BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 8 2011, pages 1108–1112
doi:10.1093/bioinformatics/btr079

Systems biology

Advance Access publication February 16, 2011

Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets

Marie Brown[1], David C. Wedge[2], Royston Goodacre[2,3], Douglas B. Kell[2], Philip N. Baker[4], Louise C. Kenny[5], Mamas A. Mamas[1,6], Ludwig Neyses[1,6] and Warwick B. Dunn[1,2,3,7,*]

[1]School of Biomedicine, The University of Manchester, Manchester M13 9PT, [2]School of Chemistry, [3]Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, Manchester M1 7DN, UK, [4]Department of Obstetrics and Gynecology, Faculty of Medicine and Dentistry, University of Alberta, 2J2.01 WMC, Edmonton AB T6G 2R7, Canada, [5]The Anu Research Centre, Department of Obstetrics and Gynaecology, University College Cork, Cork University Maternity Hospital, Cork, Ireland, [6]Manchester Heart Centre, Central Manchester University Hospitals NHS Foundation Trust, Manchester Royal Infirmary and [7]Centre for Advanced Discovery and Experimental Therapeutics, York Place (off Oxford Road), Central Manchester University Hospitals NHS Foundation Trust, Manchester M13 9WL, UK

Associate Editor: John Quackenbush

ABSTRACT

Motivation: The study of metabolites (metabolomics) is increasingly being applied to investigate microbial, plant, environmental and mammalian systems. One of the limiting factors is that of chemically identifying metabolites from mass spectrometric signals present in complex datasets.

Results: Three workflows have been developed to allow for the rapid, automated and high-throughput annotation and putative metabolite identification of electrospray LC-MS-derived metabolomic datasets. The collection of workflows are defined as PUTMEDID_LCMS and perform feature annotation, matching of accurate *m/z* to the accurate mass of neutral molecules and associated molecular formula and matching of the molecular formulae to a reference file of metabolites. The software is independent of the instrument and data pre-processing applied. The number of false positives is reduced by eliminating the inaccurate matching of many artifact, isotope, multiply charged and complex adduct peaks through complex interrogation of experimental data.

Availability: The workflows, standard operating procedure and further information are publicly available at http://www.mcisb.org/resources/putmedid.html.

Contact: warwick.dunn@manchester.ac.uk

Received on November 5, 2010; revised on February 4, 2011; accepted on February 7, 2011

1 INTRODUCTION

Systems biology is applied to study the components of, and more importantly their complex interactions in, biological systems. One set of components which are studied in systems biology investigations are metabolites, either by targeted or holistic profiling experimental strategies (Dunn *et al*., 2011). Low molecular weight inorganic and organic metabolites play important roles in the operation and maintenance of biological systems. The study of metabolites (metabolomics) is increasingly being applied to investigate microbial (Bradley *et al*., 2009; MacKenzie *et al*., 2008; Mashego *et al*., 2007), plant (Allwood *et al*., 2008; Fernie and Schauer, 2009; Hall *et al*., 2008), environmental (Bundy *et al*., 2009; Viant *et al*., 2006) and mammalian (Griffin, 2008; Kenny *et al*., 2010; Lewis *et al*., 2008) systems. Many studies follow a hypothesis generating or inductive strategy (Kell and Oliver, 2004) and start from a small and known subset of biological knowledge. Valid experiments are designed to acquire robust and reproducible data on a wide range of different metabolites and metabolite classes from carefully selected samples. Subsequent data analysis procedures define the metabolic differences associated with a biological change related to genotype, biological perturbation or environmental intervention (for example, drug therapy). These studies employ a metabolic profiling strategy to detect a wide range of (but not all) metabolites from numerous biochemical classes to obtain maximum metabolic information rapidly. This strategy provides the detection of hundreds or thousands of metabolites. However, due to the diverse range of chemical and physical properties and the wide concentration range of metabolites within the metabolome, no single analytical technology can provide the non-biased quantitative detection of all metabolites in a biological system (Dunn, 2008). Metabolic profiling provides semi-quantitative data, typically as chromatographic peak areas, rather than absolute quantitation where metabolite concentrations would be reported. Mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) have been widely employed and provide complementary roles (Dunn *et al*., 2005, 2011). Chromatography-MS techniques provide advantages for these highly complex biological samples and include gas chromatography (GC-MS), liquid chromatography (LC-MS) and LC derivatives including ultra performance liquid chromatography (UPLC-MS). Capillary Electrophoresis-MS (Soga *et al*., 2003) and LC-MS apply electrospray ionization. Direct infusion mass spectrometry (DIMS) can also be applied, though lacks the chromatographic separation of

