Published online 10 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W71–W77***
doi:10.1093/nar/gkq329

MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data

Jianguo Xia[1] and David S. Wishart[1,2,3,*]

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

Received January 27, 2010; Revised April 2, 2010; Accepted April 17, 2010

ABSTRACT

Gene set enrichment analysis (GSEA) is a widely used technique in transcriptomic data analysis that uses a database of predefined gene sets to rank lists of genes from microarray studies to identify significant and coordinated changes in gene expression data. While GSEA has been playing a significant role in understanding transcriptomic data, no similar tools are currently available for understanding metabolomic data. Here, we introduce a web-based server, called Metabolite Set Enrichment Analysis (MSEA), to help researchers identify and interpret patterns of human or mammalian metabolite concentration changes in a biologically meaningful context. Key to the development of MSEA has been the creation of a library of ~1000 predefined metabolite sets covering various metabolic pathways, disease states, biofluids, and tissue locations. MSEA also supports user-defined or custom metabolite sets for more specialized analysis. MSEA offers three different enrichment analyses for metabolomic studies including overrepresentation analysis (ORA), single sample profiling (SSP) and quantitative enrichment analysis (QEA). ORA requires only a list of compound names, while SSP and QEA require both compound names and compound concentrations. MSEA generates easily understood graphs or tables embedded with hyperlinks to relevant pathway images and disease descriptors. For non-mammalian or more specialized metabolomic studies, MSEA allows users to provide their own metabolite sets for enrichment analysis. The MSEA server also supports conversion between metabolite common names, synonyms, and major database identifiers. MSEA has the potential to help users identify obvious as well as 'subtle but coordinated' changes among a group of related metabolites that may go undetected with conventional approaches. MSEA is freely available at http://www.msea.ca.

INTRODUCTION

Metabolomics is a field of omics science concerned with the comprehensive characterization of small molecule metabolites found in cells, tissues, biofluids, and organisms. It uses a combination of NMR spectroscopy, mass spectrometry, and/or liquid/gas chromatography to specifically identify metabolites or generate metabolic spectral profiles. Because metabolomics is concerned with looking at the small molecule products of gene, protein, and environmental interactions, it provides complementary information to what is normally obtained via genomics, transcriptomics, and proteomics. As a consequence, metabolomics is playing an increasingly important role in both systems biology and synthetic biology (1,2). It is also finding wide applications in diagnostic biomarker discovery, toxicological testing, food and beverage analysis, plant and animal phenotyping as well as drug discovery and development (1–4).

There are two routes to conducting a metabolomics experiment. One is called quantitative (or targeted) metabolomics and the other is called chemometric (or untargeted) metabolomics. In chemometric metabolomics, spectral patterns from two or more large sample sets are processed chemometrically and significant peak differences are identified. The limited number of compounds contributing to these differences is then (ideally) identified. In quantitative metabolomics, large numbers of compounds are first identified and quantified before the data are further processed. In this regard, quantitative metabolomics is more similar to a standard proteomics or transcriptomics experiment. As with any 'omics' experiment, a typical quantitative metabolomic study consists of three stages: data collection, data analysis, and data

*To whom correspondence should be addressed. Tel: +1 780 492 0383; Fax: +1 780 492 1071; Email: david.wishart@ualberta.ca

interpretation. In the data collection stage, sample spectra are first acquired using various analytical platforms (NMR, GC-MS, LC-MS, HPLC). These spectra are then processed by different software tools to facilitate compound identification and quantification, thereby generating metabolite lists. In the second (data analysis) stage, various statistical methods are applied to identify those metabolites that are changed significantly under the given study conditions. Popular methods include *t*-tests, principal component analysis (PCA), partial least square (PLS) discriminant analysis, as well as a variety of other methods. Compounds are first ranked using appropriate importance measures such as *P*-values, loading scores or variable importance in projection scores. A cut-off threshold is then applied to select the top *n* metabolites from the ranked list. In the final stage (data interpretation), the list of significant metabolites is examined, often manually, to see if any biological interesting patterns can be identified. For example, if compounds involved in a certain metabolic pathway appear to be more frequently observed than would be anticipated by random chance, then it may be reasonable to assume that this pathway is biologically or metabolically important.

