Metabolomics (2012) 8:1170–1180
DOI 10.1007/s11306-012-0422-8

ORIGINAL ARTICLE

Characterization of data analysis methods for information recovery from metabolic ^{1}H NMR spectra using artificial complex mixtures

Alexessander C. Alves · Jia V. Li · Isabel Garcia-Perez · Caroline Sands · Coral Barbas · Elaine Holmes · Timothy M. D. Ebbels

Received: 26 September 2011 / Accepted: 29 March 2012 / Published online: 27 April 2012

Abstract The assessment of data analysis methods in ^{1}H NMR based metabolic profiling is hampered owing to a lack of knowledge of the exact sample composition. In this study, an artificial complex mixture design comprising two artificially defined groups designated normal and disease, each containing 30 samples, was implemented using 21 metabolites at concentrations typically found in human urine and having a realistic distribution of inter-metabolite correlations. These artificial mixtures were profiled by ^{1}H NMR spectroscopy and used to assess data analytical methods in the task of differentiating the two conditions. When metabolites were individually quantified, volcano plots provided an excellent method to track the effect size and significance of the change between conditions. Interestingly, the Welch *t* test detected a similar set of metabolites changing between classes in both quantified and spectral data, suggesting that differential analysis of ^{1}H NMR spectra using a false discovery rate correction, taking into account fold changes, is a reliable approach to detect differential metabolites in complex mixture studies. Various multivariate regression methods based on partial least squares (PLS) were applied in discriminant analysis mode. The most reliable methods in quantified and spectral ^{1}H NMR data were PLS and RPLS linear and logistic regression respectively. A jackknife based strategy for variable selection was assessed on both quantified and spectral data and results indicate that it may be possible to improve on the conventional Orthogonal-PLS methodology in terms of accuracy and sensitivity. A key improvement of our approach consists of objective criteria to select significant signals associated with a condition that provides a confidence level on the discoveries made, which can be implemented in metabolic profiling studies.

Keywords Artificial mixtures · Data analysis · *t* test · PLS · NMR

1 Introduction

Metabolic profiling is underpinned by analytical techniques such as nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS). Spectral data are complex, typically consisting of many thousand variables providing information on hundreds of metabolites. The numerical complexity of the measurements hinders the direct quantification of significant change by visual inspection alone. Therefore, research on statistical methods for the analysis of metabolic profiles is essential to reliably identify metabolic biomarkers.

Discriminant analysis models and statistical hypothesis testing have been applied to robustly identify biomarkers that change significantly between conditions (Allen et al.

Electronic supplementary material The online version of this article (doi:10.1007/s11306-012-0422-8) contains supplementary material, which is available to authorized users.

A. C. Alves · J. V. Li · I. Garcia-Perez · C. Sands · E. Holmes (✉) · T. M. D. Ebbels (✉)
Section of Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK
e-mail: elaine.holmes@imperial.ac.uk

T. M. D. Ebbels
e-mail: t.ebbels@imperial.ac.uk

C. Barbas
CEMBIO (Center for Metabolomics and Bioanalysis), Pharmacy Faculty, Campus Monteprincipe, San Pablo-CEU University, 28668 Boadilla del Monte, Spain

2003; Bundy et al. 2009; Fiehn et al. 2000; Holmes et al. 2008). Among these methods partial least squares discriminant analysis (PLS-DA) and orthogonal PLS discriminant analysis (O-PLS-DA) are particularly common. O-PLS-DA is a latent variable regression model that has been applied to metabolic data to build predictive models identifying variables associated with particular biological conditions (Lindon et al. 2007; Trygg and Wold 2002). Discriminant algorithms which filter out variation unrelated to the biological class have proved useful, particularly for data sets in which the effect of interest may be quite subtle compared to the high degree of compositional variation often observed, for example, in human nutritional studies (Lindon et al. 2007). Although these methods have been extensively applied, their assessment using NMR spectra with known composition and concentration levels has been less explored and since the precise biochemical composition is unknown, many assumptions are made as to the behaviour of the mathematical procedures. In order to elucidate the accuracy of the statistical methods applied with respect to the analytical platform, the composition of the samples must be known a priori.

Several approaches can be used to assess statistical methodology, including simulated spectra, artificial mixtures and spike-in samples. Simulated spectra have been applied to analyse the impact of peak position and variable standardization on the analysis of metabolic profiles (Cloarec et al. 2005b; Craig et al. 2006; Dumas et al. 2006; Keun et al. 2002). However, it is hard to simulate spectra with all the chemical and statistical characteristics of real analytical data (such as peak overlap and positional variation), although relatively sophisticated software packages have recently become available for this purpose (Muncey et al. 2010). Spiking in known compounds to a biological matrix of known provenance offers an excellent approach combining both biological complexity and realistic spectral properties, while retaining knowledge about the true levels of the spiked in analytes. However, the main disadvantage of retaining the background matrix is that many compounds are already present in the sample, increasing the difficulty of compound quantification through peak overlap and the unknown levels of endogenous metabolites in the matrix. Use of stable isotope labelled standard compounds is a possible way to circumvent this problem, though acquiring them is often difficult and expensive. Artificial complex mixtures with known composition is an alternative strategy, intermediate between the former two approaches.

Previous work analysing the reproducibility of ^{1}H NMR data acquisition under different conditions and protocols (Dumas et al. 2006; Keun et al. 2002) has been based on ^{1}H NMR spectra of split biofluid samples of identical composition. Other studies have used simulated spectra to analyse the effect of spectral normalization and variable scaling (Craig et al. 2006) and the impact of matrix effects on O-PLS (Cloarec et al. 2005b). More recently, the predictive power of ^{1}H NMR spectral correlations for peak assignment were analysed using a set of known compounds in real biofluid spectra (Couto Alves et al. 2009). However, the performance of data analysis methods used in the process of biomarker discovery using artificial complex mixtures to better identify the qualitative and quantitative behaviour of metabolites analysed via various statistical algorithms, has not received much attention.

