

MetaboAnalyst 2.0—a comprehensive server for metabolomic data analysis

Jianguo Xia¹, Rupasri Mandal², Igor V. Sinelnikov², David Broadhurst³ and David S. Wishart^{1,2,4,*}

¹Department of Biological Sciences, ²Department of Computing Science, ³Department of Medicine and ⁴National Research Council, National Institute for Nanotechnology (NINT), Edmonton, AB, Canada T6G 2E8

Received February 13, 2012; Revised April 3, 2012; Accepted April 12, 2012

ABSTRACT

First released in 2009, MetaboAnalyst (www.metaboanalyst.ca) was a relatively simple web server designed to facilitate metabolomic data processing and statistical analysis. With continuing advances in metabolomics along with constant user feedback, it became clear that a substantial upgrade to the original server was necessary. MetaboAnalyst 2.0, which is the successor to MetaboAnalyst, represents just such an upgrade. MetaboAnalyst 2.0 now contains dozens of new features and functions including new procedures for data filtering, data editing and data normalization. It also supports multi-group data analysis, two-factor analysis as well as time-series data analysis. These new functions have also been supplemented with: (i) a quality-control module that allows users to evaluate their data quality before conducting any analysis, (ii) a functional enrichment analysis module that allows users to identify biologically meaningful patterns using metabolite set enrichment analysis and (iii) a metabolic pathway analysis module that allows users to perform pathway analysis and visualization for 15 different model organisms. In developing MetaboAnalyst 2.0 we have also substantially improved its graphical presentation tools. All images are now generated using anti-aliasing and are available over a range of resolutions, sizes and formats (PNG, TIFF, PDF, PostScript, or SVG). To improve its performance, MetaboAnalyst 2.0 is now hosted on a much more powerful server with substantially modified code to take advantage the server's multi-core CPUs for computationally intensive tasks. MetaboAnalyst 2.0 also maintains a collection of 50 or more FAQs and more than a dozen tutorials compiled from user

queries and requests. A downloadable version of MetaboAnalyst 2.0, along detailed instructions for local installation is now available as well.

INTRODUCTION

MetaboAnalyst is a web-based suite for high-throughput metabolomic data analysis. It was originally released in 2009 (1). As the first dedicated web server for metabolomic data processing, MetaboAnalyst quickly gained considerable traction within the metabolomics community. To date it has served more than 10 000 researchers from nearly 1500 different institutions. However, the rapidly changing nature of metabolomics (both in terms of the technology and the experimental design) has meant that some of the analytical and graphical components in the original version of MetaboAnalyst have become outdated, insufficient or inadequate. For instance, as more and more metabolomics researchers adopt quantitative or targeted metabolomic approaches (2–5), requests by our users for improved methods to perform functional or biological interpretation have continued to grow. Likewise, with increasing numbers of large-scale metabolomic studies being performed, the need for data quality control (QC) and quality assessment support has become much more apparent (6,7) and much more frequently requested by MetaboAnalyst users. Other obvious trends in metabolomic data analysis include: (i) a greater demand for tools to support time-series analysis, (ii) a growing need to support the statistical analysis of more complex experimental designs (8) and (iii) requests to offer more stringent evaluation of the results generated from chemometric analyses (9). As the popularity of MetaboAnalyst has grown, so too has its use by non-experts or statistically naïve users. This has led to numerous requests by users to simplify its interface, to improve the graphics, to accelerate the calculations and to provide more user support.

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 5305; Email: david.wishart@ualberta.ca

In response to these requests and in anticipation of upcoming analytical demands we have developed MetaboAnalyst 2.0. The new edition of MetaboAnalyst represents a substantially upgraded and significantly improved version over what was described in 2009. In particular, MetaboAnalyst 2.0 now includes a variety of new modules for data processing, data QC and data normalization. It also has new tools to assist in data interpretation, new functions to support multi-group data analysis, as well as new capabilities in correlation analysis, time-series analysis and two-factor analysis. We have also updated and upgraded the graphical output to support the generation of high resolution, publication quality images. Furthermore, almost every module in MetaboAnalyst has been rewritten, refactored and optimized for performance. For example, many CPU-intensive functions have been rewritten to take full advantage of the host server's multi-core processors. We have also added a powerful new server to support dedicated backend statistical computing. Additionally, dozens of tutorials and more than 50 FAQs have been added to the website to address common user questions and to facilitate user interactions.

