University of Florence

International Doctorate in Structural Biology

Cycle XXIII (2008-2010)

Evaluation of the meaning of the NMR metabolic profiles of healthy and sick subjects

Ph.D. thesis of

Patrizia Bernini

Tutor

Prof. Claudio Luchinat

Coordinator

Prof. Ivano Bertini

S.S.D. CHIM/03

This thesis has been approved by the University of Florence,
the University of Frankfurt and the Utrecht University

The Sleep of Reason Produce Monsters
Francisco de Goya

Si studia dai sette ai settant'anni e oltre…
Il mio maestro

1. Introduzione

Metabolomics in systems biology

"*There is an epoch in the growth of a science during which facts accumulate faster than theories can accommodate them.*"[1]

"*But one thing is clear: to understand the whole one must study the whole*."[2]

With the onset of full-scale DNA sequencing projects, scientists began to look at organisms in an entirely and globally new way. In contrast to last century's paradigm, genes are not the only target of investigations, and sequencing is just the first step in order to understand organisms at a molecular level. To really understand the function of genes, biological systems have to be analyzed at multiple levels of control both following external parameters (environment, developmental stage, molecular signals, etc.) and internal parameters (transcription and mRNA degradation, post-trascriptional modification, protein dynamics, metabolite concentrations, etc.)[3]. Therefore the decoding of human and other mammalian genomes caused a strong evolution in the field of molecular biology and new "omics" sciences born. The first born of these new sciences were transcriptomics and proteomics, which studies respectively gene expression and protein expression in different disease state, subjects, organ, tissues or cell line[4,5]. In this context both metabolomics and metabonomics are defined (see **Figure 1**).

The distinction between metabolomics and metabonomics is mainly philosophical, rather than technical[6]. The term metabonomics is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"[7]: the goal is understanding systemic change through time in complex multi cellular systems[6]. The term metabolomics was introduced later as the "comprehensive and quantitative analysis of all metabolites" in a system[8]: the aim is to characterize and quantify all the small molecules in a sample. In practice, the terms are often used interchangeably, and the analytical and statistical procedures are almost the same[6].

There were two, largely independent, starting points of metabolomics. In 1960s the idea was born of quantifying metabolite concentrations, and different methods were proposed in order to do that: gas chromatography (GC) and GC coupled to mass spectrometry (MS)[6] are the most important examples. The second starting point can be identified in the development of nuclear magnetic resonance (NMR) spectroscopy. Instead, by the mid-1980s, ^{1}H NMR was sensitive enough to identify metabolites in unmodified biological fluids[9].

Metabonomics beautifully agrees with the spirit of systems biology, because it provides a global overview of biochemistry in complex organisms, that is the final product of the cascade of interactions that take place initially between genes, therefore between proteins, and finally between metabolites in individual cell types. A problem of this approach is that each level of this cascade (gene expression, protein expression and metabolism) operates on a markedly different timescale from the others, making it difficult to find causal linkages. Moreover, environmental and lifestyle factors greatly influence metabolism, making it difficult to separate their effects from gene-related effects[6].

For a long time it was believed that the relationship between genes and environmental factors was essentially additive: genotypes explained a certain fixed amount of the population variance in disease prevalence and when environmental factors were added to genotype, the sum of heredity and environment gave the visible phenotype

$$V_{phenotype} = V_{genotype} + V_{enviroment}$$

The problem with this approach is that linear equations do not accurately reflect the complexity of nonlinear interactions at a molecular level. Most environmental factors cause epigenetic changes that affect multiple systems simultaneously[10].

Metabonomics solves these problems by monitoring the global outcome of all the influencing factors, without making assumptions about the effect of any single contribution to that outcome[6].

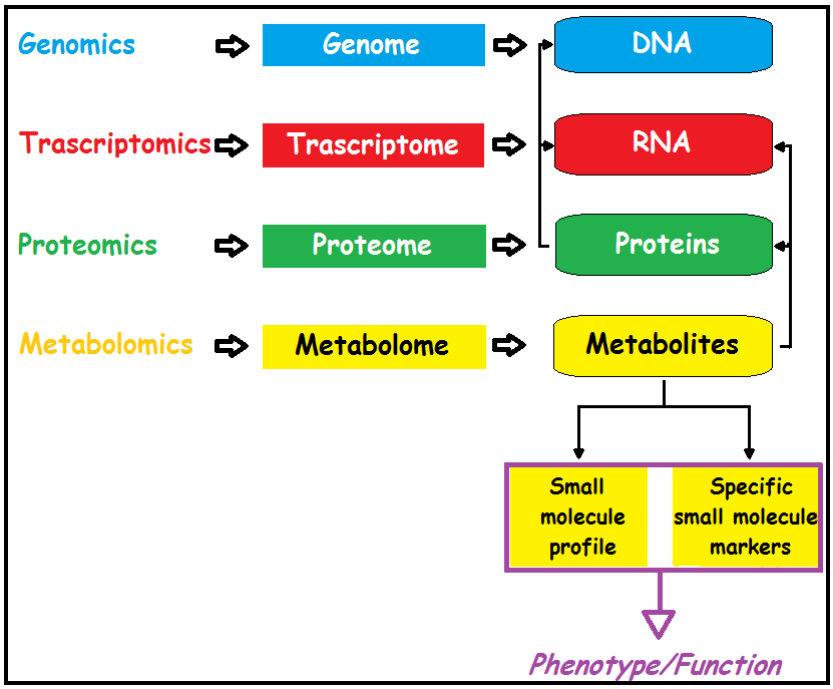


Figure 1 The "omics" sciences

Most metabolomics studies involve common biofluids as urine and serum/plasma that are easily obtainable from mammalians, especially from humans. Moreover, these types of biofluids are obtained in a non- or poorly-invasive way and they are easy to find and collect because they are commonly used for many other biological analyses. Hence they can be used for disease diagnosis and in a clinical trials setting for monitoring drug therapy. However, all biological fluids are useful for metabolomics analysis, as cerebrospinal fluid, synovial fluid, exhaled breath condensate, saliva and so on[11]. There are some problems associated to the use of these peculiar fluids for which they are not normally analyzed: poor collectable quantity, highly-invasive extraction techniques, poor metabolites information carried on. In addition, some metabolomics studies use tissue samples and their aqueous or lipid extracts[12] or *in vitro* cell systems[13]. The number of different metabolites in these human fluids is unknown; estimates range from a minimum of 2,000 to 3,000 to a maximum of around 30,000 metabolites, compared with an estimated 20,000-25,000[14] genes and about 40,000 proteins[15]. Small metabolites, that are low-molecular weight (usually less than 1 kDa[16]) compounds that serve as substrates and products in various metabolic pathways, are of particular interest for metabolomics researchers. These small molecules include compounds such as lipids, sugars, and amino acids, as well as bioactive products acting at very low concentrations in tissue signaling functions[17].

Analytical technologies in metabolomics

The main analytical techniques employed in metabolomics are NMR spectroscopy and MS, but optical techniques, such as Raman spectroscopy and infrared spectroscopy, can also be employed[37]. Both MS and NMR methods provide information on a wide range of metabolites in a single measurement, without having to preselect analytes to detect. Furthermore, both can be used to identify the metabolites' structures and to measure the relative and absolute concentrations of the molecules (although NMR is more reliable for determining concentrations)[6].

NMR spectroscopy is a non-destructive and high reproducible technique that is generally used to detect ^{1}H nuclei in metabolites providing detailed information on molecular structure of both pure compounds and complex mixtures[18]. In a typical biological-fluid sample, all hydrogen-containing molecules in the sample will give an ^{1}H NMR spectrum, as long as they are present in concentrations above the detection limit. The NMR spectrum of a biological fluid is therefore the superposition of the spectra of thousands different small molecules (up to

2500 for urine, up to 200 for serum/plasma) present in sample at concentration >1 μM[19] (**Figura 2)**. An advantage of NMR is that the biological fluid requires only a mild physical or chemical treatment prior to the analysis. A typical ^{1}H NMR spectrum is obtained using water suppression techniques and adding TSP (sodium trimethylsilyl [2,2,3,3-$^{2}H_4$]propionate) as internal reference: it contains for urine predominantly sharp lines due to small molecules, while for serum and plasma both broad and sharp signals are present due respectively to macromolecules, as lipoproteins, and low molecular weight metabolites. The broad signal contribution from protein and other high molecular weight components can be removed by applying a Carr-Purcell-Meiboom-Gill (CPMG) filter to a standard 1D sequence[20]. Conversely, signals from small molecules can be removed by appropriate diffusion-based gradient filter editing experiments[21]. One of the principal disadvantages of the NMR approach is the difficult identification of all metabolites in the samples, a process that involves other techniques like two-dimensional NMR experiments (**Figura 3)**. Indeed the ^{1}H NMR of biological fluids is very complex and even though many resonances can be directly assigned on the base of their chemical shifts, multiplicity and by addition test, various two-dimensional NMR experiments are necessary to increase, but not to complete, the identification of biomarkers in biofluids. These 2D experiments include: i) ^{1}H-^{1}H J-resolved (J-res), to attenuate macromolecules signals and to give more information about multiplicity and coupling patterns ii) ^{1}H-^{1}H correlation spectroscopy (COSY) and ^{1}H-^{1}H total correlation spectroscopy (TOCSY), to provide ^{1}H-^{1}H spin-spin coupling connectivities iii) various heteronuclear experiments that use information coming from other types of nuclei as ^{13}C, ^{15}N and, if it is present, ^{31}P, as for example ^{13}C-^{1}H HSQC, to obtain information on the direct coupling ^{13}C-^{1}H.

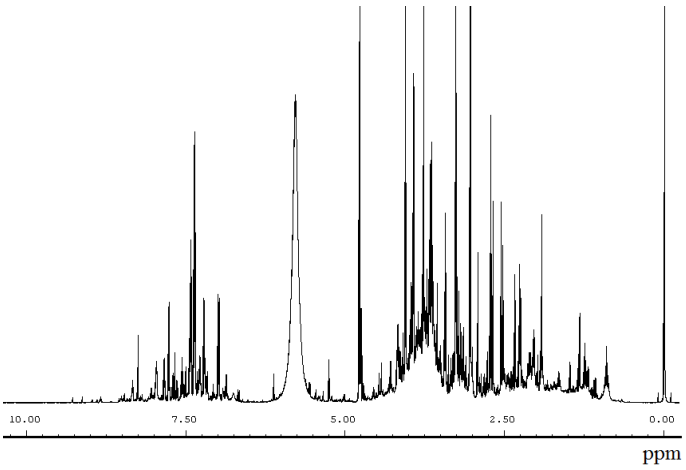


Figure 2 1D ^{1}H NMR NOESY1Dpresat spectrum of urine acquired at 600 MHz

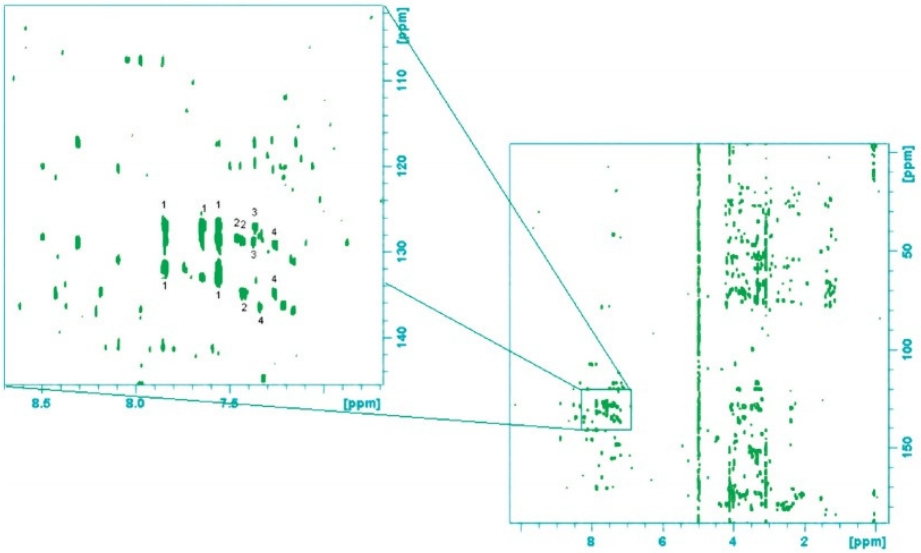


Figure 3 Example of a 2D HMBC spectrum acquired at 900 MHz. The zoomed zone represents the aromatic region. It possible to distinguish some aromatics peaks of Hippurate (1), PAG (2), Histidine (3) and mHPPA (4).

MS is a destructive technique that requires a very low quantity of sample. Over the last few years its application to mammalian study increased, especially for its great sensitivity, higher than NMR, and because it is a major technique for molecular identification, through the use of tandem methods for fragmentation studies or of Fourier transform MS for a very accurate mass determination[3]. As opposite to NMR, MS usually require that the metabolites are separated from the biological fluid before detection, typically by using GC or Liquid Cromatography (LC) (**Figura 4**).

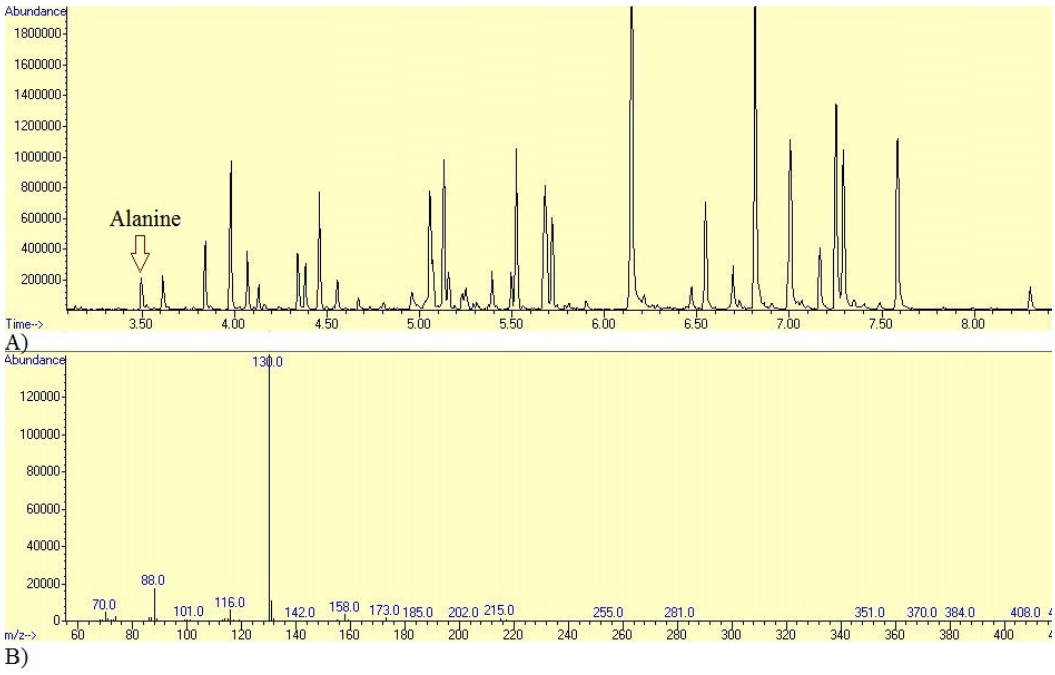


Figura 4 Gas chromatogram of a plasma sample (A) and mass spectrum of alanine

GC-MS is a robust technique for the analysis of volatile and semi-volatile compounds suitable for chemical derivatization to increase their volatility[22]. Derivation methods based on oxime and silylation reactions to replace reactive functional group such as carbonyl, hydroxyl, amino and sulfydryl located in carboxylic acid, amino acid, amine, sugar and steroid metabolites[23,24] have been predominantly employed. Electron Ionization (EI) in GC-MS is quite reproducible[22]. In contrast to GC-MS, LC-MS is especially suitable for the analysis of non-volatile and/or thermally unstable metabolites. The process of ionization in LC-MS, for biofluids as urine, is usually obtained by the electron spray (ESI) method, and then both positive and negative chromatograms are measured. A three-dimensional chromatogram (retention time, mass, intensity) is generated by applying the LC (or GC)-MS approach. The great advantage is the possibility to cut off any mass peak from interfering substance (as for example drug metabolites or contaminant) without altering the dataset[3].

Biological samples contains hundreds of thousands of metabolites which cannot be adequately separated using a single LC or GC experiment to resolve and detect all the metabolites. The separation of the largest number of compounds in a complex mixture can be accomplished using orthogonal multidimensional separation approaches, mostly 2D-cromatography. In 2-D methodology two different mechanism of separation are employed either by selecting two different column chemistries or two different mobile phases[25].

The introduction of UPLC (Ultra Performance Liquid Chromatography) and capillary LC enabled better peak resolution and further increase in sensitivity and speed analysis and it is still successfully applied to metabolomics studies[26]. Another characteristic of capillary LC is lower sample consumption.

Because NMR does not damage analytes, it is particularly useful for studying metabolite levels in intact tissues, such as biopsy samples, which can then be used in further experiments. The non-destructive nature of NMR also enables the dynamics and sequestration of metabolites in tissue samples to be observed; such information is lost in MS experiments, because this technique disrupts the structures and interactions of molecular complexes[6].

The sensitivity problems of NMR are partially solved by using cryogenic probes, where the detector coil and the pre-amplifier are cooled around 20 K, increasing the signal-noise ratio[27]; moreover, the recent development of a technique called high resolution ^{1}H magic angle spinning (MAS) made feasible the acquisition of data on little pieces of tissue without any treatment; indeed, with the rapid spinning of the sample at an angle of 54.7° relative to the

applied magnetic field, the line broadening effects and the associated loss of information[28,29,30] are reduced.

HR-MAS, in combination with multivariate analysis, clearly discriminates the metabolic profiles of cancer tissue and normal and/or adjacent tissue in different cancer types such as breast[31], colorectal[32], prostate[33], and cervical cancer[34].

The integration of multiple metabolic profiling techniques will provide complementary information. Indeed, some "hyphenated" approaches like HPLC-NMR-MS[35], in which each eluting HPLC peak is split to enable a parallel analysis by NMR and MS, are still used for a complete metabolic identification.

Metabolomics and disease

Metabolites serve critical roles in biology due to their involvement in cellular and physiological energetics, structure, and signaling. Fatty acids contribute, for instance, to all three of these central processes through fatty acid metabolism, cellular membrane formation, and nuclear receptor activation, respectively. Similar functions can be delineated for different types of metabolites, such as amino acids, sugars, and many other metabolite classes[36]. The idea that various disease states can be reflected by changes in metabolite concentrations is fundamental to metabolomics, and some has dated the principle back to the ancient Greeks[6]. Analyzing metabolic differences between unperturbed and perturbed systems, such as healthy volunteers and patients with a disease or exposed to a specific treatment, can lead to insights into the underlying pathology[37].

Within disease diagnosis metabolomics and metabolic profiling can be utilized in a number of ways. However, two extremes can be single out. On one side it is possible to base a diagnosis method purely on metabolic fingerprinting by applying pattern recognition techniques to determine if a suspected patient has the disease. This does not necessarily require much knowledge about the underlying metabolic perturbations leading to the diagnosis. If a method is developed that is sufficiently specific this strategy can potentially be useful in fast screening of larger patient populations. The other extreme case consists of using metabolomics strategies to find a single biomarker that is sufficiently specific to confidently detects the investigated disease. In reality, the successful approach will be dictated by the specific project aims and will often lie in between the two extremes. In the majority of metabolomics studies, more than a single metabolite was necessary in order to provide

acceptable specificity. This means that the diagnosis have to be based either on the whole metabolomics dataset or on a selection of biomarkers (biomarker profile), which can be quantified in a reliable way[37].

The use of sensitivity and specificity[38] for reporting results provides an easy way of comparing results between alternative methods, as well as makes it possible to assess the results even if one is not familiar with the methods used in producing the results. Results from metabolomics studies should always be seen in the light of the methods presently available: Some diseases are very difficult to diagnose, and here even methods with moderate classification success can be of interest. On the other hand a metabolomics method that diagnoses for example high blood pressure with 100% sensitivity might not in itself be very interesting to the clinical community, since this condition can be diagnosed much easier and by less expensively way.

One of the major diseases classes targeted by metabolomic researches is cancer. Metabolomics can be used for diagnosing cancer states directly in body fluids, with the aim of facilitating fast screening, or it can be applied to histological samples in order to classify different tumors using techniques such as magic angle spinning (MAS)-NMR[39]or metabolite extraction followed by MS analysis. However, one of the main goals within cancer metabolomics is to develop fast and reliable methods for non-invasive diagnosis by means of urine and blood derivates. The use of metabolomics within cancer research has been discussed in several reviews[40,15,41,42,43,44,45]. The existence of studies applying metabolomics to liver [46,47,48], bladder[49], prostate[50,51], oral[52,53,54], stomach[55], renal[56,57,58], brain[59,60,61] and colorectal cancer[62,63,64,65], as well as a selection of women's cancer[66,67,68,69,70] demonstrates the wide application potential of these methods in oncology.

Another cornerstone is the uses of metabolomics in diabetes diagnosis, treatment and research[71,72] Two different types of diabetes exist. Type 1 diabetes is characterized by lack of insulin, due to an autoimmune destruction of the insulin producing beta-cells in the pancreas. Type 2 diabetes is, at least at early stages, characterized by insulin resistance and decreased excretion, and is now becoming one of the major health problems in developed countries due to lifestyle factors[73] Early diagnosis is of big importance in both cases, since many of the complications can be avoided or decreased by early intervention: this make these diseases interesting from a metabolomics point of view. Indeed, because they are metabolic diseases,

metabolomics profiles are expected to be sensitive at very early stages. All these observation justify the several are the metabolomics studies on type 1[74,75,76,77] and type 2 diabetes[78,79].

Cardiovascular diseases comprise a group of disease states including a large number of more specific conditions, such as heart failure, coronary heart disease, atheroschlerosis and hypertension[72]. Collectively, diseases of the cardiovascular system are the leading causes of death in many industrialized countries[80,81]. Metabolomics collected a lot of results also in this field: coronary artery disease[82,83,84], cardiac events in diabetic subjects[85], myocardial ischemia[86,87], hypertension[88,89], preeclampsia[90,91].

The nervous system, including the brain, is the most complex biochemical system in humans[92]. Common neurological diseases include depression, schizophrenia, epilepsy, Parkinson's disease, Alzheimer's disease, multiple sclerosis and amyotrophic lateral sclerosis. Usually these diseases lack robust biomarkers; their diagnosis and prognosis are therefore often problematic. Metabolomic approaches have been reported in multiple studies on nervous system diseases: amyotrophic lateral sclerosis[93], mild cognitive impairment[94], Parkinson's disease patients[95], Huntington's disease patients[96], schizophrenia[97,98,99], Alzheimer's disease[100], meningitis[101].

Other metabolobics studies concern respiratory diseases like asthma[102] or chronic obstructive pulmonary disease[103]; gut diseases as celiac disease[104], inflammatory bowel diseases[105,106,107], appendicitis[108]; renal diseases like uremia[109] or glomerulonephritis[110]; other inflammatory diseases like hepatitis B[111] or HIV[112].

The application of metabolomics to all these pathologies showed the potential of metabolomics in identifying disease-related differences in one or more of the numerous endogenous metabolites found in clinical samples as biofluids and tissues. Moreover it is important highlight as global metabolomics analysis of accessible biofluids provides information on all tissues that deliver and obtain metabolites from those fluids[113]. Instead target- or metabolite-specific studies, unlike global metabolomics, searches for specific groups of metabolites (amino acids, lipids, peptides, fatty acids, steroids, sugar, extrogen, etc…) based on a *priori* knowledge of the biological system or metabolomic pathway[114,115].

Future applications

One of the long-term goals of metabolomics is clearly the understanding of the existing relationships between genetic polymorphisms of different individuals and their metabolic fingerprint, in order to completely understand the response of different organisms to external stimuli. Achieving this goal could be very important in the field of *pharmacometabolomics,* leading to personalized healthcare[116]. Thus, an individual's drug treatment could be tailored so as to achieve maximal efficacy and avoid adverse drug reactions. In order to accomplish this purpose it is necessary to exactly determine the genetic and environmental influences on the basal metabolic fingerprint of an individual, since these factors will also influence the outcome of a chemical intervention. In this direction a great improvement is obtained with the demonstration of the existence of unique individual metabolic phenotypes, called *metabotypes*[117,118]

Moreover, there is a wide range of current and emerging applications of metabolomics: i) investigation of invertebrates tissue extracts and biofluids to monitor environmental toxicity[119,120]; ii) investigation of plant biofluids[121]; iii) studying of relationship between specific changes in gene expression and alteration of biochemical process (functional genomics)[122,123]

Chemometrics methods

Once the spectra of each biofluid are collected, it is necessary to analyze these data. Both NMR spectroscopy and MS can be thought of as a multidimensional set of metabolic coordinates, whose values are the spectral intensity at each data point (ppm for NMR and retention time for mass) (**Figura 5**). Starting from these data, the principal aims of metabolomics can be summarized in four goals: i) visualization of overall differences, trends and relationships between samples and variables, ii) determination of whether there is a significant difference between groups, for instance between healthy subjects and sick subjects for a pathology, iii) highlighting all metabolites that are responsible for these differences and iv) construct a predictive model for new samples. Multivariate statistical analyses is the key to achieve these goals.

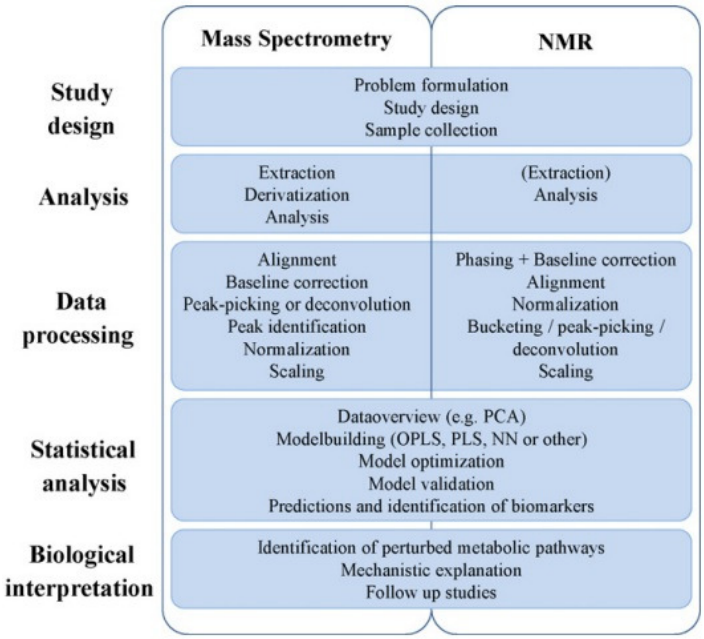


Figura 5 Workflow of metabolomics studies[37]

Starting from NMR spectra, high dimensional data sets are obtained by "bucketing". Bucketing is a procedure used to reduce the total number of variables. One bucket (or bin) is a little slice of the spectrum (**Figura 6**). The corresponding intensity of this slice is calculated in order to obtain our primary variables for each bucket. Obviously, the size of the buckets is one of the parameter to choose: 0.02 ppm is a good compromise between the necessity to reduce residual chemical shift drift and data complexity and the loss of the spectra resolution important for the following biomarker identification[118].

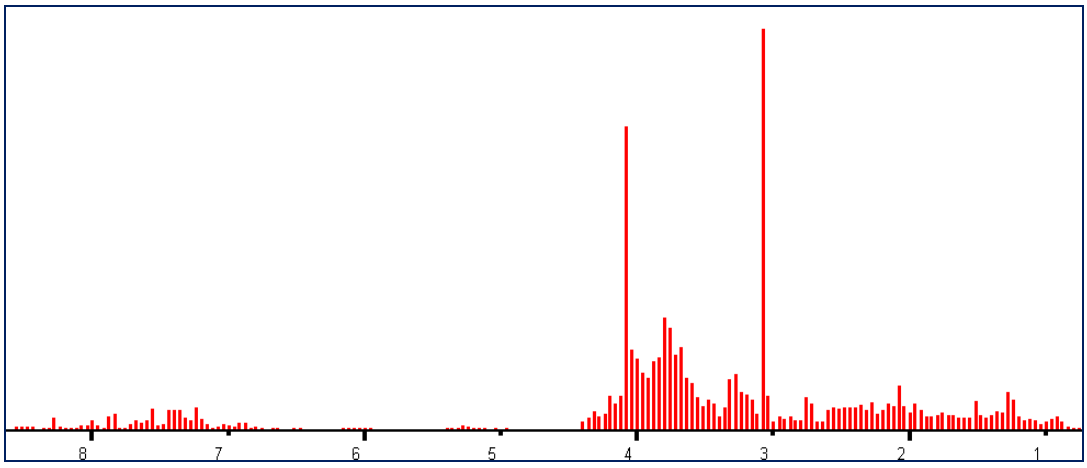


Figura 6 Result of the bucketing of a 1D ^{1}H NMR NOESY1Dpresat spectrum

Observations in a high-dimensional space are necessarily sparser and less representative than those in a low-dimensional space. Often, many of the dimensions in a data set (the measured features) are not useful in producing a model, being irrelevant or redundant. Regression and

classification algorithms may require large amounts of storage and computation time to process raw data, and even if the algorithms are successful the resulting models may contain an incomprehensible number of terms. Because of these challenges, multivariate statistical methods in metabolomics often begin with some type of dimension reduction, in which data are approximated by points in a lower-dimensional space. Dimension reduction often leads to simpler models and fewer measured variables, with consequent benefits when measurements are expensive and visualization is important. This is possible because metabolomics data are often highly collinear. One of the simplest data reduction techniques largely used in metabolomics is PCA (Principal Component Analysis) (**Figura 7**). PCA is a quantitatively rigorous method for achieving this simplification. The method generates a new set of variables, called principal components (PCs). Each PC is a linear combination of the original variables (buckets). All the PCs are orthogonal to each other, so there is no redundant information. There are an infinite number of ways to construct an orthogonal basis for several columns of a dataset: PCA is a linear technique that expresses the maximum of variance in a data set through a small number of PCs. The first PC (PC1) expresses the maximum percentage of variability, then the value of variance quickly decreases in such a way that the first PCs are the responsible for the great part of variability, while the last PCs are significantly less important and express noise variability. The trasformation of data matrix into PCs gives two new matrices: score matrix and loading matrix. The score matrix contains the coordinates of the original data in the new coordinate system defined by the PCs and it is the same size as the input data matrix. The loading matrix contains the coefficients of the linear combinations of the original variables that generate the PCs. Hence, in the loading plot, each point represents a different spectral intensity and the weight of the variables in the discriminate between samples, practically which buckets are responsible for the maximum variance[124] Moreover PCA is used for giving an overview of data, which will reveals outliers, groups and trend of data[125].

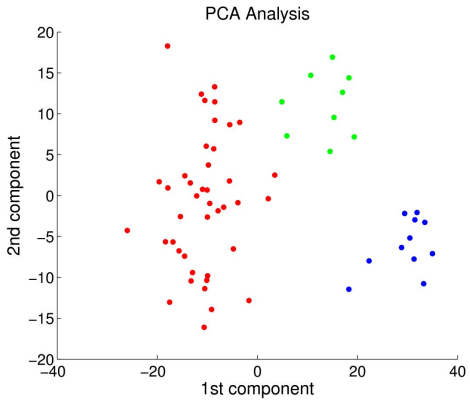


Figure 7 Plot of the first two components after the PCA analysis

After the initial overview of the data the aim is usually to build a model that is useful for classifying new samples, and at the same time allows identification of biomarkers for the studied diseases. Classification can be done using linear methods such Partial Least Square (PLS), OPLS (Orthogonal Projection to Latent Structures method), Linear Discriminant Analysis (LDA) or non linear method such as Self-Organising Maps (SOMs), Artificial Neural Networks (ANNs), Support Vector Machine (SVM) and k-Nearest-Neighbour (k-NN) rule. Linear methods have the advantage that biomarkers are readily identified from the model using the loading values. For non-linear methods it is often impossible to interpret the model biologically, since statistical significance of the coefficients and their direction may not be directly related to the importance of the corresponding variables for the model building[126].

PLS[127] is a technique used with data that contain correlated predictor variables regression such as information about the nature of the samples (healthy or sick in case of studies of pathologies). This technique is between multiple linear regression, that allow to find a combination of the predictors that best fit a response, and PCA, that permit to single out combinations of the predictors with large variance without use response values. PLS finds combinations of the predictors that have a large covariance with the response values by combining information about the variances of both the predictors and the responses, also considering the correlations among them. One recent modification of the PLS method is constituted by the OPLS[128], that is used to remove irrelevant and confusing parameters. The basic idea of OPLS is to separate the systematic variation in data matrix in two parts, one linearly related to predictor variables matrix, and the other unrelated (orthogonal) to it. This partition facilitates model interpretation and model execution on a new set of samples.

LDA[129] is based on finding linear combination of n-dimensional predictor variable values, with large ratios of between group to within group sums of square. LDA analysis constructs, for a two-classes problem, a separating hyperplane between the two datasets. PCA, OPLS and PLS can be used combined with LDA to establish the optimum of discrimination between samples[130].

The analysis of variance (ANOVA) technique takes a set of grouped data and determine whether the mean of a variable differs significantly among groups. Often there are multiple response variables, and the interest is determine whether the entire set of means is different from one group to the next. There is a multivariate version of analysis of variance that can address the problem: MANOVA technique. MANOVA function in Matlab 7.9.0 allows also to perform the Canonical Analysis (CA) that give the canonical component that allow the LDA analysis. PCA looks for the combination of the original variables that has the largest possible variation. MANOVA instead looks for the linear combination of the original variables that has the largest separation between groups. Having found that combination, the base idea is next look for the combination with the second highest separation, and so on[124].

SOMs is an unsupervised method of classification that reduces the dimensionality of the data through the creation of an array of nodes, each one described by a "codebook" vector. During the training, each sample is presented to the map and the node with the closest reference vector is selected as the final node for the sample (also called "winning node"). Clearly the number of chosen nodes is a fundamental variable; with few nodes, for instance, the map does not represent faithfully the data distribution, while with a large number of nodes phenomena as overfitting and noise susceptibility influence the analysis. Moreover, another disadvantage is associated to the final position of similar clusters, that may end up in distant parts of the SOM despite their similarity.

ANNs is one of the most popular supervised method for pattern recognition in biomedical area. It consists of a network of nodes. Each of these nodes performs an operation to give a single output. Theoretically the nodes can be divided into three layers: i) input nodes ii) hidden nodes and iii) output nodes. Generally in metabolomic studies each input node is formed by each single spectrum, while each output node represents each class of samples (even if it is better to have one more output node corresponding to an unknown class). Each hidden node receives some information from various input nodes and re-combines this information using a non-linear activation function to give a signal to the output nodes. For

their great versatility the ANNs are still applied to some metabolomic studies[131,132,133,134]. However, it is often difficult to clearly understand the connection weights to completely single the parameters (buckets in our case) responsible for identification.

In the case of SVM, a data point is viewed as a *p*-dimensional vector (a list of *p* numbers) and we can separate such points with a (*p*-1)-dimensional hyperplane: this is called a linear classifier. There are many hyperplanes that might classify the data: the best hyperplane is the one that represents the largest separation, or margin, between two or more classes in a way that distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier[135].

kNN rule is the simplest of all supervised classification approaches. Every sample is classified to the class most frequently expressed among the *k* nearest neighbour. Therefore the *k* is the fundamental parameter to set. Small values of *k* can lead to the construction of a model subject to significant statistical fluctuations, while large values of *k* reduce statistical errors but can smooth out many details of the class distribution. As *k*-means, this method gives a classification of the sample without associated visualization and interpretation of data and, therefore, it is not applied only in metabolomics[124,136].

Cluster consists in portioning a set of elements into subsets according to the differences between them. In other words, it is the process of grouping similar elements together. The main difference from the supervised classification is that, in clustering, we have no information about how many classes there are. Cluster analysis, also called data segmentation, has a variety of goals all related to grouping or segmenting a collection of objects into subsets or "cluster" in a such way that objects within each cluster are closely related to one other then objects assigned to different clusters. Sometime the goal is to arrange the cluster into a natural hierarchy: this involves successively grouping the clusters themselves so that, at each level of the hierarchy, cluster within the same group are more similar to each other than those in different groups. Central to all goals of cluster analysis is the notion of the degree of similarity (or dissimilarity) between the individual objects clusterized[130].

Hierarchical Cluster Analysis (HCA) is widely used in all areas of science and it has been recently applied to a metabolomic studies[137]. These methods groups data over a variety of scales by creating a cluster tree diagram or dendrogram, in which the relationships between

samples are expressed (**Figura 8**). The algorithm starts analyzing the similarity or dissimilarity between every pair of objects in the data set and calculating the distance between all pairs of data points. Then, it proceeds to group the objects into a binary, hierarchical cluster tree: in this step, this method links pairs of objects using the distance information generated in the previous step to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters (initially each cluster consists of a single data point) until a hierarchical tree is formed. The described algorithm is the agglomerative algorithm, but there is also a divisive algorithm that operates by successively splitting the groups, beginning with a single group and continuing until there are N groups, each of a single individual: generally, divisive algorithm are computationally inefficient[130]. The final step of the algorithm is determine where to cut the hierarchical tree into clusters: these clusters can be created by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The main problem in the application of these methods to metabolomic studies is that the reproducibility is not good; the inclusion of new data, for instance, requires a complete re-computation of the dendrogram that can lead to a new structure not necessarily similar to that generated from the previous training set. Moreover, HCA does not generate diagnostic information about what features are responsible for the classification in sub-clusters.

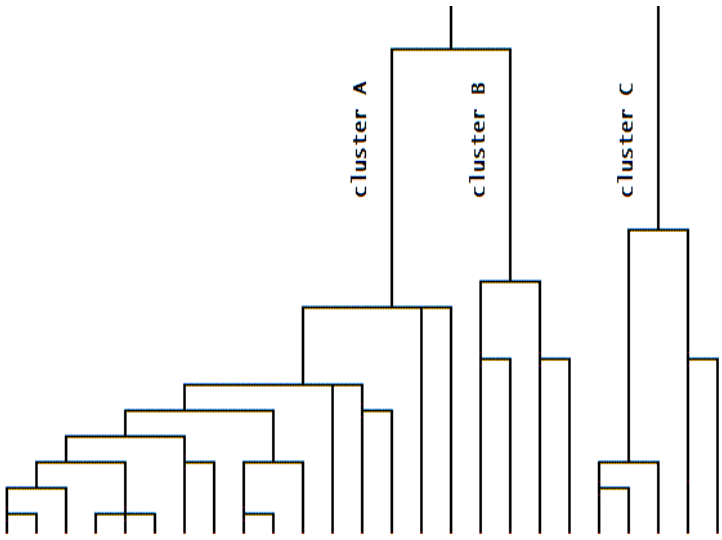


Figura 8 Example of HCA

K-means clustering is an unsupervised algorithm. It is a partitioning method: it partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. It starts to work selecting the desired number of clusters and randomly assigning their center. For each iteration a data point is classified through the assignment to one of the clusters, basing on the closest cluster center. After this addition the new cluster centers are recomputed. Unlike HCA, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters. The distinctions mean that k-means clustering is often more suitable than HCA for large amounts of data. However despite its popularity, k-means is not often applied to metabolomics. The main reasons are the absence of diagnostic tools and associated visualization. In order to avoid these disadvantages this technique can be applied in combination with other statistical procedures, as in a metabolomic study of plant and marine invertebrate extracts, with HCA[124].

Regardless of the chosen method, both statistical and biological validations are critical.
When using metabolomics in diagnosis disease the end result will often be a prediction of class belonging, using a multivariate algorithm. It is important that this prediction is carried out in a relevant way. A model that seems valid can in some cases perform badly when predicting new samples, so it is important that the model is validated. A cross-validation scheme should been employed, to make sure the model is not over fitted. In cross-validation parallel models are build with a part of the data excluded; by subsequently predicting the excluded values it is possible to get a measure of the predictive ability of the model[138]. Cross-validation is a good and widely used tool for optimizing multivariate models, however, it is prone to over fitting of the models-especially when using Leave-One-Out (LOO) cross-validation[139].
In the LOO method, the statistical model is built on the entire dataset except one sample and the sample leave out is after predicted by means of some kind of classification method. The method is repeated for each sample.

A better strategy for validation is therefore to use an external dataset for validation. Preferably a separate follow-up study analyzed separately should be used, since this will reduce the risk of undetected biases during sample collection or analysis, but alternatively a part of the data set can be excluded for validation purposes. This is the idea of the training/evaluation splitting method. In the training/evaluation splitting method the dataset is splitted into two subsets, called “training set” and “validation set”. The training set is used to build the mathematical

model which is used to predict samples belonging to the validation set. Naturally many random and variable training sets and validation sets can be built to correctly and fully investigate the data through a robust model.

Biological validation of models can include studies that independently confirm or reject the assumed hypothesis or provide new biological explanation for the observed initial results.

In order to single out the significant variables for the classification in a multivariate statistical approach, we can use Wilcoxon test or the Kruskal Wallis test according to a p-value ≤ 0.05 or ≤ 0.02.

The Wilcoxon-Whitney test, is a nonparametric test that compares two paired groups. It is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to location (median), if at all. The Wilcoxon Mann-Whitney test does not require the assumption that the differences between the two samples are normally distributed. In many applications, the Wilcoxon Mann-Whitney Test is used in place of the two sample t-test when the normality assumption is questionable.

The Kruskal-Wallis test is a nonparametric test used to compare three or more samples. It is used to test the null hypothesis that all populations have identical distribution functions against the alternative hypothesis that at least two of the samples differ only with respect to location (median), if at all. It is the analogue to the F-test used in analysis of variance. While analysis of variance tests depends on the assumption that all populations under comparison are normally distributed, the Kruskal-Wallis test places no such restriction on the comparison. It is a logical extension of the Wilcoxon-Mann-Whitney Test.

Aim of the research

The general aim of this thesis is to explore the potential of metabolomics for disease prognosis, diagnosis and therapy toxicity assessment and to highlight its necessary and irreplaceable role as a tool for understanding the biochemical mechanisms at the basis of the pathologies. Indeed metabolomics more than other omics sciences can help medical research in this sense by taking an overall picture of the ensemble of the metabolites present in biological fluids in a certain moment and extracting a disease signature from the multivariate analysis of the metabolic profiles.

All the results that I am presenting here are obtained on urine and serum, biofluids that can be collected in a non- or poorly-invasively way and easy to find because commonly used for many other clinical analyses. Considering the invasivity and the consequent stress for patients, especially if suffering of chronic diseases, of several diagnostic tools, this feature is undoubtedly a good point assessed for metabolomics.

A milestone of the whole work relies on the establishment of the existence individual metabolomics signature of healthy subjects found to be stable over the years (Chapter 3). The definition of this unique fingerprint is the starting point because it allows to unravel whether modifications occur in the metabolomic profile have a physiologic or a pathologic origin. Once defined the "solidity" of metabolomic approach in this sense, at least in urine, we defined the metabolomic signature of three large impact diseases: breast cancer (Chapter 4), celiac disease (Chapter 5) and cardiovascular risk (Chapter 6). These pathologies are investigated applying the "holistic" approach (the analysis of the fingerprinting), followed by the research of the metabolites responsible for the pathological signature with respect the "healthy" signature.

Moreover the toxicity of ozone therapy in blood is explored and also the role of metabolomics in the evaluation of toxicity of therapies is highlighted (Chapter 7).

Finally the reliability of the metabolomics approach requires that the chemical nature and the relative concentration of the metabolites present in the biofluids are neither affected by the preanalytical conditions nor by the analytical methodology. The importance of metabolomics is clarified also in the definition of operating procedures aimed at collecting and preserving biological samples according to protocols as standardized as possible also that may be shared in biobanking activities and in general for all those studies in which the impact of the results is linked to the large number of the collected samples (Chapter 8).

The specific project aims and results will be extensively treated and discussed in the dedicated chapters.

Bibliography

[1] Medawar P. (1982). Pluto's Republic. *Oxford Univeristy Press*: p29.

[2] Kacser H., Welchand G.R., Clegg J.S. 1986. The organization of cell metabolism. *Plenum Press New York*: p327.

[3] Weckwerth W. (2007). Metabolomics Methods and Protocols. *Humana Press.*

[4] Baldi P., Hatfield G.W. (2002). DNA microarrays and gene expression. *Cambrige University Press*: p230.

[5] Wasinger V.C., Cordwell S.J., Cerpa-Poljak A., Yan J. X., Gooley A.A. et al. (1995). Progress with gene-product mapping of the Mollicutes: Mycoplasma Genitalium. *Electrophoresis* 16(7): 1090-1094.

[6] Nicholson J.K., Lindon J.C. (2008). Metabonomics. *Nature* 455: 1054-1056.

[7] Nicholson J.K., Lindon J.C., Holmes E. "Metabonomics" : understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29: 1181-1189.

[8] Fiehn O. (2002). Metabolomics – the link between genotypes and phenotypes. *Plant Mol Biol*: 155-171.

[9] Nicholson J.K., Wilson I.D. (1989). High-resolution proton magnetic resonance spectroscopy of biological fluids. *Prog NMR Spectros.* 21: 449-501.

[10] Knox S.S. (2010). From 'omics' to complex disease: a systems biology approach to gene environment interactions in cancer. *Cancer Cell Int* 10: 11-23.

[11] Lindon J.C., Nicholson J.K., Everett J.R. NMR spectroscopy of biofluids. *Ann. Reports on NMR Spectrosc* 38: 1-88.

[12] Mayr M., Chung Y.L., Mayr U., Yin X.K., Ly L. et al. (2005). Proteomics and metabolomic analyses of atherosclerotic vessels from apolipoprotein E-deficient mice reveal alterations in inflammation, oxidative stress, and energy metabolism. *Arterioscl Thromb Vasc Biol* 25: 2135-2142.

[13] Waters N.J., Holmes E., Waterfield C.J., Duncan Farrant R., Nicholson J.K. (2002). NMR and pattern recognition studies on liver extracts and intact livers from rats treated with α-naphthylisothiocyanate. *Biochem Pharmacol* 64: 67-77.

[14] Levy S., Sutton G., NG P.C., Feuk L., Halpern A.L. et al. (2007). The diploid genome sequence of an individual human. *PLOS boil*, 5(10): 2113-2144.

[15] Browsing NCBI on 22/10/2010.

[16] Samuelsson L.M., Larsson D.G. (2008). Contributions from metabolomics to fish research. *Mol Biosyst* 84(10):974-9.

[17] German J.B., Hammock B.D., Watkins S.M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* 1: 3-9.

[18] Claridge T.D.W. (2004). High-Resolution NMR techniques in organic chemistry. *Elsevier*: p384.

[19] Wishart D.S., Tzur D., Knox C., Eisner R., Guo A.C. et al. (2007). HMDB: the Human Metabolome Database. *Nucleic Acids Res* 35: D521-D526.

[20] Meiboom S., Gill D. (1958). Modified spin-echo method for measuring nuclear relaxation times. *Rev Sci Instrum* 29:688-691.

[21] Wu D.H., Chen A.D., Johnson C.S. (1996). Three-dimensional diffusion-ordered NMR spectroscopy: The homonuclear COSY-DOSY experiment. *J Magnen Reson A* 123:215-218.

[22] Wu Z., Huang Z., Lehmann R., Zhao C., Xu G. (2009). The application of Chromatography-Mass spectrometry: methods to metabonomics. *Chromatographia* 69: S23-S32.

[23] Fiehn O., Kopka J., Dormann P., Altmann T., Trethewey R.N. et al. (2000). Metabolite profiling for plant functional genomics. *Nat Biotechnol* 18: 1157-1161.

[24] Jiye A., Trygg J., Gullberg J., Johansson A.I., Jonsson P. et al. (2005). Extraction and GC/MS analysis of the human blood plasma metabolome. *Anal Chem* 77(24): 8086-8094.

[25] Issaq H.G., Van Q.N., Waybright T.J., Muschik G.M., Veenstra T.D. (2009). Analytical and statistical approaches to metabolomics research. *J Sep Sci* 32: 21832199.

[26] Wilson I.D., Nicholoson J.K., Castro-Perez J., Granger J.H., Johnson K.A. et al. (2005). High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *J Proteome Res* 4: 591-598.

[27] Keun H.C., Beckonert O., Griffin J.L., Richter C., Moskau D. et al. (2002). Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies. *Anal Chem* 74: 4588-4593.

[28] Tomlins A., Foxall P.J.D., Lindon J.C., Lynch M.J., Spraul M. et al. (1998). High resolution magic angle spinning ^{1}H nuclear magnetic resonance analysis of intact prostatic hyperplastic and tumour tissues. *Anal Commun* 35: 113-115.

[29] Garrod S.L., Humpfer E., Spraul M., Connor S.C., Polley S. et al. (1999). High-resolution magic angle spinning ^{1}H-NMR spectroscopic studies on intact rat renal cortex and medulla. *Magn Res Med* 41: 1108-1118.

[30] Cheng L.L., Chang I.W., Louis D.N., Gonzalez R.G. (1998). Correlation of high-resolution magic angle spinning proton magnetic resonance spectroscopy with histopathology of intact human brain tumor specimens. *Cancer Res* 58: 1825-1832.

[31] Bathen T.F., Jensen L.R., Sitter B., Fjösne H.E., Halgunset J. et al. (2007). MR-determined metabolic phenotype of breast cancer in prediction of lymphatic spread, grade, and hormone status. *Breast Cancer Res Treat* 104: 181–189.

[32] Chan E.C Y., Koh P.K., Mal M., Cheah P.Y., Eu K.V. et al. (2009). Metabolic profiling of human colorectal cancer using high resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS). *J Proteome Res* 8: 352–361.

[33] Cheng L.L., Burns M.A., Taylor J.L., He W., Halpern E.F. et al. (2005). Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. *Cancer Res* 65: 3030–3034.

[34] Sitter B., Bathen T.F., Hagen B , Arentz C., Skjeldestad F.E. et al. (2004). Cervical cancer tissue characterized by high-resolution magic angle spinning MR spectroscopy. *Magn Reson Mater Phy* 16: 174–181.

[35] Lindon J.C., Nicholson J.K., Wilson I.D. (2000). Directly-coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development. *J Chromatog B* 748: 233-258.

[36] Vinayavekhin N., Homan E.A., Saghatelian A. (2010). Exploring Disease through Metabolomics. *Chem Biol* 5(1): 91–103.

[37] Madsen R., Lundstedt T., Trygg J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Anal Chim Acta* 659(1-2): 23-33.

[38] Altman D.G., Bland J.M. (1994). Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ* 308: 1552.

[39] Sitter B., Bathen T.F., Tessem M.B., Gribbestad I.S. (2009). High-resolution magic angle spinning (HR MAS) MR spectroscopy in metabolic characterization of human cancer. *Prog. Nucl. Magn. Reson. Spectrosc.* 54**:** 239.

[40] Chung Y.L., Griffiths J.R. (2008). Using metabolomics to monitor anticancer drugs. *Ernst Schering Foundation Symposium Proceedings* 4: 55-78.

[41] Griffin J.L., Kauppinen R.A. (2007). A metabolomics perspective of human brain tumours. *Febs J* 274(5): 1132-1139.

[42] Jain K.K. (2007). Cancer biomarkers: Current issues and future directions. *Curr. Opin. Mol. Ther.* 9(6): 563-571.

[43] Jordan K.W., Cheng L.L. (2007). NMR-based metabolomics approach to target biomarkers for human prostate cancer. *Expert Rev Proteometr* 4(3): 389-400.

[44] Kim Y.S., Maruvada P., Milner J.A. (2008). Metabolomics in biomarker discovery: Future uses for cancer prevention. *Future Oncol* 4(1):93-102.

[45] Serkova N.J., Spratlin J.L., Eckhardt S.G. (2007). NMR-based metabolomics: Translational application and treatment of cancer. *Curr Opin Mol Ther* 9(6)**:** 572-585.

[46] Gao H.C., Lu Q., Liu X., Cong H., Zhao L.C. et al. (2009). Application of ^{1}H NMR-based metabonomics in the study of metabolic profiling of human hepatocellular carcinoma and liver cirrhosis. *Cancer Sci* 100 (4): 782-785.

[47] Yang J., Xu G., Zheng Y., Kong H., Pang T. et al. (2004). Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *J Chromatogr B: Anal Technol Biomed Life Sci* 813(1-2): 59-65.

[48] Yang Y.X., Li C.L., Nie X., Feng X.S., Chen W.X. et al. (2007). Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning ^{1}H NMR spectroscopy in conjunction with multivariate data analysis. *J Proteome Res* 6(7): 2605-2614.

[49] Issaq H.J., Nativ O., Waybright T., Luke B., Veenstra T.D. et al. (2008). Detection of Bladder Cancer in Human Urine by Metabolomic Profiling Using High Performance Liquid Chromatography/Mass Spectrometry. *J Urol* 179(6): 2422-2426.

[50] Serkova N.J., Gamito E.J., Jones R.H., O'Donnell C., Brown J.L. et al. (2008). The metabolites citrate, myo-inositol, and spermine are potential age-independent markers of prostate cancer in human expressed prostatic secretions. *Prostate* 68(6): 620-628.

[51] Sreekumar A., Poisson L.M., Rajendiran T.M., Khan A.P., Cao Q. et al. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 457(7231): 910-914.

[52] Yan S.K., Wei B.J., Lin Z.Y, Yang Y., Zhou Z.T. et al. (2008). A metabonomic approach to the diagnosis of oral squamous cell carcinoma, oral lichen planus and oral leukoplakia. *Oral Oncol* 44(5): 477-483.

[53] Zhou J., Xu B., Huang J., Jia X., Xue J. et al. (2009). ^{1}H NMR-based metabonomic and pattern recognition analysis for detection of oral squamous cell carcinoma. *Clin Chim Acta* 401(1-2): 8-13.

[54] Tiziani S., Lopes V., Gunther U.L. (2009). Early stage diagnosis of oral cancer using ^{1}H NMR-Based metabolomics. *Neoplasia* 11(3): 269-276.

[55] Hirayama A., Kami K., Sugimoto M., Sugawara M., Toki N. et al. (2009). Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. *Cancer Res* 69(11): 4918-4925.

[56] Gao H.C., Dong B.J., Liu X., Xuan H.Q., Huang Y.R. et al. (2008). Metabonomic profiling of renal cell carcinoma: High-resolution proton nuclear magnetic resonance spectroscopy of human serum with multivariate data analysis. *Anal Chim Acta* 624(2): 269-277.

[57] Kim K., Aronov P., Zakharkin S.O., Anderson D., Perroud B. et al. (2009). Urine metabolomics analysis for kidney cancer detection and biomarker discovery. *Mol Cell Proteoml* 8(3): 558-570.

[58] Kind T., Tolstikov V., Fiehn O., Weiss R.H. (2007). A comprehensive urinary metabolomic approach for identifying kidney cancer. *Anal Biochem* 363(2): 185.

[59] Chen W.X., Lou H.Y., Zhang H.P., Nie X., Xiang Y. et al. (2008). WSSV infection activates STAT in shrimp. *Comp Dev Immunol* 32: 1142-1150.

[60] Monleon D., Morales J.M., Gonzalez-Darder J., Talamantes F., Cortes O. et al. (2008). Benign and Atypical Meningioma Metabolic Signatures by High-Resolution Magic-Angle Spinning Molecular Profiling. *J Proteome Res* 7(7): 2882-2888.

[61] Tan C., Chen H., Xia C. (2009). Early prediction of lung cancer based on the combination of trace element analysis in urine and an Adaboost algorithm. *J Pharm Biomed Anal* 49(3)**:** 746-752.

[62] Monleon D., Morales J.M., Barrasa A., Lopez J.A., Vazquez C. et al. (2009). Metabolite profiling of fecal water extracts from human colorectal cancer. *Nmr Biomed* 22(3)**:** 342-348.

[63] Mal M., Koh P.K., Cheah P.Y., Chan E.C.Y. (2009). Development and validation of a gas chromatography/mass spectrometry method for the metabolic profiling of human colon tissue. *Rapid Commun Mass Spectrom* 23(4)**:** 487-494.

[64] Denkert C., Budczies J., Weichert W., Wohlgemuth G., Scholz M. et al. (2008). Metabolite profiling of human colon carcinoma – deregulation of TCA cycle and amino acid turnover. *Mol Cancer* 7**:** 72-87.

[65] Chan E.C., Koh P.K., Mal M., Cheah P.Y., Eu K.W. et al. (2009). Metabolic Profiling of Human Colorectal Cancer Using High-Resolution Magic Angle Spinning Nuclear Magnetic Resonance (HR-MAS NMR) Spectroscopy and Gas Chromatography Mass Spectrometry (GC/MS). *J Proteome Res* 8(1): 352-361.

[66] Frickenschmidt A., Frohlich H., Bullinger D., Zell A., Laufer S. et al. (2008). Metabonomics in cancer diagnosis: mass spectrometry-based profiling of urinary nucleosides from breast cancer patients. *Biomarkers* 13(4): 435-449.

[67] Woo H.M, Kim K.M., Choi M.H., Jung B.H., Lee J. et al. (2009). Mass spectrometry based metabolomic approaches in urinary biomarker study of women's cancers. *Clin Chim Acta* 400(1-2): 63-69.

[68] Bullinger D., Frohlich H., Klaus F., Neubauer H., Frickenschmidt A. et al. (2008). Bioinformatical evaluation of modified nucleosides as biomedical markers in diagnosis of breast cancer. *Anal Chim Acta* 618(1)**:** 29-34.

[69] Odunsi K., Wollman R.M., Ambrosone C.B., Hutson A., McCann S.E. et al. (2005). Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *Int J Cancer* 113(5): 782-788.

[70] Denkert C., Budczies J, Kind T., Weichert W., Tablack P. at al. (2006). Mass spectrometry-based metabolomics profiling reveals different metabolite pattern in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Res* 66(22)**:** 10795-10804.

[71] Griffin J.L., Nicholls A.W. (2006). Metabolomics as a functional genomic tool for understanding lipid dysfunction in diabetes, obesity and related disorders. *Pharmacogenomics* 7(7)**:** 1095-1107.

[72] Sebedio J.L., Pujos-Guillot E., Ferrara M. (2009). Metabolomics in evaluation of glucose disorders. *Curr Opin Clin Nutr Metab Care* 12(4): 412-408.

[73] Rang H.P.D., Ritter M.M., Flower J.M., Rang R.J. (2007). Dale's Pharmacology, Elsevier, Philadelphia.

[74] Oresic M., Simell S., Sysi-Aho M., Nanto-Salonen K., Seppanen-Laakso T. et al. (2008).Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *J Exp Med* 205(13): 2975-2984.

[75] Mäkinen V.P., Forsblom C., Thorn L.M., Wadén J., Gordin D. et al. (2008). Metabolic phebotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes. *Diabetes* 57(9): 2480-2487.

[76] Makinen V.P., Soininen P., Forsblom C., Parkkonen M., Ingman P. at al. (2008). 1H NMR metabonomics approach to the disease continuum of diabetcs complications and premature death. *Mol Syst Biol* 4: 167.

[77] Mäkinen V.P., Soininen P., Forsblom C., Parkkonen M., Ingman P. at al. (2006). Diagnosing diabetic nephropathy by 1H NMR metabonomics of serum. *MAGMA* 19(6): 281-296.

[78] Yang J., Xu G., Hong Q., Liebich H.M., Lutz K. et al. (2004). Discrimination of Type 2 diabetic patients from healthy controls by using metabonomics method based on their serum fatty acid profiles. *J Chromatogr B* 813(1-2)**:** 53-58.

[79] Eriksson L.J., Kettaneh-Wold E., Wold N.S. (2001). Multi- and Megavariate Data Analysis—Principles and Applications, Umetrics Academy, Umeå.

[80] Minino A.M., Heron M.P., Murphy S.L., Kochanek K.D. (2007). Deaths: final data for 2004. *Natl Vital Stat Rep* 55(19): 1-119.

[81] http://www.statistics.gov.uk/pdfdir/hsq0506.pdf acessed on february 2 2009.

[82] Brindle J.T., Antti H, Holmes E., Tranter G., Nicholson J.K. at al. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* 8(12): 1439-1444.

[83] Kirschenlohr H.L. , Griffin G.L., Clarke S.C., Rhydwen R., Grace A.A. at al. (2006). Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nat Med* 12(6): 705-710.

[84] Vallejo M., Garcia A. , Tunon J., Garcia-Martinez D., Angulo S. at al. (2009). Plasma fingerprinting with GC-MS in acute coronary syndrome. *Anal Bioanal Chem* 394(6):1517-1524.

[85] Roussel R., Mentre F., Bouchemal N., Hadjadj S., Lievre M. at al. (2007). NMR-based prediction of cardiovascular risk in diabetes. *Nat Med* 13(4): 399-400.

[86] Sabatine M.S., Liu E., Morrow D.A., Heller E., McCarroll R. et al. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation* 112(25): 3868-3875.

[87] Barba I., de León G.D., Martín E., Cuevas A., Aguade S. et al. (2008). Nuclear magnetic resonance-based metabolomics predicts exercise-induced ischemia in patients with suspected coronary artery disease. *Magn Reson Med* 60(1): 27-32.

[88] DeMeyer T., Sinnaeve D., Van Gasse B., Tsiporkova T., Rietzschel E.R. et al. (2008). NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal Chem* 80(10): 3783-3790.

[89] Holmes E., Leng Loo R., Stamler J., Bictash M., Yap I.K.S. et al. (2008). Human metabolomic phenotype diversity and its association with diet and blood pressure. *Nature* 453: 396-400.

[90] Kenny C.L., Dunn W.B., Ellis D.I., Myers J., Baker P.N. et al. (2005). Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics* 1(3): 227-234.

[91] Turner E., Brewster J.A., Simpson N.A., Walker J.J., Fisher J. (2008). Aromatic amino acid biomarkers of preeclampsia—a nuclear magnetic resonance investigation. *Hypertens Pregnancy* 27(3): 225-235.

[92] Rang H.D.P., Ritter M.M., Flower J.M., Rang R.J. (2007). *Dale's Pharmacology*, Elsevier, Philadelphia.

[93] Rozen S., Cudkowicz M.E., Bogdanov M., Matson W.R., Kristal B.S. et al. (2005). Metabolomic analysis and signatures in motor neuron disease. *Metabolomics* 1(2): 101-108.

[94] Tukiainen T., Tynkkynen T., Mäkinen V.P.,.Jylänki P., Kangas A. et al. (2008). A multivariate analysis of serum by 1H NMR spectroscopy: early systemic signs of Alzheimer's disease. *Biochem Biophys Res Commun* 375(3): 356-361.

[95] Michell A.W., Mosedale D., Grainger D.J., Barker R.A. (2008). Metabolomic analysis of urine and serum in Parkinson's disease. *Metabolomics* 4(3): 191-201.

[96] Underwood B. R., Broadhurst D., Dunn W.B., Ellis D.I., Michell A.W. et al. (2006). Huntington disease patients and transgenic mice have similar pro-catabolic serum metabolite profiles. *Brain* 129(Pt 4): 877-886.

[97] Holmes E., Tsang T.M., Huang J.T.J, Leweke F.M., Koethe D. et al. (2006). Metabolomic profiling of CSF: evidence that early intervention may impact on disease progression and outcome in schizophrenia. *PLoS Med* 3(8): e327.

[98] Kaddurah-Daouk R., McEvoy J., Baillie R.A., Lee D., Yao J.K. et al. (2007). Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Mol Psychiatry* 12(10): 934-945.

[99] Watkins S.M., Reifsnyder P.R., Pan H.J., German J.B., Leiter E.H. (2002). Lipid metabolome-wide effects of the PPARgamma agonist rosiglitazone. *J Lipid Res* 43(11): 1809-1817.

[100] Greenberg N., Grassano A., Thambisetty M., Lovestone S., Legido-Quigley C. (2009). A proposed strategy for monitoring disease progression in Alzheimer's disease. *Electrophoresis* 30(7): 1235-1239.

[101] Coen M., O'Sullivan M., Bubb W.A., Kuchel P.W., Sorrell T. (2005). Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clin Infect Dis* 141(11): 1582-1590.

[102] Carraro S., Rezzi S., Reniero F., Heberger K., Giordano G. et al. (2007). Metabolomics applied to exhaled breath condensate in childhood asthma. *Am J Respir Crit CareMed* 175(10): 986-990.

[103] De Laurentiis G., Paris D., Melck D., Maniscalco M., Marsico S. et al. (2008). Metabolomic analysis of exhaled breath condensate in adults by nuclear magnetic resonance spectroscopy. *Eur Respir J* 32(5): 1175-1183.

[104] Bertini I., Calabro A., DeCarli V., Luchinat C., Nepi S. et al. (2009). The metabolomic signature of celiac disease. *J Proteome Res* 8(1): 170-177.

[105] Marchesi J.R., Holmes E., Khan E., Kochhar S., Scanlan P. at al. (2007). Rapid and noninvasive metabolomic characterization of inflammatory bowel disease. *J Proteome Res* 6(2): 546-551.

[106] Williams H.R.T., Cox I.J, Walker D.J., North B.V., Patel V.M. et al. (2009). Characterization of inflammatory bowel disease with urinary metabolomic profiling. *Am J Gastroenterol* 104(6): 1435-1444.

[107] Williams H.R.T., Cox I.J., Walker D.J., Cobbold J.F., Taylor-Robinson S.D. et al. (2010). Differences in gut microbial metabolism are responsible for reduced hippurate synthesis in Chron's disease. *BMC Gastroenterol* 10(1): 108.

[108] Johannesson N., Olsson L., Backstrom D., Wetterhall M., Danielsson R. et al. (2007). Screening for biomarkers in plasma from patients with gangrenous and phlegmonous appendicitis using CE and CEC in combination with MS. *Electrophoresis* 28(9): 1435-1443.

[109] Tao X.M., Liu Y.M., Wang Y.H., Qiu Y.P., Lin J.C. et al. (2008). GC-MS with ethyl chloroformate derivatization for comprehensive analysis of metabolites in serum and its application to human uremia. *Anal Bioanal Chem* 391(8): 2881-2889.

[110] Psihogios N.G., Kalaitzidis R.G, Dimou S., Seferiadis K.I., Siamopoulos K.C. et al. (2007). Evaluation of tubulointerstitial lesions'severity in patients with glomerulonephritides: an NMR-based metabolomic study. *J ProteomeRes* 6(9): 3760-3770.

[111] Yang J., Zhao X., Liu X., Wang C., Gao P. et al. (2006).. High performance liquid chromatography-mass spectrometry for metabonomics: potential biomarkers for acute deterioration of liver function in chronic hepatitis B. *J Proteome Res* 5(3): 554-561.

[112] Hewer R, Vorster J, Steffens F.E., Meyer D. (2006). Applying biofluid 1H NMR-based metabolomic techniques to distinguish between HIV-1 positive/AIDS patients on antiretroviral treatment and HIV-1 negative individuals. *J Pharm Biomed Anal* 41(4): 1442-1446.

[113] German J.B., Hammock B.D., Watkins S.M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* 1:3-9.

[114] Williams R.E., Lenz E.M., Evans J.A., Wilson I.D., Granger J.H. et al. (2005). A combined ^{1}H NMR and HPLC–MS-based metabonomic study of urine from obese (fa/fa) Zucker and normal Wistar-derived rats. *J Pharm Biomed Anal* 38: 465-471.

[115] Williams R.E., Major H., Lock E.A., Lenz E.M., Wilson I.D. et al. (2005). d-Serine-induced nephrotoxicity: a HPLC–TOF/MS-based metabonomics approach. *Toxicology* 207: 179-190.

[116] Clayton T.A., Lindon J.C., Cloarec O., Antti H., Charuel C. et al. (2006). Pharmaco-metabonomic phenotyping and personalised drug treatment. *Nature* 440: 1073-1077.

[117] Gavaghan C.L., Holmes E., Lenz E., Wilson I.D., Nicholson J.K. (2000). An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett* 484(3): 169-17.

[118] Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H. et al. (2008). Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U.S.A.* 105(5): 1420-1424.

[119] Griffin J. L., Walker L. A., Troke J., Osborn D., Shone R. F. et al. (2000). The initial pathogenesis of cadmium induced renal toxicity. *FEBS Lett* 478(1-2): 147-150.

[120] Griffin J. L., Walker L. A., Shore R. F., Nicholson J. K. (2001). High-resolution magic angle spinning 1H-NMR spectroscopy studies on the renal biochemistry in the bank vole (Clethrionomys glareolus) and the effects of arsenic (As^{3+}) toxicity. *Xenobiotica* 31(6): 377-385.

[121] Bino R. J., Hall R. D., Fiehn O., Kopka J., Saito K. et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* 9(9): 418-425.

[122] Fiehn O. (2001). Combining genomic, metabolome analysis, and biochemical modeling to understand metabolic networks. *Comp Funct Genom* 2(3): 155-168.

[123] Gieger C., Geistlinger L., Altmaier E., de Angelis M.H., Kronenberg F. et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *Plos Genetics* 4(11): e1000282.

[124] Help Matlab 7.9.0

[125] Wold S., Esbensen K., Geladi P. (1987). Principal component analysis. *Chemometr Intell Lab Syst.* 2(1-3):37-52.

[126] Rantalainen M., Bylesjo M., Cloarec O., Nicholson J.K., Holmes E. et al. (2007). Kernel-based orthogonal projections to latent structures (K-OPLS). *J Chemometr* 21(7-9): 376-385.

[127] Wold S., Eriksson L., Sjöström M. (1998). PLS in chemistry, encyclopedia of computational chemistry 2006: 16.

[128] Trygg J., Wold S. (2002). Orthogonal projections to latent structures (O-PLS). *J Chemometr* 16: 119-128.

[129] Fisher R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann Eugenetics* 7: 179-188.

[130] Larranaga P., Calvo B., Santana R., Bielza C., Galdiano J. et al.(2005). Machine learning in bioinformatics. *Brief Informatics* 7(1): 86-112.

[131] Howells S.L., Maxwell R.J., Peet A.C., Griffiths J.R. (1992). An investigation of tumor 1H nuclear magnetic resonance spectra by the application of chemometric techniques. *Magn Reson Med* 28: 214-236.

[132] Maxwell R.J., Martinez-Perez I., Cerdan S., Cabanas M.E., Arus C. et al. (1998). Pattern recognition analysis of 1H NMR spectra from perchloric acid extracts of human brain tumor biopsies. *Magn Reson Med* 39: 869-877.

