Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 1586−1595

pubs.acs.org/jpr

Optimized Phenotypic Biomarker Discovery and Confounder Elimination via Covariate-Adjusted Projection to Latent Structures from Metabolic Spectroscopy Data

Joram M. Posma,[*,†,§] Isabel Garcia-Perez,[†,¶] Timothy M. D. Ebbels,[†] John C. Lindon,[†] Jeremiah Stamler,[#] Paul Elliott,[§,⊥] Elaine Holmes,[†,⊥,‡] and Jeremy K. Nicholson[*,†,⊥,‡]

[†]Division of Integrative Systems Medicine and Digestive Diseases, Department of Surgery and Cancer, Faculty of Medicine and [‡]MRC-NIHR National Phenome Centre, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, SW7 2AZ London, United Kingdom

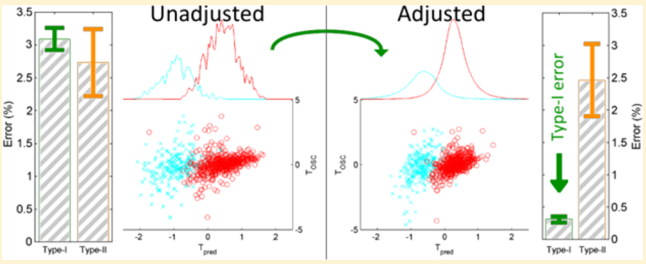
[§]Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine and [⊥]MRC-PHE Centre for Environment and Health, School of Public Health, Faculty of Medicine, Imperial College London, W2 1PG London, United Kingdom

[¶]Investigative Medicine, Department of Medicine, Faculty of Medicine, Imperial College London, W12 0NN London, United Kingdom

[#]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, United States

S *Supporting Information*

ABSTRACT: Metabolism is altered by genetics, diet, disease status, environment, and many other factors. Modeling either one of these is often done without considering the effects of the other covariates. Attributing differences in metabolic profile to one of these factors needs to be done while controlling for the metabolic influence of the rest. We describe here a data analysis framework and novel confounder-adjustment algorithm for multivariate analysis of metabolic profiling data. Using simulated data, we show that similar numbers of true associations and significantly less false positives are found compared to other commonly used methods. Covariate-adjusted projections to latent structures (CA-PLS) are exemplified here using a large-scale metabolic phenotyping study of two Chinese populations at different risks for cardiovascular disease. Using CA-PLS, we find that some previously reported differences are actually associated with external factors and discover a number of previously unreported biomarkers linked to different metabolic pathways. CA-PLS can be applied to any multivariate data where confounding may be an issue and the confounder-adjustment procedure is translatable to other multivariate regression techniques.



KEYWORDS: *biomarker discovery, chemometrics, confounder elimination, covariate adjustment, metabolic phenotyping, Monte Carlo cross-validation, multivariate data analysis, random matrix theory, reanalysis, sampling bias*

INTRODUCTION

Human metabolic phenotypes, "metabotypes",[1,2] are influenced by multiple interacting factors, such as dietary, environmental, genetic, and microbial variation,[2−6] and reflect the health status of an individual.[7] Metabotypes can be studied with metabonomics and metabolomics,[8,9] which utilize multivariate statistical methods to find relevant changes in metabolite profiles related to outcomes/responses. For this, urine and plasma/serum are the most desirable biofluids, as they can be (relatively) noninvasively obtained and they are not likely to be volume limited in humans.[10] Urine gives a homeostatic signature of all metabolic processes in a biological system, including genetic, diet, and gut microbial activity,[11] and thus, the variation in the urinary metabolic profile can be attributed to many factors other than disease risk, whereas the plasma/serum biological matrix holds information on physiological status at the specific sampling time.[7]

The measurement of metabolites in biofluid samples is most commonly performed using ^{1}H nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS), with the latter often preceded with a liquid/gas chromatography step for metabolite separation.[12] Both platforms yield data sets/matrices ($\mathbf{X}$) with thousands to ten-thousands of variables (p) with often a much lower number of samples (n). This wealth of information comes at a price: there are likely spurious findings which will need to be controlled for (type-I errors) and findings attributable to other factors rather than the response ($\mathbf{Y}$) (e.g.,

Received: December 8, 2017
Published: February 19, 2018

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

disease state). Discovering accurate information relating to etiology or pathogenesis depends on a number of aspects: correct metabolite assignment using proper analytical assays, use of appropriate statistical methods, validation of results and ensuring confounding factors do not influence the relation between **X** and **Y**. The latter is an important aspect in epidemiology; however, it is not common practice in metabonomics despite the fact that in molecular epidemiology studies there often is a wealth of meta-data available that are often gathered at the same time as sample collection.[2]

In the gene expression and proteomic literature, there are methods that aim to separate known covariations (batch/population structure) in data,[13−15] correct for unknown confounders,[16−18] or adjust for both types.[19,20] A difference between the analysis of genomic/proteomic and metabonomic data is that for the former data analysis is often done univariately, whereas for metabonomics data are often analyzed using multivariate methods as these are able to capture metabolite-metabolite interactions part of potentially perturbed pathways. Therefore, confounder correction methods used in genomics/proteomics are not necessarily suitable for metabonomics. Some metabonomic studies[6,11,21,22] have aimed to adjust for confounders using multiple linear regression (MLR) by regressing **Y** on a matrix of the covariates (**C**) and a single variable i from the data set ($\mathbf{X}_i$). This adjusts the contribution of $\mathbf{X}_i$ on **Y** because **C** is also included in the same model; however, this approach is univariate for **X** and thus does not capture metabolite-metabolite interactions. Other studies[23−25] regress each confounder on **X** and then compare the "significant" metabolites with those found from regressing **Y** on **X**. Last, there is orthogonal projections to latent structures[26] (OPLS), which is widely used in metabonomics and removes variation orthogonal to **Y** from **X** before calculating the regression coefficients. OPLS has been claimed to possibly correct for confounders;[27−29] however, confounders are not necessarily orthogonal to **Y**; thus, this method will not correct for all confounders. Nevertheless, the (O)PLS method is popular in metabonomics because it can deal with large p, a high degree of collinearity and only requires one parameter to be estimated.[30] Different methods exist that deal with confounder adjustment in kernel matrices[31,32] and those that take an unsupervised functional data analysis approach.[33] However, while inclusion in kernel matrices can be beneficial in terms of classification by including potential nonlinearity in the kernel transformation, it comes at a price as the variable importance is lost for biological interpretation.

The naïve approach of concatenating **C** and **X** and regressing **Y** on this concatenated matrix using multivariate methods will, most likely, not properly adjust for confounders as the variables in **X** will dominate the model. Regularization approaches such as the lasso[34] and elastic net[35] can be used to circumvent this problem in forcing the majority of regression coefficients (β) to 0. However, the lasso model can contain at most n nonzero βs and is known to perform poorly when data sets consist of correlated variables (as with metabonomics data). Therefore, the inclusion of a variable from **X** in the lasso model does not indicate it is not associated with **C**. The elastic net regularizes βs while simultaneously including groups of correlated variables; however, attributing which variables are associated with confounders is challenging and, in addition, it is a computationally heavy approach.

We propose here a new data analysis framework and algorithm to correct for known confounders using PLS (or orthogonal signal corrected PLS), called covariate-adjusted PLS (CA-PLS); however, in theory any multivariate regression method can be used instead of PLS. Our method mimics how MLR works in the univariate case, where **C** is used to counterweight **X** and not **Y**,[36] and still provides a level of variable importance.

MATERIALS AND METHODS

INTERMAP

The INTERMAP study investigates dietary and other factors associated with blood pressure[37] (BP), the major modifiable risk factor underlying the worldwide epidemic of cardiovascular diseases[38] (CVDs). INTERMAP surveyed a total of 4680 men and women aged 40−59 from 17 population samples in four countries (People's Republic of China (PRC), Japan, United Kingdom, and United States) at two time-points ("visits"). In this study, data from the three Chinese population samples were used to study the effect of potential confounders on metabolic profiles and compared to a previous study on these data done without any adjustment[39] (Yap et al.). These three (rural) populations are from two geographical locations, two from the north (Pinggu county, Beijing, and Yu county, Shanxi) and one from the south (Wuming county, Guangxi). Studies have shown that northern and southern Chinese are at different risks for CVD[40] and the metabolic profiles of these populations are different, with the two northern Chinese populations most similar to each other.[41] However, it is unclear whether other factors may be causing the differences in metabolic profiles instead of genetics and environment.