In this paper, an experimental design comprising artificial mixtures was developed for the purpose of comparing data analysis methods (see schematic diagram in supplementary information Fig. S1). The designed mixture incorporated endogenous and exogenous metabolites as well as xenobiotics commonly found in urine samples from human populations. Note that we do not attempt to emulate the true complexity of real biofluids, but rather to construct a sample set whose statistical characteristics are representative of multicomponent mixtures for the purpose of comparing data analysis methods. The metabolite concentration means and standard deviations were based on reference values for human urine derived from the Human Metabolome Database (HMDB) (Wishart et al. 2007). The dataset consisted of two classes of samples, assigned as ‘control’ and ‘disease’, with biologically realistic inter-metabolite correlations as well as differences in both concentrations and correlations between classes. This approach allowed us to construct a difficult separation problem with which to challenge the data analysis methods. By design, only a few metabolites in the sample had statistically significant differences between group means, analogous to real world clinical situations where disease signatures are often subtle and obscured by the high degree of compositional variability. Several metabolites without detectable ^{1}H NMR resonances were also included to simulate possible effects of a complex unobservable chemical background present in real mixtures. The acquired data were employed to evaluate the effectiveness of various data analysis methods. We cover both uni- and multivariate statistical methods, applied to both spectral and quantified metabolite data. Our comparison addresses the relative merits methods for the two different aspects of class discrimination and biomarker selection.

2 Methods

2.1 Concentrations of urinary metabolites in a normal human population

A sample of 21 metabolites representing endogenous metabolites and xenobiotics often observed in human urine were carefully chosen covering a wide range of metabolic

processes (see Table 1). The choice of metabolites was also guided by a desire to examine the relative impact of different analytical platforms on the recovery of metabolic information and associated data analysis techniques. Inter-platform comparisons will be discussed elsewhere; here we focus on the ^{1}H NMR data alone. Typical concentrations in human urine were obtained from the HMDB (Wishart et al. 2007) and the published literature (Saude et al. 2007). Mean concentrations for metabolites without reported mean levels were generated randomly from the distribution of those compounds whose mean concentrations were known. The concentration standard deviation for metabolites missing this information was predicted using a linear model fit to log transformed values for metabolites with known mean and standard deviation. In five cases, only concentration ranges, but not standard deviations were reported. For these metabolites, range values were converted to standard deviations by assuming that the range corresponded to a 95 % confidence interval around the mean concentrations using the expression: $\sigma = \frac{(x_{\max} - x_{\min})/2}{z_{95\%}}$ where is the 95 % percentile of the standard Normal distribution.

2.2 Experimental design

Two groups were constructed, 'normal' and 'disease', each comprising 30 samples. The metabolite concentrations for each sample were randomly generated and the groups differed in both their mean concentrations and correlations. (See Tables S5–S8 for details). The correlation matrices of normal and disease datasets were similar but seven metabolites had their correlations with other metabolites erased in the disease group by setting the population correlation to zero. In the absence of data from a representative human population, the correlations were generated following the distribution observed from in-house data of

Table 1 Metabolite concentration data for a human population and theoretical values for the synthetic dataset[‡]

Metabolite	Human population		Synthetic dataset			Class differences	
	(concentration μM/mM creatinine)		(Concentration μM) mean		std	Concentration	Correlation random (disease)
	Mean	std	Normal	Disease	Both	*t* test *p* value	
Hippurate	191.600	62.015	1730.148	1698.701	559.998	0.8506	No
Phenylacetylglycine (PAG)[b]	0.06	0.01	0.57	0.5	0.009	0.0004	Yes
Pipecolic acid[c]	0.030	0.026	0.307	0.424	0.120	0.0135	No
Indoxyl sulphate	14.000	2.857	126.420	117.390	25.800	0.1424	No
Trimethylamine-*N*-oxide (TMAO)	70.750	31.224	638.873	643.839	270.900	0.0610	Yes
Trimethylamine (TMA)[c]	7.700	7.550	68.628	77.658	28.896	0.0441	No
4-cresyl glucuronide[a]	0.601	0.134	5.425	5.986	1.209	0.1953	No
Paracetamol[a]	88.711	16.193	801.060	801.060	0	1.0000	No
Valine	2.650	0.561	24.020	25.764	5.068	0.8723	Yes
Alanine	33.900	6.505	306.117	320.219	58.741	0.1804	Yes
Creatinine[a]	9.030[a]	4.370[a]	9030.000	9030.000	0	1.0000	No
D-3-hydroxybutyrate (BHB)	35.600	5.765	330.498	321.468	52.061	0.3128	No
Citrate	226.700	51.684	2047.101	2047.101	466.704	0.6369	No
Succinate	12.150	2.526	109.715	114.682	22.805	0.4729	Yes
Guanidinoacetate (GAA)[c]	89.000	28.827	803.670	704.014	260.304	0.9407	No
Uric acid[c]	188.000	55.383	1697.640	2096.368	500.105	0.0163	No
Xanthine[c]	2.600	0.395	23.478	25.798	3.571	0.0000	Yes
N-methylnicotinamide[b]	11.243	2.229	101.524	111.656	20.128	0.7226	No
Isocitrate	38.900	7.339	351.267	283.596	66.268	0.0009	Yes
Ureidopropionate	2.23	0.365	20.047	21.672	3.296	0.0015	No
Tyrosine	10.900	1.913	98.427	102.791	17.277	0.9907	No

[a] Creatinine population metabolite levels are expressed in mM. The metabolites marked with 'b' were randomly generated from a probability distribution fitted to the distribution of known mean concentration levels. The population standard deviation of the metabolites marked with 'c' was estimated from range values. Paracetamol and creatinine concentration is constant across all samples. The citrate mean concentration is identical for both classes. For all metabolites, the standard deviation (std) is equal for both classes. 'Correlation random' denotes the variables having population correlation set to zero in the disease group

normal rat urine. Table 1 gives the means and standard deviations for each metabolite and provides the p value of the between-class t test as well as indicating the metabolites whose inter-metabolite correlations were erased.

In order to reduce the number of stock solutions as well as the experimental and sample preparation errors, the metabolite concentrations were quantized in ten levels with quanta optimized using the Lloyd-Max algorithm (Lloyd 2003) to minimize the mean square distortion of the quantized distribution. Of the 21 metabolites in Table 1, only 17 were expected to be observed in the NMR analysis owing either to the deliberately low concentrations (PAG, pipecolic acid and 4-cresyl glucuronide) or specific NMR-structural properties in the case of uric acid, which does not possess resolvable protons. The unobserved four compounds were retained in the design to assess the performance of other analytical platforms (not described in this paper) and to assess impact of background noise.