There are several potential problems associated with this kind of analytical protocol, particularly with the second and third stages. In the second stage, the threshold used for selecting important metabolites is usually chosen arbitrarily. Many moderate but meaningful changes may be missed if an inappropriate threshold is chosen. Indeed, compounds that are critical components of a particular biological process may be left out and the resulting patterns could become indiscernible in the third stage. Choosing a different threshold value may, therefore, lead to a different biological conclusion. In addition, most statistical methods used in compound selection assume that metabolites are being sampled independently, which is certainly not true given the intricate correlations and connections seen among metabolites in metabolic networks. In the third stage, despite obvious differences in importance measures, all selected compounds are treated equally, as only their occurrences (and not their concentrations) are considered in the analysis. This loss of information can obviously reduce the accuracy of any subsequent interpretations. In addition, since the identification of metabolically meaningful patterns is usually performed manually, this process can be very time consuming. Likewise, the final interpretation is often subject to the background knowledge or biases of individual researchers.

These data analysis/interpretation issues are not unique to metabolomics. In fact, they have long been the subject of intense debate among researchers involved in gene expression data analysis (5–8). This debate has inspired the development of a new approach for gene expression analysis, generally referred to as gene set enrichment analysis (GSEA). The key idea behind GSEA is to directly investigate the enrichment of pre-defined groups of functionally related genes (or gene sets) instead of individual genes. This group-based approach does not require pre-selection of genes with an arbitrary threshold. Instead, functionally related genes are evaluated together as gene sets, allowing additional biological information to be incorporated into the analysis process. The GSEA approach has proven to be remarkably successful in deriving new information from genome-wide expression studies, having been cited 1000′s of times since its initial description in 2005 (9). Its success has also inspired many extensions, improvements and variations (8–16).

However, to our knowledge, no tools similar to GSEA have been developed to support this group-based approach for metabolomic data analysis. This is likely because both enrichment analysis and quantitative metabolomics are relatively new techniques. However, it is also likely due to the fact that in order to use this approach, one needs an extensive and biologically meaningful metabolite set library. Such a library is very laborious and time consuming to create. To address this issue, we collected, both through text-mining and manual curation, a large body of mammalian (primarily human) metabolite and metabolic pathway information from the literature and various public databases. Using this library of metabolites and disease/pathway/tissue associations, we have implemented a web-based application, named MSEA, to support group-based enrichment analysis for human and/or mammalian metabolomic studies. The main features of MSEA include the following.

(i) A collection of five metabolite set libraries containing ~1000 biologically meaningful groups of metabolites.
(ii) Three enrichment analysis methods – overrepresentation analysis (ORA), single sample profiling (SPP), and quantitative enrichment analysis (QEA), to support common data forms generated in metabolomic studies;
(iii) Support for enrichment analysis with discrete and continuous phenotypes.
(iv) Support for enrichment analysis using customized (non-mammalian) metabolite sets.
(v) Support for conversions between metabolite common names, synonyms, and identifiers (ID) of nine major metabolomic databases.
(vi) Comprehensive analysis report generation.

In other words, with MSEA and its accompanying databases it is possible to take a list of altered metabolites from a biofluid or tissue sample and use it to suggest a biological pathway or disease condition that can be further investigated. The MSEA server and all of its accompanying databases are freely available at http://www.msea.ca.

