Journal of proteome research

Article

pubs.acs.org/jpr

Bayesian Independent Component Analysis Recovers Pathway Signatures from Blood Metabolomics Data

Jan Krumsiek,[†] Karsten Suhre,[†,‡] Thomas Illig,[§,∥] Jerzy Adamski,[⊥,#] and Fabian J. Theis*[,†,∇]

[†]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany

[‡]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, State of Qatar

[§]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Germany

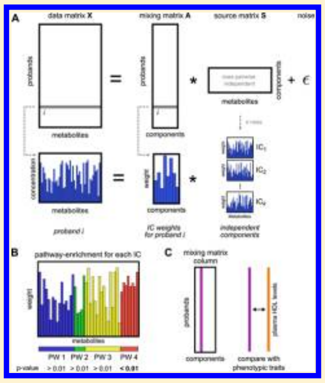
[∥]Biobank of the Hanover Medical School, Germany

[⊥]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, Germany

[#]Lehrstuhl für Experimentelle Genetik, Technische Universität München, 85350 Freising-Weihenstephan, Germany

[∇]Department of Mathematics, Technische Universität München, Germany

S *Supporting Information*



ABSTRACT: Interpreting the complex interplay of metabolites in heterogeneous biosamples still poses a challenging task. In this study, we propose independent component analysis (ICA) as a multivariate analysis tool for the interpretation of large-scale metabolomics data. In particular, we employ a Bayesian ICA method based on a mean-field approach, which allows us to statistically infer the number of independent components to be reconstructed. The advantage of ICA over correlation-based methods like principal component analysis (PCA) is the utilization of higher order statistical dependencies, which not only yield additional information but also allow a more meaningful representation of the data with fewer components. We performed the described ICA approach on a large-scale metabolomics data set of human serum samples, comprising a total of 1764 study probands with 218 measured metabolites. Inspecting the *source matrix* of statistically independent metabolite profiles using a weighted enrichment algorithm, we observe strong enrichment of specific metabolic pathways in all components. This includes signatures from amino acid metabolism, energy-related processes, carbohydrate metabolism, and lipid metabolism. Our results imply that the human blood metabolome is composed of a distinct set of overlaying, statistically independent signals. ICA furthermore produces a *mixing matrix*, describing the strength of each independent component for each of the study probands. Correlating these values with plasma high-density lipoprotein (HDL) levels, we establish a novel association between HDL plasma levels and the branched-chain amino acid pathway. We conclude that the Bayesian ICA methodology has the power and flexibility to replace many of the nowadays common PCA and clustering-based analyses common in the research field.

KEYWORDS: *metabolomics, independent component analysis, Bayesian, systems biology, bioinformatics, blood serum, population cohorts*

1. INTRODUCTION

Metabolomics is a newly arising *omics* technology aiming at the quantification of ideally all metabolites in a given tissue, cell culture, or biofluid.[1,2] The field of metabolomics has tremendously advanced in the past few years, with discoveries in epidemiology,[3,4] nutritional challenging,[5,6] and molecular cell biology mechanisms.[7,8] Understanding the functional relationships between metabolite concentrations and physiological traits, however, remains a challenging task.

The majority of previously published metabolomics studies follows a supervised data analysis approach, where metabolite levels are investigated with respect to a given phenotype, condition, or quantitative trait. *T* tests, analyses of variance (ANOVAs) and related statistical tests are frequently used to assess group-wise differences of metabolite concentrations, for instance, for medication-induced changes[9] or cancer progression.[10] Furthermore, linear regression models can be used to detect metabolic changes correlating with quantitative traits, for example, for changes in insulin action.[11] Another popular approach is the use of supervised linear mixture models. As probably one of the most prominent examples in metabolomics, partial least-squares discriminant analysis (PLS-DA) attempts to find a projection of multivariate metabolite data such that sample groups in the data are maximally separated with respect to a given phenotype. An example application is the separation of patients suffering from Parkinson's disease versus control individuals.[12]

Unsupervised data analysis techniques, in contrast, use concentration data alone to detect intrinsic relations between

Received: March 9, 2012
Published: June 20, 2012

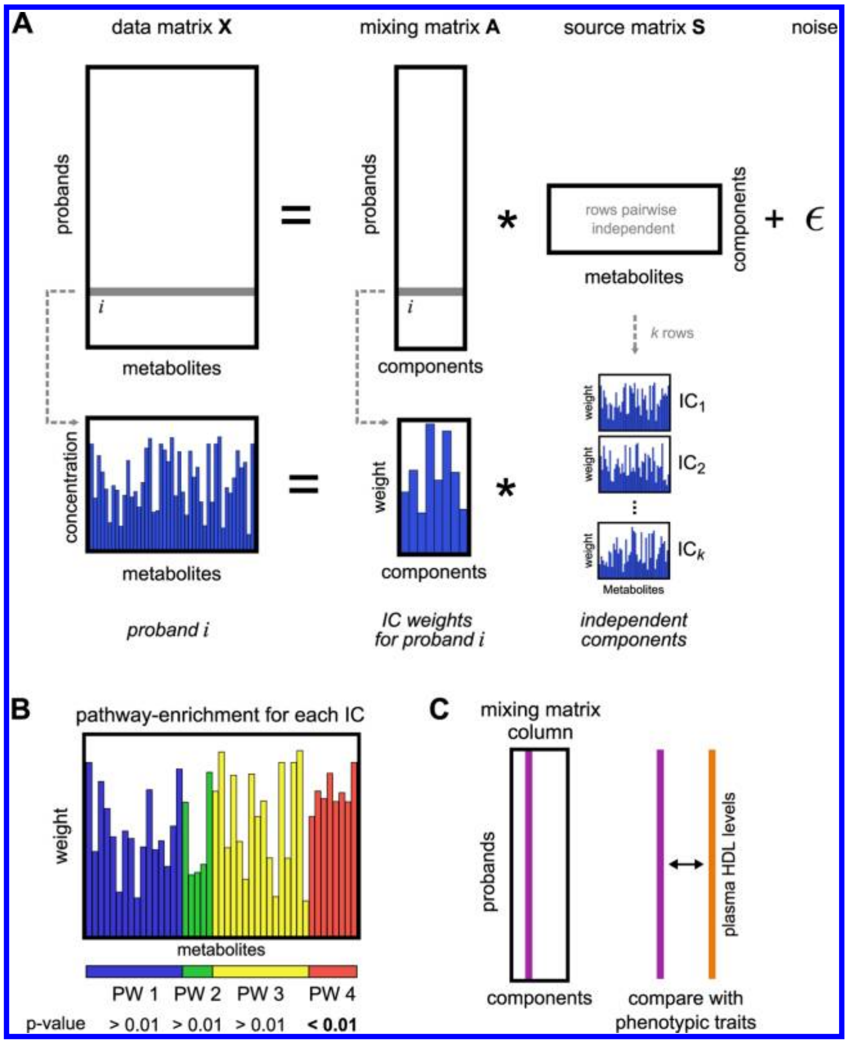


Figure 1. (A) ICA model applied to metabolomics data. The data matrix **X** is decomposed into the product of a mixing matrix **A** and a source matrix **S**, cf. eq 1 in the text. The source matrix contains statistically independent profiles of metabolites ($s_{l \cdot}$, termed "IC" = independent component throughout the manuscript), whereas the mixing matrix represents the contribution strengths of each component to the respective metabolomics sample. (B) Concept of pathway enrichment performed for each IC. We statistically assess whether the IC contributions for the metabolites from a specific pathway are higher than expected by chance. (C) Each column in the mixing matrix represents a newly derived variable in the data set that can be correlated with other proband-specific traits.

measured entities. This approach is commonly used as an important, explorative step in the understanding of multivariate *omics* observations and is followed by a subsequent supervised or correlative analysis. A well-known approach for unsupervised analysis is cluster analysis, where related groups of measured samples are determined from the data (c.f., e.g., Oresic et al.[13]). In addition, principal component analysis (PCA, another example of a linear mixture model) searches for mutually decorrelated directions in metabolite vectors that explain maximal variance in the data.[14]

While PCA is a conceptually simple and powerful tool for multivariate analysis, it only considers second-order dependencies (i.e., correlations) of metabolite variables. However, in practice, we frequently observe higher order dependencies, which may yield additional information that is otherwise neglected. Metabolomics data, for instance, do not display an entirely Gaussian distribution even after logarithmizing,[15] thus leaving multivariate dependencies, which cannot be captured by second-order statistics. In this paper, we aim at using the full-order multivariate statistics in an explorative analysis of metabolomics data; hence, we propose the use of independent component analysis (ICA) as a statistically motivated extension of PCA for metabolomics data.[16] The introduction of statistical independence here naturally generalizes the concept of decorrelation for non-normal data.

