Metabolomics of biofluids : from analytical tools to data interpretation

Nevedomskaya, E.

Metabolomics of biofluids: from analytical tools to data interpretation

Ekaterina Nevedomskaya

ISBN: 978-94-6182-038-9

The printing of this thesis was financially supported by:
Bruker BioSpin GmbH, Germany
Dionex Benelux B.V.
Beckman Coulter (Nederland) B.V.
Bruker Nederland B.V.

Cover: fragment of “To Touch The Grass - 1” by Anton Shirkin.

Printing: Off Page, Amsterdam

Metabolomics of biofluids: from analytical tools to data interpretation

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus Prof. Mr. P.F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 23 november 2011
klokke 16:15 uur

door

Ekaterina Nevedomskaya

geboren te Moskou, Rusland in 1985

PROMOTIECOMMISSIE

Promotor: Prof. dr. A.M. Deelder

Co-promotor: Dr. O.A. Mayboroda

Overige leden: Prof. dr. M.S. Gelfand
Institute for Information Transmission Problems RAS, Moscow, Russia

Dr. H. Keun
Dept. of Surgery and Cancer, Faculty of Medicine, Imperial College, London, UK

Prof. dr. T.W.J. Huizinga

Prof. dr. P. Slagboom

Prof. dr. J.T. van Dissel

Prof. dr. A.E. Gorbalenya

Моим родителям
(For my parents)

TABLE OF CONTENTS

General Introduction

Variability of genomes in populations is a necessary prerequisite for evolution. Owing to it, populations adapt to the changing environment, conquer new territories and new species evolve. However, this genetic variability can only be seen at the population level; at the level of the individual its genome is constant and static. The genome defines the possibilities of a given organism to adapt to the environment, but does not reflect its actual state. As in modern medicine there is a conceptual shift to personalized health, the attention is being redirected from populations to individuals. Though the importance of genetics is hard to overestimate, new ways of assessing human individuality, or, in other words, the phenotype, are sought. The biochemical representation of the phenotype is believed to be most closely approximated by the metabolome – a collection of low-molecular-weight (<1 kDa) compounds (metabolites) present in an organism.(1) Metabolites are the products of all the biochemical processes in the organism, which makes them a more appropriate target for phenotype-based research than transcripts and proteins, which are information messengers and executors of the biochemical reactions, respectively.

The key words that characterize metabolomics are diversity and variation. These characteristics can be both virtue and vice for the metabolomics workflow (Figure 1). At different steps variation has either to be explored and used or reduced to the possible minimum.

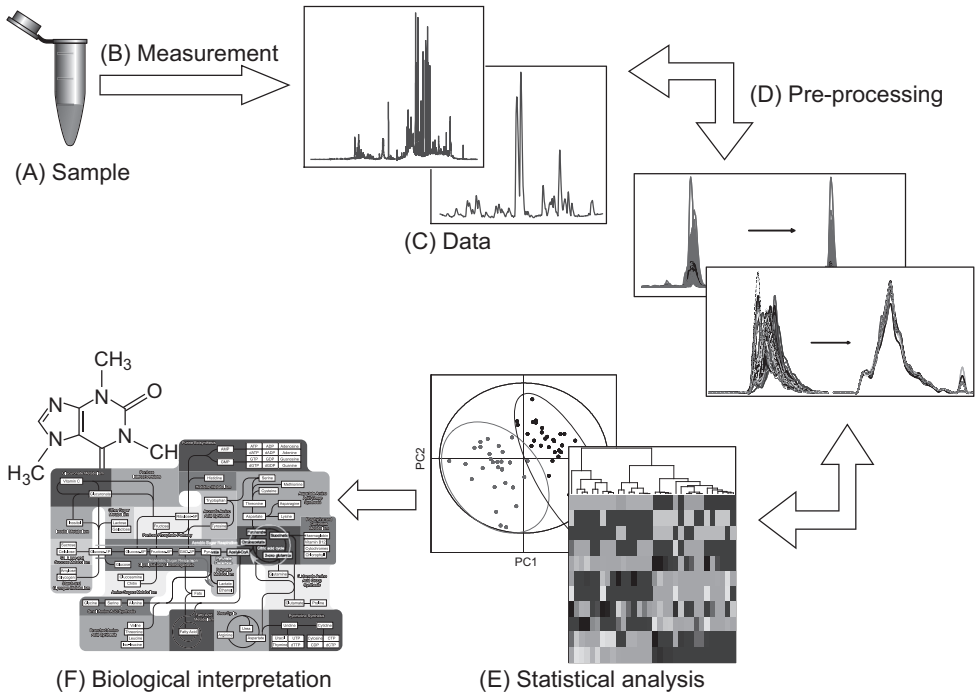

Figure 1. Typical metabolomics workflow.

The variability that occurs in metabolomics, as well as in other 'omics'-experiments, has three sources – biological, pre-analytical and analytical.(2) Unlike in animal experiments, in which it is relatively easy to standardize both the conditions under which the animals are kept and the handling of the samples, clinical experiments are much more susceptible to bias and interference. This variability affects the outcome of the existing clinical tests(3), and without question influences metabolomics results as well.

Already at the stage of planning the experiments and the study design it is important to reduce the part of biological variation which is not related to the question addressed by the study. When, for example, two groups of patients or patients and healthy controls are to be compared, it is important that the selected groups have minimum differences not related to the research aim. These can be associated with gender, age and diet. These factors have been shown to have a large effect on metabolic profiles.(4;5) It also has been shown that the differences due to the fact that samples are collected in different countries and cities can be easily detected in urine metabolic profiles.(6;7) These extremes should be avoided of course, but it is not difficult to imagine that in a typical multi-center clinical study samples come from different hospitals, than the results obtained from the data should be considered with great caution.

When samples are collected (Figure 1(A)) it is very important that the collection is highly standardized. First of all, the time of collection is an important factor to consider. Diurnal variation is not that obvious in case of blood plasma, but has a strong impact on metabolic profiles of urine.(8) The subjects should also either follow a certain diet or fast before the sample collection. Sample handling obviously also plays a role and such factors as the collection tubes, time on ice before freezing, temperature and time of storage, the number of thaw-freeze cycles can introduce a considerable bias.(9)

Assuming that the study had been properly designed and that the samples had been collected, the aim of a metabolomics experiment would be to generate a comprehensive view on the low-molecular-weight components in the samples (Figure 1(B)). At this stage another issue of variability in metabolomics is faced: which is the diversity of metabolites themselves. And this represents a major difference between metabolomics and the other 'omics' technologies. In genomics, transcriptomics and proteomics the molecules measured belong to the same chemical classes; metabolites, however, are extremely diverse in their chemical and physical properties. It is impossible to cover all of the metabolites by a single analytical technique, which is the reason why a number of analytical platforms are being used in the field.(10) Those platforms are either Mass-Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) based. Both approaches have their specific advantages and disadvantages.(11) For example, NMR has a lower concentration sensitivity than MS, but

requires less sample pretreatment and is non-destructive. MS-based technologies are very sensitive, but the range of the compounds detected is defined by the separation method used, thus making it more targeted compared to ^{1}H NMR, which is universal for all molecules containing hydrogen, which in case of biological samples means very few exceptions. The combination of analytical methods can increase the coverage of the metabolome and thus lead to better understanding of the biological process under study as well as improve identification of the compounds.(12) However, due to many factors, such as, for instance, the high costs of the different instruments, the whole range of the machines suitable for metabolomics experiments is rarely present in one lab. Though the analytical capabilities and drawbacks of each of the platforms are known, it is less known which impact they have on the recovery of biological information. This knowledge is essential for making a decision about the suitability of a given analytical method for a certain biological/clinical application.

The reproducibility of analytical methods is not perfect and this obviously can result in variability in the data. MS-based methods are more prone to this variability compared to NMR, not only due to the nature of the instrumentation, but also because the samples need more extensive pre-treatment and preparation. The reproducibility of every technique should be carefully assessed in order to understand the possible drawbacks and the extent of post-processing needed. Also each of the methods has to be optimized to give minimum variability. A way to control the instrument performance while running long sequences of measurements is to use a set of analytical standards and quality control (QC) samples.(13)

Despite all the efforts to get rid of the unwanted analytical variability in the final data, this can only be minimized. A common loss in repeatability is the drift in peak position in the spectra. In NMR this is related to the difference in pH and ion strength between the samples which is not completely abolished by the addition of buffer into the samples before the measurement. In chromatography-MS techniques misalignment of chromatographic peaks is caused by the sample matrix, pH, column ageing; the stability of the MS can also be compromised and cause run-to-run differences.(14) Peak shift is also considered a serious drawback of capillary electrophoresis.(15) A trivial method that allows overcoming the drift problem to a certain extent is binning – dividing the available data axis in short segments (bins) and integrating the signal intensity of each of them. More sophisticated extensions of the method are also available, such as, for example, adaptive binning (16) and Gaussian binning.(17) Binning is often used in case of NMR, although there are reasons to consider this method not optimal.(14) For chromatography-MS techniques it is even less advantageous due to the more complex nature of the data, the large number of variables generated and the loss of peak information. For this reason, in chromatography-MS a

combination of alignment and peak picking is most regularly used with a variety of available algorithms and software.(18-20)

After all the manipulations mentioned above the only variability that should be left in a metabolomics dataset is of biological origin. The unwanted part of it is related to the different dilution of the samples, which is most prominent in the case of urine due to different water uptake by the subjects. Normalization is the step that removes this variation. For this, a few methods are available: normalization to the total sum, to some "housekeeping" molecules (*e.g.* to creatinine) and more sophisticated ones (*e.g.* Probabilistic Quotient Normalization(21)), but none of them are optimal for all cases and the choice should be made depending on the biological context.(22)

The rest of the variation in the data has to be explored as this represents the biological variability (Figure 1(E)). The first step of the analysis is to investigate data structure in order to find patterns, natural grouping of the samples and possible outliers. This is done by means of unsupervised methods, which do not use any *a priori* information about the data and thus give an unbiased view. The most often used procedures are Principal Component Analysis (PCA) and its variations and clustering.

PCA is a projection-based method that summarizes the variation present in the data in a lower-dimension space. When applied to data containing both biological and QC samples it is possible to estimate the analytical variation in the data in comparison to the biological variability: QCs must have orders of magnitude less variation than the real samples, thus clustering tightly together. PCA is also extremely useful for identifying abnormal samples (outliers), which may have to be removed prior to any further analysis, and for detecting any grouping of samples. The latter might be related to the question of interest, addressed by the study, or might be related to other phenomena. The first case is very encouraging for continuation of the analysis. The second does not mean that subsequent analysis cannot be done, because often the studied differences are subtle and masked by other sources of biological diversity; it however implies a more careful selection of strategies for discrimination and especially for validation, as natural clustering of the samples might influence the results of cross-validation.

Clustering is a collection of unsupervised methods to assess inter-sample relations and identify natural groups of samples. There are many variants of clustering that use different measures for the distance between the samples (Euclidean, Mahalanobis, Manhattan *etc.*), different algorithms (hierarchical, partitioning *etc.*) and various initial assumptions (*e.g.* whether samples belong to only one or to multiple clusters). Clustering not only allows discovering grouping of samples, but also assessing the quality of the data.

As has already been mentioned above, in unsupervised methods the variation of interest is not necessarily reflected in the natural grouping of samples. The questions often posed in clinical metabolomics research are finding the differences between two and more groups of samples (from patients and controls, from subjects under various conditions and/or interventions) and predicting to which of the studied groups new samples belong. Statistically speaking the tasks are discrimination and classification, which are often carried out together in one method.

The abundance of discrimination/classification methods may appear confusing: projection-based methods (Partial Least Square Discriminant analysis (PLS-DA) and Orthogonal PLS-DA (OPLS-DA), Soft Independent Modelling of Class Analogies (SIMCA)), k-nearest neighbor algorithm (k-NN), artificial neural networks (ANN), support vector machine and others. The choice of a particular method for a certain application might to a large extent be based on the expertise of the user; however there are some factors that should be taken into consideration when selecting the procedure to be applied. One of them is the assumption about the distribution of the data: if there is the information available parametric methods can be used (for example, projection-based methods); without such information non-parametric methods are the preferred choice (k-NN, ANN). SIMCA, ANN, k-NN cope better with a large number of classes than discrimination methods. Discrimination methods show the best performance when the classes are tight, homogeneous in terms of dispersion and covariance structure; otherwise, classes should be modeled separately by means of SIMCA, for example. If samples belong to a number of definite classes, discrimination techniques are used; if not, class modeling techniques should be chosen.(23)

As mentioned above, most often in metabolomics and in particular in clinical metabolomics, the aim is to find the differences in profiles of two or more groups of samples. These differences might not be uncovered in a simple analysis. Thus larger groups have to be investigated in search of systematic variation and more sophisticated statistical methods are used. The latter can result in substantial overfitting and are more difficult to validate and interpret.

Modeling groups against each other averages the effects between the samples from one group and reduces the individual-specific variability. It already has been recognized by the clinical and especially the pharmaceutical community that averaging health and medication intervention effects between people is a dead end street for the development of future medicine and that more personalized approaches have to be found.(24)

The pharmaceutical industry was the first to respond to this need due to certain stagnation in the field, a decrease in development of new drug and the withdrawals of drugs

due to unforeseen side effects. With the genomics boom after the completion of the Human Genome project, the answers for personalized medical care and treatment were sought with the help of genomics and resulted in the emergence of a new discipline – pharmacogenomics. Despite all the expectations of this new research field, the number of genomics-driven drug discoveries is low.(25) A possible explanation is that, despite the importance of genes for defining the phenotype, they are not the only factors responsible for determining the actual physiological and/or pathological state of the organism, the response to treatment and the clinical outcome. The metabolome, on the other hand, is much closer to the phenotype in comparison to other 'omes'. The switch from genetic to metabolic "individuality" thus is logical and the need for such a switch is already recognized by the community.(26)

The idea of the connection of metabolism and human individuality and integrity is not new. It was first proposed and documented in the classical work of Sir Archibald E. Garrod "The incidence of alkaptonuria: a study in chemical individuality", the title of which speaks for itself. "...No two individuals of a species are absolutely identical in bodily structure," Garrod wrote as long ago as in 1902, "neither are their chemical processes carried out on exactly the same lines" .(27) The next progress in the field was made almost 50 years later by Roger J. Williams who demonstrated "evidence indicating that each individual possesses what may be called a "metabolic personality"-that is, a distinctive pattern of metabolic traits" and that these traits are maintained over a period of several months.(28)

As many other fundamental ideas, the idea of "metabolic personality", although maybe not enough appreciated at the time it appeared, came back in the 21st century. In the recent publications of Assfalg *et al.* and Bernini *et al.* the two collaborating groups elaborated and experimentally supported exactly the same basic thoughts – that "metabolic phenotypes" (the name changed slightly 60 years after R.J. Williams) do indeed exist and that they are stable over time.(29;30) As the analytical technologies have advanced enormously in the last decades, the analytical basis of the latest research is different from that of R.J. Williams – the use of NMR makes it possible to measure a large number of molecules in one run and to obtain precise quantitative information on these molecules. However, as shown by Assfalg *et al.*, the full variety of metabolites assessed by NMR is not necessary to define the individual metabolic patterns – equally good results can be obtained using only a limited set of 12 compounds. The latter are even to a certain extent overlapping with those measured by Williams. The importance of metabolic phenotypes in relation to disease, nutrition and response to various stimuli has been outlined in both studies.

All the environment and nutrition influences in, for instance, urine are superimposed on the invariant profiles represented by metabolic phenotypes. The existence and stability

of such distinct profiles are related to homeostasis. Homeostasis is defined as "a state of dynamic balance with the variables fluctuating between tolerance limits". The moment when "tolerance limits" are crossed and homeostasis lost can be considered as the moment of the onset of disease.(31) Detecting this occasion might enable early diagnosis, prognosis and possibilities for more successful intervention.

Both health and disease are dynamic entities, understanding of which would be possible only by monitoring them in time. As has been mentioned above, the key feature of the genome is that it is static. The metabolome, on the contrary, is highly dynamic, reflecting the changes happening over time and the reaction of the organism to the altering environment. As such, time-correlated changes of the metabolome will likely have more diagnostic and prognostic power than a single time point measurement.

There is an increasing awareness that metabolomics is of great importance for the medicine of the future due to its "personalized" and dynamic nature. However, to make this possible, the present strategies in metabolomics experiment design and data treatment should be changed. Currently metabolomics literature is dominated by "case-control" studies, which both average the effects between individuals and neglect the beneficial dynamic essence of metabolic profiles. Clinical metabolomics is not an exception, though the concept of "dynamic disease" has been around for a considerable time.(32) The advantages and the gain in information recovery obtained by dynamic profiling are starting to be recognized, but are not universally applied. In order to generalize its use it is important to change study design, sample collection and data analysis strategies.

Even when dynamic metabolomics data is being collected, the advantages of it are not always exploited as it is sometimes analyzed with the use of "classical" statistical methods, such as PCA and PLS-DA.(33) However, it has been clearly shown that those methods are not optimal for such data and optimized or new strategies should be applied.(34)

Metabolomics data generated by NMR and MS-based technologies is multivariate by its nature due to the large number of molecular species measured in one run. With the addition of the time dimension the data becomes also multilevel as different levels of variation, for instance, between- and within-individual, can be assessed. A collection of powerful methods for dealing with such data and for separating the levels of variability is thus called multilevel.(35;36) Exploring the between-individual variation allows neglecting the intra-individual changes, which may be non-systematic day-to-day differences that do not relate to the question of interest. On the other hand, the within-individual block comprises the time-related information and using this is a more personalized approach for data analysis as each person acts as its own control.

To sum up, metabolomics is an attractive methodology for clinical research as it is the most accurate reflection of the actual physiological and biochemical state of the organism. The dynamic and highly "individualized" nature of the metabolome is a strong indication that it could provide the means to make personalized medicine go all the way from an "elusive dream"(37), via "proof-of-principle", to real application. The current thesis does not offer a recipe how to do it, but describes a number of essential components for the development of this new type of medicine, such as robust and reproducible analytical methods, pre-processing routines, various data analysis methods and also metabolomics applications to both animal and clinical experiments that use the longitudinal study design.

REFERENCES

1. Lindon,J.C., Holmes,E., Bollard,M.E., Stanley,E.G., and Nicholson,J.K. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 9:1-31.

2. De Noo,M.E., Tollenaar,R.A., Deelder,A.M., and Bouwman,L.H. 2006. Current status and prospects of clinical proteomics studies on detection of colorectal cancer: hopes and fears. World J. Gastroenterol. 12:6594-6601.

3. Lippi,G., Guidi,G.C., Mattiuzzi,C., and Plebani,M. 2006. Preanalytical variability: the dark side of the moon in laboratory testing. Clin. Chem Lab Med. 44:358-365.

4. Holmes,E., Loo,R.L., Stamler,J., Bictash,M., Yap,I.K., Chan,Q., Ebbels,T., de,I.M., Brown,I.J., Veselkov,K.A. *et al* 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453:396-400.

5. Slupsky,C.M., Rankin,K.N., Wagner,J., Fu,H., Chang,D., Weljie,A.M., Saude,E.J., Lix,B., Adamko,D.J., Shah,S. *et al* 2007. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Anal Chem 79:6995-7004.

6. Lenz,E.M., Bright,J., Wilson,I.D., Hughes,A., Morrisson,J., Lindberg,H., and Lockton,A. 2004. Metabonomics, dietary influences and cultural differences: a ^{1}H NMR-based study of urine samples obtained from healthy British and Swedish subjects. J. Pharm. Biomed. Anal 36:841-849.

7. Holmes,E., Loo,R.L., Stamler,J., Bictash,M., Yap,I.K., Chan,Q., Ebbels,T., de,I.M., Brown,I.J., Veselkov,K.A. *et al* 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453:396-400.

8. Lenz,E.M., Bright,J., Wilson,I.D., Morgan,S.R., and Nash,A.F. 2003. A ^{1}H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. J. Pharm. Biomed. Anal 33:1103-1115.

9. Lindon,J.C., Nicholson,J.K., Holmes,E., Keun,H.C., Craig,A., Pearce,J.T., Bruce,S.J., Hardy,N., Sansone,S.A., Antti,H. *et al* 2005. Summary recommendations for standardization and reporting of metabolic analyses. Nat. Biotechnol. 23:833-838.

10. Lindon,J.C., and Nicholson,J.K. 2008. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trac-Trends in Analytical Chemistry 27:194-204.

11. Lindon,J.C., and Nicholson,J.K. 2008. Spectroscopic and Statistical Techniques for Information Recovery in Metabonomics and Metabolomics. Annual Review of Analytical Chemistry 1:45-69.

12. Crockford,D.J., Holmes,E., Lindon,J.C., Plumb,R.S., Zirah,S., Bruce,S.J., Rainville,P., Stumpf,C.L., and Nicholson,J.K. 2006. Statistical heterospectroscopy, an approach to the integrated analysis of NMR and UPLC-MS data sets: application in metabonomic toxicology studies. Anal Chem 78:363-371.

13. van der Greef,J., Martin,S., Juhasz,P., Adourian,A., Plasterer,T., Verheij,E.R., and McBurney,R.N. 2007. The art and practice of systems biology in medicine: mapping patterns of relationships. J. Proteome Res. 6:1540-1559.

14. Aberg,K.M., Alm,E., and Torgrip,R.J. 2009. The correspondence problem for metabonomics datasets. Anal Bioanal Chem.

15. Garcia-Perez,I., Vallejo,M., Garcia,A., Legido-Quigley,C., and Barbas,C. 2008. Metabolic fingerprinting with capillary electrophoresis. J. Chromatogr. A 1204:130-139.

16. De Meyer,T., Sinnaeve,D., Van,G.B., Tsiporkova,E., Rietzschel,E.R., De Buyzere,M.L., Gillebert,T.C., Bekaert,S., Martins,J.C., and Van,C.W. 2008. NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. Anal Chem 80:3783-3790.

17. Anderson,P.E., Reo,N.V., DelRaso,N.J., Doom,T.E., and Raymer,M.L. 2008. Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. Metabolomics 4:261-272.

18. Katajamaa,M., and Oresic,M. 2007. Data processing for mass spectrometry-based metabolomics. J. Chromatogr. A 1158:318-328.

19. van Nederkassel,A.M., Daszykowski,M., Eilers,P.H., and Heyden,Y.V. 2006. A comparison of three algorithms for chromatograms alignment. J. Chromatogr. A 1118:199-210.

20. Lange,E., Tautenhahn,R., Neumann,S., and Gropl,C. 2008. Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. BMC Bioinformatics 9:375.

21. Dieterle,F., Ross,A., Schlotterbeck,G., and Senn,H. 2006. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. Anal Chem 78:4281-4290.

22. Craig,A., Cloarec,O., Holmes,E., Nicholson,J.K., and Lindon,J.C. 2006. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal Chem 78:2262-2267.

23. Gonzalez,A.G. 2007. Use and misuse of supervised pattern recognition methods for interpreting compositional data. Journal of Chromatography A 1158:215-225.

24. Woodcock,J. 2007. The prospects for "personalized medicine" in drug development and drug therapy. Clinical Pharmacology & Therapeutics 81:164-169.

25. Goldsmith,P., Fenton,H., Morris-Stiff,G., Ahmad,N., Fisher,J., and Prasad,K.R. 2010. Metabonomics: a useful tool for the future surgeon. J. Surg. Res. 160:122-132.

26. Nicholson,J.K., Wilson,I.D., and Lindon,J.C. 2010. Pharmacometabonomics as an effector for personalized medicine. Pharmacogenomics 12:103-111.

27. Garrod,A.E. 1902. THE INCIDENCE OF ALKAPTONURIA : A STUDY IN CHEMICAL INDIVIDUALITY. The Lancet 160:1616-1620.

28. Williams,R.J., Berry,L.J., and Beerstecher,E. 1949. Individual Metabolic Patterns, Alcoholism, Genetotrophic Diseases. Proceedings of the National Academy of Sciences of the United States of America 35:265-271.

29. Assfalg,M., Bertini,I., Colangiuli,D., Luchinat,C., Schafer,H., Schutz,B., and Spraul,M. 2008. Evidence of different metabolic phenotypes in humans. Proceedings of the National Academy of Sciences of the United States of America 105:1420-1424.

30. Bernini,P., Bertini,I., Luchinat,C., Nepi,S., Saccenti,E., Schafer,H., Schutz,B., Spraul,M., and Tenori,L. 2009. Individual Human Phenotypes in Metabolic Space and Time. J. Proteome Res.

31. van der Greef,J., and Smilde,A. 2005. Symbiosis of chemometrics and metabolomics: past, present, and future. JOURNAL OF CHEMOMETRICS 19:376-386.

32. Glass,L., and Mackey,M.C. 1988. From clocks to chaos : the rhythms of life. Princeton University Press. Princeton, NJ.

33. Li,J.A., Wijffels,G., Yu,Y.H., Nielsen,L.K., Niemeyer,D.O., Fisher,A.D., Ferguson,D.M., and Schirra,H.J. 2011. Altered Fatty Acid Metabolism in Long Duration Road Transport: An NMR-based Metabonomics Study in Sheep. Journal of Proteome Research 10:1073-1087.

34. Smilde,A.K., Westerhuis,J.A., Hoefsloot,H.C.J., Bijlsma,S., Rubingh,C.M., Vis,D.J., Jellema,R.H., Pijl,H., Roelfsema,F., and van der Greef,J. 2010. Dynamic metabolomic data analysis: a tutorial review. Metabolomics 6:3-17.

35. Jansen,J.J., Hoefsloot,H.C.J., van der Greef,J., Timmerman,M.E., and Smilde,A.K. 2005. Multilevel component analysis of time-resolved metabolic fingerprinting data. Analytica Chimica Acta 530:173-183.

36. Westerhuis,J.A., van Velzen,E.J.J., Hoefsloot,H.C.J., and Smilde,A.K. 2010. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 6:119-128.

37. Lesko,L.J. 2007. Personalized medicine: Elusive dream or imminent reality? Clinical Pharmacology & Therapeutics 81:807-816.

Part I

Method development

Chapter 1

Gas chromatography/atmospheric pressure chemical ionization-time of flight mass spectrometry: analytical validation and applicability to metabolic profiling

Carrasco-Pancorbo A., Nevedomskaya E., Arthen-Engeland T., Zey T., Zurek G., Baessmann C., Deelder A.M., Mayboroda O.A.

Analytical Chemistry **2009**, 81, 10071–10079

ABSTRACT

Gas Chromatography (GC)-Mass Spectrometry (MS) with Atmospheric Pressure (AP) interface was introduced more than 30 years ago but never became a mainstream technique, mainly because of technical difficulties and cost of instrumentation. A recently introduced multipurpose AP source created the opportunity to reconsider the importance of AP ionization for GC. Here, we present an analytical evaluation of GC/APCI-MS showing the benefits of soft atmospheric pressure chemical ionization for GC in combination with a Time of Flight (TOF) mass analyzer. During this study, the complete analytical procedure was optimized and evaluated with respect to characteristic analytical parameters, such as repeatability, reproducibility, linearity, and detection limits. Limits of detection (LOD) were found within the range from 11.8 to 72.5 nM depending on the type of compound. The intraday and interday repeatability tests demonstrate relative standard deviations (RSDs) of peak areas between 0.7%-2.1% and 3.8%-6.4% correspondingly. Finally, we applied the developed method to the analysis of human cerebrospinal fluid (CSF) samples to check the potential of this new analytical combination for metabolic profiling.

INTRODUCTION

There are different definitions of metabolomics. However, regardless of terminology and phrasing differences, any definition implies an enormous analytical challenges to cover a wide range of polarities, concentrations, and sizes of chemical entities composing the human metabolome. In response to this challenge, more and more efforts are directed toward cross-platform analysis and integration of data obtained on different analytical platforms. At the same time, a revision and modernization of proven technologies like, for example, gas chromatography (GC) is taking place. Since it was invented by Martin and James (1) in 1952, GC became one of the most important and widely applied techniques in modern analytical chemistry. Even before the term "metabolomics" was introduced, there were a number of published studies with GC as main analytical method, which could be described as metabolomics or metabolic profiling.(2) However, only with the introduction of fused-silica capillary columns at the end of 1980s, which significantly improved the separation quality of GC, and GC-MS instrumentation, GC turned into the one of the most effective techniques for large scale metabolic profiling.(3-8) GC-MS was the first analytical technique implemented in a real metabolic profiling workflow. It includes all steps from sample preparation to the compound identification and remains flexible because of a number of options in selection of mass analyzer and ionization techniques. There are several types of mass analyzers routinely used with GC systems, namely, ion trap (IT), single (Q) and triple-quadrupoles (QqQ), and time of flight (TOF). However, the characteristics of a TOF mass analyzer are most favorable for such application as metabolic profiling. Speed, sensitivity, resolving power, and multiplex detection are clear advantages over scanning instruments, such as quadrupoles. These performance factors make TOF mass analyzers almost ideal for metabolomics, especially in combination with GC.(9) Moreover, modern TOF analyzers provide a data quality sufficient for identification of metabolites using a combination of accurate mass, isotopic distribution, and retention time.(10;11)

Most of the commercial GC-MS systems use ionization under vacuum conditions: electron impact ionization (EI) and chemical ionization (CI).(12) EI is considered to be a hard ionization technique, meaning that the energy of the electrons is high enough to produce highly reproducible fragmentation patterns of small molecules. Characteristic fragmentation patterns make GC/EI-MS a powerful analytical technique for comparing the mass spectra of unknown substances to data sets of commercial and open source 70 eV EI mass spectral libraries. However, the fragmentation of the compounds is sometimes so strong that it impairs the structural significance of the parent ion. On the contrary, CI where ions are formed because of the reaction with reagent gas is a softer ionization

technique and energy transfer usually does not exceed 5 eV. Consequently, fewer fragments are formed and information about the precursor ion is preserved. Moreover, since the fragmentation pattern depends on the properties of the reagent gas, different structural information can be obtained from different reagent gases. Atmospheric pressure ionization sources (API), which are probably the key of the "overnight success" of MS detectors in analytical sciences because of coupling with liquid chromatography, are rarely used with GC instruments. The first APCI source for GC-MS was described more than 30 years ago by Horning *et al.*(13-16) Later, several papers were published in which the effluent from a GC is ionized at atmospheric pressure with an interface coupling the GC to a ^{63}Ni ion source of a mass spectrometer built for APCI gas-phase studies.(17-19) Revelsky *et al.*(20) and Schiewek *et al.*(21) have applied GC/APPI-MS for analyzing a wide variety of volatile organic compounds, and ESI has been successfully applied for ionization of gaseous analytes separated by GC.(22;23) Even so, GC/API-MS has never become widely used, in part because of the high costs of the custom instrumentation needed for these analyses, in part because of availability of commercial "plug and play" EI and CI GC systems. Recently, Schiewek *et al.* introduced a new multipurpose API source, which for the first time offers a "user friendly" and robust solution for a GC/APCI technique.(24) In the current manuscript, we present a detailed analytical evaluation of GC/APCI in combination with a TOF mass spectrometer. In addition to the detailed examination of the analytical performance (repeatability, reproducibility, linearity, and detection limits), we demonstrate the applicability of this technique for metabolic profiling of cerebrospinal fluid (CSF).

MATERIALS AND METHODS

Chemicals. A standard solution of 17 amino acids at 1 mM each in 0.1 M HCl was purchased from Sigma-Aldrich. 4-Nitrobenzoic acid, dopamine hydrochloride, and Phe-Gly hydrate were obtained from Fluka. Sarcosine, theophylline, caffeine, nortriptyline hydrochloride, hippuric acid, creatinine, 4-*O*-methyldopamine hydrochloride, homovanillyl alcohol, benzoic acid, uric acid, and 5-hydroxyindole-3-acetic acid were acquired from Sigma. Stock standard solutions of the 31 compounds under study were prepared in methanol at a concentration of 200 µM. N,O-bis(Trimethylsilyl)trifluoroacetamide with 1% trimethylchlorosilane (BSTFA + 1% TMCS) and N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (MSTFA + 1% TMCS) from Pierce (Rockford, IL, U.S.A.) were used as derivatization reagents. These reagents were used from freshly opened 1 mL bottles. Methoxyamine hydrochloride was purchased from Supelco. Methanol (HPLC grade) was acquired from Sigma-Aldrich and pyridine (>99%, ultrapure GC grade) was from Fluka.

Biological Samples. Human CSF samples were taken by lumbar puncture. The study was approved by the ethical committee at the Leiden University Medical Center. Samples were processed within 1 h, centrifuged at 300 × g to remove cells, aliquoted and stored at -80 °C until use.

Protein Precipitation and Metabolite Extraction. 250 µL sample aliquots were taken, 600 µL of cold extraction solvent (MeOH) were added, and the sample was shaken vigorously for 20 s. The samples were placed in an ice bath for 2 h, and then centrifuged at 20,800 rcf for 15 min. The liquid supernatant was collected and evaporated in a speed vacuum concentrator before derivatization.

Derivatization. A speed vacuum concentrator or lyophilizer was used for drying the standard mixture (100 µL at 100 µM) and the CSF extracts to complete dryness. A mixture of 20 mg/mL of methoxyamine·HCl in pyridine was freshly prepared using an ultrasonicator. The dried samples were taken from store and warmed up to room temperature before starting derivatization. Methoxyamine + pyridine mixture (100 µL) was added to each GC vial, closing it immediately, and the samples were agitated for 2 min. Methoxyamination was performed at 40 °C for 60 min. After the addition of the derivatization reagent containing 1% TMCS as the catalyst (100 µL) the solution was vortexed again for 2 min. Trimethylsilylation reaction was performed at 40 °C for 30 min. A minimum of 2 h equilibration time was necessary before sample injection.

GC-MS Analysis. The derivatized samples (1 µL) were applied by splitless injection with a programmable CTC PAL multipurposesampler (CTC Analytics AG, Zwingen, Switzerland) into an Agilent 7890A GC (Agilent, Palo Alto, U.S.A.) equipped with a HP-5-MS column (30 m, 0.25 mm ID, 0.25 µm film thickness). Injection programs included sequential washing steps of the 10 µL syringe before and after the injection, and a sample pumping step for removal of small air bubbles.

The injection temperature was set at 250 °C. Helium was used as carrier gas at a constant flow rate of 1 mL/min through the column. For every analysis splitless injection time was 60 s and after this the injector was purged at 20 mL/min flow rate. The column temperature was initially kept at 70 °C for 5 min and then raised at 5 °C/min over 42 min to 280 °C and held for 10 min.

The GC transfer line to the mass spectrometer was kept at 280 °C. The APCI source and MS were operated in positive mode, temperature and flow rate of the dry gas (nitrogen) were 250 °C and 5.00 L/min, respectively. The APCI vaporizer temperature was 450 °C; the pressure of the nebulizer gas (nitrogen) was set to 2 bar, and the voltage of the corona discharge needle was 2000 nA. Capillary voltage was set at -1000 V and the end-plate offset at -1000 V.

As a detector an orthogonal-accelerated TOF mass spectrometer (oaTOF-MS) MicroTOF (Bruker Daltonik, Bremen, Germany) was used. The polarity of the APCI interface and all the parameters of TOF MS detector were optimized using the area of the MS signal for the metabolites included in the standard mixture and the chromatographic resolution as analytical parameters. The position of the column in the transfer line, the transfer line temperature, the flow rate and pressure of nebulizer gas (nitrogen), the vaporizer temperature, voltages in the corona, source and ion transfer settings: all those parameters were optimized empirically. These are essential for optimal performance of an instrument but may vary from instrument to instrument.

Data were acquired for mass range from 50 to 1000 m/z with a repetition rate of 1 Hz. DataAnalysis 4.0 software (Bruker Daltonik) was used for data processing. The SmartFormula tool within DataAnalysis was used for the calculation of elemental composition of compounds; it uses a CHNO algorithm, which provides standard functionalities such as minimum/maximum elemental range, electron configuration, and ring-plus double bonds equivalents, as well as a sophisticated comparison of the theoretical with the measured isotope pattern (Sigma-Value) for increased confidence in the suggested molecular formula.(11)

The instrument was calibrated externally using an APCI calibration tune mix. Because of the compensation of temperature drift in the mass spectrometer, this external calibration provided consistent mass values for a complete experimental sequence. Moreover, an additional internal calibration was performed using cyclic-siloxanes, a typical background in GC-MS.(25;26)

Linearity and Sensitivity. Linearity of the detector response (TOF-MS) was verified with standard solutions containing the 31 analytes under study at 5 different concentration levels over the range from the quantification limit to 100 μM. Each point of the calibration graph corresponded to the mean value from three independent replicate injections. Calibration curves were obtained for each standard by plotting the standard concentration as a function of the peak area obtained from GC/APCI-TOF MS analyses. The sensitivity of the analytical procedure was calculated by defining the limits of detection (LOD) and quantification (LOQ) for the individual analytes in standard solutions according to the IUPAC method.(27) The smallest concentration that could be detected with a reasonable certainty for our analytical procedure (LOD) was considered S/N = 3, while LOQ was S/N = 10.

Precision and Accuracy. The precision of the analytical procedure described was measured as repeatability and reproducibility. Quality control (QC) samples were tested in six replicates (at an intermediate concentration value of the calibration curve) and

calculated with calibration curves obtained daily. The precision of the analytical procedure was expressed as the relative standard deviation (RSD). The intra- and interday repeatability in the peak areas was determined as the RSD obtained for six consecutive injections of each metabolite at an intermediate concentration value of the calibration curve, carried out within the same day and on three different days.

RESULTS AND DISCUSSION

Selection of Derivatization Conditions. BSTFA (+1% TMCS) and MSTFA (+1% TMCS) were used as derivatization reagents. They react with a range of polar compounds by replacing active hydrogen in alcohols, amines, carboxylic acids, and so forth. To find optimal derivatization conditions, we studied effects of derivatization time and temperature and the concentration ratio of the derivatization reagent to the concentration of pyridine/methoxyamine.

Regardless of the derivatization reagent, changing the reagent to pyridine/methoxyamine ratio from 0.8:1.2 until 1.2:0.8 did not affect peak areas of the test mixture significantly. Thus, the ratio 1:1 was chosen for further experiments. The effect of the derivatization time on peak areas was most significant in the interval between 10-30 min. Starting from 30 min incubation peak areas remained constant and further increase of derivatization time had little impact on data quality (Supplementary Materials, Figure S1). Thus, to reduce the error and shorten time, 30 min was selected as derivatization time. The influence of temperature on peak areas was minimal, at least in the evaluated interval between room temperature and 80 °C. However, at 40 °C we observed more compounds with just one TMS derivative. Thus, the final derivatization protocol consisted of a methoxyamination step (40 °C for 60 min) and subsequent trimethylsilylation (MSTFA + 1% TMCS, at 40 °C for 30 min).

The stability of derivatized samples is an important factor for large scale metabolomics temperature and performed analysis in equal time intervals between 0 and 72 h. Data proved to be rather consistent from 0 to 65 h. However, data collected on later time points demonstrated steadily increasing variability. Nevertheless, to avoid any possible risk of derivatization-dependent variability, material should preferably be processed within the first 48 h.

GC/APCI-TOF MS Analysis of Standard Mixture. A standard mixture consisting of 31 compounds was used for the general test of performance and evaluation of analytical parameters. The compounds were selected with the aim to cover a range of polarities and molecular weights of the metabolites typically reported as components of body fluids. Table 1 represents our test mixture grouped in different chemical families, such as amines, amino

acids, organic acids, alcohols, xanthines, compounds with indole or imidazole groups, and one dipeptide.

Table 1. Compounds Included in the Standard Mixture

amino acids	alanine
	arginine
	aspartic acid
	cysteine
	glutamic acid
	glycine
	histidine
	isoleucine
	leucine
	lysine
	methionine
	phenylalanine
	proline
	serine
	threonine
	tyrosine
	valine
	sarcosine
organic acids	benzoic acid
	hippuric acid
	4-nitrobenzoic acid
alcohols	homovanillyl alcohol
xanthines and related coumpounds	caffeine
	theophylline
	uric acid
compound with indoles group	5-hydroxyindole-3-acetic acid
amines	nortriptyline hydrochloride
Compounds with hydroxyl and amine groups	dopamine hydrochloride
	4-O-Methyldopamine hydrochloride
compounds with imidazol groups	creatinine
dipeptides	Phe-Gly hydrate

Figure 1 represents a combined extracted ion chromatogram (EIC) of the standard mixture recorded with optimum GC and MS settings.

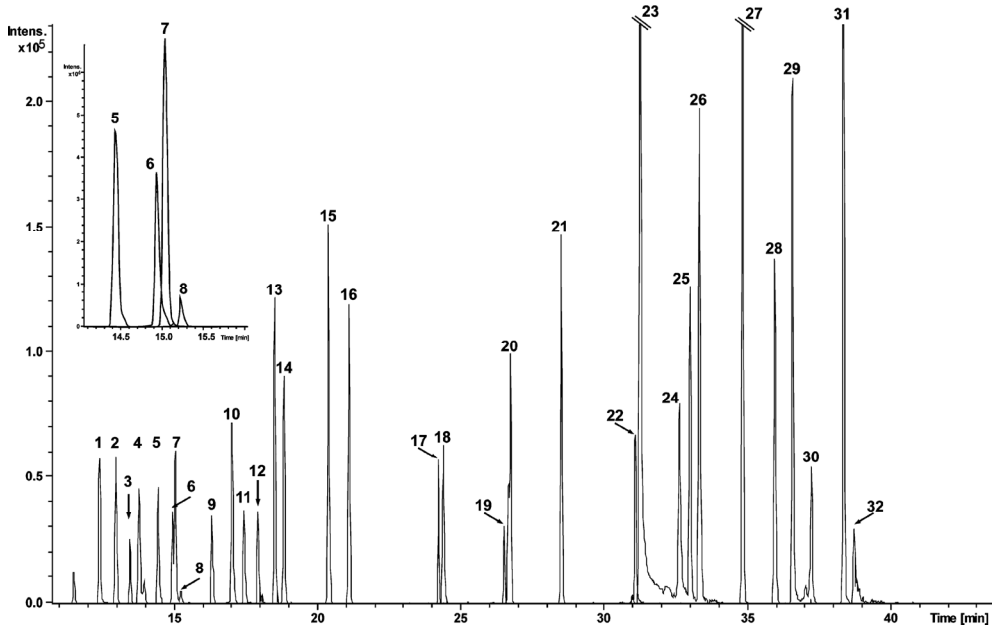

Figure 1. Extracted ion GC/APCI-TOF MS chromatograms of the 32 features corresponding to 25 compounds of the standard mix (100 μM). Numbering of compounds corresponds to Table 2.

With an analytical window of approximately 30 min, we observed 32 peaks, which could be assigned to 25 compounds. Table 2 shows all analytes detected, with their formula, retention time, measured and theoretical m/z, error (mDa) and sigma value. All values were calculated from samples with concentrations close to the LOQ; nevertheless the mass position error remained within 1.0 mDa and high quality sigma fit values (<10 mSigma) were obtained for all compounds.

However, the same table (Table 2) demonstrates that we failed in detecting a few components of our test mixture, namely, three amino acids (arginine, cysteine, and histidine), one organic acid (4-nitrobenzoic acid), homovanillyl alcohol, and creatinine. The thermal instability of amino acids, especially arginine and cysteine, is a known problem and has already been addressed in literature.(28;29) In addition, treatment with silylation reagents, even under the mild conditions generally employed in metabolite profiling, can lead to chemical conversion.(30) For example, arginine can be converted to ornithine. When studying metabolic processes in detail, particularly where the intermediate compounds may be reactive or unstable, one should always be aware of such possibilities when interpreting the results. If there is any doubt, alternative derivatization procedures for specific functional groups should be considered.

