

## A comprehensive urinary metabolomic approach for identifying kidney cancer

Tobias Kind<sup>a</sup>, Vladimir Tolstikov<sup>a</sup>, Oliver Fiehn<sup>a</sup>, Robert H. Weiss<sup>b,c,\*</sup>

<sup>a</sup> Genome Center, University of California, Davis, CA 95616, USA

<sup>b</sup> Division of Nephrology, Department of Internal Medicine, University of California, Davis, CA 95616, USA

<sup>c</sup> Department of Veterans Affairs, Northern California Health Care System, Sacramento, CA 95655, USA

Received 28 September 2006

Available online 26 January 2007

### Abstract

The diagnosis of cancer by examination of the urine has the potential to improve patient outcomes by means of earlier detection. Due to the fact that the urine contains metabolic signatures of many biochemical pathways, this biofluid is ideally suited for metabolomic analysis, especially involving diseases of the kidney and urinary system. In this pilot study, we test three independent analytical techniques for suitability for detection of renal cell carcinoma (RCC) in urine of affected patients. Hydrophilic interaction chromatography (HILIC–LC–MS), reversed-phase ultra performance liquid chromatography (RP–UPLC–MS), and gas chromatography time-of-flight mass spectrometry (GC–TOF–MS) all were used as complementary separation techniques. The combination of these techniques is best suited to cover a very large part of the urine metabolome by enabling the detection of both lipophilic and hydrophilic metabolites present therein. In this study, it is demonstrated that sample pretreatment with urease dramatically alters the metabolome composition apart from removal of urea. Two new freely available peak alignment methods, MZmine and XCMS, are used for peak detection and retention time alignment. The results are analyzed by a feature selection algorithm with subsequent univariate analysis of variance (ANOVA) and a multivariate partial least squares (PLS) approach. From more than 2000 mass spectral features detected in the urine, we identify several significant components that lead to discrimination between RCC patients and controls despite the relatively small sample size. A feature selection process condensed the significant features to less than 30 components in each of the data sets. In future work, these potential biomarkers will be further validated with a larger patient cohort. Such investigation will likely lead to clinically applicable assays for earlier diagnosis of RCC, as well as other malignancies, and thereby improved patient prognosis.

© 2007 Elsevier Inc. All rights reserved.

**Keywords:** Metabolomics; Kidney cancer; Urine analysis; Chromatography; Mass spectrometry

Most current strategies for cancer biomarker detection use a proteomics approach [1] in which plasma (or another biofluid) from cancer patients is analyzed for discrete proteins or for specific changes in their levels as compared with normal individuals. Although basic bio-

chemical techniques applied to urine analysis have long been used to aid in the diagnosis of a multitude of renal and metabolic diseases, such as the nephrotic syndrome and diabetes, examination of the urine for the diagnosis of latent malignancies has not been well investigated. Of the human malignancies, urologic cancers seem to be most suitable for developing a urinary diagnostic assay given that this organ system is in intimate contact with this biofluid. In addition, such early diagnosis is an area of obvious benefit given the poor prognosis (<10% 5-year survival) of patients with metastatic clear cell renal

\* Corresponding author. Fax: +1 530-752-3791.

E-mail address: [rhweiss@ucdavis.edu](mailto:rhweiss@ucdavis.edu) (R.H. Weiss).

cell carcinoma (ccRCC),<sup>1</sup> the most common type of kidney cancer [2].

Gas chromatography coupled to mass spectrometry (GC–MS) has been used on a large scale for more than 30 years in clinical biochemistry as well as for urine analysis [3]. Liquid chromatography coupled to mass spectrometry (LC–MS) has been used for urine analysis for more than 25 years, and a review from 1973 already cited several hundred applications for LC–MS and GC–MS [4]. Recent advances in column technology, such as hydrophilic interaction chromatography (HILIC) coupled to electrospray mass spectrometry (ESI–MS), allow the detection of highly polar compounds that also occur in urine [5]. The development of monolithic capillary columns [6] and ultra high-pressure LC–MS or ultra performance liquid chromatography (UPLC) [7] introduced high chromatographic peak resolution to LC; thus, less overlap of components and high-quality mass spectra can be obtained in a very short time. Monolithic LC columns and UPLC can have a very high column efficiency, which in the past was reached only by capillary GC columns.

In this pilot study, we describe a comprehensive metabolomic approach in which we use three distinct analytical chemistry methods to examine the complement of small molecules, including primary and secondary metabolites up to 2000 Da in size, in the urine of ccRCC patients. The urine metabolome is highly variable with respect to the composition and quantity of compounds in response to diet, medication, and metabolic state. Here we suggest using comprehensive metabolomic analysis to examine thousands of compounds without complete structural identification to discern patterns of metabolic end products that show specificity for ccRCC and ultimately to aid in the diagnosis of this disease.

It was expected that alterations in compound structures would disable metabolite target-driven analytical approaches. Hence, each chromatogram needed to be analyzed in an unbiased way with the aim of detecting as many metabolic features as possible. Many software packages are commercially available for such biomarker detection. Approximately 20 available alignment programs for GC–MS, LC–MS, nuclear magnetic resonance (NMR), and MS are known [8]. In this study, we use two freely available

<sup>1</sup> *Abbreviations used:* ccRCC, clear cell renal cell carcinoma; GC–MS, gas chromatography coupled to mass spectrometry; LC–MS, liquid chromatography coupled to mass spectrometry; HILIC, hydrophilic interaction chromatography; ESI–MS, electrospray mass spectrometry; UPLC, ultra performance liquid chromatography; NMR, nuclear magnetic resonance; FT–MS, Fourier transform mass spectrometry; FT–IR, Fourier transform infrared spectroscopy; MSTFA, N-methyl-N-trimethylsilyltrifluoroacetamide; TOF, time-of-flight; ALEX, automated liner exchange; RP, reversed-phase; PLS, partial least squares; ANOVA, analysis of variance; PCA, principal component analysis; NIPALS–PLS, nonlinear iterative partial least squares; TCA, tricarboxylic acid; SIMCA, soft independent modeling of class analogy; PLS–DA, partial least squares–discriminant analysis; O–PLS, orthogonal partial least squares; ICA, independent component analysis; hKIM-1, human kidney injury molecule-1.

tools, XCMS and MZmine, for chromatogram alignment. The Java application MZmine [8] can be used to visually examine chromatograms and has a number of noise filters, peak picking, and alignment procedures implemented. In early applications, this program was used for metabolic profiling of secondary plant metabolites. The software XCMS is implemented with the freely available R statistical language, allowing selection of important peaks and/or components, and uses a *t* test and the resulting *p* values to rank important components. This program previously has been used successfully for an enzyme knockout study and a large-scale study of 238 plasma samples from mice [9].

The ultimate goal in biomarker identification is, of course, to assign a molecular structure for each of the significant biomarkers (Fig. 1). In the case of metabolomics, this can be done using known analytical chemistry techniques such as NMR, Fourier transform mass spectrometry (FT–MS), LC–MS, GC–MS, and Fourier transform infrared spectroscopy (FT–IR) [10]. Here the approach using an orthogonal separation process (LC), as opposed to MS alone, also helps to identify these components because liquid fractions can be collected with a fraction collector and then later NMR can be used for further structure elucidation. Ultimately, such potential biomarkers will be identified and validated in large human cohort studies (“slow lane” in Fig. 1) to devise a rapid and inexpensive clinical assay for the diagnosis of early-stage RCC as well as to assess responses to specific therapeutic interventions.

