This is an open access article published under a Creative Commons Attribution (CC-BY) License, which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

Journal of proteome research

Article

pubs.acs.org/jpr

# Improving Visualization and Interpretation of Metabolome-Wide Association Studies: An Application in a Population-Based Cohort Using Untargeted $^1$H NMR Metabolic Profiling

Raphaële Castagné,[†,‡] Claire Laurence Boulangé,[§] Ibrahim Karaman,[†,‡] Gianluca Campanella,[†,‡] Diana L. Santos Ferreira,[†,‡] Manuja R. Kaluarachchi,[§] Benjamin Lehne,[†,‡] Alireza Moayyeri,[†,‡,⊥] Matthew R. Lewis,[§] Konstantina Spagou,[§] Anthony C. Dona,[§] Vangelis Evangelos,[†,¶] Russell Tracy,[#] Philip Greenland,[‖] John C. Lindon,[§,□] David Herrington,[▼] Timothy M. D. Ebbels,[§,□] Paul Elliott,[†,‡] Ioanna Tzoulaki,[†,‡] and Marc Chadeau-Hyam*[,†,‡]

[†]Department of Epidemiology and Biostatistics, School of Public Health and [‡]MRC-PHE Centre for Environment and Health, Imperial College London, W2 1PG London, United Kingdom

[§]Bioincubator Unit, Metabometrix Ltd, Bessemer Building, Prince Consort Road, South Kensington, London SW7 2BP U.K.

[⊥]Farr Institute of Health Informatics Research, University College London Institute of Health Informatics, 222 Euston Road, NW1 2DA London, United Kingdom

[¶]Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina 45110, Greece

[#]Department of Pathology and Laboratory Medicine, University of Vermont Larner College of Medicine, Burlington, Vermont 05405, United States

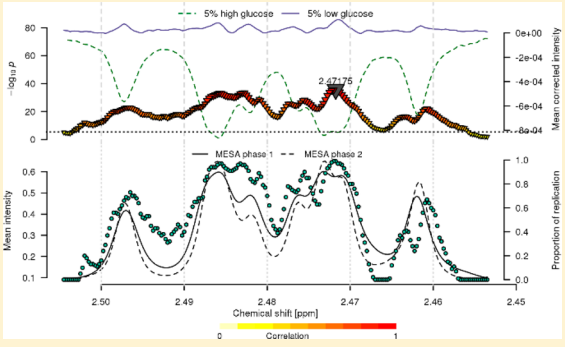
[‖]Department of Preventive Medicine and the Institute for Public Health and Medicine, Northwestern University, Chicago, Illinois 60611, United States

[□]Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, SW7 2AZ London, United Kingdom

[▼]Section on Cardiovascular Medicine, Department of Internal Medicine, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, North Carolina 27157, United States

Supporting Information

**ABSTRACT:** $^1$H NMR spectroscopy of biofluids generates reproducible data allowing detection and quantification of small molecules in large population cohorts. Statistical models to analyze such data are now well-established, and the use of univariate metabolome wide association studies (MWAS) investigating the spectral features separately has emerged as a computationally efficient and interpretable alternative to multivariate models. The MWAS rely on the accurate estimation of a metabolome wide significance level (MWSL) to be applied to control the family wise error rate. Subsequent interpretation requires efficient visualization and formal feature annotation, which, in-turn, call for efficient prioritization of spectral variables of interest. Using human serum $^1$H NMR spectroscopic profiles from 3948 participants from the Multi-Ethnic Study of Atherosclerosis (MESA), we have performed a series of MWAS for serum levels of glucose. We first propose an extension of the conventional MWSL that yields stable estimates of the MWSL across the different model parameterizations and distributional features of the outcome. We propose both efficient visualization methods and a strategy based on subsampling and internal validation to prioritize the associations. Our work proposes and illustrates practical and scalable solutions to facilitate the implementation of the MWAS approach and improve interpretation in large cohort studies.




**KEYWORDS:** *full resolution $^1$H NMR, metabolome wide association study, multiple testing correction, significance level, cohort studies, molecular epidemiology, MESA, results visualization and prioritization, high-throughput analysis, metabolic profiling*

## INTRODUCTION

Over the past 15 years, improvements in high-throughput technologies have accelerated the simultaneous measurement of large numbers of metabolites in a single sample using NMR and mass

**Received:** May 30, 2017
**Published:** August 20, 2017

 © 2017 American Chemical Society 

DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

spectrometry (MS), which are both widely used to characterize a biofluid using either untargeted or targeted profiling approaches.[1,2] Both technologies have emerged as efficient tools for identifying biomarkers of exposures as well as early disease manifestations, hence informing on molecular mechanisms involved in pathogenesis (e.g., in cancer, diabetes, cardiovascular, and neurological diseases).[3−6] Metabolic phenotyping uses robust and reliable analytical methods[7,8] that are ideally suited for untargeted profiling since prior knowledge about compounds present in a sample is not required. Such an agnostic discovery approach can inform complementary targeted methods by identifying novel chemical compounds that may contribute to the molecular pathways involved in complex phenotypes.[8−10] These screening exercises call upon the efficient analyses of high-dimensional data whose exploration poses complex methodological and interpretation problems.[11] Typically, untargeted NMR experiments in a large population study generate tens of thousands of data points for thousands of individuals. By adopting univariate approaches, the concept of the metabolome wide association study (MWAS)[12] has been proposed to analyze these data, and various statistical approaches to perform such analyses were established and have been explored and reviewed.[13−17] The complex correlation structures existing in metabolic profiles result in partially redundant signals across the metabolic spectra. These need to be accounted for while correcting for multiple testing, for instance by using a permutation-based procedure to derive the metabolome-wide significance level (MWSL) controlling the family wise error rate (FWER).[13] However, the comparative applicability of these approaches, as well as the visualization of the results they produce, has not been comprehensively explored. Here we provide, using a real-life example from the COMBI-BIO project, an extension of the MWAS methodology we initially developed using simulated data sets in a class discrimination context[13] and further extended to accommodate continuous outcomes. The current extension includes the computation of a stable and reproducible statistical significance threshold and proposes ways to estimate consistently across different types of NMR spectra as well as an internal validation procedure to assess the robustness of candidate associations. Our data set comprises two versions of $^1$H NMR nontargeted metabolic profiles in 3948 participants from the Multiethnic Study of Atherosclerosis (MESA) cohort.[18] As a proof-of-principle example, we focus on identifying the NMR spectral features associated with fasting blood serum glucose, which was measured by an independent technique (glucose oxidase method). By examining the full NMR spectrum, our analysis was designed to identify spectral regions corresponding to other metabolites beyond glucose itself that also vary with fasting serum glucose ($^1$H NMR glucose associated peaks).

## MATERIALS AND METHODS

### Study Population and Sample Selection

The MESA cohort has been described elsewhere[18] and includes 6814 participants (53% females, 47% males) aged 44−84 years (mean = 62 years) from four different ethnic groups: Chinese-American ($n$ = 803), African-American ($n$ = 1893), Hispanic ($n$ = 1496), and Caucasian ($n$ = 2622), all recruited between 2000 and 2002 at clinical centers in the United States. Participants were free of symptomatic cardiovascular disease at baseline, and demographic, medical history, anthropometric, lifestyle data, and serum samples were collected during the first examination (July 2000−August 2002), together with information on lipid or blood pressure treatment, and diabetes, and measures of systolic blood pressure. Serum samples were stored at −80 °C after collection. At enrolment, high density lipoprotein (HDL-C) was measured in EDTA plasma on the Roche/Hitachi 911 Automatic Analyzer (Roche Diagnostics Corporation, Indianapolis, IN), and low density lipoprotein (LDL-C) was calculated using the Friedewald equation,[19] together with fasting serum glucose using the glucose oxidase method on the Vitros analyzer (Johnson and Johnson Clinical Diagnostics). Ethical approval was obtained by local ethical review boards, and subsequent analysis was conducted in full accordance with the ethical approval obtained.