*To whom correspondence should be addressed.

metabolites (Dunn *et al*., 2005; Southam *et al*., 2007). Hyphenated platforms can provide (with suitable operation and mass calibration) high separation resolution, high mass resolution and mass accuracy, typical limits of detection of micromol per litre and the ability to identify metabolites through a combination of Retention Time (RT)/index, accurate mass and gas-phase fragmentation-derived mass spectra. Each of these platforms, whether hyphenated or non-hyphenated, provide different advantages and disadvantages for metabolite identification as has been reviewed previously (Dunn *et al*., 2011). The increasing use of high mass resolution LC-MS and UPLC-MS platforms provides the detection of many thousands of features [see Brown *et al*. (2009) for a comparison of the features detected related to sample type] with high mass accuracy and has led to a need to develop data handling methods for the conversion of this raw analytical data into biological knowledge.

One of the data processing procedures essential in metabolic profiling is metabolite identification which has been reviewed previously (Dunn *et al*., 2011; Wishart, 2009). Guidelines have been provided by the Metabolomics Standards Initiative to define how the different levels of metabolite identification can be reported (Sumner *et al*., 2007). A range of approaches can be applied for metabolite identification. Two generalized types of identification are achievable: putative identification and definitive identification.

Putative identification usually employs one or more molecular properties for identification, but does not compare these to the same properties of an authentic chemical standard as is performed for definitive identification. The accurate mass (or m/z) of an analyte and its associated isotopologues can be used to define molecular formulae (MFs) from which suitable metabolites can be derived by searching a range of electronic resources [e.g. PubChem (http://pubchem.ncbi.nlm.nih.gov/), HMDB (http://www.hmdb.ca/), KEGG (http://www.genome.jp/kegg/) and MMD (Brown *et al*., 2009)] and has been previously shown (Brown *et al*., 2005; Junot *et al*., 2010; Lane *et al*., 2008; Rogers *et al*., 2009). Direct matching of accurate mass (or m/z) to data in electronic resources without intermediate matching to MF can also be performed. However, structural isomers and stereoisomers have the same accurate mass and therefore require a separate, orthogonal property for identification of all potential isomers. Typically, this is chromatographic separation though separation of isomers is not always achievable. Separation of enantiomers requires a chiral chromatography column.

Definitive identification employs at least two properties (typically RT or index and fragmentation mass spectrum) and compares these properties to an authentic chemical standard analysed under identical analytical conditions. In LC-MS and UPLC-MS applications, the accurate masses of the detected ions is employed in combination with other rules (e.g. isotope ratio of ^{12}C and ^{13}C isotopic peaks to define the number of carbon atoms present in the MF; calculated as peak area ^{13}C isotopologue/peak area ^{12}C isotopologue) to generate MF and thus provide putative metabolite identification(s). Specific rules are not always applicable. For example, ^{13}C/^{12}C isotopic peaks can only be applied on instruments where accurate isotopic ratios are detected and where ^{13}C-artificially labelled metabolites have not been applied in the biological experiment. Fragmentation mass spectra (MS/MS or MS^n) are then used to provide increased confidence through comparison to authentic chemical standards or to *in silico*-derived fragmentation mass spectra to give an unequivocal metabolite identification (Wolf *et al*., 2010). It should be noted that not all authentic chemical standards are commercially available and that MS/MS and MS^n fragmentation mass spectra are not always accurate in unequivocal identification of two isomeric metabolites.

Studies by the authors have assessed the level of complexity of electrospray UPLC-MS data derived from biological extracts in a metabolic profiling strategy (Brown *et al*., 2009). This work has shown that a multitude of different ion types are observed including commonly described ions (e.g. protonated, deprotonated, sodium or potassium adducts and ^{13}C isotope). However, many other unexpected types of ions including adducts (e.g. complex combinations of sodium chloride and formate dependant on the matrix type and mobile phase), fragments, dimers, multiply charged and instrument specific ions (e.g. Fourier Transform (FT) artifact peaks) are also detected. Each different ion type is defined as a feature, whose accurate mass (or m/z) is unique but whose RT and chromatographic peak profiles are identical. Information on the type of ion should be applied in metabolite identification (Brown *et al*., 2009; Draper *et al*., 2009).