Table 2. Forms of the Different Compounds Included in the Standard Mixture (at a Concentration Close to LOQ) Detected with GC/APCI-TOF MS Method.

peak ID	compound	formula (peak found)	retention time	m/z experimental	m/z calculated	error (mDa)	mSigma value
1	Valine+1TMS+H	$C_8H_{20}NO_2Si$	12.4	190.1256	190.1258	0.21	3.4
2	Alanine+2TMS+H	$C_9H_{24}NO_2Si_2$	13	234.1338	234.134	0.2	5.1
3	Glycine+2TMS+H	$C_8H_{22}NO_2Si_2$	13.5	220.1181	220.1184	0.31	4.6
4	Sarcosine+2TMS+H	$C_9H_{24}NO_2Si_2$	13.8	234.1338	234.134	0.21	4.4
5	Leucine+1TMS+H	$C_9H_{22}NO_2Si$	14.4	204.1414	204.1414	0	1.8
6	Proline+1TMS+H	$C_8H_{18}NO_2Si$	14.9	188.1108	188.1101	-0.7	2.2
7	Isoleucine+1TMS+H	$C_9H_{22}NO_2Si$	15	204.1409	204.1414	0.49	1.8
8	Uric acid+3TMS+H	$C_{14}H_{29}N_4O_3Si_3$	15.2	385.1545	385.1542	0.31	3.4
9	Valine+2TMS+H	$C_{11}H_{28}NO_2Si_2$	16.3	262.1656	262.1653	-0.29	6.1
10	Benzoic acid+1TMS+H	$C_{10}H_{15}O_2Si$	17	195.087	195.0877	0.7	1.8
11	Serine+2TMS+H	$C_9H_{24}NO_3Si_2$	17.4	250.129	250.1289	-0.1	1.6
12	Leucine+2TMS+H	$C_{12}H_{30}NO_2Si_2$	17.9	276.1813	276.181	-0.3	8.9
13	Isoleucine+2TMS+H	$C_{12}H_{30}NO_2Si_2$	18.5	276.1802	276.181	0.8	6.4
14	Glycine+3TMS+H	$C_{11}H_{30}NO_2Si_3$	18.8	292.1578	292.1579	0.09	4.7
14	Serine+3TMS+H	$C_{12}H_{32}NO_3Si_3$	20.4	322.1681	322.1684	0.29	5.1
16	Threonine+3TMS+H	$C_{13}H_{34}NO_3Si_3$	21.1	336.1834	336.1841	0.71	2.9
17	Methionine+2TMS+H	$C_{11}H_{28}NO_2SSi_2$	24.2	294.1376	294.1374	-0.2	1.3
18	Aspartic acid+3TMS+H	$C_{13}H_{32}NO_4Si_3$	24.4	350.1631	350.1634	0.32	1.6
19	Glutamic acid+3TMS+H	$C_{14}H_{34}NO_4Si_3$	26.6	364.1786	364.179	0.4	5.5
20	Phenylalanine+2TMS+H	$C_{15}H_{28}NO_2Si_2$	26.7	310.1653	310.1653	0	5.8
21	Phenyl-Gly+H	$C_{11}H_{15}N_2O_3$	28.5	223.108	223.1077	-0.29	3.3
22	Hippuric acid+1TMS+H	$C_{12}H_8NO_3Si$	31.1	252.1047	252.105	0.3	1.8
23	Caffeine+H	$C_8H_{11}N_4O_2$	31.2	195.0835	195.0836	0.1	3.2
24	Theophylline+1TMS+H	$C_{10}H_{17}N_4O_2Si$	32.6	253.1116	253.1115	-0.1	2.9
25	Lysine+4TMS+H	$C_{18}H_{47}N_2O_2Si_4$	33	435.2699	435.2709	1	2.5
26	Tyrosine+3TMS+H	$C_{18}H_{36}NO_3Si_3$	33.3	398.1999	398.1998	-0.12	5.5
27	4-Methyldopamine hydrochlor+3Si+H	$C_{18}H_{38}NO_2Si_3$	34.8	384.2199	384.2205	0.61	4.2
28	Dopamine hydrochlor+4TMS+H	$C_{20}H_{44}NO_2Si_4$	35.9	442.2448	442.2444	-0.39	4
29	Uric acid+4TMS+H	$C_{17}H_{37}N_4O_3Si_4$	36.5	457.1939	457.1937	-0.18	9.1
30	Phenyl-Gly+2TMS+H	$C_{17}H_{31}N_2O_3Si_2$	37.2	367.1869	367.1868	-0.11	5.6
31	5-hydroxyindole-3-acetic+3TMS+H	$C_{19}H_{34}NO_3Si_3$	38.3	408.1842	408.1841	-0.08	7.7
32	Nortriptyline hydrochlor+H	$C_{19}H_{22}N$	38.7	264.1744	264.1747	0.29	9.2

Analysis of creatinine by GC requires rather selective conditions, which are optimal only for creatinine itself and a few related compounds. Creatinine can be converted, for instance, to the ethyl ester of N-(4,6-dimethyl-2-pyrimidinyl)-N-methylglycine,(31) or derivatized

with trifluoroacetic anhydride,(32) although the last one has been mainly analyzed by HPLC. The same is true for 4-nitrobenzoic acid or homovanillyl alcohol. In general, we can conclude that those two compounds are analyzed more properly by HPLC.

At a first glance, the few compounds "missing" from our test mixture might be considered as serious drawback of the total workflow. However, metabolic profiling workflows always imply a compromise between analytical limitations of the methods and their applicability. Even more, as Fiehn *et al.*(33) formulated in their validation criteria for metabolite profiling protocols, comprehensiveness is more important than inclusion of a certain metabolite, and the overall dynamic range for the majority of the compounds is more important than the detection limit for one specific substance. Thus, we measured compounds belonging to nine different chemical families within one experiment (chromatogram). Moreover, the correct elementary composition of measured compounds was calculated from data acquired at levels close to the LOQ(11;34). Considering the chromatographic behavior, mass accuracy, and isotopic distribution, the described method could distinguish between isomers (i.e., Alanine/Sarcosine; Isoleucine/ Leucine).

Analytical Parameters. Calibration curves were obtained for each standard by plotting the standard concentration as a function of the peak area obtained from GC/APCI-TOF MS analyses. The parameters of the calibration functions, LOD, calibration range, correlation coefficient, precision, and accuracy are summarized in Table 3.

To calculate the calibration functions and LOD's, we took the EIC of the most intense or base peak in the mass spectrum for each compound in the standard mixture. If the compound was represented by more than one silylated form, the one with higher linearity in the calibration range was used for calculation of analytical parameters. For example, in the case of glycine, we used glycine+3TMS+H; for isoleucine, isoleucine+1TMS+H; for leucine, leucine+1TMS+H; for serine, serine+3TMS+H; for valine, valine+1TMS+H; for uric acid, uric acid+4TMS+H; and in the case of Phe-Gly hydrate, we used Phe-Gly+H. The results summarized in Table 3 indicate that the GC/APCI-TOF MS method is a reliable approach for the analysis of a wide range of compounds. LODs were found within the range from 11.8 to 72.5 nM depending on the type of compound. To the best of our knowledge, these LOD values are considerably lower than the normal values previously described in literature for the determination of this kind of compounds by GC-MS.(29;35-37) Still, the brief overview of the values reported in literature (Supplementary Materials, Table S1) for more "classical" GC-MS systems shows how difficult it is to do a fair comparison with APCI-GC. LOD and LOQ values usually reported in the studies targeted to one or two classes of the metabolites. On the contrary, our standard mixture was designed to mimic a profiling condition and includes compounds belonging to nine different chemical families.

Table 3. Analytical Parameters for the GC/APCI-TOF MS Method Described (Positive Polarity).

Compounds	LOD (μM x 10^{-3})	LOQ (μM x 10^{-3})	calibration range (μM)	calibration curve[a]	r^2	considered ion	repeat. intra day[b]	repeat. inter day[b]	reproducibility[c]	accuracy[d]
Valine+1TMS+H	28.5	95		y = 88131x - 770846	0.9311	190.1256	1.23	5.16	5.98	95.7
Alanine+2TMS+H	48.7	162.3		y = 51606x - 1007331	0.9136	234.1338	1.51	4.1	6.39	96.4
Sarcosine+2TMS+H	55.1	183.7		y = 45585x + 1523067	0.921	234.138	1.9	3.76	7.01	98.3
Leucine+1TMS+H	24.7	82.3		y = 101739x - 1476560	0.902	204.1414	0.95	5.37	7.55	102.1
Proline+1TMS+H	72.5	241.7		y = 34612x + 45712	0.9694	188.1108	0.73	5.05	6.5	101.6
Isoleucine+1TMS+H	25.3	84.3		y = 99186x - 1232551	0.9841	204.1409	1.87	6.37	8.78	99.5
Benzoic acid+1TMS+H	39.4	131.3		y = 63729x + 201338	0.9196	195.087	1.65	4.74	7.04	98.5
Glycine+3TMS+H	25.2	84		y = 99579x + 1491920	0.9757	292.1578	1.23	6.01	6.57	99.2
Serine+3TMS+H	36.2	120.7		y = 69450x - 719359	0.9814	322.1681	2.09	4.22	6.88	96.1
Threonine+3TMS+H	35	116.7		y = 71798x - 170943	0.9233	336.1834	1.87	4.11	7.09	95.5
Methionine+2TMS+H	45.4	151.3		y = 55355x - 619794	0.9867	294.1376	1.11	5.01	6.44	98.1
Aspartic acid+3TMS+H	38.2	127.3		y = 65717x + 213067	0.9338	350.1631	0.89	5.55	6.01	98.7
Glutamic acid+3TMS+H	48.1	160.3	QL-100	y = 52232x - 760373	0.9391	364.1786	1.09	4.11	6.55	97.3
Phenylalanine+2TMS+H	32.6	108.7		y = 76927x - 488942	0.9057	310.1653	1.21	4.09	6.32	98.3
Phenyl-Gly+H	18.5	61.7		y = 135673x + 2391638	0.9052	223.108	1.56	5.65	7.11	96.1
Hippuric acid+1TMS+H	16.6	55.3		y = 151185x - 2387767	0.962	252.1047	1.76	6.01	6.21	98.2
Caffeine+H	11.8	39.3		y = 212078x - 1766041	0.967	195.0835	1.44	5.98	6.09	98.1
Theophylline+1TMS+H	14.5	48.3		y = 172904x - 4035222	0.9309	253.1166	1.32	4.89	6.81	96.1
Lysine+4TMS+H	22.2	74		y = 112802x - 853604	0.9372	435.2699	1.78	5.01	7.45	97.7
Tyrosine+3TMS+H	19.1	63.7		y = 131596x - 1252913	0.9334	398.1999	1.66	5.22	8.9	96.3
4-Methyldopamine+3TMS+H	17.5	58.3		y = 143581x + 2238266	0.9785	384.2199	1.56	5.76	7.91	95.9
Dopamine+4TMS+H	18.7	62.3		y = 134231x + 1841824	0.976	442.2448	1.01	4.33	6.56	100.8
Uric acid+4TMS+H	23.8	79.3		y = 105543x - 1922989	0.9372	457.1939	0.91	4.52	8.32	99.7
5-Hydroxyindole-3-	17	56.6		y = 147645x + 709932	0.9339	408.1842	0.9	4.25	8.76	98.4
Nortriptyline+H	65.9	219.7		y = 38120x - 758169	0.9665	264.1744	1.81	5.11	6.04	96.4

[a] A (peak area)= a + b × C (μM) for five points (n = 5).
[b] RSDs values (%) for peak areas corresponding to each compound; measured from three injections of each analyte within the same day (intra-) and on three different days (inter-).
[c] RSDs values (%) from two consecutive injections with two different technicians and within two different days.
[d] The accuracy of the assay is the closeness of the test value obtained to the nominal value. It is calculated by determining trueness and precision. (%Recovery, %RSD).

Moreover, a proper comparison of APCI and EI/CI could only be done if data are obtained on the same mass analyzer type, with the same sample preparation and derivatization strategies. At the end, the output still will not be 100% conclusive. We see as more beneficial the strategy, which will explore complementarities of both methods, combining high quality MS data generated under APCI condition with highly reproducible fragmentation spectra of EI.

Finally, we calculated the two most important parameters for evaluation of the precision of the analytical procedure: repeatability and reproducibility. In terms of repeatability; calculated RSDs did not exceed 6.37%. Reproducibility was determined by calculation the RSDs values (%) from two consecutive injections with two different technicians and within two different days and it did not exceed 8.90%.

Applicability of GC/APCI-TOF MS for Metabolic Profiling in Biological Samples. A human CSF pool was extracted, dried, derivatized, and analyzed by GC/APCI-TOF MS as described above (see Materials and Methods). At first, we compared the chromatograms of the human CSF with those obtained for the standard mixture. Confirmation of compounds identity was accomplished by comparing retention time, mass position, and isotopic pattern of standards and sample.

Figure 2A shows the metabolic profile of human CSF as base peak chromatogram. The observed complexity and richness of the chromatogram demonstrates the potential of the method. In Figure 2B we show several EICs of metabolites, which were identified in the CSF. Several of them were assigned using only mass position and isotopic distribution. Supplementary Materials, Figure S2 shows an example of such assignment for N-acetyl-aspartate.

Table 4 contains information concerning the compounds of our standard mixture found in the human CSF (formula, retention time, experimental m/z and theoretical, mass error and sigma value). Even in this case of analyzing an extremely complex biological sample, the accurate measurements (very low mass error) and the isotopic distribution evaluation (sigma value) obtained by TOF MS could confirm the identity of the analytes.

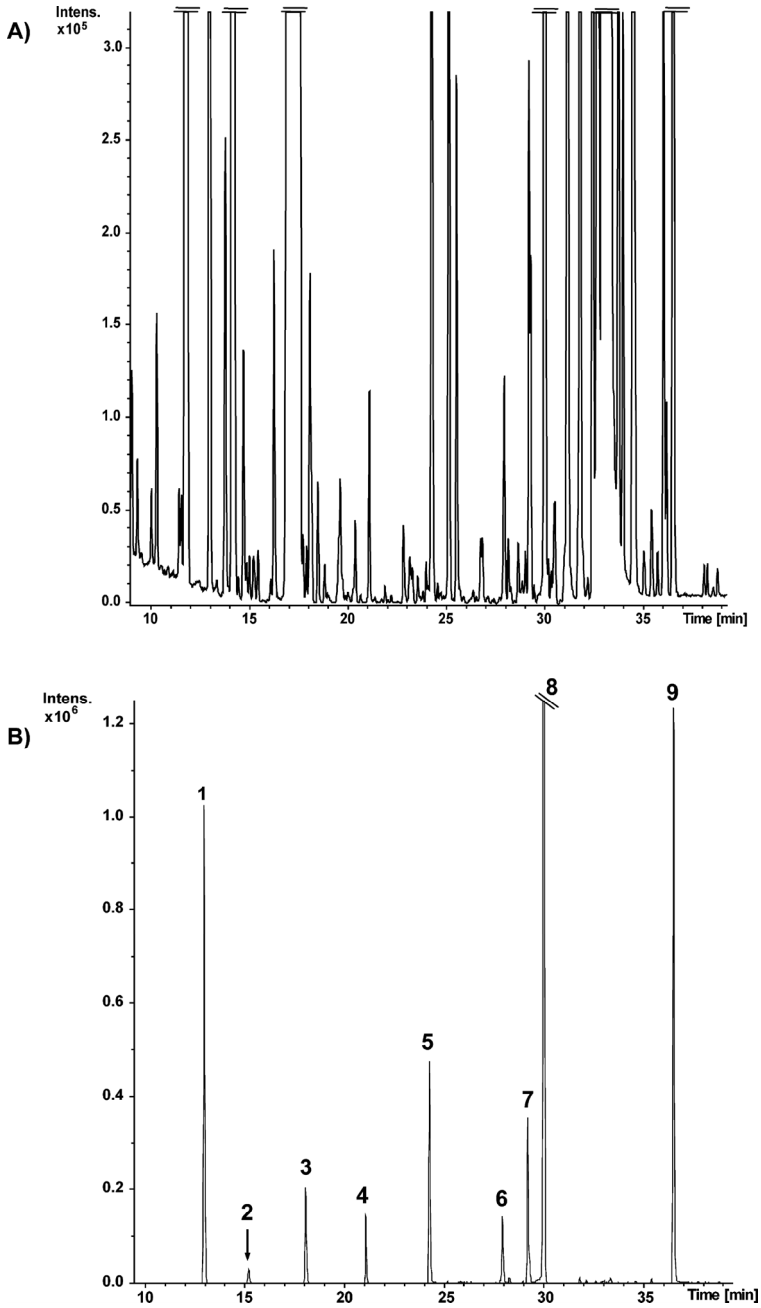

Figure 2. GC/APCI-TOF MS analysis of CSF sample: (A) Base peak chromatogram of the derivatized CSF sample. (B) EICs of several identified metabolites; peaks 1 (Glycine), 2 (Uric acid), 4 (Threonine) assigned with help of standards, peaks 3 (Glycerol), 5 (Pyroglutamic acid), 6 (N-acetyl-aspartate), 7 (Ribitol), 8 (Glutamine), 9 (Glucose) assigned using mass position and isotopic pattern.

Table 4. Compounds Included in the Standard Mixture Found in Human CSF Samples

compound	formula (peak found)	retention time	*m/z* experimental	*m/z* calculated	error (mDa)	mSigma value
Valine+1TMS+H	$C_8H_{20}NO_2Si$	12.4	190.1245	190.1258	1.29	5.2
Alanine+2TMS+H	$C_9H_{24}NO_2Si_2$	13	234.133	234.134	1	4.5
Glycine+2TMS+H	$C_8H_{22}NO_2Si_2$	13.1	220.1171	220.1184	1.3	5.1
Sarcosine+2TMS+H	$C_9H_{24}NO_2Si_2$	13.8	234.1338	234.134	0.21	2.7
Leucine+1TMS+H	$C_9H_{22}NO_2Si$	14.4	204.1404	204.1414	1	4.1
Isoleucine+1TMS+H	$C_9H_{22}NO_2Si$	15	204.1422	204.1414	-0.79	2.7
Uricacid+3TMS+H	$C_{14}H_{29}N_4O_3Si_3$	15.2	385.153	385.1542	1.19	3.7
Valine+2TMS+H	$C_{11}H_{28}NO_2Si_2$	16.3	262.1653	262.1653	0	5.3
Benzoicacid+1TMS+H	$C_{10}H_{15}O_2Si$	17	195.0869	195.0877	0.8	8.4
Serine+2TMS+H	$C_9H_{24}NO_3Si_2$	17.4	250.129	250.1289	-0.1	1.9
Leucine+2TMS+H	$C_{12}H_{30}NO_2Si_2$	17.9	276.1815	276.181	-0.49	10.1
Isoleucine+2TMS+H	$C_{12}H_{30}NO_2Si_2$	18.5	276.1823	276.181	-1.3	6.4
Glycine+3TMS+H	$C_{11}H_{30}NO_2Si_3$	18.8	292.1574	292.1579	0.5	9.3
Serine+3TMS+H	$C_{12}H_{32}NO_3Si_3$	20.4	322.1689	322.1684	-0.52	5.2
Threonine+3TMS+H	$C_{13}H_{34}NO_3Si_3$	21.1	336.1861	336.1841	-1.98	5.6
Methionine+2TMS+H	$C_{11}H_{28}NO_2SSi_2$	24.2	294.138	294.1374	-0.59	7.4
Asparticacid+3TMS+H	$C_{13}H_{32}NO_4Si_3$	24.4	350.1631	350.1634	0.32	1.6
Phenylalanine+2TMS+H	$C_{15}H_{28}NO_2Si_2$	26.7	310.1663	310.1653	-0.99	6.9
Phenyl-Gly+H	$C_{11}H_{15}N_2O_3$	28.5	223.1082	223.1077	-0.49	3.3
Hippuricacid+1TMS+H	$C_{12}H_8NO_3Si$	31.1	252.1047	252.105	0.3	1.8
Caffeine+H	$C_8H_{11}N_4O_2$	31.2	195.0835	195.0836	0.1	3.5
Theophylline+1TMS+H	$C_{10}H_{17}N_4O_2Si$	32.6	253.1116	253.1115	-0.1	2.9
Lysine+4TMS+H	$C_{18}H_{47}N_2O_2Si_4$	33	435.2697	435.2709	1.22	2.5
Tyrosine+3TMS+H	$C_{18}H_{36}NO_3Si_3$	33.3	398.1999	398.1998	0.12	5.5
Uricacid+4TMS+H	$C_{17}H_{37}N_4O_3Si_4$	36.5	457.1949	457.1937	-1.19	9.1
5-hydroxyindole-3-acetic+3TMS+H	$C_{19}H_{34}NO_3Si_3$	38.3	408.1834	408.1841	0.69	7.7
Nortriptyline+H	$C_{19}H_{22}N$	38.7	264.1734	264.1747	1.29	9.2

In total, our method was capable to determine more than 300 compounds with different isotopic features in the CSF sample. As commented before, the identity of some of those peaks could be corroborated by the standards included in our mixture, but in other cases, we used mass position and isotopic distribution to achieve the identification of the analytes present in the CSF according to their molecular formula. Some examples are included in Figure 2B, where we have shown the EICs of silylated forms of uric acid, glycerol, pyroglutamic acid, N-acetyl-aspartate, ribitol, glutamine, and glucose. The values of mass error and sigma value for the mentioned compounds were excellent, showing the capability of our GC/APCI-TOF MS method to confirm the identity of an important number of metabolites which can be found in CSF samples. However, being strict we should

discriminate between assignments validated by data from the standard mixture and those which were made solely based on sigma value calculation. If in the first case the reference to standard makes an assignment almost 100% correct, the second one is the best guess possible on the basis of available data. In Figure 3 we have shown MS spectra produced by GC/APCI-TOF MS for some compounds found in human CSF. Included compounds belong to different chemical families: amino acids, xanthines, organic acids, indoles and amines.

Valine was detected as valine+1TMS+H (m/z 190.1245), according to the reaction described above $[M+H]^+$ (in the current case $[M+1TMS+H]^+$), observing mainly the mentioned m/z signal and not its fragments. In the case of glycine and aspartic acid, the main peak in the spectrum was the amino acid+3TMS+H. Because of in source-fragmentation, some fragments were also observed. A neutral loss of 72.0387 appears after losing one of the trimethylsilane (TMS) groups, more precisely -OH replacement with -$OSi(CH_3)_3$, (=$[C_3H_8Si]$), trimethylsiloxane. The loss of two TMS groups should lead to $[M-2TMS+H]^+$, resulting in a loss of 144.0785. Moreover, for glycine we detected a fragment produced for the loss of 82.0495, and for aspartic acid, another one after losing 118.1170. The last one could be the result of losing one TMS group and three CH_3 groups. One of the xanthines, theophylline, showed in its spectrum $[M+1TMS+H]^+$ and also $[M-1TMS+H]^+$ with low intensity in comparison with $[M+1TMS+H]^+$. [5-Hydroxyindole- 3-acetic acid+3TMS+H]$^+$ was the peak we found in CSF for the compound containing an indole moiety. Again 72.0389 for the loss of one TMS group, 144.0801 for the loss of 2TMS, and 118.1171 for the loss of one TMS group and three CH3 groups were observed. The amine nortriptyline hydrochloride showed up as $[M+H]^+$ without undergoing any fragmentation.

As commented before, Table 4 includes only a small fraction of compounds detected in CSF. We have detected more than 300 distinct features even using very strict peak finding criteria. This fact in combination with the here presented analytical characteristics (LODs, repeatability, and reproducibility) demonstrates the potential of GC/APCI-TOF MS for metabolic profiling. In other words, this analytical procedure might indeed be a valuable addition to the “metabolomics toolbox”.

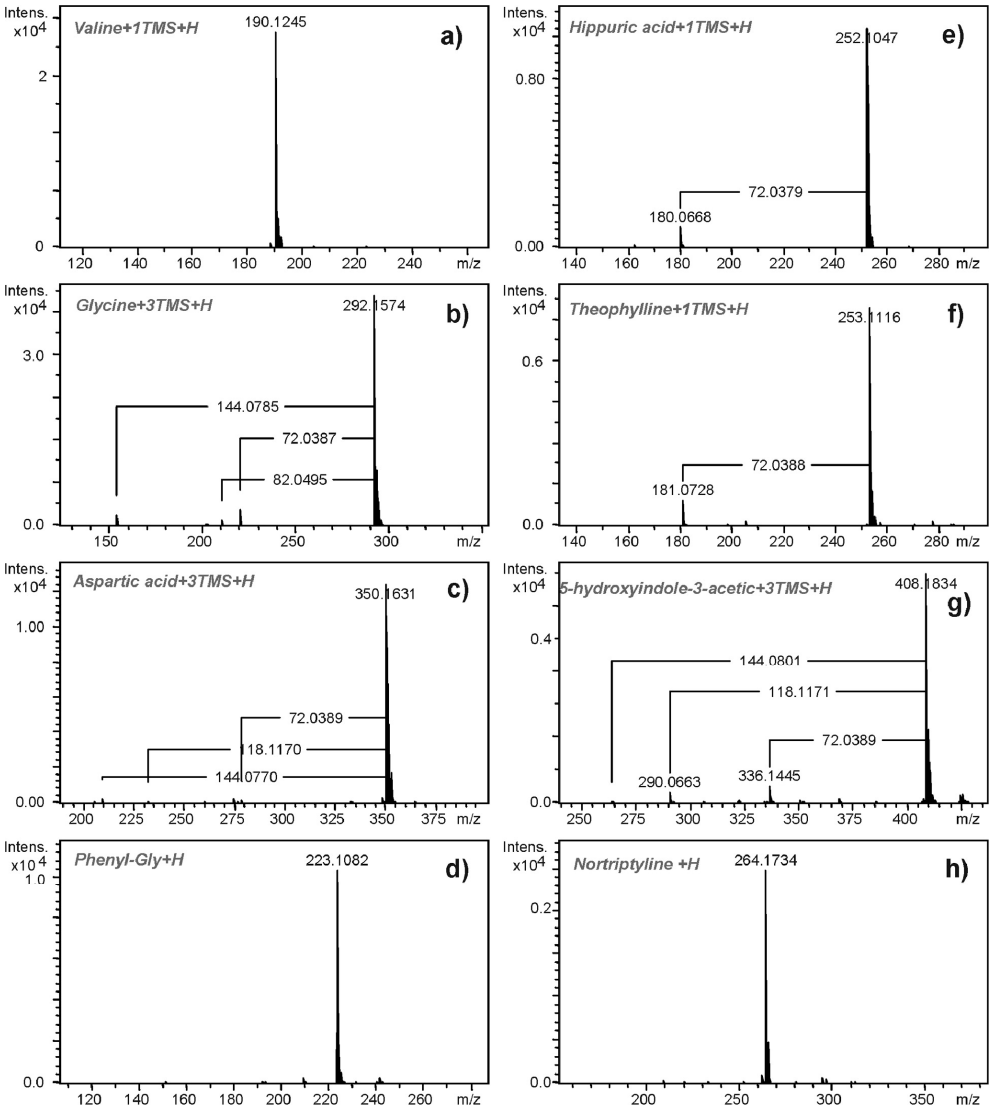

Figure 3. Typical APCI MS spectra of silylated compounds from different chemical families: amino acids (a–c), dipeptide (d), organic acid (e), xanthine (f), indole (g), and amine (h).

CONCLUSIONS

EI and CI are the ionization techniques conventionally used in GC-MS, both operating under vacuum condition. EI mass spectra are mainly characterized by numerous fragments produced during the high energy ionization process, while the CI mass spectra exhibit both the protonated molecules and intense fragment ions. Commercial and in-house database

mass spectral libraries can then be used to identify the separated compounds or at least give structural clues to support the identification process. Here, we present an alternative to the classical GC-MS methods, namely, gas phase APCI as interface in combination with orthogonal TOF-MS. A very sensitive and accurate GC/APCI-TOF MS method was developed for the automated analysis of metabolites in biological samples. At present, the analytical evaluation of the method was made by using amino acids, organic acids, alcohols, xanthines, indoles, dipeptides, compounds with imidazole groups, amines, and analytes with hydroxyl and amine groups, demonstrating that 25 analytes of the 31 present in our mixture can be reliably determined. Excellent repeatability was obtained, with relative standard deviations (RSDs) of peak areas between 0.7% and 2.1% in the intraday study, and between 3.8% and 6.4% in the interday study.

Analysis of CSF has demonstrated a rich chromatographic pattern consisting of hundreds of features. The high quality of the spectra creates an opportunity to make structural assignments of metabolites based on mass position and isotopic distribution. However, the use of more advanced mass analyzers such as hybrid quadrupole TOF will be beneficial to resolve more difficult cases and support identification by fragmentation data. In summary, GC/APCI-TOF MS is an analytical procedure, which combines the best of chromatography with one of the most robust MS interfaces, and as such, it has a potential to become one of the standard methods in metabolic profiling.

REFERENCES

1. James,A.T., and Martin,A.J. 1952. Gas-liquid partition chromatography: the separation and micro-estimation of ammonia and the methylamines. Biochem. J. 52:238-242.

2. Pauling,L., Robinson,A.B., Teranishi,R., and Cary,P. 1971. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. Proc. Natl. Acad. Sci. U. S. A 68:2374-2376.

3. Niwa,T., Asada,H., Maeda,K., Yamada,K., Ohki,T., and Saito,A. 1986. Profiling of organic acids and polyols in nerves of uraemic and non-uraemic patients. J. Chromatogr. 377:15-22.

4. Jiye A, Trygg,J., Gullberg,J., Johansson,A.I., Jonsson,P., Antti,H., Marklund,S.L., and Moritz,T. 2005. Extraction and GC/MS analysis of the human blood plasma metabolome. Anal Chem 77:8086-8094.

5. Koek,M.M., Muilwijk,B., van der Werf,M.J., and Hankemeier,T. 2006. Microbial metabolomics with gas chromatography/mass spectrometry. Anal Chem 78:1272-1281.

6. Kind,T., Tolstikov,V., Fiehn,O., and Weiss,R.H. 2007. A comprehensive urinary metabolomic approach for identifying kidney cancerr. Anal Biochem. 363:185-195.

7. Pohjanen,E., Thysell,E., Jonsson,P., Eklund,C., Silfver,A., Carlsson,I.B., Lundgren,K., Moritz,T., Svensson,M.B., and Antti,H. 2007. A multivariate screening strategy for investigating metabolic effects of strenuous physical exercise in human serum. J. Proteome Res. 6:2113-2120.

8. Fiehn,O. 2008. Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. Trends Analyt. Chem 27:261-269.

9. Begley,P., Francis-McIntyre,S., Dunn,W.B., Broadhurst,D.I., Halsall,A., Tseng,A., Knowles,J., Goodacre,R., and Kell,D.B. 2009. Development and Performance of a Gas Chromatography-Time-of-Flight Mass Spectrometry Analysis for Large-Scale Nontargeted Metabolomic Studies of Human Serum. Anal Chem.

10. van Deurse,M.M., Beens,J., Janssen,H.G., Leclercq,P.A., and Cramers,C.A. 2000. Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. J. Chromatogr. A 878:205-213.

11. Ojanpera,S., Pelander,A., Pelzing,M., Krebs,I., Vuori,E., and Ojanpera,I. 2006. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 20:1161-1167.

12. Sparkman,O.D., Penton,Z., and Kitson,F.G. 2011. Gas Chromatography and Mass Spectrometry: A Practical Guide. Elsevier Science & Technology.

13. Horning,E.C., Horning,M.G., Carroll,D.I., Dzidic,I., and STILLWEL.RN 1973. New Picogram Detection System Based on A Mass-Spectrometer with An External Ionization Source at Atmospheric-Pressure. Analytical Chemistry 45:936-943.

14. Horning,E.C., Carroll,D.I., Dzidic,I., Haegele,K.D., Lin,S.N., Oertli,C.U., and Stillwell,R.N. 1977. Development and Use of Analytical Systems Based on Mass-Spectrometry. Clinical Chemistry 23:13-21.

15. McEwen,C.N., and McKay,R.G. 2005. A combination atmospheric pressure LC/MS:GC/MS ion source: advantages of dual AP-LC/MS:GC/MS instrumentation. J. Am. Soc. Mass Spectrom. 16:1730-1738.

16. McEwen,C.N. 2007. GC/MS on an LC/MS instrument using atmospheric pressure photoionization. International Journal of Mass Spectrometry 259:57-64.

17. Mitchum,R.K., Korfmacher,W.A., Moler,G.F., and Stalling,D.L. 1982. Capillary Gas-Chromatography Atmospheric-Pressure Negative Chemical Ionization Mass-Spectrometry of the 22 Isomeric Tetrachlorodibenzo-P-Dioxins. Analytical Chemistry 54:719-722.

18. Korfmacher,W.A., Rushing,L.G., Engelbach,R.J., Freeman,J.P., Djuric,Z., Fifer,E.K., and Beland,F.A. 1987. Analysis of 3 Aminonitropyrene Isomers Via Fused-Silica Gas-Chromatography Combined with Negative-Ion Atmospheric-Pressure Ionization Mass-Spectrometry. Journal of High Resolution Chromatography & Chromatography Communications 10:43-45.

19. Kinouchi,T., Miranda,A.T.L., Rushing,L.G., Beland,F.A., and Korfmacher,W.A. 1990. Detection of 2-Aminofluorene at Femtogram Levels Via High-Resolution Gas-Chromatography Combined with Negative-Ion Atmospheric-Pressure Ionization Mass-Spectrometry. Hrc-Journal of High Resolution Chromatography 13:281-284.

20. Revelsky,I.A., Yashin,Y.S., Sobolevsky,T.G., Revelsky,A.I., Miller,B., and Oriedo,V. 2003. Electron ionization and atmospheric pressure photochemical ionization in gas chromatography-mass spectrometry analysis of amino acids. Eur. J. Mass Spectrom. (Chichester, Eng) 9:497-507.

21. Schiewek,R., Schellentrager,M., Monnikes,R., Lorenz,M., Giese,R., Brockmann,K.J., Gab,S., Benter,T., and Schmitz,O.J. 2007. Ultrasensitive determination of polycyclic aromatic compounds with atmospheric-pressure laser ionization as an interface for GC/MS. Anal Chem 79:4135-4140.

22. Wu,C., Siems,W.F., and Hill,H.H., Jr. 2000. Secondary electrospray ionization ion mobility spectrometry/mass spectrometry of illicit drugs. Anal Chem 72:396-403.

23. Brenner,N., Haapala,M., Vuorensola,K., and Kostiainen,R. 2008. Simple coupling of gas

chromatography to electrospray ionization mass spectrometry. Anal Chem 80:8334-8339.

24. Schiewek,R., Lorenz,M., Giese,R., Brockmann,K., Benter,T., Gab,S., and Schmitz,O.J. 2008. Development of a multipurpose ion source for LC-MS and GC-API MS. Anal Bioanal Chem 392:87-96.

25. Tong,H., Bell,D., Tabei,K., and Siegel,M.M. 1999. Automated data massaging, interpretation, and E-mailing modules for high throughput open access mass spectrometry. Journal of the American Society for Mass Spectrometry 10:1174-1187.

26. Keller,B.O., Sui,J., Young,A.B., and Whittal,R.M. 2008. Interferences and contaminants encountered in modern mass spectrometry. Anal Chim. Acta 627:71-81.

27. Currie,L.A. 1995. Nomenclature in Evaluation of Analytical Methods Including Detection and Quantification Capabilities (Iupac Recommendations 1995). Pure and Applied Chemistry 67:1699-1723.

28. Kaspar,H., Dettmer,K., Gronwald,W., and Oefner,P.J. 2008. Automated GC-MS analysis of free amino acids in biological fluids. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 870:222-232.

29. Namera,A., Yashiki,M., Nishida,M., and Kojima,T. 2002. Direct extract derivatization for determination of amino acids in human urine by gas chromatography and mass spectrometry. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 776:49-55.

30. Halket,J.M., Waterman,D., Przyborowska,A.M., Patel,R.K., Fraser,P.D., and Bramley,P.M. 2005. Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. J. Exp. Bot. 56:219-243.

31. Welch,M.J., Cohen,A., Hertz,H.S., Ng,K.J., Schaffer,R., Van der,L.P., and White,E. 1986. Determination of serum creatinine by isotope dilution mass spectrometry as a candidate definitive method. Anal Chem 58:1681-1685.

32. MacNeil,L., Hill,L., MacDonald,D., Keefe,L., Cormier,J.F., Burke,D.G., and Smith-Palmer,T. 2005. Analysis of creatine, creatinine, creatine-d3 and creatinine-d3 in urine, plasma, and red blood cells by HPLC and GC-MS to follow the fate of ingested creatine-d3. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 827:210-215.

33. Fiehn,O., and Kind,T. 2007. Metabolite profiling in blood plasma. Methods Mol. Biol. 358:3-17.

34. Bristow,A.W., and Webb,K.S. 2003. Intercomparison study on accurate mass measurement of small molecules in mass spectrometry. J. Am. Soc. Mass Spectrom. 14:1086-1098.

35. Rodier,C., Sternberg,R., Raulin,F., and Vidal-Madjar,C. 2001. Chemical derivatization of amino acids for in situ analysis of Martian samples by gas chromatography. Journal of Chromatography A 915:199-207.

36. Mohabbat,T., and Drew,B. 2008. Simultaneous determination of 33 amino acids and dipeptides in spent cell culture media by gas chromatography-flame ionization detection following liquid and solid phase extraction. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 862:86-92.

37. Shen,X., Deng,C., Wang,B., and Dong,L. 2006. Quantification of trimethylsilyl derivatives of amino acid disease biomarkers in neonatal blood samples by gas chromatography-mass spectrometry. Anal Bioanal Chem 384:931-938.

SUPPLEMENTARY MATERIALS

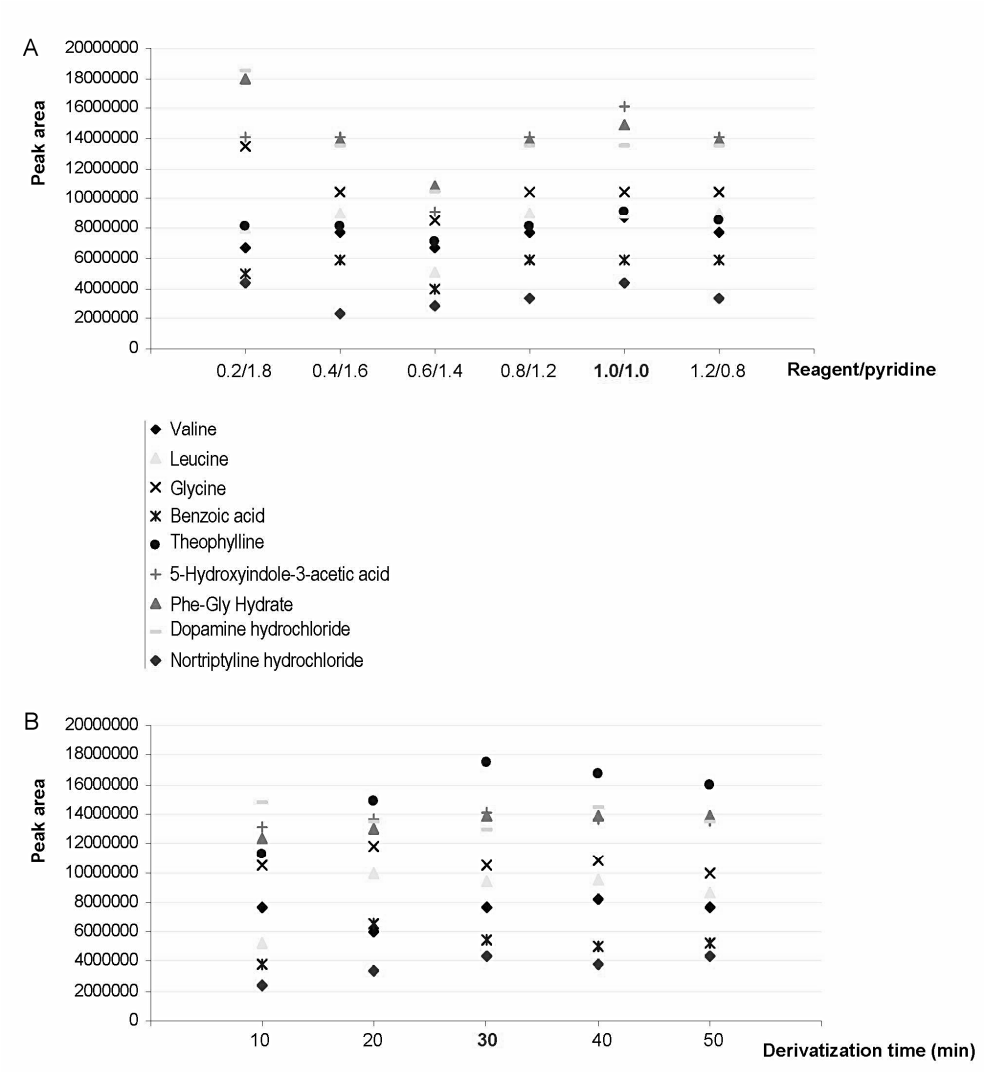

Figure S1. Effect of derivatization conditions on peak area of several metabolites included in the standard mixture. A) volume ratio of derivatization reagent (µL) and pyridine (µL); B) derivatization time.

Table S1. Some relevant information about other GC-MS methods previously published for the determination in biological samples of the same compounds as those under study in the current paper or other similar analytes.

Compounds under study	Derivatization reaction	Instrumentation used	Detection limit	Sample analyzed	Other comments	Reference
Organic acids, amino acids, sugars, polyols, purines, pyrimidines and other compounds are simultaneously analyzed and quantified	BSTFA with 1% TMCS	A Hewlett-Packard GC–MSD (HP6890/MSD5973) and a Shimadzu QP5000 GC–MS were used for GC–MS measurement	-	Human urine	Pilot study for screening of 22 target diseases in newborns conducted in Japan is described. The diagnostic procedure consists of the use of urine or filter paper urine, preincubation of urine with urease, stable isotope dilution, and GCMS.	J. Chromatogr. B 2001, 758, 3-25
Buprenorphine and norbuprenorphine	Pentafuoropropionic anhydride	HP 5890 GC with a 5971A mass selective detector	LOQ buprenorphine=0.05 ng/ml; for Norbuprenorphine=0.1 ng/ml	Human plasma	The method could be used to explore the pharmacokinetic/pharmacodynamic relationship of buprenorphine and norbuprenorphine	European J. Pharmaceutics and Biopharmaceutics 2001, 51, 147-151
Amino acids	Ethyl chloroformate	GC-MS Hewlett-Packard 5890 series II GC and 5971 A MS	0.5-1.0 µg/ml	Human urine	Several derivatisation reagents used. Threonine, serine, asparagines, glutamine, arginine not derivatized by using any tested reagent.	J. Chromatogr. B 2002, 776, 49-55
Organic acids and glycine conjugates	BSTFA with 1% TMCS	Hewlett-Packard HP5890A GC coupled to HP5970B mass-selective detector	0.4-200 nmol/l	Amniotic fluid	12 metabolites simultaneously quantified	J. Inherit. Metab. Dis. 2004, 27, 567-579
Global approach (hundreds of molecular features detected)	BSTFA with 1% TMCS	Agilent 6890 GC with FID and 5973MSD MS	-	Human urine samples	GC as complementary tool to NMR	Rapid. Commun. Mass Spectrom. 2006, 20, 2271-2280
Antidepressants and their active metabolites	Heptafluorobutyrylation	HP 6890 GC system with HP 5973 mass-selective detector	5.0-12.5 ng/ml (EI and positive CI)	Plasma	Validation of GC coupled to MS by CI and EI sources. CI offered advantages in selectivity and sensitivity	J. Chromatogr. A 2007, 1176, 236-245
Amino acids, organic acids, sugars…goblal approach	1st step: methoximation. 2nd step: MSTFA with 1% TMCS	Finnigan GC (Thermo Finnigan, USA) coupled with mass spectrometry (TRACE DSQ)	26 selected compounds could be detected at S/N ≥ 3 when the urine dilution was 0.02 (v/v, urine/urine +water): it could be defined as LOD	Rat urine	Forty-nine endogenous metabolites were separated and identified in GCMS chromatogram, of which 26 identified compounds were selected for quantitative analysis	J. Chromatogr. B 2007, 854, 20-25

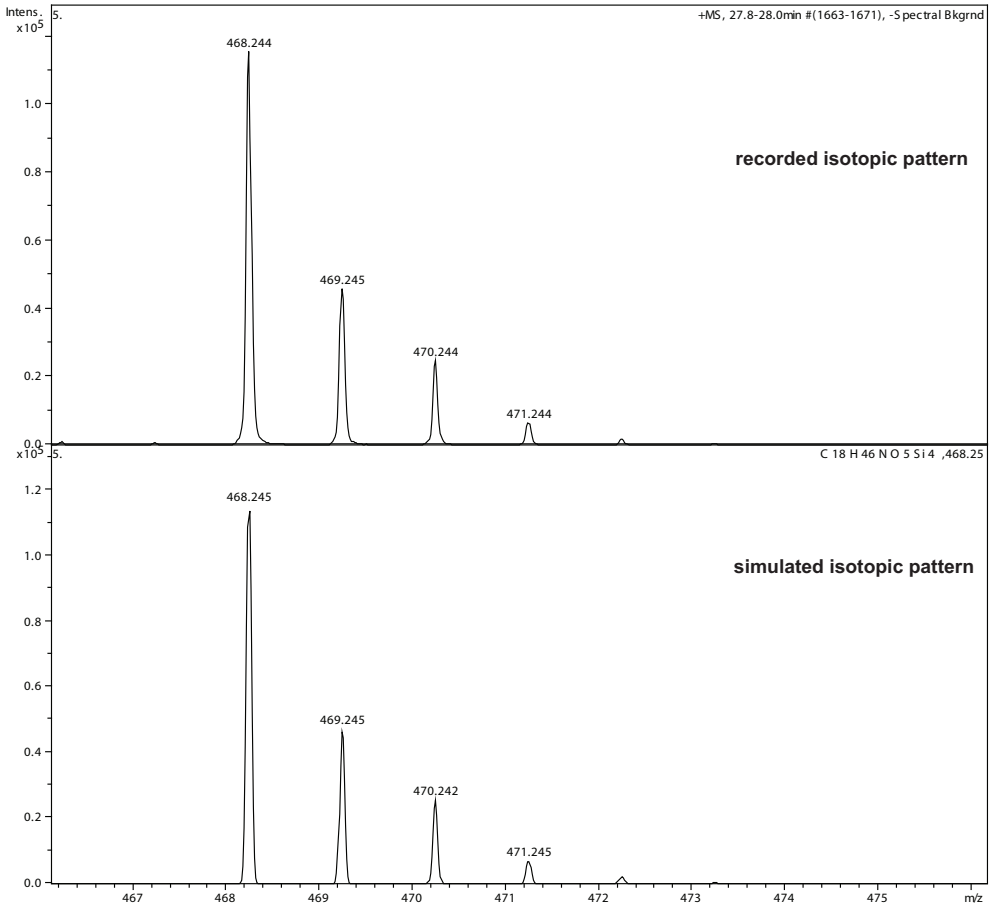

Figure S2. Assignment of N-acetyl-aspartate using accurate mass and isotopic distribution information.

Chapter 2

Alignment of capillary electrophoresis–mass spectrometry datasets using accurate mass information

Nevedomskaya E., Derks R., Deelder A.M., Mayboroda O.A., Palmblad M.

Analytical and Bioanalytical Chemistry **2009**, 395, 2527–2533

ABSTRACT

Capillary electrophoresis–mass spectrometry (CE–MS) is a powerful technique for the analysis of small soluble compounds in biological fluids. A major drawback of CE is the poor migration time reproducibility, which makes it difficult to combine data from different experiments and correctly assign compounds. A number of alignment algorithms have been developed but not all of them can cope with large and irregular time shifts between CE–MS runs. Here we present a genetic algorithm designed for alignment of CE–MS data using accurate mass information. The utility of the algorithm was demonstrated on real data, and the results were compared with one of the existing packages. The new algorithm showed a significant reduction of elution time variation in the aligned datasets. The importance of mass accuracy for the performance of the algorithm was also demonstrated by comparing alignments of datasets from a standard time-of-flight (TOF) instrument with those from the new ultrahigh resolution TOF maXis (Bruker Daltonics).