### Samples Preparation and $^1$H NMR Spectroscopic Acquisition

The full sample preparation and quality control procedure have been extensively described elsewhere.[20] Briefly, serum samples were thawed on the day prior to analysis, and 300 $\mu$L of each sample was mixed with 300 $\mu$L of phosphate buffer (NaHPO4, 0.075M, pH = 7.4). Samples were processed in two phases (each corresponding to a separate analytical batch). Eppendorfs were used for phase 1, while 96-well plates were used for phase 2. After centrifugation (12 000$g$ at 4 °C for 5 min), 550 $\mu$L of each sample-buffer mixture was manually transferred into SampleJet 5 mm diameter NMR tubes and kept at 4 °C until analysis. Different types of quality control (QC) samples were used for each phase as described elsewhere.[20] All QC pools were aliquoted in 350 $\mu$L and stored at −80 °C prior to analysis. $^1$H NMR spectra were acquired using a Bruker DRX600 spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 600 MHz. A standard water suppressed one-dimensional spectrum (usually termed NOESY) and a Carr−Purcell−Meiboom−Gill (CPMG) spectrum were obtained for each sample.[20]

### $^1$H NMR Metabolite Profiling

The metabolite profiling workflow has been detailed elsewhere.[20] For each biosample, two NMR profiles were generated: (i) a 1-D spectrum (NOESY) showing resonances from all proton-containing molecules in the sample, including broad, largely undefined bands from serum proteins, sharper and well-defined bands from serum lipoproteins (with some classification into their main groups), and sharp peaks from a range of small molecule metabolites such as amino acids, simple carbohydrates, organic acids, organic bases, and a number of osmolytes, and (ii) a Carr−Purcell−Meiboom−Gill (CPMG) spectrum that attenuates the peaks from the macromolecules and allows better definition of the small molecules.

For both CPMG and NOESY NMR data, in-house written MATLAB (Mathworks Inc., USA) routines were utilized for phasing and baseline correction. Prior to spectral peak alignment, the region $\delta$ 4.400−5.100 corresponding to the $H_2O$ resonance was removed. Spectral peak alignment was performed by the Recursive Segment-wise Peak Alignment (RSPA) algorithm.[21] Regions where the peaks of different suspected contaminations (i.e., methanol) occurred were removed from the whole spectra ($\delta$ 1.180−1.240, $\delta$ 2.244−2.261 and $\delta$ 3.660−3.710). The remaining spectral regions were normalized by probabilistic quotient normalization using the median spectrum as the reference.[22] The normalized high resolution spectra contained 30 590 data points (variables) for both CPMG and NOESY data sets. To ensure comparability across batches (i.e., across studies and phases), each variable was mean-centered[23] (we will hereafter refer to this as mean corrected intensities). Score plots


DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

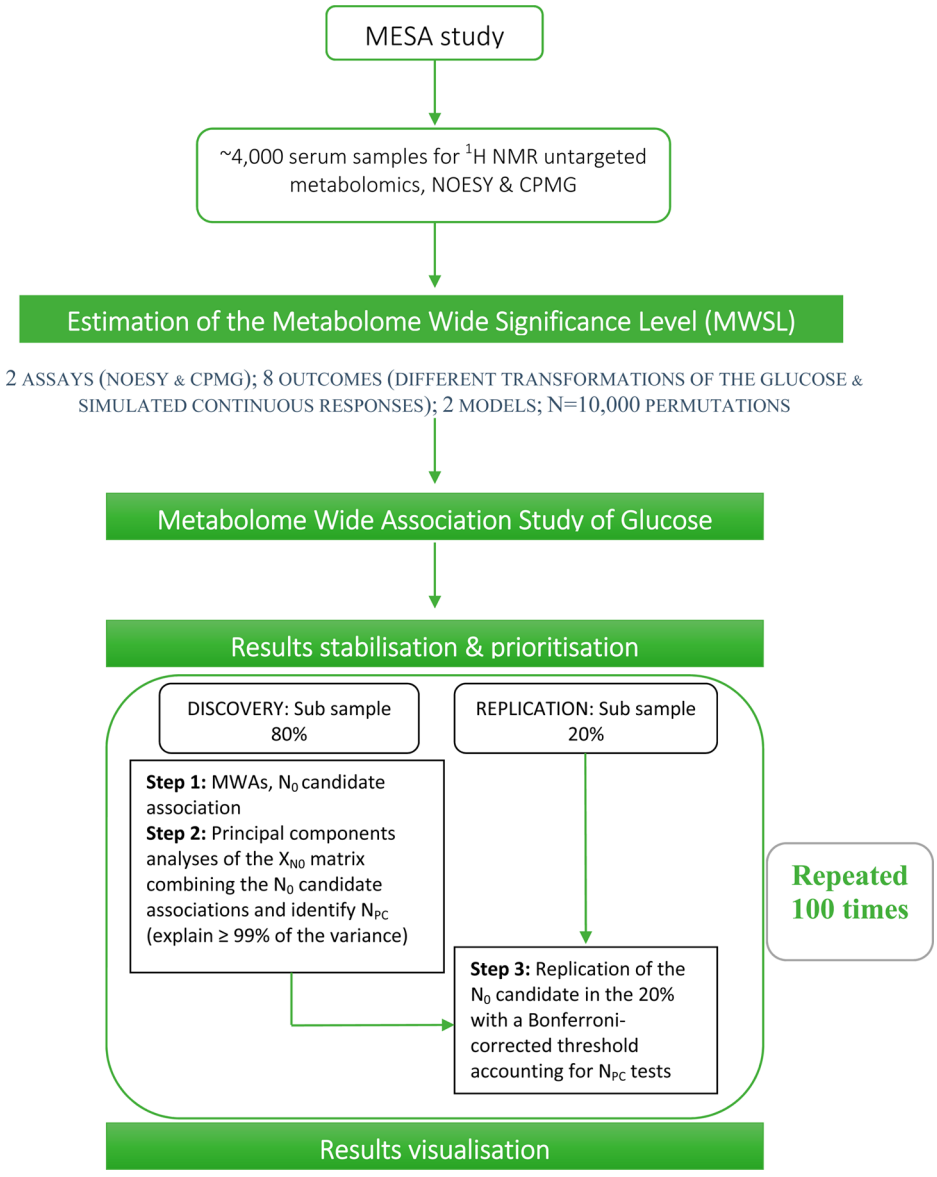



**Figure 1.** Analytical workflow.

of the first few principal components for the resulting data set were finally used to identify and remove potential outlier samples ($N$ = 3 for the CPMG and $N$ = 4 for the NOESY data) from further analyses. The MESA study population was further examined for potential outliers using score plots from the first two principal components (Figure S-1), which showed strong homogeneity in the study participants.

## Metabolite Assignments

Selected samples were analyzed using a range of 2-D NMR experiments such as total correlation spectroscopy (TOCSY) or heteronuclear single quantum coherence (HSQC) to aid molecular identification. We used approaches such as statistical total correlation spectroscopy (STOCSY) and STORM to help constrain possible molecular structures.[24,25] Peak identification in the $^1$H NMR data was supported by a semiautomatic clustering of the full resolution $^1$H NMR spectra (30 590 data points) using statistical recoupling of variables,[26] where the algorithm defines a cluster as containing 10 or more variables. Each cluster is subsequently checked by NMR experts to improve the data point grouping and identify peak overlaps. From our data, 136 and 159 clusters were identified in $^1$H NMR NOESY and CPMG data, respectively, each of them corresponding to a single or a group of peaks. Resulting spectral information was also compared to the literature[27,28] and to existing databases such as the Human Metabolome Database[29] (HMDB, http://www.hmdb.ca/) and the Biological Magnetic Resonance Data Bank[30] (BMRB, http://www.bmrb.wisc.edu). Resulting sets of NMR features were ultimately confirmed through spiking experiments using commercial standards. The level of assignment (LoA) we used was adapted from Sumner et al.[31] NMR features were also characterized by their peak multiplicity: singlet (s), doublet (d), triplet (t), quartet (q), doublet of doublet (dd), multiplet (m), broad peak (b), and noise (n). Ninety-two clusters were assigned to 44 unique metabolites for the NOESY data set, and 91 clusters were assigned to 48 unique metabolites for the CPMG data set.

## Metabolome-Wide Association Study (MWAS)

Figure 1 presents our analytical workflow. We adopted a MWAS approach using univariate linear regression models to systematically



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

screen the 30 590 data points assayed in both NOESY and CPMG profiles.

