

Insightful graphical outputs to explore relationships between two ‘omics’ data sets

Ignacio González^{*1}, Kim-Anh Lê Cao², Melissa Davis², Sébastien Déjean¹

¹Institut de Mathématiques - Université de Toulouse et CNRS, UMR 5219, F-31062 Toulouse, France

²Queensland Facility for Advanced Bioinformatics, University of Queensland, 4072 St Lucia, QLD, Australia

Email: Ignacio González - ignacio.gonzalez@math.univ-toulouse.fr; Kim-Anh Lê Cao - k.lecao@uq.edu.au; Melissa Davis - m.davis@imb.uq.edu.au; Sébastien Déjean - sebastien.dejean@math.univ-toulouse.fr;

*Corresponding author

Abstract

Background: Each omics platform is now able to generate a large amount of data. Genomics, proteomics, metabolomics, interactomics are compiled at an ever increasing pace and now form a core part of the fundamental systems biology framework. The integrative analysis of these data that are co jointly measured on the same samples represent analytical challenges to extract and visualise meaningful information.

Results: The exploratory statistical approaches ‘regularized Canonical Correlation Analysis’ and ‘sparse Partial Least Squares regression’ have been recently developed to deal with highly dimensional data, to integrate two types of ‘omics’ data and to select relevant information. Using the results of these methods, we propose further graphical developments to generate Clustered Image Map and Relevance Networks to better understand the relationships between ‘omics’ data and to better visualise the correlation structure between the different entities. We demonstrate the usefulness of such graphical outputs on several biological data sets. Using Cystoscape and GeneGo to further assess the biological relevance of such graphical tools, we show that the inferred networks are relevant to the system under study.

Conclusions: Such graphical outputs are undoubtedly useful to aid the interpretation of these promising integrative analysis tools and will certainly help in addressing fundamental biological questions and understanding systems as a whole.

Availability: The methods described in this paper are implemented in the freely available R package `mixOmics`.

Background

‘Omics’ data now form a core part of systems biology by enabling researchers to understand the integrated functions of a living organism. However, the available abundance of such data (genomics, proteomics, metabolomics, interactomics ...) is not a guarantee of obtaining useful information in the investigated system if the data are not properly processed and analyzed to highlight this useful information. A major challenge with the integration of omics data is therefore the extraction of discernable biological meaning from multiple omics data.

Recently, several authors have further improved statistical methodologies to integrate two highly dimensional data sets. Such methodologies include regularized and sparse variants of Canonical Correlation Analysis (CCA) [1–5] and Partial Least Squares (PLS) regression [6,7] - also referred as Projection to Latent Structures. However, most of the articles that present such approaches are limited to numerical results, and little attention is paid to either the interpretation of the results or the graphical outputs. The typical plot that accompanies CCA or PLS regression is a diagram of the correlations between variates and variables sometimes called *correlation circle* plot [8–12] as used with Principal Component Analysis (PCA). This graphical display allows to visualise strongly associated (or correlated) variables that are projected in the same direction and that are close the circle of radius one. However, in the high-dimensional context where many points (perhaps several thousands) are plotted on the correlation circle, the readability of such plot and the interpretability of the correlation structure of the variables can be very difficult. The variables in both data sets are intermingled as points on the plot, which interferes with clear labelling, and therefore a clear visualization. Thus, there is a need to simultaneously display the variables of the different types to visualise their association in a high dimensional setting.

We propose to generate Clustered Image Map (CIM) representation [13,14] and to infer Relevance Networks based on the results of CCA or PLS methods. Network correlation analysis has been extensively used to integrate metabolomics and transcriptomics data to identify co-regulation [15,16]. For example [17] recently proposed network cartography based on Pearson correlation to generate similarity matrices after applying PCA. We propose instead to use CCA or PLS as a first step analysis as these approaches are directly focusing on statistical integrative analysis of two highly dimensional data sets. The end products

are representations of the variables associations that enable biologists to explore and interpret the data in a natural and intuitive manner through statistical organization and graphical displays. These methods are implemented in the R package `mixOmics`¹ that is dedicated to the integrative analysis of ‘omics’ data [18].

In the following Results section, we first assess the relevance of the proposed CIM and Relevance Networks on a simulated data set. We then illustrate the use of such graphical outputs on two real data sets, provide a thorough biological interpretation of the results obtained and compare the inferred statistical networks to known biological networks using data and knowledge driven analyses. The Methods section gives a brief introduction of the two already published methodologies RCCA and SPLS which outputs are used to compute pair-wise similarity matrices. We then detail how these matrices can generate such CIM and Relevance Networks representations.

Results and Discussion

We investigate the relevance of CIM and Relevance Networks representations, firstly on a simulated data set to assess if the proposed graphical outputs are able to highlight pair-wise association structure between two data sets and to evaluate the quality of the inferred networks; and secondly on two biological data sets to assess the biological relevance of such graphical tools.

Simulated data

Data sets

We generated two data sets X and Y with an equal number of 30 observations in each data set, and applied RCCA and PLS-can (see Methods section). Both graphical representations CIM and Relevance Networks can be obtained for each methodology. A subset of relevant variables in X were associated with a subset of relevant variables in Y according to the model described below, and the remaining variables were simulated as noise. This simulation study enables to assess if the graphical representations can differentiate the associated groups of relevant variables from the noisy variables.

- The relevant X and Y variables were generated according to a normal distribution with zero mean and covariance matrix Σ defined by :

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma'_{XY} & \Sigma_{YY} \end{bmatrix}, \quad \text{with} \quad \Sigma_{XY} = \begin{bmatrix} A_{XY} & 0 & 0 \\ 0 & B_{XY} & 0 \\ 0 & 0 & C_{XY} \end{bmatrix}.$$

¹<http://www.math.univ-toulouse.fr/~biostat/mixOmics>

Details about the covariance matrices can be found in Additional file 1.

- X contains three independent sets of respectively 10, 10 and 3 cross-correlated variables: $X_A = [X_A^1, \dots, X_A^{10}]$, $X_B = [X_B^1, \dots, X_B^{10}]$ and $X_C = [X_C^1, X_C^2, X_C^3]$; and Y contains three independent sets of respectively 16, 5 and 2 cross-correlated variables: $Y_A = [Y_A^1, \dots, Y_A^{16}]$, $Y_B = [Y_B^1, \dots, Y_B^5]$ and $Y_C = [Y_C^1, Y_C^2]$. These groups of variables are associated with each other according to the cross-correlation matrix Σ_{XY} .
- The relevant variables in X_A and Y_A were generated with an absolute cross-correlation varying between 0.5 and 0.93, X_A positively correlated with $\{Y_A^k : k = 1, 2, 6, 13, 15\}$, and negatively correlated with the other variables in Y_A $\{Y_A^k : k = 3 : 5, 7 : 12, 14, 16\}$. The variables in X_B and Y_B were generated with a positive cross-correlation varying between 0.5 and 0.85; and the variables in X_C and Y_C were generated with an absolute cross-correlation varying between 0.81 and 0.93, X_C is positively correlated with Y_C^1 and is negatively correlated with Y_C^2 .
- The irrelevant (noisy) variables were simulated with a normal distribution with zero mean and covariance identity matrices and were added to the sets such that final data set contained 50 variables for X and 100 variables for Y . These variables are independent within the sets X and Y and with each other.