METHODS

Creation of metabolite set libraries

A group of metabolites are considered to constitute a metabolite set if they are known to be: (1) involved in the same biological processes (i.e., metabolic pathways, signaling pathways); (2) changed significantly under the same pathological conditions (i.e., various metabolic diseases); and (3) present in the same locations such as organs, tissues, or cellular organelles. These data were collected through manual curation from books and journals as well as through text mining of public databases. The

Table 1. Overview of MSEA's metabolite set libraries

Category	Total number	Sources	Web links
Pathway based	84	SMPDB	http://www.smpdb.ca
Disease—associated	851		
Blood[a]	398	HMDB	http://www.hmdb.ca/disease_browse
Urine[a]	335	MIC	http://www.metagene.de
CSF[a]	118	PubMed	http://www.ncbi.nlm.nih.gov/pubmed
Location based	57	HMDB	http://www.hmdb.ca/

[a]Metabolite sets were collected from multiple sources including HMDB, MIC, PubMed and SMPDB.

resulting metabolite sets were manually validated/edited and then further organized into three categories: pathway associated, disease associated, and location based. MSEA's pathway-associated metabolite library contains 84 entries based on the 84 human metabolic pathways found in the Small Molecular Pathway Database (SMPDB) (17). MSEA's disease-associated metabolite sets were mainly collected from the literature. Metabolites associated with different diseases were manually identified, merged and subsequently refined by reading the original publications listed in the Human Metabolome Database (HMDB) (18), the Metabolic Information Center (MIC), and SMPDB. Using these resources, a total of 851 physiologically informative metabolite sets were created. These disease-associated metabolite sets were further divided into three subcategories based on the biofluids in which they were measured: 398 metabolite sets in blood; 335 in urine; and 118 in cerebral–spinal fluid (CSF). MSEA's location-based library contains 57 metabolite sets based on the 'Cellular Location' and 'Tissue Location' listed in the HMDB. A summary of these metabolite set libraries is shown in Table 1.

Creation of a metabolite dictionary and concentration database

In order for the MSEA server to accept a range of metabolite names, synonyms or ID as input, it was also necessary to develop a local metabolite dictionary that could be used to perform facile name conversion or 'normalization'. Information contained in the HMDB was used to extract common names, synonyms, as well as ID used in nine major metabolomic databases [HMDB, PubChem (19), ChEBI (20), KEGG (21), BiGG (22), METLIN (23), BioCyc (24), Reactome (25), and Wikipedia]. Examples of MSEA's supported IDs are listed in Table 2. In order for MSEA to perform single sample profiling (SSP) analysis, it was also critical to obtain reference concentrations for as many metabolites as possible. These concentration data were collected primarily from the HMDB with additional values being added through manual curation. MSEA's reference concentrations are organized based on the biofluids in which they were measured. Concentrations are presented in the form of *mean (minimum – maximum)*. For concentrations reported as mean and standard deviation (SD), their 95% confidence intervals (mean ± 2 SD) were used to define the concentration ranges. One compound may have multiple concentration values as reported from different studies.

Table 2. Overview of compound labels currently supported by MSEA

Label type	Examples
Common Name	Adenosine, acetic acid, adenine, creatine
HMDB	HMDB00050, HMDB00042, HMDB00034, HMDB0006
PubChem	60961, 176, 190, 586
ChEBI	16335, 15366, 16708, 16919
KEGG	C00212, C00033, C00147, C00300
BiGG	34273, 33590, 34039, 34543
METLIN	86, 3206, 85, 7
BioCyc	ADENOSINE, ACET, ADENINE, CREATINE
Reactome	114933, 114747, 114936, 114818
Wikipedia	Adenosine, acetic acid, adenine, creatine

Implementation of enrichment analysis programs

Over the past 5 years, many different algorithms have been developed for group-based enrichment analysis, including GSEA (9), GSEA-P (26), PAGE (27), globaltest (11), SAFE (12), SAM-GS (13) and GSA (14). Based on a thorough review of the literature, we decided to adopt the *globaltest* algorithm as the backend for MSEA. There were three main reasons: (1) recent publications have indicated that *globaltest* exhibited similar or superior performance when tested against several other algorithms (28–30); (2) *globaltest* is very flexible and supports binary, multiclass, and continuous phenotype labels; and (3) *globaltest* is computationally efficient as the *P*-values can be calculated based on the *Q*-stat's asymptotic distribution, which appears to work well with both large and small sample sizes. The *globaltest* algorithm was originally designed for testing associations between gene sets and clinical outcomes (11). It uses a generalized linear model to compute a '*Q*-stat' for each gene set. For a group of m genes, the *Q*-stat is calculated as the average of the Q values ($Q_1 \ldots Q_m$) calculated for the m single genes, where Q_i is the average of the squared covariance between the gene expression pattern and the clinical outcome. Conventional ORA was implemented based on a cumulative hypergeometric distribution. Since many metabolite sets are tested simultaneously, we also implemented methods to adjust for the multiple testing problems that occur during enrichment analysis.