For a given data point, the linear model can be formulated as

$$Y^i \approx \alpha + \beta_1 X^i + \mathrm{B}_2 \mathrm{FE}^i + \varepsilon^i$$

where $X^i$ represents the normalized NMR intensity, $Y^i$ is the outcome variable; here the $\log_{10}$ transformed blood glucose concentration and $\varepsilon^i$ is the error term for an observation $i$. $\alpha$ is the intercept of the model, and ß$_1$ measures how strongly each data point influences the outcome variable. $\mathrm{FE}^i$ is a vector of fixed effect observations for individual $i$ and the vector $\mathrm{B}_2$ compiles the regression coefficients for each adjustment covariate; for model 1, age, gender, phase, and ethnicity, and for model 2, we additionally correct for LDL and HDL cholesterol, lipids and blood pressure treatment, systolic blood pressure, smoking status, and diabetes.

Most of the recently published MWAs adopted a Bonferroni correction for multiple testing,[32−34] which ignores correlation across variables and therefore does not account for the (partial) redundancy across the statistical test performed. This may lead to an overly conservative multiple testing correction. To take into account the high degree of correlation in spectral data and prevent overcorrection for multiple testing, one computationally efficient method relies on the estimation of the effective number of tests (ENT), as defined by the virtual number of independent tests that are performed across the actual $p$ tests performed. ENT measures the level of correlation within the spectral data and can be estimated through spectral decomposition of the correlation matrix of predictors.[35] From this estimate, the per-test significance level ensuring a FWER control can be defined as the Bonferroni-corrected threshold corresponding to that number of independent tests. However, this approach remains of limited use in real-life metabolomic data sets, notably because, for numerical reasons, the ENT is upper-bounded by the number of observations.[36]

As a scalable alternative, we used a permutation-based method to estimate the metabolome wide significance level (MWSL or $\alpha'$)[13] in which the outcome (glucose levels) is randomly shuffled across observations. Each permutation mimics the null hypothesis of no association, and we performed for each permuted data set a MWAS using the linear regression model described above. In that setting, and for a given permuted data set, the minimum $p$-value across all variables (denoted $q$) represents the largest significance level to be considered to ensure no false positive findings, and the MWSL $\alpha'$ controlling the FWER at a level $\alpha$ can be derived from the distribution of $q$ across the $N$ (set here to 10 000) permutations.

The effective number of tests (ENT) is then defined as the number of independent tests that would be required to obtain the same significance level using a Bonferroni correction: ENT= $\alpha/\alpha'$ and measures the level of correlation across the $p$ tests performed.

To assess the robustness of the ENT estimates and to circumvent the strong assumptions from the generalized linear model on data structure (independence of each data points, distribution of the residuals, variance structure, and linear relationship between response and predictors), we ran our permutation-based procedure for both data sets (NOESY and CPMG) for both models (1 and 2) and used different transformations of the glucose distribution: raw concentrations, truncated concentrations (129 outlying observations with more than 2 standard deviations away from the mean glucose level were discarded), and log10-transformed glucose levels (Figure S-2).

To formally assess the sensitivity of our MWSL estimates to distributional features of the outcome, we also simulated continuous responses from gamma and several Gaussian distributions and compared resulting MWSL estimates to those obtained using measured glucose levels.

### Sensitivity, Stability Analyses, and Results Prioritization

We performed further sensitivity analyses to assess the stability of the candidate associations we identified. These included a cross-validation procedure based on the independent subsampling ($N$ = 100 times) of discovery (containing 80% of the observations) and replication (comprising the remaining 20% observation) data sets.

For each 80:20 discovery and replication split, we performed a MWAS and used the MWSL to identify the candidate associations in the discovery set. For each discovery set (80% of the observations), the number of independent signals among the candidate associations was approximated by the number of principal components needed to explain more than 99% of their variance. That number of PCs was then used to compute the Bonferroni corrected MWSL used in the replication set.[37]

Our strategy could be summarized as follows for a given split:

(1) MWAS: identify ($N_0$) candidate associations in the discovery set with discovery $p$-value < MWSL.
(2) Estimate the number of independent signals these $N_0$ correspond to run a PCA on the $X_{N0,}$ the matrix combining the $N_0$ data points declared significant in the discovery set (a random 80% subsample or the full study population), and identify $N_{PC}$, the number of PC's needed to explain more than 99% of the variance in $X_{N0}$.
(3) Replication: identify from the candidate signals (step 1) those replicating in the 20% replication set at a Bonferroni-corrected significance level accounting for $N_{PC}$ tests, $\alpha/ N_{PC}$, setting $\alpha$ = 5%.

Steps 1−3 are repeated across the 100 independent splits and the proportion a given signal is identified in the discovery set, and replicated in the validation set is reported.

Statistical analyses were all performed using R v3.1.2.[38]

## RESULTS AND DISCUSSION

A total of 3948 individuals from the MESA cohort were included in the analysis, and for each individual, 30 590 serum NMR features were measured for both NOESY and CPMG spectra. The characteristics of the study population are summarized in Table 1.

We first explored the sensitivity of the MWSL estimates to (i) the parametrization of the statistical model, (ii) the type of NMR data under investigation, and (iii) distributional features of the outcome of interest. As summarized in Table 2, we then ran the permutation procedure for two different models (Models 1 and 2), for both types of NMR data sets (NOESY and CPMG) and 3 versions of glucose blood concentrations: raw, log10 transformed and truncated (removing 129 outlying observations which were outside the 2 standard deviation range from the mean glucose level). Across all models investigated, MWSL estimates for CPMG are more stringent than for NOESY spectra and, correspondingly, ENT estimates for CPMG are much greater than those for NOESY data (ranging from 17 610 to 122 460 and from 3680 to 17 010 for CPMG and NOESY, respectively). This suggests stronger correlations within NOESY data, which is plausible given that NOESY NMR spectra contain stronger broad peaks from proteins and lipoproteins than CPMG spectra, such that data point intensities are highly correlated.


DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

**Table 1. Summary Characteristics of the Study Population: Multiethnic Study of Atherosclerosis Cohort**

| | | N | % or mean (sd) |
|---|---|---|---|
| gender | men | 1951 | 49.4 |
| | women | 1997 | 50.6 |
| age (y) | all | 3948 | 62.9 (10.3) |
| phase | 1 | 1976 | 50 |
| | 2 | 1972 | 50 |
| ethnicity | Caucasian | 1521 | 38.5 |
| | Hispanic | 926 | 23.4 |
| | African-American | 968 | 24.5 |
| | Chinese-American | 533 | 13.5 |
| body mass index (kg/m$^2$) | all | 3948 | 28.2 (5.4) |
| glucose (mg/dL) | all | 3945 | 98.3 (31.1) |
| | missing | 3 | |
| LDL cholesterol (mg/dL) | all | 3884 | 117.3 (31.4) |
| | missing | 64 | |
| HDL cholesterol (mg/dL) | all | 3942 | 50.7 (14.7) |
| | missing | 6 | |
| systolic blood pressure (mmHg) | all | 3948 | 127.1 (21.3) |
| height (cm) | all | 3948 | 166.4 (10.2) |
| diabetes | no | 3387 | 85.8 |
| | yes | 561 | 14.2 |
| lipids treatment | no | 3286 | 83.2 |
| | yes | 662 | 16.8 |
| blood pressure treatment | no | 2449 | 62.1 |
| | yes | 1497 | 37.9 |
| smoking | never | 1988 | 50.4 |
| | former | 483 | 12.2 |
| | current | 1461 | 37 |
| | missing | 16 | 0.4 |

As summarized in Table 2, irrespective of the type of spectrum, MWSL estimates (and corresponding ENT) seem to be only marginally affected by the number of confounders included in the model (i.e., comparing models 1 and 2). However, a systematic but moderate increase in the ENT is observed for the fully adjusted model (Model 2). This can be explained by the fact that the fully adjusted model removes spectral signals relating to the components of the Framingham risk scores (such as the conventionally measured lipoprotein levels), and these confounders were highly likely to be driving correlations across some data points.