## Materials and methods

### Sample procurement

After obtaining approval from the institutional review board, six random urine samples from ccRCC patients of various ages and genders, and at various stages and grades, were obtained from the Cooperative Human Tissue Network at Vanderbilt University. All urine samples were obtained prior to chemotherapy, radiation, or nephrectomy and were immediately frozen at  $-20$  or  $-80$  °C. Six random control urine samples were obtained from healthy volunteers. The hypothesis behind this limited study was that the randomness of patients and controls would enhance the likelihood of finding higher probabilities for general ccRCC disease-related biomarkers than would any effort in tightly matching the individuals that invariably would still be different in diets and other environmental factors even if age, gender, and habits had been more tightly controlled. Such controlled cohort studies will be necessary in the future for validation studies that are needed to confirm the initial findings of potential biomarkers.

### Analytical data

#### GC–MS analysis

The derivatization procedure has been described previously [11–13]. Briefly, neat urine samples were lyophilized

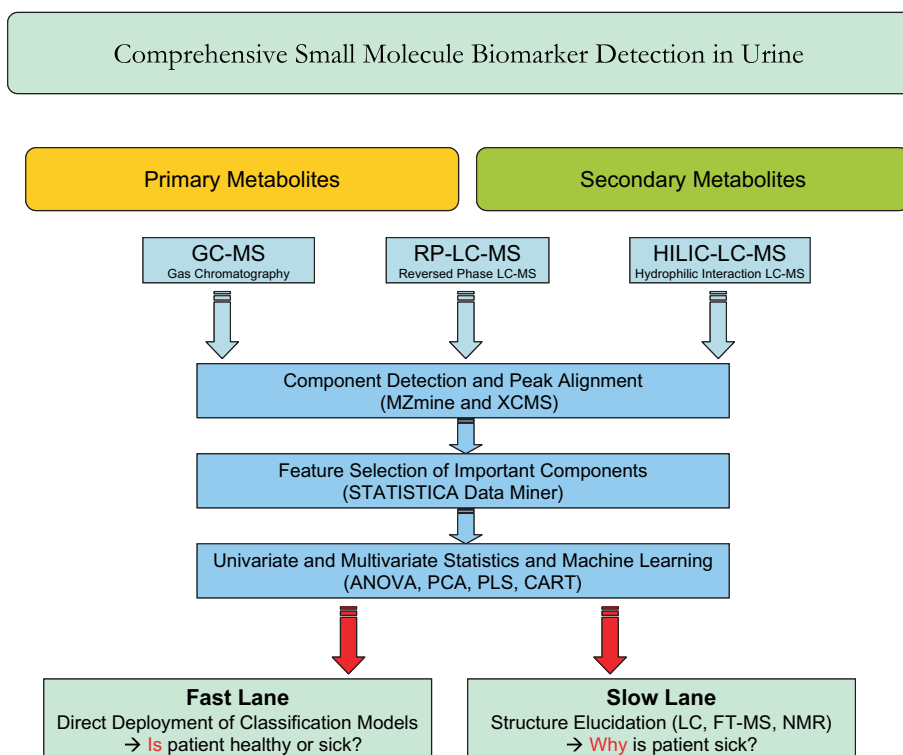


Fig. 1. Algorithm describing the metabolomic approaches for biomarker identification using complementary analytical techniques covering the whole metabolome or small molecule space.

without further pretreatment after our initial finding of severe alterations using urease treatments. To the dried samples, 20  $\mu\text{l}$  of 40 mg/ml methoxylamine hydrochloride in pyridine was added, and samples were agitated at 30  $^{\circ}\text{C}$  for 30 min. Subsequently, 180  $\mu\text{l}$  of trimethylsilylating agent *N*-methyl-*N*-trimethylsilyltrifluoroacetamide

(MSTFA) was added, and samples were agitated at 37  $^{\circ}\text{C}$  for 30 min. GC–MS analysis [11–13] was performed using a Agilent 6890 N gas chromatograph (Atlanta, GA, USA) interfaced to a time-of-flight (TOF) Pegasus III mass spectrometer (Leco, St. Joseph, MI, USA). Automated injections were performed with a programmable robotic Gerstel MPS2 multipurpose sampler (Mülheim an der Ruhr, Germany). The GC was fitted with both an Agilent injector and a Gerstel temperature-programmed injector, cooled injection system (model CIS 4), with a Peltier cooling source. An automated liner exchange (ALEX) designed by Gerstel was used to eliminate cross-contamination from sample matrix occurring between sample runs. Multiple baffled liners for the GC inlet were deactivated with 1- $\mu\text{l}$  injections of MSTFA. The Agilent injector temperature was held constant at 250  $^{\circ}\text{C}$  while the Gerstel injector was programmed (initial temperature 50  $^{\circ}\text{C}$ , hold 0.1 min, and increased at a rate of 10  $^{\circ}\text{C}/\text{s}$  to a final temperature of 330  $^{\circ}\text{C}$ , hold time 10 min). Injections of 1  $\mu\text{l}$  were made in split (1:5) mode (purge time 120 s, purge flow 40 ml/min). Chromatography was performed on an Rtx-5Sil MS column (30 m  $\times$  0.25 mm i.d., 0.25  $\mu\text{m}$  film thickness) with an Integra-Guard column (Restek, Bellefonte, PA, USA). Helium carrier gas was used at a con-

stant flow of 1 ml/min. The GC oven temperature program was 50  $^{\circ}\text{C}$  initial temperature with 1 min hold time and ramping at 20  $^{\circ}\text{C}/\text{min}$  to a final temperature of 330  $^{\circ}\text{C}$  with 5 min hold time. Both the transfer line and source temperatures were 250  $^{\circ}\text{C}$ . After a solvent delay of 350 s, mass spectra were acquired at 20 scans/s with a mass range of 50 to 500  $m/z$ . Initial peak detection and mass spectrum deconvolution were performed with ChromaTOF software (version 2.25, Leco), and later samples were exported to the netCDF format for further data evaluation with MZmine and XCMS.

#### LC–MS analysis

LC–MS analysis was performed using reversed-phase (RP) chromatography [12] and HILIC [5] coupled to ion trap MS. RP separation does not require any processing of urine samples, whereas neat urine was mixed with an equal volume of acetonitrile at room temperature for HILIC. All samples were spun for 5 min at 13,000 rpm prior to injection for particulate matter removal. LC was performed using acetonitrile (LC–MS grade, J. T. Baker, Phillipsburg, NJ, USA) (A) and 13 mM ammonium acetate buffer (pH 5.5, adjusted by acetic acid for RP separations, and pH 9.1 for HILIC separations, adjusted by ammonium hydroxide) (B) as the mobile phase at a flow rate of 0.5 ml/min at 40  $^{\circ}\text{C}$ . Ammonium acetate (extra pure, EMD Chemicals), ammonium hydroxide solution, and acetic acid (glacial, J. T. Baker) were purchased from VWR Scientific Products. Water was purified by a Milli-Q Gradient A10 system (Millipore, Billerica, MA, USA).

