Multi-omics data integration with mixOmics

A/Prof. Kim-Anh Lê Cao

NHMRC Career Development Fellow
Melbourne Integrative Genomics
School of Mathematics and Statistics

THE UNIVERSITY OF MELBOURNE

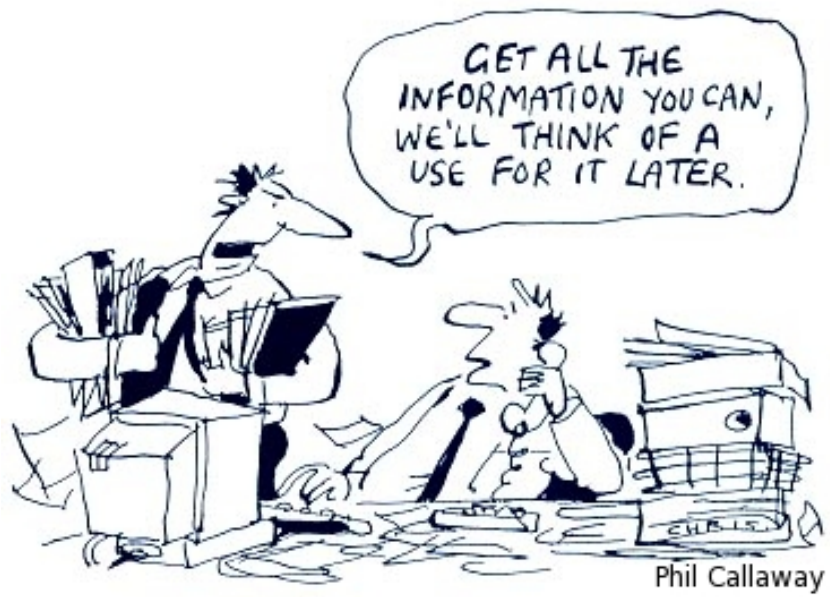
@mixOmics_team | www.lecao-lab.science.unimelb.edu.au

When biology and statistics meet

"Data don't make any sense,
we will have to resort to statistics."

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

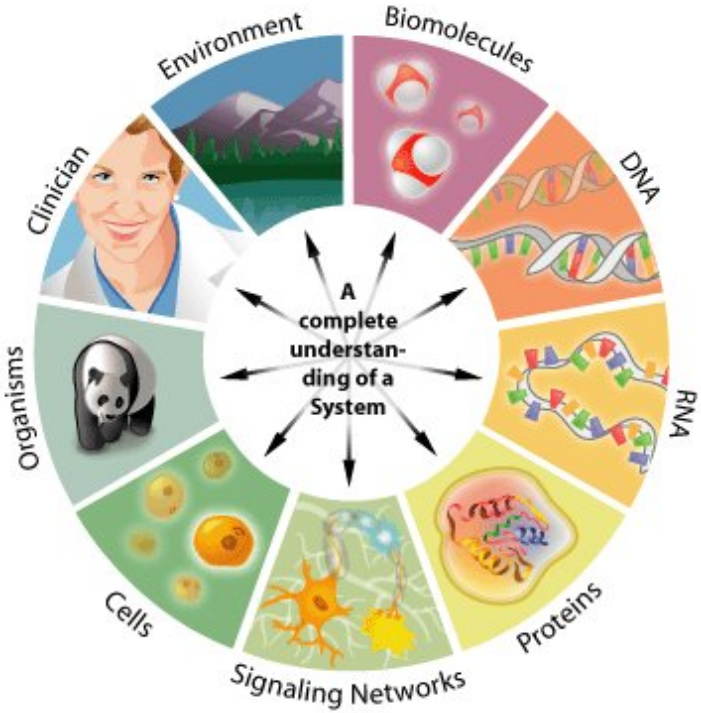
A holistic view of a biological system

From reductionism ...

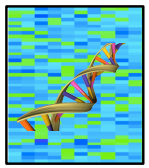
1 gene = 1 hypothesis = 1 statistical test

... to holism:

Thousands of molecules = ??

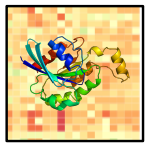


The 'omes and the biological dogma? not that straighforward



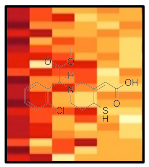
Transcriptomics
$\sim 100,000$ transcripts

↓

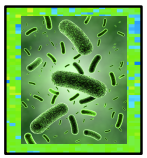


Proteomics
$\sim 1,000,000$ proteins

↓



Metabolomics
$\sim 2,500$ compounds

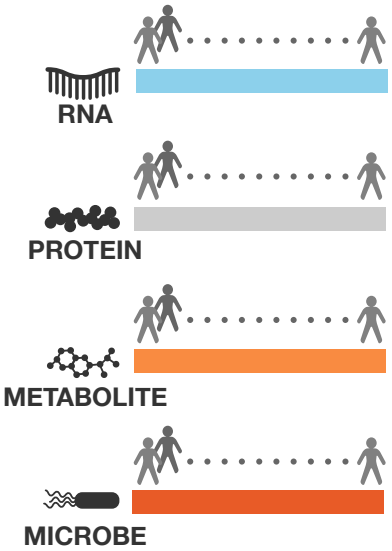


Microbiome
$\sim 1,500$ trillion microbes

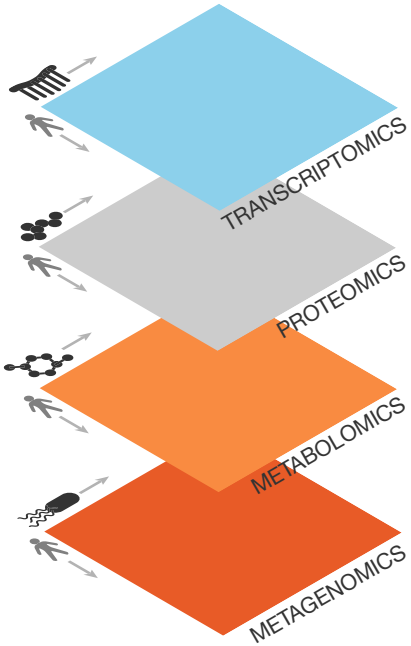
⋮

Clinical data
many

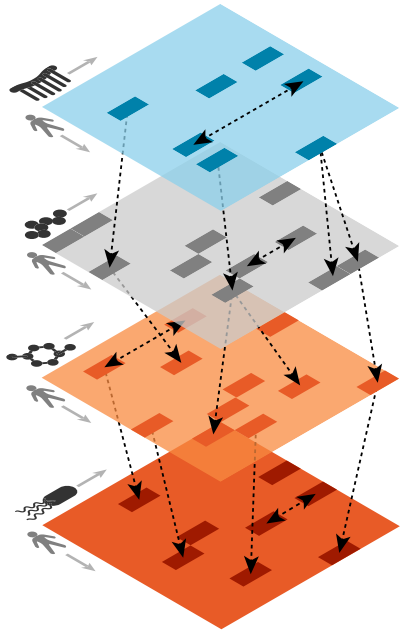
A new research field to establish



Conventional molecular biology
(Reductionist)

