



You have downloaded a document from
RE-BUŚ
repository of the University of Silesia in Katowice

Title: DNA microarray integromics analysis platform

Author: Tomasz Waller, Tomasz Gubała, Krzysztof Sarapata, Monika Piwowar, Wiktor Jurkowski

Citation style: Waller Tomasz, Gubała Tomasz, Sarapata Krzysztof, Piwowar Monika, Jurkowski Wiktor. (2015). DNA microarray integromics analysis platform. "BioData Mining" (vol. 8, iss. 8 (2015), s. 1-23), doi 10.1186/s13040-015-0052-6.



Uznanie autorstwa - Licencja ta pozwala na kopiowanie, zmienianie, rozprowadzanie, przedstawianie i wykonywanie utworu jedynie pod warunkiem oznaczenia autorstwa.



UNIwersYTET ŚLĄSKI
W KATOWICACH



Biblioteka
Uniwersytetu Śląskiego



Ministerstwo Nauki
i Szkolnictwa Wyższego



SOFTWARE ARTICLE

Open Access



DNA microarray integromics analysis platform

Tomasz Waller^{1,2*}, Tomasz Gubała², Krzysztof Sarapata³, Monika Piwowar⁴ and Wiktor Jurkowski⁵

* Correspondence: t.waller@biobyte.pl

¹Institute of Computer Science, Division of Biomedical Computer Systems, University of Silesia, Katowice, Poland

²Academic Computer Centre CYFRONET, AGH University of Science and Technology, Kraków, Poland

Full list of author information is available at the end of the article

Abstract

Background: The study of interactions between molecules belonging to different biochemical families (such as lipids and nucleic acids) requires specialized data analysis methods. This article describes the DNA Microarray Integromics Analysis Platform, a unique web application that focuses on computational integration and analysis of “multi-omics” data. Our tool supports a range of complex analyses, including – among others – low- and high-level analyses of DNA microarray data, integrated analysis of transcriptomics and lipidomics data and the ability to infer miRNA-mRNA interactions.

Results: We demonstrate the characteristics and benefits of the DNA Microarray Integromics Analysis Platform using two different test cases. The first test case involves the analysis of the nutrimum dataset, which contains measurements of the expression of genes involved in nutritional problems and the concentrations of hepatic fatty acids. The second test case involves the analysis of miRNA-mRNA interactions in polysaccharide-stimulated human dermal fibroblasts infected with porcine endogenous retroviruses.

Conclusions: The DNA Microarray Integromics Analysis Platform is a web-based graphical user interface for “multi-omics” data management and analysis. Its intuitive nature and wide range of available workflows make it an effective tool for molecular biology research. The platform is hosted at <https://lifescience.plgrid.pl/>.

Background

Transcriptomics, lipidomics and other molecular techniques produce enormous volumes of data that must be stored, analysed and interpreted using various methods. The integration of “multi-omics” data displays the potential to fully expose the molecular mechanisms occurring in an organism. A holistic approach to biomedical research may help identify new biomarkers for disease diagnostics and improve the sensitivity and specificity of the existing biomarkers [1]. The inclusion of temporal and spatial parameters enables mathematical modelling (often referred to as “systems biology”), which may produce new insights into the mechanisms of pathogenesis and support the development of novel therapies [1].

Comprehensively cataloguing the interactions between genes, lipids and other biological molecules is a highly complex task. Although the human genome contains just over 25,000 genes, the human metabolome – given the multitude of post-translational modifications – is composed of millions of different molecules. Genomes, transcriptomes, and proteomes are fundamental to the functional integrity of the organism. These metabolites reflect the important functions of gene and protein regulation; thus, genomics, transcriptomics and proteomics may provide vital information

regarding the biological status of the system. Modelling and studying the interactions between molecules belonging to different biochemical families (proteins, nucleic acids, lipids, and carbohydrates) requires large-scale computing power and specialized data analysis methods.

DNA microarrays represent a high-throughput measurement technology that is widely used in biological research, especially gene expression experiments. This technology revolutionized biological research by enabling the discovery of a large set of genes whose expression levels reflect a given cell type, treatment, disease or developmental stage [2]. In the first quarter of 2014 alone more than 1600 experiments based on DNA microarrays were uploaded to the functional genomics experiment database operated by the European Bioinformatics Institute (termed ArrayExpress [3]). Integrating DNA microarray data with datasets from external sources may improve the identification of significant biomarkers [4].

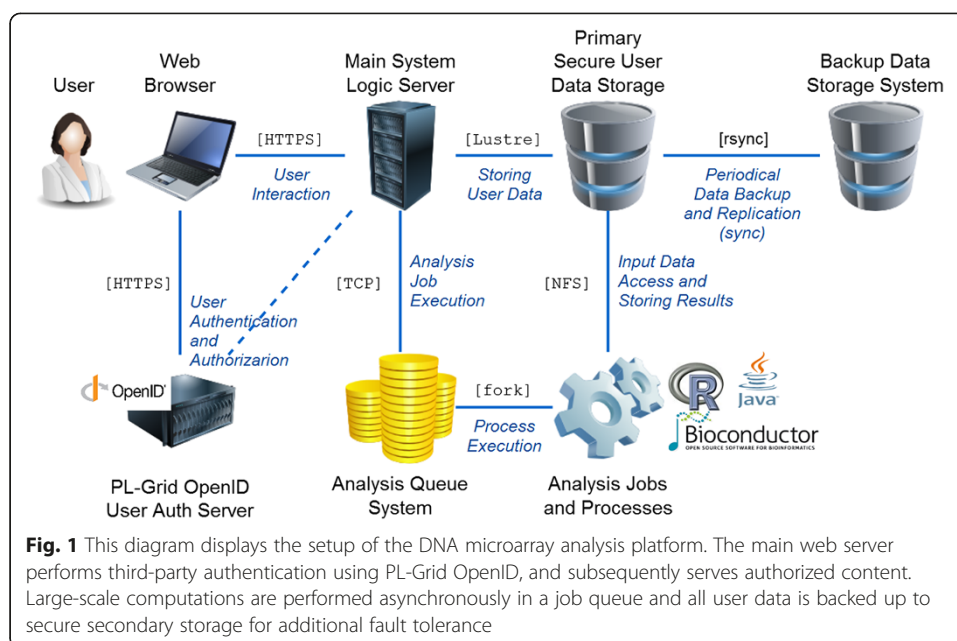
The analysis of environmental and hereditary factors begins with a snapshot of the transcriptome using a set of probes. The lipidome and the microRNAome can guide the exploration of gene-level mechanisms. Integrating the transcriptomics and lipidomics data with microRNA (miRNA)-mRNA interaction data may reveal new information about the underlying cellular processes that cannot be directly derived from any of the individual datasets.

At the heart of mRNA-miRNA interaction analysis is the correct identification of miRNA-corresponding targets. This identification is facilitated by various computational algorithms and laboratory experiments. However, the gene regulation processes involving miRNA are poorly understood, resulting in low specificity and poor accuracy of the targets identified using the available prediction methods [5]. To improve target site recognition we should exploit additional information: expression level measurements of both miRNA and mRNA, evaluation of targets obtained using other prediction methods, sequence-based information, contextual information, phylogenetics and experimentally validated databases. To integrate these sources of information scientists use regression [6, 7] and correlation methods [8–10].

Although several web-based DNA microarray analysis platforms have already been developed [11–13], most do not support integration of “multi-omics” data. In contrast, the DNA Microarray Integromics Analysis Platform permits integrated analysis of transcriptomics and lipidomics data, along with analysis of miRNA-mRNA interactions.

Implementation

The platform (see Fig. 1) is provided to users in the form of a web application. To ensure a powerful, stable and sustainable foundation for the presented system, the PL-Grid Infrastructure [14] was selected as the underlying hardware layer. Any user data uploaded to the platform is stored in a secure disk array based on Lustre technology, which is important from the point of view of efficient computation (i.e., rapid parsing and analysis of microarray data). All user data is periodically backed up using an efficient synchronization protocol (rsync) to a redundant storage system. This additional system is located on a physically separate rack server at an offsite location, providing further fault tolerance and user data safety.



The most important function of the system is to provide a rich set of available algorithms for gene expression analysis. To sustain efficient user interfaces in the web application portion (the *frontend*), the platform offloads all heavy-duty analytical tasks to background processes. All user analysis tasks are managed by a standalone queuing system that the main platform uses to schedule new analysis jobs, monitor the execution status of existing jobs and notify the user when a job is complete. Such an asynchronous mode of operation allows researchers to freely use the system while waiting for scheduled tasks to complete. The queue functions in a FIFO mode for all user jobs such that no user is privileged and all users share common computational resources according to a *fair share* policy. Currently, our platform provides a set of analyses based on R Bioconductor packages [15] and custom Java solutions. The web application portion of the system was developed using the Ruby on Rails framework. The platform is an open source project [16].

The DNA Microarray Integromics Analysis Platform supports a wide range of analyses, including primary raw data processing and detection of differential expression, as well as more advanced techniques, including data mining procedures explaining gene expression-phenotype relationships (clustering, multidimensional scaling, and construction of predictive models). The prime focus of the Analysis Platform is the integration of various ‘omics’ data – accordingly, two specific features (integration of lipidomics and transcriptomics data; integration of miRNA and mRNA data) are described in more detail in the following sections.

Low-level analysis of DNA microarray data

Low-level analysis is the first step in the analysis of DNA microarray data. The DNA Microarray Integromics Analysis Platform supports different types of Agilent (e.g., SurePrint G3 4 × 44 K or SurePrint G3 8 × 60 K) and Affymetrix (e.g., Mouse Gene 1.0 ST; Human Gene 1.0 ST; or U133A 2.0) one-channel DNA microarray data. The

platform is able to perform background adjustment, normalization and summarization of raw DNA microarray data based on widely accepted methods and algorithms. There are three options for background adjustment (no background correction, RMA [17], or GCRMA [18]) and three options for data normalization (quantile [19], scaling and variance stabilization [46]). Once the preprocessing step is complete, any differences between the microarrays are confirmed to be due to differential expression rather than printing, hybridization, or scanning artefacts.

