Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics

M. Brown,†*[a] W. B. Dunn,†*[b] P. Dobson,†[a] Y. Patel,†[a] C. L. Winder,[b] S. Francis-McIntyre,[a] P. Begley,[a] K. Carroll,[b] D. Broadhurst,[a] A. Tseng,[a] N. Swainston,[b] I. Spasic,[b] R. Goodacre[bc] and D. B. Kell[ab]

Received 20th January 2009, Accepted 27th March 2009
First published as an Advance Article on the web 9th April 2009
DOI: 10.1039/b901179j

The chemical identification of mass spectrometric signals in metabolomic applications is important to provide conversion of analytical data to biological knowledge about metabolic pathways. The complexity of electrospray mass spectrometric data acquired from a range of samples (serum, urine, yeast intracellular extracts, yeast metabolic footprints, placental tissue metabolic footprints) has been investigated and has defined the frequency of different ion types routinely detected. Although some ion types were expected (protonated and deprotonated peaks, isotope peaks, multiply charged peaks) others were not expected (sodium formate adduct ions). In parallel, the Manchester Metabolomics Database (MMD) has been constructed with data from genome scale metabolic reconstructions, HMDB, KEGG, Lipid Maps, BioCyc and DrugBank to provide knowledge on 42,687 endogenous and exogenous metabolite species. The combination of accurate mass data for a large collection of metabolites, theoretical isotope abundance data and knowledge of the different ion types detected provided a greater number of electrospray mass spectrometric signals which were putatively identified and with greater confidence in the samples studied. To provide definitive identification metabolite-specific mass spectral libraries for UPLC-MS and GC-MS have been constructed for 1,065 commercially available authentic standards. The MMD data are available at http://dbkgroup.org/MMD/

Introduction

The metabolome is defined as the quantitative complement of small molecular weight chemicals present in a biological system.[1,2] The holistic study of the metabolome, defined as metabolomics, offers distinct advantages when compared to genomic, transcriptomic and proteomic investigations.[2,3] However, the combined study of all functional levels and their complex interactions to provide a systems-wide view of biological organisation (systems biology) is beneficial.[3–5] Metabolomics is increasingly being applied in post-genomic sciences to study biological systems including microorganisms,[6,7] plants,[8,9] mammals[3,10–12] and the environment.[13,14] From an analytical perspective, metabolomics is a strategy that offers the ability to perform high-throughput studies with relatively low operating costs after initial instrument purchase.[15]

Metabolomes are complex systems to study being composed of hundreds or thousands of metabolites (yeast 1168,[16] plants 200 000 in total kingdom, many fewer per species,[1] mammals > 6500,[17] and may be present from more than one organism either from a symbiotic or pathogenic relationship[8,18]) with a wide range of physical and chemical properties such as hydrophobicity/hydrophilicity, volatility, molecular weight and size. The study of these systems requires an integrated approach or metabolome pipeline[19] and a number of strategies are applied.[15,20] Two orthogonal strategies are typically employed: metabolic profiling and targeted analysis. Metabolic profiling (sometimes referred to as untargeted analysis or metabolite profiling) provides a more or less holistic study of a metabolome with detection of hundreds or thousands of metabolites. The strategy is applied as a hypothesis-generation strategy in metabolomic studies. Although metabolic profiling has been described as unbiased and global, in reality all methods of sample preparation and all analytical platforms introduce a level of chemical bias. Consequently, a range of analytical platforms are applied in metabolomics[21–23] including gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), capillary electrophoresis-mass spectrometry (CE-MS), nuclear magnetic resonance spectroscopy (NMR), direct infusion mass spectrometry (DIMS) and FT-IR and Raman spectroscopies. Of these, chromatography-mass spectrometry and NMR are the most widely applied.

One of the limiting factors in metabolomics is that of identifying molecules from spectroscopic/spectrometric signals.[15,24] For metabolomics to be successful there is a requirement to convert raw analytical data to metabolites (named chemicals) that may confer biological knowledge. In many studies, entities

[a] *Bioanalytical Sciences Group, School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, UK M1 7DN. E-mail: M.C.Brown@manchester.ac.uk; Fax: +44 (0) 161 306 4556; Tel: +44 (0) 161 306 5145*

[b] *Manchester Centre for Integrative Systems Biology, School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, UK M1 7DN. E-mail: warwick.dunn@manchester.ac.uk; Fax: +44 (0) 161 306 4556; Tel: +44 (0) 161 306 5146*

[c] *Laboratory for Bioanalytical Spectroscopy, School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, UK M1 7DN*

† These authors contributed equally to the research.

of biological significance are reported with no chemical identification, and generally referred to as unknowns. Consequently, a number of strategies are being brought forward to assist in the chemical identification of these unknowns, including the development of metabolite-specific mass spectral libraries and databases.[17,25–28] Two types of identification are achievable, putative or preliminary 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 standard as is performed for definitive identification. Experimentally determined accurate mass or electron-impact mass spectrum are typically applied. In LC-MS and DIMS the accurate mass is used to define molecular formulae from which suitable metabolites can be derived by searching electronic resources. However, isomers have the same accurate mass and therefore require a separate, orthogonal property for definitive identification of all potential isomers. Definitive identification employs at least two properties (typically retention time or index and fragmentation mass spectrum) and provides confidence *via* the use of authentic chemical standards. Recent informatics standards for reporting the basis of metabolite identifications have been described.[29]

The application of electrospray ionisation-mass spectrometry in metabolomics has increased rapidly in the previous five years, whether as LC-MS, CE-MS or DIMS. Electrospray (ES+ and ES−) and nanoelectrospray ionisation sources enable detection of hundreds or thousands of features in a run, where features are defined as a single mass (DIMS) or a chromatographic peak defined by the same nominal or accurate mass across the peak (LC-MS, CE-MS). Electrospray ion sources can be viewed as chemical reactors possessing the ability to produce a wide range of non-covalent and ionic interactions between metabolites and other species present. The potential for covalent bond fission is also high when operating under specific instrument conditions. These processes are observed where populations of all neutral species and ions are present at atmospheric pressures and the probability of interactions are high. As a result, a single metabolite can be detected as multiple features in either positive or negative ion modes. For example, a metabolite may be detected as the protonated and sodiated ions in positive ion mode and as the deprotonated ion in negative ion mode. These multiplicities result in overestimation of the number of detected metabolites and provide difficulties in determining the molecular formula of detected metabolites as the type of ion formed is often unknown. These multiplicities are observed because the sample matrix is generally not separated from the metabolites in metabolic profiling applications and therefore the sample is a mixture of metabolites and high concentration matrix components including inorganic salts.

High mass-accuracy mass spectrometers have been applied to assist in the chemical identification of mass signals. Typically, the combination of accurate mass and isotope ratios are employed to calculate probable molecular formulae from which potential metabolites can be derived by searching metabolomic or chemical databases (for example, The Human Metabolome Database,[17] PubChem,[30] Biospider,[31] KEGG[32]). However, the number of probable molecular formulae increases exponentially as the mass increases.[26] High mass resolution and mass accuracy can be observed when employing Fourier transform ion cyclotron resonance (FT-ICR) and Orbitrap mass analysers and to a lesser extent time of flight (ToF) mass analysers.[33,34] The Orbitrap mass analyser has accurate and precise mass measurements and high mass resolution, as shown previously in metabolomic applications.[35–40] The application of statistical analysis in metabolomics to describe biological-related or analytical-related correlations have been described, including its application to mass spectrometric and other analytical data.[5,41–45] Other researchers have detailed accurate-mass lists of possible contaminants in LC-MS datasets.[29]

We here describe methodologies used to interrogate data acquired from a wide range of complex metabolomes by ultra performance liquid chromatography (UPLC) coupled to an electrospray LTQ-Orbitrap hybrid mass spectrometer. Methods to assign and correlate mass signals deriving from the same metabolite and to use these data to increase the efficiency of putative metabolite identification are described. The construction of metabolite-specific databases (using data from a range of electronic sources) and mass spectral libraries (from analysis of authentic standards) will be described. We also give methodologies to search electronic data to provide accurate mass matching of experimental data to reported metabolites (and where possible biological naming).