For CPMG data, estimates of the MWSL using raw glucose concentrations led to an estimated ENT greater than the actual number of tests (ratio around 4.00). This might be attributed to the parametric assumption in generalized linear regression model (e.g., normality of the outcome and equal variance) underlying our permutation procedure being violated. This is supported by the distribution of the blood levels of glucose, which is right skewed, with several outlying observations (Figure S-2A).

To account for this asymmetrical distribution, we first applied a $\log_{10}$ transformation to glucose levels (Figure S-2B). Although the transformed distribution remains right skewed, corresponding estimates of the MWSL were less stringent, and the effective to actual number of tests ratio dropped to around 0.75 for CPMG (and to 0.15 for NOESY). To remove the influence of outlying observations, we truncated the glucose distribution and removed observations more than 2 standard deviations away from the mean glucose level (Figure S-2C). While only a small number of observations were discarded ($N$ = 129), this removal strongly impacted the MWSL estimates for CPMG data, and the effective to actual number of test ratio dropped to less than 60% for both models.

These results suggest that MWSL estimates are sensitive to the shape of the distribution of the continuous outcome under investigation, and are specifically affected by both the relative weight of its tail, and by the presence of outlying observations. To formally assess the sensitivity of our MWSL estimates to the parametric form of the response variable, we ran a series of sensitivity analyses where we randomly sampled for each participant (and for each permutation) the glucose levels from (i) a Gamma distribution fitted on the measured glucose levels (shape = 15.90, scale = 6.18), and (ii) several Gaussian distributions (mean = 0, sd = 1; mean = 0, sd = 10; mean = 0, sd = 100, mean = mean(glucose), sd = sd(glucose). By construction, the Gamma-distributed response did not include outlying observations, but featured an inflated right tail, which provided less stringent MWSL for both NOESY. Results from the normally distributed outcomes showed consistent MWSL estimates and did not seem to be strongly affected by the parameters choices defining the Gaussian distribution (Table S-1, Figure S-3). Since the numbers of associated variables were only marginally affected by the way the ENT was computed in our example, we took forward the MWSL estimated from the Gaussian simulated outcome (mean = 0, sd = 1), which also seemed to provide the

**Table 2. Significance Threshold $\alpha'$ and Effective Number of Test (ENT) Based on a Bonferroni Correction[ab]**

| | | | | NOESY (30 590 variables) | | CPMG (30 590 variables) | |
|---|---|---|---|---|---|---|---|
| phenotype | $N$ | model | | MWSL (FWER = 5%) | | MWSL (FWER = 5%) | |
| glucose | 3945 | 1 | $\alpha'$ ($\times 10^{-5}$) | 0.31 (0.26; 0.35) | | 0.04 (0.03; 0.05) | |
| | | | ENT ($\times 10^3$) | 16.33 (14.42; 19.51) | **53.39** | 122.46 (108.67; 143.67) | **400.32** |
| | 3866 | 2 | $\alpha'$ ($\times 10^{-5}$) | 0.29 (0.25; 0.32) | | 0.04 (0.04; 0.05) | |
| | | | ENT ($\times 10^3$) | 17.01 (15.57; 19.88) | **55.60** | 121.64 (107.21; 142.23) | **397.64** |
| $\log_{10}$(glucose) | 3945 | 1 | $\alpha'$ ($\times 10^{-5}$) | 1.09 (0.96; 1.17) | | 0.22 (0.20; 0.23) | |
| | | | ENT ($\times 10^3$) | 4.59 (4.28; 5.22) | **15.01** | 23.1 (21.7; 24.4) | **75.37** |
| | 3866 | 2 | $\alpha'$ ($\times 10^{-5}$) | 1.05 (0.96; 1.13) | | 0.21 (0.19; 0.22) | |
| | | | ENT ($\times 10^3$) | 4.75 (4.43; 5.02) | **15.52** | 23.6 (22.7; 26.2) | **77.07** |
| glucose without outliers | 3816 | 1 | $\alpha'$ ($\times 10^{-5}$) | 1.36 (1.25; 1.45) | | 0.28 (0.26; 0.29) | |
| | | | ENT ($\times 10^3$) | 3.68 (3.45; 4.01) | **12.02** | 18.02 (17.36; 19.07) | **58.92** |
| | 3743 | 2 | $\alpha'$ ($\times 10^{-5}$) | 1.06 (1.00; 1.10) | | 0.28 (0.27; 0.30) | |
| | | | ENT ($\times 10^3$) | 4.70 (4.53; 4.99) | **15.36** | 17.61 (16.74; 18.74) | **57.57** |

[a]95% confidence intervals are given in parentheses. Figures are based on 10 000 permutations for each model (1 and 2) and given for the glucose, log10(glucose) and the glucose after outliers exclusion ($N$ = 129 excluded). [b]Bold figures are the ratios of effective/actual number of tests.



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

**Table 3. Number and Percentage of Associated Variables for Models 1 and 2 for Both NOESY and CPMG by Class of Metabolite[a]**

| | | NOESY | | | CPMG | | |
|---|---|---|---|---|---|---|---|
| model | | number of data points per group(%) | number of significant data points (%) | number of significant data points after results prioritization (%) | number of data points per group (%) | number of significant data points (%) | number of significant data points after results prioritization (%) |
| 1 | all[b] | 30 590 (100) | 22 066 (72) | 18 340 (83) | 30 590 (100) | 12 303 (40) | 9920 (81) |
| | amino-acids[c] | 4786 (15.65) | 3653 (76.3) | 3222 (88.2) | 4524 (14.79) | 2780 (61.5) | 2078 (74.7) |
| | carbohydrates[c] | 2394 (7.83) | 2204 (92.1) | 2192 (99.5) | 1826 (5.97) | 1763 (96.5) | 1734 (98.4) |
| | drug derivatives[c] | 191 (0.62) | 191 (100) | 116 (60.7) | 81 (0.26) | 30 (37) | 26 (86.7) |
| | lipids[c] | 4085 (13.35) | 2815 (68.9) | 2569 (91.3) | 3906 (12.77) | 2772 (71) | 2530 (91.3) |
| | nucleosides[c] | 116 (0.38) | 41 (35.3) | 35 (85.4) | 103 (0.34) | 13 (12.6) | 1 (7.7) |
| | organic acids[c] | 1383 (4.52) | 949 (68.6) | 838 (88.3) | 864 (2.82) | 390 (45.1) | 266 (68.2) |
| | others[c] | 377 (1.23) | 229 (60.7) | 173 (75.5) | 363 (1.19) | 261 (71.9) | 237 (90.8) |
| | proteins[c] | 549 (1.79) | 542 (98.7) | 500 (92.3) | 499 (1.63) | 499 (100) | 494 (99) |
| | unassigned | 15 978 (52.23) | 11 081 (69.4) | 8472 (76.5) | 10 315 (33.72) | 3414 (33.1) | 2252 (66) |
| 2 | all[b] | 30 590 (100) | 13 449 (44) | 9610 (71) | 30 590 (100) | 4909 (16) | 3355 (68) |
| | amino-acids[c] | 4786 (15.65) | 3123 (65.3) | 2168 (69.4) | 4524 (14.79) | 1048 (23.2) | 602 (57.4) |
| | carbohydrates[c] | 2394 (7.83) | 2079 (86.8) | 1906 (91.7) | 1826 (5.97) | 1695 (92.8) | 1653 (97.5) |
| | drug derivatives[c] | 191 (0.62) | 65 (34) | 36 (55.4) | 81 (0.26) | 24 (29.6) | 17 (70.8) |
| | lipids[c] | 4085 (13.35) | 1172 (28.7) | 176 (15) | 3906 (12.77) | 263 (6.7) | 116 (44.1) |
| | nucleosides[c] | 116 (0.38) | 7 (6) | 0 (0) | 103 (0.34) | 0(−) | 0(−) |
| | organic acids[c] | 1383 (4.52) | 798 (57.7) | 462 (57.9) | 864 (2.82) | 114 (13.2) | 18 (15.8) |
| | others[c] | 377 (1.23) | 112 (29.7) | 58 (51.8) | 363 (1.19) | 87 (24) | 35 (40.2) |
| | proteins[c] | 549 (1.79) | 529 (96.4) | 400 (75.6) | 499 (1.63) | 407 (81.6) | 352 (86.5) |
| | unassigned | 15 978 (52.23) | 5421 (33.9) | 4317 (79.6) | 10 315 (33.72) | 1067 (10.3) | 413 (38.7) |

[a]Results are also given after results prioritization: variables identified in the discovery set and replicated in the validation set in at least one split across the 100 splits (see Methods). [b]Figures are given for the whole spectra ($N$ = 30 590 variables) including the unassigned regions. [c]Figures are based on the achieved NMR assignment: not all variables have been assigned in the spectra.

