How to analyze and combine multiple high-dimensional datasets?

Kim-Anh Lê Cao

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

THE UNIVERSITY OF MELBOURNE

Slides + tuto: www.mixOmics.org/presentations

My background

- Diplôme d'Ingénieur, Mathematics and Modelling, PhD in Applied Statistics (Toulouse, France)
- 2009: Postdoc at the University of Queensland
- 2014: Group Leader, UQ Diamantina biomedical institute, Translational Research Institute
- April 2017: Senior Lecturer, UoM, Centre for Systems Genomics

Toulouse, a.k.a 'La Ville Rose'

Research context: Systems biology

From reductionism:

1 gene = 1 hypothesis = 1 statistical test

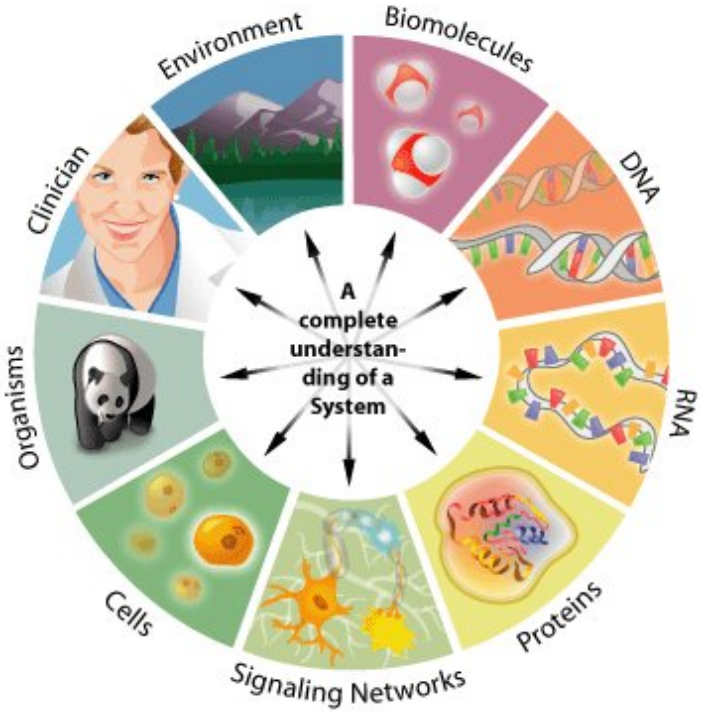
$\Downarrow$

To holism:

Thousands of molecules = ??

Biological aims:

- integrate data from different 'omics molecular levels to better understand a biological system
- postulate novel biological hypotheses to be validated in the lab



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 and explain variation in the data

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

A linear combination of variables

$$\mathbf{X} = \underset{(N \times P)}{}$$

	Var 1	Var 2
1	3.20	0.07
2	1.65	0.07
3	3.28	0.12
4	1.01	0.19
5	2.71	0.24
6	2.06	0.08
7	1.85	0.12
8	2.71	0.18
9	2.36	0.08
10	1.93	0.25

$$\mathbf{a} = \begin{pmatrix} 0.2 \\ 0.8 \end{pmatrix}$$

Assign to Variable 1 and Variable 2 the coefficients $a_1 = 0.2$ and $a_2 = 0.8$ respectively.

N = number of samples / indiv; P = number of variables; X = data matrix

	Var 1		Var 2		Linear combination
	3.20		0.07		0.69
	1.65		0.07		0.39
	3.28		0.12		0.75
	1.01		0.19		0.35
$0.2 \times$	2.71	$+ \quad 0.8 \times$	0.24	$=$	0.73
	2.06		0.08		0.48
	1.85		0.12		0.47
	2.71		0.18		0.69
	2.36		0.08		0.53
	1.93		0.25		0.59

- Linear combination (*component*) summarises the original data X.
- The weights in $\boldsymbol{a}$ (*loading vector*) defined according to a statistical criterion to optimize.
- Solved with Singular Value Decomposition algorithm for large data sets

Principal Component Analysis with a 'larger' P

data $=$
$(N \times P)$

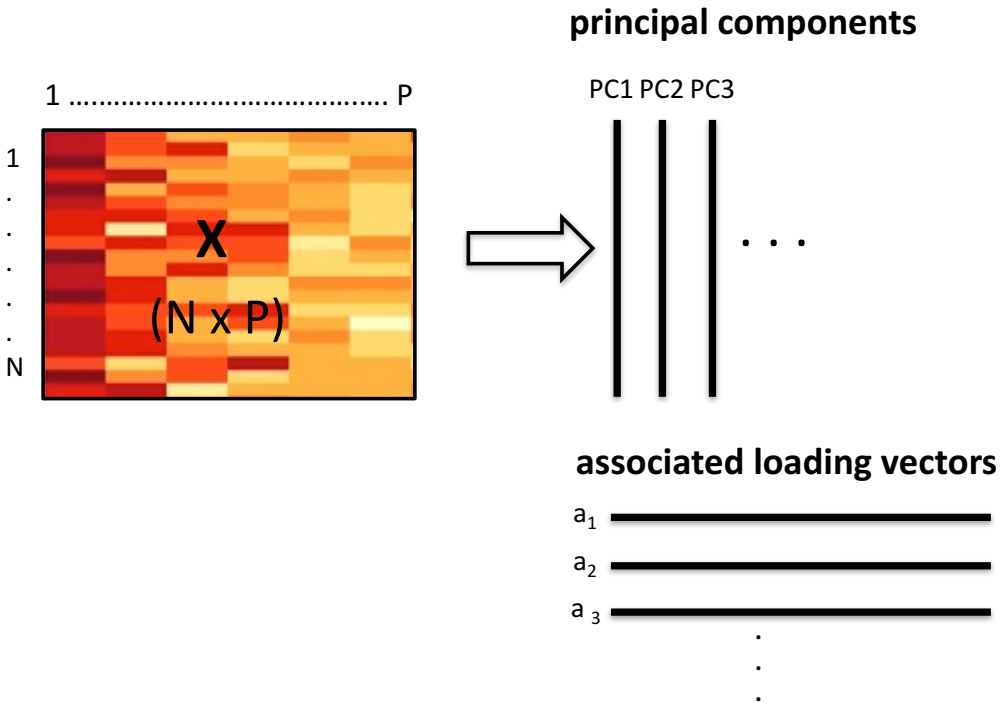
	Var 1	Var 2	Var 3	Var4	Var5
1	3.97	3.16	3.54	3.89	2.11
2	2.05	2.36	3.89	3.50	1.58
3	3.36	3.11	4.20	2.26	-0.09
4	4.15	4.79	5.43	3.30	2.00
5	2.91	3.05	3.87	4.99	1.69
6	3.44	3.96	3.54	3.89	2.11
7	3.65	2.22	3.89	3.50	1.58
8	4.40	3.60	4.20	2.26	-0.09
9	2.68	2.29	5.43	3.30	2.00
10	3.85	3.89	3.87	4.99	1.69