In addition to the original *P*-values, MSEA also reports *Bonferroni* corrected *P*-values and a false discovery rate (FDR) according to Benjamini and Hochberg (31).

Web server characteristics

MSEA's web interface was implemented using the JSF or Java Server Faces (http://java.sun.com/javaee/javaserverfaces) framework. The enrichment analysis algorithms were implemented in the R (version 2.10.0) programming language (http://www.**r**-project.org/). The communication between R and Java was established through the *Rserve* TCP/IP server (http://www.rforge.net/Rserve/). The web application is hosted on GlassFish (version 3) using a Linux operating system (Fedora Core 10). MSEA's host server is equipped with two Intel Quad Core 2 processors (3.0 GHz each) and 8 GB of physical memory. The web application is platform independent and has been tested successfully on Internet Explorer 8.0, Mozilla Firefox 3.0, and Safari 4.0.

PROGRAM DESCRIPTION

MSEA's workflow is illustrated in Figure 1. Briefly, MSEA can be described in four steps – data input, data processing, data analysis, and data download. In addition to its analysis utilities, users can directly download, browse or search MSEA's metabolite set libraries, or perform compound name and ID conversions. The details of each step are discussed below.

Step 1. Data input

MSEA accepts data in three different formats: (i) a list of compound names entered in a single-column format; (ii) a list of compound concentrations entered as two-column data with the first column corresponding to the compound names/labels and the second corresponding to the concentration values; or (iii) a concentration table containing metabolite concentration data from multiple samples. The table must contain comma-separated values (*.csv*) with rows for samples and columns for metabolites. The second column of the table is reserved for phenotype labels (binary, multiclass, or continuous). Examples of these input formats are provided on the MSEA homepage.

Step 2. Data processing

In this step, both the compound labels and the concentration values are examined for their suitability for downstream analysis. It is critical that the compound labels be recognized by the program in order to be compared with MSEA's collection of compound names in metabolite sets. Therefore, a consistency check is done with the input names or IDs against the names and IDs stored in MSEA's metabolite dictionary. Any nomenclature inconsistency is flagged and displayed to users for manual inspection and correction. For SSP, the concentrations must be provided in a standard concentration unit (μmol for blood and CSF and μmol $mmol^{-1}$_creatinine for urine) in order for the input data to be properly compared with MSEA's reference concentrations database. For QEA, the concentration values can be normalized and negative/missing values are allowed. Two widely used chemometric methods—PCA and PLS analysis—are available in MSEA to allow for data visualization, pattern identification, and outlier detection. Note that MSEA does not perform data normalization. Users are advised to visit MSEA's companion web site MetaboAnalyst (32) to access a variety of data processing and normalization options.