[133] El-Deredy W., Ashmore S.M., Branston N.M., Darling J.L., Williams S.R. et al. (1997). Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks *Cancer Res* 57: 4196-4199.

[134] Lisboa P.J., Kirby S.P., Vellido A., Lee Y.Y., El-Deredy W. (1998). Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection. *NMR Biomed* 11: 225-234.

[135] Cortes C., Vapnik V. (1995). Support-Vector Networks. *Machine Learning* 20:273-297.

[136] Oust A., Moretro T., Kirschner C., Narvhus J. A., Kohler A. (2004). FT-IR spectroscopy for identification of closely related lactobacilli. *J Microbiol Methods* 59: 149-162.

[137] Beckonert O., Bollard E., Ebbels T.M. D., Keun H.C., Antti H. et al. (2003). NMR-based ,metabonomic toxicity classification: Hierarchical cluster analysis and k-nearest-neighbor approaches. *Anal Chim Acta* 490-493.

[138] Martens H., Naes T. (1989). Multivariate Calibration, *Wiley*.

[139] Golbraikh A., Tropsha A. (2002). Beware of q^2! *J Mol Graph Model* 20(4): 269-276.

2. Materials and methods

Sample preparation

Urine samples, once collected following the standard procedures for urine colture, were processed, aliquoted and frozen as soon as possible (usually within the 4 hours) to – 80°C in order to avoid metabolites degradation. The day of analysis samples were thawed at room temperature and shaken before use.

Blood samples, once collected in the specific vacutainers for plasma and serum collection, were differently treated for the isolation of each of these blood derivates: vacutainers for plasma collection containing different kind of anticoagulant compounds (citrate, ethylenediaminetetraacetic acid or EDTA, heparine, fluoride/oxalate) were centrifugated as soon as possible at a speed around 820 RCF for 10' at 4°C, while vacutainers for serum collection were centrifugated usually after an half hour at room temperature, as required for clotting, at a speed around 1500 RCF for 10'at 25°C. In both the cases surnatant was recovered and frozen as soon as possible (usually within the 4 hours) to – 80°C. The day of analysis samples were thawed at room temperature and shaken before use.

Blood serum is blood plasma without fibrinogen and the other clotting factors.[1]

Urine. According to a commonly used protocol[2,3], urine samples were shaken before use and 630 µL were centrifuged at 14000 RCF for 5 min. 540 µL of the supernatant were added to sixty microliters of potassium phosphate buffer (0.2 M Na_2HPO_4, 0,2 M NaH_2PO_4, 30mM sodium azide (NaN_3) and 10 mM sodium trimethylsilyl [2,2,3,3-2H4] propionate, TMSP, in 100% ^2H_2O, pH 7.0). 540 µL of the mixture were pipetted into 4.25 mm NMR tubes (Bruker BioSpin srl).

Blood-derivative NMR samples. Both serum and plasma, were prepared adding 300 µl of phosphate sodium buffer (70 mM Na_2HPO_4, 38 mM NaN_3 and 55 mM TMSP in 20% ^2H2O, pH 7.4) to 300 µl of plasma or serum. A total of 450 µl of this mixture was transferred into a 4.25 mm NMR tube (Bruker BioSpin srl).

NMR experiments and data analysis

The strategy analysis of the projects described in this thesis can be synthesized by four basic steps:

- Acquisition, processing (FT, phase correction, baseline correction, spectra alignment) of ^1H-NMR spectra;

- Bucketing;
- Data mining by the PCA and unsupervised clustering;
- Supervised clustering and modeling;
- Validation of the models;
- Identification of the biomarkers.

Acquisition and processing

All ^{1}H-NMR spectra were acquired using a Bruker 600 MHz spectrometer operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm CPTCI ^{1}H-^{13}C/^{31}P-^{2}H cryo-probe including a *z*-axis gradient coil, an automatic tuning-matching (ATM) unit and an automatic sample changer. A PT 100 thermocouple served for temperature stabilization at the level of approximately ±0.1 K at the sample. Before measurement, samples were kept for at least 3 minutes inside the NMR probe head, for temperature equilibration i.e., 300.0 K for urine and serum/plasma samples with the exception of plasma samples of the project AVIS that were analysed at a temperature of 310.0 K. The latter temperature were recently resolved to allow a better detection of lipidic profiles.

Urine For each urine sample, a one-dimensional (1D) NMR spectrum was acquired with water peak suppression using a standard pulse sequence (NOESYpresat; Bruker), using 64 scans, 64 k data points, a spectral width of 12,019 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s, and a mixing time of 100 ms.

Blood-derivates For each serum/plasma sample 1D NOESYpresat spectra were recorded; 1D spectra were acquired also with the Carr-Purcell-Meiboom-Gill (CPMG; Bruker) spin-echo sequence to suppress signals arising from high molecular weight molecules. In this case NOESYpresat spectra consisted of 64 scans, 98 k data points, a spectral width of 18,028 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s and a mixing time of 100 ms. CPMG spectra consisted of 64 scans, 74 k data points, a spectral width of 12,019 Hz, an acquisition time of 3.1 s, a relaxation delay of 4 s and a mixing time of 100 ms

For the project AVIS we acquire also also^{1}H-^{1}H J-resolved (J-res) spectra in order to get more information about signal multiplicity and coupling pattern and diffusion edited sequence using a diffusion time of 120 ms.

Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated (TMSP peak at 0.00 ppm for urine samples, alanine peak at 1.5 ppm for blood) using TopSpin 2.1 (Bruker Biospin srl).

Bucketing

Each 1D spectrum in the range between 0.02 and 10.00 ppm was segmented into 0.02-ppm chemical shift bins, and the corresponding spectral areas were integrated using AMIX software (Bruker BioSpin). Regions between 6.0 and 4.5 ppm containing residual water and urea signals were removed. To compensate for the high dynamic range in metabolite concentration, the normalized data are also scaled using the mean centering method so that changes in abundant metabolites do not dominate statistical models.

All resonances of interest were assigned on template one-dimensional NMR profiles by using matching routines of AMIX 3.8.4 (Bruker BioSpin) in combination with the BBIOREFCODE (Version 2-0-0; Bruker BioSpin) reference database and published literature when available.

Data mining by the PCA and unsupervised clustering

Multivariate statistical analysis is a tool to examine relationships among a great number of statistical variables at the same time (multivariate data). Generally multivariate data are represented by a matrix (the *X* matrix). Each row of this matrix corresponds to an object, while each column is a peculiar observable of the objects, for instance a bucket or a bin in metabonomics.

Matlab is a computational environment that provide a strong programming language for the analysis of data organized in matrix. R is an open source environment for data statistical analysis. All the statistical calculations described in this chapter were performed using scripts scripts written in R[4] and in MATLAB (The MathWorks, Natick, MA) language.

Many methods can be applied in order to perform a multivariate statistical analysis. Substantially the aim of these methods is to obtain a clusterization without giving data information to the systems (unsupervised methods). The principal and most used of these methods is PCA[5]. In practice, it searches for correlating variables that allow a division of the objects in two or more classes. This technique aims to extract more information as possible from a multivariate dataset and synthesizes it in few linear combinations of the same variables. From a geometric point of view, this is a rotation process of the original data carried out in a way that the first new axis (the first principal component) is oriented in the direction of maximum percentage of variability of the data, the second axis is perpendicular to the first and in the direction of the second maximum percentage of variability of the data and so on for all the new axis. The algorithm creates the new components in a way that the explained variance from the nth component is bigger than the variance explained from the

(nth+1) component. The number of these new axis (principal components or PCs) it is equal to the number of the original variables; the matrix given from these new linear combinations is called “scores matrix”, while the matrix of the coefficient of this linear combinations is the “loading matrix”. Usually only a small number of new components, sufficient to explain data aggregation, is selected. Moreover the component analysis allow really useful graphical representations for the starting exploration of the data.

In a following step, by the mean of the “Pattern Recognition” methods, the similarity between the samples could be evaluated and the features specific of the similarity classes could be extracted. Clustering techniques, as HCA, are employed in order to reach this goal in Chapter 3[6,7].

As in the previous step, also for this step the analysis is carried out without information on the data structure “external” to these methods. This kind of analysis methods are the so called “unsupervised” methods.

HCA groups data over a variety of scales by creating a cluster tree diagram or dendrogram, in which the relationships between samples are expressed. The algorithm starts analyzing the similarity or dissimilarity between every pair of objects in the data set and calculating the distance between all pairs of data points. Then, it proceeds to group the objects into a binary, hierarchical cluster tree. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters (initially each cluster consists of a single data point) until a hierarchical tree is formed. The described algorithm is the agglomerative algorithm, but there is also a divisive algorithm that operates by successively splitting the groups, beginning with a single group and continuing until there are N groups, each of a single individual: generally, divisive algorithm are computationally inefficient[8]. The final step of the algorithm is determine where to cut the hierarchical tree into clusters: these clusters can be created by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. The main problem in the application of these methods to metabolomic studies is that the reproducibility is not good and the inclusion of new data requires a complete re-computation of the dendrogram that can lead to a new structure not necessarily similar to that generated from the previous training set. Moreover, HCA does not generate diagnostic information about what features are responsible for the classification in sub-clusters.

Supervised clustering and modeling

After that the unsupervised analysis provided a beginning picture of the intrinsic structure of the data, the following step allows to develop models by supervised regression methods like PLS (Partial Least Squares) regression, OPLS (Orthogonal Projection to Latent Structures) method, K-OPLS (Kernel-based orthogonal projections to latent structures) method, PLS-DA (Partial Least Squares Discriminant Analysis) and supervised classification method like PCA-DA (PCA followed by Discriminant Analysis), SVM (Supported Vector Machine) and k-Nearest-Neighbour (k-NN)[9]. Supervised classification methods provide mathematical models able to predict the belonging class of a sample among that singled out on the basis of the data. In order to reach this goal functional relations between the variables and one or more Y functions with the information about data structure "external" to these methods are used.

The hard core of these methods is the use of multivariate analysis techniques and of classification algorithm for single out, separate and define the common features of the classes in which data are separated and that explain the similarities and differences between the samples: these are defined clustering techniques. Clustering is defined as the study of algorithms and methods that allow to group objects with common features.

PLS[10] is a technique used with data that contain correlated predictor variables regression such as information about the nature of the samples. This technique is between multiple linear regression, that allow to find a combination of the predictors that best fit a response, and PCA, that permit to single out combinations of the predictors with large variance without use response values. PLS finds combinations of the predictors that have a large covariance with the response values by combining information about the variances of both the predictors and the responses, also considering the correlations among them. One recent modification of the PLS method is constituted by the OPLS[11], that is used to remove irrelevant and confusing parameters. The basic idea of OPLS is to separate the systematic variation in data matrix in two parts, one linearly related to predictor variables matrix, and the other unrelated (orthogonal) to it. This partition facilitates model interpretation and model execution on a new set of samples. The K-OPLS is a further modification of the PLS method with the introduction of the *kernel trick.*

The analysis of variance (ANOVA) technique takes a set of grouped data and determine whether the mean of a variable differs significantly among groups. Often there are multiple response variables, and the interest is determine whether the entire set of means is different from one group to the next. There is a multivariate version of analysis of variance that can address the problem: MANOVA technique. MANOVA function in Matlab 7.9.0 allows also to

perform the Canonical Analysis (CA). PCA looks for the combination of the original variables that has the largest possible variation. MANOVA instead looks for the linear combination of the original variables that has the largest separation between groups. Having found that combination, the base idea is next look for the combination with the second highest separation, and so on[12]. PCA and PLS can be used also combined with CA/DA to establish the optimum of discrimination.

In the case of SVM, a data point is viewed as a *p*-dimensional vector (a list of *p* numbers) and we can separate such points with a (*p*-1)-dimensional hyperplane: this is called a linear classifier. There are many hyperplanes that might classify the data: the best hyperplane is the one that represents the largest separation, or margin, between two or more classes in a way that distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum margin classifier[13].

The kNN rule is the simplest of all supervised classification approaches. Every sample is classified to the class most frequently expressed among the *k* nearest neighbour. Therefore the *k* is the fundamental parameter to set. Small values of *k* can lead to the construction of a model subject to significant statistical fluctuations, while large values of *k* reduce statistical errors but can smooth out many details of the class distribution. As *k*-means, this method gives a classification of the sample without associated visualization and interpretation of data and, therefore, it is not applied only in metabolomics[14,15].

Validation of the models

For a proper cross validation, the total data should be divided into a training set and a test set. Using the training set a model is developed and optimized. The test set is only used to test the model performance. By repeating the procedure in a way that each sample appear once and only once in the test set, the prediction error is representative for new samples. Accuracy, specificity and sensitivity were estimated using standard definitions[16].

Identification of the biomarkers

Finally to asses which buckets are significantly different among various classes a one-way analysis of variance is used. Statistical significance of difference in average values calculated over the two groups is assessed using Kruskal-Wallis or Wilcoxon test by using the Bonferroni correction on a nominal value of 0.005.

Project	Dataoverview	Modelbuilding	Classification	Model validation	Identification of biomarkers
Metabolic phenotypes	PCA\HCA	PCA-CA	kNN	Cross-validation with Monte Carlo approach	Kruskal-Wallis test
Breast cancer and metabolomics	PCA	OPLS	SVM	Cross-validation with Monte Carlo approach	Wilcoxon test
Celiac disease and metabolomics	PCA	KOPLS	SVM	Cross-validation with Monte Carlo approach	Wilcoxon test
Cardiovascular risk and metabolomics	PCA	PLS-CA	SVM	Cross-validation with Monte Carlo approach	Wilcoxon test
Ozone therapy and metabolomics	\	\	\	\	ANOVA
SOPs definition by metabolomics	PCA	PLS-CA\PCA-CA	kNN	Cross-validation with Monte Carlo approach	Kruskal-Wallis test

Table 1 Statistical methods of the projects described in this thesis

Bibliography

[1] Anthea M., Hopkins J., McLaughlin C. V., Johnson S., Quon Warner M. et al. (1993). Human Biology and Health. Englewood Cliffs, New Jersey, USA.

[2] Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H. et al. (2008) Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U.S.A. 105: 1420-1424.

[3] Bernini P., Bertini I., Luchinat C., Nepi S., Saccenti E. et al. (2009) Individual human phenotypes in metabolic space and time. J Proteome Res 8: 4264-4271.

[4] Ihaka R., Gentleman R R. (1996). A Language for Data Analysis and Graphics. J Comput Stat Graph 5:299-314.

[5] Griffin J. (2003). Metabonomic: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterization of xenobiotic toxicity and disease diagnosis. Curr opin Chem Biol 7:648-654.

[6] el-Deredy W. (1997). Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review. NMR Biomed 10(3):99-124.

[7] Bernini P., Bertini I., Luchinat C., Nepi S., Saccenti E., Schäfer H., Schütz B., Spraul M., Tenori L. (2009). Individual human phenotypes in metabolic space and time. *J Proteome Res*, 8: 4264-4271.

[8] Beckonert O., Bollard E., Ebbels T.M. D., Keun H.C., Antti H. et al. (2003). NMR-based ,metabonomic toxicity classification: Hierarchical cluster analysis and k-nearest-neighbor approaches. *Anal Chim Acta* 490-493.

[9] Lindon J.C., Holmes E., Nicholson J.K. (2001). Pattern recognition methods and applications in biomedical magnetic resonance. Prog Nucl Magn Reson Spectrosc 39:1-40.

[10] Wold S., Eriksson L., Sjöström M. (1998). PLS in chemistry, encyclopedia of computational chemistry 2006: 16.

[11] Trygg J., Wold S. (2002). Orthogonal projections to latent structures (O-PLS). *J Chemometr* 16: 119-128.

[12] Fiehn O. (2001). Combining genomic, metabolome analysis, and biochemical modeling to understand metabolic networks. *Comp Funct Genom* 2(3): 155-168.

[13] Cortes C., Vapnik V. (1995). Support-Vector Networks. *Machine Learning* 20:273-297.

[14] Gieger C., Geistlinger L., Altmaier E., de Angelis M.H., Kronenberg F. et al. (2008). Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *Plos Genetics* 4(11): e1000282.

[15] Oust A., Moretro T., Kirschner C., Narvhus J. A., Kohler A. (2004). FT-IR spectroscopy for identification of closely related lactobacilli. *J Microbiol Methods* 59: 149-162.

[16] Westerhuis J.A., Hoefsloot H.C.J., Smit S., Vis D.J., Smilze A.K. et al. (2008). Assessment of PLS-DA cross validation. Metabolomics 4:81-89.

3. Metabolic phenotypes

Background

Since its birth, metabonomics shows a great potential in various research areas like prognosis and diagnosis assessment of diseases[1], toxicology[2], pharmacology[3] and nutraceutics[4,5]. In particular the relevance of metabonomics can be greatly enhanced by the determination of the existence of a metabolic phenotype, typical for each subject and stable over time. Differences in experimental metabolic profiles due to genetic strain differences in animal models have been observed, leading to the suggestion that each individual or group of individual may be characterised by a different metabotype, defined as "the multiparametric description of an organism in a given physiological state based on metabonomic data"[6]. The major problem is that the experimental metabolic profile is influenced not only by the genotype but also by age, lifestyle, environmental factors, nutritional status, assumption of drugs, and other metabolites from symbiotic organisms, as gut microflora[7,8,9]. Consequently, changes in the metabolic profile of biologically complex organisms (as humans) in response to pathological stimuli may be difficult to distinguish from normal physiological variations. Despite these factors the experimental evidence of the existence of a metabolic phenotype has been recently collected[10], assessing the influence of possible perturbing factors on the metabolic profiles and minimizing them in order to eliminate noise due to random daily variation. Indeed the approach used for the determination of the metabotypes counts on the NMR analysis (in a projet named MetRef1) of multiple urine samples (40 for each individuals) taken in quite consecutive days (about 2-3 months) from 22 healthy subjects. Each one of these subjects avoided alcohol and drug intake the day before the collection, and, moreover, filled a complete dietary sheet in order to better determine changes due to food behavior. In these way it has been obtained for the first time a natural, stable, and invariant metabolic profile that was typical of a given subject, even if not necessarily unique.

Aim of the work

The discovery of the existence of a stable individual human phenotypes in a short time (about 2-3 months) is clearly an important improvement for metabonomics. At the same time it was necessary to unravel the behavior of the metabotypes in a more extensive time period, in order to completely help researchers in all above mentioned fields and, especially, for medical application. To reach this goal, 11 subjects of the project called MetRef1[10] were recruited again after 2 years, entering in the MetRef2 project, and 4 of them were recruited once more

time 1 year later, entering in the MetRef3 project. Nine healthy individuals, non present in MetRef1, were recruited 2 years of MetRef1 after together with the others 11 subjects. In this way the project MetRef2 was constituted by 20 individuals (9 males, 11 females), in the age range 25-55, donating 40 urine samples (first in the morning, preprandial) collected over a period of about 3 months. While MetRef3 was constituted only by 40 samples donated by 4 subjects recruited for the third time (see **Figure 1**)

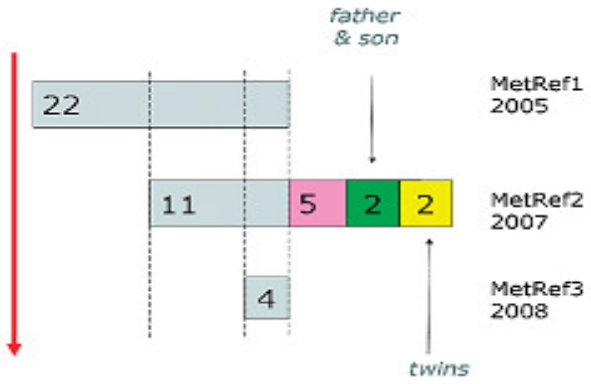


Figure 1 Collection scheme of the three MetRef projects

In the new collections some singularities were present. Even if not statistically relevant, they allowed to better understand environmental and genetic contributions to the definition of the individual metabolic phenotypes. Indeed, as it is noticeable in **Figure 1**, two new recruited subjects were homozygous twins (indicated with TA and TB codes), while other two subjects (BU and BV) were father and son. Moreover some subjects that were donating in more than one project quite drastically changed their lifestyle, as BC, that between the two collections moved from Italy (MetRef1) to Spain (MetRef2), varying the diet, and AR, that quitted to smoke.

Results and discussion

The data yet published about the first MetRef work[10] clearly showed that all subjects present a typical metabotype. This finding was confirmed through a predictive analysis that allowed to find out an average accuracy value of 99.7% for MetRef1 collection.

The same identical analysis (PCA-CA) was done on the 20 individuals of the MetRef2 projects giving a similar results with an average accuracy of about 99.6%. These results confirmed the robustness of the method used on the first work. Moreover analogous results were obtained pooling together all 31 individuals of the MetRef1 and MetRef2 collection

(clearly for the subjects that were participating to both collection the analyzed spectra were about 80 per each one). This result clearly highlight not only the existence of the individual metabolic phenotypes, but above all show their stability even a period of two years.

In order to definitively confirm the time stability of the metabotypes, spectra of the same individuals presents in different collection (11 for MetRef1 and MetRef2, 4 for MetRef3) were used in a single vote classification as both training and test set (see **Table 1**).

MetRef1 → MetRef2		MetRef1 → MetRef3		MetRef2 → MetRef3	
AI	99.971%	AR	99.522%	AR	100.000%
AO	100.000%	AS	94.439%	AS	99.915%
AR	97.242%	AU	99.944%	AU	97.377%
AS	98.754%	AW	98.746%	AW	92.699%
AU	100.000%	**MEAN =**	**98.163%**	**MEAN =**	**97.498%**
AW	99.712%				
BC	99.434%				
BF	97.132%				
BG	98.645%				
BH	100.000%				
BI	98.723%				
MEAN =	**99.056%**				

Table 2 Single vote classification. Different collections are used as training and test sets

As it is noticeable, the average values were high and substantially similar, even if, using MetRef3 as training set and MetRef2 as test set, the value was slightly lower. The high percentages reported in the previous table demonstrate that the metabotypes substantially remain stable in a time scale of about two-three years and this is a significant basis to the medical implication of the metabotypes. Indeed this justifies that drugs are differently metabolized by different subjects that may show different responses and/or different adverse effects on the basis of their own metabolism. Moreover these findings suggest the necessity to develop all medical therapies in a personal way according to the various existing metabotypes. Moreover the robustness of this “healthy” metabotype allows, together with the understanding of its peculiar features, to unravel whether modifications occur in the metabolomic profile have a physiologic or a pathologic origin.

Some considerations about the environmental and genetic contributions to metabotypes can be done by examining the obtained accuracy values for each subject and their respectives distances in the PCA-CA subspace. In particular TA and TB (homozygous twins) presented yet the shortest distance in the PCA-CA subspace, followed by BU and BV that are father and son, indicating the presence of a genetic component. This finding was also confirmed by the use of single vote classification; indeed the few misdiagnosed samples for TA and TB were

assigned to one another, non casually assigned to others subjects as it happened for all individuals.

Finally it is necessary to do a consideration on the variability of the spectra. Even if in this work it was clearly shown the time stability of the metabotypes, the detailed visual analysis of the spectra revealed some differences between various collections. These differences were classified as “spikes”, “waves” and “jumps” and they seemed principally due to diet effects. In particular, i) spikes are signals that appears in a single spectra having a markedly differences in intensity between preceding and following days, ii) waves are signals that have a variation in intensity more gradual but that persist for several days, iii) jumps are signals that markedly appear in a collection as spikes but they remain practically unchanged for all the collection and, sometimes, they are carried to next collection and are reversible. Spikes are principally due to particular food intake, for instance the peak of TMAO (TriMethylAmine-N-Oxide)[11] clearly appeared in spectra of urine collected the day after a fish diet intake, as well mannitol appeared after chewing gum intake. Nevertheless a particular diet is one of the causes of some jumps, in particular for individual AW it was observed a jump for xanthosine[12] due to an excessive meat consumption during a period of more days. Great part of jumps and waves are really due to the modification of gut microflora. Indeed it was noted that the peaks involved in waves and jumps were principally due to metabolites as hippurate, *meta*-hydroxyphenylpropionate, formate and phenylacetylglycine that were usually products of the metabolism of intestinal bacteria[13,14,15,16,17] and, therefore, linkable to activity variations of gut microflora. This last consideration lead us to consider the individual metabolic phenotypes as a metagenomic entity strongly affected by both host genotype, environmental factors and gut microbiome. This concept is even stronger under the light of the last findings that showed that humans have ten times more bacteria than human cells[18,19] that encode 100-fold more genes than human genome. Moreover the majority of microbes reside in the gut itself. It is showed their profound influence on human physiology and nutrition, and are crucial for human life[20,21]. Furthermore, the gut microbes contribute to energy harvest from food, and changes of gut microbiome may be associated with bowel diseases or obesity[22,23,24,25,26]. This concept it will be even more clear in Chapter 5 where the gut microbioma metabolism will be highlighted as one of the main component of the celiac fingerprint.

Bibliography

[1] Madsen R., Lundstedt T., Trygg J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Anal Chim Acta* 659(1-2): 23-33.

[2] Beger R.D., Sun J., Schnackenberg L.K. (2010). Metabolomics approaches for discovering biomarkers of drug-induced hepatotoxicity and nephrotoxicity. *Toxicol Appl Pharmacol* 243(2):154-166.

[3] Clayton T.A., Lindon J.C., Antti H., Charuel C., Hanton G. et al. (2006). Pharmaco-metabonomic phenotyping and personalised drug treatment. *Nature* 440: 1535-1542..

[4] Rezzi S., Ramadan Z., Fay L. B., Kochhar S. (2007). Nutritional metabonomics: applications and perspectives. *J Proteome Res* 6: 513-525.

[5] Kussmann M., Raymond F., Affolter M.(2006). OMICS-driven biomarker discovery in nutrition and health. *J Biotechnol* 124: 758-787.

[6] Gavaghan C.L., Holmes E., Lenz E., Wilson I.D., Nicholson J.K. (2000). An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57BL10J and Alpk:ApfCDmouse. *FEBS Lett* 484: 169-174.

[7] Nicholson J.K., Wilson J.D. (2003). Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat Rev Drug Discovery* 2: 668-676.

[8] Nicholson J.K., Holmes E., Lindon J.C., Wilson I.D. (2004). The challenges of modeling mammalian biocomplexity. *Nat Biotechnol* 22: 1268-1274.

[9] Tiret L. (2002). Gene environment interaction: a central concept in multifactorial disease. *Proc Nutr Soc* 61: 457-463.

[10] Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H., et al. (2008). Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci U.S.A.* 105: 1420-1424.

[11] Rabinowitz L., Berlin R., Yamauchi H. (1987). Plasma potassium and diurnal cyclic potassium excretion in the rat. *Am J Physiol* 253(22): F1178:F1181.

[12] Young D.S., Epley J.A., Goldman P. (1971). Influence of chemically defined diet on the composition of serum and urine. *Clin Chem* 17(8): 765-773.

[13] Phipps A.N., Stewart J., Wright B., Wilson I.D. (1998). 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* 28: 527-537.

[14] Williams R.E., Eyton-Jones H.W., Farnworth M.J., Gallagher R., Provan W.M. (2002). Effect of intestinal microflora on the urinary metabolic profile of rats: a (1)H-muclear magnetic resonance spectroscopy study. *Xenobiotica* 32: 783-794.

[15] Goodwin B.L., Ruthven C.R., Sandler M. (1994). Gut flora and the origin of some urinary aromatic phenolic compounds 47(12): 2294-2297.

[16] Nicholls A.W., Mortishire-Smith R.J., Nicholson J.K. (2003). NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem Res Toxicol* 16(11): 1395-1404.

[17] Samuel B.S., Gordon J.I. (2006). A humanized gnotobiotic mouse model of host-archeal bacterialmutualism. *Proc Natl Acad Sci U.S.A.* 103: 10011-10016.

[18] Sears C.L. (2005). A dynamic partnership: celebrating our gut flora. *Anaerobe* 11 (5): 247–251.

[19] Ley R.E., Peterson D.A., Gordon J.I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell* 124: 837–848.

[20] Backhed F., Ley R. E., Sonnenburg J.L., Peterson D. A., Gordon J.I.(2005). Host-bacterial mutualism in the human intestine. *Science* 307: 1915–1920.

[21] Hooper L.V., Midtvedt T., Gordon J.I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu Rev Nutr* 22: 283–307.

[22] Ley R.E., Turnbaugh P.J., Klein S., Gordon J.I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature* 444: 1022–1023.

[23] Turnbaugh P.J., Ley R.E., Mahowald M.A., Magrini V., Mardiset E.R. et al. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* 444: 1027–1031.

[24] Ley R.E., Bäckhed F., Turnbaugh P., Lozupone C.A., Knigh R.D. et al. (2005). Obesity alters gut microbial ecology. *Proc Natl Acad Sci USA* 102: 11070–11075

[25] Zhang H., DiBaise J.K., Zuccolo A., Kudrna D., Braidotti M. et al. (2009). Human gut microbiota in obesity and after gastric bypass. *Proc Natl Acad Sci USA* 106: 2365–2370.

[26] Turnbaugh P.J., Hamady M., Yatsunenko T., Cantarel B.L., Duncan A. et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457: 480–484

4. Breast cancer and metabolomics

Introduction and classification

In the worldwide breast cancer it is the second most common type of cancer after the lung cancer, with an incidence of 10.4 % (calculated on both sexes) and the fifth most common cause of cancer death (about 519.000 deaths in the 2004, equal to 7% of all cancer deaths)[1]. Breast cancer is about 100 times more frequent in women than in men, even if the rates of survival are practically equivalent[2]. Women are at higher risk of breast cancer because they have much more breast tissue than men do and because estrogens promote its development. Breast cancer is the most common malignances in women: 32% of all female cancers[3].

Each breast has about 20 sections called lobes, which have many smaller sections called lobules. The lobes and lobules are connected by ducts (**Figure1**). The most common type of breast cancer is ductal cancer. Cancers are classified as non-invasive (*in situ*) and invasive (infiltrating). The term *in situ* refers to cancer that has not spread past the area where it initially developed. Invasive breast cancer has a tendency to spread (metastasize) to other tissues of the breast and/or other regions of the body. A less common type of breast cancer is the inflammatory breast cancer, characterized by general inflammation of breast.[4]

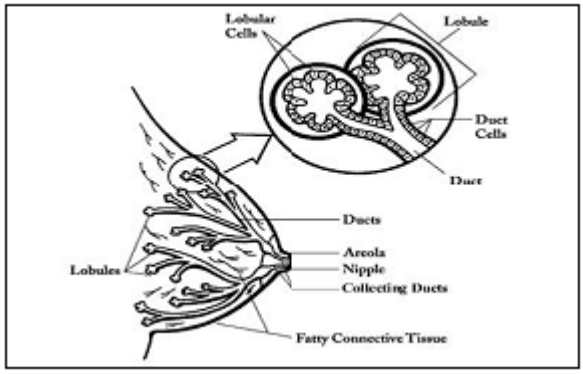


Figure 1 The structure of the female breast

Beyond of the type of pathology, other classifications exist for the breast cancer.

Grade of tumor: It is determined by the pathologist using a microscopy and the Bloom-Richardson-Elston staging system[5,6]. With the microscopy it is assigned a score ranging from 1 to 3 to the three followed features: i) percentage of tumor with normal ducts, ii) number of observable mitotic figures and iii) characteristic of cell nuclei. Therefore the final score will be included between 3 (well differentiated, best prognosis) and 9 (poorly

differentiated, worst prognosis). In details tumor with scores between 3 and 5 are grade 1, between 6 and 7 are grade 2 and between 8 and 9 are grade 3.

Genetic and Proteic Expression: These tests are usually done by immunohistochemistry on tissue from a biopsy. The breast cancer cells are tested for expression of some genes, as estrogen receptor (ER) and progesterone receptor (PR), and of some protein, as human epidermal grow factor (EGF) or receptor 2 (HER2).

In both pre- and postmenopausal women, levels of steroid receptors ER and PR can predict which women are likely to benefit from hormone treatment. Most breast cancers in post-menopausal women are ER-positive, meaning that they require estrogen to grow. These ER-positive breast cancers are less aggressive than ER-negative breast cancers, which are found generally in pre-menopausal women.

HER2 is overexpressed in breast cancers as a result of HER 2 proto-oncogene amplification. It is measured in the tissue from a biopsy either by immunological assays of the protein or PCR. The presence of HER-2 is generally associated with a more aggressive growth and poorer prognosis or breast and ovarian cancer[7]. It can also help to determine treatment options and for resistance to some conventional therapies.

Stage of tumor: TNM classification is based on the clinical size and extent of invasion of the primary tumor (T), the clinical absence or presence of palpable axillary lymph nodes and evidence of their local invasion (N), together with the clinical and imaging evidence of distant metastases (M). This is then translated into TNM classification which has been subdivided into:

- Stage 0, called carcinoma *in situ*.
- Stage I, early stage breast cancer where the tumor is less than 2 cm across and has not spread beyond the breast.
- Stage II – early stage breast cancer where the tumor is either less than 2 cm across and has spread to the lymph nodes under the arm; or the tumor is between 2 and 5 cm (with or without spread to the lymph nodes under the arm); or the tumor is greater than 5 cm and has not spread outside the breast.
- Stage III – locally advanced breast cancer where the tumor is greater than 5 cm across and has spread to the lymph nodes under the arm; or the cancer is extensive in the underarm lymph nodes; or the cancer has spread to lymph nodes near the breast bone or to other tissues near the breast.

- Stage IV – metastatic breast cancer where the cancer has spread outside the breast to other organs in the body.

Etiology

Even if the first work on the epidemiology and etiology of breast cancer was published on 1926 by Janet E. Lane-Claypon for British Ministry of Health, it is still not possible to definitely establish epidemiological risk factors and etiology for every type of breast cancer. Every woman is at risk for developing breast cancer. Several relatively strong risk factors for breast cancer that affect large portions of the general population have been identified. However, the majority of breast cancer cases occur in women who have no identifiable risk factors other than their gender[8].

About 5% of new breast cancers are attributable to hereditary factors and family history of breast cancer, while the etiology of the other 95% is unknown[9]. The risk of breast cancer is higher among women who have a close blood relative with the disease, especially if the relative developed breast cancer before the age of 50[10]. However, most women who get breast cancer (approximately 80%) have no such family history[11]. Recently, it has been shown that germline mutations in the BRCA1 and BRCA2 genes account for a large portion of cases of hereditary breast cancer[12]. In addition to mutations in the BRCA1 and BRCA2 genes, there are unidentified genetic defects that predispose to breast cancer development[13,14] and additional studies may help in identifying these genes in the future.

The other "established" risk factors for breast cancer beyond female gender are advanced age[2,15], previous breast cancer[16], benign breast disease[17], early age at menarche[17], late age at menopause[2], late age at first full-term pregnancy[17], postmenopausal obesity[18,19], low physical activity[20,21], race/ethnicity (non-Hispanic white, Hawaiian, and black women ethnicity) [22] and high-dose exposure to ionizing radiation early in life[23]. For almost all these factors estrogens play a pivotal role.

Clinical manifestations and signs of breast cancer

It is considered that, in about 80% of cases[1], the breast cancer is discovered by the own patients when they find a lump that feels different from the surrounding breast tissue. Sometimes the lumps are discovered through a mammography. Moreover the presence

of a lump in the armpits (lymph nodes) is also a possible symptom of breast cancer. Obviously many other signs may be included as changes in size or shape, skin dimpling, nipple inversion. Pain (called mastodynia) is another possible symptom but it is not characteristic. Indeed it is possible to have it in many other breast pathologies such as mastitis and fibroadenoma. When breast cancer cells invade the dermal lymphatics, that are small vessels in the skin of breast, its presentation can resemble skin inflammation and thus is known as inflammatory breast cancer as previously mentioned. Another reported symptom complex of breast cancer is Paget's disease that cause eczematoid lesions of the nipple skin. Approximately half of women diagnosed with Paget's also have a lump in the breast[24]. Furthermore, when breast cancer is manifest in the metastatic form, a great number of new usually "non-specific" symptoms appear and their features depend on the location of metastasis. Metastasis of breast cancer are commonly located in bone, liver, lung and brain[25].

Diagnosis and therapeutics

The most common used screening method for diagnosis of breast cancer is a combination of x-ray mammography and clinical breast exam. In particular regular mammography is recommended in several countries in women over a certain age as a screening tool. In women at higher than normal risk, such as those with a strong family history of cancer, additional tools may include also genetic tests or breast Magnetic Resonance Imaging.

A further support in the diagnosis and follow up of breast cancer could be given also by the cancer antigen (CA) dosage, that is usually elevated in breast cancer. However this test lacks of the required sensitivity and specificity for routine detection of breast cancer and does not discriminate patients with early carcinoma from those with benign breast disease.

The common diagnoses is elaborated after a "triple test" of clinical breast examination, lead by a trained specialist, mammography, and fine needle aspiration and cytology (FNAC). These three tools can be used to diagnose breast cancer with a good accuracy. Another useful option is biopsy, that consists in the removal of either part or entire lump. As other tumors, breast cancer is treated by surgery when it is localized and, following, an adjuvant hormonal therapy, chemotherapy and/or radiotherapy is chosen on the basis of the features of the tumor and of the patient. Indeed, depending on

clinical criteria (age, type of cancer, size, metastasis), patients are collocated into risk groups and different therapy strategies are consequently applied. Adjuvant!online is a free software that is very useful to predict the risk of relapses and mortality associated to a breast tumor in not-operated subjects, if they will be not treated. Moreover the software also values the reduction of the risk in case of specific therapies. Parameters used by software are related to form and type of tumor, as size, hormonal receptor (ER) status, lymph node involvement[26], but also to some patient characteristics, as age, diet, used drugs.

Aim of the work

An obvious aim of any research on breast or other types of cancer is the improvement of the knowledge of both pathologies and available clinical therapies. Metabonomics, as a science that provides a dynamic portrait of the metabolic status, can be very helpful in these directions. Indeed in Chapther 6, we have already shown the potentiality of metabolomics in define the presence of an individual fingerprinting stable over the time. On the other hand, in the study of breast cancer as well of the other examined pathologies of this thesis (celiac and cardiovascular disease risk), the goal is that of identify the pathological modifications that may occur in the individual profile distinguishing them from the physiologicals. Moreover one of the main goals within cancer metabolomics, but also within celiac disease, is to develop a standardazeble and non-invasively methods for diagnosing the disease[27].

Particularly an improved predictivity of clinical outcomes provides a better approach to the treatment of breast cancer, permits individualized therapy and reduces the side effects associated to it. For these purposes the attention was focused on the possibility to discriminate i) pre-operative not-metastatic breast cancer subjects ii) post-operative not-metastatic breast cancer subjects and iii) post-operative metastatic breast cancer subjects on the basis of the analysis metabonomic analysis of their serum blood samples. Moreover, another important goal was to check the capability of metabonomics to single out the presence of micrometastasis in the organism. Indeed, in the treatment of early breast cancer, a critical issue is the identification of which individuals can really benefit from adjuvant intervention. As mentioned above, the patients are divided in risk groups and their therapy is decided starting from this classification. However it is common to under- or over-estimate the risk and, therefore, use to apply therapies that

are not completely correct and appropriate. One, and probably the most, important cause of this mis-assignment is the existence of unpredicted micrometastasis. The detection of micrometastasis is not a trivial task. In this work the idea was to detect whether a metabonomic fingerprint of the micrometastasis exists and to evaluate how the "metabonomic risk" correlate with that derived from the use of Adjuvant!online[28] software, which predict the 10-years mortality.

Results and discussion

The study was based on the metabolomic analysis of serum. Collected serum samples were divided in three classes: i) 44 pre-operative early breast cancer ii) 98 post-operative early breast cancer (44 of which were the same that gave also the pre-operative samples, 45 new recruited) iii) 51 metastatic breast cancer. For the early breast cancer patients the mean time between pre-operative collection and surgery was 16 days (range 2:40), while for post-operative was 33 days (range 16-55).

A good clusterization between early pre-operative and metastatic subjects was obtained above all on NOESYpresat serum spectra using O-PLS (**Figure 1**): this clearly separation highlight the existence of a metabonomic fingerprint for the metastatic pathology and was confirmed also by the clusterization between post-operative and metastatic subjects

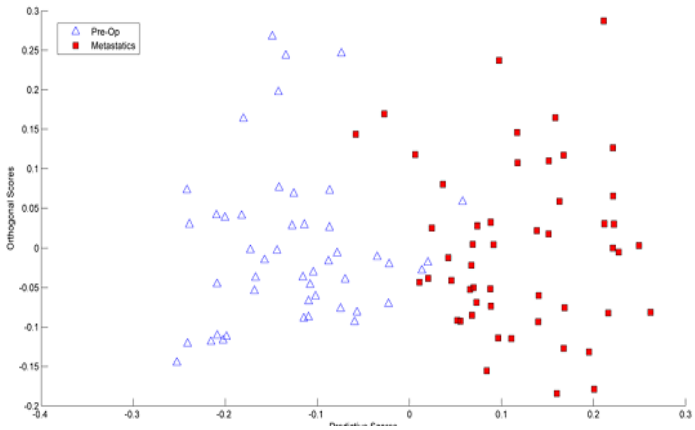


Figure 1 O-PLS of pre-operative (N=44) and metastatic (N=51) patients showing near complete separation of patient groups (NOESYpresat spectra).

Also the double cross validation scheme confirmed the goodness of this clusterization with an accuracy of 75% for CPMG and 77% for NOESYpresat. The clusterization shown in **Figure 1** suggests that some pre-operative samples present very similar

characteristic to metastatic samples. In order to better understand the origin of this misclassification, a risk factor, called "Metabolomic risk" was assigned to every not-metastatic sample. The value of each Metabolomic risk was assigned measuring the Euclidean distance of each dot from the centre of metastatic cluster. Obviously it was supposed that samples that have an higher distance from metastatic center had less metastatic characteristic and, therefore, they had a less value of risk. Practically the short is the distance, the higher is the risk, and vice versa. The comparison between Metabolomic risk and 10-year mortality by Adjuvant! (described above) showed a clear relationship between the two established risks with a percentage values of concordance of about 48%. However there was a significant discordance for three samples that were classified as having a very high metabolomic risk whilst they resulted to be not at risk for Adjuvant! To explain these data the best possible assumption is related to the presence of micrometastasis in some breast cancer hosts. Indeed, micrometastasis were not considered in calculation of risk by Adjuvant!. Therefore pre-operative patients that had a metabonomic fingerprint similar to metastatic subjects, and thus closer to the metastatic cluster, could have micrometastasis.

Perspectives

The only and better way to confirm our findings is to repeat the study in a class of subjects donating also follow up samples. In such way it may possible to understand if the discordance in the risk assessment between Adjuvant! and metabolomics is due to the presence of micrometastasis that can be the cause of relapse. In order to have this validation we are presently analyzing blood serum samples collected by the Memorial Sloan Kattering of New York within a follow-up study longer than 5 years. These samples belong to 95 metastatics subjects and to 109 post-operative patients (60 low risk for Adjuvant! and 49 high risk for Adjuvant!).

Another important approach can be the analysis of tissues of breast cancer. Analyses of tissues is used in metabonomics and led to the significant results of suggesting sarcosin[29] as a possible biomarker of prostate cancer progression. An identical approach could be used to better understand the metabolism of breast cancer and also to study in details both the role of phenylalanine in tumor progression and the importance of its pathways in the onset of breast cancer and, moreover, in its progression to metastasis.

Bibliography

[1] World Health Organization (web site www.who.int) (2006). Fact sheets N. 297, *Cancer.*

[2] Wu S.C., Hotes J., Fulton J.P., Chen V.W., Howe H.L. et al. (2002). Cancer in North America, 1995-1999. Volume III: NAACCR Combined Cancer Incidence Rates. Springfield, IL: North American Association of Central Cancer Registries.

[3] Ries L.A.G., Eisner M.P., Kosary C.L., Hankey B.F., Miller B.A. et al. (2002). SEER Cancer Statistics Review, 1973-1999, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1973_1999/, 2002.

[4] WHO. World Health Organization (1981). Histological Typing of Breast Tumors, 2nd edn. *Geneva*: WHO.

[5] Bloom H.J., Richardson W.W. (1957). Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br J Cancer* 11: 359-377.

[6] Genestie C., Zafrani B., Asselain B., Fourquet A., Rozan S. et al. (1998). Comparison of the prognostic value of Scarff-Bloom-Richardson and Nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems. *Anticancer Res* 18: 571-576.

[7] Tiwari R.K., Borgen P.I., Wong G.Y., Cordon-Cardo C., Osborne M.P. (1992). HER-2/neu amplification and overexpression in primary human breast cancer is associated with early metastasis. *Anticancer Res* 12 (2): 419-425.

[8] Kelsey J.L., Gammon M.D. (1990). Epidemiology of breast cancer. *Epidemiol Rev* 1990, 12: 228-240.

[9] Madigan M.P., Ziegler R.G., Benichou J., Byrne C., Hoover R.N. (1995). Proportion of breast cancer cases in the United States explained by well-established risk factors. *J Natl Cancer Inst* 87: 1681-1685.

[10] Claus EB, Stowe M, Carter D.(2003). Family history of breast and ovarian cancer and the risk of breast carcinoma in situ. *Breast Cancer Res Treat* 78 (1): 7-15.

[11] Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* 2001, 358(9291): 1389-1399.

[12] Ford D., Easton D.F. (1995). The genetics of breast and ovarian cancer. *Br J Cancer* 72 (4): 805-812.

[13] Jass J.R. (1997). Origins of … familial cancer: histopathological perspectives. *J Clin Pathol* 50 (11): 892-895

[14] Venkitaraman A.R. (2002). Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell* 108: 171-182.

[15] Edwards B.K., Howe H.L., Ries L.A., Thun M.J., Rosenberg H.M. et al. (2002). Annual report to the nation on the status of cancer, 1973-1999, featuring implications of age and aging on U.S. cancer burden. *Cancer* 94 (10): 2766-2792.

[16] Minami Y., Ohuchi N., Taeda Y., Takano A., Fukao A. et al. (1999). Risk of breast cancer in Japanese women with benign breast disease. *Jpn J Cancer Res* 90 (6): 600-606.

[17] Helmrich S.P., Shapiro S., Rosenberg L., Kaufman D.W., Slone D. et al. (1983). Risk factors for breast cancer. *Am J Epidemiol* 117 (1): 35-45.

[18] Brown P., Allen A.R. (2002). Obesity linked to some forms of cancer. *W V Med J* 98 (6): 271-272.

[19] Hirose K., Tajima K., Hamajima N., Takezaki T., Inoue M. et al. (2001). Association of family history and other risk factors with breast cancer risk among Japanese premenopausal and postmenopausal women. *Cancer Causes Control* 12 (4): 349-358.

[20] Bernstein L., Henderson B.E., Hanisch R., Sullivan-Halley J., Ross R.K.(1994). Physical exercise and reduced risk of breast cancer in young women. *J Nat Cancer Inst* 86:1403-1408.

[21] Friedenreich C., Bryant H.E., Courneya K.S. (2001). Case-control study of lifetime physical activity and breast cancer risk. *Am J Epidemiol* 154: 336-347.

[22] Racial/Ethnic Patterns of Cancer in the United States 1988-1992. National cancer Institute Surveillance, Epidemiology, and end results Program.

[23] Preston D.L., Mattsson A., Holmberg E., Shore R., Hildreth N.G. (2002). Boice JD Jr. Radiation effects on breast cancer risk: a pooled analysis of eight cohorts. *Radiat Res* 158 (2): 220-235.

[24] Marcus E. (2004). The management of Paget's disease of the breast. *Curr Treat Options Oncol* 5: 153-160.

[25] Lacroix M. (2006). Significance, detection and markers of disseminated breast cancer cells *Endocr Relat Cancer* 13: 1033-1067.

[26] Ravdin P.M., Siminoff L.A., Davis G.J., Mercer M.B., Hewlett J. et al. (2001). Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *J Clin Oncol* 19: 980-991.

[27] Madsen R., Lundstedt T., Trygg J. (2010). Chemometrics in metabolomics—A review in human disease diagnosis. *Anal Chim Acta* 659(1-2): 23-33.

[28] Goldstein L., Gray R., Badve S., Childs B.H., Yoshizawa C. et al. (2008). Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. *J Clin Oncol* 26: 4063-4071.

[29] Sreekumar A., Poisson L.M., Rajendiran T.M., Khan A.P., Cao Q. et al. (2009). Metabolomics profile delineate potential role for sarcosine in prostate cancer progression. *Nature* 2009 457: 910-914.

5. Celiac disease and metabolomics

Definition and etiology

Celiac disease (CD) is an autoimmune disorder of the small intestine caused by a lifelong intolerance to gluten that occurs in people of all ages. It is characterized by immune-mediated enteropathy causing maldigestion malabsorption[1]. Although no pharmaceutical treatment is actually available, a lifelong free-gluten diet allows to reverse the symptoms and permit to patients to spent a normal life[2].

The pathology is caused by an immune reaction to gliadin, a gluten protein normally present in wheat, barley and rye. Prolamines, gliadin and glutamine, are rich proline proteins. The high proline content is the cause of the resistance to complete proteolytic digestion of these proteins by the enzymes of the human intestine, that are physiological deficient in prolyl endopeptidase activity[3,4]. This can result in the accumulation of relatively large peptide fragments (as many as 50 amino acids in length) with a high proline and glutamine content in the small intestine[4,5]. Celiac subjects show an altered intestinal permeably that allows these peptides to pass through the gut mucosa[6]. After exposure to gliadin, the enzyme tissue transglutaminase (tTG) modifies these peptides, by the addition of crosslinkins between gluttammine and lysine residues, and the immune system cross-reacts with the small-bowel tissue, causing an inflammatory reaction[7] that leads to the atrophy of the villi of the small intestine. Finally this interferes with the absorption of nutrients, because the intestinal villi are responsible for absorption[8]. CD is a multifactorial disorder in which both genetic and environmental factors play a crucial role in pathogenesis[9]. Genetically it is associated with specific alleles: HLA-DQ2 and HLA-DQ8. HLA-DQ2 is expressed in more than 90% of people with CD[10,11] (**Figure 1**). However the expression of these two alleles is not sufficient to develop CD. Other studies highlight the role of the gap junctions between enterocytes and the possible role carried out by zonuline[12]. Moreover, results of studies in homozygotic twins suggest that genes are not the main cause of the disease development[13].

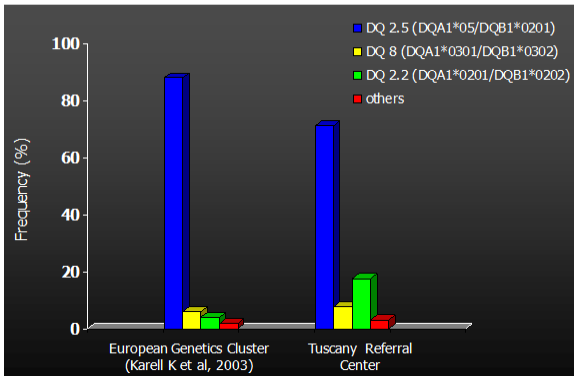


Figure 1 HLA Types in Coeliac Disease Patients

Clinical manifestation of celiac disease

Children between six and twenty-four months affected by CD tend to present bowel symptoms, abdominal distension, vomiting, growth problems, irritability and asthenia. The described symptoms are typical of classic form of the CD that starts after few months since the introduction of glutin in the diet. This form, in really rare cases, can also give the so called celiac crisis characterized by acute diarrhea, strong dehydration, hypotension and edema. Further the typical or classical form, CD give an atypical form in adults characterized by anemia caused by folic acid or iron deficiency, apathy, constipation, obesity, headache, anxiety, depression, cramps, paresthesia, but also chronic fatigue, muscular pain, athralgia, recurrent mouth ulcer and alteration of tooth enamel.

Moreover several other conditions are described associated to CD as various type of cancer (T-cell lymphoma[14,15,16], small bowel carcinoma[16]), endocrine diseases (type 1 diabetes[17], autoimmune thyroiditis[18] and Addison's disease[19]), dermatologic diseases (herpetiform dermatitis[20], psoriasi[21], chronic uticaria[22], alopecia[23]), neurologic disease (epilepsy[24], athaxia[24], drug resistant headache[24], multiple sclerosis[25], Corea di Huntington[24], peripheral neuropathy[24]), osteoarticular diseases (osteoporosis[26,27], rheumatoid arthritis[28], Sjögren syndrome[29]), psychiatric diseases (depression[30]), cardiovascular diseases and others diseases as Systemic Lupus Erythematosus (Lupus)[31], Down syndrome[32], Turner syndrome[33], Williams syndrome[34], autism[35], infertility[36], polyabortively[36].

Under the inclusive definition of celiac disease, or gluten-sensitive enteropathy, the concepts of silent, latent and potential CD have been introduced. While silent CD is marked by severe damage to the jejunal mucosa in the absence of clinical symptoms, both latent and potential CD are characterized by jejunal mucosa that would be reported as normal by most clinical pathologists in an individual on a gluten-containing diet. Latent subjects sometimes in their life have had a flat jejunal biopsy which recovered on a gluten-free diet. Latent celiac patients are often symptomatic; neither high titres of gliadin antibodies nor mucosal changes (including raised intraepithelial lymphocyte counts) are obligate features of latent CD, although the presence of elevated endomysial antibodies is probably the best predictor of progression towards villous atrophy[37].

The term potential CD has been proposed for those subjects who, as opposed to latent CD patients, do not have, and have never had, a jejunal biopsy consistent with clear CD, and yet have immunological abnormalities similar to those found in celiac patients. Good markers of potential CD include the presence of serum endomysial antibodies, a high count of

intraepithelial lymphocytes and subtle pathological alteration such as increased density of intraepithelial lymphocytes expressing gamma delta T cell receptor, signs of activated mucosal cell-mediated immunity, celiac-like intestinal antibody pattern, and positive rectal gluten challenge[37].

Prevalence and incidence

Celiac disease is one of the most common chronic disorders in Europe and western countries of European origin. Historically for celiac disease the prevalence was considered of about 0.02%[38], but the introduction of new diagnostic practices caused an increase in this value. Now the prevalence is considered to be between 0.54% and 0.94% in Italy[39,40], the 0.95% in USA[41], 0.60% in Argentina [42], 1,01% in Finland[43],1.30% in Sweden[44], the 0.40% in Australia[45] and even the 5.56%. in an Arabic population of Sahara[46]. Moreover, the prevalence is indicated to be between 0.33% and 1.06% in children (for Sahrawi people is 5.66%) and between 0.18% and 1.20% in adults[41,46]. It has been also reported a ratio male/female equal to 1/3[41].

Corrado G et al. reported in a study of 2001[47] that patients whit classical celiac symptoms show a double increased mortality during the first three years after diagnosis. The main causes of mortality are malignant diseases (as no-Hodgkin lymphoma), digestive diseases and respiratory diseases. The mortality data are confirmed also by Peters U et al. in 2003[48]: they find as main causes of death Hodking lymphoma, cancer of small intestine, rheumatoid arthritis, inflammatory bowel diseases and liver cirrhosis.

Diagnosis and therapeutics

Several tests could be used to diagnose CD, including serological and endoscopic tests. The biopsy by fiber-optic endoscopy is the test which carries the higher value of sensitivity (about 100%) and a lower frequencies of error, due to the possibility to have false positive (specificity is about 61%). The positivity of the biopsy is based on the identification of different degree of lesions (1-3C) on the basis of the Marsh Classification modified by Oberhuber[49] (**Figure 2**). However endoscopy represents a very invasive test, it is used to definitively confirm the presence of the CD in patients with positive serology and/or high-risk symptoms, as weight loss, anemia (more than 120 g/l in females and 130 g/l in males) and diarrhoea[50].

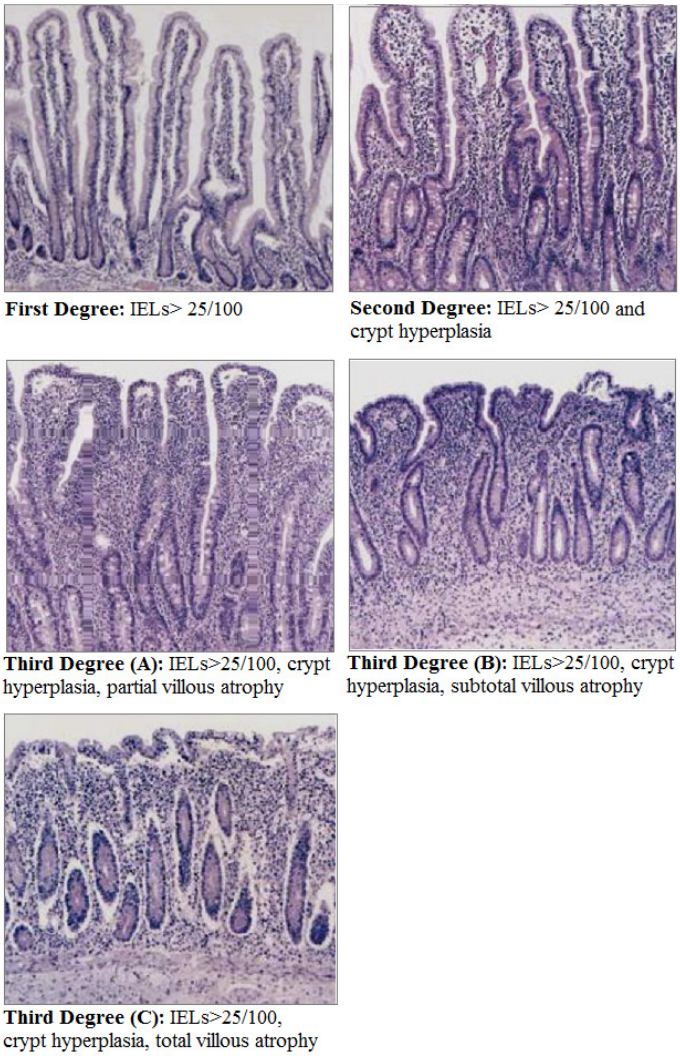


Figure 2 Marsh Classification modified by Oberhuber. IELs: intraepithelial linfocytes.

Serological tests are the first approach used to determine CD. Most used tests are the indirect immunofluorescence measures of four antibodies: IgA-anti-reticulin (ARA), IgA-anti-gliadin (AGA) and, above all, IgA-anti-endomysium (EMA) and IgA-anti-transglutaminase (tTG)[51]. In parallel total levels of all IgA are checked to avoid the possibility to have false negative results associated to celiac patients with IgA deficiency: in this case IgG antibodies could be useful to diagnose CD (25). These tests have very high value of sensitivity (98%) and specificity (about 99%)[52]. Similar value of specificity and sensitivity are associated with anti-tTG test, even if it is quicker and easier than anti-EMA test[53]. Although serology appears as the better first approach to diagnose CD it is interesting to note that some case of apparent seronegative CD occur, even if in presence of normal serum IgA[54].

It is also important remember as the described diagnostic methods have to be employed only on freed diet-subjects because the gluten-free diet cause a progressive normalization of both serologic and endoscopic tests.

Historically the only one therapeutical approach in CD is a complete e permanent gluten-free diet[2]. After a period of diet, from few weeks to some months, all the symptoms, included villous atrophy, are totally reversed except for the alteration of the tooth enamel. All other tried strategies, for example the use of bacterial prolyl endopeptidase as dietary supplement to degrade anti-gliadin peptides[55], have failed *in vivo*.

Background

A previous work[56], published in *Journal of Proteome Research*, fully shows the existence of a typical metabonomic signature for celiac disease. The work was based on ^{1}H NMR spectra of urine and serum samples of 34 CD patients and 34 healthy subjects. A CD patients' subgroup was also examined after a gluten-free diet. Using this metabolic fingerprint, it was possible to make predictions about the CD status with reasonably good accuracy (about 84%). Furthermore, Bertini et al. found that after a year on a GFD all but one of the CD patients classified with the healthy group, showing that the metabolic impact of the disease changes during treatment and can be followed using metabolomics. These results, although extremely encouraging, are not particularly helpful from a merely diagnostic point of view, since serology by blood test already achieves a sensitivity of about 98% and a specificity of over 95%[52]. However, the clear discrimination between patients and healthy subjects and patients and follow-up cases, achieved by metabolic fingerprinting, undoubtedly demonstrates the existence of a metabolic signature for CD both in serum and urine. Bertini et al. found that the metabolomic fingerprinting of CD is based on three components: i) malabsorption, ii) alteration of energy metabolism and iii) alteration of gut microflora. In this way it also suggests new features to explore and investigate some unsolved aspects of CD as chronic fatigue, that could be linked to the alteration of the energy metabolism that in CD patients seems to prefer other source of energy respect to glycolysis.

Aim of the work

Bertini et al. undoubtedly demonstrated the existence of a metabolic signature for CD both in serum and urine Nevertheless it is necessary to extend them to distinguish CD from other causes of malabsorption (small intestinal bacteria overgrowth, Crohn disease, short bowel syndrome and so on) and to well understand the capability of metabonomic to clearly highlights the presence of CD, even if for peculiar forms of the pathology, as silent or potential CD.

To solve these purposes of investigation 61 both urine and serum sample of different CD patients (11 males, 50 females, mean age 40.2 ± 13.9 yrs), 51 both urine and serum sample of different healthy subjects (19 males, 32 females, mean age 35.2 ± 14.3 yrs) and 29 both urine and serum sample of different potential celiac subjects (5 males, 24 females, mean age 37.1 ± 14.7 yrs) were collected and analyzed.

Our interest in potential CD subjects stems from the fact that they have all the typical immunoclinical features of clear CD, in the absence of any apparent intestinal damage. The aim of this work was investigate whether (and to what extent) potential celiac subjects share the same metabolic fingerprint of overt CD patients, thereby addressing an as yet open question: are the metabolic alterations associated to CD due to malabsorption (and therefore to intestinal damage) or are they independent of mucosal injury and therefore intrinsic to the metabolism of this pathology? In order to reach this goal, potential CD subjects could be an ideal subject of study because do not have villous atrophy. Moreover the comparison of the fingerprint of potential CD subjects with overt CD subjects and healthy controls can contribute to the open debate about whether or not it is necessary to put these patients on a gluten-free diet.

Moreover, in addition to metabolomic analysis, in order to better characterize the clinical phenotype of potential CD patients, citrulline levels of the three kind of study subjects were monitored and compared. Plasma citrulline level is an indirect measure of the enterocyte mass and it was proposed as a clinical biomarker for enterocyte dysfunctions in many diseases[57,58,59,60]. Although the idea of observing citrulline levels of celiac patients was not new[61], this was the first time that such analysis is done in potential celiac patients.

Results and discussion

The best clusterization of controls and untreated celiac subjects was obtained by means of k-OPLS on CPMG spectra (shown in **Figure 3**).

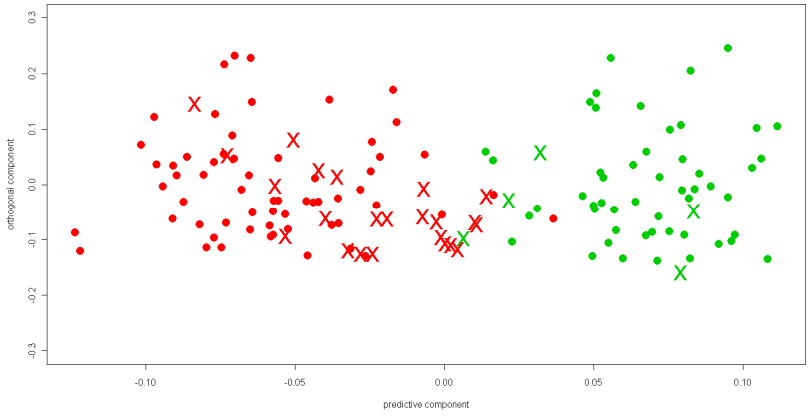


Figure 3 Clusterization of CD patients (red) and healthy controls (green) obtained with kopls for CPMG serum spectra. Potential celiac patients (cross) are projected on the kopls component, predicted by SVM and coloured accordingly to the results.

The statistical models applied proved to be effective in discriminating between patients and control subjects with values of accuracy in line with the previous work[56]: 81.2 vs 83.4 for CPMG spectra. The obtained results undoubtedly confirm the existence of the metabolic signature of CD, both in serum and in urine. Moreover if the potential celiac subjects were treated as blind samples and were subjected to classification by the previous statistical model, 24 out of 29 were classified as celiac using CPMG serum spectra(**Figure 3**). Those results strongly suggest that potential celiac subjects are more similar to overt celiac than to healthy controls. Moreover, when the statistical analysis was used to discriminate potential from overt celiac subjects, the accuracy was significantly lower (57.8%), confirming that potential CD subjects and CD patients share a similar metabotypes.

From the assessment of the significative buckets in the clusterization of spectra of serum samples and urine samples of untreated celiac patients and controls, again was confirmed that the metabolomic fingerprinting of CD is based on the three components: i) malabsorption, ii) alteration of energy metabolism and iii) alteration of gut microflora[56]. Otherwise the metabolic differences between potential CD and healthy controls in serum were quite similar to those between overt CD and controls at least for the metabolites related to energy metabolism. Especially, the key differences were in the metabolites originating from the gut microflora (m-HPPA, IS and PAG), confirming a relationship between overt CD, villous atrophy and bacterial consortia of the host. Apart from these differences, the metabolic patterns of overt CD and potential CD were largely the same, allowing us to hypothesize that CD exists as such before an evident intestinal damage occurs.

In line with previous findings, the comparison of citrulline levels clearly showed that citrulline levels in the CD patients were markedly lower than in the control group. But, more interestingly, also potential CD subjects, that from endoscopy did not show any manifest sign of villous atrophy, had lower levels of citrulline than controls. According to clinical experience, if a celiac subject has a positive antibody test but does not show macroscopic alterations of intestinal mucosa there can be only two possibilities 1) a (limited) intestinal damage is present but the endoscopy was not able to detect it (some errors in the sampling procedure are possible) or 2) the intestinal mucosa is really intact and the subject is a so-called “potential CD subject”. The interpretation of our data (both from metabolomics and from citrulline essay) suggests a third option: potential CD subjects are, indeed, non potential

at all. They already experience some subtle alterations of the enterocytes, at least at the microscopic functional level, but not at the macroscopic level. This may justify our results, showing that potential CD subjects are metabolically similar to CD, but they do not show endoscopic evidence of intestinal damage. Finally the core result from our investigation would be that metabolomics can detect CD also when its clinical manifestation is not fully evident. On this basis, putting potential CD subjects on a gluten fie diet can be extremely justified, because they are experiencing exactly the same pathological alterations as overt CD patients.

Conclusions

These results show that metabolomics can clearly identify a metabolomic fingerprinting of CD. Respect to the study on breast cancer (shown in Chapter 4) this results undoubtedly appear more robust. This finding is not at all completely unexpected because, respect to breast cancer, CD is a real metabolic pathology and, as shown by the variability of its clinical manifestation, strongly affects the functionality of the entire organism. This means that, although in Chapter 4 is shown that cancer is an interestingly subject of study for metabolomics, the field of the metabolic pathologies as CD or cardiovascular diseases (as we are going to show in Chapter 6) is that in which metabolomics seems to be the "natural" investigation tool.

Bibliography

[1] Holtmeier W., Caspary W. (2006). Celiac disease. *Orphanet J Rare Dis* 1: 1-3.

[2] Kupper C. (2005). Dietary guidelines and implementation for celiac disease. *Gastroenterology* 128: S121-S127.

[3] Hausch F., Shan L., Santiago N.A., Gray G.M., Khosla C. (2002). Intestinal digestive resistance of immunodominant gliadin peptides. *Am J Physiol Gastrointest Liver Physiol* 283:G996-G1003.

[4] Shan L., Molberg Ø., Parrot I., Hausch F., Filiz F. et al. (2002). Structural basis for gluten intolerance in celiac sprue. *Science.* 297: 2275-2279.

[5] Shan L., Qiao S.W., Arentz-Hansen H., Molberg Ø., Gray G.M. et al. (2005). Identification and analysis of multivalent proteolytically resistant peptides from gluten: implications for celiac sprue. *J Proteome Res* 4:1732-1741.

[6] Drago S., El Asmar R., Di Pierro M., Grazia Clemente M., Tripathi A. (2006). Gliadin, zonulin and gut permeability: Effects on celiac and non-celiac intestinal mucosa and intestinal cell lines. *Scand J Gastroenterol* 41(4): 408-419.

[7] Sollid L.M (2002). Coeliac disease: dissecting a complex inflammatory disorder. *Nat Rev Immunol* 2(9): 647-655.

[8] Green P.H., Jabri B. (2003). Coeliac disease. *Lancet* 362(9381) 383-391.

[9] Schuppan D. (2000). Current concepts of celiac disease pathogenesis. *Gastroenterology* 119: 234-242.

[10] van Heel D.A., Hunt K., Greco L., Wijmenga C. (2005). Genetics in coeliac disease. *Best Pract Res Clin Gastroenterol* 19: 323-339.

[11] Karell K., Louka A.S., Moodie S.J., Ascher H., Clot F. et al. (2003). Hla types in celiac disease patients not carrying the *DQA1*05-DQB1*02* (DQ2) heterodimer: results from the european genetics cluster on celiac disease. *Human Immunol* 64(4): 469-477.

[12] Clemente M.G., De Virgiliis S., Kang J.S., Macatagney R., Musu M.P. et al. (2003). Early effects of gliadin on enterocyte intracellular signalling involved in intestinal barrier function. *Gut* 52:218-223.

[13] Greco L., Romino R., Coto I., Di Cosmo N., Percopo S. et al. (2002). The first large population based twin study of coeliac disease. *Gut* 50: 624-628.

[14] Catassi C., Fabiani E., Corrao G., Barbato M., De Renzo A. et al. (2002). Risk of non-Hodgkin lymphoma in celiac disease. *J Am Med Assoc* 287: 1413.1419.

[15] Mearin M.L., Catassi C., Brousse N., Brand R., Collin P. et al. (2006). European multi-centre study on coeliac disease and non-Hodgkin lymphoma*. Eur J Gastroenterol Hepatol* 18: 187-194.

[16] Silano M., Volta U., Mecchia A., Dessì M, Di Benedetto R. et al. (2007). Delayed diagnosis of coeliac disease increases cancer risk. *BMC Gastroenterology* 7:8.

