Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data

Jan Krumsiek[1], Nikola Müller[1], Thomas Illig[2], Jerzy Adamski[3], Karsten Suhre[1,4], Fabian J. Theis[1,5]

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München
[2]Institute of Epidemiology, Helmholtz Zentrum München
[3]Institute of Experimental Genetics, Helmholtz Zentrum München
[4]Department of Physiology and Biophysics, Weill Cornell Medical College in Qatar, P.O. Box 24144, Education
[5]Department of Mathematics, Technische Universität München

Metabolomics is a newly arising field aiming at the measurement of all endogenous metabolites of a tissue or body fluid under given conditions. While classical approaches focus on the statistical association of metabolite concentrations with certain phenotypes or clinical traits (like blood pressure, disease states etc.), we here investigate dependencies between the metabolites themselves. If two metabolites are highly correlated, we expect them to share one or more steps in the underlying metabolic pathways.

A major drawback of regular correlations measures, however, is their inability to distinguish between direct and indirect associations. Correlation coefficients are generally high in large-scale *-omics* data sets, suggesting a plethora of indirect and systemic associations. *Gaussian graphical models* (GGMs) circumvent indirect association effects by evaluating *conditional* dependencies in multivariate Gaussian distributions. A GGM is an undirected graph in which each edge represents the pairwise correlation between two variables conditioned against the correlations with all other variables (also denoted as *partial* correlation coefficients). GGMs have a simple interpretation in terms of linear regression techniques. Intuitively speaking, we remove the (linear) effects of all other variables on X and Y and compare the remaining signals. If the variables are still correlated, the correlation is directly determined by the association of X and Y and not mediated by the other variables.

First, we discuss the quality of the method and possible problems and pitfalls on computer-simulated systems. We then apply GGMs to a lipid-focused targeted metabolomics data set of 1020 blood serum samples with 151 measured metabolites from the German population study KORA. The GGM is sparse in comparison to the corresponding Pearson correlation network, displays a modular structure with respect to different metabolite classes, and is stable towards changes in the underlying data set. We demonstrate that top-ranking metabolite pairs and further densely connected subgraphs in the GGM can indeed be attributed to known reactions in the human fatty acid biosynthesis and degradation pathways. In order to systematically verify this finding, we map partial correlation coefficients to the number of reaction steps between all metabolite pairs based on a literature-curated fatty acid pathway model. We observe statistically significant discriminatory features of GGMs to distinguish between directly and non-directly interacting metabolites in the metabolic network. The results have been published in Krumsiek et al., *BMC Systems Biology* 2011, **5:**21.

While the above-mentioned study primarily served as a proof-of-concept for the reconstruction method, we next used these data-driven metabolic networks to aid the interpretation of other analyses. For instance, the metabolomics GGM was employed to elucidate how gender-specific differences spread throughout the metabolic network (Mittelstrass et al., *PLoS Genetics*, 2011. 7(8): e1002215.). Furthermore, we calculated *differential* GGMs in order to detect changes in the correlation structure between treated and untreated glioblastoma cell lines (Müller, Krumsiek et al., *Independent Component Analyses, Wavelets, Neural Networks, Biosystems, and Nanoengineering IX, SPIE,* 2011.)

We believe our work is relevant for the bioinformatics community for both methodological and biological reasons: (1) GGMs have a statistically sound background and can easily be transferred to other types of *-omics* data, like mRNA or proteomics measurements. (2) Detectable footprints of cellular metabolism in high-throughput blood metabolomics data are a non-trivial finding, and the results of this analysis should be considered for the interpretation of metabolomics data in general.

Krumsiek et al. BMC Systems Biology 2011, 5:21
http://www.biomedcentral.com/1752-0509/5/21

BMC Systems Biology

RESEARCH ARTICLE Open Access

Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data

Jan Krumsiek[1], Karsten Suhre[1,2], Thomas Illig[3], Jerzy Adamski[4], Fabian J Theis[1,5*]

Abstract

Background: With the advent of high-throughput targeted metabolic profiling techniques, the question of how to interpret and analyze the resulting vast amount of data becomes more and more important. In this work we address the reconstruction of metabolic reactions from cross-sectional metabolomics data, that is without the requirement for time-resolved measurements or specific system perturbations. Previous studies in this area mainly focused on Pearson correlation coefficients, which however are generally incapable of distinguishing between direct and indirect metabolic interactions.

Results: In our new approach we propose the application of a Gaussian graphical model (GGM), an undirected probabilistic graphical model estimating the conditional dependence between variables. GGMs are based on partial correlation coefficients, that is pairwise Pearson correlation coefficients conditioned against the correlation with all other metabolites. We first demonstrate the general validity of the method and its advantages over regular correlation networks with computer-simulated reaction systems. Then we estimate a GGM on data from a large human population cohort, covering 1020 fasting blood serum samples with 151 quantified metabolites. The GGM is much sparser than the correlation network, shows a modular structure with respect to metabolite classes, and is stable to the choice of samples in the data set. On the example of human fatty acid metabolism, we demonstrate for the first time that high partial correlation coefficients generally correspond to known metabolic reactions. This feature is evaluated both manually by investigating specific pairs of high-scoring metabolites, and then systematically on a literature-curated model of fatty acid synthesis and degradation. Our method detects many known reactions along with possibly novel pathway interactions, representing candidates for further experimental examination.

Conclusions: In summary, we demonstrate strong signatures of intracellular pathways in blood serum data, and provide a valuable tool for the unbiased reconstruction of metabolic reactions from large-scale metabolomics data sets.

Background

Metabolomics is a newly arising field aiming at the measurement of all endogenous metabolites of a tissue or body fluid under given conditions [1-3]. The resulting *metabolome* of a biological system is considered to provide a readout of the integrated response of cellular processes to genetic and environmental factors [4]. Understanding the complex biochemical interplay between hundreds of measured metabolite species is a daunting task, which can be approached by combining advanced computational methods with data from large population-based studies. On the biochemical level, metabolite concentrations are determined by a set of specific metabolic enzymes. Variabilities in both enzyme activity and metabolite exchange rates - induced by a continuous spectrum of metabolic states throughout measured samples - give rise to characteristic patterns in the metabolite profiles which are directly linked the underlying biochemical reaction network [5,6]. Although human metabolism has been extensively characterized in

* Correspondence: fabian.theis@helmholtz-muenchen.de
[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany
Full list of author information is available at the end of the article

the past decades [7], the reconstruction of metabolic networks from such metabolite patterns is a key question in the computational research field. Previous attempts focused on linear metabolite associations measured by Pearson correlation coefficients. These include studies utilizing time-course measurements and clustering [8], theoretical approaches relating metabolite fluctuations to properties of the dynamical system [5] and metabolic control analysis to derive effects of enzyme variability [6]. Other reconstruction methods rely on specific perturbations of the biological system, like the induction of concentration pulses for certain metabolites [9].

A major drawback of correlation networks, however, is their inability to distinguish between direct and indirect associations. Correlation coefficients are generally high in large-scale *omics* data sets, suggesting a plethora of indirect and systemic associations. For example, transcriptional coregulation amongst many genes will give rise to indirect interaction effects in mRNA expression data [10]. Similar effects can be observed in metabolic systems which, in contrast to genetic networks, contain fast biochemical reactions in an open mass- flow system. Metabolite levels are supposed to be in quasi-steady state compared to the time scales of upstream regulatory processes [11]. That is, metabolites will follow changes in gene expression and physiological processes on the order of minutes and hours, but will appear unchanged on the order of seconds. These properties, even though substantially different from mRNA expression mechanisms, also give rise to indirect, system-wide correlations between distantly connected metabolites.

Gaussian graphical models (GGMs) circumvent indirect association effects by evaluating *conditional* dependencies in multivariate Gaussian distributions [10]. A GGM is an undirected graph in which each edge represents the pairwise correlation between two variables conditioned against the correlations with all other variables (also denoted as *partial* correlation coefficients). GGMs have a simple interpretation in terms of linear regression techniques. When regressing two random variables X and Y on the remaining variables in the data set, the partial correlation coefficient between X and Y is given by the Pearson correlation of the residuals from both regressions. Intuitively speaking, we remove the (linear) effects of all other variables on X and Y and compare the remaining signals. If the variables are still correlated, the correlation is directly determined by the association of X and Y and not mediated by the other variables. Partial correlations have recently been applied to biological data sets for the inference of association networks from mRNA expression data [12-15], and for the elucidation of relationships between genomic features in the human genome [16]. One previous study used second-order partial correlations of genetic associations to elucidate genetically determined relations between metabolites [17].

In this manuscript we now study the capabilities of GGMs to recover metabolic pathway reactions solely from measured metabolite concentrations. First, we discuss the quality of the method and possible problems and pitfalls on computer-simulated systems. We then apply GGMs to a lipid-focused targeted metabolomics data set of 1020 blood serum samples with 151 measured metabolites from the German population study KORA [18,19]. The GGM is sparse in comparison to the corresponding Pearson correlation network, displays a modular structure with respect to different metabolite classes, and is stable towards changes in the underlying data set. We demonstrate that top-ranking metabolite pairs and further densely connected subgraphs in the GGM can indeed be attributed to known reactions in the human fatty acid biosynthesis and degradation pathways. In order to systematically verify this finding, we map partial correlation coefficients to the number of reaction steps between all metabolite pairs based on a literature-curated fatty acid pathway model. We observe statistically significant discriminatory features of GGMs to distinguish between directly and non-directly interacting metabolites in the metabolic network. In addition, low-order partial correlations turned out to be a suitable alternative to full-order GGMs for the present dataset. Finally, we will summarize and discuss the relevance of GGMs for metabolomics data sets, point out limitations of the method and suggest future steps. All metabolomics data used in this study, the generated correlation networks, model files and metabolite annotations are available online at http://hmgu.de/cmb/ggm.

Results and Discussion

GGMs delineate direct relationships in artificial reaction systems

Computer-simulated reaction systems are a valuable tool for the evaluation of correlation-based measures prior to their application to real metabolomics data sets. Previous works focused on the modeling of biological replicates with intrinsic noise on the metabolite levels [5]. In contrast, we here investigate the effects of variation of enzymatic activity in a human population cohort. Such variation might be genetically determined or, more likely, be the result of distinct regulatory effects and metabolic states between individuals. All reaction systems were implemented as ordinary differential equations with simple mass-action kinetics rate laws and reversible Michaelis-Menten-type enzyme kinetics (see Methods). In order to account for the above-mentioned enzymatic variability we applied a log-normal noise model, which has been previously described to be a reasonable approximation of cellular rate parameter

distributions [20]. The standard deviation σ was set to a value of 0.2 for the underlying normal distribution (note that the results are insensitive to the magnitude of σ). For each parameter sample we calculated the metabolite steady state concentrations on log-scale, and subsequently estimated the GGM by calculating partial correlation coefficients. All analyzed systems exhibit single, unique steady states independent of the respective parameter values. This feature was structurally verified using the ERNEST toolbox [21] for all networks except the negative feedback system. For the latter one, we employed empirical initial state sampling to ensure monostability in the given parameter range (see Additional file 1, section 1).

The first network we analyzed consists of a linear chain of three metabolites with different variants of reaction reversibility (Figure 1A-C). We observe high pairwise correlations for metabolites in mutual equilibrium due to reversible reactions (Figure 1A). This is in accordance with previous findings from [6], where correlation-generating mechanisms in metabolic reaction networks were identified. Furthermore, this simple example demonstrates how partial correlation coefficients in GGMs discriminate between directly and indirectly related metabolites. If only irreversible reactions are employed in the chain, neither regular correlation networks nor GGMs can distinguish between direct and indirect effects (Figure 1B). Species A is the only input metabolite in the system, and thus completely determines the levels of both B and C. This leads to generally high and non-distinguishable correlations between the three metabolites. However, if we introduce exchange reactions for all species, the GGM again correctly describes the network connectivity (Figure 1C). Such exchange mechanisms are likely to be present for most intracellular metabolites, which usually participate in multiple metabolic pathways (see e.g. KEGG PATHWAY online). Note that for this third case both regular

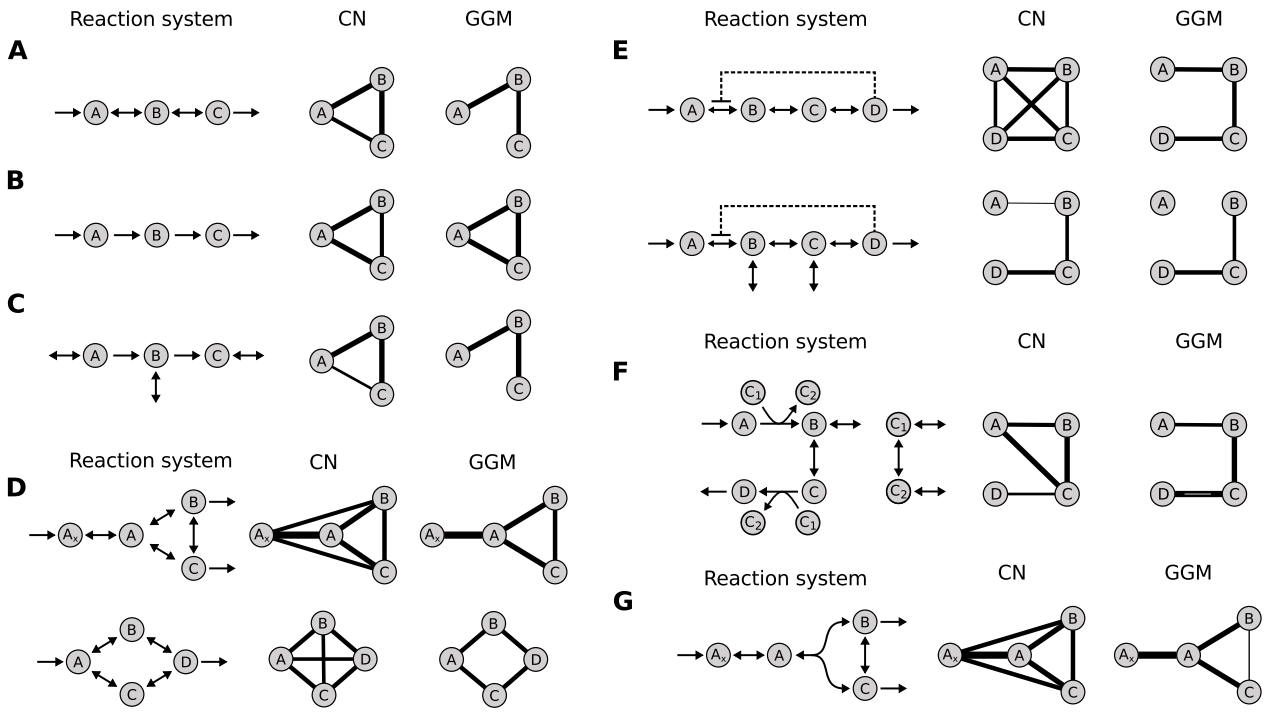


Figure 1 Evaluation of correlation networks (CN) and Gaussian graphical models (GGM) on artificial systems. Line widths represent relative edge weights in the respective networks (scaled to the strongest edges). **A:** Linear chain of three metabolites with reversible intermediate reactions. While the standard Pearson correlation network (CN) is fully connected, implying an overall high correlation of all metabolites, the GGM correctly discriminates between direct and indirect interactions. **B:** Linear chain with irreversible intermediate reactions. Neither CN nor GGM can distinguish direct from indirect effects, as metabolite A equally determines the levels of both B and C. **C:** Linear chain with irreversible reactions and input/output reactions for each metabolite. Although the edge weights for both CN and GGM are generally lower, the GGM now correctly predicts the network topology. **D:** Branched-chain first-order networks are correctly reconstructed by the GGM. **E:** End-product inhibition modules. When modeled as an open system, *A* is decoupled from the other metabolites and reconstruction fails at this point. Dashed lines mark enzyme inhibition interactions, larger arrows to the right indicate faster forward than backward reactions. **F:** Cofactor-driven network resembling the first three reactions from the glycolysis pathway. A correlation network fails to predict the correct pathway relationships. **G:** Non-linear system with a bi-molecular reaction. The GGM predicts only a only weak interaction between B and C. This is due to counterantagonistic processes of isomerization and substrate participation in the same reaction.

and partial correlation values are notably lower than for the first two chain variants. In addition to linear chains, pathway modules consisting of branched topologies with first-order, reversible reactions are correctly reconstructed by our method (Figure 1D). An overview of the reconstruction accuracy of GGMs on various types of first-order networks with different variants of reaction reversibility can be found in Additional file 1, section 2.

Interestingly, for some reaction setups, the accuracy of the method improves drastically with an increasing amount of external noise. Specifically, if the metabolite transport towards a pathway is subject to higher fluctuations, the GGM edge weight difference between directly and indirectly connected metabolites becomes larger. For a detailed discussion of this finding we refer the reader to Additional file 1, section 3. The second question we addressed with artificial reaction networks was the influence of enzyme-catalyzed reactions on GGM estimation. Therefore we setup reaction chains with four metabolites incorporating reversible enzymatic reactions. Forward maximal reaction rates $V_{\max}$ were set twice as fast as the backward reactions in order to ensure a directed mass flow. We found that the usage of Michaelis-Menten-type enzyme kinetics instead of mass-action kinetics does not alter our general findings. When forward reaction rates exceed backward reactions by far, the GGM discrimination quality is impaired. This is in line with the observation that purely irreversible reactions cannot be distinguished in the mass-action case (see above). Other specific parameters, like the Michaelis constant K_M , did not affect GGM calculation (Additional file 1, section 4). Another important aspect of enzyme-catalyzed reactions are allosteric regulation mechanism, like end-product inhibition for instance, which constitutes a negative feedback from the end to the beginning of a pathway [22]. The reconstruction results differ depending on whether exchange reactions are included in the system for not (Figure 1E). If the inhibitory module represents a closed system (no external fluxes except for the first and last metabolite), the regulatory interaction does not in influence GGM calculation. The net metabolite turnover speed might be drastically affected, but the topological effects of this reaction chain on the correlation structure remain unchanged. In contrast, when exchange reactions are introduced (second example in Figure 1E), the inhibition decouples A from the other metabolites and the reconstruction fails for this metabolite. Detailed results for different strengths of the inhibitory interaction are presented in Additional file 1, section 5.

Next, we studied the influence of cofactor-driven reactions on the reconstruction. Cofactors are ubiquitous substances usually involved in the transfer of certain molecular moieties or redox potentials [23]. We investigated such cofactor-coupled reactions (a) because they introduce non-linearity in the simulated dynamical systems, and (b) because cofactors are usually involved in many reactions and thus generate network-wide metabolite dependencies. We set up a network resembling the first three reactions from the glycolysis pathway. It consists of four metabolites and two energy transfer-related cofactors, ATP and ADP, involved in two phosphorylation reactions [24]. Again the GGM precisely describes metabolite connectivity in the system, whereas a regular correlation graph leads to false interpretations of the network topology (Figure 1F). Cofactors were modeled with input and output reactions to the rest of the metabolic system in order to account for the above-mentioned participation of cofactors in various reactions of the system. Again, it makes a substantial difference whether such exchange reactions are included in the model or not. Since our toy model only represents a small part of a larger system, missing exchange reaction for cofactors would create a false mass conservation relation that compromises correlation calculation. Finally, we investigated the effects of rate laws with non-linear substrate dependencies in the absence of cofactors. Therefore we modeled a reversible, bimolecular split reaction with isomerization of the two substrates (Figure 1G). An example of such a reaction network can be found in the glycolysis pathway between *fructose-1,6-bisphosphate, glyceraldehyde-3-phosphate and dihydroxyacetone phosphate*. Our simulations demonstrate that again a regular Pearson correlation network cannot delineate direct from indirect relationships in the pathway. The GGM only detects a weak association between B and C. This is due to counteranta gonistic processes in this reaction setup: isomerization and other reversible reactions generally induce positive correlations, whereas coparticipation as substrates in the same reaction induces negative correlations. Such effects of correlation-generating mechanisms which cancel each other out have been described before [6] and pose a problem to all reconstruction approaches which rely on linear dependencies.