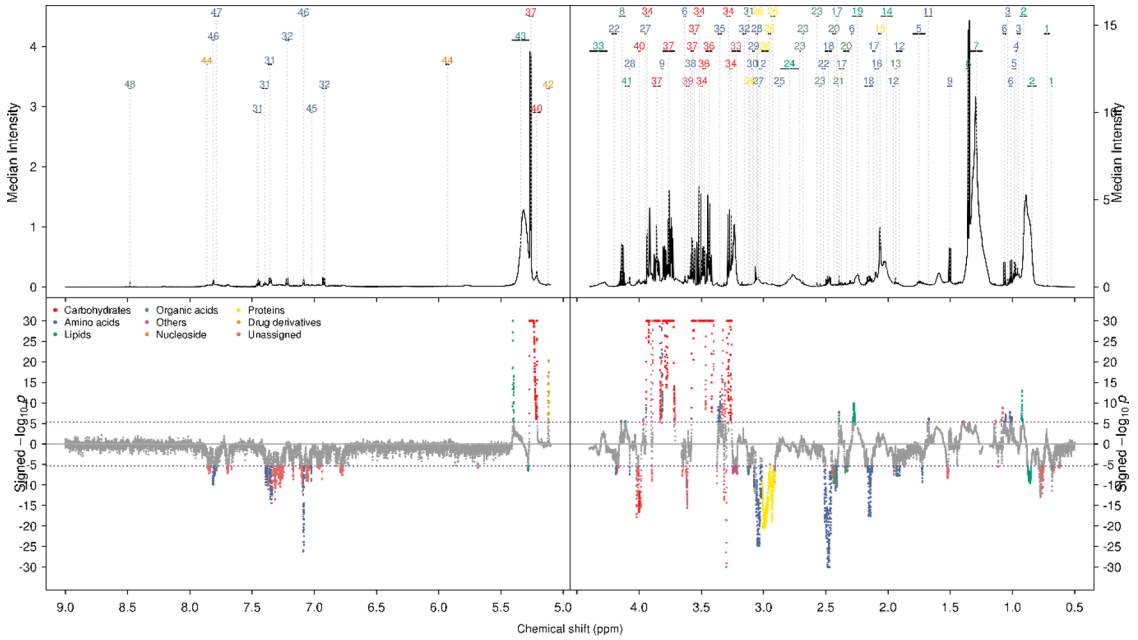



**Figure 2.** Metabolome wide study of glucose (model 2). This Manhattan plot shows the analysis of the 30 590 CPMG features. The signed negative log10 $p$-value is plotted against the chemical shift in ppm. To ease the visualization, all log $p$-value $\leq 10^{-30}$ were set to $1 \times 10^{-30}$. The horizontal dashed line indicates the $\alpha'$ per-test significance level controlling the FWER at a 5% level using the Gaussian simulated outcome. Data points are colored by class of metabolites. Components were: 1, L1; 2, L2; 3, isoleucine; 4, leucine, isoleucine; 5, leucine; 6, valine; 7, L3; 8, lactate; 9, alanine; 10, L4; 11, arginine; 12, lysine; 13, acetate; 14, L5; 15, acetylglycoproteins; 16, methionine; 17, glutamate; 18, glutamine; 19, L6; 20, 3-hydroxybutyrate; 21, pyruvate; 22, pyroglutamate; 23, citrate; 24, L7; 25, aspartate; 26, albumin; 27, creatine; 28, creatinine; 29, ornithine, tyrosine; 30, ornithine; 31, phenylalanine; 32, tyrosine; 33, choline; 34, beta-glucose; 35, proline; 36, alpha-glucose, beta-glucose; 37, alpha-glucose; 38, glycine; 39, glycerol; 40, mannose; 41, glyceryl groups of lipids; 42, APAP glucuronide; 43, L8; 44, uridine; 45, 1-Methylhistidine; 46, histidine; 47, 3-methylhistidine; 48, formate.

best balance between generalizability and simplicity (Figures S-3 and S-4). We further compared the proportion of associated variables from different multiple testing correction strategies including Bonferroni and FWER control, and Benjamini-Hochberg[39] false discovery rate (BH-FDR) procedure (Figure S-5). In all scenario considered, the largest proportion of associated



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

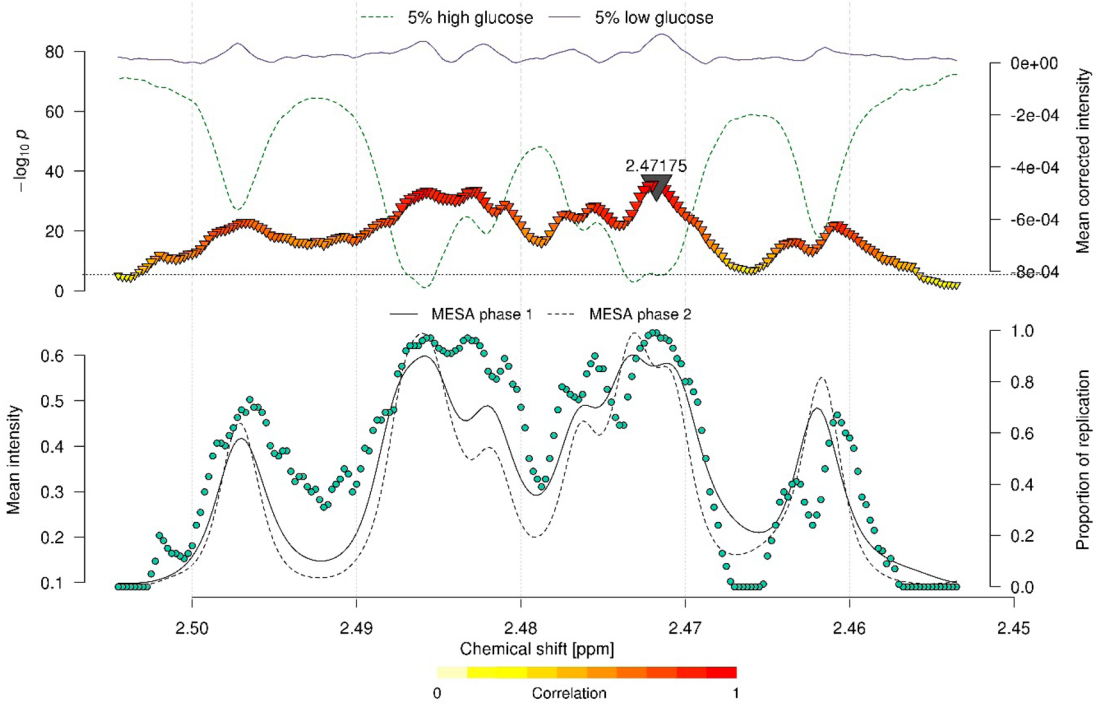



**Figure 3.** CPMG-model 2 regional association plots with $\log_{10}$ (glucose) for the glutamine. In the upper plot, the −log10 *p*-value for the features at two regions are shown on each plot. Features are colored based on their correlation with the gray hit that has the smallest *p*-value in the region. The lines show the mean corrected intensity (i.e., residuals removing the linear effect of the phase and the cohort) in the 5% of samples with high residual glucose in green and 5% of the samples with low residual glucose in blue. The bottom plot shows the mean spectral intensity in MESA phase 1 (plain line) and in MESA phase 2 (dashed line). Green circles indicate the proportion of replication after results prioritization.

variables was always observed when using a BH-FDR procedure, and the smallest proportion of associated variables was always observed when using a Bonferroni procedure (Figure S-5).