For ICA, we assume metabolite profiles to be composed of statistically independent components (ICs), whose mixture makes up the measured metabolomics profile. Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}_+^{n \times p}$ be the preprocessed data matrix, where each of the n rows corresponds to one measured study proband, and each of the p columns represents one metabolite. For a given number of components k, ICA attempts to find a factorization of the data matrix

$$x_{ij} = \sum_{l=1}^{k} a_{il} \cdot s_{lj} + \epsilon_{ij} \qquad (1)$$

where the *mixing matrix* $\mathbf{A} = (a_{il})$ is of dimension $n \times k$, the *source matrix* $\mathbf{S} = (s_{lj})$ is $k \times p$, and ϵ_{ij} represents independent, normally distributed noise (Figure 1A). The particularity of ICA is the requirement of all rows $s_{l \cdot}$ in $\mathbf{S}$ (which we will refer to as IC$_j$) to be samples of a statistically independent random vector. Interpreted biologically, each row in **S** represents a distinct metabolic process, which contributes to the overall concentration profile. The matrix **A**, on the other hand, reflects how strong each of these processes is *active* in a given sample (study proband in our case). In other words, instead of describing the metabolome of each proband by p numeric values, after ICA, we can equivalently represent the metabolome using only $k \ll p$ values. It can be shown that the decomposition into **A** and **S** is unique given sufficiently many samples.[17,18]

In biomedical research, ICA is commonly used as a method for high-dimensional data reduction and analysis. Early applications from the neuroscience field include the analysis of electroencephalographic measurements[19] and fMRI data.[20−22] For molecular biology, ICA has frequently been used to analyze transcriptomics data, for example, for cancer classification[23−25] or the investigation of cell differentiation.[26,27] Moreover, several studies already applied ICA in the context of metabolomics data, for instance, for the analysis of plant parasites[28] and toxins[29] and for metabolite fingerprinting.[30] While certainly interesting for their respective biological questions, these metabolomics studies merely used ICA as a data compression and visualization method rather than functionally investigating the reconstructed ICs in detail. The only studies that, to the best of our knowledge, performed a functional analysis of **A** and **S** are (i) Wienkoop et al.,[31] who did a joint ICA of metabolomics and proteomics data in starch metabolism, and (ii) Martin et al.,[32] who investigated the development of colitis in mice using NMR metabolomics.

In our study, we employ a Bayesian ICA approach. The key idea of Bayesian inference is to interpret each parameter as a random distribution. These distributions are then estimated using Bayes rule, for example, by Markov chain Monte Carlo methods or simply by maximum a posteriori estimation. With an inferred parameter distribution at hand, we can obtain both conventional point estimates but also parameter error estimates as provided by the respective variance. Moreover, by choosing adequate priors, we can include known information beforehand. In our case, we require nonnegative values of both the source and the mixing matrix. We argue that such nonnegativity better represents biological processes than arbitrarily negative matrix entries. In classical ICA, the choice of model parameters such as the number of components k to be reconstructed is a nontrivial problem. Usually, an ad hoc number of components is chosen, thereby accepting possible fusions of components (if too few are selected) or generation of information-free noise components.[16] A series of tools for identifying the correct model have been developed in the ICA community, mostly using heuristics, for example, based on clustering similar components.[33,34] We here evaluate the Bayesian Information Criterion (BIC) for each ICA calculation to get a trade-off between model accuracy (how close the matrix product gets to the original data matrix) and the number of parameters in the model. Finally, we select the number of components for which we obtained the highest BIC value. Methodologically, we applied a Bayesian mean-field ICA method,[35] which uses an EM-like parameter estimation scheme.

The novelty in the present study is the application of parameter-free, Bayesian, noisy ICA approach to metabolomics data, followed by a functional analysis of both independent metabolite processes in **S** as well as proband-specific signals in **A**. *Parameter-free, noisy, Bayesian* here refers to (i) avoiding a manual selection of the number of components k; (ii) obtaining an actual distributions **S**, thus providing confidence intervals for the reconstructed values; and (iii) allowing for an independently estimated noise term ϵ_{ij}.

The manuscript is organized as follows: First, we apply ICA to a large data set of human blood serum metabolomics samples of 1764 probands and 218 measured metabolites (Figure 1A) and estimate the number of components k using the above-mentioned Bayesian mean-field ICA approach. Next, we investigate the source matrix **S**, first by manual investigation and then by calculating the statistical enrichment of known metabolic pathways in each component (Figure 1B). We demonstrate that the approach outperforms PCA, k-means clustering, as well as fuzzy c-means with respect to biological pathway enrichment. In the final results part, we correlate the columns of the mixing matrix **A** to HDL (high-density lipoprotein) concentrations in blood plasma (Figure 1C). One IC correlates stronger with HDL concentrations than all metabolites in the data set alone. We thereby establish a novel connection between blood plasma HDL and branched-chain amino acids and discuss potential biological implications. Bayesian ICA calculation code and an implementation of the enrichment algorithm are available from http://cmb.helmholtz-muenchen.de/metaica.

2. MATERIALS AND METHODS

2.1. Metabolomics Data Set and Annotations

We used metabolomics data from the German KORA F4 study, as previously described in Suhre et al.[36] Briefly, metabolic profiling was performed using ultrahigh-performance liquid phase chromatography and gas chromatography separation, coupled with tandem mass spectrometry. The data set consists of 1764 fasting blood serum samples and a total of 218 measured metabolites from various pathways. For each metabolite, one of the following eight *superpathway* annotations was provided: “Lipid”, “Carbohydrate”, “Amino acid”, “Xenobiotics”, “Nucleotide”, “Energy”, “Peptide”, “Cofactors and vitamins”. Furthermore, there are a 61 *subpathway* annotations like “Oxidative phosphorylation”, “Carnitine metabolism”, or “Valine, leucine and isoleucine metabolism”. The complete set of measured metabolites and their respective pathway annotations can be found in the Supporting Information, S6.

Fatty acid metabolites are described by the number of carbon atoms, double bonds, and, if applicable, position of the last double bond. For instance, “fatty acid 18:2(n-6)” denotes a fatty acid with 18 carbon atoms and two double bonds, the last of which lies at the n-6 position (between carbon atoms 12 and 13). Phospholipids are named by the type of phospholipid and the fatty acids in both side chains. For example, PI(20:4(n-6)/0:0) represents a phosphatidylinositol containing an arachidonate residue (20 carbon atoms, four double bonds, n-6) at the sn-1 position. PC(0:0/18:0) contains a 18:0 fatty acid at the sn-2 position. Note that the current metabolite panel only measures lyso-phospholipids, that is, phospholipids with only one fatty acid chain. Phospholipid class abbreviations are as follows: PC, phosphatidylcholine; PI, phosphatidylinositol; and PE, phosphatidylethanolamine.

2.2. Bayesian ICA Model and Component Selection

For preprocessing, the data matrix **X** was column-normalized to unit variance and subsequently scaled between 0 and 1. We solved the described noisy source separation problem by probabilistic ICA.[37,38] Assuming normally distributed white noise with covariance matrix Σ, the mixing model results in the model likelihood

$$P(\mathbf{X}|\mathbf{A}, \mathbf{S}, \Sigma) = (\det 2\pi\Sigma)^{-N/2} \exp\left(-\frac{1}{2}tr(\mathbf{X} - \mathbf{AS})^{T}\Sigma^{-1}(\mathbf{X} - \mathbf{AS})\right)$$

which describes the probability of observing data **X** given mixing matrix **A**, sources **S**, and noise with covariance Σ. Instead of maximizing this likelihood, we follow a Bayesian approach and consider the model posterior $P(\mathbf{A}, \mathbf{S}, \Sigma|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{A}, \mathbf{S}, \Sigma)P(\mathbf{A})P(\mathbf{S})P(\Sigma)$ with (independent) priors $P(\mathbf{A})$, $P(\mathbf{S})$, and $P(\Sigma)$. Full sampling of this posterior is too time-consuming and requires more elaborate Markov Chain Monte Carlo sampling. We decided to follow a simpler two-step EM type algorithm by iteratively estimating first source posterior $P(\mathbf{S}|\mathbf{X}, \mathbf{A},\Sigma)$ and then point estimates of **A** and Σ using a MAP (maximum a posteriori) estimator. We used a mean-field-based algorithm proposed by Højen-Sørensen et al.,[35] since it allows flexible choice of source priors. We assumed nonnegative mixing matrix and exponentially distributed source weights. We then analyzed the resulting point estimates for mixing matrix and noise covariance as well as the source distributions, which are shown componentwise as mean and standard deviation.

The model assumes a fixed number k of source components. We determined the optimal number of components using the BIC.[39] It is here defined as BIC = pL −(1/2)(nk +1) log(p), where L represents the log likelihood of the fitted ICA model. We chose k with maximal BIC value.