GENERAL OVERVIEW

Fundamentally, MetaboAnalyst 2.0 is a web-based pipeline that supports step-wise metabolomic data analysis. The main analytical modules in MetaboAnalyst 2.0 are summarized in Figure 1. Briefly, after a user uploads their metabolomic data set a number of different data processing, labeling, filtering and cleansing procedures are then employed to transform the raw data into 'cleaner' data table. Users can then choose from an array of data normalization methods to prepare the data for further downstream analysis. MetaboAnalyst 2.0 has four kinds of data analysis modules: binary/multiple-group data analysis, two-factor/time-series data analysis, metabolite set enrichment analysis and metabolic pathway analysis. The latter two modules are based on literature-derived knowledge, specialized metabolic pathway databases [such as SMPDB (10)] and richly annotated metabolite databases [such as HMDB (11)]. In addition, MetaboAnalyst 2.0 also includes a number of 'general utility' functions for QC analysis, high-resolution image generation, metabolite name/ID conversion, pathway mapping, peak searching and many other functions. These modules, along with a number of newly added functions and enhanced features, are described in more detail below.

Data processing and normalization

Data processing is key to every aspect metabolomic data analysis, and this is especially true for data acquired via chemometric or untargeted metabolomic approaches. Most of the functions in this module were available in the original version of MetaboAnalyst, including peak picking, peak alignment and missing value imputation. However, a number of

new options have been added to MetaboAnalyst 2.0 including:

- (i) Data filtering: this function allows users to selectively remove low-quality data points from their metabolomic datasets. This filtering step can usually improve the performance and reduce the false discovery rates (FDR) in downstream statistical analysis (12,13). MetaboAnalyst 2.0 now offers two types of filters: (a) a low-value threshold filter to exclude data points having low values based on known detection limits, or based on sample means and/or medians; or (b) a low-variance filter to exclude data points showing little variance across experimental conditions based on their interquartile range (IQR), coefficient of variation (CV), or standard deviation. This filtering procedure is often not necessary for quantitative (i.e. data with absolute concentrations) but it is highly recommended for chemometric or non-quantitative data, which often contains a large amount of noise.
- (ii) Data editing: the data editor allows users to manually exclude specific samples or variables from their data without having to exit the program, edit their input data in a separate program and reload the data. This data editor may be used to remove data outliers after they have been identified in MetaboAnalyst's downstream analysis.

Another new feature in MetaboAnalyst 2.0's data processing module is a *color picker* tool. This particular function allows users to set colors for different group labels, thereby simplifying subsequent analyses and enhancing subsequent graphical displays. Complementing the color picker's capacity for data customization is an automated name normalization feature that facilitates data standardization. Nomenclature standardization is particularly important for any user wanting to do functional annotation or functional analysis. This name standardization function is able to recognize common compound names and synonyms and map them to specific HMDB, KEGG, PubChem and ChEBI identifiers which are, in turn, linked to MetaboAnalyst's extensive metabolite knowledgebases.

Another important part of data processing or pre-processing is data normalization. Many standard statistical algorithms [such as *t*-tests and analysis of variance (ANOVA) tests] work under the assumption that the data being analyzed are normally distributed (i.e. they follow a Gaussian or 'normal' distribution). If the data are not normally distributed or cannot be transformed into a normal distribution, then most standard statistical tests become unreliable. In addition to improving reliability and interpretability, data normalization can also help reduce any systematic bias in the data that may have arisen from instrumental or sampling problems. MetaboAnalyst 2.0 now supports 11 different procedures for data transformation and normalization (14,15) all of which can be interactively visualized and assessed using various diagnostic plotting tools. While it is often