Experimental

Chemicals

All chemicals and solvents used were of Analytical Grade purity or greater. Methanol and water were purchased from Sigma-Aldrich (Gillingham, UK) and formic acid was purchased from VWR (Loughborough, UK).

Sample preparation

Serum. Serum was obtained from 90 healthy subjects (41 male, 49 female) as part of the HUSERMET project.[46] Samples were prepared and analysed in one analytical batch for each of positive (ES+) and negative (ES−) ion modes as described previously.[36]

Urine. Urine was obtained from 40 healthy subjects as part of the UK Biobank sample collection and transport validation study.[47] Samples were deproteinised by addition of 200 μL methanol to 100 μL urine at room temperature, vortex mixing and centrifugation (13456*g*, 15 minutes). The supernatant was lyophilised (HETO VR MAXI vacuum centrifuge attached to a HETO CT/DW 60E cooling trap; Thermo Life Sciences, Basingstoke, UK) and reconstituted in 200 μL water. All samples were analysed in one analytical batch for each of ES+ and ES−.

Saccharomyces cerevisiae **intracellular extracts and metabolic footprints.** Samples were collected from a fermentor (Applikon Biotechnology, Netherlands) operating under turbidostat conditions.[48] A diploid heterozygous deletion yeast strain BY4743 hoΔ/HO, (YDL227C; MATa/MATα; his3Δ1/his3Δ1; leu2Δ0/leu2Δ0; met15Δ0/MET15; LYS2/lys2Δ0; ura3Δ0/ura3Δ0) was grown aerobically in a synthetic media[49] with a working volume of 2 L. Metabolic footprint (exometabolome) samples (n = 16) were collected by sampling 5 mL of culture followed by separation of cells and metabolic footprint by syringe filtration (0.22 μm, Sartorius, UK). Intracellular extracts (n = 49) were obtained employing a range of quenching solutions containing either 0.9% saline, 60% methanol, 90% methanol, 60% methanol and 0.85% ammonium carbonate, glycerol–water (3:2 v/v) or ethanol and extraction protocols (ethanol, methanol/chloroform and either 100% or 60% methanol) as described previously.[50] All samples were lyophilised (200 μL-footprint, 800 μL-intracellular extract), reconstituted in 200 μL water and analysed in one analytical batch for each sample type for each of ES+ and ES−.

Placental tissue metabolic footprints. Placental tissue was cultured for 96 hours in a serum-based growth medium at 3 different oxygen tensions (1, 6 and 20%) as described previously.[51] A total of 36 samples were analysed following lyophilisation of 100 μL of the metabolic footprint and reconstitution in 200 μL water. All samples were analysed in one analytical batch for each of ES+ and ES−.

Ultra performance liquid chromatography-LTQ/Orbitrap mass spectrometry

All samples were analysed using an Acquity UPLC chromatographic system (Waters, Elstree, UK) coupled to an electrospray LTQ-Orbitrap hybrid mass spectrometer (ThermoFisher Scientific, Bremen, Germany). 10 μL of each sample was injected on to the chromatographic system and was eluted as previously reported [52] on an Acquity UPLC BEH 1.7 μm-C_{18} column. The LTQ-Orbitrap hybrid mass spectrometer was operated in positive (ES+) and negative (ES−) ion modes, with a mass resolution of 30 000 (FWHM) and with operating parameters being tuned for maximum sensitivity for MRFA (Sigma-Aldrich, UK) at mass 514.28 Da in ES− and 524.26 Da in ES+. This mass was chosen because it is central in the typical mass ranges acquired of 50–1000 Da. The instrumental parameters were constant for each sample set but changed for the analysis of different sample types because of the tuning process applied. The Orbitrap mass analyser was mass calibrated each day using a calibration solution defined by the manufacturers.

Gas chromatography-TOF-mass spectrometry (GC-TOF-MS)

All samples were analysed as previously described on an Agilent 6890 GC (Agilent Technologies, Stockport, UK) coupled to a LECO Pegasus III electron impact mass spectrometer (LECO Corp., St. Joseph, MO). Lyophilised single component solutions (0.22–1.09 millimoles L^{-1}) were chemically derivatised in a two-step process. Firstly, 50 μL of 20 mg mL^{-1} O-methoxylamine in pyridine was added, vortexed, and incubated at 40 °C for 80 minutes in a dri-block heater. The second step involved addition of 50 μL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA), vortexing, and incubation at 40 °C for a further 80 minutes. On completion, 20 μL of retention index marker solution was added (0.6 mg mL^{-1} docosane, nonadecane, decane, dodecane and pentadecane in pyridine). Derivatised samples were analysed using two methods as previously described.[53]

Data processing of UPLC-MS data

Raw data processing. All raw data (in.raw file format) were converted to netCDF file format with the FileConverter program available in XCalibur (ThermoFisher Scientific, Bremen, Germany).

XCMS deconvolution. XCMS is an open-source deconvolution program available for LC-MS data.[54] Deconvolution using the XCMS program was performed using similar settings to those reported previously[36] with the exception of sn threshold = 3, step = 0.02, *m/z* diff = 0.05 and for grouping bandwidth = 10 and mzwidth = 0.05. The esi program[55] available with the XCMS software package was used to write peak output files to an annotated version (as a .csv file) which is more appropriate for these studies.[56] XCMS and esi were run using R version 2.6.0.

Subsequent analysis of the data was performed using the Matlab® scripting language 7.4.0[57] and all workflows were written using Taverna 1.7.0.[58] These processes were performed on a Windows-based PC with 2GB RAM.

Frequency of mass differences

(a) mass difference versus *Pearson correlation calculations.* The procedure described was applied to each sample set individually in positive ion mode and then repeated in negative ion mode. Taking each peak in turn in increasing mass order, all mass differences between all peaks were calculated and results were binned to mass windows of 0.025 Da in the mass range 0–1000 Da, in total 40 000 bins. Simultaneously Pearson correlations using chromatographic peak area data were calculated between all binary peak combinations and results were binned to windows of 0.05 in the range −1 to +1 (a total of 40 bins). The data for mass differences and Pearson correlations were combined in to a two-dimensional data matrix of mass difference *versus* Pearson correlation. The data matrix was completed with the number of entries for each mass difference-Pearson correlation pairing (40 000 × 40 provides a total of 1.6 million pairings).

(b) mass difference versus *retention time (RT) difference calculations.* The procedure described was applied to each sample set individually in positive ion mode and then repeated in negative ion mode. Taking each peak in turn in increasing mass order, all mass differences between all peaks were calculated and results were binned to mass windows of 0.025 Da in the mass range 0–1000 Da, in total 40 000 bins. Simultaneously, retention time differences were binned to 2 seconds, range 0–1200 for each binary peak combination, in total 600 bins. The data for mass and retention time differences were combined in to a two-dimensional data matrix of mass difference *versus* retention time difference. The data matrix was completed with the number of entries for each mass difference-retention time difference pairing (a total of 24 million pairings).

