Journal of proteome research

Article

Cite This: J. Proteome Res. 2019, 18, 3360−3368

pubs.acs.org/jpr

Automated Analysis of Large-Scale NMR Data Generates Metabolomic Signatures and Links Them to Candidate Metabolites

Bita Khalili,[†,‡,∥] Mattia Tomasoni,[†,‡,∥] Mirjam Mattei,[†,‡,∥] Roger Mallol Parera,[†,‡] Reyhan Sonmez,[†,‡] Daniel Krefl,[†,‡] Rico Rueedi,[†,‡,∥] and Sven Bergmann*[,†,‡,§,∥]

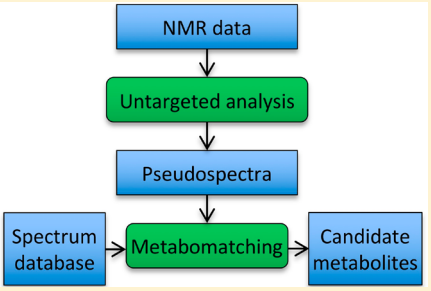
[†]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland

[‡]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[§]Department of Integrative Biomedical Sciences, University of Cape Town, Cape Town 7700, South Africa

Supporting Information

ABSTRACT: Identification of metabolites in large-scale ^{1}H NMR data from human biofluids remains challenging due to the complexity of the spectra and their sensitivity to pH and ionic concentrations. In this work, we tested the capacity of three analysis tools to extract metabolite signatures from 968 NMR profiles of human urine samples. Specifically, we studied sets of covarying features derived from principal component analysis (PCA), the iterative signature algorithm (ISA), and averaged correlation profiles (ACP), a new method we devised inspired by the STOCSY approach. We used our previously developed metabomatching method to match the sets generated by these algorithms to NMR spectra of individual metabolites available in public databases. On the basis of the number and quality of the matches, we concluded that ISA and ACP can robustly identify ten and nine metabolites, respectively, half of which were shared, while PCA did not produce any signatures with robust matches.



KEYWORDS: *1D NMR automated analysis, metabolite identification, modular analysis, STOCSY, ISA, pseudoquantification, NMR spectroscopy, untargeted metabolomics*

INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is a powerful technique for metabolomic profiling. NMR spectroscopy does not consume the sample and has high accuracy and reproducibility. Single proton NMR spectroscopy (^{1}H NMR) can be used to generate one-dimensional spectra of biofluids at high throughput and low cost, facilitating the generation of large sets of spectral data.

A first step in NMR spectral analysis is usually to identify the main metabolites giving rise to a given spectrum or set of spectra. This is nontrivial, since human biofluids typically contain a large number of individual metabolites and their corresponding peak positions may overlap and are often affected by the pH, ionic strength, and overall protein content of the fluid.

For small sets of samples, expert analysis is therefore still the most accurate means for metabolite identification, yet for large sets this approach is costly, time-consuming, and potentially less reproducible. As a result, various methods have been suggested to assist or fully automate metabolite identification.

In their landmark paper on statistical total correlation spectroscopy (STOCSY), Cloarec et al. showed that analyzing the correlation patterns between features across a sizable collection of ^{1}H NMR spectra has great potential for metabolite identification.[1] This is because features corresponding to the same molecule (or molecules whose concentrations covary) tend to be significantly correlated in large data sets. Analyzing data from 612 mouse urine samples, they observed that the correlation matrix exhibited correlated peaks of features characteristic of valeramide, glucose, hippurate, 2-oxoglutarate, 3-hydroxyphenylpropionate, citrate, as well as methylamine, dimethylamine, and trimethylamine.

Subsequent variations of STOCSY attempted to make clusters of NMR peaks to simplify the interpretation of the information stored in the correlation matrix from STOCSY analysis. Recoupled-STOCSY (R-STOCSY) employs a variable size bucketing method to reduce the dimensionality of NMR data and statistical recoupling of variables (SRV) to identify correlations between distant clusters.[2] Iterative-STOCSY (I-STOCSY) aims at separating the intermetabolite connections from intrametabolite connection by recursively applying STOCSY analysis first from a selected *driver peak* and then for all peaks correlating with the driver peak above a specific threshold.[3] Subset optimization by reference matching (STORM) selects subsets of ^{1}H NMR spectra that contain specific spectroscopic signatures of biomarkers differentiating between different human populations.[4]

Once metabolite identification has been achieved, the next challenge is to quantify metabolite concentrations. This process works robustly only for a relatively small set of metabolites, and requires expert refinement when using publicly available quantification tools such as BATMAN,[5] FOCUS,[6] BAYESIL,[7]

Received: May 6, 2019
Published: July 18, 2019

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

ASICS,[8] AQuA,[9] or rDolphin.[10] This is unsatisfactory in light of the fact that a sizable number of metabolites have been identified in human biofluids. For example, the latest version of the Human Metabolome Database (HMDB 4.0)[11] includes more than 1500 metabolites with ^{1}H NMR spectra, 179 of which have been identified in urine[12] and 67 in serum.[13,14]

There are several reasons why it remains difficult to perform fully automated quantification from large-scale NMR data for the vast majority of metabolites. First, human biofluids contain a large number of individual metabolites whose concentrations vary across several orders of magnitude. This makes it difficult to disentangle the contributions of metabolites with low concentrations, in particular when their NMR features are not unique. Second, the exact feature positions depend on the biofluid and may have been different when acquiring reference spectra. Third, while the number of reference spectra continues to grow, reference databases are certainly not yet exhaustive.

In two recent studies, we demonstrated that the limitations of targeted NMR metabolomics can be addressed by linking metabolites to external variables.[15,16] Specifically, in the context of genome-wide association studies (GWAS) applied to metabolomics (known as mGWAS) the aim is to associate metabolites with genotypic variants. We observed that the effect of a genetic variant on the concentration of a metabolite often translates into associations with all or many features of the metabolite NMR spectrum. The set of association scores with all measured features provides a *pseudospectrum* across the full range of ppm covered by the ^{1}H NMR spectra. The challenge is then to identify the metabolite underlying the most significant associations. To this end we developed the analysis tool *metabomatching*, which takes as input a pseudospectrum and a collection of reference spectra for individual metabolites found, for example, in HMDB.[11] Our previous work showed that metabomatching works well to prioritize the most likely metabolite candidates for pseudospectra derived from metabolome feature association with genotypes.[15−17]

In the present work, we tested the metabomatching methodology for identifying metabolites that vary across large collection of samples without the need for any external variables associated with this variation. We investigated three methods to identify covarying spectral features within large-scale NMR data: principal component analysis (PCA), the iterative signature algorithm (ISA), and averaged correlation profiles (ACP) inspired by the STOCSY approach. For each method, we devised a principled way for processing their output into pseudospectra. In addition, we extended metabomatching to process the respective outputs of the methods, and implemented a permutation-based robustness test to assess the quality of the matches. This allowed us to compare the matches across different methods, and assess the consistency or complementarity of the methods. We incorporated our analysis for unsupervised generation of metabolomic signatures from large-scale NMR data and integration with metabomatching (including further documentation) into the *metabomodules* software tool, which is publicly available at https://github.com/BergmannLab/metabomodules-docker.

