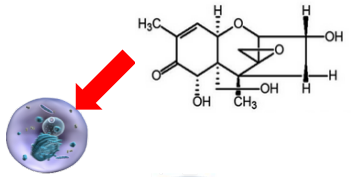
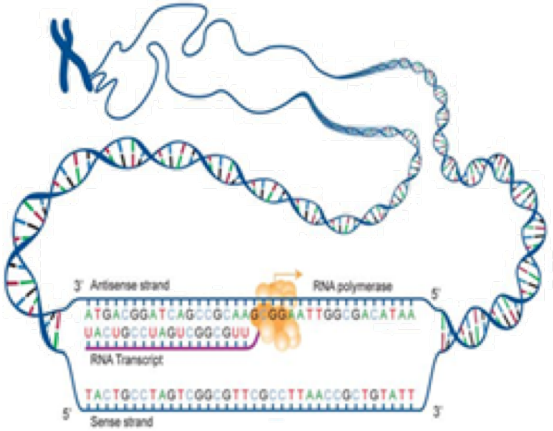
Fusion of transcriptomic and metabolomic data to achieve deeper insights into mycotoxin effects

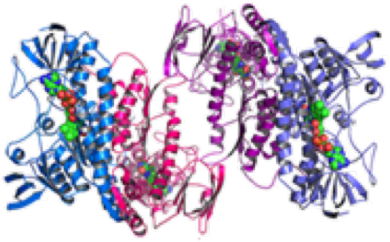
Marie Tremblay-Franco, Yannick Lippi, Cécile Canlet, Roselyne Gautier, Claire Naylies, Manon Neves, Philippe Pinton, Imourana Alassane-Kpembi, Laurent Debrauwer, Isabelle P. Oswald

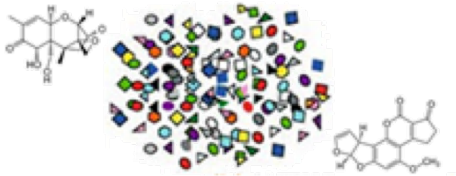
12èmes journées scientifiques du Réseau Francophone de Métabolomique et de Fluxomique, Clermont-Ferrand, 23 mai 2019

THE “OMICS” CASCADE









ADN = **GENOMICS**

Transcription

ARNm = **TRANSCRIPTOMICS**

Translation

Proteins = **PROTEOMICS**

Enzymatic reaction

Metabolites = **METABOLOMICS**

PHENOTYPE

- “omics” data = collection of data at the scale of the whole organism

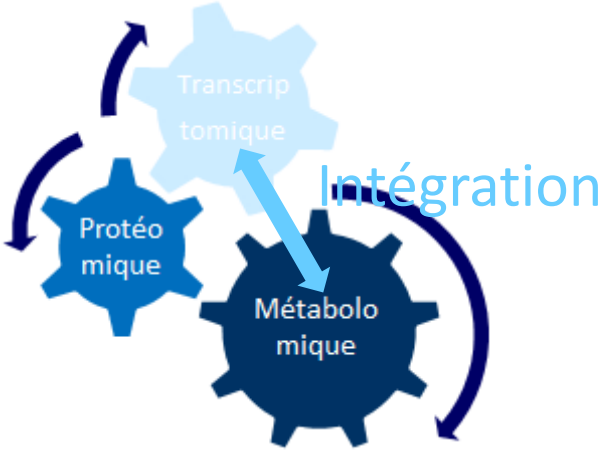
⇒ Assessment of phenotypic changes following exposure to one biological factor

DATA FUSION

- **« Combination of multiple omic datasets in order to develop multivariate models that are predictive of complex phenotypes » (Ritchie et al., 2015)**

⇒ Extraction of complementary information, on the whole biological system

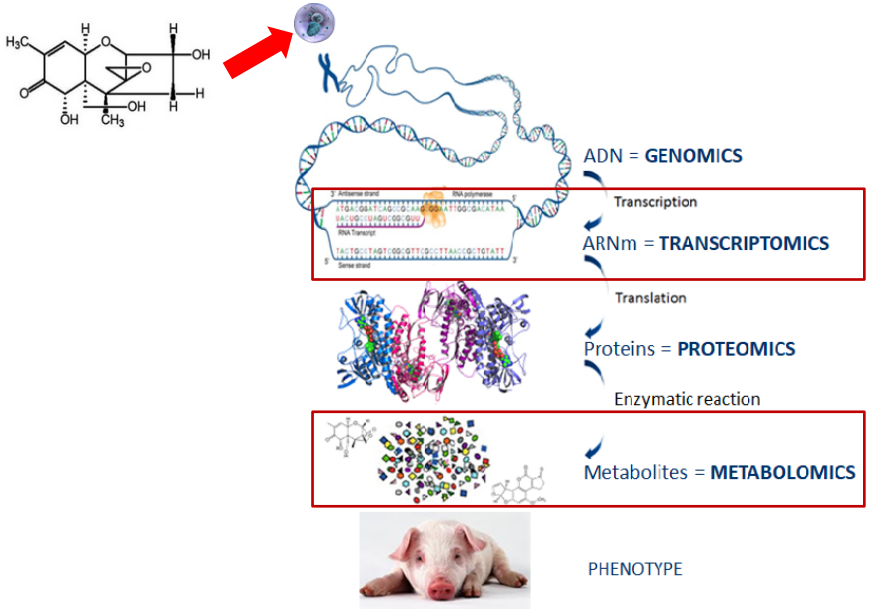
- Biological assumption (accepted): link between functional levels (Günther et al., 2014)
- *Aim: fusion of transcriptomic and metabolomic data*



⇒ Assessment of correlations between the two functional levels and identification of genes and metabolites markers of exposure to the studied factor

TRANSCRIPTOMICS / METABOLOMICS : WHY?