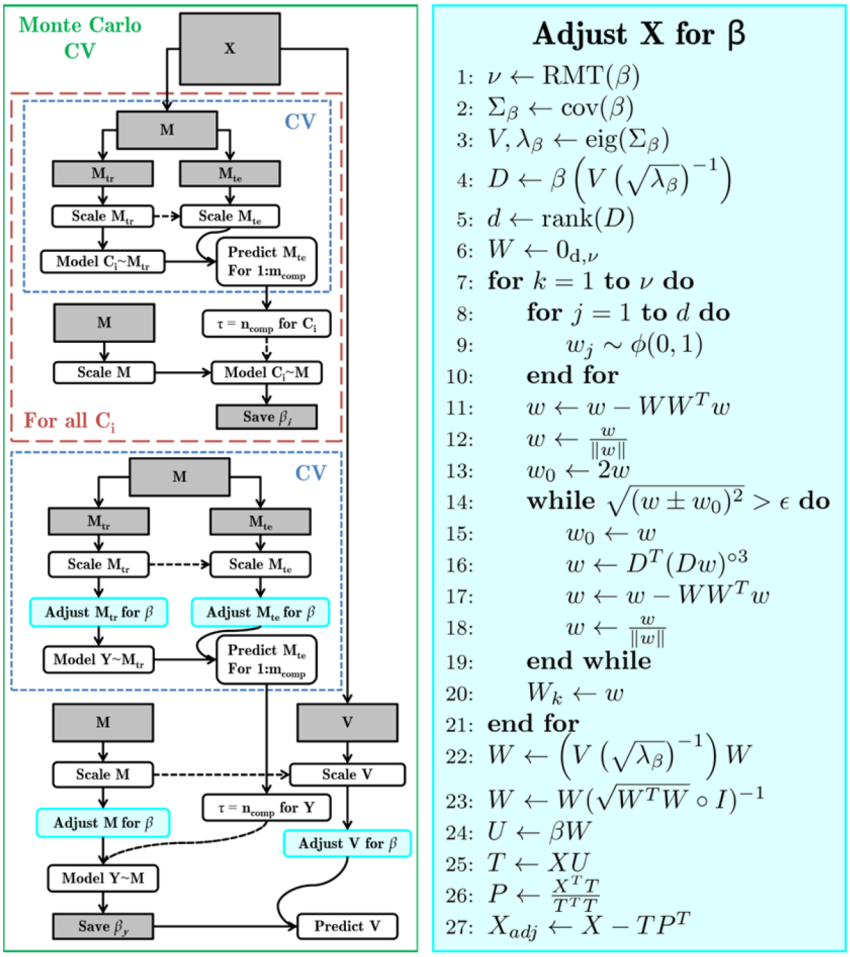
NMR Spectroscopy

Urine specimens were analyzed using a Bruker Avance-III spectrometer, operating at 600.29 MHz (^{1}H), equipped with a 5 mm, TCI, Z-gradient Cryo-probe. ^{1}H NMR spectra of urine were acquired using standard 1D pulse sequences with water presaturation during both the relaxation delay (RD = 2 s) and mixing time (t_m = 150 ms).[42] The 90° pulse length was 10 μs and total acquisition time 2.73 s. Per sample, 64 scans were collected into 32K data-points using a spectral width of 20 ppm. Free induction decays (FIDs) were multiplied by an exponential weighing function (corresponding to line broadening of 0.3 Hz) prior to Fourier-transformation.

FIDs were referenced to an internal standard (trimethylsilyl-[$^2H_4$]-propionate, TSP), baseline and phase-corrected using in-house software. Spectral regions containing water/urea (δ6.4 to 4.5), TSP (δ0.2 to −0.2), and noise (δ0.5 to 0.2, δ-0.2 to −4.5, δ15.5 to 9.5) were removed prior to median-fold change normalization.[43] Remaining variables were binned to 7100 variables using bin widths of 0.001 ppm to down-sample the total number of variables (for computation) while still retaining peak shapes. A separate study[44] showed good analytical reproducibility of the data set with 96% of split pairs correctly identified. Metabolic outliers were defined, and excluded, as participants whose principal component analysis scores, for either visit, mapped outside Hotelling's T^2 95% confidence interval (CI_{95}).[41]

Subset optimization by reference matching[45] (STORM) was used to identify metabolites using the correlation structure of ^{1}H NMR data. Localized clustering of small spectral regions was used for selecting appropriate reference spectra. Additionally, a Bruker compound library, internal databases, and extensive 2D NMR identification strategies[46] were used for identification of molecular species.

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595



Adjust X for β

$\nu \leftarrow \mathrm{RMT}(\beta)$
$\Sigma_\beta \leftarrow \mathrm{cov}(\beta)$
$V, \lambda_\beta \leftarrow \mathrm{eig}(\Sigma_\beta)$
$D \leftarrow \beta \left(V \left(\sqrt{\lambda_\beta} \right)^{-1} \right)$
$d \leftarrow \mathrm{rank}(D)$
$W \leftarrow 0_{\mathrm{d},\nu}$
for $k = 1$ **to** ν **do**
  for $j = 1$ **to** d **do**
    $w_j \sim \phi(0, 1)$
  end for
  $w \leftarrow w - WW^T w$
  $w \leftarrow \frac{w}{\|w\|}$
  $w_0 \leftarrow 2w$
  while $\sqrt{(w \pm w_0)^2} > \epsilon$ **do**
    $w_0 \leftarrow w$
    $w \leftarrow D^T (Dw)^{\circ 3}$
    $w \leftarrow w - WW^T w$
    $w \leftarrow \frac{w}{\|w\|}$
  end while
  $W_k \leftarrow w$
end for
$W \leftarrow \left(V \left(\sqrt{\lambda_\beta} \right)^{-1} \right) W$
$W \leftarrow W(\sqrt{W^T W} \circ I)^{-1}$
$U \leftarrow \beta W$
$T \leftarrow XU$
$P \leftarrow \frac{X^T T}{T^T T}$
$X_{adj} \leftarrow X - TP^T$

Figure 1. Data analysis framework and covariate-adjustment algorithm. Left panel shows different stages of the data analysis and shows how the introduction of bias is avoided by carefully splitting and scaling the data before modeling. Right panel (cyan box) outlines the covariate-adjustment algorithm that is used in the data analysis framework in the left panel in the cyan-colored boxes. The green outline indicates the entire MCCV procedure, the red dashed box the regression analysis performed for each covariate and blue dotted boxes indicate a CV loop. Here, β are regression coefficients, RMT stands for random matrix theory (see Supporting Information for algorithm) and $\circ$ denotes an element-wise operation. See Supporting Information for a glossary of mathematical operations used here.

CA-PLS Framework and Algorithm

We use a version of the SIMPLS algorithm to deal specifically with wide **X** matrices[47] (see Supporting Information). Same as for the original PLS algorithm, this algorithm can also be used to do orthogonal signal corrected[48] (OSC) PLS, for which **X** is replaced by the OSC matrix $\mathbf{X}_{osc}$.