components $=$
$(N \times H)$

	Comp1	Comp2	Comp3	Comp4	Comp5
1	0.62	0.31	-0.45	0.71	-0.15
2	0.41	-1.53	0.05	-0.34	-0.47
3	-1.91	-0.53	-0.35	-0.39	-0.07
4	-0.33	1.72	1.31	-0.08	-0.11
5	1.36	-0.24	-0.38	-0.52	0.35
6	0.69	0.64	-0.29	0.02	-0.69
7	0.07	-0.71	-0.29	0.73	0.26
8	-2.16	0.46	-0.53	0.04	0.18
9	0.14	-1.10	1.45	0.13	0.27
10	1.11	0.97	-0.52	-0.31	0.42

- Components are orthogonal to ensure no overlap of information.
- Not all components are needed to summarise the information ($H << P$) $\rightsquigarrow$ data compression

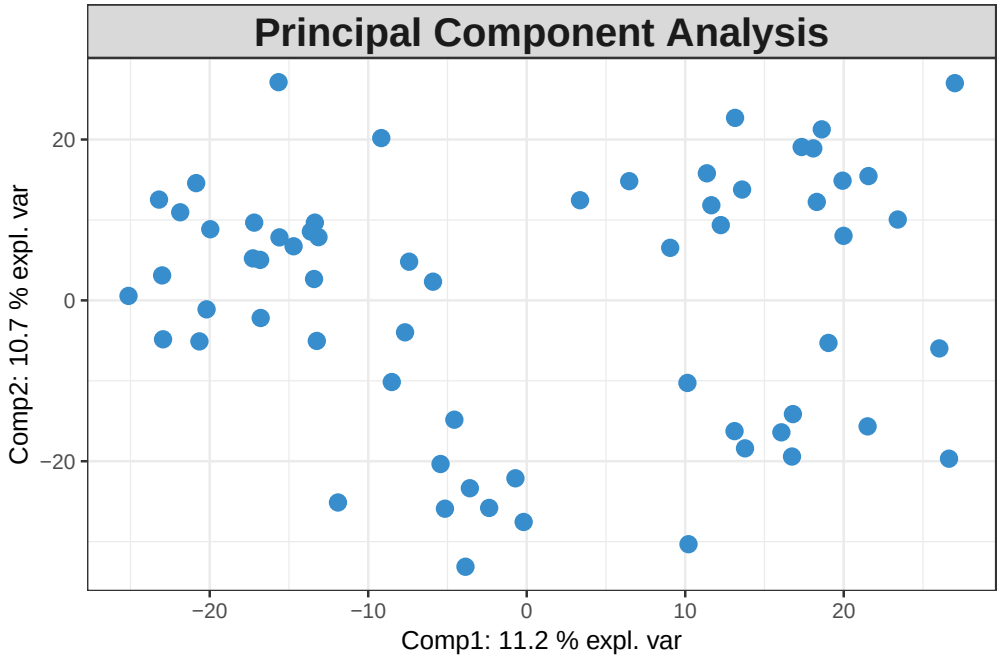
Data compression with PCA is a matrix decomposition



- Components are linear combinations of original variables, and orthogonal.
- Loading vectors indicate the weight (importance) of each variable in the linear combination.

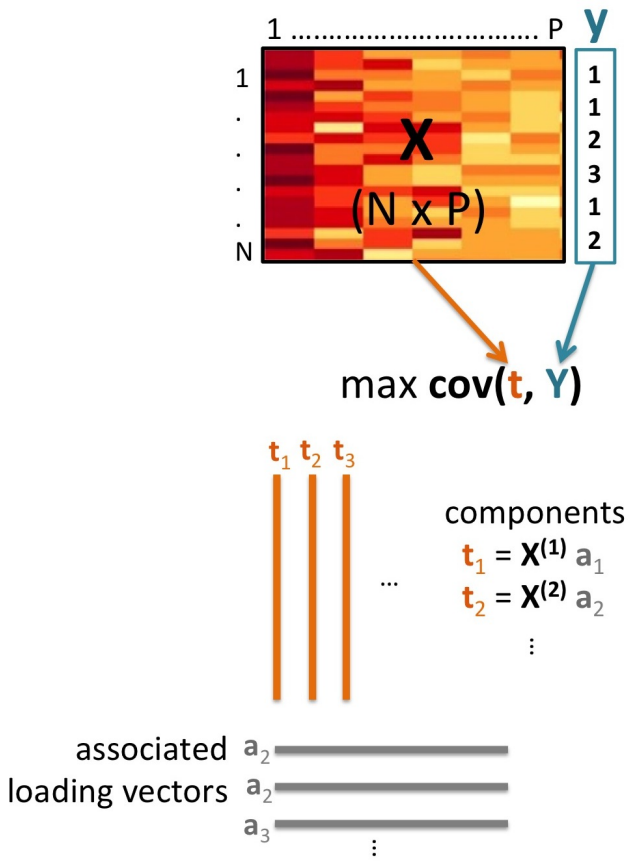
Visualisation: project data into a small subspace spanned by the components

PCA: visualisation of 63 samples x 2,300 variables (genes)



