Visualising associations between paired ‘omics’ data sets

By: Ignacio González, Kim-Anh Lê Cao, Melissa J Davis and Sébastien Déjean

October 28th Journal Club Presentation

David Yang

Problem

- To obtain useful information from ‘omics’ data, it must be properly processed and analyzed
- Hence, a major challenge with the integration of omics data involves the extraction of discernible biological meaning from the results
- Statistical methods such as Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) regression have allowed investigation of the correlation structure between datasets
- However, little attention has been given to the interpretation of the results and graphical outputs

Purpose

1. The paper revisits several graphical outputs dedicated to exploratory approaches to highlight associations
2. The relevance of these proposed graphical outputs were assessed on a simulation study
3. On two real data sets, interpretations from the graphical outputs are compared to known biological networks

Table of Content

Projection-based statistical methods

- Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) regression
- Two-block data matrices to be integrated: *X (n x p)* and *Y (n x q)*
  - *p* and *q* are the total number of variables measured on the same *n* subjects
- For example, X can contain gene expression and Y can contain metabolite concentrations

CCA

- Searches for the largest correlation between a linear combination of X and a linear combination of Y
- The first pair of canonical variates maximizes: $\rho 1 = cor(Xa^1, Yb^1)$ where $var(Xa^1) = var(Yb^1) = 1$.
- Onwards, subsequent canonical variate pairs (Xa^l, Yb^l) are maximized for up to min{p, q} times subject to the constraint that they are to be uncorrelated with the first pair of canonical variables

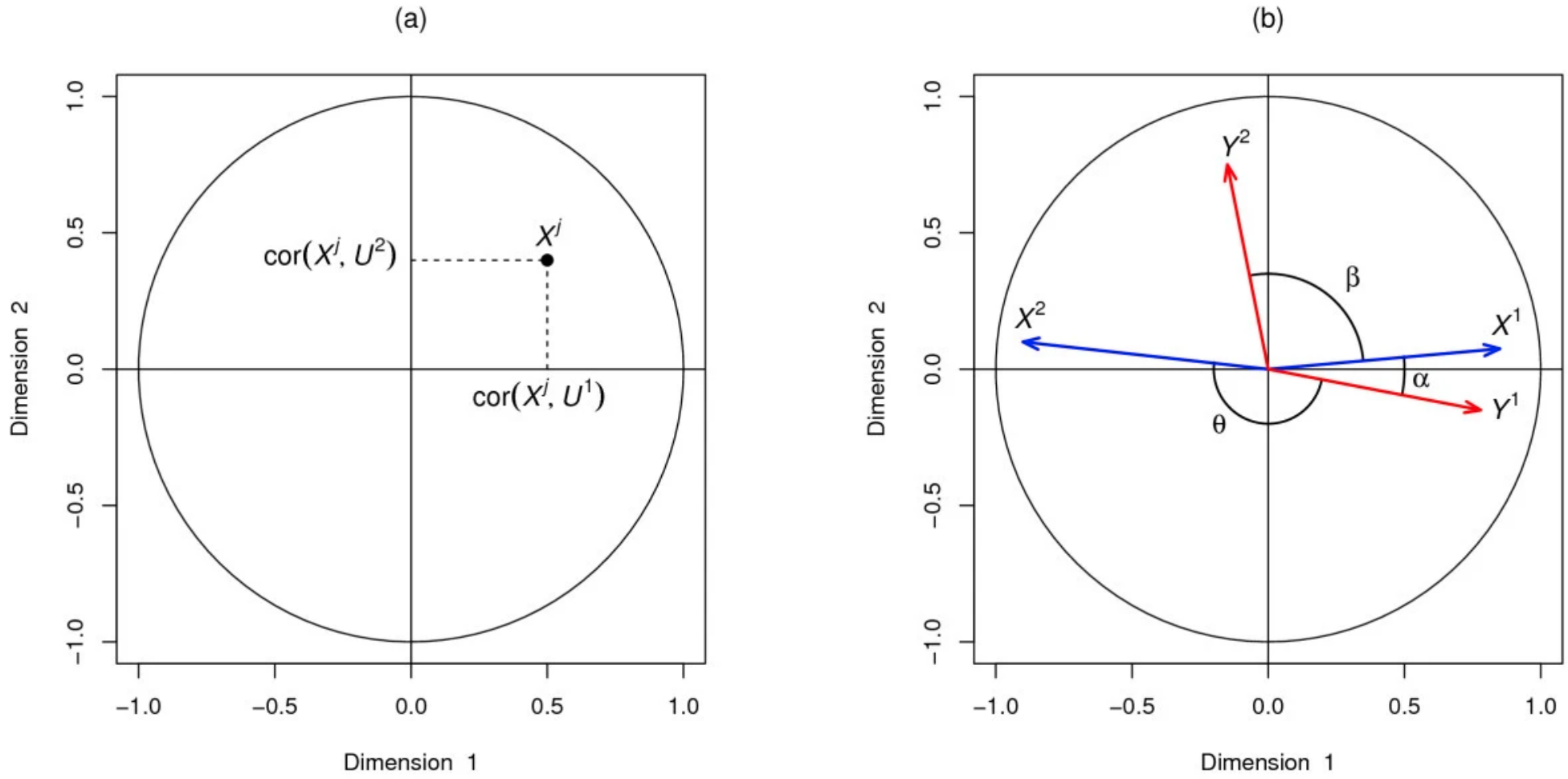
PLS

- Searches for the largest covariance between linear combinations of x and the Y variables.
- The first latent variable pair maximises cov($X\,a^{1}, Y\,b^{1}$) subject to $||a^{1}||=||b^{1}||=1$
- Onwards, subsequent pairs ($X\,a^{l}, Y\,b^{l}$) are maximized up to q times subject to the constraint that each pair is uncorrelated with the previous pair
- PLS-regression models unidirectional relationship (Y = AX)
- PLS-canonical models bidirectional relationships (AY = BX)

rCCA and PLS

- In the context of high throughput biological data, the number of variables often exceeds the number of samples
- This poses computational problems in CCA and PLS
- Therefore, regularized CCA (rCCA) and sparse PLS (sPLS) are used
- In short, rCCA adds a regularization term to the diagonal of covariance matrices while sPLS includes Lasso penalization on the loading vectors to shrink some coefficients to zero