Differential analysis of the gene expression and metabolite levels

To calculate differences in the transcript levels, the platform supplies two methods: Significance Analysis of Microarray (SAM) and *T*-test statistical analysis. Significant gene selection via SAM [21] calculates the relative differences in the transcript levels based on statistical analysis of data permutations.

Both methods can be applied separately to miRNA and mRNA expression data. The latter is also used to determine the differential levels of small-molecule concentrations as measured by mass spectrometry experiments.

In addition to the aforementioned approaches which directly depend on comparisons of individual transcript levels, the platform also enables the use of functional priors in the form of Gene Set Enrichment Analysis (GSEA), which is commonly used to test for statistically significant overrepresentation of particular groups of genes in the context of the Gene Ontology Annotation Database [24].

Data clustering

Data structure can be investigated by performing cluster analysis which categorizes unsupervised data into classes of relative similarity using various samples or experiments (represented as a dendrogram) [20]. The platform supports clustering based on samples or transcripts. The user may select different methods of distance measurement and agglomeration. Cluster analysis can be used to identify novel sample clusters and their associated transcripts, and to evaluate data quality by verifying that replicate samples under similar conditions are clustered together.

The platform also includes Principal Component Analysis (PCA) to reduce the dimensionality of the data by extracting the variables (coordinates) displaying the highest variance, which are assumed to account for the variability in the analysed dataset [23].

Phenotype classification

Conditional analysis of gene expression based on artificial neural networks (ANNs) is capable of creating and training a neural model (a perceptron) and subsequently using it to distinguish between healthy and abnormal tissue based on gene expression profile data [22]. This ANN analysis supports the simultaneous training of thousands of network architectures, computing predictions from each network architecture to pinpoint the single best network. Trained ANNs have been successfully applied to the analysis of various diseases, including the classification of cardiovascular disease and the prediction of bacteria-antibiotic interactions and colorectal cancer patient survival [22].

Analysis of biological pathways

Biological pathways can be currently accessed using the Wikipathways [25], KEGG [51] and Reactome [52] web services as an optional step following determination of the differentially expressed genes. The platform scans the gene expression datasets and the pathway repositories to identify the genes that participate in a genetic pathway. The genes in the specified pathway are graphically represented as a color-coded diagram. The discovered pathways are ranked by the percentage of genes affected.

Analysis of transcriptomics and lipidomics data

Although multivariate statistical techniques are the method of choice for studying complex data relationships, classical methods such as canonical correlation analysis (CCA) or partial least squares regression (PLS) [26] are not practical due to the specific structure of the data obtained from high-throughput platforms. For example, gene expression data typically include tens of thousands of variables measured from a small number of observations (samples taken under certain conditions). To cope with this suboptimal data structure comprising many variables and few observations, the standard methods can be modified to restructure data either by performing feature selection or introducing artificial variables. Sparse PLS (sPLS) reduces the number of variables incorporated into the model [27, 28] and can be applied to both regression and canonical correlation frameworks. Regularized canonical correlation analysis (rCCA) increases the number of pseudo-observations in the regularization procedure of covariance matrices [29, 30]. Finally, principal component regression (PCR) [31] represents the class of multidimensional scaling techniques. Despite difficulties in estimating statistical significance, appropriate application of these methods, including correctly assuming the correlations between variables, provides an approximate interpretation of heterogeneous data (e.g., metabolites and genes), which can be helpful for biological data analysis and may yield clinically relevant results [32–34].

All statistical techniques included in the presented platform are implemented in R and include classical CCA and rCCA (YACCA [35] and FRCCA [36]) and PLS [37], as well as specialized packages dedicated to lipidomics data (mixOmics [38]).

Integration of data obtained from two independent platforms is based on three statistical methods: classical canonical analysis (CCA; YACCA package in R), regularized canonical analysis (rCCA; FRCCA package in R) and PLS (see Fig. 2).

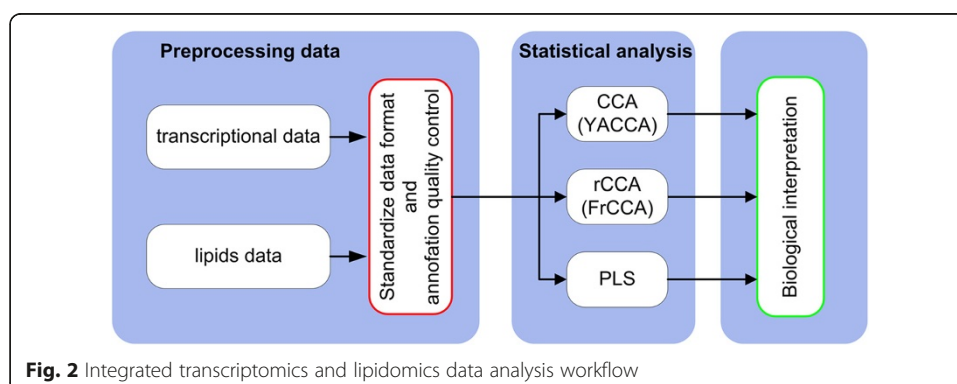


Fig. 2 Integrated transcriptomics and lipidomics data analysis workflow

Analysis of miRNA-mRNA target site recognition

The interactions between miRNAs and mRNAs play an important role in the process of gene regulation. RNA interference (RNAi) depends on appropriately recognizing specific targets of the miRNA molecules that are integrated into an RNA-induced silencing complex (RISC) [39]. RISC uses miRNA as a template for recognizing complementary transcripts. However, in metazoans this complementarity is imperfect. Generally, a miRNA-mRNA duplex occurs in the 3'UTR region, and complementarity applies only to the first 7–8 nucleotides of the miRNA, also referred to as the seed. Such short complementarity is a fundamental problem for correct recognition of the target site, as its specificity is very low. As a consequence, it is possible for any miRNA-mRNA pair to interfere. Additionally, mismatches may occur in the seed region of the miRNA during the hybridization process.

We use TargetScan [40] to predict miRNA targets. The method considers the 3'UTR regions of the transcriptome and calculates context score [41] from the type of seed match (8mer, 7mer-m8, and 7mer-1A), 3' pairing contribution (miRNA-target complementarity outside the seed region); preferentially conserved targeting (PCT) estimated as described in [42] and [43]; and multiple additional parameters.

Effective expression differentiation and miRNA/mRNA interaction detection require large sets of samples representing multiple cell types or conditions. As sufficiently large sample sets are unfortunately not always available. The platform supports the inference of miRNA-mRNA interactions using two complementary Bioconductor methods, Roleswitch [44] and TargetScore [45], which account for these limitations.

Roleswitch is able to infer probabilities of miRNA-mRNA interactions from a single sample. The Probabilities of MiRNA-mRNA Interaction Signature (ProMISE) are calculated from two paired expression vectors, M miRNA and N mRNA, and the $N \times M$ seed match matrix containing the number of target sites for each miRNA-mRNA pair. Notably, Roleswitch assumes that the total transcribed mRNA levels are higher than the observed (equilibrium) mRNAs levels. An entire family of versatile preliminary seed matrices (in which the source data is obtained from the TargetScan database) has been prepared and is provided to platform users. The pre-calculated matrices differ with respect to the following:

- conservation of miRNA: broadly conserved = conserved across most vertebrates, typically up to zebrafish; conserved = conserved across most mammals, but typically not beyond placental mammals; or poorly conserved = all others; and
- the threshold of the total context site, defined as the sum of context scores for the most favourable (most negative) miRNA in this family. For highly conserved miRNA families the total context site is complementary to preferential conservation of the sites (aggregate PCT).

The user may select targets of interest, including the expected level of conservation.

The TargetScore approach assumes input data from the experimental transfection procedure. Transfection relies on the introduction of miRNA molecules to cells containing the reagent. By determining the expression levels of transcripts prior to and following transfection one can infer the targets of the given miRNA molecule. In this model, paired microarrays of mRNA and miRNA are generated from the same

experiment to measure the cellular response to an environmental change (e.g., treatment with endotoxin lipopolysaccharides (LPS)). When multiple miRNA samples are examined a list of miRNA molecules is used as the input, as in the case of a transfection experiment, into a corresponding mRNA expression matrix in the $\log(\text{FC})$ form. Then, the miRNA-mRNA interactions are analysed using the TargetScore package and the appropriate sequence-based information (including the context score and PCT) for each selected miRNA sample.

TargetScore infers the posterior distributions of miRNA targets by probabilistically modelling the likelihood of a fold-increase in mRNA expression and sequence-based scores. The variational Bayesian Gaussian mixture model (VB-GMM) applies these logarithmic fold-change and sequence scores to determine the posterior latent variables (i.e., the miRNA targets). The sequence-based information, context score and PCT are obtained from the previously mentioned TargetScan database. The author of TargetScore have demonstrated that target prediction can be improved by simultaneously using expression fold-change and sequence-based information. The advantages of TargetScore include improved quality of target recognition via the use of sequence-based information and its deployment as an unsupervised method operating on the entire gene dataset to more closely model the overall interaction probability.

Using the presented DNA microarray analysis it is possible to investigate miRNA-mRNA interactions based on the following: 1) a single sample; 2) differential expression; 3) differential expression and sequence-based scores; or 4) a combination of the above approaches.

Given a set of paired raw miRNA and mRNA microarray expression data from each sample, the investigation workflow follows a typical normalization and expression-calling procedure, as well as steps specific to integrated miRNA and mRNA analysis.

Predictions are summarized by listing the genes/transcripts ranked by the mRNA samples displaying the maximum probabilities of interaction with any miRNAs, the miRNA samples displaying the maximum probabilities of interaction with any transcripts and the maximum probabilities of interaction between specific miRNA-mRNA pairs.