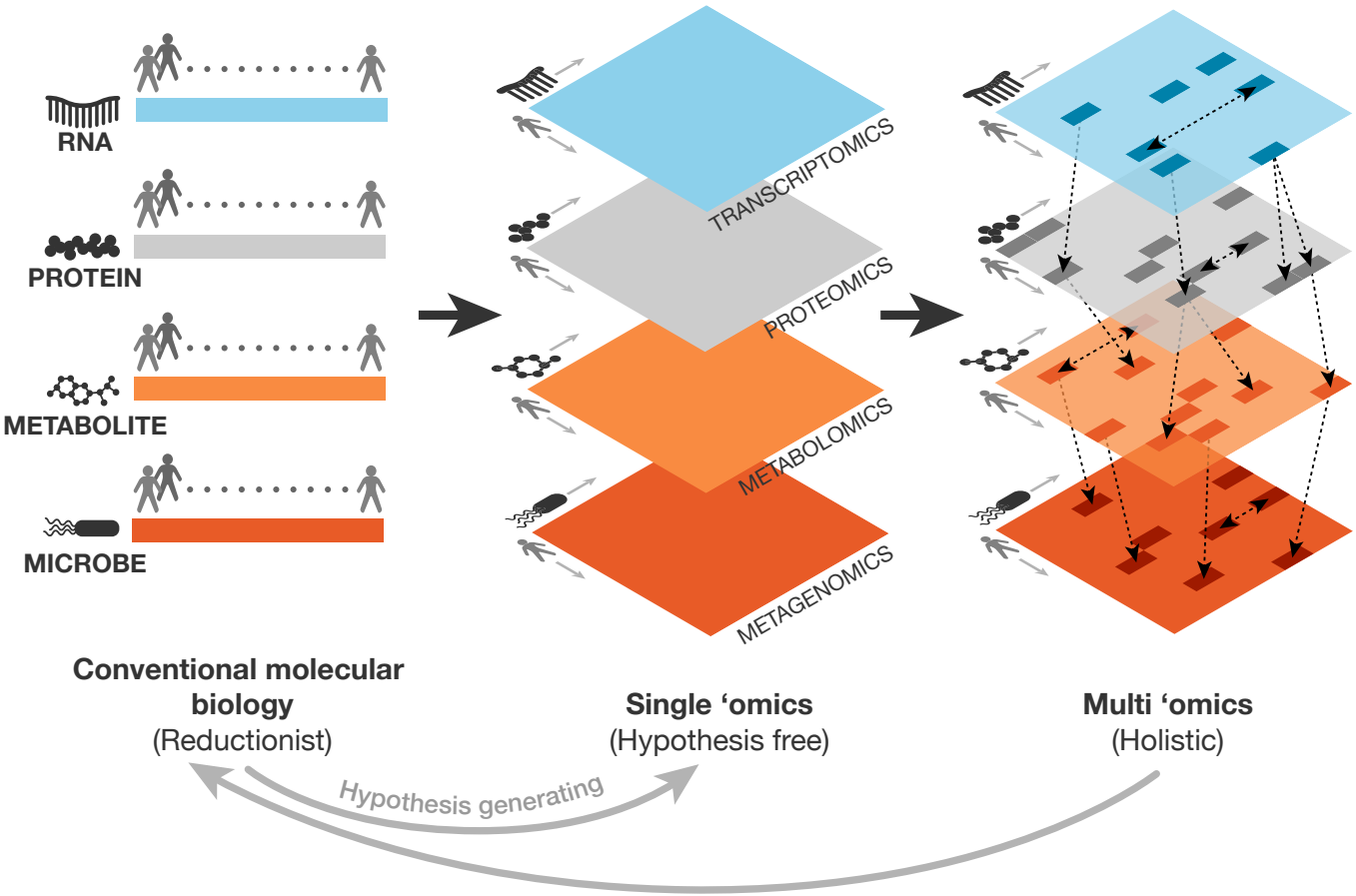


Single 'omics
(Hypothesis free)