The information content of each IC was assessed by means of kurtosis, that is, the fourth standardized moment. The kurtosis β_i of each IC$_i$ is defined as

$$\beta_i = \frac{\frac{1}{p}\sum_{j=1}^{p}\left(\mathbf{S}_{ij} - \overline{\mathbf{S}}_{i\cdot}\right)^4}{\left[\frac{1}{p}\sum_{j=1}^{p}(\mathbf{S}_{ij} - \overline{\mathbf{S}}_{i\cdot})^2\right]^2}$$

where p is the number of metabolites (i.e., the number of columns in **S**) and $\overline{\mathbf{S}}_{i\cdot}$ denotes the average value of IC i.

2.3. Weighted Enrichment Analysis

Let p again be the number of metabolites in our data set and c be the number of distinct class annotations. We investigate the class enrichment in a vector **w** of non-negative weights: $w_i \in \mathbb{R}_+$, for each metabolite $i = 1, ..., p$. Class assignments are specified in the Boolean matrix $\mathbf{B} = (b_{ij})$ of dimension $p \times c$ by

$$b_{ij} = \begin{cases} 1, & \text{if metabolite } i \text{ belongs to class } j \\ 0, & \text{else} \end{cases}$$

We now compute the class enrichment vector **e** of dimension c as $\mathbf{e} = \mathbf{B}\cdot\mathbf{w} \in \mathbb{R}^c$, that is, for each class, we simply sum up the contributions of all metabolites that belong to that specific class.

The values in **e** have no properly defined scale and can thus not be directly interpreted. Instead, we randomly shuffle the metabolite-class associations $r = 10^7$ times and recalculate a randomized vector $\mathbf{e}_r$. Let **f** contain the number of randomized values among all sampled $\mathbf{e}_r$ that are larger than the respective elements in **e**. We compute the empirical p value vector of length c as $\mathbf{p} = \mathbf{f}/r$. The result vector **p** thus contains one empirical p value for the enrichment of each class in **w**.

2.4. PCA, *k*-means, and Fuzzy *c*-means Clustering

PCA represents a standard multivariate data analysis procedure reviewed, for instance, in Shlens.[14] Briefly, similar to ICA, PCA represents a mixture model, where the data matrix **X** is split into two matrices **A** and **S** such that $\mathbf{X} = \mathbf{A}\cdot\mathbf{S}$. In contrast to ICA, **S** is here chosen such that all components are decorrelated, that is, cov $(\mathbf{S}^T) = 0$. For k-means and fuzzy c-means clustering, we used the MATLAB-integrated functions *kmeans* and *fcm*, respectively. As a second variant of the fuzzy c-means approach, we only set the highest value of each metabolite in the fuzzy clustering matrix to 1 and the rest to 0 (thus again creating a hard clustering as produced by k-means). For all methods but ICA, we logarithmized and subsequently column-normalized the data matrix.

2.5. Regression Analysis

Associations between the HDL values and the component strength vectors (columns) of the mixing matrix as well all metabolites were estimated using linear regression analysis. Before performing the actual analysis, we removed from the data (i) age effects by only taking the residuals from a linear regression of the mixing matrix and the metabolite matrix columns on age and (ii) gender-specific effects by subtracting the group-wise medians from each column in the data. We then regressed the HDL values on both the mixing matrix columns and each metabolite using the MATLAB regress function. P values were obtained from the t distribution with studentized residuals, and the explained variance is determined by the coefficient of determination R^2. For the linear model forward feature selection algorithm based on AIC (Akaike information criterion), we used the R platform function step with setting direction='forward'.

3. RESULTS

3.1. Bayesian Noisy ICA on Metabolomics Data

For data preprocessing, we normalized each column in the data matrix (1764 probands, 218 metabolites) to a standard deviation of 1 and subsequently scaled the values between 0 and 1. The following ICA calculations are based on the Bayesian mean-field ICA approach described in Højen-Sørensen et al.[35] We assumed a nonnegativity prior for **A**, an exponential distribution (and thus positive values) for **S**, and an isotropic noise model for ϵ_{ij}. To determine the number of components k to be used, we calculated the BIC for k = 2 up to k = 30 components, with 100 random initial conditions (Figure 2, showing the first 20 components). The diagram demonstrates (i) proper convergence of the algorithm due to similar BIC values in multiple runs for each k and (ii) a clear BIC peak around 7−10 components. The highest score in the analysis was achieved for one run at k = 8, so we chose this number of components for all subsequent analysis steps. For higher numbers of k, the increase in reconstruction quality was not sufficient to compensate for the penalty imposed due to more parameters in the model. To verify the stability of the choice of k with respect to changes in the underlying data set, we employed a sample bootstrapping approach. This robustness analysis did not reveal significant differences to the full data set run. Both the detailed regular analysis with 30 components and the bootstrapping results can be found in the Supporting Information, S1.

The resulting matrices **S** (with estimated parameter variance) and **A** are visualized in Figures 3 and 7, respectively, and will be

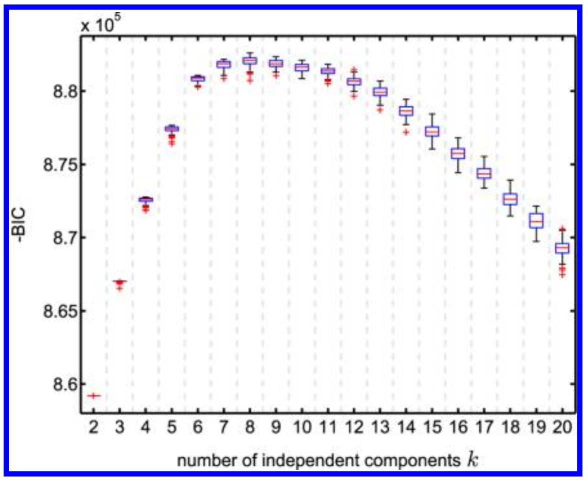


Figure 2. Selection of the number of components. The BIC of the ICA model was estimated according to Hoejen-Soerensen et al.[35] for a range of k values, with 100 random initial value conditions for each k. We observe a clear peak around 7−10 components and choose k = 8 for all subsequent analyses.

subject to detailed functional analyses in the following sections. Detailed values along with standard deviations for **S** can be found in the Supporting Information, S2.

3.2. Manual Investigation of ICs in S

While the separation of the metabolomics data set into eight ICs might be sound from a statistical point of view, we have to ask whether we can gain insights into metabolic processes underneath giving rise to the data. Each component consists of a vector $\mathbf{s}_{l.}$ of non-negative contribution strengths, that is, one value for each metabolite (Figure 3). To get an overview of the metabolic functions in which the components might be involved, we manually investigated the 15 strongest contributions for each component (Figure 4). Estimation certainty is generally high, as indicated by small error bars resulting from the probabilistic ICA approach. Functionally, we observe prominent metabolites from each IC to be biologically related. The following paragraph briefly describes each of the eight reconstructed ICs with respect to biochemical characteristics of the top-scoring metabolites.

IC_1 primarily contains amino acids and related substances. Among the top-scoring metabolites in this component are amino acids containing functional amine groups, like glutamine, histidine, arginine, and carnitine, as well as several aromatic compounds, including tryptophan and phenylalanine. The strongest metabolites in IC_2 are again primarily amino acids. We observe phenylalanine and tryptophan in the top-scoring compound list and, in particular, various branched-chain amino acids. Valine, leucine, and isoleucine constitute high contributions but also their direct degradation products 3-methyl-2-oxobutyrate, 4-methyl-2-oxopentanoate, and 3-methyl-2-oxovalerate, respectively. IC_3 exclusively contains long chain fatty acids comprising 12−20 carbon atoms among its 15 strongest metabolites. This includes fatty acids with both even numbers of carbon atoms as well as a few odd-numbered fatty acids and various levels of desaturation (i.e., number of double bonds). IC_4 represents a rather heterogeneous set of fatty acid-based lipids. These include short and medium chain fatty acids, hydroxy fatty acids, two polyunsaturated fatty acids (arachidonate and dihomolineolate), and several phospatidylinositols. IC_5 contains as its strongest entries several metabolites involved in energy homeostatic processes. This includes phosphate and acetylphosphate, lactate, and pyruvate but also carbohydrates like glucose and mannose. IC_6 contains both signals from amino acids (including glutamine, tryptophan, phenylalanine, isoleucine, valine, and proline) and from lipid metabolism including phosphatidylethanolamines and medium chain fatty acids. IC_7 also constitutes a rather mixed component with metabolites from tryptophan metabolism (glycosyltryptophane, kynurenin, and 3-indoxylsulfate), nucleotide-related substances (pseudouridine, N1-methyladenosine), carbohydrates (myo-inositol, erythronate, and erythritol), and others. Finally, IC_8 primarily represents the phosphatidylcholine (PC) lipid class, particularly lyso-PCs with a single fatty acid residue bound to either the sn-1 or the sn-2 position of the glycerol backbone. Fatty acid side chains vary from medium chain saturated 14:0 up to polyunsaturated fatty acid residues 20:4. Taken together, these results suggest that each metabolomics profile represents a mixture of statistically independent signals, each of which corresponds to a distinct part in cellular metabolism.