INTRODUCTION

Capillary electrophoresis (CE) is an ideal technique for separation of small soluble polar compounds that are present in biological fluids.(1) There are also other advantages of CE for analysis of biological fluids, such as relatively short separation times with good resolution and low sample consumption.(2) CE is often criticized for its low loading capacity. However, pre-concentration techniques such as pH-mediated stacking (3) can overcome this drawback. If a mass spectrometer is used as a detector (CE–mass spectrometry (MS)), additional information on mass and isotopic distribution (4) is provided, which enables compounds and potential biomarkers to be identified. For comparison of multiple samples, elution or migration time precision is also very important. This is a serious concern for CE which, especially when bare-fused silica capillaries are used, lacks reproducibility of migration time.(5) Low reproducibility of migration time affects not only identification of compounds and their synchronization between samples but also statistical analysis. Misalignment introduces variation in the data that will noticeably affect results of multivariate statistics (6) and for studies involving numerous samples, as typically encountered in clinical research, manually assisted alignment of CE–MS datasets is not feasible.

The data produced by CE–MS is three-dimensional: intensity as a function of time and mass-to-charge ratio. There are two main strategies for alignment of this type of data: (1) to group features together in matrices that can be further statistically analyzed or (2) to transform all time axes to a common axis with further analysis of aligned signals. (7) The former works well for protein and peptide data, as some peaks can be identified and used as internal standards for quantification and correction, but is problematic in case of metabolomics.(8) In addition, for complex and overlapped electropherograms, peak assignment is less reliable.(9) For the second strategy, there are already many algorithms and software packages available. However, most of these programs have been developed for liquid chromatography- mass spectrometry (LC-MS) and cannot deal with the large and irregular time shifts typically encountered in CE–MS. Furthermore, a majority of these programs use only chromatographic information, aligning base peak or total ion chromatograms using different time warping procedures. As has been mentioned by Daszykowski *et al.*(10), the next step in the development of alignment methods should take advantage of mass as well as chromatographic information.

Another issue for data processing tools is to have them platform independent, to be able to share results within the scientific community. Commercial software often works with data from certain instruments using their specific formats, making them vendor dependent. Free software is working with data formats that can be generated by a large number of tools

and programs for format conversion, such as mzXML.(11) Currently, mzXML is the preferred format to generate aligned CE–MS data as many programs exist which can read this format for further analysis. In this paper we describe the adaptation and application of an algorithm originally developed for LC–MS and LC–MS/MS (12) for alignment of CE–MS datasets—msalign2. Previously published algorithm was developed for alignment of LC–MS and LC–MS/MS data generated by two different mass analyzers (for example, high resolution data of FTICR and low resolution data of ion-trap). The latter was used for confident identification of peptides, masses of which were then matched to masses in LC–MS dataset. The new msalign2 is an alignment method for hyphenated MS applications. It is not limited to only capillary electrophoresis but can be as well used for any hyphenated technique, for instance LC–MS. CE–MS was chosen as it represents the most challenging task for alignment, and there is a demand for this type of software.

The algorithm and ancillary software is implemented in C and R and is available as open source (http://www.ms-utils.org/msalign2/). The algorithm has been shown to work on real CE–MS datasets of urine that are representative of data from a biomedical study. The results have been compared with another alignment tool in the open-source package XCMS (13) in terms of efficiency and relevance of further statistical analysis, visually inspecting and comparing principal component analysis (PCA) results. Our algorithm showed reduced variance in the data and performed better for multivariate analysis.

THEORY

Two CE–MS datasets can be aligned by matching masses across samples and fitting a curve to these matches. The curve represents the relation between electropherograms.

The shifts in migration time are not linear.(14) Non-linearity can be introduced by changes in conductivity and electroosmotic flow or the sheath liquid flow driven by a mechanical pump. That is why the natural solution to alignment problem in CE–MS (and LC–MS) is a piece-wise function of time that can cope with these irregularities.

The problem of finding a function best fitting measured data is an optimization problem. Genetic algorithms are one class of methods for solving this problem. These algorithms were developed in the 1960s from earlier published computer simulations of evolution and artificial selection.(15) A genetic algorithm (GA) is able to find exact or approximate solutions for optimization problems.

GA operates on a population of possible solutions for a problem, called chromosomes. The starting population is created randomly and then goes through a number of generations being transformed by the operators of inheritance: mutation, selection, and recombination. Chromosomes are encoded in such a way that they are suitable for applying

these operators. A function for computing the quality of each chromosome is required. Using this fitness function best candidate solutions are selected in each generation and allowed to reproduce. In the end the global optimum or its close approximation is found.

For aligning CE–MS datasets, a candidate solution is set by breakpoints of the piece-wise function. The fitness function $F(s_i)$ for a candidate solution s_i is calculated as:

$$F(s_i) = \sum_{m=1}^{N} e^{-\frac{(y_m - y(x_m))^2}{2\sigma^2}} - kn_i \qquad (1)$$

where m are peaks out of all N matches with retention times x_m in the dataset to be aligned, y_m in the reference dataset; n_i is the number of breakpoints in chromosome s_i, k is a cost per breakpoint, and σ^2 is the residual variance between the datasets. The fitness function is the sum of likelihoods of observing peak m at a retention times x_m and y_m in the two datasets with the piece-wise function evaluating $y(x_m)$. The cost for a breakpoint is introduced so that the number of breakpoints in the alignment is not too large. The residual variance can be provided by the user based on the knowledge of the analytical system used, or, if absent, is estimated automatically by the software, as previously described.

In each generation, half of the chromosomes in the population with the lowest fitness are replaced by copies of half of the chromosomes with the highest fitness, applying three types of mutations: insertion (randomly adding a breakpoint anywhere in the alignment interval), deletion (removing a breakpoint), and shifting a breakpoint by a small random amount in both dimensions. After a single pass through the GA, the solution of the highest fitness was chosen as the alignment of the two datasets.

The genetic algorithm was run for 1,000 generations with a population size of 300 candidate solutions with maximum number of breakpoints of 12 and the cost for breakpoint set to 0.5.

MATERIALS AND METHODS

Chemicals. Methanol (MeOH) HPLC-grade (Biosolve B.V., Netherlands), ultrapure water (18.2 MΩ/cm), and formic acid (FA) (Fluka, Germany) were used for solvent preparation. NH_4OH was from Sigma-Aldrich, NaOH from J.T. Baker.

Urine samples. Urine was collected and pooled from two groups of mice wild-type 129 and Swiss mice and stored at −20 °C. Urine (4 μL) was mixed with 4 μL of MeOH, 11 μL of water, and 1 μL of BGE, centrifuged to eliminate any possible sediment remaining, and put into vials for injection into CE instrument.

Instrumentation. CE was performed on a PA 800 (Beckman Coulter, Fullerton, CA, USA) as described before.(16) Uncoated fused silica capillaries (BGB-Analytik, Germany)

of total length of 100 cm with 50 μm inner diameter were used for separation. MeOH (20%) with 2 M FA was used as background electrolyte. Sample injection was performed hydrodynamically with pH-mediated stacking: a small plug (50 mbar, 9 s) of 12.5% NH_4OH was injected before the sample plug (50 mbar, 90 s).

The second set of mouse urine has been measured with 0.1 M NaOH washing step between the runs.

MS was performed using two types of time-of-flight (TOF) mass spectrometers: the micrOTOF (Bruker Daltonics, Bremen, Germany) and the new ultrahigh resolution TOF (UHR-TOF), maXis from the same vendor. The acquisition and spraying parameters were optimized so that the total areas on both instruments were identical. Transfer parameters were optimized by direct infusion of an ESI tuning mix (Agilent Technologies, Waldbronn, Germany). Spectra were collected with a time resolution of 1 s. CE–MS coupling was realized by a co-axial sheath liquid interface (Agilent Technologies, Waldbronn, Germany) with methanol–water–formic acid (50:50:0.1, *v/v/v*) as sheath liquid. The following spray conditions were used: sheath liquid flow, 4 μL/min; dry gas temperature, 180 °C; nitrogen flow, 4 L/min; nebulizer pressure, 0.5 bar. Electrospray in positive ionization mode was achieved and ESI voltage was −4.5 kV.

Data analysis. Electropherograms were aligned using in-house-developed genetic algorithm running 1,000 generations and by XCMS (The Scripps Research Institute, La Jolla, USA). Data were normalized using non-parametric algorithm as described earlier.(17) The results of alignment were analyzed by PCA in SIMCA-P+ software (Umetrics, Umeå, Sweden). All calculations were performed on a standard office PC (Core 2 Quad, 2.4 GHz, 2 GB RAM). The alignment of 20 datasets took about 40 min and used up to 200 MB RAM. This time is tens of times less than the time needed for acquisition of the data, so it does not represent a bottleneck in the whole analysis pipeline.

RESULTS AND DISCUSSION

The algorithm works pair-wise, operating on two CE–MS datasets in the mzXML format, which makes it platform independent and suitable for alignment of data generated by any type of CE–MS instrumentation. To demonstrate the alignment performance of the algorithm, 20 electropherograms were aligned. Figure 1 represents the result of alignment.

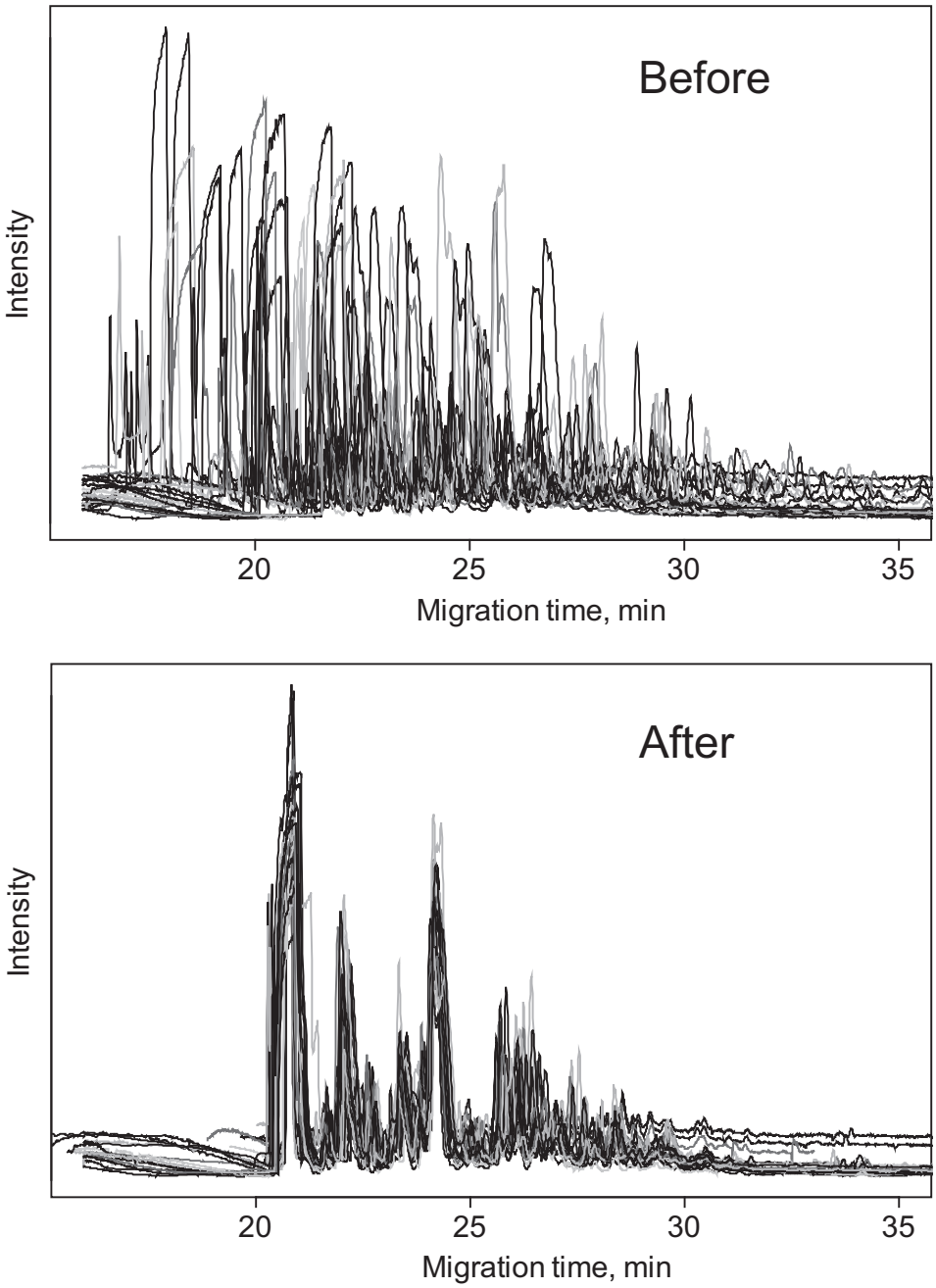

Figure 1. Total ion electropherograms of 20 CE–MS datasets before and after alignment. In the region where compounds migrate the alignment works for each single dataset.

A visual inspection already shows a significant improvement in peak positions. As previously mentioned, migration time shifts are not linear, so the evaluation of alignments should be done not on a single peak but better on several peaks in different parts of the electropherogram. The relative standard deviations (RSD) in migration time were calculated for three peaks at the beginning, middle, and end of the electropherogram and ranged from 5.6% to 6.5% before alignment. The RSD was higher for peaks, which are closer to the end of electropherogram. This happens because toward the end of the run adsorption on the walls of the capillary accumulates, and shifts in migration times increase.(18) After the alignment, the RSD significantly improved and varied from 0.12% to 0.99%. Only analytical window was used for alignment, excluding the time at the beginning of the run where the sodium clusters are migrating. It can be seen on the aligned chromatograms that at the beginning of the run, the variation in migration time is still present, whereas in the rest of the electropherogram, peaks are very well aligned.

To show how alignment improves further statistical analysis, a case study consisting of two groups of measurements was selected. Two-group scenario of data analysis is common in clinical research where there are typically groups of samples of patients and controls or patients at different time points or with different treatments. For our study we used two groups, each consisting of nine electropherograms of pooled mouse urine from two strains of mice (129 and Swiss). The samples from the second group were measured with a NaOH washing step between samples, which introduces a systematic difference between the groups and allows to show the robustness of the algorithm. The washing step is important in CE–MS. When a biological matrix is introduced, the capillary is contaminated, leading to large shifts in migration time, clogging, or even breakdown of the capillary. This is especially the case when bare-fused silica in a low pH system is used. There are two principal solutions to this problem. The first is changing the capillaries after a certain number of runs, introducing some additional variation, and then aligning all the datasets using one of the available algorithms. Alternatively, it is possible to regenerate the capillary with sodium hydroxide after each run. The second option is more time consuming as it requires washing steps with NaOH followed by water. However as we also show here, the washing strategy gives better analytical results with less variation in the data. The variation in the data without the washing step cannot be entirely eliminated even when advanced alignment techniques are used.

We compared the alignment of 18 electropherograms by our piece-wise alignment and by one of the existing methods — XCMS. XCMS was chosen because it is a very powerful and quite widely used package, for which our tool can be a useful complement. The

workflow of the alignment in XCMS is based on completely different principles and includes peak detection and matching prior to time correction.

The ubiquitous creatinine peak was used for visual inspection and control of the alignment efficiency by two programs. The extracted ion chromatograms of creatinine (m/z 114.12±0.05) are shown in Fig. 2. Both algorithms perform quite satisfactorily but it is apparent that XCMS leaves some of the peaks unaligned, whereas msalign2 successfully aligned all datasets.

This may be because the large number of changeable parameters in XCMS makes it difficult for the user to optimize the process and find the ideal settings for given sets of data. In contrast, msalign2 has a minimum number of parameters, most of which never have to be changed between one sample and the next. The free parameters are mass measurement error, background threshold, start and end scans for the time interval for alignment, and expected variance in elution time. The last is optional, and if the user does not supply it, the program will automatically estimate the variance. Mass measurement error is given in ppm and depends on the type of mass spectrometer being used and not the separation technique. The background threshold should be chosen such that the number of matched masses across samples is on the order of 1,000 to get reliable alignments with a reasonable computational time. Start and end scan numbers are used to align only the informative part of the electropherograms, disregarding intense signals that commonly appear at the beginning and end of each run, such as sodium clusters and compounds that have attached to the capillary wall and are released during the washing step. The web-application contains, besides the alignment algorithm, a tool to estimate the background threshold needed to get specified number of matched features for the alignment.

Another possibility why some misalignment appears in the case of XCMS is that the package was primarily designed for LC–MS and can give suboptimal results for CE–MS, which may have much larger shifts in time domain. Time shifts are crucial for the matching step performed in XCMS before the alignment. Nevertheless it does not decrease the applicability and usefulness of the XCMS package as it includes not only an alignment algorithm but also powerful peak picking that might for instance be used after performing alignment with msalign2.

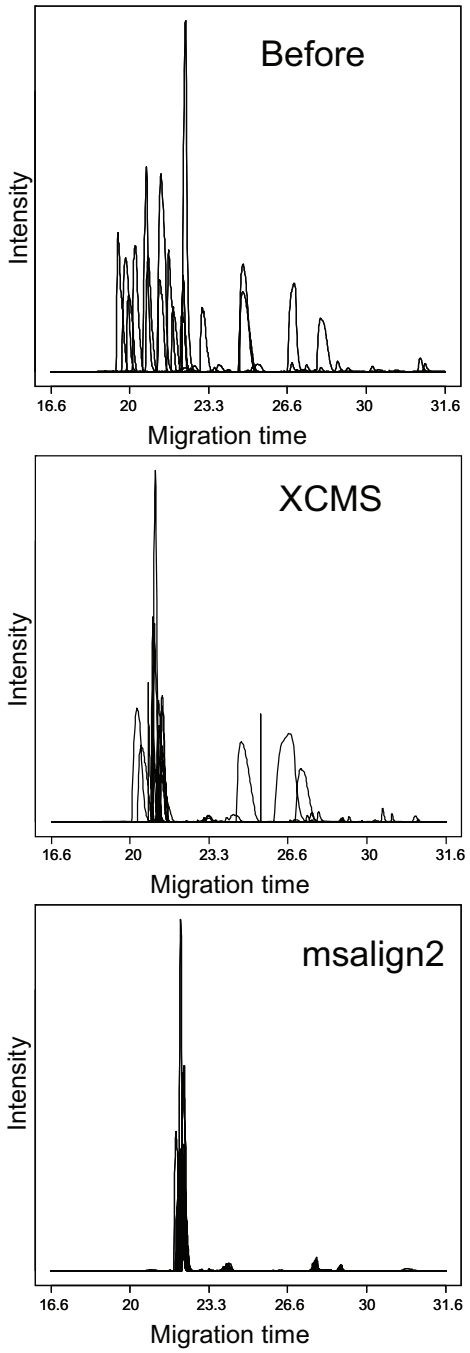

Figure 2. Extracted ion electropherograms of creatinine before alignment, aligned with XCMS, and aligned with msalign2.

The output of XCMS contains a table with detected features (chromatographic peaks) and their intensities throughout all samples. It can be directly used for analysis with any statistical package, for instance SIMCA-P+.(19) The output of our algorithm is an mzXML file with corrected retention times. This makes it easy to explore how well the alignment works. To apply further statistics, one can either perform binning or peak picking from these mzXML files. Here we used the peak picking from the XCMS package, so that the only difference between the two methods is the alignment step. The resulting tables from both alignments were normalized as described above and imported to the SIMCA-P+ software package for multivariate analysis.

PCA is a usual first step in analyzing multivariate data. It shows the correlation structure present in the data and the directions of the largest variance. Tables produced by peak picking step were normalized and used as input for PCA. The scores plots are shown in Fig. 3.

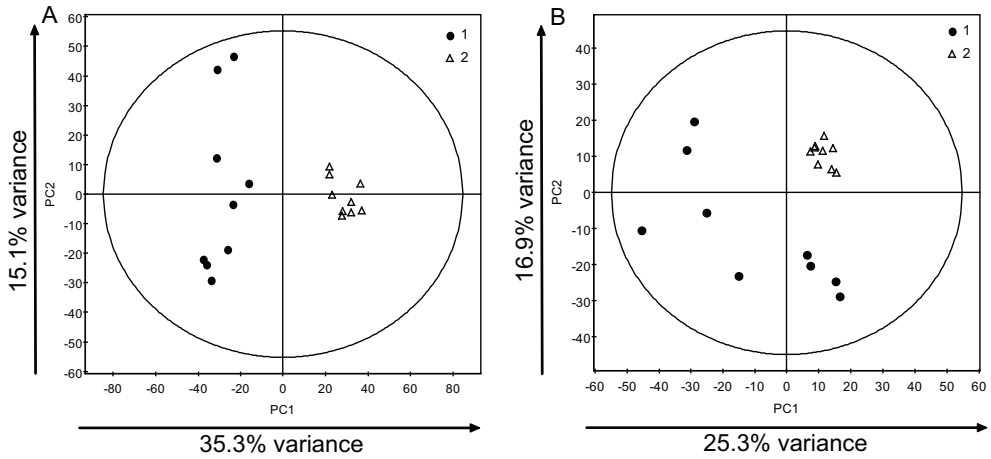

Figure 3. PCA scores plots: A data from peak picking by XCMS after genetic algorithm alignment and B data obtained from XCMS alignment. The variance explained by the components is indicated on arrows along the axes. Group 1 (without washing step) is indicated by black dots; group 2 (with washing step) is indicated by triangles.

The overall picture is the same in both cases, with the two sample groups nicely separated. The datasets acquired with the washing step between each sample show significantly smaller variance. However, more variation (>50%) is explained by first two principal components with the msalign2 than with the XCMS alignment (<50%). This is due to the additional variation introduced by misalignment in XCMS. It is important to reduce systematic variability in the data as far as possible, and not introduce additional

variability by alignment or normalization procedures that can obscure the chemical species correlating with or being responsible for the actual phenomena under study.

As mentioned above, an important feature of our algorithm is that it uses accurate mass information. In theory, the better the mass accuracy, the easier the alignment task. To test this hypothesis we generated electropherograms of the same pooled urine using two different mass spectrometers: a standard orthogonal TOF (Bruker micrOTOF) and the recently released ultrahigh resolution TOF instrument (Bruker maXis). Two main differences between these instruments are the resolving power (40,000 vs 20,000 at *m/z* 600) and mass measurement precision (0.8 vs 3 ppm).

Mass electropherograms generated using these machines were indeed found to differ in resolution and mass accuracy. msalign2 performed well on both pairs of data, but as can be seen in Fig. 4, there are significantly fewer mismatched features between the maXis datasets (7%) than between the micrOTOF datasets (more than 40%).

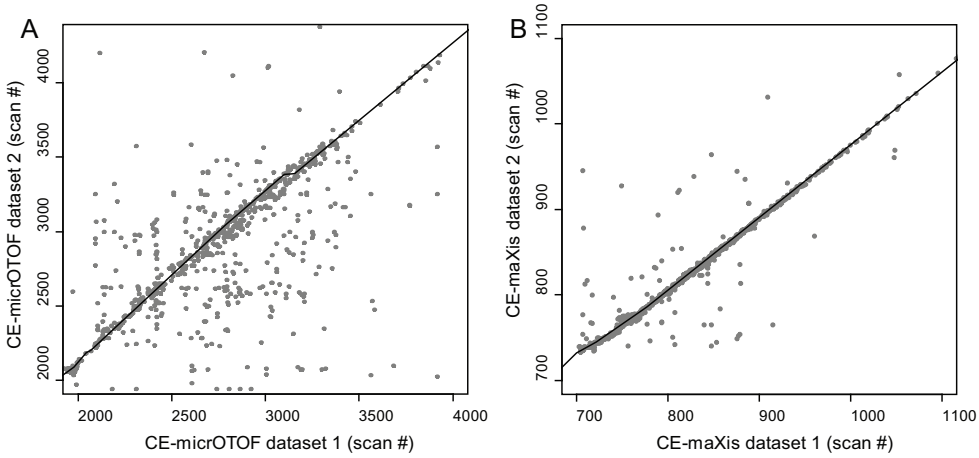

Figure 4. Alignments of CE–MS datasets from two types of instruments—micrOTOF (a) and maXis (b). The matched features are presented as gray dots; black line represents the trend line found by genetic algorithm.

These results demonstrate that more compounds can be correctly identified based on accurate mass with the new UHR-TOF instrument compared to the standard TOF. The alignment problem is easier to solve with better mass accuracy, but the genetic algorithm is sufficiently robust to find the correct global alignment also between the micrOTOF datasets.

The genetic algorithm and supporting software is implemented in C, R and is available as open source on http://www.ms-utils.org/msalign2.

Here we presented the application of our algorithm to CE–MS alignment, which is the worst case among separation techniques in terms of time reproducibility. As our method does not depend on chromatographic parameters and quality, it can as well be used for alignment of LC–MS and GC–MS data.

The analysis of large amounts of data generated by metabolomic, proteomic, peptidomic, or other types of "omic" experiments includes many steps, most of which need sophisticated algorithms and computational implementation. Small mistakes and imperfections on each of these steps can lead to incorrect data interpretation and misleading results. The consequences may be even more dramatic in the field of system biology when the data from different analytical platforms and from different levels of biological organization have to be combined. Alignment of chromato- (electrophero-) grams is just one of the steps of data analysis but is an important one and should be the subject of careful examination and optimization, as was performed in this study.

CONCLUSIONS

Here we present a platform-independent, open-source algorithm for alignment of complex CE–MS datasets. In contrast to other available alignment algorithms it efficiently uses mass information. Performance of MS instrumentation, mass accuracy, and resolving power positively affect alignment results. However, we have clearly shown that the algorithm is robust enough and performs even with relatively "low cost" MS instrumentation. It is shown also that the variation should be reduced not only by means of data processing but also by selecting proper experimental conditions.

As a tool for alignment of CE–MS data, our algorithm outperforms such packages as XCMS resulting in reduced variation that appears in multivariate analysis. On the other hand, alignment is only a step in the data processing pipeline and as such our algorithm is fully complimentary to XCMS.

Although in this paper we have focused on CE–MS as this approach represents one of the most challenging tasks for alignment, the algorithm can obviously as well be used for alignment of LC–MS and GC–MS datasets.

REFERENCES

1. Monton,M.R.N., and Soga,T. 2007. Metabolome analysis by capillary electrophoresis-mass spectrometry. Journal of Chromatography A 1168:237-246.

2. Song,E.J., Babar,S.M., Oh,E., Hasan,M.N., Hong,H.M., and Yoo,Y.S. 2008. CE at the omics level: towards systems biology--an update. Electrophoresis 29:129-142.

3. Neususs,C., Pelzing,M., and Macht,M. 2002. A robust approach for the analysis of peptides in the low femtomole range by capillary electrophoresis-tandem mass spectrometry. Electrophoresis 23:3149-3159.

4. Ojanpera,S., Pelander,A., Pelzing,M., Krebs,I., Vuori,E., and Ojanpera,I. 2006. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. 20:1161-1167.

5. Garcia-Perez,I., Vallejo,M., Garcia,A., Legido-Quigley,C., and Barbas,C. 2008. Metabolic fingerprinting with capillary electrophoresis. J. Chromatogr. A 1204:130-139.

6. van Nederkassel,A.M., Xu,C.J., Lancelin,P., Sarraf,M., Mackenzie,D.A., Walton,N.J., Bensaid,F., Lees,M., Martin,G.J., Desmurs,J.R. *et al* 2006. Chemometric treatment of vanillin fingerprint chromatograms. Effect of different signal alignments on principal component analysis plots. J. Chromatogr. A 1120:291-298.

7. Katajamaa,M., and Oresic,M. 2007. Data processing for mass spectrometry-based metabolomics. J. Chromatogr. A 1158:318-328.

8. Aberg,K.M., Alm,E., and Torgrip,R.J. 2009. The correspondence problem for metabonomics datasets. Anal Bioanal Chem.

9. Johnson,K.J., Wright,B.W., Jarman,K.H., and Synovec,R.E. 2003. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. J. Chromatogr. A 996:141-155.

10. Daszykowski,M., Danielsson,R., and Walczak,B. 2008. No-alignment-strategies for exploring a set of two-way data tables obtained from capillary electrophoresis-mass spectrometry. J. Chromatogr. A 1192:157-165.

11. Pedrioli,P.G., Eng,J.K., Hubley,R., Vogelzang,M., Deutsch,E.W., Raught,B., Pratt,B., Nilsson,E., Angeletti,R.H., Apweiler,R. *et al* 2004. A common open representation of mass spectrometry data and its application to proteomics research. Nat. Biotechnol. 22:1459-1466.

12. Palmblad,M., Mills,D.J., Bindschedler,L.V., and Cramer,R. 2007. Chromatographic alignment of LC-MS and LC-MS/MS datasets by genetic algorithm feature extraction. J. Am. Soc. Mass Spectrom. 18:1835-1843.

13. Smith,C.A., Want,E.J., O'Maille,G., Abagyan,R., and Siuzdak,G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78:779-787.

14. America,A.H., Cordewener,J.H., van Geffen,M.H., Lommen,A., Vissers,J.P., Bino,R.J., and Hall,R.D. 2006. Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional LC-MS. Proteomics 6:641-653.

15. Fraser,A. 1957. Simulation of genetic systems by automatic digital computers. I. Introduction. Australian J. Biol. Sci. 10:484-491.

16. Mayboroda,O.A., Neususs,C., Pelzing,M., Zurek,G., Derks,R., Meulenbelt,I., Kloppenburg,M., Slagboom,E.P., and Deelder,A.M. 2007. Amino acid profiling in urine by capillary zone electrophoresis - mass spectrometry. J. Chromatogr. A 1159:149-153.

17. Sidorov I.A., Hosack D.A., Gee D., Yang J., Cam M.C., Lempicki R.A., and Dimitrov D.S. 2002. Oligonucleotide microarray data distribution and normalization. Information Sciences 146:67-73.

18. Stutz,H. 2009. Protein attachment onto silica surfaces - a survey of molecular fundamentals, resulting effects and novel preventive strategies in CE. Electrophoresis 30:2032-2061.

19. Nordstrom,A., O'Maille,G., Qin,C., and Siuzdak,G. 2006. Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum. Anal Chem 78:3289-3295.

Part II

Application to animal studies

Chapter 3

CE-MS for metabolic profiling of volume-limited urine samples: application to accelerated aging TTD mice

Nevedomskaya E., Ramautar R., Derks R., Westbroek I., Zondag G., van der Pluijm I., Deelder A.M., Mayboroda O.A.

Journal of Proteome Research **2010**, *9*, 4869–4874

ABSTRACT

Metabolic profiling of biological samples is increasingly used to obtain more insight into the pathophysiology of diseases. For translational studies, biological samples from animal models are explored; however, the volume of these samples can be a limiting factor for metabolic profiling studies. For instance, only a few microliters of urine is often available from small animals like mice. Hence, there is a need for a tailor-made analytical method for metabolic profiling of volume-limited samples. In the present study, the feasibility of capillary electrophoresis time-of-flight mass spectrometry (CE-ToF-MS) for metabolic profiling of urine from mice is evaluated. Special attention is paid to the analytical workflow; that is, such aspects as sample preparation, analysis, and data treatment are discussed from the metabolomics viewpoint. We show that metabolites belonging to several chemical families can be analyzed in mouse urine with the CE-ToF-MS method using minimal sample pretreatment and an in-capillary pre-concentration procedure. This exemplifies the advantages of CE-ToF-MS for metabolic profiling of volume-limited samples as loss of material is minimized. The feasibility of the CE-ToF-MS-based workflow for metabolic profiling is illustrated by the analysis of urine samples from wild-type as well as from TTD mutant mice, which are a model for the accelerated aging, with osteoporosis being one of the main hallmarks.

INTRODUCTION

Metabolomics is a rapidly developing field in the "postgenomic" research that focuses on the global profiling of small endogenous molecules present in body fluids.(1) As end-points of biochemical processes, endogenous metabolites are directly associated with the phenotype of the organism.(2) Consequently, an overview of the metabolic composition of body fluids and urine in particular can be helpful for the phenotypic characterization of genetically modified animals and eventually for the understanding of the link between phenotypic trait and genetic background.

At present, high-resolution ^{1}H NMR spectroscopy is one of the key technologies for body fluid investigations as it is capable of producing fast and highly reproducible metabolic profiles in body fluids without the need for the preselection of analytical parameters or sample derivatization procedures. For volume-limited samples, a miniaturized NMR probe coil can be implemented for the analysis of volumes as low as a few microliters.(3) However, a major limitation of NMR still is the relatively poor concentration sensitivity compared to MS-based techniques. For GC-MS, 10-50 μL of sample is often a minimum volume requirement for sample pretreatment steps (such as for example, derivatization), and in LC-MS, injection volumes of 5-10 μL are commonly used for metabolic profiling studies.

With regard to miniaturization, capillary electrophoresis-mass spectrometry (CE-MS) (4-6) is a well-suited method, as recently illustrated by single-cell and even subcellular level analyses.(7) Moreover, it is possible to simultaneously concentrate and separate analytes in a biological sample without any sample pretreatment, which is very advantageous for volume-limited samples.(8)

In the present study, we describe an analytical workflow based on CE-MS for metabolic profiling of urine samples from mice. As a model we use TTD mice: fast-aging mice which carry a mutation in the XPD gene that is involved in the Nucleotide Excision Repair (NER) pathway.(9) At a first glance, the phenotype of these animals would appear to require no additional analysis. They exhibit a series of clear phenotypical changes such as osteoporosis and kyphosis, osteosclerosis, early greying, cachexia, infertility, and reduced life-span. However, such a complex phenotype may conceal some basic physiological changes in animal metabolism essential for understanding the effect of the mutation. Using a previously developed CE-MS method for amino acid profiling in human urine as a starting point, (10) we here extend this method to metabolic profiling of urine samples from mice, including all steps required for the processing and statistical analysis of complex samples.(11;12)

MATERIALS AND METHODS

Chemicals. Methanol (MeOH) (HPLC-grade, Biosolve B.V., The Netherlands), ultrapure water (18.2 MΩ/cm), and formic acid (FA) (Fluka, Germany) were used for solvent preparation. A standard solution of 17 amino acids at 1 mM each in 1 M HCl was purchased from Sigma (Sigma-Aldrich, Germany). Dopamine hydrochloride, folic acid, and Phe-Gly hydrate were purchased from Fluka (Germany). Sarcosine, theophylline, caffeine, nortriptyline hydrochloride, creatinine, 4-O-methyldopamine hydrochloride, homovanillyl alcohol, glutathione, thyroxin, and 5-hydroxyindole-3-acetid acid (5-HIAA) were acquired from Sigma (Germany). Stock solutions of the 30 reference compounds were prepared in water at a concentration of 200 μM.

Mouse Urine Samples. Animals were housed at the Animal Resource Center (Erasmus University Medical Center), which operates in compliance with the "Animal Welfare Act" of the Dutch government, using the "Guide for the Care and Use of Laboratory Animals" as its standard. As required by Dutch law, formal permission to generate and use genetically modified animals was obtained from the responsible local and national authorities. All animal studies were approved by an independent Animal Ethical Committee.

Experiments were performed in accordance with the "Principles of laboratory animal care" (NIH publication no. 86-23) and the guidelines approved by the Erasmus University animal care committee. The generation of XPD^{TTD} alleles has been previously described.(13) XPD^{TTD} homozygous mutant animals were obtained by crossing $XPD^{TTD/+}$ with $XPD^{TTD/+}$ mice in a pure C57Bl6J background. Wild-type littermates were used as controls. Mice were housed in individual ventilated cages with *ad libitum* access to AIN93 M synthetic food.

In total, 75 mice were used; 38 TTD mutants and 37 wild-type. Groups were gender matched. The urine was collected at 4 time points, 26, 45, 52, and 65 weeks. Each mouse urine was collected on a piece of Parafilm between 11.00 and 13.00 h and stored at -70 °C. Four microliters of urine was mixed with 4 μL of MeOH, 11 μL of water, and 1 μL of background electrolyte (BGE); centrifuged to eliminate any possible sediment remaining; and put into vials for injection into the CE instrument.

Instrumentation. CE was performed on a P/ACE ProteomeLab PA 800 (Beckman Coulter, Fullerton, CA). Fused silica capillaries (BGB-Analytik, Germany) with total length of 100 cm with 50 μm inner diameter were used for separation. Twenty percent MeOH with 2 M FA was used as background electrolyte. Sample injection was performed hydrodynamically with pH-mediated stacking: a small plug (50 mbar, 9 s) of 12.5% ammonium hydroxide (NH_4OH) was injected before the sample plug (50 mbar, 90 s). The sample plug injected corresponds to a volume of ca. 69 nL, which corresponds to 3.9% of total capillary volume. A washing step of 5 min with 0.1 M NaOH was included between the

runs. During rinsing with 0.1 M NaOH, the end plate voltage, capillary voltage, and the nebulizer were set to 0, which prevented the solution to enter the vacuum part of the MS. The separation voltage was +30 kV (yielding a current of ca. 23 μA).

MS was performed using a micrOTOF (Bruker Daltonics, Bremen, Germany). Transfer parameters were optimized by direct infusion of an ESI tuning mix (Agilent Technologies, Waldbronn, Germany). Spectra were collected with a time resolution of 1 s. CE-MS coupling was realized by a coaxial sheath liquid interface (Agilent Technologies, Waldbronn, Germany) with methanol-water-formic acid (50:50:0.1, *v/v/v*) as sheath liquid. The following spray conditions were used: sheath liquid flow, 4 μL/min; dry gas temperature, 180 °C; nitrogen flow, 4 L/min; nebulizer pressure, 0.5 bar. Electrospray in positive ionization mode was achieved at -4.5 kV. The analytical performance of the CE-MS method was evaluated by the analysis of a standard mixture of metabolites throughout the experiments.

Tandem MS analysis was performed using a micrOTOF-Q (Bruker Daltonics, Bremen, Germany) instrument.

Data Analysis. All data files were recalibrated on the masses of sodium formiate clusters. For estimation of the number of detected compounds, Find Molecular Features algorithm within DataAnalysis (DA) software package (Bruker Daltonics) was used (signal-to-noise cutoff set to 10, correlation coefficient 0.9).

The alignment of electropherograms was performed using XCMS software (The Scripps Research Institute, La Jolla, CA).(14) The resulting table included the detected ion features and their peak areas, which were then normalized using nonparametric normalization.(15) Afterwards, data was imported into Simca-P+ software package, version 12.0 (Umetrics, Umeå, Sweden) for further multivariate analysis. Following principal component analysis (PCA), partial least-squares discriminant analysis (PLSDA), and orthogonal projections to latent structures discriminant analysis (OPLS-DA), compounds responsible for group separation were found. To identify metabolites of interest, rational chemical formulas were generated based on internally calibrated monoisotopic masses within 10 mDa mass error, using the SmartFormula tool within the DA. The chemically reasonable formulas were submitted to metabolome databases: Kyoto Encyclopedia of Genes and Genomes (KEGG) ligand database,(16) the Human Metabolome Database (HMDB),(17) and the METLIN database.(18) The isotopic distribution patterns of the matched metabolite candidates were simulated with the Simulate Pattern tool of DA and compared with observed mass spectra to reduce further the number of potential elemental compositions.(19)

The results of MS/MS analysis confirmed previously acquired identity of the compounds with the help of public databases: HMDB, METLIN, and MassBank.(20)

RESULTS AND DISCUSSION

Recently, we have developed a CE-ToF-MS method for the highly efficient and sensitive analysis of amino acids in human urine.(10) The application of this method for phenotype characterization using metabolic profiling demands broader coverage of the metabolites. Therefore, we first evaluated whether this method can be used for the profiling of diverse metabolites. For this, we have analyzed a standard mixture composed of compounds from different chemical families: amino acids, alcohols, xanthines, amines, dipeptides, and so forth (Supplementary Materials, Table S1). Under the conditions used for separation, pH 1.8, all basic and amphoteric compounds are positively charged, allowing their migration toward the MS detector. Except for glycine, which is outside the selected mass window, all the compounds in the standard mix have been analyzed within 30 min (Supplementary Materials, Figure S1). Compounds that are doubly charged under the used conditions (*e.g.*, lysine, creatinine, histidine) migrated first, followed by amino acids carrying a single charge (*e.g.*, threonine, serine, valine, *etc.*) as well as small peptides (e.g., phenylalanyl-glycine, glutathione). The next group is formed by compounds containing phenyl group as well as acidic groups (*e.g.*, thyroxin and folic acid). The clear separation of compounds present in our standard mix indicates that this CE-MS method can be used for the analysis of various classes of metabolites as required in metabolic profiling studies.

For the evaluation study and testing of the CE-MS method for volume-limited samples, a cohort of urine samples from accelerated aging TTD mice was chosen. One of the model's features is significantly reduced body weight: as a consequence the sampling of body fluids is difficult and only volume-limited samples can be collected. Thus, CE-MS is the method of choice for this particular set of samples. Our workflow included CE-ToF-MS measurement, data preprocessing (alignment and normalization), data analysis, and identification of metabolites of interest (Figure 1).

By using pH-mediated stacking as an in-capillary pre-concentration step, around 600 compounds were detected on average in mouse urine (see Supplementary Materials, Figure S2), which is an improvement of 50% compared to the analysis without pH-mediated stacking. A substantial fraction of those compounds could be identified on the basis of their accurate mass and isotopic pattern.

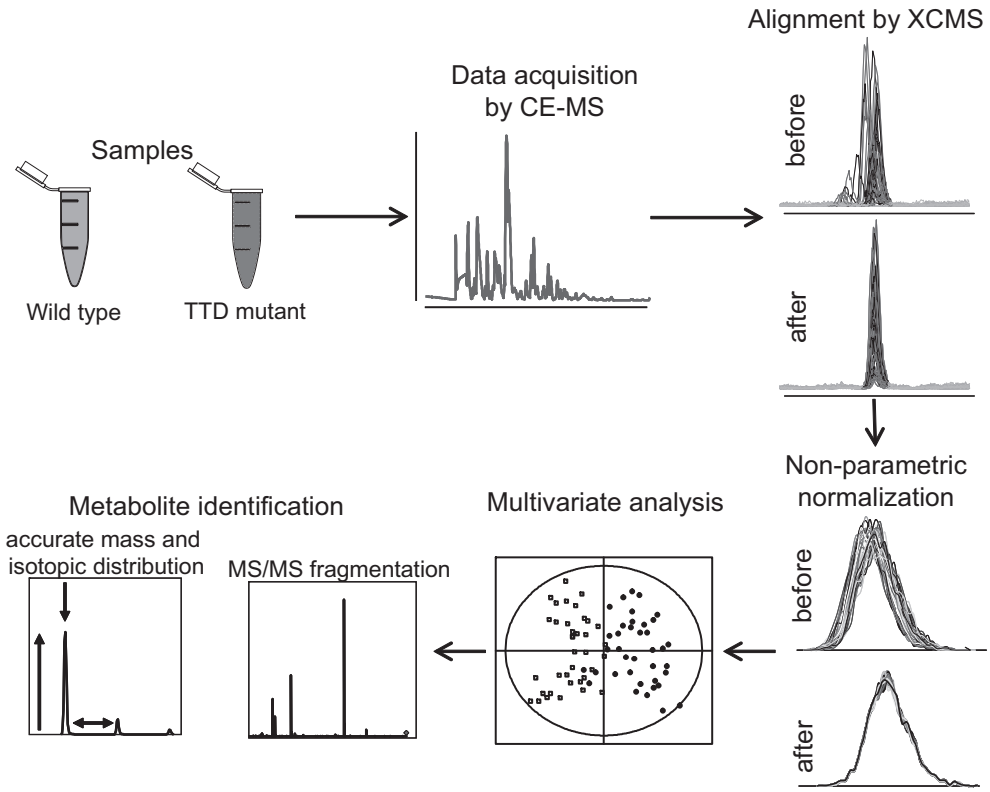

Figure 1. Analytical workflow.

One of the frequently discussed drawbacks of CE-MS using bare fused silica capillaries is the lack of migration time reproducibility. Therefore, XCMS, which is a package combining alignment, peak picking, and statistical analysis, was used for data analysis.(14) The output contains detected ion features and their areas across all samples. To find features responsible for the separation of the groups of animals, a multivariate data analysis approach was used, including unsupervised as well as supervised methods. Prior to statistical analysis, nonparametric normalization was done, as previously described, (15) which unifies distributions of variables. Performing principal component analysis (PCA) on the set of standards run between blocks of mouse urine samples, it is clear that the variation present in the group of biological samples is much larger than the variation between standards (Figure 2A). This means that the analytical variation does not interfere with natural variation between the samples under study. PCA models were also computed for the whole cohort (Supplementary Materials, Figure S3) as well as for samples of male and female mice separately (Supplementary Materials, Figure S4). None of the PCA scores plots showed clustering of the samples according to genotype. This is not unusual in metabolic

profiling studies as the differences of interest might be small and obscured by the relatively large, intrinsic variation between biological samples. To reveal this information, supervised methods were used. Partial least squares discriminant analysis (PLS-DA) model for the whole cohort was quite poor (R2Y = 0.71, Q2 = 0.28). However PLS-DA models, computed for males and females separately, showed that for females the model was quite satisfactory with R2Y and Q2 parameters equal to 0.977 and 0.5, respectively, unlike for males with very low predictive ability (R2Y) 0.78, Q2 < 0, Supplementary Materials, Figure S5). Female mice were chosen for modeling differences between wild-type and mutant phenotypes and retrieval of compounds responsible for separation between them. Next, orthogonal PLS-DA (OPLS-DA) was used to separate systemic noise from variation correlated with the studied classes' discrimination. The scores plot for OPLS-DA model (Figure 2B) shows that the wild-type and mutants are clearly separated from each other.