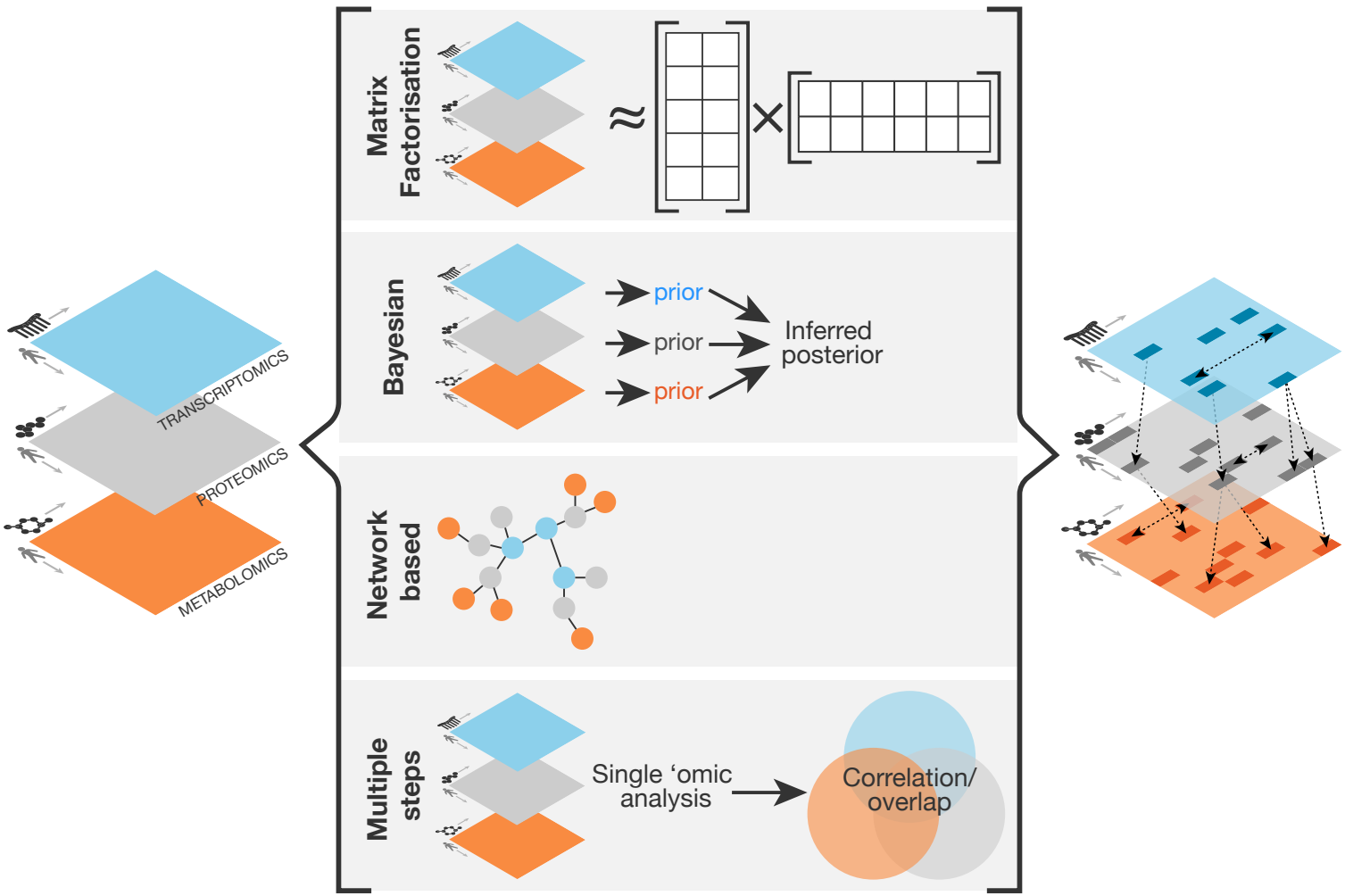
Multi 'omics
(Holistic)

A new research field to establish



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



Aims

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

Multivariate analysis:

- 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

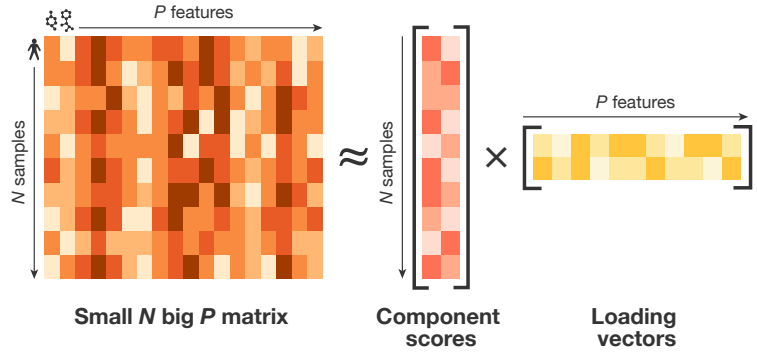
Expectations from 'omics data integration

- ☐ An overview of the role of each 'omics in a biological system
- ☐ A better understanding of the relationship between 'omics types
- ☐ A molecular signature with insights into molecular mechanisms
- ☐ A predictive analytical model
- ☐ All of the above!

Matrix factorisation methods

Build components that aggregate observable *variables* (e.g. genes, transcripts, proteins) to summarise sources of variation in the data

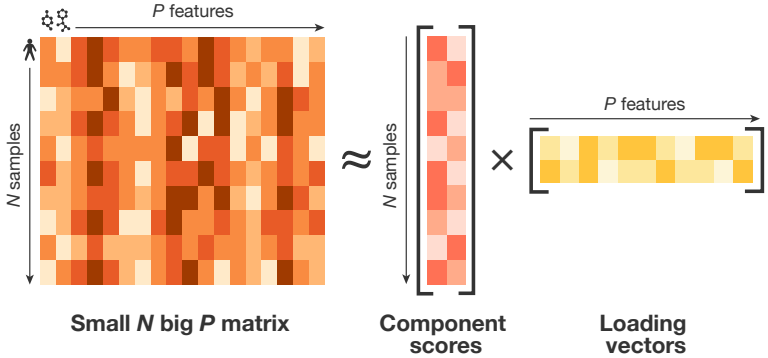
- Reduce data dimension
- Handle data characteristics (difficult otherwise with conventional statistical methods)
- Capture and explain variation (information) in the data



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

Matrix decomposition & dimension reduction

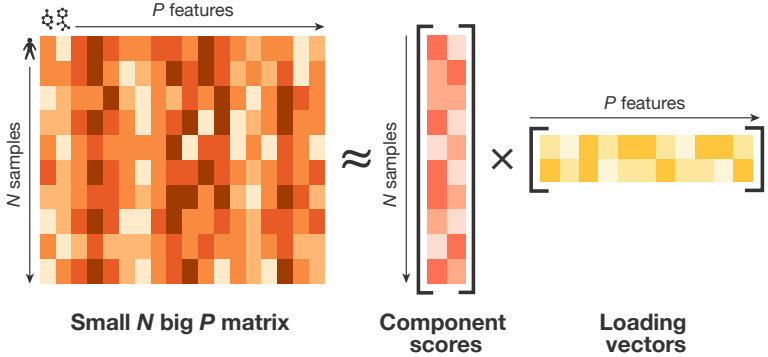
PCA: Data exploration



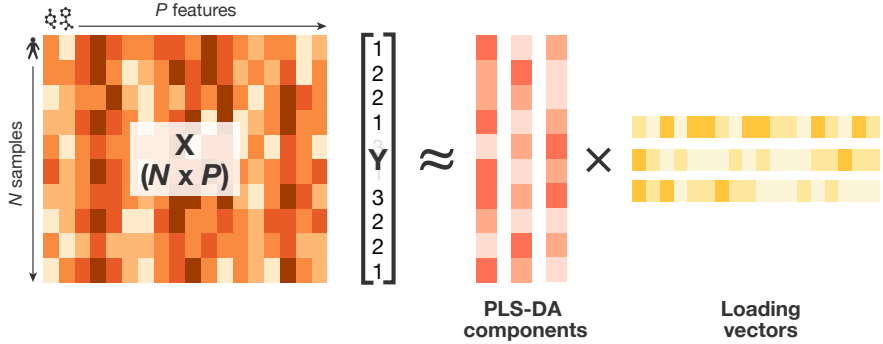
- Component scores are linear combination of features ($\ll P$)
- Loading weights indicate the importance of each feature in the linear combination
- Feature selection: weights shrunk to zero using Lasso penalisation