[17] Cronin C.C., Feighery A., Ferriss J.B., Liddy C., Shanahan F. et al.(1997). High prevalence of celiac disease among patients with insulin-dependent (type I) diabetes mellitus. *Am J Gastroenterol*: 92: 2210-2212.

[18] Ch'ng C.L., Jones M.K., Kingham J.G. (2007). Celiac disease and autoimmune thyroid disease. *Clin Med Res* 5: 184-192.

[19] Myhre A.G., Aarsetoy H., Undlien D.E., Hovdenak N., Aksnes L. et al. (2003). High frequency of coeliac disease among patients with autoimmune adrenocortical failure*. Scand J Gastroenterol* 38: 511-515.

[20] Reunala T. (1998). Dermatitis herpetiformis: Coeliac disease of the skin. *Ann Med* 5: 416-418.

[21] Abenavoli L., Leggio L., Gasbarrini G., Addolorato G. (2007). Celiac disease and skin: Psoriasis association. *Gastroenterology* 13(14): 2138-2139.

[22] Gabrielli M., Candelli M., Cremonini F., Ojetti V., Santarelli L. et al. (2005). Idiopathic chronic urticaria and celiac disease. *Digest Dis Sci* 50(9): 1702-1704.

[23] Corazza G.R., Andreani M.L., Venturo N., Bernardi M., Tosti A. et al. (1995). Celiac disease and alopecia areata: report of a new association. *Gastroenterology* 109(4): 1333-1337.

[24] Bushara K.O. (2005). Neurologic presentation of celiac disease. *Gastroenterology* 128(4): S92-S97.

[25] Ghezzi A., Zaffaroni M. (2001). Neurological manifestations of gastrointestinal disorders, with particular reference to the differential diagnosis of multiple sclerosis. Neurol Sci 22:S117-S122.

[26] Vasquez H., Mazure R., Gonzalez D., Flores D., Pedreira S. et al. (2000). Risk of fractures in celiac disease patients: a cross-sectional, case-control study. *Am J Gastroenterol* 95: 183-189.

[27] Riches P.L., McRorie E, Fraser W.D., Determann C., van't Hof R. et al. (2009). Osteoporosis Associated with Neutralizing Autoantibodies against Osteoprotegerin. New Engl J Med 361:1459-1465.

[28] Lubrano E, Ciacci C., Ames P.R.J., Mazzacca G., Ordente P. et al. (1990). The arthritis of coeliac disease prevalence and pattern in 200 adult patients. *Rheumatology* 35(12): 1314-1318.

[29] Collin P., Reunala T., Pukkala E., Laippala P., Keyriläinen O. et al. (1994). Celiac disease - associated disorders and survival. Gut 35 (9):1215-1218.

[30] Pynnönen P.A., Isometsä E.T., Aronen E.T., Verkasalo M.A., Savilahti E. et al. (2004). Mental disorders in adolescents with celiac disease. Psychosomatics 45:325-335.

[31] Rensch M.J., Szyjkowski R., Shaffer R.T., Fink S., Kopecky C. et al. (2001). The prevalence of celiac disease autoantibodies in patients with systemic lupus erythematosus . *Am J Gastroenterol* 96(4):1113-1115.

[32] Bonamico M., Mariani P., Danesi H., Crisogianni M., Pinella F. et al. (2001). Prevalence and clinical picture of celiac disease in Italian Down syndrome patients: a multicenter study. *J Pediatr Gastr Nutr* 33(2): 139-143.

[33] Bonamico M., Bottaro G., Pasquino A.M., Caruso-Nicoletti M., Mariani P. et al. (1998). Celiac disease and Turner syndrome. *J Pediatr Gastr Nutr* 26(5): 496-499.

34 Giannotti A., Tiberio G., Castro M., Virgilii F., Colistro F. et al. (2001). Coeliac disease in Williams syndrome. *J Med Genet* 38:767-768.

[35] Barcia G., Posar A., Santucci M., Parmeggiani A. (2008). Autism and celiac disease. *J Autism Dev Disord.* 38(2): 407-408-

[36] Ozgör B., Selimoğlu M.A. (2010). Celiac diseases and reproductive disorders. *Scand J Gastroenterol* 45(4): 395-402.

[37] Troncone R, Greco L, Mayer M, Paparo M, Caputo N. et al. (1996). Latent and potential coeliac disease. *Acta Paediatr* 412:10-14.

[38] Barker J.M., Liu E. (2008). Celiac disease: pathophysiology, clinical manifestations, and associated autoimmune conditions. *Adv Pediatr* 55 349-365.

[39] Catassi C., Fabiani E., Rätsch I.M., Coppa G.V., Giorgi P.L. et al. (1996). The coeliac iceberg in Italy. A multicentre antigliadin antibodies screening for coeliac disease in school-age subjects. *Acta Pediatr* 85:29-35-

40 Tommasinini A., Not T., Kiren V., Baldas V., Santon D. et al. (2004). Mass screening for coeliac disease using antihuman transglutaminase antibody assay. *Arch Dis Child* 89:512-515.

[41] Fasano A., Berti I., Gerarduzzi T.,Not T., Colletti R.B. et al. (2003). Prevalence of celiac disease in at-risk and not-at-risk groups in the United States. *Arch Intern Med* 163:286-292.

[42] Gomez J.C., Selvaggio G.S., Viola M., Pizarro B., la Motta G. et al. (2001). Prevalence of celiac disease in Argentina: screening of an adult population in the La Plata area. *Am J Gastroenterol* 96(9): 2700-2704.

[43] Mäki M., Mustalahti K., Kokkonen J., Kulmala P., Haapalahti M. et al. (2003). Prevalence of Celiac disease among children in Finland. *N Engl J Med* 348(25):2517-2524.

[44] Carlsson A.K., Axelsson I.E., Borulf S.K., Bredberg A.C., Ivarsson S.A. (2001). Serological screening for celiac disease in healthy 2.5-year-old children in Sweden. *Pediatrics* 107(1):42-45.

[45] Hovell C.J., Collett J.A., Vautier G., Cheng A.J., Sutanto E. et al. (2001). High prevalence of coeliac disease in a population-based study from Western Australia: a case for screening? *Med J Aust* 175(5):247-250.

[46] Catassi C., Ratsch I., Gandolfi L., Pratesi R., Fabiani E. et al. (1999). Why is coeliac disease endemic in the people of the Sahara? *The Lancet* 354(9179): 647-648.

[47] Corrado G., Corazza G., Bagnardi V., Brusco G., Ciacci C. et al. (2001). Mortality in patients with coeliac disease and their relatives: a cohort study. *The Lancet* 358(9279): 356-361.

[48] Peters U., Askling J., Gridley G., Ekbom A., Linet M. (2003). Causes of Death in Patients With Celiac Disease in a Population-Based Swedish Cohort. *Arch Intern Med* 163:1566-1572.

[49] Oberhuber G., Granditsch G., Vogelsang H. (1999). The histopathology of coeliac disease: time for a standardized report scheme for pathologists. *Eur J Gastroenterol Hepatol* 11:1185–1194.

[50] Hopper A., Cross S., Hurlstone D., McAlindon M., Lobo A. et al. (2007). Pre-endoscopy serological testing for coeliac disease: evaluation of a clinical decision tool. *BMJ* 334(7596): 729.

[51] Chorzelski T.P., Beutner E.H., Sulej J., Tchorzewska H., Jablonska S. et al. (1984). IgA anti-endomysium antibody. A new immunological marker of dermatitis herpetiformis and coeliac disease. *Br J Dermatol* 111: 395:402.

[52] James M.W., Scott B.B. (2000). Endomysial antibody in the diagnosis and management of coeliac disease. *Postgrad Med J* 76: 466-468.

[53] Feighery C. (1999). Fornightly review: coeliac disease. *BMJ* 319: 236-239.

[54] van Heel D.A., West J. (2006). Recent advance in coeliac disease. *Gut* 55 1037-1046.

[55] Matysiak-Budnik T., Candalh C., Cellier C., Dugave C., Namane A et al. (2005). Limited efficiency of prolyl endpeptidases in the detoxification of gliadin petides in celiac disease. Gut 129: 786-796.

[56] Bertini I., Calabro A., DeCarli V., Luchinat C., Nepi S. et al. (2009). The metabolomic signature of celiac disease. *J Proteome Res* 8(1): 170-177.

[57] Crenn P., Coudray-Lucas C., Thuillier F., Cynober L., Messing B. (2000). Postabsorptive plasma citrulline concentration is a marker of absorptive enterocyte mass and intestinal failure in humans. *Gastroenterology* 119:1496-505.

[58] Crenn P., Vahedi K., Lavergne-Slove A., Cynober L., Matuchansky C. et al. (2003). Plasma citrulline: a marker of enterocyte mass in villous atrophy-associated small bowel disease. *Gastroenterology* 124:1210-1219.

[59] Blijlevens N.M., Lutgens L.C., Schattenberg A.V., Donnelly J.P. (2004). Citrulline: a potentially simple quantitative marker of intestinal epithelial damage following myeloablative therapy. *Bone Marrow Transplant* 34:193–196.

[60] Pappas P.A., Tzakis A.G., Saudubray J.M., Gaynor J.J., Carreno M.R., et al. (2004). Trends in serum citrulline and acute rejection among recipients of small bowel transplants. *Transplant Proc* 36: 345–347.

[61] Miceli E., Poggi N., Missanelli A., Bianchi P., Moratti R. et al. (2008). Is serum citrulline measurement clinically useful in coeliac disease? 3:233-236.

6. Cardiovascular risk and metabolomics

Definition

Cardiovascular diseases (CVDs) comprise a group of disease states including a large number of more specific conditions, such as heart failure, coronary heart disease, atheroschlerosis and hypertension[1]. Collectively, diseases of the cardiovascular system are the leading causes of death in many industrialized countries[2,3]. "Each year, heart disease kills more Americans than cancer"[4]. These classes of diseases involve the heart or blood vessels (arteries and veins)[5]. While the term technically refers to any disease that affects the cardiovascular system, it is usually used to refer to those related to atherosclerosis.The morbidity, mortality, and socioeconomic importance of CVDs make timely accurate diagnosis and cost-effective management of the utmost importance. In most people, CVD stems from the combined effect of a number of causal risk factors. For this reason the scientific community coined the term "metabolomic syndrome" (for its definition see next section) to asses "a constellation of lipid-related and unrelated cardiovascular risk factors of metabolic origin"[6], tightly linked to a general disorder known as insulin resistance.

A debate is open on the utility of the metabolic syndrome diagnosis for clinicians in planning management strategy. However, several definitions based on different clinical parameters has been proposed. During the years it has been assessed the centrality of insulin resistance, that in the long time give type 2 diabetes, as amplifier factor of cardiovascular risk when the metabolic syndrome condition is present[7].

Since it was first proposed in New Zealand in 1993, estimating cardiovascular risk has become an integral part of determining the eligibility for treatment with aspirin, antihypertensives or statins[8]. To date, the Framingham cardiovascular equation remains the most widely used method of assessing cardiovascular risk. This equation was derived from a large cohort study in the United States. It determines cumulative incidence (risk) of any vascular event in individuals free from cardiovascular disease, using a combination of predictors such as age, sex, systolic blood pressure, smoking status, total cholesterol level, high-density lipoprotein (HDL) cholesterol level and diabetic status[9]. Cardiovascular risk estimation is therefore a diagnostic test that has become a part of routine clinical practice.

Risk factors

The major and independent risk factors for CVD are cigarette smoking, elevated blood pressure, elevated serum total cholesterol and low-density lipoprotein cholesterol (LDL), low serum HDL cholesterol, diabetes mellitus, and advancing age. The major risk factors are

additive in predictive power. Accordingly, the total risk of a person can be estimated by a weighted sum of the risk imparted by each of the major risk factors. Other factors are associated with increased risk for CVD. These are of two types: conditional risk factors and predisposing risk factors. The conditional risk factors (elevated serum triglycerides, elevated serum homocysteine, elevated serum lipoprotein, prothrombotic factors (as fibrinogen), inflammatory markers (as C-reactive protein)) are associated with increased risk for CVD although their causative, independent, and quantitative contributions to CVD have not been well documented. The predisposing risk factors (obesity, abdominal obesity, physical inactivity, family history of premature heart disease, ethnic characteristics and psychosocial factors) are those that aggravate the independent risk factors. The adverse effects of obesity are worsened when it is expressed as abdominal obesity, an indicator of insulin resistance[10]. Insulin resistance is a condition in which normal insulin concentrations fail to achieve a normal metabolic response: this means that higher than normal insulin concentrations are needed to achieve a normal metabolic response[11]. Abdominal obesity and insulin resistance (two significant predisposing factors) are a manifestation of the "metabolic syndrome" which in turn is a key in the evolution of cardiovascular disease. There is debate regarding whether obesity or insulin resistance is the cause of the metabolic syndrome or if they are consequences of a more far-reaching metabolic derangement.

A number of markers of systemic inflammation, including C-reactive protein, are often increased: they are fibrinogen, interleukin 6 (IL–6), tumor necrosis factor-alpha (TNFα), and others. Some have pointed to a variety of causes including increased uric acid levels caused by dietary fructose[10].

Aim of the work

A number of recent reviews have assessed results and future possibilities for metabolomics research in the field of CVDs[12,13,1415,16] However until now, no exhaustive metabolomic investigation of cardiovascular risk factors or, somewhat ironically, of the metabolic syndrome as such, has been reported, despite the major risk factors are essentially altered blood metabolites. The only exception is given by blood pressure that has been already investigated by Nicholson at al.[17]. However, because all of these alterations are of metabolic nature, we hypothesized that metabolomics would be an effective tool for a deeper investigation of the molecular mechanism and biochemical pathways involved in the composition of cardiovascular risk.

In this study, we examined the metabolic profile of 864 healthy volunteers (678 males, 186 females, mean age 40.87 ± 11.0 years) with the aim of investigating the metabolic alterations characteristic of an increased cardiovascular risk.

Results and discussion

We performed several statistical analyses in order to understand whether metabolic fingerprint of the plasma samples contains predictive information about the global status of the individuals. At a first inspection we did a classification analysis by the mean of PLS-CA, followed by SVM on the CA scores., using as predictor the full NMR profile, with respect to the total cholesterol (TC), LDL cholesterol, HDL cholesterol, triglycerides, and glycaemia. For classification purposes the groups were created by dividing samples in two classes represented by the highest and the lowest quintiles.

Values of sensitivity, specificity and accuracy, obtained from the double cross-validation procedure[18], were extremely encouraging: all the accuracy values were always higher than 95% for all classifications and types of 1-D spectra, except for glycaemia, that showed an accuracy around 85%. Interestingly this result could be also due to the fact the glucose is one of the metabolites that we found to be seriously affected by the standard operating procedures for the collection of serum/plasma (see Chapter 8). Therefore these results suggest that the whole plasma spectrum profile contains meaningful information about the metabolic status of individuals.

With these encouraging results, a more sophisticated regression analysis was also performed, with the aim of establishing a quantitative relationship between clinical data and NMR spectra. The average relative error is less than 10% and the R^2 higher the 0.80 for LDL and HDL cholesterol. These striking results were not completely unexpected, because the NMR spectra contain several strong peaks directly related to the lipidic components of the plasma.

To confirm that metabolomics is able to extract also more subtle metabolic information, we tested whether we could classify individuals with high and low values of TC or glycaemia even after removing spectral regions with peaks associated with lipid fractions (for TC analysis) and with glucose (for glycaemia analysis). Although the accuracy decreased to around 75% (both for TC and glycaemia), it was still highly significant, especially considering that we removed more than 100 spectral buckets (25% of the total) in order to be on the safe side: this result is a convincing indicator that changes in the blood composition are a global phenomena that involve several metabolic alterations and not only obvious patterns of lipids and lipoproteins peaks.

As it is well known that low levels of HDL cholesterol and high levels of TC correlate to high risk of developing CVDs, we performed the same analyses also for the TC/HDL ratio. Again the results, both for classification and regression, were extremely good, with an accuracy for classification even higher than for the two risk factors taken separately and an R^2 of 0.88 for regression.

The Framingham score for each individual was also calculated, using the equation of Wilson et al.[9] The classification gave high accuracy values also in this case (over 89%), but lowest than obtained in the previous analysis. This may due to the fact that the Framingham risk score considers also non-analytical variables, but it can derive also from the fact that metabolomics is able to extract a far richer and more complex metabolic pattern.

Finally, in order to identify the metabolites that were most characteristic of the metabolic fingerprint of cardiovascular risk, the univariate Wilcoxon test was applied on the different groups, using a significance level of $P<.001$. These results allow us to explore the metabolite composition of plasma profiles, in order to characterize what is typical of an undoubtedly a "low risk" status (high HDL levels, low LDL levels, low triglycerides and low glycaemia) and what is typical of a "high risk" status (the reverse). From the point of view of this comprehensive metabolic analysis it seems that the "low risk" pattern is mostly constituted by high levels of 3-HB and low levels of threonine and creatinine. On the other hand the "high risk" pattern is more complex, and is characterized by low levels of several metabolites, in particular α-ketoglutarate, dimethylglycine and serine.

Having in hand the results of metabolite profiling one might already attempt to elaborate new hypothesis about the origin of the metabolic syndrome and its connection with cardiovascular risk. In particular our findings seem to highlight the role of 3-HB, α-ketoglutarate, threonine and dimethylglycine as the key to define the fingerprint associated to high or low cardiovascular risk. Although a full clarification of a so complex machinery is far away, all our findings point to an imbalance in the mitochondrial redox potential (caused by oxidative stress), as the starting point that can lead to an alteration of the fatty acids beta-oxidation pathway with an increased secretion of LDL lipoproteins in the blood. (**Figure 1**). The involvement of mitochondrial dysfunction in the genesis of insulin resistance in adolescents with familiar history of type-2 diabetes has been reported also by Befroy[19], that observe a decreased rates of mitochondrial substrate oxidation of the 30% in lean, insulin-resistant offspring compared with insulin-sensitive control subjects. This observation can be easily linked to our findings, moving the starting point of obesity, insulin resistance, and of the risk of developing cardiovascular injury to the mitochondrial activity.

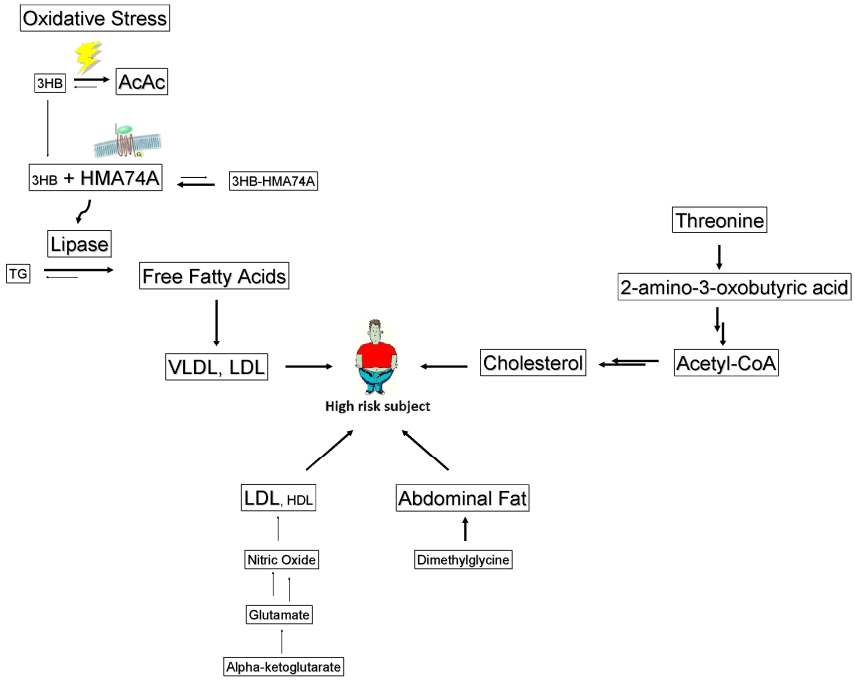


Figure 1 Metabolic effect of the discussed metabolites in high risk subjects.

Conclusions and perspectives

A metabolomic signature of cardiovascular risk exists, and its analysis may provide novel insights into the associated biochemical pathways and the current prevention therapies. Moreover, the good agreement with the calculated values of Framingham score supports the idea that metabolomic analysis of plasma samples can be an effective tool for monitoring the individual healthy status, and, in particular the risk of developing metabolic syndrome and cardiovascular disease. As discussed in the previous chapter, this good result is not completely unexpected because also here we are evaluating metabolomics as a tool for monitor the "health" of individual global metabolism. The mismatches between our prediction and the true Framingham values can be due to our ability to unravel subtle metabolic changes, which are not included in the classical risk scores that probably do not allow to take an overall picture of the living organism as well as metabolomics do. Our observations strongly suggest that metabolomic analyses should be included in future year-long prospective studies and a new (and possibly more efficient) cardiovascular score, as well as for the metabolomic risk assessed for the project on breast cancer (Chapter 4), have to be formulated on the basis of the whole metabolic fingerprint and/or on the basis of the plasma concentration of the discussed metabolites.

Bibliography

[1] Rang H.P.D., Ritter M.M., Flower J.M., Rang R.J. (2007). Dale'sPharmacology. *Elsevier, Philadelphia.*

[2] Minino A.M., Heron M.P., Murphy V.S.L. , Kochanek K.D. (2007). *Natl Vital Stat Rep* 55:1.

[3] http://www.statistics.gov.uk/pdfdir/hsq0506.pdf acessed on february 2 2009.

[4] National Vital Statistics Report (2002). Deaths: final data for 2000. 50(15).

[5] Maton A.(1993). Human Biology and Health. Englewood Cliffs, New Jersey: Prentice Hall.

[6] Reaven G.M. (2005). The metabolic syndrome: requiescat in pace. *Clin Chem* 51: 931-938.

[7] Reaven G.M.(2006). The metabolic syndrome: is this diagnosis necessary? *Am j clin nutr* 83:1237-1247

[8] Jackson R., Barham P., Maling T., MacMahon S., Bills. J et al. (1993). The management of raised blood pressure in New Zealand. *Br Med J* 307:107-110.

[9] Wilson P.W., D'Agostino R.B., Levy D., Belanger A.M., Silbershatz H. et al. (1998). Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* 97:1837-1840.

[10] Grundy S.M., Pasternak R., Greenland P., Smith S., Fuster V. (1999). Assessment of Cardiovascular Risk by Use of Multiple-Risk-Factor Assessment Equations A Statement for Healthcare Professionals From the American Heart Association and the American College of Cardiology. *J Am Coll Cardiol* 34(4):1348-1359.

[11] Olefsky J.M. (1997). Ellenberg and Rifkin's Diabetes Mellitus, 5th ed., 513.

[12] Samuel J.L., Schaub M.C., Zaugg M., Mamas M., Dunn W.B., et al. (2008). Genomics in cardiac metabolism. *Cardiovasc Res* 79(2): 218-227.

[13] Feala J.D., Coquin L., Paternostro G., McCulloch A.D. (2008). Integrating metabolomics and phenomics with systems models of cardiac hypoxia. Prog Biophys Mol Biol. 96(1-3):

209-225.

[14] Giovane A., Balestrieri A., Napoli C. (2008). New insights into cardiovascular and lipid metabolomics. J Cell Biochem 105(3):648-54.

[15] Lewis G.D., Asnani A., Gerszten R.E. (2008). Application of metabolomics to cardiovascular biomarker and pathway discovery. J Am Coll Cardiol 52(2):117-23.

[16] Mayr M., Madhu B., Xu Q. (2008). Proteomics and metabolomics combined in cardiovascular research. Trends Cardiovasc Med 17(2):43-8.

[17] Holmes E., Loo R.L., Stamler J., Bictash M., Yap I.K. et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 453(7193):396-400.

[18] Westerhuis J.A., Hoefsloot H.C.J., Smit S., Vis D.J., Smilde A.K. et al. (2008). Assessment of PLSDA cross validation. *Metabolomics* 2008;4:81-89.

[19] Befroy D.E., Petersen K.F., Dufour S., Mason G.F., de Graaf R.A. et al. (2007). Impaired mitochondrial substrate oxidation in muscle of insulin-resistant offspring of type 2 diabetic patients. *Diabetes* 56(5):1376-1381.

7. Ozone therapy and metabolomics

Applications

Several and of different nature are the research studies on reactive oxygen species (ROS), both free radicals and nonradicals species, among which ozone plays an important role (**Table 1**)[1]

Free radicals	Non radicals
Superoxide, $O_2^{\cdot-}$ ($1x10^{-3}$, $1x10^{-6}$ in presence of SOD[a])	Hydrogen peroxide, H_2O_2 (fairly stable in H_2O, 2.5 in plasma, < 1 in blood)
Hydroxyl, $HO^{\cdot}$ ($1x10^{-9}$)	Hypobromous acid, HOBr
Hydroperoxyl, $HO_2^{\cdot}$	Hypochlorous acid, HOCl (fairly stable)
Peroxyl, $RO_2^{\cdot}$ (0.1-7)	Ozone, O_3
Carbonate, $CO_3^{\cdot-}$	Singlet oxygen, $^1O_2$ ($1x10^{-6}$)
Carbon dioxide, $CO_2^{\cdot-}$	Organic peroxides, ROOH
Nitric oxide, $NO^{\cdot}$ (0.1-5)	Peroxynitrite, $ONOO^-$ (1-4)
	Peroxynitrous acid, NOOH

Table 1 Nomenclature of ROS (numbers into brackets represents half life values, in sec, at 37°C. [a]Superoxide dismutase

On the basis of previous results[2,3,4,5], ozone causes lipid peroxidation and can generate a variety of compounds.

In the stratosphere, ozone constitutes the protective layer around the planet while, at the ground level, it is a pollutant mixed with nitrogen oxides and hydrocarbons generated by human activities. At the same time for its fungicidal, virucidal and bactericidal capacities, it is used to purify the drinking water in many municipalities. The use of ozone for therapeutic purposes was reported in 1885 by doctor Charles J. Kenworthy and since that time it has been used in a rather empirical basis by many practitioners in the world. There are specific applications of ozone therapy in a number of pathologies such as vascular diseases, ulcers, acute and chronic viral diseases. On the basis of some biochemical studies[6,7,8,9] performed during the last fifteen years, a framework for understanding the basic mechanisms of action of ozone when it comes in contact with human blood at appropriate doses is now available. These data have been critically reviewed[10,11] and the following biological responses can be summarized as: a) improvement of blood circulation and oxygen delivery to ischemic tissue, owing to the concerted effect of NO and CO and a variable increase of intra erythrocytic 2,3-bisphosphoglycerate level[1213]; b) enhancement of the general metabolism [12]; c) up-regulation

of the cellular antioxidant enzymes and induction of Heme oxygenase (HO)-1 and Heat Shock Protein (HSP)-70[14,13]; d) induction of a mild activation of the immune system[6,7] and enhancement of the release of growth factors from platelets[15]; e) excellent disinfectant activity when topically used in water and oils[16,17] while this is negligible in the circulation owing to the potent blood antioxidant capacity[18]. Moreover, adequate ozone treatments do not procure acute or late side effects and often patients report a feeling of wellness[12]

Recently, the therapeutic ozone application in clinical dentistry by NMR has been studied[19].

Aim of the work

Our poupose was to investigate the effect on the metabonomic profile of normal human blood of: i) ozonation within the therapeutical concentration (0.42÷1.68 µmol/mL); ii) ozonation at high concentrations (3.36÷16.8 µmol/mL); iii) ozonation at abnormal high concentration (126 µmol/mL) associated with other stressors, as performed by Torre-Amione et al.[20] (Torre-Amione et al., 2008) in order to define specific alterations and toxicity regarding the amount of metabolites and to suggest specific markers of oxidative stress.

Results and discussion

The examination of CPMG and NOESY1Dpresat spectra of the control samples treated with pure oxygen did not yield particular differences with respect to basal samples, except a small increment of pyruvate peak at 2.35 ppm. On the other hand, samples treated with increasing amounts of ozone (from 0 to 16.8 µmol/mL) showed variations in the concentration of several metabolites. In particular, a strong increase of the peaks at 8.45 ppm (CH-formate), 5.40 ppm (CH-allantoin), 2.29 ppm (CH_3-acetoacetate) and 1.93 ppm (CH_3-acetate) were observed as well as a dramatic decrease of the 2.35 ppm peak (CH_3-pyruvate) and all the variations of these metabolites were strongly correlated with the concentration of insufflated ozone with R^2 values major than 0.95 (except for CH_3 -pyruvate) and a linear trend (except for CH-allantoni that showed an hyperbolic trend).

The variations of pyruvate and acetate are chemically correlated: acetate is formed by a direct ozone oxidation of pyruvate. Formate is the last product of the oxidative process of carbohydrates and allantoin comes from oxidation of uric acid[19,21] which behaves as a "sacrificial" molecule in the presence of a strong oxidant such as ozone[22,23]. Acetoacetate is often formed in plasma by oxidation of γ-hydroxybutyrate[24,19].

Notably, the main part of the information on the oxidative stress was obtained from the CPMG spectra: this observation suggests that, although it is known that ozone causes lipid peroxidation, ozone treatment doesn't cause visible alteration in lipid composition of blood in the range of the well-documented therapeutic concentrations of ozone[23].

Additional tests were performed in order to: i) confirm the trends of variations of each metabolites at higher concentrations of ozone, ii) confirm the specificity of the obtained results for the ozone treatment in comparison to other kinds of oxidative stress.

The increase of the ozone concentration up to 16.8 µmol/mL gas per mL of blood confirmed our findings for all the previous metabolites leading to the disappearance of the pyruvate signal from the spectrum caused by the excessive oxidative process. Moreover, following the treatment with an ozone concentration of 800 µg/mL, changes in several peaks, both in NOESY1Dpresat and in CPMG spectra were observed, leading to the hypothesis of a damage of cell membranes caused by lipid oxidation with possible cytolysis. It is important highlight as the first changes in the lipidic component appear only with high ozone concentrations not normally used for therapeutic purpose (5.04 µmol/mL upward).

A further analysis of the spectra highlighted the presence of unidentified new peaks. These peaks resonate at 2.82 and 8.08 ppm and they were not present in untreated plasma. Thus these peaks were due to specific products directly linked to the ozone activity and indeed their amounts in the plasma were related to different ozone concentrations (**Figures 1**).

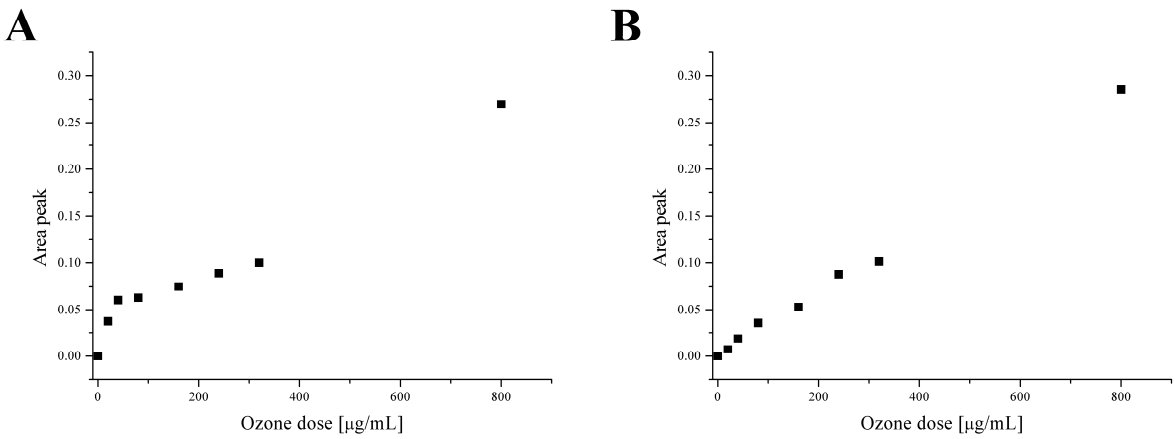


Figure 1 Linear trends of appearance of 2 unidentified peaks in relation to ozone dosages: signal areas at 2.82 ppm (A) and at 8.08ppm (B)(n ¼ 3;CV%<2). The correlations between the areas of each peak and the concentration of insufflated ozone are really high.

These data suggest that these peaks may represent markers of oxidative stress. In order to express an hypothesis about the origin of these two peaks other tests were performed.

It is well known that ozone causes lipid peroxidation yielding several aldehydes among which HNE is quantitatively relevant[25]. This aldehyde have a double bond in beta and it is not stable. It reacts easily with free thiol groups of cysteine, glutathione and other metabolites normally present in plasma, to form adducts easily eliminable from the organism with the urine[26,25]. Recent studies indicates an interaction of HNE with Cys34 present in albumin[27]. Starting from these data some samples obtained with the addition of scalar concentration of HNE (from 0.08 μM up to 3.2 μM) to a solution of 4% human albumin in saline were analyzed. After the addition of HNE to albumin, the two peaks at 2.82 ppm and 8.08 normally present after blood ozonation were absent, as well as the aldehydic signals of HNE. On the other hand, direct ozonation of a 4% albumin solution in saline allowed the detection of these two peaks which increased in intensity in relation to the amount of ozone concentrations, indicating that these two signals were due to the direct oxidant activity of ozone on the albumin. Indeed, also albumin has been categorized as a "sacrificial" molecule, although it has not been clarified whether albumin undergoes dimerization[28] or allows the formation of sulfenic acid[29].

Moreover, the experiments with albumin + uric acid (300 μM) and/or ascorbic acid (50 μM) showed an increase in the formation of allantoin. The peak at 4.5-4.6 (ascorbic acid) present in the control sample, disappeared as soon as ozone (0.42 μmol/mL) is added. These results suggest that uric acid is oxidized to allantoin and ascorbic acid is oxidized to dehydroascorbate. Additionally, the presence of formate as well as the appearance of both 2.82 and 8.08 ppm peaks were evidenced.

On the basis of the very harsh ozonation of blood (about 6 mg/mL of blood) plus UV radiation and heat (42.5 °C) recently reported[20], it was of interest to evaluate the spectra after blood treatment with: i) O_3; ii) O_3 + UV rays; iii) O_3 + heat ; iv) O_3 + UV rays + heat; v) UV rays; vi) heat; vii) UV rays + heat. No appreciable influence of heat and/or UV treatments on the metabolites present in the blood was detected, except than at 4.43 and 8.19 ppm in case of heat. These signals were not present in ozone-treated samples. The remaining other types of treatments (i-iv) showed the typical fingerprint alterations already seen in the previous tests (formate, acetoacetate, acetate, allantoin go up; pyruvate goes down; arising peaks at 2.82 ppm and 8.08 ppm) to indicate that the principal stressor is due to ozone.

To confirm metabonomic results of the effects of blood oxidative stress on the erythrocytic membrane and the possible cell damages with intraerythrocytic material release, physicochemical investigation based on hemolysis as well as viscosity were also performed. A statistically significant increase of hemolysis was observed at 16.8 μmol/mL, while

viscosity variations started from 160 µg/mL. The combination of ozone + UV (ii) as well as ozone + UV + heat (iv) yields the highest values of hemolysis, while the viscosity markedly increased only in the ozone + UV combination.

Conclusions and perspectives

The metabolic profile of blood samples treated with ozone is defined: i) the principal metabolic variation with the possible markers of oxidative stress due to ozone treatment were indicated; ii) blood alterations due to ozone were demonstrated to be strictly correlated to the concentration of insufflated ozone. However when ozone concentrations are within the therapeutic range (10 µg/mL blood, or 0.21 µmol/mL - 80 µg/mL blood, or 1.68 µmol/mL) only few alterations were detected and are not associated to significant damages. On the other hand several conditions, such as the association of ozone with heat and/or UV, appear deleterious, as showed by the physicochemical investigation based on hemolysis and viscosity that highlighted an extensive blood cell destruction. In other words, if ozonetherapy has to be safely used, it requires precise and well-calibrated ozone concentrations and the avoidance of ozone bubbling as well as additional stressor.

The present work demonstrates that the metabonomic approach is a valid instrument in the detection of alterations in plasma profile due to oxidative stress induced by ozone and other stressors. Moreover, in addition to the role in the investigation of diseases showed in the previous chapters, metabolomics shows here another promising application in the evaluation of the toxicity associated to the use of therapies and in the definition of the therapeutic windows.

Of course the further step in the investigation of ozone therapy will be the metabolomic evaluation of the effect of ozone therapeutical concentrations *in vivo*. Although we may expect that the results of the study *in vivo* will be less clear then those show in this chapter for the increased complexity of the system and for dilution effects, we are confident that also in the *in vivo* validation metabolomics will play a good role.

Bibliography

[1] Halliwell B.,Whiteman M. (2004). Measuring reactive species and oxidative damage in vivo and in cell culture: how should you do it and what do the results mean? *Br J Pharmacol* 142: 231-255.

[2] Esterbauer H., Schaur R.J., Zollner H. (1991). Chemistry and biochemistry of 4-hydroxynonenal, malonaldehyde and related aldehydes. *Free Radic Biol Med* 11: 81-128.

[3] Pryor W.A. (1993). Ozone in all its reactive splendor. *J Lab Clin Med* 122: 483-486.

[4] Pryor W.A., Squadrito G.L., Friedman M. (1995). The cascade mechanism to explain ozone toxicity: the role of lipid ozonation products. *Free Radic Biol Med* 19: 935-941.

[5] Esterbauer H. (1996). Estimation of peroxidative damage. A critical review. *Pathol Biol* 44: 25-28.

[6] Bocci V., Luzzi E., Corradeschi F., Paulesu L., Di Stefano A. (1993). Studies on the biological effects of ozone: 3. An attempt to define conditions for optimal induction of cytokines. *Lymphokine Cytokine Res* 12: 121-126.

[7] Bocci V., Valacchi G., Corradeschi F., Aldinucci C., Silvestri S. et al. (1998). Studies on the biological effects of ozone: 7. Generation of reactive oxygen species (ROS) after exposure of human blood to ozone. *J Biol Regul Homeost Agents* 12: 67-75.

[8] Bocci V., Aldinucci C. (2006). Biochemical modifications induced in human blood by oxygenation-ozonation. *J Biochem Mol Toxicol* 20: 133-138.

[9] Travagli V., Zanardi I., Silvietti A., Bocci V. (2007). A physicochemical investigation on the effects of ozone on blood. *Int J Biol Macromol* 41: 504-511.

[10] Bocci V. (2006). Is it true that ozone is always toxic? The end of a dogma. *Toxicol Appl Pharmacol* 21: 493-504.

[11] Bocci V. (2007a) The case for oxygen-ozonetherapy. *Br J Biomed Sci* 64: 44-49.

[12] Bocci V. (2005) *Ozone. A new medical drug*, Springer, Dordrecht, The Netherlands.

[13] Bocci V. (2009) The failure of the ACCLAIM trial is due to an irrational technology. *Int J Cardiol*, in press. DOI:10.1016/j.ijcard.2008.10.001.

[14] Bocci V., Aldinucci C., Mosci F., Carraro F., Valacchi G. (2007). Ozonation of human blood induces a remarkable upregulation of heme oxygenase-1 and heat stress protein-70. *Mediators Inflamm* 2007: 26785.

[15] Valacchi G., Bocci V. (1999). Studies on the biological effects of ozone: 10. Release of factors from ozonated human platelets Studies on the biological effects of ozone: 9. Effects of ozone on human platelets. *Mediators Inflamm* 8: 205-209.

[16] Sechi L.A., Lezcano I., Nuñez N., Espim M., Dupre I. et al. (2001). Antibacterial activity of ozonized sunflower oil (OLEOZON). *J Appl Microbiol* 90: 279-284.

[17] Travagli V., Zanardi I., and Bocci V. (2009) Topical applications of ozone and ozonated oils as anti-infective agents: An insight into the patent claims. *Recent Pat Antiinfect Drug Discov* 4: 130-142.

[18] Burgassi S., Zanardi I., Travagli V., Montomoli E., Bocci V. (2009) How much ozone bactericidal activity is compromised by plasma components? *J Appl Microbiol* 105: 1715-1721.

[19] Grootveld M., Silwood C.J., Lynch E. (2006) ^{1}H NMR investigations of the oxidative consumption of salivary biomolecules by ozone: relevance to the therapeutic applications of this agent in clinical dentistry. *Biofactors* 27: 5-18.

[20] Torre-Amione G., Anker S.D., Bourge R.C., Colucci W.S., Greenberg B.H. et al. (2008). Results of a non-specific immunomodulation therapy in chronic heart failure (ACCLAIM trial): a placebo-controlled randomised trial. *Lancet* 371: 228-236.

[21] Gruber J., Tang S.Y., Jenner A.M., Mudway I., Blomberg A. et al.(2009). Allantoin in human plasma, serum and nasal lining fluids as a biomarker of oxidative stress; avoiding artifacts and establishing real in vivo concentrations. *Antioxid Redox Signal*, in press.

[22] Shinriki N., Suzuki T., Takama K., Fukunaga K., Ohgiya S. et al. (1998). Susceptibilities of plasma antioxidants and erythrocyte constituents to low levels of ozone. *Haematologia* 29: 229-239.

[23] Bocci V., Borrelli E., Travagli V., Zanardi I. (2009). The ozone paradox: ozone is a strong oxidant as well as a medical drug. *Med Res Rev* 29: 646-682.

[24] Simons J.R., Theodore J., Robin E.D. (1974) Common oxidant lesion of mitochondrial redox state produced by nitrogen dioxide, ozone and high oxygen in alveolare macrophages. *Chest* 166: 9S-12S.

[25] Poli G., Schaur, R.J., Siems W.G., Leonarduzzi G. (2008). 4-Hydroxynonenal: a membrane lipid oxidation product of medicinal interest. *Med Res Rev* 28: 569-631.

[26] Mally A., Amberg A., Hard G.C., Dekant W. (2007). Are 4-hydroxy-2(E)-nonenal derived mercapturic acids and ^{1}H NMR metabonomics potential biomarkers of chemically induced oxidative stress in the kidney? *Toxicology* 230: 244-255.

[27] Aldini G., Vistoli G., Regazzoni L., Gamberoni L., Facino R.M. et al. (2008) Albumin is the main nucleophilic target of human plasma: a protective role against pro-atherogenic electrophilic reactive carbonyl species? *Chem Res Toxicol* 21: 824-835.

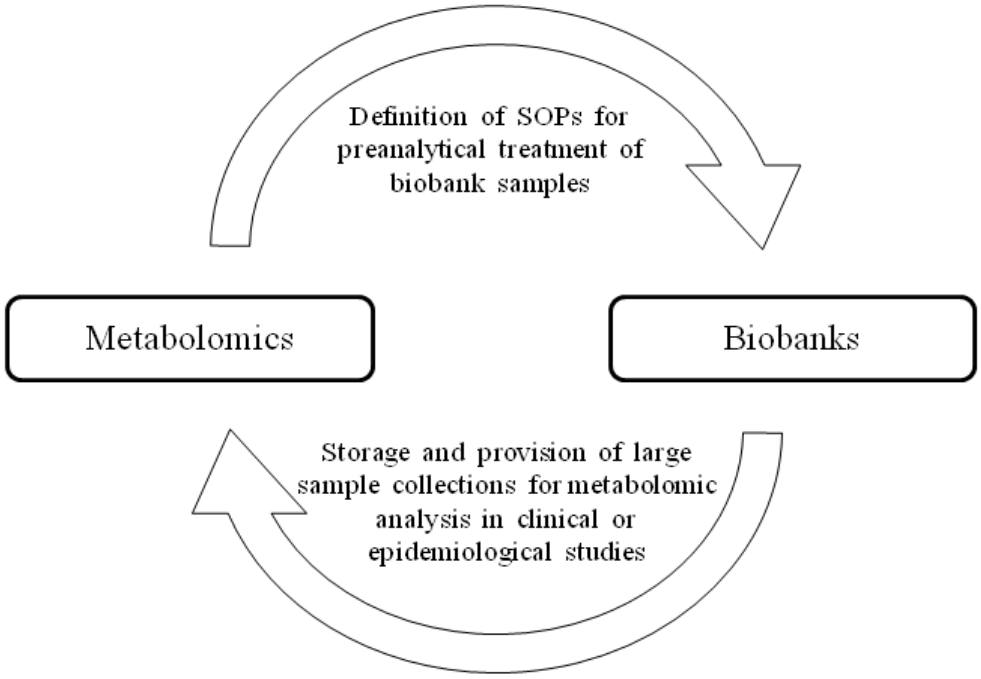
[28] Ogasawara Y., Namai T., Togawa T., Ishii K. (2006). Formation of albumin dimers induced by exposure to peroxides in human plasma: a possible biomarker for oxidative stress. *Biochem Biophys Res Commun* 340: 353-358.

[29] Carballal S., Radi R., Kirk M.C., Barnes S., Freeman B.A. et al. (2003). Sulfenic acid formation in human serum albumin by hydrogen peroxide and peroxynitrite. *Biochemistry* 42: 9906-9914.

8. SOPs definition by metabolomics

Introduction

The potential of metabolomics for disease diagnosis, prognosis and, in a clinical trial setting, for monitoring drug therapy relies on its ability to extract a disease signature from the multivariate analysis of the metabolic profiles of statistically relevant ensembles of samples derived from different donors[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19]. The reliability of the approach requires that the chemical nature and the relative concentration of the metabolites present in the biofluids are neither affected by the preanalytical conditions nor by the analytical methodology. For the great importance that this concept has in metabolomics, metabolomics itself is assuming a growing importance in the definition of operating procedures aimed at collecting and preserving biological samples according to protocols as standardized as possible. The idea of using metabolomic technologies for sample quality studies has been first proposed by the UK Biobank Pilot Study[20,21,22,23]. The definition of Standard Operating Procedures (SOPs) is essential for all those projects (often multicenter and retrospective) that are based on great numbers of samples[24]. Biobanks, that are infrastructure devoted to the collection, cataloguing and storing of biological material samples in order to make them available for medical and clinical research, represent an irreplaceable support for all those studies in which the impact of the results is linked to the large number of the collected samples. At the same time, they have to guarantee top quality biological samples for any possible future studies, including metabolomics (Scheme 1).



Scheme 1.

Among the different metabolomic tools, NMR represents a technique of choice[25] for the definition of pre-analytical procedures as it is a high-throughput tool requiring only minimal sample handling before spectra acquisition and allowing the contemporary detection of a large number of metabolites with different physiochemical properties. The variety of detectable molecules allows the monitoring of residual enzymatic activities and/or chemical reactions that may alter the NMR profile of the analyzed sample, which becomes no more representative of the metabolome before collection. These alterations could seriously bias the results of studies based on samples with a different collection, treatment and storage history, and differences in relative concentrations of the most sensitive metabolites should be considered with care.

Common biofluids in metabolomics

Most metabolomics studies, as that showed in this thesis, involve common biofluids as urine and serum/plasma that are obtainable from mammalians, especially from humans in a non- or poorly-invasively way and they are easy to find and collect because they are commonly used for many other clinical analyses.

Urine could seem a biospecimen with little scientific value, but it is assuming an increasingly importance as a bio-bankable sample thanks to metabolomics studies. Browsing PubMed (on 08/10/2010) for metabolomics/metabonomics publications resulted in 250 publications based on urine, 147 publications for serum, 185 for plasma and 185 for tissues. Use of urine as a first choice object of study is justified by the fact that the collection method is the simplest and the least invasive one: these characteristics allow the easy multiple sampling that guarantees reliability of the statistical analysis. In this thesis I showed that NMR analysis of urine is able to reveal individual metabolic phenotypes[26,27], but also to extract disease signatures (as seen for celiac disease in Chapter 5).

Blood (serum and plasma) is the second biofluid in order of importance in metabolomics studies (in terms of number of publications). The disadvantage with respect to urine is that the collection of blood samples is a little bit more invasive and it is not so easy perform multiple collections from patients and even more from healthy volunteers. On the other hand, blood is less affected by daily variations and daily diet than urine[27].

Aim of the work

In this study we focused on the most commonly used biofluids: urine, serum and plasma.

The starting point for this research are the Recommendations on Biobanking Procedures for urine processing and management recently published by the European Consensus Expert Group Report[28]. According to this document, biobanking procedures for urine should consider the following general consensus recommendations: i) cells and particulate matter should be removed (for example by centrifugation); ii) samples should be stored at -80 °C or below; iii) time limits for the processing should have been defined experimentally and should be appropriate to the analytes to be measured; iv) unless specified for a particular downstream analysis, urine samples should be stored without additives. Nevertheless, as the authors pointed out, advances in new downstream analytical platforms may impose specific requirements related to the nature of the targeted analytes, thus leading to specific revision/elaboration of the general procedures. Here we address the above critical issues in relation to the consequences for downstream metabolomics applications. Several methods of urine sample processing (centrifugation, filtration, addition of preservative as sodium azide) or combination of them as well as sample storage temperature (at -80°C or in liquid nitrogen) were tested and found to have significant effect on the metablome, as detected by ^{1}H-NMR. Presence of host or bacterial cells in urine is identified as a potential major source of alteration of the metabolic profile.

Moreover we analyzed here the quality of serum and plasma samples as a function of the time delay (0-4 h) and storage temperature (25°C and 4°C) before processing. Moreover changes in the NMR profiles of these two biofluids after serum and plasma preparation were monitored for a further 48 hours, to simulate the effect of the time between processing and freezing. A number of metabolites which are often used as disease biomarkers in metabolomics studies are heavily affected by the tested variation factors.

Results and discussion

Urine Fresh urine is characterized by the presence of human cells (erythrocytes, leucocytes, urothelial cells, epithelial cells), bacteria, fungi, sperm counts, non cellular components (mucus filaments, cylinders, cylindroids, pseudocylindres, crystals, urates). The protocol for the analysis of urinary sediment proposes a speed around 1600 RCF for the centrifugation[29] in

order to avoid the breaking of the frailest elements like cylinders. On the other hand, to avoid the presence of any particulate components in the NMR samples that will affect the spectral quality, standard protocols[30] for the immediate NMR analysis of urine require a centrifugation of the sample at 14000 RCF for 5 minutes at 4 °C. Such a high speed may induce cell breaking and release of cellular components into the biofluid. We thus analyzed the advantage of applying a mild centrifugation (pre-centrifugation) to fresh urine before the high-speed centrifugation needed by the preparation of NMR samples for immediate use (**Figure 1A**).

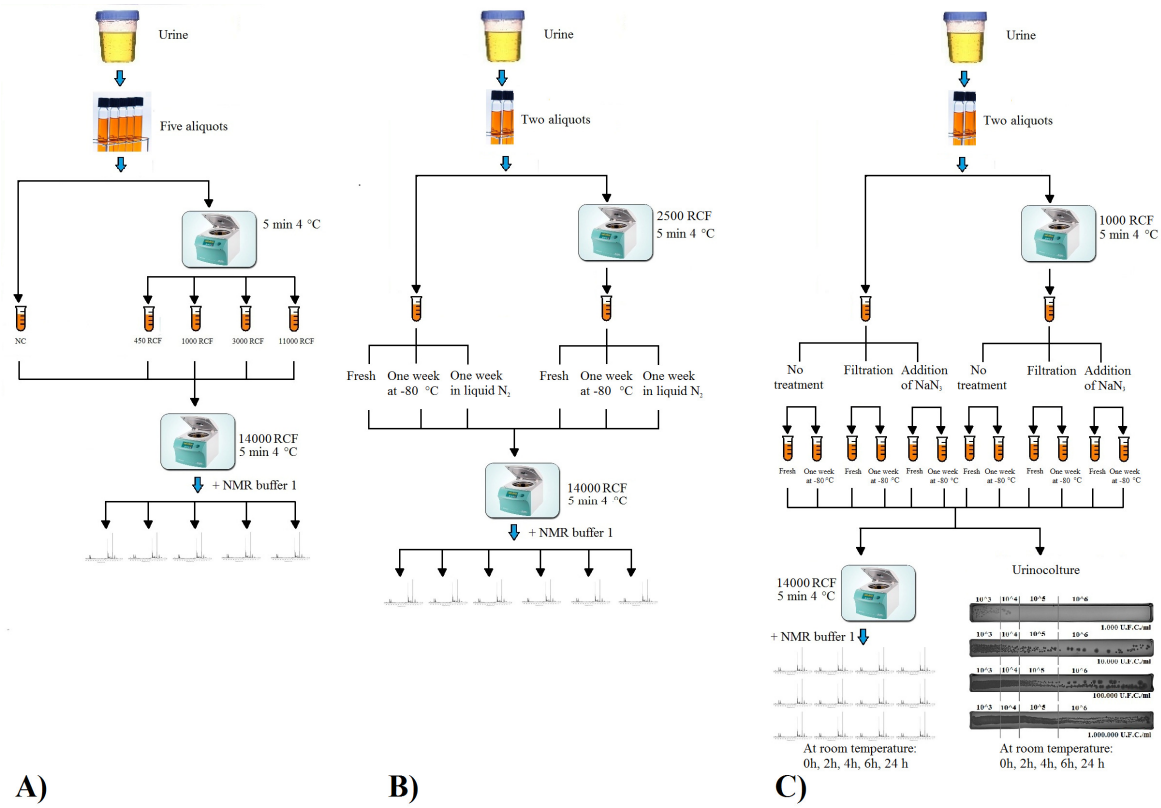


Figure 1 Experimental design for: **(A)** the evaluation of the impact of pre-centrifugation speed on the NMR metabolomic fingerprint, **(B)** the evaluation of the effect of the use of pre-centrifugation as a function of the temperature used for sample storage, **C)** the comparison of the efficiency of pre-centrifugation, filtration and addition of sodium azide in reducing bacterial content

Our results show that there is an effect of the pre-centrifugation speed. Apparently, the proposed 5-minutes 14000 RCF-centrifugation of the NMR protocol is too harsh as well as speeds ≥ 11000 RCF at the pre-centrifugation level. If a mild pre-centrifugation (between 1000-3000 RCF) is applied, the cellular components are spun down and the subsequent 14000 RCF centrifugation of the supernatant is useful for the elimination of other suspended, mainly

inorganic, particles. The importance of the pre-centrifugation step of course depends upon the content in cellular components of the analyzed samples, and may therefore be very different from one individual to another.

Upon inspection of the NMR spectra it appeared that the above discussed differences between precentrifuged at 2500 RCF or non-precentrifuged samples discussed above were due to chemical shift changes in signals that are extremely sensitive to pH variations. However, no resonances attributable to new metabolites or increase in intensity of existing signals were observed, suggesting that soluble components released by cells in fresh urine are below the detection limit.

The presence of cellular components has an impact on the NMR profiles even if non-fresh urine samples are analyzed, and the importance of the observed change depends upon the storage temperature (**Figure 1B**). NMR metabolic profiles of non pre-centrifuged samples stored for a week at –80 °C did differ from those that have undergone a mild (600-2500 RCF) pre-centrifugation before freezing at the same temperature. The effect is less severe if urine samples are stored in liquid nitrogen. This result is not unexpected because rapid freezing in liquid nitrogen, below the critical ice crystal temperature (–130 °C) avoids crystal formation which may cause cell breaking.

Finally, we evaluated the two most commonly used preservation methods[24]: filtration and addition of sodium azide. They were used as such or in combination with a pre-centrifugation step at 1000 RCF, as summarized in **Figure 1C**. Differently treated aliquots of the same sample were analyzed either fresh or after storage for a week at -80 °C, that is the temperature most commonly used in biobanking. Spectral variations were followed for 24 hours after urine processing or thawing after a week of storage at -80°C. Changes over time were observed for pH-sensitive metabolites that undergo chemical shift variations. Changes in the relative concentration of some molecules were also observed: succinate and acetate increase with time; urea, lactate and glutamate/glutamine decrease. The extent of their variations depends upon the preservation method, pre-centrifugation and storage, but, consistent with available literature data[24], filtration is the method that makes the NMR spectra more stable over time.

With the aim of evaluate the origin of the changes that affect the NMR spectra of samples over time, we tested three working hypothesis

- The chemical hypothesis: reactions (e.g. oxidations) occur with time.

- The bacterial hypothesis: the bacteria present in urine grow over time with the consequent production of bacterial metabolites.
- The enzymatic hypothesis: the enzymatic activities of urine cause the consumption of certain metabolites and the increase of others.

The three working hypotheses suggested above helped us in understanding the observed variations of the NMR signals, but it is also important to highlight that all three mechanisms are present and their effects may partially overlap. Instead, our results show that, in order to quench bacterial/enzymatic activity in urine a mild precentrifugation combined with filtration results to be the safest way to avoid contamination of the metabolome with soluble molecules derived from cellular components. Urine did not result to be sensitive to the presence of oxygen, but a number of enzymatic reactions occur in this biofluid. Addition of inhibitors is able to quench/slow-down such activities but invariably alters the metabolomic profile as it introduces in the spectra signals of the added molecules, whose presence may also induce changes in pH, ionic strength, further altering the whole spectral properties. In this sense, keeping the sample at the lowest possible temperature from collection throughout the analysis is extremely important.

With respect to the general recommendations in Yuille et al.[28], our studies permitted the definition of the best practices for the removal of cells and particulate matter. Nevertheless, as complete removal of cells is never achieved, long-term storing temperatures below the critical ice crystal temperature would be advisable. Changes in metabolite concentrations at room temperatures are relatively fast; reduction of time between collection and processing and the temperature at which urine samples are kept during this time delay is critical for a reliable NMR metabolic fingerprint. Annotation of time delays and temperatures for the entire history of each sample will add statistical value.

Blood According to the same recommendations document mentioned in the urine section[28], in the plasma/serum processing and management one should consider the following aspects: i) use of EDTA or citrate as anticoagulant, not heparin; ii) sample storage at a temperature of -80 °C, or below if e.g. critical ice crystal formation is problematic; iii) recording of the time from collection through processing; iv) experimental definition of time limits that are appropriate to the analytes to be measured. In that article the ‘‘gold standard’’ for all downstream applications was considered to be the extraction of DNA for genetic analysis. Here we validate the above procedures and evaluate whether additional procedures may be needed for the optimal maintainance of the original metabolome.

The time limits for processing whole blood were examined (**Figure 2A**) by evaluating the effect on the NMR metabolic profile of serum and plasma-EDTA samples.

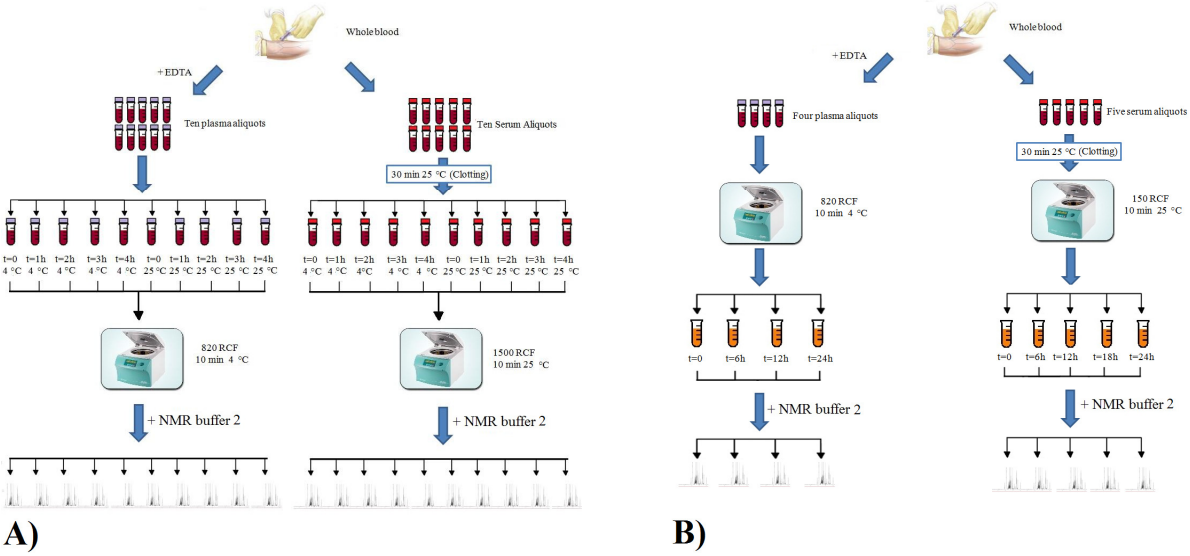


Figure 2 Experimental design for: (**A**) the evaluation of the effect of the time between blood collection and processing on serum and plasma NMR spectra; (**B**) the evaluation of the effect of the time between processing and freezing on the quality of NMR spectra of plasma and serum.

The PLS-CA analysis of CPMG spectra showed that the spectral patterns of serum and plasma samples processed at different times carry meaningful differences. Degradation processes are time-dependent and temperature-dependent both for serum and plasma samples: in both cases incubation at 25 °C causes deeper changes in the NMR profile. In order to unravel the main mechanisms at the basis of the degradation processes occurring from collection through processing, we singled out the significant buckets ($P \leq 0.05$) by Kruskal-Wallis test. They correspond to the resonances of glucose, lactate and pyruvate. The decrease in glucose concentration is more important in serum, most probably because EDTA in plasma tubes exerts an inhibitory effect on metalloenzymes and metal-dependent enzymes involved in glycolysis. An increase in lactate, the end product of the anaerobic glycolysis is observed, which is steeper than the decrease in glucose, roughly in agreement with the expected stechiometry of the reaction. Otherwise less clear is the behaviour of pyruvate that is an intermediate metabolite of glycolysis.

In clinical analyses, the standard for the measurement of glycaemia is represented by fluoride/oxalate coated vacutainers. It has been reported that glucose concentrations in plasma from blood collected into EDTA coated vacutainers showed no significant differences with glucose concentrations in plasma from fluoride/oxalate coated vacutainers (which represent

the standard for the measurement of glycaemia) up to 36 h[22,23]. Comparison of the changes in concentration of glucose and its derivatives during the time from collection through processing of citrate, EDTA and fluoride/oxalate plasma in the present work did not show meaningful differences. Therefore, none of these three anticoagulants solves the problem of glucose degradation and we can just suggest that for a reliable evaluation of glucose, lactate and pyruvate whole blood should be processed within 2 h from collection, keeping it at 4 °C, although the latter may still not be the best procedure for pyruvate.

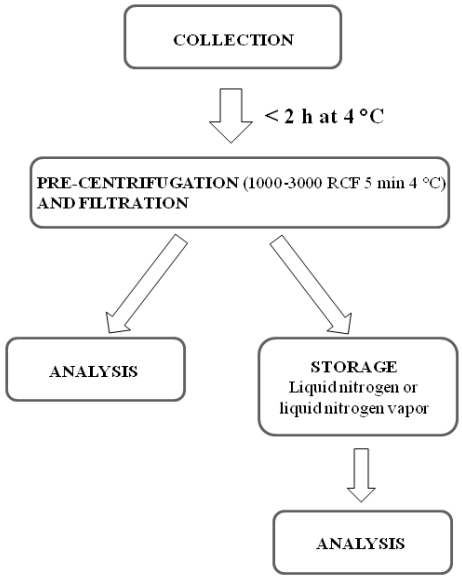
Possible degradation processes occurring in blood derivatives after processing were analyzed using 5 plasma and 5 serum samples of different patients (**Figure 2B**). Each sample was split in multiple aliquots and NMR spectra of each aliquot were acquired in the time range 0-24 hours keeping the samples at room temperature.

The PLS-CA analysis of the CPMG spectra showed that NMR spectra gradually shift in the same direction during the 24 h of monitoring time, the effects are larger for serum. In order to single out the significant buckets ($P \leq 0.05$), and thus the corresponding metabolites, that are responsible of this time progression we performed the Kruskal-Wallis test. The results of the test for serum identified the following metabolites as markers of the time changes: triglycerides, proline, choline, citrate and histidine; albumin and LDL/VLDL (low-density-lipoproteins/very-low-density lipoproteins) also contribute to the variations of the NMR profile. The Hε1 signal of His shifts with time, consistently with a pH variation increase of about 0.1 units. The signals of the other metabolites, but citrate, and those of albumin and LDL/VLDL decrease in intensity. The decrease for fatty acids and LDL/VLDL is attributable to oxidation reactions, as the effect is attenuated if samples are kept under inert atmosphere. The chemical shift variations observed for the signals of citrate are consistent with literature data about saliva[31] where the observed behavior has been rationalized in terms of a competition between calcium(II) and magnesium(II) binding to this chelator. Consistently, no chemical shift changes are detected in the present work for citrate in plasma EDTA-samples. The decrease in concentration observed in serum of all these molecules, but proline and choline, parallels that observed in plasma: otherwise no meaningful decrease in proline concentration is visible in plasma EDTA and citrate samples, whereas proline decreases sizably in plasma fluoride/oxalate samples. In summary, plasma appears to be slightly more stable than serum when kept at room temperature.

Conclusions and perspectives

Our findings help contribute to the definition of standard procedures for specimen collection that will be implemented at the local da Vinci European Biobank (https://www.davincieuropeanbiobank.org/) and represent a step forward with respect to the available recommendations for biobanking procedures[28].

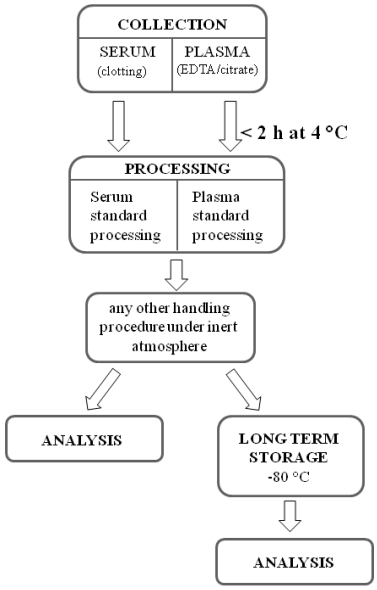
On the basis of the results of our experiments, we propose the following procedures for the optimal processing and management of urine samples to maintain as much as possible their original metabolome (Scheme 2): i) removal of cells and particulate matters through the combined use of a mild pre-centrifugation 1000-3000 RCF (5' at 4 °C) and filtration; ii) long-term storage of samples in liquid nitrogen (or liquid nitrogen vapour) to avoid breaking of residual cells; iii) fast processing (within 2 hours for collection); iv) storage at 4 °C between collection and processing. Recommendations at points iii) and iv) aim at reducing the effects of any possible enzymatic/cellular activities. Addition of additives (like enzyme inhibitors) should be avoided because the required concentrations will introduce signals in the NMR spectra covering the resonance of metabolites and may also induce changes in pH, ionic strength, etc. thus further affecting the original NMR profiles.



Scheme 1: Suggested procedure for the optimal processing and management of urine to be used for metabolomic NMR studies.

Based on the results collected on blood derivates, we propose the following procedures for the optimal processing and management of blood and its derivatives when they have to be used

for future metabolomic NMR studies (**Scheme 3**): i) either EDTA or citrate can be used as anticoagulant for plasma samples; ii) time between collection and processing should not exceed 2 h; iii) during this time delay samples should be kept at 4 °C; iv) any handling procedure of plasma and serum should be performed under inert atmosphere; v) samples should be frozen immediately after processing; vi) long term storage should be done at -80 °C (no need of lower temperature once erythrocytes have been removed).



Scheme 3.

Beyond the effective contribution that metabolomics could give to the definition of SOPs for biobanking and in general for collection of huge quantity of samples for epidemiological studies, clearly this work highlights the importance to define SOPs for the collection, treatment and storage that allow to guarantee the reproducibility of the metabolomics results themselves. Instead, most of the metabolomics studies, as well as that presented in this thesis, point to single out some significant metabolites that allow to better define the metabolomics signature of diseases. However, because several metabolites, as highlighted by this work, are really sensitive to the protocols adopted for the collection and treatment of the samples, it is necessary pay attention to the "history" of the analyzed samples in order to develop coherent discussions and conclusions on them.

The next step it will be obviously that of define by SOPs also for the less common samples studied by metabolomics as tissue, breath exhaled, saliva, cerebrospinal liquid and so on.

Bibliography

[1] Oakman C., Tenori L., Biganzoli L., Santarpia L., Cappadodona S. et al. (2010). Uncovering the metabolomic fingerprint of breast cancer. *Int J Biochem Cell Biol* in press.

[2] Sreekumar A., Poisson L.M., Rajendiran T.M., Khan A.P., Cao Q Yu J. et al. (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* 457: 910-914.

[3] Bertini I., Calabrò A., De Carli V., Luchinat C., Nepi S. et al. (2009). Metabonomic signature of celiac disease. *J Proteome Res* 8: 170-177.

[4] Akira K., Masu S., Imachi M., Mitome H., Hashimoto M. et al. (2008).^{1}H NMR-based metabonomic analysis of urine from young spontaneously hypertensive rats. *J Pharm Biomed Anal* 46:550-556.

[5] Holmes E., Loo R.L., Stamler J., Bictash M., Yap I.K.S. et al. (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 453: 396-401

[6] Bartsch T., Alfke K., Wolff S., Rohr A., Jansen O. et al. (2008) Focal MR spectroscopy of hippocampal CA-1 lesions in transient global amnesia. *Neurology* 70: 1030-1035.

[7] Gao X.X., Ge H.M., Zheng W.F., Tan R.X. (2008). NMR-based metabonomics for detection of Helicobacter pylori infection in gerbils: which is more descriptive. *Helicobacter* 13: 103-111.

[8] Fearnside J.F., Dumas M.E., Rothwell A.R., Wilder S.P., Cloarec O. et al. (2008). Phylometabonomic patterns of

adaptation to high fat diet feeding in inbred mice. *PLoS ONE* 3:e1668.

[9] Makinen V.P., Soininen P., Forsblom C., Parkkonen M., Ingman P. et al. (2008) ^{1}H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. *Mol Syst Biol* 4: 167

[10] Teichert F., Verschoyle R.D., Greaves P., Edwards R.E., Teahan O. et al. (2008) Metabolic profiling of transgenic adenocarcinoma of mouse prostate (TRAMP) Tissue by (1)H-NMR analysis: evidence for unusual

phospholipid metabolism. *Prostate* 68: 1035-1047

[11] Coolen S.A., Daykin C.A., van Duynhoven J.P., van Dorsten F.A., Wulfert F. et al. (2008) Measurement of ischaemia-reperfusion in patients with intermittent claudication using NMR-based metabonomics. *NMR Biomed* 21: 686-95.