To note is that mass differences and retention times are peak-specific whereas Pearson correlations are peak- and sample-specific. The comparatively large sample sizes applied in this study allow correlation coefficients to be calculated with narrow confidence intervals. For smaller sample sets the results are less precise with larger confidence intervals and the results can be unreliable.[44]

Results and discussion

1. Determination of typical mass signals and ion categories detected

LC-MS, CE-MS and DIMS analytical platforms are increasingly being applied with electrospray ionisation and provide thousands of mass signals.[36] A particularly intricate collection of mass signals is detected where one metabolite can be detected as multiple ions, each of a different mass. These ions of different mass are highly correlated to the parent metabolite and are detected at the same retention time when chromatography is interfaced to the mass spectrometer. This pattern of behaviour was used to identify, in five different sample types, the different types of ions that are detected frequently. Calculations to compare mass differences *versus* Pearson correlations and to compare mass differences *versus* retention time (RT) difference were performed as described in the Experimental section. Data calculated were visualised in surf plots (3D plot of the surface for the frequency (*z* axis) compared to the mass difference (*x* axis) *versus* correlation data (*y* axis) and separately for the mass difference (*x* axis) *versus* retention time difference (*y* axis)). Surf plots for ES+ data acquired by the analysis of serum are shown in Fig. 1(a) (mass difference *versus* Pearson correlation *versus* frequency) and Fig. 1(b) (mass difference *versus* retention time difference *versus* frequency) for expanded regions defining the ions detected of highest frequency.

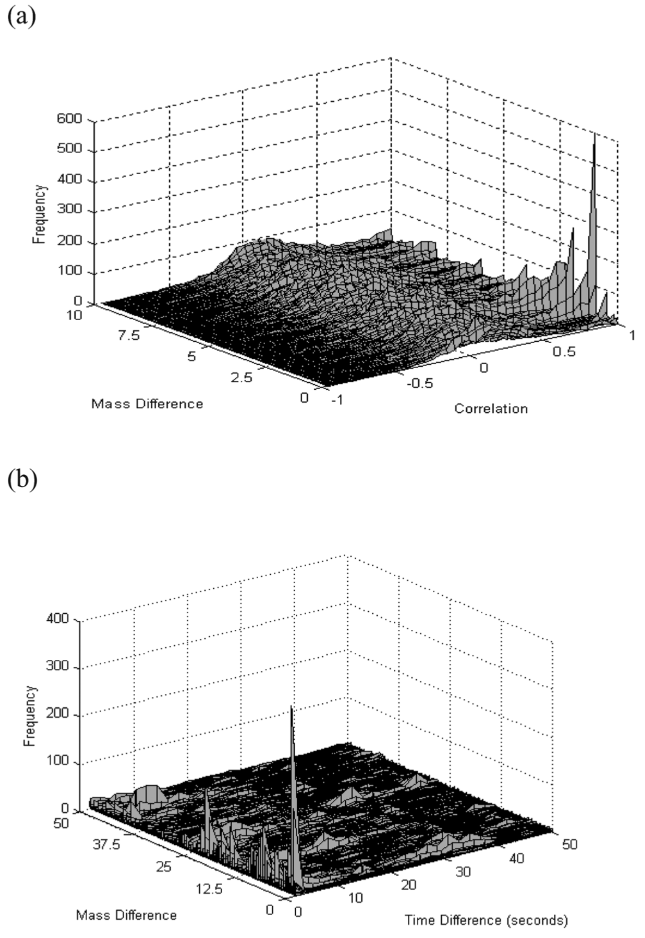

Fig. 1 Surf plots describing the frequency of mass differences detected in metabolic profiling experiments with data acquired on an Acquity UPLC coupled to an electrospray LTQ-Orbitrap hybrid mass spectrometer. Data for (a) mass difference *versus* Pearson correlation *versus* frequency and (b) mass difference *versus* retention time difference *versus* frequency are shown.

The surf plots describe the distribution of the mass differences with either correlation using peak area (Fig. 1(a)) or retention time difference (Fig. 1(b)). Some mass differences are observed more frequently above the general background of mass differences, at high positive correlations (>0.9) and at small retention time differences (<2 seconds). Only positive Pearson correlations indicated a relationship between ions originating from the same metabolite. As a separate calculation for those pairings (Pearson correlation coefficient > 0.9 and RT < 2 seconds) of higher frequency the exact mass differences were calculated from accurate mass data (not binned data) and the median mass difference reported. This was performed to increase the mass accuracy from the initial 0.025 Da bin size to an accurate mass. The bin size of 0.025 was chosen to provide an overview (while ensuring acceptable calculation speeds) from which the data for these bins only were used to calculate mass differences accurately.

Some of the most frequent mass differences (associated with retention time differences of less than 2 seconds) could be easily identified. It should be noted that other programs can undertake one or more of these processes (for example, the esi program[55]), though not employing the same method of calculation. A list of these mass differences and associated chemical identifications are shown in Table 1. Of specific interest is the high mass accuracy (mass error of less than 0.0003 Da) observed between measured and expected mass differences, even when the mass differences are measured for ions of different ion intensities (for example, for low molecular weight metabolites the ^{13}C isotope is commonly only detected at responses of less than 5% of the detected ^{12}C metabolite response, which itself can be of low intensity). The mass difference relating to the carbon isotope peak (1.0033) was the most frequent mass difference observed in all studies. Other frequently observed mass differences were those relating to sodium adducts, sulfur (^{34}S) isotope peaks, chlorine (^{37}Cl) isotope peaks and doubly/triply charged ions.

Mass differences relating to the ^{2}H, ^{15}N, ^{18}O isotopes were not detected. Theoretically the Orbitrap mass analyzer can resolve mass differences associated with the mass difference between ^{12}C and ^{13}C isotopes (mass difference = 1.0034 Da) and ^{1}H and ^{2}H isotopes (mass difference = 1.0063 Da). However, the mass difference for hydrogen isotopes was not detected because of the low natural abundance of ^{2}H (0.015%). The instrumental limit of detection restricts the possibility of detecting other low natural abundance isotopes for most metabolites which are present at low concentrations. The detection of sulfur (^{34}S) and chlorine (^{37}Cl) isotopes can assist in reducing the number of possible molecular formulae in putative identification processes. The isotope ratio can also be applied to deduce the number of elemental atoms present in the molecule to further reduce the molecular formulae possibilities.[26]

The detection of adduct ions was expected and sodium (and to a lesser degree potassium) adducts were frequent. Ammonium adducts were also detected though this could also be the loss of

Table 1 Frequently detected mass differences observed using the electrospray LTQ-Orbitrap hybrid mass spectrometry system. The experimental accurate mass difference is the median for all sample types analysed. The theoretical mass difference is defined in parenthesis. The ^{34}S and ^{37}Cl mass differences were detected in the same mass bin of width 0.025 Da

Peak	Experimental accurate mass difference (Theoretical)	Type
^{13}C isotope	1.0033 (1.0034)	Isotope
Doubly charged (C isotope)	0.5018 (0.5017)	Isotope
Triply charged (C isotope)	0.3372 (0.3344)	Isotope
^{34}S isotope	1.9956 (1.9958)*	Isotope
^{37}Cl isotope	1.9971 (1.9972)*	Isotope
	<0.3	Artifact
Sodium (Na)	21.9820 (21.9820)	Adduct
Doubly charged Na ion	10.9870 (10.991)	Charged adduct
Potassium (K)	37.9570 (37.9555)	Adduct
H_2O	18.0106 (18.0106)	Adduct/fragment
NH_3	17.0266 (17.0265)	Fragment
CO	27.9950 (27.9950)	Fragment
CO_2	43.9898 (43.9898)	Fragment
HCOOH	46.0055 (46.0054)	Fragment
HCOONa	67.9876 (67.9874)	Adduct or fragment
NaCl	57.9588 (57.9586)	Adduct
$C_3H_4O_2$	72.0206 (72.0211)	Fragment
HCOOK	83.9615 (83.9613)	Adduct or fragment
Na − K	15.9734 (15.9739)	Adduct difference
NH_3 − Na	4.9554 (4.9554)	Adduct difference
NaCl − HCOONa	10.0288 (10.0288)	Adduct difference

ammonia from amine-containing metabolites. With the method of calculation this shows that protonated and one or both salt adducts are detected which means that the number of metabolites detected is lower than the number of mass features observed. This is also observed in GC-MS data where multiple trimethylsilyl derivatives are detected.[15]