The user is able to operate not only at the gene level but also at the transcript level. Most R/Bioconductor microarray workflows function only at the gene level, by associating a single gene with a set of probes and calculating the average expression for multiple transcripts of the same gene. However, it is occasionally more effective to focus on the transcripts, such as when investigating gene isoforms or alternative splicing. When the user selects transcript-level analysis, the platform links the related transcripts to the same gene in the analysis results. In the inverse case, in which a single transcript is related to multiple genes, the system presents the user with a list of such transcript-gene pairs.

Data management

The DNA Microarray Integromics Analysis Platform supports simple and efficient data management in a variety of modalities.

No researcher works alone. To deliver experiment-sharing capabilities (an “experiment” is defined as a set of connected raw data variables, analyses and their results), the platform implements a team-based sharing mechanism. A user is able to share an experiment with any team he/she belongs to (see Fig. 3).

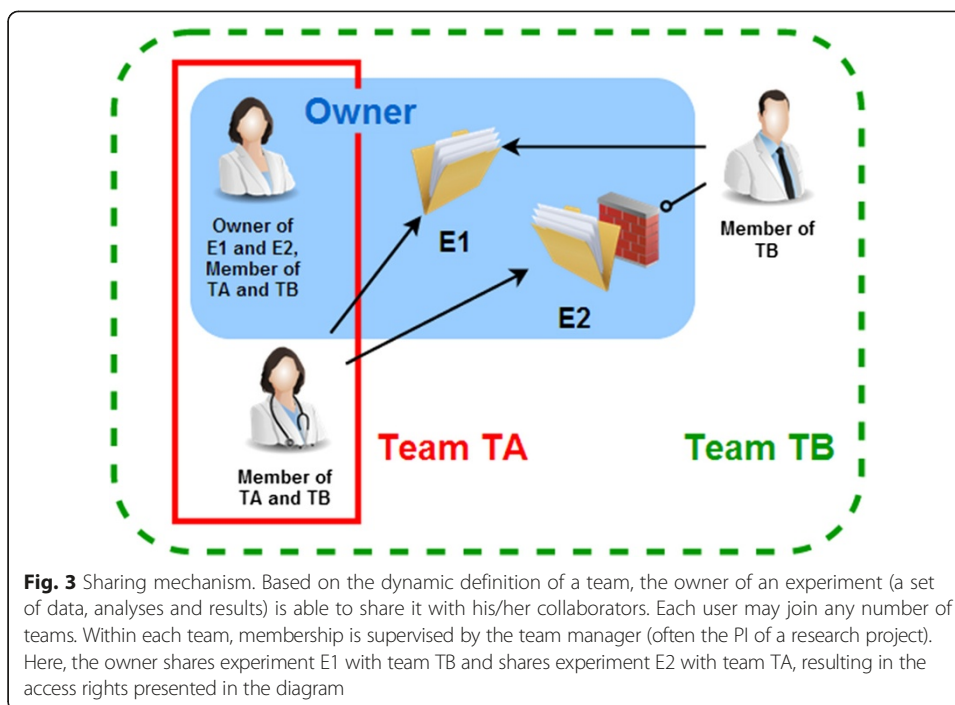


Fig. 3 Sharing mechanism. Based on the dynamic definition of a team, the owner of an experiment (a set of data, analyses and results) is able to share it with his/her collaborators. Each user may join any number of teams. Within each team, membership is supervised by the team manager (often the PI of a research project). Here, the owner shares experiment E1 with team TB and shares experiment E2 with team TA, resulting in the access rights presented in the diagram

To accommodate both users who are in the middle of their experimentation and data analysis process and those who have already published their results, the platform provides two important mechanisms, both of which are based on the Minimum Information About a Microarray Experiment (MIAME) standard [47]. First, the system enables its users to import preprocessed datasets by uploading a matrix of normalized transcript expression data containing rows representing individual transcripts and columns representing samples (individual microarrays). Using an SDRF file prepared according to the MIAME regulations, the user uploads both the preprocessed data and the experimental metadata to the system. Once the experiment has been appropriately registered, it is possible to conduct high-level analyses using the imported data.

When a user has already published his/her findings in a journal and uploaded the data to the EBI ArrayExpress [3] database, the platform facilitates importing the entire experiment directly from this database. Experimental data is downloaded in the background upon request – the users are not required to handle any files themselves. If the experimental record in the ArrayExpress database contains all of the data necessary for the system to process it and the microarray files are of a type recognized by the platform, the experiment is downloaded, and the platform notifies the user when the download process is complete. This feature is also useful for educational purposes, so that a teacher and students do not need to supply their own data (which could be cumbersome in the case of large microarray experiments), but rather can simply select an interesting publicly accessible ArrayExpress experiment and rely on the platform to download all of the necessary data for them. The system uses a high-throughput backbone network that is a component of the PL-Grid Infrastructure, which ensures rapid transfer between European laboratories and the disk array of the system.

The platform also helps export microarray experiments to the ArrayExpress database by providing a facility to easily generate the required MAGE-TAB files (specifically,

Investigation Description Format (IDF) files and Sample and Data Relationship Format (SDRF) files) according to the required MIAME data and formatting regulations. These files are useful when registering one's own microarray experiments in the ArrayExpress database.

Results

Case study 1: functional links between gene products and metabolites

Publicly available lipidomics and transcriptomics data from a murine nutrigenomics study [48] were used to demonstrate the performance of the proposed analytical protocol using two independent data sets. The data consist of the expression levels of 120 genes (X) and the concentrations of 21 fatty acids (Y), both spanning a total of 40 observations of several genotypes and diets.

Following data import, analytical procedures were executed using the Integromics platform (Fig. 4). Depending on the number of observations compared to the number of variables (dependent and independent), appropriate methods of canonical correlation (i.e., classical canonical correlation, such as CCA or rCCA) may be applied.

CCA is appropriate for cases involving a relatively large number of observations and a low number of correlated variables (dependent and independent) in the two datasets. rCCA is a modification of the classical canonical correlation method that permits analysis of a small number of observations and a large number of variables in both sets. PLS is not restricted by the number of observations or highly correlated variables.

Case study using YACCA (canonical analysis) and FRCCA (regulated canonical analysis) packages in R.

The canonical analysis procedure (CCA) generates specific data corresponding to the analysed datasets. This case study focused on the most basic parameters of canonical analysis and their interpretation. Conducting CCA initially requires the removal of correlated predictors (independent variables, or the transcriptomics data (X)) and responses (dependent variables, or the lipidomics data (Y)) as highly correlated

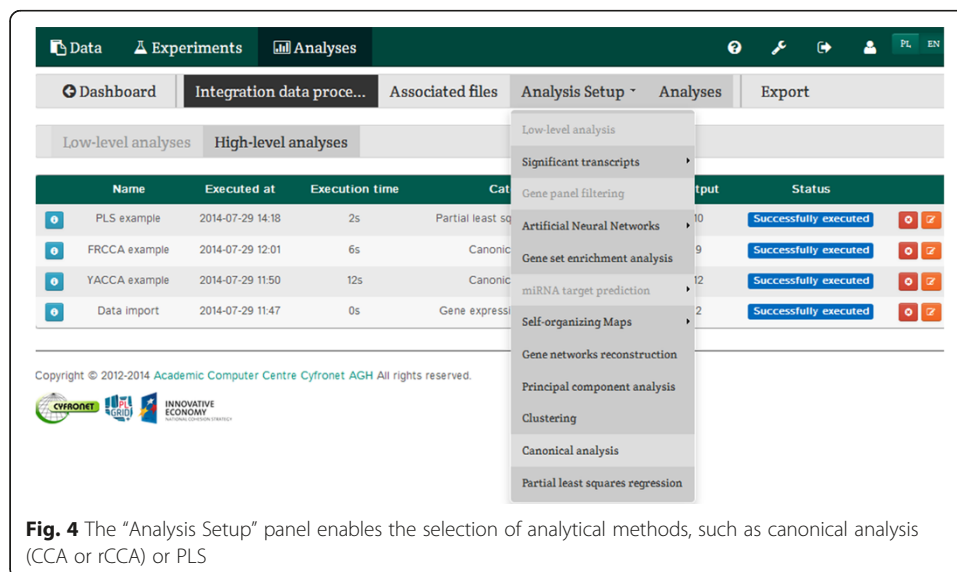


Fig. 4 The "Analysis Setup" panel enables the selection of analytical methods, such as canonical analysis (CCA or rCCA) or PLS

variables may significantly skew the results. The correlation threshold is set arbitrarily and depends on the type of data. In the life sciences domain, the correlation coefficient is considered to be greater than 0.7.

Once highly correlated data are removed from X and Y, the correlation between the dependent and independent variables can be calculated. The Integromics platform produces a correlation map for X and Y, identifying the variables that display strong correlations (Fig. 5).

The appropriately prepared data sets were analysed using the canonical correlation procedure which identifies linear combinations of variables in the analysed sets such that the correlations between them are as high as possible. The results of the CCA procedure are accessible as text (numerical data) or PNG (figures) files several seconds after the analysis is complete.

Detailed information about the CCA results, such as the canonical correlations, the X coefficients, the Y coefficients, the structural correlations (loadings) for variables X and Y, and the aggregate redundancy coefficients (total variance explained) are accessible in the `cca.fit-YACCA.txt` file. Statistics that describe datasets X and Y are saved in the `summaryX-YACCA.txt` and `summaryY-YACCA.txt` files respectively. The significance of the calculated canonical correlation is performed using the F-test (Rao's F approximation), and these results are provided in the `F_test_CCA.txt` file.

The results obtained from the murine nutrigenomics datasets reveal high canonical correlations (CV1:1, CV2:0.997, CV3:0.995). These canonical correlations were found to be statistically significant at a significance level of 0.05. The calculated values of aggregate redundancy were X | Y: 0.42 and Y | X: 0.89, indicating that Y explains X in 42 % of cases and that X explains Y in 89 % of cases.