- **Transcriptomics** = first level of integration
  - Early response
  - Understanding of cellular activity modulations (Mele et al., 2003)
- **Metabolomics** = final level of the “omics” cascade
  - Integrated status of genetic and environmental factors = “Metabolomics is a crucial element in bridging the difference between the genotype and phenotype of an organism” (Fiehn, 2002)
  - Metabolites = final phenotypic expression of an organism



BIBLIOGRAPHY

- **Unsupervised method**: assessment of correlations between features
  - Canonical Correlation Analysis (CCA; Lê Cao et al. 2009, Wilms et Croux 2016; …)
  - Self Organized Maps (SOM; Hirai et al. 2005, Stegmayer et al. 2012)

- **Supervised method**: relationship between biological factor and "omics" features = Partial Least Squares (PLS)-based methods
  - O2-PLSDA (Bylesjö et al. 2007; Bouhaddani et al. 2016)
  - Concensus Orthogonal-PLS (Boccard et al. 2013)

AIMS

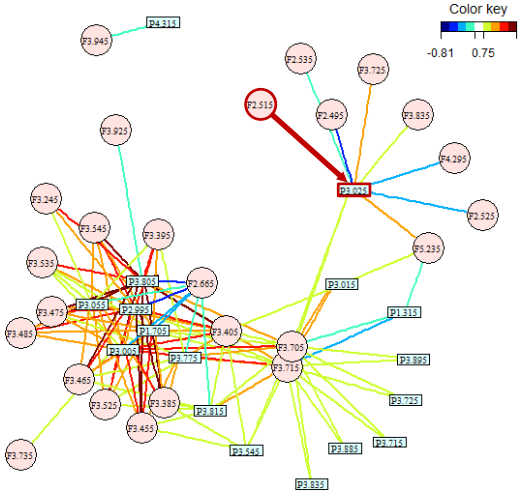
- Comparison of Canonical Correlation Analysis and Self Organized Maps to identify correlated transcriptomic and metabolomic features

- Adjust a regression model to assess relationship between factor of exposure and correlated features (as identified in the above step)

CANONICAL CORRELATION ANALYSIS (CCA)

- CCA (Hotelling, 1936) = multivariate method to assess statistical correlations between 2 datasets

⇒ **Are changes in metabolite concentrations following factor of exposure linked to changes in genes expression?**



- Maximization of the **correlation** between a latent variable from the transcriptomic block U and a latent variable from the metabolomic block *V*

$$U = \sum_{j=1}^{p} a_j X_j \; ; \; V = \sum_{k=1}^{q} b_k \, Y_k$$

⇒ Computation of weight vectors a and b such that $cor(U, V)$ $maximale$

PENALIZED CANONICAL CORRELATION ANALYSIS

- BUT…
  - High dimensionality of datasets = latent variables lack of biological interpretability
  - p >> n = computational problems

⇒ Penalization needed to select the most important features (discernable biological meaning / information)

- « **sparse** » CCA (Wilms et al., 2016)
  - Some weights equal 0 : $a_1X_1+\mathbf{0}X_2+a_3X_3+\mathbf{0}X_4+a_5X_5$

⇒ Removal of noisy features = biological interpretability of latent variables is improved

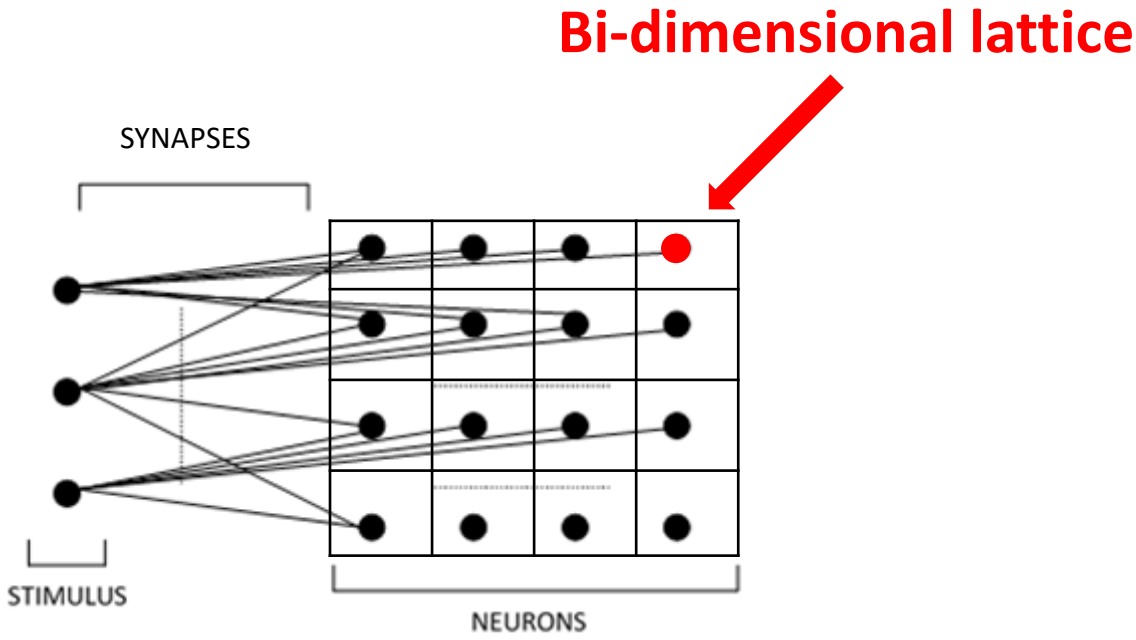
  - Computation of penalization = cross validation

- mixOmics R package

mixOmics: Exploration and Integration of Omics datasets

SELF-ORGANIZING MAPS (SOM)