[12] Claudino W.M., Quattrone A., Biganzoli L., Pestrin M., Bertini I. et al. (2007) Metabolomics: available results, current research projects in breast cancer, and future applications. *J Clin Onc* 25: 2840-2846

[13] Schnackenberg L.K., Sun J., Espandiari P., Holland R.D., Hanig J. et al. (2007) Metabonomics evaluations of age-related changes in the urinary compositions of male Sprague Dawley rats and effects of data normalization methods on statistical and quantitative analysis. *BMC Bioinformatics* 8 Suppl 7: S3

[14] Marchesi J.R., Holmes E., Khan F., Kochhar S., Scanlan P. et al. (2007). Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *J Proteome Res* 6: 546-551

[15] Griffin J.L., Scott J., Nicholson J.K. (2007) The influence of pharmacogenetics on fatty liver disease in the wistar and kyoto rats: a combined transcriptomic and metabonomic study. *J Proteome Res* 6: 54-61.

[16] Constantinou M.A., Theocharis S.E., Mikros E (2007) Application of metabonomics on an experimental model of fibrosis and cirrhosis induced by thioacetamide in rats. *Toxicol Appl Pharmacol* 218:11-19.

[17] Coen M., O'Sullivan M., Bubb W.A., Kuchel P.W. et al. (2005) Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clin Infect Dis* 41:1582-1590.

[18] Lindon J.C., Holmes E., Nicholson J.K. (2004) Toxicological applications of magnetic resonance. *Prog NMR Spectrosc* 45: 109-143.

[19] Brindle J.T., Antti H., Holmes E., Tranter G., Nicholson J.K. et al. (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* 8: 1439-1444.

[20] Barton R.H., Nicholson J.K., Elliott P., Holmes E. (2008). High-throughput ^{1}H NMR-based metabolomic analysis of human serum and urine for large-scale epidemiological studies: validation studi. *Int J Epidemiol*, 37:i31-i40.

[21] Dunn W.B., Broadhurst D., Ellis D.I., Brown M., Halsall A. et al. (2008). A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. *Int J Epidemiol*, 37:i23-i30.

[22] Jackson C., Best N., Elliott P. (2008). The UK Biobank sample handling and storage validation studies. Int J Epidemiol, 37:i2-i6.

[23] Peakman T.C., Elliott P. (2008) UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. *Int J Epidemiol* 37 :i16-i22

[24] Saude E.J., Sykes B.D. (2007) Urine stability for metabolomic studies: effects of preparation and storage. *Metabolomics,* 3: 19-27.

[25] Nicholson JK, Lindon JC (2008) Metabonomics. *Nature* 455: 1054-1056.

[26] Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H. et al. (2008) Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U.S.A. 105: 1420-1424.

[27] Bernini P., Bertini I., Luchinat C., Nepi S., Saccenti E. et al. (2009) Individual human phenotypes in metabolic space and time. *J Proteome Res* 8: 4264-4271.

[28] Yuille M., Illig T., Hveem K., Schmitz G., Hansen J. et al. (2010). Laboratory management of samples in biobanks: European consensus expert group report. *Biopreservation & Biobanking*, 8(1): 65-69.

[29] Koukoulaki M.., Donovan M.O., Pursglove S., Alexopoulou D., Hadjiconstantinou V. et al. (2008) Prospective study of urine cytology screening for BK polyomavirus replication in renal transplant recipients. *Cytopathology*, 19: 385-388

[30] Beckonert O., Keun H.C., Ebbels T.M.D., Bundy .J, Holmes E. et al. (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature*, 11: 2692-2703

[31] Silwood C.J.L., Grootveld M., Lynch E. (2002) ^{1}H NMR investigations of the molecular nature of low-molecular-mass calcium ions in biofluids. *J Biol Inorg Chem*, 7: 46-57

9. Conclusion

Conclusions

Any study aimed at the identification of relevant metabolites should be presented with reference to the normal or control population in order to be able to allow the identification of relevant metabolic changes. This it was considered as the starting point of this thesis work (Chapter 3).

Besides the perspectives opened by each project, that are reported in the appropriate section of each proper chapter, it is here generally demonstrated the feasibility of metabolomics studies for the interpretation of three large impact pathologies (breast cancer in chapter 4, celiac disease in chapter 5 and cardiovascular risk in chapter 6). In all these researches is the role of metabolomics not only as diagnostic tool, but also as a tool for the understating of the underlying biochemical mechanism is demonstrated.

In addition, as showed in this thesis, because NMR is able to provide predictive profile of different kind of diseases, its role in this field is unquestionable. But the its role is so exclusive that NMR is even presented in Chapter 8 as an election technique[1] for the definition of pre-analytical procedures. Indeed, NMR is a high-throughput methodology requiring only minimal sample handling before spectra acquisition and allowing collection of the whole metabolic profile, that means the simultaneous detection of a large number of metabolites with different physiochemical properties (hydrophobicity/hydrophilicity, acidity/basicity, redox reactivity).

The obtained results demonstrated that the NMR based metabolomic approach can be ideally extended to other pathologies. Indeed some other metabolomics projects have already started or are in a starting-phase. These new projects involve some widespread pathologies such as diabetes (type II), lung carcinoma, Obstructive and Chronic BroncoPathy, liver cirrhosis and carcinoma, Helicobacter Pylori infection or validation of health status indicators (geriatric assessment). The study of these new pathologies lead us to better define their characteristic, but also allow us to clearly understand what are the ideal targets and the limits of the metabolomic approach.

Moreover the application of metabolomics in the evaluation of toxicity of therapeutical treatments is assessed in the project on the effects of ozone therapy on blood (Chapter 7)

Furthermore it is showed the necessary to continuously improve the step of the standardization: part of the work was done for the definition of standard operative procedures

for sample handing/storage (Chapter 8). But a lot of a work aimed to asses all the possible causes of variability have to be done. This because metabolomics is a relatively new technique and because the analytical technique itself is still undergoing optimization. This is particularly true and necessary when we start to analyze new fluids, such as exhaled breath condensate, saliva or tissue extract for instance.

The same considerations apply also to statistical and NMR methods. The development of new simple 1D NMR pulse sequences, namely the 1D diffusion filtered recently implemented, can be very useful to single out some information: it allows us to obtain information only about macromolecules present in serum, and in general in a biofluid. Moreover the development of new statistical analysis could further increase the obtainable data.

Finally the role of metabolic in systems biology is confirmed (**Figure 1**). Systems biology is one of the strongest driving forces in modern biomedical research with increasing emphasis on the integration of various "omics" science to understand integrated system activity[2]. Metabonomics dovetails beautifully with the spirit of systems biology, because it provides a 'top-down', integrated view of biochemistry in complex organisms, as opposed to the traditional 'bottom-up' approach that investigates the network of interactions between genes, proteins and metabolites in individual cell types[1]. Given the clear results obtained using metabolomics for the three large impact pathologies described in this thesis and, more in general, for other several pathologies that afflict human being, as shown by literature and extensively discussed in the INTRODUCTION, I would like to think about a future, hopefully closer as possible, in which metabolomics it will be the omic science that really will carry system biology to the bed of the patient.

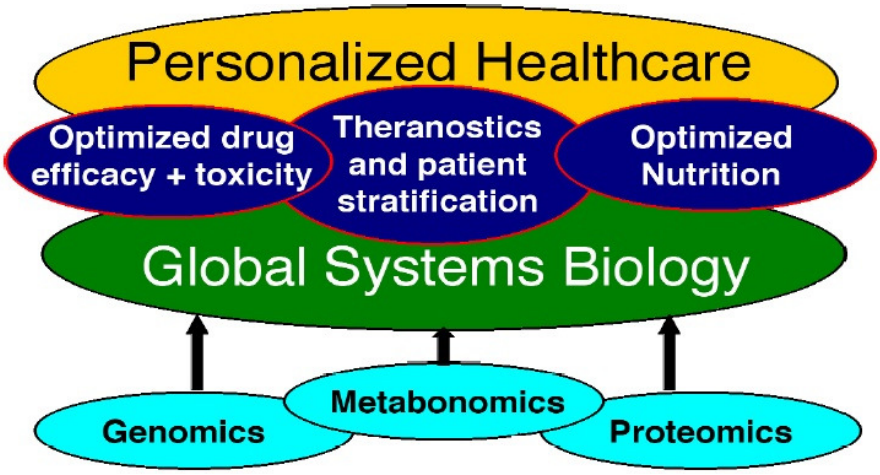


Figure 1 Metabolomics in system biology (adapted from Nicholson J.K. & Holmes E., 2006[3])

Bibliography

[1] Nicholson J.K., Lindon J.C. (2008) Metabonomics. Nature 455: 1054-1056

[2] Hood L., Heath J.R., Phelps M.E., Lin B. (2004). Systems biology and new technologies enable predictive and preventative medicine. Science 306(5696):640-643.

[3] Nicholson J.K., Holmes E. (2006). Global systems biology and personalized healthcare solutions. Discov Med 6(32): 63-70.

10. Publications

Publications list

1) Bernini P., Bertini I., Luchinat C., Nepi S., Saccenti E., Schäfer H., Schütz B., Spraul M., Tenori L. (2009). Individual human phenotypes in metabolic space and time. *J Proteome Res*, 8: 4264-4271.

2) Oakman C., Tenori L., Claudino W.M., Cappadona S., Nepi S., Battaglia A., Bernini P., Zafarana E., Saccenti E., Fornier M., Morris P.G., Biganzoli L., Luchinat C., Bertini I., Di Leo A. (2010). Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods. *Ann Oncol*, accepted.

3) Bernini P., Bertini I., Calabrò A., la Marca G., Lami G., Luchinat C., Renzi D., Tenori L. (2010). Are patients with potential celiac disease really potential? The answer of metabonomics. *J Proteome Res*, under referee judgment.

4) Bernini P., Bertini I., Luchinat C., Tenori L., Tognacini A. (2010). The metabonomic signature of cardiovascular risk. Ready for the submission to *Circulation.* No present in this chapter.

5) Travagli V., Zanardi I., Bernini P., Nepi S., Tenori L., Bocci V. (2010). Effects of ozone blood treatment on the metabolite profile of human blood. *Int J Toxicol*, 29(2):165-74.

6) Bernini P., Bertini I., Luchinat C., Nincheri P., Staderini S., Turano P. (2010). Standard Operating Procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. Ready for the submission to *J Biomol NMR.* No present in this chapter.

Individual contribution

Project	Chapter	Publication	Role
Metabolic phenotypes	**3**	**1**	Collection of NMR spectra, processing and biological interpretation
Breast cancer and metabolomics	**4**	**2**	Collection of NMR spectra, processing and biological interpretation
Celiac disease and metabolomics	**5**	**3**	Collection of NMR spectra and processing
Cardiovascular risk and metabolomics	**6**	**4**	Collection of NMR spectra, processing, part of the statistical analysis and biological interpretation
Ozone therapy and metabolomics	**7**	**5**	Collection of NMR spectra, processing, part of the statistical analysis and biological interpretation
SOPs definition by metabolomics	**8**	**6**	Collection of NMR spectra, processing, part of the statistical analysis and biological interpretation

research articles Journal of proteome research

Individual Human Phenotypes in Metabolic Space and Time

Patrizia Bernini,[†,#] Ivano Bertini,[*,†,‡] Claudio Luchinat,[†,¶] Stefano Nepi,[†,#] Edoardo Saccenti,[†,#] Hartmut Schäfer,[§] Birk Schütz,[§] Manfred Spraul,[§] and Leonardo Tenori[†,#]

Magnetic Resonance Center (CERM), University of Florence, Via L.Sacconi 6, 50019 Sesto Fiorentino, Italy, Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy, Department of Agricultural Biotechnology, University of Florence, Via Maragliano 75-77, 50144 Florence, Italy, FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy, and Bruker BioSpin GmbH, Silberstreifen, D-76287 Rheinstetten, Germany

Received April 15, 2009

Differences between individual phenotypes are due both to differences in genotype and to exposure to different environmental factors. A fundamental contribution to the definition of the individual phenotype for clinical and therapeutic applications would come from a deeper understanding of the *metabolic phenotype*. The existence of unique individual metabolic phenotypes has been hypothesized, but the experimental evidence has been only recently collected. Analysis of individual phenotypes over the timescale of years shows that the metabolic phenotypes are largely invariant. The present work also supports the idea that the individual metabolic phenotype can also be considered a metagenomic entity that is strongly affected by both gut microbiome and host metabolic phenotype, the latter defined by both genetic and environmental contributions.

Keywords: Metabolomics • NMR spectroscopy • individual metabolic fingerprint • human metabolites • gut microflora metabolites • principal component analysis

Introduction

Individual phenotypes differ because individuals are genetically different, but also because of environmental factors. It is the phenotype, rather than the genotype alone, that matters when dealing with the development of pathologies or with the individual responses to drug treatments or external stimuli.

An important contribution to the definition of an individual phenotype for medical purposes would come from the knowledge of the *metabolic phenotype*. The metabolic phenotype (or *metabotype*[1]) is defined as a "multiparametric description of an organism in a given physiological state based on metabolomic data". Life, indeed, is maintained by numerous biochemical cycles, molecular and cellular mechanisms, whose maintenance in steady states defines homeostasis. Individual metabolic phenotypes can be regarded as images of these steady states: deviations from the optimal conditions can be directly linked to a pathophysiological status.

The existence of individual metabolic phenotypes has been repeatedly postulated,[2−4] but experimental hints to its existence have been only recently obtained.[5,6] The elusiveness of metabolic phenotypes stems from the fact that the individual metabolic profiles, such as those obtained from NMR spectra of urine samples, can vary daily, and the variation is sufficiently large as to make profiles from different individuals heavily overlap. The source of variation can arise from random factors (e.g., diet) or from cyclical biochemical variations: for example, observation of cyclical variations of potassium content in human urine dates back to the 19th century. More recently, with the development of chronobiology, the periodicity of many biochemical cycles such as those linked to hormonal secretion and their circadian periodicity[7−10] has been assessed. Nevertheless, recently, the collection of multiple urine samples from each individual over a period of a few weeks provided clear evidence that beyond these numerous sources of variations there is a "stable" part of the profile, which was then found to be strongly individual-specific.[5] It was concluded that the stable part of the profile was maintained at least for the few weeks of the collection period, but a more extended analysis to unravel and characterize the intrinsic metabolic phenotype complexity was not attempted. For an individual metabolic phenotype to be of physiological relevance and medical usefulness, one should now be able to show that it is relatively stable over time for a period of at least a few years and, above all, what are the most relevant causes of its variability, if any, and their relative weight with respect to a possible invariant contribution of genetic origin.

The present research addresses the above issues by a thorough statistical analysis of sets of about 40 urine samples from each of 31 different individuals, out of which 11 provided two 40-samples sets separated by 2 years, and 4 of them provided again 40-samples sets for a third time 1 year later, for a total of 1849 samples. It is found that the individual metabolic phenotype is largely stable over the examined period

* Corresponding author: Ivano Bertini, Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy. E-mail: bertini@cerm.unifi.it. Tel.: 39-055-4574270. Fax: 39-055-4574271.
† Magnetic Resonance Center (CERM), University of Florence.
# FiorGen Foundation.
‡ Department of Chemistry, University of Florence.
¶ Department of Agricultural Biotechnology, University of Florence.
§ Bruker BioSpin GmbH.

10.1021/pr900344m CCC: $40.75

Published on Web 06/15/2009

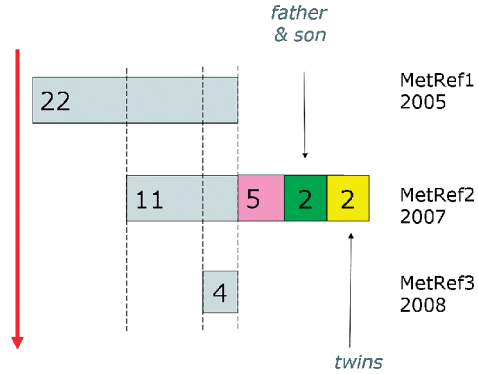


Figure 1. Collection scheme: 22 healthy individuals were involved in the previous study, with sample collection in late-spring−early-summer 2005. Eleven of them (plus nine new individuals) were recruited in late-spring−early-summer 2007 and 4 of them were again recruited for another sample collection in late-spring−early-summer 2008.

of 2−3 years, that no age-related drift is apparent, that modest changes of the phenotype may occur on a variety of timescales, and that they are mainly due to changes in lifestyle and gut microflora. Analysis of individual phenotypes over the timescale of years demonstrates that metabolic phenotypes are the results of complex interaction between environmental factors and genetics. The effects of the environment are observable in the periodical and/or sporadic metabolic fluctuations, while a combination of genetic and environmental contributions is present in the stable, invariant part of the metabolic phenotype that allows to recognize spectra as belonging to a unique subject even over a period of some years.

Materials and Methods

Sample and Metadata Collection. About 40 urine samples (first in the morning, preprandial) were collected from 20 healthy individuals (9 males, 11 females) in the age range 25−55 over a period of about 3 months, in late-spring−early-summer 2007. Eleven of them (6 males and 5 females) had taken part in a previous study involving 22 subjects,[5] with sample collection in late-spring−early-summer 2005, and 4 of them (males) were again recruited for another sample collection in late-spring−early-summer 2008. The collection scheme is summarized in Figure 1.

In total, 944 samples were analyzed. Samples were collected from each individual in sterile 15-mL propylene tubes, frozen within 4 h of collection, and stored at −80 °C. All subjects were Caucasian. Personal data were collected from every subject, including gender, age, BMI, and general habits such as practiced physical activity and normal diet. A detailed diet sheet relative to the day before each collection was also provide by each donor. Some subjects presented peculiarities: one quit smoking during the time between the first and the second collection, another one moved from Italy (2005 collection) to Spain (2007 collection). Moreover, two subjects are homozygous twins and two other subjects are father and son.

Sample Preparation. Frozen samples were thawed at room temperature and shaken before use. Aliquots of each urine sample (630 μL) were added to 70 μL of sodium phosphate buffer (0.2 M Na_2HPO_4 and 0.2 M NaH_2PO_4 in 100% $2H_2O$, pH 7.0) to minimize variations in metabolite NMR chemical shifts arising from differences in urinary pH, also containing 10 mM sodium trimethylsilyl [2,2,3,3−$2H_4$]propionate (TSP) and 30 mM sodium azide. Samples were centrifuged at 14 000*g* for 5 min to remove any solid debris; 450 μL of the supernatant was placed in 4.25-mm-outer-diameter NMR tube (for Bruker MATCH holder). All chemicals were from Sigma Aldrich.

NMR Experiments. The urine samples were analyzed by 600 MHz 1D ^{1}H NMR spectroscopy. The 1D ^{1}H NMR spectrum of each sample was acquired with water peak suppression using a standard pulse sequence (noesypresat; Bruker). For 12 selected samples taken from the four donors that participated in all three collections (one sample for each subject from each collection), a set of 2D spectra was acquired to improve metabolites identification using a 900 MHz spectrometer (Bruker BioSpin). The set of 2D spectra is constituted by the following experiments: JRES, COSY, ^{1}H−^{13}C-HMBC (see Figure 2), ^{1}H−^{13}C-HSQC, TOCSY. All experiments were performed using Bruker library sequences. Further details are reported in Supporting Information.

Data Reduction and Preprocessing of the ^{1}H NMR Spectra. The generation of input variables for statistical analysis was done by bucketing. For this, ^{1}H NMR spectra were normalized to the NMR signal intensity of the CH_3-group of creatinine which was calculated via nonlinear fitting and localized spectral deconvolution in order to take off contributions from the neighboring creatine signal (further information given in the Supporting Information). After scaling spectra, bucketing was applied to the data where the spectral regions $\delta > 9.5$, δ 6.0−4.5, and $\delta < 0.5$ were discarded before dividing the remainder of each spectrum into sequential segments ("bins") of 0.02 ppm width and obtaining an integral for each segment. Finally, prior to predictive multivariate statistics, buckets were log-transformed.

Statistical Analysis: descriptive level. Statistical data analysis was performed in MATLAB (The MathWorks, Natick, MA) using standard procedures provided in MATLAB itself and in the MATLAB STATISTICS TOOLBOX and using in-house written routines.

Principal Component Analysis (PCA) was used as dimension reduction technique and as a preparation of the data table for further multivariate statistical analyses. Data reduction was carried out by means of projection into a PCA subspace explaining 99.9% of the variance in the data.

Multivariate statistical analysis was performed using the MANOVA implementation of the Statistical Toolbox in MATLAB. The dimensionality of the relevant subspace was defined by the dimensionality of the group means provided by the MANOVA output (see comments on MANOVA in the Supporting Information). Linear combinations (canonical variables) of the original variables that reflect the largest possible separation between groups were obtained through Canonical Analysis (CA).

Hierarchical Cluster Analysis (HCA) was performed using the HCA implementation of the Statistical Toolbox in MATLAB. Euclidean distance was considered.

Pairwise Comparison of Data Sets. For each possible pair of the 46 different data sets, a Kruskal−Wallis test was applied to assess the mean differences in metabolites concentrations. A conservative estimate of significative levels of 1.288×10^{-8} was derived by using the Bonferroni correction on a nominal value of 0.005. Kruskal−Wallis test (nonparametric ANOVA test analogue) was chosen to infer differences on the biological asymptotic assumption that metabolite concentrations are not normally distributed. All resonances of interest were then manually checked and signals were assigned on template one-

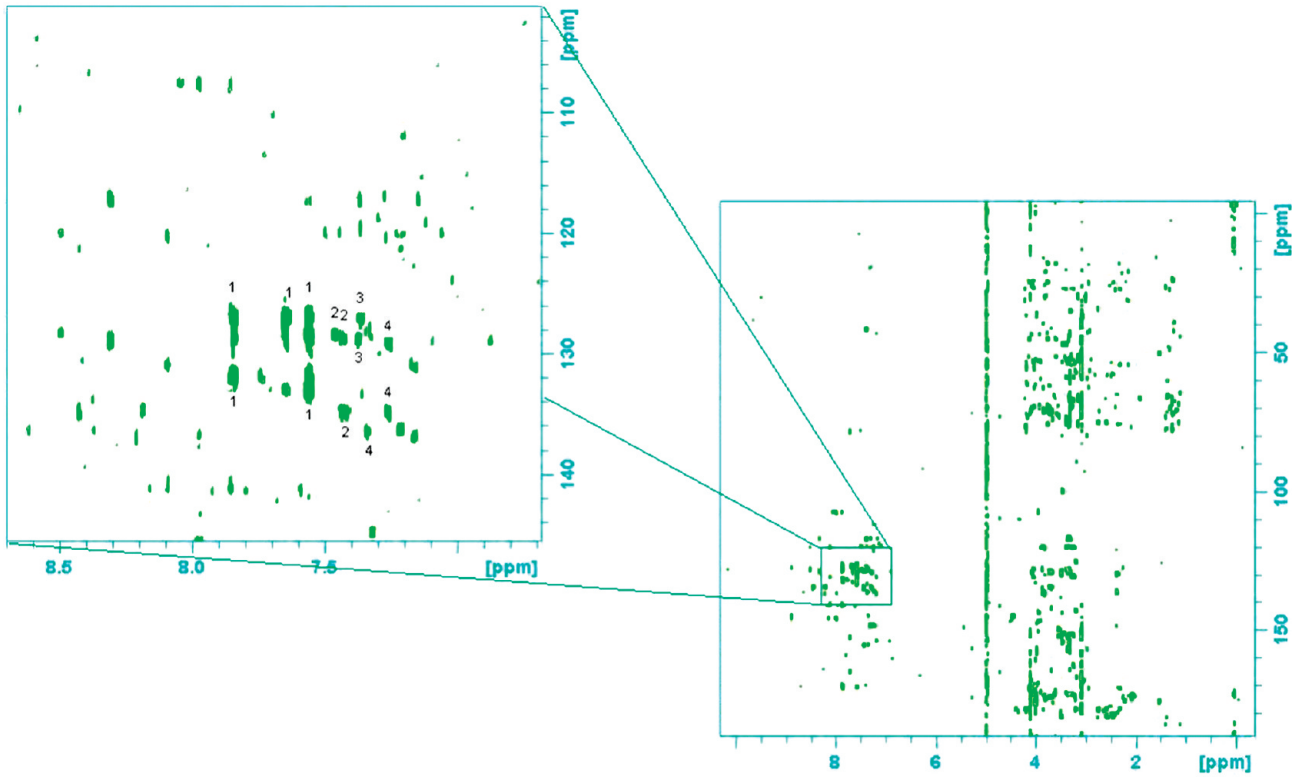


Figure 2. Example of a 2D HMBC spectrum acquired at 900 MHz. The zoomed zone represents the aromatic region. It possible to distinguish some aromatics peaks of Hippurate (1), PAG (2), Histidine (3) and mHPPA (4).

dimensional NMR profiles by using matching routines of AMIX (Bruker BioSpin GmbH Rheinstetten, Germany) in combination with the BBIOREFCODE (Version 2-0-0; Bruker BioSpin) reference database and published literature when available.

The most representative spectrum for each individual was chosen with the criterion of minimizing the total sum of square deviations from the average spectrum. The corresponding samples were used for analysis at 900 MHz NMR.

Statistical Analysis: Predictive Level. The key of the validation of the NMR predictivity on person identity relies on the Test Set Validation (TSV) approach which requires that models do not have any knowledge on the existence of any test set data. Two principal validation situations needed to be addressed, that is, within-collection validation (e.g., model set and test set both from METREF1) and cross-collection validation (e.g., model set from METREF1, test set from METREF2). Within-collection validation was performed via Monte Carlo embedded leave-25%-out cross-validation MC/CV. Key of cross-collection validation was straightforward test set validation performed multiple times always using new model sets obtained from Monte Carlo based selection of 75% subsets from the complete model data (e.g., 75% of METREF1 to classify METREF2 spectra). In each validation, 1000 Monte Carlo iterations were applied.

In each test set validation step, model set data were mean centered and unit variance scaled first. PCA was applied to the model set and scores of the subspace explaining 99.9% of the variance were further subjected to CA in order to define a further reduced subspace with optimum group separation; dimensionality was chosen according to the MANOVA output of the dimensionality of the respective group means. Then, test set data were mean centered and unit variance scaled using **model set mean** and **model set variance values**. Then, they were projected into the PCA/CA subspace defined by the model set. Finally, k-NN classification was applied to each test set object, evaluating the distance to the nearest model set objects. "Majority" rule was applied to classify training-set donors via a preselected number (7) of test set samples per test set donor. From this, (cross-) collection and donor specific probabilities of correct and false classification (confusion matrix) were calculated. Detailed information on the overall procedure is given in the Supporting Information.

Results and Discussion

Metabolic Space. A dendrogram representing the intersampling distances in the discriminant space of the PCA/CA components is shown in Figure 3b, with the clusterization of the 20 individuals of the 2007 collection. Figure 3a shows for comparison the analogous representation for the 22 individuals of the 2005 collection.[5] It is apparent that on the descriptive level the separation of individuals is of the same quality in the two cases. Figure 3c shows the same representation where all 31 different individuals are reported. Clearly, no worsening of the separation is observed.

On the predictive level, the average recognition rate of randomly chosen spectra in the 2007 collection using the single vote classification was 99.5%, practically identical to the 99.7% in the 2005 collection, while the MRC classification gave 100% in both cases. Table 1 reports the individual single vote scores for the two collections, where it is apparent that most individuals in both collections had scores higher than 99% and several individuals scored 100% recognition even with single vote classification. Again, pooling together the 31 different individuals did not spoil these exceptionally high percentages (99.6% average recognition rate). These results suggest that the number

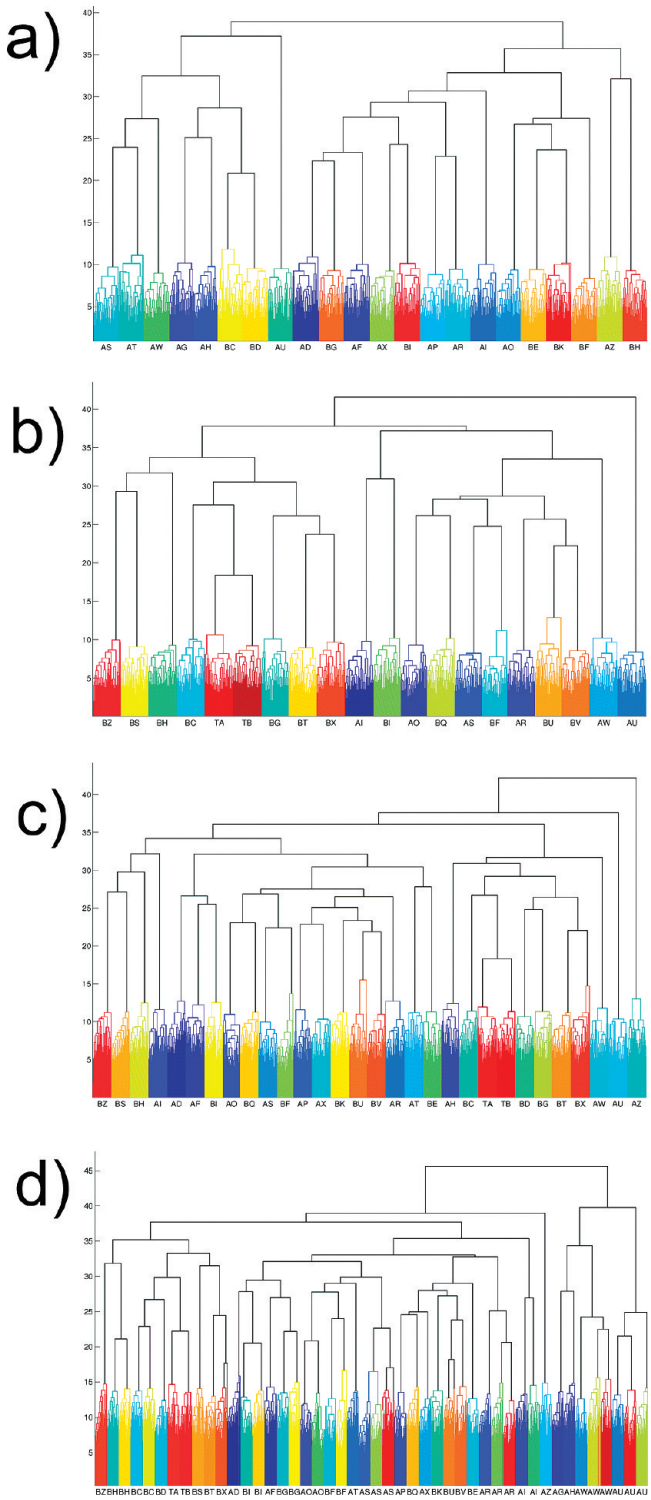


Figure 3. Dendrograms relative to cluster analysis. (a) On the 21-dimensional PCA/CA subspace for the 22 individuals of MetRef1 (2005 collection). (b) On the 19-dimensional PCA/CA subspace for the 20 individuals of MetRef2 (2007 collection). (c) On the 30-dimensional PCA/CA subspace for the 31 different individuals of both MetRef1 and MetRef2. (d) On the 45-dimensional PCA/CA subspace for the 46 different pseudoindividuals of MetRef1, MetRef2 and MetRef3; taking the two/three times donor as different pseudoindividuals. For every figure, clustering according to donor is obvious.

of significantly different metabotypes in humans is much higher than a few tens, and yet to be determined. The larger the number of human metabotypes, the more meaningful will be their perspective use in biology and medicine.

Two- and 3-year Recognition. The spectra of the same individuals in different collections (11 in the first and second collection, 4 in the third) were then used as training and test sets, respectively, and subjected to single vote classifications as described above. The results are reported in Table 2. On average, the percentage of recognition was slightly lowered (but still higher than 97%) for the 11 individuals after 2 years with respect to their scores within each collection, and for the four individuals after 3 years, the lowest values being higher than 92.5%. The fact that all individuals could be almost equally well reassigned to their own metabolic phenotype after 2 to 3 years as they were within each collection period of 3 months demonstrates that the metabolic phenotype is largely constant over a time frame of a few years. This is a very relevant finding in view of the physiological meaning and perspective medical usefulness of metabolic phenotypes information.

Figure 4 shows the color-coded matrix of the distances in the metabolic space among all the individuals involved in at least two collections. The shortest off-diagonal distances are always between samples of the same individuals in different collections. In particular, individual AS shows the shortest distance between collections; that is, it is the most constant subject. This correlates with the slightly lower recognition percentage reported for AS in Table 1d, due to the fact that samples of AS belonging to different collections are occasionally confused with one another.

To better illustrate the implications of these findings, the statistical analysis described above was repeated on the descriptive level by considering each set of about 40 spectra as originating from a different individual (a *pseudoindividual*) even if 15 (11 + 4) of these sets were actually repetitions of collections from the same individuals. A total of 46 pseudoindividuals were thus defined, and subjected to PCA/CA. The results are illustrated in Figure 3d. It is striking to note that pseudoindividuals corresponding to the same individual always cluster together. It should be noted that by this procedure the distance in metabolic space of each individual from itself is even exaggerated, as the present PCA/CA method is built to maximize the distances within any ensemble of predefined groups. Therefore, the similarity of the data sets for the same individual emerges whatever attempt is made to differentiate them, and so confirming the high stability of the metabotypes over a period of years.

Metabolic Jumps and Periodicities. The clear-cut demonstration of the strong persistence of metabolic phenotypes, together with the availability of the present large collection of data sets, hints to the possibility that the sporadic and modest fluctuations of these phenotypes over time frames ranging from days to years can be analyzed in more detail and possibly qualitatively understood. By visual inspection of spectral data sets for each individual, either represented by single, double or triple collection, it is possible to identify various types of changes that, for simplicity, can be described as “spikes”, “waves” and “jumps”. Spikes are signals that are observed in single spectra having markedly different intensity from the preceding and following days: they are most often caused by sporadic assumption of particular food components (like mannitol, contained, for example, in chewing-gum, or TMAO,[11] contained in fish) or unusually high alcohol intake (ethanol signal) or meat (carnosine[12,13] and xanthosine[14]). Other spikes are due to extraordinary physical activity (lactate). Interestingly,

Table 1. The Individual Single Vote Scores for (a) the First Collection, (b) the Second Collection, (c) the 31 Different Donors, and (d) All 46 Pseudodonors

(a) 2005		(b) 2007		(c) 31-donors		(d) 46-pseudodonors	
AD	99.905%	AI	100.000%	AD	99.895%	AD	99.990%
AF	100.000%	AO	100.000%	AF	100.000%	AF	100.000%
AG	100.000%	AR	99.995%	AG	99.997%	AG	100.000%
AH	99.922%	AS	99.941%	AH	99.919%	AH	99.989%
AI	99.760%	AU	100.000%	AI	100.000%	AI	99.775%
AO	97.500%	AW	99.998%	AO	100.000%	AO	97.500%
AP	97.500%	BC	99.705%	AP	97.500%	AP	97.500%
AR	100.000%	BF	99.629%	AR	100.000%	AR	100.000%
AS	100.000%	BG	100.000%	AS	99.992%	AS	99.792%
AT	99.995%	BH	100.000%	AT	100.000%	AT	100.000%
AU	100.000%	BI	98.462%	AU	100.000%	AU	99.379%
AW	100.000%	BQ	99.998%	AW	100.000%	AW	100.000%
AX	100.000%	BS	100.000%	AX	100.000%	AX	100.000%
AZ	100.000%	BT	99.983%	AZ	100.000%	AZ	100.000%
BC	99.998%	BU	96.647%	BC	99.965%	BC	98.630%
BD	100.000%	BV	99.998%	BD	100.000%	BD	100.000%
BE	100.000%	BX	99.998%	BE	100.000%	BE	100.000%
BF	100.000%	BZ	100.000%	BF	99.497%	BF	100.000%
BG	99.933%	TA	98.450%	BG	100.000%	BG	98.838%
BH	100.000%	TB	98.235%	BH	100.000%	BH	98.845%
BI	100.000%	**MEAN =**	**99.552%**	BI	97.846%	BI	99.323%
BK	99.470%			BK	99.193%	BK	99.855%
MEAN =	**99.726%**			BQ	100.000%	AI	99.845%
				BS	100.000%	AO	99.716%
				BT	100.000%	AR	97.649%
				BU	95.750%	AS	86.669%
				BV	100.000%	AU	99.998%
				BX	100.000%	AW	99.848%
				BZ	100.000%	BC	99.988%
				TA	97.668%	BF	97.411%
				TB	99.688%	BG	94.456%
				MEAN =	**99.578%**	BH	99.378%
						BI	99.049%
						BQ	100.000%
						BS	100.000%
						BT	100.000%
						BU	94.982%
						BV	100.000%
						BX	100.000%
						BZ	100.000%
						TA	97.648%
						TB	99.900%
						AR	99.233%
						AS	70.103%
						AU	99.525%
						AW	91.617%
						MEAN =	**98.183%**

"anti" spikes have been also observed, that is, the sporadic disappearance of the signal of a metabolite usually present, as in the case of citrate (individual BF), possibly due to metal complexation.

Waves are signals whose variation in intensity is more gradual and persists for a number of days. Waves are mainly found in the aromatic region of the spectra, a region whose overall profile has been demonstrated to be strongly dependent on the activity of the gut microflora. In several individuals, waves can be observed in the concentration of hippurate (benzoyl glycine) and *m*-hydroxyphenylpropionic acid, the two major gut microbial mammalian co-metabolites,[15,16] together with phenylacetylglycine and formate.[17−19] These variations are likely due to drifts of the endogenous gut microflora population, whose constancy apparently contributes to establish a stable urinary profile. *N*-Methylnicotinamide, a metabolite involved in energy metabolism, also showed a wave (individual BF).

Finally, jumps are sudden and marked changes in signal intensities that, once occurred, persist for the entire collection period and, in some cases, are even carried over to the next collection. Jumps have been observed in the hippurate aromatic signals, thus, indicating a net and sudden (re)modulation in the activity and/or composition of the gut microflora (individual AU). A jump for individual AS is observed in a signal around 8.79 ppm, again in the aromatic region of the urine spectra, dominated by gut microflora-related metabolites. Other jumps have been observed for individual AW at 4.40 ppm (identified as xanthosine, and in line with a higher meat consumption,[14] as declared in the dietary sheet) and for individual AR in the region around 5.52 ppm (possibly due to

Table 2. Single Vote Classification Different Collections As Training and Test Sets[a]

MetRef1 → MetRef2		MetRef1 → MetRef3		MetRef2 → MetRef3	
AI	99.971%	AR	99.522%	AR	100.000%
AO	100.000%	AS	94.439%	AS	99.915%
AR	97.242%	AU	99.944%	AU	97.377%
AS	98.754%	AW	98.746%	AW	92.699%
AU	100.000%	**MEAN =**	**98.163%**	**MEAN =**	**97.498%**
AW	99.712%				
BC	99.434%				
BF	97.132%				
BG	98.645%				
BH	100.000%				
BI	98.723%				
MEAN =	**99.056%**				

[a] These results show how all individuals could be almost equally well reassigned to their own metabolic phenotype after 2 to 3 years.

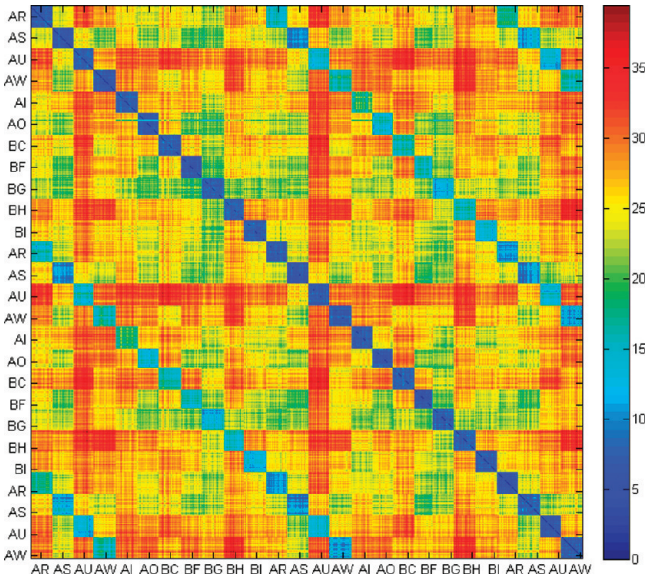


Figure 4. Distances in metabolic space for all the individuals that participated in at least two collections. The pseudoindividuals corresponding to the same individual show lower distances with respect to the others (blue squares).

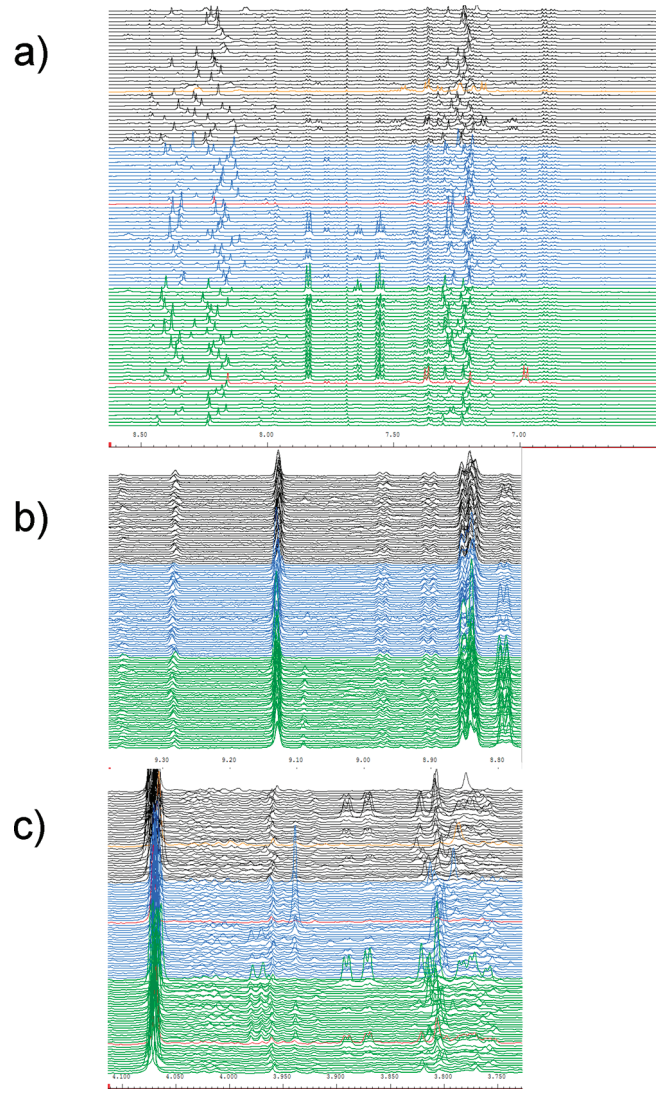


Figure 5. (a) Metabolic "jump" (between 7.90 and 7.55 ppm) for individual AU. (b) Metabolic "wave" (between 9.00 and 8.90 ppm) for individual AS. (c) Metabolic "spike" (at 3.94 ppm) for individual AU.

a drug or to some unidentified food additive, although no particular notes are present in the dietary sheet).

A typical jump is shown in Figure 5a for individual AU, where the spectra of the 2005 collection and the first 23 NMR spectra of the 2007 collection are clearly distinct from the first two-thirds of the 2008 collection, while a wave is shown in Figure 5b and a spike in Figure 5c.

Interestingly, suboptimal recognition rates within the same collection period cannot generally be ascribed to spikes, waves or metabolic jumps.

Jumps and waves are hardly related to significant changes in lifestyle (which would have appeared in the metadata provided by each donor). Excluding assumption of medicines, which was also accurately monitored, it is likely that these jumps are due, directly or indirectly, to changes in the gut microflora, which may occur even in the absence of major changes in diet. It is also interesting to note that jumps can be reversible, as it has been noticed in the case of individual AU during the 2008 collection (Figure 5a).

Despite and beyond all these sources of intraindividuals variations (i.e., sporadic food component intake, persistent or temporary changes in life-style, remodulation of gut microflora populations) interindividual discrimination is still possible and very strong. What defines the individual metabolic phenotype is an invariant part that is very likely to descend both from systematic life and dietary habits and from genetics.

The presence of a non-negligible genomic component is hinted by the results of the study of subjects with common genetic traits to different extents, like father and son and homozygous twins. The study subjects TA and TB are female homozygous twins, and the examination of Figure 3b or 3c shows that on the descriptive level TA and TB are actually very close to one another, although still distinct. In the respective PCA/CA subspaces, all interindividual distances in Figure 3 are much larger than the intraindividual distances, but strikingly, Figure 3b shows that TA and TB are by far the two individuals with the shortest interindividual distance in metabolic space, followed by father and son BU and BV.

The average recognition percentage for the twins sisters shown in Table 1 (98.3%) is on the low side, and also deserves a comment. In contrast to most other individuals, that when wrongly assigned are assigned to a variety of different other

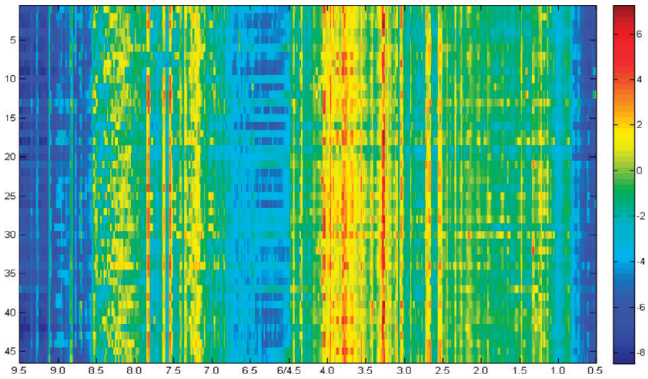


Figure 6. Intrinsic variability of each region of the NMR fingerprint for all analyzed spectra. Within a single individual, the larger variability is in the region between 4.0 and 2.5 ppm, the aliphatic side chains region, and between 8.0 and 7.5 ppm, the aromatic region.

individuals, TA and TB, when wrongly assigned, are mostly assigned to one another (Supplementary Figure S8). In other words, TA and TB are the only case of some confusion between two individual metabolic phenotypes. Undoubtedly, this confusion is of genetic origin, confirming that the genotype has a non-negligible influence on the metabotype. While it is not possible to quantify the extent of such influence from a single observation of this type, it is instructive to compare the distances in the PCA/CA space among all 46 pseudoindividuals by inspecting the complete dendrogram (Figure 3d). It appears again that the distance between TA and TB is one of the shortest, and even shorter than several of the distances between pseudoindividuals corresponding to different collections of the same individual. In other words, the present homozygous twins are metabolically as alike as is an individual with him/herself after 2 to 3 years. All in all, genetics contributes to the interindividual separation but also contributes to the actual (pseudo)individuals clustering, by forcing them to cluster together when they are analyzed after 2 or 3 years. The similarity of one individual with him/herself is maintained even when a significant change in lifestyle, such as modification in dietary habits, quality and kind of food intake due to moving to a different country has occurred, as is the case for individual BC.

It can be concluded that the uniqueness of the metabotype of an individual rests on both environmental factors and the genotype.

Relevant Metabolites. Individual data sets have been used to assess which regions of the spectra are more variable within and between different subjects. Each matrix data set was collapsed into a vector whose components are the standard deviation of each column, that is, the standard deviation of each integrated bucket. This allowed us to highlight which spectral regions are more intrinsically variable. From Figure 6, it appears that the more variable regions *within* a single individual are in the range 8.0−7.5 ppm and 4.0−2.5 ppm.

By means of a pairwise comparison among all the 46 different data sets, 52 metabolites have been found to have different mean concentration values at statistically significant level, and are reported in Supplementary Table 1. Although a systematic analysis of patterns of significantly different metabolites is beyond the scope of this work, it can be noted that some of these metabolites are related to fundamental metabolic pathways, like glycolysis (lactate, pyruvate, acetoacetate), ketonic bodies catabolism (acetoacetate, 3-hydroxyisobutyrate), urea cycle (citrulline, ornithine, arginine), and Krebs cycle (citrate, alpha-ketoglutarate, *trans*-aconitate, succinate). Other metabolites, such as essential amino acids (isoleucine, leucine, valine, lysine, threonine), choline and caproate, are introduced with the diet. Some other metabolites, like hippurate, meta-[hydroxyphenyl]propionic acid (mHPPA),[15,16] phenylacetylglycine (PAG),[17,18] formate,[19] indoxyl-sulfate (IS)[20−22] and valerate[23] are strongly related to gut microflora.[24]

Conclusions

Evidence of the existence of individual metabolic phenotypes had been obtained by means of the analysis of multiple samples collection of urine from a panel of healthy volunteers.[5] In this study, we analyzed the stability of the individual phenotypes with respect to different timescales. Time resulted to play a pivotal role in defining to which extent the individual metabolic phenotype can be considered as a dynamic entity and why. Although the modulation effects of symbiotic gut microbes have been reported in the definition of human phenotypes,[6] timescale analysis of individual phenotypes as derived from NMR spectra of human urines indicates that although the genotype plays a role in defining the individual metabolic phenotype, the latter is modulated by both gut microflora and by response to external stimuli. In this light, individual metabolic phenotypes describe the fact that each individual possesses an unique adaptive response to the environment. The constancy of the phenotype within a time span of years is an important result that was not obvious to predict. It strengthens the idea that an individual is stably associated to a part of its metabolome. This justifies, for instance, that drugs (i) may have different effects on different individuals and/or (ii) may be metabolized differently, allowing one to predict the adverse effects of drugs through metabolomic analysis of urine of patients.[25]

Our analysis on such a relatively long timescale by means of multiple collections allowed to minimize the effects of the day-to-day variability of the individual metabolic fingerprint, providing a dynamical picture of the contributions of genetics and/or gut microflora to these phenotypes, and not only mere snapshots. We have been able to identify fluctuations in gut microflora contributions, described as waves and jumps in metabolites, like hippurate and *m*-HPPA, that are directly related to the activities of symbiotic microorganisms. We have also shown that this kind of modulation is relatively minor with respect to the stable core of the individual metabolic phenotype. By means of hierarchical cluster analysis, we showed that genetically related individuals, like father and son and twins, are closer in the component space than unrelated individuals, and in the case of the twins, they are as close as is the same individual to him/herself between different collections. Confusion between metabolic phenotypes of the two twin sisters is certainly to be ascribed to genetic influence, but may be also due to the fact that, for genetic reasons, identical twins could show a much higher similarity in gut microbiota structures than genetically unrelated individuals. The present work also strongly supports the idea that the individual metabolic phenotype can also be considered a metagenomic entity that links gut microbiome and host metabolic phenotype;[26,27] that is, the composition of the gut microflora is compatible with the genome of the host.

Acknowledgment. This work was supported by grants from Ente Cassa di Risparmio di Firenze to the FiorGen

Foundation and by a fellowship from Boehringer Ingelheim Italia (to L.T. through the FiorGen Foundation). The authors declare no competing financial interest.

Supporting Information Available: Supplementary Table 1, most relevant metabolites with respect to their variability *between* subjects; Supplementary Materials and Methods; testing of the predictivity of the PCA-CA-KNN classification approach; detailed listing of results of predictivity tests; figures of confusion matrices. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Gavaghan, C. L.; Holmes, E.; Lenz, E.; Wilson, I. D.; Nicholson, J. K. An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57BL10J and Alpk:ApfCD mouse. *FEBS Lett.* **2000**, *484*, 169–174.

(2) Rezzi, S.; Ramadan, Z.; Martin, F. P.; Fay, L. B.; van Bladeren, P.; Lindon, J. C.; Nicholson, J. K.; Kochhar, S. Human metabolomic phenotypes link directly to specific dietary preferences in healthy individuals. *J. Proteome Res.* **2007**, *6* (11), 4469–4477.

(3) Kussmann, M.; Raymond, F.; Affolter, M. OMICS-driven biomarker discovery in nutrition and health. *J. Biotechnol.* **2006**, *124*, 758–787.

(4) Clayton, T. A.; Lindon, J. C.; Cloarec, O.; Antti, H.; Charuel, C.; Hanton, G.; Provost, J.; Le Net, J.; Baker, D.; Walley, R. J.; Everett, J. R.; Nicholson, J. K. Pharmaco-metabonomic phenotyping and personalized drug treatment. *Nature* **2006**, *440* (7087), 1073–1077.

(5) Assfalg, M.; Bertini, I.; Colangiuli, D.; Luchinat, C.; Schäfer, H.; Schütz, B.; Spraul, M. Evidence of different metabolic phenotypes in humans. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (5), 1420–1424.

(6) Holmes, E.; Loo, R. L.; Stamler, J.; Bictash, M.; Yap, I. K. S.; Chan, Q.; Ebbels, T.; De Iorio, 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*, 396–400.

(7) Moore Ede, M. C.; Brennan, M. F.; Ball, M. R. Circadian variation of intercompartmental potassium fluxes in man. *J. Appl. Physiol.* **1975**, *38* (l), 163–170.

(8) Moore Ede, M. C.; Herd, J. A. Renal electrolyte circadian rhythms: independence from feeding and activity patterns. *Am. J. Physiol.* **1977**, *232* (2), F128−F135.

(9) Rabinowitz, L.; Berlin, R.; Yamauchi, H. Plasma potassium and diurnal cyclic potassium excretion in the rat. *Am. J. Physiol.* **1987**, *253* (22), Fll78−F1181.

(10) Steele, A.; De Veber, H.; Quaggin, E.; Scheich, A.; Ethier, J.; Halperin, M. L. What is responsible for the diurnal variation in potassium excretion. *Am. J. Physiol.* **1994**, *267* (36), R554−R560.

(11) Bell, J. D.; Lee, J. A.; Lee, H. A.; Sadler, P. J.; Wilkie, D. R.; Woodham, R. H. Nuclear magnetic resonance studies of blood plasma and urine from subjects with chronic renal failure: identification of trimethylamine-N-oxide. *Biochim. Biophys. Acta* **1991**, *1096* (2), 101–107.

(12) Park, Y. J.; Volpe, S. L.; Decker, E. A. Quantitation of carnosine in humans plasma after dietary consumption of beef. *J. Agric. Food. Chem.* **2005**, *53*, 4736–4739.

(13) Bauchart, C.; Savary-Auzeloux, I.; Mirand, P. P.; Thomas, E.; Morzel, M.; Remond, D. Carnosine concentration of ingested meat affects carnosine net release into the portal vein of minipigs. *J. Nutr.* **2007**, *137*, 589–593.

(14) Young, D. S.; Epley, J. A.; Goldman, P. Influence of chemically defined diet on the composition of serum and urine. *Clin. Chem.* **1971**, *17* (8), 765–773.

(15) 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*, 527–537.

(16) Williams, R. E.; Eyton-Jones, H. W.; Farnworth, M. J.; Gallagher, R.; Provan, W. M. Effect of intestinal microflora on the urinary metabolic profile of rats: a (1)H-nuclear magnetic resonance spectroscopy study. *Xenobiotica* **2002**, *32*, 783–794.

(17) Goodwin, B. L.; Ruthven, C. R.; Sandler, M. Gut flora and the origin of some urinary aromatic phenolic compounds. *Biochem. Pharmacol.* **1994**, *47* (12), 2294–2297.

(18) Nicholls, A. W.; Mortishire-Smith, R. J.; Nicholson, J. K. NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem. Res. Toxicol.* **2003**, *16* (11), 1395–1404.

(19) Samuel, B. S.; Gordon, J. I. A humanized gnotobiotic mouse model of host-archaeal bacterialmutualism. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 10011–10016.

(20) Bueschkens, D. H.; Stiles, M. E. Escherichia coli variants for gas and indole production at elevated incubation temperatures. *Appl. Environ. Microbiol.* **1984**, *48*, 601–605.

(21) Gao, X. X.; Ge, H. M.; Zheng, W. F.; Tan, R. X. NMR-based metabonomics for detection of Helicobacter pylori infection in gerbils: which is more descriptive. *Helicobacter* **2008**, *13*, 103–111.

(22) Enomoto, A.; Takeda, M.; Tojo, A.; Sekine, T.; Seok, H. C.; Khamdang, S.; Takayam, F.; Aoyam, I.; Nakamura, S.; Enodu, H.; Niwa, T. Role of organic anion transporters in the tubular transport of indoxyl sulfate and the induction of its nephrotoxicity. *J. Am. Soc. Nephrol.* **2002**, *13*, 1711–1720.

(23) Patel, P.; Nankova, B. P.; La Gamma, E. F. Butyrate, a gut-derived environmental signal, regulates tyrosine hydroxylase gene expression via a novel promoter element. *J. Dev. Brain Res.* **2005**, *160*, 53–62.

(24) Bertini, I.; Calabrò, A.; De Carli, V.; Luchinat, C.; Nepi, S.; Porfirio, B.; Renzi, D.; Saccenti, E.; Tenori, L. The metabonomic signature of celiac disease. *J. Proteome Res.* **2009**, *8* (1), 170–177.

(25) Wederson, M. C.; Quattrone, A.; Biganzoli, L.; Pestrin, M.; Bertini, I.; Di Leo, A. Metabolomics: available results, current research projects in breast cancer, and future applications. *J. Clin. Oncol.* **2007**, *25* (19), 2840–2846.

(26) Li, M.; Wang, B.; Zhang, M.; Rantalainen, M.; Wang, S.; Zhou, H.; Zhang, Y.; Shen, J.; Pang, X.; Zhang, M.; Wei, H.; Chen, Y.; Lu, H.; Zuo, J.; Su, M.; Qiu, Y.; Jia, W.; Xiao, C.; Smith, L. M.; Yang, S.; Holmes, E.; Tang, H.; Zhao, G.; Nicholson, J. K.; Li, L.; Zhao, L. Symbiotic gut microbes modulate human metabolic phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (6), 2117–2122.

(27) Martin, F. P.; Wang, Y.; Sprerngrer, N.; Yap, I. K.; Lundstedt, T.; Lek, P.; Rezzi, S.; Ramadan, Z.; Van Bladeren, P.; Fay, L. B.; Kochhar, S.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Probiotic modulation of symbiotic gut microbial-host metabolic interactions in a humanized microbiome mouse model. *Mol. Syst. Biol.* **2008**, *4* (157), 1–15.

PR900344M

Research article

Title: Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at varying risks of disease relapse by traditional prognostic methods.

C Oakman (1 *), L Tenori (2,3 *), W. M. Claudino (1 *), S. Cappadona (1), S. Nepi (2,3), A. Battaglia (1), P. Bernini (2), E. Zafarana (1), E. Saccenti (2), M. Fornier (4), P.G. Morris (4), L. Biganzoli (1), C. Luchinat (2,5), I. Bertini (2,5), A. Di Leo (1 **)

(1) “Sandro Pitigliani” Medical Oncology Unit, Department of Oncology, Hospital of Prato, Istituto Toscano Tumori, Prato, Italy

(2) Magnetic Resonance Center (CERM), University of Florence, Via L.Sacconi 6, 50019 Sesto Fiorentino, Italy

(3) FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

(4) Breast Cancer Medicine Service, Memorial Sloan-Kettering Cancer Center, New York, USA

(5) Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

(*) These authors have equally contributed to this work

(**) Corresponding author: Angelo Di Leo, “Sandro Pitigliani” Medical Oncology Unit, Hospital of Prato, Istituto Toscano Tumori, Piazza Ospedale 2, 59100, Prato, Italy. Tel +39 0574 434766; fax +39 0574 29798, email address: adileo@usl4.toscana.it

SUMMARY

Background: Prognostic tools in early breast cancer are inadequate. The evolving field of metabolomics may allow more accurate identification of patients with residual micrometastases.

Patients and Methods: Forty-four early breast cancer patients with pre- and post-operative serum samples had metabolomic assessment by Nuclear Magnetic Resonance. Fifty-one metastatic patients served as control. Differential clustering was identified and used to calculate individual early patient 'metabolomic risk', calculated as inverse distance of each early patient from the metastatic cluster baricenter. Metabolomic risk was compared with Adjuvant!online 10-year mortality assessment.

Results: Innate serum metabolomic differences exist between early and advanced patients. Pre-operative patients were identified, with 75% sensitivity, 69% specificity and 72% predictive accuracy. Comparison with Adjuvant!online revealed discordance. Of 21 patients assessed as high risk by Adjuvant!online, 10(48%) and 6(29%) were high risk by metabolomics in pre- and post-operative settings, respectively. Of 23 low risk patients by Adjuvant!online, 11(48%) pre- and 20(87%) post-operative patients were low risk by metabolomics.

Conclusion: This study identifies metabolomic discrimination between early and advanced breast cancer. Micrometastatic disease may account for metabolomic misclassification of some early patients as metastatic. Metabolomics identifies more patients as low relapse risk compared with Adjuvant!online. Further exploration of this metabolomic fingerprint is warranted.

KEYWORDS

Breast Cancer; Metabolomics; Micrometastasis; Prognosis

INTRODUCTION:

In early breast cancer, a critical issue is identification of individuals likely to benefit from adjuvant intervention. Better prediction of clinical outcomes is required to optimize and individualize our therapeutic approach. Current approaches, using traditional clinicopathological features or gene profiling, assess the primary tumor and estimate risk of recurrence based on the presumption of micrometastatic disease. These tools have limitations. Consequently, an individual's risk may be over or underestimated.

The pivotal Milan adjuvant chemotherapy trial compared classical cyclophosphamide, methotrexate and fluorouracil (CMF) with surgery alone in node-positive patients. Thirty-year follow up confirms a sustained benefit from chemotherapy, but also showed long term, relapse-free survival (RFS) in a significant number of high risk patients treated with surgery alone, RFS 29% vs 22% respectively (1). Similarly, 20-year follow up of intravenous CMF compared with surgery alone in node-negative, estrogen receptor (ER) negative disease reveals long term RFS in approximately 40% patients treated with surgery alone (1). The 21-gene OncotypeDx assay was assessed in 355 placebo-treated patients from the NSABP-B14 trial in node-negative, ER positive disease. Ten-year distant-recurrence free survival for these patients treated with surgery alone, was 86%, 62% and 69% for low, intermediate and high recurrence scores respectively (2). The 70-gene Mammaprint applied to 151 lymph node-negative patients, only 10 of whom received any adjuvant therapy, showed differential 10-year distant metastases free survival between good and poor prognosis signatures, at 87% and 44% respectively (3). A striking feature of these studies is that some individuals, despite apparent high risk disease, clearly have excellent long term outcome. This reflects heterogeneity of disease, host and risk, and highlights overestimation of risk by current prognostic tools.

An alternative to presuming residual disease is actual measurement of micrometastases. Studies of micrometastatic disease are intriguing, particularly isolated

tumor cells (ITC) in the bone marrow and circulating tumor cells (CTC) (4-11). Of particular interest is that not all patients with ITC or CTC develop clinically detectable metastatic disease. Possible hypotheses include lack of autonomous proliferation, unfavorable stromal environment and clearance of tumor by host immune mechanisms. Thus tumor survival depends on both favorable tumor and host characteristics. Indeed, assessment of this dynamic, multifactorial interaction is a strength of the evolving field of metabolomics.

Metabolomics is a science that provides a dynamic portrait of metabolic status, assessing end products of gene expression and environmental influences. Metabolomics may be defined as 'the quantitative analysis of all the metabolites of an organism or specified biological sample'(12). A review of this evolving field was previously presented by our group (13)

Transformed human cells exhibit profound metabolic shifts, particularly reflecting induction of cell membrane phospholipids biosynthesis and breakdown, and preferential use of glucose through non oxidative pathways (14-20). Metabolomic analyses of patient serum and urine samples have been shown to delineate between healthy, benign and malignant conditions (21-25). Specifically with breast cancer there is cell line evidence of metabolomic distinction between normal and malignant, and even more specifically identification of malignant breast cell lines with greater metastatic potential. (26-30). With breast tissue, metabolomic analyses distinguish normal tissue, benign disease, carcinoma in situ and invasive carcinoma (31-34). The subsequent challenge is to capture the malignant metabolomic signal amongst the complex serum metabolomic fingerprint for an individual (35).

A metabolomic fingerprint (36) may be determined from a variety of biological samples. Nuclear Magnetic Resonance (NMR)-based metabolomics in its most popular version consists of recording and analyzing one dimensional (1D) proton NMR profiles of serum and urine (37). Key elements, specifically hydrogen nuclei (^{1}H), are influenced by

an external magnetic field. This magnetic influence is measurable and recorded as an ^{1}H-NMR spectrum. ^{1}H-NMR spectra of biological fluids are extremely complex, essentially the superposition of the ^{1}H-NMR spectra of a vast number of single metabolites, plus a number of food components and occasionally drug or drug-derived molecules, present in a sample at > 1 μM concentration (38). Huge volumes of qualitative and quantitative data are obtained by primary analysis and chemometric methods then transform this to manageable and comparable data.

At variance with knowledge-guided search of metabolites in pathological samples, NMR-based metabolomics makes no assumptions about the identity of the metabolites that are relevant for the selected pathology. Information on the metabolite pattern alterations that can be significantly associated to the pathology is directly obtained through statistical analysis of the NMR profiles (39).

A metabolomic fingerprint may exist for micrometastatic disease. More specifically a fingerprint may exist which identifies the interaction between host and any residual disease. We have performed and compared metabolomic analyses in breast cancer patients with early and metastatic disease. Prognostic ability of the fingerprint has been explored by comparison with 10-year mortality rates determined by the current prognostic tool Adjuvant!online. The pilot model, developed in 44 early breast cancer patients, was then validated in a second cohort of 45 early breast cancer patients.

PATIENTS and METHODS

Patient selection

This series was recruited at our institution, The Department of Oncology, Hospital of Prato, Italy, between March 2006 and November 2008. Approval was granted by the Hospital of Prato Human Investigations Ethics Committee. All patients provided written, informed consent. All patients were older than 18 years old with histological confirmation of

breast cancer. Eligible early breast cancer patients had an Eastern Cooperative Oncology Group (ECOG) performance status between 0 and 2, no detectable macro-metastatic disease and no prior anti-cancer treatment. Eligible metastatic patients could have received chemotherapy and/or endocrine therapy prior to study entry.

Samples were collected from 3 patient categories: (1) 44 patients with early disease with pre-operative and post-operative samples; (2) 51 patients with metastatic disease, as the control group; (3) 45 patients with early disease provided only a post-operative sample, as the validation series.

Sample collection and analysis

All patients recorded their dietary intake and use of any medication on the day before sample collection. Ten mL of venous blood were collected after an overnight fast in serum-separator tubes with gel. The tubes were centrifuged for 10 minutes at 1500xg at room temperature. The serum was collected and transferred in 1 ml aliquots to pre-labelled 2ml cryovials (Nalgene, Rochester, USA). Three aliquots per patient were immediately frozen at -80°C until used. For the early patients, mean time elapsing between pre-operative serum collection and surgery was 16 days (range: 2 – 41 days) whilst for post-operative patients, the mean time between surgery and post-operative serum collection was 33 days (range 16 – 55). No patients received adjuvant systemic therapy prior to blood collection. For the metasatic patients, 22 had received prior chemotherapy for advanced disease. Mean time elapsing between last chemotherapy administration and sample collection was 93 days (range: 5 - 484 days).

Frozen serum samples were thawed at room temperature and shaken before use. A phosphate buffer (70 mM, pH 7.4) was added in ratio 1:1 before analysis. One-dimensional ^{1}H-NMR spectra of serum samples were measured on a spectrometer operating at 600 MHz proton Larmor frequency using standardized protocols (NOESY1D and CPMG sequences).

Data reduction was obtained by means of Orthogonal Projection to Latent Structure (OPLS) (40). For the purpose of classification we used the Support Vector Machines (SVM) method (41) applied on the OPLS scores. To assess the prediction ability of the model, a double cross-validation scheme was adopted (42). Accuracy, sensitivity and specificity were estimated using standard definitions (43). All calculations were made using R (44). Details of analytical and statistical methods applied are given in Supplemental Material.

'Metabolomic Risk'

OPLS provided differential clusterization of early and metastatic patients. Two dimensional cluster plots gave a simple, visual representation of what is a complex multidimensional comparison. Each patient spectrum became one dot on the cluster plot. The position of each patient was determined by multispatial analysis of metabolites. Initially both NOESY1D and CPMG revealed similar results and subsequently only CPMG plots were used.

We hypothesized that early patients with micrometastatic disease would be closer to the metastatic patients in their clustering position; conversely patients with fingerprints most distant from the metastatic patients would not have micrometastatic disease. 'Metabolomic risk' for each early breast cancer patient was measured as their inverse Euclidean distance from the baricenter of the metastatic population. The distance for each patient was inversely proportional to their risk of micrometastatic disease. Thus, a patient with a short distance from the metastatic baricenter has a high metabolomic risk; a patient at a long distance has a low metabolomic risk.

Metabolomic risk showed concordance with accuracy of disease stage prediction, i.e. early patients with low metabolomic risk were correctly classified as having early disease, whilst early patients with high metabolomic risk were misclassified as metastatic. This validated subsequent use of metabolomic risk as a surrogate for relapse.

Comparison with Adjuvant!online

Comparison of metabolomic risk with Adjuvant!online (Standard Version 8.0) was undertaken (45). Adjuvant!online is a web-based program that provides prognostic and predictive data based upon a patient's clinicopathologic profile. It provides estimates of 10-year risk of disease relapse and death (46). Although Adjuvant!online has limitations, particularly lack of the human epidermal growth factor receptor 2 (Her-2) and ki-67 status, poor characterization of small tumors and focus on ductal histopathology, it is a widely used, validated tool for risk assessment. To assess concordance, pre-operative metabolomic risk was plotted against 10-year risk of cancer mortality (in absence of adjuvant therapy) as calculated by Adjuvant!online.

Using arbitrary thresholds of a 10% 10-year mortality by Adjuvant!online and the median metabolomic risk from the pre-operative samples, patients were categorized as high or low risk for each, with four possible resulting combinations: low_A/low_M (i.e. low risk by Adjuvant!online and metabolomics), $low_A/high_M$, $high_A/low_M$ or $high_A/high_M$.

Metabolomic risk shift

The best control for each post-operative specimen was its corresponding pre-operative specimen. Using each patient as their own control had the distinct advantage of eliminating the confounder of inter-individual variation. This analysis was to explore the effect of surgical removal of the primary tumor on the metabolomic fingerprint. Each patient had their shift in metabolomic risk assessed, noting particularly whether their risk category, as described above, altered.

Validation series

45 post operative patients had serum analysis by ^{1}H-NMR and calculation of metabolomic risk, using the same metastatic control group as described. 10-year mortality by Adjuvant!online was compared with metabolomic risk, using the same Adjuvant!online and metabolomic thresholds as defined above, to validate the findings of the pilot group.

RESULTS

Patient characteristics

Patient characteristics are displayed in Table 1. Most patients were over 50 years of age, with ductal histology, ER-positive and Her-2 negative tumors. Review of dietary and medication history did not highlight any remarkable differences.

