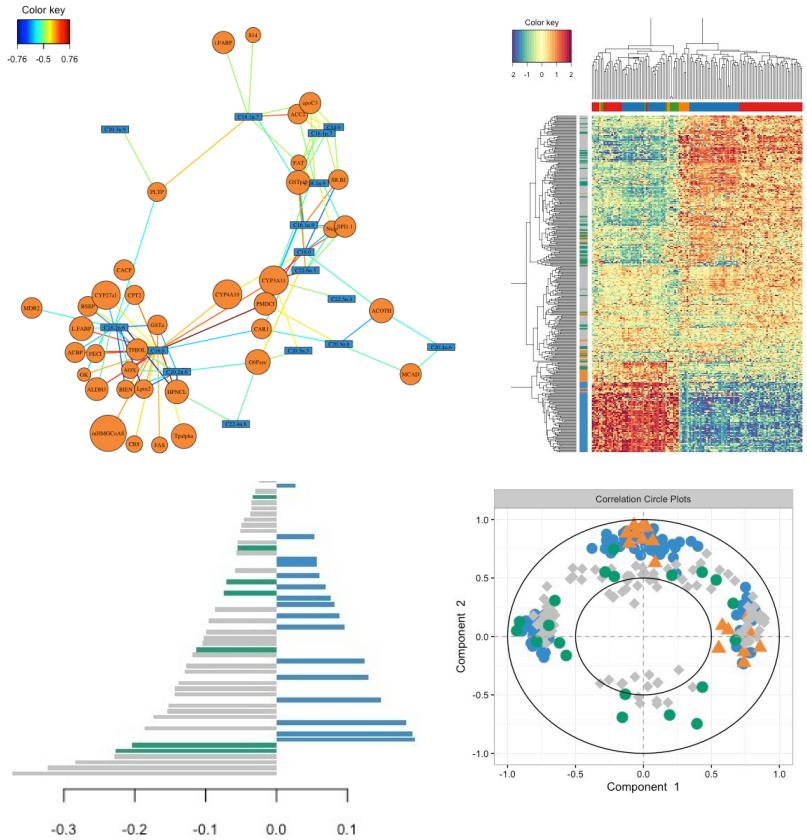


Visualisation using components

Sample plot



Variable plots



Pros and Cons of multivariate methods

Pros

- Handle $n << p$ and multicollinear data
- Few assumptions on data distribution, allowing large data range and variability
- Sparse methods to identify molecular signatures
- Visualisation outputs

Cons

- Linear relationship between predictors
- Accuracy measures (ROC, performance), no p-value
- Still limited by very small n

Integration of multiple data sets

n-integration (same n)

- Integrate different 'omics,
  e.g. mRNA, miRNA, proteomics and methylation data
- Identify a multi-'omics signature to predict outcome.

Integration of multiple data sets

n-integration (same n)

- Integrate different 'omics,
  e.g. mRNA, miRNA, proteomics and methylation data
- Identify a multi-'omics signature to predict outcome.

p-integration (same p)

- Integrate several studies on the same 'omics
  e.g. microarrays, RNA-seq from $\neq$ consortia
- Identify a gene signature agnostic of platform type to predict outcome.

Integration of multiple data sets

n-integration (same n)

- Integrate different 'omics,
  e.g. mRNA, miRNA, proteomics and methylation data
- Identify a multi-'omics signature to predict outcome.

p-integration (same p)

- Integrate several studies on the same 'omics
  e.g. microarrays, RNA-seq from $\neq$ consortia
- Identify a gene signature agnostic of platform type to predict outcome.

$\rightsquigarrow$ unwanted systematic variation due to platform artefacts decreases statistical power and leads to spurious results

Table of Contents

1 mixOmics R toolkit

2 n-integration

3 p-integration

Multi-'omics breast cancer study

- Training $n = 379$; Test $n = 610$ (TCGA consortium)
- Outcome $\boldsymbol{Y}$: lumA, lumB, HER2, basal cancer subtypes

- Four 'omics:
  - mRNA* ($p_1 = 3,000$)
  - miRNA ($p_2 = 184$)
  - proteomics ($p_3 = 142$)
  - methylation CpG* ($p_4 = 3,000$)

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

* Prefiltered based on sd

powered by mixOmics

Amrit Singh (UBC, Vancouver),
Benoit Gautier & Florian Rohart (UQDI)

Maximize the sum of covariances between two components at a time

$$\max_{\boldsymbol{a}^1,\ldots,\boldsymbol{a}^J} \sum_{j,k=1,j\neq k}^{J} c_{kj}\ \mathrm{cov}(\boldsymbol{X}_j\boldsymbol{a}^j, \boldsymbol{X}_k\boldsymbol{a}^k)$$

$$\text{subject to } ||\boldsymbol{a}^j||_2 = 1 \quad \text{and} \quad ||\boldsymbol{a}^j||_1 \leq \lambda_j$$