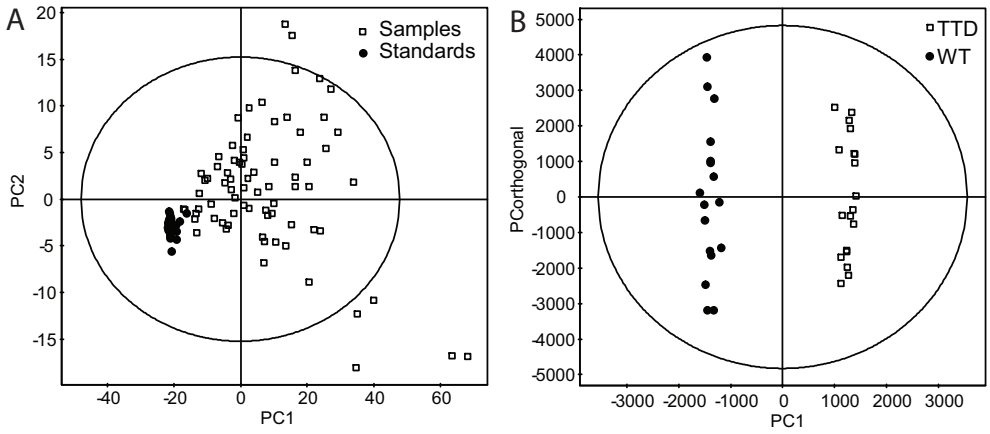

Figure 2. Multivariate data analysis of obtained CE-MS data. (A) PCA scores plot on samples (white boxes) and standards (black circles). First two principal components cover 56.8% of the variation. (B) OPLS-DA scores plot discriminating wild-type female mice (black circles) from mutant TTD female mice (white boxes).

The compounds responsible for discrimination between classes have been chosen based on an S-plot, which shows the compounds' importance and reliability (modeled covariance higher than 0.13 or less than -0.13, modeled correlation more than 0.2 or less than -0.2). Seven compounds fulfilled these criteria and were used for further analysis as possible markers of TTD phenotype (Table 1).

Table 1. Identified classifiers between female TTD mutants and female wild type animals.

measured mass	migration time, min (RSD, %)[a]	identified compound	formula	theoretical mass	error, mDa	mSigma	fold change[b]	p-value[c]	modelled covariation	modelled correlation
399.1459	22.5 (2)	S-Adenosyl-L-Methionine	$C_{15}H_{23}N_6O_5S$	399.1445	-1.4	25	9.5	0.0009	0.22	0.51
156.0769	22.3 (2.5)	Histidine	$C_6H_{10}N_3O_2$	156.0768	-0.2	5.2	3.2	0.014	0.14	0.36
189.1589	22.05 (1.7)	Trimethyl-L-Lysine	$C_9H_{20}N_2O_2$	189.1589	0.8	8.4	6.4	0.012	-0.16	-0.4
204.1349	28.3 (1.8)	Gly-Lys, Gly-Lys	$C_8H_{17}N_3O_3$	204.1343	-0.6	10.8	2.5	0.034	0.12	0.3
147.113	21.3 (1.3)	Lysine	$C_6H_{14}N_2O_2$	147.1128	-0.2	7.1	-1.9	0.064	-0.13	-0.27
188.1747	20.8 (1.7)	N-acetylspermidine	$C_9H_{21}N_3O$	188.1757	1	10.4	1.5	0.086	0.11	0.24
151.1443	20.5 (3.5)	Unidentified					-1.6	0.107	-0.09	-0.2

[a] RSD – relative standard deviation in migration time

[b] calculated as difference in mean areas in TTD mutant relative to wild type animals (negative values indicate that the compound is decreased in TTD compared to WT, positive that it is increased in TTD compared to WT)

[c] unpaired t-test using a Benjamini–Hochberg correction for the p-values

Identification of compounds of interest was performed using the SmartFormula tool within the Bruker Daltonics' DataAnalysis software package and various databases. As some of the tentatively identified compounds were also present in the standard mixture, their migration times were compared and confirmed. The chemical structures of these compounds have been confirmed by tandem mass spectrometry. A sample MS/MS spectra for the compound with the mass 399.1459 (S-Adenosyl-L-Methionine) is shown in Figure 3. Only one compound with mass 151.1443 has not been identified.

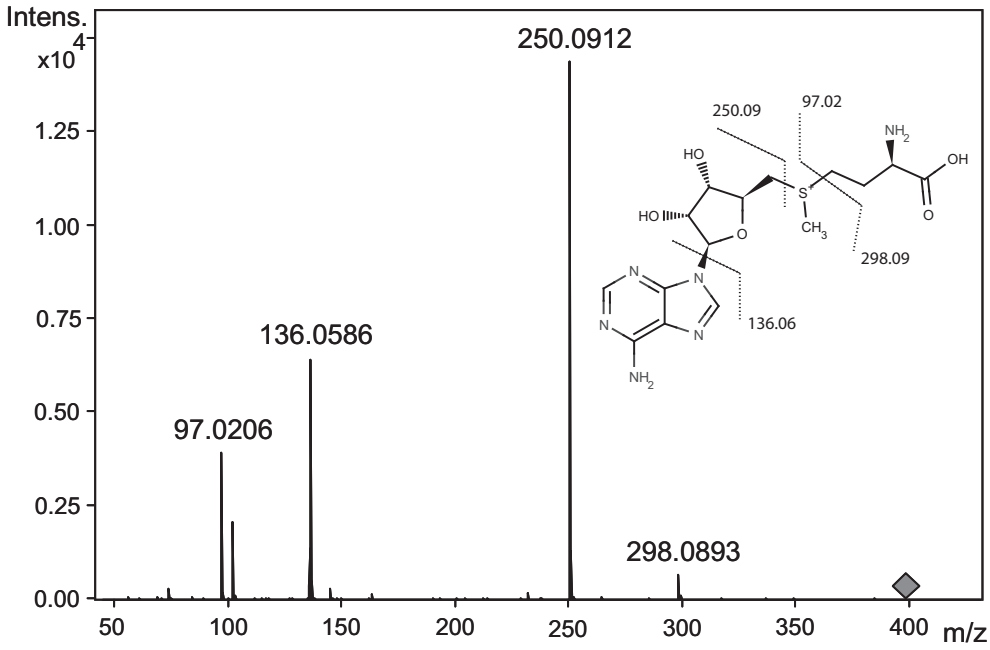

Figure 3. MS/MS spectrum for the compound of interest with mass 399.1459.

Regardless of the analytical method used (NMR, LC-MS, GC-MS), the interpretation of the biological significance of metabolic profiles is based only on a fraction of the metabolome accessible to a particular technique, and CE-MS is not an exception. Despite the mentioned above bias toward polar, positively charged metabolites, our method was capable of revealing gender specific effects of the mutation. The identified classifiers contribute most significantly to the stability of OPLS model, but an attempt to build a biological interpretation solely on those classifiers could be misleading. The measured perturbation in the metabolic composition of mouse urine is a result of complex interplay of many physiological processes. Thus, it is more practical to treat those metabolites as a part of the metabolic "signature of the phenotype" rather than the fingerposts to the

particular pathways. Nevertheless, some of the identified metabolites deserve at least a brief comment.

For example, one of these compounds, histidine, is a classical "usual suspect".(21) Such compounds are often reported as differential metabolites in many pathological states (*e.g.*, (22;23)), but because they are involved in a large number of biochemical processes, it is difficult to assign them to a specific pathway. L-lysine has been associated with effects on bone health, as it stimulates intestinal calcium absorption and cross-linking of bone collagen. Because TTD mice exhibit spontaneous development of osteoporosis, this particular molecule and its relation to bone density, for example, would be interesting to investigate further. N-acetylspermidine and trimethyl-L-lysine are linked in their biosynthesis to another identified classifier, S-adenosyl-L-methionine. Their simultaneous change can indicate the existence of an underlying mechanism regulating them. N-Acetylspermidine is involved in processes of cell growth and differentiation, in adaptation of the cell to a range of stress conditions, and in protection of DNA, lipids, and proteins from oxidative damage. S-Adenosyl-L-methionine is a methyl donor; methylation is one of the mechanisms of regulation of cell growth and differentiation as well as the main epigenetics gear. It is also important for generation of an antioxidant glutathione. As these TTD mice suffer from oxidative damage of DNA that is not repaired by DNA-repair mechanism, up-regulation of the compounds related to antioxidant defense may indicate activation of other protective mechanisms.

To summarize, we have shown that CE-ToF-MS is a very attractive method for metabolic profiling in urine especially in studies with mice where the sample amount is rather limited. Using this method, we have demonstrated differences in metabolic composition between wild-type and mutant animals. These differences were found to be more prominent within female mice, which is in accordance with other phenotypical observations. The identified discriminating compounds would appear to be an interesting group of molecules and need further investigation.

CONCLUSIONS

In this work, we have outlined an analytical workflow based on CE-MS for metabolic profiling of volume-limited samples, that is, mouse urine. We have shown that with a limited amount of sample a wide array of metabolite classes can be covered in mouse urine with the CE-MS method using minimal sample pretreatment and in-capillary preconcentration. The CE-MS method outlined here is especially suited for the profiling of cationic metabolites. In the near future, we will extend this methodology to the profiling of

acidic compounds at high pH separation conditions to further increase the detection coverage of metabolites in urine.

The feasibility of the CE-MS-based workflow for small-sized samples was shown for urine samples from TTD mutant mice. Multivariate analysis of the preprocessed data showed that changes in cationic metabolites are more prominent in female mice. A number of classifiers were identified based on high-resolution MS data as well as fragmentation pattern in MS/MS experiments. These compounds are interesting and the alterations in their concentrations can be related to oxidative defense in the mutant animals.

ACKNOWLEDGEMENTS

This work was supported by Senter Novem (ScreenOP IS061028; All parties), EU FP 7 (MarkAge 200880; Dept. of Genetics and DNage) and EU FP6 (Lifespan, (EC-LSHG-CT-2007-036894); Dept. of Genetics). Gerben Zondag and Ingrid van der Pluijm work for DNage B.V. The authors declare no conflict of interest.

REFERENCES

1. Nicholson,J.K., Lindon,J.C., and Holmes,E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29:1181-1189.

2. Lindon,J.C., Holmes,E., Bollard,M.E., Stanley,E.G., and Nicholson,J.K. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 9:1-31.

3. Webb,A.G. 2005. Microcoil nuclear magnetic resonance spectroscopy. Journal of Pharmaceutical and Biomedical Analysis 38:892-903.

4. Lapainis,T., Rubakhin,S.S., and Sweedler,J.V. 2009. Capillary electrophoresis with electrospray ionization mass spectrometric detection for single-cell metabolomics. Anal Chem 81:5858-5864.

5. Hirayama,A., Kami,K., Sugimoto,M., Sugawara,M., Toki,N., Onozuka,H., Kinoshita,T., Saito,N., Ochiai,A., Tomita,M. *et al* 2009. Quantitative Metabolome Profiling of Colon and Stomach Cancer Microenvironment by Capillary Electrophoresis Time-of-Flight Mass Spectrometry. Cancer Research 69:4918-4925.

6. Leon,C., Rodriguez-Meizoso,I., Lucio,M., Garcia-Canas,V., Ibanez,E., Schmitt-Kopplin,P., and Cifuentes,A. 2009. Metabolomics of transgenic maize combining Fourier transform-ion cyclotron resonance-mass spectrometry, capillary electrophoresis-mass spectrometry and pressurized liquid extraction. J. Chromatogr. A 1216:7314-7323.

7. Chiu,D.T., Lillard,S.J., Scheller,R.H., Zare,R.N., Rodriguez-Cruz,S.E., Williams,E.R., Orwar,O., Sandberg,M., and Lundqvist,J.A. 1998. Probing single secretory vesicles with capillary electrophoresis. Science 279:1190-1193.

8. Osbourn,D.M., Weiss,D.J., and Lunte,C.E. 2000. On-line preconcentration methods for capillary electrophoresis. Electrophoresis 21:2768-2779.

9. de Boer,J., Andressoo,J.O., de Wit,J., Huijmans,J., Beems,R.B., van Steeg H., Weeda,G., van der Horst,G.T., van Leeuwen,W., Themmen,A.P. *et al* 2002. Premature aging in mice

deficient in DNA repair and transcription. Science 296:1276-1279.

10. Mayboroda,O.A., Neususs,C., Pelzing,M., Zurek,G., Derks,R., Meulenbelt,I., Kloppenburg,M., Slagboom,E.P., and Deelder,A.M. 2007. Amino acid profiling in urine by capillary zone electrophoresis - mass spectrometry. J. Chromatogr. A 1159:149-153.

11. Shulaev,V. 2006. Metabolomics technology and bioinformatics. Brief Bioinform. 7:128-139.

12. Weckwerth,W., and Morgenthal,K. 2005. Metabolomics: from pattern recognition to biological interpretation. Drug Discov. Today 10:1551-1558.

13. de Boer,J., de Wit,J., van Steeg,H., Berg,R.J., Morreau,H., Visser,P., Lehmann,A.R., Duran,M., Hoeijmakers,J.H., and Weeda,G. 1998. A mouse model for the basal transcription/DNA repair syndrome trichothiodystrophy. Mol. Cell 1:981-990.

14. Smith,C.A., Want,E.J., O'Maille,G., Abagyan,R., and Siuzdak,G. 2006. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78:779-787.

15. Sidorov I.A., Hosack D.A., Gee D., Yang J., Cam M.C., Lempicki R.A., and Dimitrov D.S. 2002. Oligonucleotide microarray data distribution and normalization. Information Sciences 146:67-73.

16. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H., and Kanehisa,M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 27:29-34.

17. Wishart,D.S., Tzur,D., Knox,C., Eisner,R., Guo,A.C., Young,N., Cheng,D., Jewell,K., Arndt,D., Sawhney,S. et al 2007. HMDB: the human metabolome database. Nucleic Acids Research 35:D521-D526.

18. Smith,C.A., O'Maille,G., Want,E.J., Qin,C., Trauger,S.A., Brandon,T.R., Custodio,D.E., Abagyan,R., and Siuzdak,G. 2005. METLIN - A metabolite mass spectral database. Therapeutic Drug Monitoring 27:747-751.

19. Kind,T., and Fiehn,O. 2006. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics 7.

20. Taguchi,R., Nishijima,M., and Shimizu,T. 2007. Basic analytical systems for lipidomics by mass spectrometry in Japan. Methods Enzymol. 432:185-211.

21. Robertson,D.G. 2005. Metabonomics in toxicology: A review. Toxicological Sciences 85:809-822.

22. Lamers,R.J., van Nesselrooij,J.H., Kraus,V.B., Jordan,J.M., Renner,J.B., Dragomir,A.D., Luta,G., van der,G.J., and DeGroot,J. 2005. Identification of an urinary metabolite profile associated with osteoarthritis. Osteoarthritis Cartilage 13:762-768.

23. Qiu,Y., Cai,G., Su,M., Chen,T., Liu,Y., Xu,Y., Ni,Y., Zhao,A., Cai,S., Xu,L.X. *et al* 2010. Urinary metabonomic study on colorectal cancer. J. Proteome Res. 9:1627-1634.

SUPPLEMENTARY MATERIALS

Table S1. Compounds included in the standard mixture.

Amino acids	Alanine
	Arginine
	Aspartic acid
	Cysteine
	Glutamic acid
	Glycine
	Histidine
	Isoleucine
	Leucine
	Lysine
	Methionine
	Phenylalanine
	Proline
	Serine
	Threonine
	Tyrosine
	Valine
	Sarcosine
	Thyroxine
Alcohols	Homovanillyl alcohol
Pterins	Folic acid
Xanthines and related coumpounds	Caffeine
	Theophylline
Compound with Indoles group	5-hydroxyindole-3-acetic acid
Amines	Nortriptyline
Compounds with Hydroxyl and Amine groups	Dopamine
	4-O-Methyldopamine
Compounds with Imidazol groups	Creatinine
Dipeptides	Phe-Gly
Tripeptides	Glutathione

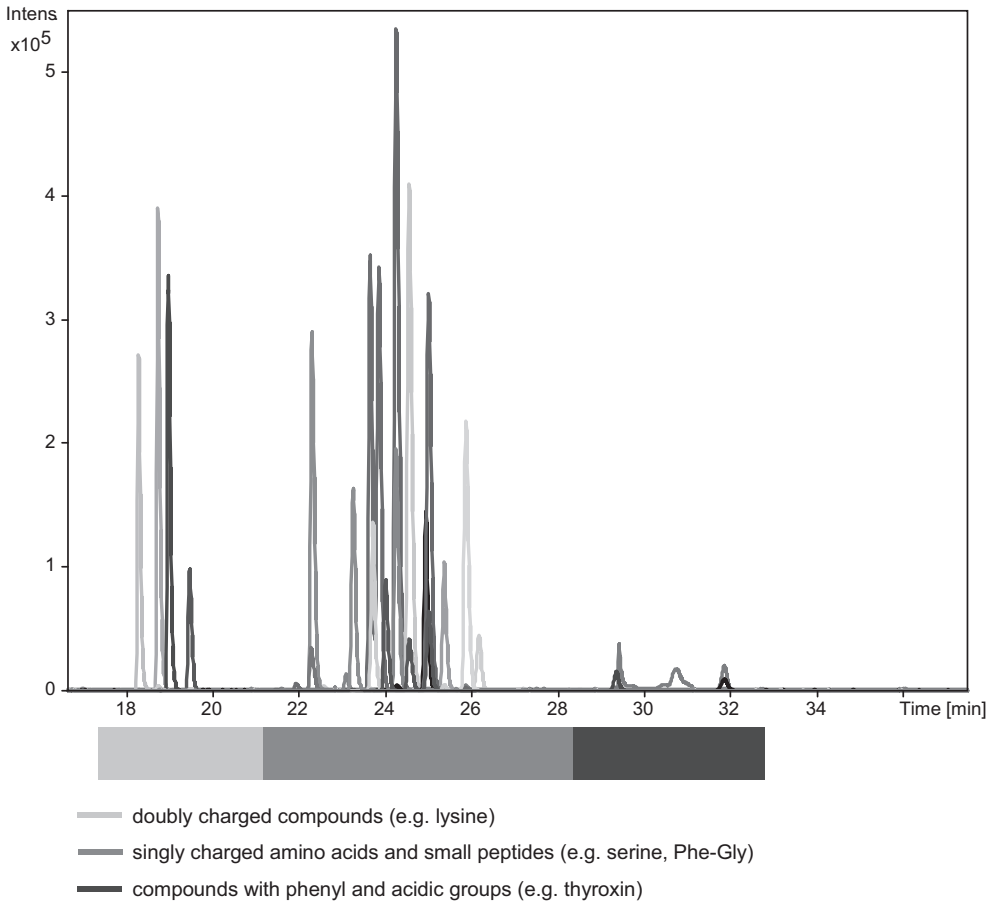

Figure S1. Combined extracted ion electropherograms of compounds present in standard mixture.

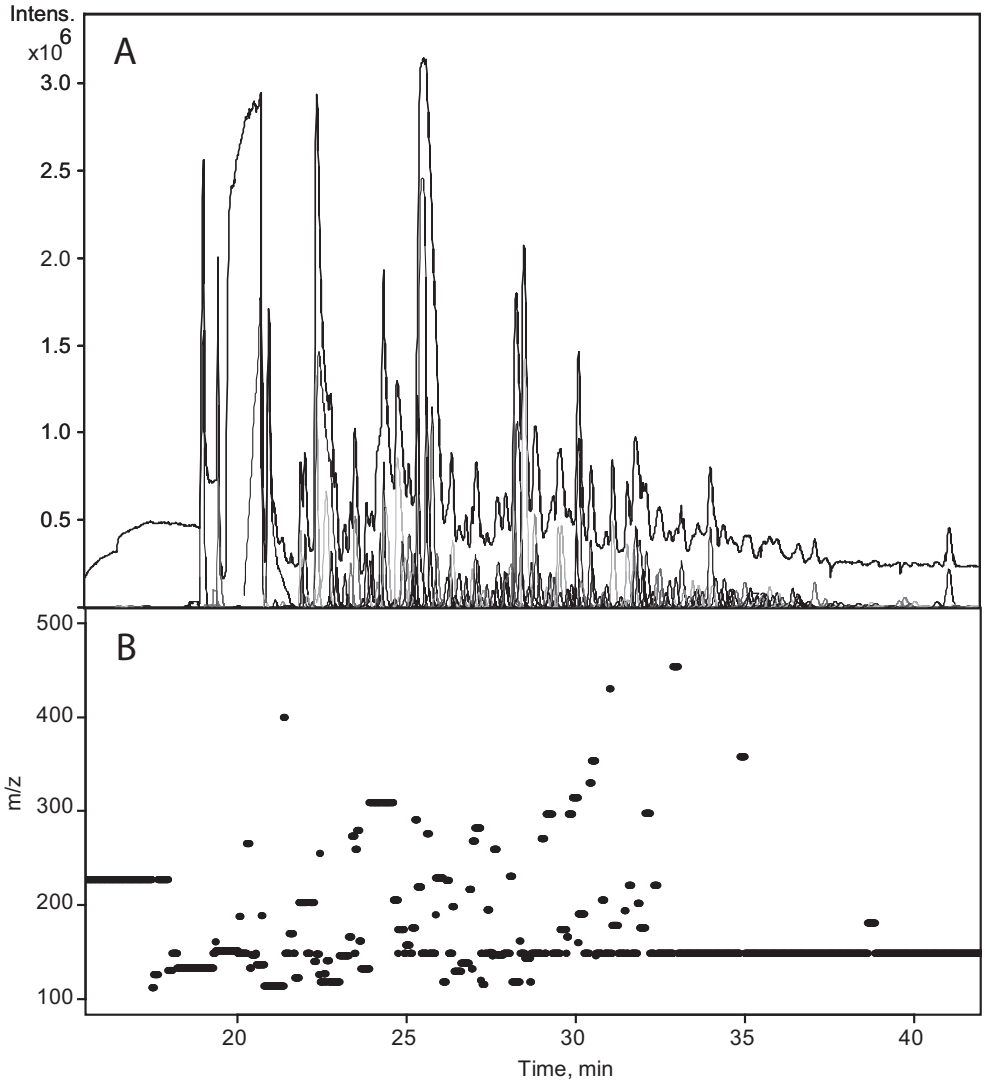

Figure S2. Typical metabolic profile of mouse urine by CE-MS. (A) Total ion electropherogram and extracted ion electropherograms. (B) Masses of the most intense peaks in the spectrum.

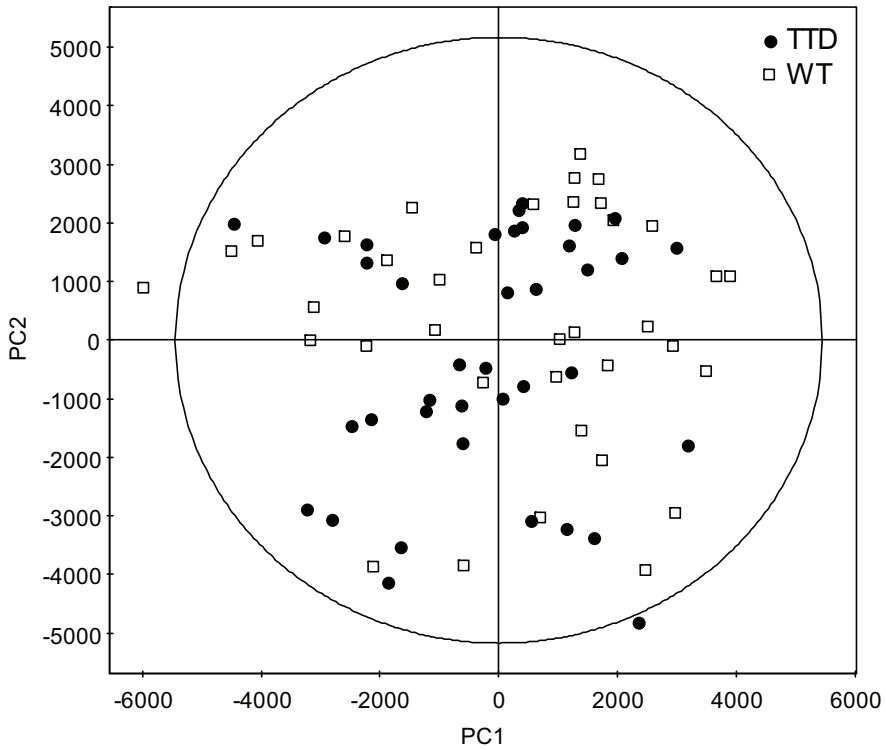

Figure S3. PCA scores plot for 75 mouse urine samples.

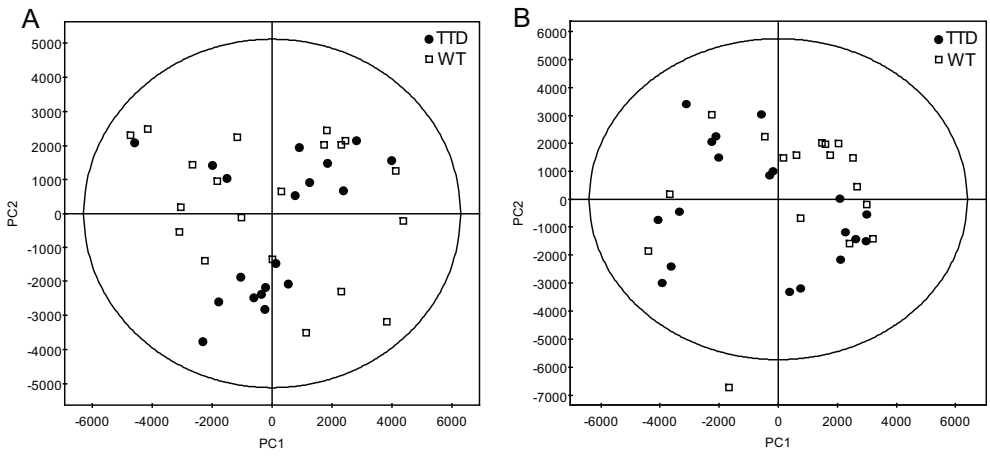

Figure S4. PCA scores plot for male (A) and female (B) mice.

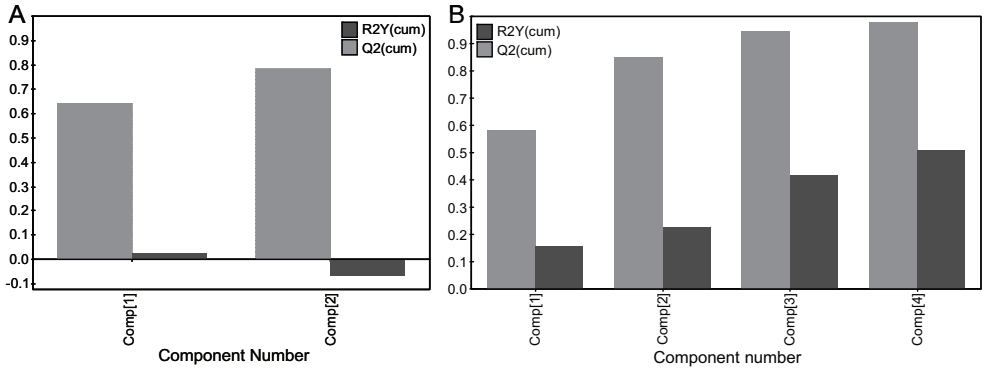

Figure S5. Overview of PLS-DA models for male (A) and female (B) mice.

Chapter 4

Metabolic Profiling of Accelerated Aging ERCC1$^{d/-}$ Mice

Nevedomskaya E., Meissner A., Goraler S., de Waard M., Ridwan Y., Zondag G., van der Pluijm I., Deelder A.M, Mayboroda O.A.

Journal of Proteome Research **2010**, 9, 3680–3687

ABSTRACT

Aging is a fundamental biological process for which the mechanism is still largely unknown due to its complex and multifactorial nature. Animal models allow us to simplify this complexity and to study individual factors separately. As there are many causative links between DNA repair deficiency and aging, we studied the ERCC1$^{d/-}$ mouse, which has a modified ERCC1 gene, involved in the Nucleotide Excision Repair, and as a result has a premature aging phenotype. Profiling of these mice on different levels can give an insight into the mechanisms underlying the aging phenotype. In the current study, we have performed metabolic profiling of serum and urine of these mice in comparison to wild type and in relation to aging by ^{1}H NMR spectroscopy. Analysis of metabolic trajectories of animals from 8 to 20 weeks suggested that wild type and ERCC1$^{d/-}$ mutants have similar age-related patterns of changes; however, the difference between genotypes becomes more prominent with age. The main differences between these two genetically diverse groups of mice were found to be associated with altered lipid and energy metabolism, transition to ketosis, and attenuated functions of the liver and kidney.

INTRODUCTION

Aging is a complex biological process that involves multiple systems at different regulation levels. Understanding of its mechanisms will allow better understanding of the impairment of human health that occurs at old age.(1) Despite many efforts, the etiology of aging is still largely unknown. It is evident that it cannot be explained by a single or a couple of mechanisms but rather by a complex network of interdependent processes involved.

Several theories have been proposed for mechanisms of aging, one of them being "free radical theory of aging".(2;3) According to this theory, free radicals represent the greatest danger for DNA, which is the blueprint for all genes and therefore RNAs and proteins. If DNA is irreversibly damaged, this has severe consequences for human health, which is illustrated by several human inherited DNA repair deficient disorders, which all show signs of premature aging. Examples of these diseases are Trichothiodystrophy and Cockayne syndrome, which both bear mutations in genes involved in nucleotide excision repair (NER) and as a result develop a plethora of premature aging symptoms including, but not limited to, neuronal problems, growth retardation, vision and hearing impairment, and a greatly reduced lifespan.(4) This illustrates the importance of DNA repair and hints at the fact that unrepaired DNA damage results in an aging phenotype.

Due to its complex nature and the time involved to develop the phenotype, aging is difficult to study in humans. A possible alternative is the use of animal models such as mice, nematodes, and others.(5) The ERCC1$^{-/-}$ mouse with a single gene knockout in nucleotide excision repair system is one of such model animals.(6) The ERCC1 gene encodes a protein that is part of an endonuclease complex essential for both NER and interstrand cross-link repair (ICLR).(7) These mutant mice have a significantly reduced lifespan of up to 4 weeks only and a severe premature aging phenotype with a range of features including aging-like skin abnormalities, reduced growth, liver and kidney dysfunction, as well as others.(8) This extremely short lifetime makes it difficult to monitor the progression of aging, and the highly anomalous phenotype does not only reflect fast aging. The ERCC1$^{d/-}$ mouse combines knockout of ERCC1 in one allele and a truncated ERCC1 allele (9) and helps to overcome those extreme effects, as this delays the display of symptoms of aging and prolongs the lifespan. The resulting phenotype shows premature aging features such as neurological problems, impaired vision and hearing, growth retardation, a shortened lifespan of about 6 months (as compared to 2.5/3 years for a wild type mouse), and also accumulation of somatic mutations.(10)

ERCC1$^{d/-}$ mice aging phenotype is expected to be reflected in the composition of biofluids. This can be the subject of investigation by proteomics, glycomics, or metabolomics. The latter is of particular interest as it is focused on studying small

molecules that are end-points of biochemical processes and can give insight into changes happening in the whole organism.(11;12)

One of the established methods in metabolomics is nuclear magnetic resonance (NMR), which allows measurement and recovery of molecular information over a wide range of small compounds.(13;14)

Two main questions arise with regard to metabolic profiling of ERCC1$^{d/-}$ mice. The first one is how different the metabolome of a particular biofluid of the mutant mice is from that of the wild type. The other one is how the metabolic profiles change with aging of ERCC1$^{d/-}$ animals in general and in comparison to changes that occur in normal mice. These questions imply that the experimental design should include both mutants and wild type mice as well as follow the same animals to monitor metabolic changes with aging in a longitudinal design. To answer these types of questions on the basis of NMR, data multivariate statistical analysis tools (*e.g.*, Principal Component Analysis, Partial Least Squares *etc.*) are needed.(15;16)

In the current study, we have analyzed cohorts of serum and urine samples from wild type and ERCC1$^{d/-}$ mutant mice by ^{1}H NMR and have detected compounds that differ between the groups as well as specific age-related changes. The biological significance of these findings is discussed.

MATERIALS AND METHODS

Sample Collection. Experiments were performed in accordance with the "Principles of laboratory animal care" (NIH publication no. 86-23) and the guidelines approved by the Erasmus University animal care committee. The generation of ERCC1^{-} and ERCC1^{d} alleles has been previously described.(9) ERCC1$^{d/-}$ mice were obtained by crossing ERCC1^{-} with ERCC1$^{d/+}$ mice of C57Bl6J and FVB backgrounds to yield ERCC1$^{d/-}$ with C57Bl6J/FVB hybrid background. Wild-type littermates were used as controls. Mice were housed in individual ventilated cages with ad libitum access to standard mouse food (CRM pellets, SDS BP Nutrition Ltd.; gross energy content 18.36 kJ/g dry mass, digestible energy 13.4 kJ/g) and water. Food intake was measured for animals individually; no difference for the studied group of ERCC1$^{d/-}$ mutants compared to wild type animals was observed for the studied age group relative to the body weight (table with body weights of groups of animals, Supplementary Materials, Table S1).

Longitudinal serum samples were collected from 20 animals, 10 wild type and 10 ERCC1$^{d/-}$ mutants, gender matched, at 4 time points - 8, 12, 16, and 20 weeks. Blood was collected via extraction from the tail vein, after which serum was collected by centrifugation at 6000 rpm and the supernatant was transferred to a fresh tube and stored at -70 °C. A few

samples were missing or had too small a volume, so that the total number of samples measured was 70.

Urine samples were collected from 13 wild type animals and from 13 ERCC1$^{d/-}$ mutants, with unbiased selection of gender and age between 8 and 16 weeks. Urine of animals was collected on a piece of Parafilm between 11.00 and 13.00 h for each mouse and stored at -70 °C.

Blood glucose was measured using a Freestyle mini blood glucose measurement device (Abbott Diabetes Care).

NMR Sample Preparation. For sample preparation, buffer was added to all samples to reduce variability in pH and supply a deuterated lock solvent. The specific buffers for urine and serum are described below.

For serum preparation, 60 µL of 75 mM phosphate buffer in H_2O/D_2O (80/20) at pH 7.4 containing 6.15 mM NaN_3 and 4.64 mM sodium 3-[trimethylsilyl] d4-propionate (TSP) was added to 20 µL of serum and manually transferred into Bruker 1.7 mm NMR Match tubes.

For urine preparation, 40 µL of 0.20 M phosphate buffer in D_2O at pH 7.0 containing 0.26 mM NaN_3 and 0.53 mM TSP was added to 40 µL urine and manually transferred into Bruker 1.7 mm NMR Match tubes.

NMR Spectroscopy. All NMR experiments were acquired on a 600 MHz Bruker Avance II spectrometer (Bruker BioSpin, Karlsruhe, Germany) equipped with a 5 mm TCI cryogenic probe head with Z-gradient system and automatic tuning and matching. Temperature calibration was done prior to each batch of measurements using the method of Findeisen *et al.*(17)

For urine, one-dimensional ^{1}H NMR spectra were recorded at 300 K using the first increment of a NOESY (18) pulse sequence with presaturation (γB_1 = 50 Hz) during a relaxation delay of 4 s and a mixing time of 10 ms for efficient water suppression. A total of 32 768 data points were recorded with 32 scans covering a sweep width of 12336 Hz. The free induction decay (FID) was zero-filled to 65 536 complex data points prior to Fourier transformation and an exponential window function was applied with a line broadening factor of 1.0 Hz.

For serum, 1D NOESY and 1D diffusion edited(19) experiments were recorded at 310 K. Presaturation with an effective field of 50 Hz during a relaxation delay of 4 s was applied and a total of 98 304 data points were recorded covering a sweep width of 18 029 Hz for both experiments. For 1D NOESY, a total of 32 scans were accumulated with water resonance saturation (γB_1 = 50 Hz) during a mixing time of 10ms. For the diffusion edited 1D experiment, a gradient echo delay of 116 ms was utilized and a total of 64 scans were recorded for each sample. For both experiments, the FID was zero-filled to 131 072 complex

data points prior to Fourier transformation and an exponential window function was applied with a line broadening factor of 1.0 Hz. All spectra were manually phase and baseline corrected using Topspin 2.1 (Bruker BioSpin, Karlsruhe, Germany) and automatically referenced to TSP (0.0 ppm).

Data Analysis. Each spectrum was integrated using 0.04 ppm integral regions between 11 and -1 ppm, excluding the residual water region in serum spectra from 5.0 to 4.6 ppm, the residual water and urea region in urine spectra from 5.0 to 4.7 ppm and from 6.2 to 5.6 ppm respectively, and the TSP signal from 0.025 to -0.025 in both sets of spectra. The citrate region in urine spectra was integrated using spectral regions from 2.76 to 2.66 ppm and from 2.60 to 2.52 ppm to avoid effects introduced by positional noise of the peaks due to differences in ion strength and pH between the samples. To account for any difference in concentration between samples, each spectrum was normalized to its total area.

Data sets were imported into SIMCA-P+ 12.0 (Umetrics, Umeå, Sweden) to perform multivariate statistics: principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA). The partial least squares (PLS) method was used for analysis of time changes in spectra with aging of the animals. Pareto scaling was used for all the statistical models.

Identification of metabolites was facilitated by using the Statistical Total Correlation Spectroscopy (STOCSY) approach (20) using in-house developed routines written in R statistical language (http://www.r-project.org/). This method determines and visualizes correlations between peaks in sets of NMR spectra, allowing annotation of peaks belonging to the same molecule. Annotation of peaks was performed based on reference spectra from the Bruker Bioref database (Bruker BioSpin, Karlsruhe, Germany).

For visualization of differentially expressed metabolites, spectra were aligned using the COW algorithm (21) and averaged.

For glucose quantification, all non-normalized spectra were referenced to the doublet in the anomeric region. Quantification was performed by deconvolution and subsequent integration of glucose anomeric signal (5.25-5.22 ppm) using an in-house developed automation routine (see Supplementary Materials). The absolute concentrations were calculated based on internal reference TSP with correction for TSP protein binding and line-broadening effects.

RESULTS

Profiling of Serum Metabolites in ERCC1$^{d/-}$ Mice by ^{1}H NMR. PCA analysis of the integrated (binned) NOESY data reflects both effects of genotype (differences between wild type and mutants) (Figure 1A) and age-related differences (Figure 1B). The first two

principal components cover more than 65% of the variation and reflect these biological processes. No gender-related separation was observed.

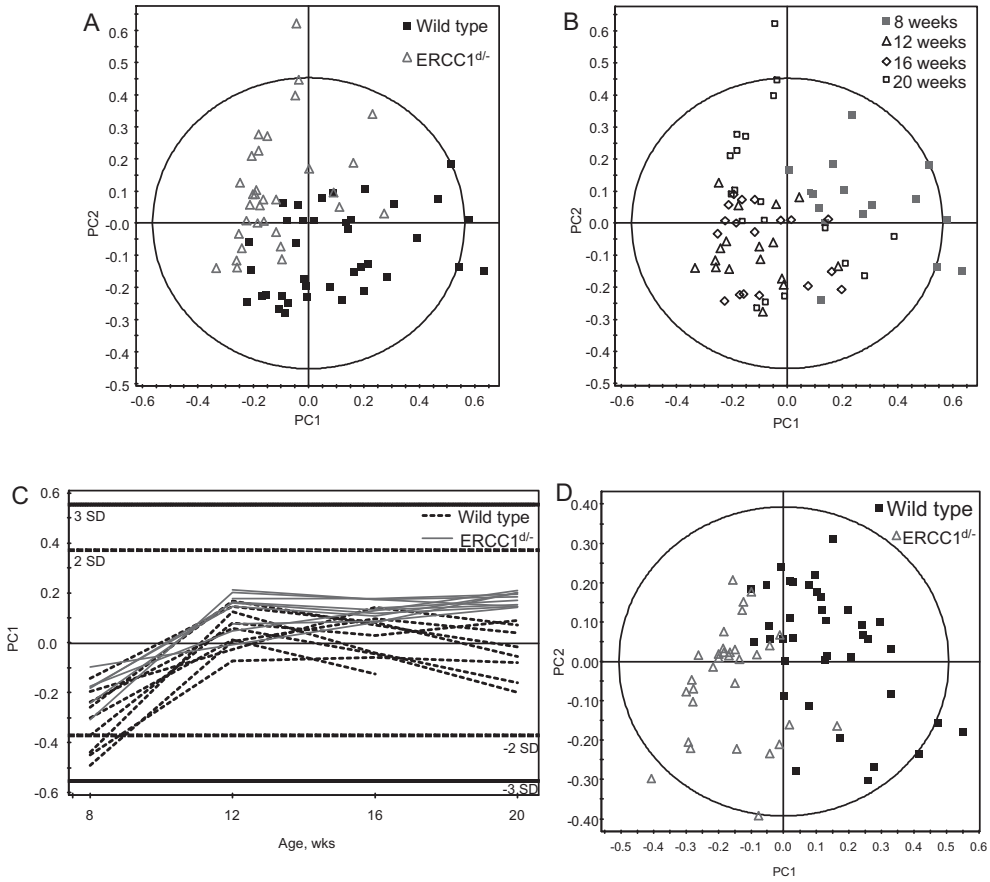

Figure 1. Multivariate statistical analysis of serum NMR data. (A) PCA scores plot, wild type (■) and ERCC1$^{d/-}$mutant mice (Δ). (B) PCA scores plot, 8 weeks (■), 12 (Δ), 16 (◊) and 20 (□) weeks mouse samples. (C) Batch PLS scores plot of serum samples mapped across the different ages. Dashed horizontal lines show two and three standard deviations for the data set. Dashed lines represent wild type, grey lines represent ERCC1$^{d/-}$ mutants. (D) PLS-DA scores plot (for first two principal components R2Y = 60.3%, Q2 = 56.9%) showing discrimination between wild type (■) and ERCC1$^{d/-}$mutant mice (Δ).

The PCA scores plot (Figure 1B) shows that profiles of both mutants and controls at 8 weeks considerably differ from profiles at other ages. Dependence on the age was studied using PLS batch analysis; the clear trend on the age could only be seen along the first

principal component (Figure 1C), while along the other components variation is not related to the lifetime. Compounds associated with the age component were found to be mainly related to lipids and lipoproteins (Supplementary Materials, Figure S1).

In addition, Figure 1C shows that while at 8 and 12 weeks there is overlap between wild type and mutant animals along the first component, at 16 weeks they start to separate and at 20 weeks mutant animals are completely distinct from the wild type and show less intragroup variation. This observation was also confirmed by building separate PLS-DA models at all four ages. The separation efficiency and model quality increased with age; the explained variation of the response variable (R2Y) values for the models were from 50% at 8 weeks to 87% at 20 weeks, while the variation of the response variable predicted by the model (Q2) increased from 10% to 80%.

Further investigation of the differences between wild type animals and mutants and identification of the molecular discriminators was performed using a PLS-DA model built with the genotype as response variable (Figure 1D). Model validation using permutation test showed a good fit and validity of the model (Supplementary Materials, Figure S2). Spectral regions responsible for separation of the two groups were selected on the basis of VIP values (variable importance in the projection); variables with values over 1.5 were selected. The identity of the underlying compounds was investigated based on the chosen spectral regions. However, identification of chemical structures based only on a single peak that falls into a specified bin interval can be difficult. To assist annotation of key discriminators, the STOCSY approach was used. In this method correlations between a selected peak and all other peaks in the spectra are explored and visualized; peaks characterized by high correlations belong to the same molecule and can be confidently annotated using reference spectra (Supplementary Materials, Figure S4).

As can be seen from Table 1, most of the compounds responsible for separation of mutant mice from wild type are related to lipoprotein distribution and concentrations and some of these regions were also found to be associated with age trajectories.

In addition to changes in the lipid profile, lactate is decreased and alanine is increased. Averaged spectra from both of the groups with selected compounds highlighted are shown in Figure 2.

Investigation of the line shapes of diffusion edited spectra in the region of aliphatic CH_3 and CH_2 lipid resonances allows us to estimate relative changes of different classes of lipoproteins.(22) The differences in the lipid region between the two groups indicate that ERCC1$^{d/-}$ mutants show an increased level of high density lipoproteins (HDL) and a decreased level of low and very low density lipoproteins (LDL and VLDL respectively) compared to wild type (Figure 3).

Table 1. Integrated regions (bins), which show different concentrations between serum profiles of wild type and mutant animals.

spectral region, ppm	fold change[a]	identity	p-value[b]
1.52-1.48	1.1	Alanine	4.81E-12
0.68-0.64	1.14	Unidentified	2.11E-10
1.6-1.56	-1.15	Lipid **CH$_2$**CH$_2$CO	1.53E-09
2.28-2.24	-1.24	Lipid **CH$_2$**CO	2.13E-09
2.08-2.04	-1.16	N-acetyl glycoproteins	3.84E-09
1.12-1.08	1.09	Unidentified	6.62E-09
0.92-0.88	-1.21	Lipoproteins	1.93E-08
1.56-1.52	1.08	Unidentified	2.41E-08
1.32-1.28	-1.49	Lipid $(CH_2)_n$	3.89E-08
0.88-0.84	1.18	Lipoproteins	8.65E-08
1.36-1.32	-1.28	Lactate	1.90E-07
3.24-3.2	1.3	Choline $-N(CH_3)_3$	1.09E-06
1.4-1.36	-1.16	Lipoproteins	1.46E-06
2.12-2.08	-1.05	Unidentified	1.67E-06
5.36-5.32	-1.34	Unsaturated lipids	5.89E-06
1.24-1.2	1.16	Lipoproteins	1.57E-05
2-1.96	1.05	Lipid	5.33E-05
1.28-1.24	1.16	Lipoproteins	5.64E-05
4.16-4.12	-1.1	Lactate	6.29E-05

[a] Calculated as difference in mean levels in ERCC1$^{d/-}$ mutant relative to wild type animals. The + and - indicate the direction of the change, i.e. - for reduced level in ERCC1$^{d/-}$ samples, + for increased level in ERCC1$^{d/-}$ samples compared to wild type.