3.3. Systematic Analysis and Statistical Enrichment

Motivated by the findings of our manual investigation, we next asked the question whether this signal can be systematically verified. More specifically, we evaluated whether the reconstructed ICs indeed represent distinct subparts of cellular metabolism. For this purpose, we designed a weighted class enrichment algorithm. Regular hypergeometric enrichment

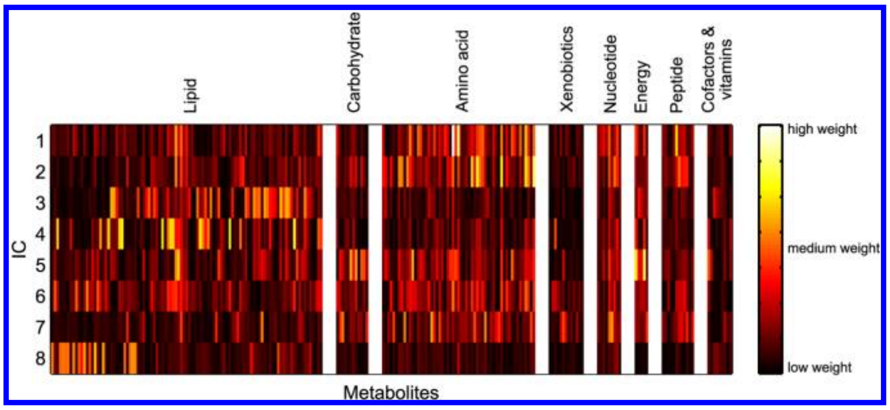


Figure 3. Source matrix **S**, grouped by the eight metabolic *superpathways* in our data set. Rows are pairwise statistically independent and contain the contributions of all metabolites to the respective component. Already from this visual inspection, we can see enrichments for specific pathways in each component, e.g., *Amino acid* in IC_1 and IC_2 and *Lipid* in IC_4 and IC_8.

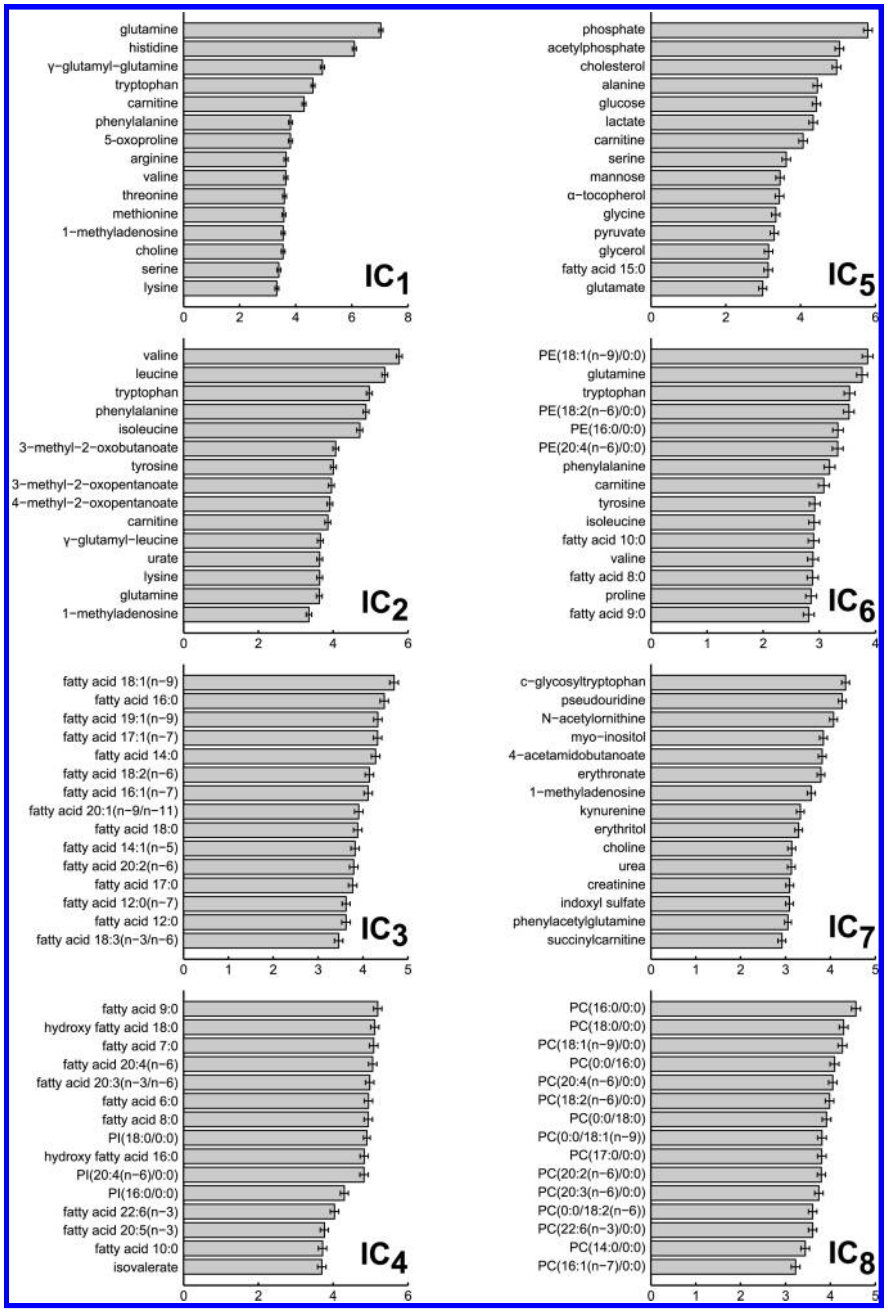


Figure 4. Top 15 metabolite contributions for each IC in **S**. For most components, we observe strong tendencies toward specific parts of cellular metabolism. For instance, IC_2 contains branched-chain amino acids and their degradation product among its highest contributing metabolites. IC_8 contains phosphatidylcholines for various chain lengths and desaturation grades, and so on. Error bars indicate standard deviations from the estimation algorithm. For a detailed description of lipid naming conventions, see the Materials and Methods.

tests like *gene set enrichment analysis* (GSEA)[40] and *metabolite set enrichment analysis* (MSEA)[41] analyze discrete yes/no assignments of each analyzed item (metabolite in our case) to one or more classes. Our approach, in contrast, takes into account the weight of each item in the group (in our case the contribution of each metabolite to each IC) to calculate the corresponding enrichment. For a formal description of the algorithm, see the Materials and Methods.

For each measured metabolite, we have annotations for *superpathway* and *subpathway*, representing two different granularities of metabolic pathway assignments (see the Materials and Methods). In the following analysis, we first determined whether each IC significantly enriches metabolites from one of the superpathways ($p \leq 0.01$). For each enriched superpathway, we then investigated whether the component also enriches one of the subpathways (Table 1). Further confirming the manual analysis, we observe strong enrichments for amino acids, lipids, and energy metabolism. In particular, ICs separate histidine, branched-chain amino acid (valine, leucine, and isoleucine) and tryptophan-related processes in the amino acid superpathway class. For the lipid class, we observe two mixed components involving various types of fatty acids as well as a third, glycerolipid-centered component. The energy-related component splits into oxidative phosphorylation and central carbon metabolism (glycolysis, gluconeogenesis, and pyruvate metabolism).

Table 1. Statistical Enrichment of Metabolic Pathways in the ICs[a]

	superpathway	p	subpathway	p
IC_1	amino acid	3.0×10^{-7}	histidine metabolism	4.6×10^{-3}
IC_2	amino acid	$<1.0 \times 10^{-7}$	valine, leucine, and isoleucine metabolism	8.0×10^{-7}
IC_6	amino acid	4.0×10^{-3}	valine, leucine, and isoleucine metabolism	3.5×10^{-3}
IC_7	amino acid	5.4×10^{-4}	tryptophan metabolism	4.0×10^{-3}
IC_3	lipid	$<1.0 \times 10^{-7}$	fatty acid, saturated, even	2.3×10^{-4}
			fatty acid, monoene	4.0×10^{-7}
			fatty acid, monoene, odd	4.3×10^{-4}
			fatty acid, polyene	6.6×10^{-4}
			carnitine metabolism	7.1×10^{-3}
IC_4	lipid	3.9×10^{-5}	fatty acid, saturated, even	2.0×10^{-3}
			fatty acid, saturated, odd	7.2×10^{-5}
			fatty acid, polyene	1.2×10^{-4}
			fatty acid, saturated, monohydroxy	1.0×10^{-3}
IC_8	lipid	$<1.0 \times 10^{-7}$	glycerolipid metabolism	$<1.0 \times 10^{-7}$
IC_5	energy	2.0×10^{-4}	oxidative phosphorylation	$<1.0 \times 10^{-7}$
	carbohydrate	2.4×10^{-3}	glycolysis, gluconeogenesis, pyruvate metabolism	1.5×10^{-3}

[a] We employed a weighted enrichment test that makes use of the actual contributions of each metabolite in the ICs (see main text). As suggested by our manual investigation, we find strong enrichment for different parts of metabolism, e.g., amino acid pathways, lipid-specific pathways, and energy-related processes. Interestingly, except for a few overlaps, each IC specifically enriches a distinct major pathway.