Other mass differences required further investigation to determine their origin. Fragmentation of molecular ions including the loss of water, ammonia, carbon monoxide, carbon dioxide and formic acid was observed. The loss of formic acid can imply decarboxylation of a carboxylic acid or the loss of formate from a formate adduct ion. Of interest is that loss of glycine from conjugated metabolites was detected at a low frequency. The loss of taurine from conjugated metabolites was not observed. Results are inconclusive as to whether fragmentation of conjugated metabolites occurs under the instrument conditions employed. The frequency with which salts bind non-covalently to charged ions (adductation) was of interest. For example, the binding of sodium formate to hydroxybutanoic acid was detected. Other common adductations were observed involving sodium, formate, sodium formate and sodium chloride addition to metabolites.

The detection of multiply-charged ions was observed. The multiply charged metabolites were generally observed for metabolites of higher molecular weight (mass > 450 Da) where the charge can be distributed across the molecule. Dimers ($[2M + H]^+$ and $[2M - H]^-$) will not be reported when applying the described methodology but were observed from manual interpretation of electrospray data. This has been observed previously.[59] Any dimers will have a frequency of one as the mass difference is unique for each metabolite and therefore were not highlighted as frequently observed using the process applied. Of these there are no retention time differences and therefore dimerisation is a process which occurs in the electrospray process and not prior to chromatography (where a different RT may be expected).

An instrument-specific artifact was observed for the Orbitrap mass analyser which required further investigation. The Orbitrap mass spectrometer operates by measuring ion oscillations within its mass analyser, where the frequencies of ion oscillations are inversely proportional to the square root of the mass-to-charge ratio. Ions of the same *m*/*z* will oscillate in phase at the same frequency (in a packet). The oscillation signal is detected as an image current on a pair of plates across the mass analyser. The resulting interferogram (consisting of a superposition of sine waves) is converted into the frequency domain, and hence the *m*/*z* domain, using a Fourier transform algorithm. If there is an overabundance of ions of a certain *m*/*z* then the detector is seen to overload and clip the associated oscillation sinusoid. This artifact can easily be seen after the Fourier transform as a symmetrical pattern of decaying low intensity peaks centred round the main high intensity peak with a range of mass differences of less than 0.3 Da. We describe these extra peaks as Fourier Artifact (FA) peaks and they must not be misinterpreted as real ion detection.

In-source fragmentation of metabolites is of concern, though it can provide greater structural information without the requirement for tandem mass spectrometry (and more expensive instrumentation). The extent of fragmentation is instrument-specific and instrument-tuning-specific. The instrumental parameter which has a high potential to cause fragmentation of molecular ions is the tube lens voltage. The types of metabolites which fragment are metabolite-class specific (for example, loss of ammonia and/or carbon dioxide from amino acids). To test whether the tube lens voltage was creating significant fragmentation we analysed a single serum sample in negative ion mode at five different tube lens voltages, with eight injections for each voltage. The results are described in Table 2. As the voltage increases the number of peaks detected increases. However, the

Table 2 Comparison of detection of different ion types as a relationship with tube lens voltages using accurate mass and retention time differences of less than 2 seconds. Correlation analysis was not used as sample sizes were low (n = 8). Data for negative ion mode only for serum samples analysed in a single batch are shown with the number of peaks detected for each ion type and tube lens voltage. Detected peaks is defined as all peaks detected in a minimum of 75% of all samples at a specific voltage

	20 V Negative ion mode (1835 peaks)	40 V	60 V	80 V	100 V	All
No. of samples	8	8	8	8	8	40
Detected peaks	901	1040	1321	1655	1697	—
Fourier peaks	30	40	42	42	40	44
Isotope peaks	296	313	322	348	341	350
Doubly charged	1	4	2	4	4	4
Triply charged	0	0	0	0	0	0
Dimers	5	5	5	5	5	5
Adducts	0	0	0	0	0	0
(Na, K)	13	13	19	19	19	19
Fragments + others	226	237	256	263	261	264
Total classified	571	612	646	681	670	687

percentage of any ion type does not increase or decrease significantly, though there is a small (less than 10%) increase in the number of fragment ions detected as the tube lens voltage increases. Increasing the tube lens voltage provides increased efficiency of ion transfer from electrospray source to mass analyser though without significant increases in the amount of fragmentation observed. Most probable is that the sub-atmospheric pressure in the tube lens reduces the probability of ion-molecule collisions required for fragmentation. The number of ions detected and the frequency of fragmentation requires balancing to maximise the detected coverage of the metabolome while minimising the fragmentation of molecular ions. Operating the UPLC system at a column temperature of 50 °C may also cause dissociation of metabolites though it would be expected that these fragments would be detected at different retention times and hence not be highlighted as fragment ions with the methodologies applied by the authors.

2. Chemical annotation of detected mass signals

As described above, a range of different types of ions can be detected for a single metabolite. The frequencies of mass differences were determined for a range of different sample types which were analysed using a UPLC-LTQ/Orbitrap platform. These included serum, urine, yeast intracellular extracts, yeast metabolic footprints and placental explant metabolic footprints.

For all data the frequencies (as a percentage of total number of detected features) of each ion type (protonated and deprotonated metabolite ions, isotopes, multiply charged ions, FA peaks, adducts, fragment ions and dimers) were calculated. For a single metabolite these different ion types have similar retention times (±2s) and a Pearson correlation greater than 0.9. The results are shown in Table 3.

Of specific interest is that a single metabolite can be detected as multiple ions of different mass. The frequency of metabolites detected as multiple ions range from 14.0% to 33.1% of all peaks detected for the sample types analysed and for the mass spectrometer employed. The percentage describes the number of features (relative to all detected features) which can be reported as a single metabolite. For placental footprints, 1 in 3 or 33.1% metabolites were detected as multiple mass ions in one analytical run. 2271 detected features relate to 1519 metabolites being present in the samples. These ranges can be expected to change for different instrument types and specific instrument operating conditions. Data also showed that the number of multiple features for a single metabolite is not concentration dependant, but the relative ratio of responses for each feature can be concentration dependant.

No ion type was uniquely observed in one ion mode only, though the frequency can be much higher for one ion mode. For example, sodium and potassium adduct ions are observed more frequently in positive ion mode. However, sodium formate and sodium chloride adduct ions are more frequently observed in negative ion mode as singly charged ions. This exhibits the non-covalent addition of ionic or neutral species to metabolites. The percentage of total peaks which are defined as sodium adducts is dependent on the sample type and therefore it could be influenced by the concentration of sodium in the sample. Serum contains 142.6 mmol L^{-1} sodium according to HMDB and ES+ has the greatest frequency of sodiated adduct ions. Urine has lower concentrations of sodium (14.7 mmol L^{-1}) and lower frequencies of sodiated adduct ions. Yeast intracellular extracts and footprint samples were observed to have similar levels of sodiated adduct ions. Of interest is that sodiated ions are observed not only in ES+ but also in ES− where sodium can create a non-covalent interaction with the metabolite.

3. Metabolite database and mass spectral libraries

The results described above highlight the complexity of the kind of raw analytical data typically obtained. We perform two iterative processes to provide, where possible, the chemical identification of a wide range of metabolites detected when applying GC-MS and UPLC-MS analytical platforms. These processes have involved the construction of (a) The Manchester Metabolomics Database (MMD) containing accurate mass data for all potential metabolites detected and other chemical and physical properties and (b) mass spectral metabolite libraries constructed using authentic standards.