Metabolomic fingerprints: Early patients vs. metastatic patients

Metabolomic fingerprints of pre-operative patients and metastatic patients showed significant differential clustering, with near complete separation of the two groups. The clustering was provided by both the CPMG and NOESY1D spectra (Figure 1a). Using the 2:8 ratio exclusion (detailed in supplementary material) double cross validation scheme, we were able to determine the percentage of times an individual sample was correctly classified as metastatic or pre-operative. Most individuals were reliably identified most of the time. Some patients however were consistently misclassified, that is, based on their metabolomic fingerprint, some early patients are consistently misclassified as metastatic, and vice versa. These analyses revealed a discrimination sensitivity of 75%, a specificity of 69% and a predictive accuracy of 72% using CPMG. Similar figures were obtained using NOESY1D, specifically sensitivity of 77%, a specificity of 68% and a predictive accuracy of 73%.

Metabolomic risk was calculated for each pre-operative patient. This risk was plotted against disease stage prediction accuracy revealing reasonable concordance. That is, pre-operative patients with a high metabolomic risk were reliably misclassified as metastatic patients (Figure 1b).

Correlation between metabolomic risk and 10-year breast cancer mortality estimates (in absence of systemic therapy) by Adjuvant!online

10-year mortality risk as calculated by Adjuvant!online and pre-operative retabolomic risk show poor concordance (Figure 2). Using the arbitrary thresholds previously described, patients were divided into 4 risk categories using Adjuvant!online 10% mortality risk and the median pre-operative metabolomic risk of 49.86. Using these 4 risk categories it is seen that in the pre-operative setting, of 21 patients assessed as high risk by Adjuvant!online, 10 are also high risk by metabolomics (48%). Of 23 patients assessed as low risk by Adjuvant!online, 11 are also low risk by metabolomics (48%). This pre-operative metabolomic fingerprint is taken in the presence of the primary tumor, with or without lymph node involvement and/or micrometastases. There was no pattern seen between likelihood of concordance with tumor size, grade, lymph node involvement, age, menopausal status, diet or medications (data not shown).

Impact of tumor removal on the correlation between Metabolomics and Adjuvant!online in defining risk

Within the 4 risk categories, the impact of primary tumor removal is presented in Figure 3. The main observation is that in 19 of 22 patients with high pre-operative metabolomic risk, surgical removal of the primary tumor reduced the metabolomic risk (86%). As shown, all 11 patients assessed as low_A/low_M pre-operatively remain low_A/low_M (100%), while 11 of 12 patients assessed as $low_A/high_M$ had a metabolomic risk reduction (92%). Similarly 8 of 10 patients assessed as $high_A/high_M$ reveal a reduction in their metabolomic risk post surgery (80%). Four of 11 patients from the $high_A/low_M$ group curiously experienced a sizeable increase in metabolomic risk post surgery (92%).

Post surgery, of 21 patients assessed as high risk by Adjuvant!online, only 6 (29%) are also assessed as high risk by metabolomics. Conversely of 23 patients assessed as low risk by Adjuvant!online, concordance markedly increased with 20 (87%) also low risk by metabolomics. Discordance persists for 15 cases with low metabolomic risk and high Adjuvant!online risk, and 3 cases with high metabolomic risk and low Adjuvant!online risk.

Validation of metabolomic risk in an independent cohort of early breast cancer patients

Comparison of metabolomic risk of 45 post-operative patients with 10-year mortality by Adjuvant!online confirmed the discrepancy seen in the pilot group.(See Figure 4). Of 14 patients assessed as low risk by Adjuvant!online, 11 (79%) are also low risk by metabolomics. Of 31 patients assessed as high risk by Adjuvant!online, only 10 (32%) are also high risk by metabolomics.

Metabolomic differences between early and metastatic patients

Differential metabolites between metastatic and early patients were identified using a Wilcoxon test with Bonferroni correction applied to a nominal P value of 0.05. Metastatic subjects were characterized by higher values of phenylalanine, glucose, proline, lysine and N-acetyl cysteine, and lower values of lipids, when compared to the spectra of both post-operative and pre-operative patients.

DISCUSSION

The appeal of metabolomics is concurrent assessment of tumor and host. Indeed survival of a specific tumor in a specific host relies on a dynamic interaction, with evasion of normal host immunity and favorable stromal environment for metastatic deposits as key factors. A strength of metabolomics, as compared to current prognostic tools, may be confirmation rather than assumption of micrometastatic disease.

Our results reveal differential metabolomic fingerprints for most early and metastatic breast cancer patients. Amongst the normal noise of the metabolomic fingerprint, most patients were distinguished based on metabolomic analysis of one serum sample.

In our study, metabolomic analysis assigns more patients to low risk than are assigned by Adjuvant!online. Similarly, when compared with conventional clinical and pathological factors, prognostic gene expression signatures generally identify more

patients of low risk. The 21-gene OncotypeDx shows direct concordance of 36% in relapse risk stratification compared with an adjusted Adjuvant!online (47). The 70-gene Mammaprint, when compared with Adjuvant!online, had stronger predictive power and provided lower risk estimates for more patients (48). These low risk patients may be spared or receive less intensive adjuvant treatment.

A small number of patients (N=4) with low metabolomic risk pre-operatively actually became high metabolomic risk post-operatively. One potential explanation for this observed increase, despite seeming counterintuitive, is that surgical removal of the primary tumor may increase risk of future disease. There is data suggesting this may occur (49-51). Analysis of post-operative serum and wound drainage fluid reveal growth factors that induce proliferation (51). Acute surgical stress induced immunosuppression and abundance of growth stimulating factors may trigger tumor growth (52,53). Of interest, all 4 patients were assessed as high risk by Adjuvant!online

The great challenge of metabolomics is innate variability of tumor and host. Variation exists between people for age, gender, race and hormonal status. For any one person, the profile may be affected by diet, diurnal rhythm, gastointestinal microflora, medication, toxins, disease, stress and exercise (35). The critical issue is whether the malignant metabolomic signature is identifiable among the vast array and variation of metabolites from blood or urine.

A potential mechanism to overcome the challenge of variability is collection of more than one sample for each patient (36). Multiple samples may reveal an invariant signal which is characteristic for each individual, with less variation or 'noise'. Other options are to examine tissue or perform a directed metabolite search, rather than examine systemic profiles from serum or urine. Certainly metabolomic analysis of tissue has been shown to distinguish normal, benign and malignant processes in several tissue types, including breast (16-18,31-34). Also targeted metabolite searches are reported, such as targeted

urine analysis for identification of a specific metabolite linked with prostate cancer with a modest predictive value for progression (18). These latter processes limit variability at the expense of eliminating analysis of the host.

Our results require validation in larger, independent cohorts to confirm their reproducibility. The discordance between Adjuvant!online and metabolomics must be explored by application of Adjuvant!online and metabolomics to a cohort with at least 10-year follow up data, to examine which is a more accurate tool. From this pilot study we will move forward to validation with an external cohort of patients from the Memorial Sloan-Kettering Cancer Center with long term follow up. Although our validation series confirmed the trend seen in the pilot group, our patients lack follow-up.

In conclusion, we believe that the benefit of metabolomics is the incorporation of both a specific tumor profile with metastatic features and a specific host profile conducive to tumor growth. We have presented a preliminary exploration in a limited number of patients of a potential role for the evolving field of metabolomics in assessment of micrometastatic disease in early breast cancer. Clearly this approach requires refinement and validation, but the distinction identified between early and late disease and the prognostic role of the metabolomic fingerprint provide an exciting platform for further work.

Funding

This work was supported by Associazione Italiana Ricerca Cancro, Milan, Italy; Breast Cancer Research Foundation, New York, USA; Associazione “Sandro Pitigliani”, Prato, Italy; Ente Cassa di Risparmio di Firenze to the FiorGen Foundation; Boehringer Ingelheim Italia [to L.T. through the FiorGen Foundation] and SFMET European Project [contract no.201640].

Acknowledgements

The authors wish to acknowledge the contributions of Associazione Italiana Ricerca Cancro, Milan, Italy; Breast Cancer Research Foundation, New York, USA; and Associazione “Sandro Pitigliani”, Prato, Italy. This work was also supported by grants from Ente Cassa di Risparmio di Firenze to the FiorGen Foundation, by a fellowship from Boehringer Ingelheim Italia (to L.T. through the FiorGen Foundation) and by the SFMET European Project (contract no.201640). We thank Hartmut Schaefer from Bruker BioSpin for helpful discussions.

Conflict of Interest

The authors have declared no conflict of interest

References

(1) Bonadonna G, Moliterni A, Zambetti M, et al: 30 years' follow up of randomised studies of adjuvant CMF in operable breast cancer: cohort study. BMJ 2005; 330: 217.

(2) Paik S, Shak S, Tang G, et al: Expression of the 21 genes in the Recurrence Score assay and tamoxifen clinical benefit in the NSABP study B-14 of node negative, estrogen receptor positive breast cancer J Clin Oncol 2005; 23: s510 [abstract].

(3) van de Vijver MJ, He YD, van't Veer LJ, et al: A gene expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; 347: 1999-2009.

(4) Braun S, Naume B. Circulating and Disseminated Tumor Cells. J Clin Oncol 2005; 23:1623-1626.

(5) Janni W, Rack B, Schindlbeck C, et al: The persistence of isolated tumor cells in bone marrow from patients with breast carcinoma predicts an increased risk for recurrence. Cancer 2005; 103: 884-891.

(6) Braun S, Vogl FD, Naume B, et al: A pooled analysis of bone marrow micrometastasis in breast cancer. N Engl J Med 2005; 353: 793-802.

(7) Cristofanilli M, Budd GT, Ellis MJ, et al: Circulating tumor cells, disease progression, and survival in metastatic breast cancer. N Engl J Med 2004; 351: 781-791.

(8) Hayes DF, Cristofanilli M, Budd GT, et al: Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival. Clin Cancer Res 2006 ; 12: 4218-4224.

(9) Stathopoulou A, Vlachonikolis I, Mavroudis D, et al. Molecular detection of cytokeratin-19- positive cells in the peripheral blood of patients with operable breast cancer: evaluation of their prognostic significance. J Clin Oncol 2002; 20: 3404-3412.

(10) Xenidis N, Vlachonikolis I, Mavroudis D, et al: Peripheral blood circulating cytokeratin-19 mRNA-positive cells after the completion of adjuvant chemotherapy in patients with operable breast cancer. Ann Oncol 2003; 14: 849-855.

(11) Harris L, Fritsche H, Mennel R, et al. American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. J Clin Oncol 2007; 25: 5287–5312.

(12) Nicholson J. Global systems biology, personalized medicine and molecular epidemiology. Mol Syst Biol 2006; 2: 52.

(13) Claudino W, Quattrone A, Biganzoli L, et al. Metabolomics: Available Results, Current Research Projects in Breast Cancer, and Future Applications. J. Clin. Oncol 2007; 19: 2840-2846.

(14) Boros LG, Cascante M, Lee W-NP. Metabolic profiling of cell growth and death in cancer: applications in drug discovery. Drug Discovery Today, 2002; 7: 364-372.

(15) Griffin J, Shockcor J. Metabolic Profiles of Cancer Cells. Nature Reviews/Cancer 2004; 4: 551 – 561.

(16) Cheng LL, Lean CL, Bogdanova A, et al. Enhanced resolution of proton NMR spectra of malignant lymph nodes using magic angle spinning. Magn Reson.Med 1996; 36: 653-658.

(17) Cheng LL, Burns MA, Taylor JL, et al: Metabolic characterization of human prostate cancer with tissue magnetic resonance spectroscopy. Cancer Res 2005 ; 65: 3030-3034.

(18) Sreekumar A, Poisson L, Rajendiran T, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 2009 ; 457 : 910-914.

(19) Denkert C, Budczies J, Weichert W, et al. Metabolite profiling of human colon carcinoma--deregulation of TCA cycle and amino acid turnover. Mol Cancer 2008; 7: 72.

(20) Ippolito J, Xu J, Jain S, et al. An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. Proc Natl Acad Sci U S A 2005; 102: 9901-9906.

(21) Fossel ET, Carr JM, McDonagh J. Detection of malignant tumors. Water-suppressed proton nuclear magnetic resonance spectroscopy of plasma. N Engl J Med 1986; 315: 1369-1376.

(22) Odunsi K, Wollman RM, Ambrosome CB et al. Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. Int J Cancer 2005; 113: 782-788.

(23) Yang J, Xu G, Zheng Y, et al: Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. J Chromatogr B Analyt Technol Biomed Life Sci 2004; 813: 59-65.

(24) Bathen TF, Engan T, Krane J, Axelson D. Analysis and classification of proton NMR spectra of lipoprotein fractions from healthy volunteers and patients with cancer or CHD. Anticancer Res 2000; 20: 2393-2408.

(25) Kind T, Tolstikov V, Fiehn O, Weiss RH. A comprehensive urinary metabolomic approach for identifying kidney cancer Anal Biochem 2007; 363: 185-195.

(26) Aboagye E, Bhujwalla Z. Malignant Transformation Alters Membrane Choline Metabolims of Human Mammary Epithelial Cells. Cancer Research 1999; 59: 80-84.

(27) Whitehead T, Monzavi-Karbassi B, Jousheghany F, et al. ^{1}H-NMR metabolic markers of malignancy correlate with spontaneous metastases in a murine mammary tumor model. Int J Oncol 2005; 27: 257-263.

(28) Yang C, Richardson A, Smith J, Osterman A: Comparative Metabolomics of Breast Cancer. Pac Symp Biocomput 2007; 181-192.

(29) Belouche-Babari M, Jackson L, Al-Saffar, et al. Identification of magnetic resonance detectable metabolic changes associated with inhibition of phosphoinositide 3-kinase signalling in human breast cancer cells. Mol Cancer Ther 2006; 5:187-196.

(30) Katz-Brul R, Seger D, Rivenson-Segal D et al. Metabolic markers of breast cancer: enhanced choline metabolism and reduced choline-ether-phospholipid synthesis. Cancer Res 2002; 62: 1966-1970.

(31) Sitter B, Lundgren S, Bathen T, et al. Comparison of HR MAS MR spectroscopic profiles of breast cancer tissue with clinical parameters. NMR Biomed 2006; 19: 30-40.

(32) Cheng LL, Chang IW, Smith BL et al. Evaluating human breast ductal carcinomas with high-resolution magic-angle spinning proton magnetic resonance spectroscopy. J Magn Reson 1998; 135:194-202.

(33) Mackinnon WB, Barry PA, Malycha PL, et al. Fine-needle biopsy specimens of benign breast lesions distinguished from invasive cancer ex vivo with proton MR spectroscopy. Radiology 1997; 204: 661-666.

(34) Mountford CE, Somorjai RL, Malycha P et al. Diagnosis and prognosis of breast cancer by magnetic resonance spectroscopy of fine needle aspirates analysed using a statistical classification strategy. Br J Surg 2001; 88: 1234-1240.

(35) Bollard ME, Stanley EG, Lindon JC, et al. NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. NMR Biomed 2005; 18: 143-162.

(36) Assfalg M, Bertini I, Colangiuli D, et al. Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U S A 2008; 105: 1420-1424.

(37) Griffin JL. Metabonomics: NMR spectroscopy and pattern recognition analysis of body fluids and tissues for characterisation of xenobiotic toxicity and disease diagnosis. Curr Opin Chem Biol 2003; 7: 648-654.

(38) Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res 2007; 35: D521-D526.

(39) Bertini I, Calabrò A, De Carli V, et al. The metabonomic signature of celiac disease. J. Proteome Res 2009; 8: 170-177.

(40) Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J. Chemometrics 2002; 16: 119-128.

(41) Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

(42) Westerhuis J, Hoefsloot H, Suzanne Smit S, et al. Assessment of PLSDA cross validation. Metabolomics 2008; 4: 81-89.

(43) Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994; 308: 1552.

(44) Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Stat Graph 1996; 5: 299-314.

(45) Ravdin PM, Siminoff LA, Davis GJ, et al: Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001; 19: 980–991.

(46) Olivotto I, Bajdik C, Ravdin P, et al: Population-Based Validation of the Prognostic Model ADJUVANT! for Early Breast Cancer. J Clin Oncol 2005; 23: 2716-2725.

(47) Goldstein L, Gray R, Badve S, et al: Prognostic Utility of the 21-Gene Assay in Hormone Receptor–Positive Operable Breast Cancer Compared With Classical Clinicopathologic Features. J Clin Oncol 2008; 26: 4063-4071.

(48) Buyse M, Loi S, van't Veer L, et al: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006; 98: 1183-1192.

(49) Baker D, Masterson T, Pace R, et al. The influence of the surgical wound on local tumor recurrence. Surgery 1989; 106: 525-532.

(50) Hofer S, Shrayer D, Reichner J, et al. Wound-induced tumor progression: a probable role in recurrence after tumor resection. Arch Surg 1998; 133: 383-389.

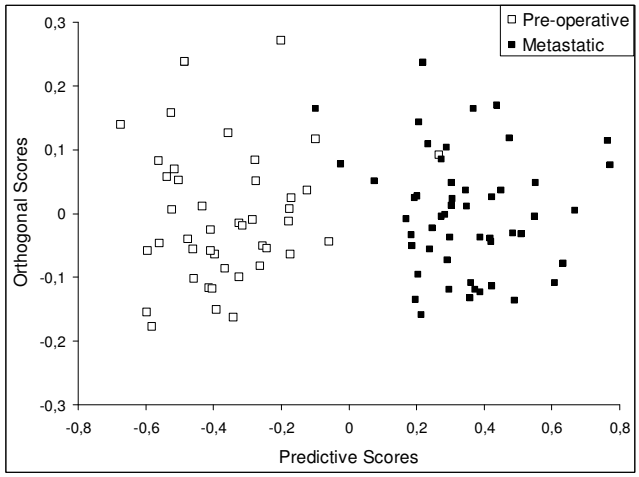
(51) Taglibue E, Agresti R, Carcangui M, et al. Role of HER2 in wound-induced breast carcinoma proliferation. The Lancet 2003; 362: 527-533.

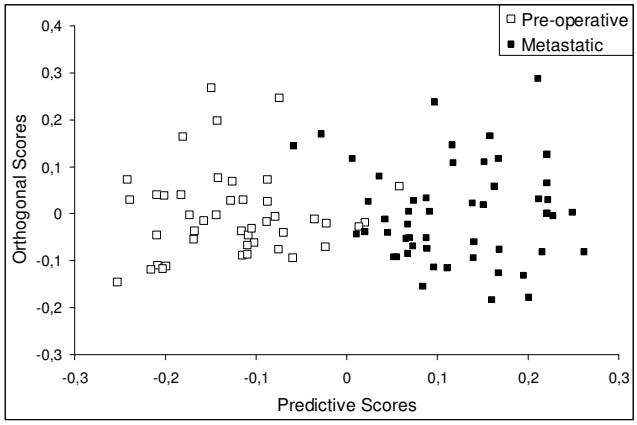
(52) Cole WH. The increase in immunosuppression and its role in the development of malignant lesions. J Surg Oncol 1985; 30: 139-144.

(53) Fisher B, Gunduz N, Coyle J, et al. Presence of a growth-stimulating factor in serum following primary tumor removal in mice. Cancer Res 1989; 49: 1996-2001.

FIGURES

(A)





(B)

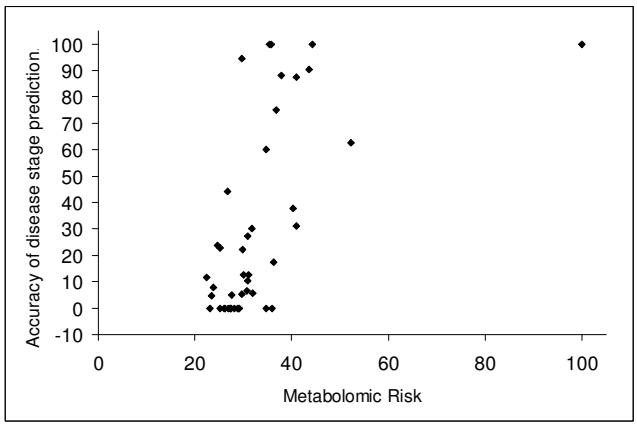


FIGURE 1

(A). Orthogonal Projection to Latent Structure (OPLS) of pre-operative and metastatic patients showing near complete separation of patient groups; CPMG (above) and NOESY1D (below) techniques

(B). Concordance between accuracy of classification of disease stage (y-axis, %) and metabolomic risk (x-axis, MR) in pre-operative patients

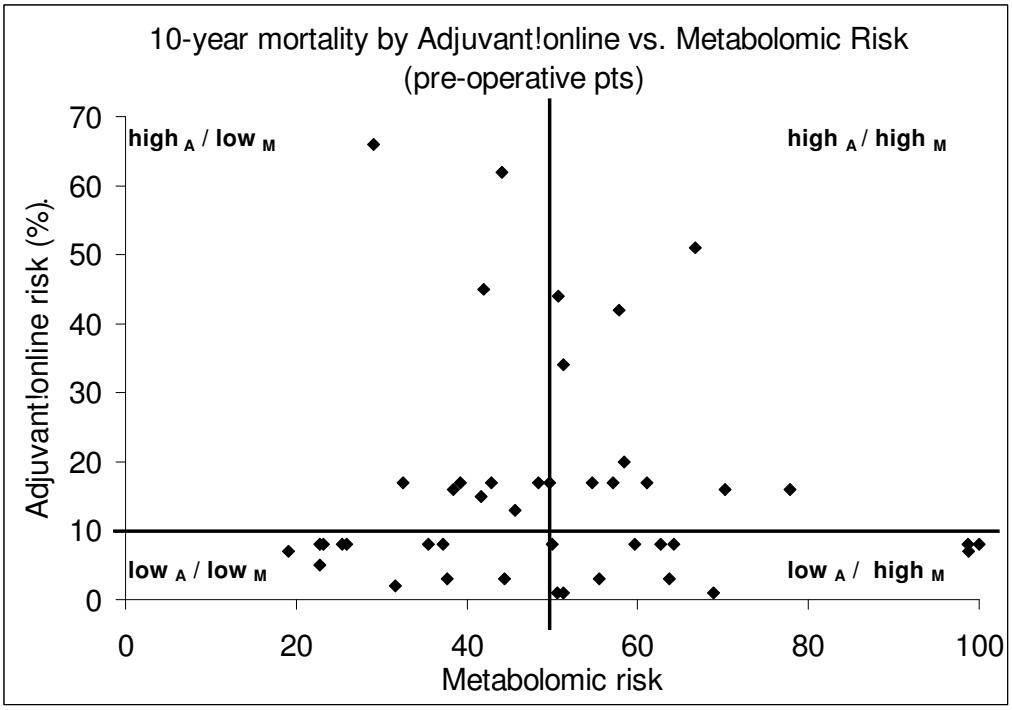


FIGURE 2.

Correlation between 10-year breast cancer mortality estimates by Adjuvant!online (A) and metabolomic risk (M) for pre-operative patients (pts).

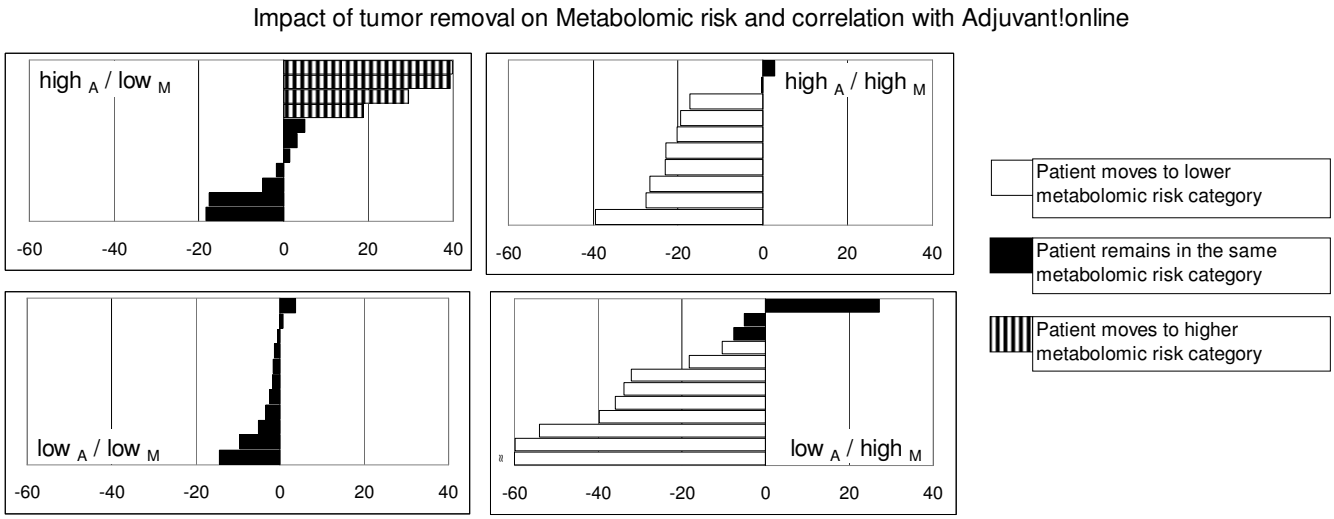


FIGURE 3

Impact of surgical removal of primary tumor on metabolomic risk. Each bar represents an early breast cancer patient. Bar length indicates absolute change in metabolomic risk between the pre- and post-operative samples. Bar pattern indicates effect on risk category. Most patients remain low risk or become low risk by metabolomics.

A: Adjuvant!online; M: metabolomic risk

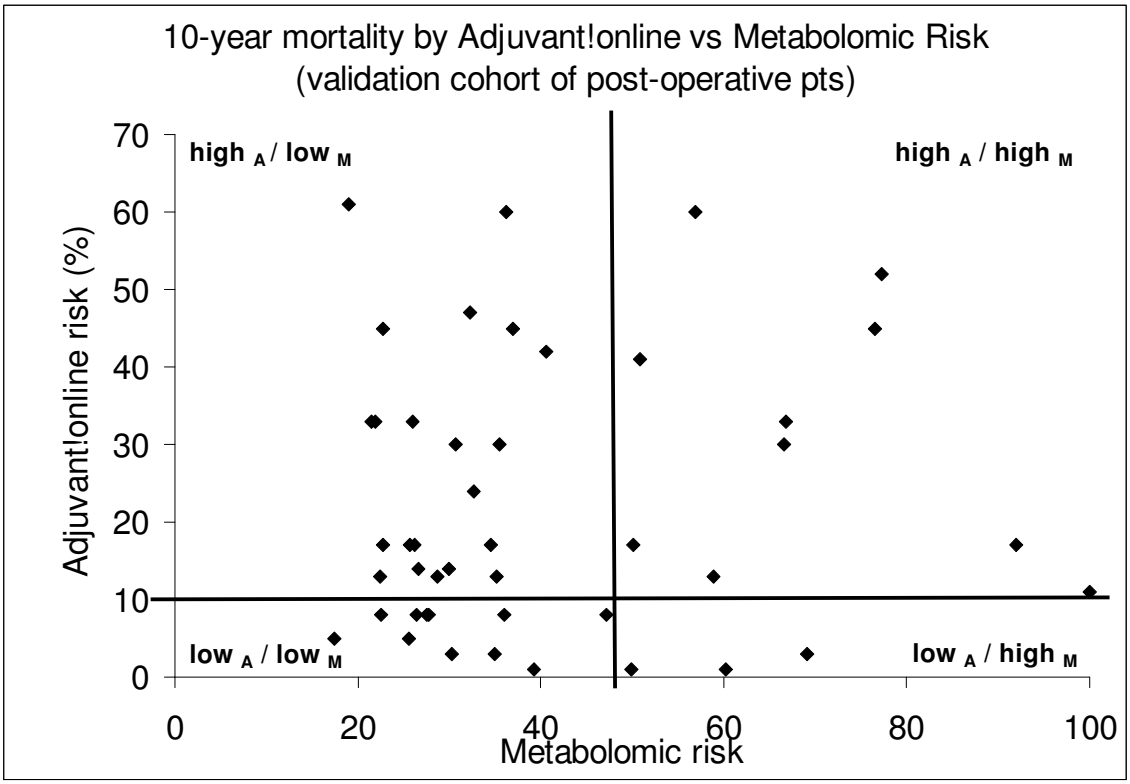


FIGURE 4.

Correlation between 10-year breast cancer mortality estimates by Adjuvant!online and metabolomic risk for independent cohort of 45 post-operative patients. The discordance seen in the pilot group between high risk patients by Adjuvant!online and low risk by metabolomics is also seen in this cohort.

TABLE 1 **PATIENT CHARACTERISTICS**	**EARLY - Pilot** **(pre and post operative)** **N=44 (%)**	**METASTATIC - Control** **N=51(%)**	**EARLY - Validation** **(post operative)** **N=45 (%)**
Age (years)			
Mean	58	59	54
Range	34 – 76	38 – 86	30-85
<40	1 (2)	3 (5.5)	4 (9)
40-50	10 (23)	10 (20)	15 (33)
≥50	33 (75)	38 (74.5)	26 (58)
Post menopausal	35 (80)	41 (80)	26 (58)
Pre menopausal	9 (20)	10 (20)	19 (42)
Size (mm)			
Mean	15		17
Range	1.6 – 45		3.5 - 45
≤ 20	33 (75)		35 (78)
> 20	11 (25)		10 (22)
Grade			
1	4 (9)		3 (7)
2	31 (71)		19 (42)
3	9 (20)		23 (51)
Estrogen receptor status			
Positive	37 (84)	39 (76)	35 (78)
Negative	7 (16)	12 (24)	10 (22)
Progesterone receptor status			
Positive	29 (66)	27 (53)	13 (29)
Negative	15 (34)	23 (45)	32 (71)
n/a		1 (2)	
Her2			
Positive	4 (9)	15 (29)	9 (20)
Negative	40 (91)	32 (63)	36 (80)
n/a		4 (8)	
Histology:			
Ductal	42 (95)	44 (86)	40 (89)
Lobular	2 (5)	3 (6)	5 (11)
Other		2 (4)	
n/a		2 (4)	
No. positive lymph nodes:			
0	28 (64)		24 (53)
1-3	13 (29)		14 (31)
≥ 4	3 (7)		7 (16)
ECOG performance status		Range 0-3	
0		25 (49)	
1		17 (33)	
2		7 (14)	
3		2 (4)	
Visceral involvement		31 (61)	
No visceral involvement		20 (39)	
Prior adjuvant chemotherapy		30 (59)	
Prior adjuvant endocrine therapy		25 (49)	
Prior chemotherapy for advanced disease			
No. lines of treatment:			
0		29 (57)	
1		8 (15.5)	
2		8 (15.5)	
≥ 3		6 (12)	
Prior endocrine therapy for advanced disease			
No. of lines of treatment			
0		26 (51)	
1		10 (20)	
≥2		15 (29)	

TABLE 1. Summary of patient characteristics and clinical data; n/a = not available

SUPPLEMENTAL MATERIAL AND METHODS

NMR Sample preparation and spectra acquisition

Frozen serum samples were thawed at room temperature and shaken before use. 300 µl of a phosphate sodium buffer (70 mM Na_2HPO_4; 20% v/v ^2H_2O; 0.025% v/v NaN_3; 0.8% w/v sodium trimethylsilyl [2,2,3,3-$^2H_4$]propionate (TSP); pH 7.4) were added to 300 µl of each serum sample and the mixture was homogenized by vortexing for 30 seconds (s). 450 µL of this mixture were transferred into a 4.25 mm NMR tube (Bruker BioSpin srl) for analysis.

^{1}H-NMR spectra for all samples were acquired using a Bruker 600 MHz metabolic profiler (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5mm CPTCI ^{1}H-^{13}C/^{31}P-^{2}H cryoprobe including a z axis gradient coil, an automatic tuning-matching (ATM) and an automatic sample changer. A BTO 2000 thermocouple served for temperature stabilization at the level of approximately 0.01 K at the sample. Before measurement, samples were kept for at least 3 minutes inside the NMR probehead, for temperature equilibration (300 K).

For each sample, two one-dimensional NMR spectra were acquired with water peak suppression: (i) a standard pulse sequence (NOESY1Dpresat; Bruker), using 64 free induction decays, 64k data points, a spectral width of 20.0306 ppm, an acquisition time of 2.7 s, a relaxation delay of 4 s, and a mixing time of 100 ms; (ii) a Carr-Purcell-Meiboom-Gill (CPMG; Bruker) spin-echo sequence to suppress signals arising from high molecular weight molecules and a standard pulse sequence (same acquisition parameters as with NOESY1Dpresat)

Spectral Processing.

Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were manually corrected for phase and baseline distortions and calibrated (TMSP peak at 0.00

ppm) using TopSpin (Version 2.1, Bruker). Each 1D spectrum in the range between 0.02 and 10.00 ppm was segmented into 0.02-ppm chemical shift bins and the corresponding spectral areas were integrated using AMIX software (version 3.8.8, Bruker BioSpin). Regions between 6.0 and 4.5 ppm containing residual water signal was removed. The total spectral area was calculated on the remaining bins and normalization was carried out on the data prior to pattern recognition. Standard binning reduced the dimension of the system to 416 bins (variables)

Statistical Analysis

Data reduction was obtained by means of Orthogonal Projection to Latent Structure (OPLS) (1) on the mean-centered data without scaling, using the algorithm implemented in the R-library “kopls” (2) and using a linear kernel.

For the purpose of classification we used the Support Vector Machines (SVM) method (3) applied on the OPLS scores. The “libsvm” module (LIBSVM: a Library for Support Vector Machines) of the R-library “e1071” was used (4).

To assess the prediction ability of the model, a double cross-validation scheme was adopted (5). In order to avoid bias in the estimation of the prediction error, we performed two nested loops of cross-validation. The original data sets were split according to a 2:8, that is, randomly removing 20% of samples, used as test set. Parameter selection (best number of predictive and orthogonal components for OPLS and kernel parameters for SVM) was carried out in the inner loop by means of 7-fold cross validation on the remaining 80%. With the best parameters we created a model which was then used to predict the excluded 20%. This procedure was repeated 100 times. We averaged the error on the test sets in order to obtain an estimation of the accuracy of the model and an accuracy for each individual. Accuracy, sensitivity and specificity were estimated using standard definitions (6).

To assess which bins were significantly different between the metastatic and early patients, a non parametric Wilcoxon test was used. A conservative estimate of significative levels of 1.201×10^{-4} was derived by using the Bonferroni correction on a nominal value of 0.05. The bins whose values appeared to be significantly different from the above statistical analysis were manually checked to identify the associated resonance peak(s). All resonances of interest were then manually checked and signals were assigned on template one-dimensional NMR profiles by using matching routines of AMIX 3.8.8 (Bruker BioSpin) in combination with the BBIOREFCODE (Version 2-0-0; Bruker BioSpin) reference database and published literature when available.

All calculations were made using R (7) scripts developed in-house.

SUPPLEMENT REFERENCES

(1) Trygg J., Wold S., Orthogonal projections to latent structures (O-PLS). J. Chemometrics 2002 ; 16: 119-128.

(2). Bylesjö M, Rantalainen M, Nicholson JK, et al. K-OPLS package: kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space, BMC Bioinformatics 2008; 19: 106

(3) Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

(4) Chang and Lin, 2001, LIBSVM: a Library for Support Vector Machines

(5) Westerhuis J, Hoefsloot H, Suzanne Smit S, et al. Assessment of PLSDA cross validation. Metabolomics 2008; 4: 81-89.

(6) Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994; 308:1552.

(7) Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J Comput Stat Graph 1996; 5: 299-314.

research articles Journal of proteome research

Are Patients with Potential Celiac Disease Really Potential? The Answer of Metabonomics

Patrizia Bernini,[†,§] Ivano Bertini,[*,†,‡] Antonio Calabrò,[||,⊥] Giancarlo la Marca,[#,▽] Gabriele Lami,[||] Claudio Luchinat,[*,†,‡] Daniela Renzi,[||,⊥] and Leonardo Tenori[§]

Magnetic Resonance Center (CERM), University of Florence, Via L.Sacconi 6, 50019 Sesto Fiorentino, Italy, Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy, FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy, Gastroenterology Unit, Department of Clinical Pathophysiology, University of Florence, Viale Pieraccini 6, 50139 Florence, Italy, Tuscany Referral Center for Adult Celiac Disease, Viale Morgagni 85, 50134 Florence, Italy, Department of Pharmacology, University of Florence, Viale Pieraccini 6, 50139 Florence, Italy, and Mass Spectrometry, Clinical Chemistry and Pharmacology Labs, Clinic of Pediatric Neurology, A. Meyer University Children's Hospital, Viale Pieraccini 24, 50134 Florence, Italy

Received September 1, 2010

Celiac disease (CD) is an autoimmune disorder caused by a permanent sensitivity to gluten in genetically susceptible individuals. Accurate diagnosis of CD at an early stage and its treatment with a gluten-free diet (GFD) are important for optimum treatment and prognosis. Recently, by employing a noninvasive metabonomic approach, we have shown that CD has a well-defined metabonomic signature. Here we address potential CD patients, defined as subjects who do not have, and have never had, a jejunal biopsy consistent with clear CD, and yet have immunological abnormalities similar to those found in celiac patients. Sixty-one overt CD patients at diagnosis, 29 patients with potential CD, and 51 control subjects were examined by ^{1}H NMR of their serum and urine: out of 29 potential CD patients, 24 were classified as CD and 5 as control subjects. Potential CD largely shares the metabonomic signature of overt CD. Most metabolites found to be significantly different between control and CD subjects were also altered in potential CD. Our results demonstrate that metabolic alterations may precede the development of small intestinal villous atrophy and provide a further rationale for early institution of GFD in patients with potential CD, as recently suggested by prospective clinical studies.

Keywords: Potential Celiac Disease • Metabolomics • Metabonomics • Nuclear Magnetic Resonance

Introduction

Celiac disease (CD) is a common autoimmune disorder triggered by the ingestion of gluten and related prolamins, plant storage proteins found in wheat, barley, and rye, in genetically susceptible individuals.[1] Patients display various degree of intestinal inflammation, ranging from mere increase in the number of intraepithelial lymphocytes (lymphocytic enteritis) to elongation of crypts (crypt hyperplasia), severe mononuclear cell infiltration of the lamina propria and overt villous atrophy. Accordingly, clinical manifestations range from completely asymptomatic forms (so-called silent CD patients) to symptoms and signs of global malabsorption.

In the majority of cases, a gluten-free diet (GFD) leads to disappearance of clinical symptoms, recovery of normal mucosal histology, disappearance of the serological signs, and the prevention of complications, while reintroduction of gluten into the diet, at any stage, causes reactivation of the disease.[2−7]

Small-bowel mucosal villous atrophy has traditionally been considered the conditio sine qua non for diagnosing CD.[8] As a logical consequence, although there is a general consensus on dietary treatment for symptomatic subjects with enteropathy, it is not so for potential cases (i.e., patients with positive CD-associated serology, but no enteropathy).[9] However, accumulating evidence suggests that celiac patients may suffer from gluten-sensitive symptoms even before villous atrophy has developed,[10−13] and an increased risk of osteoporosis has been reported also for this group of patients.[10]

In a recent work, Bertini et al. have examined a cohort of CD patients, before and after gluten free diet GDF, and healthy controls, by ^{1}H NMR metabolic profiling of their serum and urine samples.[14] Through multivariate analysis, they were able to demonstrate the existence of a clear metabonomic[15] signa-

* Corresponding authors: Ivano Bertini, Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy. E-mail, bertini@cerm.unifi.it; tel., 39-055-4574270; fax, 39-055-4574271. Claudio Luchinat, Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy. E-mail, luchinat@cerm.unifi.it; tel., 39-055-4574270; fax, 39-055-4574271.
† Magnetic Resonance Center (CERM), University of Florence.
§ FiorGen Foundation.
‡ Department of Chemistry, University of Florence.
|| Department of Clinical Pathophysiology, University of Florence.
⊥ Tuscany Referral Center for Adult Celiac Disease.
Department of Pharmacology, University of Florence.
▽ A. Meyer University Children's Hospital.

ture of CD and to identify a number of serum and urine metabolites whose levels were significantly different in CD at diagnosis as compared to controls. The metabonomic signature as well the levels of most metabolites reverted to normal within 12 months of a strict GFD. The metabonomic fingerprint of CD has two components not directly linked to malabsorption: one related to energy metabolism, and the other related to alterations of gut microflora. Using this metabolic fingerprint, it is possible to make predictions about the CD status wit reasonably good accuracy (about 84%).

These results, although extremely encouraging, are not particularly helpful from a merely diagnostic point of view, since serology by blood test already achieves a sensitivity of about 98% and a specificity of over 95%.[16] Moreover, professional guidelines recommend that a positive blood test should still be followed by an endoscopy, and this would pick up false positives as well as offer alternative explanations for the symptoms.[17−19] However, the clear discrimination between patients and healthy subjects and patients and follow-up cases, achieved by metabolic fingerprinting, undoubtedly demonstrates the existence of a metabolic signature for CD both in serum and urine.

Under the inclusive definition of celiac disease, or gluten-sensitive enteropathy, the concepts of silent, latent, and potential CD have been introduced over 15 years ago by Professor Anne Ferguson.[20] While silent CD is marked by severe damage to the jejunal mucosa in the absence of clinical symptoms, both latent and potential CD are characterized by jejunal mucosa that would be reported as normal by most clinical pathologists in an individual on a gluten-containing diet. Latent subjects sometimes in their life have had a flat jejunal biopsy which recovered on a gluten-free diet. Latent CD patients are often symptomatic; neither high titers of gliadin antibodies nor mucosal changes (including raised intraepithelial lymphocyte counts) are obligate features of latent CD, although the presence of elevated anti-endomysial antibodies (EmA) is probably the best predictor of progression toward villous atrophy.[21]

The term potential CD has been proposed for those subjects who, as opposed to latent CD patients, do not have, and have never had, a jejunal biopsy consistent with clear CD, and yet have immunological abnormalities similar to those found in celiac patients. Good markers of potential CD include the presence of serum endomysial antibodies, a high count of intraepithelial lymphocytes, and subtle pathological alteration such as increased density of intraepithelial lymphocytes expressing gamma delta T cell receptor, signs of activated mucosal cell-mediated immunity, celiac-like intestinal antibody pattern, and positive rectal gluten challenge.[21]

Our interest in potential CD subjects stems from the fact that they have all the typical immunoclinical features of clear CD, in the absence of any apparent intestinal damage. The aim of this work is to investigate whether (and to what extent) potential CD patients share the same metabolic fingerprint of overt CD patients, thereby addressing an as yet open question: are the metabolic alterations associated to CD due to malabsorption (i.e., intestinal damage) or are they independent of mucosal injury and therefore intrinsic to the pathology? In other words, our aim is to clarify if we are effectively looking at a metabolic signature of CD or if we are simply following the metabolic modification due to intestinal malabsorption. To this end, a suitable cohort of potential CD subjects is ideal because, by definition, they do not have villous atrophy.

Recent prospective studies on both adults and children classified as potential CD patients have addressed the question of whether GFD prevents development of villous atrophy and/or alleviates the symptoms. There is increasing evidence that GFD has a beneficial effect, although the effect can be more[22,23] or less[24] marked. The comparison of the fingerprint of potential CD subjects with overt CD subjects and healthy controls can thus contribute to the open debate about whether or not it is necessary to put these patients on a GFD.

Materials and Methods

Volunteers Recruitment and Samples Collection. The volunteers participating in this study were recruited by the Tuscany Referral Center for Adult CD. In total, 141 adult (35 males, 106 females, mean age 38.3 ± 13.9 years) participated in this study. Among the celiac group, 61 patients (10 males, 51 females, mean age 39.8 ± 13.9 years) were affected by overt CD, and 29 were diagnosed as having potential CD (6 males, 23 females, mean age 34.4 ± 14.3 years). The diagnosis of CD was based on positive serology and confirmed by histological examination of small bowel biopsies taken from the distal part of the duodenum. Small intestinal mucosal damage was graduated according to the classification of Marsh modified by Oberhuber.[25] According to the hystology, 7 patients with overt CD had partial villous atrophy (Marsh 3A), 7 subtotal villous atrophy (Marsh 3B), and 16 total villous atrophy (Marsh 3C). Patients with potential CD were all positive for anti-tissue transglutaminase antibodies (tTGA); all but one were also EmA positive; histological examination revealed an apparently normal mucosa (Marsh 0) in 9 subjects, just an increase in intraepithelial lymphocytes (Marsh 1) in 12 and increased intraepithelial lymphocytes coupled with crypt hyperplasia (Marsh 2) in 8 patients. The control group consisted of 51 nonceliac subjects (19 males, 32 females, mean age 36.6 ± 14.2 years): 34 (13 males, 21 females, mean age 37.6 ± 14.8 years) were healthy controls from the medical staff of the Gastrointestinal Unit and 17 (6 males, 11 females, mean age 30.9 ± 12.5 years) were patients with clinical suspicion of celiac disease; they were all EmA and tTGA negative. Therefore, for ethical reasons, the 34 controls that were from the medical staff did not undergo intestinal biopsy; in the remaining 17 cases, gastrointestinal endoscopy was performed only when it was considered necessary to the diagnosis. The clinical characteristics of all the subjects, including those examined in the previous work[14] are reported in Supporting Table S1.

Each subject had fasted overnight, and urine and blood were collected in the morning preprandial. Venous blood samples were collected into plastic serum tubes (6 mL), with increased silica act clot activator, silicone-coated interior (Becton Dickinson, Plymouth, U.K.). Samples were allowed to clot by standing tubes vertically at room temperature (22 °C) for 60 min. Tubes were centrifuged at 1800 RCF for 10 min at room temperature. Within 15 min of centrifugation, the supernatant (serum) was transferred in 500 μL aliquots to prelabeled 1 mL cryovials (Bruker BioSpin, Milan, Italy). Three aliquots per patient were immediately frozen and stored at −80 °C until used.

First morning preprandial urine void was used for the collection period. Patients were supplied with appropriate collection instructions and information on fasting, diet and medication restrictions when necessary. The urine samples were collected into prelabeled sterile collections cups. One milliliter of urine samples was transferred into prelabeled 1 mL

sterile cryovials (Nalgene, Rochester, NY). Three aliquots per patient were immediately frozen and stored at −80 °C until used.

Additionally, CD patients and healthy controls were asked to record their dietary intake and the use of any medication (either prescribed or self-administered) on the day before each visit, and to fast from midnight until blood samples and urine were collected the following morning.

Genotyping. Potential CD patients were analyzed for CD-related genes following the procedure already reported.[14] They were all genetically susceptible: in particular 22 were DQ2, 3 DQ8, and 4 DQ2.2 positive (i.e., they carried the HLA genes DRB1 *07, DQB1 *0202 and DQA1 *0201, see Supporting Table S2). In the course of our previous work,[14] we had checked that among healthy subjects those who were DQ2/DQ8 positive were not statistically distinguishable from the negative ones, while they were distinguishable from the CD patients.

Antibody Testing. Anti-tissue transglutaminase antibodies were measured by a commercially available enzyme-linked immunosorbent assay kit (EutTG, Eurospital, Trieste, Italy) that employs human recombinant tTG as antigen; EmA were determined by indirect immunofluorescence, using tissue section of monkey esophagus (Eurospital), as previously described.[26]

Citrulline Assay. Labeled standards of citrulline were purchased from Cambridge Isotope Laboratories, Andover, MA; a stock solution was made in methanol. The standard concentration was 500 μmol/L. To obtain working solutions, daily dilutions (1:100) were made using methanol−water 85:15. All chemicals and solvents were of the highest purity available from commercial sources and were used without any further purification.

A dried serum spot was punched into a 1.5 mL tube or a 96 well plate and 200 μL of methanol containing labeled standard was added. The sample was shaken on a vortex system for 20 min, and was then dried under a nitrogen flow at 50 °C. The extracted amino acid was derivatized to butyl-ester using *n*-butanol and HCl (3 M) at 65 °C for 25 min. After derivatization, the sample was dried under nitrogen flow at 55 °C and then recovered using 200 μL of water/acetonitrile (1:1) containing 0.1% formic acid. A total of 40 μL of the diluted sample was injected in FIA (Flow Injection Analysis) mode for the MS/MS experiments.

An Applied Biosystems-Sciex (Toronto, Canada) API 4000 triple-quadrupole mass spectrometer equipped with a TurboIonSpray source was employed for this study. The TurboIonSpray source was operated in positive ion mode with a needle potential of +5900 V and with a "turbo" gas flow of 10 L/min of air heated at 400 °C (nominal heating-gun temperature). Mass calibration and resolution adjustments on the resolving quadrupoles were performed automatically by using a PPG 10^{-7} mol/L solution introduced via the built-in infusion pump. The peak width was set on both resolving quadrupoles at 0.7 Th (measured at 1/2 height) for all MS and MS/MS experiments. Collision Activated Dissociation (CAD) MS/MS was performed in the LINAC Q2 collision cell, operating with 10 mTorr pressure of nitrogen as collision gas. The declustering potential (DP) and collision energy (CE) were automatically optimized for amino acids using the Analyst 1.3.2 software. The resulting DP was +20 eV, and optimal CE was found to be 25 eV.

MS and MS/MS spectra were collected in continuous flow mode by connecting the infusion pump directly to the TurboIonSpray source. The quantitation experiments were performed using a Series 1100 Agilent Technologies (Waldbronn, Germany) CapPump coupled to an Agilent Micro ALS autosampler, both fully controlled from the API 4000 data system. Experimental flow rate was 50 μL/min using water/acetonitrile (1:1) mixture containing 0.1% formic acid. The eluent from the column was directed to the TurboIonSpray probe. The acquired data were processed using the Analyst 1.4.1 proprietary software including the "Explore" option (for chromatographic and spectral interpretation) and the ChemoView software (for quantitative information generation).

NMR Sample Preparation. Frozen serum samples were thawed at room temperature and shaken before use. A total of 300 μL of a phosphate sodium buffer (70 mM Na_2HPO_4; 20% (v/v) ^2H_2O; 0.025% (w/v) NaN_3; 0.8% (w/v) sodium trimethylsilyl [2,2,3,3-$^2H_4$]propionate (TSP); pH 7.4) was added to 300 μL of each serum sample, and the mixture was homogenized by vortexing for 30 s. A total of 550 μL of this mixture was transferred into a 4.25 mm NMR tube (Bruker BioSpin srl) for analysis.

Frozen urine samples were thawed at room temperature and shaken before use. Samples were centrifuged at $1.4 \times 10^4 g$ for 5 min and 540 μL of the supernatant was added to 60 μL of sodium phosphate buffer (0.2 M Na_2HPO_4; 0.2 M NaH_2PO_4 in 100% ^2H_2O; pH 7.0) also containing 10 mM TSP, and 30 mM NaN_3. A total of 450 μL of the mixture was pipetted into 4.25 mm NMR tubes (Bruker BioSpin srl).

NMR Analysis. ^{1}H NMR spectra for all samples were acquired using a Bruker 600 MHz spectrometer (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm CPTCI ^{1}H−^{13}C/^{31}P−^{2}H cryo-probe including a *z*-axis gradient coil, an automatic tuning-matching (ATM) and an automatic sample changer. A PT 100 thermocouple provided temperature stabilization at the level of approximately 0.1 K for the sample. Before measurement, samples were kept for at least 3 min inside the NMR probehead, for temperature equilibration (300 K).

For each urine sample, a one-dimensional NMR spectrum was acquired with water peak suppression using a standard pulse sequence (NOESYpresat; Bruker), using 64 free induction decays (FIDs), 64k data points, a spectral width of 12 019 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s, and a mixing time of 100 ms.

One-dimensional spectra of serum samples were acquired using a Carr−Purcell−Meiboom−Gill (CPMG; Bruker) spin−echo sequence to suppress signals arising from high molecular weight molecules and a standard pulse sequence (NOESYpresat, same acquisition parameters as in the case of urine samples).

Spectral Processing. Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were manually corrected for phase and baseline distortions and calibrated (TMSP peak at 0.00 ppm) using TopSpin (Version 2.1, Bruker). Each 1-D spectrum in the range between 0.02 and 10.00 ppm was segmented into 0.02-ppm chemical shift bins, and the corresponding spectral areas were integrated using AMIX software (Bruker BioSpin). Regions between 6.0 and 4.5 ppm containing residual water and urea signals were removed. The total spectral area was calculated on the remaining bins and normalization was carried out on the data prior to pattern recognition.

Statistical Analysis. Data reduction was obtained by means of Kernel Orthogonal Projection to Latent Structure (KOPLS),[27]

using the algorithm implemented in the R-library "kopls"[28] using a Gaussian kernel.

For the purpose of classification, we used the Support Vector Machines (SVM) method[29] applied on the KOPLS scores. The "libsvm" module (LIBSVM: a Library for Support Vector Machines) of the R-library "e1071" was used.[30]

The statistical procedure used is a combination of KOPLS for data reduction (i.e., to reduce the initial number of variables and to have a clear depiction of the class separation) followed by SVM on the KOPLS scores. SVM are used with the default "C-classification" settings.

Accuracy, specificity, and sensitivity were estimated using standard definitions. The global accuracy for classification was assessed by means of double cross-validation scheme.[31] The original data set was spit in a training set (90%) and a test set (10%) prior to any other step of statistical analysis. Parameter selection was carried out by 3-fold cross-validation on the 90% training set. The whole procedure was repeated 100 times

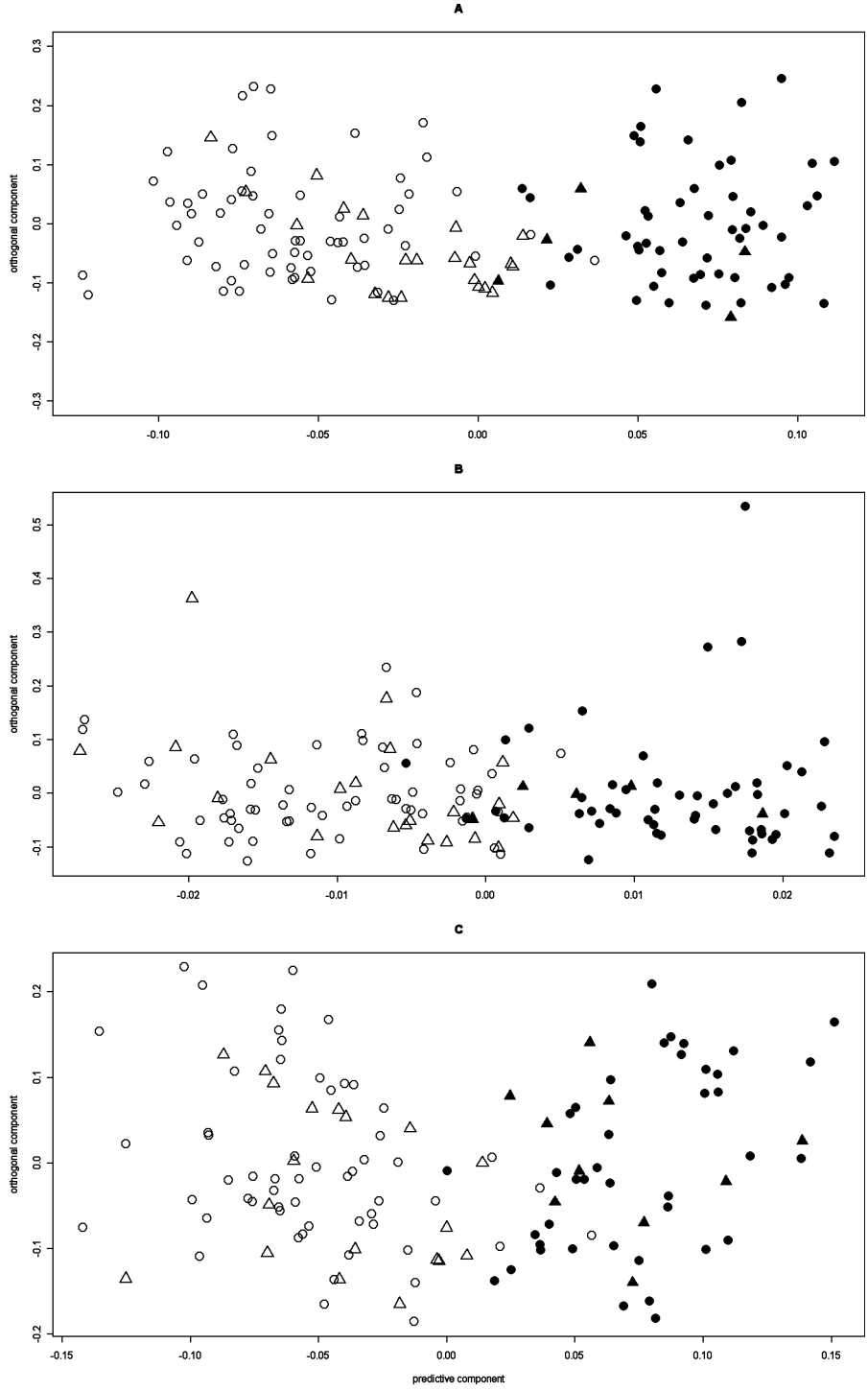


Figure 1. Clusterization of overt CD patients (open circles) and healthy controls (filled circles) obtained with kopls for cpmg serum spectra (A), NOESY serum spectra (B), and NOESY urine spectra (C). Potential CD patients (triangles) are projected on the kopls component, predicted by SVM, and filled or not accordingly to the results.

Table 1. Cross-Validation Classification Results Obtained Using CPMG and NOESY Serum Spectra and NOESY Urine Spectra of Overt CD Patients and Nonceliac Controls (NC)

	overt CD	NC	accuracy
	Serum CPMG Spectra		
overt CD	80.5%	19.5%	81.2%
NC	18.0%	82.0%	
	Serum NOESY Spectra		
overt CD	77.4%	22.6%	75.5%
NC	26.7%	73.3%	
	Urine NOESY Spectra		
overt CD	77.5%	22.5%	74.7%
NC	29.0%	71.0%	

Table 2. Cross-Validation Classification Results Obtained Using CPMG and NOESY Serum Spectra and NOESY Urine Spectra of CD Patients and Potential CD Subjects

	overt CD	potential CD	accuracy
	Serum CPMG Spectra		
overt CD	71.1%	28.9%	57.8%
potential CD	70.2%	29.8%	
	Serum NOESY Spectra		
overt CD	67.6%	32.3%	55.2%
potential CD	71.1%	28.9%	
	Urine NOESY Spectra		
overt CD	81.4%	18.6%	57.4%
potential CD	89.7%	10.3%	

inside a Monte Carlo cross-validation scheme. From this procedure, we obtain unbiased generalization error estimation. To avoid overtraining and to speed up the calculations, in the inner loop of the cross-validation (parameter selection), the SVM was kept with the default settings, varying only the number of components and the kernel sigma parameter for KOPLS.

To assess which buckets (that is, resonance peaks) were significantly different between different groups, a univariate wilcoxon test was used. A P-value ≤ 0.05 was considered statistically significant. The buckets whose values appeared to be significantly different from the above statistical analysis were manually checked to identify the associated resonance peak(s). All resonances of interest were then manually checked and signals were assigned on template one-dimensional NMR profiles by using matching routines of AMIX 6.8 (Bruker BioSpin) in combination with the BBIOREFCODE (Bruker BioSpin) reference database and published literature when available. All calculations were made using homemade scripts written in R language.[32] Serum citrulline concentrations in the three groups were analyzed by two-way ANOVA, followed by Tukey's multiple comparison test, and a P-value of less than 0.05 was accepted to denote statistical significance.

Ethical Issue. The study was approved by the Internal Review Board of the Department of Clinical Pathophysiology of the University of Florence, and informed consent was obtained from all patients and healthy volunteers.

Results

The clusterization of controls and overt untreated CD patients for the three sets of experiments (serum CPMG, serum NOESY, and urine NOESY) is shown in Figure 1. It appears that the best clusterization is provided by the serum CPMG spectra.

The statistical models applied proved to be effective in discriminating between patients and control subjects. These results are in line with our previous work,[14] where for the first time the existence of a metabolic fingerprint of celiac disease was demonstrated, both in serum and urine. Although the best recognition accuracy is slightly less than previously reported (81.2 vs 83.4 for serum CPMG spectra), these new data undoubtedly confirm the existence of the metabolic signature of CD, both in serum and in urine. In Table 1, the values for classification accuracy for each experiment are reported.

If the potential CD patients are treated as blind samples and are subjected to classification by the previous statistical model, 24 out of 29 are classified as CD patients using CPMG serum spectra. Those results strongly suggest that potential CD subjects are more similar to overt CD than to healthy controls. Moreover, when the statistical analysis is used to discriminate potential from overt CD patients, the accuracy is significantly lower (57.8%). Furthermore, as shown in Table 2, this result is highly asymmetric with high sensitivity (>70%) but poor specificity (ca. 20%). Again potential CD subjects show a behavior that is much more similar to overt CD.

When comparing spectra of serum samples of overt untreated CD patients and nonceliac controls, we find that CD patients are characterized ($P < 0.05$) by lower levels of choline, creatinine, isoleucine, lactate, leucine, methionine, valine, lipids, and glycoproteins (from their *N*-acetyl signal) and higher levels of glucose, citrate, serine, and phenylalanine.

In the case of urine, overt CD patients are characterized by higher levels of acetoacetate, choline, glutamate/glutamine, indoxyl sulfate (IS), meta-hydroxyphenyl-propioni acid (mHPPA), phenylacetylglycine (PAG), 3-hydroxybutyric acid, methylamine, valeric acid, sarcosine, *p*-cresolsulfate, and 1-methylnicotinamide. These data are summarized in Table 3. Even though minor differences are present (due to the different data set and slightly different statistical analysis), this pattern is coherent with the previous work and corroborates the previous findings.[14]

Comparing the serum spectra of overt CD with those of potential CD, we find that potential CD patients are characterized by lower levels of dimethylglycine and higher levels of glutamate and lipids ($P < 0.05$); no other metabolites are found significantly different in serum. In urine, more differences are found, in particular potential CD patients shows lower levels of glutamate/glutamine, IS, mHPPA, PAG, *p*-cresolsulfate, and citrate (see Table 3).

The metabolic differences between potential CD and nonceliac controls in serum are quite similar to those between overt CD and controls: potential CD subjects have lower levels of creatinine, leucine, dimethylglycine, arginine, and lipids and higher levels of glucose, proline, and glutamate. In urine, potential CD subjects show lower levels of hippurate and higher levels of dimethylamine, 1-methylnicotinamide, and citrate. These data are summarized in Table 3.

Several of the potential CD patients analyzed here were put on a GFD, although solely based on medical considerations and not on the results from metabolomics. At the present stage, we have been able to examine again 13 of them, after an average period of GFD of 6 months (Supporting Table S2): all but three had negative TGA. By processing the follow-up serum samples with the same statistical model built for the discrimination, we found that all of them, except for one of the three positive to TGA, had reverted to a normal metabonomic profile.

In addition to metabolomic analysis, with the aim of obtaining more information able to better characterize the clinical

Table 3. Metabolites Found To Be Statistically Different ($P < 0.05$) in Sera (A) and Urine (B) of Untreated Overt CD Patients, Potential CD Patients, and Nonceliac Controls (CD)[a]

metabolite	CD patients vs NC	potential CD vs NC	overt CD vs potential CD
	(A) Serum		
Choline	−		
Creatinine	−	−	
Glucose	+	+	
Glycoproteins	−		
Isoleucine	−		
Lactate	−		
Leucine	−	−	
Lipids	−	−	−
Methionine	−		
Valine	−	−	
Citrate	+		
Serine	+		
Phenilalanine	+		
Glutamate	+	+	+
Dimethylglycine		−	+
Proline		+	
Arginine		−	
	(B) Urine		
Acetoacetate	+		
Choline	+		
Glutamate/Glutamine	+		+
Indoxyl Sulfate	+		+
mHPPA	+		+
PAG	+		+
3-OH-butirate	+		
Methylamine	+		
Valerate	+		
Sarcosine	+		
p-cresolsulfate	+		
1-methylnicotinammide	+	−	
Citrate	+	−	+
Hippurate		−	

[a] The + and − indicate, respectively, higher and lower levels of metabolites concentrations in the first group of each classification, with respect to the other.

phenotype of potential CD patients, we monitored and compared the citrulline levels of the three kind of subjects (30 overt CD, 51 controls, and 29 potential CD). Serum citrulline level is in fact an indirect measure of the enterocyte mass and it was proposed as a clinical biomarker for enterocyte dysfunctions in many diseases.[33−36] Although the idea of observing citrulline levels of celiac patients is not new,[37] this is, at the best of our knowledge, the first time that such analysis is done in potential CD patients. The results are extremely interesting because a statistically significant difference ($P < 0.001$) is obtained not only between CD patients and the control group (as expected), but also between potential CD subjects and controls. Moreover, the difference between CD patients and potential CD subjects is not statistically significant (Figure 2).

Discussion

The present work classifies potential CD subjects (29) on the basis of a learning set of controls and CD subjects that has been considerably expanded with respect to the previous study[14] (61 + 51). The global accuracy obtained in discriminating nonceliac control subjects and patients is good enough (81.2%) to allow us to try to make predictions about the metabolic status of potential CD subjects. It is important to stress here that in doing

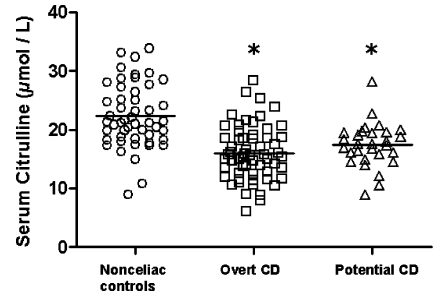


Figure 2. Serum citrulline concentrations in the three groups of subjects. In overt and potential CD, the mean serum citrulline concentration was significantly lower (15.97 + 4.62 SD and 17.38 + 3.70 SD) than in nonceliac controls (22.30 + 5.39 SD, $*p < 0.001$). One-way analysis of variance followed by Tukey's post test was applied to the results.

so we are not proposing an alternative diagnostic test for the screening of difficult cases, sometimes challenging also for expert clinicians. Rather, this exercise is useful to have a qualitative answer to the title question. Using CPMG serum spectra, the result of the prediction is remarkably clear: out of 29 potential CD subjects, 24 are classified as CD and 5 are classified as controls. Furthermore, three of the latter are located near the cluster boundary (e.g., they are borderline subjects) and only two are clearly classified as controls (Figure 1A). Although it is not possible to give a rationale for these differences, the answer of metabolomics is straightforward: potential CD subjects are extremely similar to overt CD patients, at least in their metabolic profile.

If we build a statistical model optimized for the discrimination between celiac patients and potential subjects, the classification accuracy is very low: 57.8% for serum CPMG spectra, 55.1% for serum NOESY spectra, and 57.4% for urine. In all cases, we have high sensitivity but a high rate of false discovery, so the specificity is very poor (29%, 29%, and 10%). In other words, it is impossible to separate overt CD from potential CD subjects, based on metabolomics.

The analysis of the metabolic profiles reveals interestingly similar behavior of overt CD and potential CD. Only few serum metabolites differentiate potential from overt CD, and none of these metabolites are relative to energy metabolism. In particular, it appears that, as in untreated CD patients, glycolysis is somehow impaired. Impairment of glycolysis explains both a lowering of lactate levels and an increase of glucose levels in blood. This is confirmed by the metabolic pattern that differentiates potential CD and nonceliac controls: most of the significant metabolites are the same that differentiate overt CD from control group.

In urine, there are many more metabolites that differentiate potential CD from overt CD than from controls. Especially, the key differences are in the metabolites originating from the gut microflora (m-HPPA, IS, and PAG), suggesting a relationship between overt CD, villous atrophy, and bacterial consortia of the host. It is also reported[38] that particular bacterial strains may work in conjunction with gluten to cause an increased immune response which is responsible of the development of CD. A distinctive 'microbial signature' in celiac pediatric patients is also reported.[39]

The similarity of the metabolic pattern in serum, and the dissimilarity in urine, between overt and potential CD subjects allow us to hypothesize that while the alterations in the urine profile may follow intestinal damage, gluten intolerance exists as such before an evident intestinal damage occurs.

To clarify this aspect and to shed more light on the present findings, we performed a serum citrulline assay. In humans, citrulline is an amino acid involved in intermediary metabolism and is not incorporated in proteins.[40−42] Circulating citrulline is mainly produced by enterocytes of the small bowel.[43] For this reason, plasma or serum citrulline concentration has been proposed as a biomarker of residual small bowel mass and function.[44]

Our data, in line with previous findings, clearly show that citrulline levels in the CD patients are significantly lower than in the control group. But, more strikingly, also potential CD subjects, that from endoscopy do not show any manifest sign of villous atrophy, have lower levels of citrulline than controls. Indeed no statistically significant difference is found between citrulline levels in CD patients and potential CD subjects (Figure 2).

According to clinical experience, if a celiac subject has a positive antibody test but does not show macroscopic alterations of intestinal mucosa, there can be only two possibilities: (1) a (limited) intestinal damage is present but the endoscopy was not able to detect it (i.e., we admit some error in the sampling procedure) or (2) the intestinal mucosa is really intact and the subject is a so-called "potential CD subject".

The interpretation of our data (both from metabonomics and from citrulline assay) suggests a third option: potential CD subjects are, indeed, not potential at all. They already experience some subtle alteration of the enterocytes (at the microscopic functional level but not at the macroscopic level) and metabolically appear similar to overt CD also without any histological evidence of intestinal damage. If this hypothesis holds, the core result from our investigation would be that metabolomics can detect CD also when its clinical manifestation is not fully evident. Moreover, these findings may bear the important medical implication that the metabolic and enterocyte alterations of CD arise *before* macroscopic intestinal damage and/or villous atrophy and precede the observable modification of jejunal mucosa. In this context, the observation that potential CD subjects are characterized by a similar serum metabonomic fingerprint as overt CD patients provides a further, more biological, rationale for the medical evidence-based choice of putting potential CD subjects on a GFD. Moreover, if the metabolic changes are (at least partially) independent from the bowel malabsorption (i.e., metabolites of gut microflora and energetic pathway), a deeper analysis of this fingerprint can be helpful to infer more information on the biochemistry of the development of the disease.

Acknowledgment. This work was partly supported by a fellowship from Boehringer Ingelheim Italia (to L.T. through the FiorGen Foundation). The authors declare no competing financial interest.