Analysis process

RCCA was applied to these data sets with regularization parameters $\lambda_1 = 0.889$ and $\lambda_2 = 0.889$. The regularization parameters were chosen using 10-fold cross-validation procedure on a regular grid of size 10×10 defined on the region $0.001 \leq \lambda_1 \leq 1$, $0.001 \leq \lambda_2 \leq 1$. To graphically represent the results of RCCA, we chose the first three dimensions as the canonical correlations values were of 0.959, 0.925, and 0.881, followed by much lower values. The tuning of the regularization parameters and the number of components is detailed in [19].

Figure 1 displays the CIM obtained with RCCA. The pair-wise association matrix was computed (see section Methods) for the first 3 dimensions. The Euclidian distance and the average agglomeration method were used for the hierarchical clustering. In the CIM display, each coloured block represents an association between subsets of the X -variables and the Y -variables. The red colour indicates that the X and Y clusters are positively correlated (cluster X_A and $\{Y_A^k : k = 1, 2, 6, 13, 15\}$, cluster X_B and Y_B , and cluster

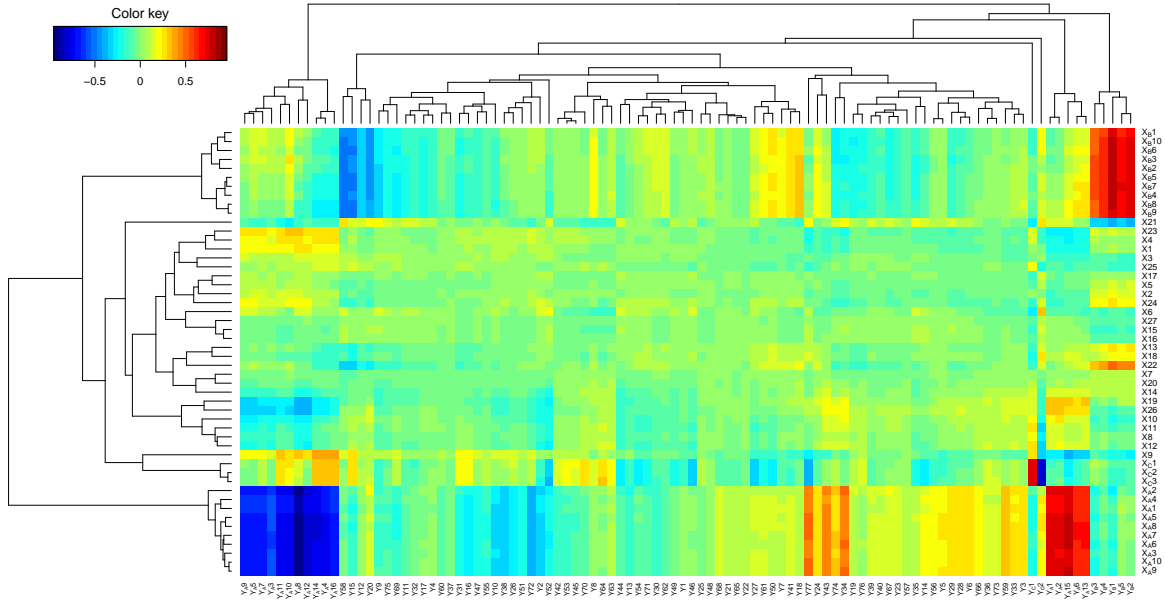


Figure 1: **CIM from simulated data.** CIM was derived using the `cim` function in the `mixOmics` package on the simulated data with the RCCA method. The red and blue colours indicate strong positive and negative correlations respectively, whereas yellow or green indicate weaker correlation values.

X_C and Y_C^1), and the blue colour indicates a negative correlation in the X - Y cluster (cluster X_A and $\{Y_A^k : k = 3 : 5, 7 : 12, 14, 16\}$, and cluster X_C and Y_C^2), whereas yellow or green indicate weaker correlation values. The dendrograms on the top and the left hand side of the map indicate how the clusters join, the longer the distance, the sharper the boundary between the coloured blocks.

The Relevance Networks obtained with PLS-can are displayed in Figure 2. Similarly to RCCA, the pair-wise association matrix was computed (see section Methods) for the first 3 dimensions. The Relevance Networks were produced using the `network` function in the `mixOmics` package, with a fixed threshold set to 0.53. Three relevant networks were obtained. It can be seen that each network links the corresponding correlated subsets: X_A with Y_A , X_B with Y_B and X_C with Y_C . Similar networks were obtained with RCCA.

Quality of the inferred network

We then investigated the accuracy of the generated networks with this same simulation setup. We considered as positive edges a simulated correlation between two variables (represented as nodes) greater than 0.5 in absolute value and negative otherwise. False positive occurs in the resulting network when an

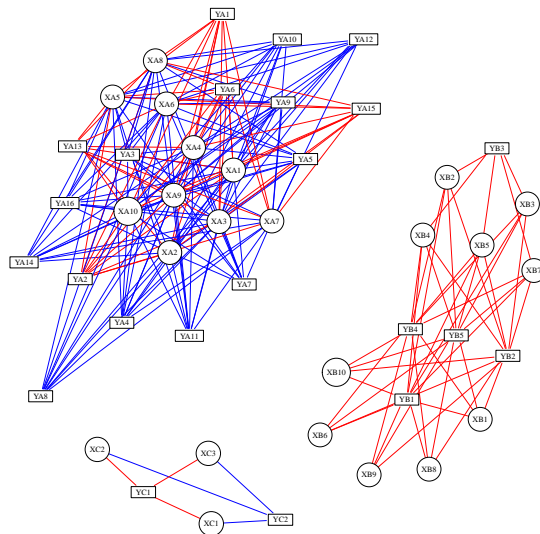


Figure 2: **Relevance networks from simulated data.** Relevance networks obtained with SPLS-can on the simulated data using the `network` function in the `mixOmics` package. Red and blue edges indicates positive and negative correlation respectively. X and Y variables are represented respectively as circles and rectangles.

edge links two variables with a correlation less than 0.5. False negative occurs when two variables with a correlation less than 0.5 are linked in the network. Five hundred simulations with 30 samples were performed. For each simulated X and Y variables, networks were inferred from the first three components of PLS-can for a threshold ranging from 0 to 1 with a step of 0.025. Positive Predictive Values (PPV, the proportion of correctly identified edges among all positive edges) and sensitivity (the proportion of positives edges correctly identified) were averaged over the 500 inferred networks for each threshold value.