[b] Unpaired t-test using a Benjamini–Hochberg correction for the p-values.

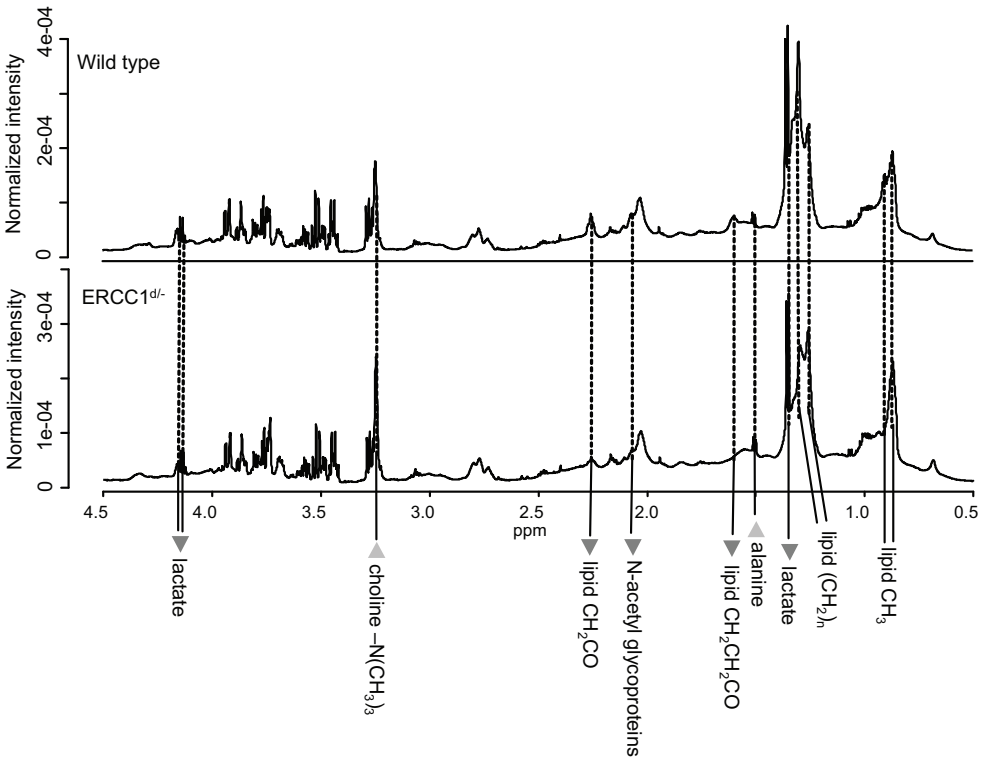

Figure 2. Averaged ^{1}H NMR spectral regions of serum from wild type (above) and ERCC1$^{d/-}$ mutant (below) animals. Differential metabolites are highlighted; ▼stands for down-regulated, ▲ for up-regulated compounds.

Glucose was measured in blood samples by a glucose measurement device (Freestyle mini, Abbott Diabetes Care), and its concentrations were found to be different between the animals with different genetic backgrounds (Supplementary Materials, Figure S5A), p-value obtained by t test is 0.001. However, ppm regions corresponding to glucose did not appear to have an influence in multivariate analysis. The reason for this might be that there are broad signals that overlay the glucose peaks due to high protein content of the samples. That is why proper deconvolution and integration were needed to correctly quantify glucose. Results obtained from NMR data were similar to those received by a conventional method (Supplementary Materials, Figure S5B). The trend was the same: glucose is lower in ERCC1$^{d/-}$ animals than in wild type (p-value obtained by t test is <0.001).

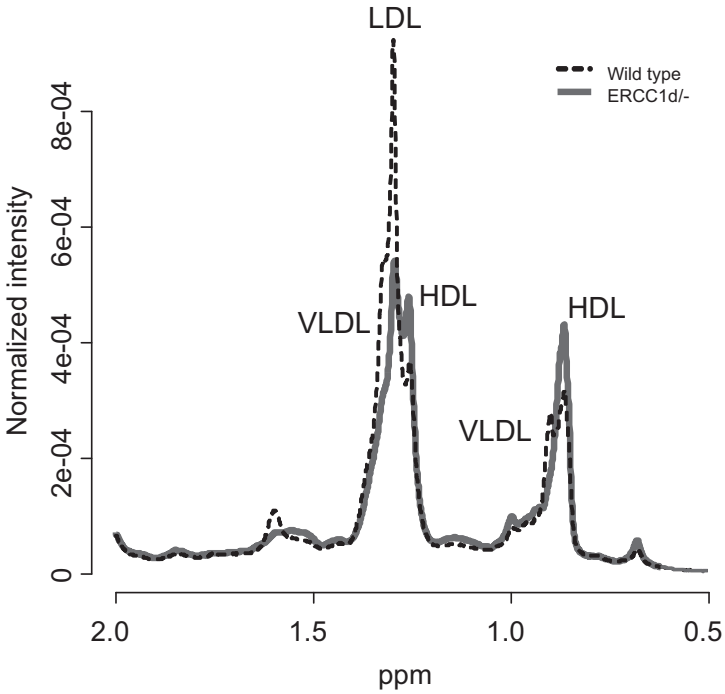

Figure 3. Spectral region corresponding to lipoproteins, averaged diffusion edited spectra of wild type (dashed line), and ERCC1$^{d/-}$mutants animals (grey line).

Profiling of Urine Metabolites in ERCC1$^{d/-}$ Mice by ^{1}H NMR. Global changes in biochemical processes in the organism should be reflected not in one particular biological fluid, such as blood, but should also be seen in others, for instance urine. Longitudinal collection of sufficient quantities of urine from the same animals from which serum was collected was not feasible due to the small size of the mutant animals and difficulties with their urination. However, we performed analysis of urine obtained from a different cohort.

As a first step, a PCA model was built that showed clear separation of wild type animals from ERCC1$^{d/-}$ mutants (Figure 4A). PLS-DA was performed, which gave a very good model with R2Y = 91.7 and Q2 = 85.3% for the first two components. On the basis of this model, variables responsible were identified and annotated (Table 2). The average spectra from both of the groups with these compounds highlighted are shown in Figure 4B. There are other differential signals visible in averaged spectra, but those were not found to be statistically significant.

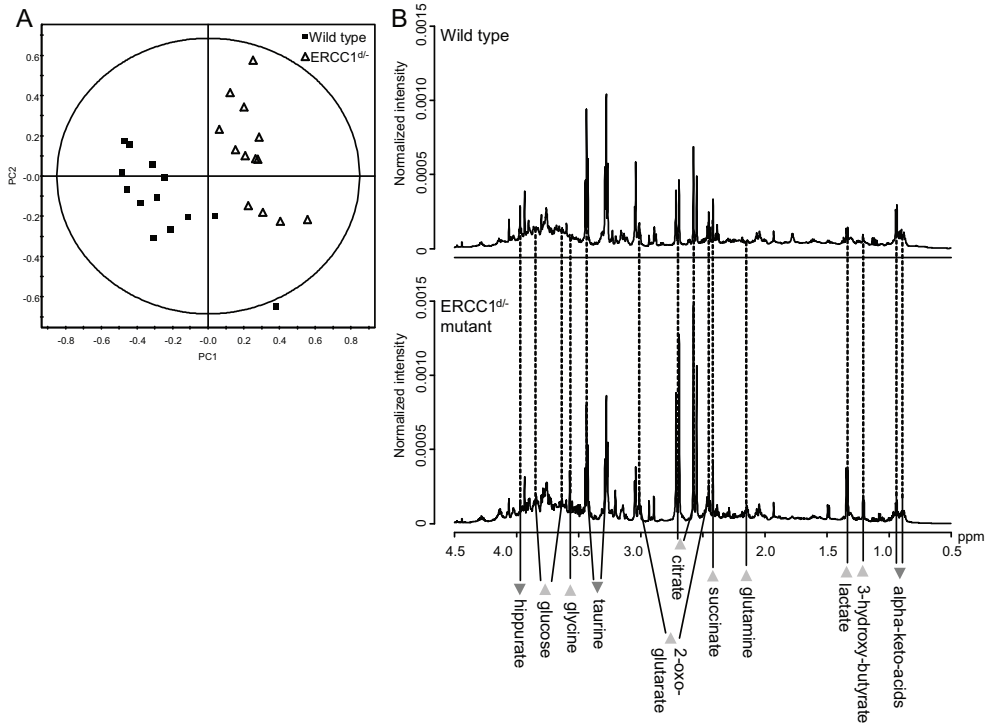

Figure 4. Urine ^{1}H NMR analysis. (A) PCA scores plot on urinary metabolic profiles scores plot, first two principal components cover 35.7% and 23.3% of variability, respectively. Separation between wild type (■) and ERCC1$^{d/-}$ mutant mice (Δ) is visible. (B) Averaged ^{1}H NMR spectral regions of urine samples of wild type (above) and ERCC1$^{d/-}$ mutant (below) animals. Differential metabolites are highlighted; ▼ down-regulated compounds, ▲ up-regulated compounds.

Table 2. Integrated Regions (Bins), which Show Different Concentrations between Urine Profiles of Wild Type and Mutant Animals.

ppm	fold change[a]	identity	p-value[b]
0.96-0.92	-1.56	Alpha-keto acids	7.48E-11
2.64-2.6	-1.59	Ketoleucine	2.08E-10
1.16-1.12	-1.7	Alpha-keto acids	1.60E-09
2.76-2.64	2.18	Citrate	1.60E-05
7.28-7.24	-2.07	Unidentified	4.46E-05
0.92-0.88	-1.49	Alpha-keto acids	5.62E-05
2.6-2.52	2.24	Citrate	9.03E-05
2.16-2.12	1.76	Glutamine	<0.0002
3.24-3.2	1.53	Unidentified	<0.0002
1.36-1.32	2.12	Lactate	0.001
3.56-3.52	1.39	Glucose	0.001
7.56-7.52	-2.51	Hippuric acid	0.001
7.84-7.8	-2.68	Hippuric acid	<0.002
3.28-3.24	1.32	TMAO	0.002
4-3.96	-1.46	Hippuric acid	0.003
3.32-3.28	-1.74	Taurine	0.006
3.52-3.48	1.57	Glucose	0.007
3.6-3.56	1.33	Glycine	0.02
5.4-5.36	-1.19	Allantoin	0.03
2.48-2.44	1.38	2-Oxoglutarate	0.05
2.44-2.4	1.15	Succinate	0.14
1.24-1.2	1.03	3-Hydroxybutyrate	0.43

[a] Calculated as difference in mean levels in ERCC1$^{d/-}$ mutant relative to wild type animals. The + and - indicate the direction of the change, i.e. - for reduced level in ERCC1$^{d/-}$ samples, + for increased level in ERCC1$^{d/-}$ samples compared to wild type.

[b] unpaired t-test using a Benjamini–Hochberg correction for the p-values

DISCUSSION

ERCC1$^{d/-}$ mutant mice represent a model which can provide insight into the biological processes involved in aging.(9) While phenotypic changes that occur in these animals have been characterized before, this is the first study in which the global profiling on the metabolite level was done by NMR.

For both serum and urine, PCA was performed as a first step of data analysis to evaluate the structure present in the data. Unsupervised analysis of ^{1}H NMR serum spectra showed that the major variance in the data matrix reflects two biological phenomena: animal genotype (Figure 1A) and their lifetime (Figure 1B); both are represented with the first two principal components covering together more than 65% of the variation. To explore the effects of those phenomena on metabolic profiles of the body fluids of the animals, supervised methods such as PLS-DA, for example, were employed.

Batch PLS analysis was chosen to explore more in-depth age-related changes. This analysis as well as PCA showed that 8 weeks old mice are considerably different from mice at older age, and this is true for both wild type and mutant mice. This might be due to the fact that at around 10 weeks the maturity of mice sets in and this is expected to involve serious changes in the overall metabolism which are reflected in the NMR spectra. It was also found that the difference between the sample groups is more prominent at older ages, while at younger ages metabolic profiles of mutant and wild type animals are more similar. This fact indicates that ERCC1$^{d/-}$ mutant animals develop more or less normally until the point of a sexual maturity, but begin to exhibit accelerated aging after reaching maturity and hence are a very good model for senescence and biological aging.

The effect of genotype on metabolic composition of serum was studied by PLS-DA which revealed a number of compounds altered between the groups.

Most of the differences in serum between wild type and mutant animals were found to be associated with lipids, either increased or decreased in mutants compared to wild type mice. It has been hypothesized that fast aging mice, which have disruptions in the NER pathway, might have the corresponding adaptive "survival" response of the organism similar to that of caloric restriction.(23;24) In this respect, analysis of the relative changes in lipid and lipoproteins might be of special interest, because in caloric restriction a specific pattern of changes in lipoprotein composition of blood has been shown.(25) Line shapes in the lipoprotein region in diffusion-edited spectra indicate that in ERCC1$^{d/-}$ mutant mice LDL and VLDL are decreased and HDL is increased, and this pattern of changes indeed resembles the state of caloric restriction.

Both the conventional method and NMR-based quantification showed that glucose is decreased in serum of ERCC1$^{d/-}$ mutants compared to wild type (Supplementary Materials, Figure S5). This is another indication for a phenotype that resembles caloric restriction. Low levels of another compound derived from the glucose metabolism, lactate, were observed in serum samples of ERCC1$^{d/-}$ mutants compared to controls; its decrease might be related to a decrease in the Cori cycle.

It is obviously of particular interest to see how the changes in serum are comparable with alterations in urine composition, if any of the biochemical processes reflected in one of the biofluids can as well be seen in the other. Therefore, we performed analysis of urine from a smaller cohort to investigate the involvement of the biochemical processes obtained by serum analysis.

In contrast to serum, glucose and lactate were found to be elevated in urine of mutants compared to wild type animals. Compounds of the TCA cycle: citrate, succinate, and 2-oxoglutarate - also showed higher levels in urine of ERCC1$^{d/-}$ mice compared to wild type animals. These compounds are involved in a large number of biochemical processes, and their alterations are difficult to interpret as they might occur due to a variety of reasons.(26) One of the possible explanations for the observed opposite changes of glucose metabolites in blood and urine is the kidney dysfunction, which leads to an impaired reabsorption of these molecules in renal tubules.(27)

In urine, there is an indication of the altered energy metabolism as well; it was found to be switched to fatty-acids utilization, shown by the presence of 3-hydroxybutyrate, one of the ketone bodies. This compound was found present only in ERCC1$^{d/-}$ mice while in wild type animals there is another unidentified compound present with a singlet in the same region as the doublet of 3-hydroxybutyrate (Supplementary Materials, Figure S6). This would explain the low fold-change value in Table 2 for 3-hydroxybutyrate as the values are calculated for integrated intensities in the binned region. Together with low glucose in blood, the presence of 3-hydroxy-butyrate is an indication of ketosis.(28) in the mutant mice. It is important to note that no difference in food intake relative to the body weight was observed in ERCC1$^{d/-}$ mutants compared to wild type controls. This means that the switch to ketosis in these animals occurs not as a response to food deficiency.

Other compounds observed that might be changing due to ketosis are alpha-keto acids (2-oxo-3-methylbutanoic, and 2-oxo-3-methylpentanoic acids), which can be used in liver as a source of energy and ketoleucine that can also be utilized in the liver, resulting in the production of ketone bodies.(29)

However, the comparison of our findings with previous studies on caloric restriction(30;31) reveals not only the resemblance but some important differences such as dissimilar changes in lactate, hippurate, succinate, and other compounds. Thus, metabolic phenotype of ERCC1$^{d/-}$ mice cannot be reduced to the caloric restriction but reflects a complex, systemic effect of a mutation.

The significantly increased level of citrate in urine of ERCC1$^{d/-}$ mice might point to metabolic alkalosis in these mice.(32) This supposition is strengthened by the fact that

glutamine, which is also related to maintaining acid-base balance,(33) was also found to be elevated in urine.

Kidney malfunction may be the reason for the decrease of hippuric acid secretion, which normally occurs through renal tubules. The decrease of taurine in urine, an important component of bile, as well as the decrease of allantoin in urine may reflect liver dysfunction.(34) Hepatic dysfunction might also be indicated by increased alanine in blood. Although pronounced kidney and liver dysfunctions develop in ERCC1$^{d/-}$ mice at a much older age (after 30 weeks), these compounds found in urine point at attenuated function of these organs already at an earlier age.

In conclusion, besides changes associated with malfunctions of some organs, the NMR data indicated that in ERCC1$^{d/-}$ mice a specific "survival" response is activated that primarily alters energy metabolism and leads to ketosis. These results are in line with the previous observations for the double knockout of ERCC1 gene in comparison to caloric restriction that showed almost identical changes in transcription and biochemistry mediated by insulin pathway.(8)

CONCLUSIONS

Using profiling by ^{1}H NMR and subsequent multivariate statistical analysis, differences in metabolic composition of both serum and urine between wild type and ERCC1$^{d/-}$ mice were found. Dependence of the profiles on age was clearly present, showing that a major change happened between 8 and 12 weeks in both of the genetically different classes of animals, which most probably reflects the time of their sexual maturity.

Differences in molecular composition assessed by NMR in serum and urine indicate a relative change of lipoproteins (decrease in LDL and VLDL, increase in HDL in mutants compared to controls), a shift of the energy metabolism to ketosis, as well as kidney and liver malfunction and possibly metabolic alkalosis in mutant mice.

REFERENCES

1. Kirkwood,T.B., and Austad,S.N. 2000. Why do we age? Nature 408:233-238.

2. Droge,W. 2002. Free radicals in the physiological control of cell function. Physiological Reviews 82:47-95.

3. Beckman,K.B., and Ames,B.N. 1998. The free radical theory of aging matures. Physiological Reviews 78:547-581.

4. Lehmann,A.R. 2003. DNA repair-deficient diseases, xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy. Biochimie 85:1101-1111.

5. Guarente,L., and Kenyon,C. 2000. Genetic pathways that regulate ageing in model organisms. Nature 408:255-262.

6. McWhir,J., Selfridge,J., Harrison,D.J., Squires,S., and Melton,D.W. 1993. Mice with DNA repair gene (ERCC-1) deficiency have elevated levels of p53, liver nuclear abnormalities and die before weaning. Nat. Genet. 5:217-224.

7. Niedernhofer,L.J., Odijk,H., Budzowska,M., van,D.E., Maas,A., Theil,A.F., de,W.J., Jaspers,N.G., Beverloo,H.B., Hoeijmakers,J.H. *et al* 2004. The structure-specific endonuclease Ercc1-Xpf is required to resolve DNA interstrand cross-link-induced double-strand breaks. Mol. Cell Biol. 24:5776-5787.

8. Niedernhofer,L.J., Garinis,G.A., Raams,A., Lalai,A.S., Robinson,A.R., Appeldoorn,E., Odijk,H., Oostendorp,R., Ahmad,A., van,L.W. *et al* 2006. A new progeroid syndrome reveals that genotoxic stress suppresses the somatotroph axis. Nature 444:1038-1043.

9. Weeda,G., Donker,I., deWit,J., Morreau,H., Janssens,R., Vissers,C.J., Nigg,A., vanSteeg,H., Bootsma,D., and Hoeijmakers,J.H.J. 1997. Disruption of mouse ERCC1 results in a novel repair syndrome with growth failure, nuclear abnormalities and senescence. Current Biology 7:427-439.

10. Dolle,M.E., Busuttil,R.A., Garcia,A.M., Wijnhoven,S., van,D.E., Niedernhofer,L.J., van der,H.G., Hoeijmakers,J.H., van,S.H., and Vijg,J. 2006. Increased genomic instability is not a prerequisite for shortened lifespan in DNA repair deficient mice. Mutat. Res. 596:22-35.

11. Fiehn,O. 2002. Metabolomics--the link between genotypes and phenotypes. Plant Mol. Biol. 48:155-171.

12. Holmes,E., Wilson,I.D., and Nicholson,J.K. 2008. Metabolic phenotyping in health and disease. Cell 134:714-717.

13. Lindon,J.C., and Nicholson,J.K. 2008. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trac-Trends in Analytical Chemistry 27:194-204.

14. Nicholson,J.K., and Lindon,J.C. 2008. Systems biology: Metabonomics. Nature 455:1054-1056.

15. Trygg,J., Holmes,E., and Lundstedt,T. 2007. Chemometrics in metabonomics. Journal of Proteome Research 6:469-479.

16. Nicholson,J.K., Lindon,J.C., and Holmes,E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29:1181-1189.

17. Findeisen,M., Brand,T., and Berger,S. 2007. A ^{1}H-NMR thermometer suitable for cryoprobes. Magn Reson. Chem 45:175-178.

18. Kumar,A., Ernst,R.R., and Wuthrich,K. 1980. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochem. Biophys. Res. Commun. 95:1-6.

19. Wu,D.H., Chen,A.D., and Johnson,C.S. 1995. An Improved Diffusion-Ordered Spectroscopy Experiment Incorporating Bipolar-Gradient Pulses. Journal of Magnetic Resonance Series A 115:260-264.

20. Cloarec,O., Dumas,M.E., Craig,A., Barton,R.H., Trygg,J., Hudson,J., Blancher,C., Gauguier,D., Lindon,J.C., Holmes,E. *et al* 2005. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. Analytical Chemistry 77:1282-1289.

21. Nielsen,N.P.V., Carstensen,J.M., and Smedsgaard,J. 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. Journal of Chromatography A 805:17-35.

22. Lindon,J.C., Nicholson,J.K., and Everett,J.R. 1999. NMR spectroscopy of biofluids. Annual Reports on Nmr Spectroscopy, Vol 38 38:1-88.

23. Schumacher,B., van,d.P., I, Moorhouse,M.J., Kosteas,T., Robinson,A.R., Suh,Y., Breit,T.M., van,S.H., Niedernhofer,L.J., van,I.W. *et al* 2008.

Delayed and accelerated aging share common longevity assurance mechanisms. PLoS. Genet. 4:e1000161.

24. van de Ven,M., Andressoo,J.O., Holcomb,V.B., von Lindern,M., Jong,W.M., De Zeeuw,C.I., Suh,Y., Hasty,P., Hoeijmakers,J.H., van der Horst,G.T. *et al* 2006. Adaptive stress response in segmental progeria resembles long-lived dwarfism and calorie restriction in mice. PLoS Genet. 2:e192.

25. Anderson,R.M., Shanmuganayagam,D., and Weindruch,R. 2009. Caloric restriction and aging: studies in mice and monkeys. Toxicol. Pathol. 37:47-51.

26. Robertson,D.G. 2005. Metabonomics in toxicology: A review. Toxicological Sciences 85:809-822.

27. Nicholson,J.K., Timbrell,J.A., and Sadler,P.J. 1985. Proton Nmr-Spectra of Urine As Indicators of Renal Damage - Mercury-Induced Nephrotoxicity in Rats. Molecular Pharmacology 27:644-651.

28. Mcgarry,J.D., and Foster,D.W. 1980. Regulation of Hepatic Fatty-Acid Oxidation and Ketone-Body Production. Annual Review of Biochemistry 49:395-420.

29. Harper,A.E., Miller,R.H., and Block,K.P. 1984. Branched-Chain Amino-Acid-Metabolism. Annual Review of Nutrition 4:409-454.

30. Rezzi,S., Martin,F.P., Shanmuganayagam,D., Colman,R.J., Nicholson,J.K., and Weindruch,R. 2009. Metabolic shifts due to long-term caloric restriction revealed in nonhuman primates. Exp. Gerontol. 44:356-362.

31. Wang,Y.L., Lawler,D., Larson,B., Ramadan,Z., Kochhar,S., Holmes,E., and Nicholson,J.K. 2007. Metabonomic investigations of aging and caloric restriction in a life-long dog study. Journal of Proteome Research 6:1846-1854.

32. Gordon,E.E. 1963. Effect of Acute Metabolic Acidosis and Alkalosis on Acetate Anc Citrate Metabolism in Rat. Journal of Clinical Investigation 42:137-&.

33. Taylor,L., and Curthoys,N.P. 2004. Glutamine metabolism - Role in acid-base balance. Biochemistry and Molecular Biology Education 32:291-304.

34. Wishart,D.S. 2005. Metabolomics: The principles and potential applications to transplantation. American Journal of Transplantation 5:2814-2820.

35. Kriat,M., Confort-Gouny,S., Vion-Dury,J., Sciaky,M., Viout,P., and Cozzone,P.J. 1992. Quantitation of metabolites in human blood serum by proton magnetic resonance spectroscopy. A comparative study of the use of formate and TSP as concentration standards. NMR Biomed. 5:179-184.

SUPPLEMENTARY MATERIALS

Table S1. Mean body weight, g (SD) of mice with different genetic background and at different ages.

Age, wks \ Genotype	Wild type	ERCC1d/-
8	23.7(4.1)	14.7(1.7)
12	25.7(5.3)	14.1(1.4)
16	26.0(6.1)	14.3(1.8)
20	30.3(7.5)	12.9(1.4)

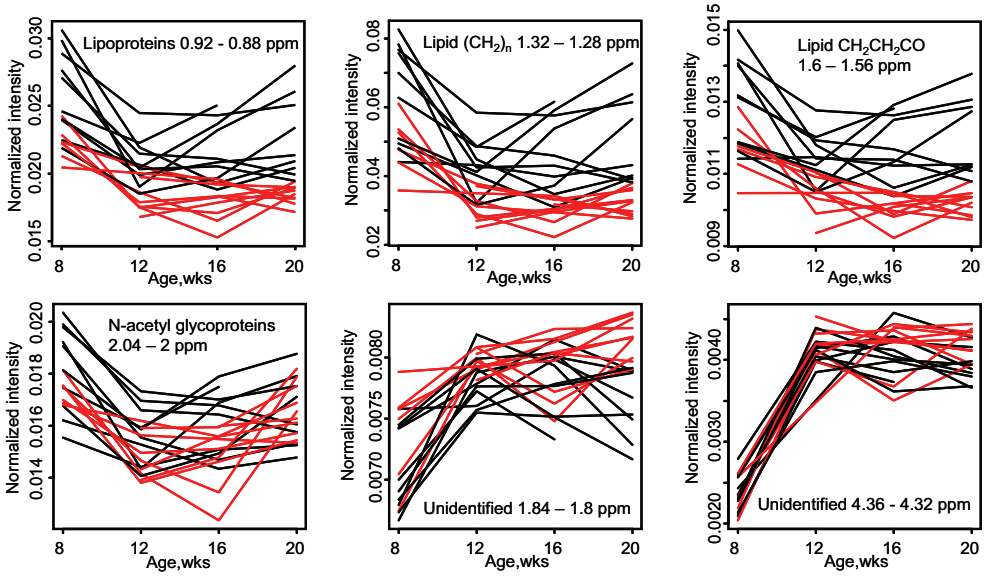

Figure S1. Metabolic compounds of serum related to age and their changes with age. Black lines represent changes in intensity of selected variables in wild type, red in ERCC1$^{d/-}$ mutants.

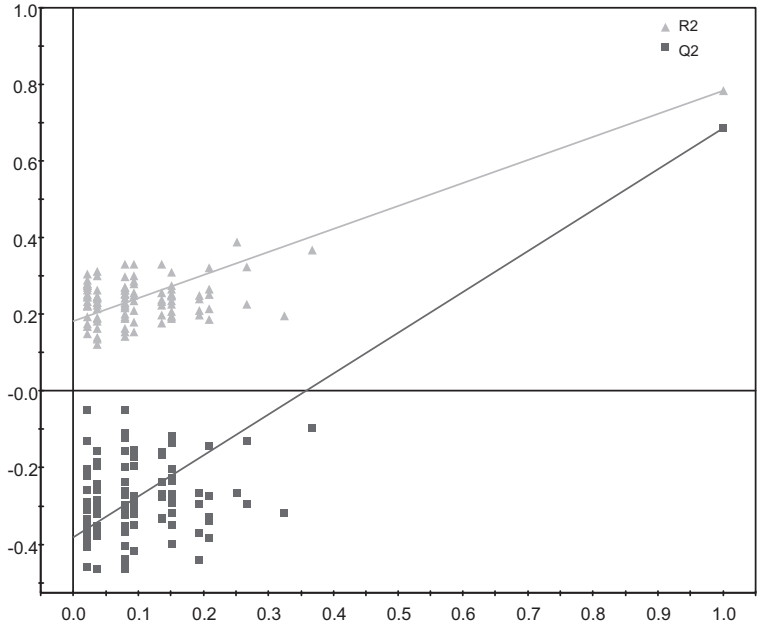

Figure S2. Validation plot of PLS-DA model for mouse serum NMR data with 100 permutations; squares show the Q^2 values and triangles are R^2 values.

Absolute Glucose Quantification. In general TSP and DSS are not suitable as reference for absolute metabolite quantification in serum and plasma due to their interaction with serum albumin which results in exchange broadening of the reference signal(35). However, for relatively similar plasma/serum samples (like metabonomics samples from individuals/animals collected under controlled conditions) a correction factor for the integral of the TSP signal can be calculated from comparison of samples with known TSP concentration in the presence and absence of serum/plasma. In first approximation assuming similar relaxation behavior between the metabolite in serum and free TSP, the metabolite concentration can then directly be calculated from the corrected TSP integrals in serum/plasma.

For determination of the TSP correction factor an additional sample of 150 L pooled rat serum with addition of 150 µl of 1.5 M phosphate buffer in H_2O/D_2O (90/10) at pH 7.4 containing 4% NaN_3 and 2 mM TSP was prepared. As reference for free TSP a second sample was prepared by adding 150 µl of 1.5 M phosphate buffer in H_2O/D_2O (90 /10) at pH 7.4 containing 4% NaN_3 and 2 mM TSP to 150 µl of 0.9% NaCl saline solution. Both samples were measured in 3mm NMR Match tubes using the first increment of a NOESY pulse sequence as described in the experimental section.

For quantification the deconvoluted TSP signals were used in order to avoid any interference with overlapping broad protein resonances in the case of the serum spectrum. The region between 0.2 and -0.2 ppm was baseline corrected automatically by subtraction of a 1st order polynomial removing any broad signals from proteins in this region. The TSP signal was then deconvoluted by fitting a mixed Lorentzian/Gaussian function (60/40) to the peak using the build-in MDCON command in Topspin (Version 2.1 pl4, Bruker Biospin). The parameters were adjusted for peak position and half width of the corresponding signal. The obtained deconvoluted signals were then quantified based on the area under the curve as determined by the MDCON algorithm. A correction factor of 1.93 for TSP in serum compared to saline was determined after correction for differences in nc_proc (Supplementary Materiaks, Figure S3).

Based on the determined correction factor, the absolute Glucose concentrations were calculated using the following formula:

$$c_G = \frac{I_G * C_T * H_T}{1.93 * I_T * H_G}$$

I_G : Integral -anomeric proton of Glucose

H_G: Number of protons of deconvoluted Glucose signal (0.36H assuming anomeric equilibrium – note only anomeric proton is used for quantification)

I_T: Integral TSP signal

H_T: Number of protons of deconvoluted TSP signal (9H)

C_T: Concentration of TSP in sample

Glucose concentrations in serum were calculated taking into account 4-time dilution of the serum sample with buffer .

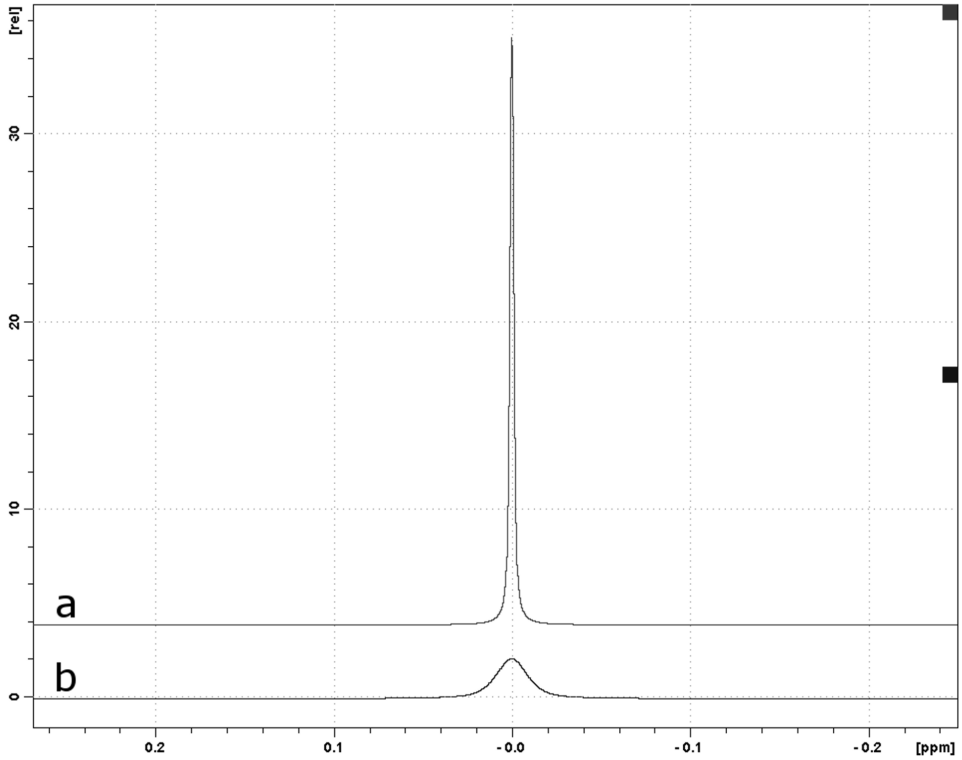

Figure S3. Comparison of deconvoluted TSP signal from (a) 1mM TSP in Saline/Buffer solution (line width = 1.5 Hz / integral = 378.5) and (b) 1 mM TSP in Serum/Buffer solution (line width = 12.4 Hz / integral = 196.0).

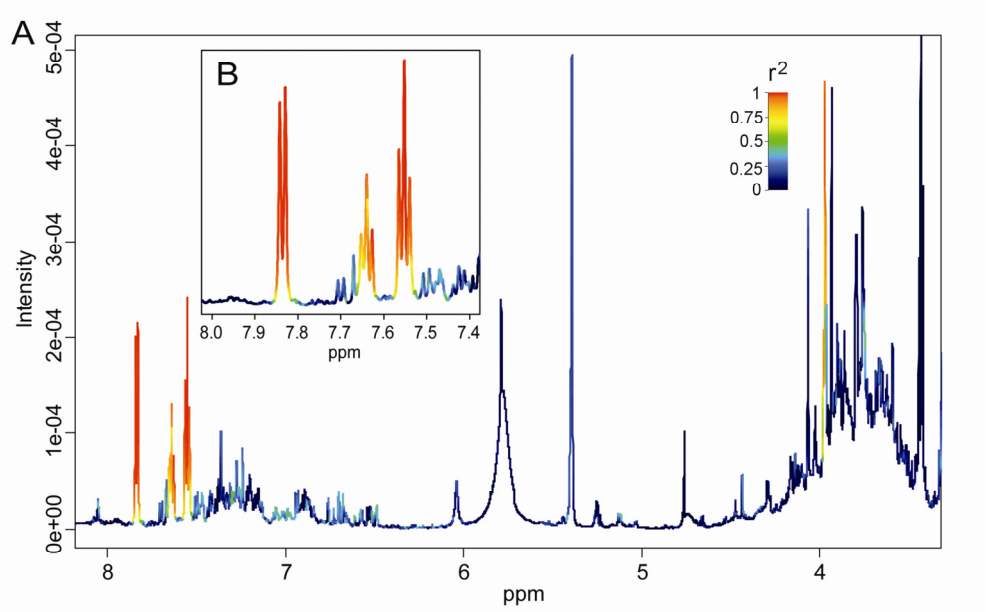

Figure S4. One-dimensional STOCSY analysis for the selected variable from the urine spectrum corresponding to 7.55 ppm. The degree of correlation across the spectrum has been color coded and projected on the spectrum that has the maximum for this variable. A) spectrum between 8 and 3.5 ppm B) zoomed in region between 8 and 7.4 ppm.

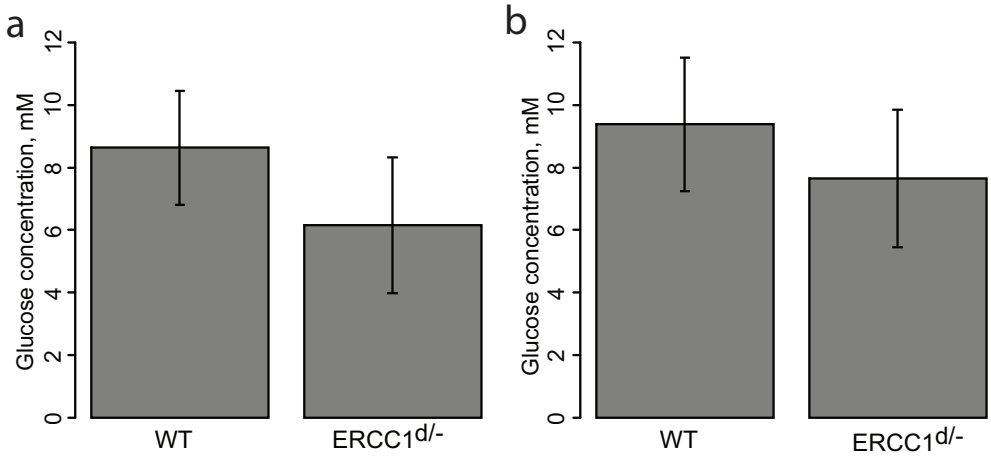

Figure S5. Glucose concentration in blood (a) measured by a conventional method (Freestyle mini blood glucose measurement device) (b) quantified based on NMR spectra. Error bars indicate standard deviations.

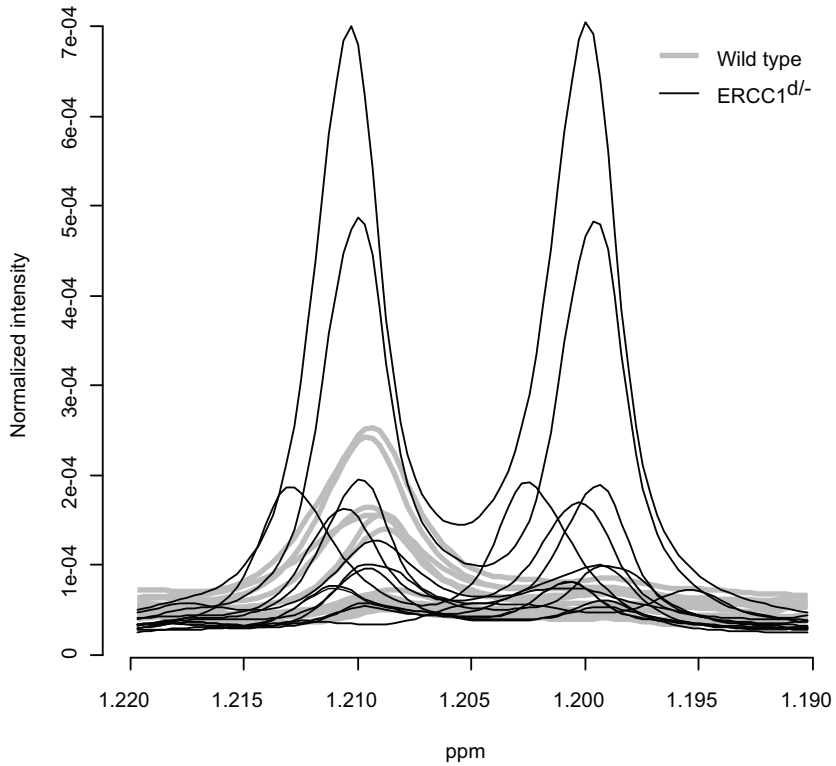

Figure S6. Spectral region of 3-hydroxybutyric acid doublet in urine samples, spectra from wild type animals are represented by black lines and from ERCC1$^{d/-}$ mutants by red lines.

Part III

Application to human studies

Chapter 5

Integrating study design and clinical data into metabolic profiling of urinary tract infection

Nevedomskaya E., Pacchiarotta T., Artemov A., Meissner A., van Nieuwkoop C., van Dissel J.T., Mayboroda O.A., Deelder A.M.

Manuscript in preparation

ABSTRACT

Urinary Tract Infection (UTI) encompasses a variety of clinical syndromes that can range from mild to life-threatening conditions. As such, it represents an interesting model for the development of an analytically based scoring system of disease severity and/or host response. Here we test the feasibility of this concept using ^{1}H NMR based metabolomics as the analytical platform. Using an exhaustively clinically characterized cohort and taking advantage of the multi-level study design, which opens possibilities for case-control and longitudinal modeling, we were able to identify molecular discriminators that characterize UTI patients. Moreover, we show that using such a design allows not only a better validation of the statistical models, but also helps dissecting various biological processes and, most importantly, significantly improves biological interpretation of the obtained results.

INTRODUCTION

Despite the progress made in understanding the mechanistic basis of many diseases in the last century, medicine is still essentially "more an art than a science".(1) Specific and sensitive biological markers are important contributors to the improved diagnostic methods as well as to patient care and drug discovery. Advanced "-omics" technologies, such as genomics, proteomics and metabolomics, enable identification of such markers. Of our particular interest is metabolomics that focuses on the analysis of metabolites present in biological fluids. Metabolites are end-points of all the biochemical processes of the organism and thus their collection – the metabolome is the closest approximation of the physiological phenotype and as such has a great potential for uncovering the biology underlying diseases and providing valuable markers of pathology.(2;3)

The biological interpretation of results from metabolomics studies is rather complex and still in an early phase of development(4). The human body is a "super-organism" that unites its own network of interconnected tissues and organs with multiple colonies of microorganisms.(5) Interpretation of changes in concentration of metabolites found in biological fluids can readily be performed based on the underlying metabolic pathway; however, it is not always possible to link the observed change in systemic metabolite concentrations to a specific tissue or organ.(6) Especially in the case of disruption of highly abundant metabolites, *e.g.* from energy or amino acid metabolism, additional information would be required in order to interpret the data in respect to the tissue of origin. In addition, a change of such metabolites does not always improve the knowledge about the underlying cellular mechanisms and biology. A way to facilitate the interpretation of clinical metabolomics data is to integrate a plethora of available clinical parameters and to utilize a multilevel study design that should provide the opportunity to access the various levels of biological processes.

One of the examples of a complex and heterogeneous clinical entity, for which current diagnostic methods are not straightforward, is Urinary Tract Infection (UTI)(7). Clinical manifestations of UTI can cover the range from mild cystitis to advanced pyelonephritis potentially leading to urosepsis and multiple-organ failure. Physical symptoms may vary from patient to patient and be similar to a number of other diseases, mainly of infectious origin. Thus, the presence of bacteria and leucocytes in urine can not be considered as a sole common denominator for UTI and even if it was, the criterion for the colony count is variable and anyway considered insensitive(8). The correct and timely diagnosis relies on effective joint work of clinicians and microbiologists(8). All of this explains the considerable interest in providing new, specific and sensitive markers for UTI and for the uropathogen involved. The focus of the available metabolomics studies on UTI in the literature has so far

been on the identification of pathogens: in the work of Gupta *et al.* a beautiful method with the use of ^{1}H NMR was proposed.(9-11) However, regrettably the method is not quantitative nor does it provide any information about the localization of the infection within the urinary tract, morbidity and preferred strategy of treatment.

In the current study we investigated possibilities of using urinary metabolic profiles to monitor the health state of UTI patients, the degree of infection and the recovery process of UTI patients in the context of febrile, complicated UTI. We used a selection of samples from an exhaustively characterized cohort, with multiple urine samples available per individual and with the main pathogen identified as *Escherichia coli*, which is the most common pathogen for UTI. Samples from a group of age- and gender- matched UTI symptom-free subjects were included as control. The longitudinal design allowed studying various biological processes: not only the difference between the patients and controls, but also the recovery process, using each patient as its own control.

MATERIALS AND METHODS

Samples. The study protocol was approved by the ethical committee of the Leiden University Medical Center and all included patients gave written informed consent.

Urine samples were collected at the Emergency Department and Primary Care Department. The sampling was carried out at several time points: the first urine samples were collected at the day of enrolment as baseline samples (t=0). Clean midstream-catch urine cultures were obtained and were analyzed using local standard microbiological methods. Three-four (t=4) and thirty days (t=30) after the day of enrolment, urine samples of the same patients were collected and new bacterial culture tests were performed (Supplementary Materials, Figure S1).

For the current study, from a database of about 700 subjects enrolled, 40 subjects, for which urine culture confirmed *E.coli*-positive complicated febrile urinary tract infection that recovered after antibiotic treatment, were selected. Samples from age- and gender-matched subjects with low bacterial culture in urine and without evidence of inflammatory diseases were used as controls (Table 1). A number of samples were missing, a few removed from the analysis due to either insufficient spectra quality or high glucose content (Supplementary Materials, Figure S1). In the end the study included four classes of samples originating from UTI symptom-free (N = 35) at day 0 (baseline control), UTI patients (N = 32) at day 0 (baseline), UTI patients (N= 29) at day 4 and UTI patients after recovery from infection (N = 37) at day 30 (Supplementary Materials, Figure S1).

Table 1. Characteristics of the studied patients and controls groups at baseline (t=0).