## Metabolome Wide Association Study of Glucose: Visualization and Prioritization

We performed the MWAS of blood levels of glucose on both CPMG and NOESY spectra using both models and setting the MWSL to that estimated using the Gaussian simulated outcome (mean = 0, sd = 1, Table S-1). As expected, irrespective of the model and of the type of spectrum, a very large number of spectral features were found associated with blood concentration of glucose (Table 3): for NOESY data, 72% and 44% of the spectral variables were found significantly associated with glucose level for models 1 and 2, respectively. These proportions were 40% and 16% for models 1 and 2, respectively, in CPMG data.

Analyses by classes of metabolite for model 1 revealed that nearly all variables assigned to drug derivative (100%), proteins (98.7%), and carbohydrates (92.1%) were associated with glucose in NOESY. Similarly, all variables assigned to proteins (100%), carbohydrates (96.5%), and others (71.9%, which include choline and glycerol) were associated with glucose in CPMG. The proportion of unassigned associated variables was higher for NOESY (69.4%) compared to CPMG (33.9%).

Adjusting for Framingham risk score (FRS) variables in model 2 reduced the total number of associations for both NOESY and CPMG spectra. This reduction was mostly observed for the lipids class (NOESY, 68.9% vs 28.7%; CPMG, 71% vs 6.7%) and the "others" class (NOESY, 60.7% vs 29.7%; CPMG, 71.9% vs 24%). As expected, the proportion of carbohydrate associated variables was one of the least affected classes by the FRS adjustment (NOESY, 92.1% vs 86.8%; CPMG, 96.5% vs 92.8%).

The utility of the MWAS approach in metabolic phenotyping relies on explicit visualization of the results. One primary output to help identify relevant spectral regions borrows from the field of genome-wide association studies and reports for each spectral variable the−log10 *p*-value multiplied by the sign of the corresponding regression coefficient. The resulting signed Manhattan plots (Figure 2 and Figure S-6 for CPMG and NOESY, respectively) offer a global view of the spectral regions associated with the outcome of interest, and can be further informed by the annotation of spectral features. Assigned metabolites found associated with conventional serum glucose measurements are reported in Tables S-2 and S-3 for CPMG and NOESY, respectively. For CPMG, 47 unique metabolites were associated with glucose for model 1, of which 34 were still associated after controlling for the FRS variable (44 and 41 for NOESY).

High-resolution visualization is also key to enable peak validation and subsequent annotation through the inspection of the shape and multiplicity of the spectral features in the neighborhood of the associated regions. Such visualization could be provided by regional plots as exemplified in Figure 3, which focuses on the glutamine region ([2.4535−2.5045] ppm). The upper panel represents the high resolution (unsigned) Manhattan plot where the $-\log_{10}$ *p*-value (left Y axis) measuring the strength of association each between peak height and the $\log_{10}$ transformed blood glucose level is represented by a triangle (down pointing for negative associations). For the region under investigation, we define the reference feature as the strongest association (here 2.47175 ppm, *p*-value $<2.10^{-16}$, represented in gray) and color-code the pairwise correlation of each spectral variable with this reference. The mean-corrected intensity (i.e., residuals removing the effect of possible confounders: phase and cohort, see Methods) is also represented (right Y axis)



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

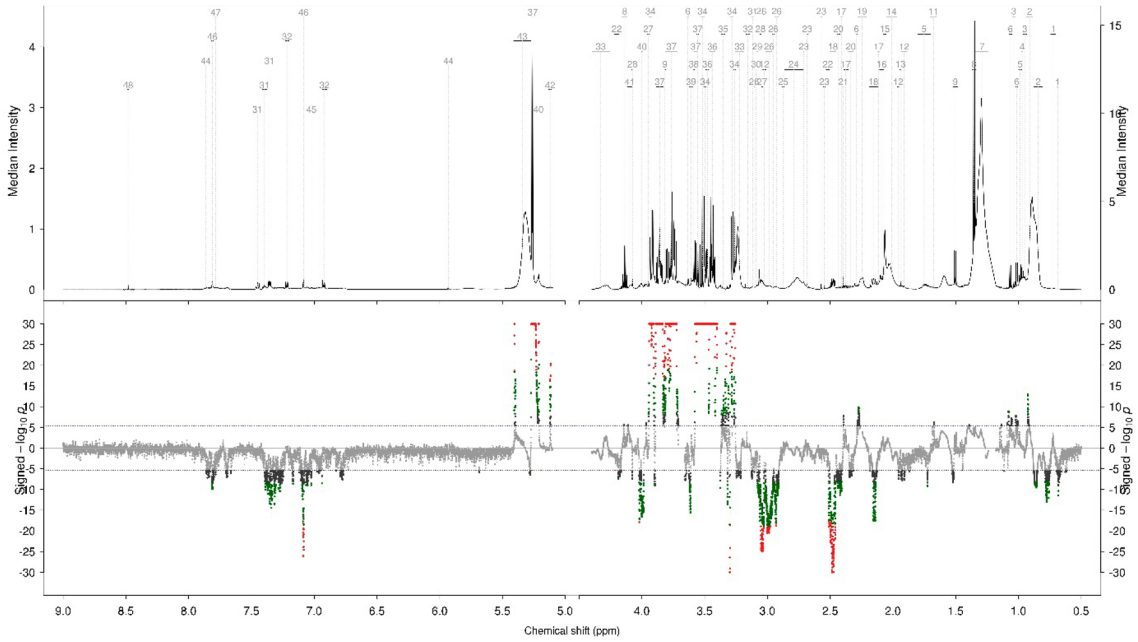



**Figure 4.** Comparison of results from the analysis in MESA to those from the 80:20 split strategy. Results are presented for the CPMG ($N$ = 30 590 data points) metabolome wide association study of glucose using model 2. To ease the visualization, all $p$-value $\leq 10^{-30}$ were set to $1 \times 10^{-30}$. The −log10($p$-value) is signed by the direction of the effect size estimate and is plotted against the chemical shift. The horizontal dashed line indicates the per-test significance level controlling the FWER at a 5% level. Variables found from the analyses in MESA are presented in black, those discovered and replicated at least once across the 100 splits are presented in green and those discovered and replicated in 50% of the split are presented in red. Components were: 1, L1; 2, L2; 3, isoleucine; 4, leucine, isoleucine; 5, leucine; 6, valine; 7, L3; 8, lactate; 9, alanine; 10, L4; 11, arginine; 12, lysine; 13, acetate; 14, L5; 15, acetylglycoproteins; 16, methionine; 17, glutamate; 18, glutamine; 19, L6; 20, 3-hydroxybutyrate; 21, pyruvate; 22, pyroglutamate; 23, citrate; 24, L7; 25, aspartate; 26, albumin; 27, creatine; 28, creatinine; 29, ornithine, tyrosine; 30, ornithine; 31, phenylalanine; 32, tyrosine; 33, choline; 34, beta-glucose; 35, proline; 36, alpha-glucose, beta-glucose; 37, alpha-glucose; 38, glycine; 39, glycerol; 40, mannose; 41, glyceryl groups of lipids; 42, APAP glucuronide; 43, L8; 44, uridine; 45, 1-methylhistidine; 46, histidine; 47, 3-methylhistidine; 48, formate.

for the samples in the fifth (blue line) and 95th (green line) quantiles of the residuals $\log_{10}$ glucose levels after controlling for the FRS variables.

From this plot, it is clear that there are strong, and exclusively positive correlation coefficients among $^1$H NMR data points throughout the region as indicated by the color-coded pairwise correlation coefficients (75% are >0.77). As expected, because of the local correlation structures in NMR spectra (caused by the finite peak width and the resonance being split by the J couplings), data points neighboring the reference signal exhibit higher correlation levels. More distant spectral variables may also show strong correlation with the reference peak (e.g., 2.5045, Spearman correlation = 0.26; 2.4535, Spearman correlation = 0.30), and irrespective of their location in the spectrum, intensities that are the most correlated to the reference peak exhibit the strongest associations with glucose levels. This arises because a given metabolite often has several NMR peaks and this procedure offers a way of finding such linked resonances as an aid to molecular annotation. For these variables (dark orange and red triangles), the difference in the mean corrected intensity in participants with the highest and lowest glucose levels are larger compared to the spectral variables less correlated to the reference peak (yellow triangles).