2.3 Artificial mixture sample preparation

All chemicals were purchased from Sigma-Aldrich (Poole, UK). The 21 individual compounds were carefully weighed and dissolved separately in 500 ml of water to form 21 'mother stock' solutions. Each mother stock was then diluted to obtain 10 stock solutions corresponding to the 10 quantized levels mentioned above. For each artificial mixture, 2 ml of the appropriate metabolite stock solution was added according to the mixture design, giving a final volume for each sample of 42 ml. A 400 µl aliquot of each sample was added to 200 µl of 0.2 M sodium phosphate buffer (D_2O:H_2O = 9:1, v:v, including 0.01 % of sodium 3-(trimethylsilyl) propionate-2,2,3,3-d4 [TSP] as a chemical shift reference, and 3 mM sodium azide preservative, pH = 7.4) and frozen at −80 °C prior to NMR analysis. Since the aim was to simulate metabolic behaviour in real biofluids, the pH of the sample mixtures was not adjusted.

2.4 NMR spectroscopy and preprocessing

^{1}H NMR spectra were acquired in a randomized order at 600.13 MHz with suppression of the water resonance using a standard pre-saturation pulse sequence 1D NOESY (90°-3 µs-90°-100 ms-90°-acquire). An exponential line-broadening filter (0.3 Hz) was applied prior to Fourier transformation and spectra were phased, baseline corrected and referenced automatically using an in-house MATLAB routine (version 7.4, The MathWorks, Natick, Massachusetts). MATLAB was used for all subsequent procedures. Unless otherwise stated, all spectra were normalized using probabilistic quotient normalization (Dieterle et al. 2006) prior to all analyses, including resonance quantification. Resonances were quantified (relative, not absolute) using peak height as this method had better performance than both curve fitting and numeric peak integration for this data set (see supplementary information Metabolite Quantification section). Multivariate data analysis was conducted using ^{1}H NMR spectra and quantified metabolites scaled to unit variance, unless otherwise stated. In the following we use the term 'quantified data' to refer to data in which each variable corresponds to a quantified peak. Similarly, we use the term 'spectral data' to refer to data in which each variable corresponds to an individual data point in the preprocessed spectrum.

2.5 PLS regression models and selection of influential variables

The following logistic regression extensions to PLS were assessed using Fort's matlab package (Fort 2005): RPLS combines PLS and Ridge penalized logistic regression (Fort and Lambert-Lacroix 2005); NR combines PLS to reduce dimensionality with proportional hazard regression model for survival analysis (Nguyen and Rocke 2002); IRPLS extends PLS into the framework of generalized linear regression using iterated weighted least-squares algorithm (Marx 1996); IRPLSF extends PLS in the context of generalized linear regression using iteratively reweighted PLS (Ding and Gentleman 2004). All models were fitted with three components and a classification threshold of 0.5. The number of PLS components was chosen through R^2_y curve analysis on a subset of 80 % of the sample size. The number of components used in all models was the same to make comparisons meaningful and also because results were very similar as all algorithms are essentially different implementation of PLS logistic regression. These models were compared to standard PLS-DA and O-PLS-DA implemented by in-house matlab scripts**.** O-PLS-DA (Trygg and Wold 2002) was parameterized with one predictive component and two orthogonal components (Cloarec et al. 2005b) using the same procedure. Prediction performance was assessed by the jackknife estimate of the coefficient of determination R^2_y, by predictivity and classification error rate (ER), both calculated using leave one out cross-validation. ER was estimated on the test set and is defined as the ratio between the total number of classification errors and the total number of samples classified.

The statistical significance of the regression coefficients was calculated using a one-sample t test with null hypothesis that the coefficient is zero using the jackknife estimate of the standard error. Multiple hypothesis testing correction was applied using Storey's false discovery rate. Variables were deemed influential if their corresponding coefficients were statistically significant (at $\alpha = 0.05$ and/or FDR = 0.1) and

their effect sizes were meaningful ($|B| > 0.01$). For high resolution ^{1}H NMR spectra, an additional constraint is required to consider a metabolite to be influential due to the high dimensionality and because metabolite resonances correspond to several spectral variables. Only metabolites with more than two influential ($\alpha = 0.05$ and/or FDR $= 0.1$, $|B| > 0.01$) spectral variables on the same resonance are considered to be putatively relevant for the description of the model of the condition. Regardless of this additional constraint, all influential spectral variables were analysed and depicted in figures because their visualisation is necessary to decide the minimum meaningful effect size. That is, varying a threshold on the absolute regression coefficient until the number of spurious signals on the base line is decimated can be used to mitigate false positives and self-calibrate the minimum effect size that can be detected above noise level. In this study, to make results comparable across methods we have kept this threshold constant. The accuracy of the variable selection process was judged by comparing results obtained using volcano plots on quantified data with expected results from the experimental design. We have judged success in two ways: first comparing the univariate analyses (including volcano plots) with the theoretical design; secondly we have used the volcano plots as the 'truth' to judge success of the multivariate methods. That is, we expect multivariate methods to select influential variables from the set of metabolites that show significant and meaningful changes on the observed data as opposed to using only theoretical data. Multivariate regression methods aim at identifying a set of variables that can accurately predict a dependent variable (class membership in this case) while univariate *t* test aims to identify mean differences between groups of data. Although some correspondence of results is expected, it does not necessarily follow that multivariate regression will select all variables found to be significant in a univariate test.

3 Results

3.1 NMR spectra and metabolite peak assignment

Figure 1 illustrates overlays of the metabolite resonances observed in the NMR spectra. Of the identified metabolites, 13 could be accurately quantified (76 % of all identified metabolites). The remaining four (indoxyl sulphate, xanthine, ureidopropionate and tyrosine) produced resonances that were either too small or overlapped. Details of quantified metabolites are shown in supplementary information Table S1.