The drawbacks of correlation-based methods discussed in this section, especially inhibitory mechanisms with exchange reactions and antagonistic mechanism, have to be kept in mind when attempting to reconstruct metabolic reactions from steady state data. For the present study, however, we assume the primarily linear lipid pathways not to contain such problematic reaction motifs.

A GGM inferred from a large-scale population-based data set displays a sparse, modular and robust structure

In the following we estimated a Gaussian graphical model using targeted metabolomics data from the

German population study KORA [18] ("Kooperative Gesundheitsforschung in der Region Augsburg"). We used a subset of the data set previously evaluated in a genome-wide association study [19], containing 1020 targeted metabolomics fasting blood serum measurements with 151 quantified metabolites. The metabolite panel includes acyl-carnitines, four classes of phospholipid species, amino acids and hexoses (see Methods). Both regular Pearson correlation coefficients and partial correlation coefficients (inducing the GGM) were calculated on the logarithmized metabolite concentrations. All edges corresponding to correlation values significantly different from zero now induce the networks displayed in Figure 2A+B. In order to exclude correlation effects generated by genetic variation in the study cohort, we investigated the in influence of SNP allele data from [19] on the GGM calculation. We found genetic effects to be neglectable (see Additional file 2), indicating that GGMs capture intrinsic biochemical properties of the system.

Pearson correlation coefficients show a strong bias towards positive values in our data set (Figure 2C); a typical feature of high-throughput data sets, also observed e.g. in microrarray expression data, which can be attributed to unspecific or indirect interactions [10]. We obtain 5479 correlation values significantly different from zero with $\tilde{\alpha} = 8.83.10^{-7}$ (α = 0.01 after Bonferroni correction), yielding an absolute significance correlation cutoff value of 0.1619 (see Methods). In contrast, the GGM shows a much sparser structure with 417 significant partial correlations after Bonferroni correction (Figure 2D). Most values center around a partial correlation coefficient of zero, whereas we observe a clear shift towards positive significant values. Note that negative partial correlations provide particular information that will be discussed later in this manuscript.

The GGM displays a modular structure with respect to the seven metabolite classes in our panel, while the class separation in the correlation network appears rather blurry (Figure 2E+F). We observe a clear separation of the amino acids and acyl-carnitines from all other classes. The four groups of phospholipids (diacyl-PCs, lyso-PCs, acyl-alkyl-PCs, and sphingomyelins) still showed locally clustered structures, but are strongly interwoven in the network. This is probably an effect of the dependence of all phospholipids on a similar fatty acid pool and, subsequently, the biosynthesis pathway acting on this substrate pool. In order to get an objective quantification of this observation, we calculated the group-based modularity Q on all significantly positive GGM edges according to [25] (see Methods). The same measure was calculated for 10^5 randomized GGM networks (random edge rewiring). For the original GGM we obtain a modularity of Q = 0.488, and the random networks yield Q = 0.118 ± 0.016, resulting in a highly significant z-score of z = 23.49. Furthermore, the modularity value induced by using the metabolite classes was compared to a partitioning optimized by simulated annealing. The optimized modularity is only slightly higher with Q = 0.557 and the resulting partitioning is very similar to the metabolite classes (see Additional file 3). Performing the modularity analysis with the full, weighted partial correlation matrix produces equivalent results (also shown in S3).

An important question for a multivariate statistical measure such as partial correlations is the robustness with respect to changes in the underlying data set. Furthermore, the dependence of the measure on the size of the data set needs to be addressed. To answer these questions, we performed two types of perturbations of our data set. First, we applied sample bootstrapping with 1000 repetitions and compared the resulting partial correlations to the original data set (Additional file 4, Figure S1). We observe small mean differences with low standard deviation ($0.03 \pm 8.2 \cdot 10^{-4}$). This indicates that for a large data set with n = 1020 samples, GGMs are robust against the choice of samples. We assume that each distinct metabolic state in the cohort is captured by a bootstrap sample, and thus all information required to calculate the GGM is contained. In addition to the bootstrap analysis, we estimated partial correlations for continuously decreasing sample sizes (Additional file 4, Figure S2). For each data set size we randomly picked samples from the original data set and repeated the procedure 100 times. The analysis shows that the GGM is stable even under decrease of the sample number. For instance, for a data set containing only around half of the original samples (n = 530) we get a partial correlation difference of $0.03 \pm 6.9\ 10^{-4}$. Only when the number of samples gets close to the number of variables (m = 151) the correlation matrix becomes ill-conditioned and strong differences from the original partial correlations occur. These problems of smaller metabolomics studies could be dealt with by regularization approaches or the usage of low-order partial correlation [26]. Taken together, our results demonstrate that the analyzed metabolomics data set is sufficient to robustly elucidate relationships between the measured metabolites.

Strong GGM edges represent known metabolic pathway interactions

The next step in our analysis was the manual investigation of metabolite pairs displaying strong partial correlation coefficients. Remarkably, we are able to provide pathway explanations for most metabolite pairs in the

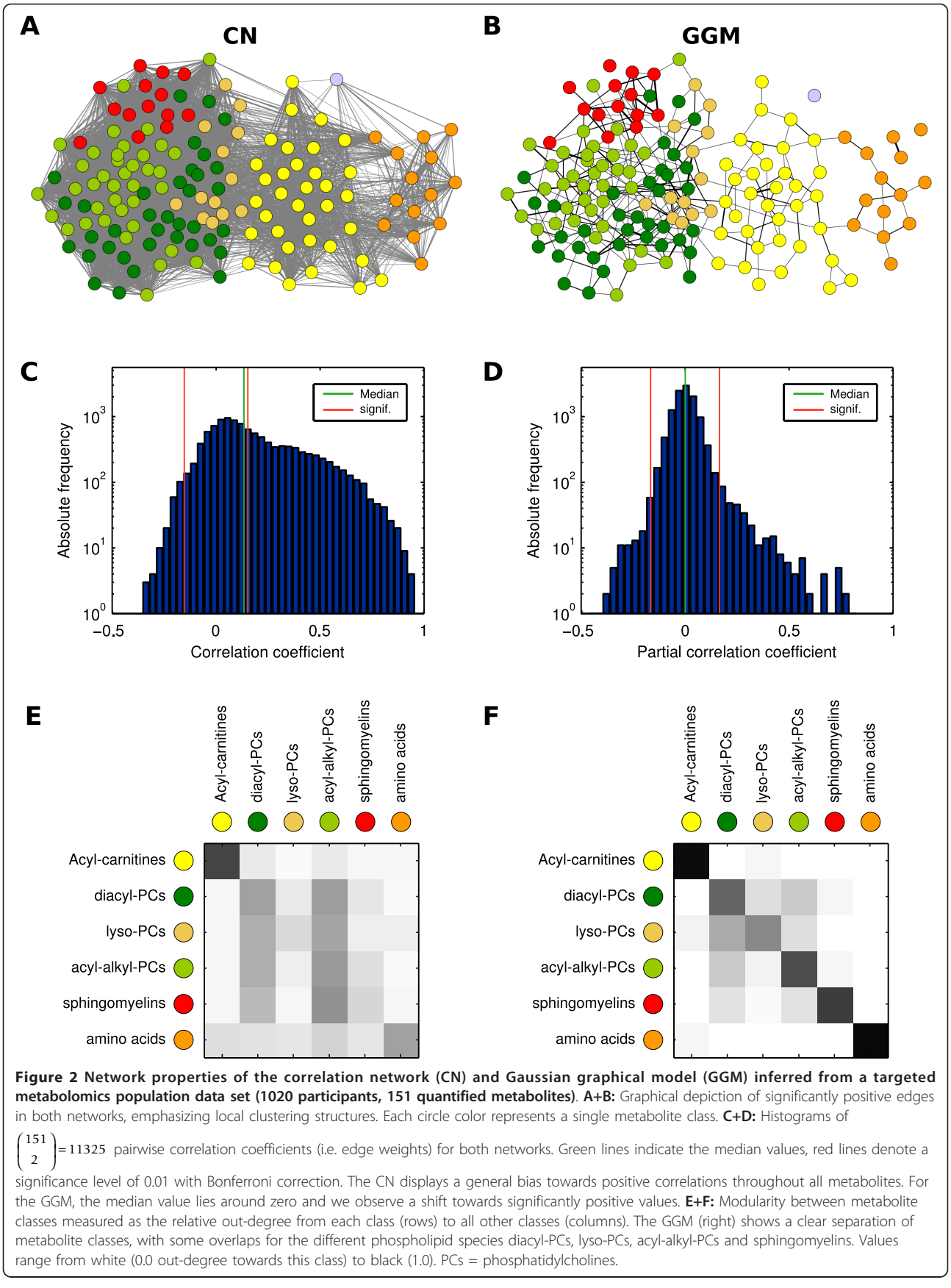


Figure 2 Network properties of the correlation network (CN) and Gaussian graphical model (GGM) inferred from a targeted metabolomics population data set (1020 participants, 151 quantified metabolites). **A+B:** Graphical depiction of significantly positive edges in both networks, emphasizing local clustering structures. Each circle color represents a single metabolite class. **C+D:** Histograms of $\binom{151}{2} = 11325$ pairwise correlation coefficients (i.e. edge weights) for both networks. Green lines indicate the median values, red lines denote a significance level of 0.01 with Bonferroni correction. The CN displays a general bias towards positive correlations throughout all metabolites. For the GGM, the median value lies around zero and we observe a shift towards significantly positive values. **E+F:** Modularity between metabolite classes measured as the relative out-degree from each class (rows) to all other classes (columns). The GGM (right) shows a clear separation of metabolite classes, with some overlaps for the different phospholipid species diacyl-PCs, lyso-PCs, acyl-alkyl-PCs and sphingomyelins. Values range from white (0.0 out-degree towards this class) to black (1.0). PCs = phosphatidylcholines.

Table 1 Top 20 positive GGM edge weights (i.e. partial correlation coefficients, PCC) in our data set along with proposed metabolic pathway explanations

Metabolite 1	Metabolite 2	PCC	Comment
Val	xLeu	0.821	Branched-chain amino acids
SM C18:0	SM C18:1	0.767	SCD/SCD5 desaturation
SM C16:1	SM C18:1	0.765	ELOVL6
PC ae C34:2	PC ae C36:3	0.752	2 reaction steps
SM (OH) C22:1	SM (OH) C22:2	0.743	sphingolipid-specific desaturation?
PC aa C34:2	PC aa C36:2	0.735	ELOVL1/ELOVL6 elongation
C10:0-carn	C8:0-carn	0.735	β-oxidation step
lysoPC a C16:0	lysoPC a C18:0	0.731	ELOVL6 elongation
PC aa C38:6	PC aa C40:6	0.709	ACOX1/3 + various ELOVLs
SM (OH) C14:1	SM (OH) C16:1	0.686	sphingolipid-specific elongation?
PC aa C36:4	PC aa C38:4	0.672	ACOX1/3 + various ELOVLs
PC aa C32:1	lysoPC a C16:1	0.661	C16:0/C16:1 phospholipid association
PC aa C38:5	PC aa C40:5	0.653	various ELOVLs
PC ae C34:3	PC ae C36:5	0.607	at least 3 reaction steps
PC aa C36:5	PC aa C38:5	0.596	ACOX1/3 + various ELOVLs
SM C24:0	SM C24:1	0.577	sphingolipid-specific desaturation?
PC ae C32:1	PC ae C32:2	0.574	SCD/SCD5 desaturation
SM (OH) C22:2	SM C24:1	0.567	possible elongation intermediate
C18:1-carn	C18:2-carn	0.561	β-oxidation intermediate

Most metabolite pairs can be directly linked to reactions in the fatty acid biosynthesis pathway, the β-oxidation pathway or amino acid-associated pathways.

top 20 positive partial correlations (Table 1). In the following, we will specifically discuss interesting, high-scoring metabolite pairs along with their responsible enzymes in the metabolic pathways.

The highest partial correlation in the data set with ζ = 0.821 is found for the two branched-chain amino acids Valine and xLeucine, where the latter compound represents both Leucine and Isoleucine (which have equal masses and are not distinguishable by the present method). The three metabolites are in close proximity in the metabolic network concerning their biosynthesis and degradation pathways. Further related amino acid pairs that display significant partial correlations are Histidine and Glutamine (ζ = 0.383), Glycine and Serine (ζ = 0.326) as well as Threonine and Methionine (ζ = 0.298).

Clear-cut signatures of the desaturation and elongation of long chain fatty acids can be seen for various sphingomyelins and lyso-PCs (Figure 3A). For example, SM C18:0 and SM C18:1 strongly associate with ζ = 0.767, most probably representing the initial Δ9 desaturation step of the polyunsaturated fatty acid biosynthesis pathway from C18:0 to C18:1-Δ9 by SCD (*Steaoryl-CoA desaturase*). The similarly high partial correlation between SM C16:1 and SM C18:1 (ζ = 0.765) as well as lysoPC a C16:1 and lysoPC a C18:1 (ζ = 0.315) can be attributed to the ELOVL6-dependent elongation from C16:1-Δ 9 to C18:1-Δ 11. Interestingly, this reaction is not contained in the public reaction databases but has been previously described by [27].

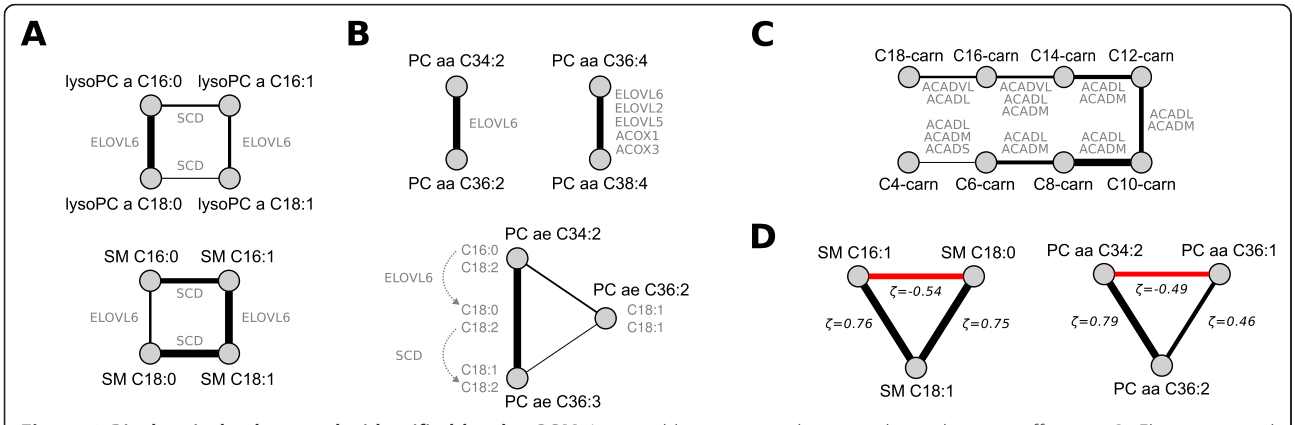


Figure 3 Biochemical subnetworks identified by the GGM. Line widths correspond to partial correlation coefficients. **A:** Elongation and desaturation signatures, most likely mediated by ELOVL6 and SCD, for C16 and C18 fatty acids incorporated in lyso-PCs and sphingomyelins. **B:** Top: Diacyl-phosphatidylcholine (PC aa) species with elongation and peroxisomal β-oxidation associations. Several combinatorial variants of side chain compositions are possible for C36:4 and C38:4, and thus different enzymes could mediate this connection. Bottom: Alkyl-acyl-phosphatidylcholines (PC ae) with supposedly distinct side chain composition, giving rise to a low association with a directly connected species (C36:2). **C:** Recovered β-oxidation pathway from C18 down to C4. Four enzymes with overlapping substrate specificities catalyze the rate-limiting reactions of this pathway. **D:** Two high-scoring triads, where metabolite pairs with a pathway distance of two constitute strong partial correlations. This feature of partial correlations aids in the reconstruction of the network topology beyond the direct neighborhood of each metabolite.

We identify a variety of strong GGM edges between diacyl-PC (lecithins, PC aa) and acyl-alkyl-PC (plasmalogens, PC ae) metabolite pairs (Figure 3B). For instance, PC aa C34:2 and PC aa C36:2 associate strongly with $\zeta = 0.735$, and PC aa C36:4 and PC aa C38:4 show a partial correlation of $\zeta = 0.672$. While the first pair can be precisely explained by an elongation from C16:0 to C18:0 by ELOVL6, different combinatorial variants come into play for the PC aa C36:4/PC aa C38:4 pair. Our mass-spectrometry technique only measures *brutto* compositions, that is the bulk side chain carbon content and total degree of desaturation. Depending on the exact composition of both fatty acid residues in the respective lipids, this association could be caused by long-chain elongations (C14 to C16 and C16 to C18 through fatty acid synthase and ELOVL6, respectively), by very-long-chain elongations (C22:4 to C24:4 through ELOVL2 or ELOVL5) and even by peroxisomal β -oxidation of fatty acids (through ACOX1 or ACOX3). An interesting situation arises for the phospholipids PC ae C34:2, PC ae C36:3 and PC ae C36:2. From its brutto formula the latter species could represent an intermediate step between the other two metabolites. However, it associates poorly with both other phospholipids, which in turn display a strong partial correlation ($\zeta = 0.752$). This finding can be explained by distinct fatty acid side chain compositions, showing differential incorporation of C18:0, C18:1 and C18:2 (Figure 3B, bottom).

For the acyl-carnitine group we observe a remarkably high partial correlation of $\zeta = 0.735$ for C8-carn and C10-carn and further acyl-carnitine pairs with a carbon atom difference of two (Figure 3C). These associations can be attributed to the β-oxidation pathway, i.e. the catabolic breakdown of fatty acids in the mitochondria [23]. During this degradation process, C_2 units are continuously split off from the shrinking fatty acid chain. Four *acyl-CoA dehydrogenases*, ACADS, ACADM and ACADL, ACADVL, catalyze the rate limiting reactions of β-oxidation for different fatty acid chain lengths [28,29]. Our interpretation of acyl-carnitine correlations as signatures of mitochondrial β-oxidation is in accordance with [19], where we identified associations between C8+C10, C12 and C4 with genetic variation in the ACADM, ACADL and ACADS loci, respectively.

We observe several associations that were not directly attributable to enzymatic interactions in the fatty acid biosynthesis or degradation pathways. For instance, lysoPC a 18:1 and lysoPC a 18:2 share a strong GGM edge ($\zeta = 0.543$) although the Δ12-desaturation step from oleic acid to linoleic acid is known to be missing in humans [30]. This missing reaction gives rise to the *essentiality* of fatty acids in the ω-6 unsaturated fatty acid pathway. A functional explanation could be a systemic equilibrium between the two fatty acids or remodeling processes specific for the lyso-PC metabolite class. Further examples are high partial correlations between the hydroxy sphingomyelins SM (OH) C22:1 and SM (OH) C22:2 ($\zeta = 0.743$) as well as the sphingomyelins SM C24:0 and SM C24:1 ($\zeta = 0.577$). To the best of our knowledge, there is no evidence for such fatty acid desaturation reactions in humans. The detected associations might therefore represent novel pathway interactions recovered by the Gaussian graphical model.

Negative values play a particular role in the interpretation of partial correlations coefficients. On the one hand, they obviously occur whenever regular negative correlations are involved. Mechanisms giving rise to negative correlations are, for example, coparticipation in the same reaction (cf. Figure 1E), mass conservation relations [6] or opposing regulatory effects. It is to be noted, however, that negative correlations are rare in our specific metabolomics data set (cf. Figure 2C). On the other hand, due to the mathematical properties of partial correlation coefficients negative partial correlation coefficients occur whenever two metabolites *A* and *B* have a strong correlation with a third metabolite *C*, but do not share a high correlation value with each other. Two examples from our data set are shown in Figure 3D. First, SM C18:0 is negatively partially correlated with SM C16:1, and both of these in turn are highly positively partially correlated with SM C18:1. The fatty acids C16:1 and C18:0 have no direct connection in the pathway, causing the strong negative partial correlation value. A similar situation can be found for three diacyl-PCs: PC aa C34:2 and PC aa C36:1 show a high partial correlation with PC aa C36:2, but a negative partial correlation with each other. Again, there is no possible direct reaction from a C34:2 lipid species to a C36:1 species. Not all metabolite triads in the network show such a one-negative/two-positive motif. But if present, they provide another step in the reconstruction of metabolic pathways (beyond the direct neighborhood of each metabolite) by detecting metabolites which are exactly two steps apart.