METHODS

Preprocessing

For this study, we used 968 ^{1}H NMR spectra acquired from urine samples from the CoLaus cohort.[15] The samples are referenced to the TSP signal, phase-corrected, and baseline-corrected.

We used the FOCUS[6] tool to align and bin the spectra to a resolution of about 0.02 ppm, and a correspondingly large number of 687 peaks. To obtain this resolution, we set the downsampling frequency parameter *window.fs* to 1 (no downsampling), the sliding window length for spectral segmentation *window.length* to 0.03 ppm, the minimum peak width parameter *peak.DFL* to 0.02 ppm, and the peak sample frequency parameter *peak.pS* to 0.2, keeping the rest of the parameters at their default values.

To normalize the data, we log-transformed, standardized across features (thereby normalizing the concentration of each sample), then standardized across samples (thereby making intensities comparable).

Confounding

In order to allow for the identification of metabolites that may be hidden by confounding, we additionally generated a data set of residuals, created by regressing out the confounders from the feature metabolome. The main confounding factors of the NMR data that we investigated here are age, sex, serum creatinine, and urine creatinine.[15,16]

Metabomatching

Our original metabomatching method was designed to match the NMR spectra of individual metabolites recorded in a database with pseudospectra from the association between metabolome features and an external variable, typically a SNP genotype. For a metabolite m, metabomatching computes the sum

$$s(F_m) = \sum_{f \in F_m} z_f^2 \tag{1}$$

where F_m is the set of N_f features that fall within a neighborhood of any peak of m according to the database, and z_f denotes the significance value for feature f, and is given by $z_f = \hat{\beta}_f / \widehat{SE}_f$, where $\hat{\beta}_f$ and $\widehat{SE}_f$ refer to the point estimates of the effect size and its standard error, respectively. Under the null hypothesis of normally and independently distributed z_f, the sum s follows a χ^2-distribution with N_f degrees of freedom, and metabomatching defines the score for m as the negative logarithm of the nominal p-value for the sum. This score is then used to rank all tested metabolites as metabolites with more similar NMR spectra to a given pseudospectrum achieve higher scores.

In addition to pseudospectra from regression analysis, provided as columns headed by *beta*, *se*, and *p*, we extended metabomatching to accept pseudospectra produced by PCA, ACP, and ISA as columns headed by *pca*, *cr*, and *isa* respectively. For ACP pseudospectra, metabomatching translates a correlation c to a z-score with the Fisher transformation $z = \lambda \operatorname{arctanh}(c)$. For independent features, $\lambda = \sqrt{N-3}$ produces z-scores with unit standard deviation, where N is the number of samples across which the correlations are computed. However, since proximal features are usually not independent, metabomatching allows for a user-provided estimate for λ (obtained from the pairwise feature−feature correlation matrix), or re-estimates λ from the given correlations. For ISA and PCA pseudospectra, metabomatching standardizes the loadings or module scores.

We used the plus/minus mode of metabomatching since features are z-scored and have positive and negative signs. This allows for detecting metabolites corresponding either to the

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

negative or positive features (see metabomatching documentation at https://github.com/rrueedi/metabomatching for more details).

We also introduced a measure of the quality of a match, which allows to compare matches between different pseudospectra. This *adjusted score* is obtained by reshuffling the pseudospectrum, and defining a heuristic p-value by the number N_p of all N_r reshuffled pseudospectra that produce a metabomatching score (for any reference spectrum) higher than the metabomatching score of the input pseudospectrum with the highest ranked reference spectrum. This p-value is defined as $(N_p + 1)/(N_r + 1)$, and the adjusted score as $-\log(p)$. We used $N_r = 9999$, which sets the upper limit for the adjusted score to 4.

For the reshuffling to be consistent with the structure of NMR spectra, metabomatching identifies *cut points* that separate the pseudospectrum into peak-preserving clusters of features and only reshuffles these clusters. The cut points are obtained as follows. Let f_i be the positions on the chemical shift axis of the metabolome features, sorted such that $f_i < f_{i+1}$, and C the set of cut points. First, metabomatching populates C with features bordering a *gap*, that is a region absent from the spectrum and larger than δ_{gap} (i.e., $f_i > f_{i-1} + \delta_{gap}$), with $\delta_{gap} = 0.3$ ppm as default value. Next it sorts the remaining features by their corresponding absolute-valued z-scores. Starting from the feature with the lowest absolute z-score it adds features to C provided they have a distance greater than δ_{min} to any features already assigned to C and an absolute z-score below a threshold z_{min}. Default values are 0.04 ppm for δ_{min} and the standard deviation of all absolute z-scores across the features of a given pseudospectrum for z_{min}.

ACP: Averaged Correlation Profile

ACP is a greedy approach to generate a list L of feature pairs and their corresponding correlation profiles c as input for metabomatching: (1) We compute and sort all pairwise correlations C_{ij} between features f_i and f_j separated by at least 0.1 ppm. (2) Starting with the feature pair $P = (i,j)$ with the highest correlation, we successively add feature pairs to L unless there is already a feature pair in L whose features are within 0.1 ppm of f_i and f_j, respectively. (3) For each feature pair in L, we define an *averaged correlation profile* as the average of the correlation profiles of f_i and f_j: $c_k = (C_{ik} + C_{jk})/2$. The correlation profiles of strongly correlated features are similar, consequently their average is similar to both of them. Crucially, the average does not contain an element equal to 1, as $c_i = c_j = (1 + C_{ij})/2 < 1$ given that $C_{ij} < 1$ in real data. For our analysis, we limit L to 179 averaged correlation profiles, 179 being the number of spectra in UMDB, the reference database on which metabomatching will run.