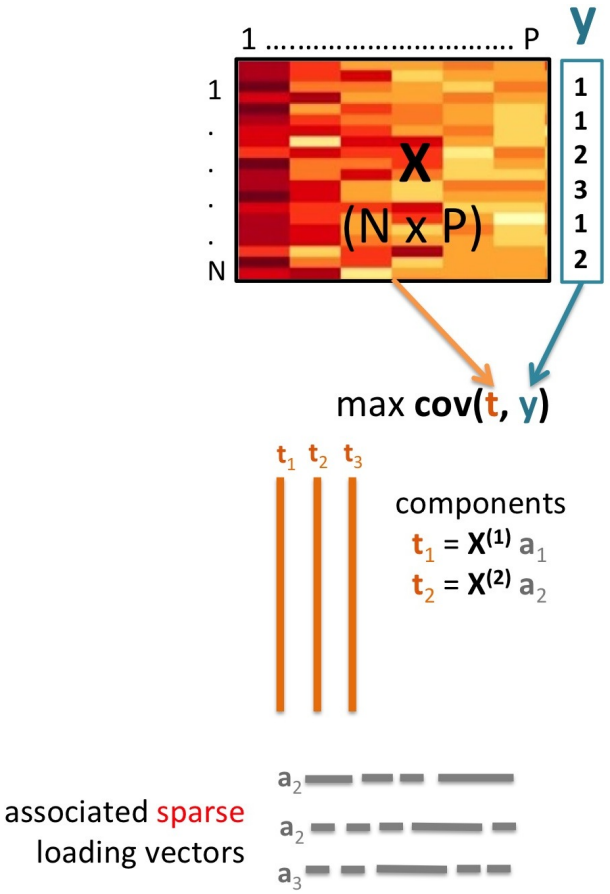
- 'Similar' samples (based on their variable values) cluster
- Unsupervised exploratory analysis: no information about sample groups included in the model

PLS-Discriminant Analysis (PLS-DA) includes sample group information



- covariance is an unstandardized version of the correlation
- decomposition of the data matrix X in relation with the outcome y with a set of components and loading vectors
  $\rightarrow$ dimension reduction

Further dimension reduction with variable selection

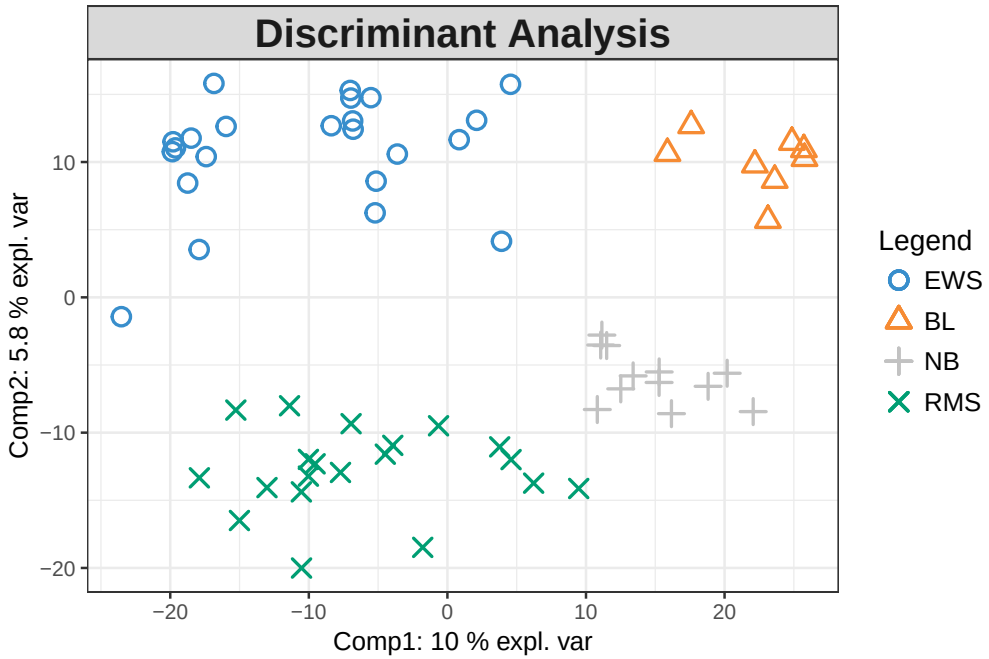


- set some variable coefficients to zero (variable selection) for each component
- each component is built on its selected variables only

Lê Cao et al. (2011). Sparse PLS Discriminant Analysis. *BMC Bioinformatics* 12:253.

Visualisation: project data into a small subspace spanned by the components

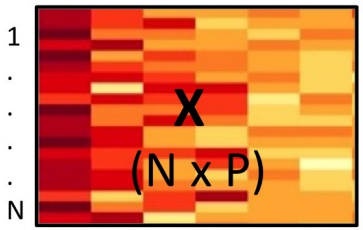
PLS-DA: visualisation of 63 samples x 2,300 genes according to their tumour subtype



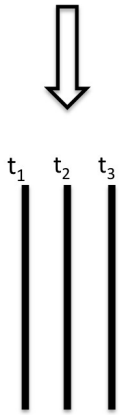
- Samples cluster according to their group
- Supervised analysis: the aim is to separate/discriminate sample

Multivariate methods achieve matrix decomposition

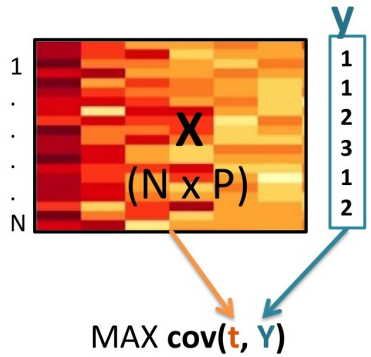
PCA maximises the **variance**



MAX **var(t)** components



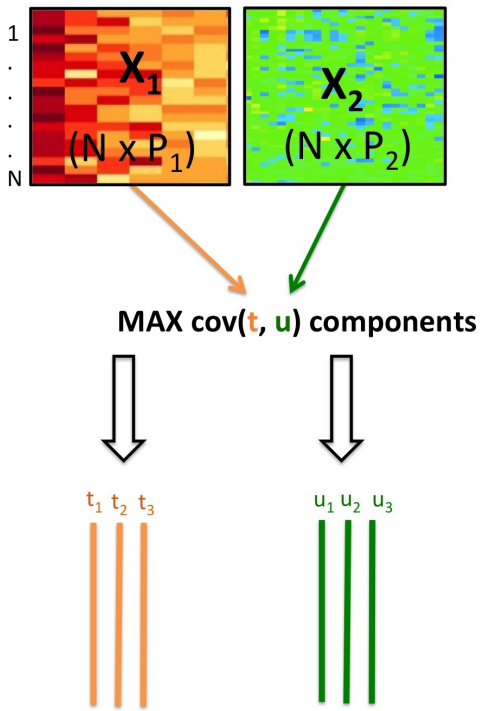
PLSDA maximises the **covariance** with a categorical outcome