Partial correlation coefficients discriminate between directly and indirectly connected metabolites in a literature-curated fatty acid pathway model

The analyses from the previous section strengthened our conception that a GGM inferred from blood serum metabolomics data represents true metabolite associations. To systematically assess how GGM edges and pathway proximity between our lipid metabolites are related, we generated a literature-based model of fatty acid biosynthesis (Figure 4A). This model includes reactions from the public databases BiGG (H. sapiens Recon 1) [7], the Edinburgh Human Metabolic Network [31]

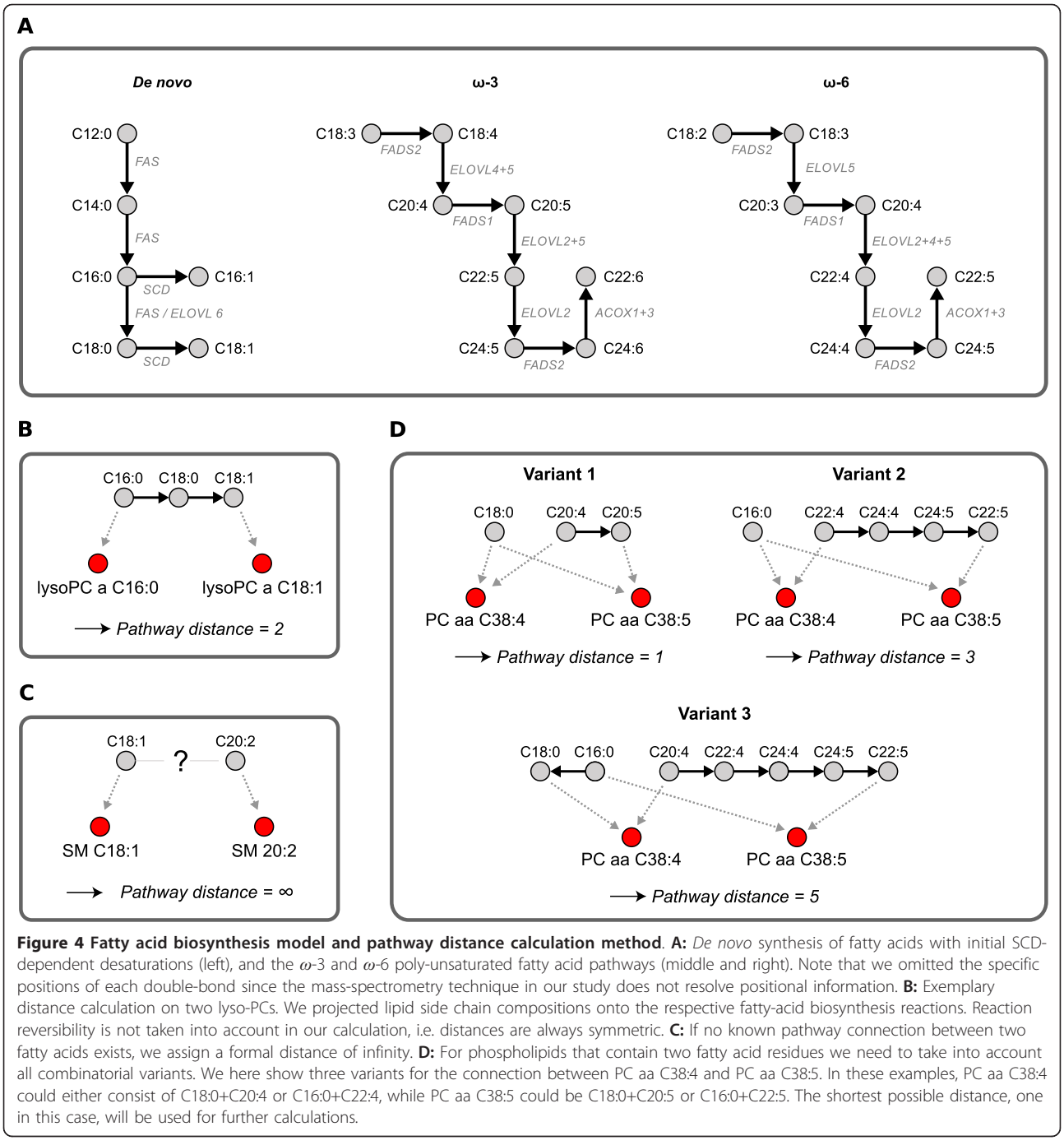
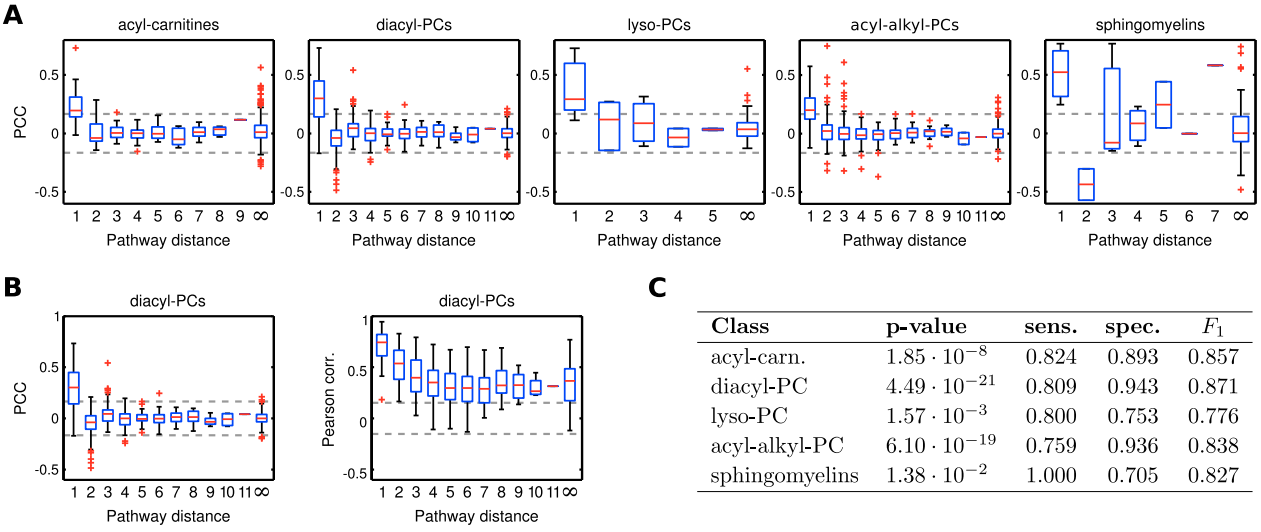


Figure 4 Fatty acid biosynthesis model and pathway distance calculation method. **A:** *De novo* synthesis of fatty acids with initial SCD-dependent desaturations (left), and the ω-3 and ω-6 poly-unsaturated fatty acid pathways (middle and right). Note that we omitted the specific positions of each double-bond since the mass-spectrometry technique in our study does not resolve positional information. **B:** Exemplary distance calculation on two lyso-PCs. We projected lipid side chain compositions onto the respective fatty-acid biosynthesis reactions. Reaction reversibility is not taken into account in our calculation, i.e. distances are always symmetric. **C:** If no known pathway connection between two fatty acids exists, we assign a formal distance of infinity. **D:** For phospholipids that contain two fatty acid residues we need to take into account all combinatorial variants. We here show three variants for the connection between PC aa C38:4 and PC aa C38:5. In these examples, PC aa C38:4 could either consist of C18:0+C20:4 or C16:0+C22:4, while PC aa C38:5 could be C18:0+C20:5 or C16:0+C22:5. The shortest possible distance, one in this case, will be used for further calculations.

and KEGG PATHWAY [29]. We then mapped the partial correlation coefficients from the KORA data set onto the minimal number of reaction steps between each pair of metabolites (*pathway distance*). Since our metabolite panel contains fatty-acid based lipids, we project the respective lipid compositions onto the fatty acid biosynthesis pathway (Figure 4B-D). For the analysis of acyl-carnitines we implemented a model of the β-oxidation pathway, consisting of a linear chain of C2 degradation steps (C10→C8→C6 etc.).

We observe a strong tendency towards significantly positive partial correlations for a pathway distance of one, i.e. directly connected metabolite pairs, for all five metabolite classes (Figure 5A). In total, 86 out of 130 partial correlations (66%) for a pathway distance of one are significantly positive. For instance, for the lyso-PC



C

Class	p-value	sens.	spec.	F_1
acyl-carn.	$1.85 \cdot 10^{-8}$	0.824	0.893	0.857
diacyl-PC	$4.49 \cdot 10^{-21}$	0.809	0.943	0.871
lyso-PC	$1.57 \cdot 10^{-3}$	0.800	0.753	0.776
acyl-alkyl-PC	$6.10 \cdot 10^{-19}$	0.759	0.936	0.838
sphingomyelins	$1.38 \cdot 10^{-2}$	1.000	0.705	0.827

Figure 5 Systematic evaluation of partial correlation coefficients versus pathway distances. Dashed lines in A and B indicate a significance level of 0.01 with Bonferroni correction. **A:** Pathway distances from our consensus model against partial correlation coefficients for the five lipid-based metabolite classes in our data set. We observe an enrichment of significant partial correlations for a pathway distance of one, which rapidly drops for an increasing number of pathway steps. **B:** Comparison of partial correlation coefficients and Pearson correlation coefficients. Pearson correlation coefficients are generally high, independent of the actual pathway distance, indicating for systemic coregulation effects throughout the lipid metabolism. **C:** Wilcoxon rank sum test p-values between the partial correlation distributions of directly and indirectly connected pairs, and sensitivity/specificity/F_1 values measuring the discriminatory power to distinguish direct from indirect pairs.

class (Figure 5A) nearly all partial correlation coefficients for a pathway distance of one are above significance level, whereas most values for a distance of two or larger remain insignificant. Some outliers from this observation, however, require closer inspection: First, for some metabolite classes we observe negative partial correlation values for metabolite pairs that are exactly two steps apart in the metabolic pathway: 10 of 73 partial correlations in the diacyl-PC class and 2 of 2 partial correlations in the sphingomyelin class are significantly negative for a distance of two. These negative values are effects of the coregulated metabolite triads described previously in this text. Second, we find 91 of 932 (~9:8%) unconnected metabolite pairs (pathway distance = ∞) with a partial correlation above significance level. These pairs represent potentially novel pathway predictions, missing interactions in the model or effects upstream of the metabolic network like enzyme coregulation.

A direct comparison of both partial and Pearson correlation coefficients for the diacyl-phosphatidylcholine class is shown in Figure 5B. As described earlier in this manuscript, we observe a general over-abundance of significant Pearson correlations independent of the actual pathway distance. Even for the metabolites without a known pathway connection, 1394 of a total of 1569 Pearson correlations are significant (88.85%, over all classes), in contrast to 131 out of 1569 for the partial correlations (8.35%).

The significantly different correlation value distributions between directly and indirectly linked metabolites (Figure 5A+B) barely provide a good quantification of the actual discrimination accuracy of this feature. Therefore we assessed the discriminative power of partial correlations to tell apart direct from indirect interactions by means of *sensitivity* and *specificity*. The sensitivity evaluates which fraction of directly connected metabolites in the pathway are recovered by significant GGM edges, whereas the specificity states how many of the significant edges actually represent a direct connection. A commonly used trade off measure between sensitivity and specificity is the F_1 score, which is defined as the harmonic mean of both quantities [32] (see Methods). Figure 5C lists sensitivity, specific city and F_1 for all 5 metabolite classes along with an evaluation of partial correlation distribution differences between directly and indirectly linked metabolites (determined by Wilcoxon's ranksum test). F_1 values over 0.75 and significant p-values for the ranksum test indicate a strong discrimination effect of partial correlation coefficients concerning direct vs. indirect pathway interactions. Possible reasons for non-perfect sensitivity and specific city values will be discussed in detail at the end of this text.

Low-order partial correlations

The data set from our present study contained enough samples to calculate full-order partial correlations, that is to calculate pairwise correlations conditioned against

all other n-2 metabolites. However, previous studies demonstrated that low-order partial correlation approaches can already be sufficient to elucidate direct interactions [12,16]. In order to assess how these measures perform in comparison to the full-order GGM, we calculated first-, second- and third order partial correlations using the approach developed by [12] for both computer-simulated networks and the metabolomics data (Additional file 5). The toy systems reveal clear cases where low-order approaches fail, for instance in the *diamond* motif displayed in Figure 1D. Surprisingly, however, especially first-order partial correlations worked remarkably well in discriminating direct from indirect interactions in the real data (F_1 values close to those displayed in Figure 5C). This result provides two valuable pieces of information. First, low-order partial correlation approaches, which require much less samples to obtain stable estimates, appear to be a suitable alternative to GGMs for the metabolite panel used in this study. Second, the high relative scoring of first-order partial correlations provides insights into the correlation structures in the data set. In particular, this result indicates that the underlying metabolic pathways are primarily composed of acyclic, linear chains, which fits well to the fatty acid pathways dominating our measured lipid species.

Conclusions

In this paper we addressed the reconstruction of metabolic pathway reactions from high-throughput targeted metabolomics measurements. Previous reconstruction approaches employed pairwise association measures, primarily standard Pearson correlation coefficients, to infer network topology information from metabolite profiles [5,6,8,33]. We here demonstrated the usefulness of Gaussian graphical models and their ability to distinguish direct from indirect associations by estimating the *conditional* dependence between variables. GGMs are based on partial correlation coefficients, that is the Pearson correlation between two metabolites corrected for the correlations with all other metabolites.

From computer simulations of metabolic reaction networks we deduced a set important aspects to be considered when interpreting partial correlation coefficients in reaction systems: (a) Metabolites in equilibrium due to reversible reactions can readily be recovered, whereas irreversible reactions pose a substantial problem to association-based reconstruction attempts (in concordance with [6]). (b) Input and output reactions for intermediate metabolites, however, improve the reconstruction accuracy. Such exchange reactions are likely to be present for most naturally occurring metabolites due to highly interconnected metabolic pathways. (c) With an increasing amount of fluctuations on the input reaction, the partial correlation difference between direct and indirect interactions increases for certain network topologies (e.g. for the irreversible linear metabolite chains). This indicates that a high heterogeneity of metabolic states in a population data set like the KORA cohort might be beneficial rather than problematic for our approach. (d) Metabolite connectivity in cofactor-driven networks can be accurately reconstructed. The presence of exchange reactions for cofactors, as they are likely to be present in real systems, has substantial impact on the reconstruction quality. The connectivity of the cofactors themselves, however, remains spurious. (e) Saturation effects in enzyme-catalyzed reactions do not pose a problem for the reconstruction process. However, inhibitory influences in metabolic modules that include exchange reactions might decouple certain metabolites and lead to false negative results. (f) Non-linear rate laws and antagonistic, correlation-generating mechanisms might impair reconstruction quality.

In the next step we inferred both a GGM and a regular correlation network from a large-scale metabolomics data set with 1020 strictly standardized samples from overnight fasting individuals measured by state-of-the art metabolomics technologies [19]. We investigated the influence of the 15 genome-wide-significant SNPs from this study on our GGM and demonstrated that genetic variation in the general population is neglectable for partial correlation calculation. We found that the GGM displays a much sparser structure than regular correlation networks. Only around 400 partial correlation values were above significance level (~3.6%), whereas half of all Pearson correlation values were significant after Bonferroni correction. This depicted the nature of partial correlation coefficients to neglect indirect associations between distantly related metabolites. We detected a strongly modular structure in the GGM with respect to the different metabolite classes, except for the four types of phospholipids which appear slightly interwoven. This provides a unique picture of the separation of metabolic pathways (synthesis, degradation and amino acid metabolism), but also the interaction between different lipid classes dependent on a single intracellular fatty acid pool. Finally, GGMs were stable with respect to both choice and number of samples in the data set. Even a smaller data set with only a few hundred samples would have been sufficient to achieve the results from this study. The estimation of GGMs for data sets with less samples than metabolites is possible [26], but notable deviations from the true partial correlation coefficient shave to be expected.

Manual investigation of high-scoring substructures in the GGM revealed groups of metabolites that could be directly attributed to reaction steps from the human fatty acid biosynthesis and degradation pathways. We

detected effects of ELOVL-mediated elongations and FADS-mediated desaturations of fatty acids as well as signatures of the catabolic β-oxidation pathway. For instance, our method successfully recovered a direct elongation from C16:1 to C18:1, which has been experimentally shown by [27] but is not present in the public reaction databases. Furthermore, we identified highly negative partial correlations as an indication for a pathway distance of two, serving as a further hint in the reconstruction of metabolic network topology. In order to systematically evaluate whether high partial correlations represent direct interactions, we generated a consensus model of fatty acid biosynthesis reactions from three publically available reaction databases. By mapping partial correlation coefficients to the number of reaction steps between two metabolites we observed a statistically significant enrichment of high values for a pathway distance of one. We calculated a high accuracy for partial correlations to discriminate between directly and indirectly associated metabolites, as measured by sensitivity, specificity and the F_1 measure. Interestingly, we could show that the discrimination quality of low-order partial correlations [12], especially the first-order variants, is close to the full-order GGM. Even though this might be a feature specific to the metabolite panel used in this study, low-order partial correlations represent a suitable alternative especially for studies with only few samples. If more samples than variables are available, however, we recommend GGMs as an unbiased approach conditioning against as many parameters as possible.

Taken together, our results demonstrate that GGMs inferred from metabolomics measurements in blood plasma samples reveal strong signatures of intracellular and even inner-mitochondrial processes. Previous studies on blood plasma samples detected similar relationships with cellular processes based on genetic associations [19] and case/control drug trials [34]. In this work we could now show that metabolite profiles alone are sufficient to capture the dynamics of metabolic pathways.

However, GGMs can never provide a perfect reconstruction of the underlying system. There are several factors that lead to the absence of high partial correlations between interacting metabolites, that is false negative edges in the GGM: (a) Counterantagonistic correlation-generating processes and bimolecular reactions (see above) might lead to the elimination of pairwise association; cf. [6]. (b) The respective enzyme might not be active in the current metabolic state, or its effects on the respective metabolite pools are neglectable. (c) Contrary to our general finding that even blood plasma metabolites carry strong signatures of metabolic pathways, the signal might be diminished for certain types of metabolites. Furthermore, the actual origins of blood plasma metabolites, e.g. in terms of measured cell types or causal tissue activity, still remain to be unraveled. The above-mentioned mechanisms are possible explanations for the non-perfect sensitivity values observed in Figure 5C. False positive GGM edges, on the other hand, provide interesting new metabolic pathway hypothesis. The presence of strong partial correlations in the absence of known metabolic connections could point out missing pathway information or regulatory effects not captured in a simple stoichiometric representation of the pathway.

Conclusively, this study presented Gaussian graphical models as a valuable tool for the recovery of biochemical reactions from high-throughput targeted metabolomics data. The present work could be extended by comparing high partial correlation coefficients with enzyme activity or expression data, or by the experimental validation of promising interaction candidates. We suggest using GGMs as a standard tool of investigation in future metabolomics studies, utilizing the upcoming wealth of metabolic profiling data to form a more comprehensive picture of cellular metabolism.