RP separations were accomplished with the use of an ACQUITY UPLC system (Waters, Milford, MA, USA). An ACQUITY UPLC BEH shield RP18 column (150 × 2.1 mm, 1.7 μm particle size, Waters) was used at 10,000 to 12,000 psi. After a 0.5-min isocratic run at 0% A, a gradient to 35% A was completed at 8 min. The injection volume was set to 5 μl. HILIC separations were performed with a Surveyor HPLC module (Thermo Electron, Boston, MA, USA) and a normal phase silica column (Luna NH2, 150 × 3 mm, 3 μm particle size, Phenomenex, Torrance, CA, USA). After a 5-min isocratic run at 20% B, a gradient to 35% B was concluded at 20 min, and then a gradient to 90% B was completed at 30 min. The injection volume was set to 3 μl. HPLC columns were connected to the electrospray interface of the Finnigan LTQ (Thermo Electron) linear ion trap mass spectrometer without splitting. Nitrogen sheath gas pressure was set to 7 bar at a flow rate of 2 to 3 L/min. Spray voltage was set to 5 kV. The temperature of the heated transfer capillary was maintained at 350 °C. Full scan mass spectra were acquired from 80 to 800 Da and unit mass resolution in both positive and negative modes.

#### *Raw data processing and statistics*

All chromatograms were evaluated in the same manner by the freely available software packages MZmine [8] and XCMS [9], both of which perform peak finding in an automated and unbiased way using the common MS netCDF file format that enables a unique way of data export irrespective of different instrument platforms (LC–ion trap MS or GC–TOF MS). For the Finnigan LTQ instruments, Xcalibur software usually is used for general data handling. Converting the Xcalibur (\*.raw) files to netCDF (\*.cdf) format was done with the included Xconvert program. For GC–MS data, the netCDF export function from the Leco ChromaTOF software was used. For MZmine, the *m/z* bin size was set to 0.01, the chromatographic threshold level was set to 0.5, the absolute intensity threshold was set to 2500, the tolerance in *m/z* values was set to 0.5, the tolerance in intensity was set to 1.0, and the minimum peak length was set to 2 s. The XCMS parameters for the R language were implemented in an automation script and can be downloaded ([http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak\\_Alignment](http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics/Peak_Alignment)) The XCMS group bandwidth was set to 30, the minimum fraction was set to 0.5, the minimum sample parameter was set to 1, the width of overlapping *m/z* values was set to 0.5, and the maximum number of groups in a single *m/z* slice was set to 0.5. After processing and peak picking, mass spectral features were retrieved from MZmine and XCMS as a TXT file and imported to Excel and then later to STATISTICA Data Miner. Normalized peak areas and unique identifiers were used for further statistical evaluation. Statistical analysis was performed using the STATISTICA Data Miner (version 7, [www.statsoft.com](http://www.statsoft.com)) for partial least squares (PLS) supervised multivariate analysis and for breakdown one-way analysis of variance (ANOVA) univariate statistics.

For each of the three methods (GC–MS, RP–LC–MS, and HILIC–LC–MS) and the two alignment procedures (MZmine and XCMS), feature selection, ANOVA, PLS, and principal component analysis (PCA) were performed. Because the alignment files contained between 150 and 3000 components, only the 50 most significant components were selected for further analysis. This was done using the STATISTICA Data Miner built-in feature selection unit. Based on ANOVA, for categorical-dependent variables, the program computes the ratio of the between-category variance to within category variance and further computes a chi-square statistic and a *P* value for each predictor variable [14]. The *P* level represents a decreasing index of the reliability of a result. In many areas of research, the *P* level of 0.05 is customarily treated as a “borderline acceptable” error level. The *F* value for a variable indicates its statistical significance in the discrimination between groups; that is, it is a measure of the extent to which a variable makes a unique contribution to the prediction of group membership. For each of the significant components, a breakdown ANOVA was performed to analyze statistics by groups (breakdowns) and a whisker plot was included to show the mean value and standard deviation for each class. The ANOVA *P* statistics for two classes results in the same number of significant values as simple *t* statistics. Using the 50 most important features, a nonlinear iterative partial least squares (NIPALS–PLS) with fivefold cross-validation was performed based on the class assignment sick (S) or healthy (H). A scree plot (to show the optimal number of eigenvalues), a score plot (to show the most important principal components and visually detect clusters), and a loading plot (to show positive and negative correlations of components) were included for each analysis. The PCA was performed in unsupervised mode (without class assignment). Cluster analysis of the PCA scores was performed using joining tree clustering with Euclidean distances.

All calculations were performed with Windows XP on a Monarch Computer Dual Opteron 254 (2.8 GHz, 4 GB memory), allowing memory transfer rates of more than 11 GB/s, equipped with an Areca ARC-1120 Raid-5 array, allowing hard disk burst read/write transfer rates of more than 500 MB/s. Using fast hardware was important because of the time-consuming deconvolution and alignment process.

## **Results**

### *Urease treatment and analysis of multiple metabolites*

Prior to column chromatography, urine is commonly pretreated with urease [15] to reduce the risk of column overloading, peak distortions or matrix effects, and ion suppression for both GC–MS and LC–MS analysis [16,17]. This approach is widely accepted, especially in the case of target analysis. However, with any of the three metabolite-profiling tools (GC–TOF, RP–LC–ion trap MS, and HILIC–ion trap MS), a dramatic alteration of profiles was

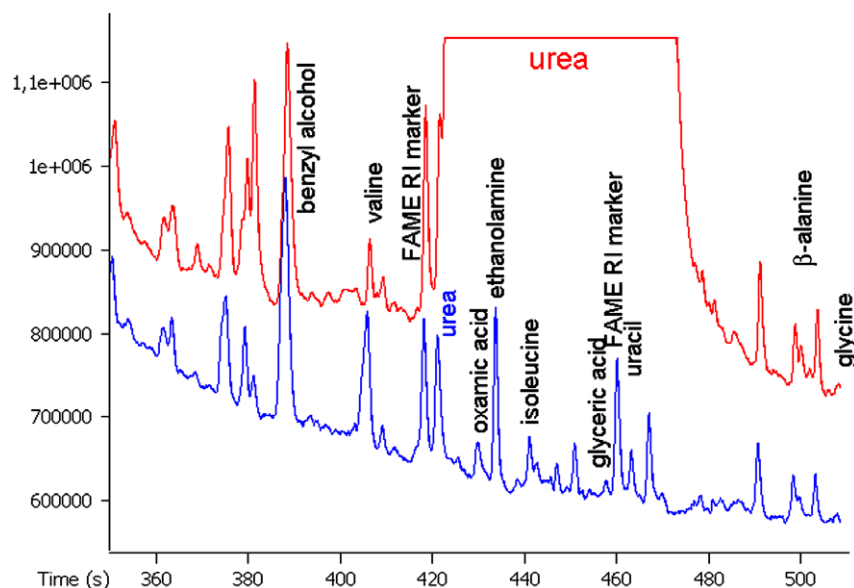


Fig. 2. GC-TOF-MS partial chromatogram showing untreated (red) and urease pretreated (blue) profiles of the same urine sample. FAME, fatty acid methyl esters. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

observed when comparing urine analysis with and without urease treatment. Urea was effectively eliminated by urease treatment, as demonstrated by GC-TOF analysis (Fig. 2). However, many other metabolites, including acotinic acid, hypoxanthine, and the tricarboxylic acid (TCA) intermediates aconitate, citrate, succinate, ascorbate, tyrosine, and glycerol, were found to be greatly diminished with urease treatment. Similar effects were observed in RP-MS and HILIC-MS profiles, mostly for unknown compounds (data not shown). As a consequence of this finding, urine used for this study was not pretreated with urease. In addition, initial gradients used for HILIC-MS were modified for elution of urea close to the void volume to minimize potential ion suppression and peak distortion effects.