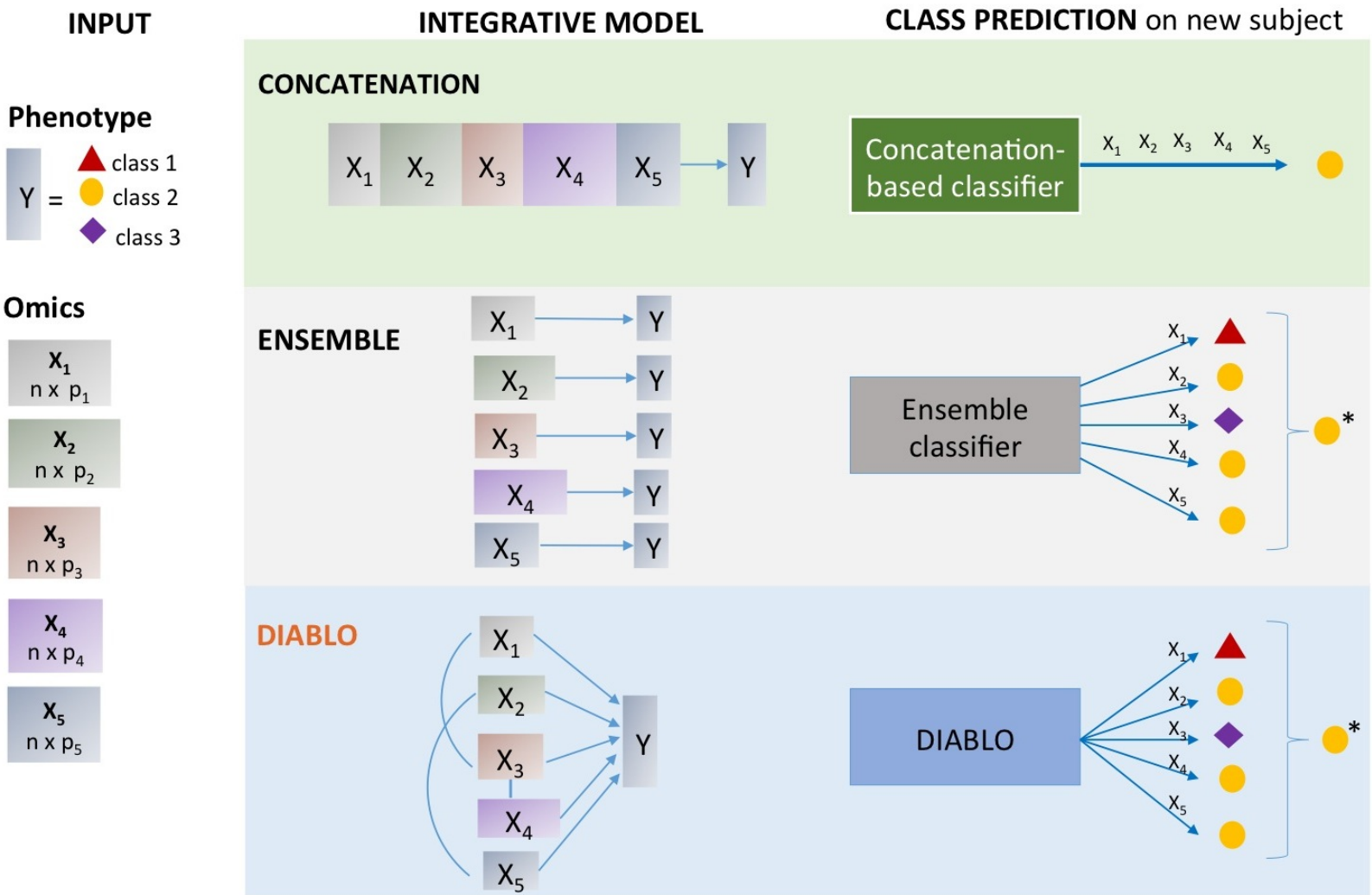
$\rightsquigarrow$ supervised analysis: substitute one $\boldsymbol{X}_j$ with dummy outcome matrix $\boldsymbol{Y}$

$\boldsymbol{X}_j\boldsymbol{a}^j$ is a component - linear combination of variables from data set $\boldsymbol{X}_j$
$\boldsymbol{a}^j$ is the variable coefficient vector for $\boldsymbol{X}_j$

Parameters:
λ_j is the lasso parameter shrinking $\boldsymbol{a}^j$ for data set $\boldsymbol{X}_j$
$\boldsymbol{C}$ is an indicator design $(J \times J)$ matrix, J is the # of 'omics data sets

DIABLO: **D**ata **I**ntegration **A**nalysis for **B**iomarker discovery using **L**atent components for **O**mics studies (variant of Generalised Canonical Correlation Analysis)

General integrative frameworks for prediction



Performance

Mean classification error rate (%), 50 x 5-fold cross-validation:

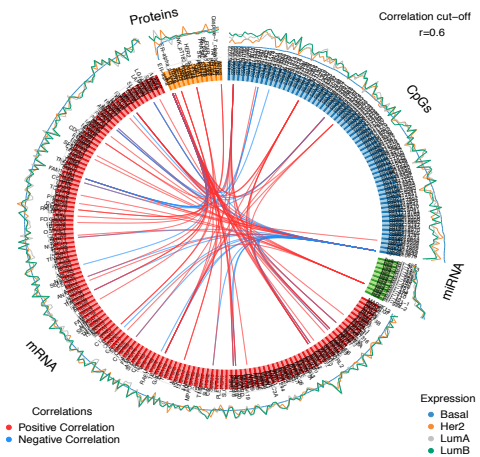
	Single 'omics (E-net)				Multiple 'omics		
	mRNA	miRNA	CpGs	Proteins	Concat	Ensemble	DIABLO
Training n = 379	14.0±1	26.0±2.0	27.0±2.0	22.0±1.0	14.0±1.0	22.0±1.0	13.5±1.3
Test n = 610	15.4	22.1	31.0	NA*	NA*	15.6	12.2

*Protein not available in test set

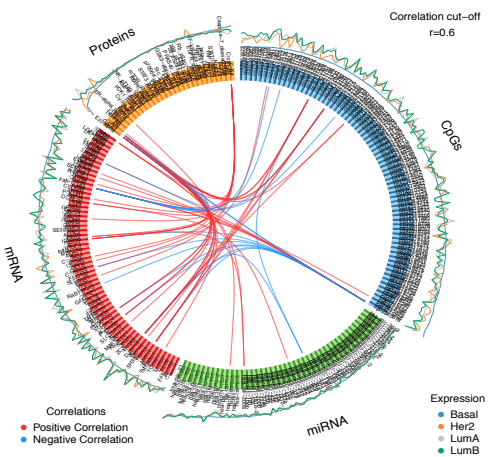
- External test set essential for unbiased assessment
- Prediction accuracy improved with integrative multi-'omics methods
- DIABLO identifies a highly predictive multi-'omics signature.