Methods

In silico simulation of artificial reaction networks

Let $x = (x_1,..., x_r)$ be a vector of metabolite concentrations and $S \varepsilon \mathbb{Z}^{m\times r}$ the stoichiometry matrix of a dynamical system with m metabolites and r reactions. Each column in S represents the compound stoichiometry of a single reaction, with negative values for the educts of a reaction and positive values for its products (cf. [35]). Furthermore, we define an *educt stoichiometry matrix* S^e, which only contains the negative values from S. The reaction rate laws v can be written as $v(x, k) = \mathrm{diag}(k)c(x)$, where $k := (k_1,...,k_r)$ represents a vector of elementary rate constants and $c_j(x) := \prod_{i=1}^{m} x_i^{-s_{ij}^e}$, $j = 1,..., r$ contains the products of substrate concentrations according to the law of mass action[36]. For example, for the reaction $x_1 + x_2 \rightarrow x_3$ we obtain $c = x_1x_2$, and $2x_1 + 3x_2 \rightarrow 2x_3$ yields $c = {x_1}^2{x_2}^3$. For enzyme-catalyzed reactions i, the corresponding entries in v are formulated using reversible Michaelis-Menten-type kinetics [37,38] instead of the mass-action term above:

$$v_i = \frac{\frac{V_{\max}^{+}}{K_M^{s}} \cdot [S] - \frac{V_{\max}^{-}}{K_M^{p}} \cdot [P]}{1 + \frac{[S]}{K_M^{s}} + \frac{[P]}{K_M^{p}}} \quad (1)$$

Where $V_{\max}^{+}$ and $V_{\max}^{-}$ are the product and substrate formation constants, respectively, K_M^s and K_M^p represent the Michaelis constants for substrate and product,

[S] represents the substrate concentration and [P] represents the product concentration. Note that we omitted reaction-specific parameter indices for simplicity here. Allosteric regulation was modeled using a mixed inhibition mechanism, which extends the rate law from equation (1) as follows:

$$v_i = \frac{\frac{V^+_{\max}}{K^s_M}\cdot[S] - \frac{V^-_{\max}}{K^p_M}\cdot[P]}{1 + \frac{[I]}{K_i} + \left(\frac{[S]}{K^s_M} + \frac{[P]}{K^p_M}\right)\left(1 + \frac{[I]}{K_{ii}}\right)}$$

with [I] being the inhibitor concentration, K_i the binding rate of the inhibitor to the enzyme and K_{ii} the binding rate of the inhibitor to the substrate-enzyme (or product-enzyme) complex. In a simple mixed (*non-competitive*) inhibition scenario, we assume $K_i = K_{ii}$.

The ordinary differential equations describing the temporal evolution of the system are now given as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = S \cdot v(x, k) \tag{2}$$

To introduce variability each parameter is subject to fluctuations according to a log-normal distribution with mean 1 and changing variances: $k_i \sim \mathrm{LogN}\left(1, \sigma_i^2\right)$. Finally, for fixed S and k, Pearson and partial correlations (see below) are calculated by drawing the vector k multiple times from the parameter distribution, calculating the corresponding metabolite steady state concentrations and logarithmizing the obtained values. If the system contains only zeroth-order and first-order reactions (i.e. input reactions and reactions with only one substrate), the steady state concentrations for a given k can be readily computed by equating (2) to zero and solving for c using linear algebra techniques. On the other hand, if higher order reactions are present, the ODEs are integrated numerically and simulated until equilibrium to get corresponding steady states. For this purpose, a variable-order solver for stiff differential equations (ode15s) from MATLAB was used [39]. The presence of a unique, single positive steady state was shown for each network individually using the ERNEST toolbox [21], or by empirical evaluation (parameter and initial value sampling). For a detailed analysis we refer the reader to Additional file 1, section 1.

Computation of correlation network and Gaussian graphical model

Let $X = (x_{kl})$ be the $\mathbb{R}^{n\times m}$ matrix of logarithmized metabolite concentrations (either measured data samples or computer-simulated steady states), where n is the number of samples, and m again represents the number of metabolites. Then the standard Pearson product-moment correlation coefficients $P = (\rho_{ij})$ between metabolites are calculated as

$$\rho_{ij} = \frac{\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)^2}\cdot\sqrt{\sum_{k=1}^{n}\left(x_{kj} - \bar{x}_j\right)^2}}$$

where $\bar{x}_i$ represents the mean value of metabolite i. Since we use a Gaussian graphical model, the conditional distributions are also Gaussian. Their width and the corresponding partial correlation coefficients can be calculated as

$$Z = (\zeta_{ij}) = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}} \quad \text{with} \quad (\omega_{ij}) = P^{-1}$$

A partial correlation value ζ_{ij} denotes the pairwise correlation of metabolites i and j corrected for the effects of all remaining metabolites. Since our study design contains more samples than measured variables, the correlation matrix has full rank and its inverse can be straightforwardly determined. First-,second-, and third-order partial correlations were calculated using the software published in [12]. To assess the significance of partial correlations, p-values $p(\zeta_{ij})$ were calculated using Fisher's z-transform [40]:

$$\begin{aligned} z(\zeta_{ij}) &= \frac{1}{2}\ln\left(\frac{1+\zeta_{ij}}{1-\zeta_{ij}}\right) \\ p(\zeta_{ij}) &= \left(1 - \phi\left(\sqrt{n-(m-2)-3}\cdot z(\zeta_{ij})\right)\right)\cdot 2 \end{aligned} \tag{3}$$

where φ stands for the cumulative distribution function of the standard normal distribution. In order to account for multiple testing, Bonferroni correction was applied to obtain an estimate of the significance level. Note that Bonferroni correction is the most conservative approach for multiple testing; it assumes independence of all tested values, which is certainly not the case for partial correlation coefficients. Based on a nominal significance level of $\alpha = 0.01$, we retrieve an adjusted level of $\tilde{\alpha} = 0.01 / 11325 = 8.83\cdot 10^{-7}$ after Bonferroni correction. Solving equation (3) for ζ_{ij} yields a minimum absolute partial correlation coefficient of 0.1619 for the given significance level. That is, all partial correlations smaller than -0.1619 or larger than 0.1619 are considered significant.

Bootstrapping was performed by randomly drawing 1020 samples with replacement from the original data set. For the second stability analysis, the investigation of different data set sizes, the respective number of samples was randomly drawn from the original data set.

The whole procedure was repeated 100 times to get a stable estimate of the deviation.

Network modularity calculation

We define the adjacency matrix ξ_{ij} of a new unweighted, undirected graph induced by all significantly positive partial correlations in ζ_{ij}:

$$\xi_{ij} := \begin{cases} 1, & \text{if } \zeta_{ij} \geq \tilde{\alpha} \\ 0, & \text{else} \end{cases}$$

where $\tilde{\alpha}$ represents the significance level after multiple testing correction. Now let $(V_1,...,V_6)$ be the partitioning of the metabolites into the six metabolite classes: acyl-carnitines, diacyl-PCs, lyso-PCs, acyl-alkyl-PCs, sphingomyelins and amino acids (the hexose is left out as only a single metabolite belongs to that class). We calculated the *relative out-degree* $R_{ij} \in \mathbb{R}^{6\times 6}$ from each class to the other classes, (i.e. the proportion of its edges each class shares with the other classes) as:

$$R_{ij} := \frac{\mathcal{A}(V_i, V_j)}{\mathcal{A}(V_i, V)}$$

where $\mathcal{A}(V', V'') = \sum_{i\in V', j\in V''} \xi_{ij}$ represents the total number of edges between V' and V'', and $V = \mathrm{U}V_i$ contains all metabolites in the network. The total network modularity Q of the network can be quantified according to [41] as:

$$Q := \sum_{i=1}^{6} \left[\frac{\mathcal{A}(V_i, V_i)}{\mathcal{A}(V, V)} - \left(\frac{\mathcal{A}(V_i, V)}{\mathcal{A}(V, V)} \right)^2 \right] \quad (4)$$

Intuitively, this measure compares the within-class edges with the edges to the rest of the network. The more edges there are within each class in comparison to the other classes, the higher Q will be. Note that equation (4) can be applied to both weighted and unweighted graphs. To assess the significance of the observed value, we performed graph randomization by edge rewiring [42,43] and subsequent calculation of Q. During the rewiring process we randomly pick two edges from the network and exchange the target nodes of each edge. In order to achieve sufficient randomization, this operation is repeated $5 \cdot e$ times, where e represents the number of edges in the graph. To perform edge reshuffling on weighted graphs, we decided on a neighbor-preserving variant as described in [44].

Study cohort and metabolite panel

KORA (Kooperative Gesundheitsforschung in der Region Augsburg) is a research platform in southern Germany with a primary focus on cardiovascular diseases, Diabetes mellitus type 2, and genetic epidemiology [18]. Fasting serum concentrations from n = 1020 individuals in the KORA F4 were determined by electrospray ionization tandem mass spectrometry (ESI-MS/MS) using the *Biocrates AbsoluteIDQ*™targeted metabolomics kit technology. These samples represent a subset of the data set previously evaluated in a genome-wide association study in [19].

A total of m = 151 metabolites were measured in the experiments: 14 amino acids including 13 proteinogenic amino acids and ornithine; hexose (sugars with 6 carbon atoms, e.g. glucose and fructose); 23 acylcarnitines [Cx:y-carn] (with x carbon atoms and y double bonds), 7 hydroxy-acylcarnitines [Cx:y-OH-carn], 6 dicarboxy-acylcarnitines [Cx:y-DC-carn], and 2 methylated dicarboxy-acylcarnitines variants [Cx:y-M-DC-carn]; 9 sphingomyelins [SM Cx:y] and 5 hydroxy-sphingomyelins [SM Cx:y-OH]; and 87 phosphatidylcholines (PC). These glycerophospholipids are further subdivided with respect to the presence of ester and ether bonds of fatty acid residues with the glycerol moiety. The set contains 36 diacyl-PCs with two esterified fatty acid residues [PC aa Cx:y], 38 acyl-alkyl-PCs with one ether-bond at the sn-2 position [PC ae Cx:y] and 13 lyso-PCs with only one esterified fatty acid residue at the sn-1 or sn-2 position [lysoPC a Cx:y]. Our mass spectometry technology cannot distinguish between the side chains of diacyl-phospholipids. The measured compounds are thus associated with the sum of carbon atoms and double bounds for both fatty acid residues. To ensure log-normality, we compared QQ-plots against normal distributions [45] for both non-logarithmized and logarithmized metabolite concentrations. All distributions were closer to log-normality than to regular normality (not shown), so we logarithmized the metabolite concentrations for the following analysis steps.

Sensitivity and specificity

In order to objectively evaluate the discrimination between directly and indirectly connected metabolites, we calculated sensitivity and specificity as:

$$\text{sens} := \frac{\text{TP}}{\text{TP} + \text{FN}}, \text{spec} := \frac{\text{TN}}{\text{TN} + \text{FP}}$$

with TP true positives, FP false positives, TN true negatives, FN false negatives [46].

A metabolite pair is considered true positive if it exhibits a partial correlation above the threshold and has a direct pathway connection; a false positive represents a metabolite pair also above the threshold but with no direct pathway connection; a false negative pair lies below the threshold but does have a direct pathway

connection; and finally a true negative pair lies below the threshold and also has no direct pathway connection. The F_1 score was calculated as the harmonic mean of both quantities:

$$F_1 := 2 \cdot \frac{\text{sens} \cdot \text{spec}}{\text{sens} + \text{spec}}$$

Pathway model

Pathway reactions in the human fatty acid metabolism were drawn from three independent databases: (1) *H. sapiens Recon 1* from the BiGG databases (confidence score of at least 4) [7], (2) the Edinburgh Human Metabolic Network reconstruction [31] and (3) the KEGG PATHWAY database [29] as of July 2010. A complete list of all curated reactions and the corresponding database identifiers can be found in Additional file 6. The reaction set was subdivided into two groups: (1) Fatty acid biosynthesis reactions which apply to the metabolite classes lyso-PC, diacyl-PC, acyl-alkyl-PC and sphingomyelins. (2) β-oxidation reactions representing fatty acid degradation to model reactions between the acylcarnitines. The β-oxidation model consists of a linear chain of C2 degradation steps (C10→C8→C6 etc.).

Fatty acid residues with identical masses, that cannot be distinguished by our mass-spectrometry technology, are merged into a single metabolite in the reaction set. For instance, the polyunsaturated fatty acids C20:4Δ8,11,14,17 from the omega-3 pathway and C20:4 Δ5,8,11,14 from the omega-6 pathway have identical numbers of carbon atoms and double bonds and are thus merged into a single metabolite C20:4.

Additional material

Additional file 1: Further results on computer-simulated networks.

Additional file 2: Effects of genetic variation on GGM calculation.

Additional file 3: Modularity: Optimized partitioning and weighted calculation.

Additional file 4: Stability of the GGM with respect to changes in the underlying data set.

Additional file 5: comparison with low-order partial correlation approaches.

Additional file 6: Literature-curated pathway model of human fatty acid biosynthesis and degradation.

Acknowledgements

The authors thank the anonymous reviewers for valuable comments and suggestions to improve the original manuscript. This research was partially supported by the Initiative and Networking Fund of the Helmholtz Association within the Helmholtz Alliance on Systems Biology (project CoReNe), by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center Diabetes Research (DZD e.V.), and by the BMBF-funded "Medizinische Systembiologie - MedSys" initiative (subproject SysMBo, project label 0315494A). Jan Krumsiek is supported by a PhD student fellowship from the "Studienstiftung des Deutschen Volkes". Thanks to Harold Gutch for critically proofreading and correcting this manuscript.

Author details

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München, Germany. [2]Faculty of Biology, Ludwig-Maximilians-Universität, Planegg-Martinsried, Germany. [3]Institute of Epidemiology, Helmholtz Zentrum München, Germany. [4]Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Zentrum München, Germany. [5]Department of Mathematics, Technische Universität München, Germany.

Authors' contributions

JK, KS and FJT conceived this data analysis project. TI and JA performed the sample preparation and data acquirement. JK performed the analysis and wrote the primary manuscript. All authors approved the final manuscript.

Received: 1 October 2010 Accepted: 31 January 2011
Published: 31 January 2011

References

1. Tweeddale H, Notley-McRobb L, Ferenci T: **Effect of slow growth on metabolism of Escherichia coli, as revealed by global metabolite pool ("metabolome") analysis.** *J Bacteriol* 1998, **180(19)**:5109-5116.
2. Wenk MR: **The emerging field of lipidomics.** *Nat Rev Drug Discov* 2005, **4(7)**:594-610[http://dx.doi.org/10.1038/nrd1776].
3. Griffin JL: **The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball?** *Philos Trans R Soc Lond B Biol Sci* 2006, **361(1465)**:147-161[http://dx.doi.org/10.1098/rstb.2005.1734].
4. Fiehn O: **Metabolomics-the link between genotypes and phenotypes.** *Plant Mol Biol* 2002, **48(1-2)**:155-171.
5. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks.** *Bioinformatics* 2003, **19(8)**:1019-1026.
6. Camacho D, de la Fuente A, Mendes P: **The origin of correlations in metabolomics data.** *Metabolomics* 2005, **1**:53-63[http://dx.doi.org/10.1007/s11306-005-1107-3].
7. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO: **Global reconstruction of the human metabolic network based on genomic and bibliomic data.** *Proc Natl Acad Sci USA* 2007, **104(6)**:1777-1782[http://dx.doi.org/10.1073/pnas.0610772104].
8. Arkin A, Shen P, Ross J: **A Test Case of Correlation Metric Construction of a Reaction Pathway from Measurements.** *Science* 1997, **277(5330)**:1275-1279[http://www.sciencemag.org/cgi/content/abstract/277/5330/1275].
9. Vance W, Arkin A, Ross J: **Determination of causal connectivities of species in reaction networks.** *Proc Natl Acad Sci USA* 2002, **99(9)**:5816-5821[http://dx.doi.org/10.1073/pnas.022049699].
10. Schäfer J, Strimmer K: **Learning Large-Scale Graphical Gaussian Models from Genomic Data.** *In Proc Natl Acad Sci USA, Volume 776, AIP* 2005, 263-276[http://link.aip.org/link/?APC/776/263/1].
11. Lee JM, Lee JM, Gianchandani EP, Eddy JA, Papin JA: **Dynamic analysis of integrated signaling, metabolic, and regulatory networks.** *PLoS Comput Biol* 2008, **4(5)**:e1000086[http://dx.doi.org/10.1371/journal.pcbi.1000086].
12. de la Fuente A, Bing N, Hoeschele I, Mendes P: **Discovery of meaningful associations in genomic data using partial correlation coefficients.** *Bioinformatics* 2004, **20(18)**:3565-3574[http://dx.doi.org/10.1093/bioinformatics/bth445].
13. Magwene PM, Kim J: **Estimating genomic coexpression networks using first-order conditional independence.** *Genome Biol* 2004, **5(12)**:R100[http://dx.doi.org/10.1186/gb-2004-5-12-r100].
14. Schäfer J, Strimmer K: **An empirical Bayes approach to inferring large-scale gene association networks.** *Bioinformatics* 2005, **21(6)**:754-764[http://dx.doi.org/10.1093/bioinformatics/bti062].
15. Wille A, Zimmermann P, Vranová E, Fürholz A, Laule O, Bleuler S, Hennig L, Prelic A, von Rohr P, Thiele L, Zitzler E, Gruissem W, Bühlmann P: **Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana.** *Genome Biol* 2004, **5(11)**:R92[http://dx.doi.org/10.1186/gb-2004-5-11-r92].