Characteristics		UTI patients n = 40	Controls n = 40	p
Age, years, median (sd)		59 (14.6)	58 (17.9)	0.9
Female, n (%)		22 (55)	22 (55)	1
Smoking, n (%)		5 (12)	5 (12)	1
Co-morbidity, n (%)				
	Urinary tract disorder	4 (10)	4 (10)	1
	Malignancy	4 (10)	1 (3)	0.17
	Heart failure	5 (13)	3 (8)	0.46
	Renal insufficiency	1 (4)	0 (0)	0.13
	Diabetes mellitus	6 (15)	2 (5)	0.14
	Immunocompromised	1 (3)	1 (3)	1
Urine dipstick results				
	Nitrate	26/37 (75)*	0/37 (0)*	< 0.001
	Leucocyte esterase	35/37 (95)*	5/37 (14)*	< 0.001

* 3 missing values

Sample preparation. Samples were thawed, transferred into 96 deep-well plates and centrifuged at 3000g for 15 minutes at 4°C to remove any precipitate. For sample preparation 520 µL urine were mixed with 60 µL of pH 7.0 phosphate buffer (1.5 M) in 100% D_2O containing 4 mM sodium 3-trimethylsilyl-tetradeuteriopropionate (TSP) and 2mM NaN3 in a 96 deep-well plate using a Gilson 215 liquid handler controlled by a Bruker Sample Track LIMS system (Bruker BioSpin, Karlsruhe, Germany).

NMR experiments and processing. ^{1}H NMR data were collected using a Bruker 600 MHz AVANCE II spectrometer equipped with a 5 mm TCI cryogenic probehead and a z-gradient system; a Bruker BEST (Bruker Efficient Sample Transfer) system was used in combination with a 120 µL CryoFIT™ flow insert for sample transfer. One-dimensional (1D) ^{1}H NMR spectra were recorded at 300 K using the first increment of a NOESY pulse sequence(12) with presaturation (γB_1=50 Hz) during a relaxation delay of 4 s and a mixing time of 10 ms for efficient water suppression(13). Eight scans of 65,536 points covering 12,335 Hz were recorded and zero filled to 65,536 complex points prior to Fourier transformation, an exponential window function was applied with a line-broadening factor of 1.0 Hz. The spectra were manually phase and baseline corrected and automatically

referenced to the internal standard (TSP = 0.0 ppm). Phase offset artifacts of the residual water resonance were manually corrected using a polynomial of degree 5 least square fit filtering of the free induction decay (FID) (14). In order to monitor proper filling of the NMR flow cell and for quality control 1D gradient profiles (15) along the z-axis were recorded for each sample prior and post data acquisition. Duration of 90 degree pulses were automatically calibrated for each individual sample using a homonuclear-gated nutation experiment(16) on the locked and shimmed samples after automatic tuning and matching of the probe head.

Statistical analysis. Each spectrum was integrated (binned) using 0.014 ppm integral regions between 10 and 1 ppm, the residual water and urea region between 6 and 4.5 ppm was excluded, resulting in 550 data points used for the analysis. To account for any difference in concentration between the samples, each spectrum was normalized to a total area of 1. Absolute values were log-transforsmed. All pre-processing was done using in-house developed routines in R statistical environment (http://www.r-project.org/). Variables were centered and unit variance scaled prior to statistical analysis in SIMCA-P+ (version 12.0; Umetrics, Sweden) software package. For initial analysis and outlier detection, principal component analysis (PCA) was performed using 10 components. After the initial PCA analysis the following regions corresponding to paracetamol and its metabolites were excluded from the analysis: 7.5 – 6.75, 3.95 – 3.8, 3.7 – 3.45, 2.2 – 2.14 and 1.84-1.88 ppm according to (17). For partial least squares-discriminant analysis (PLS-DA) (18) samples were categorized based on classes as defined by the study design. PLS model was built using 5 categories according to logarithm of bacterial count as a Y variable. Statistical models from supervised multivariate data analysis were validated by random permutation of the response variable and comparison of the goodness of fit (R^2Y and Q^2) (19;20). For random permutation tests 100 models were calculated and the goodness of fit was compared with the original model in a validation plot. Spectral regions responsible for the separation between classes in supervised models were identified based on the Variable Influence on Projection (VIP) values, which correspond to the importance of the variables (bins) for the model. The variables with a VIP value larger than 1.8 were considered significant and used for further analysis and identification of the responsible peak(s) within the spectrum. Prediction of class membership of samples by PLS-DA model was based on the predicted Y variable with the cut-off of 0.5.

For multilevel components analysis (MCA) using an in-house developed script in R as described by Jansen *et al.*(21) data were not log-transformed.

Univariate tests were performed to assess the statistical significance of the spectroscopic regions found using multivariate analysis: unpaired t-test was performed for the regions

found as discriminating between UTI patients and controls by PLS-DA; ANOVA was performed on the regions that showed association with bacterial count in PLS; paired t-test was carried out on the regions identified in multilevel analysis. All the corresponding p-values were adjusted for multiple testing using Benjamini-Hochberg correction.

Identification of compounds of interest. Annotation of identified peaks was performed based on reference spectra from the Bruker Bioref database and in-house reference data. Confident identification was facilitated by the use of Statistical Total Correlation SpectroscopY method (STOCSY)(22).

Quantification of paracetamol. Quantification was performed by deconvolution and subsequent integration of paracetamol-glucuronide resonance at 5.10 ppm (d, 7.1 Hz) using an in-house developed automation routine. The absolute concentrations were calculated based on internal reference TSP. Values were not corrected for differential attenuation of the signals caused by relaxation during the mixing time and rapid-pulsing saturation effects.

RESULTS

The initial PCA on baseline samples revealed a trend in separation between UTI patients and controls in the scores plot of the first two principal components as shown in Figure 1A. The loadings plot of this model was dominated by the spectral regions that belonged to one of the most commonly used over-the-counter analgesic, paracetamol (Supplementary Materials, Figure S2). The absolute concentration of paracetamol-glucuronide was used to stratify samples in the PCA plot: the direction of increase of paracetamol-glucuronide was found to match the direction of controls-patients separation (Figure 1B). As paracetamol is not an infection or morbidity marker, the further analysis was performed after the exclusion of the regions corresponding to the drug and its metabolites.

The PCA analysis of the baseline samples after the removal of spectral regions of paracetamol and its metabolites did not show separation between UTI patients and controls within the scores plot of the first two principal components; however, a clear trend was identified along the third principal component (Figure 2), which means that inter-individual variability is to a certain extent more prominent than the disease effect. No outliers were detected based on distance to the model (DModX).

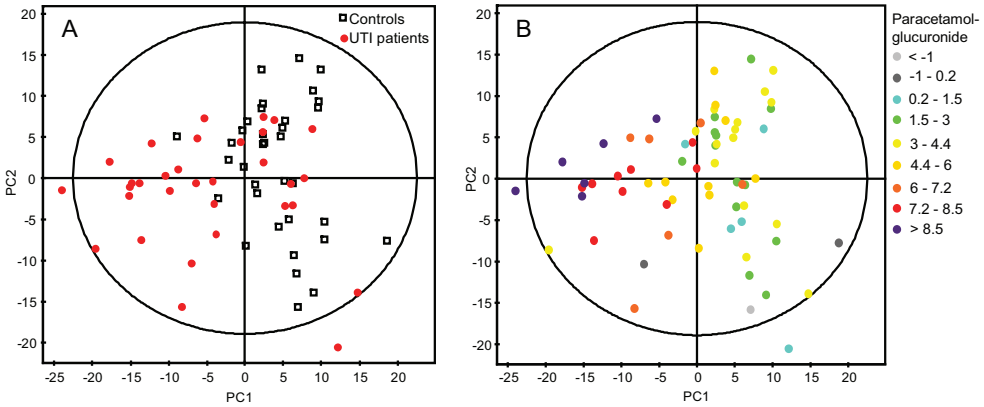

Figure 1. PCA scores plot of ^{1}H NMR data from controls and UTI patients urine samples at baseline, first two principal components covering 14.5 and 10.2% of variation respectively. (A) Colored according to controls (□) and UTI patients (●). (B) Colored according to the logarithm of absolute concentration of paracetamol-glucuronide.

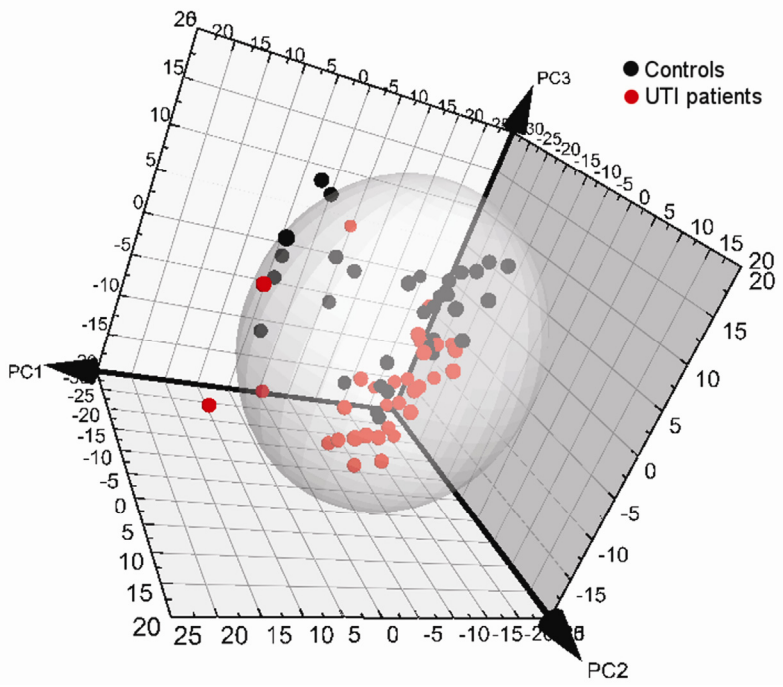

Figure 2. PCA scores plots of ^{1}H NMR data from controls (black) and UTI patients (red) urine samples at baseline after removal of the regions corresponding to paracetamol and its metabolites. First principal component covers 11.7%, second 11.2% and third 9.8% of variation.

In the next step a supervised PLS-DA model was built for t=0 using UTI/controls as a response variable. In the scores plot of the resulting model the two groups were well separated (Figure 3). Cumulative explained variance (R2Y) of 0.88 and cross validated predictive fraction (Q2) of 0.63 were calculated for the model; the model validation plot showed intercepts of the R2Y and Q2 regression lines with the vertical axis at 0.63 and -0.11, respectively, indicating a valid model. Molecular discriminators were identified based on relevant regions as identified by the corresponding VIP. A list of those regions, along with the p-values based on t-test (corrected for multiple testing), the direction of change and identities of the corresponding metabolites are summarized in Table 2.

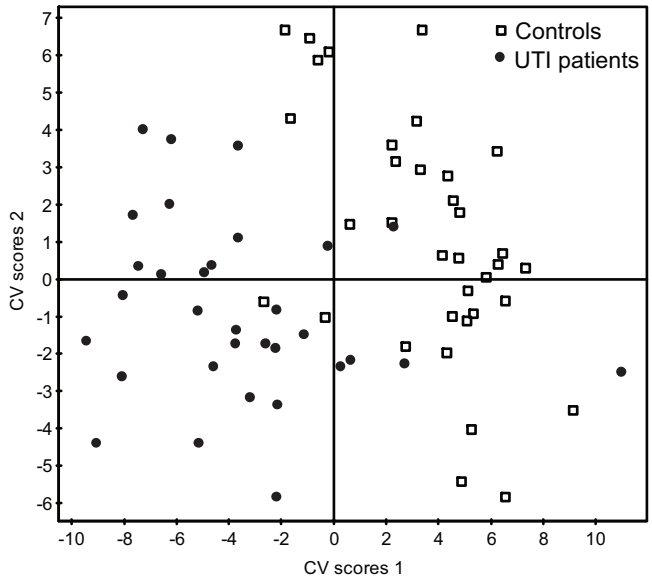

Figure 3. Cross-validated PLS-DA scores plot of urine ^{1}H NMR spectra of controls (□) and UTI patients at baseline (●), R2Y = 0.88, Q2 = 0.63.

The advantage of PLS-based models is that they can easily be used to predict the class membership of new samples. Data of the UTI patients at t=4 were predicted using the two-class PLS-DA model that was built as described above. Of a total of 29 urine samples included in the prediction set, 19 (65.5%) were classified as controls, whereas 10 (34.5%) samples were classified as UTI (Figure 4). Besides using data from the 4-days time point as prediction set, we also performed a separate analysis for the 30-days time point (Figure 4). In this case, out of 37 samples collected, 32 (86.5%) were attributed to the group of controls and 5 (13.5%) were categorized as UTI.

Table 2. Spectroscopic regions that appear as influential in various statistical models and statistical significance of the corresponding univariate tests adjusted for multiple testing using Benjamini-Hochberg method.

		Controls vs. UTI patients[a]		Bacteria concentration[b]		Recovery from t=0 to t=30[c]	
ppm region	Identity	t-test p-value	change	ANOVA p-value	change	paired t-test p-value	change
9.291 - 9.277	1-methylnicotinamide	<0.0001	-	<0.001	-		
9.277 - 9.264	1-methylnicotinamide	<0.01	-				
8.977 - 8.964	1-methylnicotinamide	<0.01	-				
4.491 - 4.477	1-methylnicotinamide	<0.01	-	<0.01	-		
1.941 - 1.927	Acetic acid	<0.01	+	<0.01	+		
1.927 - 1.914	Acetic acid	<0.0001	+	<0.0001	+		
3.196 - 3.182	Acetylcarnitine	<0.01	+				
2.568 - 2.555	Citric acid	<0.01	-				
2.541 - 2.527	Citric acid	<0.01	-				
4.082 - 4.068	Creatinine	0.03	-				
3.073 - 3.059	Creatinine	<0.01	-	0.07	-		
3.059 - 3.045	Creatinine	0.09	-				
7.709 - 7.696	Furoylglycine					<0.01	+
7.696 - 7.682	Furoylglycine	<0.01	-	<0.01	-		
3.959 - 3.946	Glycolic acid derivative	<0.001	-	<0.01	-	<0.0001	+
7.859 -7.846	Hippuric acid	<0.01	-	<0.01	-		
7.668 - 7.655	Hippuric acid	<0.001	-	<0.01	-		
7.655 - 7.641	Hippuric acid	0.01	-	0.02	-		
7.586 - 7.573	Hippuric acid	<0.01	-	0.05	-		
3.973 - 3.959	Hippuric acid	0.01	-	0.03	-		
8.555 - 8.541	Hippuric acid (amide)	<0.01	-				
8.541 - 8.527	Hippuric acid (amide)	<0.001	-	<0.01	-		
1.341 - 1.327	Lactic acid	<0.01	+	<0.01	+		
7.764 - 7.75	Para-aminohippuric					<0.001	+
3.332 - 3.318	Scyllo-inositol					<0.01	+
3.455 - 3.441	Taurine	<0.0001	+	<0.001	+	<0.0001	-
3.441 - 3.427	Taurine	<0.0001	+	<0.001	+	<0.0001	-
3.427 - 3.414	Taurine	<0.0001	+	<0.01	+		
3.264 - 3.250	Taurine	<0.001	+				
8.855 - 8.541	Trigonelline					0.01	+
4.45 - 4.436	Trigonelline					<0.01	+
2.896 -2.881	Trimethylamine	<0.0001	+	<0.0001	+		
8.486 - 8.473	Unknown					<0.01	+
7.968 - 7.955	Unknown					<0.001	+
7.75 - 7.736	Unknown					<0.01	+
7.518 - 7.505	Unknown			<0.01	+		
6.686 - 6.673	Unknown					<0.0001	+
6.509 - 6.496	Unknown					0.04	+
3.168 - 3.155	Unknown			<0.01	-		

[a] two-group t-test for the healthy controls and UTI patients at baseline; positive direction of change corresponds to intensity of the region being higher in UTI patients compared to controls, negative – region intensity is lower in UTI patients compared to controls

[b] ANOVA analysis for the number of bacteria present in urine; direction corresponds to the correlation to the number of bacteria: positive corresponds to the raise of the region intensity with the increase of the number of bacteria, negative - to the decrease of the region intensity with the increase of the number of bacteria

[c] paired t-test for the UTI patients at baseline and 30 days; positive direction of change corresponds to intensity of the region being higher at 30 days compared to baseline, negative – region intensity is lower at 30 days compared to baseline

An important parameter characterizing UTI patients is the number of bacteria in urine; however, bacteria can also be present in urine of the individuals, who do not exhibit any symptoms of UTI(25). We built a PLS regression model from NMR data of urine at baseline using the result of bacterial culture as response variable. Since bacterial count and UTI classification do not fully correlate we expected to obtain a slightly different model as compared to the model built based on UTI classification for this timepoint. Using 2 components a cumulative R2Y = 0.78 and Q2 = 0.44 were obtained and model validation showed intercepts of the R2Y and Q2Y regression lines with the vertical axis at 0.63 and -0.12, respectively, in the model validation plot. As can be seen from the PLS scores plot (Figure 5) the samples with the highest bacteria concentration in urine were very distinct from the rest forming a separate cluster, whereas the rest of the samples were overlapping. The spectral regions responsible for the correlation of the ^{1}H NMR data and bacterial count were chosen on the basis of the corresponding VIP. A list of those regions, along with the p-values derived from ANOVA (corrected for multiple testing), the direction of change and identities of the corresponding metabolites are summarized in Table 2.

To better understand the process of patient recovery and to find the spectroscopic regions that correlate with this process, we took advantage of the longitudinal study design. One of the statistical methods suitable for such analysis is multilevel component analysis (MCA) that separates variation present in the data into two levels: between-individual and within-individual. We performed this analysis on the 29 patients for which both the data from the baseline and from the 30-days time point were available and concentrated on the within-individual information. This should best reflect the recovery from the baseline, when patients are diagnosed as infected, to 30 days, when they are considered UTI symptom-free. PCA scores plot of the first two principle components that cover 15.8 and 14.8% of the variation, respectively, showed good separation between baseline and t=30 time points (data not shown). The PLS-DA model of this data had high quality parameters (R2Y = 0.98, Q2 = 0.78 for four components), performs significantly better then random models ($p<10^{-15}$) and perfectly separated the two time points (data not shown). The NMR spectral regions responsible for the separation between baseline and the t=30 time point were identified based on VIP values. The underlying metabolites as well as the p-values from paired t-test (corrected for multiple testing) and the direction of change are summarized in Table 2.

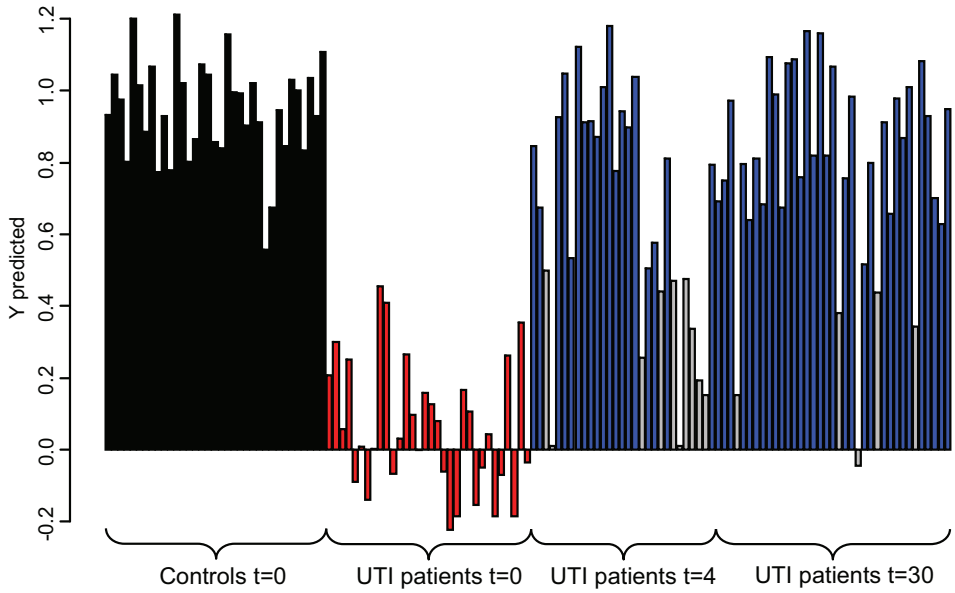

Figure 4. Predicted response value for two-class PLS-DA model based on controls (black bars) and UTI patients (red bars) at baseline: blue bars are the t=4 and t=30 classified as controls, grey are the t=4 and t=30 samples classified as UTI patients at t=0.

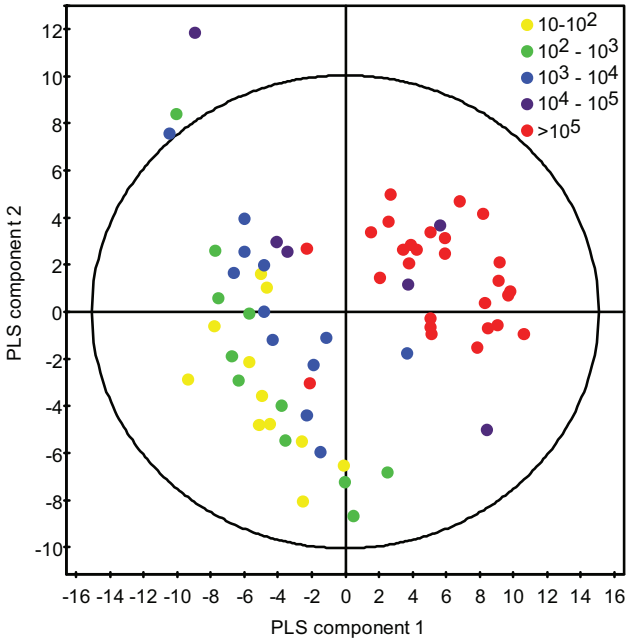

Figure 5. Scores plot of the PLS model of urine ^{1}H NMR spectra at baseline vs. the number of bacteria (CFU/mL) found in urine (R2Y = 0.78, Q2 = 0.44). Colored by the number of bacteria.

DISCUSSION

UTI represents a complex clinical entity, for which diagnostics is not straightforward and based on consensus criteria (7). In the current paper we identified metabolites that characterize UTI and its pathology with the use of ^{1}H NMR. We demonstrate how the use of clinical data and multiple samples per individual can enrich the biological interpretation of the findings. To reduce the heterogeneity typically posed by UTI research, as a first attempt the smaller selection of UTI subjects from a bigger cohort was used, with similar diagnosis and with the major pathogen being *E.coli*. A set of matched controls was also available.

Unlike in animal experiments, in clinical research assigning people to certain groups is not always unconditional. The diagnosis of a disease can be fuzzy and defining the "healthy" group is even more difficult, as there is hardly a definition of healthy. Thus, it may be very advantageous to supplement a traditional "case-control" design with a more complex study design and the use of additional clinical data. When used without extra information, "case-control" analysis might even be misleading. For example, the separation of the control and UTI groups was seen in the first two principal components of PCA; however, this discrimination was not disease-related, but the result of patients taking the antipyretic and analgesic drug paracetamol. An analysis strategy for such type of data is to identify all of the spectroscopic regions that contain signals from drug-related compounds and to exclude them prior to further analysis. However, it is not feasible to account for the whole range of the medication used and, more importantly within the context of clinical metabolomics studies in general, to account for drug-related shifts in metabolism, especially in the case of long-term treatment regimes of chronic conditions. It is essential to consider such effects when developing the study design in order to minimize or control such influences.

Samples from 4 days after admission, when the patients were still under therapy, but on the way to recovery, were used to check if the modeled differences were related to the effect of medication or not. The fact that the majority of those samples were classified as healthy by the model built on baseline samples is an indication that the model is not reflecting therapy/drug intake, but is indeed related to the clinical difference between the groups.

The samples from the 30-days time point, when UTI patients were symptom-free, could also be used to gain additional information on the performance of the model as well as to get insight into the underlying biology. When predicted using the PLS-DA model built on the baseline UTI infected and UTI symptom-free samples, most of the 30-days samples (86.5%) were projected to the control group. Those few, which were still predicted as infected UTI patients, may have another condition (as we do not know at this point how

specific our model is) or have asymptomatic UTI. On the other hand, they can be healthy and be false positives, as the predictive ability of our model, estimated by cross-validation was 63%. Despite that, considering the prediction of 30-days samples as an independent statistical test for our model, it gives very satisfactory results.

Pair-wise analysis for baseline and 30-days samples from the same individuals was conducted in order to monitor the recovery process. It revealed a number of classifiers and improved their statistical significance. The identified metabolites overlapped with the compounds from the model discriminating healthy and UTI subjects, however a few of them were unique (para-aminohippuric acid, scyllo-inositol and a few unidentified compounds).

Besides the multilevel design, the advantage of the current study was the exhaustive clinical characterization of the patients. Among the variety of clinical parameters available, the number of bacteria in urine was of specific importance. We performed regression-based analysis of the relation between the ^{1}H NMR data and the bacterial load in urine as determined by bacterial culture. The classifiers that emerged from this analysis were to a certain extent overlapping with the classifiers derived from the discriminative model on baseline samples. This was no surprise, since UTI is generally characterized by the presence of bacteria in urine.

When comparing the lists of discriminators obtained from the different models (discriminating UTI patients from controls, modeling the recovery process and modeling the data against the degree of bacterial contamination of urine) it is evident that there is a large overlap which makes biological interpretation of the results feasible. For instance, some of the overlapping metabolites were already known from the literature to be related to the bacterial contamination of urine: acetate, lactate and trimethylamine (9). Others, if they were found only in the comparative analysis of the two groups, could be attributed based on previous studies to certain phenomena. Hippuric acid, for example, is often associated with the gut microflora (26) and taurine with liver toxicity (27). However, our findings suggest that they are also associated with the bacterial contamination of urine, which obviously does not mean that they are not related to the mentioned physiological processes as well, but that a complex network of interconnected factors is involved. The metabolites that appear to be related to the recovery process might be considered as potential morbidity markers. One of them, para-aminohippuric acid, is a well-established diagnostic marker for renal plasma flow and glomerular filtration.(28) The recovery from the complicated, tissue-invasive UTI is associated with the resumption of the kidneys' function, so the positive change in para-aminohippuric acid corroborates our assumption that some of the markers discovered in the paired analysis are the markers of morbidity.

CONCLUSIONS

In the current paper we used a metabolomics approach to profile Urinary Tract Infection, which is on the one hand one of the most common infectious diseases among the adults, and on the other hand a disease that still lacks markers of morbidity. Using ^{1}H NMR profiles of urine we generated various statistical models: a) discriminating UTI patients and control subjects, b) following the recovery process of UTI patients and c) associating urine metabolic content with bacterial contamination. The discriminative model was able to classify most of the independent samples correctly according to their diagnosis, which indicates its high predictive ability. Comparing the sets of molecules derived from different analyses, we concluded that some of the compounds (*e.g.* trimethylamine and acetate) can be attributed to the effect of bacterial contamination of urine, others (*e.g.* para-aminohippuric acid, scyllo-inositol) can be considered markers of morbidity.

ACKNOWLEDGEMENTS

The authors would like to thank Sibel Göraler M.Sc. for the analytical work.

REFERENCES

1. Woodcock,J. 2007. The prospects for "personalized medicine" in drug development and drug therapy. Clinical Pharmacology & Therapeutics 81:164-169.

2. Lindon,J.C., Holmes,E., Bollard,M.E., Stanley,E.G., and Nicholson,J.K. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 9:1-31.

3. Holmes,E., Wilson,I.D., and Nicholson,J.K. 2008. Metabolic phenotyping in health and disease. Cell 134:714-717.

4. Mendes,P., Camacho,D., and de la,F.A. 2005. Modelling and simulation for metabolomics data analysis. Biochem. Soc. Trans. 33:1427-1429.

5. Nicholson,J.K., Holmes,E., and Wilson,I.D. 2005. Gut microorganisms, mammalian metabolism and personalized health care. Nat. Rev. Microbiol. 3:431-438.

6. Adourian,A., Jennings,E., Balasubramanian,R., Hines,W.M., Damian,D., Plasterer,T.N., Clish,C.B., Stroobant,P., McBurney,R., Verheij,E.R. *et al* 2008. Correlation network analysis for data integration and biomarker selection. Molecular Biosystems 4:249-259.

7. Wilson,M.L., and Gaido,L. 2004. Laboratory diagnosis of urinary tract infections in adult patients. Clinical Infectious Diseases 38:1150-1158.

8. Johnson,J.R. 2004. Laboratory diagnosis of urinary tract infections in adult patients. Clinical Infectious Diseases 39:873.

9. Gupta,A., Dwivedi,M., Mahdi,A.A., Gowda,G.A., Khetrapal,C.L., and Bhandari,M. 2009. 1H-nuclear magnetic resonance spectroscopy for identifying and quantifying common uropathogens: a metabolic approach to the urinary tract infection. BJU. Int. 104:236-244.

10. Gupta,A., Dwivedi,M., Nagana Gowda,G.A., Ayyagari,A., Mahdi,A.A., Bhandari,M., and Khetrapal,C.L. 2005. (1)H NMR spectroscopy in the diagnosis of Pseudomonas aeruginosa-induced urinary tract infection. NMR Biomed. 18:293-299.

11. Gupta,A., Dwivedi,M., Gowda,G.A., Mahdi,A.A., Jain,A., Ayyagari,A., Roy,R., Bhandari,M., and Khetrapal,C.L. 2006. ^{1}H NMR spectroscopy in the diagnosis of Klebsiella pneumoniae-induced urinary tract infection. NMR Biomed. 19:1055-1061.

12. Kumar,A., Ernst,R.R., and Wuthrich,K. 1980. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochem. Biophys. Res. Commun. 95:1-6.

13. Price,W.S. 1999. Water signal suppression in NMR spectroscopy. Annual Reports on Nmr Spectroscopy, Vol 38 38:289-354.

14. Coron,A., Vanhamme,L., Antoine,J.P., Van,H.P., and Van,H.S. 2001. The filtering approach to solvent peak suppression in MRS: a critical review. J. Magn Reson. 152:26-40.

15. Vanzijl,P.C.M., Sukumar,S., Johnson,M.O., Webb,P., and Hurd,R.E. 1994. Optimized Shimming for High-Resolution Nmr Using 3-Dimensional Image-Based Field-Mapping. Journal of Magnetic Resonance Series A 111:203-207.

16. Wu,P.S., and Otting,G. 2005. Rapid pulse length determination in high-resolution NMR. J. Magn Reson. 176:115-119.

17. Bales,J.R., Nicholson,J.K., and Sadler,P.J. 1985. Two-dimensional proton nuclear magnetic resonance "maps" of acetaminophen metabolites in human urine. Clin. Chem 31:757-762.

18. Nocairi,H., Qannari,E.M., Vigneau,E., and Bertrand,D. 2005. Discrimination on latent components with respect to patterns. Application to multicollinear data. Computational Statistics & Data Analysis 48:139-147.

19. Westerhuis,J.A., Hoefsloot,H.C.J., Smit,S., Vis,D.J., Smilde,A.K., van Velzen,E.J.J., van Duijnhoven,J.P.M., and van Dorsten,F.A. 2008. Assessment of PLSDA cross validation. Metabolomics 4:81-89.

20. Lindgren,F., Hansen,B., Karcher,W., Sjostrom,M., and Eriksson,L. 1996. Model validation by permutation tests: Applications to variable selection. Journal of Chemometrics 10:521-532.

21. Jansen,J.J., Hoefsloot,H.C.J., van der Greef,J., Timmerman,M.E., and Smilde,A.K. 2005. Multilevel component analysis of time-resolved metabolic fingerprinting data. Analytica Chimica Acta 530:173-183.

22. Cloarec,O., Dumas,M.E., Craig,A., Barton,R.H., Trygg,J., Hudson,J., Blancher,C., Gauguier,D., Lindon,J.C., Holmes,E. et al 2005. Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. Analytical Chemistry 77:1282-1289.

23. Lin,K., and Fajardo,K. 2008. Screening for asymptomatic bacteriuria in adults: evidence for the U.S. Preventive Services Task Force reaffirmation recommendation statement. Ann. Intern. Med. 149:W20-W24.

24. Swann,J., Wang,Y., Abecia,L., Costabile,A., Tuohy,K., Gibson,G., Roberts,D., Sidaway,J., Jones,H., Wilson,I.D. et al 2009. Gut microbiome modulates the toxicity of hydrazine: a metabonomic study. Mol. Biosyst. 5:351-355.

25. Nicholson,J.K., Lindon,J.C., and Holmes,E. 1999. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29:1181-1189.

26. REUBI,F.C. 1953. Glomerular filtration rate, renal blood flow and blood viscosity during and after diabetic coma. Circ. Res. 1:410-413.

SUPPLEMENTARY MATERIALS

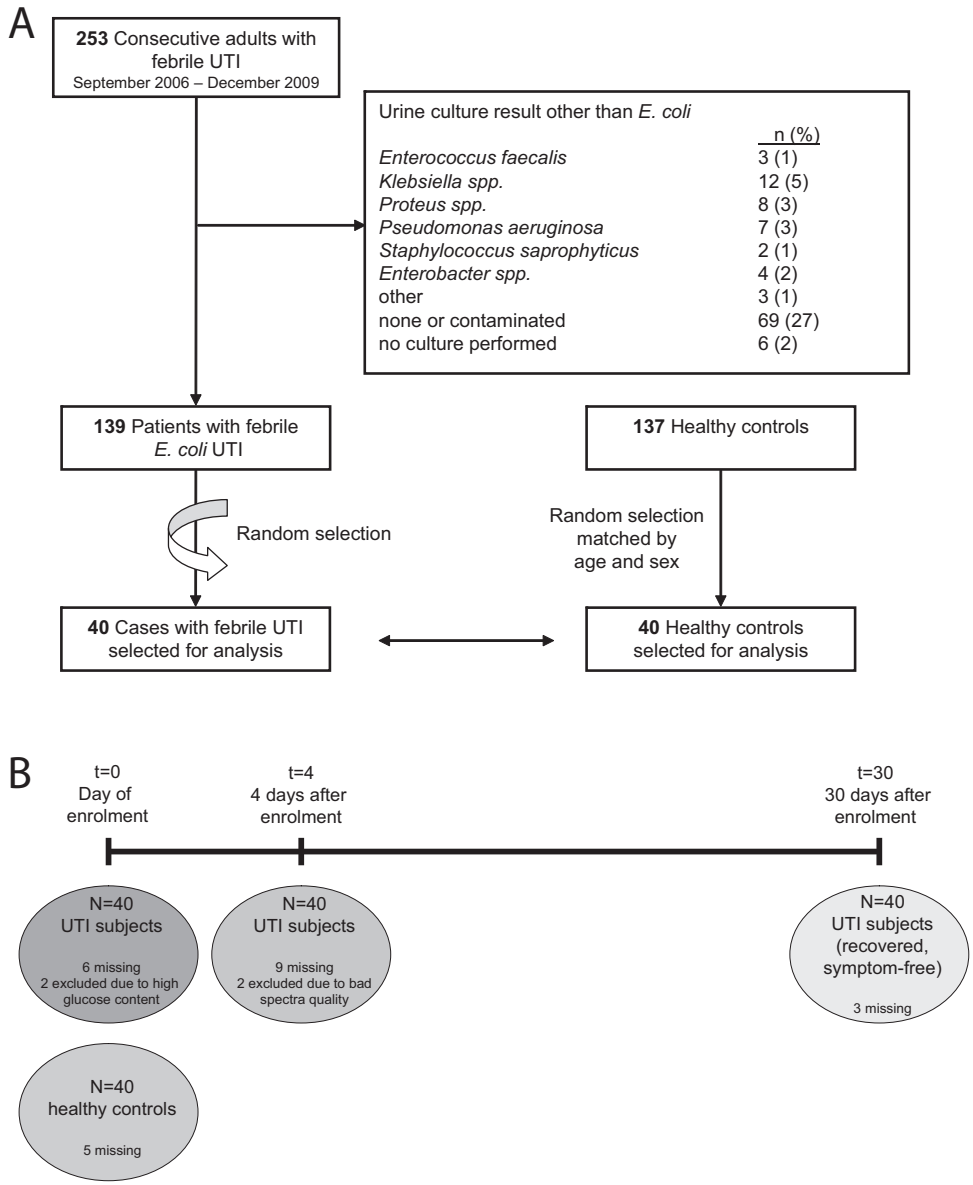

Figure S1. Design of the study.

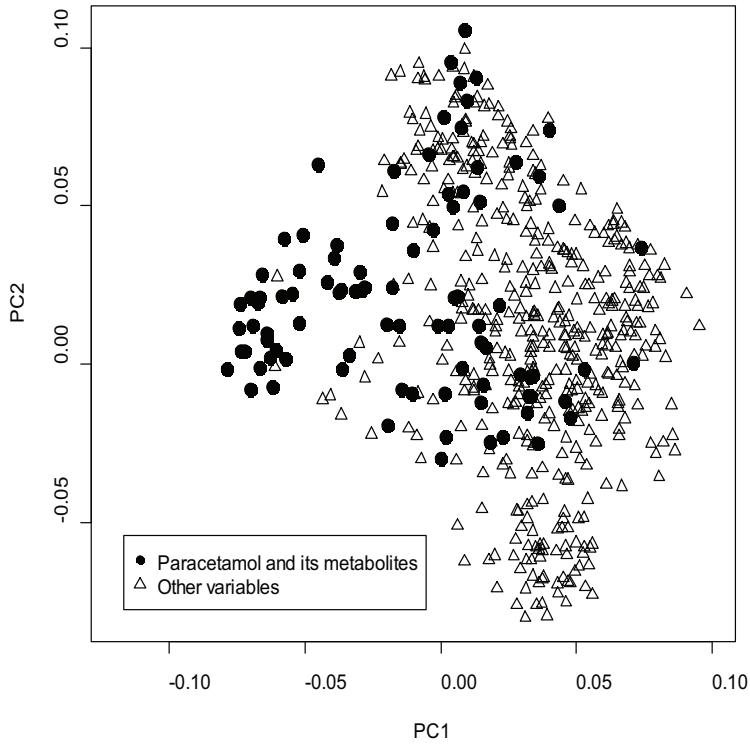

Figure S2. Loadings plot of the PCA model created using urine spectra of samples at baseline. Dots indicate variables that correspond to the spectral regions of paracetamol and its metabolites, triangles represent all the other variables.

Chapter 6

Cross-platform analysis of longitudinal data in metabolomics

Nevedomskaya E., Mayboroda O.A., Deelder A.M.

Molecular Biosystems, **2011**, DOI: 10.1039/c1mb05280b

ABSTRACT

Metabolic profiling is considered to be a very promising tool for diagnostic purposes, for assessing nutritional status and response to drugs. However, it is also evident that human metabolic profiles have a complex nature, influenced by many external factors. This, together with the understanding of the difficulty to assign people to distinct groups and a general move in clinical science towards personalized medicine, raises the interest to explore individual and variable metabolic features for each individual separately in longitudinal study design. In the current paper we have analyzed a set of metabolic profiles of a selection of six urine samples per person from a set of healthy individuals by ^{1}H NMR and reversed-phase UPLC-MS. We have demonstrated that the method for recovery of individual metabolic phenotypes can give complementary information to another established method for analysis of longitudinal data—multilevel component analysis. We also show that individual metabolic signatures can be found not only in ^{1}H NMR data, as has been demonstrated before, but also even more strongly in LC-MS data.

INTRODUCTION

Metabolomics is a post-genomics technology, the aim of which is "profiling metabolism in complex systems".(1) The reasoning behind metabolomics experiments is that metabolites, compared to genes, transcripts and proteins, offer the closest representation of the phenotype.(2) Thus, they can contain valuable information on a disease development, in contrast to genetics that gives insight into predisposition to a disease. Conducting this type of research assumes the existence of specific biomarkers or metabolic signatures that can distinguish between pathological states. The potential of metabolomics to reveal signatures of pathological conditions has been demonstrated on a number of neurological disorders (Huntington's disease,(3) Parkinson's disease,(4) multiple sclerosis(5) *etc.*), various cancers (ovarian and breast,(6) pancreatic,(7) colorectal(8) and others), cardiological abnormalities (*e.g.* ischemia(9)) and many other diseases.

It is, however, evident, that metabolic profiles reflect not only the disease/healthy state of the organism but, as a representation of a given phenotype, they strongly depend on factors such as gender, age and our daily habits, like, for example, diet, drugs and alcohol intake.(10,11) Another extremely important factor affecting metabolic profiles is gut microbiota. It has been shown that even for genetically identical laboratory animals gut microflora is influenced by environment and dietary factors.(12–14) For humans, in which genetic variation is enormous, gut microflora is much more diverse and is additionally affected by factors such as, for instance, stress.(15)

Therefore, human metabolic profiles are highly dependent on environmental factors and may vary from day to day due to turbulent conditions of our fast and highly stressful modern lives. A way to overcome possible negative effects of this variability on the interpretation of metabolomics data is multiple sampling over time per individual. The main advantage of this approach is the possibility to get an insight into the biological processes, which are usually missed by a simple, static comparison of "diseased vs. healthy". The feasibility of the longitudinal sampling has been demonstrated more than once in toxicology experiments in animals,(16) in human intervention studies(17) and in monitoring cyclic dynamics in healthy women.(18)

A number of statistical methods that can deal with longitudinal(19) or paired(20) data exist. One of the methods suitable for the analysis of such data is the multilevel approach that separates levels of variation present in the data into inter- and intra-individual.(21) Moreover, it has been shown that, despite the multiple sources of variability, present in human biofluids and particularly in urine, there are constant individual metabolic signatures, which are probably to a great extent determined by genetics.(22) Assfalg et al. used a combination of established statistical methods for individual classification, or in

other words person recognition, on the basis of Nuclear Magnetic Resonance (NMR) spectra of urine. The core of this approach was a variant of Principal Component Analysis (PCA), which was an innovative tool for face recognition 20 years ago,(23) and proved to be innovative for recognition of human urine metabolic signatures today. The existence of stable personal metabolic phenotypes is linked to the idea of homeostasis or, more precisely, to the idea that individual, self-regulatory, genetically controlled mechanisms maintain the homeostasis at any price. Consequently, the disruption of homeostasis means the beginning of a disease development.(24;25) Thus, monitoring of individual metabolic signatures, which represent dynamic, time-correlated changes of the phenotype, might ultimately develop into a preferred approach for practical personalized medicine.

Thus, there are different statistical methods that can be applied to the metabolic profiles of multiple samples per individual and allow having various perspectives on the data. Besides, an enormous chemical diversity of metabolites has resulted in a broad spectrum of analytical approaches used in metabolomics, especially with regards to mass spectrometry. To select an auxiliary to NMR MS method, one has to make a choice between gas chromatography (GC), capillary electrophoresis (CE) and liquid chromatography (LC). Choosing the latter, one still remains with a range of possibilities like, for example, reverse phase or hydrophilic interaction liquid chromatography (rpLC or HILIC).

In the current research we wanted to demonstrate that multilevel component analysis (MCA) and person recognition can be used in parallel, and that the information retrieved by the two methods is complimentary and together they can form a toolbox for analysis of longitudinal datasets. To this end we analyzed a set of urine samples from 8 healthy individuals (each of them donated 6 samples) by ^{1}H NMR and reversed phase UPLC-MS, which requires relatively simple sample preparation and is one of the most widely used MS techniques in metabolomics. MCA and person recognition methods were applied to the data. We here show that individual metabolic phenotypes can be identified not only on the basis of ^{1}H NMR spectra, as has been shown before, but also on the basis of LC-MS data. We also demonstrate the information extracted from this type of designed study using the two statistical approaches, based on diverse sources of variation in the data. The difference in information content of the data from the two analytical techniques is analyzed and discussed as well.

MATERIALS AND METHODS

Samples. Urine samples were collected, after written consent, from 8 self-declared healthy individuals from the same working environment (Leiden University Medical Center, the Netherlands). The volunteers were equally divided into men and women, aged

between 25 and 45 years old, all Caucasian. Each volunteer provided 6 urine samples of the first morning urine after over-night fasting on 6 different days (5 consecutive weekdays and one after the weekend). No diet restrictions were implied; none of the subjects was taking medication. Samples were collected in sterile 15 ml polypropylene tubes, kept at 4 °C, frozen within 8 h of collection, and stored for approximately 2 weeks at -20 °C until the measurement. In total 48 urine samples were analyzed.

Sample preparation. Frozen samples were thawed at room temperature and vortexed before use.

Sample preparation for ^{1}H NMR experiments. Aliquots of urine sample (1000 μl) were centrifuged at 3000g for 15 min at 4 °C to remove any precipitate. 600 μl of each sample were transferred to a 96 deep-well plate, further preparation was automated using the Bruker Sample Track system and a Gilson 215 robotic system. Here 540 μl urine were added to 60 μl of pH 7.0 sodium phosphate buffer (0.2 M) in 10% D2O containing 0.53 mM sodium 3-trimethylsilyl-tetradeuteriopropionate (TSP) and 0.26 mM NaN3, thoroughly mixed and transferred to a new 96 deep-well plate. Samples were centrifuged at 3000g for 5 min to remove any solid debris. A modified Gilson 215 robot was used to transfer 565 ml of sample from the plate into 5 mm SampleJet NMR tubes.

Sample preparation for LC-MS experiments. 150 μl of each urine sample were mixed with 450 μl of water and subsequently centrifuged at 3000 rpm for 10 min. 5 μl of sample was used for injection.

Data acquisitions.

^{1}H NMR experiments. All ^{1}H NMR experiments were performed on a 600 MHz Bruker Avance II spectrometer (Bruker BioSpin, Karlsruhe, Germany) equipped with a 5 mm TCI cryogenic probe head with Z-gradient system and automatic tuning and matching. Temperature calibration was done prior to the measurements using the method of Findeisen *et al.*(40)

One-dimensional ^{1}H NMR spectra were recorded at 300 K using the first increment of a NOESY(41) pulse sequence with presaturation (γB1 = 50 Hz) during a relaxation delay of 4 s and a mixing time of 10 ms for efficient water suppression. A total of 32 768 data points were recorded with 32 scans covering a sweep width of 12 336 Hz. The free induction decay (FID) was zero-filled to 65 536 complex data points prior to Fourier transformation and an exponential window function was applied with a line broadening factor of 1.0 Hz.