Supporting Information Available: Tables of demographic and clinical data on study patients; demographic data, genotyping, antibodies levels and metabolomic classification of the 29 potential CD subjects. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kagnoff, M. F. Celiac disease: pathogenesis of a model immunogenetic disease. *J. Clin. Invest.* **2007**, *117*, 41–49.
(2) Trier, J. S. Celiac sprue. *N. Engl. J. Med.* **1991**, *325*, 1709–1719.
(3) Schuppan, D. Current concepts of celiac disease pathogenesis. *Gastroenterology* **2000**, *119*, 234–242.
(4) Farrell, R. J.; Kelly, C. P. Celiac sprue. *N. Engl. J. Med.* **2002**, *346*, 180–188.
(5) Green, P. H.; Jabri, B. Coeliac disease. *Lancet* **2003**, *362*, 383–391.
(6) Green, P. H.; Cellier, C. Celiac disease. *N. Engl. J. Med.* **2007**, *357*, 1731–1743.
(7) Di Sabatino, A.; Corazza, G. R. Coeliac disease. *Lancet* **2009**, *373*, 1480–1493.
(8) Troncone, R.; Auricchio, R.; Granata, V. Issues related to gluten-free diet in coeliac disease. *Curr. Opin. Clin. Nutr. Metab. Care* **2008**, *11*, 329–333.
(9) Setty, M.; Hormaza, L.; Guandalini, S. Celiac disease: risk assessment, diagnosis, and monitoring. *Mol. Diagn. Ther.* **2008**, *12*, 289–298.
(10) Kaukinen, K.; Mäki, M.; Partanen, J.; Sievänen, H.; Collin, P. Celiac disease without villous atrophy. Revision of criteria called for. *Dig. Dis. Sci.* **2001**, *46*, 879–887.
(11) Kaukinen, K.; Peräaho, M.; Collin, P.; Partanen, J.; Woolley, N.; Kaartinen, T.; Nuutinen, T.; Halttunen, T.; Mäki, M.; Korponay-Szabo, I. Small bowel mucosal transglutaminase 2-specific IgA deposits in coeliac disease without villous atrophy: a prospective and randomized study. *Scand. J. Gastroenterol.* **2005**, *40*, 564–572.
(12) Troncone, R. Latent coeliac disease in Italy. *Acta Paediatr.* **1995**, *84*, 1252–7.
(13) Matysiak-Budnik, T.; Malamut, G.; de Serre, N. P.; Grosdidier, E.; Seguier, S.; Brousse, N.; Caillat-Zucman, S.; Cerf-Bensussan, N.; Schmitz, J.; Cellier, C. Long-term follow-up of 61 patients diagnosed in childhood: evolution towards latency is possible on a normal diet. *Gut* **2007**, *56*, 1379–86.
(14) Bertini, I.; Calabrò, A.; De Carli, V.; Luchinat, C.; Nepi, S.; Porfirio, B.; Renzi, D.; Saccenti, E.; Tenori, L. The metabonomic signature of celiac disease. *J. Proteome Res.* **2009**, *8*, 170–177.
(15) Nicholson, J. K.; Lindon, J. C.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29*, 1181–1189.
(16) Fasano, A.; Catassi, C. Current approaches to diagnosis and treatment of celiac disease: an evolving spectrum. *Gastroenterology* **2001**, *12*, 1527–1528.
(17) Jackson Allen, P. L. Guidelines for the diagnosis and treatment of celiac disease in children. *Pediatr. Nurs.* **2004**, *30*, 473–476.
(18) Cellier, C.; Green, P. H.; Collin, P.; Murray, J. ICCE consensus for celiac disease. *Endoscopy* **2005**, *37*, 1055–1059.
(19) Hill, I. D.; Dirks, M. H.; Liptak, G. S.; Colletti, R. B.; Fasano, A.; Guandalini, S.; Hoffenberg, E. J.; Horvath, K.; Murray, J. A.; Pivor, M.; Seidman, E. G. North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. Guideline for the diagnosis and treatment of celiac disease in children: recommendations of the North American Society for Pediatric Gastroenterology, Hepatology and Nutrition. *J. Pediatr. Gastroenterol. Nutr.* **2005**, *40*, 1–19.
(20) Ferguson, A.; Arranz, E.; O'Mahony, S. Clinical and pathological spectrum of coeliac disease: active, silent, latent and potential. *Gut* **1993**, *34*, 150–151.
(21) Troncone, R.; Greco, L.; Mayer, M.; Paparo, M.; Caputo, N.; Micillo, M.; Mugione, P.; Auricchio, S. Latent and potential coeliac disease. *Acta Paediatr. Supp.* **1996**, *l412*, 10–14.
(22) Kurppa, K.; Ashorn, M.; Iltanen, S.; Koskinen, L. L.; Saavalainen, P.; Koskinen, O.; Mäki, M.; Kaukinen, K. Celiac disease without villous atrophy in children: a prospective study. *J. Pediatr.* **2010**, *157* (3), 373–380.
(23) Kurppa, K.; Collin, P.; Viljamaa, M.; Haimila, K.; Saavalainen, P.; Partanen, J.; Laurila, K.; Huhtala, H.; Paasikivi, K.; Mäki, M.; Kaukinen, K. Diagnosing mild enteropathy celiac disease: a randomized, controlled clinical study. *Gastroenterology* **2009**, *136* (3), 816–823.
(24) Tosco, A.; Salvati, V. M.; Auricchio, R.; Maglio, M.; Borrelli, M.; Corazzo, A.; Paparo, F.; Beffardi, M.; Esposito, A.; D'Adamo, G.; Malamisura, B.; Greco, L.; Troncone, R. Most children with potential celiac disease are healthy but 1 third of them develop villous atrophy at 3-years follow-up *Clin. Gastroenterol. Hepatol.* [Online early access]. DOI: 10.1016/j.cgh.2010.09.006. Published online: Sept 17, 2010.
(25) Oberhuber, G.; Granditsch, G.; Vogelsang, H. The histopathology of coeliac disease: time for a standardized report scheme for pathologists. *Eur. J. Gastroenterol. Hepatol.* **1999**, *11*, 1185–1194.
(26) Matà, S.; Renzi, D.; Pinto, F.; Calabrò, A. Anti-tissue transglutaminase IgA antibodies in peripheral neuropathy and motor neuronopathy. *Acta Neurol. Scand.* **2006**, *114*, 54–58.

(27) Rantalainen, M.; Bylesjö, M.; Cloarec, O.; Nicholson, J. K.; Holmes, E.; Trygg, J. Kernel-based orthogonal projections to latent structures (K-OPLS). *J. Chemom.* **2007**, *21*, 376–385.
(28) Bylesjö, M.; Rantalainen, M.; Nicholson, J. K.; Holmes, E.; Trygg, J. K-OPLS package: kernel-based orthogonal projections to latent structures for prediction and interpretation in feature space. *BMC Bioinf.* **2008**, *9*, 106–110.
(29) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer-Verlag: New York, 1995.
(30) Chang C.-C.; Lin, C.-J.; LIBSVM: a Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
(31) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilze, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA cross validation. *Metabolomics* **2008**, *4*, 81–89.
(32) Ihaka, R.; Gentleman, R. R: A language for data analysis and graphic. *J. Comput. Stat. Graph.* **1996**, *5*, 299–314.
(33) Crenn, P.; Coudray-Lucas, C.; Thuillier, F.; Cynober, L.; Messing, B. Postabsorptive plasma citrulline concentration is a marker of absorptive enterocyte mass and intestinal failure in humans. *Gastroenterology* **2000**, *119*, 1496–505.
(34) Crenn, P.; Vahedi, K.; Lavergne-Slove, A.; Cynober, L.; Matuchansky, C.; Messing, B. Plasma citrulline: a marker of enterocyte mass in villous atrophy-associated small bowel disease. *Gastroenterology* **2003**, *124*, 1210–1219.
(35) Blijlevens, N. M.; Lutgens, L. C.; Schattenberg, A. V.; Donnelly, J. P. Citrulline: a potentially simple quantitative marker of intestinal epithelial damage following myeloablative therapy. *Bone Marrow Transplant.* **2004**, *34*, 193–196.
(36) Pappas, P. A.; Tzakis, A. G.; Saudubray, J. M. Trends in serum citrulline and acute rejection among recipients of small bowel transplants. *Transplant Proc.* **2004**, *36*, 345–347.
(37) Miceli, E.; Poggi, N.; Missanelli, A.; Bianchi, P.; Moratti, R.; Corazza, G. R. Is serum citrulline measurement clinically useful in coeliac disease. *Intern. Emerg. Med.* **2008**, *3*, 233–236.
(38) Sollid, M. L.; Gray, G. M. A role for bacteria in celiac disease. *Am. J. Gastroenterol.* **2004**, *99*, 894–904.
(39) Schippa, S.; Iebba, V.; Barbato, M.; di Nardo, G.; Totino, V.; Proietti Checchi, M.; Longhi, C.; Maiella, G.; Cucchiara, S.; Conte, M. P. A distinctive 'microbial signature' in celiac pediatric patients. *BMC Microbiol.* **2010**, *10*, 175.
(40) Windmueller, H. G.; Spaeth, A. E. Source and fate of circulating citrulline. *Am. J. Physiol.* **1981**, *241*, 473–480.
(41) Wakabayashi, Y.; Jones, M. E. Pyrroline-5-carboxylate synthesis from glutamate by rat intestinal mucosa. *J. Biol. Chem.* **1983**, *258*, 3865–3872.
(42) Wu, G. Intestinal mucosal amino acid catabolism. *J. Nutr.* **1998**, *128*, 1249–1252.
(43) Wakabayashi, Y.; Yamada, E.; Hasegawa, T.; Yamada, R. Enzymological evidence for the indispensability of small intestine in the synthesis of arginine from glutamate.I. Pyrroline-5-carboxylate synthase. *Arch. Biochem. Biophys.* **1991**, *291*, 1–8.
(44) Crenn, P.; Messing, B.; Cynober, L. Citrulline as a biomarker of intestinal failure due to enterocyte mass reduction. *Clin Nutr.* **2008**, *27*, 328–339.

PR100896S

Supporting Table S1. Demographic and clinical data on study patients.

	overt CD n = 61	potential CD n = 29	nonceliac controls n = 51
Female, n (%)	51 (83.6)	23 (79.3)	32 (62.7)
Mean age (SD) yrs	39.8 (13.9)	34.4 (14.3)	36.6 (14.2)
Abdominal complaints*, n (%)	38 (62.3)	13 (44.8)	10 (16.4)
Malabsorption†, n (%)	26 (42.6)	6 (20.7)	5 (9.8)
Atypical symptoms‡, n (%)	21 (34.4)	7 (24.1)	5 (9.8)
Screening of risk groups§, n (%)	6 (9.8)	3 (10.3)	2 (11.7)
Family history of CD, n (%)	19 (31.1)	6 (20.79)	2 (3.9)

*Chronic diarrhea, flatulence, abdominal pain

†Weight loss, anaemia

‡Aphthous stomatitis, cryptogenic hypertransaminasemia, polyneuropathy, migraine, arthralgia, alopecia, infertility, recurrent miscarriage

§Type 1 diabetes, autoimmune thyroid disease, Sjögrens syndrome, osteoporosis

Supporting Table S2. Demographic data, genotyping, antibodies levels and metabolomic classification of the 29 potential CD subjects. 13 of them where analyzed again after GFD.

N°	Age (yr)	Gender	HLA genotyping	tTG IgA Before GFD	EMA IgA Before GFD	CPMG classification Before GFD	tTG IgA After GFD	EMA IgA After GFD	CPMG classification After GFD
1	63	M	**DQ2/DQ2.2**	> 100	POS	CD	6	NEG	Healthy
2	36	M	**DQ2/DQ2**	> 100	POS	CD	12	Low POS	Healthy
3	50	F	**DQ2**/DQ7	17.0	POS	CD	6.4	NEG	Healthy
4	15	F	**DQ2/**DQ5	18.6	POS	CD			
5	16	F	**DQ2**/DQ7	18.6	POS	Healthy			
6	32	F	**DQ2**/DQ7	22.4	NEG	CD			
7	19	F	**DQ2/DQ8**	> 100	POS	CD	9.5	NEG	Healthy
8	18	F	**DQ2/DQ2.2**	16.6	POS	CD	4.3	NEG	Healthy
9	39	F	**DQ8**/DQ6	19.4	POS	CD			
10	44	F	**DQ2**/DQ7	38.6	POS	CD			
11	26	F	**DQ2**/DQ7	31.5	POS	CD			
12	49	F	**DQ8**/DQ7	32.3	POS	CD			
13	51	F	**DQ2**/DQ7	26.0	POS	CD			
14	23	F	**DQ2.2/**DQ5	21.8	POS	CD	4.7	NEG	Healthy
15	18	M	**DQ2/**DQ5	38.1	POS	CD			
16	34	F	**DQ2/DQ2.2**	> 100	POS	CD	14.2	POS	Healthy
17	30	F	**DQ2.2/**DQ5	28.1	POS	Healthy	3.9	NEG	Healthy
18	37	F	**DQ2**/DQ7	19.5	POS	CD	8.2	Low POS	Healthy
19	35	F	**DQ2.2/**DQ6	32.5	POS	CD			
20	45	M	**DQ2.2/**DQ5	> 100	POS	Healthy	25.2	POS	Healthy
21	44	F	**DQ2/**DQ5	24.7	POS	Healthy			
22	66	F	**DQ2**/DQ7	38.2	POS	Healthy	21	POS	CD
23	39	F	**DQ2/DQ2.2**	16.1	POS	CD	6.7	NEG	Healthy
24	48	F	**DQ2**/DQ6	18.7	POS	CD	16.9	POS	Healthy
25	20	M	**DQ2**/DQ7	46.3	POS	CD			
26	20	F	**DQ2**/DQ7	80.4	POS	CD			
27	35	M	**DQ2**/DQ7	28.6	POS	CD			
28	44	F	**DQ2**/DQ7	23.1	POS	CD			
29	38	F	**DQ2/DQ2.2**	23.7	POS	CD			

The metabonomic signature of cardiovascular risk points to oxidative stress of hepatic cell mitochondria

Patrizia Bernini, MSc; Ivano Bertini, PhD; Claudio Luchinat, PhD; Leonardo Tenori, PhD; Adriana Tognaccini, MD; (altri from PT?)

From Magnetic Resonance Center, University of Florence, Via L.Sacconi 6, 50019 Sesto Fiorentino, Italy (PB, IB, CL), Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy (IB, CL), FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy (PB, LT) and Immunohaematology and Transfusion Service of Pistoia Hospital, Piazza Giovanni XXIII, Pistoia (AT)

Correspondence to Prof. Ivano Bertini, PhD, Magnetic Resonance Center (CERM), University of Florence, Via L.Sacconi 6, 50019 Sesto Fiorentino, Italy. E-mail bertini@cerm.unifi.it, phone 39-055-4574270; fax 39-055-4574271

Total word: 5971/6000

Short title: The metabonomics signature of cardiovascular risk

Subject Codes: 135, 140

Abstract

Background: The identification and the present wide acceptance of cardiovascular risk factors such as age, sex, hypertension, hyperlipidemia, smoking, obesity, diabetes, and physical inactivity have led to dramatic reductions in cardiovascular morbidity and mortality. However, novel risk predictors present opportunities to identify more patients at risk and to more accurately define the biochemical signature of that risk. **Method and Results:** In this paper we present a comprehensive metabolic analysis of 864 plasma samples from healthy volunteers, from which we have inferred a pattern of metabolic alteration that correlates with accepted cardiovascular risk factors. From this pattern a novel link between classical cardiovascular risk indicators and oxidative stress of hepatic cell mitochondria seems to emerge. **Conclusions**: A metabolic signature of cardiovascular risk exists, and its analysis may provide novel insight into the associated biochemical pathways and the current prevention therapies.

Key Words: metabolism, risk factors, spectroscopy, cardiovascular diseases.

Introduction

Cardiovascular diseases (CVDs) are the most common cause of death in the developed world. Each year, heart disease kills more Americans than cancer[1]. The morbidity, mortality, and socioeconomic importance of CVDs make timely accurate diagnosis and cost-effective management of the utmost importance.

In most people, CVD stem from the combined effect of a number of causal risk factors. For this reason, risk estimation systems have been developed to assist clinicians to assess the effects of risk factor combinations in planning management strategies. Since it was first proposed in New Zealand in 1993, estimating cardiovascular risk has become an integral part of determining the eligibility for treatment with aspirin, antihypertensives or statins[2].

To date, the Framingham cardiovascular equation remains the most widely used method of assessing cardiovascular risk. This equation was derived from a large cohort study in the United States. It determines cumulative incidence (risk) of any vascular event in individuals free from cardiovascular disease, using a combination of predictors such as age, sex, systolic blood pressure, smoking status, total cholesterol level and high-density lipoprotein (HDL) cholesterol level and diabetic status[3]. Cardiovascular risk estimation is therefore a diagnostic test that has become a part of routine clinical practice.

The major and independent risk factors for CVD are cigarette smoking, elevated blood pressure, elevated serum total cholesterol and low-density lipoprotein cholesterol (LDL), low serum HDL cholesterol, diabetes mellitus, and advancing age. The major risk factors are additive in predictive power. Accordingly, the total risk of a person can be estimated by a weighted sum of the risk imparted by each of the major risk factors. Other factors are associated with increased risk for CVD. These are of two types: conditional risk factors and predisposing risk factors. The conditional risk factors (elevated serum triglycerides, elevated serum homocysteine, elevated serum lipoprotein, prothrombotic factors (eg, fibrinogen), inflammatory markers (eg, C-reactive protein)) are associated with increased risk for CVD although their causative, independent, and quantitative contributions to CVD have not been well documented. The predisposing risk factors (obesity, abdominal obesity, physical inactivity, family history of premature heart disease, ethnic characteristics and psychosocial factors) are those that aggravate the independent risk factors. The adverse effects of obesity are worsened when it is expressed as abdominal obesity, an indicator of insulin

resistance[4]. Abdominal obesity and insulin resistance (two significant predisposing factors) are a manifestation of the so called “metabolic syndrome” which in turn is a key in the evolution of cardiovascular disease. There is debate regarding whether obesity or insulin resistance is the cause of the metabolic syndrome or if they are consequences of a more far-reaching metabolic derangement. A number of markers of systemic inflammation, including C-reactive protein, are often increased, as are fibrinogen, interleukin 6 (IL–6), tumor necrosis factor-alpha (TNFα), and others. Some have pointed to a variety of causes including increased uric acid levels caused by dietary fructose[4].

Until now, no exhaustive metabolomic investigation of cardiovascular risk factors or, somewhat ironically, of the metabolic syndrome as such, has been reported, despite the major risk factors are essentially altered blood metabolites. The only exception is given by blood pressure that is already investigated by Nicholson at al[6]. However, because all of these alterations are of metabolic nature, we hypothesized that metabolomics would be an effective tool for a deeper investigation of the molecular mechanism and biochemical pathways involved in the composition of cardiovascular risk.

Metabolomics has established itself as a useful complement to the characterization of pathologies. The metabolome can be considered the downstream of genome, transcriptome and proteome, and it is the best representation of a healthy or diseased phenotype of an organism. The metabolome, in fact, amplifies changes[7-9] caused by a biological perturbation. As opposed to metabolomics, which places a greater emphasis on comprehensive metabolic profiling, metabonomics is more often used to describe multiple (but not necessarily comprehensive) metabolic changes caused by a biological perturbation. Metabonomics has been defined as “the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification”[10].

Nuclear Magnetic Resonance (NMR)-based metabonomics offers evident advantages in contrast with knowledge-guided search of metabolites. NMR-based metabonomics makes no assumptions on the identity of the metabolites that are relevant for the selected pathology. Information on the metabolite pattern alterations that can be significantly associated to the pathology is directly obtained through statistical analysis of the NMR profiles. Usually, metabonomics does not rely on the measurement of a single metabolite-associated peak(s) but analyze the spectrum as

whole: in principle, an NMR profile contains qualitative and quantitative information of thousands different small molecules present in the sample[11]. Small changes in enzyme concentrations can be reflected in considerable alterations in intermediate products. In addition, those metabolic networks are connected by few high crowded nodes, which can be investigated by NMR-based metabonomic analysis of biological fluids such as serum, plasma, and urine.

NMR-based metabonomic studies have provided significant information on a wide range of pathologies, like cancer[12-14], meningitis[15], neurological disorders[16], diabetes[17], as well as a variety of gastrointestinal diseases[18-20].

In this study, we have examined the metabolic profile of 864 healthy volunteers with the aim of investigating the metabolic alterations characteristic of an increased cardiovascular risk.

Methods

Volunteers Recruitment

The volunteers participating in this study were recruited by the Tuscanian section of the Italian Association of Blood Donors (AVIS). In total 864 adult healthy volunteers (678 males, 186 females, mean age 40.87 ± 11.0 yrs, **Table 1**), participated in this study.

Plasma samples were obtained from the Transfusion Service of Pistoia Hospital, following the Standard Operating Procedures (SOPs) defined for blood donation in Italy. Each subject had fasted overnight, and blood samples were collected in the morning preprandial. The SOPs require that a separate backup tube of plasma from donors is stored at +4 °C during the following 24 hours, in case of necessity of further clinical analysis. Normally these tubes are disposed after that time. These tubes where instead collected and stored at -80 °C for our metabolic analyses.

Written informed consent was obtained from all volunteers.

NMR Samples Preparation

Frozen plasma samples were thawed at room temperature and shaken before use. 300 µl of a phosphate sodium buffer were added to 300 µl of each plasma sample, and 450 µL of this mixture were transferred into a 4.25 mm NMR tube (Bruker BioSpin srl) for analysis. Full details are reported in Supplemental Material.

NMR Analysis and spectral processing

One-dimensional ^{1}H-NMR spectra of plasma samples were measured on a spectrometer operating at 600MHz proton Larmor frequency using standardized protocols (see Supplemental Material). Each spectra in the region 10.00-0.02 ppm was segmented into 0.02-ppm chemical shifts bins prior any statistical analysis.

Statistical Analysis

The statistical procedure used is a combination of Partial Least Square, followed by Canonical Correlation Analysis (CA), followed by Support Vector Machines (SVM) on the CA scores. SVM are used both for classification and for regression.

The accuracy for classification and regression was assessed by means of a double cross-validation scheme[21]. From this procedure we obtain unbiased error estimations for both classification and regression. For regression we reported also the R^2 between the true clinical values and the mean cross-validated predicted values for each sample.

For classification purposes the groups are created by dividing samples in two classes represented by the highest and the lowest quintiles, respectively, relative to the target variable: HDL <44 and >63 (42 vs. 44 individuals, respectively), LDL <100 and >160 (42 vs. 44 samples), TC <174 and >232 (153 vs.153), TC/HDL ratio <3 and >4.90 (40 vs. 40), triglycerides <60 and >131 (150 vs. 151), glycaemia <78 and >105 (148 vs. 153), Framingham score <1.25 and >7.00 (40 vs. 40). For regression analyses all the available samples were used: 206 for HDL, 199 for LDL, 715 for TC, 205 for TC/HDL ratio, 709 for triglycerides, 693 for glycaemia and 201 for Framingham score. Different numbers of individuals in each class are due to the fact that not for all individuals all clinical parameters were available.

To assess which buckets (that is, resonance peaks) were significantly different between different groups a univariate wilcoxon test was used. A *P*-value ≤ .001 (with Bonferroni correction for multiple tests) was considered statistically significant.

All calculations were made using home-made scripts written in R language[22].

Further references and information on statistical methods are reported in Supplemental Material.

Results

We have performed several through statistical analyses in order to ascertain whether metabolic fingerprint of the plasma samples contain predictive information

about the global status of the individuals. As a first inspection we have done a classification analysis, using as predictor the full NMR profile, with respect to the total cholesterol (TC), LDL cholesterol, HDL cholesterol, triglycerides, and glycaemia. The results are shown in **Tables 2-8** for CPMG and noesy spectra and in **Supplemental Tables S1-S6** for diffusion edited spectra. Values of sensitivity, specificity and accuracy, obtained from a non-biased double cross-validation procedure[21], are reported. As we can see, all the accuracy values are very good, always higher than 95% for all classifications and kind of 1-D spectra (CPMG, NOESY, diffusion edited), except for glycaemia, with accuracy around 85%. The separation obtained plotting the CA scores is also very clear (**Figure 1**). With these encouraging results, a more sophisticated regression analysis was also attempted, with the aim of establishing a quantitative relationship between clinical data and NMR spectra. **Figures 2** and **3** show the agreement between the predicted values of HDL and LDL cholesterol with the "true" values given by the clinical analysis. The mean relative error is less than 10% in both cases. These striking results are not completely unexpected, because the NMR spectra (especially NOESY and diffusion edited) contain several strong peaks directly related to the lipidic components of the plasma. To confirm that metabolomics is able to extract also more subtle metabolic information, we tested whether we could classify individuals with high and low values of TC or glycaemia even after removing from the spectra all signals arising from lipids or glucose, respectively. The good accuracy obtained (75% in both case) is a convincing indicator that the changes in the blood composition are global phenomena that involve several metabolic alterations, and that metabolomics is able to outline all these effects simultaneously.

As it is well known that low levels of HDL cholesterol and high levels of TC correlated to high risk of developing CVDs, we performed the same analyses also for the TC/HDL ratio. Again the results, both for classification and regression, are extremely good, with accuracy for classification even higher than for the two risk factors taken separately (**Table 7**), and R^2 of 0.88 for regression (**Figure 4**). Analogously good values are obtained for the HDL/LDL ratio.

Finally, having in hands the clinical data of the subjects, we decided to calculate also the well established Framingham score for each individual, using the equation of Wilson et al[3]. This score combines age, sex, systolic blood pressure, smoking status, TC level, HDL level and diabetic status in order to derive a probability of major

cardiovascular event in the following ten years. A classification was again performed by splitting the data into the quintiles with the highest score and the quintiles with the lower score. The accuracy is high (over 89% with the NOESY spectra) but lower than in the previous analysis (**Table 8**). This may due to the fact that the Framingham risk score considers also non-analytical variables, but it can derive also from the fact that we are able to extract a far richer and more complex metabolic pattern.

These findings confirm that the metabolomic approach to the study of blood composition is feasible and provides a comprehensive picture of the underlying multifactorial mechanisms arising from the interactions between genomics and environment.

In order to identify the metabolites that are most characteristic of the metabolic fingerprint of cardiovascular risk we apply a univariate wilcoxon test on the different groups, using a significance level of $P<.001$, after applying a Bonferroni correction. We find that individuals with high plasma levels of HDL are characterized by lower levels of creatinine and threonine, and by higher levels of 3-hydroxybutirrate (3-HB). Individuals with high plasma levels of LDL are characterized by lower levels of α-ketoglutarate and dimethylglycine. Individuals with high levels of TC are characterized by lower levels of α-ketoglutarate, dimethylglycine and serine. Individuals with high levels of triglycerides have a more complex pattern, with lower levels of several metabolites: acetate, creatine, creatinine dimethylglycine, 3-HB, isoleucine, methionine, phenylalanine and serine, and higher levels of threonine. Because 3-HB signals in 1-D spectra are partially covered by other signals, the significance value of 3-HB has been also confirmed using J-RES projection were this signal is well resolved.

The complete chart of the most significant metabolites found is reported in **Table 9**. Because our aim is to find a metabolic pattern that correlates with an increase of cardiovascular risk we performed the same analysis only considering subjects with high level of HDL and low levels of LDL cholesterol (i.e. the “lowest risk” people), ordered by the HDL/LDL ratio. This metabolic pattern is strongly characterized by lower levels of creatinine and threonine, and by higher levels of 3-HB.

Discussion

The statistical results obtained by classification and regression analysis are extremely encouraging, suggesting that the whole plasma spectrum profile contains

meaningful information about the metabolic status of individuals. Also CPMG spectra (that essentially do not contain peaks associated with lipoproteins) are able to predict with high accuracy the clinical data. This reinforces the view that we are effectively following global metabolic changes and not only obvious patterns of lipids and lipoproteins peaks. In order to prove this statement we also tried to completely remove spectral regions with peaks associated with lipid fractions (for TC analysis) and with glucose (for glycaemia analysis). Although the accuracy decreases around 75% (both for TC and glycaemia), it is still highly significant, especially if we consider that we have removed more than 100 spectral buckets (25% of the total) in order to be on the safe side.

These results allow us to explore the metabolite composition of plasma profiles, in order to characterize what is typical of an undoubtedly a “low risk” status (high HDL levels, low LDL levels, low triglycerides and low glycaemia) and what is typical of a high risk status (the reverse). From the point of view of this comprehensive metabolic analysis it seems that the “low risk” pattern is mostly constituted by high levels of 3-HB and low levels of threonine and creatinine. On the other hand the “high risk” pattern is more complex, and is characterized by low levels of several metabolites, in particular α-ketoglutarate, dimethylglycine and serine.

Having in hand the results of metabolite profiling one might already attempt to elaborate new hypothesis about the origin of the metabolic syndrome and its connection with cardiovascular risk. In particular our findings seem to highlight the role of 3-HB, α-ketoglutarate, threonine and dimethylglycine as the key to disentangle the fingerprint associated to high or low cardiovascular risk. All these metabolites are related to the mitochondrial activity.

3-HB is generated by fatty acid beta-oxidation, a process involving a mitochondrial oxido-reductase superfamily, with fatty acid-CoA thioesters as substrates. The enzyme 3-hydroxybutyrate dehydrogenase maintains the equilibrium between 3-HB and acetoacetate. The equation is

$$\text{3-hydroxybutyrate} + NAD^+ \leftrightarrows \text{acetoacetate} + NADH + H^+$$

It is known that the ketone bodie 3-HB has a protective role in a broad spectrum of cerebral injuries and diseases and they preserve neuronal cell integrity and stability in vitro[23-25]. From these observations it was inferred that mitochondrial

dysfunction plays a significant role in common neurological diseases[26]. In the same way our data show that high levels of circulating 3-HB correlate with a high HDL/LDL ratio that, in turn, is related to a reduced risk of cardiovascular event. Analogously, the role of mitochondria can be invoked. The acetoacetate to 3-hydroxybutyrate ratio is a good indicator of the mitochondrial NAD+/NADH redox ratio in the mitochondria of hepatic cells[27,28]. Usual values of the ratio in healthy subjects are highly variable and ranges between 0.96 to 2.55[29].

In our findings, because the level of acetoacetate seems not to vary as a function of the HDL/LDL ratio, but 3-HB show a positive correlation, the final ratio between the acetoacetate and 3-HB decreases in the case of the “low risk” profile. Although the use of ketone bodies ratio is not new in the literature (expecially for monitoring the mitochondrial activity in the liver of patients with hepatic failure[30]), at the best of our knowledge this is the first time that this figure is related with LDL and HDL levels in the blood and, in turn, with metabolic syndrome and enhanced risk of cardiovascular disease. We can hypothesize that oxidative stress enhance the oxidation of 3-HB to acetoacetate, reducing its plasma level.

The mechanism that correlates “low risk” plasma parameters with high levels of 3-HB may be related to the ability of 3-HB to bind the HM74A receptor expressed in adipocytes, resulting in inhibition of the lipolysis and thereby reducing free fatty acids plasma levels. This receptor is know to bind nicotinic acid, the first drug available for lowering LDL levels while, concomitantly, raising the HDL levels. Although the clinical use of nicotinic acid goes back to the fifties[31], remarkably, its putative mechanism of action has only been recently proposed[32]. This receptor responds to both nicotinic acid and the ketone body 3-HB, and the latter is thought to be the probable endogenous ligand for HM74A. Hence the occupancy of HM74A receptor by circulating 3-HB could represent a physiological feedback cycle for the control of lipolysis rate[33].

α-ketoglutarate is a key intermediate in the Krebs cycle, coming after isocitrate and before succinyl CoA. The cycle can be filled at this juncture by synthesizing α-ketoglutarate through the action of glutamate dehydrogenase on glutamate.

The first information about the influence of α-ketoglutarate on blood serum cholesterol was presented by Bazzano & Bazzano[34] in 1972. The authors

documented a 66% lower level of cholesterol in the blood serum of male Mongolian gerbils treated with α-ketoglutarate. Additionally, they observed a 15 % decrease in body weight in animals fed with an experimentally prepared diet containing α-ketoglutarate. More recently[35] it was proved that oral treatment with α-ketoglutarate can decrease the risk of hypercholesterolaemia developing, decrease the blood serum concentration of the LDL cholesterol, increase the concentration of HDL cholesterol, and can lower the body weight.

Because, in our findings, low levels of α-ketoglutarate are associated with increased cardiovascular risk, we can hypothesise an imbalance of the concentrations of the redox substrates of glutamate dehydrogenase due to some alteration of the mitochondrial redox potential.

In addition to these considerations, it is interesting to note that most of the aminoacids concentrations are lower with high levels of triglycerides, with the exception of threonine.

Intriguingly, the catabolism of threonine is regulated by threonine deydrogenase (THD), which is a mitochondrial enzyme which catalyzes NAD^+-dependent oxidation of threonine to 2-amino-3-oxobutyric acid, which in turn is very rapidly metabolized by 2-amino-3-oxobutyrate-CoA ligase to glycine and acetyl-CoA[36]. The formation of acetyl-CoA from ligase is so fast that acetyl-CoA in practice is considered a product of TDH[36]. Acetyl-CoA is derived from TDH, but also from fatty acids by β-oxidation. Excessive formation of acetyl-CoA by TDH would inhibit both the β-ketothiolase reaction and β-oxidation by the law of mass action, playing a biological role in the regulation of fatty acid breakdown, biosynthesis, and in ketone bodies formation. A fine regulation balances threonine concentration under different conditions and TDH is the target of selective, feedback inhibition by all compounds derived from its major product, acetyl-CoA and, remarkably, by 3-HB[36].

Dimethylglycine is a tertiary amino acid, which is an intermediary metabolite in the choline pathway, as it is formed in liver mitochondria by removal of one methyl group from betaine[37]. It can be metabolized in liver mitochondria, rendering both its methyl groups through transmethylation of tetrahydrofolate, and free glycine[38,39]. Recently it was found that a dietary supplement of dimethylglycine in chickens diet significantly reduces the deposition of dietary fat into abdominal depot tissue and the fasted non-

esterified fatty acid plasma level significantly decreased with increasing dimethylglycine level[40]. Coherently, we found that patients with high levels of triglycerides have lower levels of plasma dimethylglycine.

It is hypothesised in[41] that a mitochondrial dysfunction can be the starting point for the development of insulin resistance in adolescents with familiar history of type-2 diabetes, reporting that rates of mitochondrial substrate oxidation were decreased by 30% in lean, insulin-resistant offspring compared with insulin-sensitive control subjects. This observation can be easily linked to our findings, moving the starting point of obesity, insulin resistance, and of the risk of developing cardiovascular injury to the mitochondrial activity.

Although a full clarification of a so complex machinery is far away, we can suggest that an imbalance in the mitochondrial redox potential (caused by oxidative stress), can lead to an alteration of the fatty acids beta-oxidation pathway with an increased secretion of LDL lipoproteins in the blood. We can also hypothesize that this mechanism amplify itself by the complementary action of the discussed metabolites, as show in **Figure 5**.

Finally, the good agreement with the calculated values of Framingham score supports the idea that metabolomic analysis of plasma samples can be an effective tool for monitoring the individual healthy status, and, in particular the risk of developing metabolic syndrome and cardiovascular disease. The mismatches between our prediction and the true Framingham values can be due to our ability to unravel subtle metabolic changes, which are not included in the classical risk scores. Our observations strongly suggest that metabolomic analyses should be included in future year-long prospective studies and new (and possibly more efficient) cardiovascular score will be formulated on the basis of the whole metabolic fingerprint and/or on the basis of the plasma concentration of the discussed metabolites.

Acknowledgments

We are very grateful to AVIS Toscana (in particular to the President Luciano Franchi and Donata Marangio), AVIS Pistoia (in particular to Alessandro Pratesi) and to technical staff of Transfusion Service for supporting us in this project.

Funding Sources

This work was partly supported by a fellowship from Boehringer Ingelheim Italia (to L.T. through the FiorGen Foundation).

Disclosures

None.

References

1. National Vital Statistics Report, Vol. 50, No. 15, 2002, “Deaths: final data for 2000”
2. Jackson R, Barham P, Maling T, MacMahon S, Bills J, Birch B et al. The management of raised blood pressure in New Zealand. *Br Med J.* 1993;307:107–110.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation*. 1998;97:1837-1840.
4. Grundy SM, Pasternak R, Greenland P, Smith S Jr, Fuster V. Assessment of Cardiovascular Risk by Use of Multiple-Risk-Factor Assessment Equations A Statement for Healthcare Professionals From the American Heart Association and the American College of Cardiology. *J Am Coll Cardiol.* 1999;34(4):1348-1359.
5. Reiser S, Powell AS, Scholfield DJ, Panda P, Ellwood KC, Canary JJ. Blood lipids, lipoproteins, apoproteins, and uric acid in men fed diets containing fructose or high-amylose cornstarch. *Am J Clin Nutr*. 1989;49(5):832–839.
6. Holmes E, Loo RL, Stamler J, Bictash M, Yap IK, Chan Q, Ebbels T, De Iorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* 2008;453(7193):396-400.
7. Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol*. 2001;19(1):45-50.

8. Mendes P, Kell DB, Westerhoff HV. Why and when channelling can decrease pool size at constant net flux in a simple dynamic channel. *Biochim Biophys Acta.* 1996;1289(2):175-186.
9. Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR. Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* 2003; 4(10):989-993.
10. Nicholson, JK, Lindon, JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica.* 1999;29:1181-1189.
11. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 2007;35:D521-526.
12. Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q, Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC, Berger A, Shuster JR, Wei JT, Varambally S, Beecher C, Chinnaiyan AM. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature.* 2009;457(7231):910-914.
13. Claudino WM, Quattrone A, Biganzoli L, Pestrin M, Bertini I, Di Leo A. Metabolomics: available results, current research projects in breast cancer, and future applications. *J Clin Oncol.* 2007;25(19):2840-2846.
14. Oakman C, Tenori L, Biganzoli L, Santarpia L, Cappadona S, Luchinat C, Di Leo A. Uncovering the metabolomic fingerprint of breast cancer. *Int J Biochem Cell Biol.* 2010;[Epub ahead of print]
15. Coen M, O'Sullivan M, Bubb WA, Kuchel PW, Sorrell T. Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clin Infect Dis.* 2005;41(11):1582-1590.

16. Bartsch T, Alfke K, Wolff S, Rohr A, Jansen O, Deuschl G. Focal MR spectroscopy of hippocampal CA-1 lesions in transient global amnesia. *Neurology.* 2008;70(13):1030-1035.
17. Lanza IR, Zhang S, Ward LE, Karakelides H, Raftery D, Nair KS. Quantitative metabolomics by ^{1}H-NMR and LC-MS/MS confirms altered metabolic pathways in diabetes. *PLoS One*. 2010;5(5):e10538.
18. Bertini I, Calabrò A, De Carli V, Luchinat C, Nepi S, Porfirio B, Renzi D, Saccenti E, Tenori L. The metabonomic signature of celiac disease. *J Proteome Res.* 2009;8(1):170-777
19. Griffin JL, Scott J, Nicholson JK. The influence of pharmacogenetics on fatty liver disease in the wistar and kyoto rats: a combined transcriptomic and metabonomic study. *J Proteome Res*. 2007;6(1):54-61.
20. Marchesi JR, Holmes E, Khan F, Kochhar S, Scanlan P, Shanahan F, Wilson ID, Wang Y. Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. *J Proteome Res*. 2007;6(2):546-51.
21. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven, JPM, van Dorsten FA. Assessment of PLSDA cross validation. *Metabolomics*. 2008;4:81-89.
22. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J *Comput Stat Graph*. 1996;5:299-314.
23. Izumi Y, Ishii K, Katsuki H, Benz AM, Zorumski CF. beta-Hydroxybutyrate fuels synaptic function during development. Histological and physiological evidence in rat hippocampal slices. *J Clin Invest*. 1998;101:1121-1132.
24. Imamura K, Takeshima T, Kashiwaya Y, Nakaso K, Nakashima K. D-beta-hydroxybutyrate protects dopaminergic SH-SY5Y cells in a rotenone model of Parkinson's disease. *J Neurosci Res* 2006;84:1376-1384.
25. Kashiwaya Y, Takeshima T, Mori N, Nakashima K, Clarke K, Veech RL. D-beta-hydroxybutyrate protects neurons in models of Alzheimer's and Parkinson's disease. *Proc Natl Acad Sci.* USA 2000;97:5440-5444.
26. Cornille E, Abou-Hamdan M, Khrestchatisky M, Nieoullon A, de Reggi M, Gharib B. Enhancement of L-3-hydroxybutyryl-CoA dehydrogenase activity and circulating ketone body levels by pantethine. Relevance to dopaminergic injury. *BMC Neurosci.* 2010;11:51-61.

27. Buss GD, Constantin J, de Lima LC, Teodoro GR, Comar JF, Ishii-Iwamoto EL, Bracht A. The action of quercetin on the mitochondrial NADH to NAD(+) ratio in the isolated perfused rat liver. *Planta Med.* 2005;71(12):1118-1122.
28. Tani T, Taki Y, Aoyama H, Jikkoh A, Arii S, Ozawa K, Tobe T. Changes in acetoacetate/beta-hydroxybutyrate ratio in arterial blood following hepatic artery embolization in man. *Life Sci.* 1984; 35(11):1177-1182.
29. Iwata S, Ozawa K, Shimahara Y, Mori K, Kobayashi N, Kumada K, Yamaoka Y. Diurnal fluctuations of arterial ketone body ratio in normal subjects and patients with liver dysfunction. *Gastroenterology.* 1991;100(5 Pt 1):1371-1378.
30. Saibara T, Maeda T, Onishi S, Yamamoto Y. Plasma exchange and the arterial blood ketone body ratio in patients with acute hepatic failure. J *Hepatol.* 1994; 20(5):617-622.
31. Altschul R, Hoffer A, Stephen JD. Influence of nicotinic acid on serum cholesterol in man. *Arch Biochem* 1955;54:558–559.
32. Tunaru S, Kero J, Schaub A, Wufka C, Blaukat A, Pfeffer K, Offermanns S. PUMA-G and HM74 are receptors for nicotinic acid and mediate its anti-lipolytic effect. *Nat Med.* 2003;9(3):352-355.
33. Soudijn W, van Wijngaarden I, Ijzerman AP. Nicotinic acid receptor subtypes and their ligands. *Med Res Rev.* 2007; 27(3):417-433.
34. Bazzano G, Bazzano GS. Hypocholesterolemic effect of aketoglutarate in the Mongolian gerbil. *Proc Soc Exp Biol Med*. 1972;140:36–39.
35. Radzki RP, Bieńko M, Pierzynowski SG. Effect of dietary alpha-ketoglutarate on blood lipid profile during hypercholesterolaemia in rats. *Scand J Clin Lab Invest.* 2009; 69(2):175-180.
36. Guerranti R, Pagani R, Neri S, Errico SV, Leoncini R, Marinello E. Inhibition and regulation of rat liver L-threonine dehydrogenase by different fatty acids and their derivatives. *Biochim Biophys Acta*. 2001; 1568(1):45-52.
37. Friesen RW, Novak EM, Hasman D, Innis SM. Relationship of dimethylglycine, choline, and betaine with oxoproline in plasma of pregnant women and their newborn infants. *J Nutr.* 2007; 137(12):2641-2646.
38. MacEnzie CG, Frisell WR. The metabolism of dimethylglycine by liver mitochondria. *J Biol Chem.* 1958;232(1):417-427.
39. Slow S, McGregor DO, Lever M, Lee MB, George PM, Chambers ST. Dimethylglycine supplementation does not affect plasma homocysteine

concentrations in pre-dialysis chronic renal failure patients. Clin Biochem. 2004;37(11):974-976.

40. Kalmar ID, Cools A, Verstegen MW, Huyghebaert G, Buyse J, Roose P, Janssens GP. Dietary supplementation with dimethylglycine affects broiler performance and plasma metabolites depending on dose and dietary fatty acid profile. *J Anim Physiol Anim Nutr*. 2010; [Epub ahead of print]
41. Befroy DE, Petersen KF, Dufour S, Mason GF, de Graaf RA, Rothman DL, Shulman GI. Impaired mitochondrial substrate oxidation in muscle of insulin-resistant offspring of type 2 diabetic patients. *Diabetes.* 2007; 56(5):1376-1381.

Figure Legends

Figure 1. Clusterization of high (triangles) and low (circles) quintiles of TC (A), LDL (B), HDL (C), and triglycerides (D) and glycaemia (E) values.

Figure 2. Correlation between true values (x axis) and predicted on the basis of noesy spectra (y axis) for HDL levels. (R^2=0.83)

Figure 3. Correlation between true values (x axis) and predicted on the basis of noesy spectra (y axis) for LDL levels. (R^2=0.88)

Figure 4. Correlation between true values (y axis) and predicted on the basis of noesy spectra (x axis) for TC/HDL levels. (R^2=0.88)

Figure 5. Metabolic effect of the discussed metabolites in high risk subjects.

Table 1. Average clinical data for recruited volunteers.

Female = 186	Mean	St. Deviation	Max	Min
Age	41.81	11.92	65	19
Glycaemia	87.85	13.94	175	45
Creatinine	0.773	0.108	1.14	0.54
Cholesterol	211.8	39.04	366	139
Triglycerides	85.15	54	556	39
HDL	65.08	13.46	98	43
LDL	134	40.63	219	69
Blood pressure Max	119.1	10.48	160	110
Blood pressure Min	77.77	7.05	100	65
Cardiac Frequency	71.74	5.47	88	52

Male = 678	Mean	St. Deviation	Max	Min
Age	40.62	10.8	71	19
Glycaemia	90.14	12.54	226	65
Creatinine	0.97	0.124	1.54	0.32
Cholesterol	201.8	38.05	339	75
Triglycerides	106.87	64.53	635	30
HDL	51.6	11.35	99	29
LDL	129.86	36.39	231	42
Blood pressure Max	124.4	10.71	160	100
Blood pressure Min	81.44	7	100	60
Cardiac Frequency	69.66	6.14	90	48

Table 2. Confusion matrices for TC classification.

CPMG	Low Values (< 174)	High Values (> 232)	N = 306 (153 vs 153)	
Low	96.13%	3.87%		Sensitivity
High	3.22%	96.78%		Specificity
			96.50%	Accuracy

NOESY	Low Values (< 174)	High Values (> 232)	N = 306 (153 vs 153)	
Low	97.60%	2.40%		Sensitivity
High	1.94%	98.06%		Specificity
			98.07%	Accuracy

Table 3. Confusion matrices for LDL classification.

CPMG	Low Values (< 100)	High Values (> 160)	N = 85 (42 vs 43)	
Low	99.01%	0.99%		Sensitivity
High	3.27%	96.73%		Specificity
			98.30%	Accuracy

NOESY	Low Values (<100)	High Values (>160)	N = 85 (42 vs 43)	
Low	100%	0%		Sensitivity
High	2.04%	97.96%		Specificity
			98.94%	Accuracy

Table 4. Confusion Matrices for HDL classification.

CPMG	Low Values (< 44)	High Values (> 63)	N = 82 (42 vs 40)	
Low	92.87%	7.13%		Sensitivity
High	5.14%	94.86%		Specificity
			92.80%	Accuracy

NOESY	Low Values (<44)	High Values (>63)	N = 82 (42 vs 40)	
Low	98.87%	1.33%		Sensitivity
High	1.15%	98.85%		Specificity
			98.81%	Accuracy

Table 5. Confusion Matrices for triglycerides classification.

CPMG	Low Values (< 60)	High Values (> 131)	N = 351 (150 vs 151)	
Low	97.77%	2.23%		Sensitivity
High	3.09%	96.91%		Specificity
			97.48%	Accuracy

NOESY	Low Values (< 60)	High Values (> 131)	N = 351 (150 vs 151)	
Low	98.18%	1.82%		Sensitivity
High	2.95%	97.05%		Specificity
			97.37%	Accuracy

Table 6. Confusion Matrices for glycaemia classification.

CPMG	Low Values (<81)	High Values (>96)	N = 301 (148 vs 153)	
Low	85.89%	14.11%		Sensitivity
High	14.68%	85.32%		Specificity
			85.65%	Accuracy

NOESY	Low Values (<81)	High Values (>96)	N = 301 (148 vs 153)	
Low	76.8%	23.2%		Sensitivity
High	20.8%	79.2%		Specificity
			78.02%	Accuracy

Table 7. Confusion Matrices for TC/HDL ratio classification.

CPMG	Low Values (< 3)	High Values (> 4.90)	N = 80 (40 vs 40)	
Low	98.15%	1.85%		Sensitivity
High	1.06%	98.94%		Specificity
			98.19%	Accuracy

NOESY	Low Values (<3)	High Values (>4.90)	N = 80 (40 vs 40)	
Low	100%	0%		Sensitivity
High	2.25%	97.75%		Specificity
			98.69%	Accuracy

Table 8. Confusion Matrices for Framingham risk score classification.

CPMG	Low Values (< 1.25)	High Values (> 7.00)	N = 80 (40 vs 40)	
Low	88.40%	11.92%		Sensitivity
High	11.54%	88.07%		Specificity
			88.21%	Accuracy
NOESY	Low Values (< 1.25)	High Values (> 7.00)	N = 80 (40 vs 40)	
Low	89.17%	10.83%		Sensitivity
High	8.83%	91.17%		Specificity
			89.68%	Accuracy

Table 9. Statistically significant metabolites. Plus or minus signs means that the selected metabolite is high or low where the levels of clinical variabiles are high.

	LDL	TC	TRI	HDL	HDL/LDL
3-HB				+	+
Acetate			-		
Alfa Ketoglutarate	-	-			
Citrate			-		
Creatine			-		
Creatinine			-	-	-
Dimethylglycine	-	-	-		
Glucose			-		
Isoleucine			-		
Methionine			-		
Phenylalanine			-		
Serine		-	-		
Threonine			+	-	-
Tyrosine			-		

Figure 1. Clusterization of high (triangles) and low (circles) quintiles of TC (A), LDL (B), HDL (C), and triglycerides (D) and glycaemia (E) values.

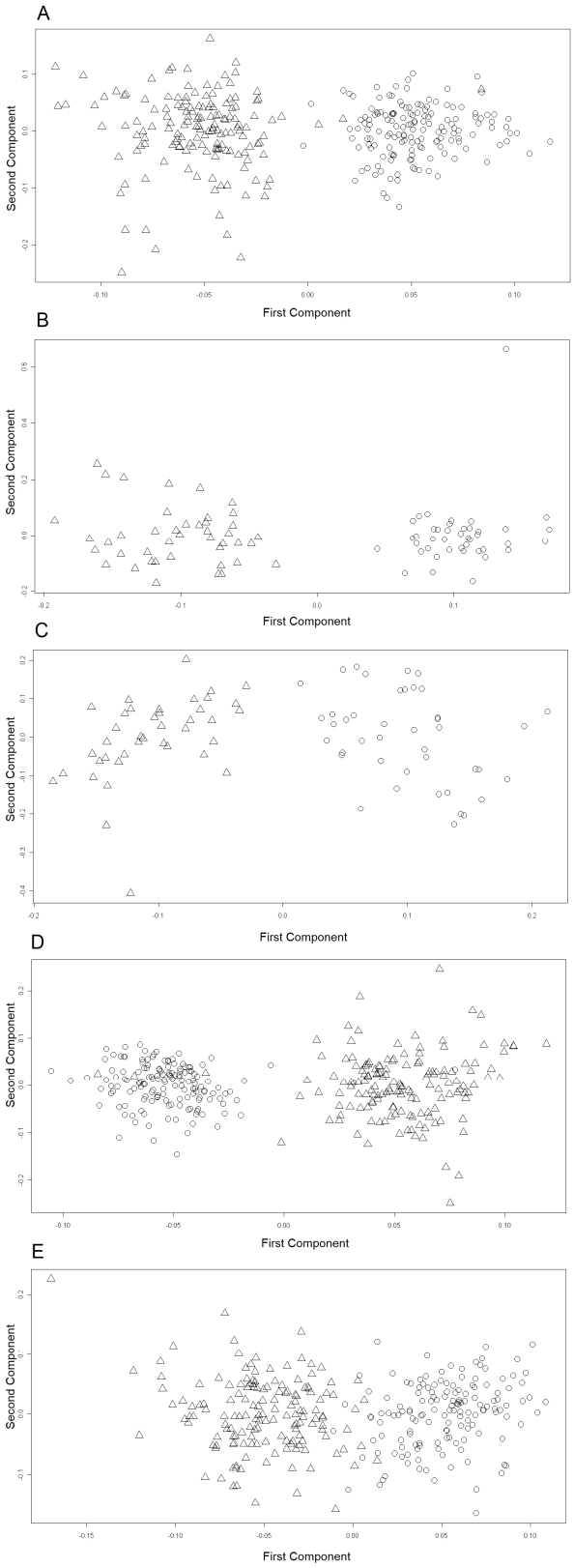


Figure 2. Correlation between true values (x axis) and predicted on the basis of noesy spectra (y axis) for HDL levels. (R^2=0.83)

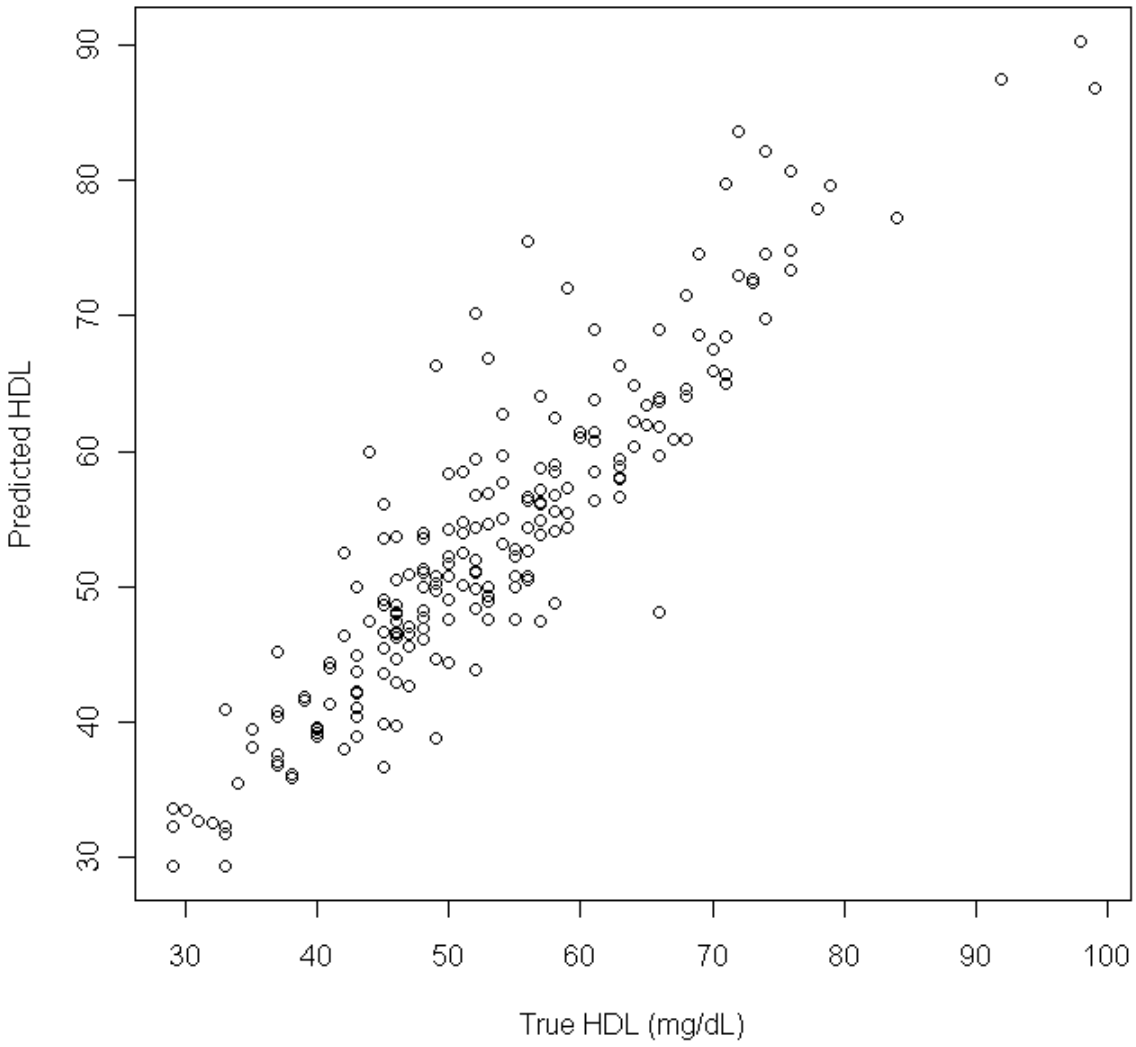


Figure 3. Correlation between true values (x axis) and predicted on the basis of noesy spectra (y axis) for LDL levels. (R^2=0.88)

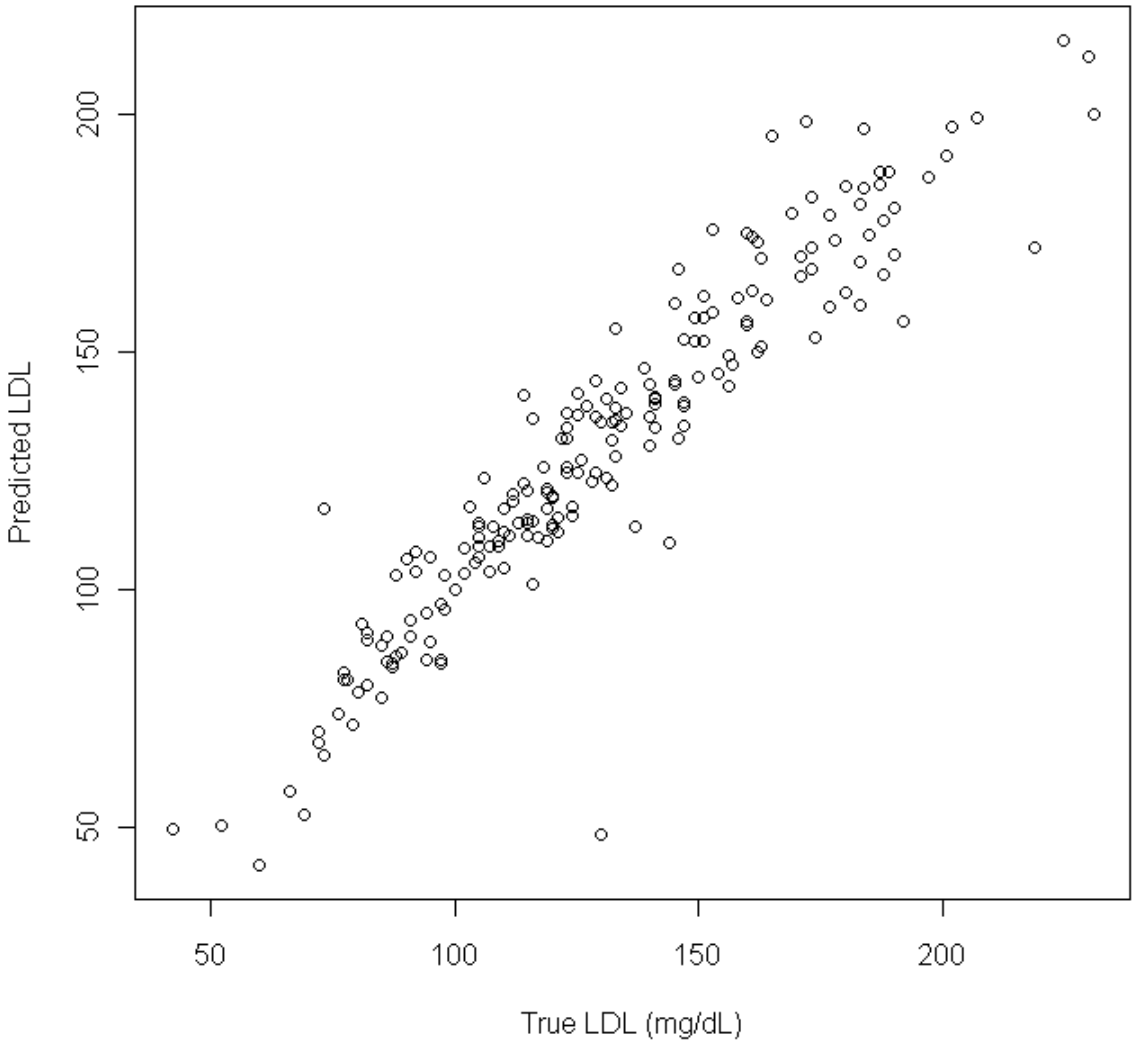


Figure 4. Correlation between true values (y axis) and predicted on the basis of noesy spectra (x axis) for TC/HDL levels. (R^2=0.88)

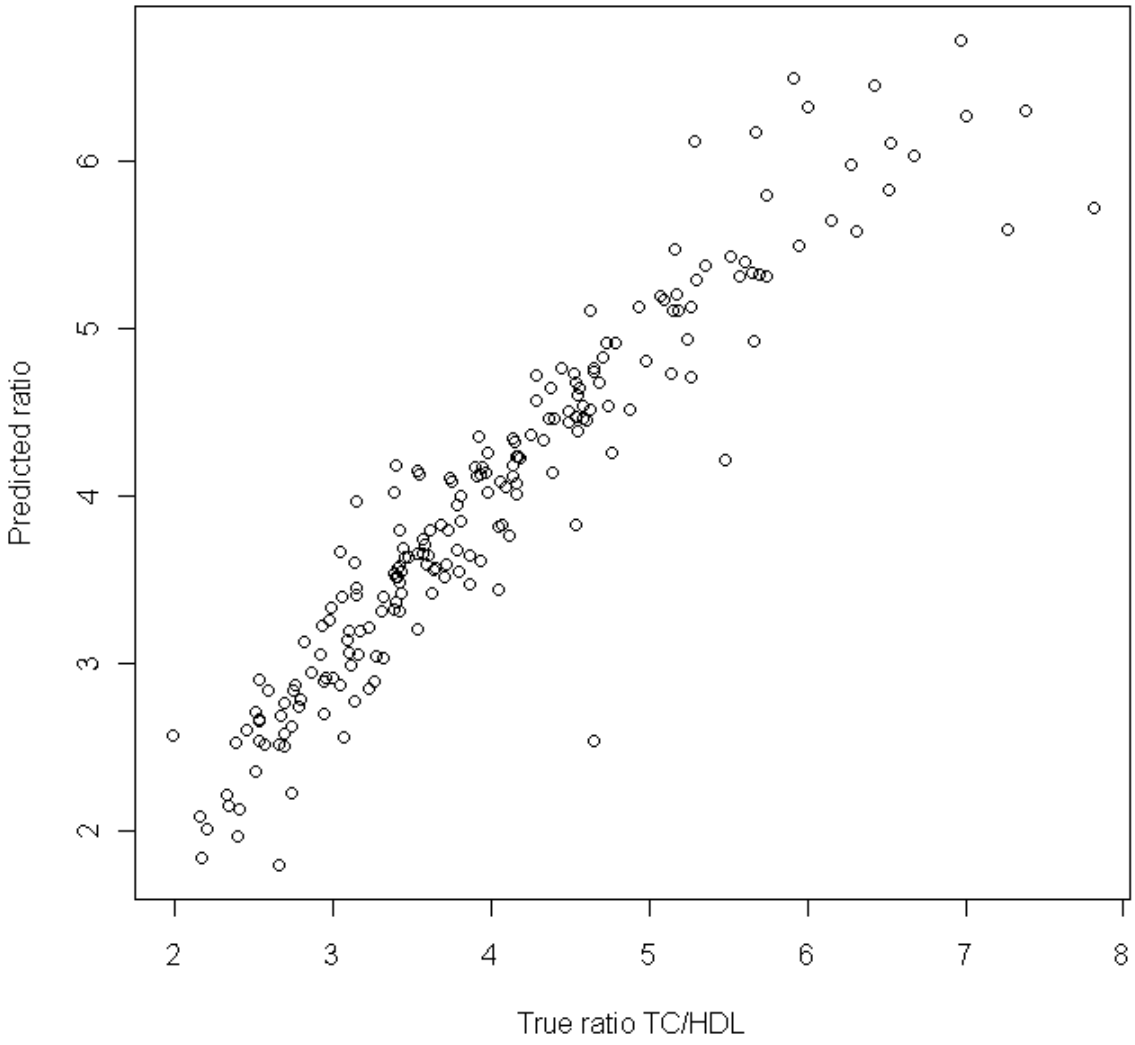
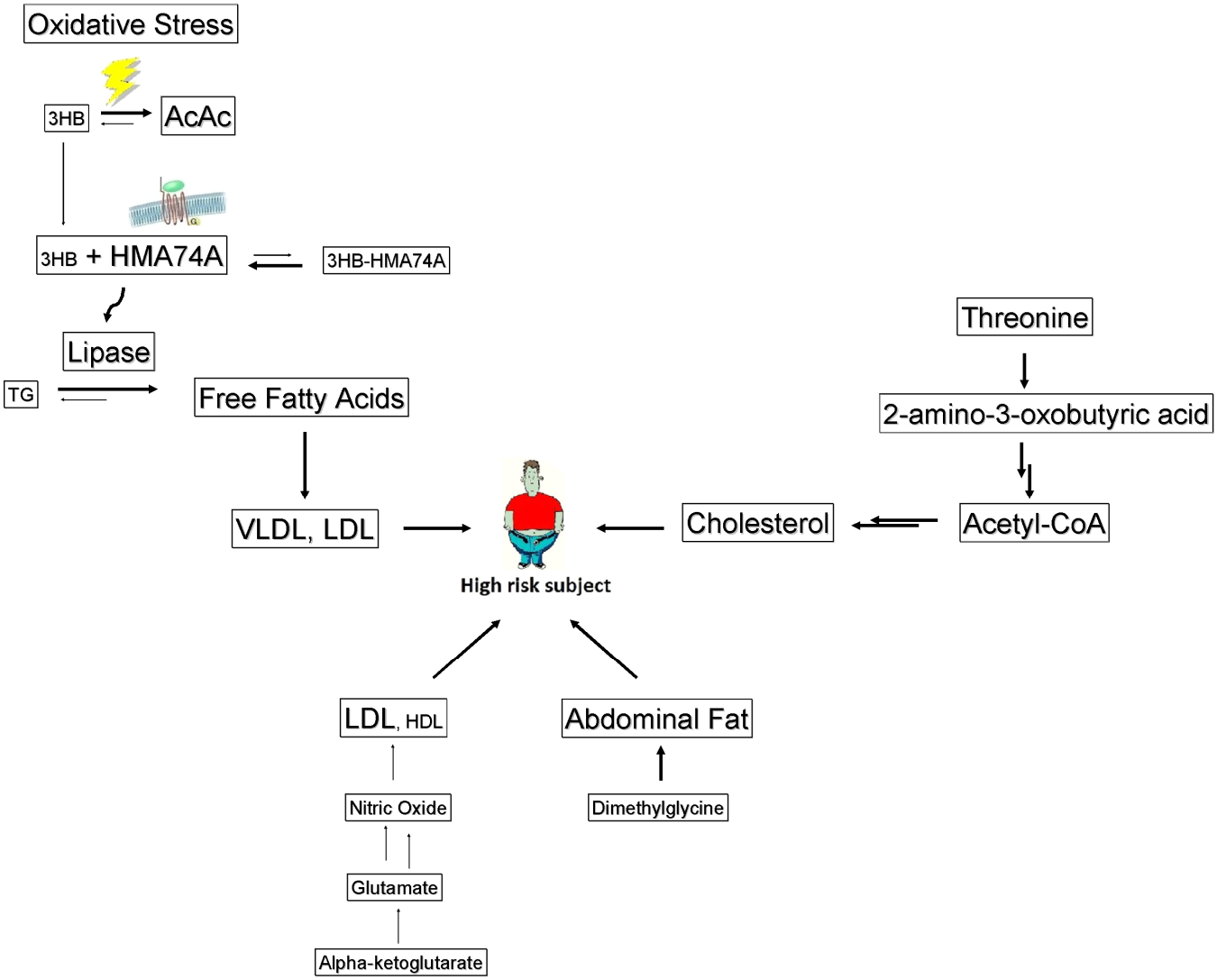


Figure 5. Metabolic effect of the discussed metabolites in high risk subjects.



Supplemental Material

Supplemental Methods

NMR Samples Preparation

Frozen plasma samples were thawed at room temperature and shaken before use. 300 µl of a phosphate sodium buffer (70 mM Na_2HPO_4; 20% v/v ^2H_2O; 0.025% v/v NaN_3; 0.8% w/v sodium trimethylsilyl [2,2,3,3-$^2H_4$]propionate (TSP); pH 7.4) were added to 300 µl of each plasma sample, and the mixture was homogenized by vortexing for 30 s. 450 µL of this mixture were transferred into a 4.25 mm NMR tube (Bruker BioSpin srl) for analysis.

NMR Analysis

1H NMR spectra for all samples were acquired using a Bruker 600 MHz spectrometer (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm CPTCI 1H-13C/31P-2H cryo-probe including a z axis gradient coil, an automatic tuning-matching (ATM) and an automatic sample changer. A BTO 2000 thermocouple served for temperature stabilization at the level of approximately 0.1 K at the sample. Before measurement, samples were kept for at least 3 min inside the NMR probehead, for temperature equilibration (310 K).

One-dimensional (1-D) spectra of plasma samples were acquired using a Carr−Purcell−Meiboom−Gill (CPMG; Bruker) spin−echo sequence to suppress signals arising from high molecular weight molecules, a standard pulse sequence (NOESYpresat), diffusion edited sequence using a diffusion time of 120 ms and fast two-dimensional (2-D) J-Resolved experiments (J-RES).

Spectral Processing

Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated (TMSP peak at 0.00 ppm) using TopSpin (Version 2.1, Bruker). Each 1-D spectrum in the range between 0.02 and 10.00 ppm was segmented into 0.02-ppm chemical shift bins, and the corresponding spectral areas were integrated using AMIX software

(Bruker BioSpin). Regions between 6.0 and 4.5 ppm containing residual water were removed. The total spectral area was calculated on the remaining bins and normalization was carried out on the data prior to statistical analysis. The projection of each 2-D J-RES spectrum was segmented and processed as described below for the 1-D spectra.