Automated software or workflows for high-throughput identification of large metabolomic datasets are not freely available. Currently, metabolite identification is a manual or semi-automated process assessing those features of biological interest and not the complete set of detected features (Dunn, 2008). For metabolomics to be successful it is essential to derive biological knowledge from analytical data, a view emphasized by a recent Metabolomics ASMS Workshop Survey 2009 which found that the biggest bottlenecks in metabolomics were thought to be identification of metabolites and assigning of biological interest (http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics-Survey-2009).

To fill the gap in requirements, three workflows have been written to perform for the first time integrated, automated and high-throughput annotation and putative metabolite identification of electrospray LC-MS and UPLC-MS metabolomic datasets in a freely available package. The workflows were developed in the Taverna Workflow Management System to provide flexibility in their operation and the ability to rapidly and simply integrate with web services and other Taverna workflows in the future [for example, see Li *et al*. (2008)], so as to provide integrated data analysis and bioinformatics or cheminformatics packages. Examples are available on myExperiment, a repository of workflows freely available to the scientific community (http://www.myexperiment.org/), including a workflow to perform data pre-processing with XCMS and a workflow to perform *in silico* fragmentation applying MetFrag. The achievement of this level of integration would be more technically demanding and would require significantly greater expertise and time if coded in many other programming languages. Taverna is also easy to operate for relative novices with minimal training as the process involves defining parameter values and files only.

2 METHODS AND IMPLEMENTATION

The current lack of freely available workflows or software to process deconvoluted data acquired from electrospray LC-MS experiments led the authors to develop three workflows. The workflows have been developed in Taverna (Hull *et al*., 2006) using Beanshell, a Java scripting language, which is enabled in Taverna and can perform data manipulation, parsing and formatting. Taverna can be downloaded from http://www.taverna.org.uk. The workflows were developed under Windows using Taverna v1.7.0

and subsequently tested using Taverna Workbench 2.2.0. In combination, the workflows perform the automated, high-throughput annotation and putative metabolite identification of electrospray LC-MS and UPLC-MS metabolomic datasets. The software has been coded as a series of separate workflows to allow more flexibility in the analysis of data by obviating the need to re-run the whole pipeline when altering one of the workflow parameters, such as the mass tolerance or database. This approach also reduces the likelihood of memory problems when handling large datasets on computers with small RAM.

The workflows, related files and standard operating procedure (SOP) are available to the user community on http://www.mcisb.org/resources/putmedid.html and will also be placed on MyExperiment (http://www.myexperiment.org/). In general, the input and output files are tab-delimited (*.txt) files and are sorted by ascending accurate mass or MF as appropriate (ordered as C, H, N, O, P, S, Br, Cl, F, Si in ascending alphanumeric form as is standard for PubChem). Internal checks are made within the workflows to ensure that the number of features in both peak and data files match (workflow for correlation analysis) and that the study and reference input files are sorted either by accurate *m/z* (workflow for metabolic feature annotation) or MF (workflow for metabolite annotation). Termination of the process and reporting of an informative error message occurs if this is not the case. The three workflows are described in detail in the available SOP and briefly below.

2.1 Workflow for correlation analysis

The workflow for correlation analysis (List_CorrData) allows the user to calculate either Pearson or Spearman rank correlations or read in previously calculated correlation data. The correlation calculations allow for NaN, Inf and 0 in the input data and are equivalent to using the Matlab (http://www.mathworks.co.uk/) corr function with the following parameters:

corr(Xdata, 'rows', 'pairwise', 'type', 'Pearson') or
corr(Xdata, 'rows', 'pairwise', 'type', 'Spearman')

2.2 Workflow for metabolic feature annotation

The workflow for metabolic feature annotation (annotate_MassMatch) uses correlation coefficient information calculated in the workflow for correlation analysis, accurate *m/z* difference, RT and median peak area data to group together and annotate features with the type of ion (isotope, adduct, dimer, others) originating from the same metabolite. The same metabolite can be detected as different ion types each with different *m/z*. Following annotation, the experimentally determined accurate *m/z* are matched to the accurate *m/z* of unique MF in a reference file within a specified *m/z* tolerance.