Multi-'omics signature

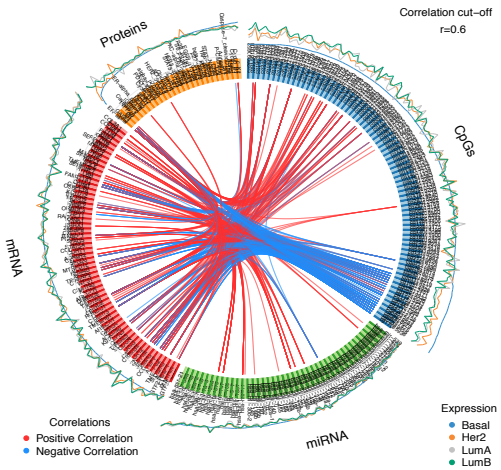
Concatenation



Ensemble



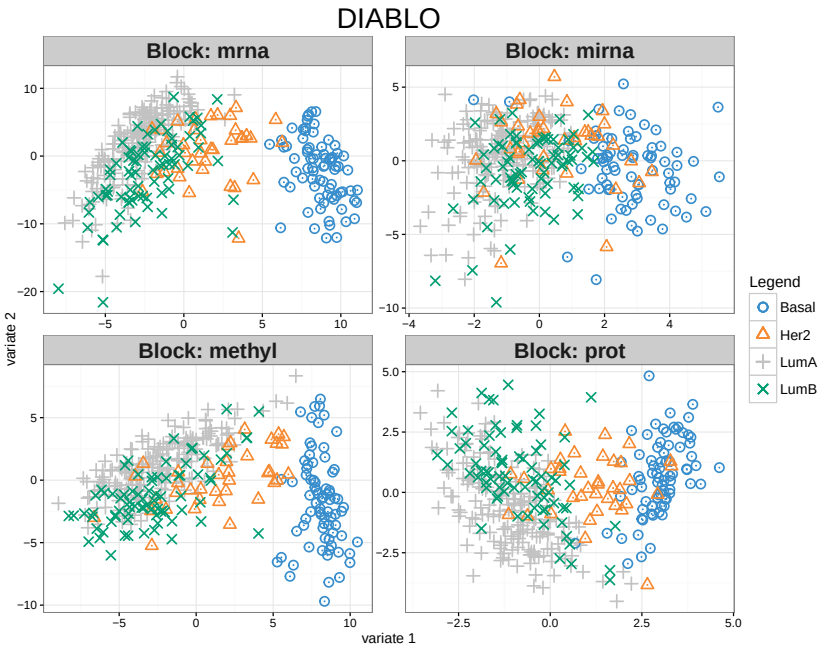
DIABLO



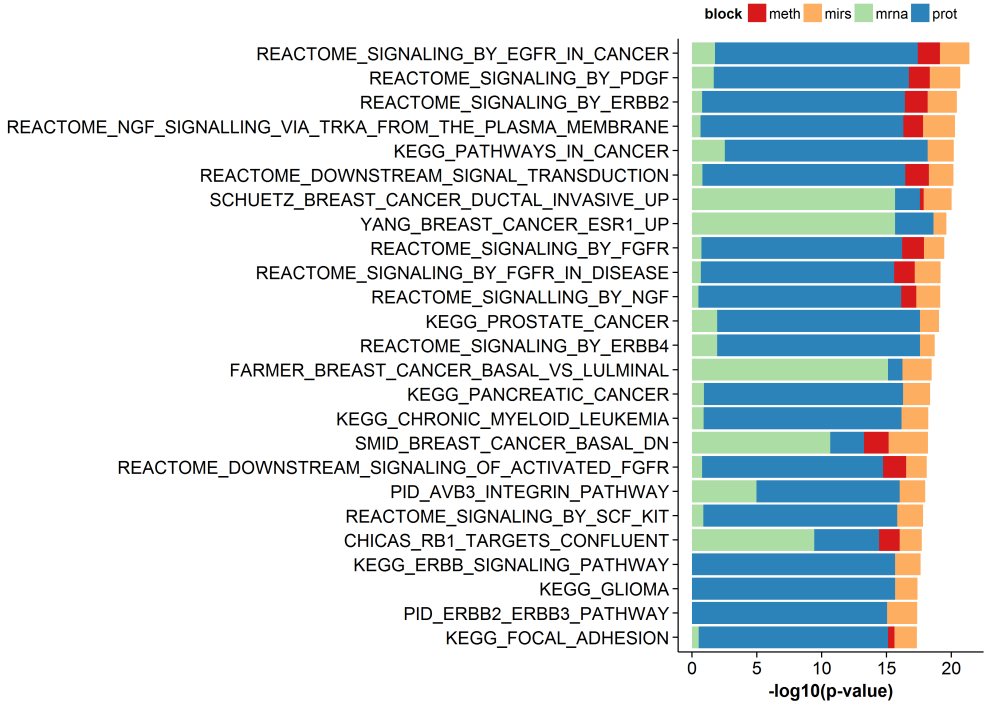
- Concatenation method: unbalanced signature
- DIABLO signature: highly correlated multi-'omics signature (111 mRNA, 84 miRNA, 144 CpG, 57 proteins)

circos plots: Michael Vacher (UWA)

Multi-'omics signature



Sample plot
(agreement btw data sets)



Pathway analysis
(complementarity of multi-'omics signature)

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

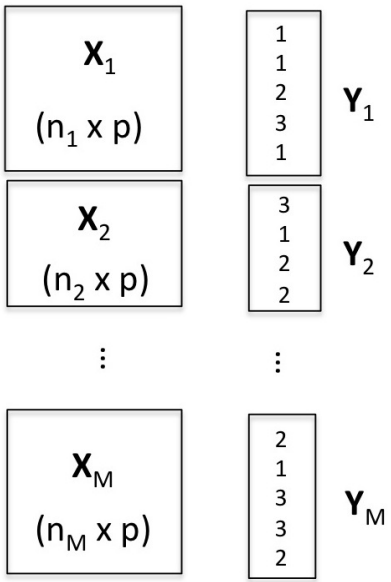
Table of Contents

1 mixOmics R toolkit

2 n-integration

3 p-integration

Multi-transcriptomics breast cancer studies



	Study	Platform	Basal	Her2	LumA	LumB
Train	METABRIC Discov	Illumina HT-12 v3	118	87	466	268
	METABRIC Valid	Illumina HT-12 v3	213	153	255	224
	TCGA RNA-seq	illumina HiSeq 2000	188	80	549	213
Test	TCGA microarray	Agilent custom 244K	57	31	99	67

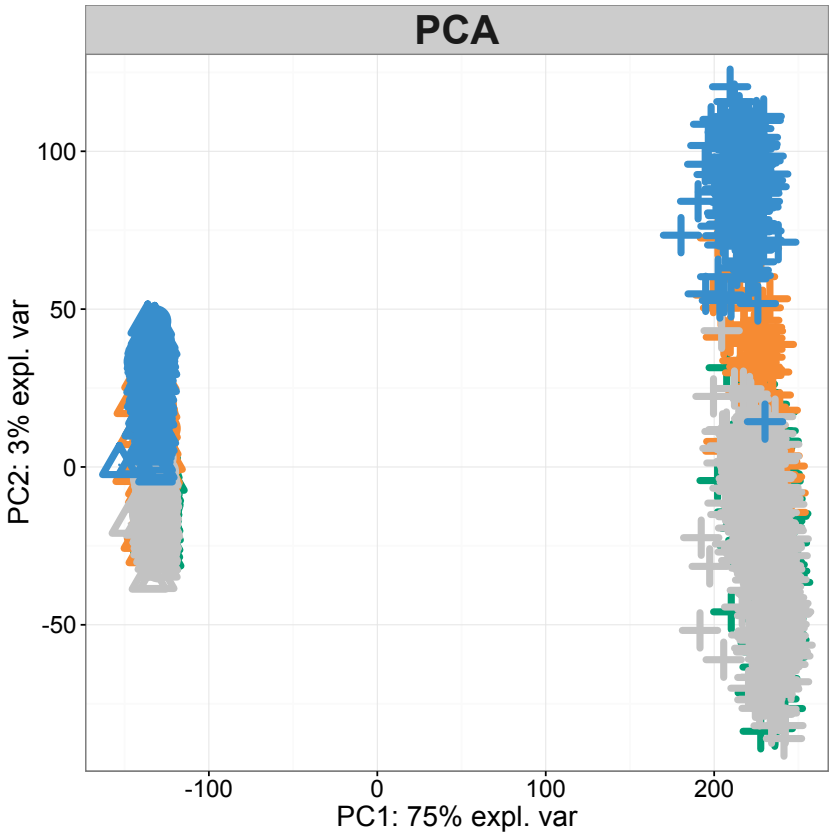
- Training set: $p = 15,803$, $n = 2,814$
- Test set: $p = 15,803$, $n = 254$
- Outcome $\boldsymbol{Y}$: lumA, lumB, HER2, basal cancer subtypes