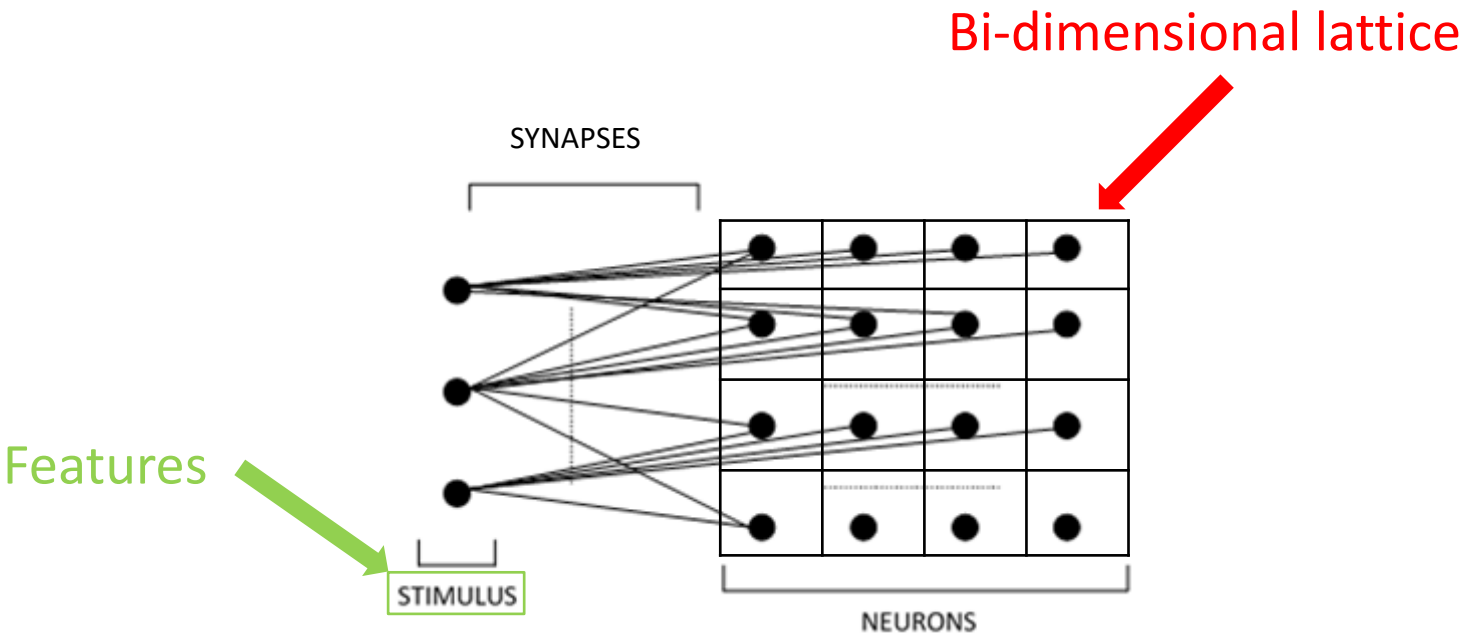
- SOM (Kohonen, 1982): unsupervised method for projection and classification of objects, based on neural networks
  - **Bi-dimensional lattice** (units = neurons)



From Bessai et al. (2002)

SELF-ORGANIZING MAPS (SOM)

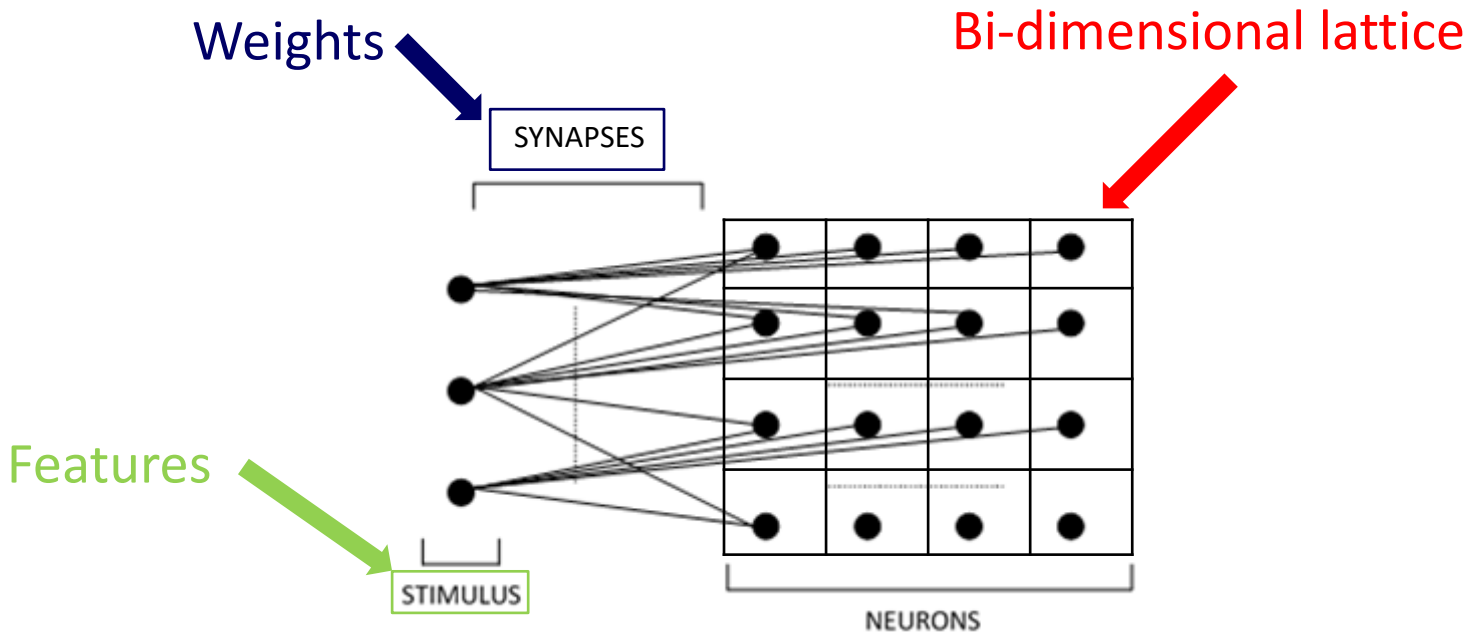
- SOM (Kohonen, 1982): unsupervised method for projection and classification of objects, based on neural networks
  - **Bi-dimensional lattice** (units = neurons) onto which **features** are projected / clustered
    - To each feature is associated a vector containing measured values for individuals (**stimulus**)



From Bessai et al. (2002)

SELF-ORGANIZING MAPS (SOM)

- SOM (Kohonen, 1982): unsupervised method for projection and classification of objects, based on neural networks
  - **Bi-dimensional lattice** (units = neurons) onto which **features** are projected / clustered
    - To each feature is associated a vector containing measured values individuals (**stimulus**)
    - To each unit is associated a vector of **weights** (prototype = synapses)