MAX **cov(t, Y)**

Other variants for data integration

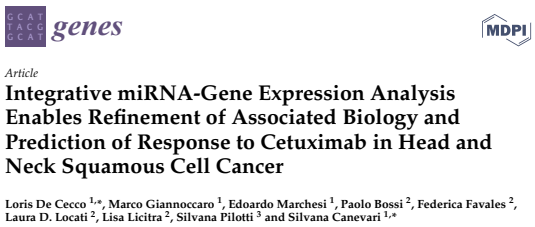
PLS maximises the **covariance** between two data sets



Wold (1966). Estimation of principal components and related models by iterative least squares. *In Krishnaiaah, P.R. Multivariate Analysis.*

Other variants for data integration

Examples of 2 'omics integration case studies using sparse PLS[1]


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,*]

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,*]

ARTICLE

doi:10.1038/nature16942

Plankton networks driving carbon export in the oligotrophic ocean

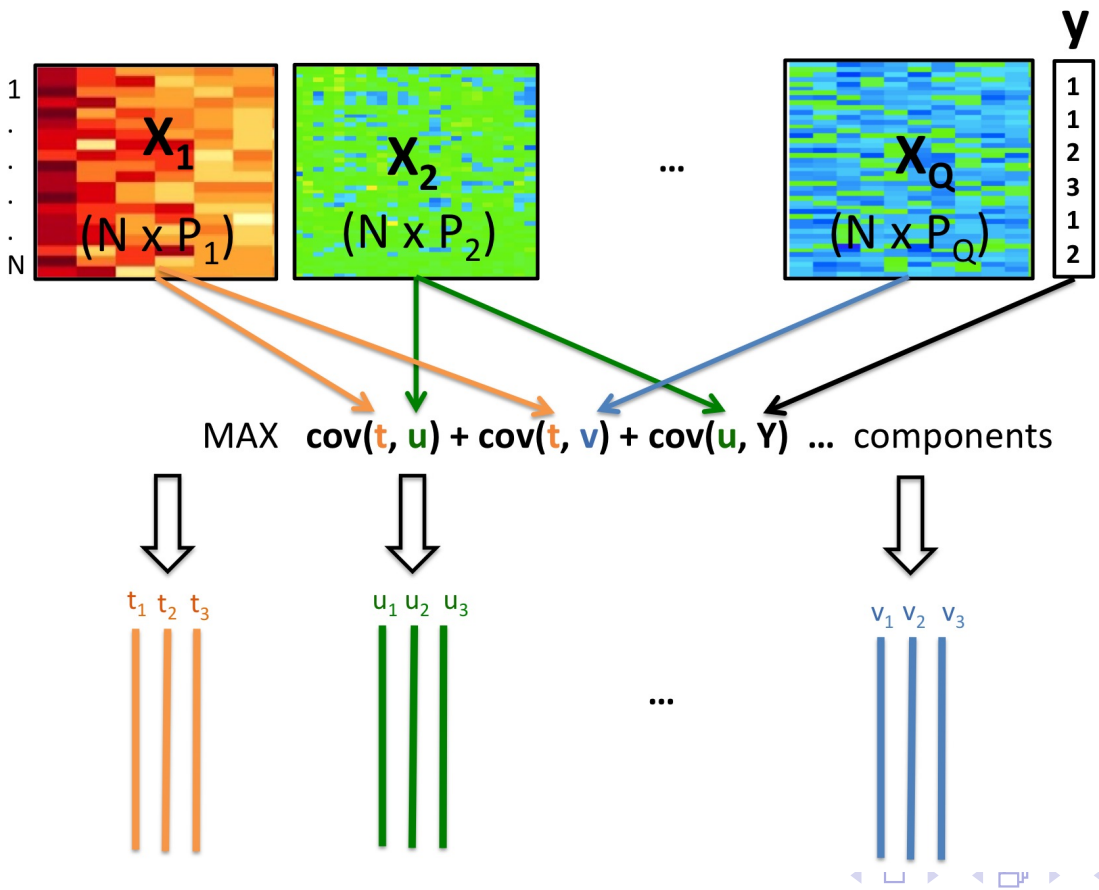
Lionel Guidi[1,2*], Samuel Chaffron[3,4,5*], Lucie Bittner[6,7,8*], Damien Eveillard[9*], Abdelhalim Larhlimi[9], Simon Roux[10†], Youssef Darzi[3,4], Stephane Audic[8], Léo Berline[1†], Jennifer R. Brum[10†], Luis Pedro Coelho[11], Julio Cesar Ignacio Espinoza[10], Shruti Malviya[7†], Shinichi Sunagawa[11], 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[1,4], 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]

$\Longrightarrow$ we recently extended sparse Generalised Canonical Correlation Analysis[2] for multiple data sets integration[3] (> 2 data sets).

[1] Lê Cao et al. (2008) A Sparse PLS for Variable Selection when Integrating Omics data, *SAGMB* **7**(1).
[2] Tenenhaus A, Lê Cao K-A. et al. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics* **15**(3).
[3] 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.

Other variants for data integration

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



Different analyses to answer different types of questions

	One data set	**Multiple data sets** measured on the same samples
Unsupervised (No response Y provided)	**Data exploration, data mining, dimension reduction & visualisation** Select a subset of correlated variables	
Supervised classification (Categorical Y response)	Select a subset of discriminative variables	