As expected, some additional signals were observed, possibly due to impurities and/or contaminants; these are not included in the statistical analysis. Due to the high concentration of creatinine, its ^{13}C satellites for both the 3.04 and 4.05 ppm resonances could be clearly identified. An unknown peak with very low intensity was found in ten samples (singlet at 3.71 ppm), this could be due to contaminants in the laboratory environment or NMR tubes.

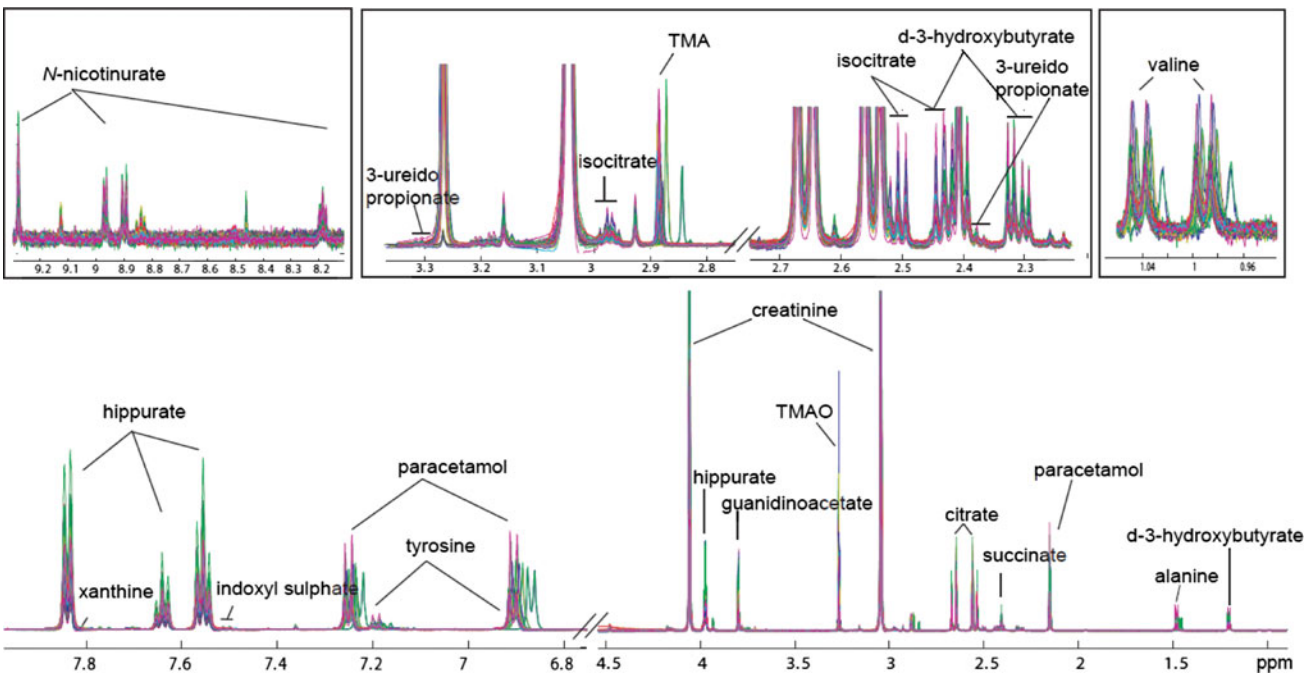


Fig. 1 ^{1}H NMR spectra of all samples of the artificial mixture with assigned resonances. Abbreviations: *TMA* trimethylamine; *TMAO* trimethylamine *N*-oxide

Two samples exhibited large peak position shifts of some metabolites (e.g. citrate and paracetamol) owing to their high pH, as expected from their known composition.

3.2 Differential analysis of quantified metabolite concentration levels

A range of statistical tests were applied to the quantified and theoretical concentrations to determine the extent to which the designed class differences in mean concentrations were reproduced in the experimental data. The results for the Welch-Student t test showed a good agreement between both datasets as shown in Table 2.

The statistical significance was correctly inferred for 12 out of 13 metabolites. Only paracetamol is unexpectedly declared as a differentially concentrated metabolite. The difference between group mean intensities was relatively small, corresponding to less than 5 % of the mean value (1.21 and 1.27 for normal and disease respectively) and the standard deviation very low (as expected), producing the significant p value. This anomalous finding could be explained in part by spectral normalization since paracetamol quantified in non normalised spectra was not significantly different ($p = 0.11$) and also by chance since random fluctuations could produce a significant result if 21 metabolites are analysed at a significance threshold of $\alpha = 0.05$ (probability of at least one significant result $= 1 - (1 - 0.05)^{21} = 0.66$). After applying a Storey false discovery rate correction (FDR $= 0.05$), it was possible to correctly identify all metabolites with significant differences. Other tests, including Mann–Whitney, bootstrap t and permutation tests gave broadly similar results but were not as successful in identifying the differential metabolites (see supplementary information Table S2). The standard deviation in both normal and disease classes was almost identical as intended in the original design.

Figure 2 shows volcano plots comparing the $\log_2$ fold change and the statistical significance from the t test for the

Table 2 Differential concentration analysis of the quantified and theoretical concentrations

Metabolite	p value (t test)		Quantified mean		Quantified standard deviation	
	Theoretical	Quantified	Normal	Disease	Normal	Disease
Valine	0.872	0.505	0.071	0.074	0.02	0.02
D-3-hydroxybutyrate	0.313	0.935	1.044	1.040	0.18	0.18
Alanine	0.180	0.075	0.851	0.932	0.16	0.18
Succinate	0.473	0.217	1.045	1.104	0.17	0.20
Citrate	0.637	0.950	4.794	4.812	1.13	1.16
Creatinine	1.000	0.253	49.848	50.986	3.66	3.96
TMAO	0.061	0.090	12.278	10.124	4.97	4.71
Guanidinoacetate	0.941	0.652	3.269	3.410	1.36	1.01
TMA	**0.044**	**0.010**	0.815	1.045	0.33	0.34
Isocitrate	**0.001**	**0.005**	0.193	0.163	0.04	0.04
Hippurate	0.851	0.829	2.337	2.381	0.84	0.73
Paracetamol	1.000	**0.02**	1.210	1.270	0.09	0.10
NMND	0.723	0.571	0.069	0.065	0.03	0.03