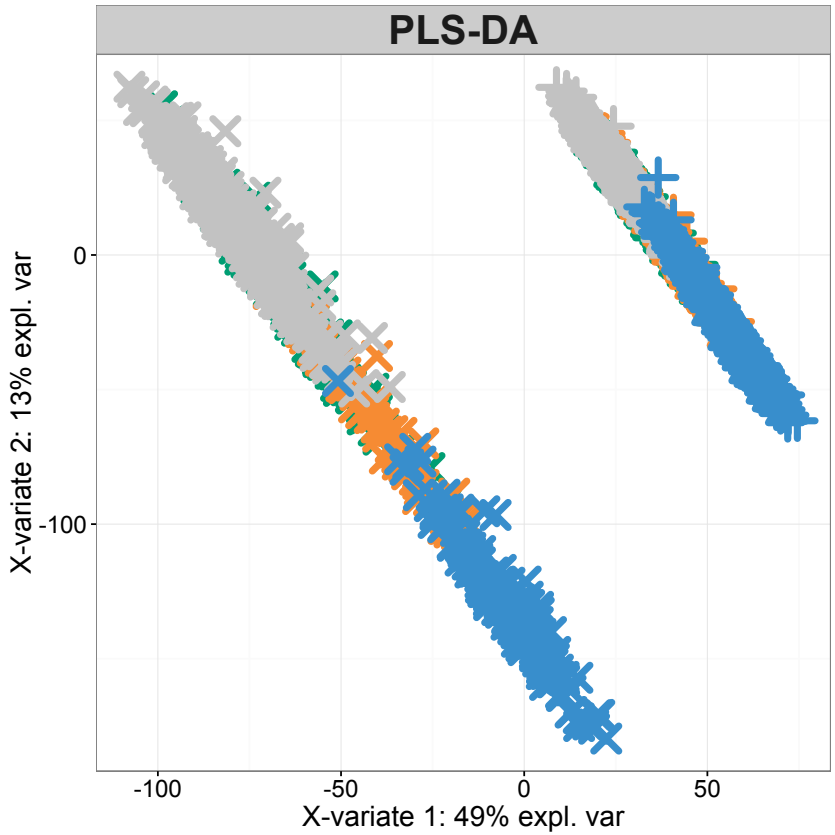
$\rightsquigarrow$ Identify platform agnostic gene signature to explain/predict $\mathbf{Y}$.

powered by mixOmics

Florian Rohart (UQDI)

Classical multivariate analysis highlight a study effect





Unsupervised analysis

Supervised analysis

Basal **Her2** **LumA** **LumB**

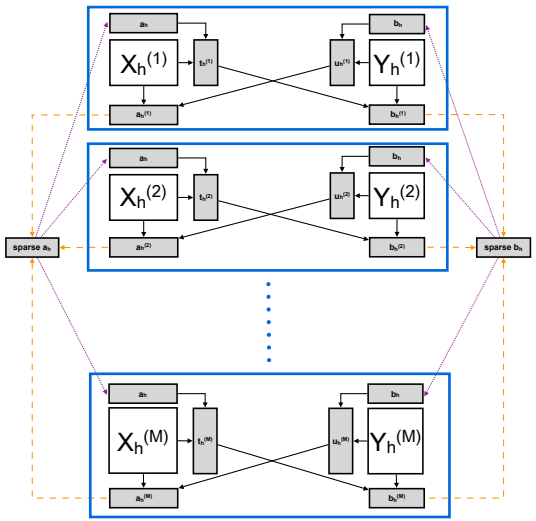
Study-specific and global components maximise the agreement between two data sets at a time

$$\max_{\boldsymbol{a}} \sum_{m=1}^{M} n_m cov(X_m \boldsymbol{a}, Y_m)$$

$$\text{subject to } ||\boldsymbol{a}||_2 = 1 \quad \text{and} \quad ||\boldsymbol{a}||_1 \leq \lambda$$

$\boldsymbol{X_m a}$ is the global component - linear combination of variables from data set $\boldsymbol{X}_m$
$\boldsymbol{a}$ is the global variable coefficient vector for $\boldsymbol{X}_m$

Parameters:
λ is the lasso parameter shrinking each $\boldsymbol{a}$ for each data set $\boldsymbol{X_m}$.