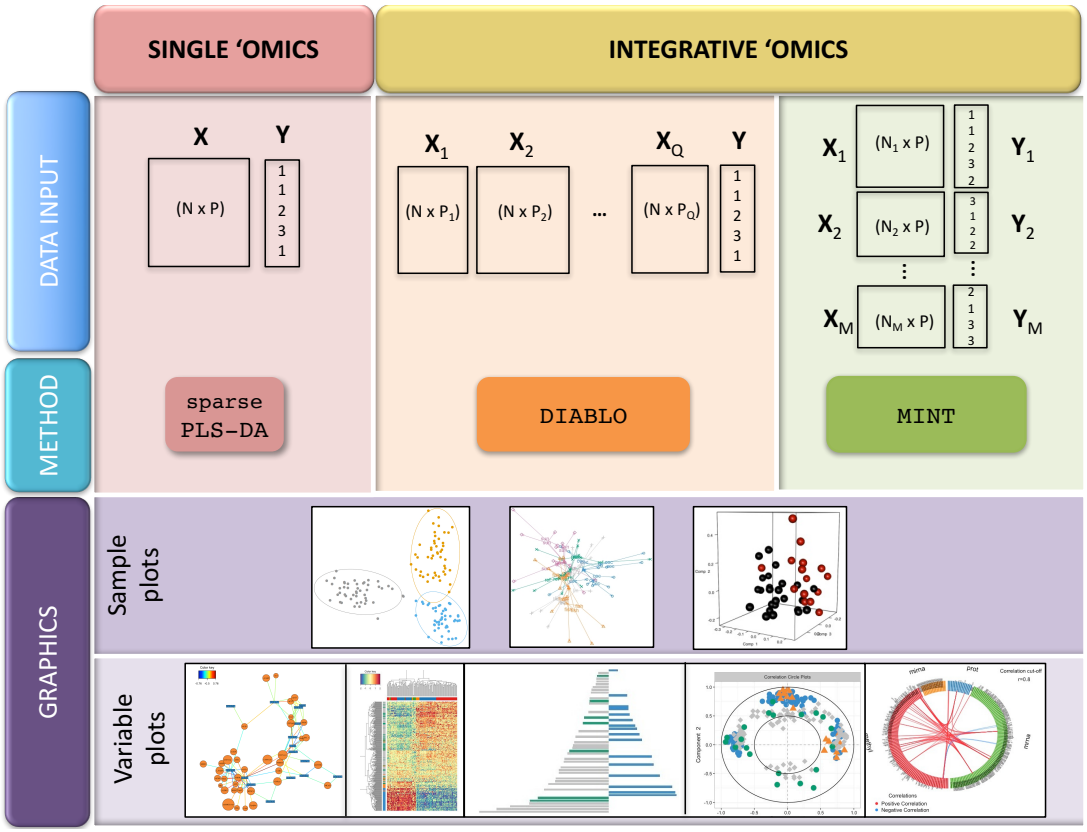
$\rightarrow$ my research focus is to make sense of large scale data via dimension reduction and identification of subset of key variables in these large data sets.

Univariate and multivariate statistics are complementary

Univariate (e.g. t-test/linear model)	**Multivariate (e.g. PLS-DA)**
• Data distribution assumption	• Relaxed / no assumptions
• Assess one variable at a time	• All variables at once
• p-value per variable	• No p-value, unless permutations
• No visualisation	• Visualisations (samples, variables) based on components
• Adjust for covariates	• Not possible (in progress)
• Prediction generally not possible if $P > N$	• Phenoype prediction for a new sample

Sparse multivariate methods identify a small subset of key variables and make downstream univariate analysis amenable.

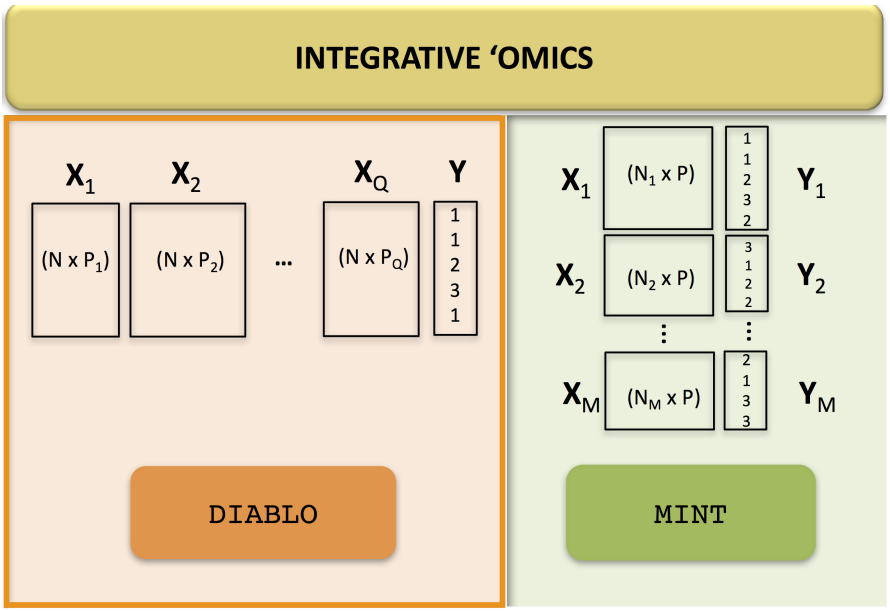
Supervised 'omics data analysis framework in mixOmics



⇝ Visualisations based on components and selected features

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

Integrating 'omics data with mixOmics: challenges



- Small sample size (50-100), large number of variables ($>$ 1,000)
- Unwanted systematic variation due to differences between platform, labs, or protocols!
- N-integration: data sets measured on the same N samples
  P-integration: data sets measured on the same P variables from independent studies (not covered today)

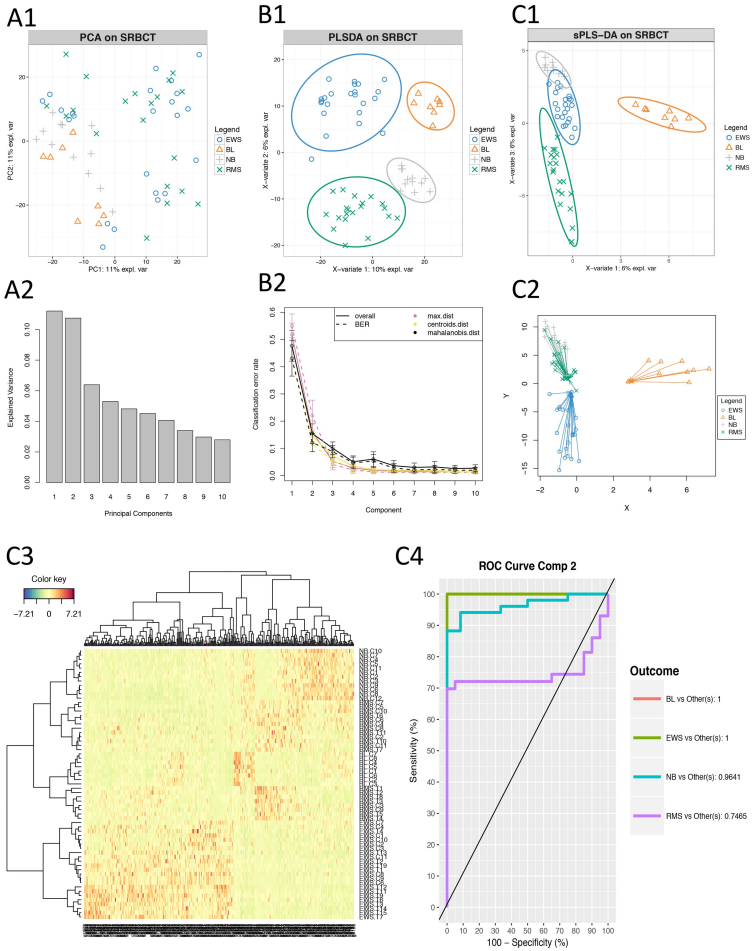
$\rightsquigarrow$ multivariate integrative methods project each data set into a space spanned by the components that model common information between data sets.