(a) The Manchester Metabolomics Database. Molecular structures of metabolites from a range of sources have been

Table 3 Summary of the frequency of different ion types detected in all sample types investigated. All results are reported as a percentage of the total number of ions detected for ES− (upper) and ES+ (lower). % Annotated describes the percentage of all detected peaks which have been identified as FA, isotope, doubly or triply charged, dimmers, adducts or fragment peaks

	Serum	Yeast intracellular extract	Yeast footprint	Placenta	Urine
No. of samples	90	49.0	16.0	36.0	108
No. of peaks	4513	595	595	2271	4804
FA peaks	1.60	4.00	4.20	1.06	2.58
Isotope peaks (C, Cl, S)	18.1	16.1	10.3	17.79	7.50
Doubly charged	3.35	<0.01	0.00	0.57	0.44
Triply charged	<0.02	<0.01	<0.01	0.00	0.10
Dimers	0.01	0.00	0.00	0.13	0.17
Adducts (Na, K)	1.60	0.67	0.50	0.97	0.71
Fragments/adducts	8.10	2.00	0.67	11.6	2.45
% Annotated	32.8	22.8	15.7	33.1	14.0

	Serum	Yeast intracellular extract	Yeast footprint	Placenta
No. of samples	100	49.0	16.0	36.0
No. of peaks	2079	979	979	1906
FA peaks	0.77	3.90	4.60	1.47
Isotope peaks (C, Cl, S)	13.7	14.6	11.5	13.27
Doubly charged	8.13	0.00	0.00	1.52
Triply charged	0.80	<0.01	<0.01	<0.01
Dimers	<0.01	0.00	0.00	0.00
Adducts (Na, K)	3.40	2.76	1.84	2.57
Fragments/adducts	1.92	5.40	4.80	5.67
% Annotated	28.7	26.7	22.8	24.5

collated in a cheminformatics workflow environment[60] (Pipeline Pilot[61]). Human metabolite structures were retrieved from the HMDB[62] and Lipid Maps[63] databases, plus human metabolic network reconstructions.[64–66] *Saccharomyces cerevisiae* metabolite structures were obtained from the yeast metabolic network reconstruction of Herrgard *et al.*.[16] Further metabolic structures not assigned to a specific species were retrieved from KEGG[67] and BioCyc.[68] Drug structures, which may also appear in samples, were taken from DrugBank[69] and KEGG Drug.[67]

Duplicate molecules, as identified by equivalent canonical SMILES[70] (a unique line notation of molecular structure) were merged to a single entry. Where records contained salts in the structure, a further record was generated without the salt. For charged metabolites a further record representing the uncharged form was generated (*e.g.*, glutamate and glutamic acid) and are represented as two separate entries in the database. The provenance of these derived molecules is preserved by linking back to the original molecule from which they were generated. Each metabolite record contains a range of chemical, physical and other properties, including preferred names and synonyms, molecular formula and accurate molecular masses to allow putative identification of metabolites from accurate mass data acquired on UPLC-MS instrumentation, SMILES and InChI strings depicting structure, internal identifiers, and links back to the source databases. SMILES and InChI are unambiguous representations of structure, including steroisomerism, and are database independent and we recommend the use of these identifiers.[16] Construction of the database is an ongoing process as metabolomics is a relatively new scientific discipline and resources are frequently updated. For example, HMDB was updated from version 1.0 to 2.0 in November 2008 and the new data are included in the MMD. The current version of the MMD database contains 42 687 records. This number is high (compare to HMDB 2.0 which contains 6500 metabolites) due to the inclusion of non-species specific metabolites and structural isomers (for example D and L isomers), but this strategy provides greater coverage.

Metabolites can be identified by many names. For example, D-glucose has 23 named entries in HMDB, including ‘grape juice’ and ‘corn sugar’. This illustrates how different researchers and databases use different names for the same metabolite. It is useful to agree on names as much as possible, and to this end we have defined a methodology for naming from the synonyms found in the source databases. Where possible we have adopted the name used in the yeast metabolic network reconstruction as these we know to have been assessed rigorously by multiple parties (further emphasising the utility of a community approach to Systems Biology[16]). Our second preference is for names derived from KEGG, then HMDB, BioCyc, LipidMaps, and the human metabolic network reconstructions. Drug names are taken principally from KEGG Drug and then DrugBank. Clearly this does not result in perfect naming, but all known synonyms are also stored for searching. IUPAC names are only provided if they are present as the unique name or synonym from the sources used. Further work may provide the matching of species to ChEBI.[71]

(b) Experimentally derived mass spectral metabolomic libraries. The first stage of creating the Manchester Metabolomic Database resulted in 4915 unique metabolites. Definitive identification of metabolites requires the matching of at least two orthogonal properties of the detected metabolite to that of an authentic standard. All available authentic standards were purchased to enable the construction of mass spectral libraries for GC-MS and UPLC-MS. SMILES strings were applied to define metabolites which were available to be purchased as authentic standards, with Sigma-Aldrich as the supplier (DiscoveryCPR[72]). 1068 metabolites were found to be available commercially and were purchased. Each metabolite was analysed using our standard analytical methods[36,53] on two analytical platforms, GC-MS and UPLC-MS.

High quality GC-MS data for 637 oxime/trimethylsilylated metabolites were acquired, resulting in 794 metabolite peak entries, since multiple products were formed during trimethylsilylation derivatisation. Retention index, electron impact mass spectrum and InCHI identifier data have been collated in a single mass spectral library using the LECO ChromaTOF software (v2.24) in a format applicable for use on multiple software packages including the NIST MS Search 2.0 program. This library is applied for metabolite identification in GC-MS datasets acquired by the authors using retention indices and electron-impact mass spectra as two orthogonal properties. We also apply other libraries including the NIST/EPA/NIH02 library and The Golm Metabolome Database[27] for preliminary identification. The inclusion of drugs present in the Sigma-Aldrich LOPAC library[73] is currently being performed. The library contains many metabolites previously reported (for example in the NIST/EPA/NIH02 library), though some metabolites have not previously been reported or have not been reported as data acquired on TOF instruments. The novelty of the library is three-fold; (a) inclusion of retention indices for a specific instrumental method (b) the method in which the metabolites were chosen was a logical bioinformatical process and (c) the application of InCHI identifiers which allows direct connectivity to other electronic resources. Metabolite names can not always provide this connectivity (see ref. 16 for further information).

A total of 788 metabolites were detected using UPLC-MS in positive ion, negative ion or in both ion modes. These data have been collated with accurate mass and molecular formula data as described in the MMD to allow preliminary identification of metabolites using two properties (retention time and accurate mass). Currently MS/MS spectra are being acquired at two normalized collision energies of 20% and 40% and included in the database and a separate library which can be employed in the NIST MS Search 2.0 program to allow definitive identification of those metabolites available as authentic standards. These data being acquired are for the protonated and deprotonated ions only at present. The inclusion of other ions such as adducts to the library is not currently planned though is of importance as these types of ions are also detected in biological samples. Definitive identification using libraries is only possible for those metabolites commercially available. Approximately, 78% of known metabolites are not commercially available from Sigma-Aldrich and therefore methodologies to identify such metabolites putatively are required. The first stage of this process for UPLC-MS data is the construction of the MMD containing all reported metabolites, which can then be interrogated.

4. Matching of UPLC-MS acquired accurate mass data to molecular formulae and the Manchester Metabolomics Database

After raw data deconvolution and grouping of UPLC-MS data using XCMS, a list of mass peaks with accurate mass (median) and mass range with median RT and RT range is reported. Following processing of the raw data and univariate or multivariate analysis to identify peaks of biological interest we applied a two-step process (Fig. 2) using Taverna workflows[74–76] which allows putative identification of metabolites in an automated manner.