The framework (Figure 1, left panel) is designed to minimize the effect of sampling/selection bias and avoid overfitting models. We perform the calculations in a Monte Carlo cross-validation[49] (MCCV) procedure; specifically, we perform 1000 iterations for the MCCV. We randomly partition the data in a model (**M**) and validation (**V**) set and use 1/7th of the data for **V** to mimic the partitioning of Yap et al. The important aspect of this framework is that **V** is completely set aside and thus not used in scaling, parameter estimation or modeling in any way to avoid biasing model prediction. The MCCV framework that is used here has previously been used in a repeated-measures design to show that dietary patterns could be predicted using analysis of a urine sample.[50]

The first step of the framework is to find the optimal parameter settings (τ) using cross-model-validation[51] (CMV) by partitioning **M** into multiple training and test sets for modeling each covariate. τ is intended to be specifically vague as the types and numbers of parameters are different for each regression method that this framework can be extended to. Hence, τ is unknown and is different for each regression method, for example, the number of components (as used here) for PLS or λ for ridge regression[52] and lasso[34] and λ and α for elastic net.[35] If the optimal number of components is k, then for PLS k components are calculated whereas for OSC-PLS a model is calculated with one predictive and k−1 orthogonal components.[53] For each partitioning, the training set is autoscaled (mean-centered, divided by standard deviation) to match Yap et al. The test set is scaled using the parameters from the training set to avoid introducing bias. Each covariate is also autoscaled to ensure regression coefficients (β) for the covariates have the same scale. τ is found by evaluating the cross-validated error-of-prediction (Q^2) (eq 1), where $\hat{\mathbf{Y}}$ is the predicted response and $\bar{\mathbf{Y}}$ the mean response for the test set.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \quad (1)$$

If, for binary responses, the predicted value is bigger (in absolute sense) than its true value, we do not penalized this

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

"error" and replace $\hat{\mathbf{Y}}$ with its true value[54] (eq 2), here "sgn" is the sign operator. The goodness-of-fit (R^2) is calculated identically, except using training instead of test data:

$$\hat{Y}_i = Y_i \,\forall\; \mathrm{sgn}(\hat{Y}_i - Y_i) = \mathrm{sgn}(Y_i) \tag{2}$$

When τ is found, β is calculated using **M**. This process is repeated for each covariate. However, not all covariates may be accurately modeled using the data. To avoid adjusting for covariates that cannot be modeled properly, we place some constraints for which covariates are adjusted. We use a lower threshold (0.10) for Q^2 and another threshold (0.25) for the robustness of cross-validation (RCV) (eq 3) to avoid adjusting for covariates that do not generalize well:

$$RCV = \frac{Q^2}{R^2} \tag{3}$$

We propose to use the RCV because often determining whether a model generalizes well is done by judging, highly subjective, whether the Q^2 is positive and "high enough". RCV can be seen as measure of how the model generalizes with respect to the optimal fit. A permutation scheme can be used to find a suitable lower bound for RCV (analogous to the permutation test for Q^2-values.[55] Here we simply use a hard threshold for RCV. However, care must be taken to deal with negative Q^2-values for calculating the RCV, for instance by setting a lower limit of 0 for the Q^2 for calculating the RCV. Low (or negative) Q^2-values indicate poor model predictive ability and in cases where R^2 is high but Q^2 is low this means the model is overfitting the data. This indicates that the correct τ has not been found.

The next step of the framework and algorithm is to adjust the data for covariates (cyan-colored boxes in the left panel and pseudocode in the right panel of Figure 1) that pass the thresholds and model **Y** on the adjusted data matrix. **M** is again split into training and test sets, these are scaled as before and then the data is adjusted for covariates using the algorithm shown in the right panel of Figure 1. As covariates may be correlated, the resulting βs will also be correlated and are thus nonorthogonal. βs of covariates must therefore be adjusted for their autocorrelation ($\mathbf{r}_B$) prior to adjusting the data. In this procedure, a Jacobian matrix is numerically computed from $\mathbf{r}_B$ using random matrix theory[56] (RMT, see Supporting Information). Then the number of uncorrelated components (υ, line 1) is defined as the number of eigenvalues from the decomposition of r_B that are larger than the largest eigenvalue from the Jacobian matrix (see Supporting Information). Once υ is known, β is decorrelated and a new set of decorrelated βs (**D**) are obtained by decomposing the covariance of β (Σ_B) and retaining eigenvectors/eigenvalues that explain at least 95% of the total variance (subject to there being at least υ retained components) (lines 2−4). Next, υ components that span the space of **D** (where $\upsilon \leq \mathrm{rank}(\mathbf{D}) \leq c$) are sought in an iterative process and saved in columns of **W** (lines 7−21). **W** is then normalized and transformed to span the space of β resulting in uncorrelated regression coefficients (**U**) (lines 22−24). **U** is used to adjust **X** for the βs from **C** (lines 25−27). Note that the same **U** is used for all adjustments, so need only be calculated once (only lines 25−27 need to be repeated for each new data matrix to replace "**X**": **M**, $\mathbf{M}_{tr}$, $\mathbf{M}_{te}$, **V**). The optimal number of components for Y ("τ") is found using the adjusted training **M** matrices. Regression coefficients for **Y** (β_Y) are then found by adjusting **M** using **U**. Using β_Y (and **U**) the validation set **V** that has not been used at any stage, can now be predicted free from bias.

This entire process is repeated in the MCCV. To find the variable contributions across all models, we recalculate each β_Y 25 times by bootstrapping[57] **Y** and **M**. Thus, after MCCV, two matrices are obtained with βs, those of each model (n = 1000) and those of the bootstrap models (n = 25 000). The mean of model βs and variance of the bootstrap βs are calculated and from these t-scores, and subsequently *P*-values, are calculated for each variable in the multivariate model.[58] *P*-values are corrected for multiple testing using the False Discovery Rate[59] (*Q*-value). We allow 5% false discoveries. Only variables whose βs are the same sign and $Q < 0.05$ are considered to be consistently and-similarly contributing ("significant") in the MCCV. Precompiled code to run CA-PLS is available from the first author's Web site.

Variable Significance

Variable significance is shown in plots as $S_i = -\frac{\beta_i}{|\beta_i|}\log_{10} q_i$, defining S_i as "significance", β_i the regression coefficient, and q_i the *Q*-value for variable *i*. A variable has to be "significant" in both visits and have the same sign.

The Supporting Information contains information about how we simulated data sets (Supplementary Figures 1 and 2) to show the difference between consistently and similarly contributing variables between (OSC)PLS and CA-(O)PLS. Calculations were performed in MATLAB (R2014a, The Mathworks, USA).

RESULTS AND DISCUSSION

Method Comparison Using Simulated Data

We compare CA-(O)PLS with PLS and OSC-PLS for the simulated data sets (see Supporting Information) with confounders introduced into the data sets. Here, CA-(O)PLS adjusts either for confounder 1 (nonorthogonal to **Y**) or confounder 2 (almost orthogonal to **Y**). Supplementary Table 1 shows how methods performed in finding consistently contributing variables associated with the case/control status for the data sets with an effect size of 1 (for inducing differences between groups). It shows the percentage of false negative (type-II error) and false positive (type-I error) findings. All methods find between 1−3% type-II errors; however, the differences between them are observed for the amount of type-I errors. CA-(O)PLS (correcting for confounder 1) finds a lower number, 0.31% (CI_{95} [0.26, 0.36]), of type-I errors compared to the other methods (1.35−3.09%). As expected, OSC-PLS and CA-(O)PLS (for confounder 2) perform similarly; however, PLS has less, 1.35% (CI_{95} [1.26, 1.44]) type-I errors than OSC-PLS/CA-(O)PLS (confounder 2) (3.09%/2.91%), which is surprising. However, Supplementary Table 1 shows the OSC-PLS model finds more variables significant that correlate to case/control status, whereas PLS finds more variables uncorrelated to case/control status significant. Similar, but less pronounced, results were found for a data set with less overlap (effect size of 1.645) between groups (Supplementary Table 2).

Unadjusted Model

It has been shown that the prevalence of CVD in general, and hypertension (HBP) specifically, is higher in the north of PRC[40] and that northern and southern Han Chinese are genetically different.[60] We find significant differences (Supple-