16. Freudenberg J, Wang M, Yang Y, Li W: **Partial correlation analysis indicates causal relationships between GC-content, exon density and recombination rate in the human genome.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:S66[http://dx.doi.org/10.1186/1471-2105-10-S1-S66].
17. Keurentjes JJB, Fu J, de Vos CHR, Lommen A, Hall RD, Bino RJ, van der Plas LHW, Jansen RC, Vreugdenhil D, Koornneef M: **The genetics of plant metabolism.** *Nat Genet* 2006, **38(7)**:842-849[http://dx.doi.org/10.1038/ng1815].
18. Holle R, Happich M, Löwel H, Wichmann HE, Group MONICAORAS: **KORA-a research platform for population based health research.** *Gesundheitswesen* 2005, **67(Suppl 1)**:S19-S25.
19. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, Prehn C, Altmaier E, Kastenmüller G, Kato BS, Mewes HW, Meitinger T, de Angelis MH, Kronenberg F, Soranzo N, Wichmann HE, Spector TD, Adamski J, Suhre K: **A genome-wide perspective of genetic variation in human metabolism.** *Nat Genet* 2010, **42(2)**:137-141[http://dx.doi.org/10.1038/ng.507].
20. Liebermeister W, Klipp E: **Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data.** *Theor Biol Med Model* 2006, **3**:42[http://dx.doi.org/10.1186/1742-4682-3-42].
21. Soranzo N, Altafini C: **ERNEST: a toolbox for chemical reaction network theory.** *Bioinformatics* 2009, **25(21)**:2853-2854[http://dx.doi.org/10.1093/bioinformatics/btp513].
22. Winicov I, Pizer LI: **The mechanism of end product inhibition of serine biosynthesis. IV. Subunit structure of phosphoglycerate dehydrogenase and steady state kinetic studies of phosphoglycerate oxidation.** *J Biol Chem* 1974, **249(5)**:1348-1355.
23. Berg JM, Tymoczko JL, Stryer L: In *Biochemistry* Edited by: Freeman WH , sixth 2006 [http://www.worldcat.org/isbn/0716787245].
24. Hynne F, Danø S, Sørensen PG: **Full-scale model of glycolysis in Saccharomyces cerevisiae.** *Biophys Chem* 2001, **94(1-2)**:121-163[http://linkinghub.elsevier.com/retrieve/pii/S0301-4622(01)00229-0].
25. Newman MEJ, Girvan M: **Finding and evaluating community structure in networks.** *Phys Rev E* 2004, **69(2)**:026113.
26. Schäfer J, Strimmer K: **A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.** *Statistical applications in genetics and molecular biology* 2005, **4**: [http://dx.doi.org/10.2202/1544-6115.1175].
27. Matsuzaka T, Shimano H, Yahagi N, Kato T, Atsumi A, Yamamoto T, Inoue N, Ishikawa M, Okada S, Ishigaki N, Iwasaki H, Iwasaki Y, Karasawa T, Kumadaki S, Matsui T, Sekiya M, Ohashi K, Hasty AH, Nakagawa Y, Takahashi A, Suzuki H, Yatoh S, Sone H, Toyoshima H, ichi Osuga J, Yamada N: **Crucial role of a long-chain fatty acid elongase, Elovl6, in obesity-induced insulin resistance.** *Nat Med* 2007, **13(10)**:1193-1202[http://dx.doi.org/10.1038/nm1662].
28. Eaton S, Bartlett K, Pourfarzam M: **Mammalian mitochondrial beta-oxidation.** *Biochem J* 1996, **320(Pt 2)**:345-357.
29. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
30. Spector A: **Essentiality of fatty acids.** *Lipids* 1999, **34(0)**:S1-S3[http://dx.doi.org/10.1007/BF02562220].
31. Ma H, Sorokin A, Mazein A, Selkov A, Selkov E, Demin O, Goryanin I: **The Edinburgh human metabolic network reconstruction and its functional analysis.** *Mol Syst Biol* 2007, **3**:135[http://dx.doi.org/10.1038/msb4100177].
32. Van Rijsbergen CJ: *Information Retrieval.* 2 edition. Dept. of Computer Science, University of Glasgow; 1979 [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.2325].
33. Steuer R: **Review: on the analysis and interpretation of correlations in metabolomic data.** *Brief Bioinform* 2006, **7(2)**:151-158[http://dx.doi.org/10.1093/bib/bbl009].
34. Altmaier E, Ramsay SL, Graber A, Mewes HW, Weinberger KM, Suhre K: **Bioinformatics analysis of targeted metabolomics-uncovering old and new tales of diabetic mice under medication.** *Endocrinology* 2008, **149(7)**:3478-3489[http://dx.doi.org/10.1210/en.2007-1747].
35. Palsson BO: *Systems Biology: Properties of Reconstructed Networks.* 1 edition. Cambridge University Press; 2006 [http://www.worldcat.org/isbn/0521859034].
36. Famili I, Mahadevan R, Palsson BO: **k-Cone analysis: determining all candidate values for kinetic parameters on a network scale.** *Biophys J* 2005, **88(3)**:1616-1625[http://dx.doi.org/10.1529/biophysj.104.050385].
37. Michaelis L, Menten ML: **Die Kinetik der Invertinwirkung.** *Biochem Z* 1913, **49(333-369)**:352.
38. Dräger A, Kronfeld M, Ziller MJ, Supper J, Planatscher H, Magnus JB, Oldiges M, Kohlbacher O, Zell A: **Modeling metabolic networks in C. glutamicum: a comparison of rate laws in combination with various parameter optimization strategies.** *BMC Syst Biol* 2009, **3**:5[http://dx.doi.org/10.1186/1752-0509-3-5].
39. Shampine LF, Reichelt MW: **The MATLAB ODE Suite.** *SIAM Journal on Scientific Computing* 1997, **18**:1-22.
40. Fisher RA: **The Distribution of the Partial Correlation Coefficient.** *Metron* 1924, **3**:329-332[http://hdl.handle.net/2440/15182].
41. White S, Smyth P: **A Spectral Clustering Approach To Finding Communities in Graphs.** *In SIAM International Conference on Data Mining* 2005 [http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.8978].
42. Maslov S, Sneppen K: **Specificity and stability in topology of protein networks.** *Science* 2002, **296(5569)**:910-913[http://dx.doi.org/10.1126/science.1065103].
43. Wong P, Althammer S, Hildebrand A, Kirschner A, Pagel P, Geissler B, Smialowski P, Blöchl F, Oesterheld M, Schmidt T, Strack N, Theis FJ, Ruepp A, Frishman D: **An evolutionary and structural characterization of mammalian protein complex organization.** *BMC Genomics* 2008, **9**:629[http://dx.doi.org/10.1186/1471-2164-9-629].
44. Hartsperger ML, Blöchl F, Stümpflen V, Theis FJ: **Structuring heterogeneous biological information using fuzzy clustering of k-partite graphs.** *BMC Bioinformatics* 2010, **11**:522[http://dx.doi.org/10.1186/1471-2105-11-522].
45. Thode HC: *Testing for normality* CRC Press; 2002.
46. Altman DG, Bland JM: **Diagnostic tests. 1: Sensitivity and specificity.** *BMJ* 1994, **308(6943)**:1552.

doi:10.1186/1752-0509-5-21
Cite this article as: Krumsiek *et al.*: Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology* 2011 **5**:21.

OPEN ACCESS Freely available online

PLoS GENETICS

Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers

Kirstin Mittelstrass[1☯], Janina S. Ried[2☯], Zhonghao Yu[1☯], Jan Krumsiek[3], Christian Gieger[2], Cornelia Prehn[4], Werner Roemisch-Margl[3], Alexey Polonikov[5], Annette Peters[6], Fabian J. Theis[3], Thomas Meitinger[7,8], Florian Kronenberg[9], Stephan Weidinger[10], Heinz Erich Wichmann[11,12,13], Karsten Suhre[3,14,15], Rui Wang-Sattler[1], Jerzy Adamski[4,16¶*], Thomas Illig[1¶*]

1 Unit of Molecular Epidemiology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **2** Institute of Genetic Epidemiology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **3** Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **4** Institute of Experimental Genetics, Genome Analysis Center, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **5** Department of Biology, Medical Genetics, and Ecology, Kursk State Medical University, Kursk, Russia, **6** Institute of Epidemiology II, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **7** Institute of Human Genetics, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **8** Institute of Human Genetics, Klinikum Rechts der Isar, Technische Universität München, Munich, Germany, **9** Division of Genetic Epidemiology, Department of Medical Genetics and Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria, **10** Department of Dermatology, Venereology, and Allergy, University Hospital Schleswig-Holstein, Kiel, Germay, **11** Institute of Epidemiology I, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany, **12** Institute of Medical Informatics, Biometry, and Epidemiology, Chair of Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany, **13** Klinikum Grosshadern, Munich, Germany, **14** Faculty of Biology, Ludwig-Maximilians-Universität, Planegg-Martinsried, Germany, **15** Weill Cornell Medical College in Qatar, Qatar Foundation, Education City, Doha, Qatar, **16** Lehrstuhl für Experimentelle Genetik, Technische Universität München, Munich, Germany

Abstract

Metabolomic profiling and the integration of whole-genome genetic association data has proven to be a powerful tool to comprehensively explore gene regulatory networks and to investigate the effects of genetic variation at the molecular level. Serum metabolite concentrations allow a direct readout of biological processes, and association of specific metabolomic signatures with complex diseases such as Alzheimer's disease and cardiovascular and metabolic disorders has been shown. There are well-known correlations between sex and the incidence, prevalence, age of onset, symptoms, and severity of a disease, as well as the reaction to drugs. However, most of the studies published so far did not consider the role of sexual dimorphism and did not analyse their data stratified by gender. This study investigated sex-specific differences of serum metabolite concentrations and their underlying genetic determination. For discovery and replication we used more than 3,300 independent individuals from KORA F3 and F4 with metabolite measurements of 131 metabolites, including amino acids, phosphatidylcholines, sphingomyelins, acylcarnitines, and C6-sugars. A linear regression approach revealed significant concentration differences between males and females for 102 out of 131 metabolites (*p-values*$<3.8\times10^{-4}$; Bonferroni-corrected threshold). Sex-specific genome-wide association studies (GWAS) showed genome-wide significant differences in beta-estimates for SNPs in the *CPS1* locus (carbamoyl-phosphate synthase 1, significance level: $p<3.8\times10^{-10}$; Bonferroni-corrected threshold) for glycine. We showed that the metabolite profiles of males and females are significantly different and, furthermore, that specific genetic variants in metabolism-related genes depict sexual dimorphism. Our study provides new important insights into sex-specific differences of cell regulatory processes and underscores that studies should consider sex-specific effects in design and interpretation.

Citation: Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, et al. (2011) Discovery of Sexual Dimorphisms in Metabolic and Genetic Biomarkers. PLoS Genet 7(8): e1002215. doi:10.1371/journal.pgen.1002215

Editor: Mark I. McCarthy, University of Oxford, United Kingdom

Received October 25, 2010; **Accepted** June 17, 2011; **Published** August 11, 2011

Funding: The KORA research platform (KORA: Cooperative Research in the Region of Augsburg) and the MONICA Augsburg studies (monitoring trends and determinants on cardiovascular diseases) were initiated and financed by the Helmholtz Center Munich, National Research Center for Environmental Health, which is funded by the German Federal Ministry of Education, Science, Research, and Technology and by the State of Bavaria. Part of this work was financed by the German National Genome Research Network (NGFN) to the Institute of Epidemiology. This study was supported 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.) and by grants from the "Genomics of Lipid-associated Disorders – GOLD" of the "Austrian Genome Research Programme GEN-AU." Our research was supported within the Munich Center of Health Sciences (MC Health) as part of LMUinnovativ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: illig@helmholtz-muenchen.de (TI); adamski@helmholtz-muenchen.de (JA)

☯ These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Author Summary

The combination of genomic and metabolic studies during the last years has provided astonishing results. However, most of the studies published so far did not consider the role of sexual dimorphism and did not analyse their data stratified by sex. The investigation of 131 serum metabolite concentrations of >3,300 population-based samples (KORA F3/F4) revealed significant differences in the metabolite profile of males and females. Furthermore, a genome-wide picture of sex-specific genetic variations in human metabolism (>2,000 subjects from KORA F3/F4 cohorts) was investigated. Sex-specific genome-wide association studies (GWAS) showed differences in the effect of genetic variations on metabolites in men and women. SNPs in the *CPS1* (carbamoyl-phosphate synthase 1) locus showed genome-wide significant differences in beta-estimates of sex-specific association analysis (significance level: 3.8×10^{-10}) for glycine. As global metabolomic techniques are more and more refined to identify more compounds in single biological samples, the predictive power of this new technology will greatly increase. This suggests that metabolites, which may be used as predictive biomarkers to indicate the presence or severity of a disease, have to be used selectively depending on sex.

Introduction

Metabolomics provides a powerful tool to analyse physiological and disease-induced biological states on the molecular level, taking into account both the organism's intrinsic properties, i.e. genetic factors, and the effects of lifestyle, diet, and environment. The development of sophisticated analytic platforms and modern computational tools to handle increasingly complex data now enables the quantification of hundreds of metabolites from complex biological samples with a high throughput rate. These advancements support the integration of metabolomic profiles with genetic, epigenetic, transcriptomic and proteomic data for holistic systems biology approaches. Recently, common genetic variants have been demonstrated to exert large effects on individual metabolic capacities called "genetically determined metabotypes" [1,2]. Therefore genetic variants in metabolism-related genes led to specific and clearly differentiated metabolic phenotypes [1,3]. Knowledge on such genetically determined metabotypes is of crucial importance to understand the contribution and complex interaction of genes, proteins and metabolites in health and disease. Consequently, genetic studies can help to elucidate the direction of effects between metabolites and a specific disease. Thus, the combination of genetic and metabolic markers is an important emerging approach for biological research. To uncover potentially confounding influences on the interpretation of metabolic results, it is important to minimize the occurring confounders on human serum metabolites in a population-based study that has not been subjected to lifestyle and dietary controls. Pointed out recently, gender inequalities are another increasingly recognized problem in both basic research and clinical medicine [4]. Nevertheless, many published studies did not analyse their data stratified by sex [4–6] although there is a strong correlation between sex and the incidence, prevalence, age at onset, symptoms and severity of a disease, as well as the reaction to drugs [7,8]. A survey of studies published in 2004 of nine different medical journals found that only 37% of participants were women (24% when restricted to drug trials), and only 13% of studies analysed data by sex [4]. Therefore we systematically assessed the effect of sex on serum metabolites in a large population-based cohort [9]. Furthermore, we investigated whether there are sex-specific differences in the genetic determination of metabotypes.

Results

Phenotypic Metabotype Differences between Males and Females

All phenotypic analysis steps were performed on population-based cohort data from KORA F4 (1452 males, 1552 females) and KORA F3 (197 males, 180 females, Figure S1) with fasting serum concentrations of 131 metabolites. The metabolites covered a biologically relevant panel that could be divided into five subgroups such as amino acids, sugars, acylcarnitines and phospholipids. Further information concerning the study population, sampling methods and the metabolite panel are described in the Material and Methods section and in the Tables S1, S6, and S7 and Figure S3.

A Partial Least Square (PLS) analysis [10] of all metabolites showed that there were major differences of mean serum metabolite concentrations between males and females, as the values for the first two PLS components clustered clearly for men and women (Figure 1).

Motivated by the global gender differences in metabolite concentrations shown by PLS analysis, furthermore, the effect of sex on each metabolite was analysed using linear regression analysis. For each metabolite we calculated a linear regression with the logarithm of the metabolite concentration as dependent and sex as explanatory variable besides the covariates age, BMI and internal batch. The test whether the explanatory variable sex has a significant effect on the logarithm of the metabolite concentration revealed significant effects of gender in 102 of 131 metabolites (*p-value* below the Bonferroni-corrected significance level of 3.8×10^{-4}). At least one metabolite of each subgroup including amino acids, acylcarnitines, phosphatidylcholines, lysophosphatidylcholines and sphingomyelins showed significant sex-specific differences in metabolite concentrations. In Table 1 the results of the linear regression analysis for representatives of each subgroup are presented, the complete list of results can be found in the Table S2.

The linear regression analysis showed that the concentrations of most amino acids were significantly higher in males except for the concentrations of glycine (effect of sex: $ß=-0.13$, $p\text{-value}=2.3\times10^{-46}$) and serine (effect of sex: $ß=-0.13$, $p\text{-value}=1.0\times10^{-12}$) which displayed higher concentrations in females. (Table 1, Table S2). The relative sex-specific difference for glycine was $\Delta=-14\%$ (Table S8). That means that the mean concentration in men was 14% lower than in women (see Material and Methods). The levels of most serum acylcarnitines were significantly higher in males compared to females (Table 1, Table S2 and S8). The concentrations of phosphatidylcholines (PC ae Cx:y or PC aa Cx:y) tended to be significantly lower in males compared to females. The most significant difference between gender could be seen for the phosphatidylcholine PC aa C32:3 ($\Delta=-17.9\%$, $p\text{-value}=4.4\times10^{-108}$), whereas lysophosphatidylcholine (lysoPC a Cx:y) concentrations were higher in males compared to females. In contrast, the concentrations of most sphingomyelins were significantly lower in men than in women (Table 1, Table S2 and S8). The concentration of H1 which is the sum of C6-sugars, was significantly higher in males compared to females ($\Delta=7.3\%$, $p\text{-value}=6.2\times10^{-27}$) (Table 1, Table S2 and S8). Figure 2 systematically reviews the sex-specific metabolite variations identified in this study.

The adjustment for different covariates (e.g.: waist-hip ratio (WHR), HDL (high density lipoprotein), LDL (low density

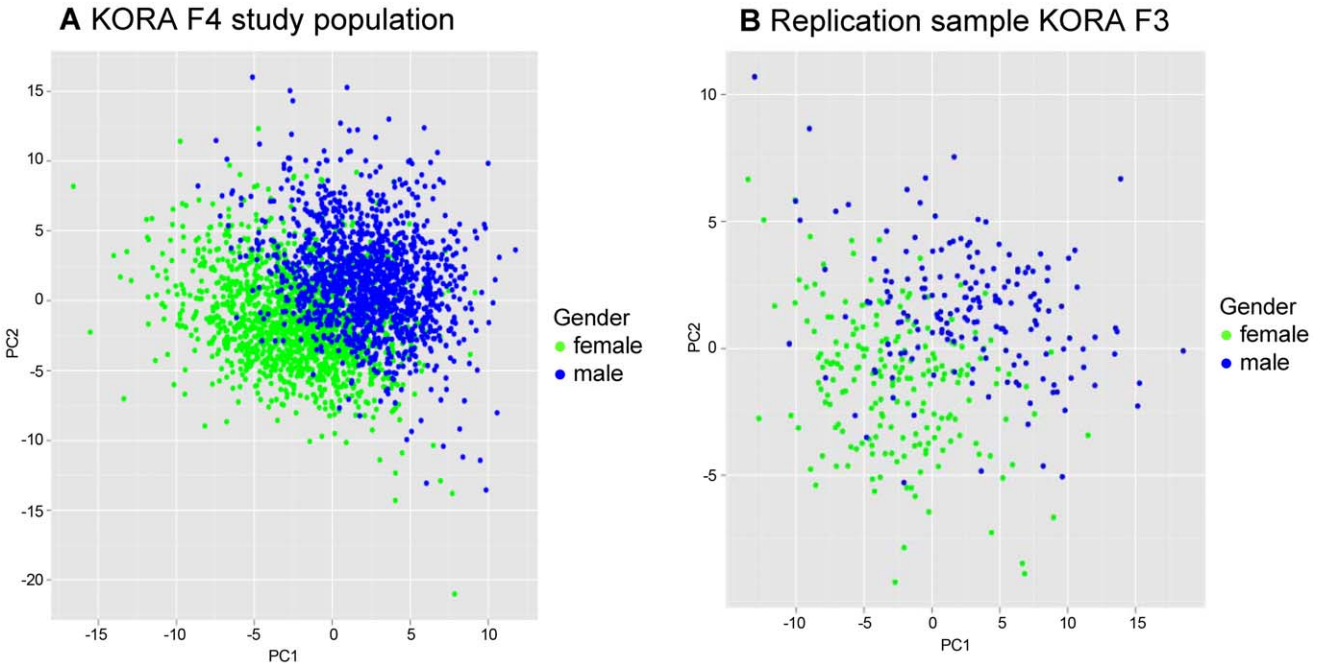


Figure 1. Two dimensional partial least square (PLS) analyses showing the contribution of 131 metabolites in males and females.
doi:10.1371/journal.pgen.1002215.g001

lipoprotein), triglycerides, type 2 diabetes, smoking and high alcohol consumption) did not affect the sex-specific differences in the metabolite concentrations extensively. The majority of the high significant sex-effects remained significant. In particular, adjustments for lipid parameter (HDL, LDL and triglycerides), type 2 diabetes, smoking and high alcohol consumption did not influence our main findings. If WHR was included into the linear regression model as covariate replacing BMI or as additional covariate besides BMI, the *p-values* of the sex-effect on metabolites changed, but for most metabolite concentrations the gender-differences remained significant. Interestingly, there were seven PC aa Cx:ys and Lyso PC a C17:0 that showed significant differences between sexes while adjusting for age and WHR but not for BMI and age adjustment. We refer the interested reader to Table S2.

As replication the same linear regression approach (covariates: age, BMI) was applied to the KORA F3 cohort which included 377 individuals (Figure S1). Despite this smaller sample size for 63 of 102 metabolites with a significant effect of sex in KORA F4, the effect of sex in KORA F3 had the same direction and a significant *p-value* lower than the Bonferroni-corrected replication significance level corrected for the 102 metabolites taken forward to replication ($0.05/102 = 4.9\times10^{-4}$). That means 61.8% of the sex-specific differences could be replicated (Table 1, Table S3).

A combined meta-analysis of KORA F4 and KORA F3 revealed 113 metabolites with a significant effect of sex (Bonferroni-corrected meta-analysis significance level: $p<3.8\times10^{-4}$) (Table S3).

Sex-Specific Effects in the Metabolic Network

We further investigated how groups of metabolites share pairwise correlations, that mean similar effects, and how the sex-specific effects propagate through the metabolic network. Therefore we calculated a partial correlation matrix between all metabolites, corrected against age, sex and BMI [11]. The resulting network, which is also referred to as a Gaussian graphical model (see Material and Methods), was annotated with the results from the linear regression analysis to get a comprehensive picture of sex-effects in this data-driven metabolic network (Figure 3). We applied a cut-off of $r = 0.3$ (r = partial correlation coefficient) in order to emphasize strong inter-metabolite effects. We observe a general structuring of the network into members from similar metabolic classes, e.g. the amino acids, the phosphatidylcholines, sphingomyelins and acylcarnitines (Figure 3). Direct correlations between metabolites, as represented by partial correlation coefficients, are rare in this metabolite panel with only around 1% of all partial correlations showing a strong effect above $r = 0.3$ (Figure S4 and S5). For this specific cut-off we obtained 14 non-singleton groups, which can be regarded as independently regulated phenotypes within the measured metabolite panel. Detailed description of the distribution of partial correlations and the group structure in the network can be found in Figure S4 and S5. The low connectedness of the network is in line with findings from *Krumsiek et al. 2011* [11] who demonstrated that Gaussian graphical models are sparsely connected on the one hand, but specifically exclude indirect metabolic interactions on the other hand.