2.3 Workflow for metabolite annotation

In the workflow for metabolite annotation (matchMMF_MF), the MF from the output file calculated in the workflow for metabolic feature annotation is matched to the MF from the Reference file of metabolites (trimMMD_sortMF.txt or other appropriate reference file). The metabolite information for all matched MFs is added to the input data and output data are generated in three formats, each of which can be saved as tab-delimited files by the user.

3 RESULTS

An assessment has been made of the workflows' ability to perform putative metabolite identification using two reference files: (i) a listing of unique accurate mass/MF data from PubChem using specific elements (C, H, N, O, P, S, Br, Cl, F, Si only; file downloaded from http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/) selected to give a wider selection of MFs and (ii) The Manchester Metabolomics Database (Brown *et al*., 2009) constructed with data from genome-scale metabolic reconstructions, HMDB, KEGG, LIPIDMAPS, BioCyc and DrugBank. Data from all these sources are included to provide a comprehensive set of metabolites. For example, HMDB does not contain all lipids that are theoretically present in human biofluids and tissues and therefore inclusion of data from LIPIDMAPS provides greater complementary metabolite coverage.

Table 1. Distribution of annotated peaks in negative and positive ion mode

Features summary	Negative ion mode	Positive ion mode
No. of features	2173	4348
No. of correlations (>0.7, RT ± 5 s)	11 867	50 867
Invalid RT (40 s < RT > 1200 s)	224	487
FT artifact peaks	61	66
Isotopes (^{13}C, ^{34}S, ^{37}Cl)	455	1170
Multiply charged ions	22	444
Salt ions (not adducted to metabolites)	40	31
Total no. of excluded features	802 (36.9%)	2198 (50.6%)
No. of features remaining for identification	1371 (63.1% of all detected features)	2150 (49.4% of all detected features)

A clinical dataset of fasting blood serum samples were taken from participants according to ethical guidelines and stored before being analysed with quality control samples in a random order and within 48 h of reconstitution using an UPLC (Waters UPLC Acquity, Elstree, UK) coupled on-line to an electrospray LTQ-Orbitrap hybrid mass spectrometer (ThermoFisher Scientific, Bremen, Germany). The collection and storage of serum samples and the UPLC and mass spectrometer methods applied have been previously described (Dunn *et al*., 2008; Zelena *et al*., 2009). Independent samples (118 in total) were analysed in both positive and negative ion mode. Raw data files (.RAW) were converted to the NetCDF format using the File converter program in XCalibur (ThermoFisher Scientific, Bremen, Germany). Deconvolution of data was performed using XCMS, running on R version 2.6.0, an open-source deconvolution program available for LC-MS data (Smith *et al*., 2006) using identical settings to those reported previously (Dunn *et al*., 2008). This produced a list of features with associated RT, accurate *m/z* and chromatographic peak area. For these data, the mass accuracy was assessed using a set of 35 and 50 metabolites commonly detected in serum and plasma in positive and negative ion modes, respectively. Shown in Table 1 is the distribution of annotated features found in the dataset and excluded from further mass matching.

In positive ion mode >50% of features were marked for exclusion from further metabolite identification, which was considerably higher than in negative ion mode (36.9%) due in part to the much greater occurrence of multiply charged ions (~10% of all features). It should be noted that multiply charged ions can be peptides, proteins or high molecular weight metabolites capable of carrying multiple charges. Approximately 25% of these excluded features are isotopic peaks. These are annotated and linked to the related molecular ion. In the workflow for metabolite annotation, all isotope and FT artifact peaks are labelled with the accurate identification observed for the molecular or adduct ions.