#### Urinary metabolite analysis

An important step in biomarker identification, whether “fast lane” or “slow lane” (Fig. 1), is statistical evaluation. Both software packages produce a list of several thousand peaks or components. Even if the components are already ranked according to their *P* value statistics, it is important to apply further statistical tests for deeper understanding. For example, there may be overlapping or coeluting components that cannot be detected by both alignment programs alone but rather can be detected only by other statistical tests.

One of the first statistical approaches in biomarker finding is to reduce dimensionality with feature selection. Several methods can be used for such an approach—ANOVA, PLS, fuzzy techniques, genetic algorithms, and several other methods [18]. Using only the most important features for further analysis will also reduce computational time and thus improve accuracy because in this case useful components are not overlaid by unimportant features or noise.

The next statistical step is a breakdown and one-way ANOVA that gives a detailed overview about which components are significant (*P* test and *F* test) and shows the variation within the groups using whisker plots. Depending on their *P* values, the most significant variables are then used for further calculations. Employing these significant variables, one can use multivariate methods such as PCA, soft independent modeling of class analogy (SIMCA), PLS, partial least squares-discriminant analysis (PLS-DA), orthogonal partial least squares (O-PLS), NIPALS-PLS, independent component analysis (ICA) [19], or any other classification methods, such as machine learning methods (support vector machines and naive Bayes classifiers) [20], neural networks, and classification tree models, to further examine the given problem [14]. Here visualizing methods, such as score plots and scatter plots of the loading factors from PCA and PLS, help to understand the connection between variables and the impact of certain features. It is also always important to switch back to the original chromatograms to identify problems in the raw data. A visualized model in this case is also extremely helpful.

Classification methods can be used to identify sick (i.e., RCC) patients from healthy (i.e., control) volunteers using the developed classification models. Such a deployment process must be carefully evaluated to set up a robust cross-validated model, avoiding misclassification or overfitting of data [21]. The sample number in all assigned classes (sick, healthy, and borderline) should be large enough to provide a sufficient number of independent test samples so that the performance of the developed models can be determined in an independent way. The question of how many samples are needed to make the study statistically significant can be answered by using the calculated analytical variance of the data set and a power analysis [22].

Urine samples from 6 patients with ccRCC (various stages and grades) and 6 healthy controls were obtained and analyzed by GC–TOF–MS, RP–LC–MS, and HILIC–MS techniques. The complement of these techniques can be assumed to cover a large portion of the urine metabolome, with GC–MS particularly being suited for analysis of primary metabolites with masses up to 500 Da (prior to derivatization) and the two LC–MS methods enabling the detection of secondary and glucuronidated metabolites up to a mass of 4000 Da. A query for urinary metabolites in the Human Metabolite Database ([http://redpoll.pharmacy.ualberta.ca/~aguo/www\\_hmdb\\_ca/HMDB](http://redpoll.pharmacy.ualberta.ca/~aguo/www_hmdb_ca/HMDB)) resulted in 1316 matches for the complement of identified urine compounds so far. Based on the number of peak markers detected by the three MS methods and unbiased peak detection, we estimate that there are at least an equal number of unknown components in addition to the compounds already described in the Human Metabolome Database. In this study, we did not attempt to identify all of the detected peaks but rather focused on evaluation of the use of the mass spectrometric and peak processing techniques for the development of diagnostic tests for RCC (“fast lane” in Fig. 1). The hypothesis was that a large group of potential biomarkers was more likely to evolve patterns for disease recognition than was any single compound, be it identified or not.

For unbiased analysis and peak picking, chromatograms need to be aligned so as to encompass small drifts in retention time or mass spectral quality. Therefore, we chose two freeware software programs (MZmine and XCMS) that showed promising results when compared with alignment and peak finding for LC–MS and GC–MS data. Both programs use the open source MS exchange format netCDF, which is readily available as export format for most mass spectrometers. The graphical interface in MZmine visualizes data in a comprehensive manner to enable interrogating metabolic features. Using the 12 urine samples from both RCC patients and controls, the number of total components resulting from each analysis ranged from 150 to 3000 metabolic features (Table 1). Depending on the performance of the peak picking algorithm, these features either can represent mass/charge ( $m/z$ ) values that were found to differ from background noise or can be real components or metabolites if the chromatography already yielded a complete separation.

Why are the resulting parameters (Table 1) so different? This is clearly the influence of the different applied decon-

volution and alignment parameters. To show this influence, we optimized parameters for both XCMS and MZmine for GC–MS data. Unit mass resolution was assumed, and the width of overlapping  $m/z$  slices was set to 1. It can be seen in this case that the same number of components is obtained with both programs (Table 1). For further proof, we checked the correct deconvolution of peaks with AMDIS software. A known peak (TMS derivative of hippuric acid, peak  $m/z=236$ ,  $m/z=206$ , and  $m/z=105$  at RT=10.85 min) was identified via an internal retention index and mass spectral database. The programs AMDIS and ChromaTOF correctly picked one peak. XCMS and MZmine, operating with optimized parameters for GC, detected three single compounds, each assigned to a specific  $m/z$  value. Programs that are optimized for GC–MS peak picking, such as AMDIS and ChromaTOF, usually perform better during deconvolution procedures, but these programs have no built-in alignment routines. Such a result of multiple component picking will not disturb the statistical analysis because that can be easily detected from the multivariate statistics (see PLS loading plots in Fig. 4). Overlapping components in the PLS loading plots usually are obtained due to peak picking problems. A full investigation of all false-positive or false-negative candidates is beyond the scope of this article and would be performed in subsequent studies using the “slow lane” approach (Fig. 1).