Correlation Circle plots

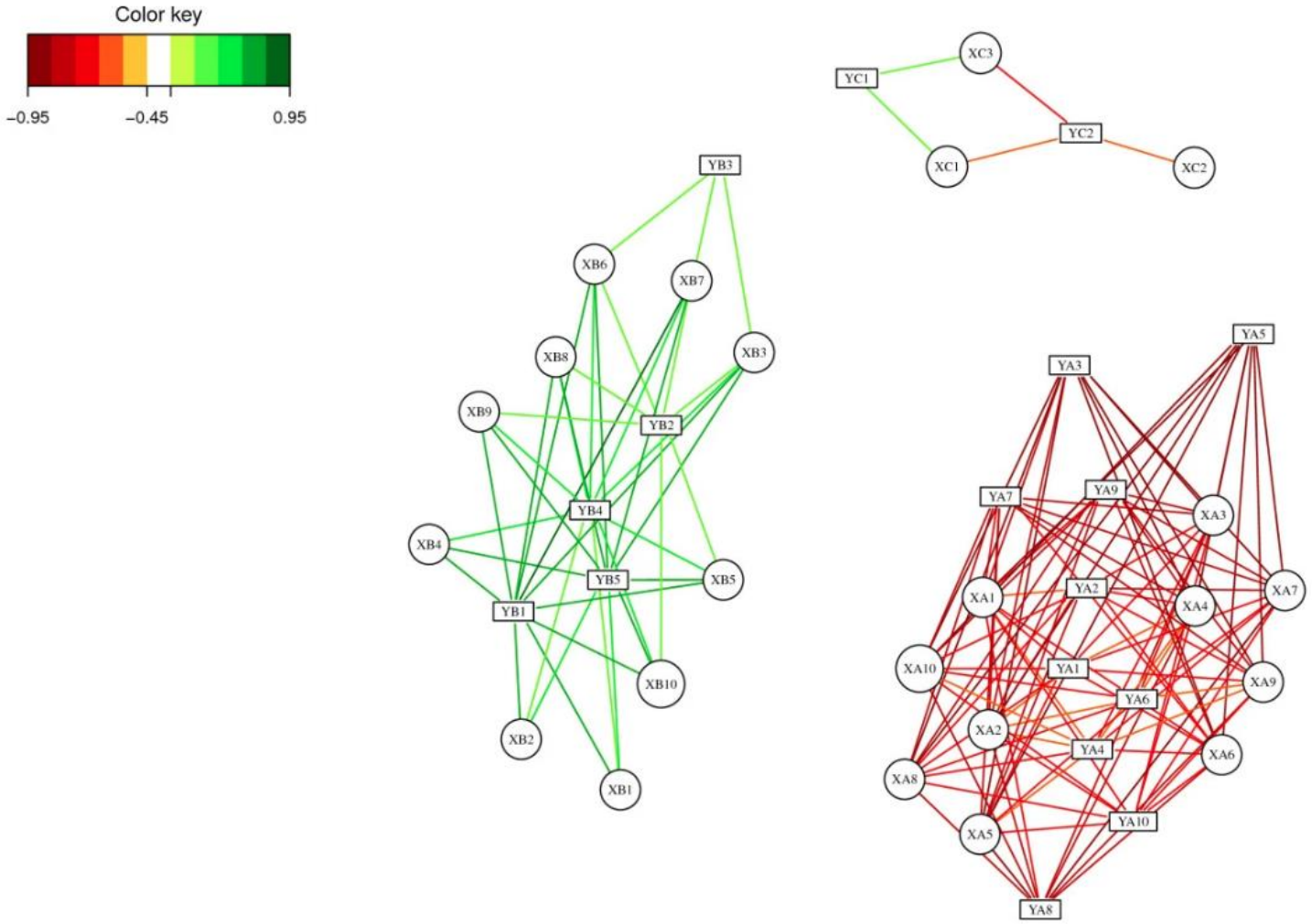


Correlation Circle plot. a) Coordinates of the X-variables on the plane defined by the first two variates U^1 and U^2. **b)** The correlation between two variables is positive if the angle is sharp $\cos(\alpha)>0$, negative if the angle is obtuse $\cos(\theta)<0$, and null if the vectors are perpendicular $\cos(\beta)\approx 0$.

Correlation Circle plots

- Enables a graphical examination of the relationships between variables and variates
- As variables are standardized and centered, the correlation obtained between each variable and component is simply the projection of the variable on the axis defined by the component
- The relationship (correlation) between the two types of variables can be approximated by the inner product between the associated vectors $A \cdot B = |A||B| \cos\theta$
- In conclusion:
  - If angle between two variables are acute, there is a positive correlation
  - If angle between two variables are obtuse, there is a negative correlation
  - If angle between two variables are right, there is no correlation