Strikingly, sex-specific effects appear to be localized with respect to metabolic classes and connections in the partial correlation matrix. For instance, while most sphingomyelin concentrations have been shown to be higher in females, we also observe them to be a connected component in the GGM. Similarly, acylcarnitines are higher in males and also share partial correlation edges, mostly with other acylcarnitines (Figure 3). Interestingly, we observed three metabolite pairs from the PC aa and lyso-PC classes, respectively, which constitute a side chain length difference of 18 carbon atoms (yellow shaded metabolite pairs, Figure 3).

Genotypic Metabotype Differences between Males and Females

For the identification of differences in genetically determined metabotypes, we used a subpopulation of 1809 participants of the

Table 1. Phenotypic metabotype differences between males and females of the discovery set (KORA F4) and the replication study (KORA F3).

	Discovery			Replication			Metaanalysis	
metabolites	**ß**	***p-value***	**r^2**	**ß**	***p-value***	**r^2**	**ß**	***p-value***
acylcarnitines								
C18	0.146	5.6E-57	21.1%	0.092	3.6E-04	8.4%	0.140	2.5E-61
C10	0.089	2.3E-10	7.9%	0.068	1.0E-01	7.4%	0.087	5.8E-11
C10.1	0.088	5.2E-14	15.9%	0.061	1.0E-01	10.2%	0.085	1.3E-14
amino acids								
xLeu	0.206	1.6E-190	30.2%	0.165	1.1E-15	22.9%	0.202	3.8E-235
Val	0.142	1.9E-78	23.9%	0.096	2.4E-07	18.6%	0.136	5.4E-88
Gly	−0.130	9.1E-46	10.9%	−0.112	2.4E-06	11.1%	−0.128	3.4E-52
phosphatidylcholines								
PC aa C32:3	−0.192	1.4E-106	15.6%	−0.272	1.4E-23	24.5%	−0.200	1.3E-138
PC aa C28:1	−0.133	1.1E-53	8.5%	−0.219	4.7E-18	18.8%	−0.143	1.8E-71
PC ae C40:3	−0.160	5.0E-99	18.7%	−0.177	2.6E-14	16.0%	−0.161	3.0E-120
PC ae C30:2	−0.152	9.1E-53	8.1%	−0.214	1.1E-22	22.8%	−0.164	4.2E-77
lysophosphatidylcholines								
lysoPC a C20:4	0.191	5.4E-62	10.8%	0.125	9.7E-05	8.6%	0.184	2.1E-67
lysoPC a C18:2	0.183	6.2E-55	22.6%	0.136	4.7E-05	17.6%	0.178	1.8E-60
lysoPC a C18:1	0.145	1.4E-41	12.7%	0.106	1.9E-04	16.3%	0.140	1.5E-45
sphingomyelins								
SM (OH) C22:2	−0.228	1.1E-124	19.6%	−0.274	3.5E-25	27.3%	−0.234	1.7E-163
SM C18:1	−0.200	1.3E-101	20.1%	−0.266	3.4E-26	27.0%	−0.209	1.1E-136
SM C20:2	−0.283	7.5E-100	17.7%	−0.280	6.8E-26	25.8%	−0.282	1.0E-135
hexoses								
H1	0.065	6.2E-27	10.5%	0.029	1.6E-01	7.4%	0.062	3.0E-27

P-values were calculated by a linear regression model with metabolites as dependent variable and sex as explanatory variable adjusted for age and BMI. Presented is a set of results of highly significant metabolite concentration differences between males and females of each metabolite subclass out of the 131 tested metabolites. A full list of results for all metabolites and additional information on the complete metabolite panel is provided as supplementary data (Table S2 and S3). Significance level after Bonferroni-correction is *p-value* = 3.8×10^{-4}.

C5 = valerylcarnitine, C0 = carnitine, C18 = octadecanoylcarnitine, xLeu = isoleucine+leucine, Val = valerine, Gly = glycine, PC aa Cx:y = phosphatidylcholine diacyl x:y, PC ae Cx:y = phosphatidylcholine acyl-akyl Cx:y, LysoPC a Cx:y = lysophosphatidylcholine acyl Cx:y, SM (OH) Cx:y = hydroxyshingomyeline Cx:y, SM Cx:y = shingomyelin Cx:y; ß = beta-estimate of linear regression, r^2 = explained variance.

doi:10.1371/journal.pgen.1002215.t001

KORA F4 (Geno-KORA F4) study. The replication was done in an independent subsample of KORA F4 (Rep-KORA F4) and in a subsample of the KORA F3 (Rep-KORA F3) cohort, see Material and Methods for details and Figure S1.

Sex-stratified genome-wide association analysis adjusted for age and BMI were performed for logarithmized concentrations of all metabolites. In order to reveal gender differences we tested the estimated SNP effects for heterogeneity between men and women (see Material and Methods, Figure 4 and Figure S2). We applied a Bonferroni-corrected genome-wide significance level of $5\times10^{-8}/131 = 3.8\times10^{-10}$. All SNPs with a minor allele frequency (maf) lower than 1% in men or in women were excluded. Moreover, SNPs with a low quality of imputation (rsq<0.4) were also excluded. Eight SNPs on chromosome 2 showed genome-wide significant differences in SNP effects (beta-estimates) between men and women for association with glycine (Table 2). The absolute beta-estimates of all eight significant SNPs were higher in women compared to men. The strongest gender difference was seen at SNP rs715 with a genome-wide significant *p-value* of 3.65×10^{-10} for the test of beta-estimate differences. For men the observed effect of rs715 was -0.067 and for women −0.2 (Table 2). SNP rs715 is part of the 3′ UTR region of the *CPS1* gene. SNP rs7422339 with a *p-value* of 3.24×10^{-11} for the test of beta-estimate differences is in a non synonymous coding region of *CPS1*. All other significant SNPs are intergenetic but located in the same region (Table S5). Local association plots for the association of this region with glycine for males and females are presented in Figure S3. The differences in beta-estimates remained significant for GWAs with adjustment for WHR instead of BMI or BMI and WHR combined (Table S4).

The significant differences in beta-estimates between men and women for the association of the eight reported SNPs with glycine were replicated in two independent cohorts Rep-KORA F4 and Rep-KORA F3, including 788 women and 758 men. In the first replication cohort Rep-KORA F4 (583 men, 635 women) the absolute beta-estimate for SNP rs7422339 was also higher in women (beta = −0.225) compared to men (beta = −0.081). The absolute difference of the beta-estimates for SNP rs7422339 was with 0.144 similar to the difference observed in the discovery sample. The *p-value* of the test for difference in beta-estimates of the replication was 1.3×10^{-13}. The other seven SNPs were not available for the Rep-KORA F4 cohort due to other genotyping

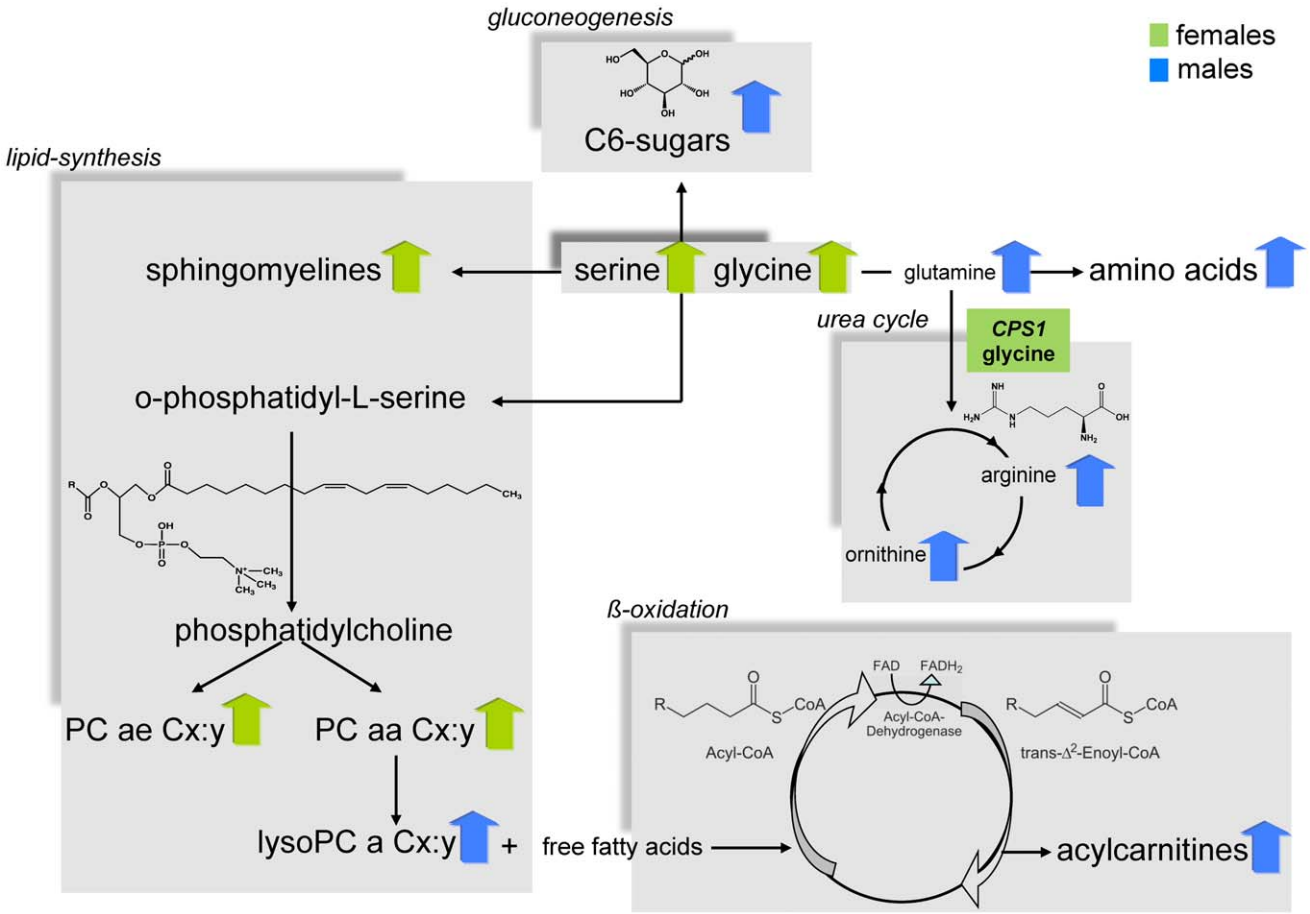


Figure 2. Systematic view of metabotype variations in the metabolism of males and females. It also shows the suggestive locus that is located in a gene encoding an enzyme that is central in human metabolism. *CPS1* is related to the amino acid metabolism. For this locus the metabolite with the strongest association is provided (green box). A blue arrow indicates metabolite concentrations which are higher in men than in women; green arrows vice versa.
doi:10.1371/journal.pgen.1002215.g002

methods. In the second but smaller Rep-KORA F3 cohort seven of the eight SNPs were available. For SNP rs7422339 we observed that the absolute beta-estimate in men (beta = −0.115) was also lower than the absolute beta-estimate in women (beta = −0.229). The absolute difference of the beta-estimates for SNP rs7422339 was 0.144 similar to the absolute difference in beta-estimates observed in the discovery and the first replication cohort (Rep-KORA F4). The *p-value* for the test of difference in beta-estimates was not significant in Rep-KORA F3 (*p-value* = 0.032). For the remaining six SNPs, which were taken forward for replication in Rep-KORA F3 the beta-estimates were also lower in males compared to females but the *p-values* of the test of beta-differences between men and women were not significantly replicated in the Rep-KORA F3 cohort (Table 2).

Discussion

There have been only few studies addressing metabolic differences between males and females, and most of these studies were rather small in sample size and determined only a small number of metabolites [5,12]. We investigated a large population-based study with sufficient statistical power to examine associations within subgroups and a large number of metabolites. Our findings shed light on sex-specific architecture of the human metabolome and provide clues on biochemical mechanisms that might explain observed differences in susceptibility and time course of the development of common diseases in males and females. Our data provided new insights into sex-specific metabotype differences. Combining results from linear regression with partial correlation analysis (resulting in a Gaussian graphical model) yielded interesting insights into how sex-specific concentration differences spread over the metabolic network (Figure 3). The analysis suggests that sex-specific concentration differences affect whole metabolic pathways rather than being randomly spread over the different metabolites. In addition, we found three interesting inter-class associations between PCaa/PCae species and lyso PC species (highlighted in yellow in Figure 3). Those pairs shared a strong partial correlation but displayed differential concentration patterns with respect to gender effects. Furthermore, these pairs displayed a fatty acid residue difference of C18:0, indicating that this fatty acid species might be a key compound giving rise to opposing metabolic gender effects.

Direct experimental evidence indicated a role for sphingolipids (sphingomyelins and ceramides) in several common complex chronic disease processes including atherosclerotic plaque formation, myocardial infarction (MI), cardiomyopathy, pancreatic beta cell failure, insulin resistance, coronary heart disease and type 2 diabetes (T2D) [13,14]. Already young children (between birth and 4 years old, with low levels of sex-hormones) may reveal significant sex-specific differences in plasma sphingolipid concentrations [15]. Our observations described new sex-specific

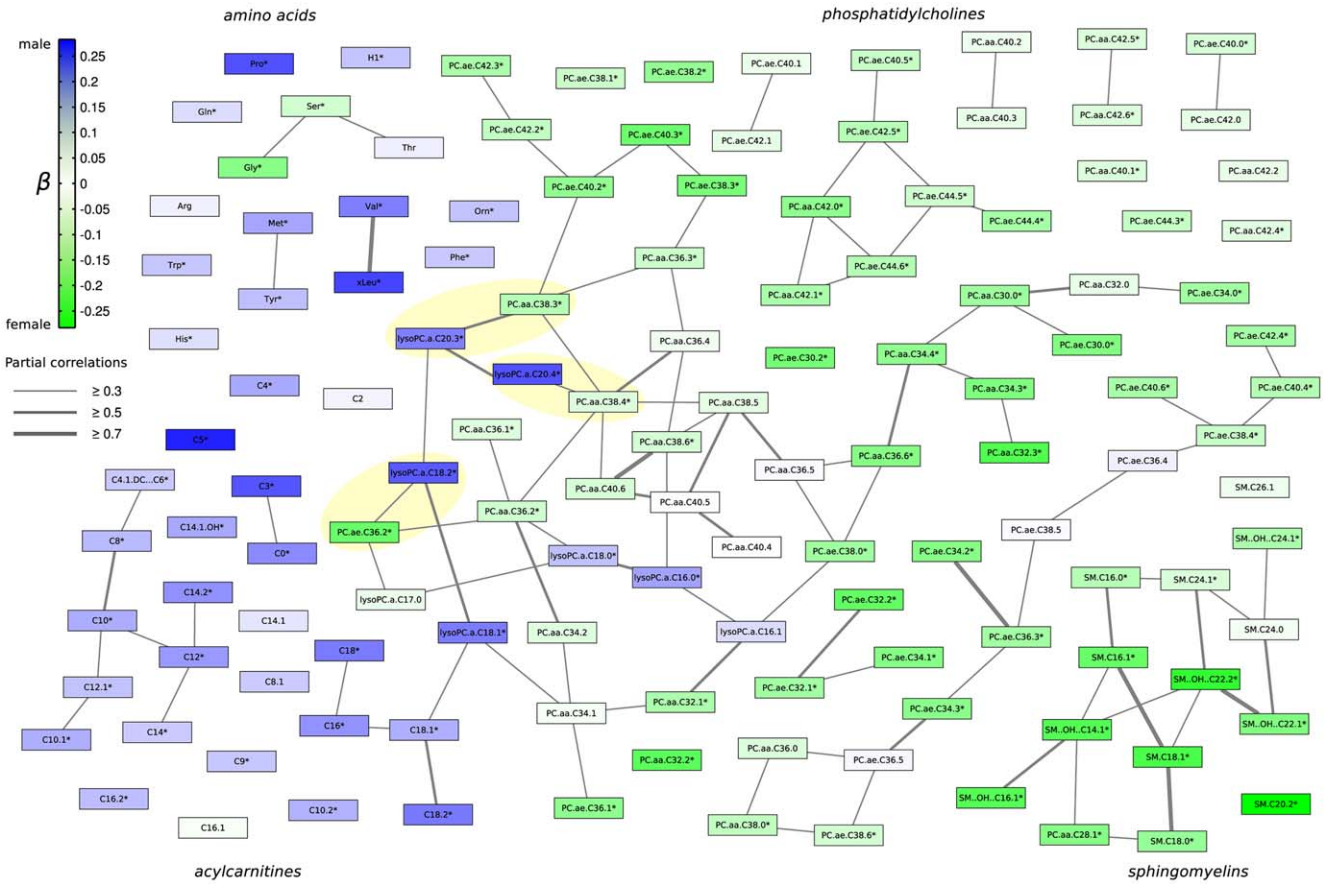


Figure 3. Gaussian graphical model of all measured metabolites illustrating the correlation strength and the propagation of gender-specific effects through the underlying metabolic network. Each node represents one metabolite whereas edge weights correspond to the strength of partial correlation. Only edges with a partial correlation above r = 0.3 are shown. Node colouring represents the strength of association (measured using β from linear regression analysis) towards either males or females. Metabolite names marked with a star * represent significantly different metabolites between genders. Yellow highlighted pairs of metabolites differ by a C18:0 fatty acid residue.
doi:10.1371/journal.pgen.1002215.g003

differences, while other lipid-derived molecules, like bile acids, were already demonstrated not to be sex-specific [16]. Therefore sphingomyelins represent important intermediate phenotypes. The concentration differences between males and females of acylcarnitines described in this study coincide with previous findings showing that carnitine (C0) and acetylcarnitine (C2) concentrations were higher in males than in females[17,18]. Phosphatidylcholines, as demonstrated in this study, are another gender-specific phenotype. Ghrelin (controlling energy homeostasis and pituitary hormone secretion in humans) levels have been shown to be similar in men and women and did not vary by menopausal status or in association with cortisol levels [19]. These findings of our and other studies urgently suggest when using metabolites for disease prediction gender has to be strictly taken into account.

A previously published non sex-stratified GWA study on metabolites based on the same population Geno-KORA F4 reported 15 loci which showed genome-wide significant associations with at least one metabolite concentration or ratio [1]. Besides others the locus *CPS1* was found to have a significant effect on glycine concentrations. But the findings of this sex-stratified genome-wide association analysis revealed that the genetic determination of *CPS1* differs significantly between males and females. Therefore it is important to analyse the data stratified by sex. SNP rs74223369 on chromosome 2 in the 3′ UTR region of the gene *CPS1* showed a genome-wide significant difference in beta-estimates between men and women for association with glycine. The gender-specific effect of SNP rs7422339 was significantly replicated as the difference between the beta-estimates of men and women was of the same direction in the discovery sample and in the replication cohort Rep-KORA F4 and the *p-value* of the test of differences was lower than the replication significance niveau (0.05/8). The other SNPs of the *CPS1* gene region also showed significant gender-specific effects but these effects could not be replicated in the Rep-KORA F3 cohort. As the effect-sizes and differences for the SNP rs7422339 are similar and at least for the other SNPs are pointing into the same direction as in the discovery set, the failed replication in Rep-KORA F3 might be a problem of power due to the smaller sample size.