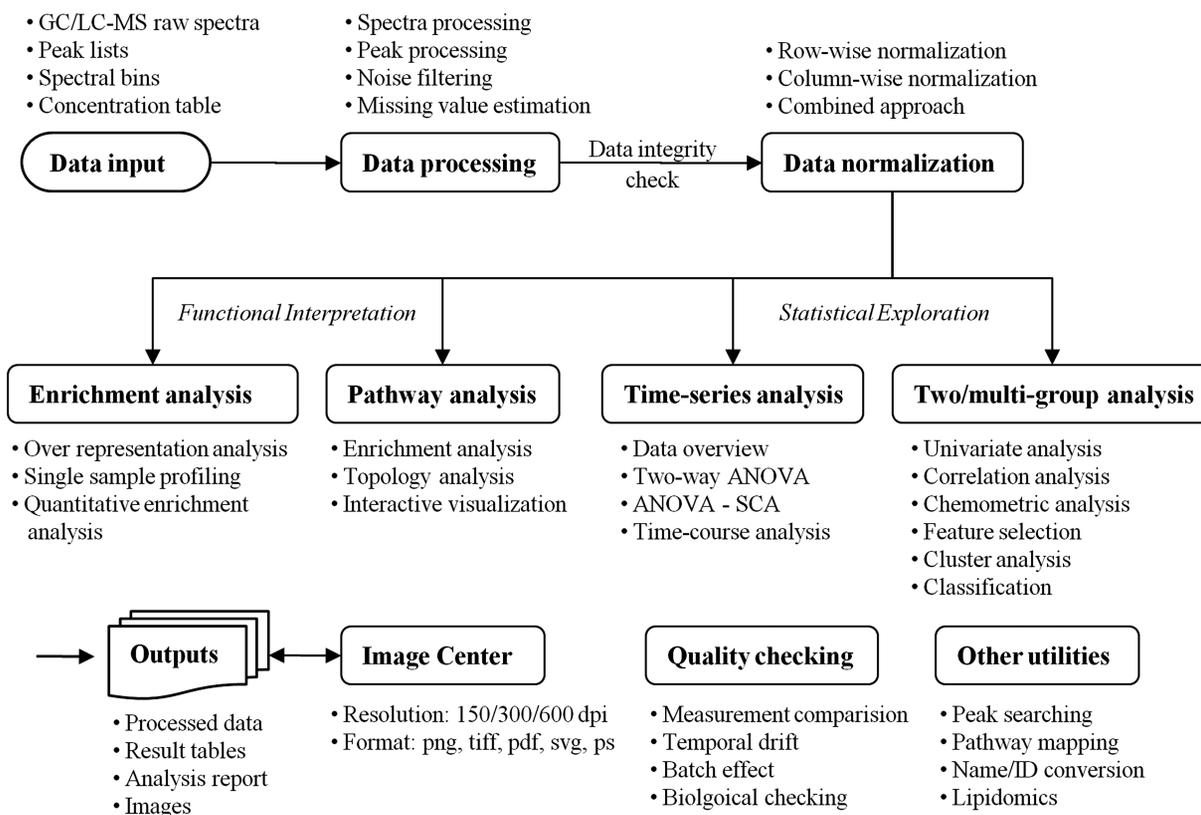


Figure 1. Overview of MetaboAnalyst 2.0. This figure summarizes the main modules, functions and features in MetaboAnalyst 2.0. Black arrows show the flow of information from the input data to the data processing module to the four major analytical modules including: (i) multi-group statistical analysis; (ii) time-series analysis; (iii) enrichment analysis and (iv) pathway analysis. Also illustrated are a number of utilities for data quality checking, metabolite name/ID conversion, peak searching, and compound annotation.

impossible to predict which normalization step is most appropriate for a given data set, FAQs and tutorials in MetaboAnalyst 2.0 now provide a number of helpful suggestions and useful guidance.