#### Statistical analysis of metabolites

The development of diagnostic tests demands that statistical analysis determine suitable candidate markers in a “fast lane” approach without necessarily identifying all of the candidate markers (Fig. 1). For validation of such tests, the chemical structures of components need to be determined so as to support mechanistic studies and to gain support and approval by the clinical medicine community. De novo compound identification is still a cumbersome “slow lane” approach (Fig. 1) but is not required for the initial proof-of-concept studies provided here. The first statistical step in biomarker detection is to reduce complexity (or dimensionality) with feature selection. This process also determines which peaks from the MS analyses are significant discriminators between classes of comparisons. The number of important features from this analysis using MZmine was the lowest in GC–MS analysis, followed by

Table 1  
Results for significant features/components found in urine samples from three different analytical techniques and two different alignment procedures

Alignment	Technique	Total features found	Significant features ( $P < 0.05$ )
MZmine	HILIC–LC–MS	3225	12
MZmine	RP–LC–MS	5034	18
MZmine	GC–MS	342	7
XCMS	HILIC–LC–MS	1525	13
XCMS	RP–LC–MS	136	27
XCMS	GC–MS	1894	5
MZmine optimized	GC–MS	326	28
XCMS optimized	GC–MS	417	29

Note. The number of significant components is strongly reduced after feature selection and ANOVA.

HILIC–LC–MS and RP–LC–MS (Table 1). From the data quality itself, GC–MS profiles were deemed best and components were sharply separated. Although XCMS already gives a result list that is sorted according to importance ( $P$  statistics), the selected features by ANOVA and PLS were different (Table 1). This is due to the fact that the breakdown ANOVA is based on class assignments and the PLS that is fed with these variables also has a built-in feature selection that can be accessed via the importance plot from the principal component model of the NIPALS–PLS.

The data from the ANOVA analyses (Tables 2 and 3) can be used to access the single components in the chro-

Table 2

Statistical results from HILIC–LC–MS analyzed by MZmine and breakdown ANOVA showing the most significant features/components that can be used for class discrimination (healthy and sick individuals)

HILIC–MZmine component	$F$ value	$P$ value
Var1776	29.6	0.001
Var997	13.7	0.006
Var3061	13.4	0.006
Var1766	13.3	0.006
Var2558	12.6	0.008
Var2027	11.4	0.010
Var1336	10.8	0.011
Var178	10.7	0.011
Var1301	10.2	0.013
Var2146	9.2	0.016
Var1724	8.5	0.020
Var1045	8.0	0.022

Note. Components with a high  $F$  value and a  $P$  level  $< 0.05$  can be used for class discrimination.

Table 3

Statistical results from GC–MS analyzed by XCMS and breakdown ANOVA showing the most significant components that can be used for class discrimination (healthy and sick individuals)

GC–MS–XCMS component	$F$ value	$P$ value
M317T833	16.8	0.002
M110T560	13.3	0.004
M145T547	12.6	0.005
M355T833	7.3	0.022
M152T383	5.1	0.047

Note. Components with a high  $F$  value and a  $P$  level  $< 0.05$  can be used for class discrimination.

matograms as primary targets for structure elucidation. The tables show a highly separable selection of significant features from two different methods (HILIC–MS in Table 2 and GC–MS in Table 3) and their  $F$  and  $P$  values. The differences seen in the box and whisker plots for two representative metabolites (Fig. 3) could be the result of two possible scenarios. In the first scenario, a component is up-regulated in healthy individuals as compared with those with ccRCC (left box plot in Fig. 3), for example, an antioxidant such as glutathione or vitamin E that is higher in healthy volunteers [23]. In the second scenario, a feature that is up-regulated in ccRCC patients (right box plot in Fig. 3) could be a true marker substance or biomarker of ccRCC. Both components can be used for model building using techniques discussed below. There is no need to identify a true biomarker if the down-regulated markers are very specific for a certain cancer type. This will become apparent once more samples are obtained.

The results from such a feature selection process are used as input for classification and machine learning algorithms. In general, unsupervised techniques such as PCA are less powerful because such methods separate samples based on overall variance that may be due to chemical noise or random unrelated variation of metabolites. For that reason, we used NIPALS–PLS as the preferred supervised classification algorithm. A PLS loading plot of the significant features as determined in Tables 2 and 3 is shown in Fig. 4. This type of plot is useful to reveal relationships either between individual components or among components that can be clustered. Such a plot also helps to identify peak selection problems during the alignment procedure; if points are overlapping and have similar or the same retention times and  $m/z$  values, the peak picking algorithm may have failed. The variables that are far away from the coordinate center have the most discrimination power and, thus, are more important and should be further examined, if not already included during the ANOVA test. Components on opposite sides of the coordinate center are negatively correlated, for example, features C3061 and C1776 (cf. Figs. 3 and 4).

To determine, using urine metabolomic analysis, whether ccRCC patients are separable from healthy controls, we once again used PLS, PCA, and cluster analysis

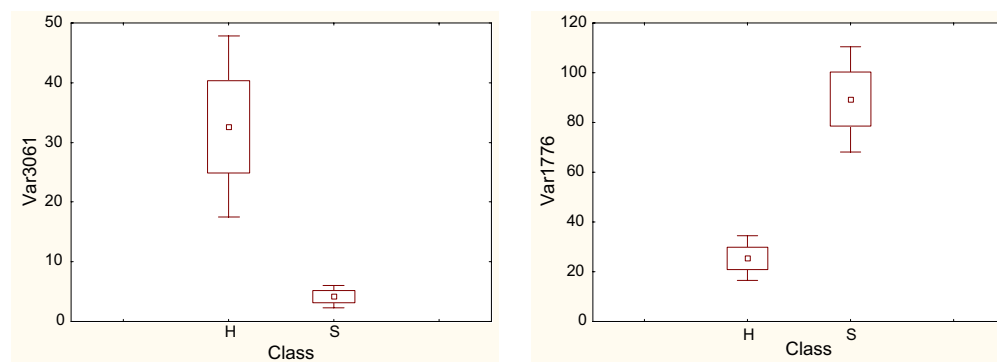


Fig. 3. Categorized box and whisker plot for HILIC–LC–MS (hydrophilic interaction) features C3061 and C1776. The left plot shows a component that is suppressed in RCC (S, sick) patients as compared with controls (H, healthy). The right plot shows a component that is highly elevated in RCC patients. The plots show the variable means  $\pm 1.96$  times the variable standard deviations (95% confidence interval).

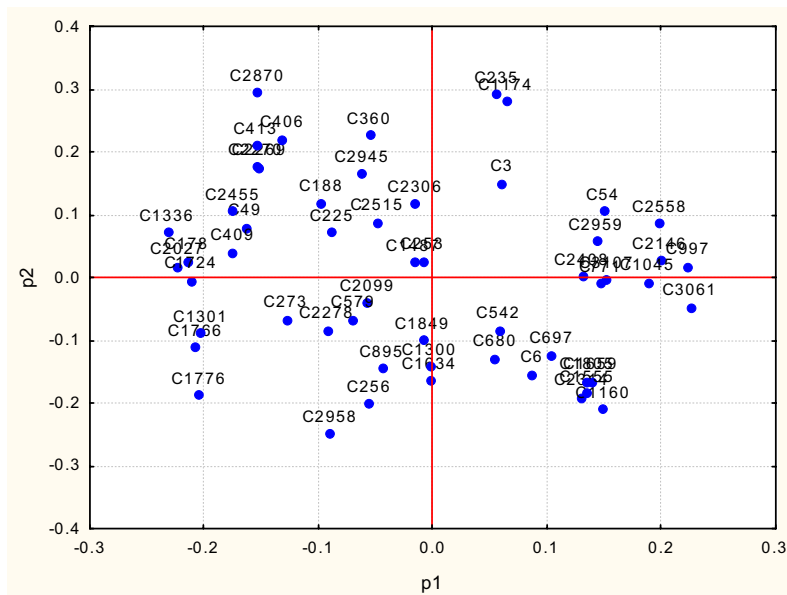


Fig. 4. PLS loading plots for HILIC–LC–MS (hydrophilic interaction). Deconvolution errors can be detected by observing overlapping points. Components very outside the coordinate center (0,0) have high discrimination power and are more important. Components on the left (e.g., C1776 [see Fig. 3]) are anti-correlated to components on the right (e.g., C3061).