CPS1, which encodes the mitochondrial enzyme *CPS1*, plays a pivotal role in protein and nitrogen metabolism catalyzing the first committed step of the hepatic urea cycle. Once ammonia has entered the mitochondria via glutamine or glutamate, CPS1 adds the ammonia to bicarbonate along with a phosphate group to form carbamoyl phosphate. Carbamoyl phosphate is then put into the urea cycle. The hepatic urea cycle is responsible for the elimination of ammonia in the form of urea as well as the synthesis of arginine. Among others, Döring *et al.* 2008 could show that there is strong evidence that, in addition to environmental components, a strong genetic and sex-specific control influences the regulation of blood uric acid concentration. They showed that the proportion of the

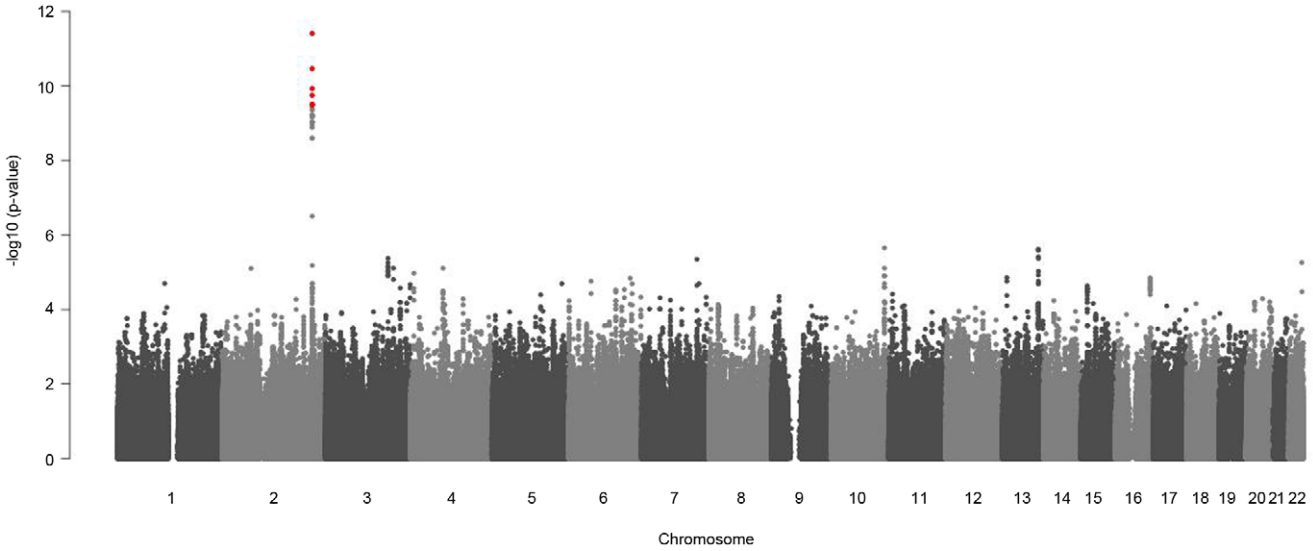


Figure 4. Manhattan plots for gender-specific genome-wide beta-differences for the metabolite glycine. Genome-wide significant beta-differences are plotted in red (significance level $p<3.8\times10^{-10}$).
doi:10.1371/journal.pgen.1002215.g004

variance of serum uric acid concentrations explained by *SLC2A9* genotypes was about 1.2% in men and 6% in women [12]. Brandstätter *et al.* 2010 also observed a sex-specific interaction with genetic association of atherogenic uric acid concentrations [20]. Paré *et al.* 2009 described that the *CPS1* SNP rs7422339, which encodes the substitution of asparagine to threonine (T1405N) in the region critical for N-acetyl-glutamate binding resulting in 20% to 30% higher enzymatic activity [21], is associated with homocysteine also in a sex-specific manner in their study [22]. Also in an Asian population the effects of the genetic variations of the *CPS1* gene were stronger in women than in men [23].

Interestingly a meta-analysis of genome-wide association data in 67,093 individuals also of European ancestry identified recently *CPS1* to affect creatinine production and secretion in Chronic Kidney Disease (CKD) [24]. Hicks *et al.* 2009 performed a GWAS of different circulating sphingolipids in five diverse European populations [3]. They could show associations of genetic loci with several lipid species but did not analyse their data stratified by sex. There is just some evidence for loci with differential sex-effects influencing classical lipids like HDL [25]. Therefore identification of sex-specific genetic variants, that alter the homeostasis of key metabolites in males and in females, will lead to a better functional understanding of the genetics of complex disorders.

As global 'omics'-techniques are more and more refined to identify more compounds in single biological samples, the predictive power of these new technologies will greatly increase. Metabolite concentration profiles and genomic data can be used as predictive biomarkers to indicate the presence or severity of a disease depending on gender. Our study provides new important insights into sex-specific differences of cell regulatory processes and underscores that studies should consider gender-specific effects in design and interpretation. Our findings help to understand biochemical mechanisms underlying sexual dimorphism, a phenomenon which may explain the differential susceptibility to common diseases in males and females.

Materials and Methods

Ethics Statement

Written informed consent has been given by each participant. The study, including the protocols for subject recruitment and assessment and the informed consent for participants, was reviewed and approved by the local ethical committee (Bayerische Landesärztekammer).

Study Population

The KORA S4 survey, an independent population-based sample from the general population living in the region of Augsburg, Southern Germany, was conducted in 1999/2001. The standardized examinations applied in the survey (4261 participants) have been described in detail elsewhere [1,9,26]. A total of 3080 subjects participated in a follow-up examination of S4 in 2006-08 (KORA F4), comprising individuals who, at that time, were aged 32–81 years (Figure S3). In a sample of 3061 individuals metabolomics data was available. This subgroup was used as discovery sample of the phenotypic analysis. For the GWAS the 3061 KORA F4 individuals with metabolomics measurements were divided into two subgroups. First, a subgroup of 1809 individuals from KORA F4 with, who were genotyped using a genome-wide SNP-array (see section genotyping and imputation): Geno-KORA F4. Second, a subgroup of 1218 individuals from KORA F4 with genotyping data generated by Metabo-Chip from Illumina: Rep-KORA F4. Since Geno-KORA F4 and Rep-KORA F4 are non overlapping subgroups of individuals from the KORA F4 cohort they can be considered independent. Therefore it was possible to take Geno-KORA F4 as discovery sample and Rep-KORA F4 as first replication sample for the GWAS.

The KORA F3 cohort is a ten years follow-up survey of the KORA S3 survey examined in 1994–1995 as described previously [26,27]. For the replication step of the phenotypic analysis randomly selected 197 males and 180 females (aged 55–79 years) from the KORA F3 cohort were taken. For 328 individuals (175 males, 153 females) genome-wide genotypes were available. These were used as second replication cohort for the GWAS. No evidence of population stratification was found in multiple published analyses using the KORA cohort [28]. The KORA F3 and F4 surveys are completely independent with no overlap of individuals (Figure S3).

Blood Sampling

Blood samples for metabolic analysis were collected during the years 2006 and 2008 in parallel with the KORA F4 examinations

Table 2. List of SNPs with significant differences in beta-estimates between men and women for association with glycine observed in Geno-KORA F4.

		Geno-KORA F4			Rep-KORA F4			Rep-KORA F3			combined		
SNP	**effect allele**	**effect men**	**effect women**	**pval (beta diff)**	**effect men**	**effect women**	**pval (beta diff)**	**effect men**	**effect women**	**pval (beta diff)**	**effect men**	**effect women**	**pval (beta diff)**
rs715	T	−0.067± (0.012)	−0.206± (0.016)	3.65E-12	-	-	-	-	-	-	−0.067± (0.012)	−0.206± (0.016)	3.65E-12
rs7422339	C	−0.078± (0.013)	−0.22± (0.017)	3.24E-11	−0.081± (0.012)	−0.225± (0.015)	1.30E-13	−0.115± (0.031)	−0.229± (0.043)	0.03151	−0.082± (0.009)	−0.223± (0.011)	2.12E-24
rs10172053	T	−0.043± (0.012)	−0.172± (0.016)	1.12E-10	-	-	-	−0.113± (0.033)	−0.113± (0.047)	1	−0.051± (0.011)	−0.166± (0.015)	1.19E-09
rs7424145	G	−0.041± (0.011)	−0.165± (0.016)	1.70E-10	-	-	-	−0.109± (0.032)	−0.161± (0.045)	0.34633	−0.048± (0.01)	−0.165± (0.015)	2.17E-10
rs10490325	G	0.043± (0.012)	0.169± (0.016)	2.98E-10	-	-	-	0.107± (0.032)	0.133± (0.045)	0.63774	0.051± (0.011)	0.165± (0.015)	1.29E-09
rs2160847	T	−0.036± (0.012)	−0.162± (0.016)	2.98E-10	-	-	-	−0.114± (0.034)	−0.126± (0.046)	0.83384	−0.045± (0.011)	−0.158± (0.015)	1.77E-09
rs2216405	G	0.043± (0.012)	0.169± (0.016)	2.98E-10	-	-	-	0.107± (0.032)	0.127± (0.045)	0.7172	0.051± (0.011)	0.164± (0.015)	1.62E-09
rs4673546	T	0.038± (0.011)	0.155± (0.015)	3.18E-10	-	-	-	0.105± (0.031)	0.149± (0.044)	0.41365	0.046± (0.01)	0.154± (0.014)	6.13E-10

Replication results for these SNPs in Rep-KORA F4 and Rep-KORA F3 are also presented. Not all SNPs were available for all replication cohorts, because different genotyping and imputation methods were used. For the combined analysis the sex-specific effects of all three studies are metaanalyzed and the beta-difference is calculated based on these sex-specific meta-analysis beta-estimates.

doi:10.1371/journal.pgen.1002215.t002

as described in [1,2], and were deep frozen at −80°C until metabolomic analysis. To avoid variation due to circadian rhythm, the blood samples were drawn in the morning between 8:00 and 10:00 am after overnight fasting. Material was immediately horizontal shaken (10 min), followed by 40 min resting at 4°C to obtain complete coagulation. The material was then centrifuged (2000 g; 4°C). Serum was aliquoted and kept for 2–4 hours at 4°C, after which it was deep frozen to −80°C until sampling.

Metabolite Measurements

Metabolomic analysis was performed on 3061 subjects from the population-based cohort KORA F4 and on 377 subjects of the population-based cohort KORA F3. Men and women were collected in a random order and samples were randomly put on plates to exclude batch effects.

Liquid handling of serum samples (10 µl) was performed with Hamilton Star (Hamilton Bonaduz AG, Bonaduz, Switzerland) robot and prepared for quantification using the *AbsoluteIDQ* kit (BIOCRATES Life Sciences AG, Innsbruck, Austria) as described previously [1]. Sample analysis were done on API 4000 QTrap LC/MS/MS System (Applied Biosystems, Darmstadt, Germany) equipped with Schimadzu Prominence LC20AD pump and SIL-20AC auto sampler. The complete analytical process was monitored with the *MetIQTM* software package, which is an integral part of the AbsoluteIDQ^{TM} kit.

Metabolite Panel

In total, 163 different metabolites were quantified. More information about the metabolite panel can be found in Text S1. Metabolite measurements of the 3061 samples were performed in three batches, with two and three months time lapse in between, respectively. Within each kit, there are three different quality controls (QCs) representing gender mixed human plasma samples provided by the manufacturer. In accordance with the kit instructions, concentration of each metabolite was adjusted based on the three QCs to minimize the potential batch effects.

To ensure data quality, metabolites had to meet three criteria: (1) average value of coefficient of variance (CV) of the three QCs should be smaller than 25%. (2) 90% of all measured sample concentrations should be above the limit of detection (LOD). (3) Correlation coefficients between two duplicated measurements of 144 re-measured samples should be above 0.5 (Table S6). In total, 131 metabolites passed the three quality controls. To detect sample outliers, the data of the 131 metabolite concentrations were first scaled to zero mean and unity standard deviation and were projected onto the unit sphere and Mahalanobis distances were then obtained. Robust principal components algorithm was used in the process [29]. Mean and variance were then calculated for the distances. A cut-off was set at 3 times variance plus mean distance. Any individual, whose distance was greater than this cut-off, was marked as an outlier and removed. Outliers were detected separately for males and females. 131 Metabolites and 3004 samples remained in the dataset. Missing values were using the R package ''mice''. Metabolite concentrations were logarithmized for all subsequent analysis steps.

Genotyping and Imputation

In KORA F4 genome-wide genotyping was done using the Affymetrix 6.0 GeneChip array (Geno-KORA F4). The algorithm Birdseed2 was used for calling. Genotyped SNPs were filtered for an individual call rate of 0.93, SNP call rate 0.93 and Hardy-Weinberg equilibrium (*P*HWE >0.001). All remaining SNPs (651,596) were used for imputation with MACH (v1.0.15). HapMap CEU version 22 was used as reference population for calling and imputation. The GWAS replication cohort Rep-KORA F4 was genotyped on Metabo-Chip, with calling algorithm GenomeStudio. The second GWAS replication sample KORA F3 was genotyped with the Affymetrix 500 K array. The calling was performed by BRLMM with reference population HapMap CEU 21. After filtering for individual call rate 0.93 and SNP call rate 0.9 and Hardy-Weinberg equilibrium (P_{HWE} >0.001) the remaining SNPs were imputed with MACH v1.0.9 using HapMap CEU version 21 as reference population.

Statistical Analysis

Partial least squares (PLS). PLS, or projection to latent structures by means of partial least squares, and is a method to relate a matrix X to a vector y (or to a matrix Y).The x-data are transformed into a set of a few intermediate linear latent variables (components). PLS analysis [10] was carried out using the *R* package *pls* to investigate the metabolic profiles of males and females. Data was visualized by plotting the scores of the first two components against each other, where each point represented an individual serum sample. For this analysis, metabolite concentrations were normalized to have a mean of zero and a standard deviation of one.

Delta (difference in concentration means for men and women). For comparison of metabolite concentrations between men and women we used the delta (Δ), as it describes the difference in concentration means for men and women for a specific metabolite relative to the mean metabolite concentration in men. Therefore the difference of mean metabolite concentration in men and mean metabolite concentration in women is calculated and than divided by the mean metabolite concentration in men. For example, a value of $\Delta = 50\%$ means, that the mean metabolite concentration in women is 50% lower than the metabolite concentration in men.

Linear regression. Metabolite concentration differences between males and females were investigated by linear regression analysis. The basic model contains the log-transformed metabolite as dependent, sex as explanatory variable and both age and BMI as covariates. Moreover, an internal batch variable is included to account for possible systematic differences that might have been caused by the metabolite measuring process. To correct for multiple testing Bonferroni-correction was applied. That means the influence of sex on a specific metabolite was called significant, if the *p-value* of the corresponding test of sex having no effect on (log-transformed) metabolite is lower than $0.05/131 = 3.8\times10^{-4}$. For replication we also applied Bonferroni-correction. That means a difference in sex on a specific metabolite is called significant, if the direction of the effect in consistent between discovery and replication cohort and the *p-value* for sex having no effect on the metabolite is lower than 0.05 corrected for the number of metabolites taken forward for replication.

We analysed the influence of anthropometric phenotypes, diseases and environmental factors by including different covariates to the linear regression and comparison of the structure of the results. Four models which differ in the use of one or more additional covariates were performed. The covariates in each model beside age are waist hip ratio (WHR), lipid parameters (HDL and LDL cholesterol, triglycerides), type 2 diabetes, alcohol consumption and smoking. All calculations were performed in R with standard procedures (lm). Furthermore, a meta-analysis of the discovery and the replication sample with a fixed effect model was analyzed to reveal the sex-specific effects of metabolite concentrations.

Partial correlation analysis. In order to investigate how strong the different metabolites correlate with each other and the

sex-specific effects propagate through the underlying metabolic network, we calculated full-order partial correlation coefficients (r) between all pairs of metabolites. The resulting partial correlation networks are commonly referred to as Gaussian graphical models (GGMs), which we have previously demonstrated to be useful for the analysis of direct metabolite-metabolite effects in the same population cohort [11]. The GGM was coloured and annotated according to the β-values and *p-values* from linear regression analysis and then exported and visualized using the free yEd graph editor.

Genome-wide association studies (GWAS). We calculated GWAS for all 131 metabolites with mach2qtl (v1.0.8) for men and women separately. We applied an additive model with covariates age, BMI and an internal variable accounting for batch effects.

Genome-wide test for sex-specific differences in beta-estimates. We tested each SNP and metabolite for equality of the beta-estimates for the SNP calculated in the sex-specific GWAS. Therefore we used an approximately normally distributed test statistic [30]:

$$\frac{\beta_{men}-\beta_{women}}{\sqrt{se(\beta_{men})^2+se(\beta_{women})^2}}$$

To take our 131 phenotypes into account we used Bonferroni correction. Therefore the genome-wide significance level is $5\times10^{-8}/131 = 3.8\times10^{-10}$ [31].

Replication of sex-specific differences in genetic effects. We confirm a genetic sex-specific difference as replicated, if the proportion of the absolute SNP effects in men and women is the same as in the discovery sample, and the *p-value* for the test for difference in effects is lower than the adjusted *p-value*.

Replication in KORA F4. We used PLINK for the calculation.

Further analysis. All further analyses were performed in R. For both subgroups, men and women, we calculated the frequencies of each SNP. The explained variance (R^2) for each SNP and metabolite was calculated as the difference of the coefficients of determination of the model with SNP and without SNP. The metaanalysis of SNP association with glycine was performed with METAL (http://www.sph.umich.edu/csg/abecasis/Metal/index.html) for men and women separately using the inverse variance weighting.