We compared the weighted enrichment algorithm with hypergeometric enrichment as used in GSEA and MSEA. The weighted approach displays a slightly higher sensitivity for the detection of enriched pathways, but the results of weighted and hypergeometric enrichment are generally comparable (Supporting Information, S3). Importantly, however, hypergeometric enrichment requires a hard yes/no assignment of metabolites to each component, that is, whether it can be considered "present" in the component or not. This introduces an additional cutoff parameter that needs to be defined before the analysis. Weighted enrichment, on the other hand, works parameter-free and directly uses the actual strength of each metabolite in the components.

We furthermore complemented the functional enrichment analysis from an information theoretical point of view, by inspecting the information content in each IC. ICA seeks for maximal non-Gaussianity, a feature commonly measured by the fourth central distribution moment (*kurtosis*). Decreasingly ordered kurtosis values for all eight components are displayed in Figure 5. Interestingly, the two components containing the least amount of information, namely, IC_6 and IC_3, are those that displayed a significant overlap in functional enrichment with other components (IC_2 and IC_4, respectively). This indicates that kurtosis can be used to sort out components containing rather little biological information, an approach that has been employed in previous studies already.[30,31] On the other hand, components displaying significant, distinct associations with biological processes also contain a high amount of information (e.g., IC_8 and IC_1). This finding establishes an appealing bridge between the statistical information content in the reconstructed components and the biological information content encoded therein.

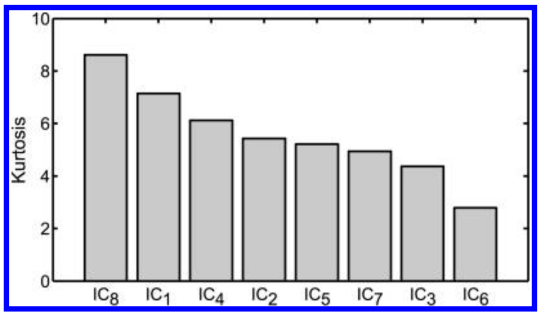


Figure 5. Kurtosis as a measure of information content for each IC. Remarkably, those components with a high information content also tend to display strong functional enrichment of a metabolic pathway.

3.4. Comparison with PCA and *k*-means Clustering

To get an objective view of the quality of our ICA approach, we compared the weighted enrichment results obtained using Bayesian ICA with commonly used data analysis techniques. We ran the enrichment calculations on the results of PCA and *k*-means clustering with the same number of components (or clusters); see Figure 6. Furthermore, we introduce the concept of *consistent* and *inconsistent* subpathway enrichments. The enrichment of a subpathway is considered inconsistent, if the superpathway that this subpathway belongs to is not enriched in the same component. For ICA, we detect one inconsistent enrichment of the γ-glutamyl peptide pathway for IC_2, which enriches the amino acid superpathway.

PCA yields seven out of eight enriched components, with a total of three distinct enriched superpathways. For the subpathway enrichment, six enrichments can be considered inconsistent since the respective superpathways are not enriched in the same component. Several components display

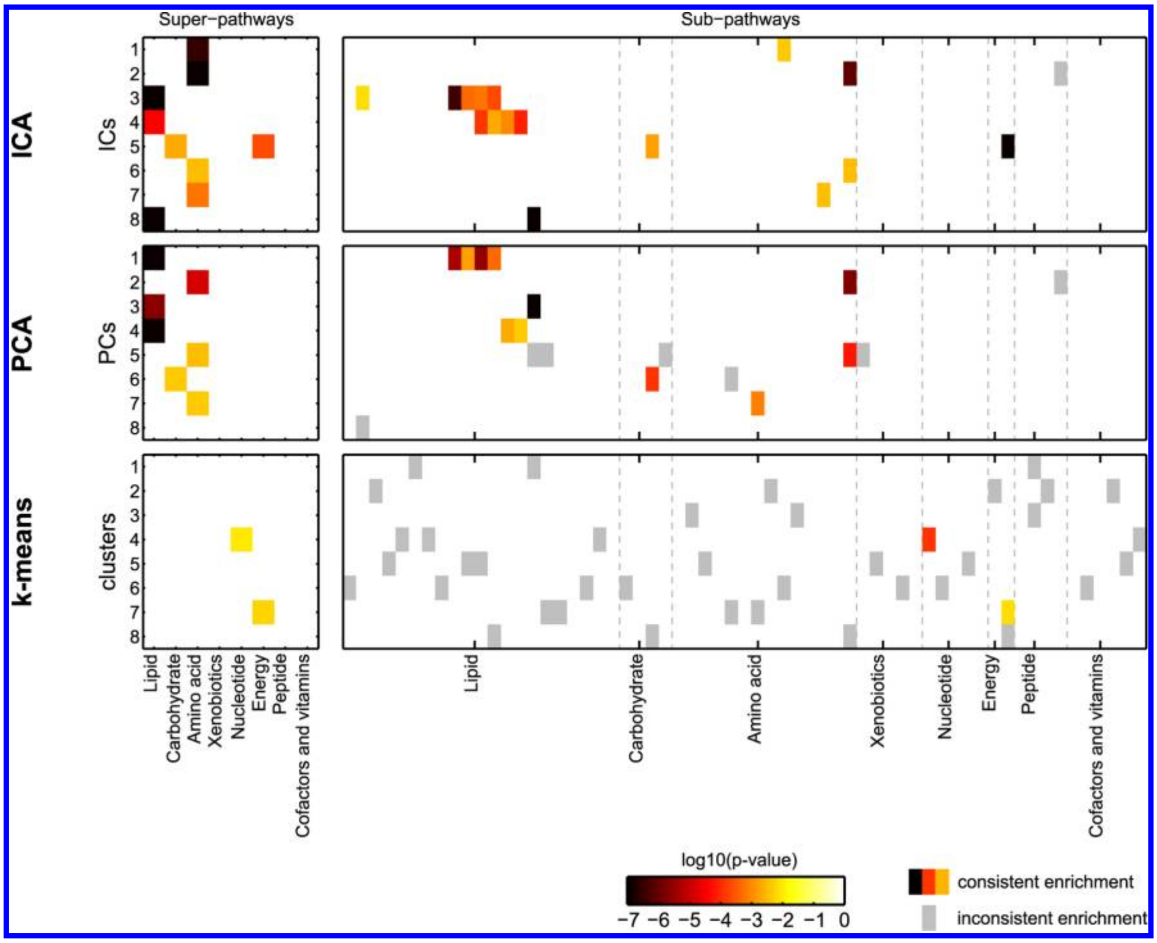


Figure 6. Comparison of pathway enrichment for ICA, PCA, and *k*-means clustering. ICA and PCA produce generally comparable results, but ICA appears more sensitive (enriches more superpathways), more specific (less inconsistent enrichments), and displays lower association *p* values. Note that the components are not comparable in order, e.g., IC_1 does not correspond to PC_1.

similar enrichments as ICs from the ICA. Specifically, IC_2/PC_5 as well as IC_6/PC_2 enrich branched-chain amino acids, IC_3/PC_1 as well as IC_4/PC_4 show specific fatty acid pathway enrichments, IC_5/PC_6 enrich the glycolysis pathway, and finally IC_8/PC_3 enrich the glycerolipids. PCA does not detect enrichments of histdine metabolism (IC_1), oxidative phosphorylation (IC_5), and tryptophane metabolism (IC_7). Furthermore, *p*-values for PCA enrichment are generally higher in comparison to ICA (colors in Figure 6), for example, with three out of seven enriched superpathways, which are only borderline significant. *k*-means clustering produces a substantial number of enrichments for subpathways that are mostly inconsistent. In other words, *k*-means recovers parts of the metabolism, which, however, do not belong to the same superpathway and cannot be considered as specific metabolic signals.

To further compare ICA with a regular clustering algorithm that supports weighted cluster assignments, we applied fuzzy *c*-means clustering. The analysis produced no significantly enriched clusters with respect to the superpathways and only few enriched subpathways. Finally, *c*-means clustering with subsequent selection of the clusters displaying the highest contribution for each metabolite (see the Materials and Methods) yields similar results as the *k*-means approach. Detailed enrichment results of Bayesian ICA, PCA, *k*-means, and the two variant of *c*-means clustering are collected in Supporting Information, S4.