Linear multivariate methods

- Seek for components summarising the information from the data
- Each component is a linear combination of the original variables
- LASSO penalties to perform *variable selection* ('sparse' 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

Single data set analysis with PLS-DA

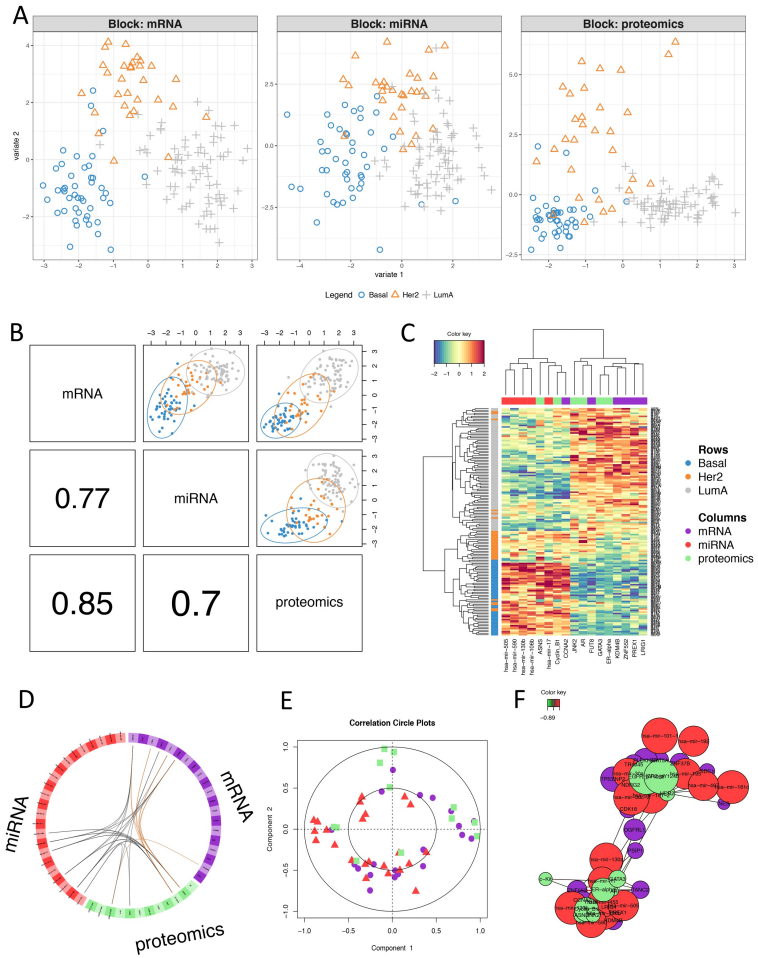


Small Round Blue Cell Tumors (SRBCT) of childhood data:

- 63 tumours
- Data matrix 63 x 2308 genes
- 4 tumor subtypes
- PCA, PLS-DA and sPLS-DA

Full example: http://mixomics.org/case-studies/srbct-example/

N-integration of multiple data sets

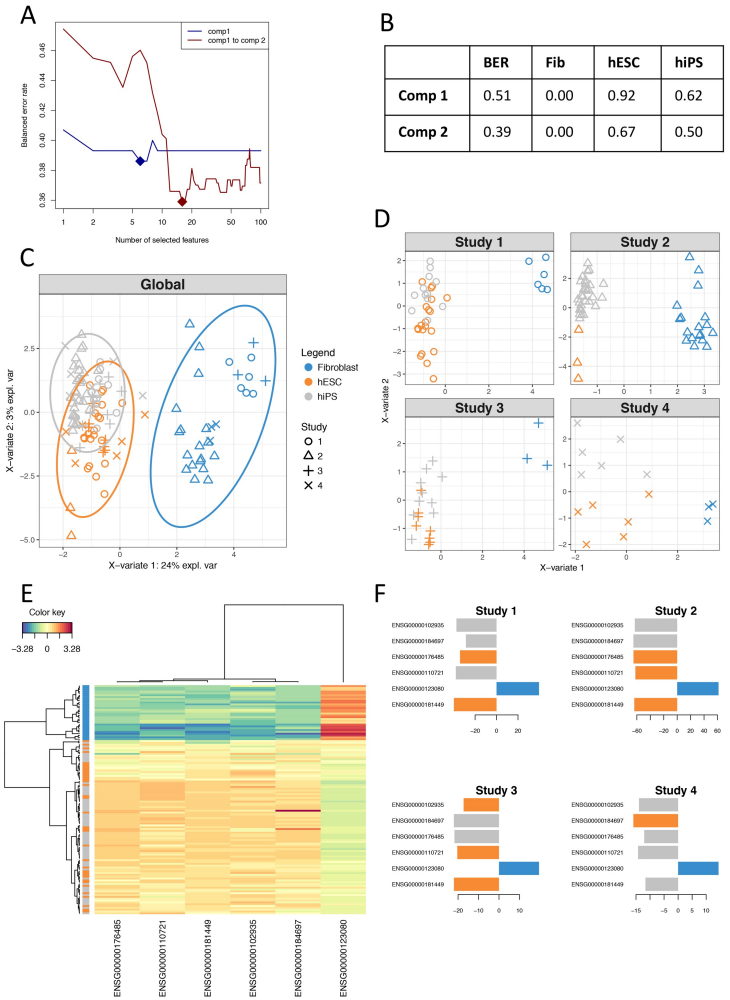


Breast Cancer from TCGA:

- subset of 150 individuals
- 3 data sets: mRNA (200), miRNA (184) and proteomics (142)
- 3 tumor subtypes
- DIABLO (`block.splsda`)

Full example: http://mixomics.org/tcga-example/

P-integration of multiple independent data sets



	BER	Fib	hESC	hiPS
Comp 1	0.51	0.00	0.92	0.62
Comp 2	0.39	0.00	0.67	0.50

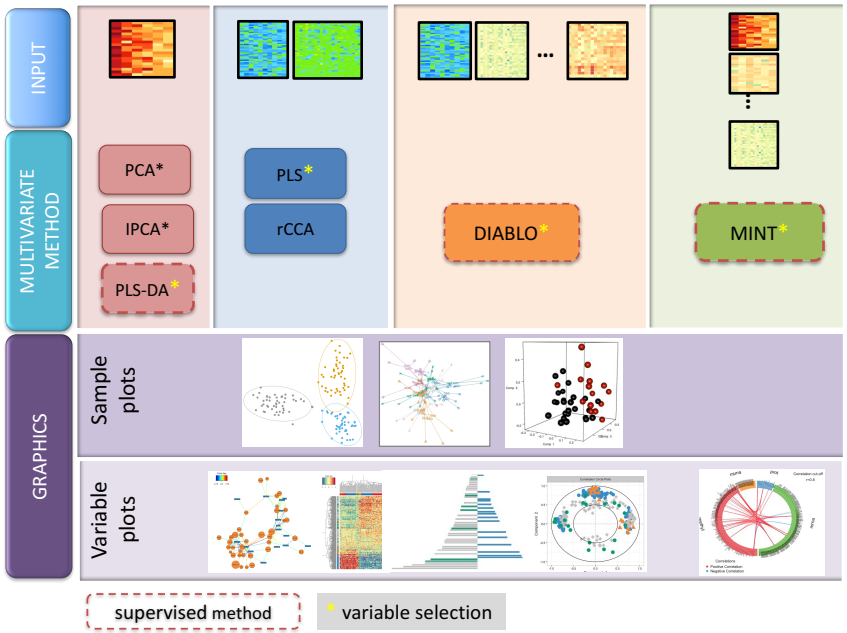
Stem cells from Stemformatics:

- subset of 125 stem cells
- 4 individual studies (38, 51, 21 and 15 cells each)
- 3 cell types
- MINT (`mint.splsda`)

See: http://mixomics.org/stemcells-example/

Vision & philosophy

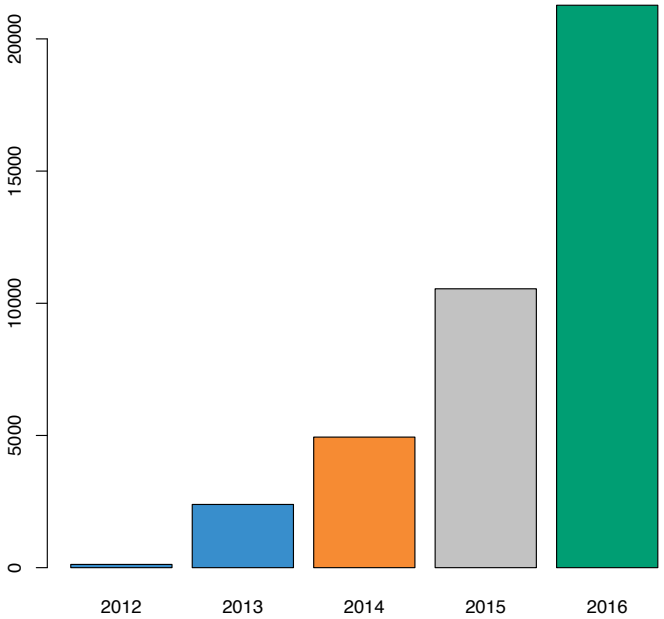
An R toolkit for multivariate data analysis of biological 'omics data, with a specific focus on data integration, feature selection and visualisation.



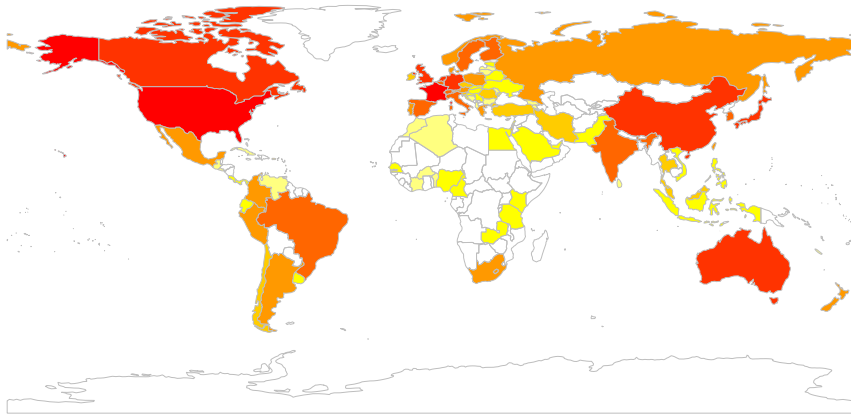
Today: 17 multivariate methods (12 are our own methods)

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

Tracking down interest: unique IP downloads



 # downloads



 # downloads per country for Sem 2 2016

 # downloads per package version for Sem 2 2016

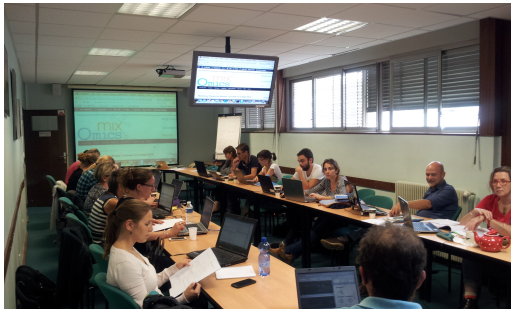
Version	5.0-1	5.0-2	5.0-3	5.0-4	5.1.2	5.2.0	6.0.0	6.1.0	6.1.1
# dl	17	1	108	32	2	1054	5066	2648	3760

Dissemination

Website: www.mixOmics.org, now organised in different modules

Dissemination

Hands-on workshops by invitation, except for conferences.



