Biomarkers, 2011; 16(1): 83–88

ORIGINAL ARTICLE

Meeting-in-the-middle using metabolic profiling – a strategy for the identification of intermediate biomarkers in cohort studies

Marc Chadeau-Hyam[1], Toby J. Athersuch[1,2,3], Hector C. Keun[2,3], Maria De Iorio[1], Timothy M.D. Ebbels[2], Mazda Jenab[4], Carlotta Sacerdote[5,6], Stephen J Bruce[2], Elaine Holmes[2,3], and Paolo Vineis[1,3,6]

[1]*Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, London, UK,* [2]*Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK,* [3]*MRC-HPA Centre for Environment and Health, School of Public Health, Faculty of Medicine, Imperial College London, London, UK,* [4]*International Agency for Research in Cancer (IARC-WHO), Lyon, France,* [5]*CPO-Piemonte, Torino, Italy, and* [6]*HuGeF Foundation, Torino, Italy*

Abstract

Background: Predictive disease risk biomarkers that can be linked to exposure have proved difficult to identify in case-control studies.

Methods: Parallel statistical analysis of the correlation between ^{1}H NMR profiles from plasma samples collected before disease onset (EPIC cohort), versus exposure to dietary compounds, and follow-up disease endpoints (colon and breast cancer) was performed.

Results: Metabonomic signatures associated with colon cancer and dietary fiber intake (a protective factor according to epidemiological studies) were identified.

Conclusion: This implementation of the novel "*meet-in-the-middle*" analytical strategy indicates how case-control studies nested in prospectively collected cohorts may reveal intermediate biomarkers linking exposure and disease.

Keywords: *cohort studies; metabonomics; nested design; metabolic profiles; intermediate biomarkers*

Introduction

Metabolic profiling, incorporating metabonomics (Nicholson *et al.*, 1999) and metabolomics (Fiehn, 2002) is a promising high-throughput technology that can identify thousands of metabolic signals and can lead, if appropriately used, to the identification of new biomarkers (Keun and Athersuch, 2007, Lindon *et al.*, 2005, Coen *et al.*, 2008). In a previous study, metabolic profiling has been shown to be able to distinguish different population subgroups and to identify potential biomarkers of diet-related hypertension (Holmes *et al.*, 2008). While this study used urine samples, several large epidemiological cohort studies that are available today have collected blood as the primary biological sample. It is therefore important to validate the use of blood samples for the development of metabolome-wide profiles in cancer studies nested in cohorts. Such profiles may lead to the identification of both biomarkers of exposure (e.g. dietary components, environmental pollutants) and of markers of early damage (e.g. early disease-specific metabolic changes), notably for carcinogenesis.

Prospective cohort studies are conceptually suitable for the identification of new biomarkers of both exposure and early biological effect, since they are based on

Address for Correspondence: MRC-HPA Centre for Environment and Health, School of Public Health, Faculty of Medicine, Imperial College London, Norfolk Place, W2 1PG, London UK, e-mail p.vineis@imperial.ac.uk tel +44 (0)20 75943372

(Received 30 July 2010; revised 13 October 2010; accepted 16 October 2010)

ISSN 1354-750X print/ISSN 1366-5804 online © 2011 Informa UK, Ltd.
DOI: 10.3109/1354750X.2010.533285

pre-clinical biological samples that are not influenced by the inherent metabolic changes due to the disease itself.

Identifying the overlap between markers of exposure and predictive markers of disease outcome has previously been defined as a "*meet-in-the-middle*" approach (Vineis and Perera, 2007). According to that original concept, the finding that preclinical biomarkers shown to be related to particular exposures in prospective studies are also elevated in certain subclasses of disease would strengthen causal links between exposures and disease. Prospective studies working from exposure to preclinical response would directly complement retrospective studies backtracking from clinical disease to preclinical response to exposure. This *"meet-in-the-middle approach"* has potential to open new avenues for prevention by identifying the specific environmental factor(s) involved in the disease process (Vineis and Perera, 2007).