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

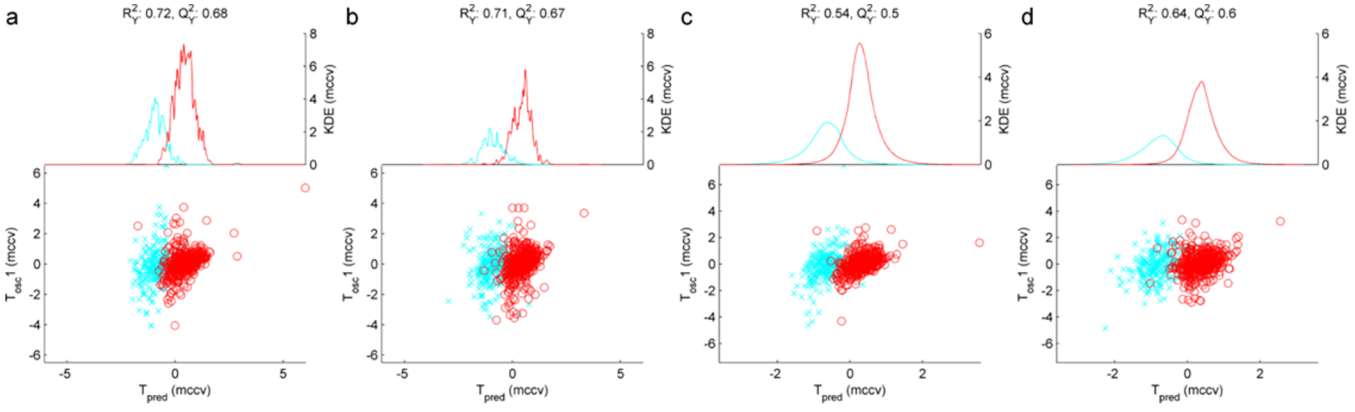


Figure 2. Score plots of the MCCV models of predictive and first orthogonal components with kernel density estimate (KDE), R^2 and Q^2 shown for the predictive axis. North Chinese individuals (Beijing and Shanxi) are shown as red circles and south Chinese (Guangxi) as cyan crosses. (a) Unadjusted model of urine collection 1. (b) Unadjusted model of urine collection 2. (c) Covariate-adjusted model of urine collection 1. (d) Covariate-adjusted model of urine collection 2. Age, gender, BMI, (on medication for) HBP, smoking status, physical activity, Na/K ratio, and total intake of fats were adjusted for in the CA-(O)PLSDA models (c and d).

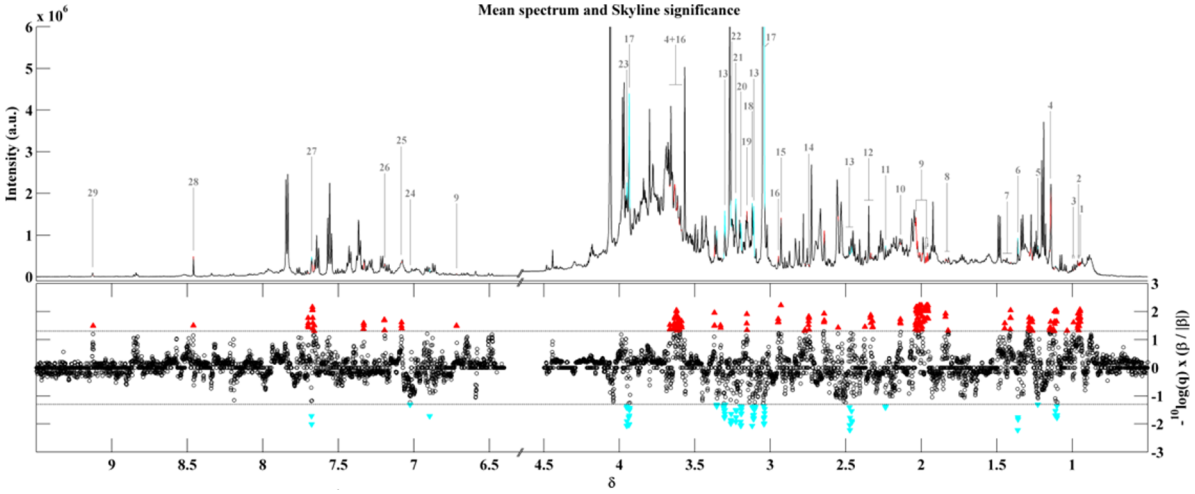


Figure 3. Top shows the average ^{1}H NMR spectrum from the first visit. The bottom panel shows the variable contribution across MCCV models. Models were adjusted for age, gender, HBP/medication, BMI, physical activity, smoking status, Na/K-ratio, and total fat intake. Labels: 1, 2-oxoisocaproate; 2, leucine; 3, valine; 4, unknown (1.15(s), 3.49(d), 3.61(d), 3.67(m), 3.83(m)); 5, ethylglucuronide; 6, 2-hydroxyisobutyrate; 7, unknown (1.42(d), 1.46(d), 1.51(d)); 8, unknown (1.82(m), 3.52(s)); 9, *N*-acetyl-*S*-(1Z)-propenyl-cysteine-sulfoxide; 10, glutamine; 11, acetone; 12, unknown (2.32(d), 2.34(d), 2.38(d), 2.40(d), 3.52(m)); 13, prolinebetaine; 14, sarcosine; 15, dimethylglycine; 16, unknown (1.84(m), 2.78(m), 2.95(s), 3.36(m), 3.59(m), 3.62(m)); 17, creatine; 18, *N6,N6,N6*-trimethyllysine; 19, dimethylsulfone; 20, *O*-acetylcarnitine; 21, carnitine; 22, taurine; 23, 4-hydroxyhippurate; 24, 1-methylhistidine; 25, histidine; 26, tyrosine; 27, pseudouridine; 28, formate; 29, *N*-methylnicotinic acid. Supplementary Figure 5 shows the results for the unadjusted model.

mentary Table 3) between our Chinese populations for a number of dietary, lifestyle, metabolic, and population risk factors for CVD for which we aim to adjust. However, we first compare the results from Yap et al., who used *t* tests to test significance of each ^{1}H NMR variable, with results obtained using our framework and assessment of variable contributions. Figure 2a and b show the resulting MCCV score plots of the mean predictions. For the model of visit 1 ("model 1"), the R^2 is 0.72 with the Q^2 being 0.68, resulting in an RCV of 0.95. For the model of visit 2 ("model 2") the R^2, Q^2, and RCV are 0.71, 0.67, and 0.95, respectively. This highlights the overall quality of the models. To indicate the spread of the predictions we include the kernel density estimate (KDE) of predicted values for each class. To obtain the KDE, we calculate for each sample the mean and standard deviation of its prediction when it was part of a test set in MCCV. Summing the distribution of each sample per class then yields the KDE as shown. The local sharp peaks of the KDE indicate large between-person variability.

To analyze sensitivity of models, we used each model to predict the data set from the other visit, resulting in goodness-of-external-predictions of 0.61 for model 1 (predicting visit 2) and 0.64 for model 2 (predicting visit 1) (Supplementary Figure 3). While both data sets do consist of the same individuals, the similarity between spectra from same individuals across visits is not high, indicated by an R_v-coefficient[61] of 0.31, where R_v = 1, indicates perfect similarity and 0 indicates dissimilarity. This is another reason why we use both data sets in determining consistently contributing variables to avoid capturing visit-specific variability.

Adjusted Model

Next, we picked a number of significantly different or important factors from Supplementary Table 3 (age, gender, body mass