Data quality checking

With larger studies becoming the norm in metabolomic research, data quality assurance (QA) and QC is becoming an increasingly important issue. This is especially critical for large-scale, collaborative metabolomic projects involving multiple personnel, multiple analytical platforms or large-scale analyses conducted over a long period of time. MetaboAnalyst 2.0's new quality checking (QA/QC) module allows users to examine their data for systematic variations or abnormal patterns that are unrelated to the experimental design such as temporal drift, instrumental drift, batch effects, sample preparation effects or sample decay. This module currently contains four functions:

- (i) Pair-wise comparison of two measurements: this function is designed to compare the measurement consistency between two different protocols, two different instrument models or even two different platforms. A scatter plot is presented with a user-adjustable threshold to label data points with

large deviations from the acceptable range of variation.

- (ii) Detection and correction of temporal drift: this data quality checking function helps determine whether there is a significant drift in the measurements using the same instrument over a long period of time. Data points are first divided into a user-adjustable number of time frames and pair-wise *P*-values are calculated between different segments. Scatter plots and box plots are produced to help visualize the results. Additionally, a LOWESS (Locally Weighted Scatterplot Smoothing) curve is generated and the LOWESS corrected data values can be downloaded and used for subsequent analysis.
- (iii) Detection of batch effects: this data quality checking method offers both univariate and multivariate approaches to help evaluate whether there are significant systematic variations among samples measured in different batches. For multivariate approaches, Principal component analysis (PCA) and heatmaps are used to visualize all data points across different batches. The univariate approach allows users to inspect individual samples or variables presented in both scatter and box plots across all batches. In either case, QC samples (if present) are highlighted to assist researchers in making their judgments.

- (iv) Checking against reference concentration ranges: this method, which is restricted to human data only, compares the measured concentration values in a user's data set with the corresponding reported values (for a given biofluid such as blood, urine or cerebrospinal fluid) from the literature. This approach is particularly useful to assess sample quality, sample contamination or to identify potentially mislabeled or misidentified metabolites

Identification of important compounds

The identification of significantly perturbed metabolites or metabolite peaks is probably the most basic task in metabolomic data analysis. A variety of strategies have been implemented in MetaboAnalyst 2.0 to help researchers select or identify compounds of interest and to facilitate different research objectives.

- (i) Differential expression analysis: this module, which has been extended to support multiple group analysis, can be used by users to identify compounds that are significantly different between two or more sets of experimental conditions or two or more populations under study. MetaboAnalyst 2.0 supports both ordinary univariate methods (i.e. *t*-tests, ANOVA with *post hoc* analysis) and moderated *t*-statistic methods [i.e. SAM (16) and EBAM] to compare means or medians of one variable across two or more groups. Because of the multiple-testing issue, FDR or Bonferonni corrected *P*-values are also computed for these functions.
- (ii) Coexpression analysis: this method, which is new to MetaboAnalyst 2.0, aims to help researchers identify compounds that either share similar or show opposite patterns of concentration changes across different conditions. These trends can be visualized through correlation heatmaps. Users can also perform correlation analysis against a query/target compound to identify other compounds that are either positively or negatively correlated to the query compound. Many similarity measures can be used, including Euclidean distance, Pearson's correlation, Spearman's rank correlation and Kendall's tau test.
- (iii) Pattern search: this function, which is also new to MetaboAnalyst 2.0, allows users to specify and search for specific patterns of interest in their data using a template matching method (17). The results are a ranked list of variables with similarity scores and *P*-values. The template patterns can be specified as a series of integers indicating the expected (relative or absolute) concentration levels in different groups or at different time points.

Clustering and classification

Clustering and classification methods are routinely used in almost all metabolomic studies. Consequently these functions have been among the most widely used tools in

MetaboAnalyst. Two different options are offered, traditional multivariate statistical methods (called Chemometric analysis) and newer machine learning methods.