As an alternative approach, we tried agglomerative clustering of features, iteratively joining features (or sets of features) whose correlation was above a threshold C_{min}. At each step, we averaged joined (sets of) features into a metafeature and recomputed its correlation to all remaining (meta)features. We then built a correlation profile for each feature cluster by averaging the correlations profiles of the component features.

ISA: Iterative Signature Algorithm

ISA is a biclustering method first developed for modular analysis of gene expression data.[18,19] ISA uses a heuristic iterative procedure starting with random features to refine *modules*, consisting of self-consistent subsets of features *and* samples. Each module is defined for a set of two thresholds, determining how extreme the features and samples are allowed to be. Importantly, scanning through an array of thresholds usually identifies a set of modules (or module families) that is smaller than the number of samples or features.

We first ran the ISA algorithm to generate modules from the NMR data using the default values for the parameters except the following: (1) we changed both row and column thresholds from the default values {1, 2, 3} to {1, 2, 3, 4, 5, 6} to produce modules containing fewer rows or columns that are more likely to represent single metabolites, (2) we increased the number of seeds from 100 to 250, and (3) we lowered the correlation threshold below which ISA considers two modules equal to one another from 0.95 to 0.50 to favor diversity in modules.

We allowed feature scores to be either positive or negative, while sample scores were always positive. This means that modules can include features which have on average higher or lower intensities in the selected samples than for the remaining samples. Modules which include such a mixture are likely to represent (at least) two metabolites whose concentrations are inversely related to each other (like a substrate and its product).

This procedure generated 216 modules. To select the 179 modules as an input for metabomatching, we sorted them by the size of their basin of attraction. To measure the basin size, we ran ISA a second time with the same parameters, but on 10 000 seeds, turning off the *sweeping* option, and keeping all converged modules (by setting the *purge* option to false). We then assigned a run 2 module to the basin of the run 1 module with which it had the highest correlation (provided that correlation be greater than 0.5). We assumed that the run 1 modules attractor basin size is approximately equal to the count of modules from run 2 which were assigned to them in the previous step. Finally, we passed to metabomatching the 179 modules from run 1 with largest basins.

PCA: Principal Component Analysis

We used the sklearn library for Python (2.7) to compute all principal components of the preprocessed data. Specifically, we used the *decomposition*.PCA object, with *n_components* set to 687 and *svd_solver* set to *full*.

Identification of Metabolites

After running metabomatching on the pseudospectra generated by ISA, ACP and PCA, we applied a filtering algorithm to select only the most robust matches among all pseudospectra. The filtering passes only those pseudospectra which achieve an adjusted score above 2 with their top metabolite match (ensuring the pseudospectra finds a reasonable match by metabomatching) and have at least one peak with z-score above 4 (ensuring there is a strong signal). Note that multiple pseudospectra can match with the same metabolite NMR spectrum.

Out of the 179 pseudospectra from ACP, 10 pseudospectra matching different metabolites passed the filtering. Among these, only for one pseudospectrum, i.e., 3.87 and 3.75, the top match to hydroxypropionate (Figure S24) did not look convincing because of slightly worse matches to mannitol and arabitol (Figure S24).

Out of the 179 pseudospectra from ISA, 19 pseudospectra matching different metabolites passed the filtering. Among these, we discarded 9 for one or more of the following three reasons: (1) There are several metabolites which all achieve high adjusted scores (Figures S25 and S26). (2) There is at least one strong peak in the pseudospectrum that does not match with any of the spectra of the best matching metabolites (see Figures S27−S32). This may happen for pseudospectra with a large

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

number of peaks. These pseudospectra are not necessarily biologically irrelevant and might carry a signature for two or more related metabolites. (3) The pseudospectrum passed the filtering, yet did not appear to have sufficiently strong signals at the peak positions of its putative matching metabolites (Figures S33 and S34).

We analyzed all 687 pseudospectra from PCA and observed an elevation in metabomatching scores for the last principal components (all between components #505−687, which jointly explain only 1% of variation; Figures S1A, S35−S39). However, since these components explain almost no variation in the metabolome and the last 9 principal components matched the same metabolite, hippurate (Figure S1D), we investigated whether these matches rely on numerical instability. Indeed, when we removed one feature from all feature pairs that correlated above 0.95 (i.e., 8 features from 34 feature pairs all belonging to hippurate multiplets regions 7.54−7.56, 7.63−7.65, 7.83−7.84, and 3.96 ppm), and ran metabomatching on all pseudospectra generated from the principal components of the remaining features, only five passed the filtering (Figure S1F). However, none of these seemed convincing when applying the same curation as for the ISA pseudospectra.

Pseudoquantification of Metabolite Concentrations

We use the term pseudoquantification as this approach should not be confused with traditional quantification approach which sometimes rely on experiments that target a specific metabolite and often require the use of proprietary software operated by an expert.

We perform our pseudoquantification by using discrete integration to estimate relative metabolite concentrations, according to

$$c_i = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H_k}\sum_{l_k \le s_j \le r_k} I_{ij}\Delta_j \qquad (2)$$

where K is the number of multiplets, H_k the number of protons in multiplet k, $[l_k, r_k]$ the range of multiplet k, s_j the chemical shift of feature j, I_{ij} the intensity of this feature in individual i, and Δ_j the width of the bin at s_j. For example, for hippurate $K = 4$, $H = [2, 2, 1, 2]$, $l = m - 0.025$, $r = m + 0.025$, where $m = [3.98, 7.54, 7.65, 7.84]$. We then evaluated this relative concentration first using the peak positions from the reference spectrum as listed in UMDB, and second using the peak positions as suggested by the pseudospectrum found by the modular approach.

To perform the pseudoquantification based on the pseudospectra of the modules, we defined a set of multiplet positions to use in eq 2 for each module that robustly identified a metabolite. This set is composed of all the chemical shifts from the module of interest with z-scores above 3 and within 0.025 ppm of the multiplet positions of the matching metabolite in reference database. It includes all relevant peaks from the metabolite detected by the modular analysis.

■ ANALYSIS AND RESULTS

In this work, we show that metabomatching can be used with pseudospectra capturing the internal structure of large-scale NMR data rather than their correlation with external variables. Specifically, our premise is that in sizable sample collections there is sufficient power for methods identifying coherent features that may point to the same metabolite.

We used three methods for identifying weighted sets of covarying spectral features from large-scale NMR data that can be used as input for metabomatching (see Methods for more details and Figure 1 for an illustration of the workflow).