A sample ^{1}H NMR spectrum can be found in Supplementary Materials, Figure S1a.

LC-MS experiments. The samples were analyzed on a UPLC-ESI-UHR-ToF system. The injection scheme was randomized and included quality control samples (mix of all of the urine samples, prepared in the same way as the individual samples), as well as a set of

analytical standards (mix of pesticides, see Supplementary Materials, Table S1) to ensure the robustness of the workflow and to evaluate the analytical variability. Quality control (QC) and analytical standards were injected at the beginning and at the end of the sequence, as well as every four biological samples. In total 28 QC runs were acquired. The UPLC (Ultimate 3000 RS tandem LC system, Dionex, Amsterdam, The Netherlands) was equipped with a pre-column (Acclaim 120 C18, 5 mm, 120 Å, 2 × 10 mm) and two analytical columns (Acclaim RSLC 120 C18, 2.2 mm, 120 Å, 2.1 × 100 mm) working alternatively to speed up the acquisition series. The UPLC flow was set at 400 μl min^{-1} and the mobile phases were water + 0.1% formic acid *v/v* (Phase A) and methanol+0.1% formic acid *v/v* (Phase B). The gradient was as follows: 1 min 0% phase B, then in 1 min to 10% phase B, held for 1 min at 10% phase B, and subsequently in 6,5 min to 100% phase B and held for 3 min at 100% phase B. Before each chromatographic run, a calibrant solution of sodium formate was injected in Flow Injection Analysis mode.

The ESI-UHR-ToF (maXis, Bruker Daltonics, Bremen, Germany) was operated in the positive ionization mode and acquired data in the mass range from *m/z* 50 to 1500 with a spectra rate of 1 Hz. The capillary was set at 2500 V, the End Plate offset at -500 V, the Nebulizer gas at 2 bar and the dry gas at 8 l min^{-1} at 180 °C.

A sample LC-MS chromatogram can be found in Supplementary Materials, Figure S1b.

Data pre-processing. *^{1}H NMR data pre-processing.* All spectra were manually phase- and baseline-corrected using Topspin 2.1 (Bruker BioSpin, Karlsruhe, Germany) and automatically referenced to TSP signal (0.0 ppm). Each spectrum was integrated (binned) using 0.0095 ppm integral regions between 0.5 and 10 ppm, the residual water and urea region between 4.5 and 6 ppm was excluded, resulting in 842 bin regions used for the analysis. To account for any difference in concentration between the samples, each spectrum was normalized to its total area and subsequently by Probabilistic Quotient Normalization (PQN) (42) using average spectrum as a reference.

LC-MS data pre-processing. All data files were recalibrated on the masses of sodium formiate clusters. The alignment of chromatograms and peak picking was performed using open-source XCMS software (The Scripps Research Institute, La Jolla, CA).(43) Finding peaks was performed using the "centWave" algorithm with *m/z* deviation set to 5 ppm, and the scan range between 20 and 700 scans. Grouping of peaks was done with parameters minsamp set to 28 (number of QC samples) and bandwidth to 10. Retention time correction was done with default parameters. The resulting table included the detected ion features and their peak areas. The peaks were filtered on the basis of QC samples: the peak was retained in the analysis if it was present in all the QC samples and relative standard deviation of the area in QC samples was less than 20%. The final table contained 965 ion

features, which areas were normalized on total areas of the samples and subsequently by PQN(42) with an average of QC samples taken as a reference.

The consistency of the data and the absence of column-bias were checked using Principal Component Analysis (Supplementary Materials, Figure S2).

Statistical data analysis. *Principal Component Analysis* was performed on logarithmically transformed, mean-centered and unit variance scaled data using the NIPALS algorithm.(44)

Person recognition. The person recognition approach used in the current paper was based on the previously published classification method.(22) Among the classification methods used by Assfalg *et al.* the combination of Principal Component Analysis (PCA) for data reduction and canonical discriminant analysis (CA) was chosen as the most effective one. The accuracy of classification was assessed using test-set validation: in each round of validation one randomly selected sample per donor was taken out into the test-set, and a model was built based on the remaining samples. The test-set samples were projected into the PCA–CA subspace and classified according to the minimum distance to the mean of the discriminated groups. The resulting class labels were compared to the real ones and the number of correct classifications was evaluated. The validation was performed in 1000 rounds and the results averaged throughout all the rounds (Supplementary Materials, Figure S3a).

Recognition accuracies were also assessed in 100 rounds of Subject ID permutations and compared with the actual accuracies, statistical significance of the difference was assessed using the Mann–Whitney test.

Multilevel Components Analysis (MCA). MCA is an effective method for separating the variation between- and within- individuals and analyzing them by different submodels. The method was implemented as described by Jansen *et al.*(21) PCA were performed on the data matrix corresponding to the between-individual variation and on the within-individual variation for each individual separately (Supplementary Materials, Figure S3b).

All the data-preprocessing and statistical analysis were performed in a R statistical software environment (http://www.r-project.org/) using in-house developed scripts and open-source packages.

RESULTS

A set of longitudinal urine samples from healthy individuals was analyzed by ^{1}H NMR and rpUPLC-MS and subsequently by various statistical methods in order to compare the information that can be retrieved from the data by different analytical and statistical approaches. PCA was performed on ^{1}H NMR and LC-MS data. The scores plot of the first

two principal components for each of the techniques revealed some clustering by individuals; however no separation between the genders was observed (Fig. 1a and c). Difference between the genders was visible on the scores plot of the third and fourth principal components in the case of the ^{1}H NMR data (Figure 1b), and of the second and third principal components in the case of the LC-MS data (Figure 1d). At a first glance, there appeared to be a similarity between the score plots of the first two principal components on ^{1}H NMR and LC-MS data, for example, subject 7 is separated from the rest of the people. A way to give a numerical value to this similarity is the use of the RV-coefficient, which is a multivariate extension of correlation coefficient. This has already been used before for estimation of the overlap of metabolomics data matrices, but in that case both matrices were derived from MS-based experiments.(26) For all the 8 principal components of the PCA the RV-coefficient was found to be not that high, not exceeding 0.46. The RV-coefficient does not increase anymore after the fourth component, which explains 44 and 53% of the variation in the ^{1}H NMR and LC-MS data respectively (Table 1). Hence, the relative positions of data points in the PCA subspace are different for ^{1}H NMR and LC-MS, with little overlap.

It was evident from the PCA analysis that there are different sources of variation present in the data, which this method is not capable of separating. One of the ways to dissect variations present at different levels in the data (*e.g.* between and within individuals) is to use multilevel analysis which has been successfully applied for a number of applications in social sciences, geography, public health (27) and recently also in metabolomics.(20,21) This method was applied to the data so that the variation at two levels—between individuals and within each of the individuals—was explored. It allows identification of spectral regions or peaks variable between individuals and peaks/regions variable between the time points for each individual. In Figure 2 the loading plots for the between and within-individual models are shown on the example of the ^{1}H NMR data, demonstrating those variable areas. The RV-coefficient, calculated for the results of multilevel analysis (on the between-individual score matrices), is higher than that for the PCA analysis, but is very stable even with the growth of the explained variance, again indicating that the two analytical techniques explain different relations between the samples (Table 2).

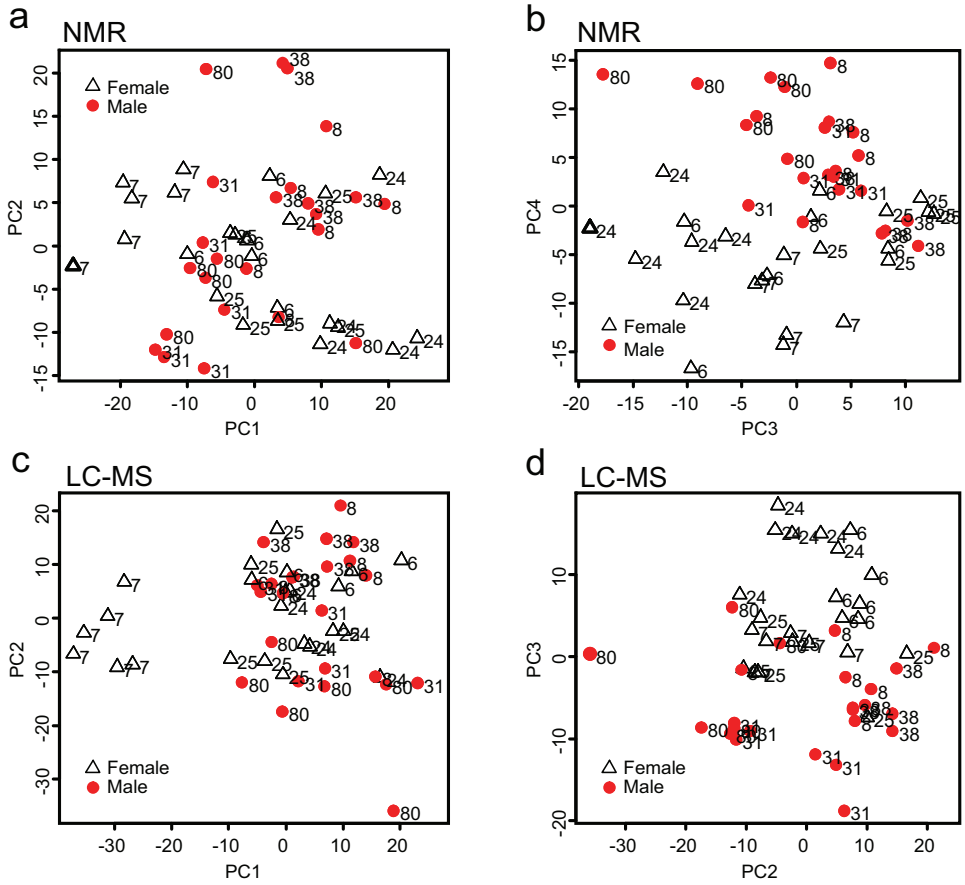

Figure 1. Scores plots of the PCA analysis of ^{1}H NMR and LC-MS data from urine samples of 8 individuals sampled at six different time points. Samples are labeled by individuals' IDs. Triangles represent urine samples of females, dots of males. (a) First two principal components of the PCA on NMR data cover 16.1 and 9.6% of variation respectively. (b) Third and fourth principal components of the PCA on NMR data cover 7.3% and 6.6% of variation, respectively. (c) First two principal components of the PCA on LC-MS data cover 21.1 and 12.3% of variation, respectively. (d) Second and third principal components of the PCA on LC-MS data, the third component covers 7.5% of variation.

Table 1. Summary of principal component analysis performed on NMR and LC-MS data and multivariate correlation (RV-coefficient), calculated on the resulting score matrices.

	Explained variation, %		
PC No.	NMR	LC-MS	RV
1	16.1	21.2	0.2
2	25.7	33.5	0.24
3	33	41	0.28
4	39.6	47.9	0.4
5	44.3	53	0.41
6	48.3	57.7	0.42
7	52.1	61.9	0.44
8	55.6	65.9	0.46

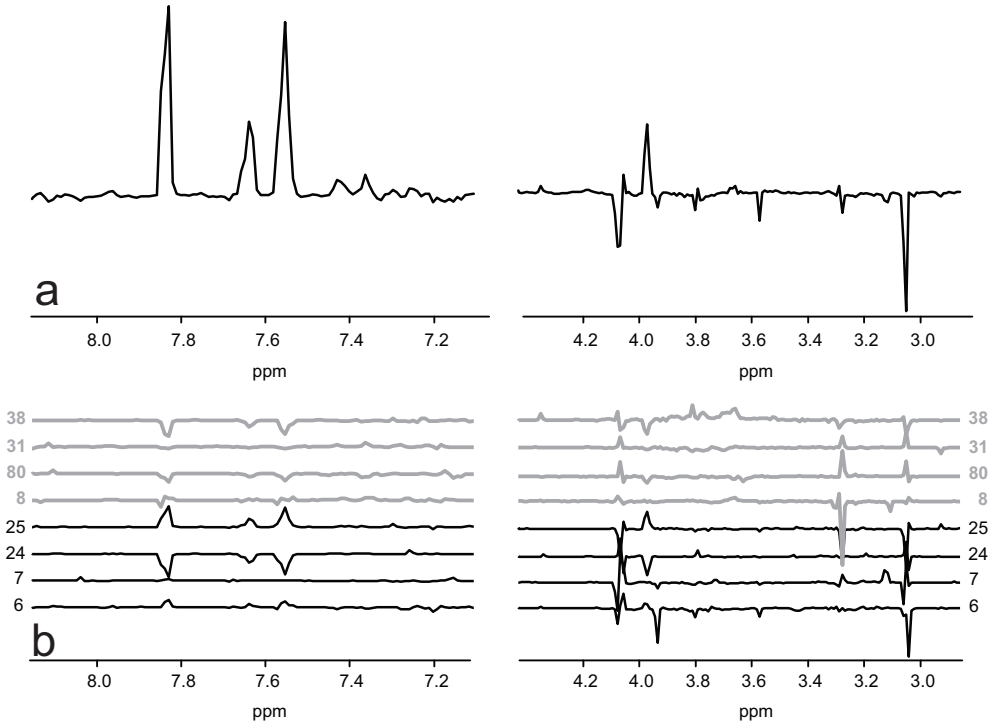

Figure 2. Loadings plot of the multilevel component analysis of ^{1}H NMR data of urine samples from 8 individuals sampled at six different time points: (a) loadings of the first component of the between-individual model, (b) loadings of the first component of the within-individual model for each person, colored by gender (grey—male, black—female) and labeled by individuals' IDs.

Depending on the research question, one might not be interested in the most variable regions for each individual, but in the most constant ones, person-specific features. Those unique patterns can be used to recognize each person from the rest. The existence of such fingerprints in ^{1}H NMR data and a method to assess them were demonstrated previously (22) using an innovative combination of classical statistical methods—PCA and canonical discriminant analysis with thorough validation. We performed person recognition on the ^{1}H NMR data evaluating the accuracy with which each of the people is predicted. The recognition accuracy ranged from 59.5 to 99.5% which matches the estimated probability of correct classification for the same number of samples in the model described in the previous work.(22) The mean recognition was 84%, which is quite high taking into account that in each validation step the model is built only on 5 spectra. The accuracy of recognition was also calculated on the set with permuted person labels and it appeared to be significantly lower (mean accuracy 13%, p-value < 0.001) than the real recognition results (Supplementary Materials, Figure S4a).

One of the advantages of the person recognition method is that it is possible to perform back-projection of scores in the canonical subspace into the PCA scores subspace and then into the original variables. As a result of this procedure individual metabolic phenotypes are obtained (Figure 3a). These metabolic phenotypes represent the characteristic spectral regions for each person and are, unlike the original profiles, easily clustered by *e.g.* hierarchical clustering per person (Supplementary Materials, Figure S5).

Table 2. Summary of multilevel component analysis (between individuals) performed on NMR and LC-MS data and multivariate correlation (RV-coefficient), calculated on the resulting score matrices.

	Explained variation, %		
PC No.	NMR	LC-MS	RV
1	39.4	49.8	0.64
2	70	78.1	0.6
3	87.8	88.1	0.6
4	93.8	94.6	0.62
5	96.9	97.3	0.63
6	99	99.2	0.65

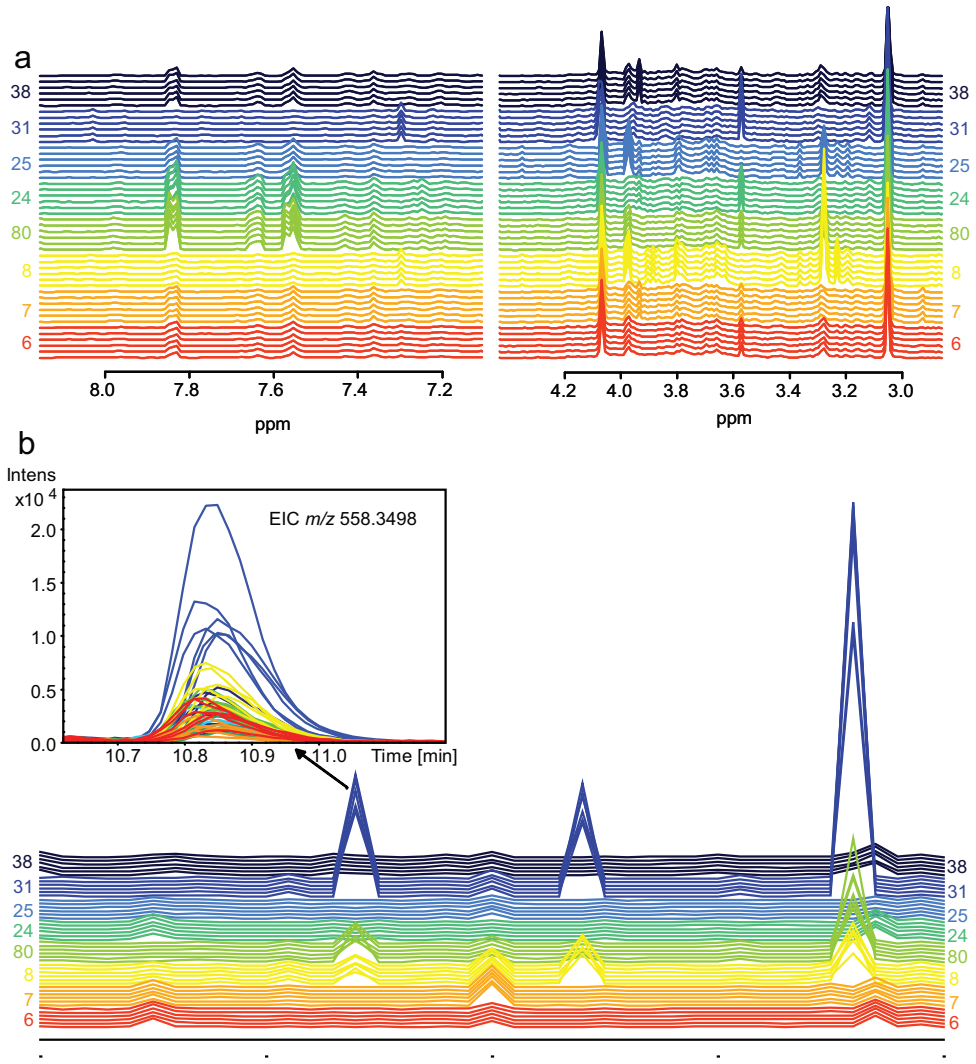

Figure 3. Individual metabolic phenotypes found within urine samples of 8 individuals sampled at six different time points. (a) Based on the ^{1}H NMR spectra. (b) Based on LC-MS data; one of the differential peaks (m/z of 558.3498) was used to construct the corresponding extracted ion chromatograms (inlet). Colored and labeled by individuals' IDs.

To illustrate the results and differences of person recognition and multilevel analysis methods we here show an example for one of the participants, using the ^{1}H NMR spectra and its analysis. In Figure 4 the original binned ^{1}H NMR spectra are shown together with

the individual metabolic profile of the person and his within- and between-individual loadings. As can be seen from the picture, there are peaks that contribute differently to the different types of analyses. For instance, the peak at 3.97 ppm appears in all of the six urine spectra, but its intensity is variable, so it contributes to both the individual metabolic profile and within-individual loadings. This peak is found to be also characteristic for separation between individuals, contributing to the between-individual loadings. Another peak at 3.81 ppm appears only in one of the six spectra, and thus it contributes to the within-individual loadings, indicating its variability through the time course of the study, however, it is not characteristic of the person (does not appear in the individual metabolic profile), neither is responsible for the separation between the subjects. Another example is the peak at 3.57 ppm which is present in all the samples and has quite a consistent intensity; it does not contribute to the within individual loadings, but is characteristic for the individual and drives the separation between the individuals.

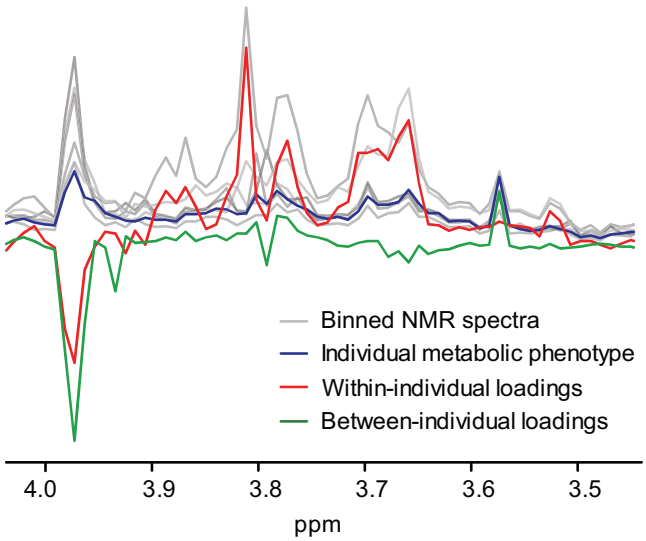

Figure 4. Comparison of the statistical analyses of the urinary ^{1}H NMR data from one individual. Grey lines show original (binned) ^{1}H NMR spectra. Blue line indicates the mean individual metabolic profile, red is within-individual loadings of the first component, green—between-individual loadings of the first component.

The individual metabolic profiles were found to exist not only in ^{1}H NMR spectra, but also in UPLC-MS data. Moreover, the accuracies of the person recognition performed on LC-MS data were found to be significantly higher than those derived from ^{1}H NMR data: the range of accuracies was between 92 and 100% and the mean recognition was 98.3% (Supplementary Materials, Figure S4b). Recognition accuracies in the randomized

experiment were significantly lower (mean 12.6%, p-value < 0.001) than the real values. In the same way as for ^{1}H NMR, the LC-MS peaks specific for an individual can be found by back-projection (Figure 3b) and traced back in the original data. Indeed they show a differential profile across individuals. The recognition accuracies based on ^{1}H NMR and LC-MS are not correlated across the subjects (correlation coefficient is -0.012): the individuals that are relatively badly recognized on the basis of ^{1}H NMR may be well-recognized on the basis of LC-MS and *vice versa*.

Thus, in the current study a small longitudinal cohort of urine samples from healthy individuals was analyzed, and even with this limited sample set it was evident that various sources of variation are confounded. There are statistical methods available to extract this variation separately and examine the data from a different perspective. Those methods (multilevel component analysis and person recognition) can be applied to both ^{1}H NMR and LC-MS data. The use of multiple analytical platforms also widens the information extracted from a study.

DISCUSSION

The importance of a personalized approach in health-related research is being widely accepted by the scientific community, (28) and metabolic profiling is recognized as a valuable tool for personalized medicine.(29) However in the majority of metabolomics studies a traditional "case–control" design is applied, although it is well-known that the definition of these groups, and especially the group of healthy individuals, is very vague.(30) In contrast, the definition of the physical individuality is very clear and monitoring an individual reaction to perturbations and its development in time is a promising approach for medicine and pharmacology.

In metabolomics the advantages of the longitudinal study design that implies multiple sampling per individual have clearly been demonstrated for interventional studies where the dynamic response of the organism to the drug or other substance can be monitored.(31;32) It has also been demonstrated that "classical" data analysis methods used n metabolomics, such as PCA and PLS-DA, are suboptimal(20) for such dynamic data and other methods, separating levels of variation, should be used. The longitudinal design offers possibilities for differential analysis; depending on the question of interest one may focus on differences between the subjects, on variations within each subject or identification of unique profiles for each subjects.

To illustrate some of the possibilities for the analysis of longitudinal metabolomics data we applied a series of statistical methods to the ^{1}H NMR and LC-MS data of urines of 8 individuals, 4 females and 4 males, each contributing 6 urine samples for the study. As no

special diet or life style restrictions were applied, it was obvious that the data would contain a lot of variation due to day-to-day differences, between-subjects diversity and certain grouping of the samples due to, for example, gender, age *etc.* This, indeed, was confirmed by PCA, which summarizes the variation present in the data. The clustering according to person was evident; however day-to-day variation for most of the people was much higher than the differences between individuals. Gender distinction was also present, but not in the first two components of PCA, suggesting that the between-gender difference is overruled by all the other sources of variation.

In the PCA analysis LC-MS data showed more variation covered in the first principal components, than the ^{1}H NMR data (Table 1). There was some similarity visible for the position of the data points along the first principal components in the two datasets; however RV-coefficients calculated on the principal component subspaces for the two techniques were rather low (Table 1). This suggests that the two analytical methods reveal different biological phenomena reflected in the metabolic composition of the same biological samples.

PCA modeling has shown that the metabolic data with underlying design contain information from different sources—from the variation between the subjects, as well as the variation between the samples for each person. There is a class of multilevel statistical models which can separate the data into levels and as such are perfectly applicable to our data. In the case of MCA, applied in the current study, the overall variation present in the study was divided into between-individual and within-individual and separate analysis was performed on each block. This method reveals spectral regions differential between the people, as well as regions which are variable for each of the people through the time course of the analysis.

Another method used, namely, person recognition, focuses on different parts of the spectra, which are consistent for the individual and thus characteristic. Before, this method was successfully applied to ^{1}H NMR spectra, revealing the existence of individual metabolic signatures, which were found to be extremely stable over time and could be largely explained by genetics.(22,33) We successfully applied the described method to our data and observed a recognition accuracy corresponding to the number of samples used. We also demonstrated on an example how the individual metabolic profiles give information complementary to that derived from multilevel analysis.

As can be seen from the analysis, different levels of variation can be recovered from a set of spectra. The choice of an appropriate method for statistical analysis should be based on the question posed by the investigation. The right answers can only be derived from a carefully designed and analyzed study. Consequently, longitudinal design offers possibilities

for real personalized medicine such as exploration of the effects not averaged between people, elimination of day-to-day variation, focusing on intrinsic individual properties reflected in the metabolic composition of urine.

The person recognition strategy, to the best of our knowledge, has so far not been applied to LC-MS data, which is one of the most commonly used analytical techniques in metabolomics experiments.(34) One study evaluating the amount of "personalized" information present in a set of LC-MS data has been conducted,(35) however this study only described the features, unique for an individual (i.e. appearing in one set of the spectra), but not the unique patterns of the features as in a person recognition approach. Thus our report appeared to be the first ever attempt to do the person recognition analysis on LC-MS data. Despite the fact that the LC-MS has somewhat lower analytical reproducibility than ^{1}H NMR, (36) person recognition accuracy was substantially higher in the case of LC-MS (all individuals were recognized with accuracy more than 92%, compared to 59% in ^{1}H NMR). There was absolutely no correlation between recognition of people in ^{1}H NMR and in LC-MS again pointing at the fact that the two techniques most probably provide different information concerning the samples.

Of course, the differences in the metabolome coverage between ^{1}H NMR and LC-MS come as no surprise. ^{1}H NMR is a universal approach capable of detecting all the compounds that contain hydrogens, whereas MS-based methods are more targeted due to the selectivity of the separation and ionization techniques used.(36) On the other hand, ^{1}H NMR has slightly lower sensitivity in comparison to MS. Thus, there is a certain "bias" in metabolomics experiments performed on a single analytical platform: the coverage of the metabolites is either limited by the sensitivity, or by the separation method. Thus, the observed "personalized" content of LC-MS data in comparison to ^{1}H NMR might be a result of such analytical bias. This, however, has to be further explored; here we can make only a few assumptions about what is driving this difference.

In general, ^{1}H NMR-based metabolomics studies result in a systemic view on the studied phenomenon, due to the fact that a lot of the detected compounds are related to energy metabolism: TCA cycle intermediates, amino acids, *etc.* These molecules are highly abundant in biofluids, are also day-to-day variable depending on the diet and are also involved in many biochemical pathways. The latter means that they change in many states of the organism, which leads to the problem of "usual suspects"(37) with many of the same metabolites discovered to be differential in a number of conditions.(38)

In contrast to NMR, LC-MS is more specific due to the inherent selectivity of the separation method and high sensitivity of the detection. Reversed-phase UPLC-MS is highly suitable for separation of medium polar and non-polar compounds.(39) Most of the

molecules related to energy metabolism, amino and other organic acids will not be retained. Thus, the part of the metabolome, covered by rpUPLC-MS, might be less affected by diet and gut microflora and might provide a closer approximation of the phenotype.

This phenomenon certainly would need more extensive investigation and might be an extremely important issue in the decision how to conduct a study using a certain analytical platform, depending on the study design and the question of interest.

In total, the amount of biologically relevant information that can be derived from metabolomics experiments is enormous. However, the quality of this information depends on a clear definition of the goals and the study design as well as on the selection of the analytical platform and the subsequent statistical analysis. All of these factors are extremely important for obtaining successful results and generating a relevant hypothesis. Consequently, performing costly, labor-intensive metabolomics experiments with the sole aim to distinguish "diseased from healthy" might be seen as a suboptimal use of manpower, instrumental resources and, the most importantly patient material. The power of a longitudinal design and the flexibility of various statistical methods to analyze such a design may open new possibilities. Individual metabolic signatures, that represent dynamic, time-correlated changes of phenotype, may actually be used as a phenotype-readout essential for practical personalized medicine.

CONCLUSIONS

In the metabolomics-related literature somewhat controversial ideas are present: on the one hand that metabolites can provide unique diagnostic information, and on the other hand that their concentrations are very sensitive to non-systemic external factors and vary even from day to day for one individual. However, it has been demonstrated that highly individual metabolic signatures exist in for instance urine, on top of which the other variation is superimposed. The available methods for the analysis of time-resolved data can focus either on variation between people or within the time course for an individual. In the current paper we have demonstrated the use and complementarity of the extracted information of some of these statistical methods on a set of data from healthy individuals.

We have also shown that the detection of individual metabolic profiles is not solely the property of ^{1}H NMR, but is also possible based on UPLC-MS data, interestingly—even with a higher accuracy. Based on this limited data set, it would appear that the parallel analysis of ^{1}H NMR and LC-MS indicates that the two techniques explain different phenomena in the data. The higher accuracy of person recognition in LC-MS further suggests that the method might be more sensitive to unique, individual-specific features, while ^{1}H NMR might reflect a more systemic response.

ACKNOWLEDGEMENTS

The authors would like to thank all the volunteers for the supplied samples, Sibel Göraler M.Sc. and Ing. Bart Schoenmaker for the analytical work, Dr Hartmut Schäfer for his help with implementing the person recognition method, Dr Paul J. Hensbergen for fruitful discussion.

REFERENCES

1. Lindon, J. C.; Holmes, E.; Bollard, M. E.; Stanley, E. G.; Nicholson, J. K. Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. Biomarkers 2004, 9 (1), 1-31.

2. Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. Nat. Rev. Drug Discov. 2002, 1 (2), 153-161.

3. Underwood, B. R.; Broadhurst, D.; Dunn, W. B.; Ellis, D. I.; Michell, A. W.; Vacher, C.; Mosedale, D. E.; Kell, D. B.; Barker, R. A.; Grainger, D. J.; Rubinsztein, D. C. Huntington disease patients and transgenic mice have similar pro-catabolic serum metabolite profiles. Brain 2006, 129 (Pt 4), 877-886.

4. Bogdanov, M.; Matson, W. R.; Wang, L.; Matson, T.; Saunders-Pullman, R.; Bressman, S. S.; Flint, B. M. Metabolomic profiling to develop blood biomarkers for Parkinson's disease. Brain 2008, 131 (Pt 2), 389-396.

5. Simone, I. L.; Federico, F.; Trojano, M.; Tortorella, C.; Liguori, M.; Giannini, P.; Picciola, E.; Natile, G.; Livrea, P. High resolution proton MR spectroscopy of cerebrospinal fluid in MS patients. Comparison with biochemical changes in demyelinating plaques. J. Neurol. Sci. 1996, 144 (1-2), 182-190.

6. Slupsky, C. M.; Steed, H.; Wells, T. H.; Dabbs, K.; Schepansky, A.; Capstick, V.; Faught, W.; Sawyer, M. B. Urine metabolite analysis offers potential early diagnosis of ovarian and breast cancers. Clin. Cancer Res. 2010, 16 (23), 5835-5841.

7. Nishiumi, S.; Shinohara, M.; Ikeda, A.; Yoshie, T.; Hatano, N.; Kakuyama, S.; Mizuno, S.; Sanuki, T.; Kutsumi, H.; Fukusaki, E.; Azuma, T.; Takenawa, T.; Yoshida, M. Serum metabolomics as a novel diagnostic approach for pancreatic cancer. Metabolomics 2010, 6 (4), 518-528.

8. Wang, H.; Tso, V. K.; Slupsky, C. M.; Fedorak, R. N. Metabolomics and detection of colorectal cancer in humans: a systematic review. Future. Oncol. 2010, 6 (9), 1395-1406.

9. Barba, I.; de Leon, G.; Martin, E.; Cuevas, A.; Aguade, S.; Candell-Riera, J.; Barrabes, J. A.; Garcia-Dorado, D. Nuclear magnetic resonance-based metabolomics predicts exercise-induced ischemia in patients with suspected coronary artery disease. Magn Reson. Med. 2008, 60 (1), 27-32.

10. Holmes, E.; Loo, R. L.; Stamler, J.; Bictash, M.; Yap, I. K.; Chan, Q.; Ebbels, T.; de, I. M.; Brown, I. J.; Veselkov, K. A.; Daviglus, M. L.; Kesteloot, H.; Ueshima, H.; Zhao, L.; Nicholson, J. K.; Elliott, P. Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 2008, 453 (7193), 396-400.

11. Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Anal Chem 2007, 79 (18), 6995-7004.

12. Friswell, M. K.; Gika, H.; Stratford, I. J.; Theodoridis, G.; Telfer, B.; Wilson, I. D.; Mcbain, A.

J. Site and Strain-Specific Variation in Gut Microbiota Profiles and Metabolism in Experimental Mice. Plos One 2010, 5 (1).

13. Martin, F. P.; Sprenger, N.; Montoliu, I.; Rezzi, S.; Kochhar, S.; Nicholson, J. K. Dietary modulation of gut functional ecology studied by fecal metabonomics. J. Proteome Res. 2010, 9 (10), 5284-5295.

14. Phipps, A. N.; Stewart, J.; Wright, B.; Wilson, I. D. Effect of diet on the urinary excretion of hippuric acid and other dietary-derived aromatics in rat. A complex interaction between diet, gut microflora and substrate specificity. Xenobiotica 1998, 28 (5), 527-537.

15. Rezzi, S.; Martin, F. P.; Alonso, C.; Guilarte, M.; Vicario, M.; Ramos, L.; Martinez, C.; Lobo, B.; Saperas, E.; Malagelada, J. R.; Santos, J.; Kochhar, S. Metabotyping of Biofluids Reveals Stress-Based Differences in Gut Permeability in Healthy Individuals. Journal of Proteome Research 2009, 8 (10), 4799-4809.

16. Keun, H. C.; Ebbels, T. M. D.; Bollard, M. E.; Beckonert, O.; Antti, H.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. Chemical Research in Toxicology 2004, 17 (5), 579-587.

17. van Velzen, E. J.; Westerhuis, J. A.; van Duynhoven, J. P.; van Dorsten, F. A.; Grun, C. H.; Jacobs, D. M.; Duchateau, G. S.; Vis, D. J.; Smilde, A. K. Phenotyping tea consumers by nutrikinetic analysis of polyphenolic end-metabolites. J. Proteome Res. 2009, 8 (7), 3317-3330.

18. Wallace, M.; Hashim, Y. Z. H. Y.; Wingfield, M.; Culliton, M.; McAuliffe, F.; Gibney, M. J.; Brennan, L. Effects of menstrual cycle phase on metabolomic profiles in premenopausal women. Human Reproduction 2010, 25 (4), 949-956.

19. Smilde, A. K.; Westerhuis, J. A.; Hoefsloot, H. C. J.; Bijlsma, S.; Rubingh, C. M.; Vis, D. J.; Jellema, R. H.; Pijl, H.; Roelfsema, F.; van der Greef, J. Dynamic metabolomic data analysis: a tutorial review. Metabolomics 2010, 6 (1), 3-17.

20. Westerhuis, J. A.; van Velzen, E. J. J.; Hoefsloot, H. C. J.; Smilde, A. K. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. Metabolomics 2010, 6 (1), 119-128.

21. Jansen, J. J.; Hoefsloot, H. C. J.; van der Greef, J.; Timmerman, M. E.; Smilde, A. K. Multilevel component analysis of time-resolved metabolic fingerprinting data. Analytica Chimica Acta 2005, 530 (2), 173-183.

22. Assfalg, M.; Bertini, I.; Colangiuli, D.; Luchinat, C.; Schafer, H.; Schutz, B.; Spraul, M. Evidence of different metabolic phenotypes in humans. Proceedings of the National Academy of Sciences of the United States of America 2008, 105 (5), 1420-1424.

23. Turk, M.; Pentland, A. Eigenfaces for Recognition. Journal of Cognitive Neuroscience 1991, 3 (1), 71-86.

24. van der, G. J.; Stroobant, P.; van der, H. R. The role of analytical sciences in medical systems biology. Curr. Opin. Chem Biol. 2004, 8 (5), 559-565.

25. van der Greef, J.; Smilde, A. Symbiosis of chemometrics and metabolomics: past, present, and future. JOURNAL OF CHEMOMETRICS 2005, 19 (5-7), 376-386.

26. Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van-der Vat; Jellema, R. H. Fusion of mass spectrometry-based metabolomics data. Analytical Chemistry 2005, 77 (20), 6729-6736.

27. Diez-Roux, A. V. Multilevel analysis in public health research. Annual Review of Public Health 2000, 21, 171-192.

28. Weston, A. D.; Hood, L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J. Proteome Res. 2004, 3 (2), 179-196.

29. Schnackenberg, L. K.; Kaput, J.; Beger, R. D. Metabolomics: a tool for personalizing medicine? Personalized Medicine 2008, 5 (5), 495-504.

30. Elliott, R.; Pico, C.; Dommels, Y.; Wybranska, I.; Hesketh, J.; Keijer, J. Nutrigenomic approaches for benefit-risk analysis of foods and food components: defining markers of health. Br. J. Nutr. 2007, 98 (6), 1095-1100.

31. Holmes, E.; Antti, H. Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. Analyst 2002, 127 (12), 1549-1557.

32. van der Greef, J.; Hankemeier, T.; McBurney, R. N. Metabolomics-based systems biology and personalized medicine: moving towards n=1 clinical trials? Pharmacogenomics 2006, 7 (7), 1087-1094.

33. Bernini, P.; Bertini, I.; Luchinat, C.; Nepi, S.; Saccenti, E.; Schafer, H.; Schutz, B.; Spraul, M.; Tenori, L. Individual Human Phenotypes in Metabolic Space and Time. J. Proteome Res. 2009.

34. Lindon, J. C.; Nicholson, J. K. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trac-Trends in Analytical Chemistry 2008, 27 (3), 194-204.

35. Johnson, J. M.; Yu, T. W.; Strobel, F. H.; Jones, D. P. A practical approach to detect unique metabolic patterns for personalized medicine. Analyst 2010, 135 (11), 2864-2870.

36. Lindon, J. C.; Nicholson, J. K. Spectroscopic and Statistical Techniques for Information Recovery in Metabonomics and Metabolomics. Annual Review of Analytical Chemistry 2008, 1, 45-69.

37. Robertson, D. G. Metabonomics in toxicology: A review. Toxicological Sciences 2005, 85 (2), 809-822.

38. Dunn, W. B.; Broadhurst, D. I.; Atherton, H. J.; Goodacre, R.; Griffin, J. L. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chem Soc. Rev. 2011, 40 (1), 387-426.

39. Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. Global metabolic profiling procedures for urine using UPLC-MS. Nature Protocols 2010, 5 (6), 1005-1018.

40. Findeisen, M.; Brand, T.; Berger, S. A ^{1}H-NMR thermometer suitable for cryoprobes. Magn Reson. Chem 2007, 45 (2), 175-178.

41. Kumar, A.; Ernst, R. R.; Wuthrich, K. A two-dimensional nuclear Overhauser enhancement (2D NOE) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules. Biochem. Biophys. Res. Commun. 1980, 95 (1), 1-6.

42. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. Anal Chem 2006, 78 (13), 4281-4290.

43. Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 2006, 78 (3), 779-787.

44. Wold, H. Estimation of principal components and related models by iterative least squares. In Multivariate Analysis, Krishnaiah, P. R., Ed.; Academic Press: New York, 1966; pp 391-420.

SUPPLEMENTARY MATERIALS

Table S1. The mix of pesticides used as the analytical standard.

Name	Molecular formula	m/z [M+H]	Retention time, min
Pymetrozine	$C_{10}H_{11}N_5O$	218.1036	4.7
Formetanate	$C_{11}H_{15}N_3O_2$	222.1237	4.88
Fenuron	$C_9H_{12}N_2O$	165.1022	4.89
Carbendazim	$C_9H_9N_3O_2$	192.0768	6.4
2-Hydroxyatrazine	$C_8H_{15}N_5O$	198.1349	6.99
Atrazine-Desisopropyl	$C_5H_8ClN_5$	174.0541	7.21
Metamitron	$C_{10}H_{10}N_4O$	203.0927	7.71
Acetamiprid	$C_{10}H_{11}ClN_4$	223.0745	7.86
Chloridazone	$C_{10}H_8ClN_3O$	222.0429	7.88
Crimidine	$C_7H_{10}ClN_3$	172.0636	7.89
Pirimicarb	$C_{11}H_{18}N_4O_2$	239.1503	8.13
Atrazine-Desethyl	$C_6H_{10}ClN_5$	188.0697	8.16
Atraton	$C_9H_{17}N_5O$	212.1506	8.29
Metoxuron	$C_{10}H_{13}ClN_2O_2$	229.0738	8.54
2-4-Dimethylphenylformamide	$C_9H_{11}NO$	150.0913	8.75
Metolcarb	$C_9H_{11}NO_2$	166.0863	8.75
Nicosulfuron	$C_{15}H_{18}N_6O_6S$	411.1081	8.96
Carbofuran	$C_{12}H_{15}NO_3$	222.1125	9.08
Carboxin	$C_{12}H_{13}NO_2S$	236.074	9.26
Fenpropidin	$C_{19}H_{31}N$	274.2529	9.32
Fosthiazate	$C_9H_{18}NO_3PS_2$	284.0538	9.45
Cyprazin	$C_9H_{14}ClN_5$	228.101	9.69
DEET (diethyltoluamide)	$C_{12}H_{17}NO$	192.1383	9.71
Diuron	$C_9H_{10}Cl_2N_2O$	233.0243	9.76
Cycluron	$C_{11}H_{22}N_2O$	199.1805	9.82
Phenmedipham	$C_{16}H_{16}N_2O_4$	301.1183	9.87
Azoxystrobin	$C_{22}H_{17}N_3O_5$	404.1241	10.01
Isoxaben	$C_{18}H_{24}N_2O_4$	333.1809	10.23
Methoxyfenozide	$C_{22}H_{28}N_2O_3$	369.2173	10.27
Chromafenozide	$C_{24}H_{30}N_2O_3$	395.2329	10.42
Metolachlor	$C_{15}H_{22}ClNO_2$	284.1412	10.65
Fenothiocarb	$C_{13}H_{19}NO_2S$	254.1209	10.78
Pencycuron	$C_{19}H_{21}ClN_2O$	329.1415	11.05

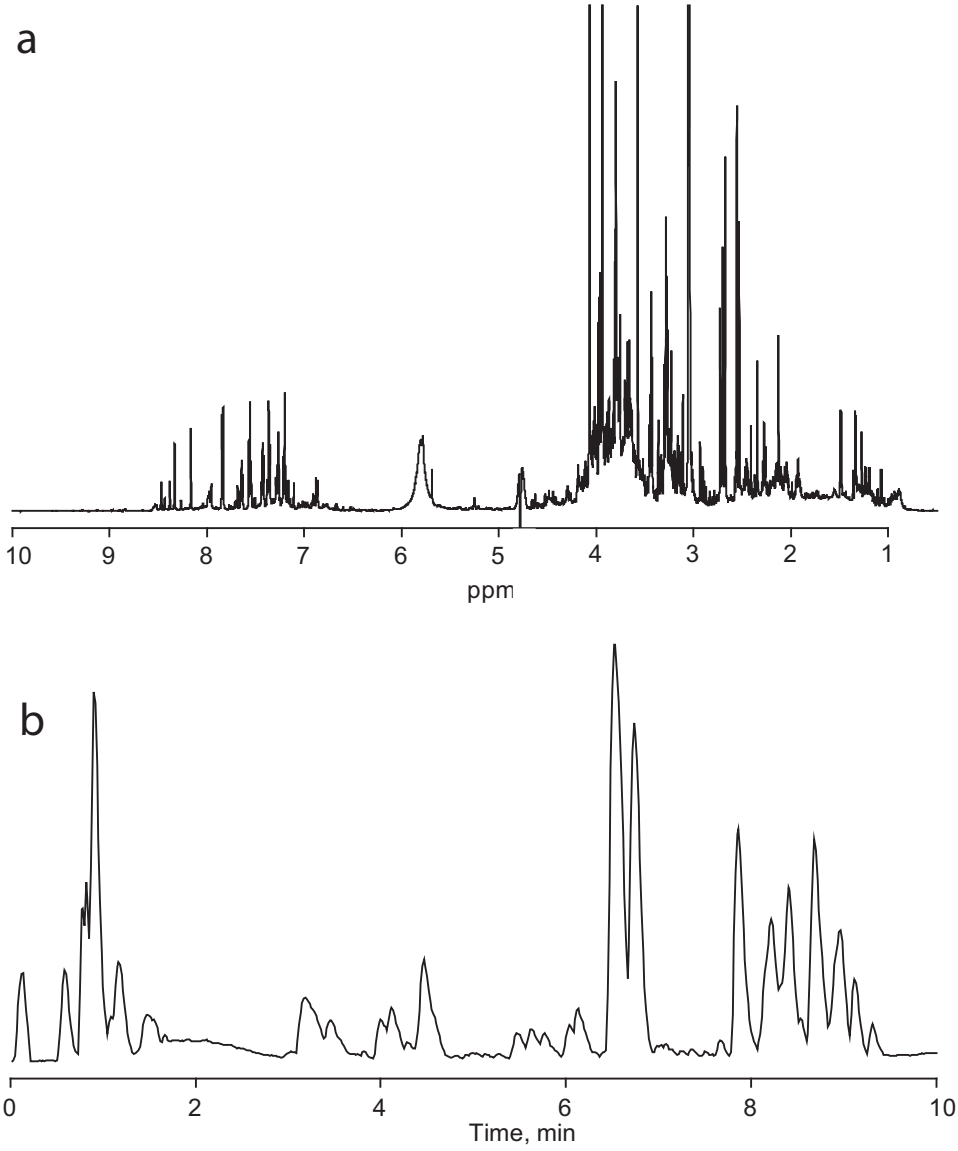

Figure S1. Sample ^{1}H NMR spectrum of urine (a) and LC-MS base peak chromatrogram from the same urine sample (b).