Step One matches the experimentally observed accurate mass (defined as median mass from XCMS) to the mass of a molecular formula present in PubChem as has been described by Kind and Fiehn.[26] This list (sorted to increasing mass) contains all known molecular formulae (as defined in PubChem) and accurate masses in the mass range 50–1000 Da for all common elements (including H, C, N, O, P, S, F, Cl, Si, but not Na or K). Isomers have identical molecular formulae and so are grouped as a single molecular formula. The PubChem list contains endogenous metabolites, drugs and other chemicals not generally classified as endogenous or exogenous metabolites. No checks are made for validity of the molecular formula (MF) *e.g.* valency or nitrogen rules as all molecular formulae relate to known compounds. During this process 'potential' isotope, charged and adduct peaks are identified based on accurate mass difference and retention time similarity (these can be confirmed using correlation analysis or scatter plots of the raw data). Using the median

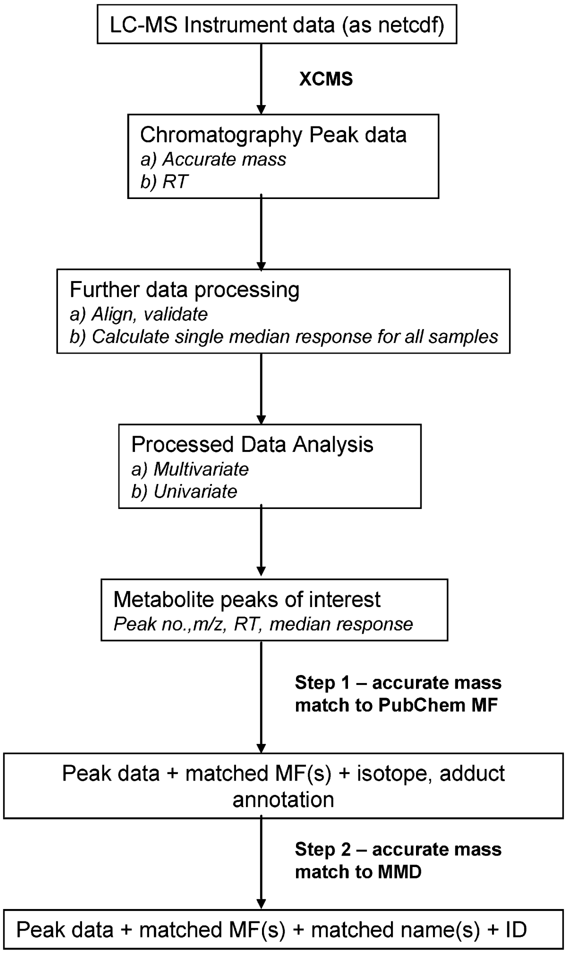
LC-MS Instrument data (as netcdf)
XCMS
Chromatography Peak data
a) Accurate mass
b) RT
Further data processing
a) Align, validate
b) Calculate single median response for all samples
Processed Data Analysis
a) Multivariate
b) Univariate
Metabolite peaks of interest
Peak no.,m/z, RT, median response
Step 1 – accurate mass match to PubChem MF
Peak data + matched MF(s) + isotope, adduct annotation
Step 2 – accurate mass match to MMD
Peak data + matched MF(s) + matched name(s) + ID

Fig. 2 Flow diagram describing the protocol applied to determine chemical and biological identification of detected mass features.

peak response for each peak the ratio for potential $^{12}C/^{13}C$, $^{35}Cl/^{37}Cl$ and $^{32}S/^{34}S$ isotope peaks is calculated and the number of C, Cl or S atoms is included in the output data. For each sample set the workflow reports all matched molecular formula within a user-defined mass error range for protonated and deprotonated ions and a range of commonly-detected adduct ions (sodium, potassium, ammonium, sodium chloride, sodium formate, potassium formate). Table 4 lists for all the sample sets the number of protonated, deprotonated and adduct ions which matched with at least one molecular formula in PubChem using a mass tolerance of 5ppm. The percentage of peaks assigned as isotope or Fourier artifact peaks ranged from 15–32% and were highest for serum in ES+ and ES−. These assignments are based on accurate mass and retention time differences only and do not include correlation-based analysis. The data presented in Table 4 applies these two parameters and correlation analysis to assign isotope and FA peaks and therefore small differences are observed in the results described in Tables 3 and 4. The percentage of the total peaks matched with at least one molecular formula in PubChem using a mass tolerance of 5 ppm ranged from 53–67% and was dependent on the sample type.

This is complementary to experimentally derived metabolite library or MMD searches as these libraries are dependent on their source and currently not all metabolites (both endogenous and exogenous) are publicly available (for example, drug metabolites, exogenous metabolites from food and environment). Therefore the ability to provide a molecular formula without a definitive match is still helpful in a range of investigations, and the absence of a match can highlight an atypical ion which can be, for example, a fragment ion. n-alkyl ions were observed in positive ion mode and were being matched to MF but would not be matched to a metabolite library. These ions are probably created by cleavage of alkyl-containing metabolites to produce the alkyl group as a charged species.

In Step Two, a Taverna workflow has been implemented to match the observed accurate mass (median) to the accurate mass of a metabolite in the MMD described above for protonated, deprotonated ions and adduct ions (sodium, potassium, ammonium, sodium chloride, sodium formate, potassium formate). A match to a user-defined mass accuracy (usually set to 5 ppm) provides a result. Common name, internal unique identifier, molecular formula and retention time are reported as a minimum. Matching to accurate mass and not molecular formula is performed because of the much greater search speed achieved using this approach. The additional information relating to isotopes, adducts and atom numbers passed through from Step One helps reduce the number of false positive matches for molecular formulae or named metabolites giving more confidence in the putative identification. Both Taverna workflows will be available at myExperiment.[77] The percentage of the total peaks matched with at least one molecular formula in MMD using the same mass tolerance of 5 ppm ranged from 39–52% where serum contained the lowest number of metabolites identified and yeast the highest number of metabolites identified. This shows the probable greater knowledge of metabolites present in yeast metabolomes compared to mammalian metabolomes. In total 74–90% of all mass ions detected were identified as either an isotope peak, FA peak, multiply-charged peak or matched to a molecular formula. These results show that further

Table 4 Description of the percentage of all detected mass signals matched to a molecular formula in PubChem or accurate mass of a metabolite present in the Manchester Metabolomics Database. Data is shown for all sample types investigated in ES+ and ES−. Excluded peaks are defined as the combination of isotope and Fourier artifact peaks. The percentage of peaks identified is calculated for the total number of peaks detected

	Serum	Yeast intracellular extract	Yeast footprint	Placenta	Urine	Serum	Yeast intracellular extract	Yeast footprint	Placenta
	Negative ion mode					Positive ion mode			
No. of samples	100	49	16	36	108	100	49	16	36
No. of peaks	4513	595	595	2271	4804	2079	979	979	1906
Isotopes %	24.6	15.5	15.5	24.6	11	16	13.5	13.5	14.9
Fourier artifact peaks %	1.5	5	5	1	2.5	0.7	4.1	4.1	1.4
Multiply-charged peaks %	1.2	0	0	0	1	15.2	0	0	3.7
Excluded peaks %	27.3	20.5	20.5	25.6	14.5	31.9	17.6	17.6	20
% PubChem MF (5 ppm)	62.8	58.2	53.3	64.2	62.1	58.8	66.7	60.6	63. 9
% MMD (5 ppm)	40.4	49.9	45.5	47.1	45.4	42.5	52.1	45	38.9

experimental and bioinformatics work is still required to provide a molecular identification to metabolites, and not just a molecular formula. The application of accurate mass and isotope ratio data for precursor ions and for corresponding product ions derived from MS^n experiments, coupled with retention time data, would provide greater accuracy of metabolite identification. Currently, no methods for automated processing of all data to acquire the information are available.