Figure 3 displays the corresponding PPV and sensitivity. For this simulation setup the PPV is very close to 1 for a threshold higher than 0.45. This indicates that if an edge is built in the network then the probability that it actually corresponds to a true edge is very high. Regarding sensitivity, Figure 3 shows that the network builds almost all or all true positives edges for a threshold higher than 0.4.

This simulation study shows that Relevance Networks and CIM derived from PLS-can and RCCA are able to highlight the relevant variables amongst the noisy ones and pinpoint the pair-wise association structure between the two data sets. In the following, we illustrate the use of such graphical outputs on highly

dimensional data sets and discuss the biological relevancy of the networks obtained.

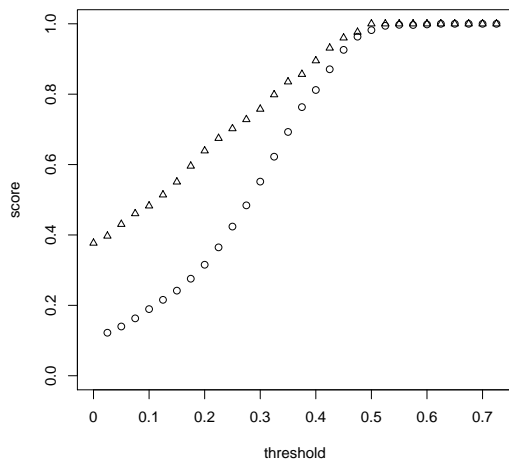


Figure 3: **Results of the accuracy study.** Sensitivity (triangle), Specificity (diamond), Positive Predictive Value (circle) for $n = 15$ (left) and $n = 30$ (right).

Biological data

Data sets

These data sets are publicly available in the `mixOmics` package [18].

Nutrimouse data. The data come from a nutrigenomic study [20] in which 40 mice from two genotypes (wild-type and $PPAR\alpha$ $-/-$ deficient) were fed with five diets with contrasted fatty acid compositions. Oils used for experimental diets preparation were corn and colza oils (50/50) for a reference diet (REF), hydrogenated coconut oil for a saturated fatty acid diet (COC), sunflower oil for an Omega6 fatty acid rich diet (SUN), linseed oil for an Omega3 rich diet (LIN) and corn/colza/enriched fish oils (43/43/14) for the FISH diet. Expression of 120 genes in liver cells were acquired through microarray experiment and concentrations of 21 hepatic fatty acids were measured by gas chromatography.

Liver toxicity data. The data come from a liver toxicity study [21] in which 64 male rats of the inbred strain Fisher F344/N were exposed to low (50 mg/kg or 150 mg/kg) or to high (1500 mg/kg or 2000 mg/kg) doses of acetaminophen (paracetamol) in a controlled experiment. Necropsies were performed at 6, 18, 24 and 48 hours after exposure and the mRNA from the liver was extracted. Ten clinical chemistry

measurements of variables containing markers for liver injury are available for each subject and the serum enzymes levels are numerically measured.

Analysis process

To take into account the biological question of each study, we applied SPLS-can to the Nutrimouse data and SPLS-reg to Liver Toxicity (see the Methods section for a description of these methodologies). In the Nutrimouse study, it cannot be assumed that variations in one set of variables can cause variations in the other one as we do not a priori know if gene expression changes imply fatty acid concentrations changes or inversely. Therefore, the use of SPLS-can is justified to perform a symmetric analysis [1]. On the contrary, in the Liver Toxicity data, an asymmetric (regression-based) analysis was performed as we attempt to predict the clinical parameters Y with the gene expression matrix X (as was also performed in [22]).

For both data sets, we arbitrarily chose to select 50 variables on each dimension. This can be justified by the illustrative purpose of this section, as well as the need to select a sufficient number of variables in order to assess their biological relevance with a Gene Ontology (GO) analysis. Regarding the choice of the number of dimensions, we chose to keep the first 3 dimensions in Nutrimouse, as was suggested by [1]. In the Liver Toxicity study, [7] showed that 3 dimensions seemed to be sufficient to explain most of the correlation or the covariance structure of the data. Therefore, the similarity matrices were computed on the basis of the selected variables on the first 3 components in both data sets.

To highlight the strongest variable associations only, variables with an association score greater than 0.6 in absolute value were chosen to infer the Relevance Networks. This threshold was arbitrarily chosen in order to obtain biologically interpretable networks that were neither too sparse nor too dense. The obtained networks were then used as an input to Cytoscape [23] for visualization and GeneGo [24] and topGO [25,26] were used to assess the biological relevancy of the inferred associations between the different types of variables (see Additional File 2 for the R script used). This analysis is similar to the one performed by [27] who assessed the results of RCCA in a metabolic syndrome study. We then compared the obtained inferred networks to known biological networks through data driven and knowledge driven biological analyses.

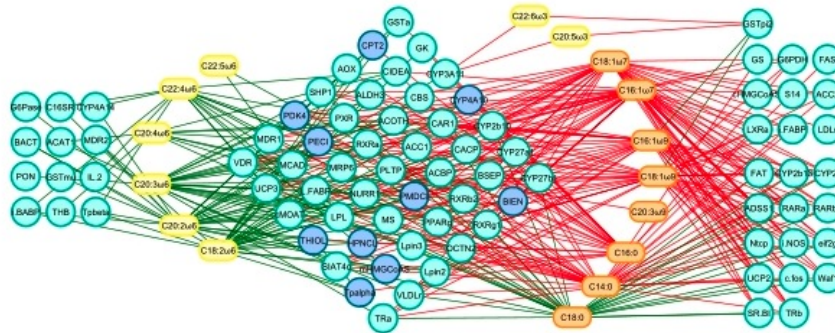


Figure 4: **Relevance networks from Nutrimore data.** Relevance networks generated with Cytoscape based on the output from `network` function in the `mixOmics` package. Red and green edges indicate positive and negative correlation respectively.

Application to Nutrimore data.