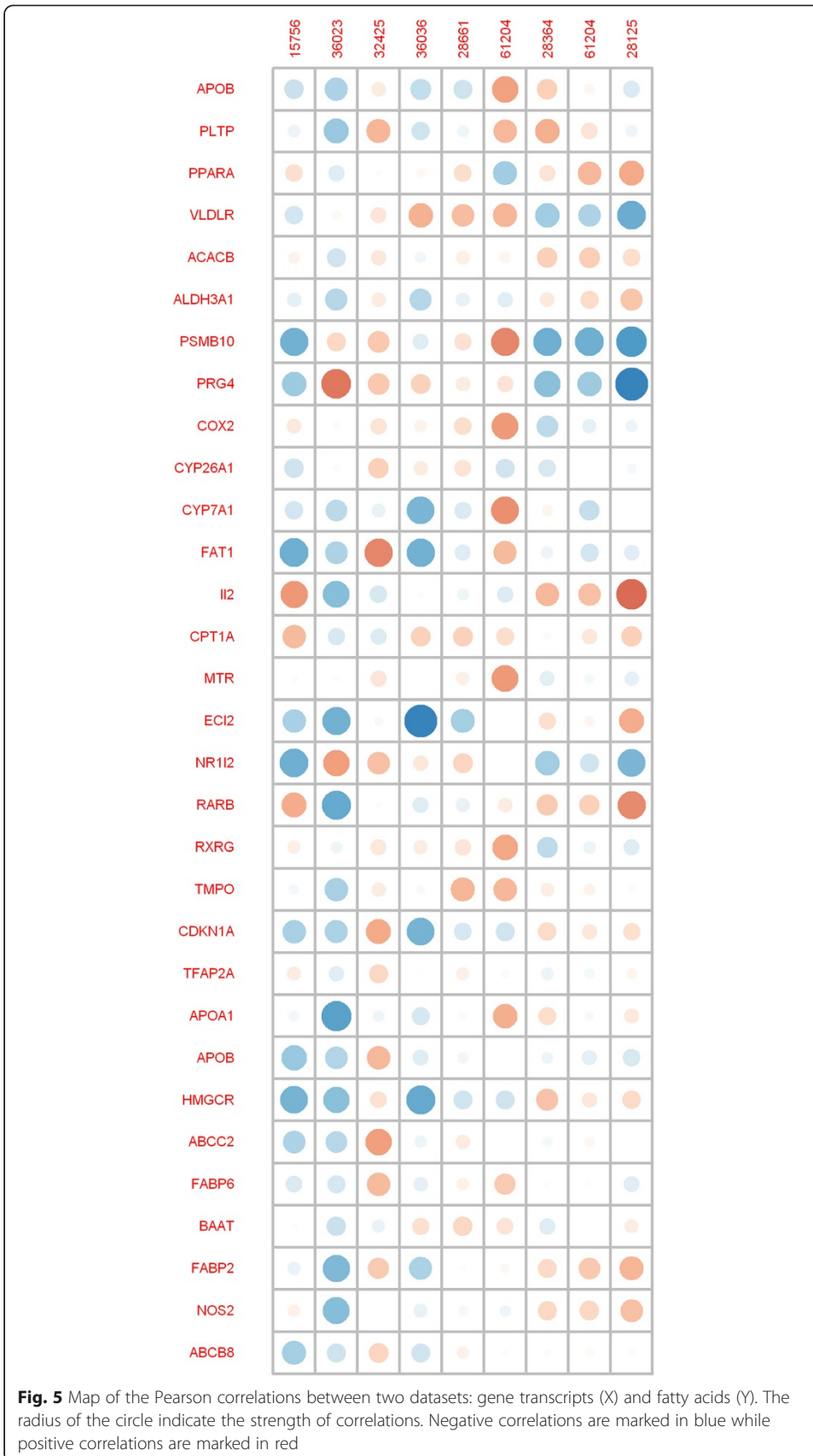
The platform supplies graphical files which present the results of the analyses. For example, a diagram of the impact loadings in both sets is presented in Fig. 6.

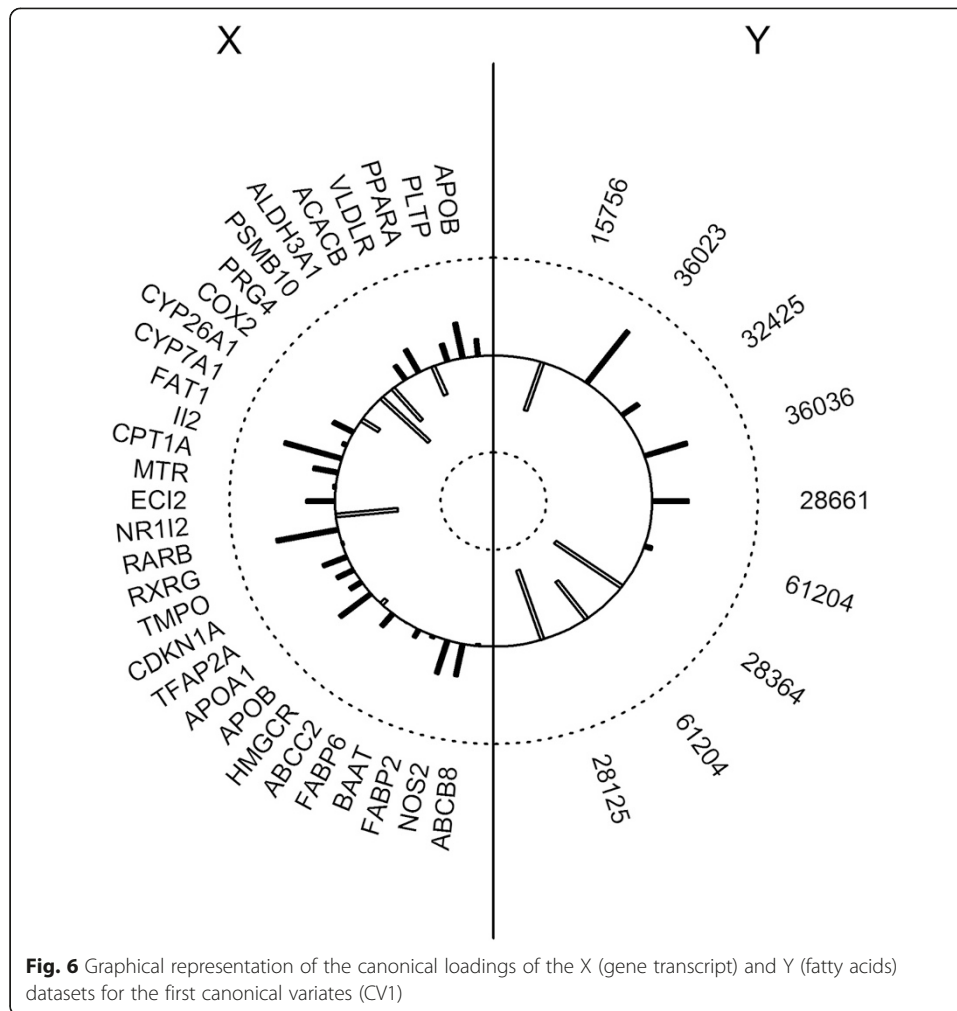
High canonical loading of a particular variable indicates the significance of that variable in the interpretation of the canonical variate. In this case, among the gene transcripts, the greatest impact on this model was obtained for genes PRG4 (-0.637), IL2 (0.56), NR112 (-0.58), and RARB (0.59) and for fatty acids 28364 (-0.78), 28125 (-0.75), and 36023 (0.66). These results are reflected by the heights of the bars in the graph.

As mentioned above, CCA should be performed for a large number of observations. If the number of observations is lower than the number of transcriptomics or lipidomics variables, classical canonical analysis is not appropriate (the number of observations should be 10 times the number of variables). In the life sciences domain the obtained data often displays a structure in which the number of observations (samples) is small while the number of variables (e.g., transcripts of genes) is large. In such situations, a modified classical canonical analysis method (rCCA) should be used.

For the analysed dataset, CCA can be used when the correlation threshold is at least 0.7 for the transcriptomics and lipidomics data. When $r = 0.8$, CCA is not performed due to the insufficient number of observations compared to the number of X and Y variables. In such cases rCCA can be performed instead. The results of the rCCA calculation are available in a similar form to the CCA results.

The canonical correlation and the canonical weights and loadings for X and Y are accessible in the `my_res-FRCCA.txt` file. Descriptive statistics are provided in the



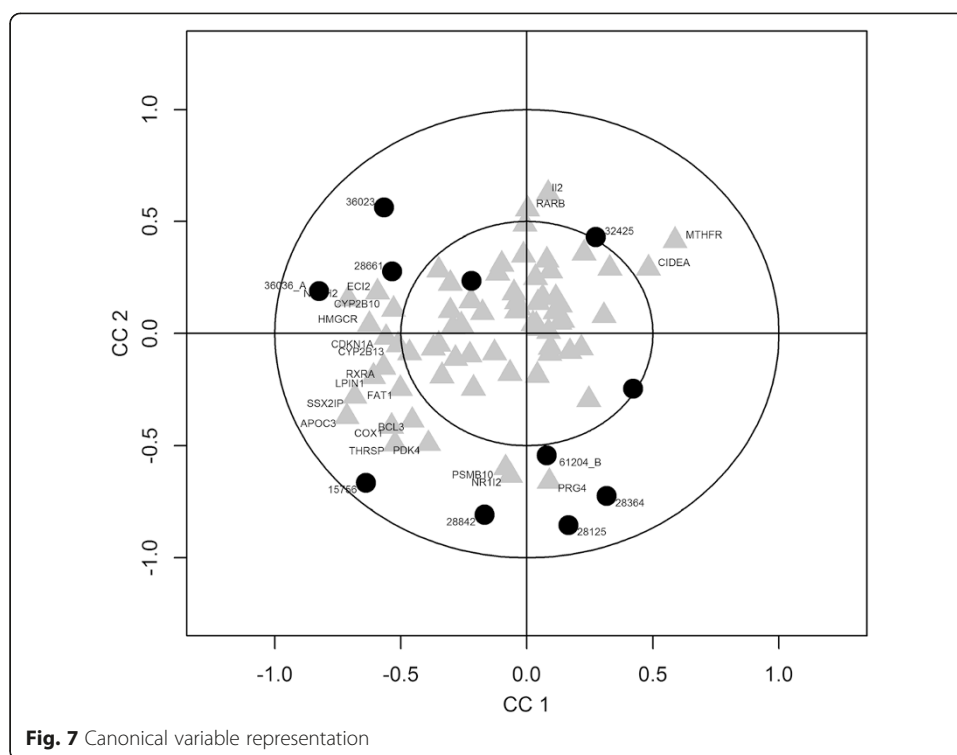


summaryX-FRCCA.txt and summaryY_FRCCA.txt files. In FRCCA the results do not include information regarding the significance calculations. The values of the canonical correlations for the first three sets of canonical variables are as follows: CV1: 0.99; CV2: 0.98; and CV3: 0.98. The graphical representation depicts the variables as points on a two-dimensional plane in which the axes represent the canonical loadings, as shown in Fig. 7.