Conclusions

The research described provides significant advances in the ability to identify the chemicals causing mass spectrometric signals detected in metabolic profiling applications, and thus to provide greater biological significance to results obtained in metabolomic investigations.

The studies described have provided greater insights into the complexity of electrospray mass spectrometry data and their application in metabolomics research. Methodologies have been applied to acquired experimental data to describe the multiple types of ions that can be detected for a single metabolite and the frequency of which these ions are typically detected in a range of different sample types. On average 14.0–33.1% of ions detected are multiplicities of a single metabolite showing the over-estimation of the number of metabolites detected. The type and frequency of ions detected has been shown to be sample type-dependent and can be expected to be instrument-dependent.

In this study related peaks were grouped together using both correlation and small retention time differences. However, methods can be adapted to obtain 'useful' information when either but not both are available or calculated. For example, when sample sets are small ($n < 10$) where correlations are unstable at these small sample sizes. Alternatively, this methodology could be applied to DIMS data where no retention time data are available but the high-throughput nature allows hundreds of samples to be analysed and therefore correlation coefficients are valid where sample sizes are large.[78]

In parallel, the Manchester Metabolomics Database (MMD) was constructed from the accumulation of metabolite-specific data from a range of sources including genome scale metabolic reconstructions, HMDB, KEGG and Lipid Maps. In combination with the knowledge acquired on the complexity of electrospray mass spectrometry data, greater numbers of detected features could be putatively identified than was previously achievable and with a greater confidence of identification. To provide definitive identifications mass spectral-based metabolite libraries (UPLC-MS and GC-MS) are being constructed which apply orthogonal chromatographic (retention time/index related to polarity or volatility) and mass spectral (mass spectrum or accurate mass related to metabolite structure) for confident metabolite identification. However, the research highlights the difficulties in providing definitive identifications of metabolites, where the majority of metabolites are not commercially available and therefore other methods are required including the combination of NMR and mass spectral elucidation of metabolites in either complex or purified samples.

Acknowledgements

The authors wish to thank the following for provision of samples; UK Biobank (urine), Alexander Heazell and Stephanie Worton (University of Manchester, placental tissue footprint samples) and the HUSERMET project (serum). MB wishes to thank The British Heart Foundation and Tommy's the Baby Charity for financial assistance. WBD, YP, CW, KC, NS, RG and DBK wish to thank the BBSRC and EPSRC for financial support of The Manchester Centre for Integrative Systems Biology (http://www.mcisb.org/).

References

1 O. Fiehn, *Plant Mol. Biol.*, 2002, **48**, 155–171.
2 R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell, *Trends Biotechnol.*, 2004, **22**, 245–252.
3 D. B. Kell, *Drug Discov. Today*, 2006, **11**, 1085–1092.
4 J. I. Castrillo, L. A. Zeef, D. C. Hoyle, N. Zhang, A. Hayes, D. C. J. Gardner, M. J. Cornell, J. Petty, L. Hakes, L. Wardleworth, B. Rash, M. Brown, W. B. Dunn, D. Broadhurst, K. O'Donoghue, S. S. Hester, T. P. J. Dunkley, S. R. Hart, N. Swainston, P. Li, S. J. Gaskell, N. W. Paton, K. S. Lilley, D. B. Kell and S. G. Oliver, *Journal of Biology*, 2007, **6**.
5 J. van der Greef, S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E. R. Verheij and R. N. McBurney, *J. Proteome Res.*, 2007, **6**, 1540–1559.
6 D. A. MacKenzie, M. Defernez, W. B. Dunn, M. Brown, L. J. Fuller, S. de Herrera, A. Guenther, S. A. James, J. Eagles, M. Philo, R. Goodacre and I. N. Roberts, *Yeast*, 2008, **25**, 501–512.
7 J. Smedsgaard and J. Nielsen, *J. Exp. Bot.*, 2005, **56**, 273–286.
8 J. W. Allwood, D. I. Ellis, J. K. Heald, R. Goodacre and L. A. J. Mur, *Plant J.*, 2006, **46**, 351–368.
9 R. D. Hall, *New Phytol.*, 2006, **169**, 453–468.
10 W. Dunn, D. Broadhurst, S. Deepak, M. Buch, G. McDowell, I. Spasic, D. Ellis, N. Brooks, D. Kell and L. Neyses, *Metabolomics*, 2007, **3**, 413–426.
11 L. C. Kenny, D. Broadhurst, M. Brown, W. B. Dunn, C. W. G. Redman, D. B. Kell and P. N. Baker, *Reprod. Sci.*, 2008, **15**, 591–597.
12 J. C. Lindon, E. Holmes and J. K. Nicholson, *FEBS J.*, 2007, **274**, 1140–1151.
13 Y. Tanaka, T. Higashi, R. Rakwal, J. Shibato, E. Kitagawa, S. Murata, S. I. Wakida and H. Iwahashi, *Advanced Environmental Monitoring*, 2008, 325–337.
14 M. R. Viant, *Mar. Ecol.-Prog. Ser.*, 2007, **332**, 301–306.
15 W. B. Dunn, *Phys. Biol.*, 2008, **5**.
16 M. J. Herrgard, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Bluthgen, S. Borger, R. Costenoble, M. Heinemann, M. Hucka, N. Le Novere, P. Li, W. Liebermeister, M. L. Mo, A. P. Oliveira, D. Petranovic, S. Pettifer, E. Simeonidis, K. Smallbone, I. Spasic, D. Weichart, R. Brent, D. S. Broomhead, H. V. Westerhoff, B. Kirdar, M. Penttila, E. Klipp, B. O. Palsson, U. Sauer, S. G. Oliver, P. Mendes, J. Nielsen and D. B. Kell, *Nat. Biotechnol.*, 2008, **26**, 1155–1160.
17 D. S. Wishart, C. Knox, A. C. Guo, R. Eisner, N. Young, B. Gautam, D. D. Hau, N. Psychogios, E. Dong, S. Bouatra, R. Mandal, I. Sinelnikov, J. Xia, L. Jia, J. A. Cruz, E. Lim, C. A. Sobsey, S. Shrivastava, P. Huang, P. Liu, L. Fang, J. Peng, R. Fradette, D. Cheng, D. Tzur, M. Clements, A. Lewis, A. De Souza, A. Zuniga, M. Dawe, Y. Xiong, D. Clive, R. Greiner, A. Nazyrova, R. Shaykhutdinov, L. Li, H. J. Vogel and I. Forsythe, *Nucleic Acids Research*, 2009, **37**, D603–D610.
18 R. Goodacre, *J. Nutr.*, 2007, **137**, 259S–266S.
19 M. Brown, W. B. Dunn, D. I. Ellis, R. Goodacre, J. Handl, J. D. Knowles, S. O'Hagan, I. Spasic and D. B. Kell, *Metabolomics*, 2005, **1**, 39–51.
20 W. Lu, B. D. Bennett and J. D. Rabinowitz, *J. Chromatogr. B*, 2008, **871**, 236–242.