3.6. Analyzing the Mixing Matrix A—Associations with HDL

Up to this point, we have demonstrated that to a certain extent, metabolomics profiles may be interpreted as a mixture of independent processes from different parts of the metabolic pathways. We next sought to investigate whether the mixing matrix **A** contains biologically interesting information as well. Recall that **A** gives us another eight variables for each sample (proband in the study cohort) in addition to the metabolite concentrations. These eight variables encode how strong each IC, that is, each recovered biological process, contributes to the respective metabolite profile. As can be seen in the clustering displayed in Figure 7, the IC weights certainly contain proband-specific information suitable for further analysis. The question now is how to determine whether these weights represent biologically meaningful descriptors. A straightforward approach is to correlate the columns of **A** with other, sample-specific parameters and measurements (Figure 1C). One such example is provided in a transcriptomics ICA study by Schachtner et al.,[27] where the mixing matrix columns were compared with so-called *design vectors*—which essentially encode the different conditions in which cells in that particular study were cultured.

We here chose blood plasma HDL levels, which represent a complex quantitative trait influenced by a variety of metabolic and physiological parameters.[42] HDL belongs to the class of lipoproteins, small particles circulating in the blood responsible for the transport of insoluble lipids through the body. We

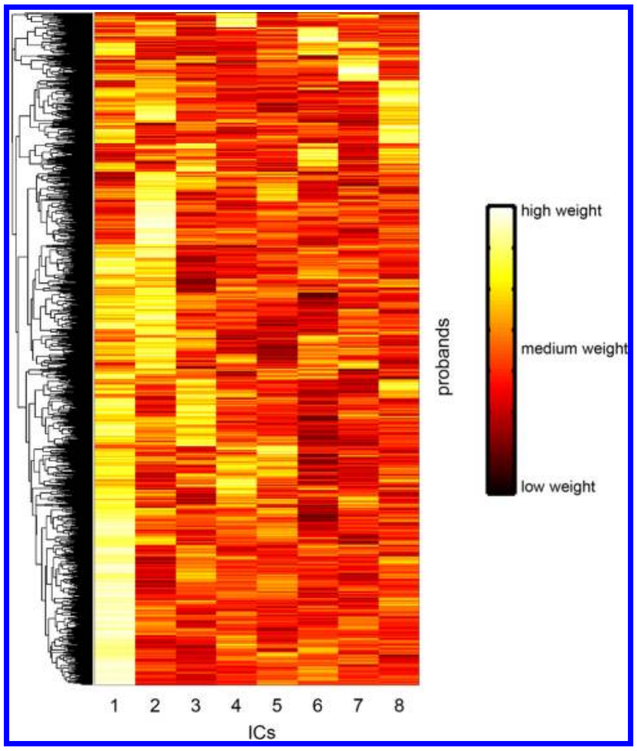


Figure 7. Mixing matrix **A**. Rows represent the strengths of each IC's contribution to the respective proband metabolome. The hierarchical clustering in the proband direction demonstrates the presence of clear-cut groups reconstructed from the ICA. Each column in the matrix is then subjected to correlation with plasma HDL levels in the next step.

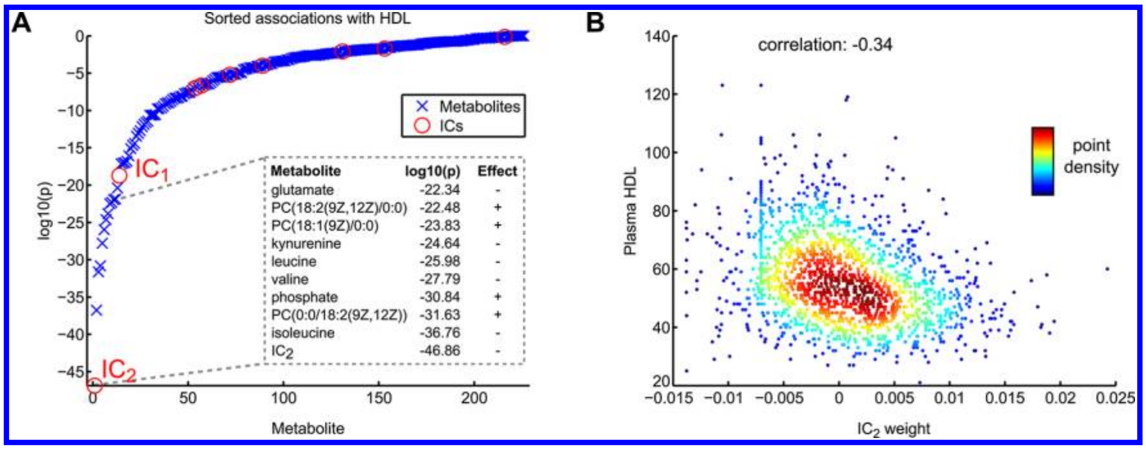
conducted a linear regression analysis of both metabolites and IC strengths against HDL levels and corrected for gender and age effects (Figure 8A). Detailed results are collected in Supporting Information, S5. Associations with HDL are generally high throughout the data set, with 88 out of 218 metabolites and five out of eight ICs displaying statistically significant associations ($\alpha = 0.05$ after Bonferroni correction). Two ICs, IC_2 and IC_1, show profound signals with p values below 10^{-17}. Remarkably, IC_2 even constitutes the strongest association throughout all analyzed variables. As described above, IC_2 primarily contains signatures of the three branched-chain amino acids valine, leucine, and isoleucine as well as their respective degradation products.

We detect a *negative* effect on plasma HDL levels for both branched-chain amino acids alone, and for the IC_2 contribution strength (Effect's column in Figure 8.8A, and Figure 8.8B). This means that a stronger contribution of this component, and thus higher values of the involved metabolites, coincides with lower values of HDL. This finding represents a novel connection between branched-chain amino acids and blood plasma HDL levels (see Discussion). For comparison, we performed the HDL comparison with loadings from PCA instead of ICA. The branched-chain amino acid principal component displays a profoundly weaker association with HDL than IC_2 ($p = 3.28 \times 10^{-5}$). The strongest association of a principal component with HDL ranks number 20 in the sorted association list. Detailed results can be found in the Supporting Information, S5.

To get an additional comparison with common regression-based approaches, we generated a linear model with multiple metabolite predictor variables. To this extent, we ran a forward feature selection approach based on AIC (Akaike information criterion, see the Materials and Methods). The results of this analysis can be found in the Supporting Information, S5. Interestingly, when ordering the metabolites by their importance for the overall model performance, isoleucine is the only branched-chain amino acid-related metabolite appearing among the top hits. This is an effect of high correlations between metabolites: Once isoleucine is added to the model, the other branched-chain amino acid compounds cannot improve model performance any further. Hence, while such a multipredictor linear regression model might produce a reasonably good description of HDL levels, the interpretation of metabolites with high weights in this model might be misleading.

4. DISCUSSION

In this paper, we evaluated a Bayesian ICA approach as a tool for the investigation of a population-based metabolomics data set containing 1764 probands and 218 metabolites. The Bayesian framework provides several advantages over a regular ICA: (1) We can implement distribution priors (a non-negativity constraint in our case) to construct a biologically meaningful factorization of the data matrix. (2) Because we get



Metabolite	log10(p)	Effect
glutamate	-22.34	-
PC(18:2(9Z,12Z)/0:0)	-22.48	+
PC(18:1(9Z)/0:0)	-23.83	+
kynurenine	-24.64	-
leucine	-25.98	-
valine	-27.79	-
phosphate	-30.84	+
PC(0:0/18:2(9Z,12Z))	-31.63	+
isoleucine	-36.76	-
IC_2	-46.86	-

Figure 8. Linear regression of plasma HDL levels on metabolite levels and IC contributions, corrected for gender and age effects. (A) The strongest association of all variables is constituted by IC_2, followed by the branched-chain amino acids, other amino acids, and several phosphatidylcholines. (B) Negative correlation between the plasma HDL and the contribution strength of IC_2 (which primarily contains contributions from branched-chain amino acids). Note that negative values for the ICA occur due to the correction for gender and age.

distributions of fitted parameters, we obtain information on the estimation certainty for each entry in **S**. (3) Using a BIC-based model selection approach, we can automatically determine the number of components to be reconstructed from the data.

We evaluated the source matrix **S** of statistically independent metabolite profiles from a biological point of view and demonstrated strong enrichment of distinct metabolic pathways in the reconstructed components. This implies that the human blood metabolome represents a mixture of overlaying, statistically independent signals, each of which can be attributed to a specific set of metabolic pathways. While this concept is quite similar to the idea of *eigengenes* and *eigenmetabolites*,[43] our approach extends the standard ICA approach by a Bayesian, noisy framework, which allows for the estimation of confidence intervals for the reconstructed values.