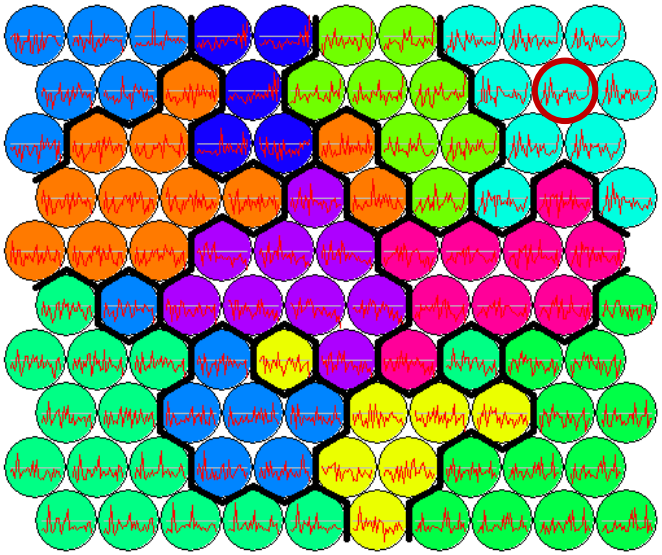
From Bessai et al. (2002)

SELF-ORGANIZING MAPS (SOM): ALGORITHM

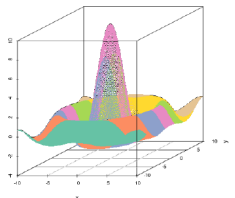
- Iterative algorithm

- Preservation of the original topology of the data: close features in the input space are clustered together into the same unit or into neighbor units on the map)

$\Rightarrow$ clustering of co-expressed genes and co-accumulated metabolites in the same unit

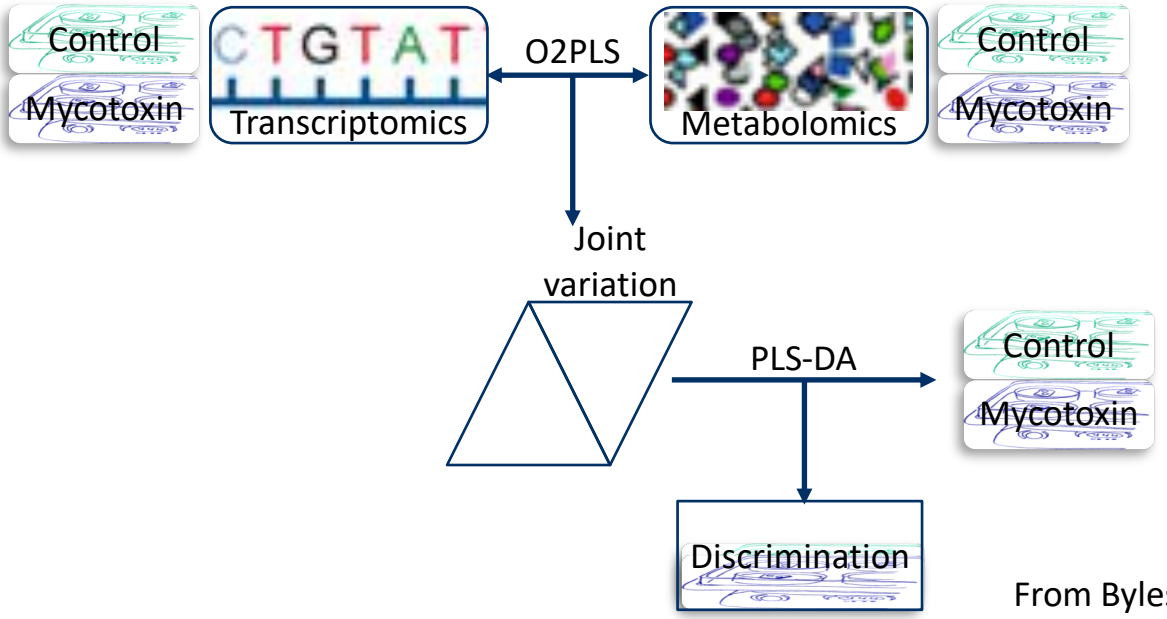


- Package R SOMbrero (Olteanu et al. 2015)



PARTIAL LEAST SQUARES – BASED METHODS

- **O2PLS**: generalization of O-PLS to two datasets
  - Separate the joint variation (e.g. used to predict metabolite levels from transcript profiles, and vice versa)
  - Orthogonal : removal of confounding variability (biological, experimental, sample collection, …)
  - PLS-DA using joint variation to model factor of exposure



From Bylesjö et al. (2007)

⇒ Discrimination of observations depending on mycotoxin exposure and list of discriminant transcripts and metabolites

MONTE CARLO SIMULATION

- Random generation of artificial data using a defined model = known structure of data

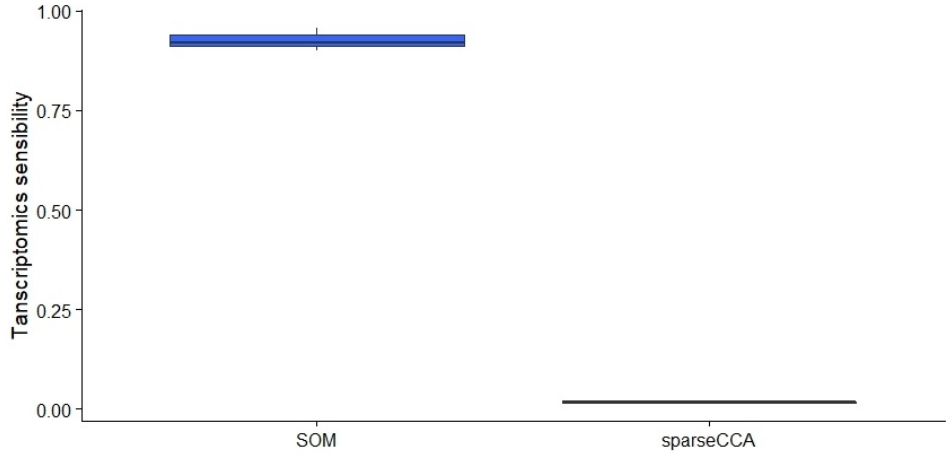
⇒ **Assessment of ability of methods to recover this structure**