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

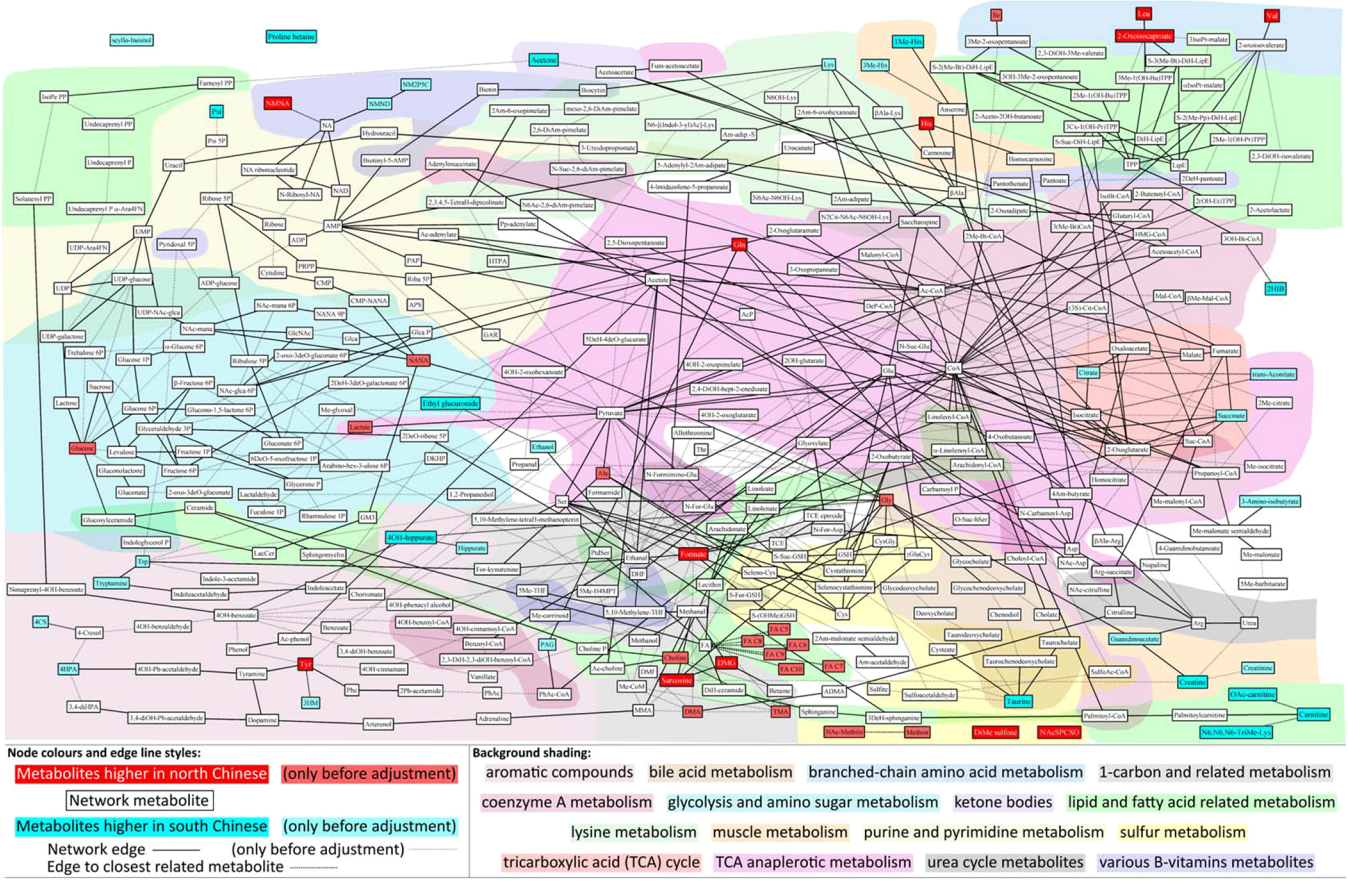


Figure 4. Perturbations to a living system often instigate changes to multiple pathways simultaneously; we show here a condensed multicompartmental metabolic reaction network of the homeostatic urinary signature of differences between north and south Chinese individuals for the human supra-organism, created using MetaboNetworks. A link is shown between two metabolites if the reaction is listed in KEGG and can occur in *Homo sapiens* (solid lines) or the most abundant endosymbionts (dotted lines). Metabolites not connected in the network, and those not listed in KEGG, were connected to the closest related metabolite in the network, indicated by a dashed line. The background shading illustrates different types of metabolism based on the closest affinity with some overlap between groups. A table with full names for the abbreviated metabolite names can be found in Supplementary Table 5.

index (BMI, kg × m^{-2}), (on medication for) HBP, smoking status, physical activity, Na/K-ratio, and total intake of fats) to adjust for using our CA-(O)PLS algorithm. The choice between HBP/medication status and individual measurements of systolic/diastolic BP was made as both measurements of BP are lowered by medication while there still is an underlying condition. The CA-(O)PLS algorithm determines, as described, which covariates are modeled accurately enough to be included in the adjustment. After the adjustment procedure, the geographical location was modeled as binary outcome variable and the resulting score plots (Figure 2c,d) of the MCCV predictions remain to show good separation; however, they are lower compared to the unadjusted model, with Q^2 values of 0.50 and 0.60 for urine collection 1 and 2, respectively. The resulting RCVs are 0.92 and 0.93, indicating the validation procedure is robust, again also demonstrated by predicting the other visits, with goodness-of-external-predictions of 0.46 for model 1 (predicting visit 2) and 0.57 for model 2 (predicting visit 1) (Supplementary Figure 4). Interestingly, here the KDEs are smooth distributions of the two groups, caused by the removal of specific variation in the data related to covariates/potential confounders. While it may seem counterintuitive to obtain models with a (slightly) lower predictive ability after correcting for confounders, it is a logical consequence of the covariates correlating with the outcome. However, the drawbacks of models with a lower predictive value (due to correlation between the outcome variable and covariates) are more than made up for by improved interpretability as the important variables relate to the part of the data that is not affected by covariates and only to the outcome. Figure 3 shows metabolites that consistently contribute to models. Unidentified metabolites are only included if their STORM[45] pseudospectrum showed clear interpretable patterns (Supplementary Figure 6).

Discriminatory Metabolites

We compare in Supplementary Table 4 the metabolites reported previously[39] and those found using the unadjusted and confounder-adjusted procedures. Our unadjusted procedure finds the same metabolites previously reported, plus a number of new associations. A large number of these metabolites are no longer significant after the confounder-adjustment and thus are likely related to one or more of the covariates.

In every iteration of the MCCV, the covariates are remodeled and only included if they were sufficiently predictive. In theory this depends on sampling of training/test sets; however, in practice we find there is a high consistency in which covariates could be accurately modeled. Gender and smoking status were modeled accurately for all models and Na/K-ratio in 94.4%

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

(visit 1) and 100% (visit 2) of models. Age, fat intake, and HBP/medication could not be modeled accurately, and BMI and physical activity were only included in 4−11% of models. To give a rough estimate of the association of metabolites no longer significant (after covariate-adjustment) and the covariates themselves, we calculated a correlation network (Supplementary Figure 7). The correlations were adjusted for multiple testing using a Bonferroni correction of $P < 1.9 \times 10^{-04}$ for both visits.

Dietary Na/K-ratio has a large number of correlations, which appear to have the inverse sign of the correlations of physical activity with metabolite levels. A logical reason is that Na/K-ratio and physical activity are inversely associated themselves. In general, south Chinese individuals are physically more active and consume more potassium and less sodium, and individuals who are more active have lower Na/K-ratios (Supplementary Table 3, Supplementary Figure 8).

Metabolic Reaction Network Analysis

Using metabolites differentially expressed between the Chinese populations, we constructed a condensed multicompartmental metabolic reaction network using MetaboNetworks.[62] It calculates the shortest paths (number of reactions) between metabolites, only considering reactions that can occur in the *Homo sapiens* supra-organism. We found a number of gut microbial cometabolites, and thus included reactions that can occur in species from the phyla firmicutes, bacteroidetes, *alpha*-proteobacteria, *beta*-proteobacteria, *gamma*-proteobacteria, *delta*-proteobacteria, and actinobacteria. These phyla make up 99% of phylotypes found in the human gut.[63] Figure 4 shows the resulting metabolic reaction network for the urinary metabolic differences in Chinese populations. Reactions between metabolites are indicated by a solid line for those that are spontaneous or due to human enzymes and by dotted lines for reactions that occur only in gut microbiota. The background colors indicate different types of conventional metabolic pathways. Figure 4 highlights the interconnectivity between many of the metabolites found to be differentially expressed between northern and southern Chinese.

Branched-chain amino acids (BCAAs) and derivatives (leucine, valine, 2-oxoisocaproate) are found higher in the north compared to the south, which may indicate a difference in energy metabolism, potentially also reflected by the tricarboxylic acid (TCA) cycle intermediates citrate and succinate, and isoleucine found higher in the north without adjustment. Aside from BCAAs, amino acids histidine, tyrosine, and glutamine are also found higher in the north. Glutamine is involved in TCA anaplerotic metabolism and histidine in muscle metabolism. While tyrosine partly links into the TCA anaplerotic metabolism via a microbial conversion to pyruvate, it is an aromatic amino acid. Other aromatic compounds were found to be higher in the south before covariate adjustment (4-cresylsulfate, phenylacetylglutamine, hippurate, 4-hydroxyphenylacetate, 3-hydroxymandelate). The fact that 4-cresylsulfate, phenylacetylglutamine, and hippurate are no longer significant after adjustment is related to gender and body weight differences[11,64] (Supplementary Figure 7). Another aromatic compound found higher in the south is 4-hydroxyhippurate, which has been linked to citrus fruit intake[65] and healthy eating in general.[50] We also find the most common biomarker of citrus fruit intake,[65] prolinebetaine, in higher concentrations in southern Chinese individuals. Aside from being a biomarker for citrus fruit intake, prolinebetaine is also considered an osmoprotectant, as are carnitine and trimethyllysine, which are also found in higher concentrations in the south. Also, the intracellular concentration of taurine (found higher in the south) increases when the extra-cellular fluid is hypertonic,[66] this may indicate that southern Chinese individuals (lower Na/K-ratios) are under less osmotic stress, reflected by the excretion of these metabolites. It should be noted however that the excretion of carnitine and *O*-acetylcarnitine are also linked to meat intake[67] and that taurine is a major metabolite in the conjugation of bile acids, which may be related to the higher intake of fats in the southern Chinese, indicating differences in lipid/fatty acid metabolism between the regions as well as being indicative of cataplerosis.[68] Acetone is a byproduct of breakdown of lipids/fatty acids for energy release. It has been noted that incomplete fatty acid oxidation and fat excess in skeletal muscle tissue can perturb energy anaplerosis and cause diabetes.[69]