- (i) Chemometric analysis: PCA and partial least squares discriminant analysis (PLS-DA) were both available in the original release of MetaboAnalyst, but for MetaboAnalyst 2.0, all the graphical display tools have been re-engineered to improve image quality, image content and rendering efficiency. Furthermore, based on user feedback, a number of new options (i.e. PLS-DA permutation tests and several other graphical display options) have been added to improve the presentation and user-friendliness.
- (ii) Machine learning approaches MetaboAnalyst 2.0 continues to support hierarchical clustering with heatmaps as well as self-organizing maps and *k*-means clustering that were available in the original version of MetaboAnalyst. However the visualization utilities for these unsupervised clustering methods have been substantially improved in MetaboAnalyst 2.0. For supervised analysis, both Random Forest and Support Vector Machines continue to be offered but with the addition of several new analysis options and a number of graphical enhancements.

Analysis of data from time-course and two-factor design

An increasing number of metabolomic studies focus on time-series analysis (i.e. studying treatment effects over different time points for different plant or animal strains) or general two-factor comparisons (i.e. studying different plant or animal strains under different growth conditions). In response to these trends, MetaboAnalyst 2.0 now offers a new module specifically designed to support these kinds of analyses. Three different functions are available in this module.

- (i) Two-way data visualization: this sub-module generates a two-way, two-dimensional heatmap to help users visualize hierarchical clustering results. It also supports an interactive three-dimensional plot to help users visualize PCA results. Both approaches allow users to easily explore trends or metabolite distribution patterns according to different factors or over a range of time points.
- (ii) Two-factor data analysis: this sub-module allows users to perform ordinary two-way within-subject or between-subject analysis of variance (ANOVA). It also supports its multivariate extension known as ANOVA-simultaneous component analysis (ASCA) (18) for two-factor and time-series data analysis.
- (iii) Multivariate time-course profiling this function offers users a multivariate empirical Bayes statistical approach for time-course analysis that has been specifically designed for analyzing omic data generated from replicated longitudinal time course experiments (19).

All three of these methods, including their strengths, their limitations and their implementation have been discussed in more detail in a recent paper by our group (20).

Functional analysis of quantitative metabolomic data

The increasing use of quantitative metabolomic techniques (i.e. measurement of absolute concentrations of metabolites) has opened the door to significantly improved functional analysis and biological interpretation. For MetaboAnalyst 2.0 we have added two modules to help researchers extract functional or biologically relevant information from their metabolomic data. These include: (i) metabolite set enrichment analysis (MSEA) and (ii) metabolic pathway analysis (MetPA). These two general approaches, along with their four specific functions have been discussed in more detail in previous papers by our group (21,22).

- (i) Overrepresentation analysis (ORA): this method allows one to identify biologically meaningful patterns given a list of significantly altered metabolites.
- (ii) Single sample profiling (SSP): this method compares the metabolite concentrations measured from a human biofluid sample (urine, blood or CSF) to their normal reference values reported in literature. Compounds outside their normal ranges are then further processed using ORA.
- (iii) Quantitative enrichment analysis (QEA): this method performs enrichment analysis directly from a compound concentration table using an approach similar to the widely-used gene set enrichment analysis (23,24);
- (iv) Metabolic pathway analysis (MetPA): this module combines functional enrichment analysis and network topology analysis to help identify metabolic pathways that are most likely to be associated with the condition under study. Associated pathways can be visualized and explored via a Google-map style interactive visualization framework.

As noted earlier, the above functional analyses are only applicable to metabolomic data containing metabolite concentration measurements. Furthermore, the enrichment analyses (options i–iii) are restricted to human-only data as the underlying metabolite set libraries were derived from human-only databases. On the other hand, the pathway analysis function (option iv) currently contains pathway information from 16 different model organisms.

Data formats, data handling and data privacy

MetaboAnalyst 2.0 uses a very simple input tabular format, namely a comma-separated value (.csv) plain text file in which the biosamples or patients are listed in rows and the compounds listed in columns (the default format). The first column should contain unique sample identifiers and the second column should contain phenotype labels. Transposed input tables are also accepted in MetaboAnalyst 2.0. The server also accepts other data formats commonly produced from untargeted metabolomic studies, including spectral bins, peak

intensity tables, peak lists and raw spectra. The 'Data Formats' page in MetaboAnalyst 2.0 provides detailed instructions, screenshots as well as downloadable sample data sets for all the data formats that it supports.