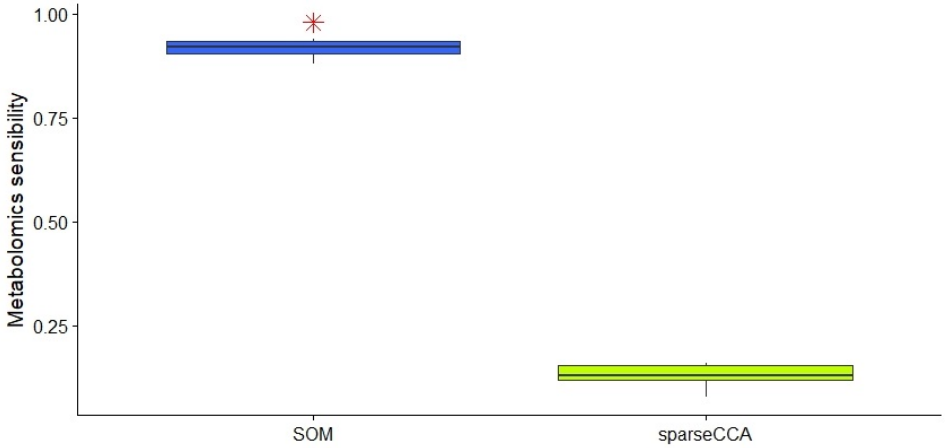
- Dataset sizes
  - n=10 observations / group

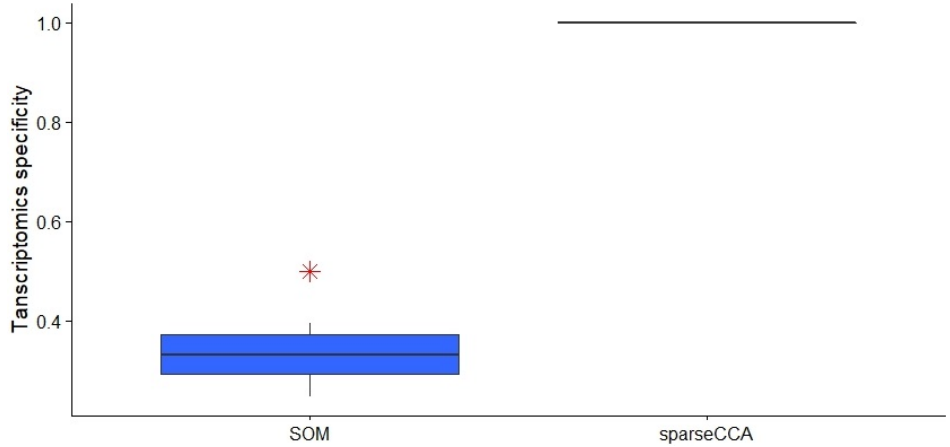
P transcripts	q NMR features
1000	100
5000	500
10000	700
12000	788

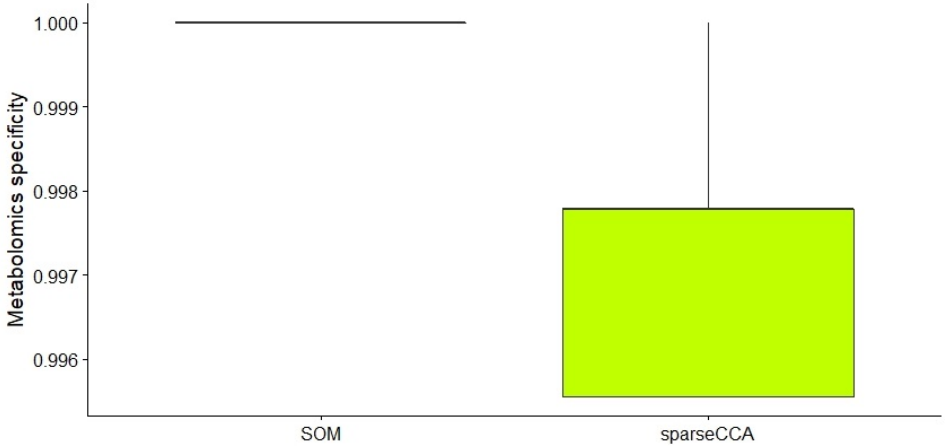
- Criteria
  - Sensitivity: ability of a test to give a positive result when an hypothesis is true (true positives)
  - Specificity: ability of a test to give a negative result when an hypothesis is false (true negatives)
  - R^2: proportion of explained variance
  - MSEP: prediction error = how well does the model classify individuals into the right group?