Using a mass tolerance of 3 p.p.m., ~6% of the remaining features were not matched to any unique MFs in the PubChem reference file (301507 entries). Seventy-six percent in negative and 60% in positive ion mode of the remaining features (791 and 1292 features) were fully annotated and given putative metabolite identification using a revised version of the MMD database (31648 entries). The reference data file is based on molecules and parent compounds carrying no charge and is derived from the MMD which contains an array of information from a wide variety of electronic resources. The MMD data file was revised by removal (or modification) of charged species of salts e.g. sodium ascorbate, calcium citrate, metamphetamine hydrochloride. Obvious duplicates of data were removed and for many common metabolites e.g. amino acids and sugars only a single stereochemical form of the compound was retained. The included form usually related to the one most well described in HMDB, and if not present in HMDB then as described in KEGG, and if not present in HMDB and KEGG then as described in LIPIDMAPS and resulted in a much cleaner dataset for putative metabolite identification. A fairly stringent mass tolerance of 3 p.p.m. was used in this analysis and in a number of cases the annotation of adducts is based on strong evidence, but the exact mass matching may be outside the allowed tolerance. This is certainly seen for metabolites such as tryptophan where many ions/adducts are detected, some of which are within the 3 p.p.m. mass tolerance and others, particularly K and NaCl/HCOONa adducts of low response, are in the mass error range of 3–10 p.p.m. Increasing the mass tolerance would result in more of these adducts being correctly matched but would greatly increase the overall number of putative metabolite identifications through matching to a greater number of MFs. This is possible if further data are acquired to reduce the number of potential hits (e.g. $^{13}C/^{12}C$ isotope ratios or MS/MS fragmentation). Additionally, neither reference file had 'complete' information relevant to the human metabolome—the MMD is from a wide variety of sources and includes drugs (but not drug metabolites), and the PubChem reference file contains a limited number of elements and limited numbers of atoms per element.

The workflow for correlation analysis processed correlation data in 3–20 min depending on the option selected and the number of features. The workflow for metabolic feature annotation processed the negative ion data in ~3 min for the PubChem reference file. In positive ion mode with 100% more peaks and nearly 5 times as many correlations, the processing time was <20 min. The number of correlations is the rate determining factor and in most cases processing is <30 min and frequently <5 min. The workflow for metabolite annotation matched MF derived from experimental data to MF in MMD rapidly when fewer than 5000 matches were present. This process took just over 1 min to perform in each ion mode. However, when using large reference files such as PubChem, typically up to 30 000 matches, processing time may be of the order of 1 h.

The workflows developed are automated, rapid, open-source and freely available with all features fully annotated or given putative metabolite identifications. The approach is flexible, it is independent of chromatographic deconvolution method and analytical instrument applied. Additional adducts can be added to the adducts file for user-specific instruments, and data and organism-specific metabolite reference files can be used. Two additional differently formatted outputs are available for all matched features and the workflows have the potential to be integrated with other Taverna workflows.

Large numbers of features are recognized as artifact, isotope, salt and multiply charged ions and removed from the metabolite identification process. This in combination with data with good mass accuracy (in raw data or following post-acquisition mass alignment) results in a greatly reduced number of putative metabolite identifications within a specified mass tolerance (typically 3 p.p.m. for the ThermoFisher Scientific LTQ-Orbitrap acquired data). Using this approach, putative metabolite identification does not depend on a high level of experience in dealing with MS data and can be used as a starting point for subsequent definitive identification.

A number of false positives are always found although they are significantly reduced using this approach by annotation of ion type prior to assignment of accurate *m/z* to MF or metabolite. Wide variation in *m/z* or RT range or missing values following deconvolution can result in any or all of the following: (i) mass difference can be outside mass tolerance limits (e.g. missed adduct); (ii) RT difference between two features may be outside given value (missed grouping); and (iii) correlation between two features may be below specified limit (missed adduct, missed grouping). Within the software, allowance is made for this, for example, a feature (sodium adduct) that has a correlation with the parent metabolite below the correlation limit will not be grouped with this feature. However, if the *m/z* is within the mass tolerance it will still be putatively identified as the appropriate sodiated ion. Additional grouping information based on correlation is present in the output file and can assist the user when two or more MFs matches are reported for a feature.

4 CONCLUSIONS

The workflows presented are rapid and high-throughput and greatly reduce the number of false positives by eliminating the inaccurate matching of many artifact, isotope and complex adduct peaks. Subsequent definitive identification employing at least two properties of sample-derived metabolite and an authentic chemical standard (typically RT and fragmentation mass spectrum) can then be performed. Additional information based on similarity measures (e.g. metabolite class or metabolite pathway) are being incorporated into the Manchester Metabolomics Database and will allow in time for further interrogation of the biological changes of interest within microbial, plant and mammalian metabolomic studies. Further developments are planned to amalgamate the separate workflows together and to integrate with separate workflows to increase the applicability and ease of data analysis and interpretation.

Funding: UK Biotechnology and Biological Sciences Research Council (BBSRC) (BBC0082191); The Wellcome Trust (088075/A/08/Z); Johnson and Johnson, Cancer Research UK; The Manchester National Institute for Health Research (NIHR) Biomedical Research Centre.

Conflict of interest: none declared.