To get further insights into the nature (shape and multiplicity) of the variables found associated with glucose we represent the mean spectrum in the bottom panel (left Y axis). Mean spectra are plotted for each of the analytical phases to assess both potential technically induced bias across experimental batches, and possible population heterogeneity across samples assayed in each phase. While the mean spectrum from phase 2 appears to have more marked variables (with higher modes), both data sets yield consistent subspectra in terms of alignment, multiplicity and overall peak shape.

To accommodate the expected large number of associations with glucose levels, efficient signal prioritization strategies are key. One established way to identify robust and replicable associations is to seek external replication in an independent data set. However, owing to the specificity of experimental protocols and the nature of the measurements (which are relative and not absolute concentrations), external validation is challenging in metabolic phenotyping, but could however be sought for using either standardized or annotated data. As a workable alternative and complementary approach prioritizing relevant metabolic features, we propose to seek for internal validation and randomly split the study population in a discovery (80% of the full population) and a validation (the remaining 20%) set. The robustness of an association identified in the full population is then quantified by the number of times discovered and replicated over the ($N$ = 100) independent splits. This proportion is represented on the regional plot (green dots on the bottom panel of Figure 2). Owing to the strong signals we identify in our glucose MWAS, we report very high replication proportions in glucose-associated regions. As a first approach, prioritized associations were defined as those discovered and validated in at least one split (i.e., with a proportion of replication >0). As illustrated in Table 3 and in Figures 4 and S-7 for CPMG and NOESY, this prioritization strategy reduced the number of significant associations for model 2 by almost 50 and 30%, respectively. As illustrated in Figure 4, the associated spectral variables are strongly clustered and define clear regions that are associated with the blood levels of glucose.



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

Our prioritization strategy identifies overall the same regions along the spectra but selects preferentially the strongest associated variables within each region and regions that have been assigned. This is even more apparent when more stringent prioritization strategies, for instance, in Figure 4 we plot the signed Manhattan plot for the features discovered and replicated in a least half of the splits (in red).

## CONCLUSION

The identification and interpretation of metabolic features that contribute to a physiological and/or disease-induced outcome from full resolution NMR data is challenging for many reasons related to the dimensionality and complexity of metabolic profiles. The main aim of the present work is to address some of these issues through the development of a robust multiple testing strategy, intuitive visualizations and objective ways to prioritize results. The MWAS approach requires accurate correction for the large number of tests performed, and therefore needs to appropriately account for the strong and complex correlation structures within the NMR spectra. All methods for performing multiple testing correction assume a valid statistical model that captures dependencies in the data. While approaches controlling the FDR usually provide less stringent multiple testing correction, these have been reported to misperform in cases of high correlations among predictors.[40] As an extension of the MWAS approach to accommodate continuous variables in a regression context, we proposed an estimation of the metabolome wide significance level adopting the same permutation strategy and investigated the sensitivity of the estimates to the data structure and to the model parametrization. Our results suggest that in the case of highly correlated variables (spectral variables) that are strongly associated with an outcome (here glucose levels), permutations do not succeed in destroying the predictor-outcome relationship, hence yielding ENT estimates greater than the actual number of tests performed. Sensitivity analyses removing extreme values, and/or log transforming glucose levels showed that outlying observations are driving this unexpected estimate of the ENT. In the presence of strong correlation (i) between blood glucose and most of the assayed metabolites, (ii) among the metabolites, and (iii) between metabolites and adjustment variables, assumptions (e.g., observations exchangeability under the null) on which permutation inferences are based may be violated, and especially in the presence of strong outlying observations. From our data, MWSL estimates appeared robust to model parametrization (i.e., marginally affected by the set of confounding variables considered), but clearly depended on the correlation structure in the data, which are population and platform specific. This suggests that the MWSL should be tailored and re-estimated for each data set. MWSL estimates were also found to be sensitive to the distribution of the outcome and especially in the case of heavy tailed distributions due to the presence of outlying observations. While numerical transformations and truncation could be a way forward, one more general and conservative option could be to calculate the MWSL once using a virtual predictor sampled from a Gaussian distribution (Figure S-4).

Using this MWSL approach, we propose extensions of existing visualization tools for the MWAS. This includes full resolution signed Manhattan plots included functional annotation and higher resolution regional plots displaying the per-variable strength of association as well as spectral summary features including correlation patterns across spectral variables. In our proof-of-principle example, we chose glucose as the outcome of interest which defined a challenging context in terms of results interpretability as a very large number of strong associations were identified and the assessment of their relevance went beyond the observed strong positive correlation for the expected glucose peaks along the $^1$H NMR spectra. This called for the definition of strong result prioritization strategy, which, in less extreme situations, is also critical to identify the most relevant associations that are worth dedicating resources for molecular assignment. Our approach relies on subsampling strategy where discovery (80%)-replication (20%) splits are used to identify associations that internally replicate. This strategy was found to be efficient to prioritize the most robust associations, which were not only those with the strongest $p$-values. We performed our internal validation at the metabolome-wide level, which was computationally intensive. To scale this approach in real-life studies, one alternative would consist in restricting the discovery-replication split strategy for the MWAS candidates that have been identified in the full population. The overall analytical strategy presented here provides a general framework for the analysis of cohort studies where large number of samples are profiled using untargeted technologies (e.g., high-resolution NMR, mass spectrometry). Overall, we believe the strategy and approach we present are generalizable and scalable and may therefore be relevant to aid the MWAS approach, particularly improving the interpretation of results.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00344.

Plots of first two PCA scores for CPMG and NOESY data in MESA; distributions of three different transformations of glucose used to investigate MWSL; percentage of effective/actual number of tests and 95% confidence intervals for CPMG and NOESY; percentage of associated variables for CPMG and NOESY derived from each simulated continuous response; percentage of associated variables for CPMG and NOESY derived from different multiple testing correction strategies; metabolome wide study of glucose from analysis of 30 590 NOESY features; comparison of results from analysis in MESA to those from 80:20 split strategy from NOESY metabolome wide association study of glucose using model 2; significance threshold a′ and ENT based on Bonferonni correction; CPMG and NOESY metabolic features associated with $\log_{10}$ (glucose) in MESA at metabolome-wide significance level for models 1 and 2 (PDF)

## AUTHOR INFORMATION


### Corresponding Author

*E-mail: m.chadeau@imperial.ac.uk.

### ORCID

Claire Laurence Boulangé: 0000-0002-9022-0159
Ibrahim Karaman: 0000-0001-9341-8155
Paul Elliott: 0000-0002-7511-5684
Marc Chadeau-Hyam: 0000-0001-8341-5436


### Notes

The authors declare no competing financial interest.
A Step-by-step R tutorial used to estimate the effective number of tests and to produce Figures 2, and 3 is publicly available at



DOI: 10.1021/acs.jproteome.7b00344
J. Proteome Res. 2017, 16, 3623−3633

https://figshare.com/articles/Tutorial_for_MWAS_and_regional_plots/5319247.

## ■ ACKNOWLEDGMENTS

This work has been carried out as part of the of the FP7 project COMBI-BIO [305422 to P. E.]. MESA was supported by Contract Nos. HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by Grant Nos. UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 from NCATS. P.E. is Director of the MRC-PHE Centre for Environment and Health and acknowledges support from the Medical Research Council and Public Health England (MR/L01341X/1). P.E. acknowledges support from the NIHR Biomedical Research Centre at Imperial College Healthcare NHS Trust and Imperial College London, and the NIHR Health Protection Research Unit in Health Impact of Environmental Hazards (HPRU-2012–10141). This work used the computing resources of the UK MEDical BIOinformatics partnership (UKMED-BIO) supported by the Medical Research Council (MR/L01632X/1). The authors wish to thank all the centres that took part in the study and the additional members of the COMBI-BIO consortium. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