Matrix decomposition & dimension reduction

PCA: Data exploration



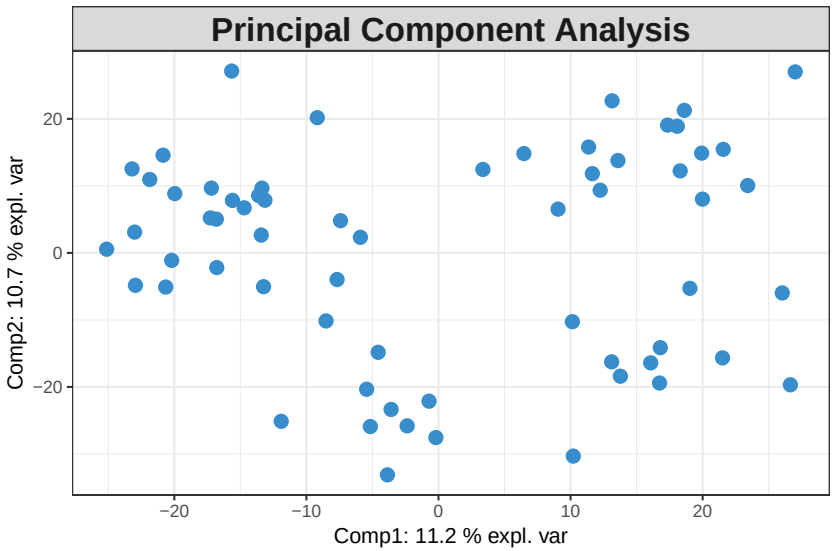
PLS - Discriminant Analysis

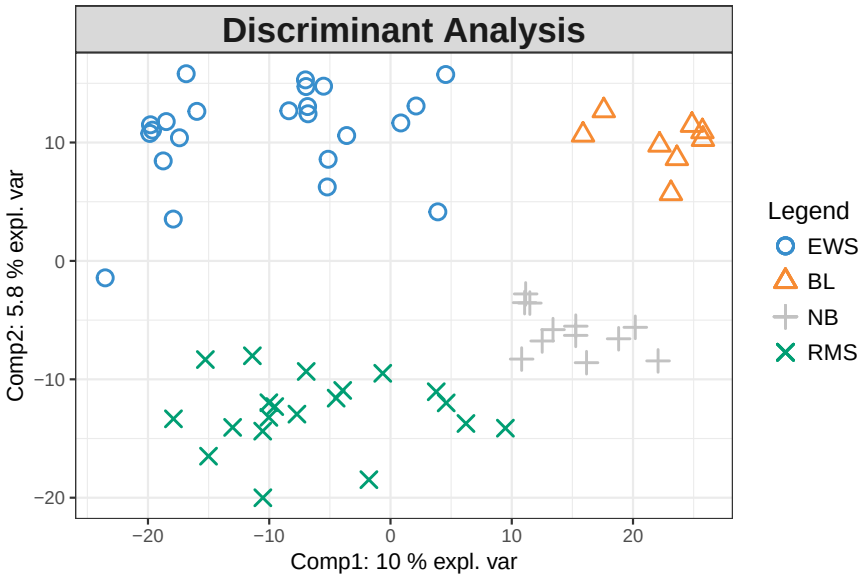


- Component scores are linear combination of features ($\ll P$) that discriminate the sample groups
- Loading weights indicate the importance of each feature in the discriminative linear combination
- Feature selection: weights shrunk to zero using Lasso penalisation

Sample visualisation in a reduced space

Example: SRBCT data with 63 samples and 3,116 genes

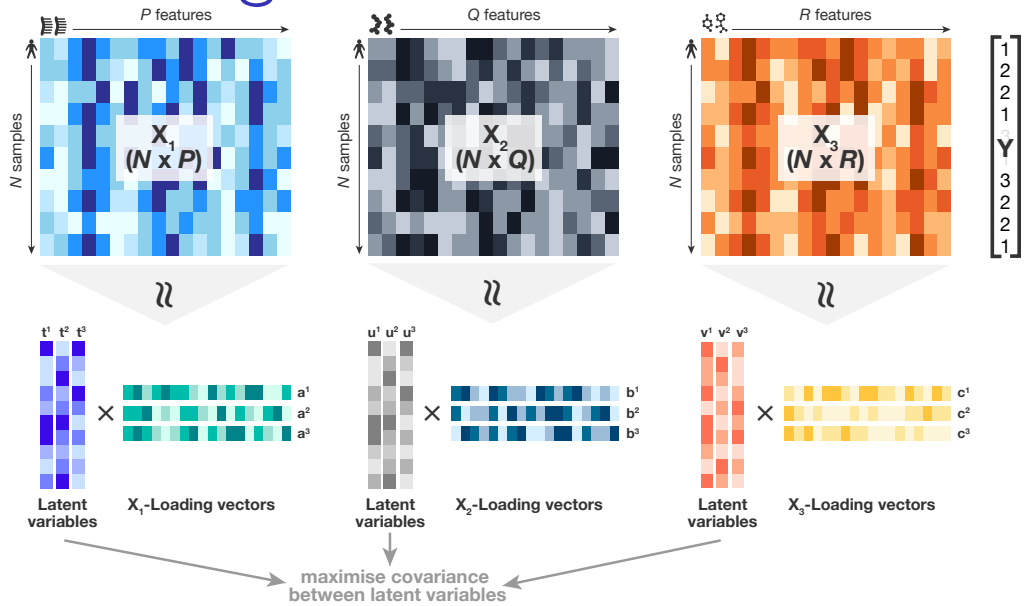




- Unsupervised exploratory analysis: 'similar' samples cluster but no information about sample groups is included in the analysis