RESULTS : sparse CCA / SOM

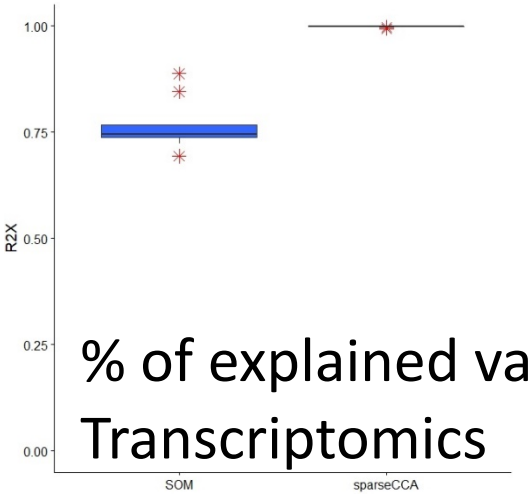




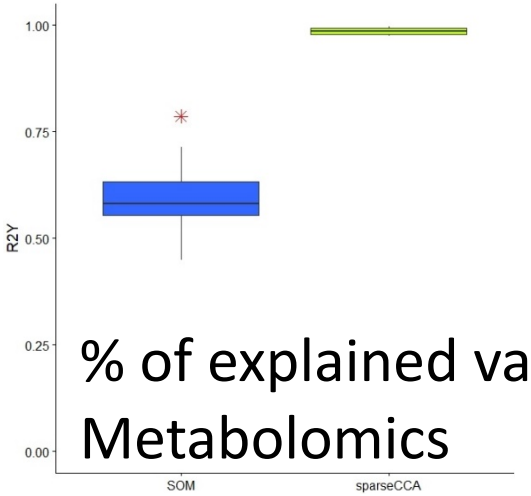




RESULTS : O2-PLSDA

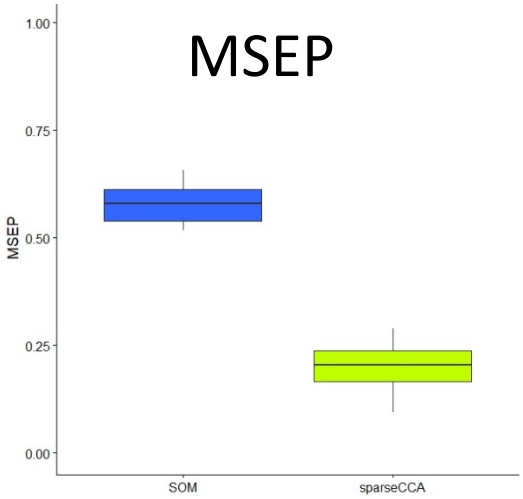


% of explained variance – Transcriptomics

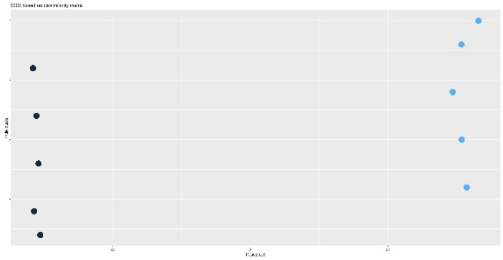


% of explained variance – Metabolomics

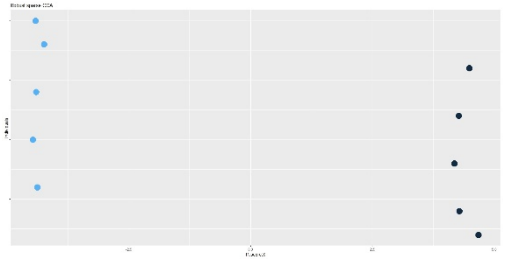
MSEP



SOM



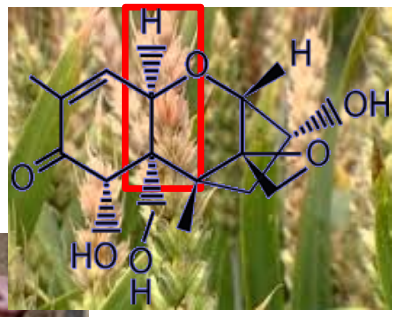
Sparse CCA



Biological application

CONTEXT

- Pig: rich-cereal food



- Fusarium: contaminant fungus of cereal

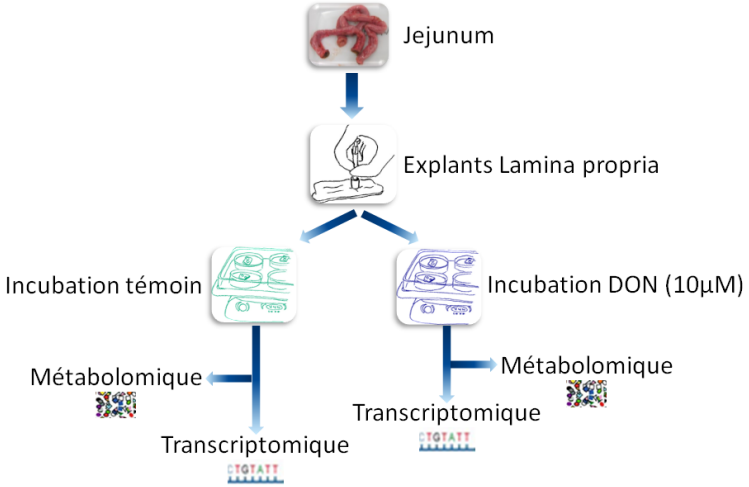
- DON: secondary metabolite of Fusarium
  - Acute and chronic disruptions on animals (gastro-intestinal tract)

⇒ Pigs are particularly exposed to DON

⇒ Identification of markers of exposure to mycotoxins is important for animal healthcare

EXPERIMENTAL DESIGN / DATA

- n=8 animals
  - Jejunal explants (ex vivo)
  - Exposition Control / Mycotoxin (10µM)



- **Transcriptomics**
  - Agilent porcine-specific microarray (60305 spots)
  - Raw data processing (signal median intensity): filtering, log2 transformation and normalization (quantiles method, Bolstad et al. 2003)

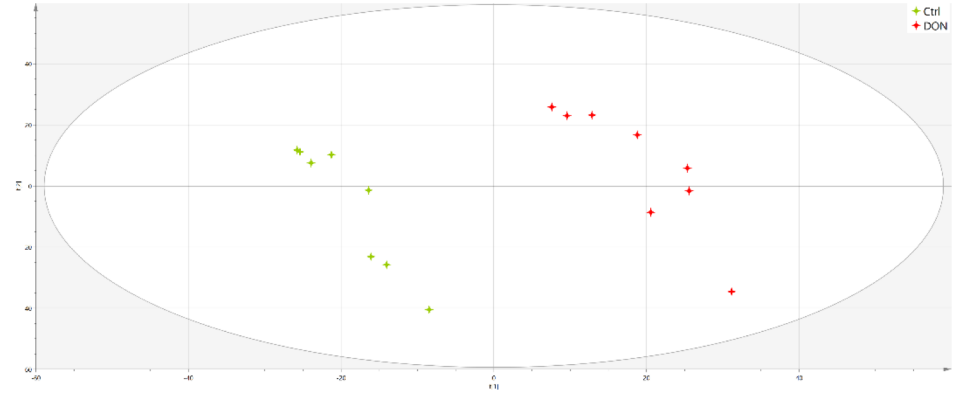
  ⇒ *p=41336 features*

- **Metabolomics**
  - ^{1}H-NMR
  - Processing: bucketing/integration et normalization (total intensity)