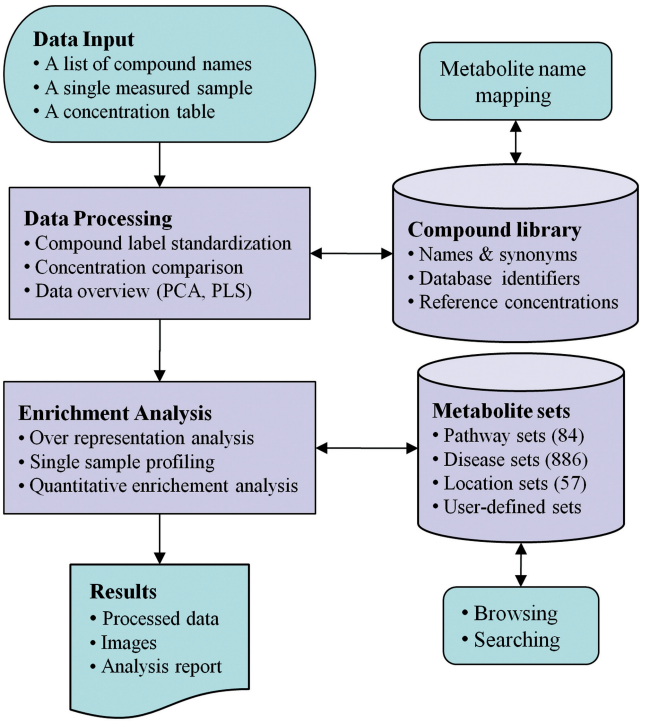

Figure 1. MSEA workflow. MSEA consists of four steps—data input, data processing, data analysis, and data download. Different analysis procedures are performed for different input types. MSEA allows users to directly browse and search its metabolite set libraries as well as to perform metabolite name mapping between different names and database ID.

Step 3. Enrichment analysis

Depending on the type of user input, MSEA offers three kinds of enrichment analysis: ORA; SSP, and QEA. These analysis modules are described in more detail below.

ORA. ORA is used to evaluate whether a particular set of metabolites is represented more than expected by chance within a given compound list. ORA is performed when the user provides only a list of compound names. Such a list can be obtained using standard feature selection methods that statistically rank all the compounds and select those scoring above a certain threshold. ORA is also very useful for analyzing a group of compounds exhibiting similar concentration changes or patterns. Such a list can be obtained from standard clustering analysis. Many commonly used feature selection and feature clustering methods are available from our companion web application MetaboAnalyst (32). The *P*-value from ORA indicates the probability of seeing at least a particular

number of metabolites from a certain metabolite set in a given compound list. The *Bonferroni* corrected *P*-value and FDR are also presented to account for problems associated with multiple comparisons. Users can click the 'View' link in the Details column of any of MSEA's metabolite sets to see all its constituent metabolites with matched ones highlighted in red, as well as pathway images (when available).

SSP. For common human biofluids such as blood, urine, or CSF, normal concentration ranges are known for many metabolites. In clinical metabolomic studies, it is often desirable to know whether certain metabolite concentrations in a given sample are significantly higher or lower than their normal ranges. MSEA's SSP module is designed to provide this kind of analysis. In particular, SSP is performed when the user provides a two-column list of both compounds and concentrations. When called, the SSP module will compare the measured concentration values of each compound to its recorded normal reference ranges of the corresponding biofluid (Figure 2A). By default, only compounds with concentrations above or below *all* the reported normal ranges will be selected for further investigation. Users can manually select or deselect compounds to override this default selection by inspecting the concentration comparison plots generated by this module (Figure 2B).

QEA. QEA is performed when the user uploads a concentration table containing metabolite concentration data from multiple samples. QEA is based on the *globaltest* algorithm to perform enrichment analysis directly from raw concentration data and does not require a list of significantly changed compounds. With QEA, enriched metabolite sets can be identified when only a few compounds are significantly changed or when many compounds are only slightly (but consistently) changed. The QEA algorithm uses a generalized linear model to estimate a '*Q*-stat' for each metabolite set. The *Q*-stat describes the correlation between compound concentration profiles, X, and phenotype labels, Y. In addition to the *Q*-stat values, the QEA module also provide *P*-values, Bonferroni-corrected *P*-values, and estimates of FDR. Figure 2C shows a screenshot of the output table from a typical QAE. Users can click the image icon of any matched metabolite set to view a detailed graphical summary of the contributions of individual metabolites (Figure 2D).

Step 4. Data download

When users finish an enrichment analysis, a comprehensive report is generated with detailed descriptions of each step performed, embedded with graphical and tabular results. The processed data, images, R scripts, as well as the R command history are also available for download. Users familiar with R can easily reproduce the results on their local machine after installing the R packages and the corresponding metabolite set libraries (available on the Resources Download page).