The Relevance Network generated for the Nutrimore data at a threshold 0.6 highlighted two clusters of fatty acids, and three clusters of genes (Fig. 4). Considering first the fatty acids, the yellow cluster on the left-hand side contained all the $\omega 6$ fatty acids from the data set (C18:2 $\omega 6$, C20:2 $\omega 6$, C20:4 $\omega 6$, C20:3 $\omega 6$, C22:5 $\omega 6$, and C22:4 $\omega 6$). The second group of fatty acids consisted of those in the $\omega 9$, $\omega 7$, and saturated fatty acid groups, along with the two $\omega 3$ fatty acids included in the data set. These clusters made sense in the context of lipid biosynthetic pathways – one biosynthetic pathway leads to the production of $\omega 6$ lipids, while the $\omega 9$, $\omega 7$ and saturated lipids are the product of an alternative lipid biosynthetic pathway (orange nodes). The $\omega 3$ group was the exception in our analysis – it was generated by a pathway related to the $\omega 6$ pathway (yellow nodes), but based on the connectivity in our network, these fatty acids partitionned with the $\omega 7$, $\omega 9$ and saturated fatty acid group [28].

The three gene sets defined by network topology were: (1) a set of genes that were negatively correlated with only the $\omega 6$ lipid group; (2) a set of genes that were negatively correlated with the $\omega 6$ group, but largely positively correlated with the other lipid group; and (3) a gene set that was only associated with the second lipid group, with positive correlations to the $\omega 3$, $\omega 7$, $\omega 9$, and saturated fatty acids C14:0 and C16:0, but negatively correlated with the C18:0.

The $\omega 6$ group showed only negative correlations with genes selected by SPLS-can. This was consistent with the observations made by [20] that feeding mice a diet rich in $\omega 6$ fatty acids lead to the down regulation of

several genes on the array.

The second group of genes contained many targets of PPAR α , a nuclear receptor transcription factor associated with the high-level regulation lipid metabolism (dark blue nodes). PPAR α targets are expected to be associated with long-chain polyunsaturated fatty acids from the ω 3 family, while the final subset of genes involved in lipid biosynthesis is expected to be closely associated with the saturated and monosaturated fatty acids of the ω 7 and ω 9 families. Both of these associations were apparent in the network. An in-depth analysis of the Nutrimouse data is behind the scope of this article. The reader can refer to [20,28] for more details about the underlying biological interpretation.

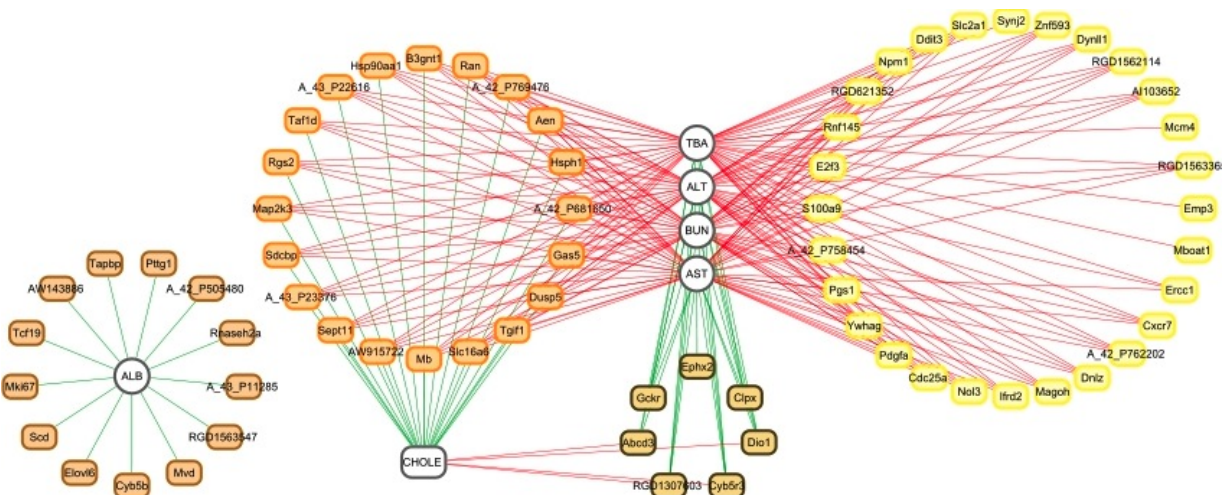


Figure 5: **Relevance networks from Liver Toxicity data.** Relevance networks generated with Cytoscape based on the output from `network` function in the `mixOmics` package. Red and green edges indicate positive and negative correlation respectively.

Application to Liver Toxicity data.

Visualization of the extracted genes with Cytoscape. Relevance networks for the Liver Toxicity data were generated from the results obtained with the SPLS-reg method. The selected variables with a pair-wise association score greater than 0.6 in absolute value were used as an input to Cytoscape (Fig. 5). This network contained three clusters of clinical chemistry measurements and four clusters of genes. Considering first the chemistry measurements (grey nodes), cluster 1 and 2 only consisted of cholesterol [CHOLE] and albumin [ALB] levels respectively. The third cluster contained indicators of liver injury (Alanine

Aminotransferase [ALT] and Aspartate aminotransferase [AST]), indication of renal injury (urea nitrogen [BUN]), and assessment of cholestasis – bile flow interruption (total bile acids [TBA]).

The four gene clusters defined by network topology (Fig. 5) were: (1) a set of genes that were positively correlated with the cholesterol levels but negatively correlated with the third cluster of clinical chemistry measurements (dark brown nodes); (2) a set of genes that were negatively correlated with ALB levels only (brown nodes); (3) a set of genes largely positively correlated with the third cluster of chemistry measurements but negatively correlated with the cholesterol levels (orange nodes); and (4) a gene set with only positive correlations to the third cluster of chemistry measurements (yellow nodes).

Biological relevance of the extracted genes. Hierarchical clustering (heatmap) of the biological samples on the extracted genes is displayed in Figure 6. This clustering reveals a very good grouping of the rats that underwent different doses of acetaminophen (also found in [21]). Clusters labelled (coloured at the top of the heatmap) with either no (violet), moderate (cyan) or severe (magenta) necrosis of the centrilobular region of the rat liver was obtained by using the expression values of the genes extracted from the network. Levels of the clinical chemistry measurements on each group of samples are given in Additional File 3. Figure 6 also highlights the differences in gene expression profiles between each gene cluster (coloured in dark brown, brown, orange and yellow at the left side of the heatmap). Gene expression differences are clearly observed between the clusters.