for visualization. PLS score plots were calculated for each of the three analytical methods (Figs. 5, 6 and 7), which were examined by two alignment tools. These plots confirm that the class assignment (RCC [sick] vs. control [healthy])

is correct because individual classes clearly are separated from each other. To further validate this outcome, we used an unsupervised PCA and performed a cluster analysis on the PCA score results (Fig. 8). The example is given here for

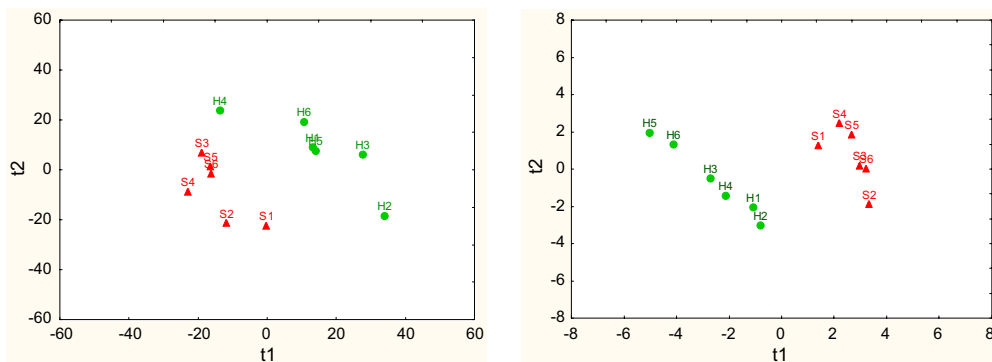


Fig. 5. PLS score plots for GC–MS urine analysis showing clear separation of RCC patients and controls by 50 important features. S1 to S5 represent RCC patients; H1 to H6 are controls. Left aligned by XCMS; right aligned by MZmine.

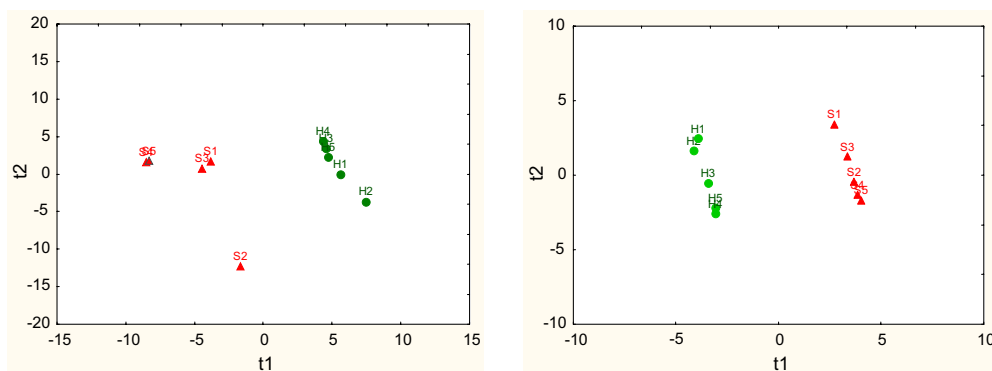


Fig. 6. PLS score plots for RP–LC–MS urine analysis showing clear separation of RCC patients and controls by 50 important features. S1 to S5 represent RCC patients; H1 to H6 are controls. Left aligned by XCMS; right aligned by MZmine.

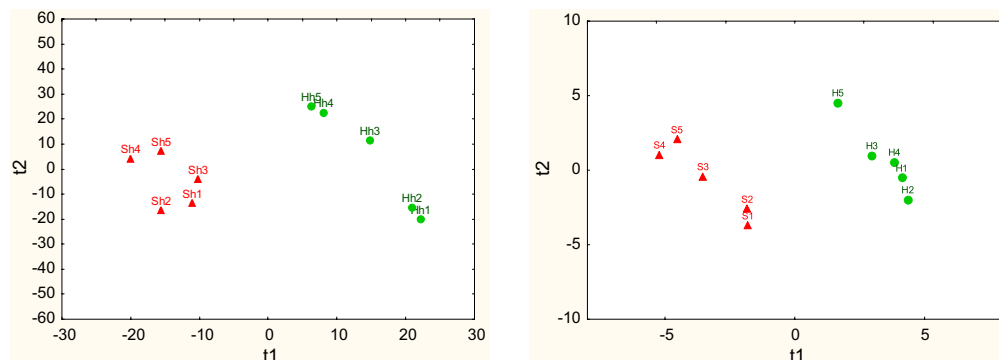


Fig. 7. PLS score plots for HILIC–LC–MS (hydrophilic interaction) urine analysis showing separation of RCC patients and controls by 50 important features. S1 to S5 represent RCC patients; H1 to H6 are controls. Left aligned by XCMS; right aligned by MZmine.

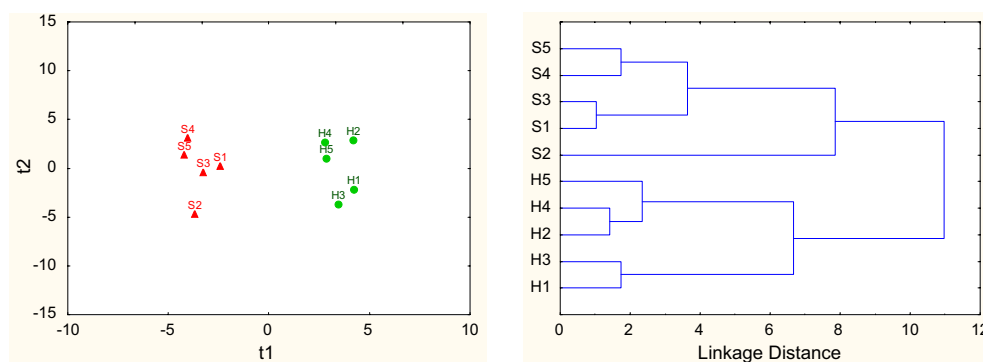


Fig. 8. PCA score plots for RP–LC–MS urine analysis aligned by MZmine showing separation of RCC patients and controls by 50 important features (left panel) and a cluster analysis from the PCA scores (right panel). S1 to S5 represent RCC patients; H1 to H5 are controls.

the RP–LC–MS data set aligned by MZmine. Two distinct clusters are obtained at a linkage distance of 11. The data for class separation and class prediction are also contained in an internal PLS deployment code. This code could be used for prediction of unknown samples. There are minor visible deviations between the two alignment procedures, which are affected by the parameters used during the peak picking and alignment procedure.