Other features

The MSEA web server also offers a number of other features to facilitate metabolomic data analysis, including (1) a compound name and ID mapping tool; (2) a browser for metabolite sets; and (3) a facility for custom metabolite set uploads. Given the fact that no consensus exists in labeling compounds in current metabolomic studies, we implemented a utility in MSEA to convert between common compound names, synonyms and the ID codes used in nine major metabolite databases (see Table 2 for details). This converter can also deal with spelling errors using an approximate text matching algorithm. In addition to this name/ID converter, MSEA also provides a browser to view MSEA's collection of metabolite set libraries. These libraries can provide a valuable source of information to investigate the biological implications of any metabolite sets identified after enrichment analysis. The browser implemented in the MSEA web server allows users to easily scan and search its metabolite set libraries. Each entry contains the metabolite set name, its constituent compounds, and links to original references. Given the incompleteness of MSEA's metabolite-set libraries, researchers may want to perform enrichment analysis using customized or self-defined metabolite sets other than the ones provided by the server. MSEA supports this option by allowing users to upload their own metabolite set library. The library file should be in a simple.*csv* file with the first column for metabolite set names and the second for compound members.

Limitations

Unlike genomics or transcriptomics, metabolomics has not yet achieved total metabolite coverage. Whereas Next-Gen DNA sequencers and modern microarrays routinely cover entire genomes, most metabolomic technologies only offer 5–10% coverage of a sample's metabolome (1–2). This makes many metabolomic studies intrinsically biased. Since most of the metabolite sets in MSEA's libraries are also derived from experimental studies, they tend to suffer the same sampling bias. Fortunately, these biases tend to cancel each other out, as essentially the same metabolite population (the fraction of the metabolome that are 'detectable' by current analytical technologies) is probed to generate both metabolite sets and user data. Nevertheless, users should always take note of their experimental conditions or technological limitations when interpreting the results from enrichment analysis.

Another key limitation to MSEA is its bias to human and/or mammalian metabolomics. This is a limitation that we are working to overcome through the addition of other metabolite sets from plants and microbes. However, until these databases and data sets can be completed (likely in two years time) we would encourage researchers who are engaged in metabolomic studies of non-mammalian species to create their own customized metabolite sets for enrichment analysis and to contribute these sets to the MSEA server for public use.

Figure 2. Enrichment analysis and visualization. Results from MSEA's enrichment analysis are presented both in tables as well as through graphical summaries. (**A**) The comparison between the measured concentrations and reference concentrations using the SSP module. The top part of (**B**) shows a graphical summary of the concentration comparison for a single compound when users click an image icon in Figure 2A. The bottom part of Figure 2B shows all the corresponding publications that reported these concentrations. (**C**) The results generated by the QEA module. The top part of (**D**) is a metabolite-set plot indicating the influence of an individual compound on each of the selected metabolite sets. The bottom part of Figure 2D shows all its constituent metabolites with matched ones highlighted in red.

CONCLUSIONS

Over the past few years, a number of software tools have been developed to address the bioinformatic needs of metabolomics. However, most of these programs were designed for spectral data processing and compound identification. More recently, several freely available software tools for the statistical analysis of metabolomic data have started to appear, such as MetaboAnalyst (32) and MeltDB (33). As yet, no publicly available tools have been made available to assist in the functional or biological interpretation of metabolomic data. To address this issue, we have developed a web server, named MSEA, designed to help researchers identify and interpret patterns of metabolite concentration changes in a biologically meaningful context. MSEA performs three kinds of enrichment analysis including ORA, SSP, and QEA. When only a list of compounds is available ORA is performed. When both compound names and concentrations are available the SSP module is called. When concentration data are available from multiple samples, MSEA performs QEA. The enrichment analyses performed by MSEA are based on five carefully compiled metabolite libraries consisting of ~1000 entries. In addition to its enrichment analysis capabilities, MSEA allows custom metabolite sets to be uploaded for more specialized (non-mammalian) studies. MSEA also supports conversion between metabolite common names, synonyms and major database ID. We believe that, over time, the MSEA approach will become more powerful as analytical technologies for metabolomics continue to improve their metabolite coverage and as the

metabolomics community develops improved standards and ontologies (34). In the long run, we would like to turn the MSEA server into a resource for metabolomic annotation, visualization, and integrated discovery much as the DAVID server (35) has become just such a resource for microarray data analysis.