- Supervised analysis: Samples cluster according to their respective group

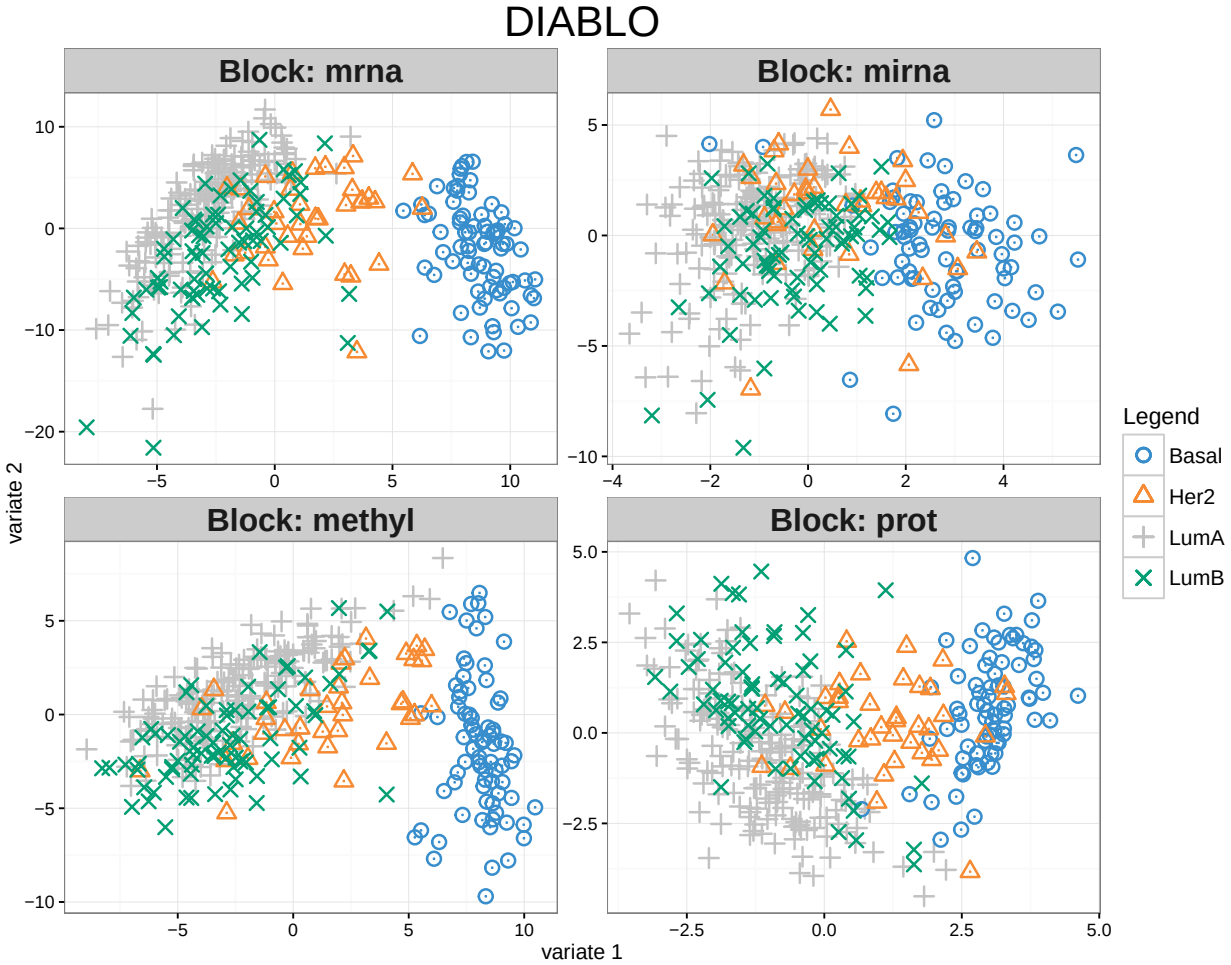
Multi-omics data integration with DIABLO



- Component scores are linear combination of a subset of selected features that are highly correlated across omics
- Loading weights indicate the importance of each selected feature in the linear combination

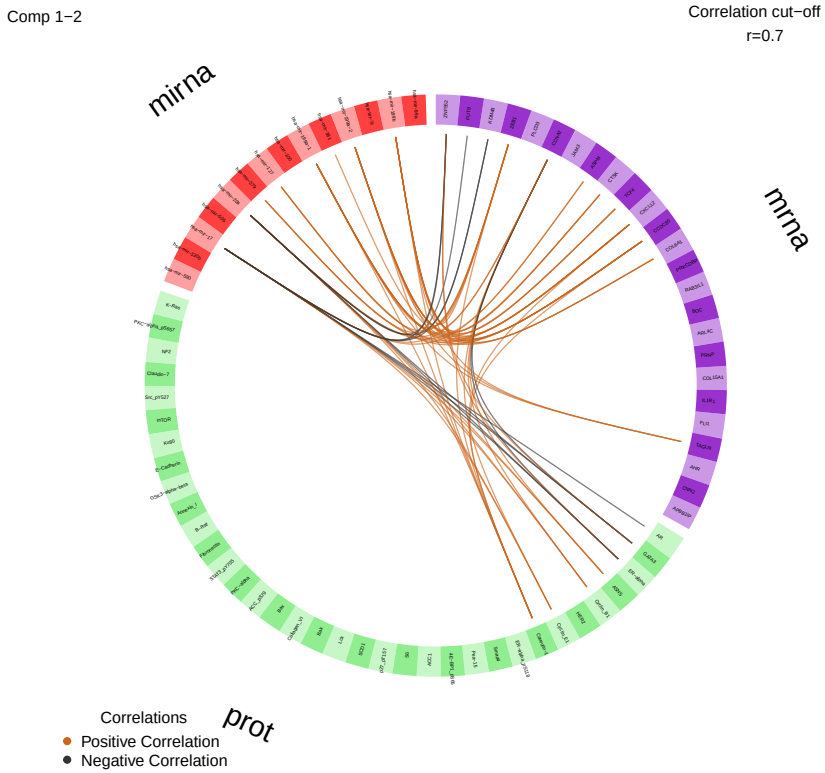
Singh A, Gautier B, Shannon C, Vacher M, Rohart F, Tebbutt S, Lê Cao K-A (2019). DIABLO: identifying key molecular drivers from multi-omic assays, an integrative approach. *Bioinformatics* 35 (17).