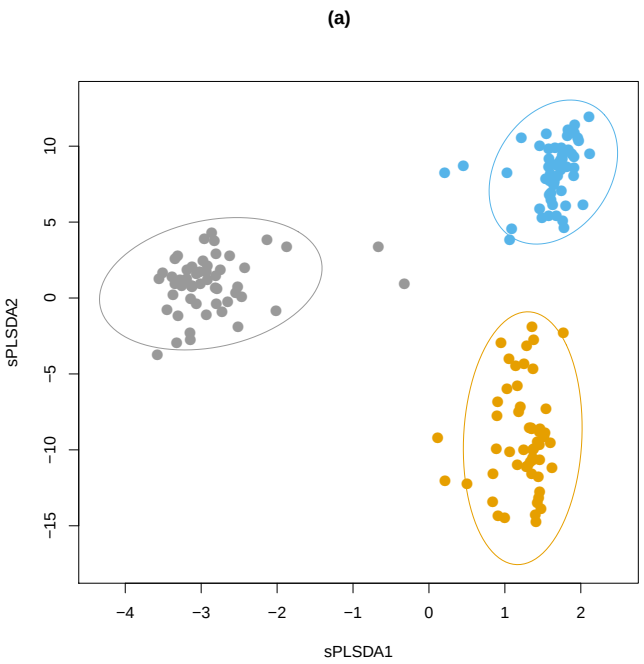
Workshop trainings

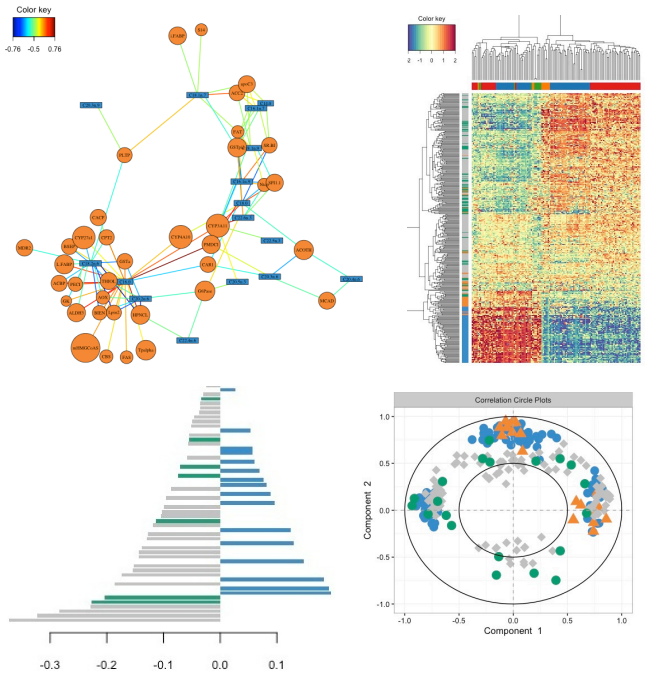
(1, 2 and 3-day workshops, lecture material ∼ 200 pages):

	2014	2015	2016	2017	2018
# workshops	2	5	5	3	2
# days	3	10	9	7	4
# attendees	46	135	96	76	60
country	FR	NZ, FR	FR, AU	FR	SP, AU

Branding

Logo, a set of color-blind friendly color palette, flyers, goodies





How does it work? Rely on the community

- Small team: 3 core members, 1 developer, contributors
- Collaborators / advanced users
- Open access and open community: R CRAN, bioRxiv papers and website early release
- Project management: bitbucket (git) and branching
- Advice / consulting: $\sim$ 5 - 10 emails / week (often new users)
- Main ambassadors: collaborators, attendees, early adopters
- Debug: bitbucket open issue

Bitbucket open issue: our issues are transparent!

```
> library(mixOmics)

Loaded mixOmics 6.1.3

Visit http://www.mixOmics.org for more details about our methods.
Any bug reports or comments? Notify us at mixomics at math.univ-toulouse.fr or https://bitbucket.org/klecao/package-mixomics/issues

Thank you for using mixOmics!
```

Kim-Anh Le Cao / package-mixOmics

Create issue

Issues

FILTERS: All | Open | My issues | Watching

Advanced search | Find issues

Issues (1–19 of 19)

Title	T	P	Status	Votes	Assignee	Created	Updated
#98: Issue with variable's label when exporting network			NEW			2017-06-23	2017-06-23
#97: Suggestions for Mfold, relevance network, explained variance, and general help.			NEW			2017-06-16	2017-06-16
#96: Error in perf function returned when running "loo"			NEW			2017-06-16	2017-06-16
#89: constant variable in block.spls with scale = TRUE			NEW			2017-05-12	2017-06-13
#87: prediction of new sample by Diablo with several components			NEW			2017-04-11	2017-06-13
#95: cim in block.pls			NEW			2017-06-12	2017-06-12
#90: network: correlation with missing values			NEW			2017-05-12	2017-05-12
#84: GCCA/Diablo versus MB-PLS			NEW			2017-03-02	2017-05-08

Where we can further improve

Our main cruxes:

- A key mixOmics publication (almost there)
- The R programming language
- Us (statisticians)
- Funding for software development

On our to-do list:

- Bioconductor reconversion
- Shiny web-interface for non-R users and non-R workshops

My tips for a successful R package

- A good name
- Keep it small and simple (3-10 functions), or gigantic (150+ functions)
- Stick to your philosophy when expanding
- Align the package with your research /activities
- Upgrade on managing and Business Development skills
- Be user-oriented (approachable, available, considerate)
- If you cannot fund a developer, ensure the knowledge remains within the core manager (you)
- Build a strong, motivated, supportive core team.

Conclusions

- Flexibility of multivariate linear methods:
  $\rightarrow$ data exploration; classification; integration of multiple data sets
  $\rightarrow$ provide a first level of understanding of a biological system
- Variable selection suggest novel hypotheses, downstream statistical analyses & experimental validations.

- `mixOmics` development and vision has been instrumental in the dissemination of my research and has enabled many others to make sense of their large datasets.

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

Wells lab: Stemformatics, Christine Wells (UQ/UoM)

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

two senior postdocs, computational biology / statistics, University of Melbourne

Australian Government
Australian Research Council

Australian Government
National Health and Medical Research Council