As a proof-of-concept, we describe here a possible implementation of the "*meet-in-the-middle*" approach, based on a small pilot study, using a nested case-control design within a cohort of healthy individuals enrolled in the European Prospective Investigation into Cancer and Nutrition (EPIC), with an average of 7 years follow-up information relating to colon and breast cancer disease outcome. In this work, we have used the approach to identify a list of putative intermediate biomarkers that link exposure and disease outcome.

Methods

Available data from EPIC Torino sub-cohort

In a nested case-control pilot study of colon (24 cases and 23 controls) and breast (19 cases and 20 controls) cancers, a total of 86 plasma samples from the Turin (Italy) sub-cohort of EPIC (referred to throughout as the EPIC Torino cohort) were processed using standard ^{1}H NMR metabolic profiling protocols (Beckonert *et al.*, 2007). The methods and rationale of EPIC, a large prospective cohort study with over 520,000 volunteer subjects enrolled from 23 centres in 10 Western European countries, have been detailed elsewhere (Riboli, 1992, Riboli *et al.*, 2002). For the EPIC Torino component, lifestyle and personal history data was collected using standardized questionnaires used in other sub-cohorts. Anthropometric data were collected from most participants at recruitment (Bingham and Riboli, 2004). For both cancers, cases and controls were matched on gender and age. Plasma samples were collected on average 7 years before cancer onset. Information on the dietary intake of the volunteers was captured using two different dietary assessment instruments. First, diet over the previous 12 months was measured at recruitment by means of validated country-specific questionnaires, designed to ensure high compliance and better measures of local dietary habits in each of the EPIC centres (Bingham and Riboli, 2004). Second, diet over the 24 hours prior to interview (24-hour dietary recall) was collected in a subset using a standardized software (Bingham and Riboli, 2004). In each of the 23 EPIC centers, blood samples were drawn from most participants, stored at 5–10°C, protected from light, and transported to local laboratories for processing and aliquot preparation. Blood was separated into 0.5 mL fractions (serum, plasma, red blood cells, and buffy coat for DNA extraction). Each fraction was placed into straws, which were heat-sealed and stored in liquid nitrogen at -196°C. Sodium citrate was used as anticoagulant and straws are made of a ionomeric resin (*CRYO BIO SYSTEM*). The time between collection and storage in this particular centre (Torino) was monitored and ranged between 2 and 5 hours. EPIC has devoted much effort to establish an adequate sample collection and storage (Peluso *et al.*, 2005), and internal studies have been conducted to assess sample stability over time (Teahan *et al.*, 2006).

Measurements for a large number of dietary exposures (e.g. estimates of intake of specific food groups, foods and nutrients) and anthropometric and lifestyle data were considered, including a wide range of potential or known risk factors for cancer (smoking, BMI, and physical activity) as well as intake estimates of a panel of nutrients potentially involved in carcinogenesis (a total of 30 dietary variables). All dietary intake estimates were adjusted on total energy intake, as previously described in EPIC analyses (Vergnaud *et al.*, 2010).

Metabolomics methodology

High-resolution one-dimensional ^{1}H NMR spectral profiles for the individual EPIC plasma samples (n = 86) were obtained at 600 MHz (^{1}H operating frequency) using a Bruker AVANCE 600 spectrometer (Bruker BioSpin, Rheinstetten, Germany) according to established sample preparation and acquisition protocols (Beckonert *et al.*, 2007). Briefly, each sample was thawed and a 300 µL aliquot diluted with 300 µL saline solution (80:20 H2O:D2O (v/v), 0.9% NaCl (w/v)) giving a total volume of 600 µL. Prior to analysis, each sample was centrifuged (12,000 g, 5 min, 4 C) to remove any suspended, insoluble material, and 550 uL of the supernatant was subsequently transferred to a 5 mm glass NMR tube (GPE Scientific Ltd, Leighton Buzzard, Bedfordshire, England). Samples were submitted for ^{1}H NMR analysis immediately after preparation (i.e. no additional freeze-thaw cycles). One-dimensional, ^{1}H Carr-Purcell-Meiboom-Gill (CPMG) spectra were acquired as described previously (Beckonert *et al.*, 2007) with suppression of the water resonance achieved by radiofrequency presaturation (δH = 4.701). Spectra were

then phased and baseline-corrected using in-house automated MATLAB routines (supplied by Drs T Ebbels, O Cloarec, HC Keun & JTM Pearce). Calibration of all spectra using the glucose doublet (δH = 5.233), was conducted as described previously (Pearce *et al.*, 2008). The spectral resolution of the metabonomic profiles indirectly impacts the power of the study. Typically, higher spectral resolution results in more statistical tests being performed and therefore to a more stringent multiple testing correction. Given the small number of observation available in our study, spectral intensities were then integrated over 275 contiguous regions, whith a corresponding spectral resolution of 0.04 ppm / region, which has already been shown to be biochemically relevant (Ebbels *et al.*, 2007).