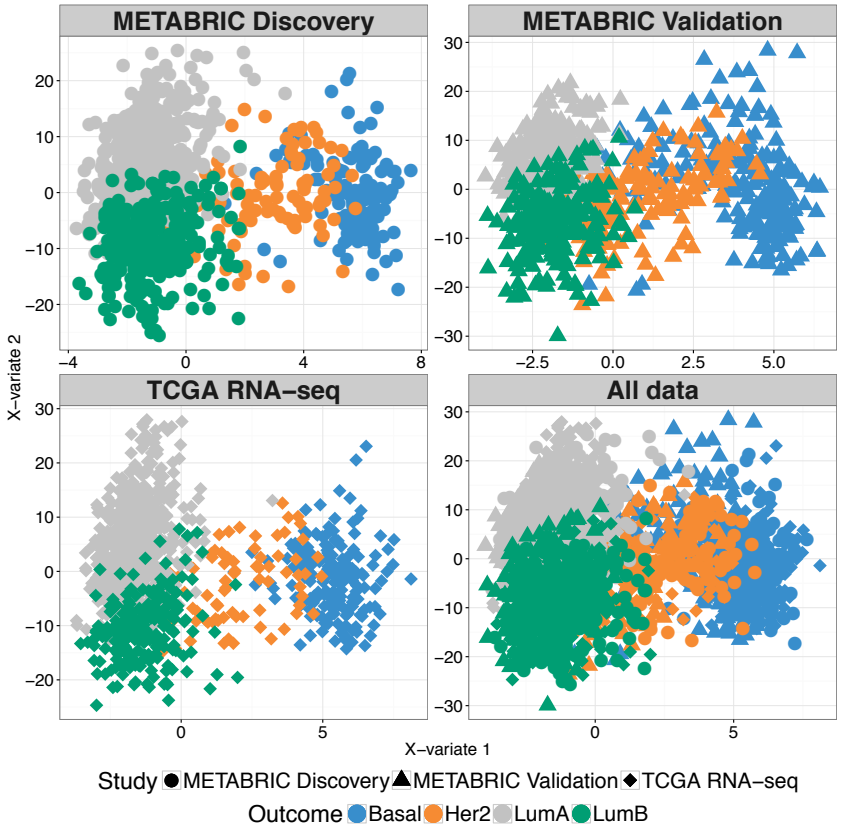
MINT: **M**ultivariate **INT**egration across independent experiments

Performance

Balanced error rate & accuracy per group:

Study	BER	Basal	Her2	LumA	LumB
METABRIC Discovery	14.5	91.5	80.5	94.2	75.7
METABRIC Validation	22.3	77.0	65.4	93.3	75.0
TCGA RNA-seq	15.7	98.4	70.0	95.8	72.8
overall	18.0	88.1	70.6	94.7	74.6

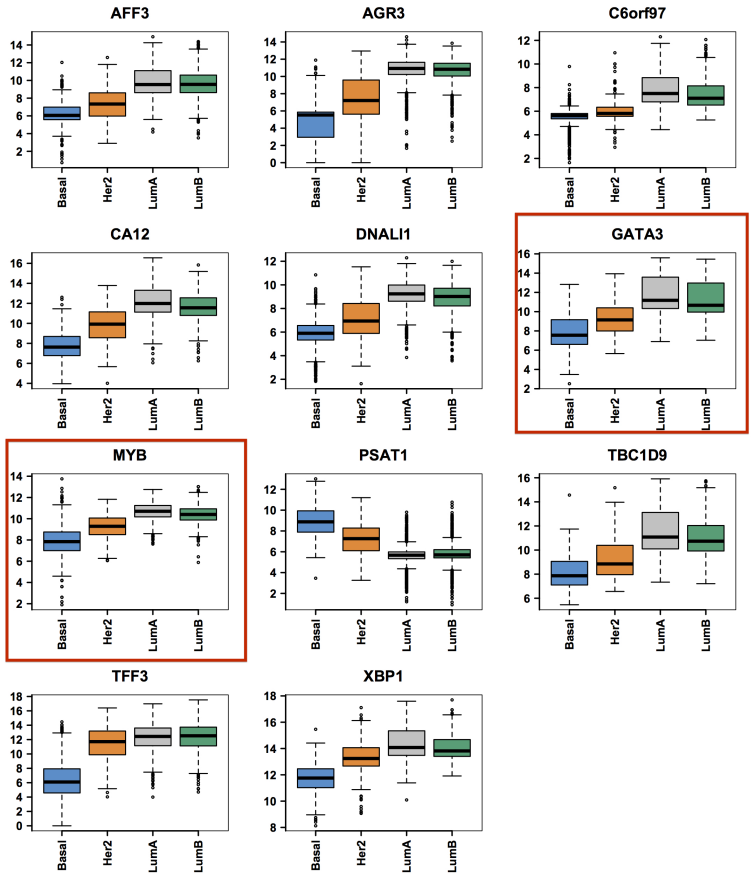
- BER = 11.2 on external Test set
- MINT performance superior to existing sequential methods



Study-specific and global sample plots (agreement btw studies)

Study agnostic gene signature

Rank		
1	GATA3	✓✓✓
2	AGR3	✓
3	CA12	
4	TBC1D9	
5	PSAT1	
6	XBP1	✓
7	TFF3	✓
8	DNALI1	
9	MYB	✓✓
10	CCDC170	
11	AFF3	



Rohart F, Eslami A, Bougeard S and Lê Cao K-A. MINT: A multivariate integrative approach to identify a reproducible biomarker signature across multiple experiments and platforms. bioRxiv, 070813

Biological questions drive novel statistical developments

- Flexibility of multivariate linear methods:
  $\rightarrow$ data exploration; classification; integration of multiple data sets
- Variable selection suggest novel hypotheses, further investigations & experimental validations.