The results obtained from the investigation of **S** are in general accordance with previously published results on Gaussian graphical models (GGMs) of metabolomics data.[15,44] While GGMs only evaluate pairwise associations instead of whole groups as in the ICA approach, the recovery of functionally related metabolites from blood plasma metabolomics samples is similar for both approaches. This fosters the idea of an actual *snapshot* of an organism's metabolism in the blood, rather than mere signatures of transportation and disposal processes in this biofluid.

Correlating the columns of the mixing matrix **A** with plasma HDL levels, we detected a possibly novel association between branched-chain amino acids and HDL blood plasma levels. HDL represents a complex, heterogeneous phenotype that is still poorly understood and associated with a variety of biological processes.[45,46] The metabolic process encoded by IC_2 in our study now adds an additional piece of functional information for the interpretation of plasma HDL. Interestingly, both HDL levels and branched-chain amino acids are well-known to be strongly connected with obesity, insulin resistance, and diabetes type II. On the one hand, branched-chain amino acid levels are altered as a direct consequence of changed insulin sensitivity and have been shown to be markers for the prediction of future diabetes type II.[47,48] Furthermore, leucine is known to directly interact on a cellular level with the insulin signaling cascade.[49] On the other hand, the pathological phenotype is known to lower HDL blood plasma levels, a condition that severely increases the risk for cardiovascular disease.[50] Using cross-sectional metabolomics data from a population cohort, we could now establish the additional association between branched-chain amino acids and HDL, irrespective of a diabetic phenotype. Interestingly, we could recover this association despite the unsupervised approach taken by ICA. In other words, IC_2 has not been specifically tailored to explain HDL levels but rather seems to reflect an intrinsic metabolic process around branched-chain amino acids that strongly associates with HDL. The only (biologically motivatable) assumption going into the ICA model is the independence of metabolite profiles to hold throughout all samples in the data.

We systematically compared the ICA results with commonly used multivariate data analysis methods like PCA and *k*-means clustering. The comparison with PCA was of particular interest here, since it is widely used for metabolomics data and, similar to ICA, also represent a linear mixture model separating the data matrix into a source and a mixing matrix. While PCA produced a series of enriched components with direct IC counterparts, ICA appeared to be more sensitive. Specifically, ICA enrichments were generally stronger in comparison to PCA and detected several pathway enrichments that could not be observed for PCA. Moreover, our findings from the HDL analysis could not be reproduced in the PCA approach. These results could be due to the rather arbitrary constraint of orthogonal basis vectors in PCA, which can hardly be biologically motivated. The notion of statistically independent processes acting in the system, as recovered by the ICA, can directly be interpreted in the context of a metabolic system.

Taken together, Bayesian ICA on metabolomics data can be used both to reconstruct meaningful metabolic profiles, which underly the measured concentrations, and to detect novel relationships with complex phenotypic traits like plasma HDL levels.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Figures and tables of source matrix mean values from mean-field ICA calculation, weighted enrichment of metabolic pathways in each reconstructed component, forward variable selection based on AIC ("step" function in R computing platform), and metabolite superpathways and subpathways. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: fabian.theis@helmholtz-muenchen.de.

Author Contributions

J.K. and F.J.T. conceived this data analysis project. K.S., T.I., and J.A. performed the sample preparation and data acquirement. J.K. performed the computational analysis. J.K. and F.J.T. wrote the primary manuscript. All authors approved the final manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.), by the European Research Council (starting grant "LatentCauses"), by BMBF Grant No. 0315494A (project "SysMBo"), and by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (project "CoReNe"). J.K. is supported by a Ph.D. student fellowship from the "Studienstiftung des Deutschen Volkes". We thank the anonymous reviewers for valuable comments and suggestions to improve the original manuscript.

■ REFERENCES

(1) Griffin, J. L. The cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos. Trans. R. Soc.,B* **2006**, *361* (1465), 147−161.

(2) Kaddurah-Daouk, R.; Kristal, B. S.; Weinshilboum, R. M. Metabolomics: A global biochemical approach to drug response and disease. *Annu. Rev. Pharmacol. Toxicol.* **2008**, *48*, 653−683.

(3) Suhre, K.; Meisinger, C.; Döring, 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.

(4) Hu, F. B. Metabolic profiling of diabetes: From black-box epidemiology to systems epidemiology. *Clin. Chem.* **2011**, *57* (9), 1224−1226.

(5) Fav, G.; Beckmann, M. E.; Draper, J. H.; Mathers, J. C. Measurement of dietary exposure: A challenging problem which may be overcome thanks to metabolomics? *Genes Nutr.* **2009**, *4* (2), 135−141.

(6) Bondia-Pons, I.; Nordlund, E.; Mattila, I.; Katina, K.; Aura, A. M.; Kolehmainen, M.; Oresic, M.; Mykkanen, H.; Poutanen, K. Postprandial differences in the plasma metabolome of healthy finnish subjects after intake of a sourdough fermented endosperm rye bread versus white wheat bread. *Nutr. J.* **2011**, *10* (1), 116.

(7) Fendt, S. M.; Buescher, J. M.; Rudroff, F.; Picotti, P.; Zamboni, N.; Sauer, U. Tradeoff between enzyme and metabolite efficiency maintains metabolic homeostasis upon perturbations in enzyme capacity. *Mol. Syst. Biol.* **2010**, *6*, 356.

(8) Heiden, M. G. V. Targeting cancer metabolism: a therapeutic window opens. *Nat. Rev. Drug Discovery* **2011**, *10* (9), 671−684.

(9) Altmaier, E.; Ramsay, S. L.; Graber, A.; Mewes, H. W.; Weinberger, K. M.; Suhre, K. Bioinformatics analysis of targeted metabolomics−uncovering old and new tales of diabetic mice under medication. *Endocrinology* **2008**, *149* (7), 3478−3489.

(10) Sreekumar, A.; Poisson, L. M.; Rajendiran, T. M.; Khan, A. P.; Cao, Q.; Yu, J.; Laxman, B.; Mehra, R.; Lonigro, R. J.; Li, Y.; Nyati, M. K.; Ahsan, A.; Kalyana-Sundaram, S.; Han, B.; Cao, X.; Byun, J.; Omenn, G. S.; Ghosh, D.; Pennathur, S.; Alexander, D. C.; Berger, A.; Shuster, J. R.; Wei, J. T.; Varambally, S.; Beecher, C.; Chinnaiyan, A. M. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **2009**, *457* (7231), 910−914.

(11) Huffman, K. M.; Shah, S. H.; Stevens, R. D.; Bain, J. R.; Muehlbauer, M.; Slentz, C. A.; Tanner, C. J.; Kuchibhatla, M.; Houmard, J. A.; Newgard, C. B.; Kraus, W. E. Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women. *Diabetes Care* **2009**, *32* (9), 1678−1683.

(12) Johansen, K. K.; Wang, L.; Aasly, J. O.; White, L. R.; Matson, W. R.; Henchcliffe, C.; Beal, M. F.; Bogdanov, M. Metabolomic profiling in lrrk2-related parkinson's disease. *PLoS One* **2009**, *4* (10), e7551.

(13) Oresic, M.; Hyotylainen, T.; Herukka, S. K.; Sysi-Aho, M.; Mattila, I.; Seppanan-Laakso, T.; Julkunen, V.; Gopalacharyulu, P. V.; Hallikainen, M.; Koikkalainen, J.; Kivipelto, M.; Helisalmi, S.; Lotjonen, J.; Soininen, H. Metabolome in progression to Alzheimer's disease. *Transl. Psychiatry* **2011**, *1*, e57.

(14) Shlens, J. *A Tutorial on Principal Component Analysis*; Systems Neurobiology Laboratory, Salk Institute for Biological Studies: La Jolla, CA, 2005.

(15) Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **2011**, *5* (1), 21.

(16) Hyvärinen, A.; Karhunen, J.; Oja, E. Adaptive and learning systems for signal processing, communications, and control. *Independent Component Analysis*; J. Wiley: New York, 2001.

(17) Comon, P. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287−314.

(18) Theis, F. Uniqueness of real and complex linear independent component analysis revisited. *Proc. European Signal Processing Conference (EUSIPCO)*; Vienna, Austria, 2004; pp 1705−1708.

(19) Makeig, S.; Bell, A. J.; Jung, T. P.; Sejnowski, T. J. Independent Component Analysis of Electroencephalographic Data. In *Advances in Neural Information Processing Systems*; Touretzky, D. S., Mozer, M. C., Hasselmo, M. E., Eds.; The MIT Press: Cambridge, MA, 1996; Vol. *8*, pp 145−151.