MetaboAnalyst 2.0 is primarily designed for quantitative metabolomic studies and readily handles metabolomic data matrices of 100 samples \times 1000 quantified compounds without any compromise in performance. Untargeted metabolomic data sets consisting of 100 or more samples having more than 1000 spectral bins or up to 5000 spectral features each, are also easily accommodated by the server. If very large data sets need to be processed (e.g. 1000 samples 5000 features) users should download and install a local copy of MetaboAnalyst 2.0 according to the detailed instructions given on the website.

Each time a user starts a MetaboAnalyst session a temporary folder with a random 16-digit name is created. This is where the uploaded user data and all analytical results are stored. Users are expected to download all their analyses, tables, graphs and reports upon completion of a session. User data will remain on the server for 72 h and then is automatically deleted. All user data handled by MetaboAnalyst 2.0 is treated as private and confidential. We leave it to the users to release, share or publish their MetaboAnalyst results in any manner they wish.

Data visualization and image center

Data visualization is the key to understanding and interpreting high-dimensional omics data. Since its first release in 2009, a significant amount of time and effort has been invested to improve MetaboAnalyst's data visualization features. For MetaboAnalyst 2.0, all the images are now anti-aliased using the Cairo graphics library (<http://cairographics.org>). Additionally many of MetaboAnalyst's original graphical displays have been completely redesigned to be more informative, to provide more options for label and display modification and to support more interactive exploration. MetaboAnalyst 2.0's *Image Center* now allows images to be generated over a range of resolutions (150/300/600 dpi), sizes and formats (PNG, TIFF, PDF, PostScript, or SVG) for publication or other presentation purposes.

Implementation and local installation

MetaboAnalyst 2.0 was built using the JavaServer Faces (JSF) technology. The majority of the backend computations are carried out by over 500 functions written in R. The communication between Java and R is established through TCP/IP using the *Rserve* program (<http://www.rforge.net/Rserve>). MetaboAnalyst can handle concurrency (multi-users logging on at the same time) and can be configured to run on multiple CPUs to achieve better performance. To cope with the heavy user load, the MetaboAnalyst 2.0 server is currently hosted on two dedicated machines that are maintained and updated regularly. For researchers having particularly large amounts of data or for those requiring secure data handling, copies of MetaboAnalyst 2.0 are now available for download and local installation. Detailed installation

instructions are available on the 'Resources' page. MetaboAnalyst 2.0 compiles easily on Linux or Mac OS, and can be adapted to run on Windows OS. MetaboAnalyst 2.0 has been successfully installed in a number of national metabolomic research centers as well as in many individual labs/companies. As with any large and complex program, some computer skills are required to install the required underlying R packages.

Tutorials and training programs on MetaboAnalyst

MetaboAnalyst 2.0 has been designed as a user-friendly server to help researchers carry out complex data analysis tasks using a self-guided menu, interactive displays and a series of simple mouse clicks. While user-friendliness and simplicity have been paramount in MetaboAnalyst's overall design, it is still important that users understand basic statistical principles and the meanings behind key parameters in order to make effective use of its many functions and tools. To address this need, we have published two comprehensive tutorials to provide some statistical background together with detailed step-by-step instructions using several examples based on publicly available datasets (25,26). An additional collection of on-line tutorials has also been prepared for this release of MetaboAnalyst 2.0 along with approximately 50 FAQs covering many common topics and questions.

DISCUSSION

MetaboAnalyst 2.0 is a substantially upgraded version of MetaboAnalyst that has been designed to support robust statistical analysis, data exploration, data visualization, as well as functional interpretation of metabolomic data. The latest version now includes substantially improved graphics and graphical displays, as well as new procedures for data filtering, data editing and data normalization. MetaboAnalyst 2.0 also supports multi-group data analysis, two-factor analysis as well as time-series data analysis. Additionally, MetaboAnalyst 2.0 now includes a new data QC module, a functional enrichment analysis module and a metabolic pathway analysis module. Furthermore, all new and previously existing functions have been substantially enhanced with the development of more efficient code, the use of more/faster processors and the creation of more extensive on-line tutorials. Most of these new tools, functions, FAQs and tutorials were developed in response to specific requests made by our users or in anticipation of clearly emerging trends based on recently published metabolomic studies.