REFERENCES

Allwood,J.W. *et al*. (2008) Biomarker metabolites capturing the metabolite variance present in a rice plant developmental period. *Physiol. Plant.*, **132**, 117–135.

Atherton,H.J. *et al*. (2009) Metabolomics of the interaction between PPAR-alpha and age in the PPAR-alpha-null mouse. *Mol. Syst. Biol.*, **5**, 259.

Bradley,P.H. *et al*. (2009) Coordinated concentration changes of transcripts and metabolites in Saccharomyces cerevisiae. *PLoS Comput. Biol.*, **5,** e1000270.

Brown,M. *et al*. (2009) Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics, *Analyst*, **134**, 1322–1332.

Brown,S.C. *et al*. (2005) Metabolomics applications of FT-ICR mass spectrometry. *Mass Spectrom. Rev.*, **24**, 223–231.

Bundy,J.G. *et al*. (2009) Environmental metabolomics: a critical review and future perspectives. *Metabolomics*, **5**, 3–21.

Draper,J. *et al*. (2009) Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinformatics*, **10**, 227.

Dunn,W.B. (2008) Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Phys. Biol.*, **5,** 011001.

Dunn,W.B. *et al*. (2005) Measuring the metabolome: current analytical technologies. *Analyst*, **130**, 606–625.

Dunn,W.B. *et al*. (2008) Metabolic profiling of serum using Ultra Performance Liquid Chromatography and the LTQ-Orbitrap mass spectrometry system. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **871**, 288–298.

Dunn,W.B. *et al*. (2011) Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* **40**, 387–426.

Dunn,W.B. *et al*. (2005) Evaluation of automated electrospray-TOF mass spectrometry for metabolic fingerprinting of the plant metabolome. *Metabolomics*, **1**, 137–148.

Fernie,A.R. and Schauer,N. (2009) Metabolomics-assisted breeding: a viable option for crop improvement? *Trends Genet.*, **25**, 39–48.

Hall,R.D. *et al*. (2008) Plant metabolomics and its potential application for human nutrition. *Physiol. Plant.*, **132**, 162–175.

Hull,D. *et al*. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.

Junot,C. *et al*. (2010) Fourier transform mass spectrometry for metabolome analysis. *Analyst*, **135**, 2203–2219.

Kell,D.B. and Oliver,S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, **26**, 99–105.

Kenny,L.C. *et al*. (2010) Robust early pregnancy prediction of later preeclampsia using metabolomic biomarkers. *Hypertension*, **56**, 741–749.

Lane,A.N. *et al*. (2008) Isotopomer-based metabolomic analysis by NMR and mass spectrometry. *Methods Cell. Biol.*, **84**, 541–588.

Lewis,G.D. *et al*. (2008) Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. *J. Clin. Invest.*, **118**, 3503–3512.

Li,P. *et al*. (2008) Performing statistical analyses on quantitative data in Taverna workflows: an example using R and maxdBrowse to identify differentially-expressed genes from microarray data. *BMC Bioinformatics*, **9**, 334.

MacKenzie,D.A. *et al*. (2008) Relatedness of medically important strains of Saccharomyces cerevisiae as revealed by phylogenetics and metabolomics. *Yeast*, **25**, 501–512.

Mashego,M.R. *et al*. (2007) Microbial metabolomics: past, present and future methodologies. *Biotechnol. Lett.*, **29**, 1–16.

Rogers,S. *et al*. (2009) Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*, **25**, 512–518.

Smith,C.A. *et al*. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Soga,T. *et al*. (2003) Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.*, **2**, 488–494.

Southam,A.D. *et al*. (2007) Dynamic range and mass accuracy of wide-scan direct infusion nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry-based metabolomics increased by the spectral stitching method. *Anal. Chem.*, **79**, 4595–4602.

Sumner,L.W. *et al*. (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.

Viant,M.R. *et al*. (2006) Toxic actions of dinoseb in medaka (Oryzias latipes) embryos as determined by in vivo P-31 NMR, HPLC-UV and H-1 NMR metabolomics. *Aquat. Toxicol.*, **76**, 329–342.

Wishart,D.S. (2009) Computational strategies for metabolite identification in metabolomics. *Bioanalysis*, **1**, 1579–1596.

Wolf,S. *et al*. (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, **11**, 148.

Zelena,E. *et al*. (2009) Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal. Chem.*, **81**, 1357–1364.