21 M. Bedair and L. W. Sumner, *Trac-Trends Anal. Chem.*, 2008, **27**, 238–250.
22 W. B. Dunn and D. I. Ellis, *Trac-Trends Anal. Chem.*, 2005, **24**, 285–294.
23 J. C. Lindon and J. K. Nicholson, *Trac-Trends Anal. Chem.*, 2008, **27**, 194–204.
24 S. Moco, R. J. Bino, R. C. H. De Vos and J. Vervoort, *Trac-Trends Anal. Chem.*, 2007, **26**, 855–866.
25 H. Horai, M. Arita and T. Nishioka, Bmei 2008: Proceedings of the International Conference on Biomedical Engineering and Informatics, 2008, Vol. 2, pp. 853–857.
26 T. Kind and O. Fiehn, *BMC Bioinformatics*, 2006, **7**.
27 J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A. R. Fernie and D. Steinhauser, *Bioinformatics*, 2005, **21**, 1635–1638.
28 C. A. Smith, G. O'Maille, E. J. Want, C. Qin, S. A. Trauger, T. R. Brandon, D. E. Custodio, R. Abagyan and G. Siuzdak, *Ther. Drug Monit.*, 2005, **27**, 747–751.
29 B. O. Keller, J. Sui, A. B. Young and R. M. Whittal, *Analytical Chimica Acta*, 2008, **627**, 71–81.
30 http://pubchem.ncbi.nlm.nih.gov/.
31 http://biospider.ca/.
32 http://www.genome.jp/kegg/.
33 S. C. Brown, G. Kruppa and J. L. Dasseux, *Mass Spectrom. Rev.*, 2005, **24**, 223–231.
34 R. H. Perry, R. G. Cooks and R. J. Noll, *Mass Spectrom. Rev.*, 2008, **27**, 661–699.
35 R. Breitling, A. R. Pitt and M. P. Barrett, *Trends Biotechnol.*, 2006, **24**, 543–548.
36 W. B. Dunn, D. Broadhurst, M. Brown, P. N. Baker, C. W. G. Redman, L. C. Kenny and D. B. Kell, *J. Chromatogr. B*, 2008, **871**, 288–298.
37 J. C. L. Erve, W. DeMaio and R. E. Talaat, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 3015–3026.
38 M. A. Kamleh, Y. Hobani, J. A. T. Dow and D. G. Watson, *FEBS Lett.*, 2008, **582**, 2916–2922.
39 A. D. Southam, T. G. Payne, H. J. Cooper, T. N. Arvanitis and M. R. Viant, *Anal. Chem.*, 2007, **79**, 4595–4602.
40 E. Werner, J. F. Heilier, C. Ducruix, E. Ezan, C. Junot and J. C. Tabet, *J. Chromatogr. B*, 2008, **871**, 143–163.
41 E. Werner, V. Croixmarie, T. Umbdenstock, E. Ezan, P. Chaminade, J. C. Tabet and C. Junot, *Anal. Chem.*, 2008, **80**, 4918–4932.
42 G. T. Gipson, K. S. Tatsuoka, B. A. Sokhansanj, R. J. Ball and S. C. Connor, *Metabolomics*, 2008, **4**, 94–103.
43 R. Steuer, *Brief. Bioinform.*, 2006, **7**, 151–158.
44 D. Camacho, A. de la Fuente and P. Mendes, *Metabolomics*, 2005, **1**, 53–63.
45 D. J. Crockford, J. C. Lindon, O. Cloarec, R. S. Plumb, S. J. Bruce, S. Zirah, P. Rainville, C. L. Stumpf, K. Johnson, E. Holmes and J. K. Nicholson, *Anal. Chem.*, 2006, **78**, 4398–4408.
46 http://www.husermet.org/.
47 W. B. Dunn, D. Broadhurst, D. I. Ellis, M. Brown, A. Halsall, S. O'Hagan, I. Spasic, A. Tseng and D. B. Kell, *Int. J. Epidemiol.*, 2008, **37**, 23–30.
48 H. M. Davey, C. L. Davey, A. M. Woodward, A. N. Edmonds, A. W. Lee and D. B. Kell, *Biosystems*, 1996, **39**, 43–61.
49 J. Allen, H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver and D. B. Kell, *Nat. Biotechnol.*, 2003, **21**, 692–696.
50 C. L. Winder, W. B. Dunn, S. Schuler, D. Broadhurst, R. Jarvis, G. M. Stephens and R. Goodacre, *Anal. Chem.*, 2008, **80**, 2939–2948.
51 A. E. P. Heazell, M. Brown, W. B. Dunn, S. A. Worton, I. P. Crocker, P. N. Baker and D. B. Kell, *Placenta*, 2008, **29**, 691–698.
52 E. Zelena, W. B. Dunn, D. Broadhurst, S. Francis-McIntyre, K. M. Carroll, P. Begley, S. O'Hagan, J. D. Knowles, A. Halsall, HUSERMET Consortium, I. D. Wilson and D. B. Kell, *Anal. Chem.*, 2009, **81**, 1357–1364.
53 S. O'Hagan, W. B. Dunn, M. Brown, J. D. Knowles and D. B. Kell, *Anal. Chem.*, 2005, **77**, 290–303.
54 C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Anal. Chem.*, 2006, **78**, 779–787.
55 http://msbi.ipb-halle.de/msbi/esi/.
56 Ralf Tautenhahn, Christoph Böttcher and S. Neumann, in *1st International Conference on Bioinformatics Research and Development*, Springer, 2007.
57 http://www.mathworks.com.
58 http://taverna.sourceforge.net/.
59 W. B. Dunn, S. Overy and W. P. Quick, *Metabolomics*, 2005, **1**, 137–148.
60 R. Romero, J. Espinoza, F. Gotsch, J. P. Kusanovic, L. A. Friel, O. Erez, S. Mazaki-Tovi, N. G. Than, S. Hassan and G. Tromp, *BJOG*, 2006, **113**, 118–135.
61 http://accelrys.com/products/scitegic/.
62 D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. MacInnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel and L. Querengesser, *Nucleic Acids Res.*, 2007, **35**, D521–D526.
63 M. Sud, E. Fahy, D. Cotter, A. Brown, E. A. Dennis, C. K. Glass, A. H. Merrill, R. C. Murphy, C. R. H. Raetz, D. W. Russell and S. Subramaniam, *Nucleic Acids Research*, 2007, **35**, D527–D532.
64 N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas and B. O. Palsson, *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 1777–1782.
65 H. W. Ma and I. Goryanin, *Drug Discovery Today*, 2008, **13**, 402–408.
66 H. W. Ma, A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin and I. Goryanin, *Molecular Systems Biology*, 2007, **3**.
67 M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu and Y. Yamanishi, *Nucleic Acids Research*, 2008, **36**, D480–D484.
68 R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang and P. D. Karp, *Nucleic Acids Research*, 2008, **36**, D623–D631.
69 D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucleic Acids Research*, 2008, **36**, D901–D906.
70 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
71 http://www.ebi.ac.uk/chebi/.
72 http://www.sigmaaldrich.com/chemistry/chemical-services/discovery-cpr.html.
73 http://www.sigmaaldrich.com/chemistry/drug-discovery/validation-libraries/lopac1280-navigator.html.
74 T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat and P. Li, *Bioinformatics*, 2004, **20**, 3045–3054.
75 T. Oinn, M. Greenwood, M. Addis, M. N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. R. Pocock, M. Senger, R. Stevens, A. Wipat and C. Wroe, *Concurr. Comput.-Pract. Exp.*, 2006, **18**, 1067–1100.
76 T. Oinn, P. Li, D. B. Kell, C. Goble, A. Goderis, M. Greenwood, D. Hull, R. Stevens, D. Turi and J. Zhao, in *Workflows for e-Science: scientific workflows for Grids*, ed. E. D. I. J. Taylor, D. B. Gannon and M. Shields, Springer, Guildford, Editon edn., 2007, pp. 300–319.
77 http://www.myexperiment.org/.
78 D. I. Broadhurst and D. B. Kell, *Metabolomics*, 2006, **2**, 171–196.