(20) Mckeown, M. J.; Makeig, S.; Brown, G. G.; Jung, T. P.; Kindermann, S. S.; Kindermann, R. S.; Bell, A. J.; Sejnowski, T. J. Analysis of fmri data by blind separation into independent spatial components. *Hum. Brain Mapping* **1998**, *6*, 160−188.

(21) Karvanen, J.; Theis, F. J. Spatial ica of fmri data in time windows. *Proceedings: Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop*, Garching, Germany, 25−30 July 2004; American Institute of Physics: Melville, NY, 2004; Vol. *735* of AIP conference proceedings, pp 312−319.

(22) Keck, I. R.; Theis, F. J.; Gruber, P.; Lang, E.; Specht, K.; Puntonet, C. G. 3d spatial analysis of fmri data on a word perception task. In *Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22−24, 2004 Proceedings*; Puntonet, C. G., Ed.; Springer: Berlin, 2004; Vol. *3195* of Lecture Notes in Computer Science, pp 977−984.

(23) Zhang, X. W.; Yap, Y. L.; Wei, D.; Chen, F.; Danchin, A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur. J. Hum. Genet.* **2005**, *13* (12), 1303−1311.

(24) Huang, D. S.; Zheng, C. H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **2006**, *22* (15), 1855−1862.

(25) Teschendorff, A. E.; Journée, M.; Absil, P. A.; Sepulchre, R.; Caldas, C Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* **2007**, *3* (8), e161.

(26) Lutter, D.; Ugocsai, P.; Grandl, M.; Orso, E.; Theis, F.; Lang, E. W.; Schmitz, G. Analyzing m-csf dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. *BMC Bioinformatics* **2008**, *9*, 100.

(27) Schachtner, R.; Lutter, D.; Knollmüller, P.; Tomé, A. M.; Theis, F. J.; Schmitz, G.; Stetter, M.; Vilda, P. G.; Lang, E. W. Knowledge-based gene expression classification via matrix factorization. *Bioinformatics* **2008**, *24* (15), 1688−1697.

(28) Hofmann, J.; Ashry, A. E. N. E.; Anwar, S.; Erban, A.; Kopka, J.; Grundler, F. Metabolic profiling reveals local and systemic responses of host plants to nematode parasitism. *Plant J.* **2010**, *62* (6), 1058−1071.

(29) Führs, H.; Götze, S.; Specht, A.; Erban, A.; Gallien, S.; Heintz, D.; Dorsselaer, A. V.; Kopka, J.; Braun, H. P.; Horst, W. J. Characterization of leaf apoplastic peroxidases and metabolites in vigna unguiculata in response to toxic manganese supply and silicon. *J. Exp. Bot.* **2009**, *60* (6), 1663−1678.

(30) Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite fingerprinting: Detecting biological features by independent component analysis. *Bioinformatics* **2004**, *20* (15), 2447−2454.

(31) Wienkoop, S.; Morgenthal, K.; Wolschin, F.; Scholz, M.; Selbig, J.; Weckwerth, W. Integration of metabolomic and proteomic phenotypes: Analysis of data covariance dissects starch and rfo metabolism from low and high temperature compensation response in arabidopsis thaliana. *Mol. Cell Proteomics* **2008**, *7* (9), 1725−1736.

(32) Mtin, F. P. J.; Rezzi, S. I. M.; Philippe, D.; Tornier, L.; Messlik, A.; Hollzlwimmer, G.; Baur, P.; Quintanilla-Fend, L.; Loh, G.; Blaut, M.; Blum, S.; Kochhar, S.; Haller, D. Metabolic assessment of gradual development of moderate experimental colitis in il-10 deficient mice. *J. Proteome Res.* **2009**, *8* (5), 2376−2387.

(33) Himberg, J.; Hyvärinen, A.; Esposito, F. Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* **2004**, *22* (3), 1214−1222.

(34) Keck, I.; Theis, F.; Gruber, P.; Lang, E.; Specht, K.; Fink, G.; Tomé, A.; Puntonet, C. Automated clustering of ICA results for fMRI data analysis. *Proc. Computational Intelligence in Medicine and Healthcare (CIMED)*; Lisbon, Portugal, 2005; pp 211−216.

(35) Højen-Sørensen, P. A. R.; Winther, O.; Hansen, L. K. Mean-field approaches to independent component analysis. *Neural Comput.* **2002**, *14* (4), 889−918.

(36) Suhre, K.; Shin, S. Y.; Petersen, A. K.; Mohney, R. P.; Meredith, D.; Wägele, B.; Altmaier, E.; CARDIoGRAM; Deloukas, P.; Erdmann, J.; Grundberg, E.; Hammond, C. J.; de Angelis, M. H.; Kastenmüller, G.; Köttgen, A.; Kronenberg, F.; Mangino, M.; Meisinger, C.; Meitinger, T.; Mewes, H. W.; Milburn, M. V.; Prehn, C.; Raffler, J.; Ried, J. S.; Römisch-Margl, W.; Samani, N. J.; Small, K. S.; Wichmann, H. E.; Zhai, G.; Illig, T.; Spector, T. D.; Adamski, J.; Soranzo, N.;

Gieger, C. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **2011**, *477* (7362), 54−60.

(37) Belouchran, A.; Cardoso, J. F. Maximum likelihood source separation by the expectation-maximization technique: Deterministic and stochastic implementation. *Proc. International Symposium on Nonlinear Theory and its Applications (NOLTA)*, 1995; pp 49−53.

(38) Hansen, L. K. *Advances in Independent Components Analysis*; Springer-Verlag: London, New York, 2000; Chapter: Blind separation of noisy image mixtures, pp 165−187.

(39) Fahrmeir, L.; Kneib, T.; Lang, S. *Regression. Modelle, Methoden und Anwendungen*, 2nd ed.; Springer: Heidelberg, 2009.

(40) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (43), 15545−15550.

(41) Xia, J.; Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using metaboanalyst. *Nat. Protoc.* **2011**, *6* (6), 743−760.

(42) Lusis, A. J.; Pajukanta, P. A treasure trove for lipoprotein biology. *Nat. Genet.* **2008**, *40* (2), 129−130.

(43) DiLeo, M. V.; Strahan, G. D.; den Bakker, M.; Hoekenga, O. A. Weighted correlation network analysis (wgcna) applied to the tomato fruit metabolome. *PLoS ONE* **2011**, *6* (10), e26683.

(44) Mittelstrass, K.; Ried, J. S.; Yu, Z.; Krumsiek, J.; Gieger, C.; Prehn, C.; Roemisch-Margl, W.; Polonikov, A.; Peters, A.; Theis, F. J.; Meitinger, T.; Kronenberg, F.; Weidinger, S.; Wichmann, H. E.; Suhre, K.; Wang-Sattler, R.; Adamski, J.; Illig, T. Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet.* **2011**, *7* (8), e1002215.

(45) Camont, L.; Chapman, M. J.; Kontush, A. Biological activities of hdl subpopulations and their relevance to cardiovascular disease. *Trends Mol. Med.* **2011**, *17* (10), 594−603.

(46) Petersen, A. K.; Stark, K.; Musameh, M. D.; Nelson, C. P.; Römisch-Margl, W.; Kremer, W.; Raffler, J.; Krug, S.; Skurk, T.; Rist, M. J.; Daniel, H.; Hauner, H.; Adamski, J.; Tomaszewski, M.; Döring, A.; Peters, A.; Wichmann, H. E.; Kaess, B. M.; Kalbitzer, H. R.; Huber, F.; Pfahlert, V.; Samani, N. J.; Kronenberg, F.; Dieplinger, H.; Illig, T.; Hengstenberg, C.; Suhre, K.; Gieger, C.; Kastenmüller, G. Genetic associations with lipoprotein subfractions provide information on their biological nature. *Hum. Mol. Genet.* **2012**, *21*, 1433−1443.

(47) Felig, P.; Marliss, E.; Cahill, G. F. Plasma amino acid levels and insulin secretion in obesity. *N. Engl. J. Med.* **1969**, *281* (15), 811−816.

(48) Wang, T. J.; Larson, M. G.; Vasan, R. S.; Cheng, S.; Rhee, E. P.; McCabe, E.; Lewis, G. D.; Fox, C. S.; Jacques, P. F.; Fernandez, C.; O'Donnell, C. J.; Carr, S. A.; Mootha, V. K.; Florez, J. C.; Souza, A.; Melander, O.; Clish, C. B.; Gerszten, R. E. Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **2011**, *17* (4), 448−453.

(49) Layman, D. K.; Walker, D. A. Potential importance of leucine in treatment of obesity and the metabolic syndrome. *J. Nutr.* **2006**, *136* (1 Suppl.), 319S−323S.

(50) Betteridge, D. J. Lipid control in patients with diabetes mellitus. *Nat. Rev. Cardiol.* **2011**, *8* (5), 278−290.