Relevance network

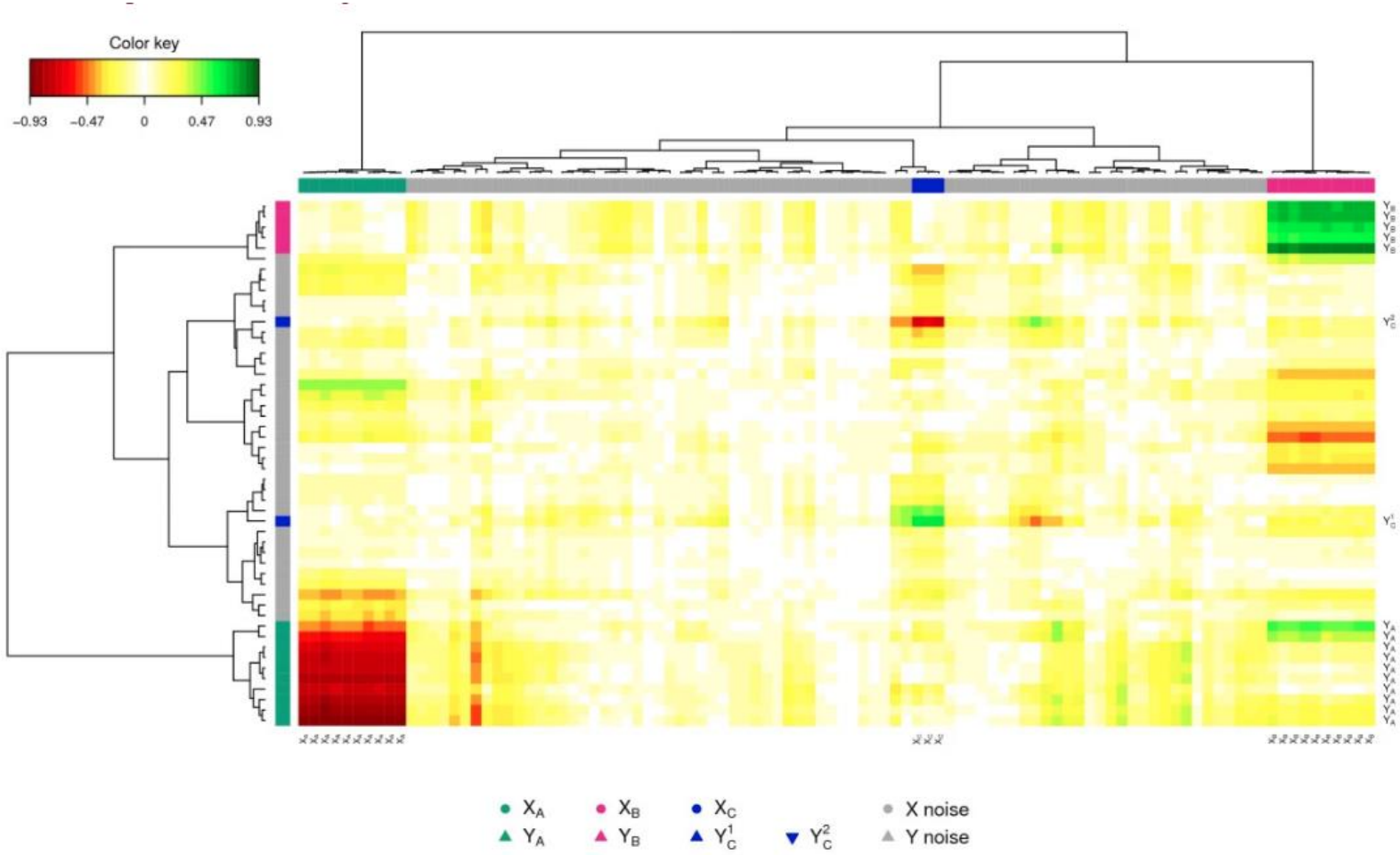


Relevance Networks for the simulation study. Relevance Networks obtained with sPLS-can on the simulated data using the `network` function in the `mixOmics` package. Green and red edges indicates positive and negative correlation respectively. *X* and *Y* variables are represented respectively as circles and rectangles.

Relevance networks

- In the graphs, nodes represent variables and edges represent variable associations
- In our case, we focus on the representation between variables of two different kinds, where nodes of X can only be connected to nodes of Y
- These bipartite networks are inferred using a pair-wise similarity matrix obtained from the statistical methods

Clustered Image Maps



CIM for the simulation study. CIM on the simulated data with the PLS-can method. The green and red colours indicate positive and negative correlations respectively, whereas yellow indicate small correlation values. The clusters of variables are colored on the top and left side of the CIM as in Figure 2. The variables with blank names indicate variables with weak correlations (irrelevant variables).

Clustered Image Maps

- Also called ‘clustered correlation’ or ‘heat maps’
- Each entry of the matrix is coloured based on its value in the pair-wise similarity matrix
- Rows and columns are ordered according to the hierarchical clustering and arranged in dendrograms (tree diagrams) as clusters

Simulation study: Data sets

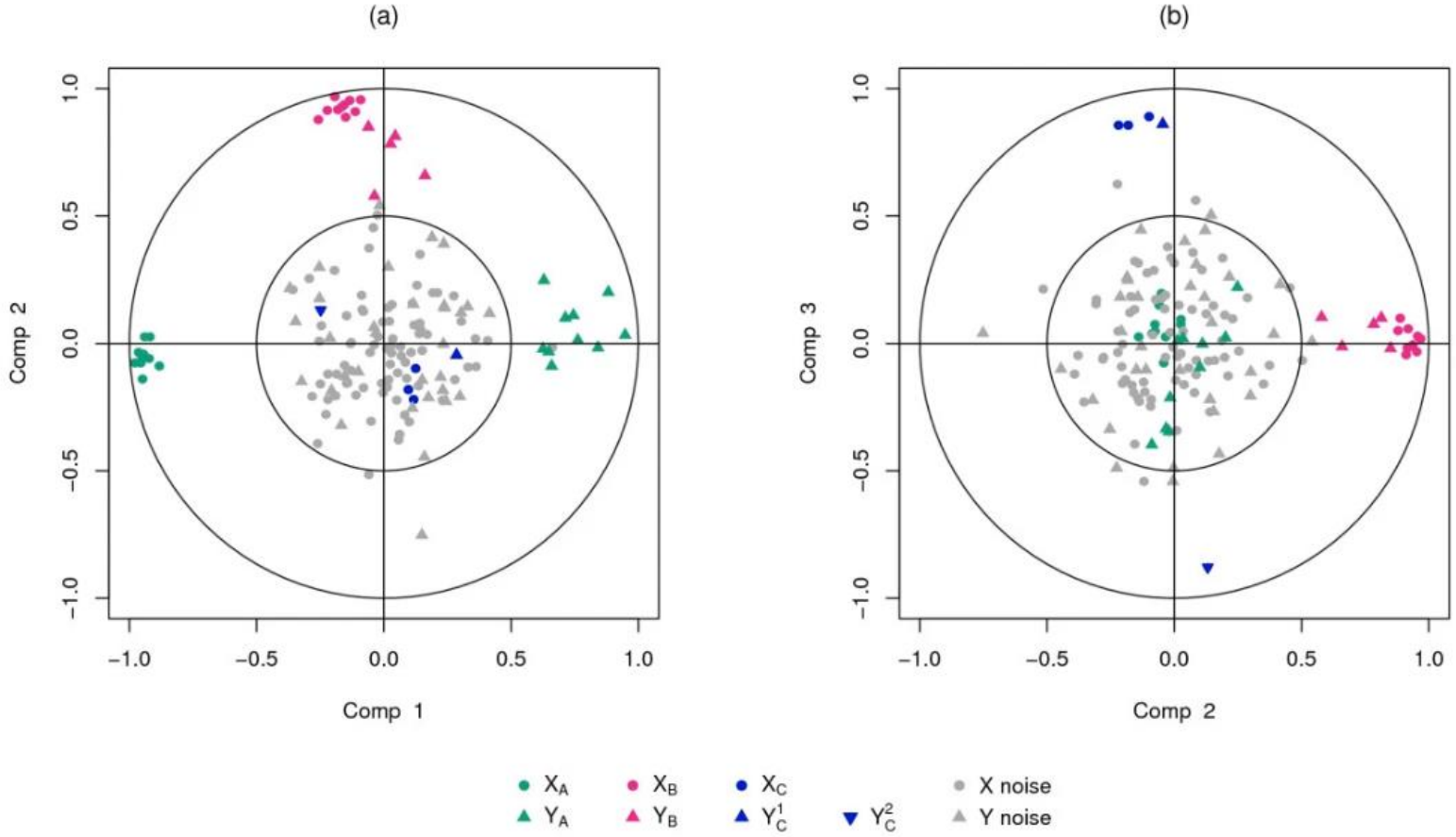
- Two data sets X (30 x 100) and Y (30 x 50) with an equal number of 30 were generated
- A subset of relevant variables in X were associated with a subset of relevant variables in Y according to the model described in the table
- The remaining variables were simulated as noise $\sim Normal(0,1)$

Variable set 1:	**Variable set 2:**	**Cross-correlation**
X_A	Y_A	-0.93 to -0.51
X_B	Y_B	0.5 to 0.85
X_C	Y_C^1	0.81 to 0.93
X_C	Y_C^2	-0.81 to -0.93