  ⇒ *q=751 NMR features*

INDIVIDUAL ANALYSIS

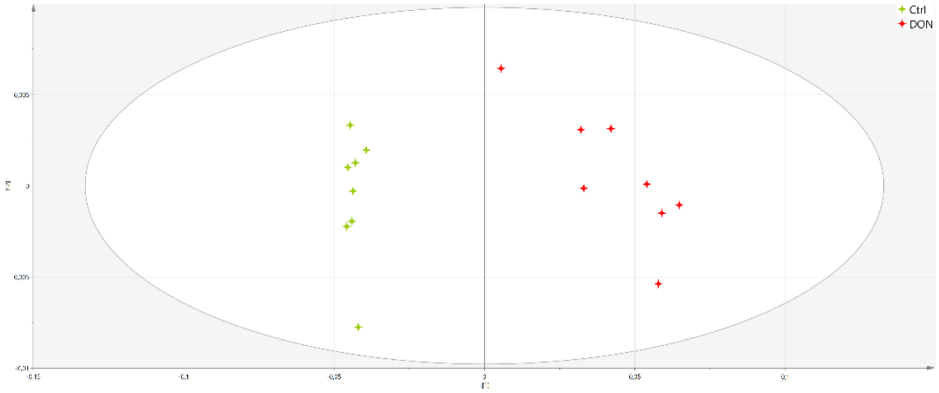
- Transcriptomics



1480 discriminant transcripts

⇒ Modulation of immunity/inflammation related genes

- Metabolomics



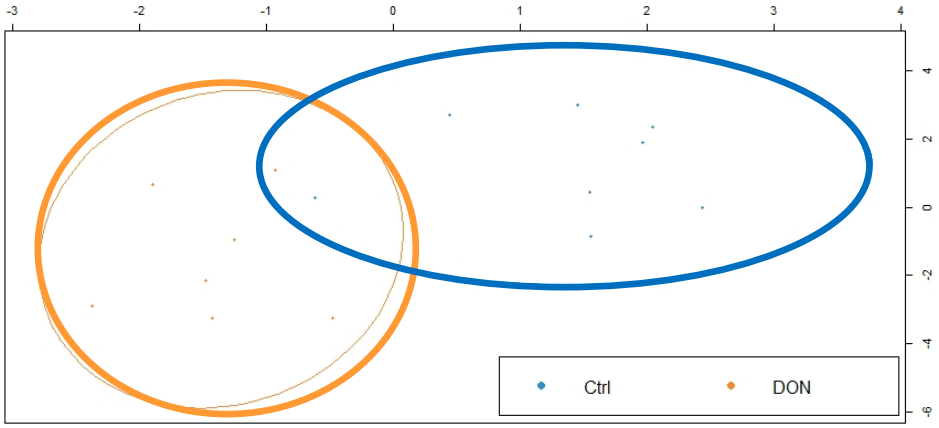
3 discriminant metabolites

⇒ Alanine et Lactic acid

Data fusion: identification of pathways linked to process changes involving metabolism of both metabolites?

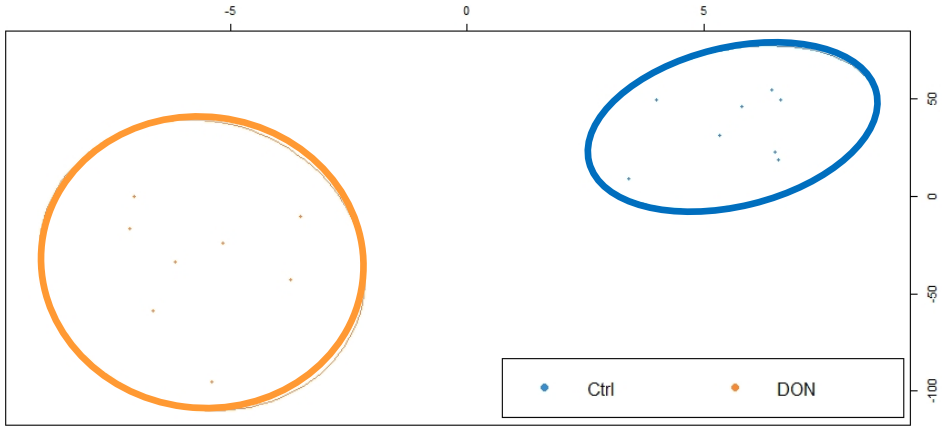
DATA FUSION (1)

- Transcripts selection: 15000 with highest standard deviation
- Sparse CCA-O2PLSDA