PLS procedure

For the integration of partially correlated transcriptomics and lipidomics data the PLS procedure was used. The first step of the PLS procedure involves constructing a predictive model. For this purpose the entire dataset was incorporated into the PLS model.

To demonstrate the prediction values, a subset of murine nutrigenomics data were used (50 transcriptomics and 10 lipidomics variables). The PLS model was fit to the 10 latent variables, followed by cross-validation. The root mean squared error of prediction (RMSEP) parameters were calculated for the specific lipidomics variables (fit method: kernelPLS; number of components considered: 10). This step yields information about the quality of the fit of the model to particular components. In the analysed dataset, some



variables were classified as poor (RMSEP too high), such as ChEBI 16196, in which RMSEP was >4.5 for $ncomp = 8$ and even higher for $ncomp > 8$ (Fig. 8). Other variables remained good, such as ChEBI 15756, exhibiting the lowest RMSEP in 5 components.

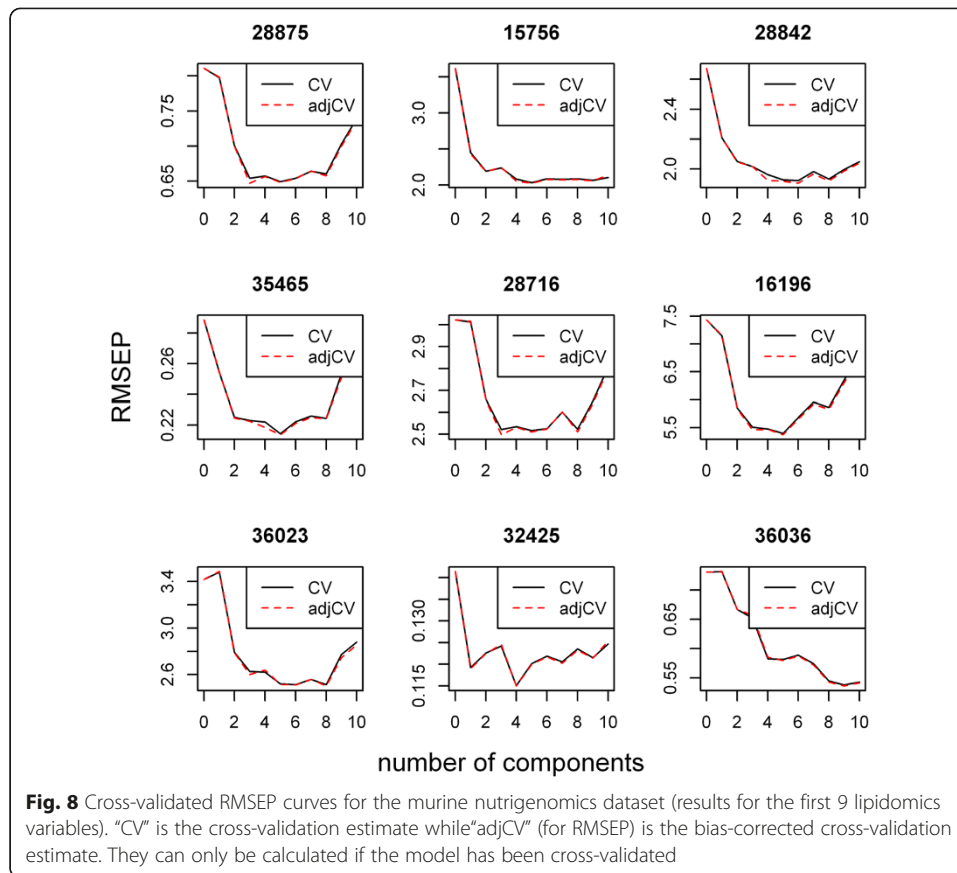
To confirm the quality of the variables included in this model, cross-validated predictions using 5 components (the components for which most of the variables displayed the lowest RMSEP) were visually compared to the measured values (Fig. 9).

The charts of some of the variables exhibited poor alignment to the corresponding target profiles – e.g., fatty acid ChEBI ID 36036 (Fig. 9). This result suggests that the X variables did not appropriately describe the given lipidomics variable. Another graphical representation of the results of the PLS procedure consists of a pairwise plot of score values. Such plots help reveal patterns, groups and outliers. In our example (5 components plotted), there was no clear indication of any grouping or outliers (Fig. 10).

The explained variances for each component in this example were as follows: Comp1 = 32 %, Comp2 = 19.5 %, Comp3 = 5.5 %, Comp4 = 13.2 %, Comp5 = 6.7 %, Comp6 = 4.2 %, Comp7 = 3.6 %, Comp8 = 2.1 %, Comp9 = 1.4 %, and Comp10 = 1.1 %. The canonical loadings are crucial for the interpretation step as higher loading values indicate greater contribution of the given variable to the model. Figure 11 presents the top 10 genes, ranked according to their loading values.

Case study 2: discovery of significant miRNA interactions

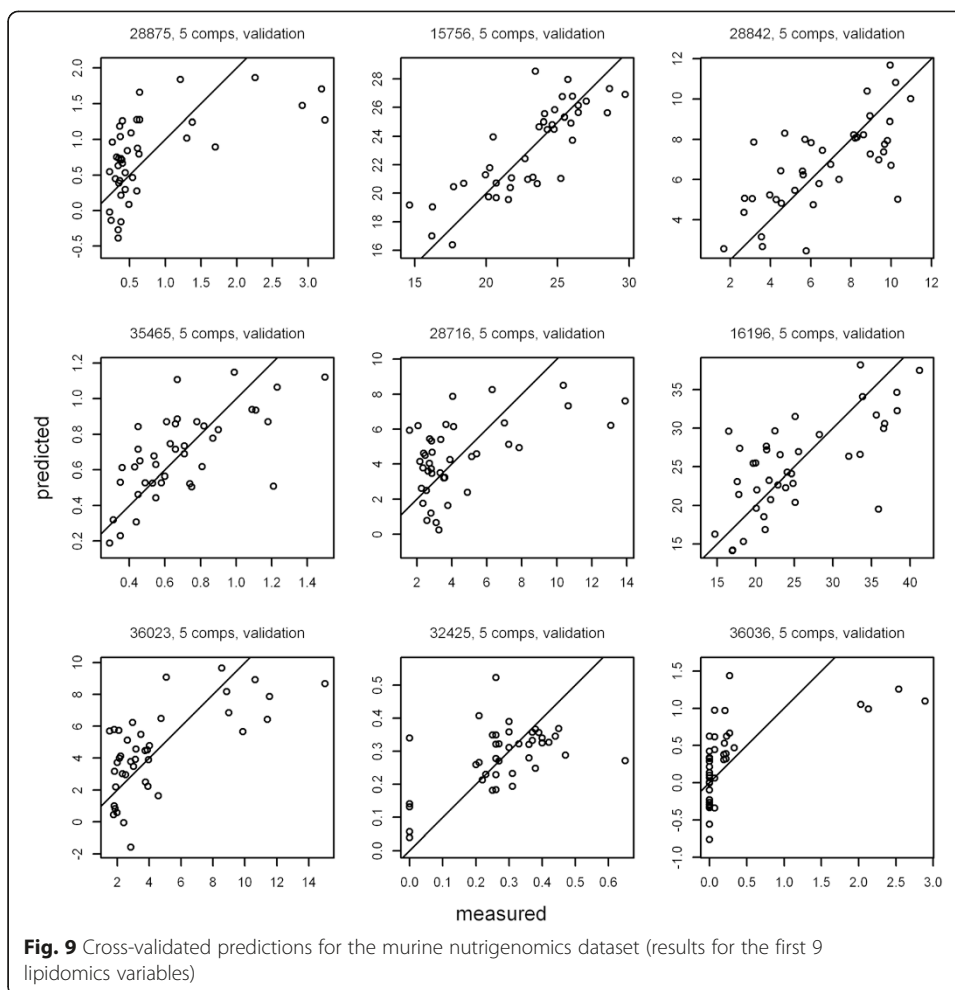
The other use case concerns the Normal Human Dermal Fibroblasts (NHDF) experiment [49]. This study compared the differential expression of the entire transcriptome, including miRNAs, between different cell lines (12 samples): a control group of NHDF cells (NHDF), NHDF cells treated with LPS (NHDF LPS), NHDF cells infected with porcine endogenous retroviruses (PERVs) (NHDF PK15) and NHDF PERV-infected



cells treated with LPS (NHDF LPS PK15). For each sample, the paired expression levels of both mRNA and miRNA were measured using AffymetrixGeneChip miRNA 2.0 and HGU133A2 microarrays. This study measured the changes in the expression levels of both mRNAs and miRNAs between specific groups and, when changes were found, attempted to establish which RNAi process regulated their expression in PERV-infected cells either with or without LPS treatment.