DIABLO: sample visualisation

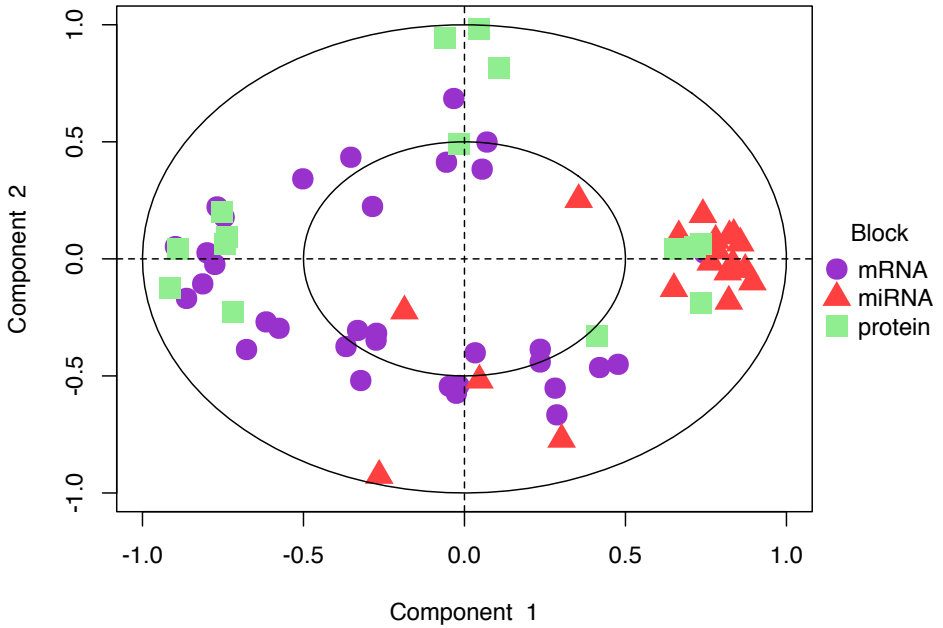


Consensus plot also possible

DIABLO: multi-omics signatures

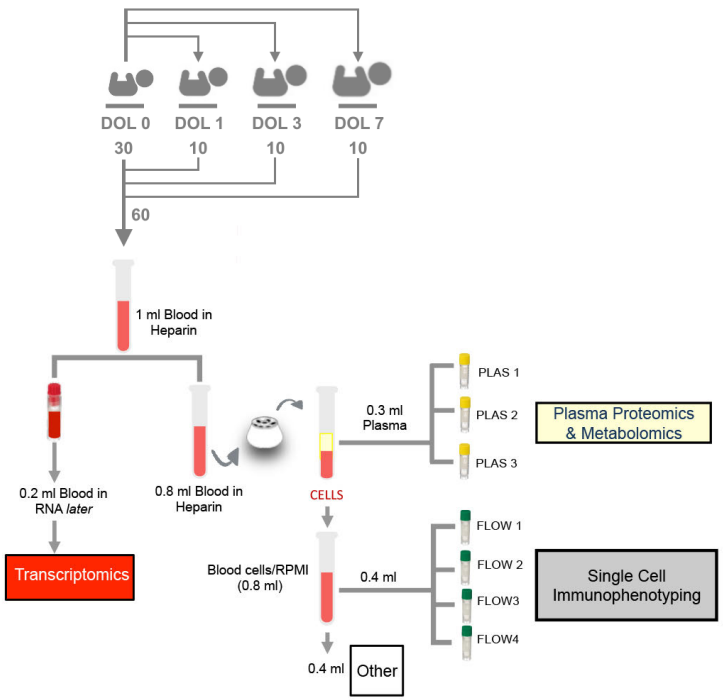


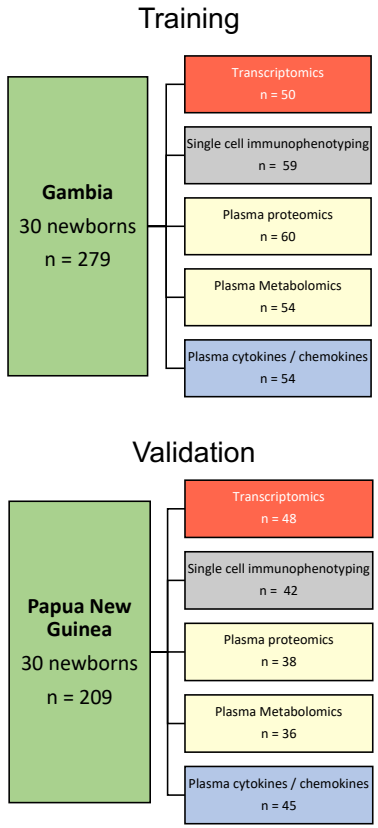
Circos plot



Correlation circle plot

#smallbig study: the first week of human life

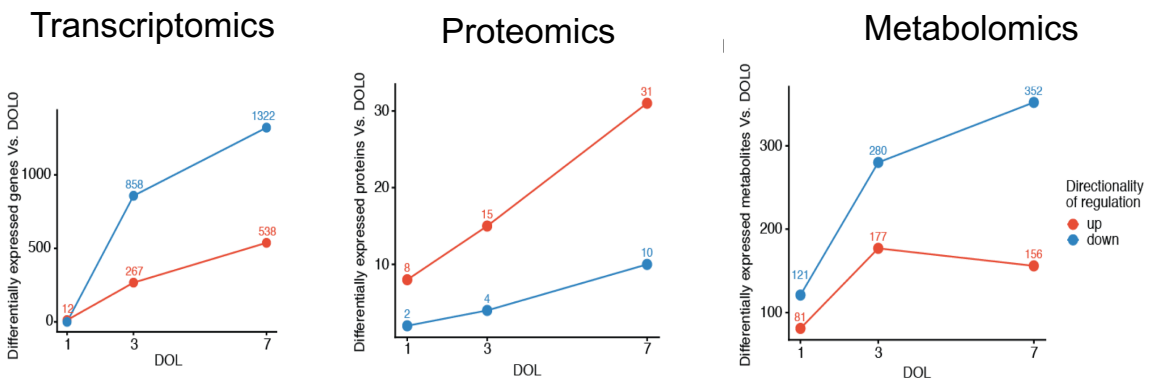


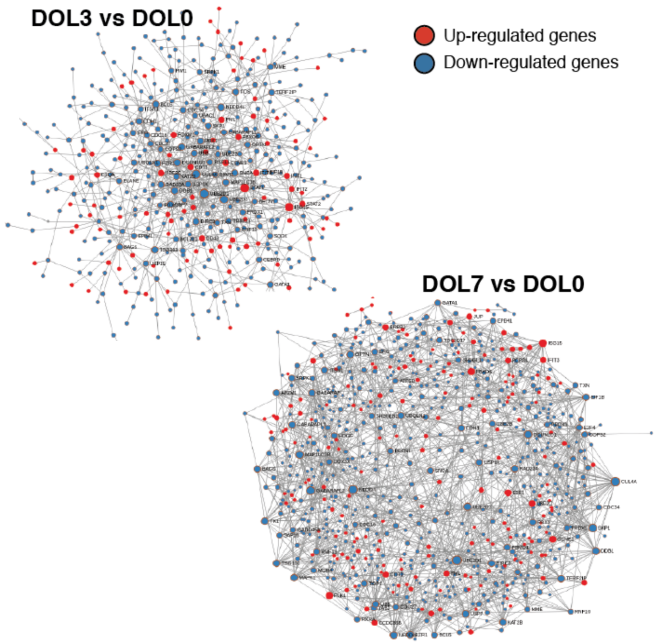


Small biosample amount: < 1ml of blood from newborns, five data types