Simulation study: Analysis process

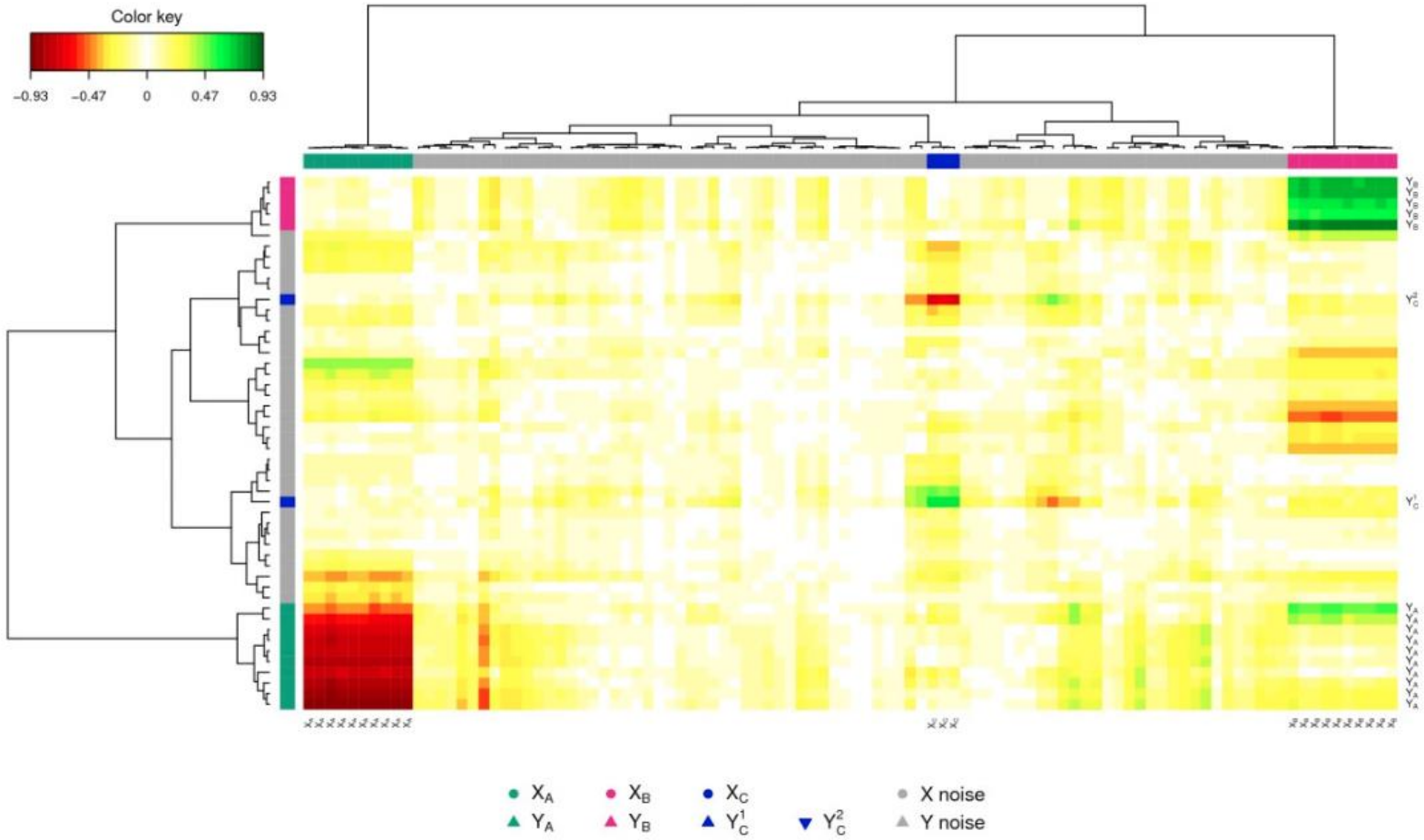
- i) PLS canonical mode was applied to these datasets
  - The first three dimensions were chosen for the graphical outputs
  - The correlation values between latent variables equal to 0.97, 0.94 and 0.95 respectively on each dimension, before decreasing for the following dimensions
- ii) rCCA approach was also applied to the datasets
  - Regularization parameters for the first three dimensions were $\lambda_1 = 0.889$ and $\lambda_2 = 0.889$
  - The canonical values obtained were of 0.959, 0.925, and 0.881 on each dimension respectively, followed by much lower values.

Simulation study: Graphical outputs



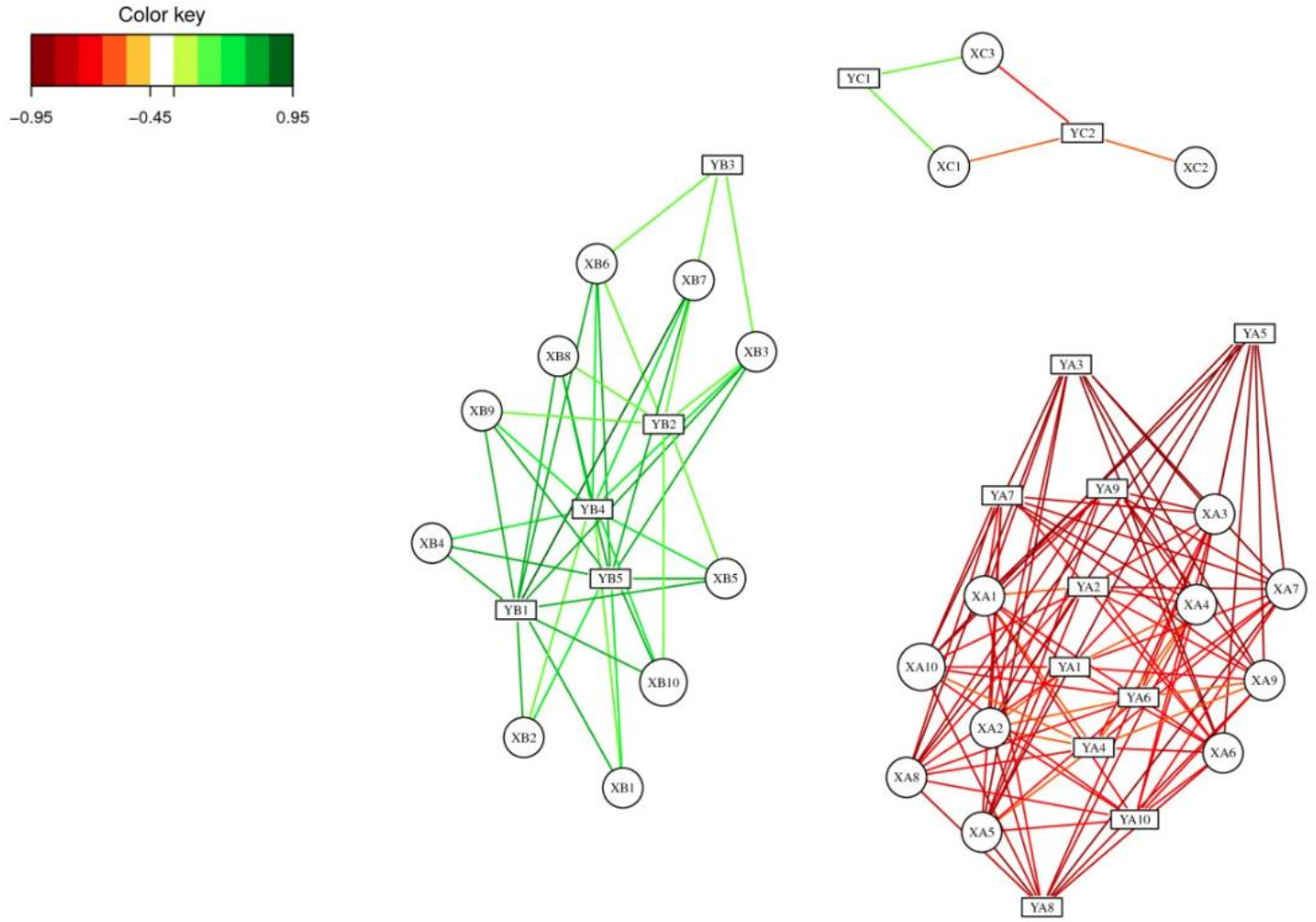
Correlation Circle plots for the simulation study. Correlation Circle plots for dimensions 1 and 2 **(a)**, and 2 and 3 **(b)**. The *X* and *Y* variables are represented by thick points and triangles respectively. The subsets of correlated variables are colored according to the legend. Expression profiles of some positively and negatively correlated variables across samples **(c)**.