Statistical Analysis

Data reduction was obtained by means of Partial Least Square (PLS) on the mean-centred data without scaling, using the classical SIMPLS algorithm[1] as implemented in the R-library "plsgenomics". Canonical Correlation Analysis (CA) was performed using standard R functions. For the purpose of regression and classification we used the Support Vector Machines (SVM) method[2] applied on the CA scores. The "libsvm" module[3] of the R-library "e1071" was used.

The statistical procedure used is a combination of PLS for data reduction (i.e. to reduce the initial number of variable and denoise the data) followed by CA (for further reduction and to have a clearer picture of the discrimination rule) followed by SVM on the CA scores. SVM are used both for classification and for regression ("C-classification" and "eps-regression" settings)

Sensitivity, specificity and accuracy, were estimated using standard definitions[4]. The accuracy for classification and regression was assessed by means of a double cross-validation scheme[5]. The original data set was split in a training set (90%) and a test set (10%) prior to any step of statistical analysis. Parameter selection (best number of PLS components, kernel type and cost of constraints violations for SVM) was carried out by 7-fold cross validation on the 90% training set. The whole procedure was repeated 100 times inside a Monte Carlo cross validation scheme. From this procedure we obtain unbiased error estimations for both classification and regression. For regression we reported also the R^2 between the true clinical values and the mean cross-validated predicted values for each sample.

For classification purposes the groups are created by dividing samples in two classes represented by the highest and the lowest quintiles, respectively, relative to the target variable: HDL <44 and >63 (42 vs. 44 individuals, respectively), LDL <100 and >160 (42 vs. 44 samples), TC <174 and >232 (153 vs.153), TC/HDL ratio <3 and >4.90 (40 vs. 40), triglycerides <60 and >131 (150 vs. 151), glycaemia <78 and >105 (148 vs. 153), Framingham score <1.25 and >7.00 (40 vs. 40). For regression

analyses all the available samples were used: 206 for HDL, 199 for LDL, 715 for TC, 205 for TC/HDL ratio, 709 for triglycerides, 693 for glycaemia and 201 for Framingham score. Different numbers of individuals in each class are due to the fact that not for all individuals all clinical parameters were available.

To assess which buckets (that is, resonance peaks) were significantly different between different groups a univariate wilcoxon test was used. A *P*-value ≤ .001 (with Bonferroni correction for multiple tests) was considered statistically significant. All resonances of interest were then manually checked, and signals were assigned on template one-dimensional NMR profiles by using matching routines of AMIX 7.3.2 (Bruker BioSpin) in combination with the BBIOREFCODE (Version 2-0-0; Bruker BioSpin) reference database and published literature when available.

All calculations were made using home-made scripts written in R language[6].

Supplemental References

1. De Jong S. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics Intell Lab Syst*. 1993;18:251-263.
2. Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.
3. Chang and Lin, 2001, LIBSVM: a Library for Support Vector Machines
4. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 1994;308:1552.
5. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, van Velzen EJJ, van Duijnhoven, JPM, van Dorsten FA. Assessment of PLSDA cross validation. *Metabolomics*. 2008;4:81-89.
6. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. J *Comput Stat Graph*. 1996;5:299-314.

Supplemental Tables

Table S1. Confusion matrices for TC classification using diffusion edited spectra.

DIFF	Low Values (< 174)	High Values (> 232)	N = 305 (152 vs 153)	
Low	97.75%	2.25%		Sensitivity
High	1.81%	98.19%		Specificity
			97.91%	Accuracy

Table S2. Confusion matrices for LDL classification using diffusion edited spectra.

DIFF	Low Values (<100)	High Values (>160)	N = 85 (42 vs 43)	
Low	98.97%	1.03%		Sensitivity
High	2.88%	97.12%		Specificity
			98.85%	Accuracy

Table S3. Confusion matrices for HDL classification using diffusion edited spectra.

DIFF	Low Values (<44)	High Values (>63)	N = 82 (42 vs 40)	
Low	97.52%	2.48%		Sensitivity
High	2.03%	97.97%		Specificity
			97.83%	Accuracy

Table S4. Confusion matrices for triglycerides classification using diffusion edited spectra.

DIFF	Low Values (< 60)	High Values (> 131)	N = 350 (149 vs 151)	
Low	98.28%	1.72%		Sensitivity
High	2.73%	97.27%		Specificity
			97.42%	Accuracy

Table S5. Confusion matrices for TC/HDL ratio classification using diffusion edited spectra.

DIFF	Low Values (<3)	High Values (>4.90)	N = 80 (40 vs 40)	
Low	98.68%	1.32%		Sensitivity
High	0.17%	99.83%		Specificity
			99.35%	Accuracy

Table S6. Confusion matrices for Framingham risk score classification using diffusion edited spectra.

DIFF	Low Values (< 1.25)	High Values (> 7.00)	N = 80 (40 vs 40)	
Low	87.73%	12.27%		Sensitivity
High	9.63%	90.37%		Specificity
			88.91%	Accuracy

Author Query Form

Journal Title : IJT
Article Number : 360069

Dear Author/Editor,

Greetings, and thank you for publishing with SAGE Publications. Your article has been copyedited, and we have a few queries for you. Please address these queries when you send your proof corrections to the production editor. Thank you for your time and effort.
Please assist us by clarifying the following queries:

Query No.	Query	Remarks
1	"INFORMATION ON ALL OF" WHAT? THE SMALL MOLECULES?	
2	PROVIDE LOCATION OF BRUKER	
3	CLARIFY WHAT YOU MEAN BY "K": THIS IS USUALLY USED AS THE SYMBOL FOR KELVIN	
4	PROVIDE LOCATION OF PERKIN ELMER	
5	PROVIDE LOCATION OF CBC EUROPE LTD	
6	PROVIDE LOCATION OF KARTELL	
7	IS "PERCENTAGE COEFFICIENT OF VARIANCE" CORRECT FOR CV%?	
8	NOT CLEAR WHAT YOU MEAN BY "STAR SYMBOL" AND "SHARP SYMBOL": * AND #?	
9	IN TABLE 2, YOU GIVE "OXIDIZED ALBUMIN" TWICE; IS THIS CORRECT?	

International Journal of Toxicology
000(00) 1-10

DOI: 10.1177/1091581809360069
http://ijt.sagepub.com
SAGE

Effects of Ozone Blood Treatment on the Metabolite Profile of Human Blood

Valter Travagli,[1] Iacopo Zanardi,[1] Patrizia Bernini,[2] Stefano Nepi,[2] Leonardo Tenori,[2] and Velio Bocci[1]

Abstract

Metabonomic characterization of the effects caused by ozone and other stressors on normal human blood was performed. Samples of blood obtained from healthy subjects were treated ex vivo with increasing concentrations of ozone and/or with UV radiation and heat. ^{1}H-NMR analysis of plasma samples after treatments showed the quantitative variation of some metabolites and the formation of new metabolites normally absent. Both the increment of some metabolites like formate, acetoacetate, and acetate and the decrement of pyruvate were of particular interest. Moreover, the oxidation of ascorbic acid and the transformation of uric acid into allantoin after ozonation within the therapeutic concentration range were observed. In the ozonated spectra, 2 unidentified peaks appeared at 2.82 ppm and 8.08 ppm. They are related to the direct antioxidant activity of albumin in the presence of ozone and they could be considered as specific markers of the blood ozonation.

Keywords

ozone toxicity, ^{1}H-NMR, reactive oxygen species, oxidative stress, metabonomic profile

Reactive oxygen species (ROS) represent an area of intensive research and include both free radicals and nonradical species, among which ozone plays an important role (Table 1).[1] In the stratosphere, ozone constitutes the protective layer around the planet, whereas at ground level ozone is a pollutant. Previous results[2-5] show that ozone causes lipid peroxidation and can generate a variety of compounds, as summarized in Figure 1. At the same time, because of its fungicidal, virucidal, and bactericidal properties, ozone is used to purify the drinking water in many municipalities.[6]

Ozone for therapeutic purposes has been used on a rather empirical basis by many practitioners in the world. There are specific applications of ozone therapy in a number of pathologies, such as vascular diseases, ulcers, and acute and chronic viral diseases. On the basis of some biochemical studies[7-10] performed during the last 15 years, a framework for understanding the basic mechanisms of action of ozone when it comes in contact with human blood at appropriate doses is now available. These data have been critically reviewed,[11,12] and the following biological responses can be summarized: (a) improvement of blood circulation and oxygen delivery to ischemic tissue, owing to the concerted effect of nitric oxide and CO and a variable increase of intraerythrocytic (2,3-diphosphoglycerate) 2,3-DPG level[13,14]; (b) enhancement of the general metabolism[13]; (c) upregulation of the cellular antioxidant enzymes and induction of heme oxygenase 1 (HO-1) and heat shock protein-70 (HSP-70)[14,15]; (d) induction of a mild activation of the immune system[7,8] and enhancement of the release of growth factors from platelets[16]; and (e) excellent disinfectant activity when topically used in water and oils,[17,18] although this is negligible in the circulation owing to the potent blood antioxidant capacity.[19] Moreover, adequate ozone treatments do not induce acute or chronic side effects, and often patients report a feeling of wellness.[13] However, an in-depth evaluation of ozone impact on blood by studying the formation of metabolites appears to be valuable.

In this specific case, metabonomics has been often used to describe multiple metabolic changes caused by a biological perturbation. *Metabonomics* has been defined as "the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification."[20]

Nuclear magnetic resonance (NMR)–based metabonomics offers evident advantages in contrast with knowledge-guided search of metabolites in biofluid samples. NMR-based metabonomics makes no assumptions about the identity of the metabolites that are relevant for the investigated aspects. Information on the metabolite pattern alterations is directly obtained through statistical analysis of the NMR profiles. Usually, metabonomics does not rely on the measurement of a single metabolite-associated peak but analyzes spectra as a whole: metabonomic profiles essentially are the superposition of the

[1] University of Siena, Siena, Italy
[2] University of Florence, Sesto Fiorentino, Italy

Corresponding Author:
Velio Bocci, viale Aldo Moro, 2 - 53100 Siena, Italy.
Email: bocci@unisi.it

Table 1. Nomenclature of Reactive Oxygen Species

Free Radicals	Nonradicals
Superoxide, $O_2\bullet^-$ (1×10^{-3}, 1×10^{-6} in presence of superoxide dismutase)	Hydrogen peroxide, H_2O_2 (fairly stable in H_2O, 2.5 in plasma, <1 in blood)
Hydroxyl, $HO\bullet$ (1×10^{-9})	Hypobromous acid, HOBr
Hydroperoxyl, $HO_2\bullet$	Hypochlorous acid, HOCl (fairly stable)
Peroxyl, $RO_2\bullet$ (0.1-7)	Ozone, O_3
Carbonate, $CO_3\bullet^-$	Singlet oxygen, $^1O_2$ (1×10^{-6})
Carbon dioxide, $CO_2\bullet^-$	Organic peroxides, ROOH
Nitric oxide, $NO\bullet$ (0.1-5)	Peroxynitrite, $ONOO^-$ (1-4)
	Peroxynitrous acid, NOOH

Numbers in parentheses represent half-life values, in seconds, at 37°C.

^{1}H-NMR spectra of tens to thousands of different small molecules (up to 2500 in the case of urine and up to 200 in the case of blood) present in the sample at greater than 1 μM concentration.[21] In principle, an NMR profile contains qualitative and quantitative information on all of them. Small changes in enzyme concentrations can cause considerable alterations in intermediate products, and owing to the fact that metabolic networks are interconnected, NMR-based metabonomic analysis of biological fluids such as serum, plasma, and urine can be very useful.

NMR-based metabonomic studies have provided significant information on a wide range of pathologies, such as cancer,[22] as well as a variety of gastrointestinal pathologies.[23] Not surprisingly, the latter are more likely to be associated with metabolic alterations, among them *Helicobacter pylori* infection.[24] Recently, therapeutic ozone application in clinical dentistry by high-resolution ^{1}H-NMR has been studied.[25]

The purpose of the present study was to investigate ex vivo the effect on the plasma metabonomic profile as obtained by (1) blood ozonation within the therapeutic concentration, (2) blood ozonation at high concentrations, and (3) blood ozonation at abnormally high concentration associated with other stressors, as performed by Torre-Amione et al.[26] The present investigation appears to be significant in that it defines definite alterations and toxicity regarding the amount of metabolites and suggests specific markers of oxidative stress induced by ozone blood treatment. In a future work, after reinfusing ozonated blood in the donor, we will investigate whether these metabolites appears in the circulation or in excreta.

Methods

Ozone Generation

Ozone was generated from medical-grade oxygen using electrical corona arc discharge, by 2 generators:

1. Ozonline ECO_3 (Torino, Italy), which delivers ozone concentrations between 0.1 and 3.0 μg/mL of gas, with a gas flow ranging from 100 mL/min (15 L/h) up to 1 L/min (60 L/h).
2. Ozonosan PM 100K (Hansler GmbH, Iffezheim, Germany), which can deliver to either blood, plasma, or other samples ozone concentrations up to 80 μg/mL with a gas flow ranging between 1 and 8 L/min.

In all cases, the ozone concentration is monitored continuously by photometry at 600 nm (Chappuis band), periodically checked by iodometric titration, as recommended by the Standardization Committee of the International Ozone Association. The photometer was periodically checked by using the iodometric titration in observance of the rules established by International Ozone Association. Blood ozonation must be carried out only using medical oxygen and not filtered air because it contains 78% of nitrogen with the inherent formation of nitrogen oxides.

Single-use silicon treated polypropylene syringes (ozone-resistant) and tygon polymer tubing were used throughout the reaction procedure to ensure containment of ozone and consistency in concentrations.

Collection of Human Blood

Several blood samples of 60 mL were taken from 2 healthy, nonsmoking, male blood donors in the morning. Either sodium citrate 3.8% (1 mL/9 mL of blood) or EDTA (final concentration, 1.35 mg/mL) was used as the anticoagulant, and blood samples were immediately subdivided and introduced in ozone-resistant syringes (5 mL) or in a quartz vessel (10 mL) with a thermostat. The experimental program was allowed by the Ethical Committee on the basis of informed consent signed by 2 (V.T. and I.Z.) of the experimenters.

Sample Treatments

In experiment A, blood was treated with a single dose of ozone (concentration per volume) ex vivo as follows: A predetermined volume of a gas mixture composed of oxygen (95%-99%) and ozone (1%-5%), at ozone concentrations 20, 40, 80, 160, 240, 320, and 800 μg/mL (from 0.42 to 16.8 μmol/mL), was collected with a syringe and immediately introduced via a multidirectional stopcock into another syringe containing the blood sample. When possible, a blood sample/gas volume at a 1:1 ratio was used. The final gas pressure remained at normal atmospheric pressure. Controls were represented by both

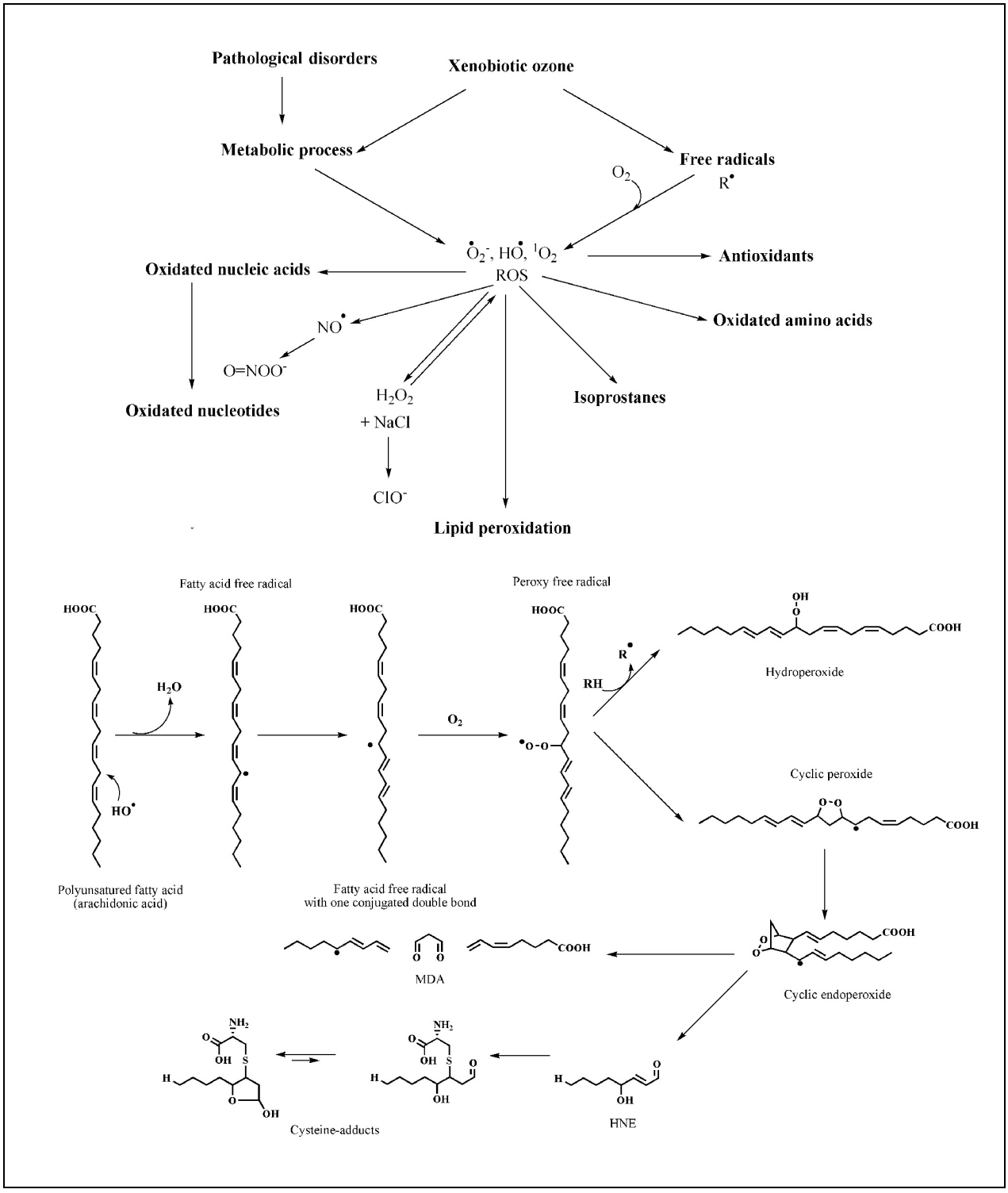


Figure 1. The cascade of compounds generated during ozone interaction in a biological environment. In particular, the sequence of reactions occurring during lipid peroxidation with production of malondialdehyde (MDA) and 4-hydroxy-2,3 trans-nonenal (HNE) is shown.

untreated blood samples and as treated with medical-grade oxygen. All control samples were normal.

In experiment B, samples of human albumin (4%, Kedrion, Barga, Italy) alone or in the presence of physiological concentrations of uric acid (300 μM) and/or ascorbic acid (50 μM) were treated with ozone. Gas delivery was performed as experiment A, at ozone concentrations of 5, 10, 20, 40, 80, and 160 μg/mL (from 0.105 to 3.6 μmol/mL). Control samples were treated with medical-grade oxygen. Samples containing albumin (4%) together with scalar concentrations of

Table 2. Correlation Factor (Expressed as r^2 Parameter) for the Variable Metabolites

Peak	r^2 Value	Trend
Acetate (1.93 ppm)	0.96063	Linear
Acetoacetate (2.29 ppm)	0.97061	Linear
Pyruvate (2.35 ppm)	0.86745	Linear
Oxidized albumin (2.82 ppm)	0.95027	Linear
Allantoin (5.40 ppm)	0.94049	Hyperbolic
Oxidized albumin (8.08 ppm)	0.99535	Linear
Formate (8.47 ppm)	0.99244	Linear

9

4-hydroxynonenal (4-HNE) (from 0.08 to 3.2 μM) were also studied by NMR analysis.

In experiment C, treatments were performed in a thermostatted quartz vessel equipped with a gas inlet, a baffle plate at about half-height to limit blood foaming, and a silicon stopper. For each experiment, the device was loaded with 10 mL of blood, and 3 stressful agents were continuously (20 minutes) supplied, alone or in combination: (1) ozone, (2) ozone + UV rays, (3) ozone + heat, (4) ozone + UV rays + heat, (5) UV rays, (6) heat, and (7) UV rays + heat. The gas outlet was connected to an ozone destructor. Specifically, the temperature was set at 42.5°C, the total amount of ozone insufflated was 60 mg, and the source of UV rays (Multirays, Helios Italquartz, Milan, Italy) consisted of 10 lamps (254 nm) of 15 W each (length 43 cm, diameter 2.6 cm). These lamps were fixed in a special lamp-holder at a distance of about 10 cm from the sample.

NMR Analysis

Blood was immediately centrifuged at 2000*g* for 5 minutes at room temperature to sediment erythrocytes and to collect the plasma as supernatant. Plasma samples as well as samples of experiment B were stored frozen (–80°C) until NMR examination.

Frozen plasma samples were thawed at room temperature and vortexed before use. Next 300 μL of a phosphate sodium buffer (70 mM Na_2HPO_4; 20% vol/vol ^2H_2O; 0.025% vol/vol NaN_3; 0.8% wt/vol sodium trimethylsilyl [2,2,3,3-$^2H_4$]propionate [TSP]; pH 7.4) was added to 300 μL of each plasma sample, and the mixture was homogenized by vortexing for 30 seconds. Then 450 μL of this mixture were transferred into a 4.25-mm NMR tube (Bruker BioSpin srl) for analysis.[23]

2

^{1}H-NMR spectra for all samples were acquired using a Bruker 600 MHz spectrometer (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency and equipped with a 5-mm CPTCI ^{1}H-^{13}C/^{31}P-^{2}H probe including a z-axis gradient coil, automatic tuning matching (ATM), and an automatic sample changer. A Pt100 thermoresistance served for temperature stabilization at the level of approximately 0.01 K at the sample. Before measurement, samples were kept for at least 3 minutes inside the NMR probe head for temperature equilibration (300 K). One-dimensional spectra were acquired using a standard pulse sequence (nuclear Overhauser enhancement spectroscopy; NOESYpresat; Bruker), using 64 free induction decays (FIDs), 64k data points, a spectral width of 12 019 Hz, an acquisition time of 2.7 seconds, a relaxation delay of 4 seconds, a mixing time of 100 milliseconds, and a Carr-Purcell-Meiboom-Gill (CPMG; Bruker) spin-echo sequence to suppress signals arising from high-molecular-weight molecules with the same acquisition parameter of NOESYpresat.

3

Spectral Processing

Free induction decays were multiplied by an exponential function equivalent to a 1.0-Hz line-broadening factor before applying Fourier transform. Transformed spectra were manually phased and corrected using cubic spin lines with TopSpin 2.1 (Bruker BioSpin srl), taking the CH_3-doublet of lactate as reference (centered at 1.330 ppm).[23] All resonances of interest were assigned on template 1-dimensional NMR profiles by using matching routines of AMIX 3.8.4 (Bruker BioSpin) in combination with the BBIOREFCODE (version 2-0-0; Bruker BioSpin) reference database and published literature when available.

All spectra were normalized using the area of the doublet of CH_3-alanine peak (between 1.50 and 1.47 ppm), which after scaling to the total area results fixed in intensity and shift for all samples analyzed, except the albumin spectra that are normalized to the area of a doublet of albumin (between 2.76 and 2.73 ppm). Peak area measurements of the metabolites are obtained using R 2.8.1[27] with house-made script for the calculation of the integrals of interest.

Spectroscopic Determination

Blood was immediately centrifuged at 2000*g* for 5 minutes at room temperature to sediment erythrocytes and to collect the plasma as supernatant. Plasma samples were stored frozen (–80°C) until determination.

The absorbance was determined with a Perkin Elmer Lambda 2B UV-VIS spectrophotometer equipped with 10-mm quartz cells. The absorption spectra were obtained over the range 700 to 400 nm. A data interval of 1 nm was selected, with a signal averaging time of 0.5 seconds for each data point (resulting in a scan time of 150 seconds). All data were acquired with UV WINLAB software. Precisely, the adopted wavelengths for the determination of hemoglobin concentration were 414, 541, and 576 nm, whereas methemoglobin was evaluated at 630 nm.[10] Before any determinations, samples were centrifuged at 3000*g* for 20 minutes in order to read the clear supernatants, and, when necessary, the samples were appropriately diluted. All scans were performed by the same operator and carried out at 21°C ± 0.5°C.

4

Viscosity Evaluation

Viscomate VM10AL (CBC Europe Ltd) is an innovative torsional oscillating viscometer characterized by constant shear

5

stress systems driven by a piezoelectric ceramic source. This instrument measures viscosity by sensing a change in oscillation amplitude of a liquid-immersed detector, based on constant input voltage. An original phase-locked loop circuit maintains instrument resonant frequency of 1 kHz; the detector oscillation amplitude with no resistance is 1 μm. Angular acceleration of the detector is measured and reported as dynamic viscosity with the range 0.400 to 1000 mPa/s. The probe dimension was 9 mm with respect to the diameter.[28]

The viscosity of treated blood (experiments A and C) was immediately evaluated. All the determinations were conducted into polystyrene Technicon sample cups (Kartell, nominal capacity 2 mL). Temperature was accurately monitored during the experiments (37.0°C ± 0.1°C). Viscosity values were recorded for 6 minutes (data collection every 5 seconds) by PC connection through an RS-232 port.

Statistical Analysis

Results were expressed as the mean (percentage coefficient of variance [CV%] <2) of at least 3 independent measurements. Statistical evaluations were performed by a 1-way analysis of variance (ANOVA) using a statistics software (Instat software, version 3.0 GraphPAD Software, San Diego, Calif). Bonferroni test was used after ANOVA to evaluate statistical difference between individual means. Significance was defined as a P value less than .05. The correlation graphs are obtained using OriginPro 8, the correlation factors are expressed with the r^2 parameter.

Results

A preliminary overview of the changes caused by blood ozonation is shown in Figure 2. Both CPMG and NOESY1D spectra were analyzed. The examination of spectra of the control samples treated with pure oxygen did not yield particular differences with respect to basal samples, except a small increment of pyruvate peak at 2.35 ppm. However, samples treated with increasing amounts of ozone (from 0 to 16.8 μmol/mL) showed variations in the concentration of several metabolites (Figure 2).

In particular, a strong increase of the peaks at 8.45 ppm (CH-formate), 5.40 ppm (CH-allantoin), 2.29 ppm (CH_3-acetoacetate), and 1.93 ppm (CH_3-acetate) were observed as well as a drastic decrease of the 2.35 ppm peak (CH_3-pyruvate), which disappeared at an ozone concentration of 320 μg/mL. All the variations of these metabolites are strongly correlated with the concentration of insufflated ozone (Figure 3 and Table 2).

The variations of pyruvate and acetate are chemically correlated, in fact, acetate is formed by a direct ozone oxidation of pyruvate; formate is the last product of the oxidative process of carbohydrates, and allantoin comes from oxidation of uric acid,[25,29] which behaves as a "sacrificial" molecule in the presence of a strong oxidant such as ozone.[14,30] Acetoacetate is often formed in plasma by oxidation of γ-hydroxybutyrate.[25,31]

Additional tests were performed to (1) confirm the trends of variations of each metabolites at higher concentrations of ozone and (2) confirm the specificity of the obtained results for the ozone treatment in comparison to other kinds of oxidative stress (treatment C).

The increase of the ozone concentration up to 800 μg/mL of gas per 1 mL of blood confirmed our findings for all the previous metabolites. For the nontherapeutic, high ozone concentrations (240, 320, and 800 μg/mL), the disappearance of the pyruvate signal from the spectrum was caused by the excessive oxidative process.

A further analysis of the spectra highlighted the presence of unidentified new peaks. These peaks occurred at 2.82 and 8.08 ppm and were not present in untreated plasma. Thus, these peaks were due to specific products directly linked to the ozone activity and indeed their amounts in the plasma are related to different ozone concentration (Figures 2 and 4).

The correlations between the areas of each peak and the concentration of insufflated ozone were very high. These data suggest that these peaks may represent markers of oxidative stress.

To express a hypothesis about the origin of these 2 peaks, other tests were performed (experiment B). It is well known that ozone causes lipid peroxidation, yielding several aldehydes among which HNE is quantitatively relevant.[32] This aldehyde has a double bond in β and it is not stable. It reacts easily with free thiol groups of cysteine, glutathione, and other metabolites normally present in plasma to form adducts easily eliminable from the organism with the urine.[32,33] A recent study indicated an interaction of HNE with Cys34 present in albumin.[34] Starting from these data, some samples obtained with the addition of scalar concentration of HNE (from 0.08 μM up to 3.2 μM) to a solution of 4% human albumin in saline were analyzed. After the addition of HNE to albumin, the 2 peaks at 2.82 ppm and 8.08 normally present after blood ozonation were absent, as well as the aldehydic signals of HNE. On the other hand, direct ozonation of a 4% albumin solution in saline allowed the detection of these 2 peaks, which increased in intensity in relation to the amount of ozone concentrations, indicating that these 2 signals are due to the direct oxidant activity of ozone on the albumin. Indeed, albumin has been called a sacrificial molecule. Whether albumin undergoes dimerization[35] or allows the formation of sulfenic acid[36] has not been clarified.

As for the experiments with albumin plus uric acid (300 μM) and/or ascorbic acid (50 μM), an increased formation of allantoin was obtained. The peak at 4.5 to 4.6 (ascorbic acid), although present when in the control sample, disappeared as soon as ozone (20 μg/mL) was added. These results suggest that uric acid is oxidized to allantoin and ascorbic acid is oxidized to dehydroascorbate. Additionally, the presence of formate and the appearance of both 2.82 and 8.08 ppm peaks were evidenced.

On the basis of the very harsh ozonation of blood (about 6 mg/mL of blood) plus UV radiation and heat (42.5°C) recently reported,[26] it was of interest to evaluate the spectra after blood treatment with (1) O_3, (2) O_3 + UV rays, (3) O_3

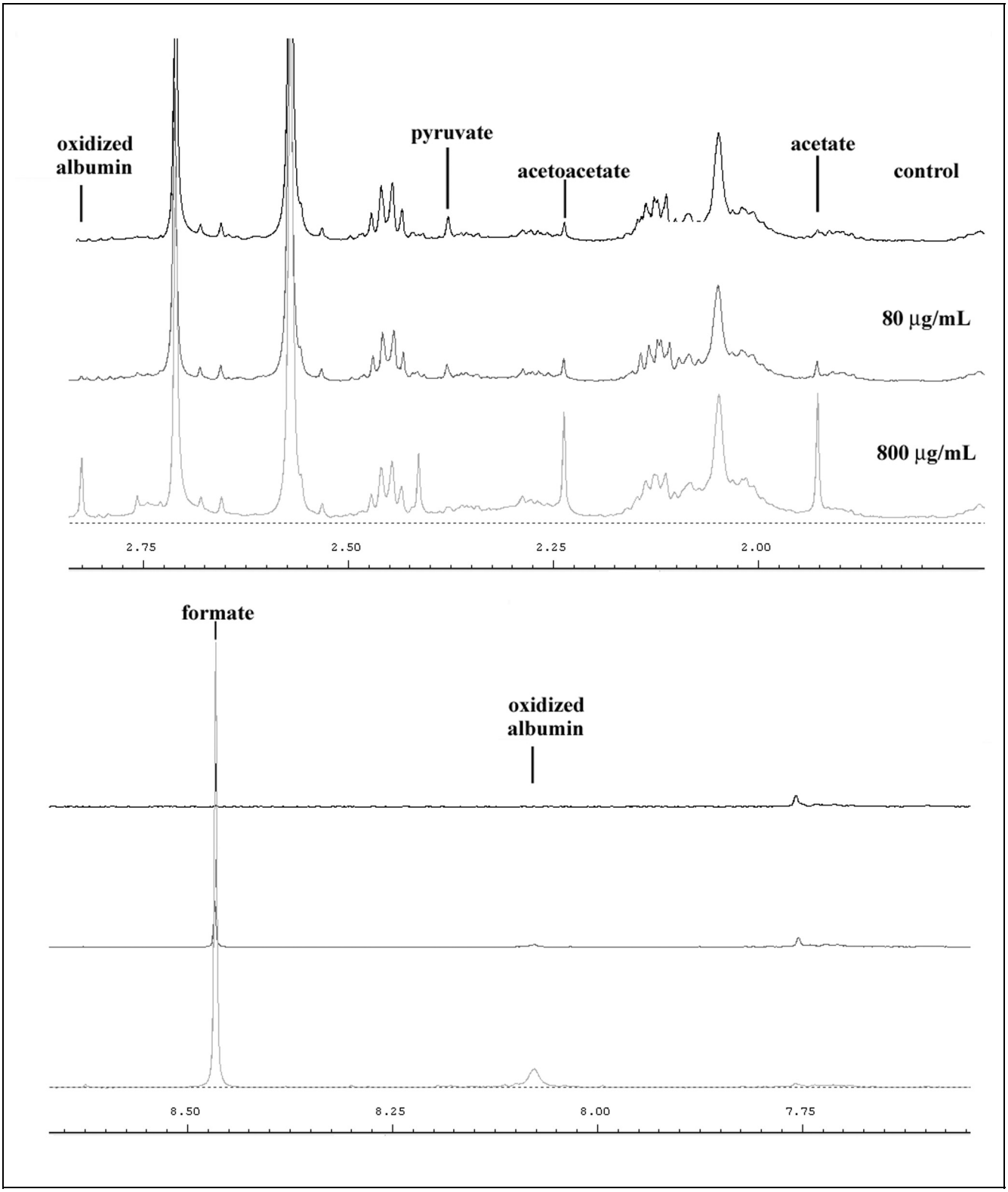


Figure 2. Snapshot of Carr-Purcell-Meiboom-Gill spectra containing some principal variable metabolites in human serum treated with ozone (control, 80 μg/mL, 800 μg/mL).

+ heat, (4) O_3 + UV rays + heat, (5) UV rays, (6) heat, and (7) UV rays + heat. No appreciable influence of heat and/or UV treatments on the metabolites present in the blood was detected, except that at 4.43 and 8.19 ppm in the case of heat. These signals were not present in ozone-treated samples. The remaining other types of treatments (1-4) showed the typical fingerprint

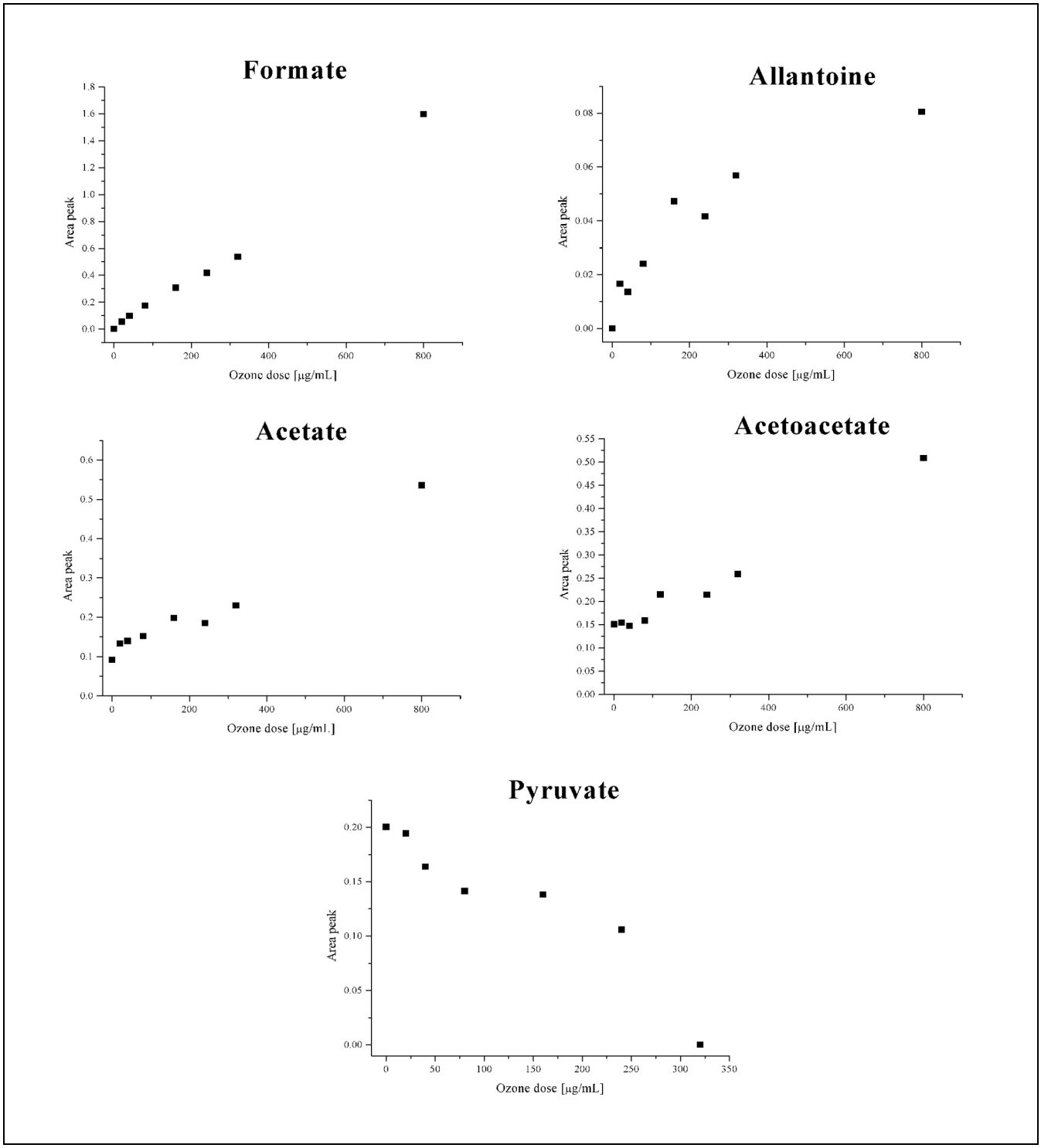


Figure 3. Variations of several metabolites in relation to increasing ozone dosages (n = 3; CV% <2).

alterations already seen in previous tests (formate, acetoacetate, acetate, and allantoin increase; pyruvate decreases; peaks occur at 2.82 ppm and 8.08 ppm) to indicate that the principal stressor is due to ozone. Considering the different areas of the previous defined marker peaks at 2.82 ppm, a classification of samples based on the insufflated quantity of ozone is possible.

To confirm metabonomic results on the effects of blood oxidative stress on the erythrocytic membrane and the possible cell damage with intraerythrocytic material release according to both treatments A and C, physicochemical investigations based on hemolysis as well as viscosity were performed. In the first set of experiments, only increasing amounts of 1 single dose of

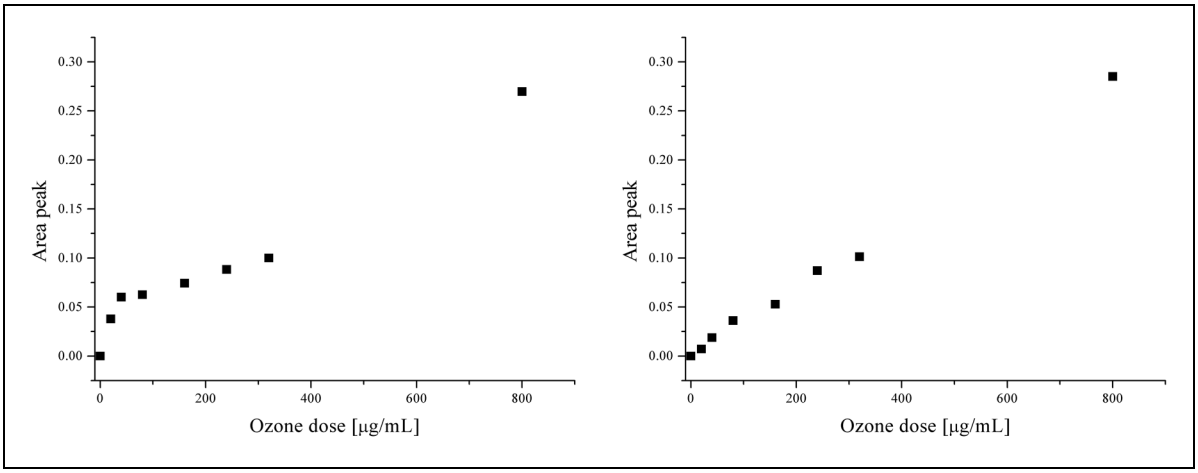


Figure 4. Linear trends of appearance of 2 unidentified peaks in relation to ozone dosages: signal areas at 2.82 ppm (left) and at 8.08 ppm (right) (n = 3; CV% <2).

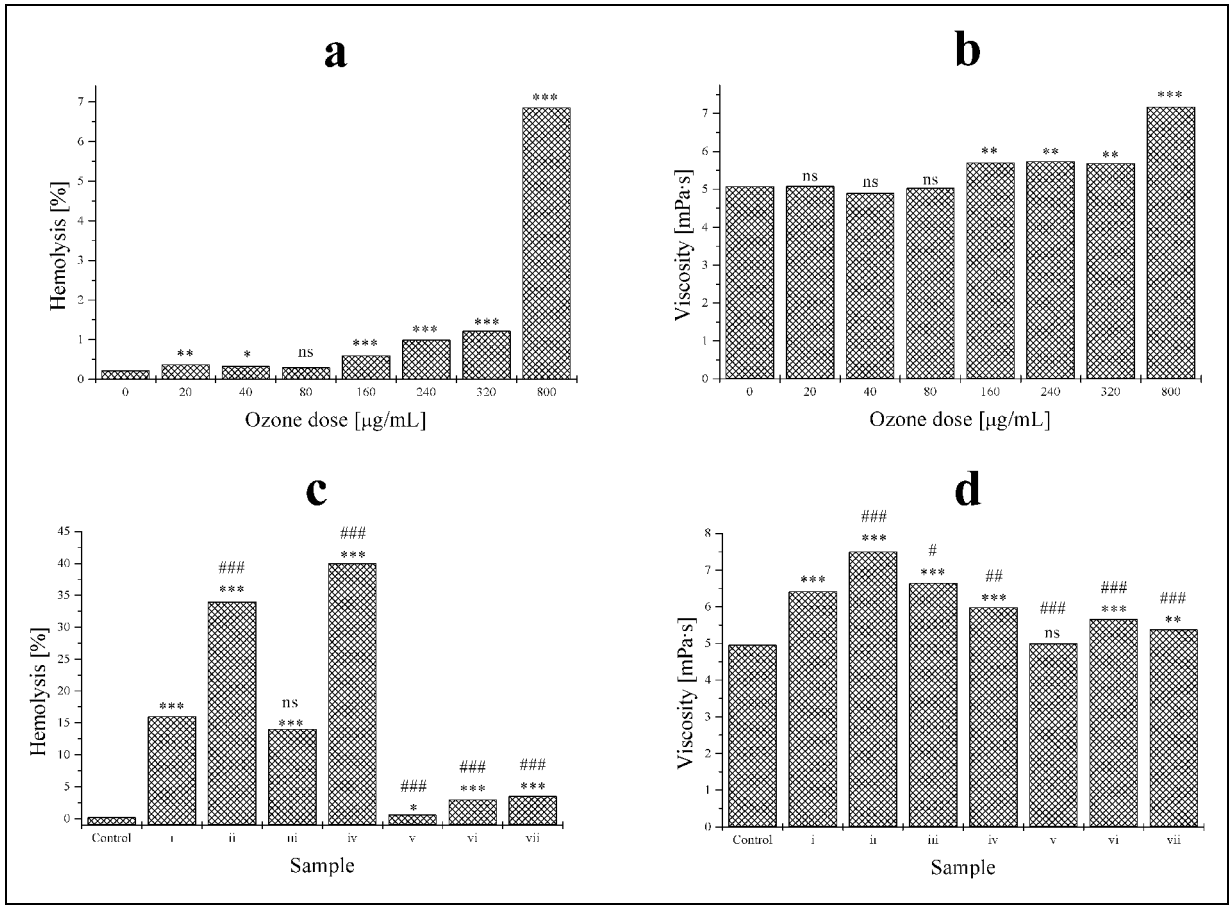


Figure 5. Hemolysis (left) and viscosity (right) variations in treatments a (top) and c (bottom). Statistical analysis: 3 symbols, $P < .001$, 2 symbols $P < .01$, 1 symbol $P < .05$, not significant $P > .05$; star symbol vs control; sharp symbol vs sample I (n = 5).

ozone, according to treatment A protocol, were evaluated. A statistically significant increase of hemolysis (>5%) was observed at 800 μg/mL, whereas viscosity variations started at 160 μg/mL (Figure 5, a and b, respectively). In the second set of experiments, variable results were obtained, depending on the type and number of stressors applied. The combination of ozone + UV (condition 2) as well as ozone + UV + heat (condition 4) yielded the highest values of hemolysis (Figure 5c, samples ii and iv). The viscosity markedly increased only in the ozone + UV combination (Figure 5d, sample ii).

Discussion

Metabonomics is very useful to evaluate large and small differences in metabolic profiles, as demonstrated in urine for oxidative kidney damage.[24] The present work demonstrates that the metabonomic approach is a valid instrument to detect alterations in plasma profile due by oxidative stress induced by ozone and other stressors. Information on the oxidative stress was obtained from the CPMG spectra, in which signals arising from large macromolecules such as lipids are suppressed. This observation suggests that although ozone causes lipid peroxidation, ozone treatment causes neither visible alteration in lipid composition of blood nor hemolysis in the range of the well-documented therapeutic concentrations of ozone, namely between 10 μg/mL blood, or 0.21 μmol/mL and 80 μg/mL blood, or 1.68 μmol/mL.[14] It is unlikely that changes in metabolites are caused by hemoglobin degradation. Moreover Shinriki et al[30] excluded any erythrocytic membrane peroxidation with ozone concentration up to 100 μg/mL per mL of blood.[30] The first changes in the lipidic component appear only with high ozone concentrations (240 μg/mL upward) not normally used for therapeutic purposes. Moreover, after treatment with an ozone concentration of 800 μg/mL, changes in several peaks, both NOESY1D and in CPMG spectra, were observed, leading us to hypothesize damage to cell membranes caused by lipid oxidation with possible cytolysis. This supposition would explain both changes in the lipid component in the spectrum, well visible in NOESY1D spectrum at the highest concentrations of ozone and in the component given by small metabolites, better visible in CPMG spectrum.

The principal product of lipid peroxidation (HNE) it doesn't alter the metabonomic profile of serum. The present alterations are only due to direct oxidation of some molecules (as pyruvate and carbohydrates) and antioxidants such as uric acids and ascorbic acid as well as to the so-called sacrificial behavior of albumin.

In this study the metabolic profile of blood samples treated with ozone has been defined: the principal metabolic variation with the possible markers of oxidative stress due to ozone treatment has been indicated; and blood alterations due to ozone have been demonstrated to be strictly correlated to the concentration of insufflated ozone. When ozone concentrations were within the therapeutic range, only a few alterations were detected and were not associated to significant damages. However, several stresses, such as the association of ozone with heat and/or UV, appear deleterious, as exemplified in Figure 5. It appears that the denominated Celacade procedure consisting of a harsh treatment of blood with an enormous ozone dose plus UV rays and heat stress[26] causes extensive blood cell destruction with no beneficial effect on patients with chronic heart disease. This meaningless procedure must be proscribed.[37] If ozone therapy is to be safely used, it requires precise and well-calibrated ozone concentrations and the avoidance of ozone bubbling as well as additional stressor.

Acknowledgments

We are very grateful to Professor C. Luchinat for allowing us to perform these experiments and for useful and critical comments. We thank Professor F. Ponticelli for the use of the Multirays device. We gratefully acknowledge the linguistic revision by Mrs. H. Carter.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

1. Halliwell B, Whiteman M. Measuring reactive species and oxidative damage in vivo and in cell culture: how should you do it and what do the results mean? *Br J Pharmacol*. 2004;142:231-255.
2. Esterbauer H, Schaur RJ, Zollner H. Chemistry and biochemistry of 4-hydroxynonenal, malonaldehyde and related aldehydes. *Free Radic Biol Med*. 1991;11:81-128.
3. Pryor WA. Ozone in all its reactive splendor. *J Lab Clin Med*. 1993;122:483-486.
4. Pryor WA, Squadrito GL, Friedman M. The cascade mechanism to explain ozone toxicity: the role of lipid ozonation products. *Free Radic Biol Med*. 1995;19:935-941.
5. Esterbauer H. Estimation of peroxidative damage: a critical review. *Pathol Biol*. 1996;44:25-28.
6. Bonacquisti TP. A drinking water utility's perspective on bromide, bromate, and ozonation. *Toxicology*. 2006;221:145-148.
7. Bocci V, Luzzi E, Corradeschi F, Paulesu L, Di Stefano A. Studies on the biological effects of ozone, 3: an attempt to define conditions for optimal induction of cytokines. *Lymphokine Cytokine Res*. 1993;12:121-126.
8. Bocci V, Valacchi G, Corradeschi F, et al. Studies on the biological effects of ozone: 7. Generation of reactive oxygen species (ROS) after exposure of human blood to ozone. *J Biol Regul Homeost Agents*. 1998;12:67-75.
9. Bocci V, Aldinucci C. Biochemical modifications induced in human blood by oxygenation-ozonation. *J Biochem Mol Toxicol*. 2006;20:133-138.
10. Travagli V, Zanardi I, Silvietti A, Bocci V. A physicochemical investigation on the effects of ozone on blood. *Int J Biol Macromol*. 2007;41:504-511.
11. Bocci V. Is it true that ozone is always toxic? The end of a dogma. *Toxicol Appl Pharmacol*. 2006;21:493-504.
12. Bocci V. The case for oxygen-ozonetherapy. *Br J Biomed Sci*. 2007;64:44-49.
13. Bocci V. Ozone. *A New Medical Drug*. Springer, Dordrecht: The Netherlands; 2005.
14. Bocci V, Borrelli E, Travagli V, Zanardi I. The ozone paradox: ozone is a strong oxidant as well as a medical drug. *Med Res Rev*. 2009;29:646-682.
15. Bocci V, Aldinucci C, Mosci F, Carraro F, Valacchi G. Ozonation of human blood induces a remarkable upregulation of heme

oxygenase-1 and heat stress protein-70. *Mediators Inflamm*. 2007; 2007:26785.
16. Valacchi G, Bocci V. Studies on the biological effects of ozone: 10. Release of factors from ozonated human platelets. *Mediators Inflamm*. 1999;8:205-209.
17. Sechi LA, Lezcano I, Nuñez N, Espim M, Dupre I, Pinna A. Antibacterial activity of ozonized sunflower oil (OLEOZON). *J Appl Microbiol*. 2001;90:279-284.
18. Travagli V, Zanardi I, Bocci V. Topical applications of ozone and ozonated oils as anti-infective agents: An insight into the patent claims. *Recent Pat Antiinfect Drug Discov*. 2009;4:130-142.
19. Burgassi S, Zanardi I, Travagli V, Montomoli E, Bocci V. How much ozone bactericidal activity is compromised by plasma components? *J Appl Microbiol*. 2009;105:1715-1721.
20. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*. 1999;29:1181-1189.
21. Wishart DS, Tzur D, Knox C, et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res*. 2007;35:D521-D526.
22. Claudino WM, Quattrone A, Biganzoli L, Pestrin M, Bertini I, Di Leo A. Metabolomics: available results, current research projects in breast cancer and future applications. *J Clin Oncol*. 2007;25: 2840-2846.
23. Bertini I, Calabrò A, De Carli V , et al. The metabonomic signature of celiac disease. *J Proteome Res*. 2009;8:170-177.
24. Gao XX, Ge HM, Zheng WF, Tan RX. NMR-based metabonomics for detection of Helicobacter pylori infection in gerbils: which is more descriptive. *Helicobacter*. 2008;13:103-111.
25. Grootveld M, Silwood CJ, Lynch E. 1H NMR investigations of the oxidative consumption of salivary biomolecules by ozone: relevance to the therapeutic applications of this agent in clinical dentistry. *Biofactors*. 2006;27:5-18.
26. Torre-Amione G, Anker SD, Bourge RC, et al. Results of a nonspecific immunomodulation therapy in chronic heart failure (ACCLAIM trial): a placebo-controlled randomised trial. *Lancet*. 2008;371:228-236.
27. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat*. 1996;5:299-314.
28. Travagli V, Zanardi I, Boschi L, et al. Comparison of blood viscosity using a torsional oscillation viscometer and a rheometer. *Clin Hemorheol Microcirc*. 2008;38:65-74.
29. Gruber J, Tang SY, Jenner AM, et al. Allantoin in human plasma, serum and nasal lining fluids as a biomarker of oxidative stress; avoiding artifacts and establishing real in vivo concentrations. *Antioxid Redox Signal*. 2009;11:1767-1776.
30. Shinriki N, Suzuki T, Takama K, et al. Susceptibilities of plasma antioxidants and erythrocyte constituents to low levels of ozone. *Haematologia*. 1998;29:229-239.
31. Simons JR, Theodore J, Robin ED. Common oxidant lesion of mitochondrial redox state produced by nitrogen dioxide, ozone and high oxygen in alveolar macrophages. *Chest*. 1974;166: 9S-12S.
32. Poli G, Schaur RJ, Siems WG, Leonarduzzi G. 4-Hydroxynonenal: a membrane lipid oxidation product of medicinal interest. *Med Res Rev*. 2008;28:569-631.
33. Mally A, Amberg A, Hard GC, Dekant W. Are 4-hydroxy-2(E)-nonenal derived mercapturic acids and 1H NMR metabonomics potential biomarkers of chemically induced oxidative stress in the kidney? *Toxicology*. 2007;230:244-255.
34. Aldini G, Vistoli G, Regazzoni L, et al. Albumin is the main nucleophilic target of human plasma: a protective role against pro-atherogenic electrophilic reactive carbonyl species? *Chem Res Toxicol*. 2008;21:824-835.
35. Ogasawara Y, Namai T, Togawa T, Ishii K. Formation of albumin dimers induced by exposure to peroxides in human plasma: a possible biomarker for oxidative stress. *Biochem Biophys Res Commun*. 2006;340:353-358.
36. Carballal S, Radi R, Kirk MC, Barnes S, Freeman BA, Alvarez B. Sulfenic acid formation in human serum albumin by hydrogen peroxide and peroxynitrite. *Biochemistry*. 2003; 42:9906-9914.
37. Bocci V. Non-specific immunomodulation in chronic heart failure. *Lancet*. 2008;371:2083.

Standard Operating Procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks

Patrizia Bernini[⊥¶], Ivano Bertini*[⊥#], Claudio Luchinat[⊥#], Paola Nincheri[¶], Samuele Staderini[¶], Paola Turano[⊥#]

[⊥]Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.

[#]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy.

[¶]FiorGen Foundation, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy.

Corresponding author:
Ivano Bertini
Magnetic Resonance Center (CERM)
University of Florence
Via L. Sacconi, 6
50019 Sesto Fiorentino, Italy
e-mail: bertini@cerm.unifi.it
Tel.: 39-055-4574270
Fax: 39-055-4574271

Abstract

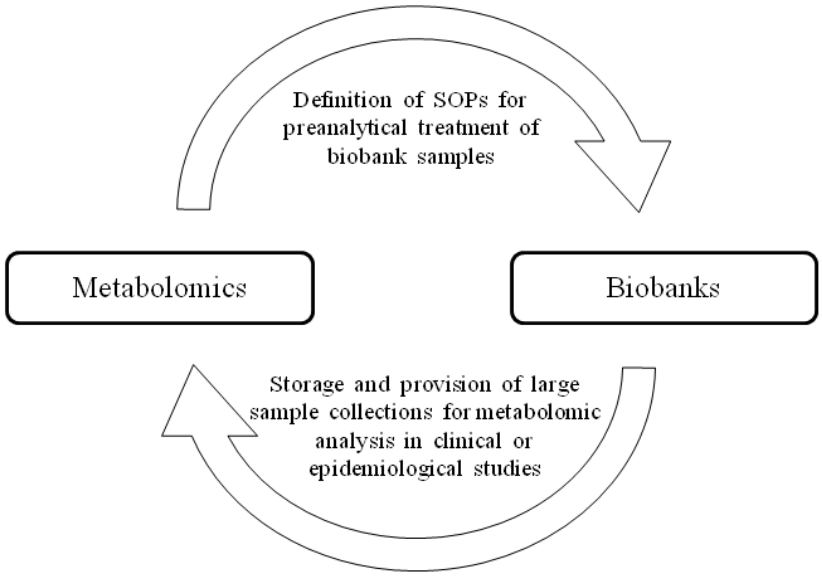
^{1}H NMR metabolic profiling of urine, serum and plasma has been used to monitor the impact of the pre-analytical steps on the sample quality and stability in order to propose standard operating procedures (SOPs) for deposition in biobanks. We analyzed the quality of serum and plasma samples as a function of the elapsed time (t = 0- 4 hours) between blood collection and processing and of the time from processing to freezing (up to 24 h). The stability of the urine metabolic profile over time (up to 24 h) at various storage temperatures was monitored as a function of the different pre-analytical treatments like pre-storage centrifugation, filtration, and addition of the bacteriostatic preservative sodium azide. Appreciable changes in the profiles, reflecting changes in the concentration of a number of metabolites, were detected and discussed in terms of chemical and enzymatic reactions for both blood and urine samples. Appropriate procedures for blood derivatives collection and urine preservation/storage that allow maintaining as much as possible the original metabolic profile of the fresh samples emerge, and are proposed as SOPs for biobanking.

Keywords: metabolomics, NMR spectroscopy, serum, plasma, urine, biobanks

Introduction

Metabolomic research is an emerging area focused on measuring the ensemble of the metabolites in biospecimens (Fiehn 2002; Nicholson et al. 1999). The objects of interest of metabolomic researches are the low-molecular weight compounds (MW < 1500 Da) that serve as substrates and/or products in various metabolic pathways. These small molecules include compounds such as lipids, sugars, amino acids, nucleotides, and a number of different organic molecules, that are the reactants, intermediates or products of biochemical reactions as well as building blocks for all other biochemical species including proteins, nucleic acids and cell membranes (German et al. 2005). The number of different metabolites in humans is unknown; estimates range from a minimum of 2,000-3,000 to a maximum of around 20,000 metabolites (browsing the Human Metabolome Database, HMBD, www.hmbd.ca), to be compared with estimated 30,000 genes (from www.ensembl.org) and about 40,000 proteins (browsing the protein database of the National Center for Biotechnology Information, NCBI, www.ncbi.nih.gov and disregarding splicing variants and post-translational modifications). Most metabolomic studies involve common biofluids as urine and serum/plasma that are obtainable from mammals, especially from humans, in a non- or minimally-invasive way and are easily available because commonly used for many other clinical analyses. A standard approach in metabolomics is that of measuring monodimensional ^{1}H NMR spectra at medium/high magnetic fields. Such spectra, often called metabolic profile, constitute a "fingerprint" of the NMR detectable part of the whole metabolome. The potential of metabolomics for disease diagnosis, prognosis and, in a clinical trial setting, for monitoring drug therapy relies on its ability to extract a disease signature from the multivariate analysis of the metabolic profiles of statistically relevant ensembles of samples derived from different donors (Oakman et al. 2010; Sreekumar et al. 2009; Bertini et al. 2009; Akira et al. 2008; Holmes et al. 2008; Bartsch et al. 2008; Gao et al. 2008; Fearnside et al. 2008; Makinen et al. 2008; Teichert et al 2008; Coolen et al. 2008; Claudino et al. 2007; Schnackenberg et al. 2007; Marchesi et al. 2007; Griffin et al. 2007; Constantinou et al. 2007; Coen et al. 2005; Lindon et al. 2004; Brindle et al. 2002). Clearly, the reliability of the approach requires that the chemical nature and the relative concentration of all the metabolites present in the biofluids are neither affected by the preanalytical treatment used to store the samples nor by the analytical methodology. Conversely, as depicted in Scheme 1, metabolomics itself is assuming a growing importance in the definition of operating procedures aimed at collecting and preserving biological samples and for standardizing protocols (Barton et al. 2008; Dunn at al. 2008; Jackson at al. 2008; Peakman et al. 2008; Saude et al. 2007). The definition of suitable Standard Operating Procedures (SOPs) is therefore essential to allow data comparison worldwide. Biobanks, that are infrastructures devoted to the collection, cataloguing and storing of biological samples in order to

make them available for medical and clinical research, represent an irreplaceable support for all those studies in which the impact of the results is linked to the large number of the collected samples. At the same time, they have to guarantee that the quality of the stored biological samples remains as close as possible to the fresh sample for any possible future studies, including metabolomics.



Scheme 1.

Among the different metabolomic tools, NMR represents the technique of choice (Nicholson et al. 2008) for the definition of pre-analytical procedures. Indeed, NMR is a high-throughput methodology requiring only minimal sample handling before spectra acquisition and allowing collection of the whole metabolic profile, i.e. the simultaneous detection of a large number of metabolites with different physiochemical properties e.g., hydrophobicity/hydrophilicity, acidity/basicity, redox reactivity. The variety of detectable molecules allows the monitoring of residual enzymatic activities and/or chemical reactions that may alter the NMR profile of the analyzed sample, which becomes not representative any longer of the metabolome before collection. These alterations could seriously bias the results of studies based on samples having different collection, treatment and storage histories.

Standardization and improvement of pre-analytical tools and procedures for in-vitro molecular diagnostics is the main goal of the European Union FP7 project SPIDIA (www.spidia.eu), which involves a consortium of public research organizations and private companies, and a standards organization. As a partner of the consortium, we performed the present research focusing on the most commonly studied biofluids: urine, serum and plasma.

Urine might be regarded as a biospecimen with little scientific value, but it is assuming an increasing importance as a bio-bankable sample thanks to metabolomics studies. Browsing PubMed for metabolomics/metabonomics publications resulted in 250 publications based on urine, 147 publications on serum, 185 on plasma and 185 on tissues. Use of urine as a first choice object of study is justified by the fact that the collection method is simple and is the least invasive one: these characteristics permit multiple collection schemes that greatly improve the reliability of the statistical analysis. Indeed, some of us have recently shown that NMR fingerprinting of multiple urine samples from the same donors is able to reveal individual metabolic phenotypes (Assfalg et al. 2008, Bernini et al. 2009). Several methods of urine sample processing (centrifugation, filtration, addition of preservatives) as well as sample storage temperature (at -80 °C or in liquid nitrogen) were tested in several combinations and found to have significant effects on the metabolome, as detected by ^{1}H-NMR fingerprinting. The presence of host cells or bacterial cells in urine is identified as a potential major source of alteration of the metabolic profile, and appropriate countermeasures emerge.

Blood derivatives like serum and plasma are also very common biofluids in metabolomics studies. The disadvantage with respect to urine is that the collection of blood samples is slightly more invasive. Therefore, it may be more difficult to obtain multiple collections from patients, and even more from healthy volunteers. On the other hand, blood is less affected by daily variations and daily diet than urine. We analyzed here the quality of serum and plasma samples as a function of the time delay (0-4 h) and storage temperature (25 °C and 4 °C) before processing. Moreover, changes in the NMR profiles of these two biofluids after serum and plasma preparation were monitored for a further 24 hours, to simulate the effect of the time between processing and freezing. A number of metabolites which are often used as disease biomarkers in metabolomics studies are heavily affected by the tested variation factors, and again, appropriate countermeasures emerge.

The present results, taken together, contribute to the definition of standard operating procedures for specimen collection that represent a step forward with respect to the available recommendations for biobanking procedures (Yuille et al. 2010). These SOPs are being implemented at the local da Vinci European Biobank (www.davincieuropeanbiobank.org) and proposed for adoption by the European

Biobanking and Biomolecular Resources Research Infrastructure (BBMRI, www.bbmri.eu) presently in the preparatory phase.

Materials and methods

Sample collection

Urine samples were collected from 6 healthy donors (3 females and 3 males) following the standard procedure for urine culture and kept not more than a half hour at 2-8 °C before further processing (as detailed in the Results section). Urine from each donor was aliquoted in 1.0 mL fractions, and each aliquot was treated independently. Different centrifugation speeds were tested (see Results and Discussion); during centrifugation samples were kept at 4 °C. For filtration a 0.20 µm filter was used; such a cut-off is small enough to ensure removal of cells and large particles still avoid problems of obstruction of the pores that may slow down urine processing and lead to sample loss. Sodium azide was added to the sample until a final concentration of 3 mM, as described in our previous metabolomics studies (Weckwerth et al. 2007, Assfalg et al. 2008, Bertini et al. 2009, Bernini et al. 2009)

Blood samples were withdrawn from healthy donors using different BD Vacutainers® for plasma or serum isolation: K2E 5.4 mg and SST™ II *Adavance* respectively. Multiple 5 mL aliquots were taken at the same time from each patient and treated in parallel afterwards. We tested different time delays (0-4 h) between blood withdrawal and processing for serum/plasma preparation; during these time intervals the vacutainers were incubated at 4 °C and 25 °C. In the case of serum t=0 was considered to be 30' at room temperature after blood collection, as required for clotting. After the incubation, vacutainers for plasma collection were centrifuged at 820 RCF (Relative Centrifugal Force) for 10' at 4 °C, while vacutainers for serum collection were centrifuged at 1500 RCF for 10' at 25 °C. In both cases, after centrifugation, the supernatants were collected and used for the preparation of NMR samples.

NMR sample preparation

According to a commonly used protocol (Assfalg et al. 2008, Bernini et al. 2009), urine samples were shaken before use and 630 µL were centrifuged at 14000 RCF for 5 min. 540 µL of the supernatant were added to sixty microliters of potassium phosphate buffer (buffer 1: 1.5 M K_2HPO_4 and 10 mM sodium trimethylsilyl [2,2,3,3-2H4] propionate, TMSP, in 100% ^2H_2O, pH 7.4). 540 µL of the mixture were pipetted into 4.25 mm NMR tubes (Bruker BioSpin srl).

Blood-derivative NMR samples were prepared adding 300 µl of phosphate sodium buffer (buffer 2: 70 mM Na_2HPO_4, 38 mM NaN_3 and 55 mM TMSP in 20% ^2H2O, pH 7.4) to 300 µl of plasma or serum. A total of 450 µl of this mixture was transferred into a 4.25 mm NMR tube (Bruker BioSpin

srl). In the follow-up of blood derivates over time, we used a buffer with the same composition and concentration described above.

NMR spectra

All ^{1}H-NMR spectra were acquired using a Bruker 600 MHz spectrometer operating at 600.13 MHz proton Larmor frequency and equipped with a 5 mm CPTCI ^{1}H-^{13}C/^{31}P-^{2}H cryo-probe including a *z*-axis gradient coil, an automatic tuning-matching (ATM) and an automatic sample changer. A PT 100 thermocouple served for temperature stabilization at the level of approximately ±0.1 K at the sample. Before measurement, samples were kept for 3-5 min inside the NMR probehead, for temperature equilibration i.e., 300.0 K for urine samples, 310.0 K for serum/plasma samples; the highest temperature for blood derivatives is used for better detection of lipidic profiles (Suna et al. 2006).

For each urine sample, a one-dimensional (1D) NMR spectrum was acquired with water peak suppression using a standard pulse sequence (NOESYpresat; Bruker), 64 scans, 64 k data points, a spectral width of 12,019 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s and a mixing time of 100 ms. ^{1}H-^{1}H J-resolved (J-res) spectra were also acquired in order to get more information about signal multiplicity and coupling patterns. For each serum/plasma sample 1D NOESYpresat spectra were recorded; 1D spectra were acquired also with the Carr-Purcell-Meiboom-Gill (CPMG; Bruker) spin-echo sequence to suppress signals arising from high molecular weight molecules. In this case NOESYpresat spectra consisted of 64 scans, 98 k data points, a spectral width of 18,028 Hz, an acquisition time of 2.7 s, a relaxation delay of 4 s and a mixing time of 100 ms. CPMG spectra consisted of 64 scans, 74 k data points, a spectral width of 12,019 Hz, an acquisition time of 3.1 s, a relaxation delay of 4 s and a mixing time of 100 ms.

Spectral processing and analysis

Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated using TopSpin 2.1 (Bruker Biospin srl). In the case of serum and plasma, spectra were calibrated assigning a value of 1.5 ppm to the alanine peak. Calibration of urine spectra was achieved by aligning the signal of the pH-invariant resonance of trigonelline at 9.12 ppm; with this approach, the other signals of trigonelline and those of other pH-insensitive metabolites such as 1-methylnicotinammide and hippurate do not shift with pH. Each 1D spectrum in the range between 0.02 and 10.00 ppm was segmented into 0.02-ppm chemical shift bins, and the corresponding spectral areas were integrated using AMIX software (Bruker BioSpin). Regions between 6.0 and 4.5 ppm containing residual water and urea signals

were removed. The total spectral area was calculated on the remaining bins and normalization was carried out on the data prior to pattern recognition.

All resonances of interest were assigned on template one-dimensional NMR profiles by using matching routines of AMIX 3.8.4 (Bruker BioSpin) in combination with the BBIOREFCODE (Version 2-0-0; Bruker BioSpin) reference database and published literature when available. All calculations were made using MATLAB 7.6.0 and AMIX 3.8.4.

Statistical analysis

Various kinds of multivariate statistical techniques were applied on the obtained buckets: PLS (Partial Least Squares)-regression and PCA (Principal Component Analysis) for data analysis, manova (One Way analysis with Multiple Responses) for data reduction, kNN (k-Nearest Neighbour) as classification method, Kruskal-Wallis test (for the determination of the meaningful metabolites) (Madsen at al. 2010).

Urine culture tests

Urine culture tests were performed by using Linearcount6® Urinocoltura: this is an integrated system that allows the bacterial count, the bacterial isolation and the presumed identification of bacteria present in the urine.

Results and discussion

Urine The starting point for this research are the Recommendations on Biobanking Procedures for urine processing and management recently published by the European Consensus Expert Group Report (Yuille et al. 2010). According to this document, biobanking procedures for urine should consider the following general consensus recommendations: i) cells and particulate matter should be removed (e.g. by centrifugation); ii) samples should be stored at -80 °C or below; iii) time limits for the processing should have been defined experimentally and should be appropriate to the analytes to be measured; iv) unless specified for a particular downstream analysis, urine samples should be stored without additives. As the authors pointed out, advances in new downstream analytical platforms may impose specific requirements related to the nature of the targeted analytes, thus leading to specific revision/elaboration of the general procedures. Here we address the above critical issues in relation to the consequences for downstream metabolomics applications.