Aside from taurine, two other sulfur-containing metabolites are found, dimethylsulfone and *N*-acetyl-*S*-(1Z)-propenyl-cysteine-sulfoxide. Both are biomarkers of onion consumption[46,70] with the later validated in a controlled clinical trial.[46,50] Another metabolite possibly linked to dietary intake is ethylglucuronide, which is a long-term marker of alcohol consumption and component of rice wine.[71] *N*-methylnicotinic acid is a metabolite linked to many different sources, among which coffee consumption[72] and peas,[46] and is a major metabolite of niacin (vitamin B3). In the metabolic network (Figure 4), there are multiple domains related to B-vitamins, such as thiamin (B1), panthothenate (B5), pyridoxal (B6), biotin (B7), folate (B9), and cobalamin (B12). These play roles in many processes, including lipid metabolism.

Closely linked to the lipidic domain through choline metabolism are formate, dimethylglycine, and sarcosine. These were all found higher in the north and are part of 1-carbon metabolism. Sarcosine is also linked to creatine and urea formation via a microbial link. Creatine is, among many other processes, related to muscle metabolism and, like 1-methylhistidine, found higher in the south. This reflects differences in muscle metabolism between populations, possibly a long-term effect from physical activity (Supplementary Table 3). We also find 2-hydroxyisobutyrate higher in the south which is a product of *n*-butyrate producing bacteria,[5] the same bacterium (*F. prausnitzii*) is associated with higher levels of *β*-aminoisobutyrate, taurine, and dimethylamine, and lower levels of lactate and glycine. With the exception of dimethylamine, these metabolites are similarly expressed in the southern Chinese in the unadjusted model, indicating a possible difference in *n*-butyrate producing bacteria. Last, pseudouridine, a marker of tRNA turnover, is higher in the south.

CONCLUSION

Adjusting data for confounders may lead to a loss of predictive power; however, the number of spurious findings is reduced (type-I errors), thereby greatly improving model interpretability. The CA-(O)PLS framework leads to finding more robust sets of biomarkers and more accurate predictions by (1) reducing sampling bias, independent scaling and MCCV, (2) optimizing parameter settings using CMV, (3) removing layers of confounding information from data, and (4) evaluating variable importance across multiple models instead of calculating a single[53] model.

We recommend testing whether each covariate can be modeled accurately before including them. However, if

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

covariates cannot be accurately modeled, they are not adjusted for and therefore do not influence models. If this is the case for all covariates, CA-(O)PLS defaults to (OSC−)PLS. While many factors can be adjusted for simultaneously, this will ultimately lead to loss of power, regardless of analysis method (univariate/multivariate). However, including highly (anti)-correlated covariates does not pose a problem for CA-(O)PLS as it finds an orthonormal set of factors from the covariate models to adjust data sets with. The benefit of CA-PLS is that it directly adjusts the data which allows *a posteriori* interpretation of metabolic signatures associated with covariates, opposed to other methods that work on a kernel matrix[32] in which this is lost.

We showed that confounders that differ between northern and southern Chinese individuals influence metabolite associations. We find that some previously reported associations are primarily associated with potential confounders. The metabolites that we found to be consistently contributing to the models highlight important underlying processes, most noticeably lipid, energy and gut-microbial metabolism, potentially of interest in determining what drives the differences in prevalence of CVDs in the Chinese population.

The multivariate confounder-adjustment framework we describe is easily translatable to other multivariate regression techniques and the potential benefit is not limited to metabolic phenotyping, but in theory it is applicable in any field, for example, other "omics" technologies, drug discovery, ecology, and potentially finance, where changes in collinear multivariate data can be attributed to confounders.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00879.

Glossary of mathematical operations; PLS algorithm; random matrix theory algorithm; generation of simulated data sets; effect size and overlap of simulated data sets; PCA pairs plot; score plots of external validation sets; mean spectrum and consistently contributing metabolites; statistical identification of metabolites; correlation network; dietary sodium/potassium ratio and physical activity; results of consistently contributing variables; descriptive data of Chinese INTERMAP population; consistently contributing metabolites; abbreviations and full names of metabolites (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: jmp111@ic.ac.uk.
*E-mail: j.nicholson@imperial.ac.uk. Fax: +44 (0)20 7594 3226.

ORCID

Joram M. Posma: 0000-0002-4971-9003
Paul Elliott: 0000-0002-7511-5684
Elaine Holmes: 0000-0002-0556-8389

Author Contributions

J.M.P. wrote the manuscript, developed algorithms, performed data analysis, and interpreted results. I.G.P. performed and interpreted spectroscopic experiments. T.M.D.E. provided statistical advice. J.C.L. assisted with interpretation of spectroscopic experiments. J.S. and P.E. supervised the INTERMAP study. J.K.N., E.H., and P.E. led the INTERMAP-metabonomics study. J.K.N. supervised the study. J.K.N. and E.H. provided support with biological interpretations. All authors read and approved the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Elaine Maibaum for carrying out the ^{1}H NMR spectroscopy of the INTERMAP urine samples and our colleagues listed in refs 37 and 39 for data collection. J.M.P. is supported by a Rutherford Fund Fellowship at Health Data Research (HDR) UK (MR/S004033/1). The INTERMAP study is supported by the US National Heart, Lung, and Blood Institute (R01-HL50490 and R01-HL84228), the Ministry for Education, Science, Sports, and Culture (Japan, grant [A] 090357003), a project grant from the West Midlands National Health Service Research and Development Programme, and grant R2019EPH from the Northern Ireland Chest, Heart and Stroke Association (UK), and infrastructure support was provided by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC) and the UK MEDical BIOinformatics partnership (MR/L01632X/1). P.E. acknowledges support from the NIHR BRC at Imperial College Healthcare NHS Trust and Imperial College London. P.E. is an NIHR Senior Investigator.

■ 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* (3), 169−74.

(2) Nicholson, J. K. Global systems biology, personalized medicine and molecular epidemiology. *Mol. Syst. Biol.* **2006**, *2* (1), 52.

(3) Go, V. L.; Nguyen, C. T.; Harris, D. M.; Lee, W. N. Nutrient-gene interaction: metabolic genotype-phenotype relationship. *J. Nutr.* **2005**, *135* (12 Suppl), 3016S−3020S.

(4) Gill, S. R.; Pop, M.; Deboy, R. T.; Eckburg, P. B.; Turnbaugh, P. J.; Samuel, B. S.; Gordon, J. I.; Relman, D. A.; Fraser-Liggett, C. M.; Nelson, K. E. Metagenomic analysis of the human distal gut microbiome. *Science* **2006**, *312* (5778), 1355−9.

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

(6) Ellis, J. K.; Athersuch, T. J.; Thomas, L. D.; Teichert, F.; Perez-Trujillo, M.; Svendsen, C.; Spurgeon, D. J.; Singh, R.; Jarup, L.; Bundy, J. G.; Keun, H. C. Metabolic profiling detects early effects of environmental and lifestyle exposure to cadmium in a human population. *BMC Med.* **2012**, *10*, 61.

(7) Holmes, E.; Wilson, I. D.; Nicholson, J. K. Metabolic Phenotyping in Health and Disease. *Cell* **2008**, *134* (5), 714−717.

(8) 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* (11), 1181−1189.

(9) Fiehn, O. Metabolomics–the link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48* (1−2), 155−71.