Simulation study: Graphical outputs



CIM for the simulation study. CIM on the simulated data with the PLS-can method. The green and red colours indicate positive and negative correlations respectively, whereas yellow indicate small correlation values. The clusters of variables are colored on the top and left side of the CIM as in Figure 2. The variables with blank names indicate variables with weak correlations (irrelevant variables).

Simulation study: Graphical outputs



Relevance Networks for the simulation study. Relevance Networks obtained with sPLS-can on the simulated data using the `network` function in the `mixOmics` package. Green and red edges indicates positive and negative correlation respectively. *X* and *Y* variables are represented respectively as circles and rectangles.

Biological study: data sets

Nutrimouse data:

- 40 mice from two genotypes (WT and PPARα deficient) were fed with five diets
- Expression of 120 gene in liver cells and concentrations of 21 hepatic fatty acids were quantified

Liver Toxicity data:

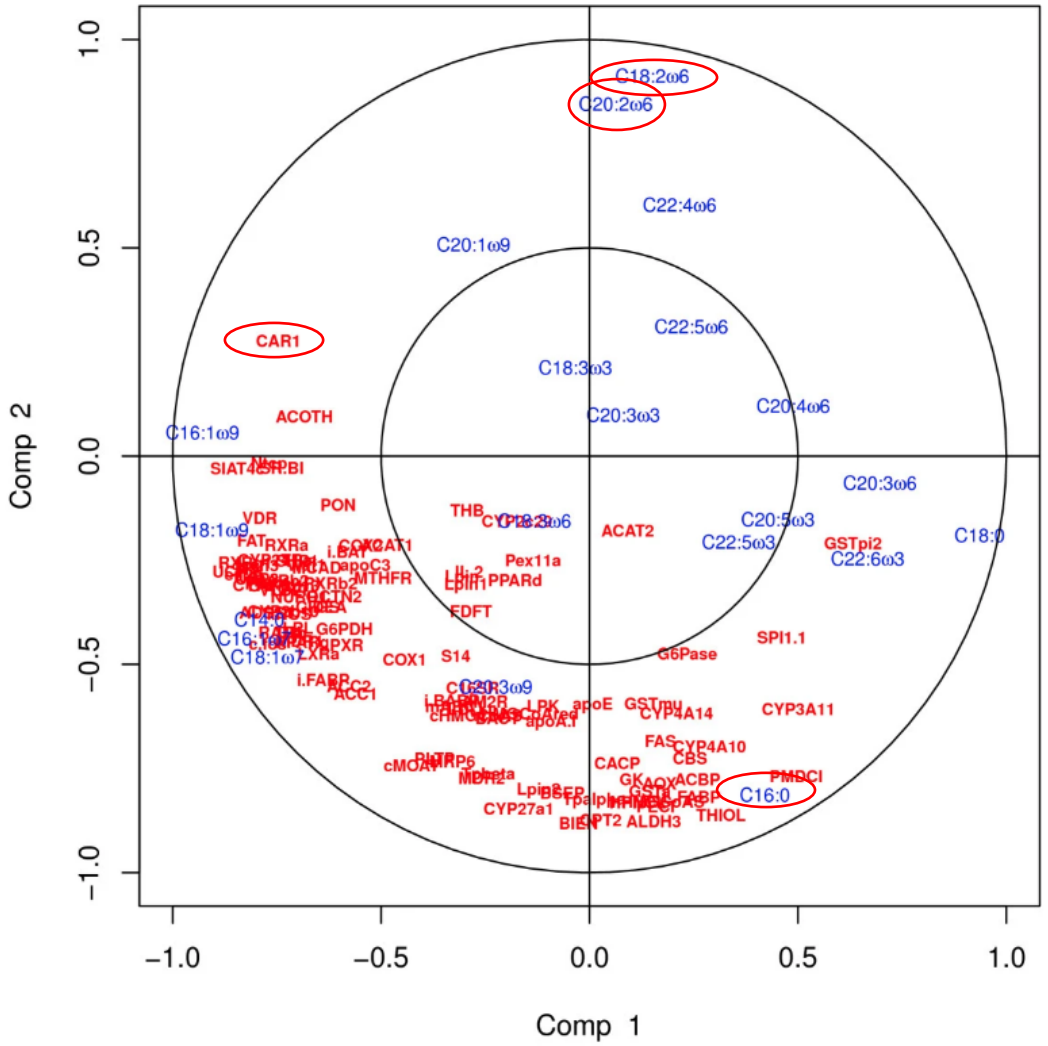
- 64 male rats of the inbred strain Fisher F344/N were exposed to 50 mg/kg, 150 mg/kg, 1500 mg/kg or 2000 mg/kg doses of acetaminophen
- Liver cell gene expression matrix (64 x 3116) and clinical measurements matrix (64 x 10)

Treatment:	**Composition of Diet:**
reference diet	corn and colza oils
saturated fatty acid diet	hydrogenated coconut oil
Omega6 fatty acid ice diet	sunflower oil
Omega3 rich diet	linseed oil
FISH diet	corn/colza/enriched fish oil