The extracted genes were uploaded into topGO [25,26]. A Gene Ontology (GO) enrichment analysis from the gene list was then performed. GO terms significantly enriched include biological processes related to nitric oxide metabolism and cellular stress responses, including responses to unfolded proteins. The top GO molecular functions enriched in the gene set relate to protein binding, nucleotide binding, and enzyme activity (eg. hydrolase, phosphatase, decarboxylase). Cellular component GO terms enriched in the set mostly relate to very general locations, however both an endopeptidase complex and the peroxisome are also present in the list, reinforcing the association of the selected gene products with proteolysis and the response to stress and unfolded proteins.

The individual gene clusters in the SPLS-reg network (Fig. 5) may also be examined for GO enrichment, as we have done for the larger cluster 4. For example, while examining the biological process terms associated this cluster, we saw an enrichment for processes involving xenobiotic transport, and interesting

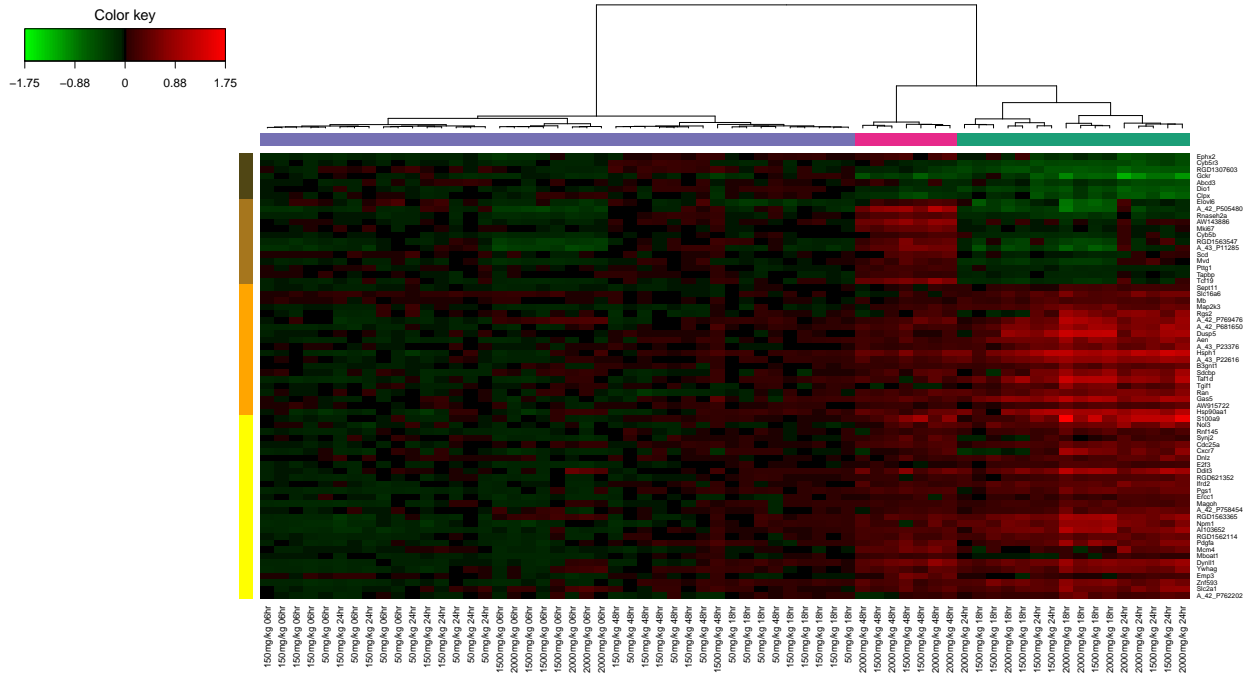


Figure 6: **Liver Toxicity heatmap.** Hierarchical clustering of the biological samples using the extracted genes from SPLS-reg network. Agglomerative hierarchical clustering was derived using the Euclidean distance as the similarity measure and the Ward methodology. The resulting heatmap contains the genes as the rows and samples as the columns. Red colour indicates up regulation, green down regulation and black indicates no change. On the top of the heatmap, clusters of the biological samples are coloured in violet, cyan and magenta for no, moderate or severe necrosis respectively. On the left-hand side of the heatmap, gene clusters are displayed (dark brown, brown, yellow and orange).

functional enrichments such as positive regulation of mesenchymal cell proliferation, a process that was previously observed to occur in other tissues in response to epithelial damage signalling to the underlying mesenchyme to initiate proliferation and tissue remodelling [29], and negative regulation of CREB transcription factor activity, interesting due to the previous association of CREB transcription factor with responses to cytotoxic stress [30,31], particularly in renal tubular cells [32].

Analysis of the gene list using the GeneGo [24] network analysis algorithm identified a total of 14 networks with a significant enrichment of genes in the Relevance Network. The top five networks were (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. A summary of these networks can be found in Additional file 4.

Conclusions

Several methodologies have been recently improved to jointly analyse two data sets. Therefore, the developments or the improvements of graphical tools are now crucial to better visualise and understand such complex biological data. In the omics era in particular, the deluge of data can make the interpretation of the results extremely difficult. In this paper, we proposed two types of graphical displays to complement the graphics usually used in CCA and PLS related methods. Both CIM and Relevance Networks representations are based on the evaluation of a pair-wise similarity measure. These graphical outputs are implemented in the R package `mixOmics` that is freely available; their biological relevancy was further assessed using GO analysis. The results obtained on simulated and real data sets illustrated very well the usefulness of these graphical outputs to further explore the relationships between two omics data sets. The thorough biological interpretation of the obtained inferred networks also demonstrated the relevancy of the approach.

Methods

We consider two approaches for visualizing correlation structures between two data sets: CIM and Relevance Networks. Both graphical displays require an estimated largescale association or pair-wise similarity matrix M as an input. Previously, several similarity measures have been proposed, including Pearson correlation coefficient [13, 33–35], entropy and mutual information [36]. We propose instead to compute the pair-wise similarity matrix using the results of either PLS or CCA approaches.

This section is organized as follows: we first give the user some background about the PLS and CCA methodologies and associated variants recently developed for the highly dimensional case, we then describe how to compute the pair-wise similarity matrix based on the results obtained via these integrative approaches in order to construct Relevance Networks and CIM.

Background: CCA and PLS based methods

Notations

We focus on two-block data matrices denoted $X(n \times p)$ and $Y(n \times q)$ where the p variables X^j and q variables Y^k are of two types and are measured on the same samples or observations n . Throughout the article, we will adopt the following notations: M_k^j represents the element of the k th row and j th column of the matrix M .