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

(10) Nicholson, J. K.; Lindon, J. C. Systems biology: Metabonomics. *Nature* **2008**, *455* (7216), 1054−6.
(11) Elliott, P.; Posma, J. M.; Chan, Q.; Garcia-Perez, I.; Wijeyesekera, A.; Bictash, M.; TM, D. E.; Ueshima, H.; Zhao, L.; van Horn, L.; Daviglus, M.; Stamler, J.; Holmes, E.; Nicholson, J. K. Urinary metabolic signatures of human adiposity. *Sci. Transl. Med.* **2015**, *7* (285), 285ra62.
(12) Lenz, E. M.; Wilson, I. D. Analytical strategies in metabonomics. *J. Proteome Res.* **2007**, *6* (2), 443−58.
(13) Kerr, M. K.; Martin, M.; Churchill, G. A. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **2000**, *7* (6), 819−837.
(14) Harrington, P. D.; Vieira, N. E.; Espinoza, J.; Nien, J. K.; Romero, R.; Yergey, A. L. Analysis of variance-principal component analysis: A soft tool for proteomic discovery. *Anal. Chim. Acta* **2005**, *544* (1−2), 118−127.
(15) de Haan, J. R.; Wehrens, R.; Bauerschmidt, S.; Piek, E.; van Schaik, R. C.; Buydens, L. M. C. Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* **2007**, *23* (2), 184−190.
(16) Leek, J. T.; Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **2007**, *3* (9), 1724−1735.
(17) Teschendorff, A. E.; Zhuang, J.; Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **2011**, *27* (11), 1496−1505.
(18) Chakraborty, S.; Datta, S.; Datta, S. Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics* **2012**, *28* (6), 799−806.
(19) Listgarten, J.; Kadie, C.; Schadt, E. E.; Heckerman, D. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (38), 16465−70.
(20) Fusi, N.; Stegle, O.; Lawrence, N. D. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.* **2012**, *8* (1), e1002330.
(21) Lawton, K. A.; Berger, A.; Mitchell, M.; Milgram, K. E.; Evans, A. M.; Guo, L. N.; Hanson, R. W.; Kalhan, S. C.; Ryals, J. A.; Milburn, M. V. Analysis of the adult human plasma metabolome. *Pharmacogenomics* **2008**, *9* (4), 383−397.
(22) Suhre, K.; Meisinger, C.; Doring, A.; Altmaier, E.; Belcredi, P.; Gieger, C.; Chang, D.; Milburn, M. V.; Gall, W. E.; Weinberger, K. M.; Mewes, H. W.; de Angelis, M. H.; Wichmann, H. E.; Kronenberg, F.; Adamski, J.; Illig, T. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS One* **2010**, *5* (11), e13953.
(23) Slupsky, C. M.; Rankin, K. N.; Wagner, J.; Fu, H.; Chang, D.; Weljie, A. M.; Saude, E. J.; Lix, B.; Adamko, D. J.; Shah, S.; Greiner, R.; Sykes, B. D.; Marrie, T. J. Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. *Anal. Chem.* **2007**, *79* (18), 6995−7004.
(24) Rocha, C. M.; Carrola, J.; Barros, A. S.; Gil, A. M.; Goodfellow, B. J.; Carreira, I. M.; Bernardo, J.; Gomes, A.; Sousa, V.; Carvalho, L.; Duarte, I. F. Metabolic Signatures of Lung Cancer in Biofluids: NMR-Based Metabonomics of Blood Plasma. *J. Proteome Res.* **2011**, *10* (9), 4314−4324.
(25) Garcia-Perez, I.; Villasenor, A.; Wijeyesekera, A.; Posma, J. M.; Jiang, Z.; Stamler, J.; Aronson, P.; Unwin, R.; Barbas, C.; Elliott, P.; Nicholson, J.; Holmes, E. Urinary Metabolic Phenotyping the slc26a6 (Chloride-Oxalate Exchanger) Null Mouse Model. *J. Proteome Res.* **2012**, *11* (9), 4425−35.
(26) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16* (3), 119−128.
(27) Wang, Y. L.; Tang, H. R.; Nicholson, J. K.; Hylands, P. J.; Sampson, J.; Holmes, E. A metabonomic strategy for the detection of the metabolic effects of chamomile (Matricaria recutita L.) ingestion. *J. Agric. Food Chem.* **2005**, *53* (2), 191−196.
(28) Yap, I. K. S.; Clayton, T. A.; Tang, H.; Everett, J. R.; Hanton, G.; Provost, J. P.; Le Net, J. L.; Charuel, C.; Lindon, J. C.; Nicholson, J. K. An integrated metabonomic approach to describe temporal metabolic disregulation induced in the rat by the model hepatotoxin allyl formate. *J. Proteome Res.* **2006**, *5* (10), 2675−2684.
(29) De Iorio, M.; Ebbels, T. M. D.; Stephens, D. A. Statistical Techniques in Metabolic Profiling. In *Handbook of Statistical Genetics*; John Wiley & Sons, Ltd, 2008; pp 347−373.
(30) Trygg, J.; Holmes, E.; Lundstedt, T. Chemometrics in metabonomics. *J. Proteome Res.* **2007**, *6* (2), 469−79.
(31) Li, L.; Rakitsch, B.; Borgwardt, K. ccSVM: correcting Support Vector Machines for confounding factors in biological data classification. *Bioinformatics* **2011**, *27* (13), i342−8.
(32) Moore, D. E.; Fluette, K. A.; Milne, H. J.; Shedlock, A. M.; Anderson, P. E. ccKOPLS: Confounder-Correcting Kernel-based Orthogonal Projections to Latent Structures. *Proceedings 2015 IEEE International Conference on Bioinformatics and Biomedicine*; IEEE, 2015; pp 897−903.
(33) Jiang, C. R.; Wang, J. L. Covariate Adjusted Functional Principal Components Analysis for Longitudinal Data. *Annals of Statistics* **2010**, *38* (2), 1194−1226.
(34) Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **1996**, *58* (1), 267−288.
(35) Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **2005**, *67* (2), 301−320.
(36) Rubin, D. B. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* **1974**, *66* (5), 688−701.
(37) Stamler, J.; Elliott, P.; Dennis, B.; Dyer, A. R.; Kesteloot, H.; Liu, K.; Ueshima, H.; Zhou, B. F. INTERMAP: background, aims, design, methods, and descriptive statistics (nondietary). *J. Hum. Hypertens.* **2003**, *17* (9), 591−608.
(38) Danaei, G.; Finucane, M. M.; Lin, J. K.; Singh, G. M.; Paciorek, C. J.; Cowan, M. J.; Farzadfar, F.; Stevens, G. A.; Lim, S. S.; Riley, L. M.; Ezzati, M. National, regional, and global trends in systolic blood pressure since 1980: systematic analysis of health examination surveys and epidemiological studies with 786 country-years and 5.4 million participants. *Lancet* **2011**, *377* (9765), 568−77.
(39) Yap, I. K. S.; Brown, I. J.; Chan, Q.; Wijeyesekera, A.; Garcia-Perez, I.; Bictash, M.; Loo, R. L.; Chadeau-Hyam, M.; Ebbeis, T.; De Iorio, M.; Maibaum, E.; Zhao, L. C.; Kesteloot, H.; Daviglus, M. L.; Stamler, J.; Nicholson, J. K.; Elliott, P.; Holmes, E. Metabolome-Wide Association Study Identifies Multiple Biomarkers that Discriminate North and South Chinese Populations at Differing Risks of Cardiovascular Disease INTERMAP Study. *J. Proteome Res.* **2010**, *9* (12), 6647−6654.
(40) Zhao, L. C.; Stamler, J.; Yan, L. J. L.; Zhou, B. F.; Wu, Y. F.; Liu, K.; Daviglus, M. L.; Dennis, B. H.; Elliott, P.; Ueshima, H.; Yang, J.; Zhu, L. G.; Guo, D. S. Blood pressure differences between northern and southern Chinese: Role of dietary factors the international study on macronutrients and blood pressure. *Hypertension* **2004**, *43* (6), 1332−1337.
(41) 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. C.; Nicholson, J. K.; Elliott, P. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature* **2008**, *453* (7193), 396−400.
(42) Holmes, E.; Loo, R. L.; Cloarec, O.; Coen, M.; Tang, H. R.; Maibaum, E.; Bruce, S.; Chan, Q.; Elliott, P.; Stamler, J.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K. Detection of urinary drug metabolite (Xenometabolome) signatures in molecular epidemiology studies via statistical total correlation (NMR) spectroscopy. *Anal. Chem.* **2007**, *79* (7), 2629−2640.
(43) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in H-1 NMR metabonomics. *Anal. Chem.* **2006**, *78* (13), 4281−4290.
(44) Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B. F.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.;