Preliminary miRNA microarray data preparation was performed using the Affymetrix miRNA QCTool [50]. The procedure included background detection, background adjustment (BC-CG Adjust), quantile normalization, addition of small constants (CVStab), and summarization using a median polishing approach. The presented platform can be used to import such pre-normalized data for further analysis. Normalization of the mRNA microarray data were performed using the presented platform, including RMA background correction and quantile normalization.

The next step involved microarray significance analysis to calculate significant differences in the transcript levels. Statistically significant differences in the miRNA levels were found for only two pairs of cell groups, NHDF PK15 (P) versus NHDF (K) giving hsa-let-7e, hsa-miR-199a-5p, hsa-miR-99a and NHDF LPS PK15 (LP) versus NHDF LPS (L) giving hsa-miR-3197. The resulting list of statistically significant miRNA subsets was selected using a Median of False Discovery Rate (FDR) threshold of 0 (the threshold used is relative, which means that genes are significantly different at a level greater than 0.05) for the prediction of putative miRNA targets among the differentially expressed mRNAs.

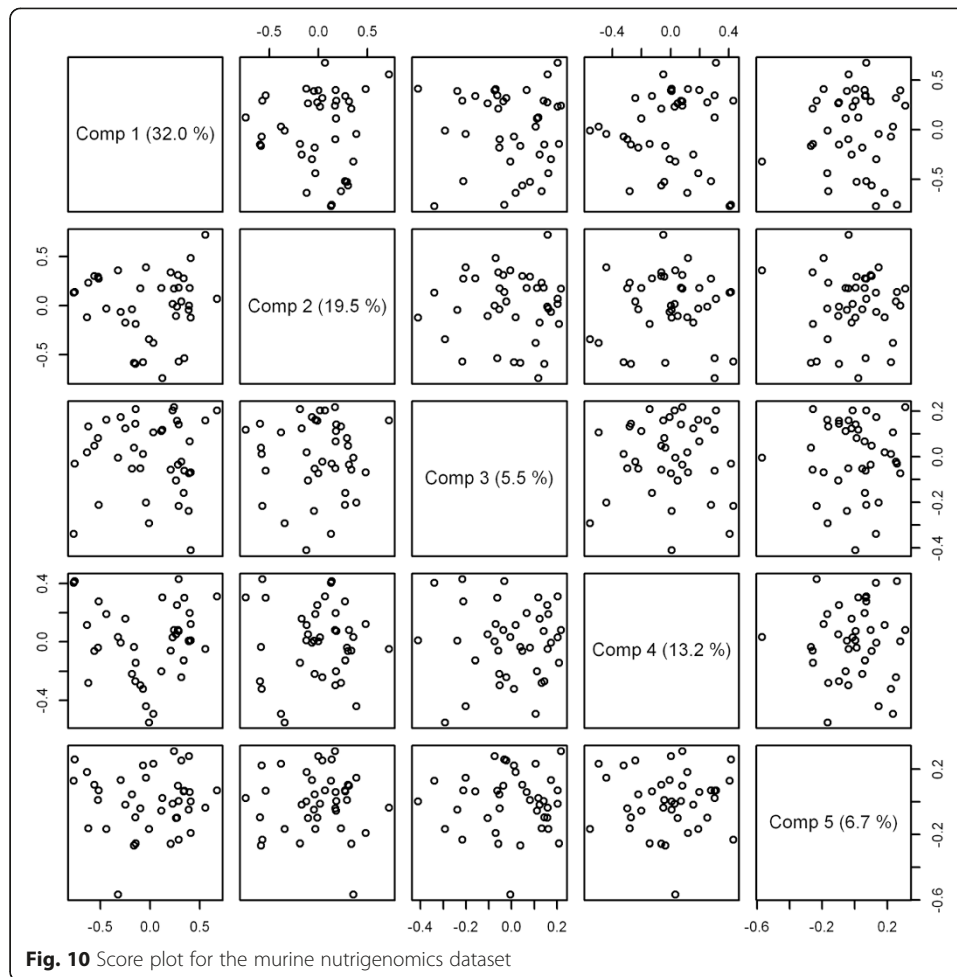


The Integromics platform supports two miRNA target prediction methods. The first is based on a single expression sample (group). Due to the low diversity of the compared groups we selected the most liberal settings for the conservation of miRNA and target sites (Fig. 12). However, the results were found to be insignificant (Fig. 13).

The available parameters of this procedure include:

- the number of cells representing the highest probability. This parameter affects the dimension of the output matrix; e.g., a value of 10 results in an output matrix with diagonal length 10;
- the granularity level of the probe set (either transcript or gene);
- VB-GMM: the threshold that defines termination/convergence of the VB expectation maximization (VB-EM) algorithm, which is used to optimize the parameters;
- the maximum number of EM iterations of VB-GMM.

The ProMISe output tables for NHDF PK15 (P), NHDF LPS (L) and NHDF LPS PK15 (LP) were very similar to the NHDF results (see Fig. 13). The maximum probability (approximately 0.2) was obtained for the same pair (TMSB10 and hsa-miR-1469) from each sample.

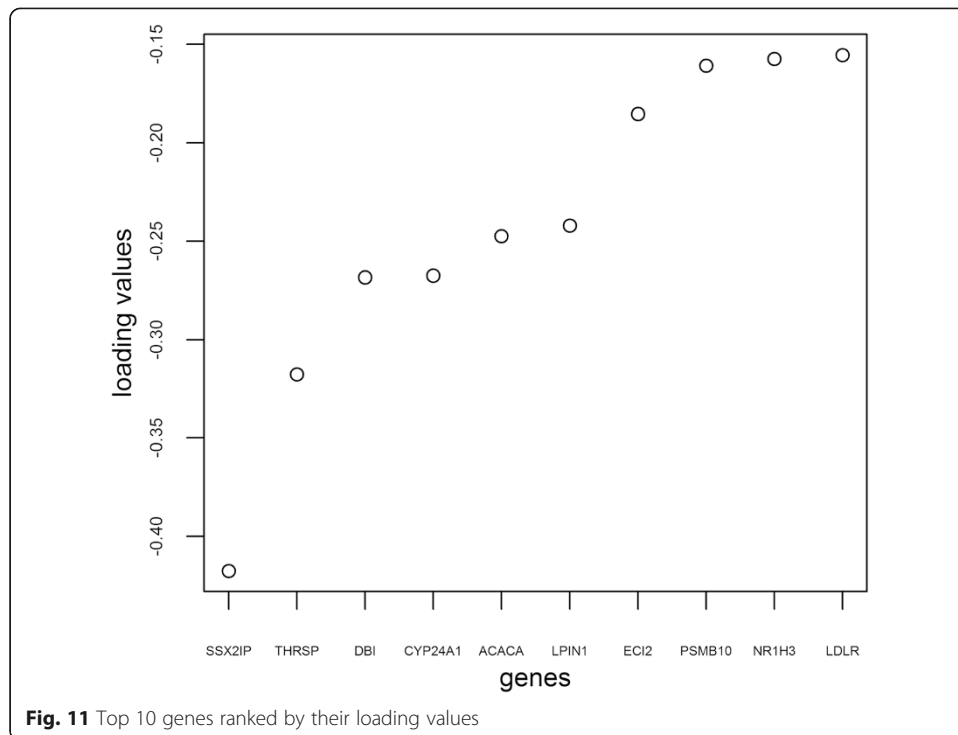


The final component of the workflow used TargetScore for miRNA target prediction based on the variability of the expression of the compared groups. Using the TargetScore, our system prepares sequence-based information, the context score and PCT for each miRNA in the selected list. The exact approach depends on the selected level of analysis at either the gene or transcript level. The system informs the user whether sequence-based information was obtained for a given miRNA sample.

The applied settings and parameters are presented above. We selected the transcript analysis level and the default values for the remaining settings (convergence threshold of the VB-EM algorithm and maximum number of EM iterations). The peak size of the resulting matrix was set at 20 (Fig. 14).

As a result of the above steps, we expected to answer the following questions: 1) which transcript (gene) displays the highest probability of being a target of any miRNAs; 2) which miRNA displays the highest probability of interfering with any transcripts; and 3) which miRNA-transcript (gene) pair displays the highest probability of interacting with each other?

The resulting tables present the peak probability values for the genes/transcripts. The attached histograms help distinguish this limited “peak” dataset from the entire population of results. If the “transcript” option is selected, the table also contains other transcripts associated with a specific gene according to the described multi-transcript protocol.