FUNDING

Alberta Ingenuity Fund, the Alberta Advanced Education and Technology; the Canadian Institutes for Health Research; Genome Alberta, a division of Genome Canada. Funding for open access charge: Canadian Institutes for Health Research

Conflict of interest statement. None declared.

REFERENCES

1. Wishart,D.S. (2008) Quantitative metabolomics using NMR. *Trends Anal. Chem.*, **27**, 228–237.
2. Hollywood,K., Brison,D.R. and Goodacre,R. (2006) Metabolomics: current technologies and future trends. *Proteomics*, **6**, 4716–4723.
3. 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.
4. Gieger,C., Geistlinger,L., Altmaier,E., Hrabe de Angelis,M., Kronenberg,F., Meitinger,T., Mewes,H.W., Wichmann,H.E., Weinberger,K.M., Adamski,J. *et al.* (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS Genet.*, **4**, e1000282.
5. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
6. Pan,K.H., Lih,C.J. and Cohen,S.N. (2005) Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. *Proc. Natl Acad. Sci. USA*, **102**, 8961–8965.
7. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
8. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
9. 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.
10. Lee,H.K., Braynen,W., Keshav,K. and Pavlidis,P. (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.
11. 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.
12. Barry,W.T., Nobel,A.B. and Wright,F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.
13. Dinu,I., Potter,J.D., Mueller,T., Liu,Q., Adewale,A.J., Jhangri,G.S., Einecke,G., Famulski,K.S., Halloran,P. and Yasui,Y. (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242.
14. Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**.
15. Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Muller,R., Meese,E. and Lenhof,H.P. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.
16. Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.
17. 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.
18. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D., Sawhney,S. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Res.*, **35**, D521–D526.
19. Austin,C.P., Brady,L.S., Insel,T.R. and Collins,F.S. (2004) NIH Molecular Libraries Initiative. *Science*, **306**, 1138–1139.
20. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcantara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
21. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–280.
22. Feist,A.M., Herrgard,M.J., Thiele,I., Reed,J.L. and Palsson,B.O. (2009) Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.*, **7**, 129–143.
23. Smith,C.A., O'Maille,G., Want,E.J., Qin,C., Trauger,S.A., Brandon,T.R., Custodio,D.E., Abagyan,R. and Siuzdak,G. (2005) METLIN – A metabolite mass spectral database. *Ther. Drug Monit.*, **27**, 747–751.
24. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.*, **33**, 6083–6089.
25. Matthews,L., Gopinath,G., Gillespie,M., Caudy,M., Croft,D., de Bono,B., Garapati,P., Hemish,J., Hermjakob,H., Jassal,B. *et al.* (2009) Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, D619–D622.
26. Subramanian,A., Kuehn,H., Gould,J., Tamayo,P. and Mesirov,J.P. (2007) GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics*, **23**, 3251–3253.
27. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
28. Hulsegge,I., Kommadath,A. and Smits,M.A. (2009) Globaltest and GOEAST: two different approaches for Gene Ontology analysis. *BMC Proc.*, **3(Suppl 4)**, S10.
29. Song,S. and Black,M.A. (2008) Microarray-based gene set analysis: a comparison of current methods. *BMC Bioinformatics*, **9**, 502.
30. Liu,Q., Dinu,I., Adewale,A.J., Potter,J.D. and Yasui,Y. (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.
31. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*, **57**, 289–300.
32. 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.
33. 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.
34. Sansone,S.A., Fan,T., Goodacre,R., Griffin,J.L., Hardy,N.W., Kaddurah-Daouk,R., Kristal,B.S., Lindon,J., Mendes,P., Morrison,N. *et al.* (2007) The metabolomics standards initiative. *Nat. Biotechnol.*, **25**, 846–848.
35. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, P3.