Implementation of the 'meet-in-the-middle' approach

An outline of our analytical strategy is illustrated in Figure 1. First relationships between spectral data and each disease end point were explored (colon and breast cancers were considered separately).

Spectral regions representing the water resonance (4.40 – 5.00 ppm) and those below 0 ppm were excluded, resulting in a total of 235 spectral regions used as predictor variables in subsequent chemometric analysis. Spectral data were first normalized so that the cumulative spectral intensities over the 235 variables summed to 1 for each individual. Spectral data were subsequently analyzed using an O2PLS approach that has been well established for biomarker identification in metabolic profiling (Trygg and Wold, 2003, Cloarec *et al.*, 2005). For each of the 235 spectral regions, the O2PLS approach provided estimates of their regression coefficient, representing their contribution to case-control discrimination. A bootstrap resampling technique was used to estimate the uncertainty of each regression coefficient, and approximate Student's t-tests were performed to estimate the p-value for each spectral variable in relation to disease outcomes (Holmes *et al.*, 2008). Bonferroni correction was applied as a conservative adjustment of the significance criteria for individual pair-wise comparisons to account for multiple testing: with an overall family wise error rate of 0.05, the per-test significance level was set to $\alpha = 0.05/235 = 2.1\text{x}10^{-4}$.

In parallel, the associations between exposure estimates (diet, anthropometry, lifestyle) and spectral regions were assessed using either an analysis of variance (ANOVA) for categorical variables or (ii) a non-parametric Spearman's rank correlation test for continuous variables.

From these two parallel analyses, we obtained lists of putative markers of (i) the disease outcome, and (ii) exposure. These were compared in a second step in order to identify possible intersecting signals, therefore defining potential intermediate biomarkers.

Plausible plasma metabolites with characteristic resonances in the spectral regions identified in the analysis described above were characterized using published literature data and database resources (Nicholson *et al.*, 1995, Wishart *et al.*, 2007), thus producing a list of putative intermediate molecules that may link exposure and disease outcome, thus providing leads for future investigation and validation.