In the p values column, the entries in *bold* are statistically significant at $\alpha = 0.05$

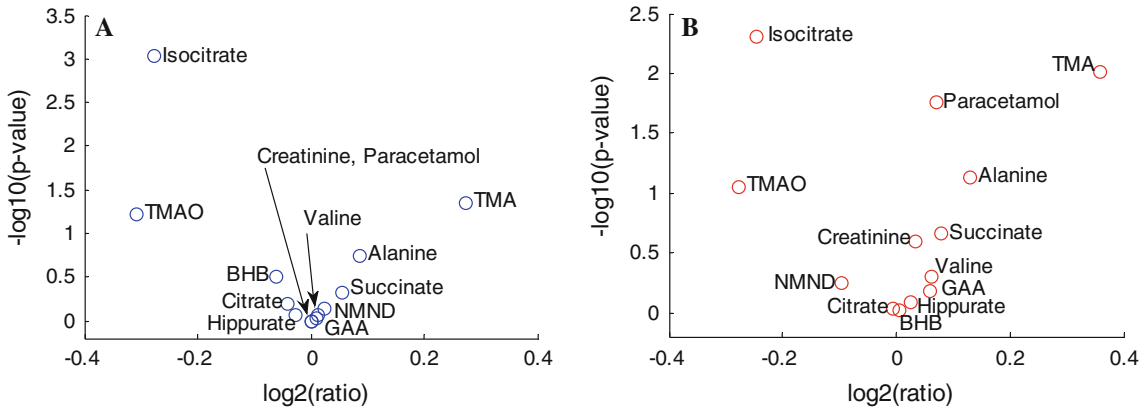


Fig. 2 Analysis of theoretical versus quantified effect size and p value. *Volcano plot* of the log intensity ratio versus log p value: **a** theoretical and **b** quantified concentration using Welch t test

Table 3 Comparison of regression models performance and variable selection accuracy on quantified concentrations

Metabolite	PLS		O-PLS		RPLS		IRPLSF		NR		IRPLS	
	ER 0.33 B^J	$Q^2_Y(R^2_Y)$ 0.11 (0.40) *p* value (FDR)	ER 0. 33 B^J	$Q^2_Y(R^2_Y)$ **0.15** (0.40) *p* value (FDR)	ER 0.33 B^J	$Q^2_Y(R^2_Y)$ 0.13 (0.39) *p* value (FDR)	ER 0.33 B^J	$Q^2_Y(R^2_Y)$ 0.08 (0.41) *p* value (FDR)	ER 0.33 B^J	$Q^2_Y(R^2_Y)$ 0.02 (0.42) *p* value (FDR)	ER **0.32** B^J	$Q^2_Y(R^2_Y)$ 0.053 **(0.45)** *p* value (FDR)
Valine	0.02	0.82 (0.58)	0.07	0.49 (0.49)	0.06	0.82 (0.82)	0.04	0.92 (0.69)	0.09	0.86 (0.71)	0.11	0.86 (0.82)
BHB	−0.01	0.85 (0.58)	−0.01	0.93 (0.62)	−0.04	0.85 (0.82)	−0.06	0.86 (0.69)	−0.06	0.89 (0.71)	0.04	0.95 (0.82)
Alanine	0.08	0.26 (0.38)	0.18	**0.04** (*0.08*)	0.31	0.27 (0.62)	0.36	0.35 (0.60)	0.54	0.34 (0.63)	0.70	0.33 (0.67)
Succinate	0.00	0.97 (0.62)	0.12	0.13 (0.20)	−0.01	0.97 (0.82)	−0.05	0.89 (0.69)	−0.09	0.87 (0.71)	−0.10	0.89 (0.82)
Citrate	−0.09	0.22 (0.38)	0.01	0.95 (0.62)	−0.37	0.20 (0.57)	−0.55	0.26 (0.60)	−0.76	0.23 (0.63)	−1.03	0.14 (0.42)
Creatinine	0.02	0.65 (0.56)	0.12	0.20 (0.26)	0.09	0.65 (0.80)	0.14	0.52 (0.60)	0.13	0.74 (0.71)	−0.20	0.71 (0.82)
TMAO	−0.03	0.70 (0.56)	−0.17	**0.05** (*0.1*)	−0.11	0.70 (0.80)	−0.11	0.75 (0.69)	−0.10	0.84 (0.71)	−0.15	0.82 (0.82)
Guanidinoacetate	−0.06	0.37 (0.43)	0.05	0.64 (0.53)	−0.24	0.37 (0.63)	−0.34	0.36 (0.60)	−0.53	0.37 (0.63)	−0.49	0.28 (0.67)
TMA	0.14	**0.01** (*0.03*)	0.26	**0.01** (*0.03*)	0.55	**0.02** (*0.08*)	0.73	0.06 (0.20)	0.98	**0.04** (0.16)	1.16	0.06 (0.25)
Isocitrate	−0.22	**9.3×10^{-6}** (4×10^{-5})	−0.28	**0.03** (*0.08*)	−0.90	**4.5×10^{-5}** (4.5×10^{-5})	−1.12	**0.01** (0.11)	−1.68	**0.01** (0.12)	−1.85	**0.04** (0.22)
Hippurate	0.05	0.42 (0.43)	0.02	0.83 (0.62)	0.19	0.42 (0.63)	0.25	0.49 (0.60)	0.36	0.46 (0.63)	0.41	0.53 (0.82)
Paracetamol	0.15	**0.01** (*0.03*)	0.24	**2.7×10^{-3}** (*0.02*)	0.61	**0.01** (*0.04*)	0.76	**0.02** (0.11)	1.13	**0.02** (0.12)	1.20	**0.02** (0.18)
NMND	−0.05	0.44 (0.43)	−0.06	0.58 (0.53)	−0.20	0.44 (0.63)	−0.22	0.50 (0.60)	−0.36	0.45 (0.63)	−0.19	0.82 (0.82)
Intercept	0.50	**1.5×10^{-11}** (1×10^{-10})	−0.10	0.33 (0.38)	0.00	1.00 (0.82)	−0.05	0.87 (0.69)	−0.06	0.89 (0.71)	−0.10	0.86 (0.82)

ER and Q^2_Y are respectively the out-of-sample ER and predictivity calculated with leave-one-out cross validation. R^2_Y and B^J are the jackknife estimates of the coefficient of determination and the regression coefficients respectively. The best performance scores, statistically significant regression coefficients and the variables selected at least once are shown in *bold*. FDR values smaller than 0.1 are in *italics*

theoretical and quantified data. Metabolites with large effect size (isocitrate, TMAO, TMA, alanine) show similar ratios and p values in theoretical and quantified data, except for paracetamol as discussed above.