## Discussion

In most malignancies, early diagnosis has the potential to markedly improve patient survival. This paradigm is especially true in kidney cancer, where early diagnosis, prior to metastatic spread, can improve survival odds from less than 10% to greater than 90% [2]. Given the fact that most biochemical processes possess a urinary metabolic “signature,” the use of metabolomics is ideally suited for elucidation of a urinary diagnostic assay, and diseases involving the kidney are ideal for such studies. In this article, we have presented a proof-of-concept study in which we applied metabolomic analysis to RCC patient urine to define a profile or pattern of metabolites indicating that an individual is likely to have RCC. Despite a relatively small number of samples for this pilot study, we see a clear and significant separation between RCC patients and normal controls. Validation

of such initial hypotheses using a larger sample size is ongoing in our laboratories.

Using urine to diagnose RCC is a theoretically ideal approach to this problem given the fact that the kidney is intimately connected to this biofluid. Despite this ideal, metabolomic analysis of RCC using this novel approach has not been reported prior to the current study. A reasonable critique of our urinary metabolomic analysis could be that the glomerular pores might be too small to allow potential metabolites to be filtered into the urine. The ultimate composition of compounds in the urine is the result not only of glomerular filtration but also of tubular secretion and reabsorption. Normal urine contains approximately 150 mg/24 h of protein, and it is known that compounds such as inulin (5 kDa) and lysozyme (14 kDa) appear freely in the urine. Thus, it can be readily appreciated that the urine in RCC patients may well contain abundant candidate metabolites as potential tumor markers. A previous report showed that human kidney injury molecule-1 (hKIM-1), when normalized to creatinine, appears in the urine of ccRCC patients and disappears or is reduced after nephrectomy [24]. In our data, such normalization to creatinine is not required because we are examining the pattern of appearance of discrete metabolites and not the magnitude of their expression. However, the data were normalized to the total peak area. In addition, rather than focusing on a single metabolite or other urinary compound

such as hKIM-1, our approach capitalizes on the entire human metabolome and searches for patterns of expression therein.

Another potential problem arising from the small sample cohort is that the discrimination between RCC patients and healthy controls may derive from factors that are related to a drug treatment of already known sick patients. This is not applicable to samples used in our study because no patient in our study had yet received chemotherapy or radiation or undergone nephrectomy. Another confounding scenario might be that one class uniformly has another disease (e.g., diabetes, hypertension) or habit (e.g., smoking) or belongs to a racial group (e.g., African American) or gender that leads to an observed discrimination effect. Although these problems cannot be totally excluded for this proof-of-concept study (of the RCC patients, 3 were Caucasian, 3 were of unknown race, and 2 were female; of the healthy controls, 4 were Caucasian, 2 were Asian, and 2 were female), such issues do not affect the whole analytical and statistical workflow per se. Another argument against this criticism is that we used three independent analytical procedures, each covering a different metabolite pool (primary metabolites, lipophilic secondary metabolites, and hydrophilic secondary metabolites). For future studies, detailed meta-data documentation must be included, and the number of individuals, as well as urine samples per individual, will be increased.

Genomic analysis and proteomic analysis have been used for several years in studies attempting to elucidate the mechanism of oncogenesis in a variety of cancers. However, although these techniques are studied and published extensively, they are not readily amenable to discovery of diagnostic tests because the results of such analyses are not generally secreted into biofluids. Metabolomics, on the other hand, is ideally suited for such analysis because small molecule metabolites frequently appear, if transiently, in both serum and urine [25]. Thus, metabolomics research ultimately may be more clinically translatable than other “omic” endeavors [26,27]. The nature of this technology is to examine all candidate small molecules (i.e., thousands) in the urine and find patterns that are associated with the malignancy in question, in this case RCC. Chemical and structural identification of the metabolites clearly is more time-consuming and falls under the “slow lane” approach (Fig. 1). This approach will be used in subsequent studies in our laboratories.

To avoid overfitting during the statistical analysis, we applied  $v$ -fold cross-validation ( $v = 5$ ) during the PLS evaluation, where the whole data set is divided into  $v$  segments,  $v - 1$  of which are used to build the model and the rest of which are used for testing [14]. This approach is also used to find the optimal number of principal components, although we always could obtain a good model with only two principal components. To further clarify this, we did not use our statistical results for prediction of unknown external samples. To increase the overall statistical robustness, the sample size needs to be increased, and this will be

done in future studies. However, this did not hamper the whole experimental setup and the applied statistical procedures used in this study.

A critical problem during biomarker identification is the detection and removal of artifacts from the data sets used. One of the first steps taken to exclude machine drift in large sample sets is to randomize samples. Another method is to use an orthogonal separation technique such as chromatography to obtain truly resolved peaks instead of direct infusion mass spectra [28]. The reason is that even with high mass accuracy and high resolving power, in most cases no single metabolite isomer structure can be assigned to a single  $m/z$  value [29]. For GC-MS data sets, the safest approach is the use of a retention index database coupled to a mass spectral database [30], thereby allowing the detection of true metabolites and working with a data set of identified compounds instead of putative or tentatively identified compounds. We will use such an approach in a future study with a larger sample cohort. This is especially critical for GC-MS data, where an additional derivatization step is included and silylation artifacts must be excluded [31]. These artifacts are molecules built by the derivatization agent itself during the derivatization or within the injector, and their amounts can be highly variable. Also, all internal standards and retention index markers must be removed before further statistical analysis. Artifacts are also known to occur in LC-MS analysis if ultra pure solvents or highly solvent resistant plastic containers are not used. LC-MS artifacts can also occur during in-source reactions or from formed solvents/clusters [32]. One solution to this artifact problem is the measurement of “reagent blanks” that contain only the pure solvents or all used derivatization agents. An additional approach is the analysis of “method blanks” that undergo all steps during extraction and sample preparation but do so without any real sample. The daily measurement of performance checks and proper mass calibration are obligatory. All of these steps were routinely performed in our laboratories.

## Conclusion

We have reported a pilot study demonstrating the use of urine metabolomics to discriminate kidney cancer patients from normal individuals without known disease. The elegance of the comprehensive method reported here, employing three different analytical techniques and two different algorithms for data alignment, lies in the fact that, even if there is a large number of individuals under investigation in future studies and metabolite changes are still very small for one single method, all three models combined in a decision tree can produce reliable results for cancer detection. Both alignment programs, MZmine and XCMS, can be used for data analysis if the input parameters are carefully optimized and alignment results are double-checked later. Once our urine metabolomic data are confirmed in a study employing a greater number of patients and urine samples per patient, we foresee the

ultimate development of a relatively simple and clinically applicable urine test that can identify affected patients. The use of such an assay, when applied to those at highest risk for the disease or its recurrence, will have a major salutary effect on patient mortality as well as on health care economics.

### Acknowledgments

This work was supported by grant 1R21CA 91259-01A1 and funding from the Early Detection Research Network (EDRN), both from the National Cancer Institute (R.H.W.), by a grant from the University of California Cancer Research Coordinating Committee (R.H.W.), and by grant NIEHS RO1 13932 from the National Institutes of Health (O.F. and T.K.).