Biological study: Analysis process

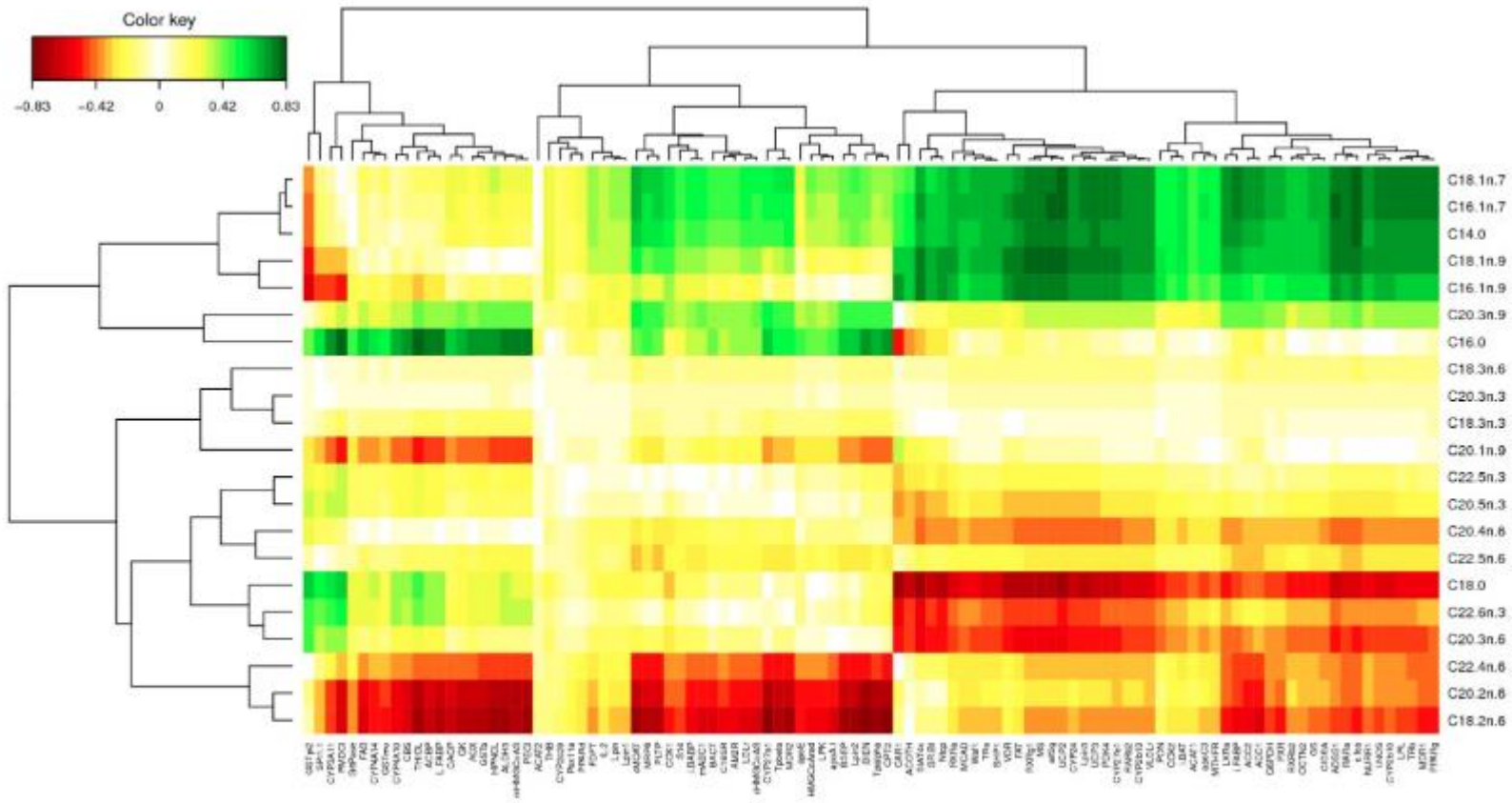
- In the Nutrimouse data, the methodology sPLS-can was applied
  - The aim was to highlight highly correlated subsets of genes and hepatic fatty acids in the two data sets
- In the Liver Toxicity data, the methodology sPLS-reg was applied
  - The aim was to highlight a subset of correlated genes which expression can predict the clinical chemistry measurements
- In addition,
  I. Obtained networks were used as an input to Cytoscape for visualization and
  II. Gene Ontology enrichment was used to assess the biological relevancy of the inferred associations between different types of variables

Nutrimouse data: results and graphical outputs



Correlation Circle plots for the Nutrimouse study. Correlation Circle plots for the first two sPLS dimensions (100 genes selected in total).