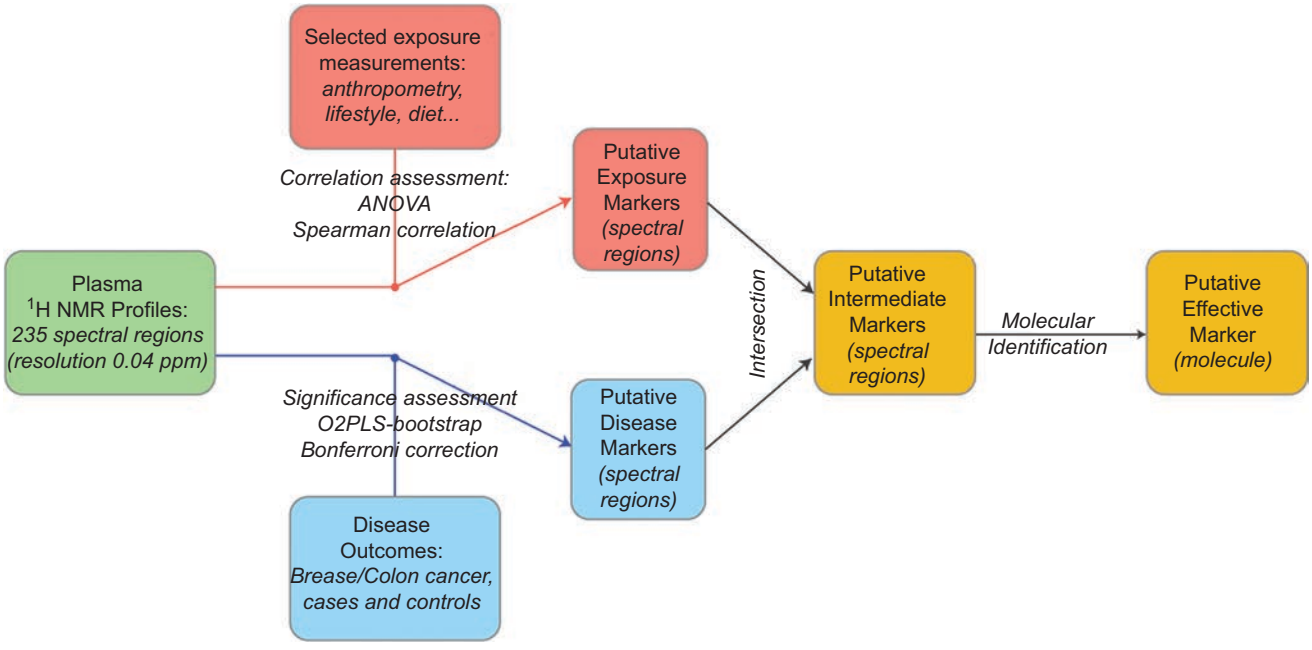

Figure 1. Schematic representation of the implementation of the 'meet-in-the middle' approach.

Results

We first compared spectra from the 24 colon cancer cases, and the 19 breast cancer cases against 43 pooled controls. Based on these samples, the O2PLS approach did not highlight any significant associations for breast cancer but found eight spectral regions significantly associated to colon cancer outcome: [3.20-3.24], [3.64-3.68], [3.68-3.72], [4.04-4.08], [4.24-4.28], [6.80-6.84], [7.00-7.04], [7.72-7.76] ppm. Basing our calculations on the standard q-value approach and R-package from Storey (Storey, 2002), we estimated the maximum false discovery rate (FDR) for the variables whose O2PLS p-values were below the Bonferroni corrected significance level: we found that the FDR for this analysis was <0.003. For both cancers, when limiting the sample size by using gender-matched case/control pairs, no statistically significant association was found. However, the ranking of the associations according to their p-values (whether significant or not) was consistent with the one based on the (larger) pooled controls population. This could be explained by the gain in power related to the use of a larger sample size, though it could lead to uncontrolled confounding for the matching variables. The results detailed below describe the analysis of colon cancer as the disease outcome compared to pooled controls.

We examined the strength of the associations between spectral regions and exposure variables. In Figure 2 we map the peak/exposure combinations showing a higher level of correlation (the top associations correspond to those with p-values <0.05).

For the vast majority of exposure measurements, the total number of such associations remains below 15, while for dietary fibers and total protein intake there are over 20 associations (21 and 38 respectively): for these two variables, larger continuous spectral regions showing a higher level of correlation can be observed.

We summarize in Table 1 the spectral variables/exposure associations involving any of the significant spectral regions primarily found to be related to case-control discrimination by the O2PLS approach (horizontal dashed lines in Figure 2).

Due to the high p-values of these associations (p-values ranging from 0.003 to almost 0.05), none would survive any realistic correction for multiple testing, and could therefore not be considered as statistically significant. However, some are well-recognized risk factors for colon cancer (e.g. smoking), while dietary fibers are the most established protective factor for colon cancer (Bingham *et al.*, 2003, Wiseman, 2008). Furthermore, lower p-values were found for exposure measurements associated to more than one putative marker of the disease outcome, therefore suggesting a greater relevance of these signals. In particular, dietary fibers intake was found to be associated to four putative markers (with corresponding p-values ranging from 0.003 to 0.02 and an Odds Ratio = 0.64). For these associations we also performed the calculation of the FDR. This showed that each of the four putative markers associated with dietary fiber intake had an FDR<0.35.

Discussion

The purpose of the present pilot study was to explore the possibility of using plasma metabolic profiles, in a prospective study on diet and cancer, to define intermediate metabolic biomarkers via a '*meet-in-the-middle*' approach as defined above. The strongest (negative) association we identified in relation to colon cancer, both in terms of statistical significance and of number spectral regions associated with cancer onset, involved dietary intake of fibers. This lends plausibility to our findings. Thiamin was also found negatively associated with three spectral regions, in turn significantly associated to colon cancer. This observation is interesting because thiamin has been recently suggested as potential protective factors for colon cancer (Bruce *et al.*, 2003).