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

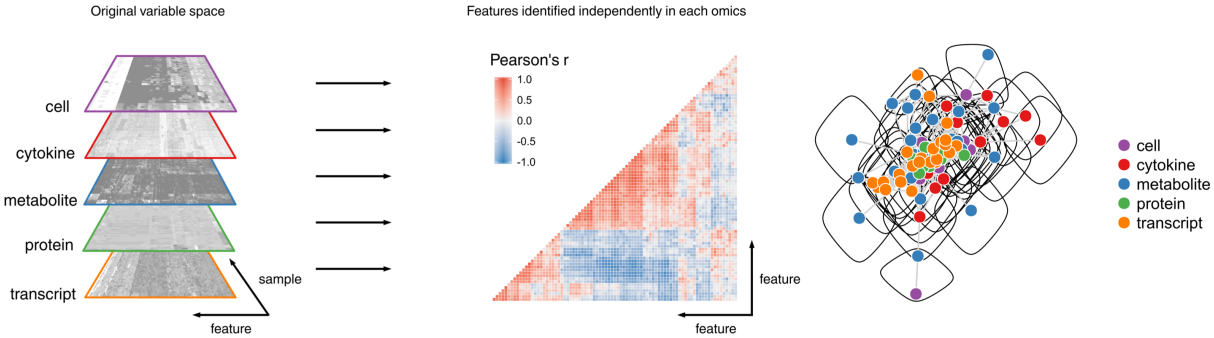




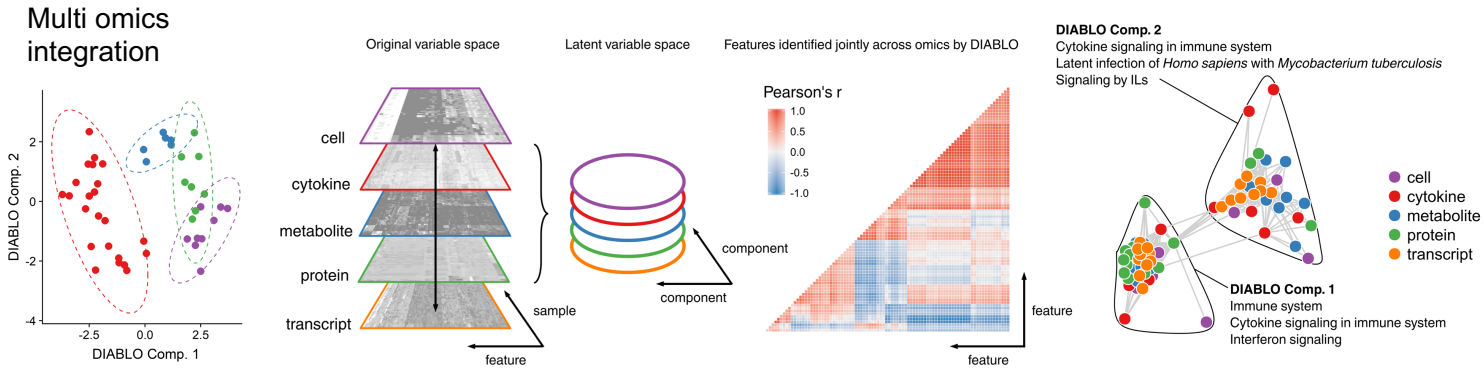
Dramatic developmental changes emerge when comparing later days of life to D0 but the correlation structure is a hairball mess!

Multi-omics integration

Single omics

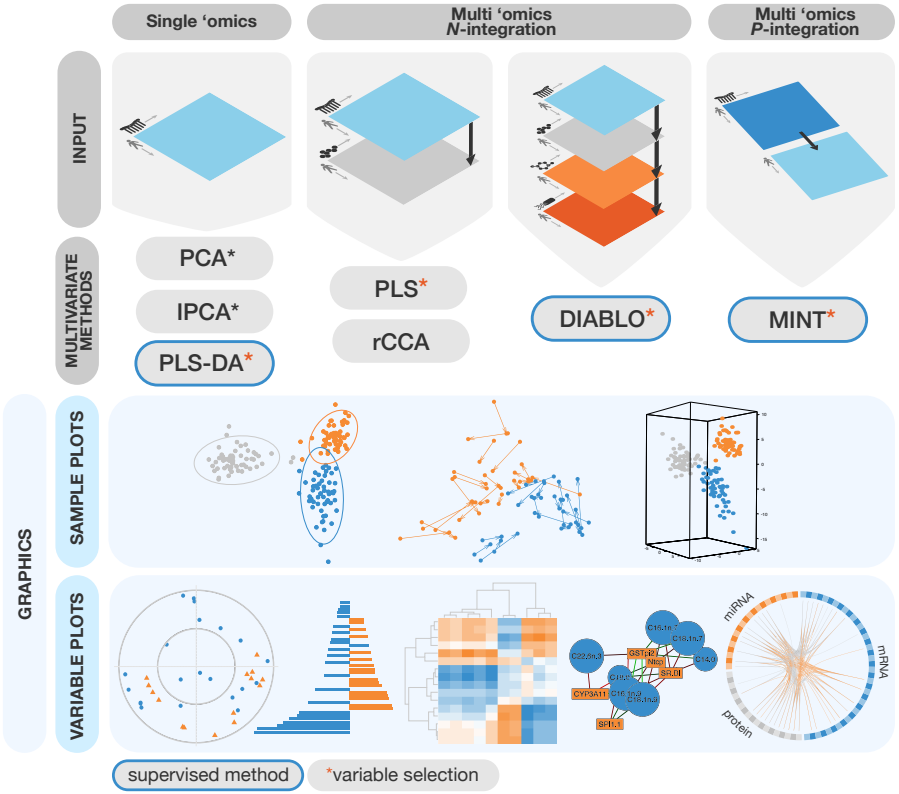


Multi omics integration



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).