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595

Holmes, E. Assessment of analytical reproducibility of H-1 NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Anal. Chem.* **2006**, *78* (7), 2199−2208.
(45) Posma, J. M.; Garcia-Perez, I.; De Iorio, M.; Lindon, J. C.; Elliott, P.; Holmes, E.; Ebbels, T. M.; Nicholson, J. K. Subset Optimization by Reference Matching (STORM): An Optimized Statistical Approach for Recovery of Metabolic Biomarker Structural Information from (1)H NMR Spectra of Biofluids. *Anal. Chem.* **2012**, *84* (24), 10694−701.
(46) Posma, J. M.; Garcia-Perez, I.; Heaton, J. C.; Burdisso, P.; Mathers, J. C.; Draper, J.; Lewis, M.; Lindon, J. C.; Frost, G.; Holmes, E.; Nicholson, J. K. Integrated Analytical and Statistical Two-Dimensional Spectroscopy Strategy for Metabolite Identification: Application to Dietary Biomarkers. *Anal. Chem.* **2017**, *89* (6), 3300−3309.
(47) Daszykowski, M.; Serneels, S.; Kaczmarek, K.; Van Espen, P.; Croux, C.; Walczak, B. TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemom. Intell. Lab. Syst.* **2007**, *85* (2), 269−277.
(48) Eriksson, L.; Trygg, J.; Johansson, E.; Bro, R.; Wold, S. Orthogonal signal correction, wavelet analysis, and multivariate calibration of complicated process fluorescence data. *Anal. Chim. Acta* **2000**, *420* (2), 181−195.
(49) Xu, Q. S.; Liang, Y. Z. Monte Carlo cross validation. *Chemom. Intell. Lab. Syst.* **2001**, *56* (1), 1−11.
(50) Garcia-Perez, I.; Posma, J. M.; Gibson, R.; Chambers, E. S.; Hansen, T. H.; Vestergaard, H.; Hansen, T.; Beckmann, M.; Pedersen, O.; Elliott, P.; Stamler, J.; Nicholson, J. K.; Draper, J.; Mathers, J. C.; Holmes, E.; Frost, G. Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. *Lancet Diabetes Endocrinol.* **2017**, *5* (3), 184−195.
(51) Anderssen, E.; Dyrstad, K.; Westad, F.; Martens, H. Reducing over-optimizm in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst.* **2006**, *84* (1−2), 69−74.
(52) Hoerl, A. E.; Kennard, R. W. Ridge Regression - Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12* (1), 55−67.
(53) Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in H-1 NMR spectroscopic metabonomic studies. *Anal. Chem.* **2005**, *77* (2), 517−526.
(54) Hastie, T.; Tibshirani, R.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, 2009; p 745.
(55) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. Assessment of PLSDA cross validation. *Metabolomics* **2008**, *4* (1), 81−89.
(56) Edelman, A.; Rao, N. R. Random matrix theory. *Acta Numerica* **1999**, *14*, 233−297.
(57) Efron, B. Nonparametric Estimates of Standard Error - the Jackknife, the Bootstrap and Other Methods. *Biometrika* **1981**, *68* (3), 589−599.
(58) Chadeau-Hyam, M.; Ebbels, T. M. D.; Brown, I. J.; Chan, Q.; Stamler, J.; Huang, C. C.; Daviglus, M. L.; Ueshima, H.; Zhao, L. C.; Holmes, E.; Nicholson, J. K.; Elliott, P.; De Iorio, M. Metabolic Profiling and the Metabolome-Wide Association Study: Significance Level For Biomarker Identification. *J. Proteome Res.* **2010**, *9* (9), 4620−4627.
(59) Storey, J. D.; Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (16), 9440−9445.
(60) Chen, J.; Zheng, H.; Bei, J. X.; Sun, L.; Jia, W. H.; Li, T.; Zhang, F.; Seielstad, M.; Zeng, Y. X.; Zhang, X.; Liu, J. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. *Am. J. Hum. Genet.* **2009**, *85* (6), 775−85.
(61) Escoufier, Y. Le Traitement des Variables Vectorielles. *Biometrics* **1973**, *29* (4), 751−760.
(62) Posma, J. M.; Robinette, S. L.; Holmes, E.; Nicholson, J. K. MetaboNetworks, an interactive Matlab-based toolbox for creating, customizing and exploring sub-networks from KEGG. *Bioinformatics* **2014**, *30* (6), 893−5.
(63) Eckburg, P. B.; Bik, E. M.; Bernstein, C. N.; Purdom, E.; Dethlefsen, L.; Sargent, M.; Gill, S. R.; Nelson, K. E.; Relman, D. A. Diversity of the human intestinal microbial flora. *Science* **2005**, *308* (5728), 1635−8.
(64) Wijeyesekera, A.; Clarke, P. A.; Bictash, M.; Brown, I. J.; Fidock, M.; Ryckmans, T.; Yap, I. K. S.; Chan, Q.; Stamler, J.; Elliott, P.; Holmes, E.; Nicholson, J. K. Quantitative UPLC-MS/MS analysis of the gut microbial co-metabolites phenylacetylglutamine, 4-cresyl sulphate and hippurate in human urine: INTERMAP Study. *Anal. Methods* **2012**, *4* (1), 65−72.
(65) Heinzmann, S. S.; Brown, I. J.; Chan, Q.; Bictash, M.; Dumas, M. E.; Kochhar, S.; Stamler, J.; Holmes, E.; Elliott, P.; Nicholson, J. K. Metabolic profiling strategy for discovery of nutritional biomarkers: proline betaine as a marker of citrus consumption. *Am. J. Clin. Nutr.* **2010**, *92* (2), 436−443.
(66) Burg, M. B. Molecular basis of osmotic regulation. *Am. J. Physiol.* **1995**, *268* (6 Pt 2), F983−96.
(67) Stella, C.; Beckwith-Hall, B.; Cloarec, O.; Holmes, E.; Lindon, J. C.; Powell, J.; van der Ouderaa, F.; Bingham, S.; Cross, A. J.; Nicholson, J. K. Susceptibility of human metabolic phenotypes to dietary modulation. *J. Proteome Res.* **2006**, *5* (10), 2780−2788.
(68) Owen, O. E.; Kalhan, S. C.; Hanson, R. W. The key role of anaplerosis and cataplerosis for citric acid cycle function. *J. Biol. Chem.* **2002**, *277* (34), 30409−12.
(69) Koves, T. R.; Ussher, J. R.; Noland, R. C.; Slentz, D.; Mosedale, M.; Ilkayeva, O.; Bain, J.; Stevens, R.; Dyck, J. R.; Newgard, C. B.; Lopaschuk, G. D.; Muoio, D. M. Mitochondrial overload and incomplete fatty acid oxidation contribute to skeletal muscle insulin resistance. *Cell Metab.* **2008**, *7* (1), 45−56.
(70) Winning, H.; Roldan-Marin, E.; Dragsted, L. O.; Viereck, N.; Poulsen, M.; Sanchez-Moreno, C.; Cano, M. P.; Engelsen, S. B. An exploratory NMR nutri-metabonomic investigation reveals dimethyl sulfone as a dietary biomarker for onion intake. *Analyst* **2009**, *134* (11), 2344−2351.
(71) Teague, C.; Holmes, E.; Maibaum, E.; Nicholson, J.; Tang, H. R.; Chan, Q. N.; Elliott, P.; Wilson, I. Ethyl glucoside in human urine following dietary exposure: detection by H-1 NMR spectroscopy as a result of metabonomic screening of humans. *Analyst* **2004**, *129* (3), 259−264.
(72) Lang, R.; Wahl, A.; Stark, T.; Hofmann, T. Urinary N-methylpyridinium and trigonelline as candidate dietary biomarkers of coffee consumption. *Mol. Nutr. Food Res.* **2011**, *55* (11), 1613−1623.

DOI: 10.1021/acs.jproteome.7b00879
J. Proteome Res. 2018, 17, 1586−1595