### References

- [1] M.T. Lotze, E. Wang, F.M. Marincola, N. Hanna, P.J. Bugelski, C.A. Burns, G. Coukos, N. Damle, T.E. Godfrey, W.M. Howell, M.C. Panelli, M.A. Perricone, E.F. Petricoin, G. Sauter, C. Scheibenbogen, S.C. Shivers, D.L. Taylor, J.N. Weinstein, T.L. Whiteside, Workshop on cancer biometrics: Identifying biomarkers and surrogates of cancer in patients—A meeting held at the Masur Auditorium, National Institutes of Health, *J. Immunother.* 28 (2005) 79–119.
- [2] R.H. Weiss, P.-Y. Lin, Kidney cancer: Identification of novel targets for therapy, *Kidney Intl.* 69 (2006) 224–232.
- [3] E. Jellum, O. Stokke, L. Eldjarn, Application of gas chromatography, mass spectrometry, and computer methods in clinical biochemistry, *Anal. Chem.* 46 (1973) 1099–1106.
- [4] M.A. Evenson, G.D. Carmack, *Clinical chemistry*, *Anal. Chem.* 51 (1979) 35R–79R.
- [5] V.V. Tolstikov, O. Fiehn, Analysis of highly polar compounds of plant origin: Combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry, *Anal. Biochem.* 301 (2002) 298–307.
- [6] V.V. Tolstikov, A. Lommen, K. Nakanishi, N. Tanaka, O. Fiehn, Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics, *Anal. Chem.* 75 (2003) 6737–6740.
- [7] R. Plumb, J. Castro-Perez, J. Granger, I. Beattie, K. Joncour, A. Wright, Ultra-performance liquid chromatography coupled to quadrupole–orthogonal time-of-flight mass spectrometry, *Rapid Commun. Mass Spectrom.* 18 (2004) 2331–2337.
- [8] M. Katajamaa, J. Miettinen, M. Oresic, MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data, *Bioinformatics* 22 (2006) 634–636.
- [9] C.A. Smith, E.J. Want, G. O’Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, *Anal. Chem.* 78 (2006) 779–787.
- [10] W.B. Dunn, D.I. Ellis, Metabolomics: Current analytical platforms and methodologies, *Trends Anal. Chem.* 24 (2005) 285–294.
- [11] W. Weckwerth, M.E. Loureiro, K. Wenzel, O. Fiehn, Differential metabolic networks unravel the effects of silent plant phenotypes, *Proc. Natl. Acad. Sci. USA* 101 (2004) 7809–7814.
- [12] V.J. Nikiforova, J. Kopka, V. Tolstikov, O. Fiehn, L. Hopkins, M.J. Hawkesford, H. Hesse, R. Hoefgen, Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of *Arabidopsis* plants, *Plant Physiol.* 138 (2005) 304–318.
- [13] M. Schad, R. Mungur, O. Fiehn, J. Kehr, Metabolic profiling of laser microdissected vascular bundles of *Arabidopsis thaliana*, *Plant Methods* 1 (2005) 2.
- [14] Statsoft Inc., *Electronic Statistics Textbook*, 2006, [www.statsoft.com/textbook/stathome.html](http://www.statsoft.com/textbook/stathome.html).
- [15] T. Kuhara, *Metabolomics: The Frontier of Systems Biology*, Springer-Verlag, Tokyo, 2005.
- [16] H. Idborg, L. Zamani, P.O. Edlund, I. Schuppe-Koistinen, S.P. Jacobsson, Metabolic fingerprinting of rat urine by LC/MS: I. Analysis by hydrophilic interaction liquid chromatography–electrospray ionization mass spectrometry, *J. Chromatogr. B* 828 (2005) 9–13.
- [17] T. Kuhara, T. Shinka, Y. Inoue, M. Ohse, X. Zhen-Wei, I. Yoshida, T. Inokuchi, S. Yamaguchi, M. Takayanagi, I. Matsumoto, Pilot study of gas chromatographic–mass spectrometric screening of newborn urine for inborn errors of metabolism after treatment with urease, *J. Chromatogr. B* 731 (1999) 141–147.
- [18] A.K. Jain, Statistical pattern recognition: A review, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (2006) 4–37.
- [19] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Metabolite fingerprinting: Detecting biological features by independent component analysis, *Bioinformatics* 20 (2004) 2447–2454.
- [20] J. Taylor, R.D. King, T. Altmann, O. Fiehn, Application of metabolomics to plant genotype discrimination using statistics and machine learning, *Bioinformatics* 18 (Suppl. 2) (2002) S241–S248.
- [21] M. Wagner, D. Naik, A. Pothen, Protocols for disease classification from mass spectrometry data, *Proteomics* 3 (2003) 1692–1698.
- [22] L.M. Friedman, C.D. Furberg, D.L. DeMets, *Fundamentals of Clinical Trials*, 3rd ed., Springer-Verlag, New York, 1998.
- [23] M.J. Fryer, Vitamin E as a protective antioxidant in progressive renal failure, *Nephrology* 5 (2000) 1–2.
- [24] W.K. Han, A. Alinani, C.L. Wu, D. Michaelson, M. Loda, F.J. McGovern, R. Thadhani, J.V. Bonventre, Human kidney injury molecule-1 is a tissue and urinary tumor marker of renal cell carcinoma, *J. Am. Soc. Nephrol.* 16 (2005) 1126–1134.
- [25] B. Perroud, J. Lee, N. Valkova, A. Dhirapong, P.Y. Lin, O. Fiehn, D. Kultz, R.H. Weiss, Pathway analysis of kidney cancer using proteomics and metabolic profiling, *Mol. Cancer* 5 (2006) 64.
- [26] C. Schmidt, Metabolomics takes its place as latest up-and-coming “omic” science, *J. Natl. Cancer Inst.* 96 (2004) 732–734.
- [27] K. Dettmer, B.D. Hammock, Metabolomics: A new exciting field within the “omics” sciences, *Environ. Health Perspect.* 112 (2004) A396–A397.
- [28] L.A. Liotta, M. Lowenthal, A. Mehta, T.P. Conrads, T.D. Veenstra, D.A. Fishman, E.F. Petricoin III, Importance of communication between producers and consumers of publicly available experimental data, *J. Natl. Cancer Inst.* 97 (2005) 310–314.
- [29] T. Kind, O. Fiehn, Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm, *BMC Bioinform.* 7 (2006) 234.
- [30] O. Fiehn, G. Wohlgenuth, M. Scholz, Automatic annotation of metabolomic mass spectra by integrating experimental metadata, *Bioinformatics* 3615 (2006) 224–239.
- [31] J.L. Little, Artifacts in trimethylsilyl derivatization reactions and ways to avoid them, *J. Chromatogr. A* 844 (1999) 1–22.
- [32] T. Benijts, W. Lambert, W. DeLeenheer, Sonic spray ionization applied to liquid chromatography/mass spectrometry analysis of endocrine-disrupting chemicals in environmental water samples, *Rapid Commun. Mass Spectrom.* 17 (2006) 1866–1872.