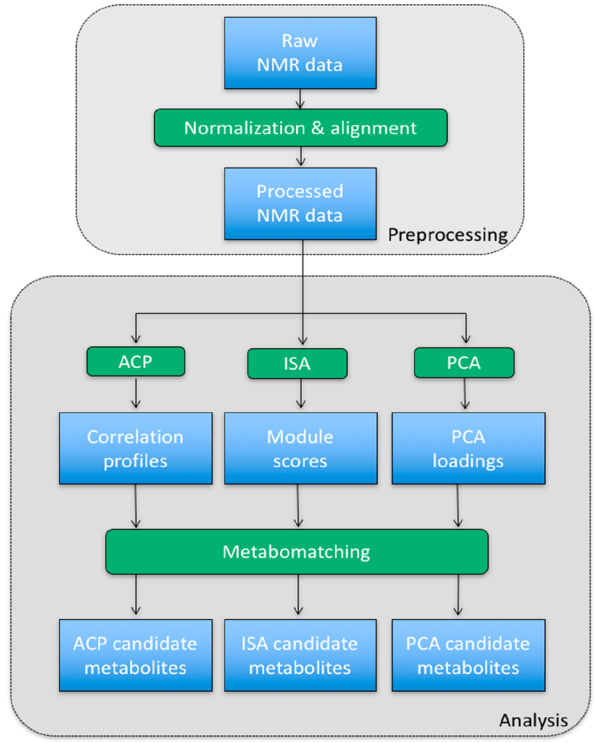


Figure 1. Workflow for unsupervised analysis of large-scale NMR data. Raw ^{1}H NMR data are normalized then aligned. These processed profiles are used as input for the averaged correlation profile (ACP), iterative signature algorithm (ISA), and principal component analysis (PCA) methods, which output correlation profiles, module scores, and PCA loadings, respectively. These outputs constitute possible pseudospectra for metabomatching, which identifies the most plausible candidate metabolites underlying the coherent feature variations.

Correlation-Based Pseudospectra

Our first approach to select covarying features was to use the correlations between features across all samples. We faced two challenges: First, using the correlations of a given feature with all features as a pseudospectrum would break the scoring algorithm in metabomatching, due to self-correlation of features that result in infinite z-scores. Second, as generating only a limited number of pseudospectra was desired ranking the input sets was necessary to select only the most relevant ones.

To address these challenges, we devised an algorithm that ranks all pairs of sufficiently distant features and computes averaged correlation profiles (ACP) for pairs of highly correlated features. Strictly speaking, ACPs are not the correlations but they are the average correlation profiles, with the premise that for highly correlated feature pairs the correlations to other features tend to be similar, while none of the averages equals to one. Within metabomatching these ACPs are then translated into z-scores using the Fisher z-transformation (see Methods for more details). We also tried hierarchical clustering to define pseudospectra from multiple highly similar features (see Methods), but this approach did not work as well.

Iterative Signature Algorithm

ISA has been designed for the unsupervised identification of coherent subsets in large-scale data.[18,19] Specifically, coherence

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

between features is *not* defined by total correlation across all samples, but rather by a subset of samples for which a set of features takes more extreme values than for the rest of the samples. Such a joint set of features and samples is called a *module*. In order to obtain a pseudospectrum for each module we averaged each feature's score (whether part of the module or not) across the samples assigned to the module. These averages were then transformed into *z*-scores. By definition the features of the module have the most extreme *z*-scores, yet other features that were just below the threshold may also have a sizable contribution. We then used these *z*-scores as input for metabomatching.

Principal Component Analysis

We also used PCA to compute the loadings of all features onto the eigenvectors of the sample−sample correlation matrix across all features. These eigenvectors (or eigensamples) characterize independent axes of variation in sample space. The corresponding eigenvalues reflect the fraction of total variation explained by each eigenvector. It was not clear a priori which principal components might characterize variation due to single metabolites. We therefore applied metabomatching to all of them. Specifically, we generated pseudospectra by standardizing the loadings corresponding to each eigensample (see Methods for more details).

Many Pseudospectra Defined by the ACP Method and ISA Match to Urine Metabolites

We observed a trend of elevated metabomatching scores for pseudospectra corresponding to principal components with small eigenvalues (starting from component #505), jointly explaining only 1% of variation (Figure S1A). The last nine principal components matched to hippurate, but disappeared when running PCA on the metabolome stripped of features that are highly correlated to other features (see Methods; Figures S1D and S1F). Additionally, the adjusted scores of all potential hits decreased significantly for the stripped metabolome (Figure S1E). We therefore concluded that PCA is not well-suited for generating robust metabolite signatures.

In contrast, our ACP method and ISA resulted in a sizable number of pseudospectra for which metabomatching produced robust matches to urine metabolites (see Figure 2 and Methods for details). Specifically, both ACP and ISA identified feature sets pointing to glucose, citrate, ethanol, hippurate, and P-hydroxyphenylacetate (Figures S2−S11). Glucose and hippurate were among the metabolites identified by Cloarec et al. with the correlation matrix of mouse urine NMR data.[1]

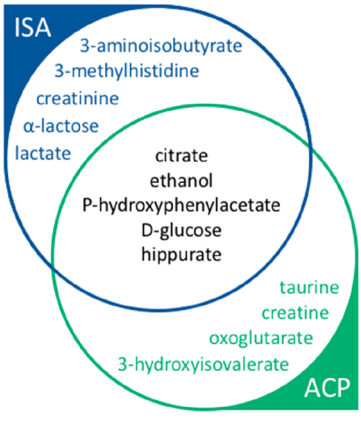


Figure 2. Urine metabolites that were robustly matched by metabomatching to pseudospectra derived from average correlation profiles (ACP, green), the iterative signature algorithm (ISA, blue), or both methods (black).

P-hydroxyphenylacetate shares an aromatic ring with 3-hydroxyphenylpropionate, another metabolite highlighted in the original STOCSY paper.[1] Both compounds are part of phenylalanine metabolism and occur as products of bacterial degradation of aromatic compounds. In human urine high concentrations of these compounds may reflect an overgrown *Clostridium* species in gut microbiota, which has been associated with autism spectrum disorders.[20]

In healthy humans, urine glucose should be low, but concentrations may be elevated due to diabetes or chronic kidney disease (CKD), conditions which are prevalent in the CoLaus population.

Citrate is an additive commonly used by the food industry and it is also synthesized as an intermediate product in the tricarboxylic acid cycle, a central pathway that releases stored energy from fat, proteins, and carbohydrates. Low urinary citrate is associated with CKD and kidney stone formation.