The design of MetaboAnalyst was actually inspired and influenced by two popular open-source transcriptomics data analysis pipelines—GEPAS (27) and GenePattern (28). The three key elements in these servers—*Analysis and Visualization*, *Data Pipelines* and *Servers*, served as thematic templates throughout the design and implementation of both MetaboAnalyst and MetaboAnalyst 2.0. As judged by the heavy use and frequent citations of GEPAS, GenePattern and MetaboAnalyst, this modular, pipelined approach to the analysis of 'omics' data has proven to be very popular with bench biologists. It has also very useful

for tool development, as each function or each module can be developed and debugged independently.

MetaboAnalyst 2.0 is still a work in progress and there are certainly features or options that could/should be added. For instance, the emerging metabolomic ontologies and reporting standards, which are still evolving, are not yet supported. Likewise MetaboAnalyst's support for processing raw GC-MS and LC-MS spectra is still quite limited. Raw spectral processing usually requires platform-specific or proprietary software along with significant manual manipulation in order to obtain satisfactory results. For now, we believe this task is best handled by locally installed programs designed and tuned for specific instruments. However, freely available tools are starting to appear, such as *XCMSOnline* (<https://xcmsonline.scripps.edu>) for LC-MS spectral processing and *MetabolomeExpress* (29) for GC-MS spectral processing. The peak lists generated by these tools can be easily uploaded into MetaboAnalyst 2.0 for further downstream analysis.

It is also important to note that MetaboAnalyst is no longer 'the only game in town'. Since it was first introduced in 2009, an increasing number of web-based or downloadable tools have become available for analyzing metabolomic data (30–34). Many of them are nicely designed and easy to use, however most are restricted to handling only one or two aspects of metabolomic data analysis (i.e. spectral processing or clustering/classification). In this regard, MetaboAnalyst 2.0 is still quite unique as it continues to offer the most comprehensive suite of web tools for metabolomic data processing, visualization and analysis.

FUNDING

Genome Alberta; Genome Canada; Canadian Institutes of Health Research; Alberta Innovates. Funding for open access charge: Genome Alberta.

Conflict of interest statement. None declared.

REFERENCES

- Xia, J., Psychogios, N., Young, N. and Wishart, D.S. (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res.*, **37**, W652–W660.
- Weljie, A.M., Newton, J., Mercier, P., Carlson, E. and Slupsky, C.M. (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal. Chem.*, **78**, 4430–4442.
- Dietmair, S., Timmins, N.E., Gray, P.P., Nielsen, L.K. and Kromer, J.O. (2010) Towards quantitative metabolomics of mammalian cells: development of a metabolite extraction protocol. *Anal. Biochem.*, **404**, 155–164.
- Griffiths, W.J., Koal, T., Wang, Y., Kohl, M., Enot, D.P. and Deigner, H.P. (2010) Targeted metabolomics for biomarker discovery. *Angewandte Chemie*, **49**, 5426–5445.
- Ewald, J.C., Heux, S. and Zamboni, N. (2009) High-throughput quantitative metabolomics: workflow for cultivation, quenching, and analysis of yeast in a multiwell format. *Anal. Chem.*, **81**, 3623–3629.
- Bijlsma, S., Bobeldijk, I., Verheij, E.R., Ramaker, R., Kochhar, S., Macdonald, I.A., van Ommen, B. and Smilde, A.K. (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal. Chem.*, **78**, 567–574.