  - R2=60.4%
  - MSEP=0.011
  - 12 transcripts & 12 metabolites were discriminant

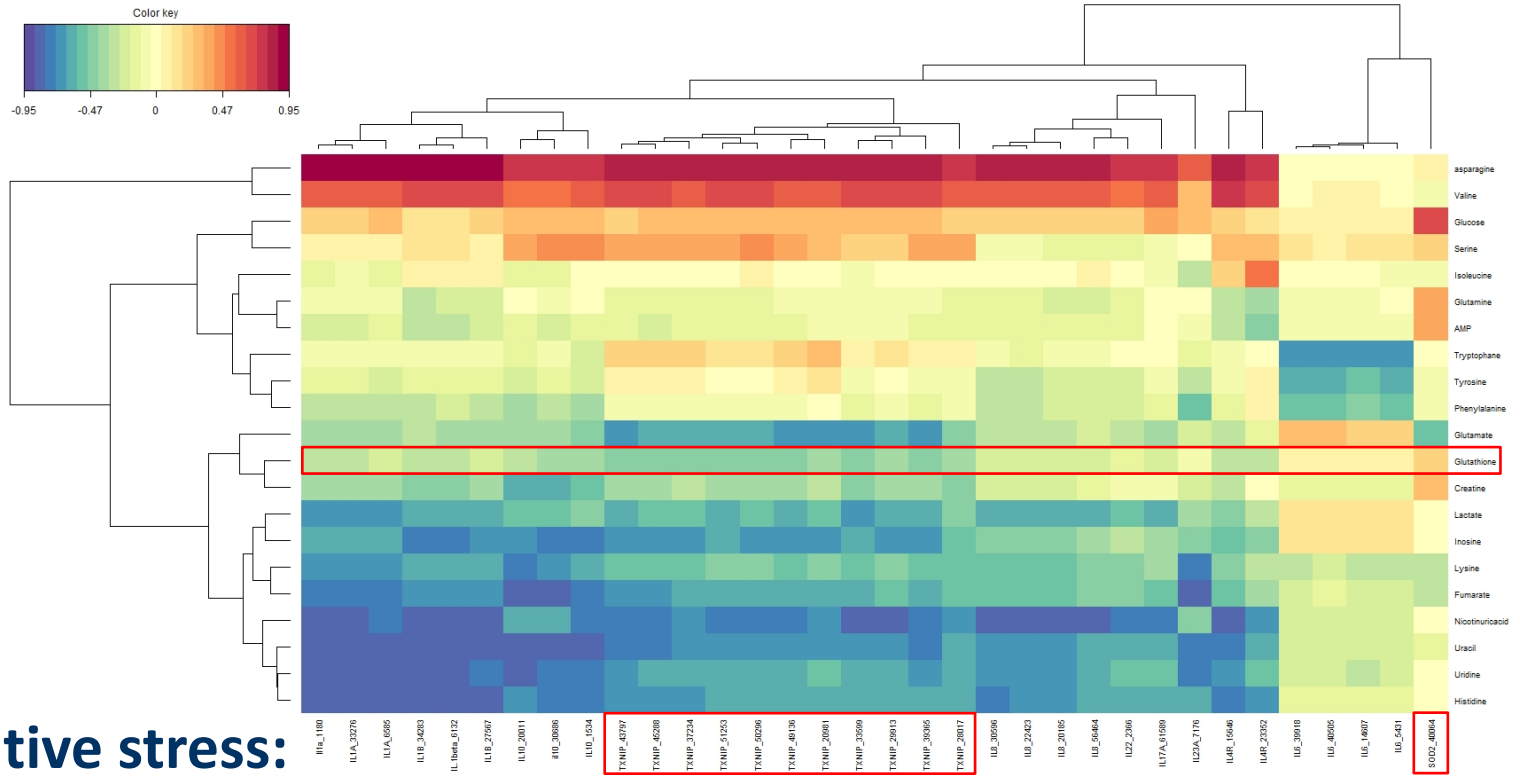
- SOM-O2PLSDA



  - R2=49.9%
  - MSEP=0.011
  - 1443 transcripts & 61 metabolites (24 identified) were discriminant

⇒ Exposed explants are better separated from Control explants with the model fitted using the SOM-selected features

DATA FUSION (2): CORRELATIONS BETWEEN DISCRIMINANT FEATURES



Oxidative stress:

- Glutathione, endogen antioxidant, correlated with:
  - TXNIP (negatively): gene encoding for a thioredoxin-binding protein. Thioredoxine (protects cells from oxidative stress): inhibition of the antioxidative function of thioredoxine $\Rightarrow$ accumulation de reactive oxygen species and cellular stress
  - SOD (superoxide dismutase, positively): antioxidant enzyme

CONCLUSION: SIMULATIONS

- **SOM**
  - High sensitivity = selection of really correlated features
  - Low specificity = selection of uncorrelated features

- **Sparse CCA**
  - High specificity but low sensitivity
  - Highly consuming-time
  - Prior selection of features

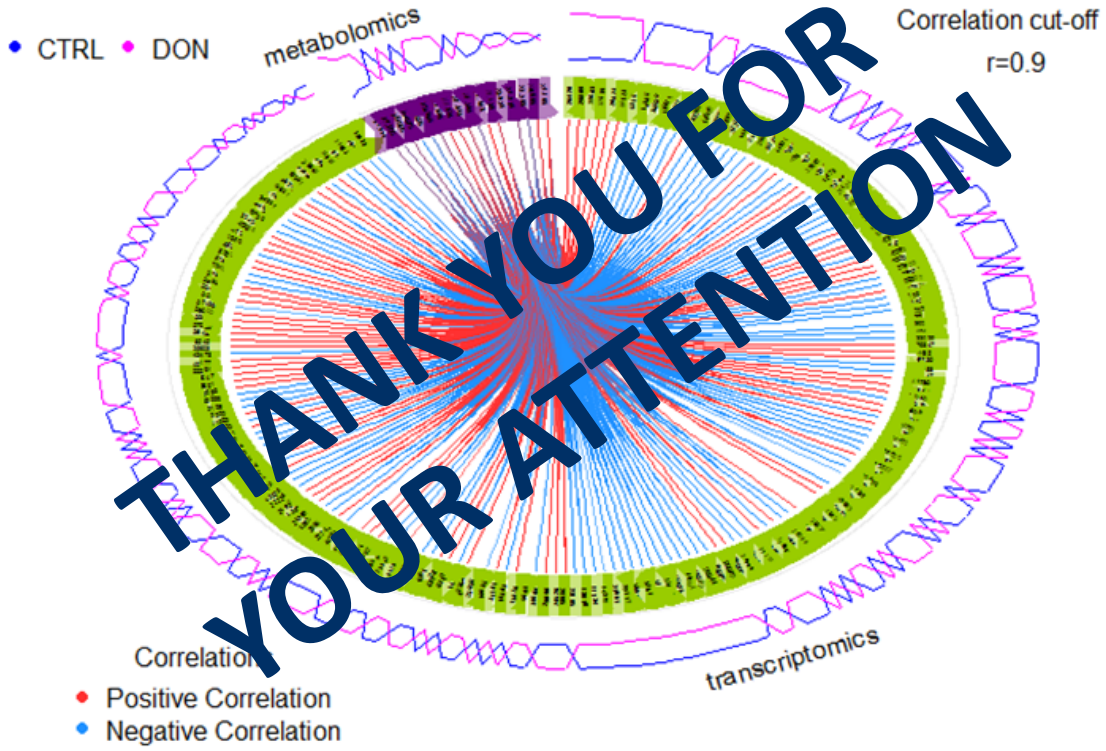
$\Rightarrow$ No universal method: combination of several methods = good alternative

CONCLUSION: BIOLOGICAL APPLICATION

- Data fusion
  - Increased number of discriminant metabolites
  - Biological link between transcripts & metabolites

- SOM
  - Best discrimination of Control observations from Mycotoxin Exposed observations
  - Biological relevance of selected features: mycotoxin exposure induces oxidative stress = reported in literature (Pierron et al. 2016) but only for the transcriptomic side

L. Debrauwer
C. Canlet
R. Gautier



THANK YOU FOR YOUR ATTENTION

Y. Lippi
C. Naylies

I. Oswald
P. Pinton
M. Neves
I. Alassane-Kpembi