There are a number of metabolites that matched to pseudospectra generated only by one of the two methods. With ISA, we found modules matching 3-aminoisobutyrate, an end product of nucleic acid metabolism that has been considered a potential biochemical marker for cancer[21] (Figure S12); creatinine, a breakdown product of creatine, whose high and stable concentration in urine is often used for normalization (Figure S13); lactose (Figure S14); and lactate, the bacterial breakdown product of lactose (Figure S15). Conversely, ACP produced correlation profiles matching to taurine (Figure S16), an organic compound widely distributed in animal tissues and a major constituent of bile; creatine (Figure S14); oxoglutarate (α-ketoglutarate), an important biological compound produced by deamination of glutamate, and an intermediate in the Krebs cycle (Figure S18); and 3-hydroxyisovalerate, a byproduct of valine, leucine, and isoleucine degradation and a marker for biotin deficiency[22] (Figure S19). These compounds are all common urine metabolites that can exist in high concentrations.

Correcting features for significant covariates, both methods also found set of features matching carnitine, which owes its name to its high concentration in meat (see Methods for more details). While it is produced in both animal and plant cells, this may explain why we could detect its signature in human urine samples.

Metabolite Concentration Pseudoquantification with NMR Features of Matched Pseudospectra

We next investigated whether the sets of weighted NMR features generated by ISA or the ACP method can not only be used to identify metabolites, but also facilitate their pseudoquantification. This pseudoquantification approach aims to estimate the relative concentration of the metabolites in untargeted ^{1}H NMR of urine samples. We performed our pseudoquantification by computing the area under the peak of each multiplet in the metabolite spectrum using discrete integration, dividing it by the number of protons associated with the multiplet and then averaging scaled areas over all multiplets in the metabolite spectrum (see Methods for more details). We hypothesized that the leading features selected by our algorithms for a certain metabolite may be better suited for pseudoquantification than the full reference spectra from public databases such as UMDB. There are two possible reasons for this. First, the exact feature positions extracted from the data by ISA or the ACP method

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

Table 1. Correlation between Pseudoquantification and Measured Biomarkers of Glucose and Ethanol

urine metabolite	feature source	multiplet positions (ppm)	related biomarker	correlation [95% CI]
glucose	UMDB	3.23, 3.40, 3.46	serum	0.46
		3.52, 3.73, 3.82	glucose	[0.41, 0.52]
		3.88, 4.63, 5.22		
glucose	ACP: 3.48 and 5.24	3.40, 3.48, 4.65	serum	0.48
		5.24	glucose	[0.43, 0.54]
glucose	ACP: 3.89 and 5.24	3.82, 3.89, 4.65	serum	0.44
		5.24	glucose	[0.38, 0.49]
glucose	ISA: module #16	3.25, 3.41, 3.48	serum	0.50
		3.50, 3.89, 4.65	glucose	[0.44, 0.55]
		5.24		
ethanol	UMDB	1.17, 3.65	serum	0.29
			CDT	[0.23, 0.35]
ethanol	ACP: 1.18 and 3.67	1.18, 3.67	serum	0.16
			CDT	[0.10, 0.22]
ethanol	ISA: module #57	1.18, 3.67	serum	0.16
			CDT	[0.10, 0.22]
EtG	Nicholas et al.[23]	1.24, 3.30, 3.52	serum	0.36
		3.71, 3.99, 4.48	CDT	[0.30, 0.42]
EtG	ISA: module #240	1.24, 3.52, 4.47	serum	0.46
			CDT	[0.40, 051]

may be more accurate, even if, by design, they fall within the margin of the matching window of reference spectrum features. Second, for metabolites with several peaks, only a subset might have been picked up by these algorithms. Indeed, both ISA and the ACP method may leave out peaks that did not contribute coherently to the peak set since their signal was too noisy (e.g., due to overlap with those from other metabolites).

We performed our pseudoquantification method (using eq 2 in Methods) to estimate concentrations of glucose and ethanol, for which relevant phenotypes were available in the cohort. For urine glucose, the phenotype was fasting blood glucose. For urine ethanol, relevant biomarkers included serum asialotransferrin and disialotransferrin, which combined are known as carbohydrate-deficient transferrin (CDT), a biomarker for heavy alcohol use. Furthermore, self-reported alcohol consumption was available. These biomarkers were measured in a different biofluid (i.e., blood), which was collected on the same day as the urine sample. We argue that detecting significant correlations between our estimated metabolite concentrations and these biomarkers provides a proof of concept that our pseudoquantification is reliable, and comparing correlations between different pseudoquantification approaches provides a means to evaluate them.

The ^{1}H NMR spectra of glucose has nine multiplets. Including all these multiplets chemical shifts from the UMDB database (Table 1) to perform pseudoquantification, we obtain a correlation of 0.46 (with a 95% confidence interval (CI) of [0.41, 0.52]) between the estimated concentration of urine glucose and fasting blood glucose. This correlation increases to 0.50 [0.44, 0.55] if the subset of seven multiplets from ISA module #16 (Figure 3A) is used for pseudoquantification (see Methods for details). From the 179 ACP pseudospectra, two robustly matched glucose, one from averaging the correlation profiles from feature pair 3.48 and 5.24 and another from averaging correlation profiles from 3.89 and 5.24 (Figure 3B,C), each capturing four out of nine multiplets of glucose (Table 1). Using the subset of peaks from 3.48 and 5.24 ACP pseudospectra we obtain a correlation of 0.48 [0.43, 0.54] between glucose pseudoquantification and fasting blood glucose while using the subset of peaks from 3.89 and 5.24 ACP pseudosspectra a correlation of 0.44 [0.38, 0.49] is obtained. Combining the peak subsets from both pseudospectra to a subset of 6 peaks did not improve the correlation beyond 0.48 [0.43, 0.54] (see Table 1 and Methods for more details on peak sets used for pseudoquantification).