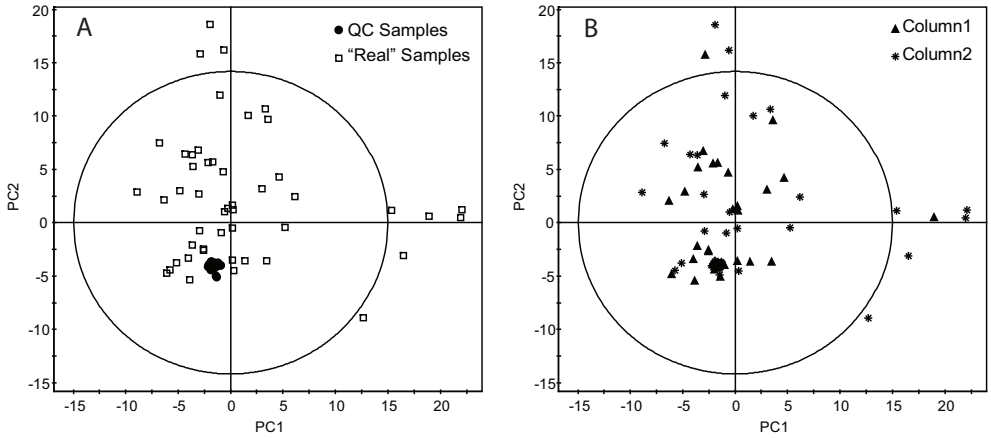

Figure S2. Scores plot of the PCA performed on the entire dataset including QC samples. (A) marked by QCs (●) and individual urine samples (□); QC samples form a tight cluster, indicating the analytical reproducibility of the method. (B) Marked by column; no separation by column is visible.

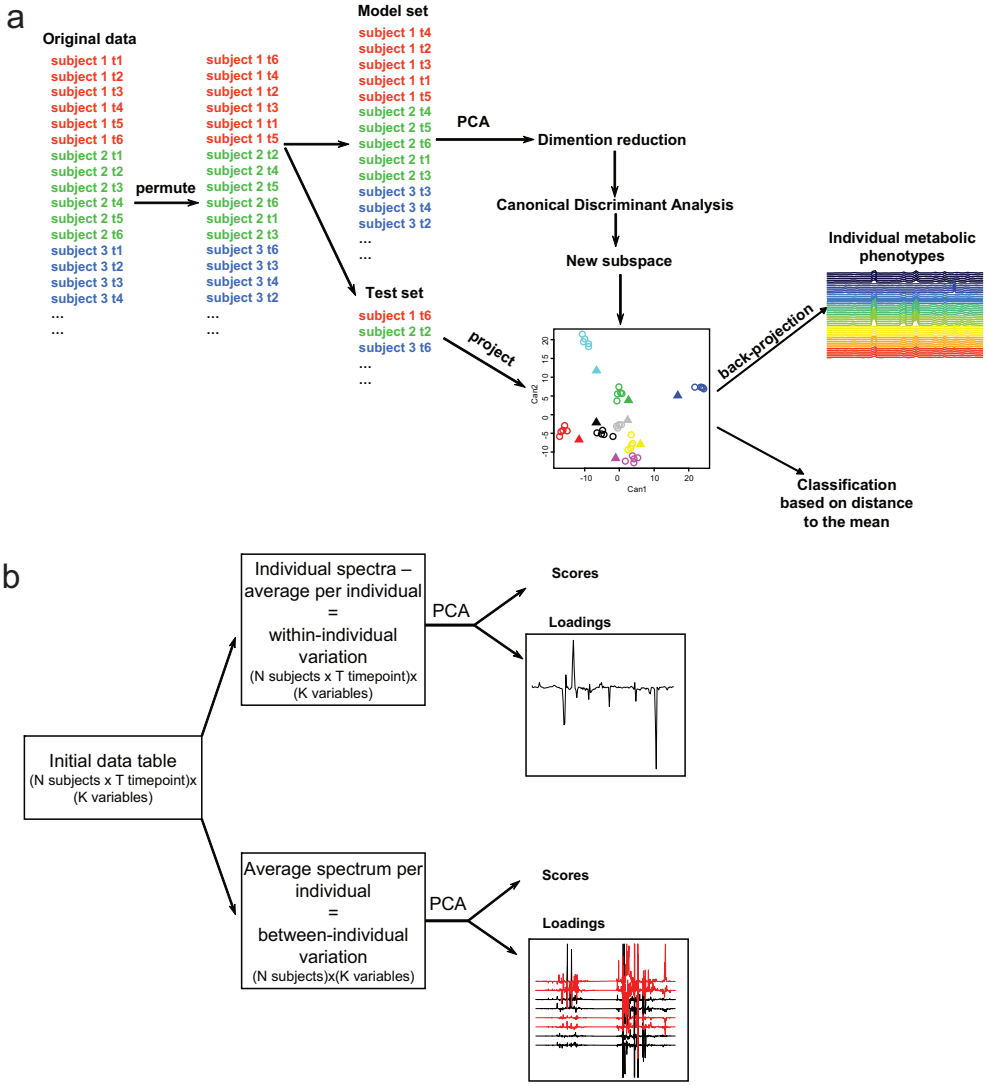

Figure S3. Schematic representation of the two statistical methods used: (a) person recognition, (b) multilevel component analysis.

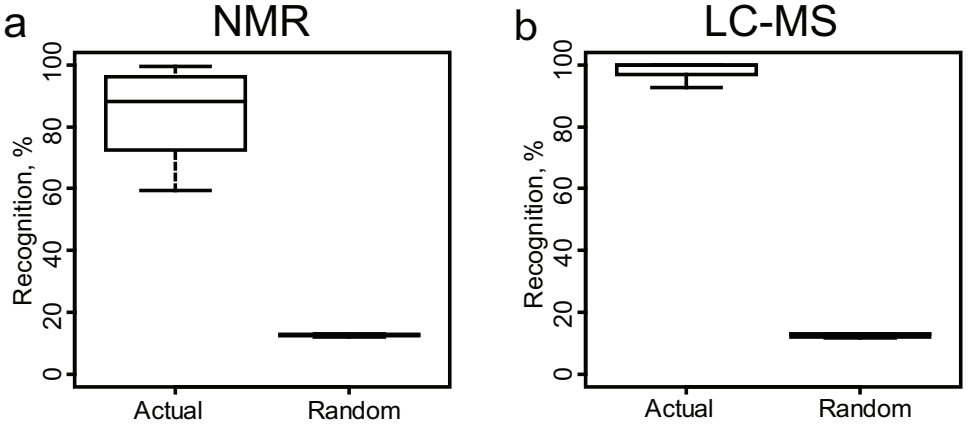

Figure S4. Person recognition based on ^{1}H NMR and LC-MS data of urine samples from 8 individuals. (a) Boxplot of the recognition accuracy based on ^{1}H NMR spectra for actual (left) and permuted (right) person labels. (b) Boxplot of the recognition accuracy based on LC-MS data for actual (left) and permuted (right) person labels.

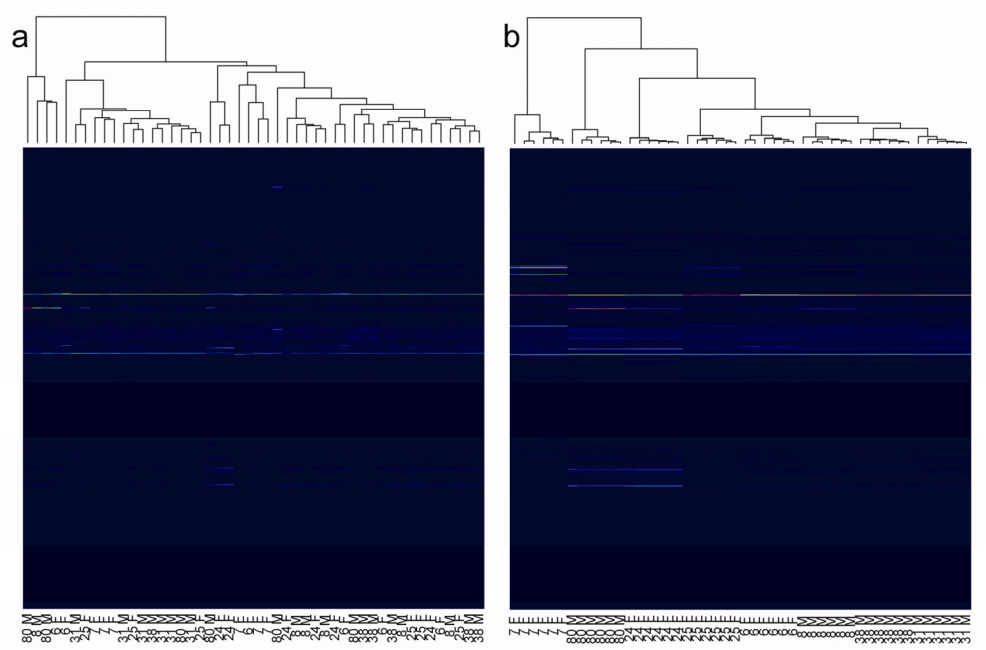

Figure S5. Heatmap and hierarchical clustering of the initial binned ^{1}H NMR table (a) and after back-projection (b). Samples are labeled by their ID and gender.

General Discussion

Metabolomics is an essential part of system biology and as such it can improve our understanding of the complex network of biochemical reactions and regulatory mechanisms in the organism.(1) It also has a great potential in medical research: it can provide new means for diagnosis and prognosis of diseases, dissection of underlying pathology and prediction of treatment outcome.(2;3)

Unlike in other disciplines that contribute to system biology, for instance, in genomics and proteomics, the object of metabolomics studies (metabolites) is characterized by an enormous diversity in physical and chemical properties, as well as in the dynamic range.(4) A universal technology that could cover all the metabolites in a single experiment does not yet exist and it is unlikely that it will be developed in the near future, if ever. Therefore a number of different analytical platforms are used in the field and additional techniques are being developed. Even within the scope of the current thesis we had to use several methods to address various metabolomics questions. Analytical evaluation of the existing and new methods is an essential step that should precede biological experiments. Each of the analytical methods has its own advantages and disadvantages. For example, Nuclear Magnetic Resonance spectroscopy (NMR) is a more universal method in comparison to hyphenated Mass Spectrometry (MS) techniques that are focused on a certain class of metabolites depending on the separation method used (polar compounds in case of capillary electrophoresis (CE), more hydrophobic in case of reversed-phase liquid chromatography (LC) and volatile in gas chromatography (GC) if no derivatization is applied). On the other hand, NMR has orders of magnitude lower sensitivity then MS-based methods.(5) These features have to be taken into account when making a choice for an optimal platform to be used for a specific metabolomics study. This choice should to a large extent be guided by the research question under study and the availability of any prior biological information. However, there might be other more practical factors involved. For instance, if samples are only available in small volumes, this can limit the selection of the analytical method. In this case one of the techniques of choice is CE-MS, which can be used for metabolic profiling for such an extremely small sample, as a single cell.(6) In this thesis we also demonstrated the feasibility of a workflow based on CE-MS for volume-limited samples on the example of urine from mutant mice.

As mentioned above, there is no single method for the fully comprehensive view of the metabolome. Therefore there is an increasing awareness that the data from different platforms have to be integrated in order to increase the coverage of metabolome, enrich biochemical information and generate diagnostic patterns with higher statistical significance. Data integration can be performed at different levels: on the level of raw data tables, on the level of extracted spectroscopic features and on the level of statistics

outputs.(7) A number of methods are available for merging metabolomics data that take into account specific features of this type of data, for instance its multivariate nature.(8;9)

Regardless of the analytical methods used, assessing the quality of the data, as well as minimizing the possible inconsistency of the data, is of great importance. This includes optimization and evaluation of the analytical reproducibility of the method and development of robust pre-processing routines. The latter include alignment of the data that should reduce the shifts in peak position resulting from difference in pH, ionic strength or any other biological matrix characteristics and are necessary for both NMR and hyphenated MS methods. Peak-picking is also an essential step in data pre-processing, along with normalization, transformation and scaling of the data. All these steps reduce the effects that are introduced by both the unwanted biological and analytical variability and prepare the data for statistical analysis.(10)

Despite the extensive developments in analytical, statistical and computer technology, which allow performing large-scale 'omics' experiments, extracting biologically relevant information from such studies is still an art. Metabolomics is no exception to this rule. A correctly and carefully organized study design with well-defined, characterized and matched groups is an essential step towards successful interpretation of the data. Unfortunately, in human studies it is impossible to fully control the influence of the environment and remove all the confounding factors. A study design with longitudinal sampling to a certain extent enables separating the influence of the environment and stable intrinsic profiles and might provide a direction towards personalized medicine and biomarker discovery.

In case of human body fluids interpreting the results of metabolic profiling studies is not straightforward due to the nature of the studied object. The human body comprises an extremely complex network of tissues, organs and microbial communities and the cross-talk between them is still poorly understood. Biological fluids are the "filtrates" of this system, and thus relating their components directly to certain biochemical processes is hardly possible and can be potentially misleading. The use of clinical parameters in combination with complex study design offer possibilities to separate the sources of information in the data and relate metabolites to certain biochemical processes.

Future prospects

The potential benefits of applying metabolomics in medical research are widely accepted. This explains the large number of applications of metabolomics to various pathological conditions. However, investigating the "healthy" metabolome should not be overlooked. Investigating "normality" has been long recognized in psychology as a "subject

really much more fascinating than abnormality, presenting incomparably greater variety and richness of material, and much more worthy of study"(11). Understanding the stability of an organism's metabolome and the limits within which it can change in the healthy state can be very helpful for dissecting the causes that drive it to abnormality, or in other words to disease.

Taking into account the extremely personalized nature of metabolic profiles and the abundance of very subtle, "silent" perturbations of metabolism, which do not bring an organism out of homeostasis and are not easily noticed, it is the question whether a healthy profile can be generalized and determined for a whole population, or whether it can only be referred to an individual. In the latter, "extreme" situation the disease can also only be related to the healthy state of the same individual and typical "case-control" studies would be of little help.

The design of the study aiming to understanding whether "health" is a population- or an individual-based characteristic should on the one hand contain a large number of subjects in order to eliminate the individual-specific variation, and on the other follow them in time to define the borders of "normality" for each person.

Hopefully, the work presented in this thesis can be helpful for the future fundamental metabolomics investigation on health. Certainly, robust analytical methods will always be the basis of such research together with effective pre-processing methods. They will allow focusing on extracting valuable biological information, without interfering analytical variation. We believe that for extracting this information datasets with multiple sampling per individual and multilevel design can be very helpful and thus have presented methods for dealing with such data.

REFERENCES

1. van der Greef,J., Stroobant,P., and van der Heijden,R. 2004. The role of analytical sciences in medical systems biology. Curr. Opin. Chem Biol. 8:559-565.

2. Goldsmith,P., Fenton,H., Morris-Stiff,G., Ahmad,N., Fisher,J., and Prasad,K.R. 2010. Metabonomics: a useful tool for the future surgeon. J. Surg. Res. 160:122-132.

3. Holmes,E., Wilson,I.D., and Nicholson,J.K. 2008. Metabolic phenotyping in health and disease. Cell 134:714-717.

4. Issaq,H.J., Abbott,E., and Veenstra,T.D. 2008. Utility of separation science in metabolomic studies. J. Sep. Sci. 31:1936-1947.

5. Lindon,J.C., and Nicholson,J.K. 2008. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. Trac-Trends in Analytical Chemistry 27:194-204.

6. Chiu,D.T., Lillard,S.J., Scheller,R.H., Zare,R.N., Rodriguez-Cruz,S.E., Williams,E.R., Orwar,O., Sandberg,M., and Lundqvist,J.A. 1998. Probing single secretory vesicles with capillary electrophoresis. Science 279:1190-1193.

7. Roussel,S., Bellon-Maurel,V., Roger,J.M., and Grenier,P. 2003. Fusion of aroma, FT-IR and UV sensor data based on the Bayesian inference. Application to the discrimination of white grape varieties. Chemometrics and Intelligent Laboratory Systems 65:209-219.

8. Forshed,J., Idborg,H., and Jacobsson,S.P. 2007. Evaluation of different techniques for data fusion of LC/MS and H-1-NMR. Chemometrics and Intelligent Laboratory Systems 85:102-109.

9. Smilde,A.K., van der Werf,M.J., Bijlsma,S., van der Werff-van-der Vat, and Jellema,R.H. 2005. Fusion of mass spectrometry-based metabolomics data. Analytical Chemistry 77:6729-6736.

10. Katajamaa,M., and Oresic,M. 2007. Data processing for mass spectrometry-based metabolomics. J. Chromatogr. A 1158:318-328.

11. Pressey,S.L., and Cole,L. 1927. Mental abnormality and deficiency: an introduction to the study of problems of mental health. The Macmillan company. New York. xii pp.

SUMMARY

Metabolomics is the quantitative measurement of metabolites present in biological samples. It can be used to characterize phenotypes that occur due to a certain genetic background, understand the organism's reaction to an intervention or toxins and in biomarker discovery. The metabolomics workflow includes study design, sample collection, data acquisition, data pre-processing, data analysis and biological interpretation of the findings. On the level of study design and sample collection the problems faced in metabolic research are similar to those in other disciplines of systems biology and in biochemical research in general: study groups have to be carefully selected and all the possible bias minimized. Data acquisition in metabolomics poses different challenges in comparison to other 'omics' fields due to the diversity of the chemical and physical properties of metabolites and the absence of a single analytical platform capable of covering the whole metabolome. The choice of the statistical method for data analysis should take specific features of metabolic profiles, such as the megavariate and highly correlated structure of the data, into account. When all the requirements at the different steps of the workflow are fulfilled, metabolomics has a great potential for clinical research not only for traditional "case-control" studies, but also for more "individualized" research ultimately aiming at personalized medicine.

In **Part I** of this thesis novel we presented and evaluated methods for obtaining and pre-processing metabolomics data. **Part II** was dedicated to metabolic profiling in animal models. **Part III** addressed applications of metabolomics in humans.

Part I Method Development

In **Chapter 1** a novel combination of a separation technique (gas chromatography), ionization (atmospheric pressure chemical ionization) and detection method (time of flight mass spectrometry (ToF-MS)) (GC/APCI-MS) was evaluated. The whole analytical method, including derivatization and acquisition parameters, was optimized. A comprehensive mix of analytical standards was used to assess linearity, limits of detection, reproducibility and repeatability of the method. The latter two are of great importance especially for clinical experiments in which cohorts of multiple samples have to be measured. And finally we demonstrated the applicability of the method for one of the biological fluids, namely cerebrospinal fluid (CSF), by showing that it was possible to detect more than 300 different molecules and confident identification of these entities based on the accurate mass and isotopic distribution.

Regardless of the analytical method applied, the generated data can not be directly used for the statistical analysis without prior pre-processing. One of the essential steps of pre-processing is peak alignment, which is necessary for metabolomics data independently of whether it is based on nuclear magnetic resonance (NMR) or hyphenated MS measurements. Various methods exist for chromatographic time correction, but very few of them use available mass information. In **Chapter 2** we introduced such a method that is based on pair-wise matching masses across samples, finding the curve best fitting to the distribution of those matches using a genetic algorithm and correcting the retention/migration times according to the curve. The algorithm can be applied to any hyphenated MS dataset, but in order to demonstrate its utility we chose one of the most difficult cases for the alignment: capillary electrophoresis (CE) coupled to MS. CE is known for its large and irregular shifts in migration time. On a set of electropherograms of mouse urines we demonstrated that our algorithm significantly improved peak positions, which was very advantageous for the subsequent statistical analysis. On the examined dataset the genetic algorithm outperformed one of the often used algorithms for alignment. As mentioned above, the described algorithm is based on matching masses across samples, thus the mass accuracy in the dataset is an important parameter. Though the genetic algorithm was robust enough to fit a curve to the matches in the case of a lower mass accuracy, the improved accuracy of the novel mass analyzer was beneficial for the performance of the alignment.

Part II Application to Animal Studies

As we have demonstrated, the major drawback of CE-MS, the low reproducibility of migration time, can be diminished by the use of advanced data processing methods. Thus, the specific advantages of the technology can be used more efficiently. These include separation efficiency and the small sample volume needed. In **Chapter 3** we demonstrated the feasibility of CE-ToF-MS for metabolic profiling of volume-limited samples. The method was able to efficiently separate compounds belonging to diverse chemical families. We discussed the whole analytical workflow, including sample preparation, analysis and data treatment. The feasibility of the approach was demonstrated on the comparative metabolic analysis of wild type mice versus accelerated aging TTD mutant mice, the latter animals delivering only very limited volumes of urine, which made this study an excellent example for our CE-MS method. The differential compounds were putatively identified using accurate mass information and isotopic distribution and subsequently their identity was confirmed by using MS/MS analysis. Some of these compounds, for instance S-Adenosyl-L-methionine, are associated with oxidative stress defense. They can be of

potential interest as biomarkers of osteoporosis, which is one of the main abnormalities affecting TTD mutants.

Metabolomics studies in experimental animals are of great importance as they allow developing and testing new methodologies and are an essential part of translational medicine. In **Chapter 4** we investigated metabolic perturbations in serum and urine of ERRC1$^{d/-}$ mutant mice, which show accelerated aging, in comparison to wild type controls by profiling with ^{1}H NMR. The advantage of the study set-up was that serum samples were collected from the same mice over time. Such a longitudinal design opened possibilities for monitoring the changes in metabolic profiles in time. We observed that the difference in metabolic composition of serum of mutant and wild type animals becomes more prominent after maturation. The differential molecules were identified in both serum and urine. Interestingly, the changes in both of the biological fluids pointed towards the same phenomenon: ERCC1$^{d/-}$ animals exhibit a similar biochemical phenotype as mice under calorie restricted diet. We found a specific change in the relative abundance of lipoprotein particles in serum (very low and low density lipoproteins decreased and high density lipoproteins increased) that is characteristic for caloric restriction. 3-Hydroxybutyrate, which is a compound released in urine in ketosis, was found in urine of mutant animals and was absent in the urine of wild type animals. This finding is in agreement with previous research done on other levels of biological regulation, such as, for instance, gene expression. Other differences observed in the mutants in our study were related to altered energy metabolism, as well as kidney and liver malfunction.

Using animal studies we demonstrated the applicability of metabolic profiling for the analysis of body fluids and its potential for obtaining biochemical information and discovery of putative biomarkers. The approach is even more beneficial when a complex study design such as a longitudinal one is used. In the following chapters the method was applied to two human studies with such an underlying design.

Part III Application to Human Studies

The object of investigation in **Chapter 5** was Urinary Tract Infection (UTI) which is the most common bacterial infection among adults. Based on the metabolic profiles of urine, obtained with ^{1}H NMR, we were able to identify molecular discriminators of UTI. Using the availability of clinical characteristics of the cohort and the longitudinal design of the study with multiple samples available for UTI patients, we associated some of those discriminators with bacterial contamination of urine (*e.g.* acetate, hippurate and trimethylamine) and found that others might be potential markers of morbidity (*e.g.* para-aminohippuric acid and scyllo-inositol). Thus, such a design offered possibilities for

improved biological interpretation of the data. The samples from the time point at which the patients were considered to be symptom-free were used to test the predictive ability of the discriminative model: most of them were classified as controls, which suggested that our model can be of potential use for predicting the status of new samples. Another time point, at which the patients were under therapy, served as proof that our model was not reflecting therapy, but a real disease-related phenomenon.

We have shown that a longitudinal design improves interpretation of metabolomics studies. With the use of various statistical methods available for the analysis of such a design it is possible to go further and look at the different levels of biological variation present in longitudinal metabolomics data. In **Chapter 6** we demonstrated this approach on a small selection of urine samples that were collected from healthy individuals over a few days. The two statistical methods applied to the data yielded complementary information: individual-specific profiles were produced with person recognition approach, while within- or between-individual variation was recovered with multilevel component analysis. We also showed for the first time that individual profiles are present not only in ^{1}H NMR, but also in liquid chromatography coupled to MS (LC-MS) data. Comparing the two analytical techniques we discovered that they may reflect different biological phenomena.

The chapters that comprised this thesis covered a broad range of subjects from analytical method development to clinical application of metabolic profiling. They were united by the facts that all of these studies aimed at analysis of biological fluids and that the presented methods and approaches may ultimately become parts of a robust metabolomics workflow that might be used in a future personalized medicine.

Nederlandse samenvatting

Metabolomics houdt zich bezig met de kwantitatieve analyse van metabolieten in biologische monsters. Het kan gebruikt worden voor de karakterisering van fenotypen in relatie tot een bepaalde genetische achtergrond, de bepaling van de reaktie van een organisme op een interventie of giftige stof en in de zoektocht naar biomarkers.

Een metabolomics onderzoek start met de opzet van de studie en het verzamelen van de monsters gevolgd door data acquisitie, data voorbewerking, data analyse en biologische interpretatie van de resultaten. Voor wat betreft de opzet van de studie en het verzamelen van de monsters gelden voor metabolomics onderzoek dezelfde regels als voor systeembiologie en biochemisch onderzoek: de studiegroepen moeten zorgvuldig geselecteerd worden en alle mogelijke bias moet zoveel mogelijk geminimaliseerd worden. Als gevolg van de grote verscheidenheid aan fysisch-chemische eigenschappen van metabolieten kent de data acquisitie binnen metabolomics, in vergelijking met andere "omics"-technologiën, grotere uitdagingen omdat er geen enkel analytisch platform beschikbaar is dat al deze metabolieten tegelijkertijd kan analyseren. Bij de keuze van de statistische methoden voor de analyse van metabolomics data moet daarnaast specifiek rekening gehouden worden met het sterk gecorreleerde en megavariate karakter van deze data. Als aan al de bovengenoemde voorwaarden is voldaan, biedt metabolomics grote mogelijkheden binnen het klinisch onderzoek, niet alleen voor traditioneel case-control onderzoek maar ook voor op het individu-gericht onderzoek met als doel om tot een persoonsgerichte geneeskunde te komen.

In **Deel I** van dit proefschrift zijn nieuwe methoden voor het verkrijgen en voorbewerken van metabolomics data gepresenteerd en geëvalueerd. **Deel II** is gericht op metabolomics analyse in diermodellen. In **Deel III** is metabolomics toegepast op patiënten studies.

Deel I Methode ontwikkeling

In **Hoofdstuk 1** is een nieuwe combinatie van technieken voor de scheiding (gaschromatografie), ionisatie (chemische ionisatie bij atmosferische druk (APCI: "Atmospheric Pressure Chemical Ionisation")) en detectie ("Time-of-Flight" massaspectrometrie (ToF-MS)) van metabolieten (GC/APCI-MS) geëvalueerd. Hiervoor hebben we de gehele analytische methode, inclusief de derivatisering en data acquisitie geoptimaliseerd. Een complex mengsel van analytische standaarden is gebruikt om de

linearitеit, gevoeligheid, reproduceerbaarheid en herhaalbaarheid van de methode te bepalen. De twee laatste parameters zijn zeer belangrijk voor klinische studies waarbij grote cohorten met meerdere monsters per individu geanalyseerd moeten worden. Als laatste hebben we de toepasbaarheid van de methode voor de analyse van een lichaamsvloeistof (CSF) aangetoond; meer dan 300 verschillende moleculen konden hierin worden geïdentificeerd op basis van hun accurate massa en isotopenverdeling.

Ongeacht de analytische methode die wordt toegepast kunnen de gegenereerde data niet direct met statistische methoden worden geanalyseerd voordat er een voorbewerking heeft plaatsgevonden. Eén van de essentiële stappen is het uitlijnen (tijds- of frequentiecorrectie ("alignen")) van de gemeten pieken, wat zowel nodig is voor metabolomics data afkomstig van nucleair magnetische resonantie (NMR) metingen als voor MS data afkomstig van systemen waarbij een scheidingstechniek is toegepast. Voor de correctie van de verschillen in chromatografische elutietijden zijn al verscheidene methoden ontwikkeld maar van de beschikbare accurate massa wordt hiervoor slechts zelden gebruik gemaakt. In **Hoofdstuk 2** wordt een dergelijke methode gepresenteerd waarbij drie stappen van belang zijn: het paarsgewijs koppelen van identieke massa's in verschillende monsters, het vinden van de curve die de distributie van deze massa's op basis van een genetisch algoritme het beste beschrijft, en vervolgens het corrigeren van de retentie- c.q. migratietijden op basis van deze curve. In principe is het ontwikkelde algoritme algemeen toepasbaar voor analyses waarbij MS aan een scheidingstechniek gekoppeld is. Wij hebben ervoor gekozen om deze aanpak te testen op een methode waar een relatief grote tijdcorrectie nodig is, capillaire electroforese (CE), vanwege de grote, onregelmatige verschuivingen in migratietijden die inherent zijn aan deze techniek. Gebruikmakend van een set aan electropherogrammen van urines van muizen konden we aantonen dat het algoritme de variatie in de positie van pieken in de verschillende runs sterk vermindert. Dit heeft een groot voordeel voor daaropvolgende statistische analyses. Daarnaast bleek dat dit nieuwe genetische algoritme voor deze dataset duidelijk beter werkte dan één van de tot dan toe meest gebruikte methoden. Zoals hierboven aangegeven, wordt het algoritme gebruikt nadat dezelfde massa's in de verschillende runs aan elkaar gekoppeld zijn. Daardoor speelt de massa-nauwkeurigheid van de dataset een belangrijke rol. Ondanks het feit dat het algoritme robuust is en toepasbaar bleek op datasets met een relatief lage massa-nauwkeurigheid, was de uiteindelijke uitlijning duidelijk beter als er van een nieuwe massaspectrometer met een hogere massa-nauwkeurigheid gebruik werd gemaakt.

Deel II Toepasbaarheid in studies van diermodellen

Zoals boven beschreven kan de belangrijkste tekortkoming van CE-MS, de lage reproduceerbaarheid van migratietijden, sterk verminderd worden door gebruik te maken van geavanceerde datavoorbewerking. Hierdoor kunnen de specifieke voordelen van de techniek, zoals het grote scheidend vermogen en de mogelijkheid tot het gebruik van hele kleine volumina, beter tot hun recht komen. In **Hoofdstuk 3** hebben we aangetoond dat het mogelijk is om CE-ToF-MS te gebruiken voor het genereren van profielen van metabolieten in monsters waarvan de hoeveelheid gelimiteerd is. De methode bleek geschikt om een grote verscheidenheid aan chemische componenten te scheiden. De hele analytische methode is onder de loep genomen, inclusief de monstervoorbereiding, analyse en dataverwerking. Daarnaast hebben we de toepasbaarheid van deze methode aangetoond in een vergelijkende studie, waarbij met deze methode metabolische profielen van wild type en snel verouderende (TTD) muizen zijn gegenereerd. Omdat van de TTD muizen slechts zeer beperkte hoeveelheden urine aanwezig waren, vormde dit een uitstekend model voor onze CE-MS methode. De identiteit van de discriminerende moleculen kon in eerste instantie op basis van accurate massa en isotopenverdeling worden voorspeld en vervolgens worden bevestigd aan de hand van MS/MS experimenten. Sommige van deze componenten, zoals S-adenosyl-L-methionine, worden geassocieerd met de oxidatieve stress response. Het zou mogelijk kunnen dienen als een biomarker voor osteoporose, wat één van de belangrijkste aandoeningen is in TTD muizen.

Metabolomics studies in diermodellen zijn van groot belang omdat ze de mogelijkheid bieden om nieuwe methodologieën te ontwikkelen en te testen, en vormen als zodanig een belangrijke bijdrage in de translationele geneeskunde. In **Hoofdstuk 4** hebben we met behulp van ^{1}H NMR de metabolische veranderingen in serum en urine van mutante, snel verouderende, ERRC$^{d/-}$ muizen onderzocht in vergelijking met wild type muizen. Het voordeel van deze studie was dat op meerdere momenten tijdens de studie serum monsters verzameld en gemeten konden worden. Een dergelijke longitudinale studie gaf nieuwe mogelijkheden om de veranderingen in de metabolische profielen te bestuderen. We konden aantonen dat de verschillen in dergelijke profielen in serum van mutante versus wild type muizen het grootst waren nadat de dieren geslachtsrijp waren. De moleculen die ten grondslag lagen aan de verschillen in de profielen van zowel serum als urine konden worden geïdentificeerd. Het was heel interessant om te zien dat de veranderingen in beide lichaamsvloeistoffen in de richting van eenzelfde verschijnsel duidden, namelijk, dat ERRC$^{d/-}$ muizen een biochemisch fenotype vertonen dat lijkt op dat van muizen op een caloriebeperkend dieet. We vonden o.a. specifieke veranderingen in de relatieve hoeveelheden van lipoproteïne deeltjes in serum (vermindering van heel lage en lage

dichtheid lipoproteïnen en verhoging van hoge dichtheid lipoproteïnen) die karakteristiek zijn voor caloriebeperkende omstandigheden. Daarnaast werd alleen in urine van mutante muizen 3-hydroxybutaraat, een molecuul dat alleen bij ketose in de urine wordt uitgescheiden, aangetoond. Dit is in overeenstemming met eerder onderzoek dat op andere niveaus van biologische regulatie, zoals genexpressie, in dit model zijn uitgevoerd. Andere verschillen die wij in onze analyse vonden waren gerelateerd aan veranderingen in energie metabolisme en verstoringen in nier- en leverfuncties.

Door gebruik te maken van de diermodellen hebben we dus laten zien dat het haalbaar is om metaboliet-profielen van lichaamsvloeistoffen te bepalen en op die manier biochemische veranderingen en potentiële biomarkers aan te tonen. Deze benadering biedt extra voordelen wanneer een complexe studie, zoals een longitudinale, wordt gebruikt en in de daaropvolgende hoofdstukken hebben we dit toegepast binnen een dergelijke studie van patiëntenpopulaties.

Deel III Toepasbaarheid in patiënten studies

In **Hoofdstuk 5** is onderzoek verricht aan urineweginfecties, de meest voorkomende bacteriële infectie bij volwassenen. Door vergelijking van de metaboliet-profielen die werden gegenereerd met behulp van ^{1}H NMR, konden we verschillende moleculen aantonen die specifiek geassocieerd waren met deze infecties. Gebruikmakend van de klinische gegevens van het studiecohort en de longitudinale studieopzet, waarbij meerdere urinemonsters van dezelfde patiënten beschikbaar waren, konden een aantal veranderingen geassocieerd worden met bacteriële verontreiniging van de urine (bv. acetaat, hippuraat en trimethylamine) terwijl andere mogelijke markers zijn voor morbiditeit (bv. para-aminohippurinezuur en scyllo-inositol). Een dergelijke studieopzet geeft dus mogelijkheden voor een meer verfijnde biologische interpretatie van de data. De urinemonsters die werden verzameld op het moment dat de patiënten als symptoomvrij werden beschouwd, werden vervolgens gebruikt om de voorspellende kracht van het model te testen; de meeste werden hierbij geclassificeerd als niet-geïnfecteerde controles wat aangeeft dat ons model mogelijk gebruikt kan worden om de status van onbekende urinemonsters te bepalen. De gegevens van een ander tijdpunt gedurende de behandeling van de patiënten werden gebruikt om aan te tonen dat ons model niet de gebruikte therapie weerspiegelde maar inderdaad de status van de infectie.

We hebben aangetoond dat een longitudinale studieopzet de interpretatie van metabolomics studies verbetert. Door gebruik te maken van verschillende statistische methoden die specifiek geschikt zijn voor een dergelijke studieopzet is het zelfs mogelijk om nog dieper op de data in te gaan en de verschillende niveaus van de biologische variatie die

in metabolomics data aanwezig zijn, te bestuderen. In **Hoofdstuk 6** is deze benadering toegepast op een relatief kleine selectie van urinemonsters die over verschillende dagen bij gezonde personen zijn verzameld. De twee gebruikte statistische methoden leverden complementaire informatie op: persoonsgebonden profielen konden worden gegenereerd door middel van een persoonsherkenning benadering terwijl de intra- en inter-individuele variatie met behulp van een "multilevel component" analyse konden worden geëxtraheerd. We konden hiermee voor het eerst laten zien dat een individu-specifiek profiel niet alleen in ^{1}H NMR maar ook in LC-MS data (vloeistofchromatografie gekoppeld aan massaspectrometrie) aanwezig is. Vergelijking van de twee analytisch technieken liet verder zien dat de technieken mogelijk verschillende verschijnselen weerspiegelen. Onze analyse suggereert dat LC-MS data meer individu-specifieke kenmerken bevat dan ^{1}H NMR data.

De hoofdstukken in dit proefschrift bestreken een breed scala aan onderwerpen met betrekking tot analytische methodeontwikkeling en de klinische toepassing van de analyse van metaboliet-profielen. Bij beide stond de analyse van biologische lichaamsvloeistoffen centraal en hopelijk kunnen de beschreven methoden en benaderingen in de toekomst onderdeel worden van een robuuste metabolomics pijplijn die een belangrijke bijdrage kan leveren aan de ontwikkeling van persoonsgebonden geneeskunde.

ACKNOWLEDGEMENTS

The four years of my PhD-studentship was an incredible experience. It was the time of professional and personal growth. I would like to thank all the people that I met and worked together with during this period.

First of all, my promotor, André. Thank you for giving me the opportunity to join your group and opening possibilities to do research on various subjects and projects. I admire the breadth of your scientific knowledge and the quickness of understanding of any given subject.

Oleg, my co-promotor, I could hardly wish for a better supervisor than you. You are always there to listen to me, you give me freedom, but also guidance. And you give me the confidence that whatever happens you will always be on my side.

I'm very grateful to the committee for the critical evaluation of my thesis.

I would also like to thank all the co-authors of my publications for their contributions and useful comments.

Thanks to all the other colleagues. You are a group of friendly, supportive people with the sense of humor that makes days go easier.

In the first year I worked in the lab, and although it is not the case anymore, I believe that experience was important for better understanding of what I am now doing behind the computer. Rico, Rawi and Bart, thank you for teaching me working with the instruments.

Axel, thanks for what you have taught me about NMR and data analysis and for your careful examination of my results and papers.

I'm very fortunate with my office-mates. Judit, Irina, Paul, Oleg, Tiziana, Anton and previously Alegría, Rocio, Artem, Jean-Marc, Janine, it's been fun! Special thanks to Paul for all the support. If you ever want a career switch, go for tax-advisor for foreigners!

Moving to another country always requires a lot of organizational things to be done. Caroline, your help was splendid!

I'm looking forward to the day of my defense to have such a beautiful paranymph by my side as Sibel. My friend and colleague, thank you for the late evenings in the lab before Christmas, for being a great companion in travelling, shopping and talking about everything.

There are many more people I would like to mention and though here I cannot dedicate some words for each and every one of them, I hope they know how much I appreciate belonging to this group: Alex, Aswin, Carolien, Crina, Dick-Paul, Dorien, Emrys, Gerhild, Hans, Kate, Kristell, Liam, Magnus, Manfred, Marco, Martin, Maurice, Ollie, Ralf, Rene, Rob, Ron, Simone, Suzanne, Yassene, Yuri.

In the Netherlands I met a lot of incredible people and I am very happy to have them as friends. The sad thing is that some of them leave the country after a while, however it's great that we stay in touch: Lisa, Roman, Polyna, I miss you!

I'm very grateful to Anton for the cover of this thesis. You have a great talent!

Dmitry, two of these four years we spent together and there is and always will be a special place for you in my heart.

My other paranymph, Eric, you are, I guess, the first person I met in the Netherlands, besides the colleagues. Being friends with you is a lot of fun!

It is incredible that, despite the distance, I still have close connection with my friends in Russia. Ромчик, спасибо тебе за твою дружбу и доверие. Катя и Никита, мои школьные друзья, с вами всегда тепло и уютно; я так рада, что у вас в семье скоро будет пополнение!

My dearest aunt and uncle, Lana and Ken, you are far away, over the Atlantic Ocean, still I always feel your support and care. I love you and miss you.

Мои дорогие мамулечка, папулечка, Лёнечка, бабушка, я вас очень люблю. Спасибо, что вы у меня есть, что поддерживаете и гордитесь мной. Безумно по вам скучаю.

Katja
Leiden, 2011

CURRICULUM VITAE

Ekaterina Nevedomskaya was born on June 2, 1985, in Moscow, Russia. After finishing secondary school in Chernogolovka, Moscow region, in 2002, she continued her education at the newly established Faculty of Bioengineering and Bioinformatics at Moscow State University. During her studies she was, together with nine other students, selected for a summer internship at the Leiden University Medical Center (LUMC), the Netherlands, where she spent one month in 2005 in the Department of Parasitology, working on the pre-processing of capillary-electrophoresis coupled to mass-spectrometry data under the supervision of Dr. O.A. Mayboroda and Dr. A. Henneman. In 2007 she graduated *cum laude* from Moscow State University. The same year she started her PhD studies in the LUMC, at the Department of Parasitology, under the supervision of Prof. Dr. A.M. Deelder and Dr. O.A. Mayboroda. Her research resulted in the current thesis entitled "Metabolomics of biofluids: from analytical tools to data interpretation". Since September 2011 she works as a post-doc in the same department, continuing her exploration of the analysis of metabolomics data, in part continuing the research initiated in collaboration with Prof. Dr. T.W.J. Huizinga (Dept. Rheumatology) and Prof. Dr. P. Slagboom (Dept. Molecular Epidemiology) on osteoarthritis and rheumatoid arthritis.

LIST OF PUBLICATIONS

Nevedomskaya E., Pacchiarotta T., Artemov A., Meissner A., van Nieuwkoop C., van Dissel J.T., Mayboroda O.A., Deelder A.M. Integrating study design and clinical data into metabolic profiling of urinary tract infection. *Manuscript in preparation.*

Pacchiarotta T., Hensbergen P.J., Wuhrer M., van Nieuwkoop C., **Nevedomskaya E.**, Derks R., Schoenmaker B., Koeleman C.A.M., van Dissel J.T., Deelder A.M., Mayboroda O.A. Fibrinogen alpha chain O-glycopeptides as possible markers of urinary tract infection. *Submitted for publication.*

Nevedomskaya E., Mayboroda O.A., Deelder A.M. Cross-platform analysis of longitudinal data in metabolomics. *Mol. Biosyst.*, **2011,** DOI: 10.1039/C1MB05280B

Ramautar R., **Nevedomskaya E.**, Mayboroda O.A., Deelder A.M., Wilson I.D., Gika H.G., Theodoridis G.A., Somsen G.W., de Jong G.J. Metabolic profiling of human urine by CE-MS using a positively charged capillary coating and comparison with UPLC-MS. *Mol Biosyst.*, **2011**, 7(1):194-9.

Pacchiarotta T., **Nevedomskaya E.**, Carrasco-Pancorbo A., Deelder A.M., Mayboroda O.A. Evaluation of GC-APCI/MS and GC-FID as a complementary platform. *J Biomol Tech.*, **2010**, 21(4):205-13.

Nevedomskaya E., Ramautar R., Derks R., Westbroek I., Zondag G., van der Pluijm I., Deelder A.M., Mayboroda O.A. CE-MS for metabolic profiling of volume-limited urine samples: application to accelerated aging TTD mice. *J Proteome Res.*, **2010**, 9(9):4869-74.

García-Villalba R., Carrasco-Pancorbo A., **Nevedomskaya E.**, Mayboroda O.A., Deelder A.M., Segura-Carretero A., Fernández-Gutiérrez A. Exploratory analysis of human urine by LC-ESI-TOF MS after high intake of olive oil: understanding the metabolism of polyphenols. *Anal Bioanal Chem.*, **2010**, 398(1):463-75.

Nevedomskaya E., Meissner A., Goraler S., de Waard M., Ridwan Y., Zondag G., van der Pluijm I., Deelder A.M., Mayboroda O.A. Metabolic profiling of accelerated aging ERCC1$^{d/-}$ mice. *J Proteome Res.*, **2010**, 9(7):3680-7.

Carrasco-Pancorbo A., **Nevedomskaya E.**, Arthen-Engeland T., Zey T., Zurek G., Baessmann C., Deelder A.M., Mayboroda O.A. Gas chromatography/atmospheric pressure chemical ionization-time of flight mass spectrometry: analytical validation and applicability to metabolic profiling. *Anal Chem.*, **2009**, 81(24):10071-9.

Nevedomskaya E., Derks R., Deelder A.M., Mayboroda O.A., Palmblad M. Alignment of capillary electrophoresis-mass spectrometry datasets using accurate mass information. *Anal Bioanal Chem.*, **2009**, 395(8):2527-33.

Ramautar R., van der Plas A.A., **Nevedomskaya E.**, Derks R.J., Somsen G.W., de Jong G.J., van Hilten J.J., Deelder A.M., Mayboroda O.A. Explorative analysis of urine by capillary electrophoresis-mass spectrometry in chronic patients with complex regional pain syndrome. *J Proteome Res.*, **2009**, 8(12):5559-67.