While it was not possible to make unambiguous associations between individual metabolites and the exposure measurements, it is interesting to note that two spectral regions have just one metabolite indicated from interrogation of metabolomic databases. One indicates a possible link between a metabolite of benzoic acid [6.80-6.84] and dietary fibers, while the other suggests an association between histidine metabolism [7.72-7.76] and BMI. Benzoic acid is predominantly derived from gut microbial fermentation of plant phenolics in the colon (Nicholson *et al.*, 2005, Phipps *et al.*, 1998, Aura, 2007), a process also plausibly linked to higher dietary fibers exposure and lower colon cancer risk. Increased blood levels and excretion of histidine metabolites such as methylhistidine are associated with a high meat intake, as is detection of carnosine, a beta-alanine-histidine dipeptide that is highly abundant in meat (Block *et al.*, 1965, Park *et al.*, 2005, Myint *et al.*, 2000, Abe *et al.*, 1993, Dragsted, 2010). It is also plausible that a signature of meat intake could positively correlate to BMI and cancer risk. One region [3.68-3.72] appears to be associated with the majority of exposure measurements: this spectral region contains signals from two of the most abundant metabolites in blood, glucose and glutamine. Both metabolites report on the combined effects of multiple energetic and dietary processes, hence these correlations are likely to be non-specific.

Taken together, our results suggest that meaningful relationships can be found using our data analysis strategy on metabolic profiling and are consistent with the epidemiological literature relating to colon cancer (Wiseman, 2008). This pilot study has provided evidence