The ^{1}H NMR spectra of ethanol has only two multiplets (as well as one singlet from the hydroxyl group, which can not be discerned in water solutions like urine). Using the reference spectrum from UMDB to perform pseudoquantification, we obtained a relatively low correlation of 0.29 [0.23, 0.35] between the estimated concentration of urine ethanol and CDT levels in serum (Table 1, see Table S1 for other alcohol markers). The ACP method produced one pseudospectra, 1.18 and 3.67, and ISA produced two modules (#57 and #240) that metabomatching matched to ethanol (Figures S6, S7, and S22). The positions of the ACP peak set (i.e., 1.18 and 3.67) were identical to those of ISA module #57 and were more similar to the ethanol spectrum (achieving a higher adjusted score in metabomatching) than those of ISA module #240 (Figure 3D−F). Nevertheless, pseudoquantification for the ACP pseudospectrum and ISA module #57 yielded a lower correlation of 0.16 [0.10, 0.22] with CDT levels than the UMDB reference peaks (0.29 [0.23, 0.35]) (Table 1). This is due to a high correlation of CDT with the features at 1.145−1.155 ppm, which are within a 0.025 ppm neighborhood of the UMDB ethanol peak at 1.17 ppm but not within the same neighborhood of the 1.18 ethanol peak from ACP and ISA module #57 (Figure S20). Yet, these peaks at 1.145−1.155 ppm are unlikely to correspond to ethanol, since their correlation with the other ethanol peak at 3.67 ppm is much weaker than the correlation between the 1.18 and 3.67 ppm peaks. Instead, they may belong to a different metabolite whose concentration is correlated with CDT (Figure S21). In contrast, summing up the intensities over all the features of ISA module #240 with a z-score above 3 as a pseudoquantification measure (in the absence of any multiplet information), we obtained a correlation of 0.51 [0.46, 0.57] with the CDT measurements.

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

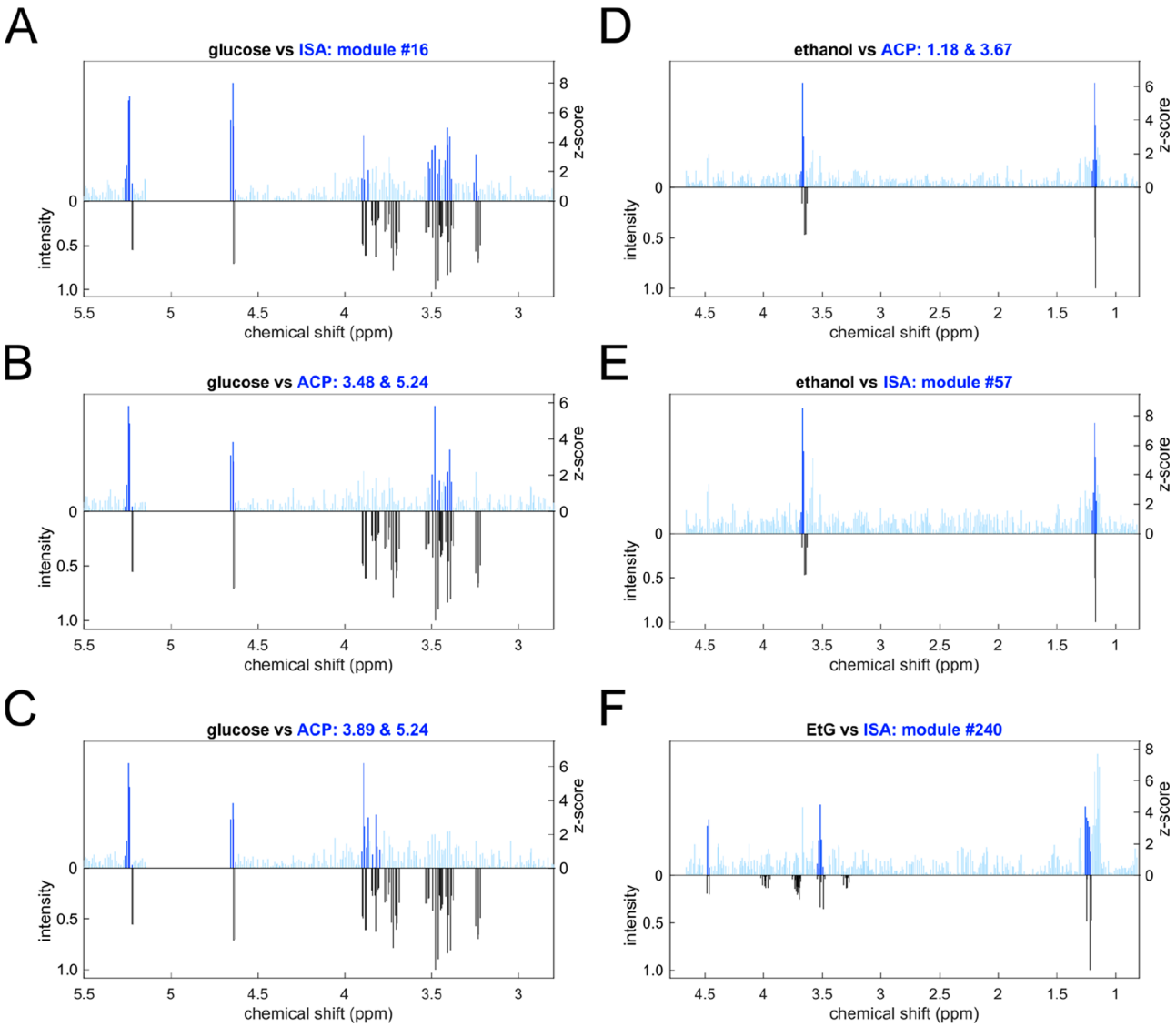


Figure 3. Pseudospectra from ACP and ISA algorithms matching glucose, ethanol, and EtG. Each plot shows the pseudospectrum in blue in the upper half and the reference spectrum from UMDB in black and in the lower half. Dark blue indicates chemical shifts and their ±0.025 ppm vicinity that were used for pseudoquantification.

To better understand why module #240 correlates more strongly to the alcohol consumption biomarker while being a worse match to ethanol than module #57 (Figure 3), we studied whether any of its features point to other compounds related to ethanol metabolism. Indeed, we found that this module contains three features, at 1.26, 3.52, and 4.47 ppm, that individually correlate more strongly to CDT (0.40 [0.34, 0.45], 0.29 [0.23, 0.35], 0.33 [0.27, 0.39], respectively) than the features mapping to ethanol. Interestingly, these features appeared to be close to those of ethyl glucuronide (EtG), a direct product of ethanol nonoxidative metabolism by conjugation with uridine diphosphate (UDP)-glucuronic acid, which had previously been detected in ^{1}H NMR spectra of liver extracts[23] and more recently in human urine of alcohol drinkers.[24] To confirm EtG as a possible match for ISA module #240, we added its features as extracted from Nicholas et al.[23] to the metabomatching library manually, since EtG had no entry in UMDB. We observed that ethanol and EtG spectra together provided a better match to ISA module #240 than ethanol alone (Figure S22). Interestingly, the distance between the two peaks corresponding to the doublet at 4.48 ppm is about 0.0126 ppm, corresponding to a coupling of 8.8 Hz (for a 700 MHz spectrometer) consistent with the coupling of 8 Hz reported in Nicholas et al.[23] (see Figure S23 and Supporting Informationfor more details).