`mixOmics` future developments

- Time course 'omics integration
- SNP analysis & integration with 'omics
- metagenomics* & integration with 'omics from the host

*Lê Cao K-A.†, Costello ME †, et al. (2016) MixMC: a multivariate statistical framework to gain insight into Microbial Communities. PLoS ONE 11(8).

Acknowledgements

mixOmics team

Sébastien Déjean	Univ Toulouse
Ignacio González	INRA Toulouse
François Bartolo	Univ Toulouse
Florian Rohart	UQ
Benoît Gautier	UQ

Ext collaborators (DIABLO)

Michael Vacher	UWA Perth
Casey Shannon	PROOF Vancouver

And many of our mixOmics users!

Lê Cao lab

Florian Rohart	Postdoc
Anne Bernard	Postdoc P/T
Nicholas Matigian	Biostats Facility

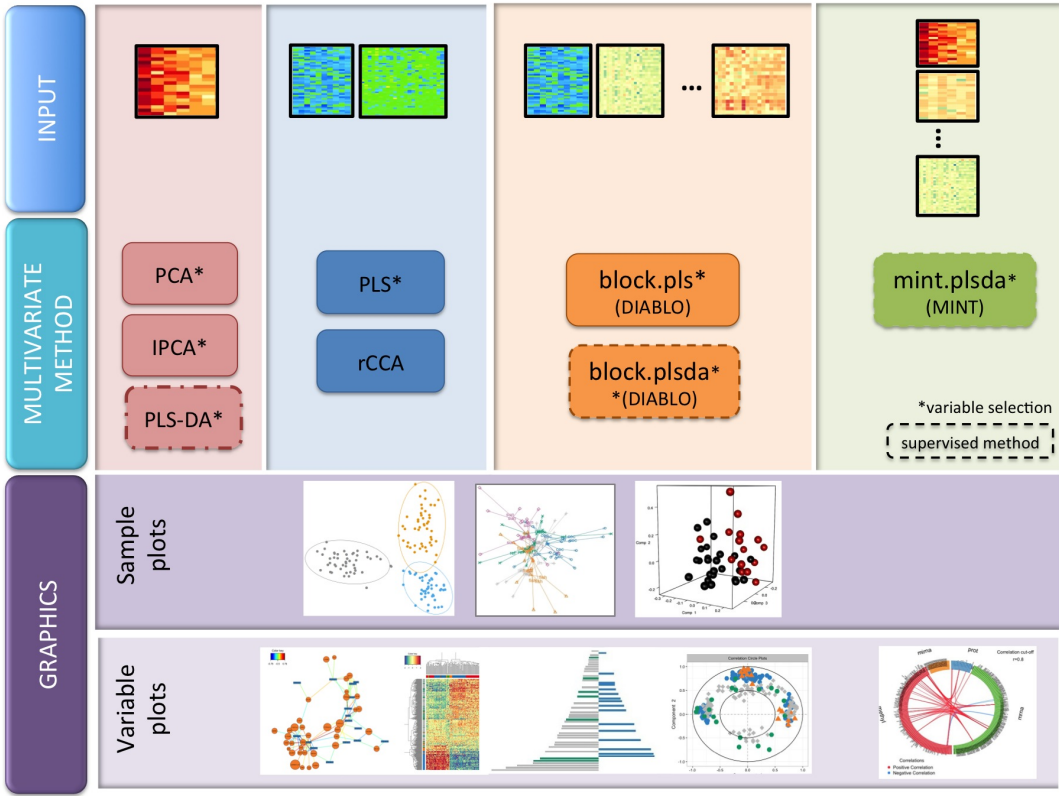
& RHD students

Amrit Singh (UBC)	**Jasmin Straube**
Chao Liu	**Ralph Patrick**
Aimee Hanson	**Vanessa Lakis**
Eva Yiwen Wang	**Nick Muller**
Nick d'Arcy	