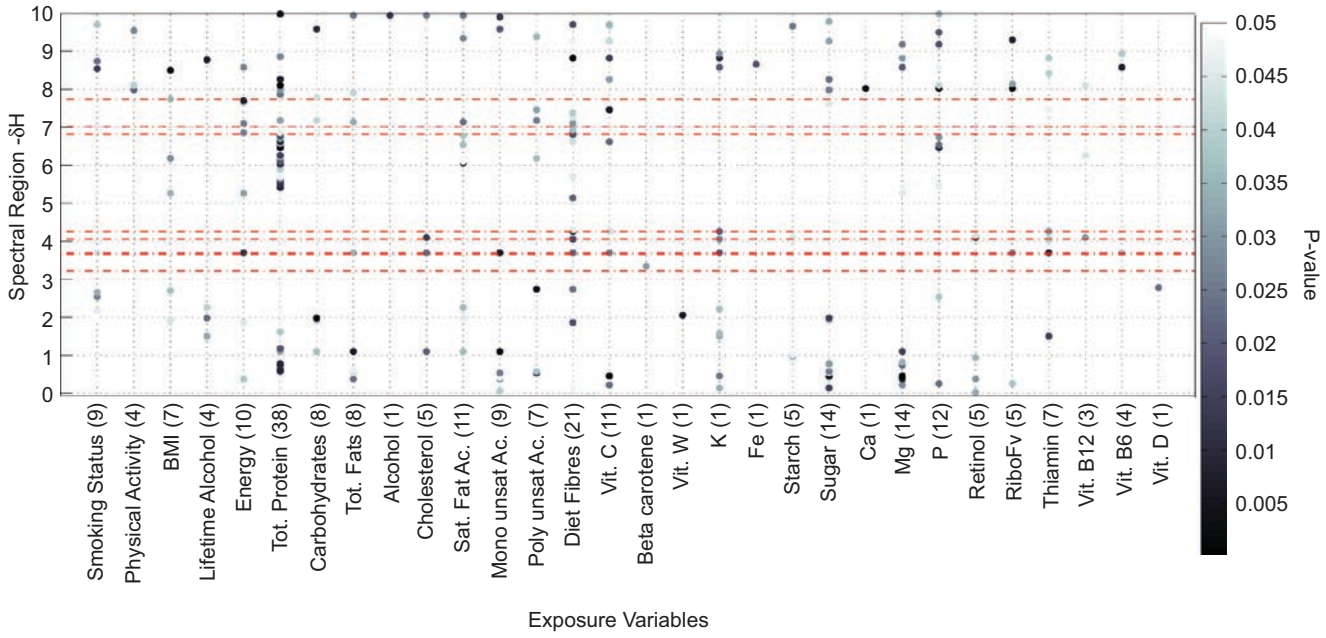

Figure 2. Map of the strongest associations between exposure variables (X-axis) and ^{1}H NMR spectral regions (Y-axis). The strength of the association for each exposure/spectral variable combination is measured by the corresponding p-value. For clarity, only associations with p-values<0.05 are plotted, and these are colored according to their p-values. The total number of such associations is shown in brackets for each exposure measurement. Horizontal dotted lines represent spectral regions in the ^{1}H NMR spectra that were found significantly associated with the disease outcome (colon cancer).

Table 1. Summary of the spectral regions that were found significantly associated with colon cancer onset by the OPLS approach, and whose association with the exposure variables had a p-value <0.05 (these p-values are reported in the table).

Exposure	[3.20-3.24]	[3.68-3.72]	[4.04-4.08]	[4.24-4.28]	[6.80-6.84]	[7.72-7.76]
Smoking Status		0.0426				
BMI						0.0351
Energy		0.0111	0.0474			
Tot. Fats		0.0334				
Cholesterol		0.0227				
Mono unsat Ac.		0.0059				
Diet. Fibres		0.0205	0.0156	0.0033	0.0141	
Vit. C		0.0249		0.0419		
K		0.0191	0.0255	0.02		
Mg	0.045					
RiboFv		0.0247				
Thiamin		0.0061	0.0375	0.0288		
Vit. B6		0.0351				
Potential Metabolite Assignments	choline arginine β-glucose	glutamine leucine citrulline α-glucose	Creatinine lipid glyceryl myo-inositol choline 2,3-diphosphoglyceric acid	Threonine lipid glyceryl	p-aminobenzoic acid	histidine

of the utility of "*meet-in-the-middle*" approaches to identify putative metabolites that may be statistically linked to both exposures and disease outcomes, thus leading to potential intermediate biomarkers. The small number of observations available in the current study was a considerable weakness and limited the confidence with which putative markers could be derived due to multiple testing considerations. Larger studies would permit analyses to be conducted with greater statistical power (and facilitate the use of high-resolution profile data) and are an obvious next step in establishing the utility of our approach. A number of omics-based profiling projects currently underway are using large numbers of biological samples from prospective cohort studies with a nested case-control design. These studies will be ideal for deriving intermediate markers of early response that link exposure and disease endpoints, through an approach similar to the one we describe here. We conclude that the "*meet-in-the-middle*" approach to biomarker discovery previously proposed (Vineis and Perera, 2007) provides

a simple and efficient strategy for biomarker discovery in molecular epidemiological research.

Declaration of interest

The authors wish to acknowledge that this work was supported in part by the European Union FP7 Project EnviroGenomarkers: Genomics Biomarkers of Environmental Health (Grant Agreement 226756). MDI and TE were partially supported by the Biotechnology and Biological Sciences Research Council (Grant Ref.BB/E20372/1).

References

Abe H, Okuma E, Sekine H, Maeda A, Yoshiue S 1993. Human urinary excretion of L-histidine-related compounds after ingestion of several meats and fish muscle. *Int. J. Biochem.,* 25, 1245–9.

Aura A-M 2007. Microbial metabolism of dietary phenolic compounds in the colon. *Phytochem Rev,* 7, 407–29.

Beckonert O, Keun HC, Ebbels TM, 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. *Nat Protoc,* 2, 2692–703.

Bingham S & Riboli E 2004. Diet and cancer--the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer,* 4, 206–15.

Bingham SA, Day NE, Luben R, et al. 2003. Dietary fibre in food and protection against colorectal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC): an observational study. *Lancet,* 361, 1496–501.

Block WD, Hubbard RW & Steele BF 1965. Excretion of Histidine and Histidine Derivatives by Human Subjects Ingesting Protein from Different Sources. *J Nutr,* 85, 419–25.