Performing the pseudoquantification of EtG using the peak set of 6 reference positions extracted from Nicholas et al.[23], we obtained a correlation of 0.36 [0.30, 0.42] with CDT levels; performing the pseudoquantification with the 3 feature subset from module #240, we obtained a correlation of 0.46 [0.40, 0.51] (Table 1). This indicates that EtG pseudoquantification correlates better with CDT than ethanol, which is in agreement with the fact that EtG is detectable in urine for a longer time window (2−5 days) than ethanol (12−24 h), and CDT is a marker for heavy alcohol use (at least five drinks a day over a period of 2 weeks before giving the sample[25]). Remarkably, the pseudoquantification facilitated by the three features of module #240 correlates even more strongly with CDT than the full set of EtG reference features, presumably because these features have the best signal-to-noise ratio and optimal position for our data. They may therefore constitute a promising urine biomarker for

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

heavy alcohol consumption. Indeed, while the correlation between EtG pseudoquantification and CDT measure increases to 0.59 [0.46, 0.72] when focusing on subjects who have self-reported heavy drinking, pseudoquantification of module #240 gives rise to a slightly higher correlation of 0.61 [0.48, 0.74].

CONCLUSIONS AND DISCUSSION

In this work, we implemented and tested new methodologies for analyzing large-scale ^{1}H NMR spectroscopy data. Building on previous ideas to use the correlation structure of such data to generate metabolomic signatures, we investigated three complementary methods for generating such signatures and benchmarked the methods in terms of how many of their signatures matched with reference spectra in public databases. By design, these approaches will only identify metabolites with at least two distinct peaks, and therefore complement peak-picking identification approaches, which tend to focus on single peak metabolites.

We found that average correlation profiles (ACP) of highly correlated feature pairs, a method inspired by STOCSY, as well as the iterative signature algorithm (ISA) identified ten and nine metabolites, respectively, five of which overlapped. In contrast, principal component analysis (PCA) did not generate any pseudospectra with robust metabomatching, likely because leading components explain variation driven by many metabolites.

While ACP is designed to pick up individual metabolites with at least two (nonproximal) features in their spectrum (or those of metabolite pairs whose concentrations are coupled), ISA is able to generate modules where many features exhibit coherent variation, yet potentially only over a subset of samples. We believe that this may be particularly useful when integrating data from a heterogeneous set of samples (e.g., including those from diseased or medicated subpopulations).

One interesting property of our modular approach is that the feature sets identified by ACP or ISA do not need to match perfectly with those of the reference spectrum of the corresponding compound. Indeed, the two ACP signatures matching glucose each only cover four and jointly six of the nine glucose peaks, while the ISA module with the best match to glucose includes seven of its peaks. Adding the "missing" peaks in our pseudoquantification slightly reduced the correlation with serum glucose, indicating there is a marginal improvement in the pseudoquantification using ppm positions only from the multiplets found by our algorithms rather than the database. Further work will be needed to substantiate this observation.

Another interesting aspect of our approach is that modular feature sets may match multiple compounds. Our current implementation of metabomatching allows simultaneous identification of up to two compounds. Indeed, our finding that ISA picked up a module whose signature mapped well to ethanol and its specific metabolic product ethyl glucuronide demonstrated the potential power of ISA to identify metabolite pairs within the same pathway. Moreover, the strong correlation of this module with the alcohol abuse marker CDT was likely driven by the fact that ISA can extract context specific covariance, which in this case is strongest in samples with particularly high alcohol consumption. This module also highlighted that using the relevant chemical shifts found by the module rather than all shifts from the reference database can lead to more accurate pseudoquantification of the underlying metabolites, due to different contribution of shifts specific to the experimental conditions in the complex urine spectra.

Extending metabomatching beyond compound pairs is challenging due to the large number of possible trios and higher order combinations, but could be feasible in future work, for example by using metabolic pathway information to limit the number of relevant metabolite combinations to test.

A critical element of our analysis was to transform the signatures generated by the different methods into a universal format (i.e., *z*-scores) as input for our metabomatching tool that we previously developed for the analysis of feature signatures generated by regression on external variables. Indeed, being able to query both internal and external signatures of large-scale NMR data against a reference data set of known spectra from individual metabolites is pivotal for exploring new methods dissecting the auto- and cross-correlation structure for integrative analyses.

In conclusion, we believe that our study using fewer than 1000 samples gives ample evidence for the potential of automated analysis of large-scale NMR data, and that increased sample sizes are likely to result in further identifications and more accurate pseudoquantifications of individual metabolites. To this end, our analysis software, metabomodules, is made publicly available on GitHub https://github.com/BergmannLab/metabomodules-docker.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.9b00295.

Supporting figures and table (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: sven.bergmann@unil.ch.

ORCID

Bita Khalili: 0000-0001-5630-1812

Author Contributions

‖B.K., M.T., and M.M. are co-first authors. R.R. and S.B. are co-last authors. S.B. and R.R. designed the study. The manuscript was written by B.K. and S.B. The correlation-based methods was implemented and applied by B.K. The modular analysis with ISA was performed by M.M. and R.R. PCA was run by M.T., D.K., and B.K. All authors discussed the results and implications, and contributed to the manuscript.

Notes

The authors declare no competing financial interest.

Our analysis software, metabomodules, is made publicly available on GitHub: https://github.com/BergmannLab/metabomodules-docker.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (grant FN 310030_152724/1) and the NIH (grant R03 CA211815).