CCA

CCA [37] looks for the largest correlation between a linear combination of the variables in the first set X and a linear combination of the variables in the second set Y . The first pair maximizes the correlation $\rho_1 = \text{cor}(Xa^1, Yb^1)$ subject to $\text{var}(Xa^1) = \text{var}(Yb^1) = 1$. The subsequent pairs (Xa^l, Yb^l) , ($l = 1, \dots, \min(p, q)$) maximize the residual correlation with the additional requirements that each pair is to be uncorrelated with the previous pairs. In the following, we will refer to a^l and b^l as the *canonical loadings* (or weights). The resulting variables $U^l = Xa^l$ and $V^l = Yb^l$ are called the *canonical variates* and ρ_l are known as the *canonical correlations*.

PLS

PLS [38] looks for a decomposition of centered (possibly standardized) matrices X and Y . The decomposition is performed using orthogonal scores, also called *latent variables* or *variates*, (U^1, \dots, U^s) and (V^1, \dots, V^s) that are n -dimensional vectors and associated *loadings*, (a^1, \dots, a^s) and (b^1, \dots, b^s) that are p and q -dimensional vectors respectively; s is the number of chosen *dimensions* or *components* of PLS. The vectors (U^1, \dots, U^s) and (a^1, \dots, a^s) are associated to the X data set, and the vectors (V^1, \dots, V^s) and (b^1, \dots, b^s) are associated to the Y data set. The optimization problem to solve is [39]:
 $\max_{\{a^l, b^l\}} \text{cov}(X_{(l-1)}a^l, Yb^l)$ subject to $\|a^l\| = \|b^l\| = 1$, where $X_{(l-1)}$ is the residual (deflated) X matrix for each PLS dimension l .

Many PLS algorithms exist, not only for different shapes of data (SIMPLS [40], PLS1 and PLS2 [38], PLS-SVD [41]) but also for different aims (predictive like PLS2, or modelling like PLS-mode A, see [2, 12, 42]). In the present paper, we will refer to a PLS approach with two different aims. *PLS-reg* (for PLS-regression mode) will be used when one wants to model an ‘asymmetric’ or uni-directional relationship between the two data sets, i.e. we want to predict the matrix Y with the data X . *PLS-can* (for PLS-canonical mode) uses a different deflation step to relate the two data sets in a ‘symmetric’ way and therefore models a bi-directional relationship. This is a similar purpose to CCA’s.

Regularized and sparse based methods

In classical CCA and PLS analysis, all variables from both sets are included in the fitted linear combinations or variates. However, in the context of high throughput biological data, the number of variables often exceeds tens of thousands. In this case, linear combinations of the entire set of features

make biological interpretability difficult as they contain too many variables to perform further tests or to generate biological hypotheses. Most importantly, the high dimensionality and the insufficient sample size lead to computational problems as CCA requires the computation of the inverse of matrices $X'X$ and $Y'Y$.

RCCA. To circumvent this problem, [1] developed a regularized (or ridge) extension of CCA (RCCA).

RCCA solves the instability of the loadings due to multicollinearity by adding a regularization term on the diagonal of the ill-conditioned matrices, i.e. the covariance matrices. Thus, highly correlated variables get similar loadings, resulting in a grouping effect. The regularization terms λ_1 and λ_2 associated to each data set are chosen by cross-validation in order to maximize the first canonical correlation.

SPLS. Several sparse PLS have been proposed in the literature to select variables [6, 7]. These approaches introduce l_1 (Lasso) penalization terms on the loading vectors to shrink some of the coefficients towards zero, thus allowing for simultaneous variables selection in the two data sets. The sparse PLS therefore solves the problem of interpretability by selecting variables from both sets and therefore providing sparse sets of associated variables. In the article, we consider the sparse PLS proposed by [7] since both regression (SPLS-reg) and canonical mode (SPLS-can, [4]) are available. For practical purposes, the two penalization parameters associated to each data set were replaced by the number of variable to select on each data set and on each SPLS dimension. More details about the tuning of these parameters can be found in [7].

Both RCCA and SPLS are implemented in `mixOmics`. These approaches require to choose the number of dimensions s and the regularization/penalization parameters associated to X and Y .

Pair-wise variable associations for CCA

The association measure that we propose to use is analogous to a correlation coefficient. Firstly, similar to a correlation circle output, the X^j and Y^k variables are projected onto a low dimensional space. Let $s \leq \min(p, q)$ the selected dimensions to adequately account for the data association, and let $Z^l = U^l + V^l$ the equiangular vector between the canonical variates U^l and V^l ($l = 1, \dots, s$). The coordinates of the variable X^j and Y^k are obtained by projecting them on the axes defined by Z^l . The projection on the Z axes seems the most natural as X and Y are symmetrically analysed in CCA. Furthermore, Saporta [11] showed that the Z variables have the property to be the closest to X and Y , i.e. the sum of their squared

multiple correlation coefficients with X and with Y is maximal.

Let $h^j = (h_1^j, \dots, h_s^j)'$ and $g^k = (g_1^k, \dots, g_s^k)'$ the coordinates of the variable X^j and Y^k respectively on the axes defined by Z^1, \dots, Z^s . These coordinates are obtained by computing the scalar innerproduct

$h_l^j = \langle X^j, Z^l \rangle$ and $g_l^k = \langle Y^k, Z^l \rangle$ ($l = 1, \dots, s$). As the variables X^j and Y^k are assumed to be of unit variance, the innerproduct is equal to the correlation between the variables X (or Y) and Z :

$$h_l^j = \text{cor}(X^j, Z^l) \text{ and } g_l^k = \text{cor}(Y^k, Z^l).$$

Then, for any two variables X^j and Y^k , a similarity score can be computed as follows:

$$M_k^j = \langle h^j, g^k \rangle = \|h^j\|_2 \|g^k\|_2 \cos \theta(h^j, g^k) = \sum_{l=1}^s h_l^j g_l^k, \quad (1)$$

where $\theta(h^j, g^k)$ is the angle between the vectors h^j and g^k , and $0 \leq |M_k^j| \leq 1$. The matrix M can be factorized as $M = GH'$ with G and H matrices of order $(p \times s)$ and $(q \times s)$ respectively. When $s = 2$, M is represented in the correlation circle by plotting the rows of G and the rows of H as vectors in a 2-dimensional Cartesian coordinate system. Therefore, the innerproduct of the X^j and Y^k coordinates is an approximation of their association score.