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council

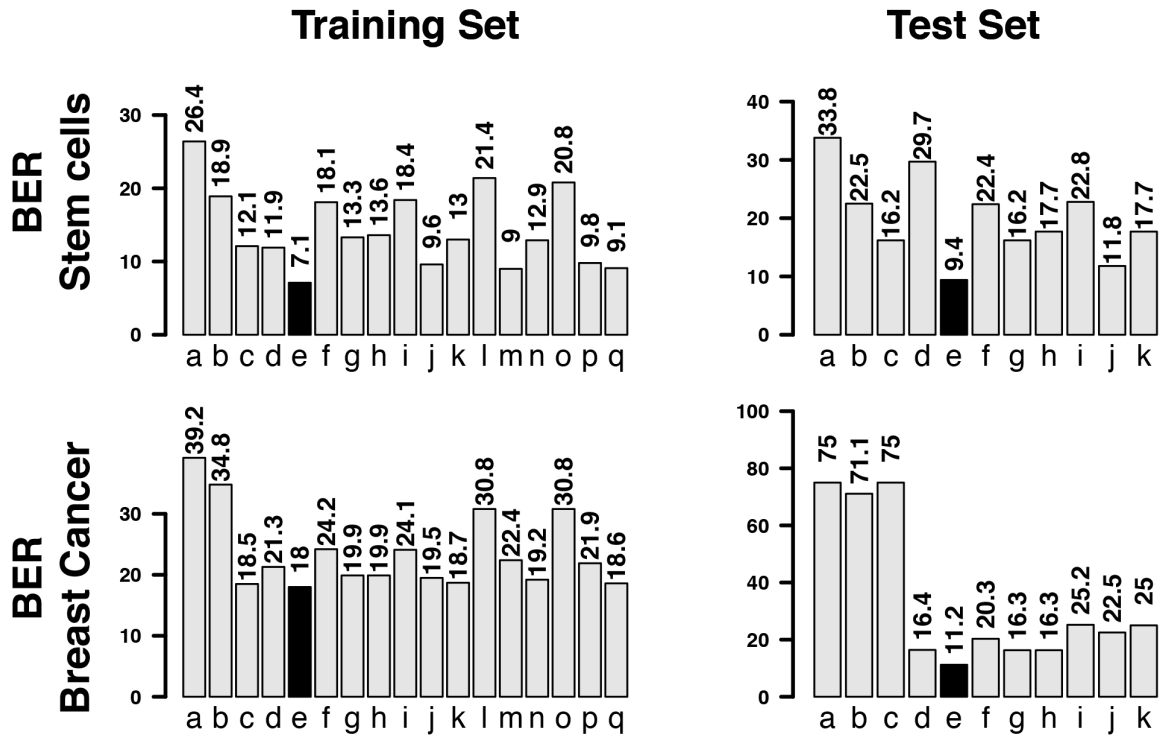
Questions?



http://www.mixOmics.org

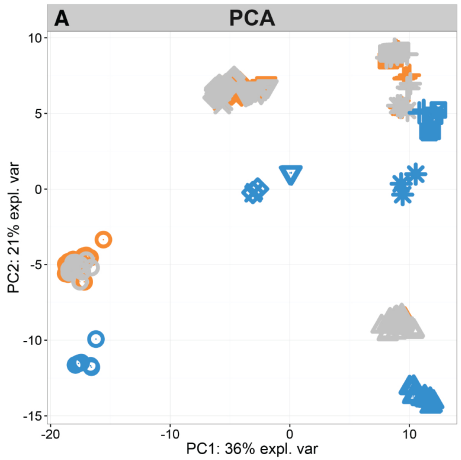
k.lecao@uq.edu.au
mixomics@math.univ-toulouse.fr

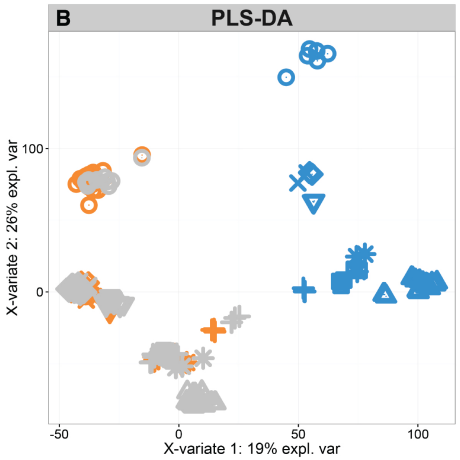
Sequential approaches vs MINT

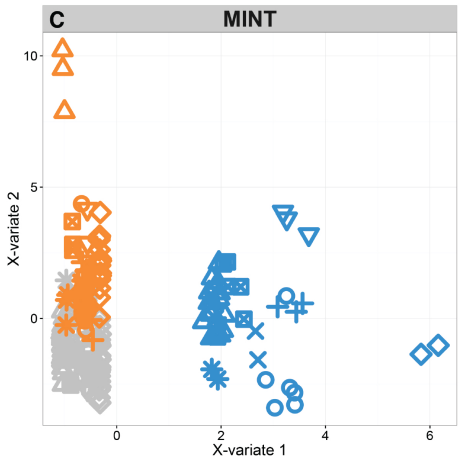


a. PLSDA
b. sPLSDA
c. RF
d. mgPLS
e. MINT
f. BMC+PLSDA
g. BMC+SPLSDA
h. BMC+RF
i. ComBat+PLSDA
j. ComBat+sPLSDA
k. ComBat+RF
l. LM+PLSDA
m. LM+sPLSDA
n. LM+RF
o. LMM+PLSDA
p. LMM+sPLSDA
q. LMM+RF

MINT on stem cell study







D

Study	BER	Fib	hESC	hiPSC
○ Bock	22.2	100	100	33.3
△ Briggs	0.00	100	100	100
+ Chung	15.0	100	75.0	80.0
× Ebert	11.1	100	100	66.7
◇ Guenther	2.0	100	94.1	100
▽ Maherali	11.1	100	66.7	100
⊠ Marchetto	0.00	100	100	100
✳ Takahashi	44.4	100	66.6	0.00
overall	7.1	100	91.9	86.7

www.stemformatics.org (Prof C Wells team)