REFERENCES

(1) 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

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368

for Latent Biomarker Identification from Metabolic 1H NMR Data Sets. *Anal. Chem.* **2005**, *77* (5), 1282−1289.
(2) Blaise, B. J.; Navratil, V.; Domange, C.; Shintu, L.; Dumas, M.-E.; Elena-Herrmann, B.; Emsley, L.; Toulhoat, P. Two-Dimensional Statistical Recoupling for the Identification of Perturbed Metabolic Networks from NMR Spectroscopy. *J. Proteome Res.* **2010**, *9* (9), 4513−4520.
(3) Sands, C. J.; Coen, M.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Data-Driven Approach for Metabolite Relationship Recovery in Biological 1H NMR Data Sets Using Iterative Statistical Total Correlation Spectroscopy. *Anal. Chem.* **2011**, *83* (6), 2075−2082.
(4) 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.
(5) Hao, J.; Astle, W.; De Iorio, M.; Ebbels, T. M. D. BATMAN—an R Package for the Automated Quantification of Metabolites from Nuclear Magnetic Resonance Spectra Using a Bayesian Model. *Bioinformatics* **2012**, *28* (15), 2088−2090.
(6) Alonso, A.; Rodríguez, M. A.; Vinaixa, M.; Tortosa, R.; Correig, X.; Julià, A.; Marsal, S. Focus: A Robust Workflow for One-Dimensional NMR Spectral Analysis. *Anal. Chem.* **2014**, *86* (2), 1160−1169.
(7) Ravanbakhsh, S.; Liu, P.; Bjorndahl, T. C.; Mandal, R.; Grant, J. R.; Wilson, M.; Eisner, R.; Sinelnikov, I.; Hu, X.; Luchinat, C.; et al. Accurate, Fully-Automated NMR Spectral Profiling for Metabolomics. *PLoS One* **2015**, *10* (5), No. e0124219.
(8) Tardivel, P. J. C.; Canlet, C.; Lefort, G.; Tremblay-Franco, M.; Debrauwer, L.; Concordet, D.; Servien, R. ASICS: An Automatic Method for Identification and Quantification of Metabolites in Complex 1D 1H NMR Spectra. *Metabolomics* **2017**, *13* (10), 109.
(9) Röhnisch, H. E.; Eriksson, J.; Müllner, E.; Agback, P.; Sandström, C.; Moazzami, A. A. AQuA: An Automated Quantification Algorithm for High-Throughput NMR-Based Metabolomics and Its Application in Human Plasma. *Anal. Chem.* **2018**, *90* (3), 2095−2102.
(10) Cañueto, D.; Gómez, J.; Salek, R. M.; Correig, X.; Cañellas, N. rDolphin: A GUI R Package for Proficient Automatic Profiling of 1D 1H-NMR Spectra of Study Datasets. *Metabolomics* **2018**, *14* (3), 24.
(11) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608−D617.
(12) Bouatra, S.; Aziat, F.; Mandal, R.; Guo, A. C.; Wilson, M. R.; Knox, C.; Bjorndahl, T. C.; Krishnamurthy, R.; Saleem, F.; Liu, P.; et al. The Human Urine Metabolome. *PLoS One* **2013**, *8* (9), No. e73076.
(13) Psychogios, N.; Hau, D. D.; Peng, J.; Guo, A. C.; Mandal, R.; Bouatra, S.; Sinelnikov, I.; Krishnamurthy, R.; Eisner, R.; Gautam, B.; et al. The Human Serum Metabolome. *PLoS One* **2011**, *6* (2), No. e16957.
(14) Nagana Gowda, G. A.; Gowda, Y. N.; Raftery, D. Expanding the Limits of Human Blood Metabolite Quantitation Using NMR Spectroscopy. *Anal. Chem.* **2015**, *87* (1), 706−715.
(15) Rueedi, R.; Ledda, M.; Nicholls, A. W.; Salek, R. M.; Marques-Vidal, P.; Morya, E.; Sameshima, K.; Montoliu, I.; Da Silva, L.; Collino, S.; et al. Genome-Wide Association Study of Metabolic Traits Reveals Novel Gene-Metabolite-Disease Links. *PLoS Genet.* **2014**, *10* (2), No. e1004132.
(16) Rueedi, R.; Mallol, R.; Raffler, J.; Lamparter, D.; Friedrich, N.; Vollenweider, P.; Waeber, G.; Kastenmüller, G.; Kutalik, Z.; Bergmann, S. Metabomatching: Using Genetic Association to Identify Metabolites in Proton NMR Spectroscopy. *PLoS Comput. Biol.* **2017**, *13* (12), No. e1005839.
(17) Raffler, J.; Friedrich, N.; Arnold, M.; Kacprowski, T.; Rueedi, R.; Altmaier, E.; Bergmann, S.; Budde, K.; Gieger, C.; Homuth, G.; et al. Genome-Wide Association Study with Targeted and Non-Targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet.* **2015**, *11* (9), No. e1005487.
(18) Ihmels, J.; Bergmann, S.; Barkai, N. Defining Transcription Modules Using Large-Scale Gene Expression Data. *Bioinformatics* **2004**, *20* (13), 1993−2003.
(19) Bergmann, S.; Ihmels, J.; Barkai, N. Iterative Signature Algorithm for the Analysis of Large-Scale Gene Expression Data. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **2003**, *67* (3), No. 031902.
(20) Xiong, X.; Liu, D.; Wang, Y.; Zeng, T.; Peng, Y. Urinary 3-(3-Hydroxyphenyl)-3-Hydroxypropionic Acid, 3-Hydroxyphenylacetic Acid, and 3-Hydroxyhippuric Acid Are Elevated in Children with Autism Spectrum Disorders. *BioMed Res. Int.* **2016**, *2016*, 9485412.
(21) Nielsen, H. R.; Killmann, S. A. Urinary Excretion of Beta-Aminoisobutyrate and Pseudouridine in Acute and Chronic Myeloid Leukemia. *J. Natl. Cancer Inst.* **1983**, *71* (5), 887−891.
(22) Ziegler, E. E., Ed.; *Present Knowledge in Nutrition*; Filer, L. J. J., Engl. Ed.; International Life Sciences Inst.: Washington, D.C., 1996.
(23) Nicholas, P. C.; Kim, D.; Crews, F. T.; Macdonald, J. M. Proton Nuclear Magnetic Resonance Spectroscopic Determination of Ethanol-Induced Formation of Ethyl Glucuronide in Liver. *Anal. Biochem.* **2006**, *358* (2), 185−191.
(24) Kim, S.; Lee, M.; Yoon, D.; Lee, D.-K.; Choi, H.-J.; Kim, S. 1D Proton NMR Spectroscopic Determination of Ethanol and Ethyl Glucuronide in Human Urine. *Bull. Korean Chem. Soc.* **2013**, *34* (8), 2413−2418.
(25) Solomons, H. D. Carbohydrate Deficient Transferrin and Alcoholism. *GERMS* **2012**, *2* (2), 75−78.

DOI: 10.1021/acs.jproteome.9b00295
J. Proteome Res. 2019, 18, 3360−3368