Pair-wise variable associations for PLS

For PLS-reg, the association score M_k^j between the variables X^j and Y^k can be obtained from an approximation of their correlation coefficient. Let r the rank of the matrix X , PLS-reg allows for the decomposition of X and Y by [43]:

$$X = U^1(\phi^1)' + U^2(\phi^2)' + \dots + U^r(\phi^r)' \quad (2)$$

$$Y = U^1(\varphi^1)' + U^2(\varphi^2)' + \dots + U^r(\varphi^r)' + E^{(r)} \quad (3)$$

where ϕ^l and φ^l , are the regression coefficients on the variates U^1, \dots, U^r , and $E^{(r)}$ is the residual matrix ($l = 1, \dots, r$). By denoting u_l the standard deviation of U^l , using the orthogonal properties of the variates and the decompositions in (2) and (3), we obtain $h_l^j = \text{cor}(X^j, U^l) = u_l \phi_j^l$ and $g_l^k = \text{cor}(Y^k, U^l) = u_l \varphi_k^l$.

Let $s < r$ the number of components selected to adequately account for the variable association, then for any two variables X^j and Y^k , the similarity score is defined by:

$$M_k^j = \langle h^j, g^k \rangle = \sum_{l=1}^s h_l^j g_l^k = \sum_{l=1}^s u_l^2 \phi_j^l \varphi_k^l \approx \text{cor}(X^j, Y^k), \quad (4)$$

where $h^j = (h_1^j, \dots, h_s^j)'$ and $g^k = (g_1^k, \dots, g_s^k)'$ are the coordinates of the variable X^j and Y^k respectively on the axes defined by U^1, \dots, U^s . When $s = 2$, a correlation circle representation is obtained by plotting h^j and g^k as points in a 2-dimensional Cartesian coordinate system.

For PLS-can, the association score M_k^j is calculated by substituting $g_l^k = \text{cor}(Y^k, V^l)$ in (4) for $l = 1, \dots, s$, as in this case the decomposition of Y is given by:

$$Y = V^1(\varphi^1)' + V^2(\varphi^2)' + \dots + V^r(\varphi^r)' + E^{(r)}$$

where φ^l ($l = 1, \dots, r$), are the regression coefficients on the variates V^1, \dots, V^r . Then,

$$\text{cor}(X^j, Y^k) \approx \sum_{l=1}^s u_l^2 \sigma_l^2 \phi_j^l \varphi_k^l = M_k^j$$

where σ_l^2 is the variance of V^l .

Constructing Relevance Networks

A conceptually simple approach for modelling net-like correlation structures between two data sets is to use *Relevance Networks*. This concept was introduced by Butte *et al.* [33] as a tool to study associations between couples of variables coming from several types of genomic data. This method generates a graph where nodes represent variables, and edges represent variable associations. The Relevance Network is built in the following simple manner. First, the correlation matrix is inferred from the data. Second, for every estimated correlation coefficients exceeding a prespecified threshold between two variables (say 0.6 in our examples), an edge is drawn between these two variables.

The construction of biological networks (gene-gene, protein-protein, etc.) with direct interactions within a variable set is of considerable interest amongst biologists, and has been extensively used in the literature. Therefore, we will not consider this case and rather focus on the representation between X and Y data sets, i.e., the representation of variables of two different types. We will thus display RCCA, SPLS-can and SPLS-reg Relevance Networks through the use of bipartite graph (or bigraph), that is, every node of one variable set X is connected to nodes of the other variables set Y only.

Bipartite networks are inferred using the pair-wise association matrix M defined in (1) and (4) for CCA and (S)PLS results respectively. Entry M_k^j in the matrix M represent the association score between X^j and Y^k variables. Then, by setting a user-defined score threshold, the pairs of variables X^j and Y^k with a

$|M_k^j|$ value greater than the threshold will be aggregated in the Relevance Network. By changing this threshold, the user can choose to include or exclude relationships in the Relevance Network. This option is proposed in an interactive manner in the `mixOmics` package [18].

Relevance networks for RCCA assume that the underlying network is fully connected, i.e. that there is an edge between any pair of X and Y variables. For SPLS-reg and SPLS-can, relevance networks are solely represented for the variables selected in the model. In this case, M_k^j pair-wise associations are calculated based on the selected variables.

Displaying CIM

CIM or heatmaps were introduced in [13,14] to represent data resulting from gene expression profiles. This type of representation is based on a hierarchical clustering simultaneously operating on the rows and columns of a real-valued similarity matrix M . The initial matrix is graphically represented as a 2-dimensional coloured image, where each entry of the matrix is coloured on the basis of its value, and where the rows and columns are reordered according to a hierarchical clustering. Dendrograms resulting of the clustering are added to the left (or right) side and to the top (or bottom) of the image. With RCCA, SPLS-can and SPLS-reg, we chose to display CIM based on the pair-wise similarity matrix M defined in (1) and in (4).

Authors contributions

IG performed the statistical analysis, the network analysis, wrote the R functions and drafted the manuscript. KALC performed the statistical analysis and helped to draft the manuscript. MD performed the network analysis. SD participated in the design of the manuscript and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Dr. Pierre Bushel (National Institute of Environmental Health Sciences) for his assistance on the Liver Toxicity study and Dr. Pascal Martin (Institut National de la Recherche Agronomique) for his feedback on the Nutrimouse study.