7. Dunn, W.B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J.D., Halsall, A., Haselden, J.N. *et al.* (2011) Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protocols*, **6**, 1060–1083.
8. Rubingh, C.M., Bijlsma, S., Jellema, R.H., Overkamp, K.M., van der Werf, M.J. and Smilde, A.K. (2009) Analyzing longitudinal microbial metabolomics data. *J. Proteome Res.*, **8**, 4319–4327.
9. Westerhuis, J.A., Hoefsloot, H.C.J., Smit, S., Vis, D.J., Smilde, A.K., van Velzen, E.J.J., van Duijnhoven, J.P.M. and van Dorsten, F.A. (2008) Assessment of PLS-DA cross validation. *Metabolomics*, **4**, 81–89.
10. Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D.D., Liu, P., Gautam, B., Ly, S., Guo, A.C. *et al.* (2010) SMPDB: The Small Molecule Pathway Database. *Nucleic Acids Res.*, **38**, D480–D487.
11. Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., Young, N., Gautam, B., Hau, D.D., Psychogios, N., Dong, E., Bouatra, S. *et al.* (2009) HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.*, **37**, D603–D610.
12. Hackstadt, A.J. and Hess, A.M. (2009) Filtering for increased power for microarray data analysis. *BMC Bioinformatics*, **10**, 11.
13. Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA*, **107**, 9546–9551.
14. van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. and van der Werf, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, **7**, 142.
15. Dieterle, F., Ross, A., Schlotterbeck, G. and Senn, H. (2006) Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabolomics. *Anal. Chem.*, **78**, 4281–4290.
16. Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
17. Pavlidis, P. (2003) Using ANOVA for gene selection from microarray studies of the nervous system. *Methods*, **31**, 282–289.
18. Smilde, A.K., Jansen, J.J., Hoefsloot, H.C., Lamers, R.J., van der Greef, J. and Timmerman, M.E. (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**, 3043–3048.
19. Tai, Y.C. and Speed, T.P. (2009) On gene ranking using replicated microarray time course data. *Biometrics*, **65**, 40–51.
20. Xia, J., Sinelnikov, I.V. and Wishart, D.S. (2011) MetATT: a web-based metabolomics tool for analyzing time-series and two-factor datasets. *Bioinformatics*, **27**, 2455–2456.
21. Xia, J. and Wishart, D.S. (2010) MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.*, **38**, W71–W77.
22. Xia, J. and Wishart, D.S. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics*, **26**, 2342–2344.
23. Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.
24. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
25. Xia, J. and Wishart, D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protocols*, **6**, 743–760.
26. Xia, J. and Wishart, D.S. (2011) Metabolomic data processing, analysis, and interpretation using MetaboAnalyst. *Curr. Protoc Bioinformatics*, **34**, 14.10.1–14.10.48.
27. Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J.M., Santoyo, J. and Dopazo, J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
28. Reich, M., Liefeld, T., Gould, J., Lerner, J., Tamayo, P. and Mesirov, J.P. (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500–501.
29. Carroll, A.J., Badger, M.R. and Harvey Millar, A. (2010) The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinformatics*, **11**, 376.
30. Kastenmuller, G., Romisch-Margl, W., Wagele, B., Altmaier, E. and Suhre, K. (2011) metaP-server: a web-based metabolomics data analysis tool. *J. Biomed. Biotechnol.*, **2011**, article ID: 839862.
31. Gao, J., Tarcea, V.G., Karnovsky, A., Mirel, B.R., Weymouth, T.E., Beecher, C.W., Cavalcoli, J.D., Athey, B.D., Omenn, G.S., Burant, C.F. *et al.* (2010) Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. *Bioinformatics*, **26**, 971–973.
32. Neuweger, H., Albaum, S.P., Dondrup, M., Persicke, M., Watt, T., Niehaus, K., Stoye, J. and Goesmann, A. (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, **24**, 2726–2732.
33. Wang, T., Shao, K., Chu, Q., Ren, Y., Mu, Y., Qu, L., He, J., Jin, C. and Xia, B. (2009) Automics: an integrated platform for NMR-based metabolomics spectral processing and data analysis. *BMC Bioinformatics*, **10**, 83.
34. Ferry-Dumazet, H., Gil, L., Deborde, C., Moing, A., Bernillon, S., Rolin, D., Nikolski, M., de Daruvar, A. and Jacob, D. (2011) MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. *BMC Plant Biol.*, **11**, 104.