Overall, the comparison shows that the volcano plot is an effective tool to distinguish the significant and important metabolites that change between conditions and that the dataset is very challenging for any statistical procedure (as intended). In the following we intend to compare methods. As expected, owing to technical issues, there are some differences between the theoretical and observed data. Therefore, instead of comparing theoretical versus observed data we consider influential metabolites to be correctly identified if they are consistent with the volcano plot analysis (i.e. the top five metabolites corresponding to $p < 0.09$ in Fig. 2).

3.3 Differential analysis of high resolution ^{1}H NMR spectral data

A similar approach as that described in the last section for identifying differentially concentrated metabolites was applied to the ^{1}H NMR spectral intensities. However, an additional threshold on the effect size was also considered; only spectral variables with absolute $\log_2$ fold change >0.5 and FDR < 0.1 were considered influential. A bootstrap as well as a parametric Welch t test with Storey's FDR correction was applied to detect significant differences between the group means of all NMR signals. Metabolites with more than two significant spectral variables in the same resonance were considered discriminatory and are labelled in supplementary information Fig. S3. Metabolites designed to be statistically significant were detected by both techniques and surprisingly the parametric Welch test produced more conservative results than the bootstrap version. The metabolites identified as differential using the Welch t test both in ^{1}H NMR spectra and quantified data were similar. Paracetamol was considered correctly identified because the result obtained in ^{1}H-NMR matches the quantified analysis. On the other hand, the bootstrap approach was more sensitive additionally detecting alanine, which is almost statistically significant ($p = 0.075$) in quantified data.

3.4 Comparison of regression models for discriminant analysis with quantified data

A comparison of PLS linear and logistic regression models was performed to select the most suitable methods for the analysis of quantified metabolic profiles. The performance criteria were out-of-sample prediction accuracy as well as the ability to accurately identify the metabolites that change concentration between conditions. Results in terms of variable selection were very satisfactory (see Table 3), and largely agree with the volcano plot of the quantified metabolites (Fig. 2).

Fig. 3 Comparison of regression models applied to ^{1}H NMR spectra ▶ with signals of the statistically significant coefficients depicted in red (FDR < 0.1). Significant peaks are identified and labelled. Identified peaks consistent with volcano plot analysis are marked with a *green tick* or with a *red cross* otherwise. All models were applied to ^{1}H NMR spectra scaled to unit variance except panel E. **a** PLS, **b** RPLS, **c** IRPLSF, **d** NR. Results for O-PLS and IRPLS are omitted because no statistically significant signals were found using unit variance scaling. **e** O-PLS DA coefficients fit to mean centred spectra using the method of (Cloarec et al. 2005a) for variable selection where the *colour scale* indicates the squared correlation to the class variable (0 or 1) and only peaks with are annotated (Color figure online)

Individually, the regression methods had distinct characteristics both in terms of performance and variable selection accuracy. PLS, O-PLS and RPLS achieved similarly high predictivity and selected coherent sets of variables with significant FDR ($\alpha < 0.05$ and FDR < 0.1). However, in terms of variable selection, O-PLS was the most sensitive, identifying five influential metabolites, which were the same top five metabolites as identified in the volcano plot (Fig. 2) with $p < 0.09$.

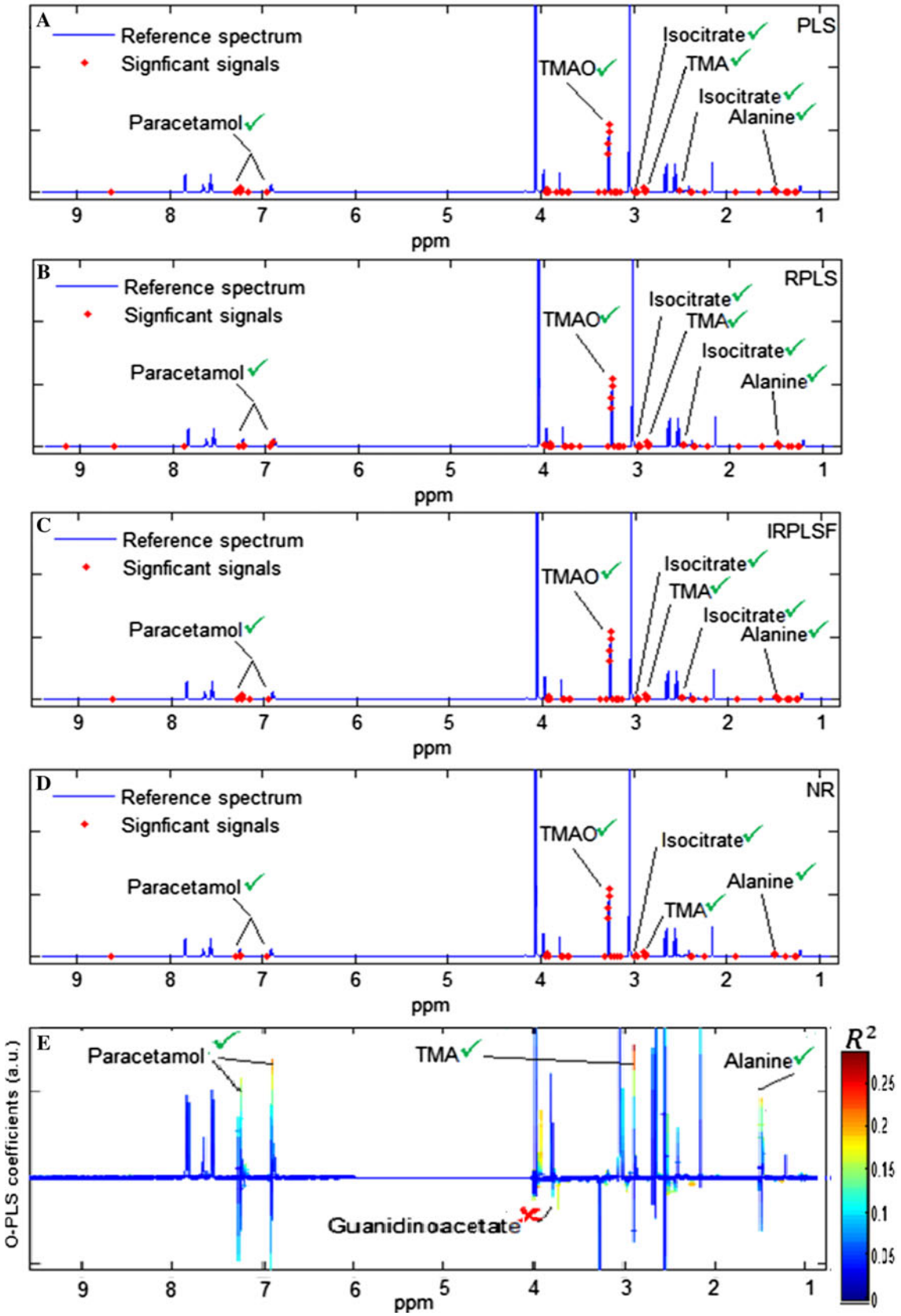
Overall, despite the very large dynamic range in mean peak height of the quantified metabolites ([0.065, 50.9] arbitrary units), all models consistently selected the differential metabolites while achieving reasonable out-of-sample classification ER. Logistic regression models did not produced significantly better out-of-sample ER or goodness of fit than linear regression.