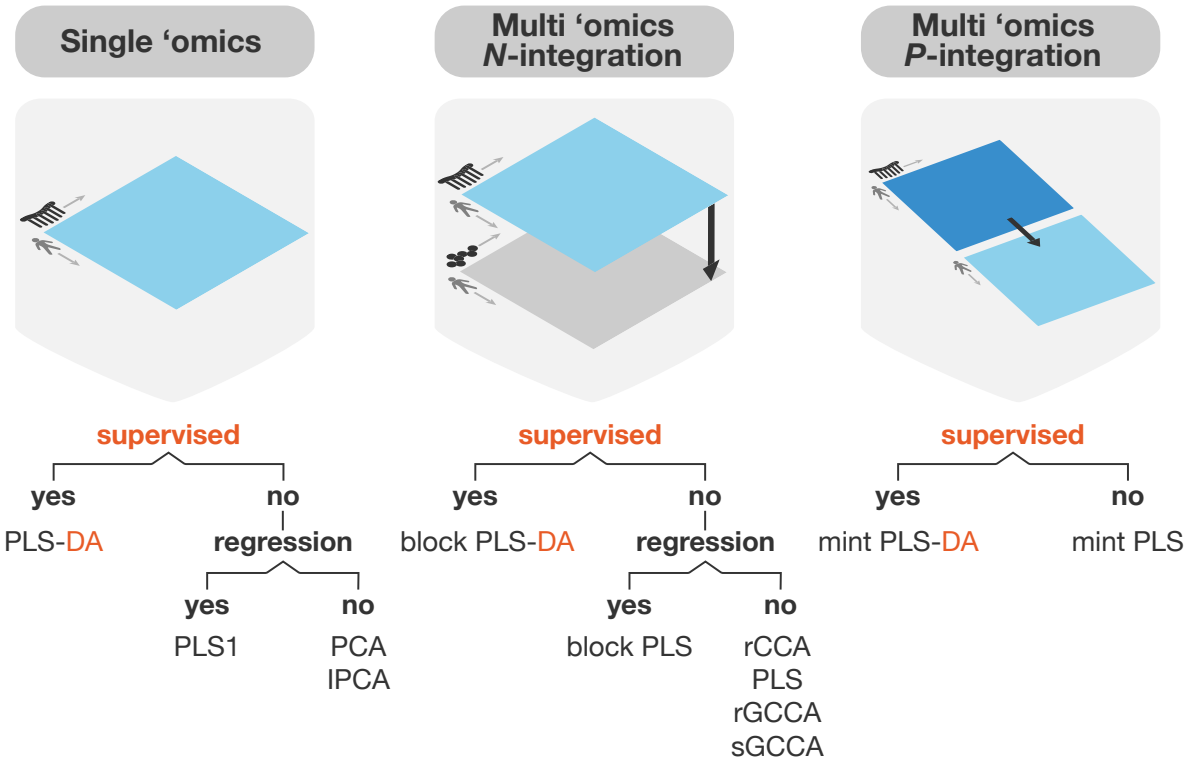
An R toolkit for multivariate data analysis and statistical integration of biological 'omics data



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

Nineteen multivariate methods (13 novel)

Each method can answer a specific biological question



↪ watch this space on www.mixOmics.org for a handbook and online courses.

Want to go further? Here are our latest extensions

- Integration of microbiome and omics
  T1D TrialNet prevention study: proteomics, 16S and metaproteomics (Gavin et al. 2018 *Diabetes Care*)
- Integration of time course omics data to identify coordinated profiles
  Method paper + application (Bodein et al. 2019 *Frontiers in Genetics*)
- Inclusion of pathway information for a hybrid data & knowledge-driven integration
  Method + application (Singh et al. 2019 DIABLO *Bioinformatics*)
- Removal of batch effects (esp. microbiome)
  Manuscript (Wang et al. *bioRxiv* 2020 358283)

Take-home messages

- We are entering a new era that first requires data-driven approaches to make sense of big biological data
- From a holistic to a hypothesis-generating approach
- Does not exclude *a priori* biological knowledge (e.g. pathway analysis) in computational methods
- New field emerges that blends computationally savvy biological scientists and vice-versa
- Development of computational tools is important to advance knowledge

Acknowledgements

mixOmics team

Sébastien Déjean, François Bartolo
K-A Lê Cao, Florian Rohart, Benoît Gautier
Amrit Singh and many mixOmics international users

Lê Cao lab alumni and members

Staff: Florian Rohart, Nicholas Matigian, Benoît Gautier, Malathi Imiyage Dona, Zitong Li, Aleks Dakic, Al Abadi , Saritha Kodikara

PhD: Amrit Singh, Jasmin Straube, Chao Liu, Ralph Patrick, Syeda Zahir, Aimee Hanson, Yiwen Wang, Isaac Virshup, Yidi Deng

Many of our great collaborators

Smallbig: Casey Shannon, Tobias Kollman, Scott Tebutt (UBC)
Rheumatoid Arthritis: Ranjeny Thomas (UQ)
Ankylosing Spondylitis: Matt Brown (KCL)
Microbial bioprocesses: Olivier Chapleur (INRAE)
Stem cells: Christine Wells (UoM)
single cell multi-omics: Matt Ritchie, Luyi Tian (WEHI), Heather Lee (UoN)
timeOmics: Antoine Bodein, Arnaud Droit (ULaval)

Australian Government Australian Research Council
Australian Government National Health and Medical Research Council
CHAN ZUCKERBERG INITIATIVE
STEMCELLS AUSTRALIA

We are open to new collaborations and scientific exchanges!

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