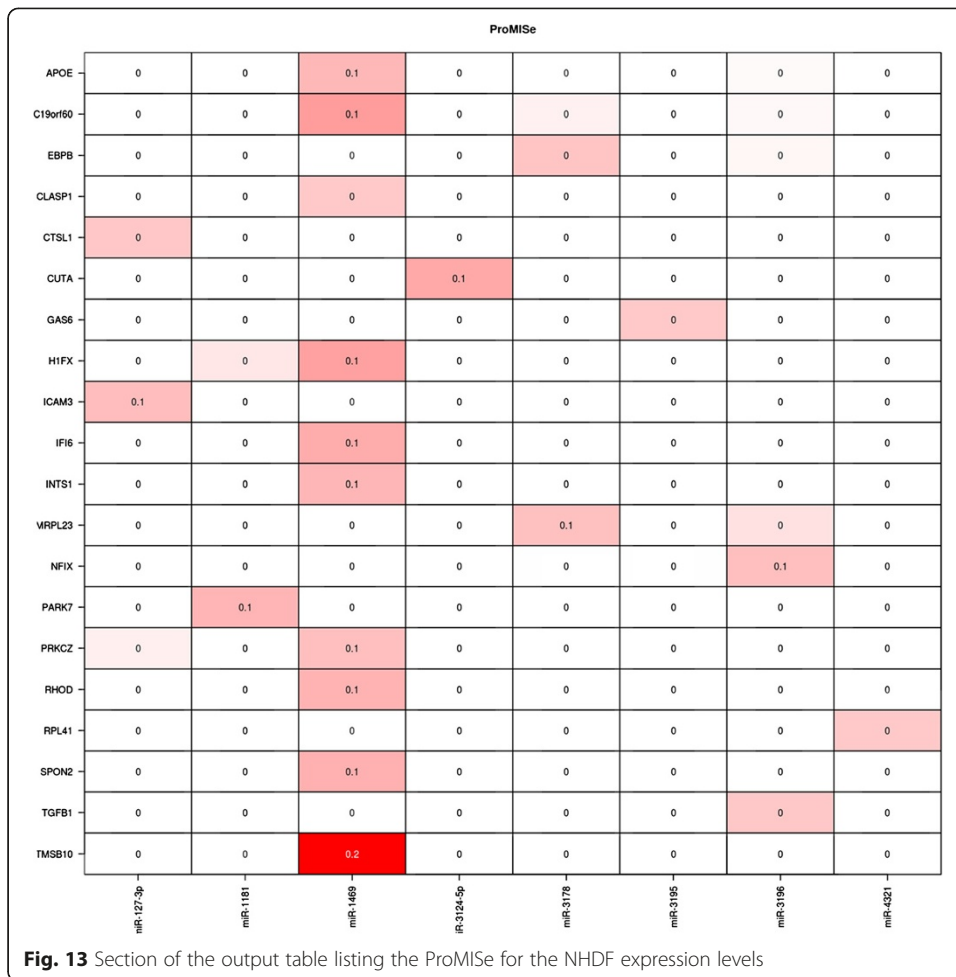
The full list of output files and miRNA target prediction charts based on the variability of expression includes TargetScore probabilities calculated only for log(FC) input:

- list of miRNA samples that contain no sequence-based information (context score or PCT) in the TargetScan database
- list of transcripts that are linked to more than one gene
- vector of the peak-probability genes computed only using log(FC)
- histogram of the TargetScore probabilities for the fold-change in expression and TargetScore probabilities calculated using the sequence-based information:
- table of cell types displaying the greatest probabilities of miRNA-mRNA interaction
- histogram of the TargetScore-computed probabilities of miRNA-mRNA interaction
- table of rows (genes/transcripts) containing the maximum values of the sum of the probabilities (see Fig. 15)

Settings of miRNA target prediction

Prediction name: <input type="text" value="K targets"/>	Target set: <input type="text" value="Conservative and nonconservative"/>
Threshold (total) context score: <input type="text" value="0"/>	Maximum diagonal of the output matrix: <input type="text" value="20"/>
Conservation of miRNA: <input type="text" value="Poorly conserved"/>	
Select samples by attribute values (groups)	
ACTING: <input type="text" value="none"/>	COCULTURE: <input type="text" value="none"/>
<input type="button" value="Predict targets"/>	

Fig. 12 Selection panel for miRNA target prediction based on a single expression sample



- histogram of the sums of the TargetScore probabilities for each row (gene/transcript)
- table of columns (miRNAs) containing the maximum values of the sums of probabilities

The output files support the following interpretation. Transcript NM_5824 (LRRC17 gene) exhibits the highest probability among mRNA/miRNA pairs (this conclusion is

Settings of miRNA target prediction

Prediction name: P vs K

Level of probe association: Transcript

Limitation of result matrices: 20

Convergence threshold of VB-EM algorithm: 1e-20

Maximum number of EM iterations: 2000

Predict targets

Fig. 14 Selection panel for miRNA target prediction based on expression variability

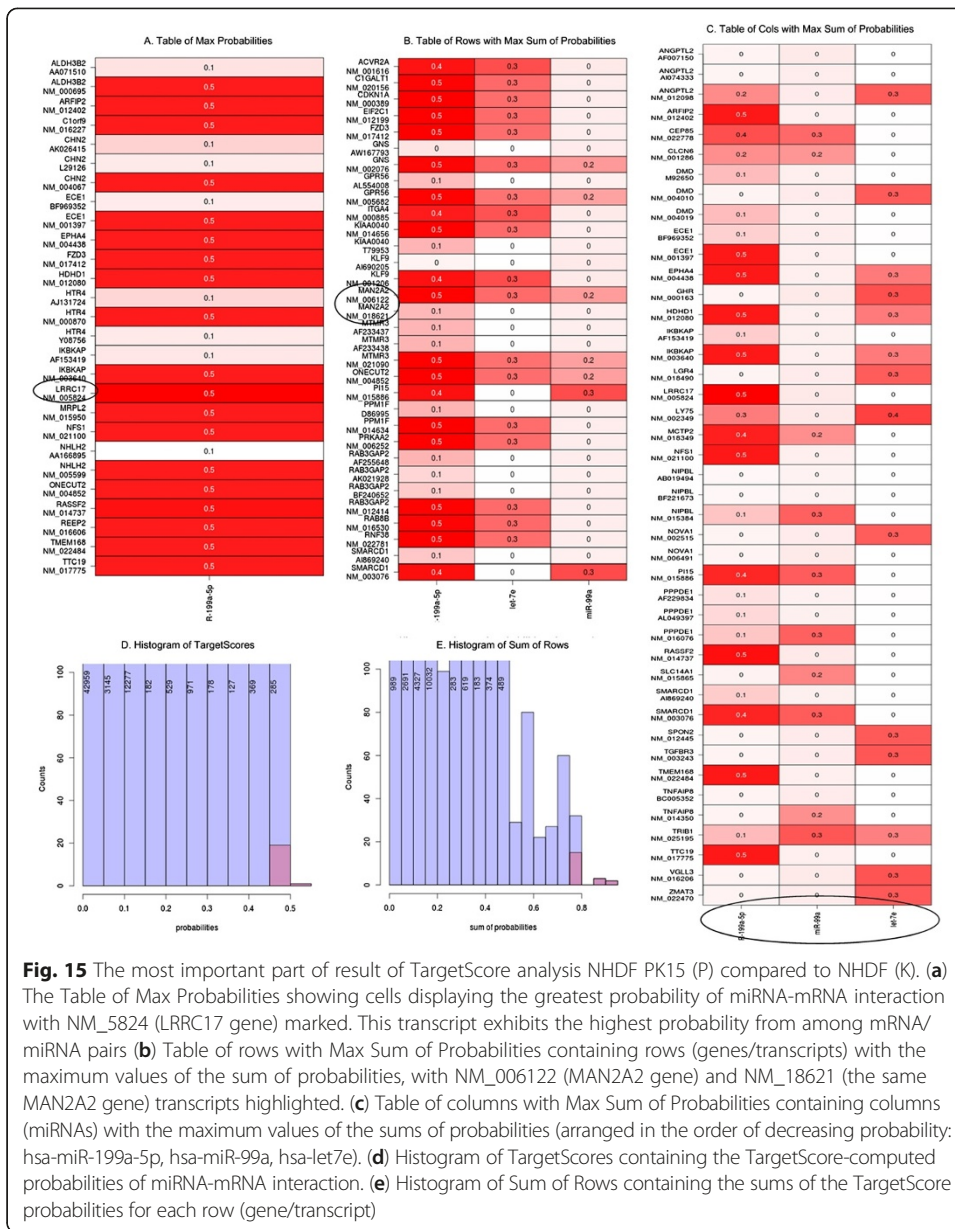


Fig. 15 The most important part of result of TargetScore analysis NHDF PK15 (P) compared to NHDF (K). (a) The Table of Max Probabilities showing cells displaying the greatest probability of miRNA-mRNA interaction with NM_5824 (LRRIC17 gene) marked. This transcript exhibits the highest probability from among mRNA/miRNA pairs (b) Table of rows with Max Sum of Probabilities containing rows (genes/transcripts) with the maximum values of the sum of probabilities, with NM_006122 (MAN2A2 gene) and NM_18621 (the same MAN2A2 gene) transcripts highlighted. (c) Table of columns with Max Sum of Probabilities containing (miRNAs) with the maximum values of the sums of probabilities (arranged in the order of decreasing probability: hsa-miR-199a-5p, hsa-miR-99a, hsa-let7e). (d) Histogram of TargetScores containing the TargetScore-computed probabilities of miRNA-mRNA interaction. (e) Histogram of Sum of Rows containing the sums of the TargetScore probabilities for each row (gene/transcript)

based on the full TargetScore results and the Table of Max Probability; see Fig. 15). The distribution of probabilities shows that the highest value is not an outlier. This suggests that we could expect many miRNA/mRNA pairs with slightly lower probability values. We select the transcripts with the best miRNA interaction probabilities from among the full TargetScore results and from the table of rows listing maximum probability sums (see Fig. 15). The selected transcript, NM_006122, is encoded by the MAN2A2 gene. The corresponding histogram facilitates assessment of the specificity of each value. By analogy, we could assess miRNAs by using maximum probability sum columns to select the highest probability of interference with any transcripts. The resulting miRNAs, arranged in the order of decreasing probability, are: hsa-miR-199a-5p, hsa-miR-99a and hsa-let7e (see Fig. 15).

Table 1 Comparison of the characteristics of popular microarray analysis tools

Platform characteristic	WebArrayDB [11]	ArrayMining [12]	Babelomics [13]	DNA Microarray integromics analysis platform
miRNA-mRNA target site recognition	-	-	-	+
Integration of transcriptomics and lipidomics	-	-	-	+
Analysis of genetic variation	-	-	+ ³	-
Functional profiling	-	-	+ ¹	+ ²
Conditional analysis of gene expression based on ANN	-	-	-	+
Imports and exports data to EBI databases	-	-	-	+
Teamwork and collaboration	-	-	-	+
License and source code availability	GNU General Public License	Free for academic and non-commercial use; source code not available	Free; source code available	Free for academic and non-commercial use; source code available

¹Includes text mining, knowledge- and network-based annotations and gene set enrichment analysis; ²includes knowledge- and network-based annotations and gene set enrichment analysis; ³includes association analysis and genotype stratification

Conclusions

The DNA Microarray Integromics Analysis Platform is a user-friendly, open-source online system that provides a panoply of computational tools for biological data analysis. Its web-based graphical user interface renders the platform more accessible to non-IT experts, enabling domain scientists to interact with it at each step of the microarray data analysis, and to integrate biological data from other sources.