Fresh urine is characterized by the presence of human cells (erythrocytes, leucocytes, urothelial cells, epithelial cells), bacteria, fungi, sperm counts, non cellular components (mucus filaments, cylinders, cylindroids, pseudocylindres, crystals, urates). The protocol for the analysis of urinary sediment proposes a speed around 1600 RCF for the centrifugation (Koukoulaki et al. 2008) in order to avoid the breaking of the frailest elements like cylinders. On the other hand, to avoid the

presence of any particulate components in the NMR samples that will affect the spectral quality, standard protocols (Beckonert et al. 2007) for the immediate NMR analysis of urine require a centrifugation of the sample at 14000 RCF for 5 minutes at 4 °C. Such a high speed may induce cell breaking and release of cellular components into the biofluid. We thus analyzed the advantage of applying a mild centrifugation (pre-centrifugation) to fresh urine before the high-speed centrifugation needed by the preparation of NMR samples for immediate use. Multiple aliquots from two different donors were collected and treated, as detailed below and summarized in Fig. 1: one aliquot per donor had no pre-centrifugation, the other aliquots were pre-centrifuged at 450, 1000, 3000 and 11000 RCF (Fig. 1A). After pre-centrifugation, the supernatant was recovered and analyzed immediately, following the protocol described in the section *Materials and Methods* for the preparation of the NMR samples. PLS (Partial Least Squares)-regression coupled with CA (Canonical Analysis) was performed on the obtained spectra as a function of the pre-centrifugation speed. To visualize the distance between the spectra of differently treated samples, the value of the first component for the non pre-centrifuged sample was taken equal to zero and the absolute value of the difference between the value of the first component in the pre-centrifuged samples and in the non pre-centrifuged sample was calculated. The plot of these values as a function of the pre-centrifugation speed (Fig. 2) shows an interesting behavior. The most distant from non pre-centrifuged samples are the aliquots pre-centrifuged at a speed in the 1000-3000 RCF range, while aliquots pre-centrifuged at >11000 RCF are slightly closer to the samples that do not undergo any pre-centrifugation. Apparently, the proposed 5-minutes 14000 RCF-centrifugation of the NMR protocol is too harsh; speeds ≥ 11000 RCF at the pre-centrifugation level induce at least partial breaking of the cellular components and release of their soluble constituents, thus affecting the NMR profile. If a mild pre-centrifugation (between 1000-3000 RCF, Fig. 2) is applied, the cellular components are spun down and the subsequent 14000 RCF centrifugation of the supernatant is useful for the elimination of other suspended, mainly inorganic, particles. On the other hand, a too mild pre-centrifugation (450 RCF) is not effective in eliminating the cellular components. The importance of the pre-centrifugation step of course depends upon the content in cellular components of the analyzed samples, and may therefore be very different from one individual to another (Fig. 2).

The differences in metabolic profile between pre-centrifuged or non pre-centrifuged samples are essentially ascribable to chemical shift changes in signals that are extremely sensitive to pH variations such as the Hε1 signal of histidine and of N^{τ}- and N^{π}-methyhistidine (Fig. S1A). For example, using a pre-centrifugation speed in the range defined by us as mild pre-centrifugation, namely 2500 RCF, the largest chemical shift variation was observed for the well resolved (in this

pH range) Hε1 signal of N^{τ}-methyhistidine and found to be statistically relevant by Kruskal-Wallis test. The observed changes are consistent with a pH increase (Fig. S1B). The extent of the observed pH variation is different from one donor to another and is larger for samples containing larger amount of cells. However, no resonances attributable to new metabolites or increase in intensity of existing signals could be observed. This suggests that soluble components released by cells in fresh urine are below the detection limit of the method (because they are low in concentration and/or are high molecular weight components) but their presence is still able to alter the physiochemical properties of the solution.

The presence of cellular components has an impact on the NMR profiles even if non-fresh urine samples are analyzed, and the importance of the observed change depends upon the storage temperature (Fig. 1B). NMR metabolic profiles of non pre-centrifuged samples stored for a week at –80 °C do differ from those that have undergone a mild (600-2500 RCF) pre-centrifugation before freezing at the same temperature. The effect is less severe if urine samples are stored in liquid nitrogen. This result is not unexpected because rapid freezing in liquid nitrogen i.e., below the critical ice crystal temperature (–130 °C) avoids crystal formation which may cause cell breaking. For example, from a PLS analysis of the spectral shift variation for samples not pre-centrifuged and pre-centrifuged at 2500 RCF and analyzed either i) fresh, or ii) after one week at -80 °C and iii) after one week at liquid nitrogen (Fig. 1B), it results that the largest difference occurs for the samples stored at -80 °C. Assigning a value of 100% to the changes observed between pre-centrifuged and not pre-centrifuged samples stored at -80 °C for a week, the corresponding changes between pre-centrifuged and not pre-centrifuged samples analyzed fresh or after a week of storage in liquid nitrogen are much smaller and of comparable extent (9% and 8%, respectively).

PLS analysis of the discussed samples shows that the recognition capability of the model, i.e. the accuracy in assigning different aliquots to the same sample, is not seriously affected by any of these treatments (Fig. S2). Nevertheless, reducing any possible external sources of variations is vital for specific studies whose results depend on revealing differences in the relative concentration of a limited set of metabolites.

Finally, we evaluated the two most commonly used preservation methods (Saude and Sykes 2007): filtration and addition of sodium azide. They were used as such or in combination with a pre-centrifugation step at 1000 RCF, as summarized in Fig. 1C. Differently treated aliquots of the same sample were analyzed either fresh or after storage for a week at -80 °C, i.e. at the temperature that is most commonly used in biobanking. Spectral variations were followed for 24 h after urine processing or thawing after a week of storage at -80°C. Changes over time are observed for pH-sensitive metabolites that undergo chemical shift variations. In particular, the shift variations for the

Hε1 N-methylhistidine (Fig. S1) are consistent with an alkalinisation of the samples; the chemical shifts of the resonances of xantine also change with time. Changes in the relative concentration of some molecules are also observed: succinate and acetate increase with time; urea, lactate and glutamate/glutamine decrease (Fig. S3). The extent of their variations depends upon the preservation method, pre-centrifugation and storage. Consistent with available literature data (Saude and Sykes 2007), filtration is the method that makes the NMR spectra more stable over time.

In order to evaluate the origin of the changes that affect the NMR spectra of samples over time, we tested three working hypothesis

- The chemical hypothesis: reactions (e.g. oxidations) occur with time.
- The bacterial hypothesis: the bacteria present in urine grow over time with the consequent production of bacterial metabolites.
- The enzymatic hypothesis: the enzymatic activities of urine cause the consumption of certain metabolites and the increase of others.

In order to test the chemical hypothesis, as far as possible oxidation reactions are concerned, the samples were kept under inert atmosphere for the entire duration of the experiment and compared with the corresponding aliquots kept under normal atmosphere. The effect of the inert atmosphere over the 24 hours in which the samples were followed is modest. The only difference is a slight reduction in the increase of succinate concentration.

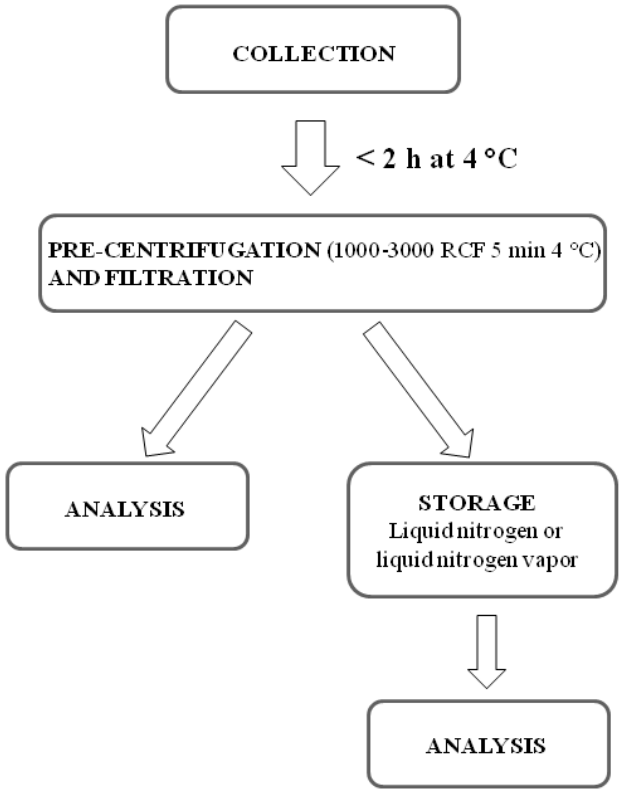
In order to test the bacterial hypothesis, in parallel with the analysis by NMR, we performed urine culture tests. The results of these tests (Tables S1 and S2) show that pre-centrifugation, addition of sodium azide and storage at -80 °C have a disinfectant effect, but they do not eliminate completely the bacterial load. The effect of sodium azide is more clearly visible after 6 hours at room temperature, whereas the effect of storage is more clearly visible when temperature-sensitive bacteria like coliforms are present. The fact that non-precentrifuged samples have the highest decrease in glutamate concentration suggests that degradation of this metabolite is mainly attributable to bacteria. The reduced degradation of lactate in samples stored at -80 °C can be explained considering that coliforms bacteria, sensitive to the low temperatures, are the main source of its transformation. Filtration, alone or in combination with pre-centrifugation, completely eliminates bacteria from urine; nevertheless NMR spectral changes are not fully quenched in filtered samples. Therefore, the bacterial growth is not the only source of the observed spectral changes occurring with temporal progression.

The enzymatic hypothesis is justified by the notion that some activities due to human enzymes such as γ-glutamyltransferase, alkaline phosphatase and N-acetylglucosaminidase are present in the urine of healthy adults (Jung and Pergande 1983). The presence of γ-glutamyltransferase is a good proof

of the disappearance of glutamate/glutamine in urine with time. On the other hand, a residual activity of urease and of isocitrate lyase, two metalloenzymes of bacterial origin, could be the cause of the decomposition of urea and of the increase of succinate, respectively. In order to test this hypothesis, we tried two different inhibitors. A 50 mM concentration of acetohydroxamic acid (AHA), a known inhibitor of urease, does inhibit the decrease of urea. A similar concentration of the bivalent metal chelator EDTA is also effective in inhibiting urea decomposition and partially inhibits the increase of succinate, acting on both urease and isocitrate lyase. The increase of acetate is also inhibited by EDTA. Because acetate is the main product of several bacterial fermentations, this result suggests that also the increase of acetate is at least in part due to metal-dependent bacterial enzymatic activities.

The three working hypotheses suggested above help us in understanding the observed variations of the NMR signals, but it is also important to highlight that all three mechanisms are present and their effects may partially overlap.

On the basis of the results of our experiments, we propose the following procedures for the optimal processing and management of urine samples to maintain as much as possible their original metabolome (Scheme 2): i) removal of cells and particulate matters through the combined use of a mild pre-centrifugation 1000-3000 RCF (5' at 4 °C) and filtration; ii) long-term storage of samples in liquid nitrogen (or liquid nitrogen vapour) to avoid breaking of residual cells; iii) fast processing (within 2 hours for collection); iv) storage at 4 °C between collection and processing. Recommendations at points iii) and iv) aim at reducing the effects of any possible enzymatic/cellular activities. Addition of additives (like enzyme inhibitors) should be avoided because the required concentrations will introduce signals in the NMR spectra covering the resonance of metabolites and may also induce changes in pH, ionic strength, etc. thus further affecting the original NMR profiles.



Scheme 2.

With respect to the general recommendations in Yuille et al., our studies permitted the definition of the best practices for the removal of cells and particulate matter. Nevertheless, as complete removal of cells is never achieved, long-term storage temperatures below the critical ice crystal temperature would be advisable. Changes in metabolite concentrations at room temperatures are relatively fast (Fig. S3); reduction of time between collection and processing and the temperature at which urine samples are kept during this time delay is critical for the maintainance of the metabolome of the fresh samples. Annotation of time delays and temperatures for the entire history of each sample will add statistical value.

Blood

According to the same recommendations document mentioned in the urine section (Yuille et al. 2010), in the plasma/serum processing and management one should consider the following aspects: i) use of EDTA or citrate as anticoagulant, not heparin; ii) sample storage at a temperature of -80 °C, or below if e.g. critical ice crystal formation is problematic; iii) recording of the time from

collection through processing; iv) experimental definition of time limits that are appropriate to the analytes to be measured. These guidelines for blood derivatives biobanking focus on the use of stored samples for future extraction of DNA for genetic analysis. Although the latter still represents the main application for biobanked serum and plasma, new analytical platforms are assuming increasing importance and specific requirements for appropriate samples for new biomarker analysis have to be defined. Here we test the validity of best practices proposed for DNA analysis in the context of metabolomics and evaluate whether additional procedures may be needed for the optimal maintainance of the original metabolome.

The time limits for processing whole blood were examined (Fig. 3A) by evaluating the effect on the NMR metabolic profile of serum and plasma-EDTA samples, prepared according to the standard protocols described in Materials and Methods. PLS-CA analysis was again used at the descriptive level and showed that the spectral patterns of serum and plasma samples processed at different times carry meaningful differences (Fig. 4A,B and D,E). The PLS analysis with information of time delays from collection through processing was performed assuming the mean of the value of the first component as zero at time zero and consequently calculating the values corresponding to the following tested times. Plots C and F in Fig. 4 show the changes in the main component of PLS for serum and plasma samples at 4 °C and at 25 °C. Degradation processes are time-dependent and temperature-dependent both for serum and plasma samples: in both cases incubation at 25 °C causes deeper changes in the NMR profile. In order to unravel the main mechanisms at the basis of the degradation processes occurring from collection through processing, we singled out the significant buckets ($P \leq 0.05$) by Kruskal-Wallis test. They correspond to the resonances of glucose, lactate and pyruvate. The intensities derived from integrals of the NMR peaks of these metabolites were normalized with respect to the alanine signal at 1.5 ppm, which has a constant intensity under the tested conditions. Average values are reported in Fig. S4. As a general trend, a decrease in glucose concentration is observed. Glycolysis is expected to represent about 90% of glucose consumption by erythrocytes (Baynes et al. 2010). The decrease in glucose concentration is more important in serum, most probably because EDTA in plasma tubes exerts an inhibitory effect on metalloenzymes and metal-dependent enzymes involved in glycolysis. An increase in lactate, the end product of the anaerobic glycolysis is observed, which is steeper than the decrease in glucose, roughly in agreement with the expected stoichiometry of the reaction. More complex is the trend in the concentration of pyruvate that is an intermediate metabolite of glycolysis: in plasma it is slightly decreasing/constant at 4 °C and it increases at 25 °C; in serum, it decreases at 4 °C and is almost constant at 25 °C. In clinical analyses, the standard for the measurement of glycaemia is represented by fluoride/oxalate coated vacutainers. It has been reported that glucose concentrations in plasma

from blood collected into EDTA coated vacutainers showed no significant differences with glucose concentrations in plasma from fluoride/oxalate coated vacutainers (which represent the standard for the measurement of glycaemia) up to 36 h (Peakman and Eliott 2008; Jackson et al. 2008). Comparison of the changes in concentration of glucose and its derivatives during the time from collection through processing of citrate, EDTA and fluoride/oxalate plasma in the present work did not show meaningful differences. Therefore, none of these three anticoagulants solves the problem of glucose degradation and we can just suggest that for a reliable evaluation of glucose, lactate and pyruvate whole blood should be processed within 2 h from collection, keeping it at 4 °C, although the latter may still not be the best procedure for pyruvate.

The choice of the anticoagulant for metabolomic studies of plasma via NMR is not so obvious and represents a study-specific additional procedure. Citrate and EDTA are commonly used but have some drawbacks. Citrate is by itself a metabolite, identified as a relevant biomarker as it is involved in the energy metabolism and found to undergo changes in our degradation studies. On the other hand EDTA is not present in blood, but its presence as an anticoagulant covers a number of metabolite resonances (choline, dimethylamine and one signal of citrate). From the point of view of NMR metabolomics, use of fluoride/oxalate coated vacutainers as plasma collection tubes would have the advantage that the anticoagulant does not introduce any additional signal in the ^{1}H NMR spectra (however, see additional discussion below).

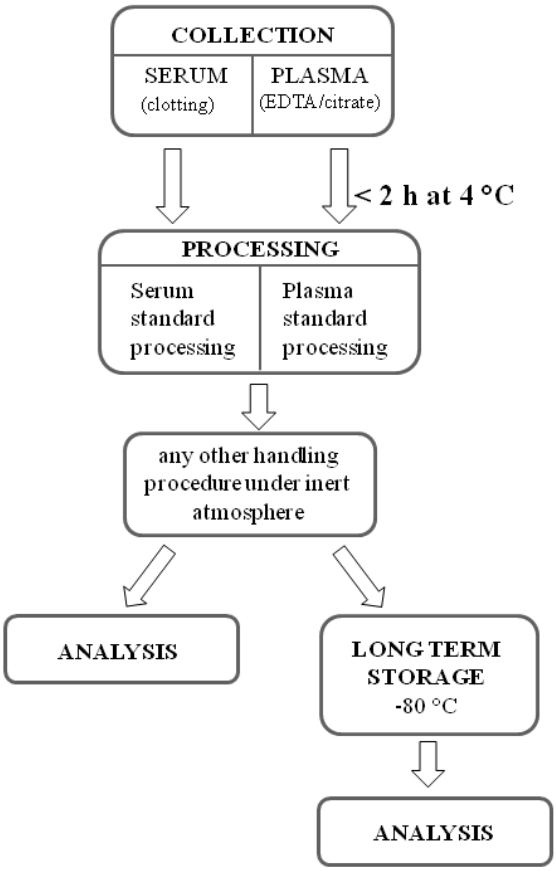
Possible degradation processes occurring in blood derivatives after processing were analyzed using 5 plasma and 5 serum samples of different patients (Fig. 3B). Each sample was split in multiple aliquots and NMR spectra of each aliquot were acquired in the time range 0-24 hours keeping the samples at room temperature.

A PLS-CA analysis of the CPMG spectra of plasma and serum samples acquired over time provided the clustering reported in Fig. 4G and H; a better discrimination is observed for serum. Although the PLS-CA first component of the spectra gradually shift in the same direction during the 24 h of monitoring time, the effects are larger for serum (Fig. 4I). In order to single out the significant buckets ($P \leq 0.05$), and thus the corresponding metabolites that are responsible of this time progression, we performed the Kruskal-Wallis test. From this procedure we identified the following metabolites as markers of the time changes (Fig. S5): triglycerides, proline, choline, citrate and histidine. The Hε1 signal of His shifts with time, consistently with a pH variation increase of about 0.1 units. The signals of the other metabolites, but citrate, decrease in intensity. Albumin and LDL/VLDL (low-density-lipoproteins/very-low-density lipoproteins) also contribute to the variations of the NMR profile in NOESY spectra; although an accurate quantitation of the observed effect for such broad signals is not possible, a clear decreasing trend is observed. The

decrease in concentration for fatty acids and LDL/VLDL is attributable to oxidation reactions, as the effect is attenuated if samples are kept under inert atmosphere. The chemical shift variations observed for the signals of citrate (Fig. S6) are consistent with literature data about saliva (Silwood et al. 2002) where the observed behavior has been rationalized in terms of a competition between calcium(II) and magnesium(II) binding to this chelator. Consistently, no chemical shift changes are detected in the present work for citrate in plasma EDTA-samples. The decrease in concentration observed in serum of all these molecules, but proline and choline, parallels that observed in plasma. No meaningful decrease in proline concentration is visible in plasma EDTA and citrate samples, whereas proline decreases sizably in plasma fluoride/oxalate samples. The latter observation indicates that EDTA or citrate are still preferable with respect to fluoride/oxalate-containing samples. The signals of choline is covered by EDTA resonances and its behavior cannot be followed. In summary, plasma appears to be slightly more stable than serum when kept at room temperature.

Collection, transport, handling, storage and analysis of serum and plasma samples may imply different times of exposure to light for different samples. In order to verify whether this translates into meaningful changes in the ^{1}H NMR metabolic profiles, spectra have been acquired to monitor possible differences between samples kept in the dark or exposed to light. A set of 7 serum samples was used: after addition of the buffer, each sample was split into two aliquots. One of them was utilized to follow the degradation mechanisms at room temperature in the light and the other while kept in the dark. Multiple and consecutive NMR spectra were acquired over a total period of 24 hours. Exposure to light was found not to induce any difference in the degradation processes that occur with time.

Based on the above findings, we propose the following procedures for the optimal processing and management of blood and its derivatives when they have to be used for future metabolomic NMR studies (Scheme 3): i) either EDTA or citrate can be used as anticoagulant for plasma samples; ii) time between collection and processing should not exceed 2 h; iii) during this time delay samples should be kept at 4 °C; iv) any handling procedure of plasma and serum should be performed under inert atmosphere; v) samples should be frozen immediately after processing; vi) long term storage should be done at -80 °C (no need of lower temperature once erythrocytes have been removed).



Scheme 3.

Conclusions

The primary objective of biobanks is not merely archiving but also distributing conserved and documented biological samples for research. The quality of biological samples is crucial for the outcome of subsequent studies. Sample quality is directly related to pre-analytical variations, whose impact depends on its end use. The molecules constituting the metabolome are generally more sensitive to handling and storage procedures than e.g. nucleic acids; in addition changes in metabolites due to residual enzymatic activity in biofluid samples can be extremely fast. The objective is to ensure that the analyzed sample is representative as much as possible of the metabolome before collection, i.e. is as close as possible to the freshly collected sample.

Validation of methods for sample collection/handling of human biofluids is essential. As samples are generally not collected in the confines of well-regulated academic laboratories for research purposes but typically in clinics or even in a domestic environment (urine) where organization and priorities are not those of future possible non-standard analytical approaches like metabolomics, time and temperature from collection through processing should be carefully recorded.

Urine and whole blood result to be “living fluids” even after collection from the patient: presence of cells and residual enzymatic activities bring about significant changes in the specimen from the point of collection through the time of analysis.

In order to quench bacterial/enzymatic activity in urine a mild pre-centrifugation combined with filtration results to be the safest way to avoid contamination of the metabolome with soluble molecules derived from cellular components. This is particularly important when urine samples need to be stored in biorepositories at -80 °C; storage in liquid nitrogen would be preferred but may become too expensive in a biobanking environment, if it has to be applied for a the long-term storage of large collections of urine samples. Urine did not result to be sensitive to the presence of oxygen, but a number of enzymatic reactions occur in this biofluid. Addition of inhibitors is able to quench/slow-down such activities but invariably alters the metabolic profile as it introduces in the spectra signals of the added molecules, whose presence may also induce changes in pH, ionic strength, further altering the whole spectral properties. In this sense, keeping the sample at the lowest possible temperature from collection throughout the analysis is extremely important.

Whole blood is generally collected in vacutainers that protect it from oxygen. The effect of time and temperature from blood collection through processing are revealed by the alterations in the relative concentration of three very important metabolites i.e. glucose, lactate and pyruvate, which are attributable to erythrocyte activity. These metabolites often play a key role in determining the signature of a disease in metabolomics studies (Warburg 1956, Garber 2004, Shaw 2006, Sreekumar et al. 2009), for example low glucose and enhanced lactate in cancer are indicators of the enhanced glycolysis expected on the basis of the Warburg effect (Qiu et al. 2009). The variability of these metabolites thus represents a real pitfall in metabolomic studies, as blood is usually collected in clinics and then transferred to analytical departments where it is processed: samples from different patients are collected the same day at different times and then delivered all together to the processing. It is also important that samples are all kept at a controlled temperature.

For serum, the processing time has to be reduced to a minimum (after the canonical 30’ at room temperature needed for clotting) and must never to be longer than 2 hours; storage for any time after clotting and before processing should be at 4 °C. The same holds for plasma.

After blood processing, glucose, lactate and pyruvate concentrations remain stable over time, but the presence of oxygen causes changes in a number of other metabolites (albumin, triglycerides, LDL/VLDL, proline, citrate and histidine). Processing and sample handling under inert atmosphere and storage and analysis under inert atmosphere would be advisable.

Acknowledgements

Research funded by the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 222916. Manfred Spraul and Hartmut Schäfer (Bruker BioSpin) are acknowledged for many discussions over the years. We thank Monica Biondi for her assistance in blood collection.

References

Akira K, Masu S, Imachi M, Mitome H, Hashimoto M, Hashimoto T (2008) ^{1}H NMR-based metabonomic analysis of urine from young spontaneously hypertensive rats. J Pharm Biomed Anal 46:550-556.

Assfalg M, Bertini I, Colangiuli D, Luchinat C, Schäfer H, Schütz B, Spraul M (2008) Evidence of different metabolic phenotypes in humans. Proc Natl Acad Sci U.S.A. 105: 1420-1424

Barton RH, Nicholson JK, Elliott P, Holmes E (2008) High-throughput ^{1}H NMR-based metabolomic analysis of human serum and urine for large-scale epidemiological studies: validation study. Int J Epidemiol 37: i31-i40

Bartsch T, Alfke K, Wolff S, Rohr A, Jansen O, Deuschl G (2008) Focal MR spectroscopy of hippocampal CA-1 lesions in transient global amnesia. Neurology 70: 1030-1035.

Baynes JW, Dominiczak MH (2010) Medical Biochemistry. Elsevier, Amsterdam.

Beckonert O, Keun HC, Ebbels TMD, Bundy J, Holmes E, Lindon JC, Nicholson JK (2007) Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. Nature 11: 2692-2703

Bernini P, Bertini I, Luchinat C, Nepi S, Saccenti E, Schäfer H, Schütz B, Spraul M, Tenori L (2009) Individual human phenotypes in metabolic space and time. J Proteome Res 8: 4264-4271

Bertini I, Calabrò A, De Carli V, Luchinat C, Nepi S, Porfirio B, Renzi D, Saccenti E, Tenori L (2009) Metabonomic signature of celiac disease. J Proteome Res 8: 170-177

Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK, Bethell HW, Clarke S, Schofield PM, McKilligin E, Mosedale DE, Grainger DJ (2002) Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. Nat Med 8: 1439-1444.

Claudino WM, Quattrone A, Biganzoli L, Pestrin M, Bertini I, Di Leo A (2007) Metabolomics: available results, current research projects in breast cancer, and future applications. J Clin Onc 25: 2840-2846

Coen M, O'Sullivan M, Bubb WA, Kuchel PW, Sorrell T (2005) Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. Clin Infect Dis 41:1582-1590

Coolen SA, Daykin CA, van Duynhoven JP, van Dorsten FA, Wulfert F, Mathot J, Scheltinga MR, Stroosma O, Vader H, Wijnen MH (2008) Measurement of ischaemia-reperfusion in patients with intermittent claudication using NMR-based metabonomics. NMR Biomed 21: 686-95

Constantinou MA, Theocharis SE, Mikros E (2007) Application of metabonomics on an experimental model of fibrosis and cirrhosis induced by thioacetamide in rats. Toxicol Appl Pharmacol 218:11-19

Dunn WB, Broadhurst D, Ellis DI, Brown M, Halsall A, O'Hagan S, Spasic I, Tseng A, Douglas BK (2008) A GC-TOF-MS study of the stability of serum and urine metabolomes during the UK Biobank sample collection and preparation protocols. Int J Epidemiol 37: i23-i30

Fearnside JF, Dumas ME, Rothwell AR, Wilder SP, Cloarec O, Toye A, Blancher C, Holmes E, Tatoud R, Barton RH, Scott J, Nicholson JK, Gauguier D (2008) Phylometabonomic patterns of adaptation to high fat diet feeding in inbred mice. PLoS ONE 3:e1668.

Fiehn O (2002) Metabolomics – the link between genotypes and phenotypes. Plant Mol Biol 48: 155-171

Gao XX, Ge HM, Zheng WF, Tan RX (2008) NMR-based metabonomics for detection of Helicobacter pylori infection in gerbils: which is more descriptive. Helicobacter 13: 103-111

Garber K (2004) Energy boost: the Warburg effect returns in a new theory of cancer. J Natl Cancer Inst 96: 1805-1806

German JB, Hammock BD, Watkins SM (2005) Metabolomics: building on a century of biochemistry to guide human health. Metabolomics 1: 3-9

Griffin JL, Scott J, Nicholson JK (2007) The influence of pharmacogenetics on fatty liver disease in the wistar and kyoto rats: a combined transcriptomic and metabonomic study. J Proteome Res 6: 54-61

Holmes E, Loo RL, Stamler J, Bictash M, Yap IKS, Chan Q, Ebbels T, Delorio M, Brown IJ, Veselkov KA, Daviglus ML, Kesteloot H, Ueshima H, Zhao L, Nicholson JK, Elliott P (2008) Human metabolic phenotype diversity and its association with diet and blood pressure. Nature 453: 396-401

Jackson C, Best N, Elliott P (2008) The UK Biobank sample handling and storage validation studies. Int J Epidemiol 37: i2-i6

Jung M, Pergande M (1983) Particulate and free enzyme activity in urine as a result of the shedding of brush-border membranes from kidney. Clin Chem 29(2): 392-393

Koukoulaki M, Donovan MO, Pursglove S, Alexopoulou D, Hadjiconstantinou V, Drakopoulos S (2008) Prospective study of urine cytology screening for BK polyomavirus replication in renal transplant recipients. Cytopathology 19: 385-388

Lindon JC, Holmes E, Nicholson JK (2004) Toxicological applications of magnetic resonance. Prog NMR Spectrosc 45: 109-143

Madsen R, Lundstedt T, Trygg J (2010) Chemometrics in metabolomics—A review in human disease diagnosis. Anal Chim Acta 659(1-2): 23-33

Makinen VP, Soininen P, Forsblom C, Parkkonen M, Ingman P, Kaski K, Groop PH, Ala-Korpela M (2008) ^{1}H NMR metabonomics approach to the disease continuum of diabetic complications and premature death. Mol Syst Biol 4: 167

Marchesi JR, Holmes E, Khan F, Kochhar S, Scanlan P, Shanahan F, Wilson ID, Wang Y (2007) Rapid and noninvasive metabonomic characterization of inflammatory bowel disease. J Proteome Res 6: 546-551

Nicholson JK, Lindon JC (2008) Metabonomics. Nature 455: 1054-1056

Nicholson JK, Lindon JC, 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

Oakman C, Tenori L, Biganzoli L, Santarpia L, Cappadodona S, Luchinat C, Di Leo A (2010) Uncovering the metabolomic fingerprint of breast cancer. Int J Biochem Cell Biol, in press.

Peakman TC, Elliott P (2008) UK Biobank Pilot Study: stability of haematological and clinical chemistry analytes. Int J Epidemiol 37 :i16-i22

Qiu YP, Cai GX, Su MM, Chen T, Zheng X, Xu Y, Ni Y, Zhao A, Xu LX, Cai S, Jia W (2009) Serum Metabolite Profiling of Human Colorectal Cancer Using GC-TOFMS and UPLC-QTOFMS. J Proteome Res 8: 4844-4850

Saude EJ, Sykes BD (2007) Urine stability for metabolomic studies: effects of preparation and storage. Metabolomics 3: 19-27

Schnackenberg LK, Sun J, Espandiari P, Holland RD, Hanig J, Beger RD (2007) Metabonomics evaluations of age-related changes in the urinary compositions of male Sprague Dawley rats and effects of data normalization methods on statistical and quantitative analysis. BMC Bioinformatics 8 Suppl 7: S3

Silwood CJL, Grootveld M, Lynch E (2002) 1H NMR investigations of the molecular nature of low-molecular-mass calcium ions in biofluids. J Biol Inorg Chem 7: 46-57

Shaw RJ (2006) Glucose metabolism and cancer. Curr Opin Cell Biol 18: 598-608

Sreekumar A, Poisson LM, Rajendiran TM, Khan AP, Cao Q Yu J, Laxman B, Mehra R, Lonigro RJ, Li Y, Nyati MK, Ahsan A, Kalyana-Sundaram S, Han B, Cao X, Byun J, Omenn GS, Ghosh D, Pennathur S, Alexander DC, Berger A, Shuster JR, Wei JT, Varambally S, Beecher C, Chinnaiyan AM (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457: 910-914

Suna T, Salminen A, Soininen P, Laatikainen R, Ingman P, Makela S, Savolainen MJ, Hannuksela ML, Jauhiainen M, Taskinen M, Kaski K, Ala-Korpela M (2006) ^{1}H NMR metabonomics of plasma lipoprotein subclasses: elucidation of metabolic clustering by self-organising-map. NMR biomed 20: 658-672

Teichert F, Verschoyle RD, Greaves P, Edwards RE, Teahan O, Jones DJ, Wilson ID, Farmer PB, Steward WP, Gant TW, Gescher AJ, Keun HC (2008) Metabolic profiling of transgenic adenocarcinoma of mouse prostate (TRAMP) Tissue by (1)H-NMR analysis: evidence for unusual phospholipid metabolism. Prostate 68: 1035-1047

Yuille M, Illig T, Hveem K, Schmitz G, Hansen J, Neumaier M, Tybring G, Wichmann E, Ollier B (2010) Laboratory management of samples in biobanks: European consensus expert group report. Biopreservation & Biobanking 8(1): 65-69

Warburg O (1956) On the origin of cancer cells. Science 123: 309-314

Weckwerth W (2007) Metabolomics methods and protocols. Humana Press Inc., New York

Captions to the figures

Fig. 1 Experimental design for: **(A)** the evaluation of the impact of pre-centrifugation speed on the NMR metabolomic fingerprint, (**B)** the evaluation of the effect of the use of pre-centrifugation as a function of the temperature used for sample storage, **C)** the comparison of the efficiency of pre-centrifugation, filtration and addition of sodium azide in reducing bacterial content

Fig. 2 Effect of the pre-centrifugation speed on the NMR spectra obtained with the experimental design of Fig. 1A. The absolute value of the difference in the first component of the PLS-CA analysis for spectra of samples pre-centrifuged (PC) at different speeds and that of the non pre-centrifuged sample (NPC) is plotted as a function of the centrifugation speed. Diamonds and circles are used to indicate urine samples from two different individuals, the former containing a larger number of cells. The comparison of the two plots shows the importance of the pre-centrifugation step in samples with high cellular components.

Fig. 3 Experimental design for: (**A)** the evaluation of the effect of the time between blood collection and processing on serum and plasma NMR spectra; (**B)** the evaluation of the effect of the time between processing and freezing on the quality of NMR spectra of plasma and serum

Fig. 4 PLS-CA clustering as a function of the time delay between blood collection and processing for plasma at (**A)** 4 °C and (**B)** 25 °C and serum at (**D)** 4 °C and (**E)** 25 °C. Dots colour coding is: red = 0 h; yellow = 1 h; green = 2 h; blue = 3 h; purple = 4 h. Change of the average value of the first PLS-CA component as a function of time at 4 °C (blue line) and 25 °C (red line) for plasma (**C)** and serum **(F)**. PLS-CA clustering as a function of time delay between processing and analysis on plasma (**G)** and serum (**H)**. Dots color coding is red = 0 h; yellow = 6 h; green = 12 h; blue = 18 h; purple = 24 h. (**I**) Change of the average value of the first component as a function of time for plasma (green line) and serum (blue line)

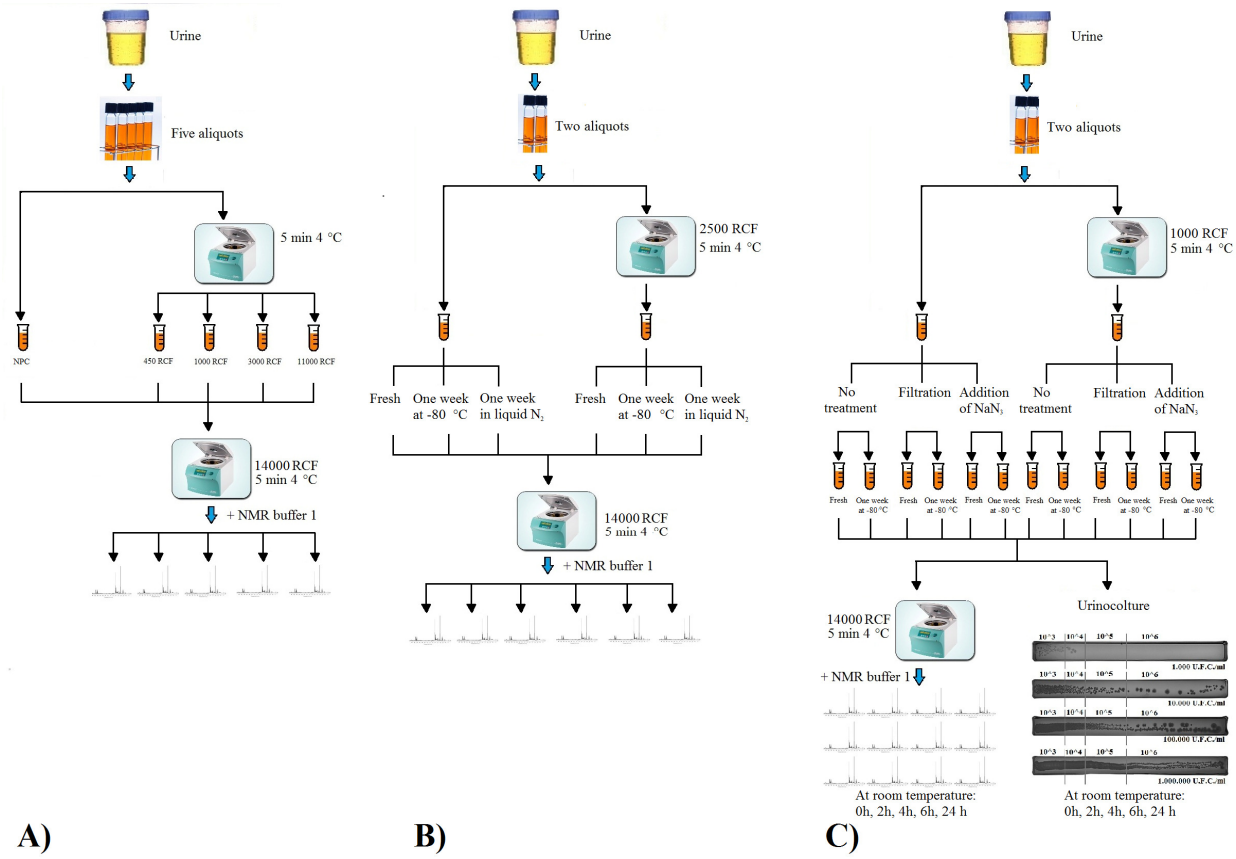

Urine
Five aliquots
5 min 4 °C
NPC
450 RCF
1000 RCF
3000 RCF
11000 RCF
14000 RCF
5 min 4 °C
+ NMR buffer 1
A)
Urine
Two aliquots
2500 RCF
5 min 4 °C
Fresh
One week at -80 °C
One week in liquid N2
Fresh
One week at -80 °C
One week in liquid N2
14000 RCF
5 min 4 °C
+ NMR buffer 1
B)
Urine
Two aliquots
1000 RCF
5 min 4 °C
No treatment
Filtration
Addition of NaN3
No treatment
Filtration
Addition of NaN3
Fresh
One week at -80 °C
14000 RCF
5 min 4 °C
+ NMR buffer 1
At room temperature:
0h, 2h, 4h, 6h, 24 h
Urinocolture
10^3
10^4
10^5
10^6
1.000 U.F.C./ml
10.000 U.F.C./ml
100.000 U.F.C./ml
1.000.000 U.F.C./ml
At room temperature:
0h, 2h, 4h, 6h, 24 h
C)

Fig. 1

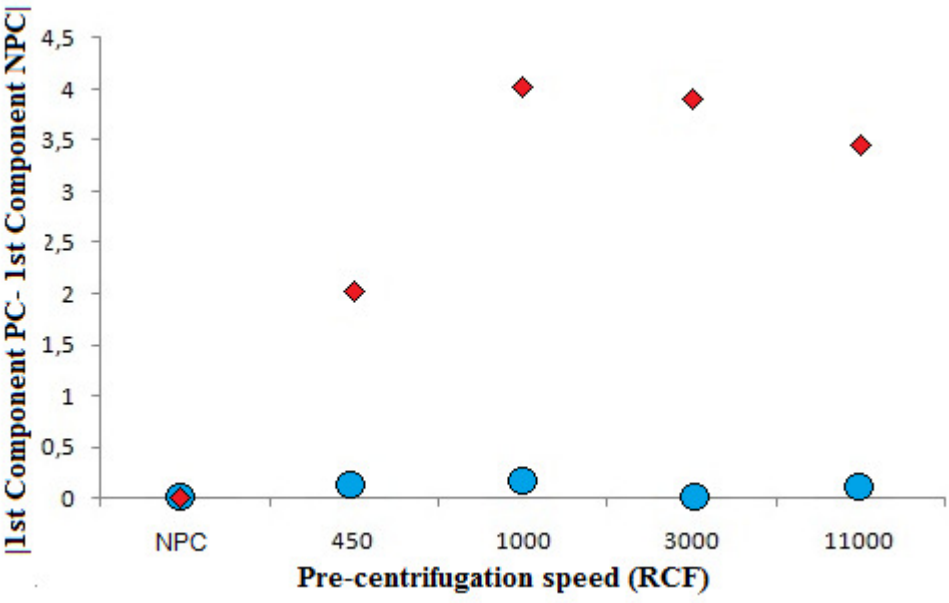

|1st Component PC- 1st Component NPC|
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
NPC
450
1000
3000
11000
Pre-centrifugation speed (RCF)

Fig. 2

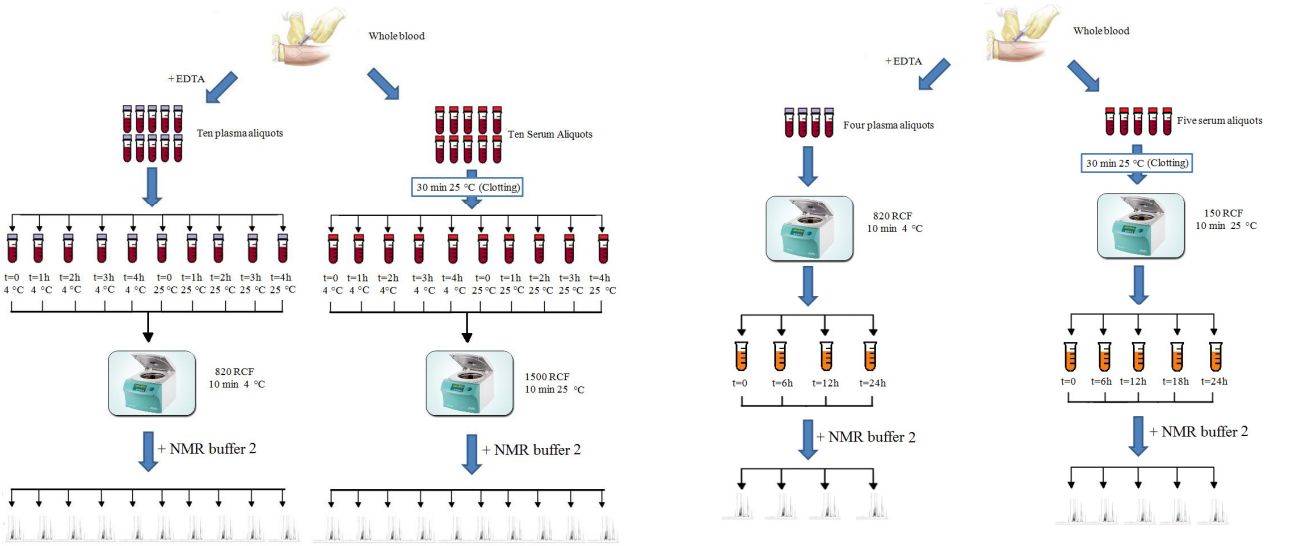

Whole blood
+EDTA
Ten plasma aliquots
Ten Serum Aliquots
30 min 25 °C (Clotting)
t=0 t=1h t=2h t=3h t=4h t=0 t=1h t=2h t=3h t=4h
4 °C 4 °C 4 °C 4 °C 4 °C 25 °C 25 °C 25 °C 25 °C 25 °C
820 RCF
10 min 4 °C
1500 RCF
10 min 25 °C
+ NMR buffer 2
A)
Whole blood
+EDTA
Four plasma aliquots
Five serum aliquots
30 min 25 °C (Clotting)
820 RCF
10 min 4 °C
150 RCF
10 min 25 °C
t=0 t=6h t=12h t=24h
t=0 t=6h t=12h t=18h t=24h
+ NMR buffer 2
B)

Fig. 3

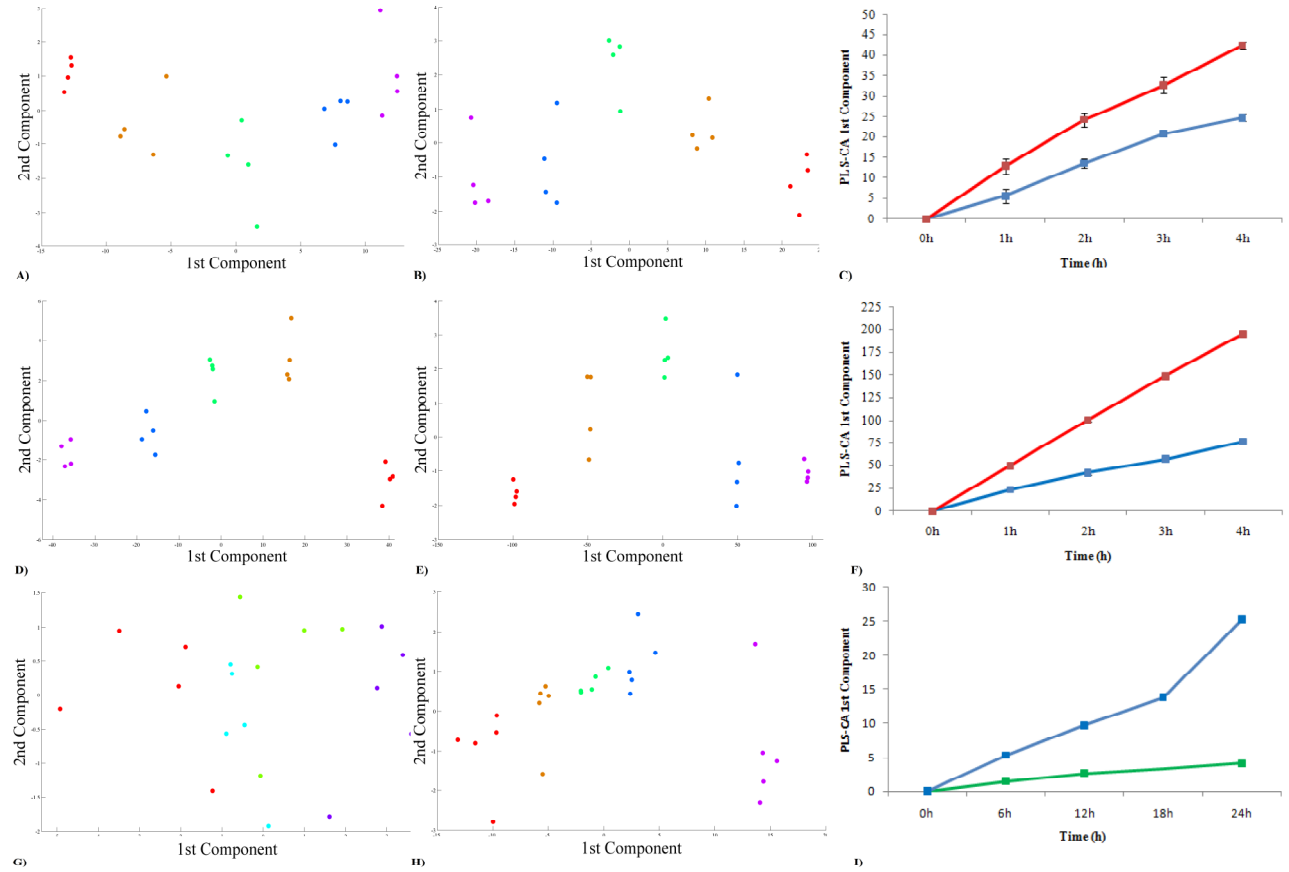

2nd Component
1st Component
A)
B)
PLS-CA 1st Component
Time (h)
0h 1h 2h 3h 4h
C)
D)
E)
F)
0h 6h 12h 18h 24h
G)
H)
I)

Fig. 4

Bernini et al. 27/11/2010 – Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks

Supplementary Material

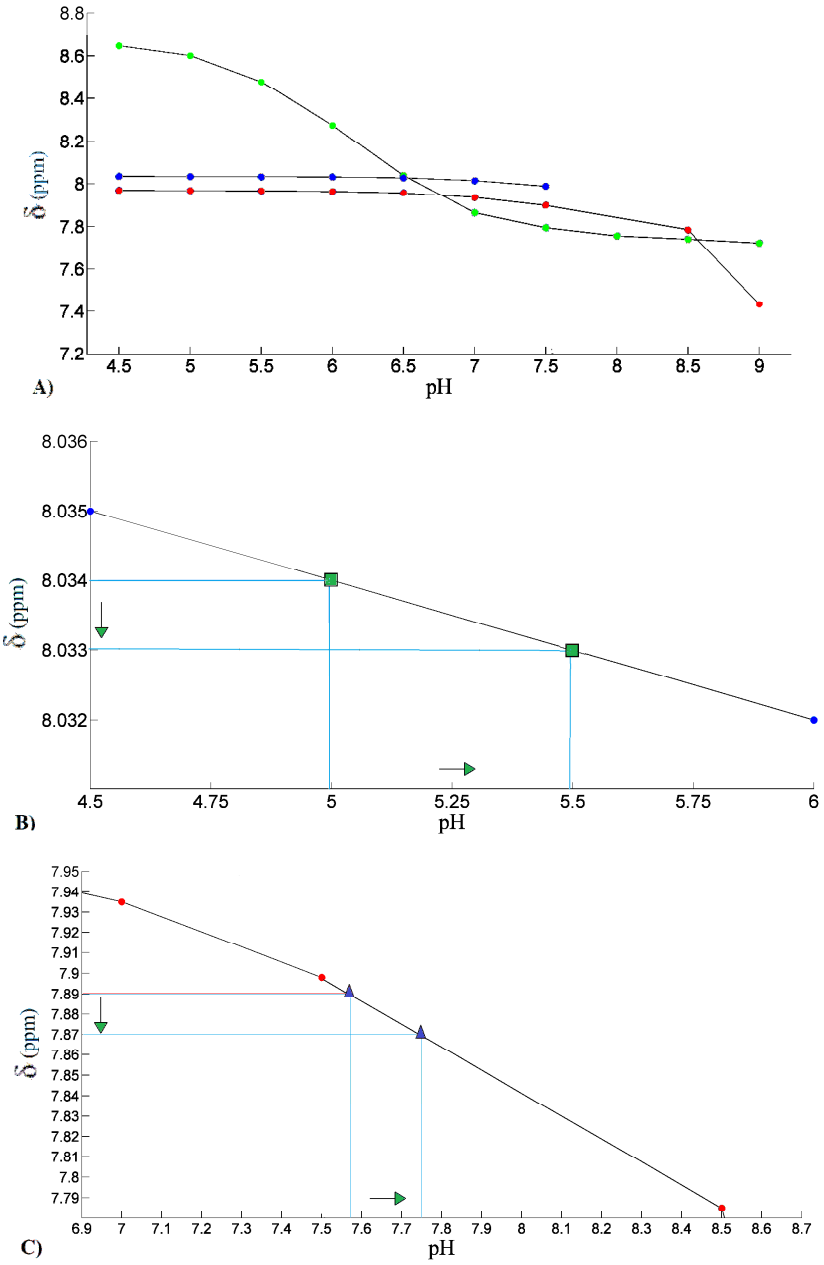


Fig. S1 **(A)** pH-dependence of the ^{1}H chemical shift of the Hε1 resonance of N-methylhistidine (N^{π} •, N^{τ} •) and histidine (•) in urine samples. Above pH 7.5 the resonance of N^{τ} could not be followed due to signal overlap. **(B)** Chemical shift values of the Hε1 of N^{τ}-methylhistidine for the non pre-centrifuged and pre-centrifuged fresh urine (■) reported on the calibration curve of panel **(A)**. **(C)** Chemical shift variations of the Hε1 of N^{π}-methylhistidine (▲) observed in pre-centrifuged urine over time (24 h) reported on the calibration curve obtained for N^{π}-methylhistidine of panel **(A)**

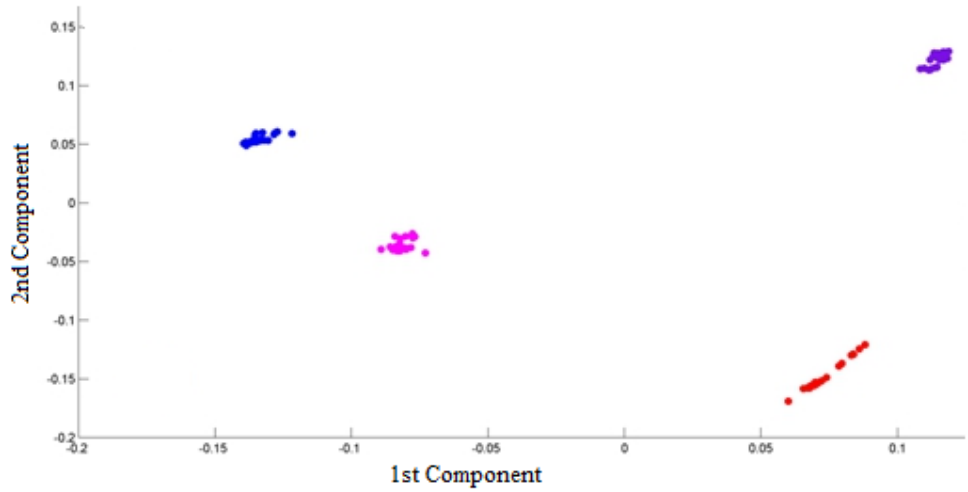


Fig. S2 PLS clustering according to donor identity of differently treated urine samples. Different aliquots have been produced from each of the four donors: eight different pre-centrifugation speeds (NPC, 600, 800, 1000, 1400, 1800, 2000, 2500 RCF) have been tested in combination with different storage conditions (fresh; one week in liq. N_2). In this plot, the same color has been used for aliquots from the same donor. A clear clustering of samples from the same individual is apparent

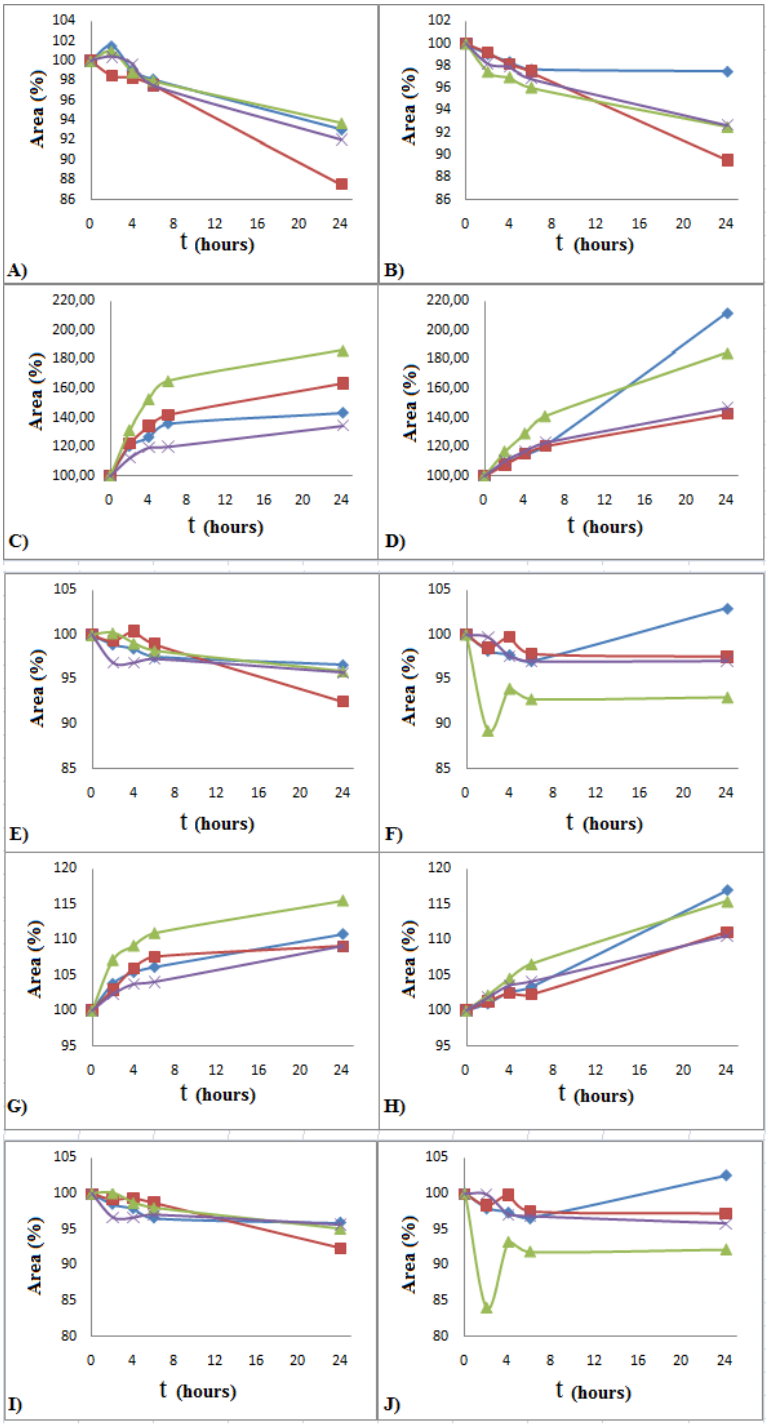
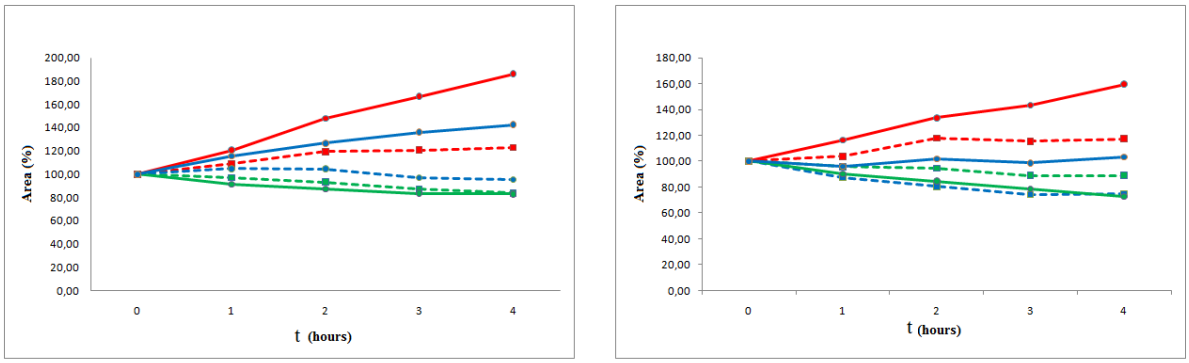


Fig. S3 Average variation in the concentration of some metabolites (glutamate in **A** and **B**; succinate in **C** and **D**, lactate in **E** and **F**, acetate in **G** and **H**, urea in **I** and **J**) as a function of time (0-24 h) in differently treated urine samples that have been analyzed either fresh (**A**, **C**, **E**, **G**, **I**) or after one week of storage at -80 °C (**B**, **D**, **F**, **H**, **J**). Signal intensities have been normalized with respect to that of the invariant CH_3 alanine resonance. Blue = pre-centrifuged; red = non-pre-centrifuged; green = pre-centrifuged and treated with NaN_3; purple = pre-centrifuged and filtered. The pre-centrifugation speed was 1000 RCF



A) **B)**

Fig. S4 Average variation in the concentration of glucose (green), lactate (red) and pyruvate (blue) as a function of time between blood collection and processing for samples kept at 25 °C (continous line) and 4 °C (dotted line) in plasma (**A**) and in serum (**B**)

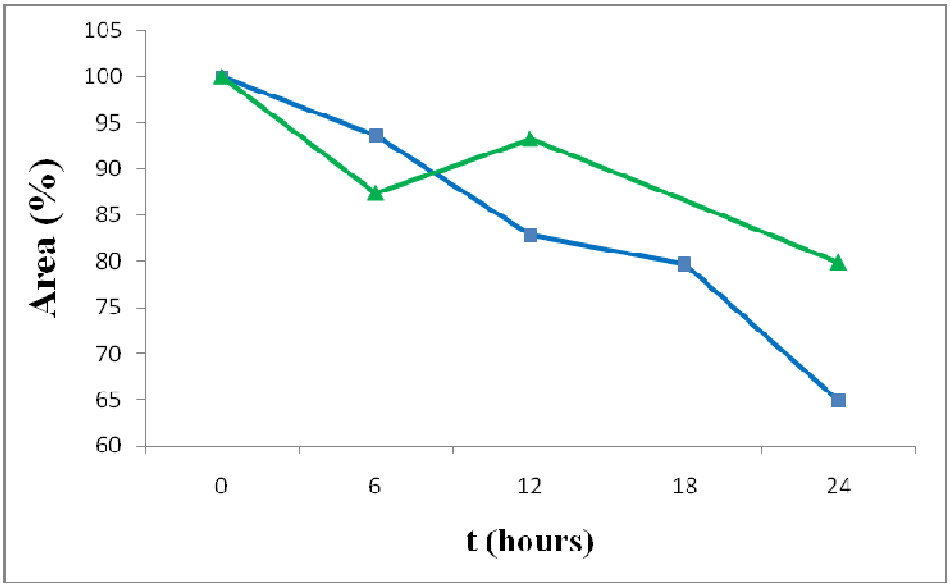


Fig. S5 Average variation in the concentration of proline in serum (blue) and plasma (green) as a function of time (0-24 hours)

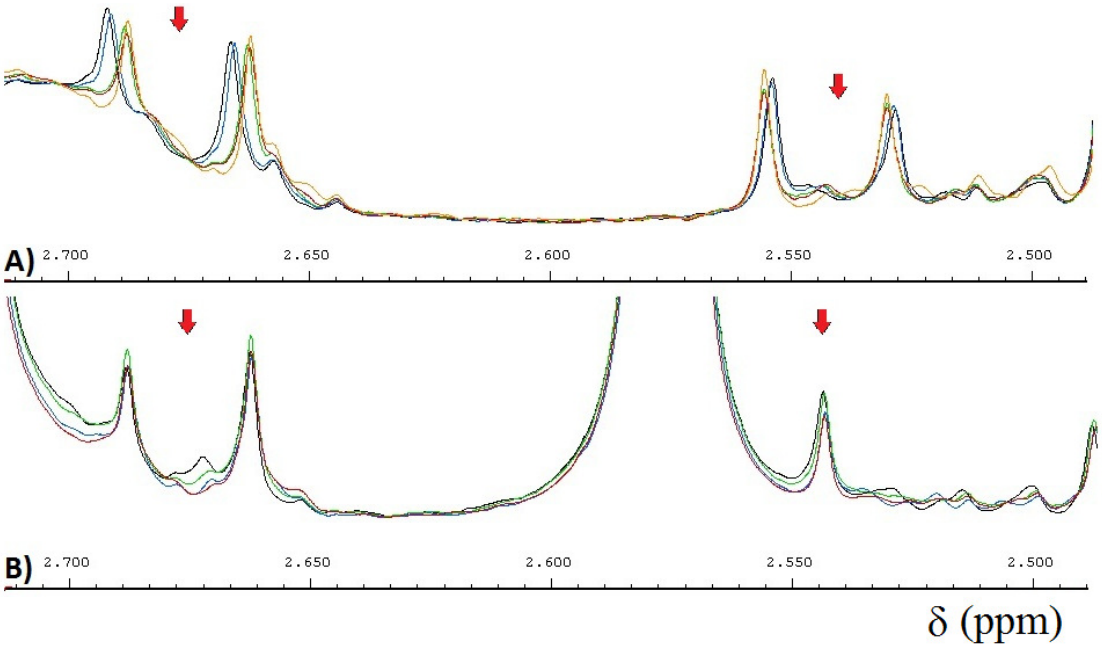


Fig. S6 Spectral variations observed for the signals of citrate in serum (**A**) and plasma (**B**) as a function of time. Color coding: **(A)** black t = 0 h, blue t = 6 h, green t = 12 h, brown t= 18 h, grey t = 24 h; **(B)** black t = 0 h, green t = 6 h, blue t = 12 h, brown t= 24 h

			Total count						Staphylococcus	
	Sample	Enterococcus	Proteus	S. Epidermidis	E. Coli	Candida	Pseudomonas	Streptococcus	S. Epidermidis	S. Aureus
T0	C (0 hour)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T0	CF (0 hour)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T1	C (2 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T1	CF (2 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T2	C (4 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T2	CF (4 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T3	C (6 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T3	CF (6 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T4	C (24 hours)	10^3 CFU*-NONE	10^4 -10^3CFU	NONE-2 CFU	10 CFU-NONE	5 CFU-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	10^4-10^3 CFU
T4	CF (24 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE

Table S1 Results of the urine culture tests performed on differently treated aliquots of a sample of Subject 1: C (pre-centrifuged at 1000 RCF), CF (pre-centrifuged and filtered). Urine culture tests were performed at time zero and after 2, 4, 6 and 24 hours, starting either from fresh urine (first value in the cell) or from urine samples stored for a week at -80 °C (second value in the cell). Notice that in the non filtered aliquots analysed after 24 hours the growth of coliforms visible in fresh aliquot is stopped by the storage at -80 °C.

			Total count						Staphylococcus	
	Sample	Enterococcus	Proteus	S. Epidermidis	E. Coli	Candida	Pseudomonas	Streptococcus	S. Epidermidis	S. Aureus
T0	C (0 hour)	NONE	10^3-10^3 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^3-10^3 CFU	NONE-1 CFU	NONE-NONE
T0	CF (0 hour)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T0	CNA (0 hour)	NONE	10^3-10^2 CFU	NONE-NONE	2 CFU-NONE	NONE-NONE	NONE-NONE	10^3-10^2 CFU	NONE-NONE	NONE-NONE
T0	NC (0 hour)	NONE	10^3-10^3 CFU	1 CFU-2 CFU	NONE-NONE	NONE-NONE	NONE-NONE	10^3-10^3 CFU	2 CFU-1 CFU	NONE-NONE
T0	NCF (0 hour)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T0	NCNA (0 hour)	NONE	10^3-10^3 CFU	1 CFU-1 CFU	NONE-1 CFU	NONE-NONE	NONE-NONE	10^3-10^3 CFU	3 CFU- 1CFU	NONE-NONE
T1	C (2 hours)	NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T1	CF (2 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T1	CNA (2 hours)	NONE	10^4_10^3 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T1	NC (2 hours)	NONE	10^5-10^4 CFU	NONE-1 CFU	NONE-NONE	1 CFU-NONE	NONE-NONE	10^4-10^4 CFU	NONE-NONE	1 CFU-NONE
T1	NCF (2 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T1	NCNA (2 hours)	NONE	10^4-10^4 CFU	1 CFU-NONE	1 CFU-2 CFU	NONE-NONE	NONE-NONE	10^4-10^4 CFU	NONE-NONE	NONE-NONE
T2	C (4 hours)	NONE	10^4-10^3 CFU	1 CFU-NONE	2 CFU-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T2	CF (4 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T2	CNA (4 hours)	NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	1 CFU-NONE
T2	NC (4 hours)	NONE	10^5-10^4 CFU	1 CFU-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^4 CFU	NEG-3 CFU	NONE-NONE
T2	NCF (4 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T2	NCNA (4 hours)	NONE	10^5-10^4 CFU	1 CFU-1 CFU	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T3	C (6 hours)	NONE	10^4-10^4 CFU	1 CFU-1 CFU	1 CFU-1 CFU	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NEG-2 CFU	NONE-NONE
T3	CF (6 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T3	CNA (6 hours)	NONE	10^4-10^4 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^4 CFU	NONE-NONE	NONE-NONE
T3	NC (6 hours)	NONE	10^5-10^4 CFU	2 CFU-NONE	2 CFU-NONE	1 CFU-NONE	NONE-NONE	10^5-10^4 CFU	3 CFU-4 CFU	NONE-1 CFU
T3	NCF (6 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T3	NCNA (6 hours)	NONE	10^4-10^4 CFU	NONE-1 CFU	NONE-NONE	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T4	C (24 hours)	NONE	10^4 -10^3CFU	NONE-2 CFU	10 CFU-NONE	5 CFU-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE
T4	CF (24 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T4	CNA (24 hours)	NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	10^3-10^3 CFU	NONE-NONE	NONE-NONE
T4	NC (24 hours)	NONE	10^5*-10^5 CFU	NONE-NONE	10^4 CFU-3 CFU	NONE-4 CFU	NONE-NONE	10^4-10^5 CFU	10^5 CFU**-NONE	10^5**-10^3 CFU
T4	NCF (24 hours)	NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE	NONE-NONE
T4	NCNA (24 hours)	NONE	10^4-10^5 CFU	NONE-6 CFU	1 CFU-1 CFU	NONE-NONE	NONE-NONE	10^4-10^3 CFU	NONE-NONE	NONE-NONE

Table S2 Results of the urine culture tests performed on differently treated aliquots of a sample of Subject 2: C (pre-centrifuged at 1000 RCF), CF (pre-centrifuged and filtered), CNA (pre-centrifuged and treated with sodium azide), NC (not pre-centrifuged), NCF (filtered), NCNA (treated with sodium azide). Urine culture tests were performed at time zero and after 2, 4, 6 and 24 hours, starting either from fresh urine (first value in the cell) or from urine samples stored for a week at -80 °C (second value in the cell).

I would like to say thank to Prof. Ivano Bertini for the tools and the knowledge provided to support my research and also to Prof. Claudio Luchinat and Prof.ssa Paola Turano. Moreover I would like to say thank to all the colleagues known during these three years at CERM: Stefano, Stefano, Leonardo, Edoardo, Dario, Michela, Paola, Giordana, Silvia, Samuele, Iacopo, Agata are only a few names… Moreover a special thank to Giacomo and to my family…

Patrizia