## ■ REFERENCES

(1) Lenz, E. M.; Wilson, I. D. Analytical strategies in metabonomics. *J. Proteome Res.* **2007**, *6* (2), 443–458.
(2) Lindon, J. C.; Nicholson, J. K. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *TrAC, Trends Anal. Chem.* **2008**, *27* (3), 194–204.
(3) Holmes, E.; Wilson, I. D.; Nicholson, J. K. Metabolic phenotyping in health and disease. *Cell* **2008**, *134* (5), 714–717.
(4) Madsen, R.; Lundstedt, T.; Trygg, J. Chemometrics in metabolomics–a review in human disease diagnosis. *Anal. Chim. Acta* **2010**, *659* (1–2), 23–33.
(5) Heather, L. C.; Wang, X.; West, J. A.; Griffin, J. L. A practical guide to metabolomic profiling as a discovery tool for human heart disease. *J. Mol. Cell. Cardiol.* **2013**, *55*, 2–11.
(6) Tzoulaki, I.; Ebbels, T. M. D.; Valdes, A.; Elliott, P.; Ioannidis, J. P. A. Design and analysis of metabolomics studies in epidemiologic research: a primer on -omic technologies. *Am. J. Epidemiol.* **2014**, *180* (2), 129–139.
(7) Dumas, M.-E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.; et al. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Anal. Chem.* **2006**, *78* (7), 2199–2208.
(8) Dona, A. C.; Jiménez, B.; Schäfer, H.; Humpfer, E.; Spraul, M.; Lewis, M. R.; Pearce, J. T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal. Chem.* **2014**, *86* (19), 9887–9894.
(9) Nicholson, J. K. Global systems biology, personalized medicine and molecular epidemiology. *Mol. Syst. Biol.* **2006**, *2*, 52.
(10) Nicholson, J. K.; Holmes, E.; Kinross, J. M.; Darzi, A. W.; Takats, Z.; Lindon, J. C. Metabolic phenotyping in clinical and surgical environments. *Nature* **2012**, *491* (7424), 384–392.
(11) Sévin, D. C.; Kuehne, A.; Zamboni, N.; Sauer, U. Biological insights through nontargeted metabolomics. *Curr. Opin. Biotechnol.* **2015**, *34*, 1–8.
(12) Bictash, M.; Ebbels, T. M.; Chan, Q.; Loo, R. L.; Yap, I. K. S.; Brown, I. J.; de Iorio, M.; Daviglus, M. L.; Holmes, E.; Stamler, J.; et al. Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. *J. Clin. Epidemiol.* **2010**, *63* (9), 970–979.
(13) Chadeau-Hyam, M.; Ebbels, T. M. D.; Brown, I. J.; Chan, Q.; Stamler, J.; Huang, C. C.; Daviglus, M. L.; Ueshima, H.; Zhao, L.; Holmes, E.; et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *J. Proteome Res.* **2010**, *9* (9), 4620–4627.
(14) De Livera, A. M.; Dias, D. A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T. P. Normalizing and integrating metabolomics data. *Anal. Chem.* **2012**, *84* (24), 10768–10776.
(15) De Livera, A. M.; Olshansky, M.; Speed, T. P. Statistical analysis of metabolomics data. *Methods Mol. Biol.* **2013**, *1055*, 291–307.
(16) Korman, A.; Oh, A.; Raskind, A.; Banks, D. Statistical methods in metabolomics. *Methods Mol. Biol.* **2012**, *856*, 381–413.
(17) Ebbels, T. M. D.; Lindon, J. C.; Coen, M. Processing and modeling of nuclear magnetic resonance (NMR) metabolic profiles. *Methods Mol. Biol.* **2011**, *708*, 365–388.
(18) Bild, D. E.; Bluemke, D. A.; Burke, G. L.; Detrano, R.; Diez Roux, A. V.; Folsom, A. R.; Greenland, P.; Jacob, D. R.; Kronmal, R.; Liu, K.; et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am. J. Epidemiol.* **2002**, *156* (9), 871–881.
(19) Friedewald, W. T.; Levy, R. I.; Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin. Chem.* **1972**, *18* (6), 499–502.
(20) Karaman, I.; Ferreira, D. L. S.; Boulangé, C. L.; Kaluarachchi, M. R.; Herrington, D.; Dona, A. C.; Castagné, R.; Moayyeri, A.; Lehne, B.; Loh, M.; et al. Workflow for Integrated Processing of Multicohort Untargeted (1)H NMR Metabolomics Data in Large-Scale Metabolic Epidemiology. *J. Proteome Res.* **2016**, *15* (12), 4188–4194.
(21) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M. D.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery. *Anal. Chem.* **2009**, *81* (1), 56–66.
(22) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* **2006**, *78* (13), 4281–4290.
(23) van Velzen, E. J. J.; Westerhuis, J. A.; van Duynhoven, J. P. M.; van Dorsten, F. A.; Hoefsloot, H. C. J.; Jacobs, D. M.; Smit, S.; Draijer, R.; Kroner, C. I.; Smilde, A. K. Multilevel data analysis of a crossover designed human nutritional intervention study. *J. Proteome Res.* **2008**, *7* (10), 4483–4491.
(24) Cloarec, O.; Dumas, M.-E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; et al. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic 1H NMR data sets. *Anal. Chem.* **2005**, *77* (5), 1282–1289.
(25) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M. D.; Nicholson, J. K. Subset optimization by reference matching (STORM): an optimized statistical approach for recovery of metabolic biomarker structural information from 1H NMR spectra of biofluids. *Anal. Chem.* **2012**, *84* (24), 10694–10701.
(26) Navratil, V.; Pontoizeau, C.; Billoir, E.; Blaise, B. J. SRV: an open-source toolbox to accelerate the recovery of metabolic biomarkers and correlations from metabolic phenotyping datasets. *Bioinformatics* **2013**, *29* (10), 1348–1349.
(27) Nicholson, J. K.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Lindon, J. C. 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. *Anal. Chem.* **1995**, *67* (5), 793–811.
(28) Merrifield, C. A.; Lewis, M.; Claus, S. P.; Beckonert, O. P.; Dumas, M.-E.; Duncker, S.; Kochhar, S.; Rezzi, S.; Lindon, J. C.; Bailey, M.; et al. A metabolic system-wide characterisation of the pig: a model for human physiology. *Mol. BioSyst.* **2011**, *7* (9), 2577–2588.


DOI: 10.1021/acs.jproteome.7b00344
*J. Proteome Res.* 2017, 16, 3623–3633

(29) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35* (Database), D521−D526.
(30) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; et al. BioMagResBank. *Nucleic Acids Res.* **2008**, *36* (Database), D402−D408.
(31) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3* (3), 211−221.
(32) Altmaier, E.; Fobo, G.; Heier, M.; Thorand, B.; Meisinger, C.; Römisch-Margl, W.; Waldenberger, M.; Gieger, C.; Illig, T.; Adamski, J.; et al. Metabolomics approach reveals effects of antihypertensives and lipid-lowering drugs on the human metabolism. *Eur. J. Epidemiol.* **2014**, *29* (5), 325−336.
(33) Sekula, P.; Goek, O.-N.; Quaye, L.; Barrios, C.; Levey, A. S.; Römisch-Margl, W.; Menni, C.; Yet, I.; Gieger, C.; Inker, L. A.; et al. A Metabolome-Wide Association Study of Kidney Function and Disease in the General Population. *J. Am. Soc. Nephrol.* **2016**, *27* (4), 1175−1188.
(34) Adkins, D. E.; McClay, J. L.; Vunck, S. A.; Batman, A. M.; Vann, R. E.; Clark, S. L.; Souza, R. P.; Crowley, J. J.; Sullivan, P. F.; van den Oord, E. J. C. G.; et al. Behavioral metabolomics analysis identifies novel neurochemical signatures in methamphetamine sensitization. *Genes Brain Behav.* **2013**, *12* (8), 780−791.
(35) Patterson, N.; Price, A. L.; Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2006**, *2* (12), e190.
(36) Schäfer, J.; Strimmer, K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **2005**, *4* (1), 1−32.
(37) Auro, K.; Joensuu, A.; Fischer, K.; Kettunen, J.; Salo, P.; Mattsson, H.; Niironen, M.; Kaprio, J.; Eriksson, J. G.; Lehtimäki, T.; et al. A metabolic view on menopause and ageing. *Nat. Commun.* **2014**, *5*, 4708.
(38) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2014; Vol. *73* (1), pp 3−36.
(39) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57* (1), 289−300.
(40) Zhang, J.; Coombes, K. R. Sources of variation in false discovery rate estimation include sample size, correlation, and inherent differences between groups. *BMC Bioinf.* **2012**, *13* (Suppl 13), S1.



DOI: 10.1021/acs.jproteome.7b00344
*J. Proteome Res.* 2017, 16, 3623−3633