Nutrimouse data: results and graphical outputs

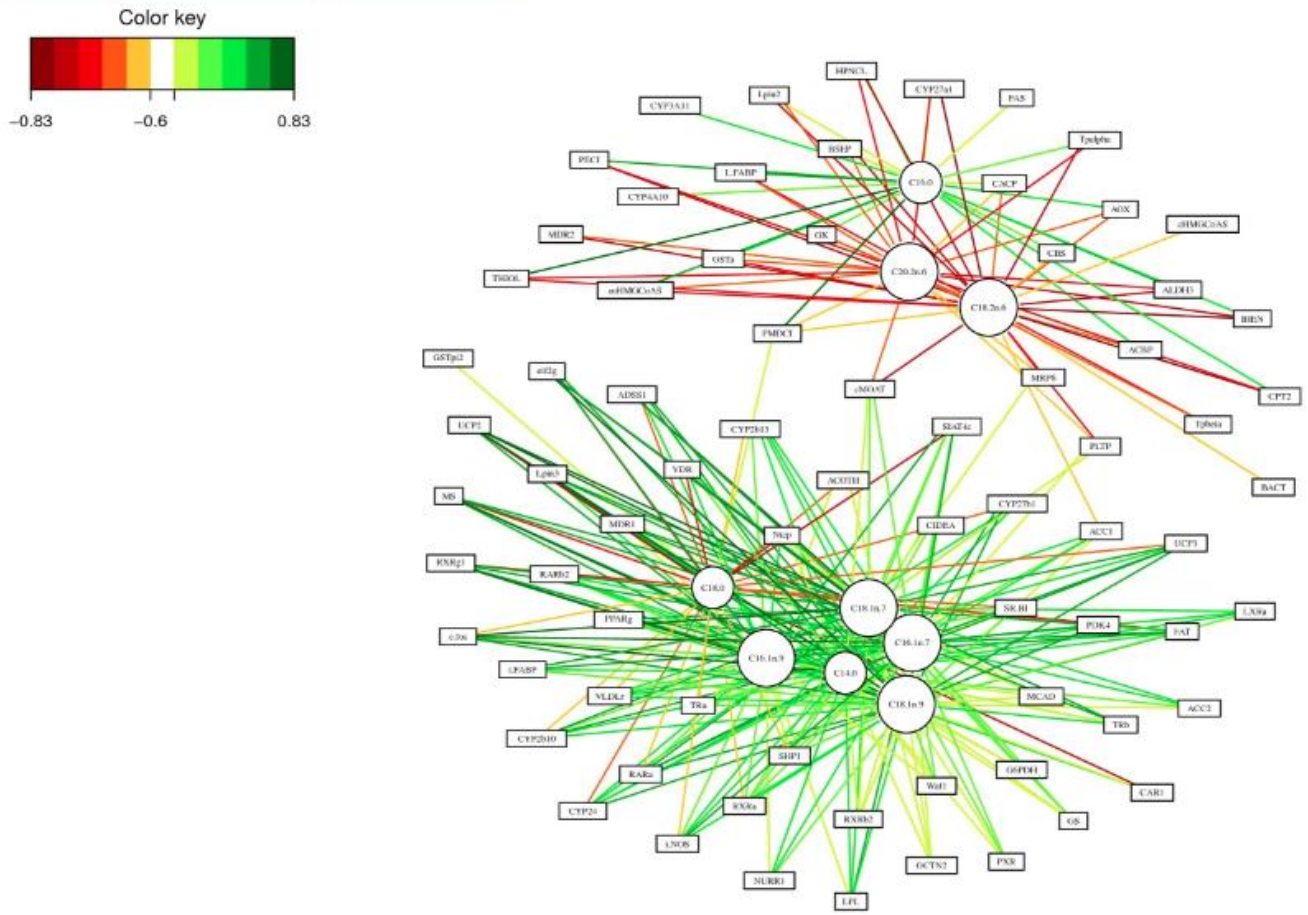


CIM for the Nutrimouse study. CIM for the first two sPLS dimensions (100 genes selected in total). Green (red) indicate high positive (negative) correlation.

Nutrimouse data: results and graphical outputs

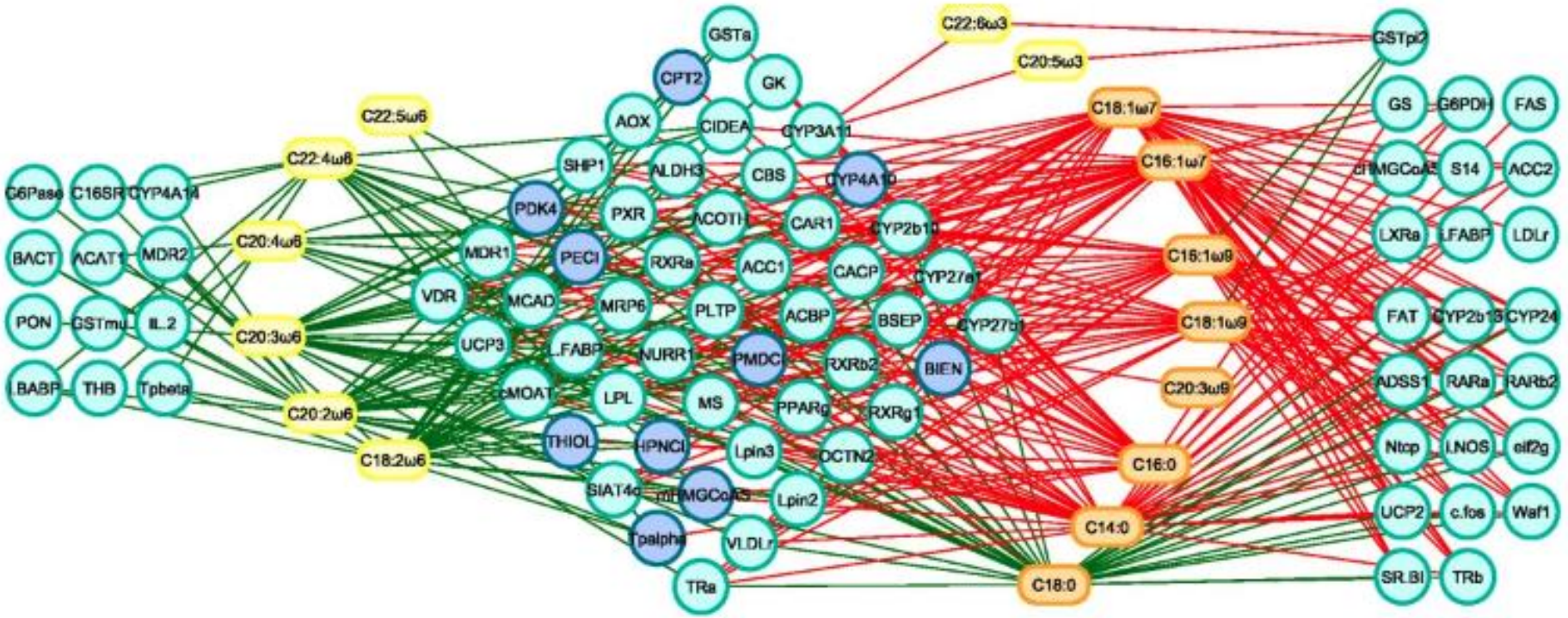
Figure 7

From: Visualising associations between paired 'omics' data sets



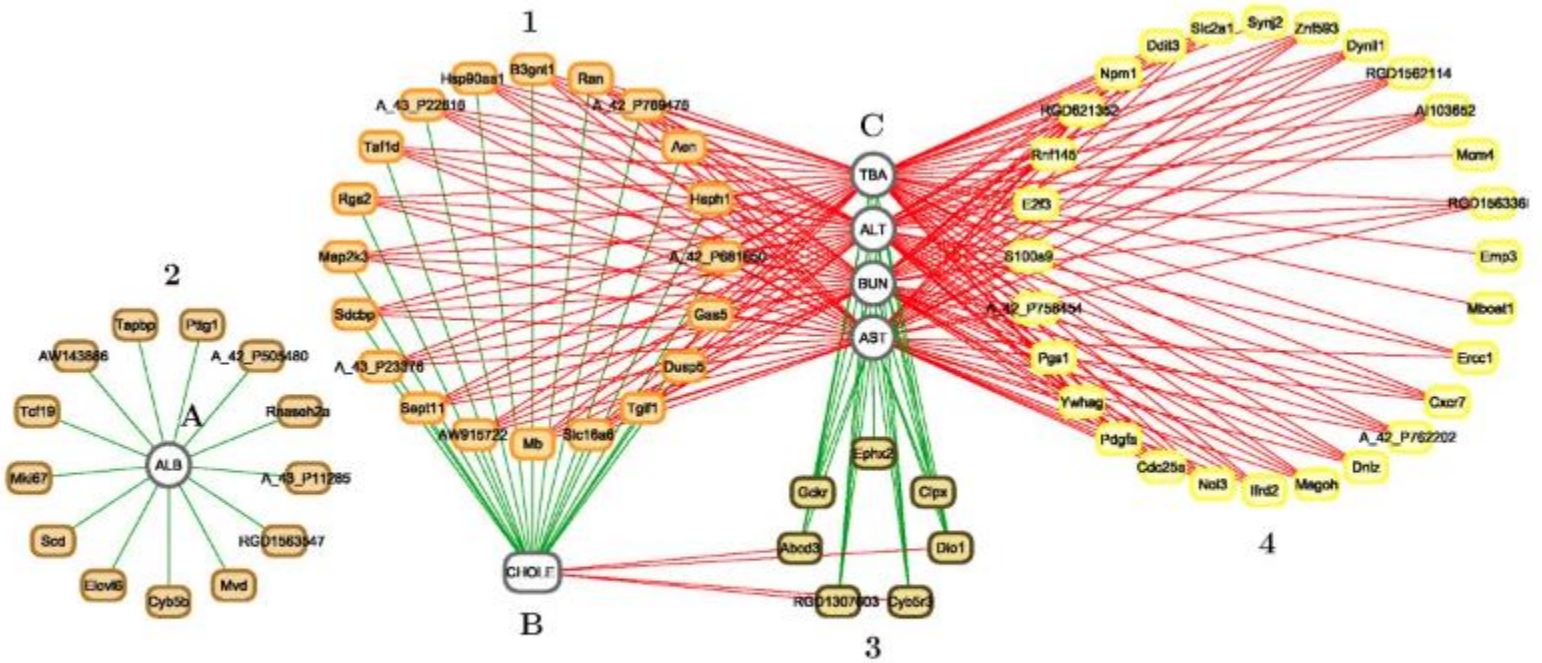
Relevance Networks for the Nutrimouse study. Relevance Networks obtained for the first two sPLS dimensions (100 genes selected in total). Green (red) indicate high positive (negative) correlation. Genes and fatty acids are represented respectively as rectangles and circles.