Bruce WR, Furrer R, Shangari N, O'Brien PJ, Medline A & Wang YP 2003. Marginal dietary thiamin deficiency induces the formation of colonic aberrant crypt foci (ACF) in rats. *Cancer Letters,* 202, 125–9.

Cloarec O, Dumas M-E, Craig A, et al. 2005. Statistical Total Correlation Spectroscopy: Äâ An Exploratory Approach for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. *Analytical Chemistry,* 77, 1282–9.

Coen M, Holmes E, Lindon JC & Nicholson JK 2008. NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology. *Chem Res Toxicol,* 21, 9–27.

Dragsted LO 2010. Biomarkers of meat intake and the application of nutrigenomics. *Meat Sci.,* 84, 301–7.

Ebbels TMD, Keun HC, Beckonert OP, Bollard ME, Lindon JC, Holmes E & Nicholson JK 2007. Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The Consortium on Metabonomic Toxicology screening approach. *J Proteome Res,* 6, 4407–22.

Fiehn O 2002. Metabolomics--the link between genotypes and phenotypes. *Plant Mol Biol,* 48, 155–71.

Holmes E, Loo RL, Stamler J, et al. 2008. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature,* 453, 396–400.

Keun HC & Athersuch TJ 2007. Application of metabonomics in drug development. *Pharmacogenomics,* 8, 731–41.

Lindon JC, Keun HC, Ebbels TM, Pearce JM, Holmes E & Nicholson JK 2005. The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics,* 6, 691–9.

Myint T, Fraser GE, Lindsted KD, Knutsen SF, Hubbard RW & Bennett HW 2000. Urinary 1-methylhistidine is a marker of meat consumption in black and in white California seventh-day Adventists. *Am. J. Epidemiol.,* 152, 752–5.

Nicholson JK, Foxall PJ, Spraul M, Farrant RD & Lindon JC 1995. 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. *Anal Chem,* 67, 793–811.

Nicholson JK, Holmes E & Wilson ID 2005. Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Microbiol.,* 3, 431–8.

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, 1181–9.

Park YJ, Volpe SL & Decker EA 2005. Quantitation of carnosine in humans plasma after dietary consumption of beef. *J. Agric. Food Chem.,* 53, 4736–9.

Pearce JT, Athersuch TJ, Ebbels TM, Lindon JC, Nicholson JK & Keun HC 2008. Robust algorithms for automated chemical shift calibration of 1D 1H NMR spectra of blood serum. *Anal Chem,* 80, 7158–62.

Peluso M, Hainaut P, Airoldi L, et al. 2005. Methodology of laboratory measurements in prospective studies on gene-environment interactions: the experience of GenAir. *Mutat Res,* 574, 92–104.

Phipps AN, Stewart J, Wright B & Wilson ID 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, 527–37.

Riboli E 1992. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol,* 3, 783–91.

Riboli E, Hunt KJ, Slimani N, et al. 2002. European prospective investigation into cancer and nutrition (EPIC): study populations and data collection. *Public Health Nutr.,* 5, 1113–24.

Storey JD 2002. A direct approach to false discovery rates. *J Roy Stat Soc B,* 64, 479–98.

Teahan O, Gamble S, Holmes E, Waxman J, Nicholson JK, Bevan C & Keun HC 2006. Impact of analytical bias in metabonomic studies of human blood serum and plasma. *Analytical Chemistry,* 78, 4307–18.

Trygg J & Wold S 2003. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *Journal of Chemometrics,* 17, 53–64.

Vergnaud AC, Norat T, Romaguera D, et al. 2010. Meat consumption and prospective weight change in participants of the EPIC-PANACEA study. *Am J Clin Nutr,* 92, 398–407.

Vineis P & Perera F 2007. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol Biomarkers Prev,* 16, 1954–65.

Wiseman M 2008. The second World Cancer Research Fund/American Institute for Cancer Research expert report. Food, nutrition, physical activity, and the prevention of cancer: A global perspective. *Proceedings of the Nutrition Society,* 67, 253–6.

Wishart DS, Tzur D, Knox C, et al. 2007. HMDB: the Human Metabolome Database. *Nucleic Acids Res,* 35, D521–6