Several web-based DNA microarray data analysis platforms currently exist. However, most of these platforms do not support the multi-omics approach as a means of enhancing insight into molecular biology mechanisms. Table 1 compares selected characteristics and features of the DNA Microarray Integromics Analysis Platform with those of other web-based tools.

Only the Babelomics [13] platform includes any algorithms for integrative analysis of transcriptomics, proteomics and genomics data. Integromics analyses for transcriptomics and lipidomics data or mRNA and miRNA data are available exclusively in the DNA Microarray Integromics Analysis Platform.

Abbreviations

ANN: Artificial neural network; GUI: Graphical user interface; RMA: Robust multi-array average; GCRMA: Guanine cytosine robust multi-array analysis; CCA: Canonical correlation analysis; PLS: Partial least squares regression.

Competing interests

The authors declare that they have no financial or non-financial competing interests.

Authors' contributions

TW designed the GUI of the platform and implemented, using the R language, low-level and high-level analyses of the DNA microarray data. TG was the primary architect and developer of the web interface and job execution components of the platform. KS generated the workflow for miRNA-mRNA interaction analysis and implemented it using R. MP and WJ designed the workflows for transcriptomics and lipidomics data analysis and coded these workflows using R. All authors have read and approved the final manuscript.

Acknowledgements

The DNA Microarray Integromics Analysis Platform was developed in the scope of the PLGrid Plus project, which was co-financed by the European Union (under the Innovative Economy Fund). The authors wish to thank Kazimierz Murzyn, GrzegorzPacoch, DariuszKról, Tomasz Bartyński and Piotr Nowakowski for their involvement in the design and development of the platform. We also thank UrszulaMazurek and her collaborators from the Medical University of Silesia for sharing their DNA microarray data.

Author details

¹Institute of Computer Science, Division of Biomedical Computer Systems, University of Silesia, Katowice, Poland.

²Academic Computer Centre CYFRONET, AGH University of Science and Technology, Kraków, Poland. ³Molecular Biology and Clinical Genetics Laboratory, Department of Medicine, Jagiellonian University, Kraków, Poland.

⁴Department of Bioinformatics and Telemedicine, Medical College, Jagiellonian University, Kraków, Poland. ⁵The Genome Analysis Centre, Norwich, UK.

Received: 1 November 2014 Accepted: 19 June 2015

Published online: 25 June 2015

References

1. Kohl M, Megger DA, Trippler M, Meckel H, Ahrens M, Bracht T, et al. A practical data processing workflow for multi-OMICS projects. *Biochim Biophys Acta*. 2014;1844(1 Pt A):52–62.
2. Zhang M, Zhang Y, Liu L, Yu L, Tsang S, Tan J, et al. Gene Expression Browser: large-scale and cross-experiment microarray data integration, management, search & visualization. *BMC Bioinformatics*. 2010;11:433.
3. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, et al. ArrayExpress update - trends in database growth and links to data analysis tools. *Nucleic Acids Res*. 2013;41(D1):D987–90.
4. Wu S, Xu Y, Feng Z, Yang X, Wang X, Gao X. Multiple-platform data integration method with application to combined analysis of microarray and proteomic data. *BMC Bioinformatics*. 2012;13:320.
5. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics (Oxford England)*. 2009;25(23):3049–55.
6. Muniategui A, Nogales-Cadenas R, Vázquez M, Aranguren XL, Agirre X, Luttun A, et al. Quantification of miRNA-mRNA interactions. *PLoS One*. 2012;7(2):e30766.
7. Hui L, Yue D, Yidong C, Gao SJ, Huang Y. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*. 2010;11(1):476.

8. Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.* 2009;37(Web Server issue):W356–362.
9. Huang JC, Frey BJ, Morris QD. Bayesian: Inference of MicroRNA Targets from Sequence and Expression Data. *J Comput Biol.* 2007;14(5):550–63.
10. Le HS, Bar-Joseph Z: Inferring interaction networks using the IBP applied to microRNA target prediction. *Adv in Neural Inf Process Syst* 2011;2011:235–243.
11. Xia XQ, McClelland M, Porwollik S, Song W, Cong X, Wang Y. WebArrayDB: Cross-platform microarray data analysis and public data repository. *Bioinformatics.* 2009;25(18):2425–9.
12. Glaab E, Garibaldi J, Krasnogor N. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics.* 2009;10:358.
13. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.* 2005;33(Web Server issue):W460–4.
14. M. Bubak, T. Szeplieniec, K. Wiatr. Building a National Distributed e-Infrastructure - PL-Grid - Scientific and Technical Achievements. Springer-Verlag Berlin Heidelberg. 2012.
15. Gentleman R, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* 2004;5:R8.
16. DNA Microarray Integromics Analysis Platform source code available at: <https://git.plgrid.pl/projects/IGR/>. If the version control repository for this project has moved, please contact the authors for an updated URL.
17. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics (Oxford, England).* 2003;4(2):249–64.
18. Wu Z, Irizarry R, Gentleman RC, Murillo FM, Spencer F. A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc.* 2004;99(468):909–17.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England).* 2003;19(2):185–93.
20. Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. *Pharmacogenomics.* 2003;4:1.
21. Larsson O, Wahlestedt C, Timmons JA. Considerations when using the significance analysis of microarrays (SAM) algorithm. *BMC Bioinformatics.* 2005;6:129.
22. Waller T, Nowak R, Tkacz M, Zapart D, Mazurek U. Familial or Sporadic Idiopathic Scoliosis – classification based on artificial neural network and GAPDH and ACTB transcription profile. *BioMed Eng Online.* 2013;12:1.
23. Wang A, Gehan EA. Gene selection for microarray data analysis using principal component analysis. *Stat Med.* 2005;24(13):2069–87.
24. Irizarry RA, Wang C, Zhou Y, Speed TP. Gene Set Enrichment Analysis Made Simple. *Stat Methods Med Res.* 2009;18(6):565–75.
25. Kelder T, Pico AR, Hanspers K, van Iersel MP, Evelo C, Conklin BR. Mining Biological Pathways Using WikiPathways Web Services. *PLoS ONE.* 2009;4:7.
26. Cserhádi T, Kósa A, Balogh S. Comparison of partial least-square method and canonical correlation analysis in a quantitative structure-retention relationship study. *J Biochem Biophys Methods.* 1998;36(2-3):131–141.
27. Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol.* 2008;7(1):35
28. Chung, Chun, & Keles. Sparse Partial Least Squares (SPLS) Regression and Classification, <http://cran.r-project.org/package=spls>.
29. Lê Cao KA, Martin PG, Robert-Granié C, Besse P. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics.* 2009;10:34.
30. Tenenhaus A, Philippe C, Guillemot V, Le Cao KA, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. *Biostatistics.* 2014;15(3):569–83.
31. Shen YF, Zhu J. Power analysis of principal components regression in genetic association studies. *Journal of Zhejiang University.* 2009;10(10):721–730.
32. Gutmann MU, Laparra V, Hyvärinen A, Malo J. Spatio-Chromatic Adaptation via Higher-Order Canonical Correlation Analysis of Natural Images. *PLoS ONE* 2014;9(2):e86481.
33. Witten DM, Tibshirani RJ. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology.* 2009;8(1):28.
34. Li X, Gill R, Cooper NG, Yoo JK, Datta S. Modeling microRNA-mRNA Interactions Using PLS Regression in Human Colon Cancer. *BMC Medical Genomics.* 2011;4:44.
35. Butts, yacca. Yet Another Canonical Correlation Analysis Package, <http://cran.r-project.org/package=yacca>.
36. Cruz-Cano R, Lee MLT. Fast regularized canonical correlation analysis. *Comput. Stat. Data Anal.* 2014;70:88–100.
37. Mevik, Wehrens, & Liland. Partial Least Squares and Principal Component regression, <http://cran.r-project.org/package=pls>.
38. González I, Cao KA, Davis MJ, Dejean S. Visualising associations between paired ‘omics’ data sets. *BioData Min.* 2012;5:19.
39. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell.* 2009;136:215–33.
40. TargetScanHumanRelease 6.2, June 2012 (<http://www.targetscaan.org/>).
41. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nat Struct Mol Biol.* 2011;18:1139–46.
42. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 2009;19(1):92–105.
43. Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature.* 2011;469(7328):97–101.

44. Yue Li: Roleswitch: Infer miRNA-mRNA interactions using paired expression data from a single sample. R package version 1.0.0. (2013) <http://www.cs.utoronto.ca/~yueli/roleswitch.html>.
45. Li Y, Goldenberg A, Wong KC, Zhang Z. A probabilistic approach to explore human microRNA targetome using microRNA-overexpression data and sequence information. *Bioinformatics*. 2014;30(5):621–8.
46. Huber W, von Heydebreck A, Suelmann H, Poustka A, Vingron M. Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression. *Bioinformatics*. 2012;18 Suppl 1:S96–S104.
47. Brazma A et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet*. 2001;29:365–71.
48. Martin PGP, Guillou H, Lasserre F, Déjean S, Lan A, et al. Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*. 2007;45:767–777.
49. Kimsa MC, Strzałka-Mrozik B, Kimsa MW, Kruszniewska-Rajs C, Gola J, Adamska J, et al. Porcine endogenous retrovirus infection changes the expression of inflammation-related genes in lipopolysaccharide-stimulated human dermal fibroblasts. *Ann Transplant*. 2013;18:576–86.
50. Affymetrix miRNA QCTool. <http://www.affymetrix.com/support/technical/byproduct.affx?product=mirnaqctool>.
51. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42:D199–205.
52. Croft D, Mundo AF, Haw R, et al. The Reactome pathway knowledgebase. *Nucleic Acids Research*. 2014;42:D472–7.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