3.5 Comparison of regression models for discriminant analysis with high resolution spectral data

Next we extended the comparison of regression models to the high resolution spectral data and compared this with the same models fit to quantified data. Interestingly, all methods performed significantly better on spectral than on quantified data in terms of predictivity, ER and coefficient of determination (see Table S3 in supplementary information).

Surprisingly, variable selection accuracy was similar to results obtained in quantified data for most methods (Fig. 3a–d). O-PLS and IRPLS did not select any significant signals (FDR < 0.1) using the proposed approach based on FDR correction of the coefficient p values and are therefore not presented. We report the reanalysis of the O-PLS model using the conventional approach with mean centred data in the next section (see Fig. 3e).

Spuriously selected variables could be easily discarded as they typically consisted of isolated data points and did not correspond to a visible NMR peak (see supplementary information Fig. S5). Although a fixed threshold on the level of the regression coefficient was applied to make


A
Reference spectrum
Signficant signals
PLS
Paracetamol
TMAO
Isocitrate
TMA
Isocitrate
Alanine
ppm
B
Reference spectrum
Signficant signals
RPLS
Paracetamol
TMAO
Isocitrate
TMA
Isocitrate
Alanine
ppm
C
Reference spectrum
Signficant signals
IRPLSF
Paracetamol
TMAO
Isocitrate
TMA
Isocitrate
Alanine
ppm
D
Reference spectrum
Signficant signals
NR
Paracetamol
TMAO
Isocitrate
Alanine
TMA
ppm
E
O-PLS coefficients (a.u.)
Paracetamol
TMA
Alanine
Guanidinoacetate
ppm
R^2
0.25
0.2
0.15
0.1
0.25
0

results comparable, varying the threshold until spurious significant results on the baseline are decimated can be used to calibrate the minimum effect size that can be detected above noise level, reducing even further the number of spurious signals. Overall, this approach to variable selection was capable of identifying all metabolites that change concentration levels between conditions significantly. Results for PLS, RPLS, IRPLSF and NR were very similar, selecting more variables in ^{1}H NMR spectra than in quantified data. These results confirm that analysis of high resolution spectral data can reveal additional sample content information even when quantified data is available.

3.6 Comparison to conventional variable selection using O-PLS discriminant analysis

Orthogonal PLS discriminant analysis models were fit to mean centred spectra and variable selection was performed according to the approach proposed by (Cloarec et al. 2005a) (Fig. 3e). Despite low predictivity ($Q^2_Y = -0.04$), the model correctly identified TMA, paracetamol and alanine. Guanidinoacetate was incorrectly selected. In order to assess the impact of variable scaling and peak alignment, the same analysis was conducted on unit-variance scaled spectra but no significant differences on variable selection accuracy were observed as opposed to peak alignment where performance declined (see supplementary Fig. S6).

Although the conventional approach to O-PLS model analysis successfully identified metabolites changing significantly, it was less sensitive and accurate than either O-PLS results on quantified data (Table 3) or the other models on ^{1}H NMR spectra (Table S3; Fig. 3). Conversely, the main advantage of the jackknife method used here is an objective cut-off for variable selection that provides confidence levels for individual signals. As with any approach, spurious results may occur, especially with low intensity signals. This can be mitigated by controlling the false discovery rate and visualising the significant variables in a spectral format as in Fig. 3. This aids the practitioner in discriminating spurious from true hits corresponding to peak resonances (as shown in the supplementary information Fig. S5). A brief comparison of the variables selected by the multivariate methods on quantified and spectral data is provided in supplementary information Table S4.

4 Discussion and conclusions

Previous work on the assessment of data analysis methods on ^{1}H NMR spectra focused mainly on simulated data or data from biological samples. These approaches can reveal important information regarding the spectral characteristics that impact data analysis methods. However, assessment of data analysis methods for metabolite selection is hampered owing to a lack of knowledge of the exact sample composition. In this study, an artificial complex mixture design was implemented. The metabolite concentrations reflect levels typically found in human urine and have a well defined correlation structure. This design addresses important biomarker discovery issues such as technical variation and biological variation, as well as the effect of concentration distribution and correlation structure in different biological conditions. This data is therefore particularly attractive to assess univariate and multivariate statistical methodologies that are applied to metabolomics datasets. The ^{1}H NMR spectra from these samples accurately reproduced the planned characteristics, as demonstrated by the analysis of the metabolite concentration profiles including statistical tests for differences between groups.