Nutrimouse data: results and graphical outputs



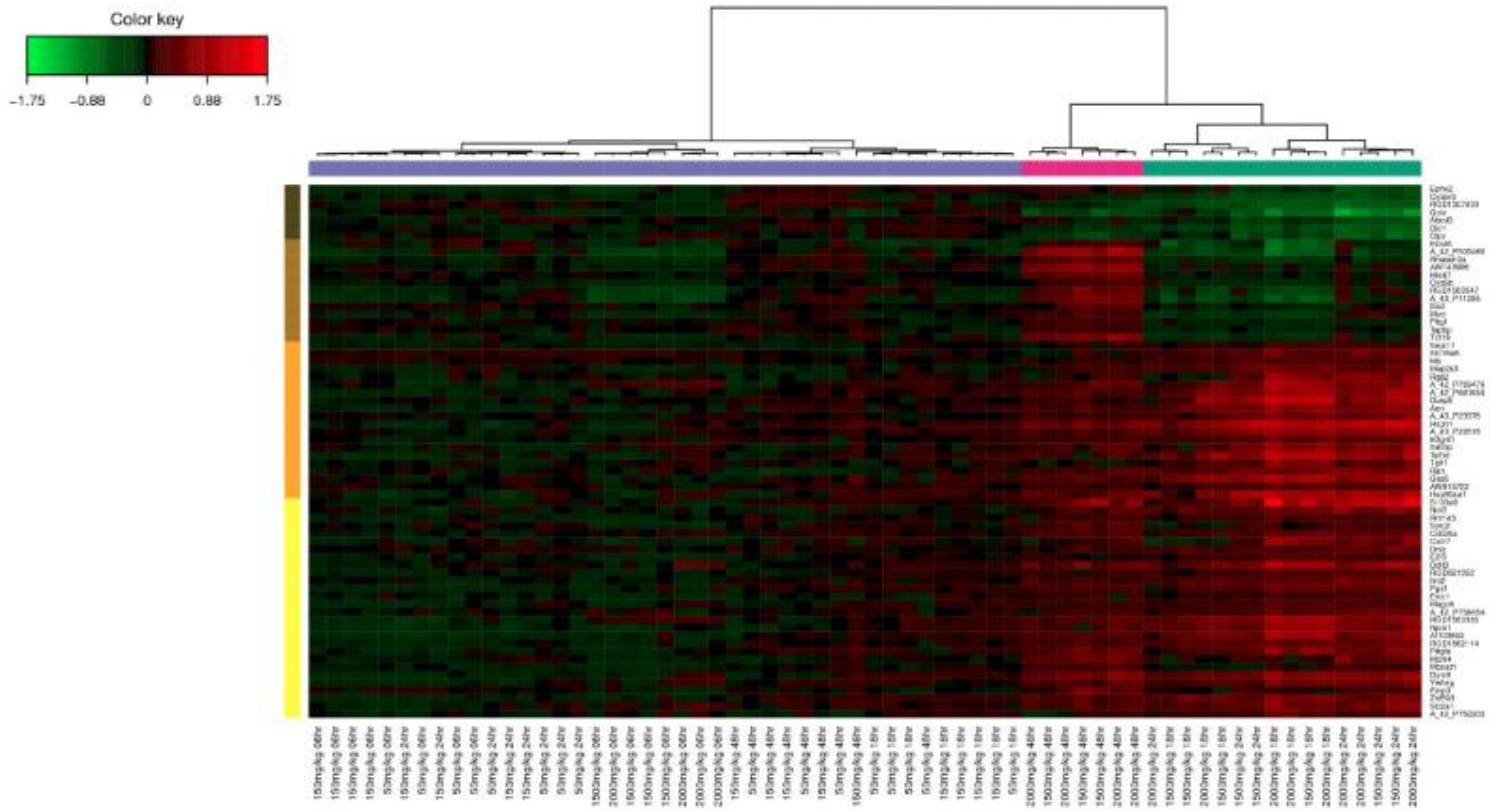
Relevance Networks for the Nutrimouse study. Relevance Networks generated with Cytoscape based on the proposed estimation of the pair-wise associations between seleted genes and fatty acids. Green and red edges indicate positive and negative correlation respectively.

Liver Toxicity data: results and graphical outputs



Relevance Networks for the Liver Toxicity study. Relevance Networks generated with Cytoscape based on the proposed estimation of the pair-wise associations between selected genes and clinical variables. Green and red edges indicate positive and negative correlation respectively. The network contained three groups of clinical chemistry measurements (white nodes): A [ALB], B [CHOLE] and C [ALT, AST, BUN, TBA] and four groups of genes (colored nodes) denoted 1, 2, 3 and 4.

Liver Toxicity data: results and graphical outputs



Hierarchical clustering of the selected genes for the Liver Toxicity study. Hierarchical clustering of the biological samples using the extracted genes from sPLS-reg network. Agglomerative hierarchical clustering was derived using the Euclidean distance as the similarity measure and Ward methodology. The resulting heatmap contains the genes in rows and samples in columns with red indicating up regulation, green down regulation and black no change. On the top of the heatmap, clusters of the biological samples are colored in violet, cyan and magenta for no, moderate or severe necrosis respectively. On the left hand side of the heatmap, gene clusters are shown (dark brown, brown, yellow and orange).

Liver toxicity data: GO enrichment

Top five networks enriched using GeneGO:

i) regulation of programmed cell death in response to stress

ii) Cell cycle and regulation of metabolism

iii) cholesterol and sterol metabolism

iv) regulation of programmed cell death in response to organic substances

v) response to stress and presentation of endogenous antigens

Conclusion

- This article described in detail several useful visualization techniques
- The value of these visualization methods were demonstrated on a simulation study and two biological studies
- Such graphical outputs are undeniably important to understanding biological questions via ‘omics’ data