References

1. González I, Déjean S, Martin P, Gonçalves O, Besse P, Baccini A: **Highlighting relationships between heterogeneous biological data through graphical displays based on regularized Canonical Correlation Analysis**. *Journal of Biological Systems* 2009, **17**(2):173–199.
2. Waaijenborg S, Verselewe de Witt Hamer PC, Zwinderman A: **Quantifying the association between gene expressions and dna-markers by penalized canonical correlation analysis**. *Statistical Applications in Genetics and Molecular Biology* 2008, **7**.
3. Parkhomenko E, Tritchler D, Beyene J: **Sparse canonical correlation analysis with application to genomic data integration**. *Statistical Applications in Genetics and Molecular Biology* 2009, **8**:1–34.
4. Lê Cao KA, Martin P, Robert-Granié C, Besse P: **Sparse canonical methods for biological data integration: application to a cross-platform study**. *BMC Bioinformatics* 2009, **10**(34).
5. Witten DM, Tibshirani R, Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009, **10**(3):515–534.
6. Chun H, Keles S: **Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis**. *Technical report, Department of Statistics, University of Wisconsin, Madison, USA* 2007.
7. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P: **A sparse PLS for variable selection when integrating omics data**. *Statistical Applications in Genetics and Molecular Biology* 2008, **7**(29).
8. Caillez F, Pages JP: *Introduction à l'analyse des données*. Paris, SMASH, Mathématiques et sciences humaines 1976.
9. van der Burg E, de Leeuw J: **Non-linear canonical correlation**. *British Journal of Mathematical and Statistical Psychology* 1983, **36**:54–80.
10. van der Geer JP: *Introduction to linear multivariate data analysis, Vol. 1*. Leiden: DSWO Press. 1986.
11. Saporta G: *Probabilités analyse des données et statistique*. Technip 2006.
12. Tenenhaus M: *La régression PLS: théorie et pratique*. Technip 1998.
13. Weinstein JN, Myers TG, O'Connor PM, Friend SH, Fornace Jr AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johnson GS, Wittes RE, Paull KD: **An information-intensive approach to the molecular pharmacology of cancer**. *Science* 1997, **275**:343–349.
14. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proceeding of the National Academy of Sciences of the USA* 1998, **95**:14863–14868.
15. Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie A: **Parallel analysis of transcript and metabolic profiles: a new approach in systems biology**. *EMBO reports* 2003, **4**(10):989–993.
16. Saito K, Hirai M, Yonekura-Sakakibara K: **Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'**. *Trends in plant science* 2008, **13**:36–43.
17. Allen E, Moing A, Ebbels T, Maucourt M, Tomos A, Rolin D, Hooks M: **Correlation Network Analysis reveals a sequential reorganization of metabolic and transcriptional states during germination and gene-metabolite relationships in developing seedlings of Arabidopsis**. *BMC Systems Biology* 2010, **4**(62).
18. Lê Cao KA, González I, S D: **integrOmics: an R package to unravel relationships between two omics data sets**. *Bioinformatics* 2009, **25**(21):2855–2856.
19. González I, Déjean S, Martin P, Baccini A: **CCA: An R Package to Extend Canonical Correlation Analysis**. *Journal of Statistical Software* 2008, **23**(12).
20. Martin P, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi JM, San Cristobal M, Legrand P, Besse P, Pineau T: **Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study**. *Hepatology* 2007, **54**:767–777.
21. Bushel P, Wolfinger RD, Gibson G: **Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes**. *BMC Systems Biology* 2007, **1**.

22. Gidskehaug: **A framework for significance analysis of gene expression data using dimension reduction methods.** *BMC Bioinformatics* 2007, **8**(346).
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Research* 2003, **13**:2498–2504.
24. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, et al.: **Gene Ontology: tool for the unification of biology.** *Nature genetics* 2000, **25**:25–29.
25. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22**:1600–1607.
26. Alexa A, Rahnenfuhrer J: *topGO: Enrichment analysis for Gene Ontology* 2010. [R package version 2.2.0].
27. Morine M, McMonagle J, Toomey S, Reynolds C, Moloney A, Gormley I, Gaora P, Roche H: **Bi-directional gene set enrichment and canonical correlation analysis identify key diet-sensitive pathways and biomarkers of metabolic syndrome.** *BMC Bioinformatics* 2010, **11**(499).
28. Guillou H, Zadavec D, Martin PGP, Jacobsson A: **The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice.** *Progress in Lipid Research* 2010, **49**:186–199.
29. Holgate S, Holloway J, Wilson S, Bucchieri F, Puddicombe S, Davies D: **Epithelial-mesenchymal communication in the pathogenesis of chronic asthma.** *Proceedings of the American Thoracic Society* 2004, **1**(2):93.
30. Holownia A, Mroz R, Wielgat P, Skiepkio A, Sitko E, Jakubow P, Kolodziejczyk A, Braszko J: **Propofol protects rat astroglial cells against tert-butyl hydroperoxide-induced cytotoxicity; the effect on histone and cAMP-response-element-binding protein (CREB) signalling.** *Journal of physiology and pharmacology* 2009, **60**(4):63–69.
31. Lee B, Cao R, Choi Y, Cho H, Rhee A, Hah C, Hoyt K, Obrietan K: **The CREB/CRE transcriptional pathway: protection against oxidative stress-mediated neuronal cell death.** *Journal of neurochemistry* 2009, **108**(5):1251–1265.
32. Arany I, Herbert J, Herbert Z, Safirstein R: **Restoration of CREB function ameliorates cisplatin cytotoxicity in renal tubular cells.** *American Journal of Physiology- Renal Physiology* 2008, **294**(3):F577.
33. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: **Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks.** *Proceedings of the National Academy of Sciences of the USA* 2000, **97**:12182–12186.
34. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN: **A Gene Expression Database for the Molecular Pharmacology of Cancer.** *Nature Genetics* 2000, **24**:236–244.
35. Moriyama M, Hoshida Y, Otsuka M, Nishimura S, Kato N, Goto T, Taniguchi H, Shiratori Y, Seki N, Omata M: **Relevance Network between Chemosensitivity and Transcriptome in Human Hepatoma Cells.** *Molecular Cancer Therapeutics* 2003, **2**:199–205.
36. Butte AJ, Kohane IS: **Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurement.** *Pacific Symposium on Biocomputing* 2000, **5**:415–426.
37. Hotelling H: **Relations between two sets of variates.** *Biometrika* 1936, **28**:321–377.
38. Wold H: **Estimation of principal components and related models by iterative least squares.** In *Multivariate Analysis, Volume 2*, 2nd edition. Edited by Krishnaiah P, New York: Wiley 1966:391–420.
39. Burnham AJ, Viveros R, MacGregor JF: **Frameworks for latent variable multivariate regression.** *Journal of Chemometrics* 1996, **10**:31–45.
40. de Jong: **Simpls: An alternative approach to partial least squares regression.** *Chemometrics and Intelligent Laboratory Systems* 1993, **18**:251–263.
41. Lorber A, Wangen L, Kowalski B: **A theoretical foundation for the PLS algorithm.** *Chemometrics* 1987, **1**:19–31.

42. Wegelin J: **A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case.** *Technical Report 371, Department of Statistics, University of Washington, Seattle.* 2000.
43. Tenenhaus M, Gauchi JP, Ménardo C: **Régression PLS et applications.** *Revue de Statistique Appliquée* 1995, **43**:7–63.

Additional Files

Additional File 1 Covariance matrices for the simulated data.

Supplemental_materials_1.pdf is a pdf file to be viewed with Adobe Acrobat.

Additional File 2 R script to generate the Relevance Networks for the Nutrilogue and Liver Toxicity data.

Supplemental_materials_2.pdf is a pdf file to be viewed with Adobe Acrobat.

Additional File 3 Levels of the clinical chemistry measurements for each group of samples from the hierarchical clustering.

Supplemental_materials_3.pdf is a pdf file to be viewed with Adobe Acrobat.

Additional file 4 Summary of the 14 networks identified with GeneGo from Liver Toxicity.

Supplemental_materials_4.xls is a xls file to be viewed with Microsoft Excel or Open Office Calc.