This work assessed methods for the analysis of metabolite concentration differences. Volcano plots provided an excellent method to track the effect size and significance of the change between conditions, allowing comparison of metabolite behaviour of the quantified and theoretical data, and offering a degree of protection against very small but nominally significant effects. On quantified data, the Welch-Student *t* test was most successful at identifying the metabolites designed to differentiate the conditions. Other tests generated a number of false positive or negative results. Interestingly, the Welch *t* test detected similar differential metabolites in both quantified and spectral data suggesting that differential analysis of full resolution ^{1}H NMR spectra using FDR correction, taking into account fold changes, is a reliable approach to detect differential metabolites in complex mixture studies.

Various multivariate regression methods were applied in discriminant analysis mode and a jackknife based strategy for variable selection was assessed on both quantified and spectral data. The results indicate that it may be possible to improve on the conventional methodology of (Cloarec et al. 2005b) in terms of accuracy and sensitivity. As expected, models fit to spectral ^{1}H NMR spectra could recover important information beyond that in the quantified data (e.g. variation of signals not incorporated in the design) and this lends weight to the strategy that discriminant analysis of quantified data should be complemented with analysis of spectral spectra. A key element of our approach consists of objective criteria to select significant signals associated with a condition that provides a confidence level on the discoveries made. These results extend earlier findings of (Chadeau-Hyam et al. 2010; Westerhuis et al. 2008) because our method does not rely on permutation analysis which is potentially faster for typical sample sizes, reduces the bias of the coefficients as a consequence of the jackknife estimator and provides confidence

intervals for the regression coefficients. The most reliable methods in quantified and spectral ^{1}H NMR data were PLS and RPLS linear and logistic regression respectively. We suggest the use of logistic regression models in discriminant analysis owing to enhanced interpretability since logistic regression coefficients provide direct information on the odds ratio of each variable. Logistic regression methods can potentially achieve lower ERs over simple linear regression methods owing to the linkage function and error model.

Some limitations are also worth noting. Even in a designed experiment several challenges were observed. The pH of some samples was inevitably high due to particular combinations of concentrations. This produced significant variation of the resonance peak position in two spectra. Local and global peak alignment methods can be used to mitigate this effect, but analyses conducted on this data suggest that peak alignment may introduce artefacts. In this experimental design using artificial mixtures without a complex biological background, peak height showed good quantification accuracy. However, in spectra of complex real biofluids, quantification methods based on peak fitting may be preferable due to variable line widths and difficulties resolving overlapped peaks. Nevertheless, the data generated by this experimental design is a formidable resource for future work including cross-platform comparison and integration of analytical techniques and for the development of new data analysis methods.

Acknowledgments Alexessander Couto Alves acknowledges an Imperial College Faculty of Medicine PhD studentship.

References

Allen, J., Davey, H. M., Broadhurst, D., Heald, J. K., Rowland, J. J., Oliver, S. G., et al. (2003). High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology, 21*, 692–696.

Bundy, J., Davey, M., & Viant, M. (2009). Environmental metabolomics: a critical review and future perspectives. *Metabolomics, 5*, 3–21. doi:10.1007/s11306-008-0152-0.

Chadeau-Hyam, M., Ebbels, T. M. D., Brown, I. J., Chan, Q., Stamler, J., Huang, C. C., et al. (2010). Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *Journal of Proteome Research, 9*, 4620–4627. doi:10.1021/pr1003449.

Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., et al. (2005a). Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Analytical Chemistry, 77*, 1282–1289. doi:10.1021/ac048630x.

Cloarec, O., Dumas, M. E., Trygg, J., Craig, A., Barton, R. H., Lindon, J. C., et al. (2005b). Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in 1H NMR spectroscopic metabonomic studies. *Analytical Chemistry, 77*, 517–526. doi:10.1021/ac048803i.

Couto Alves, A., Rantalainen, M., Holmes, E., Nicholson, J. K., & Ebbels, T. M. (2009). Analytic properties of statistical total correlation spectroscopy based information recovery in (1)H NMR metabolic data sets. *Analytical Chemistry,*. doi:10.1021/ac801982h.

Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry, 78*, 2262–2267.

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. *Analytical Chemistry, 78*, 4281–4290.

Ding, B., & Gentleman, R. (2004). Classification using generalized partial least squares. Bioconductor Project Working Papers, 5.

Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., et al. (2006). Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry, 78*, 2199–2208. doi:10.1021/ac0517085.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat Biotech, 18*, 1157–1161.

Fort, G. (2005). Inference in logistic regression models. http://perso.telecom-paristech.fr/~gfort/GLM/Programs.html.

Fort, G., & Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics, 21*, 1104–1111. doi:10.1093/bioinformatics/bti114.

Holmes, E., Loo, R. L., Stamler, J., Bictash, M., Yap, I. K. S., Chan, Q., et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature, 453*, 396–400.

Keun, H. C., Ebbels, T. M. D., Antti, H., Bollard, M. E., Beckonert, O., Schlotterbeck, G., et al. (2002). Analytical reproducibility in 1H NMR-based metabonomic urinalysis. *Chemical Research in Toxicology, 15*, 1380–1386. doi:10.1021/tx0255774.

Lindon, J., Nicholson, J., & Holmes, E. (2007). *The handbook of metabonomics and metabolomics*. Amsterdam: Elsevier Science.

Lloyd, S. (2003). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*, 129–137.

Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics, 38*, 374–381.

Muncey, H., Jones, R., De Iorio, M., & Ebbels, T. (2010). MetAssimulo: simulation of realistic NMR metabolic profiles. *BMC Bioinformatics, 11*, 496.

Nguyen, D. V., & Rocke, D. M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics, 18*, 1216–1226. doi:10.1093/bioinformatics/18.9.1216.

Saude, E. J., Adamko, D., Rowe, B. H., Marrie, T., & Sykes, B. D. (2007). Variation of metabolites in normal human urine. *Metabolomics, 3*, 439–451.

Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics, 16*, 119–128.

Westerhuis, J., Hoefsloot, H., Smit, S., Vis, D., Smilde, A., van Velzen, E., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics, 4*, 81–89. doi:10.1007/s11306-007-0099-6.

Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., et al. (2007). HMDB: the human metabolome database. *Nucleic Acids Research, 35*, D521–D526.