Supporting Information

Figure S1 KORA study populations with subsamples used in this study.
(TIFF)

Figure S2 QQ-plots for the sex-stratified GWAS with glycine. The QQ-plot shows the p-values of the sex stratified GWAS for glycine in the discovery sample Geno-KORA F4 versus the expected *p-values* under the null hypotheses of no SNP having an effect on glycine.
(JPG)

Figure S3 Regional association plots for sex-stratified GWAS with glycine around the locus *CPS1*. Association *p-values* of SNPs with glycine for men and women are presented for a region surrounding rs715, which had the strongest difference in beta-estimates between men and women. SNPs with genome-wide significant differences in beta-estimates are highlighted in blue. The level of linkage disequilibrium of rs715 with other SNPs is indicated by circle colour ranging from red $r^2>0.8$, orange $0.8>r^2>0.5$, grey $0.5>r^2>0.2$ to white $0.2>r^2$.
(TIFF)

Figure S4 Distribution of partial correlation coefficients. Partial correlations center around zero with a shift towards positive high values. When applying a correlation cutoff of $r = 0.3$, we are left with 109 out of 8515 correlation values (1.28%).
(TIFF)

Figure S5 Number of clustered groups in the GGM as a function of the absolute partial correlation cutoff. Note that we did not count singleton metabolites that is metabolites without any partial correlation above threshold, here. Most non-singleton groups emerge in the cutoff range between 0.3 and 0.7, which corresponds to the figure in the main manuscript. For our lower cutoff of 0.3, we obtain 14 groups, which can here be regarded as *independent phenotypes* in the metabolite pool.
(TIFF)

Table S1 Study population characteristics. Data are presented as mean (SD) or number of persons (N); BMI indicates body mass index; HDL high density lipoprotein; LDL low density lipoprotein; smokers: number of smokers with one or more than one cigarette/day, high alcohol intake: subjects were counted for high alcohol intake when they had an alcohol consumption of ≥0 g alcohol/day for males and ≥20 g alcohol/day for females. (A) Study populations used for phenotypic analysis. (B) Study populations used for genotypic analysis.
(DOCX)

Table S2 Phenotypic metabotype differences between males and females of the discovery sample KORA F4. *P*-values were calculated by a linear regression model with metabolite concentration as outcome and sex as explanatory variable adjusted for different covariables. Gray shaded columns show significant *p-values* for differences in the metabolite concentrations between males and females after Bonferroni correction (significance level after multiple testing correction = $p\text{-value}<3.8\times10^{-4}$).
(DOCX)

Table S3 Phenotypic metabotype differences between males and females of the replication sample KORA F3. *P*-values were calculated by a linear regression model with metabolite concentration as outcome and sex as explanatory variable adjusted for age, BMI and waist-hip ratio (WHR). Gray shaded columns show significant *p-values* for differences in the metabolite concentrations between males and females after Bonferroni correction (significance level after multiple testing = $p\text{-value}<3.8\times10^{-4}$).
(DOCX)

Table S4 Comparison of different adjustments in association of SNPs with glycine. Results for SNPs which showed a significant difference in beta-estimates for KORA F4 with the adjustment of sex-specific GWAs for BMI (age, batch), for different adjustment for waist-hip ratio (WHR) (age, batch) or adjustment for WHR and BMI (age, batch).
(DOCX)

Table S5 Detailed information for SNPs with significant gender differences in beta-estimates for association with glycine. Minor allele frequency is calculated for men and women separately in each study. Imputation quality (RSQ) respectively call rate for genotyped SNPs is calculated based on all IDs.
(DOCX)

Table S6 Full biochemical names of all 131 metabolites used for further analysis that were measured on the Biocrates Absolute

IDQ kit. Abbreviations and full biochemical names of the 131 metabolites are shown in the first and second columns, respectively. The third column lists shows the correlation coefficients (r) between two duplicated measurements of 144 re-measured samples. The following column shows percentage of 3061 individuals above limit of detection (LOD) and the last column shows the mean value of the correlation coefficient (CV) of the three quality controls for the three batches.
(DOCX)

Table S7 Excluded metabolites that were measured on the Biocrates AbsoluteIDQ kit. Abbreviations and full biochemical names of the excluded metabolites are shown in the first and second columns, respectively. The third column shows the correlation coefficients (r) between two duplicated measurements of 144 re-measured samples. The following column shows percentage of 3061 individuals above limit of detection (LOD). Mean value of the correlation coefficient (CV) of the three quality controls for the three batches is shown in the last column.
(DOCX)

Table S8 Metabolite concentrations of the study cohorts KORA F4 and KORA F3 and the relative sex-specific difference (Δ in %). $\Delta =$ (Mean(metabolite concentration of men) − mean(metabolite concentration of women)) /mean (metabolite concentration of men); difference of metabolite concentrations between men and women in %.
(DOCX)

Text S1 Metabolite panel.
(DOC)

Acknowledgments

We gratefully acknowledge the contributions of Dr. P. Lichtner, G. Eckstein, G. Fischer, Dr. T. Strom, and all other members of the Helmholtz Zentrum München genotyping staff for generating the SNP dataset; Julia Henrichs, Tamara Halex, and Arsin Sabunchi of the Genome Analysis Center Metabolomic Platform; as well as the contribution of all members of field staffs who were involved in the planning and conducting the MONICA/KORA Augsburg studies. We gratefully acknowledge the KORA study group, Institute of Epidemiology, Helmholtz Center Munich, Center for Environment and Health, Neuherberg, Germany, consisting of H. E. Wichmann (speaker), A. Peters, C. Meisinger, T. Illig, R. Holle, J. John, and their co-workers and all individuals who participated as cases or controls in this study, and the KORA Study Center and their co-workers for organizing and conducting the data collection.

Author Contributions

Conceived and designed the experiments: T Illig, K Mittelstrass, JS Ried, J Adamski, R Wang-Sattler, J Krumsiek, FJ Theis, S Weidinger. Performed the experiments: R Wang-Sattler, J Adamski, T Illig, K Mittelstrass, J Krumsiek, JS Ried, Z Yu. Analyzed the data: K Mittelstrass, C Prehn, J Krumsiek, JS Ried, F Kronenberg, Z Yu. Contributed reagents/materials/analysis tools: J Adamski, R Wang-Sattler, T Illig, T Meitinger, W Roemisch-Margl, C Gieger, K Suhre, HE Wichmann, A Polonikov, A Peters. Wrote the paper: K Mittelstrass, JS Ried, J Krumsiek.

References

1. Illig T, Gieger C, Zhai G, Römisch-Margl W, Wang-Sattler R, et al. (2010) A genome-wide perspective of genetic variation in human metabolism. Nat Genet 42(2): 137–141.
2. Gieger C, Geistlinger L, Altmaier E, Hrabé de Angelis M, Kronenberg F, et al. (2008) Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. PLoS Genet 4: e1000282. doi:10.1371/journal.pgen.1000282.
3. Hicks AA, Pramstaller PP, Johansson A, Vitart V, Rudan I, et al. (2009) Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet 5: e1000672. doi:10.1371/journal.pgen.1000672.
4. Kim AM, Tingen CM, Woodruff TK (2010) Sex bias in trials and treatment must end. Nature 2010 Jun 10;465(7299): 688–689.
5. Geller SE, Adams MG, Carnes M (2006) Adherence to federal guidelines for reporting of sex and race/ethnicity in clinical trials. J Womens Health (Larchmt) 2006 Dec;15(10): 1123–1131.
6. Tingen CM, Kim AM, Wu P-H, Woodruff TK (2010) Sex and sensitivity: the continued need for sex-based biomedical research and implementation. Womens Health (Lond Engl) 2010 Jul;6(4): 511–516.
7. Fairweather D, Rose NR (2004) Women and autoimmune diseases. Emerging Infect. Dis 2004 Nov;10(11): 2005–2011.
8. Mostertz W, Stevenson M, Acharya C, Chan I, Walters K, et al. (2010) Age- and sex-specific genomic profiles in non-small cell lung cancer. JAMA 2010 Feb 10;303(6): 535–543.
9. Holle R, Happich M, Löwel H, Wichmann HE (2005) KORA–a research platform for population based health research. Gesundheitswesen 2005 Aug;67(Suppl 1): S19–25.
10. Lorber A, Wangen LE, Kowalski BR (1987) A theoretical foundation for the PLS algorithm. Journal of Chemometrics 1987;1(1): 19–31.
11. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Systems Biology 2011;5(1): 21.
12. Döring A, Gieger C, Mehta D, Gohlke H, Prokisch H, et al. (2008) SLC2A9 influences uric acid concentrations with pronounced sex-specific effects. Nat Genet 2008 Apr;40(4): 430–436.
13. Holland WL, Summers SA (2008) Sphingolipids, insulin resistance, and metabolic disease: new insights from in vivo manipulation of sphingolipid metabolism. Endocrine reviews 2008;29(4): 381.
14. Yeboah J, McNamara C, Jiang XC, Tabas I, Herrington DM, et al. (2010) Association of plasma sphingomyelin levels and incident coronary heart disease events in an adult population: Multi-Ethnic Study of Atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology 2010;30(3): 628.
15. Nikkilä J, Sysi-Aho M, Ermolov A, Seppänen-Laakso T, Simell O, et al. (2008) Gender-dependent progression of systemic metabolic states in early childhood. Molecular systems biology 2008;4(1).
16. Rodrigues CM, Kren BT, Steer CJ, Setchell KD (1996) Formation of delta 22-bile acids in rats is not gender specific and occurs in the peroxisome. Journal of lipid research 1996;37(3): 540.
17. Slupsky CM, Rankin KN, Wagner J, Fu H, Chang D, et al. (2007) Investigations of the effects of gender, diurnal variation, and age in human urinary metabolomic profiles. Analytical chemistry 2007;79(18): 6995–7004.
18. Reuter SE, Evans AM, Chace DH, Fornasini G (2008) Determination of the reference range of endogenous plasma carnitines in healthy adults. Annals of clinical biochemistry 2008;45(6): 585.
19. Purnell JQ, Weigle DS, Breen P, Cummings DE (2003) Ghrelin levels correlate with insulin levels, insulin resistance, and high-density lipoprotein cholesterol, but not with gender, menopausal status, or cortisol levels in humans. Journal of Clinical Endocrinology & Metabolism 2003;88(12): 5747.
20. Brandstätter A, Lamina C, Kiechl S, Hunt SC, Coassin S, et al. (2010) Sex and age interaction with genetic association of atherogenic uric acid concentrations. Atherosclerosis 2010;210(2): 474–478.
21. Summar ML, Hall L, Christman B, Barr F, Smith H, et al. (2004) Environmentally determined genetic expression: clinical correlates with molecular variants of carbamyl phosphate synthetase I. Mol Genet Metab 2004 Apr;81(Suppl 1): S12–19.
22. Paré G, Chasman DI, Parker AN, Zee RRY, Mälarstig A, et al. (2009) Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population: a genome-wide evaluation of 13 974 participants in the Women's Genome Health Study. Circ Cardiovasc Genet 2009 Apr;2(2): 142–150.
23. Lange LA, Croteau-Chonka DC, Marvelle AF, Qin L, Gaulton KJ, et al. (2010) Genome-wide association study of homocysteine levels in Filipinos provides evidence for CPS1 in women and a stronger MTHFR effect in young adults. Hum Mol Genet 2010 May 15;19(10): 2050–2058.
24. Köttgen A, Pattaro C, Böger CA, Fuchsberger C, Olden M, et al. (2010) New loci associated with kidney function and chronic kidney disease. Nat Genet 2010 May;42(5): 376–384.
25. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, et al. (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 2009 Jan;41(1): 47–55.
26. Wichmann H-E, Gieger C, Illig T (2005) KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen 2005Aug ;67(Suppl 1): S26–30.
27. Löwel H, Meisinger C, Heier M, Hörmann A (2005) The population-based acute myocardial infarction (AMI) registry of the MONICA/KORA study region of Augsburg. Gesundheitswesen 2005 Aug;67(Suppl 1): S31–37.
28. Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, et al. (2006) SNP-based analysis of genetic substructure in the German population. Hum Hered 2006;62(1): 20–29.

29. Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Computational Statistics & Data Analysis 2008;52(3): 1694–1711.
30. Paternoster R, Brame R, Mazerolle P, Piquero A (1998) Using the correct statistical test for the equality of regression coefficients. Criminology 1998;36(4): 859–866.
31. Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genetic epidemiology 2008;32(4): 381–385.

Gaussian Graphical Modeling Reveals Specific Lipid Correlations in Glioblastoma Cells

Nikola S. Mueller[a], Jan Krumsiek[b], Fabian J. Theis[b], Christian Böhm[c] and Anke Meyer-Baese[d]

[a]Max Planck Institute of Biochemistry, Martinsried, Germany;
[b]Helmholz Zentrum, Neuherberg, Germany;
[c]Ludwig-Maximilians Universität, Munich, Germany;
[d]Florida State University, Tallahassee, USA

ABSTRACT

Advances in high-throughput measurements of biological specimens necessitate the development of biologically driven computational techniques. To understand the molecular level of many human diseases, such as cancer, lipid quantifications have been shown to offer an excellent opportunity to reveal disease-specific regulations. The data analysis of the cell lipidome, however, remains a challenging task and cannot be accomplished solely based on intuitive reasoning. We have developed a method to identify a lipid correlation network which is entirely disease-specific. A powerful method to correlate experimentally measured lipid levels across the various samples is a Gaussian Graphical Model (GGM), which is based on partial correlation coefficients. In contrast to regular Pearson correlations, partial correlations aim to identify only direct correlations while eliminating indirect associations. Conventional GGM calculations on the entire dataset can, however, not provide information on whether a correlation is truly disease-specific with respect to the disease samples and not a correlation of control samples. Thus, we implemented a novel differential GGM approach unraveling only the disease-specific correlations, and applied it to the lipidome of immortal Glioblastoma tumor cells. A large set of lipid species were measured by mass spectrometry in order to evaluate lipid remodeling as a result to a combination of perturbation of cells inducing programmed cell death, while the other perturbations served solely as biological controls. With the differential GGM, we were able to reveal Glioblastoma-specific lipid correlations to advance biomedical research on novel gene therapies.

Keywords: Correlation Networks, Partial Correlations, Gaussian Graphical Models, Lipidomics, Glioblastoma

1. INTRODUCTION

Despite recent progress in therapy and surgical intervention, Glioblastoma multiforms, malignant primary brain tumors, are nearly always fatal. The *in vitro* model of human Glioblastoma brain tumors is the U87 cell line, the major characteristic of which is its resistance to apoptosis (programmed cell death). Recent studies showed that the combined perturbation of gene transfection with the p53 tumor suppressor gene prior to chemotherapy with SN-38 triggers cell death in the (otherwise immortal) Glioblastoma cell line.[1,2] At first a proteomic study showed a down-regulation of Galectin-1 in response to the combined perturbation,[1] which motivated the elucidation of lipid regulations.[2] In order to measure the lipidome, a specialized mass spectrometry (MS) technique was developed.[3] On an organism-wide scale, changes in complex polar lipid levels were reliably identified. The set of all commonly regulated lipids might reveal dysregulations of e.g. metabolic pathways or functionally similar proteins. However, the molecular details of the perturbation-affected lipid coregulations still remain to be elucidated.

We aimed to identify partial correlations of lipid concentrations while accounting for the biological interpretation of the perturbation. To that end, we used Gaussian Graphical Models (GGMs), which are statistical graph models based on partial correlation coefficients. We chose to use a GGM over simple Pearson correlations since correlations are only detected for direct but not indirect dependencies.[4] Beyond conventional GGM analysis,

Further author information: (Send correspondence to A.M-B.: E-mail: amb@eng.fsu.edu)

Independent Component Analyses, Wavelets, Neural Networks, Biosystems, and Nanoengineering IX,
edited by Harold Szu, Liyi Dai, Proc. of SPIE Vol. 8058, 805819 · © 2011 SPIE
CCC code: 0277-786X/11/$18 · doi: 10.1117/12.884196

where one GGM is calculated for the entire data set, we introduce a disease-driven GGM calculation. With this here introduced differential GGM approach, we can now address the question whether a correlation in the GGM is biologically relevant or not. In general, not every identified correlation on the entire dataset is equally relevant to the disease, especially if the majority of the dataset are control measurements. While identifying only those lipids that respond to the biologically relevant perturbations but not to control perturbations, we answered the key question: Which lipids or lipid classes are co-affected by the perturbation by wild-type (wt) p53 transfection prior to SN-38 chemotherapy triggering apoptosis of the brain tumor cell lines?

2. GLIOBLASTOMA AND ITS LIPIDOME

U87 cells transfected with wt tumor suppressor gene p53 prior to treatment with the chemotherapeutic drug SN-38 underwent modest apoptosis and cell cycle arrest in G2, while chemotherapy alone did not trigger the same phenotype.[1] The reverse order of SN-38 treatment prior to p53 transfection results in almost complete apoptosis and complete G2 arrest. To analyze the lipid variations as a response to the effective perturbation, high-throughput MS/MS experiments were conducted as follows. Cell lysates of all perturbed cell lines were analyzed for variations of lipid levels (Fig. 1a).[2,3] A specialized Fourier-Transform Ion-Cyclotron-Resonance (FT-ICR) MS/MS technique was developed to separate complex lipids.[3] With the FT-ICR MS/MS, polar lipids, such as phospholipids, as well as complex glycolipids, such as gangliosides were reliably identified. Quantitative analysis of relative abundance profiles of polar lipids were obtained from cell lysates, whereby lipid levels were measured across six different perturbations and wt (without perturbation) with two technical replicates. Out of the large set of lipids, 167 polar lipids were measured with FT-ICR MS/MS across six lipid classes (varying primarily in their respective head groups). While lipid head groups can uniquely be identified with MS/MS, the associated fatty acid side chains cannot be independently resolved. An example for a complex lipid with ambiguous fatty acid side chains is PS(C36:4) that could have e.g. C18:2/C18:2 fatty acids incorporated, but also C16:0/C20:4 or C16:2/C20:2, etc.. Note, that some lipid classes, like gangliosides, have one variable and one fixed fatty acid side chain, thus, both side chains can unambiguously inferred. The MS/MS result – the matrix to be analyzed in this study – holds concentrations of lipids for each cell line for all perturbations.

Only the combined perturbation of p53 adenoviral transfection prior to SN-38 chemotherapy is biologically relevant for this study. In order to identify those lipids that specifically respond to the combined perturbation a

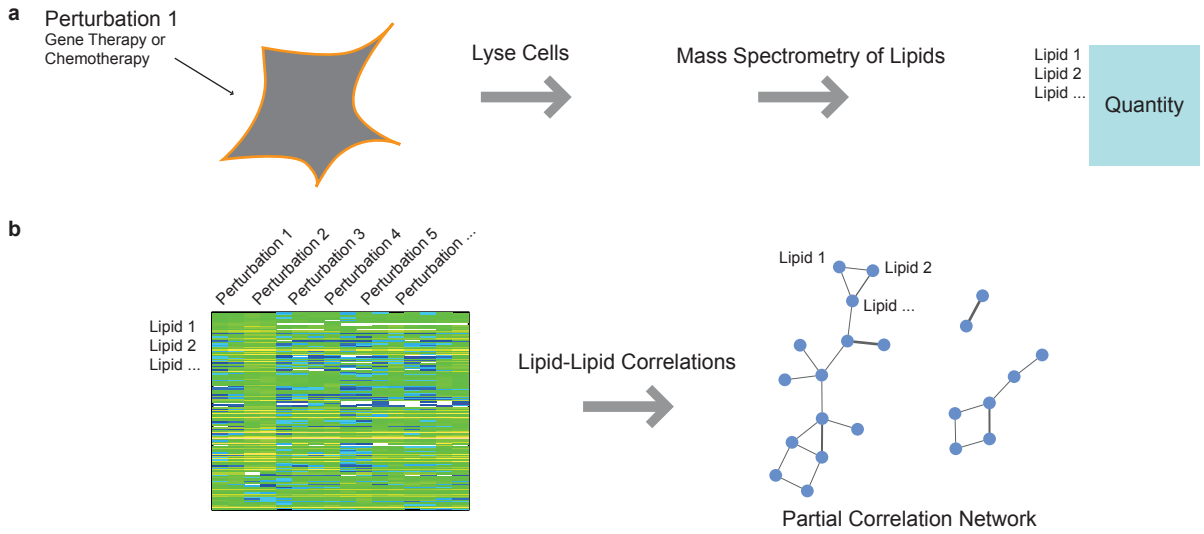


Figure 1. **From Cells to Lipid Correlations. a.** U87 cell lines were perturbed and subsequently lysed prior to MS analysis. Subsequently, lipid concentrations of 167 polar lipid species were obtained. **b.** The raw data of this study holds lipid quantifications for various perturbations (the samples). Pairwise correlations of lipids result in an undirected graph of lipid-to-lipid interactions holding the partial correlation values. Only statistically significant correlations are included in the resulting network. Edge widths indicate correlation strengths.

series of control experiments were conducted, which were permutations of the single perturbations as described previously,[2,5] e.g. SN-38 alone, empty virus transfection or empty virus transfection prior to chemotherapy. In order to unravel the lipid remodeling that effected or was affected by apoptosis of U87 cells, the comparison of wt cell lines with the p53 plus SN-38 perturbations is not sufficient. For example, lipid remodeling can be the result of singular effects, like the transfection of the empty adenovirus, only the wt p53 adenovirus or solely the SN-38 chemotherapy. Only the entire dataset with all perturbations and wt allows to statistically exploit the wealth of all perturbation effects, which might not be feasible by comparing only two biologically relevant perturbations.

3. GAUSSIAN GRAPHICAL MODELS

Traditionally, correlation networks have been used to obtain information on coregulations of variables $L = (l_1, \ldots, l_p)$, $|L| = p$ measured across all samples $S = (s_1, \ldots, s_n), |S| = n$; with $X = (x_{ls})$ the raw data matrix used for calculations. In case of the present metabolite data, a correlation coefficient will provide information on the degree of dependence between the measured variables. This pairwise correlation is thereby calculated based on the measurements across all samples – the cell lines with various perturbations (Fig. 1b).

The standard measure of pairwise correlations are Pearson product-moment correlation coefficients $P = (\rho_{ij})$, which quantify the linear dependency between two variables l_i and l_j. A common problem of Pearson correlation coefficients are indirect effects giving rise to a plethora of unspecifically high correlation coefficients throughout *omics* datasets.[4] GGMs attempt to estimate conditional dependencies between measured variables over all samples rather than marginal dependencies, thereby eliminating such indirect correlations. The derivation of partial correlation coefficients can also be explained by linear regression: The partial correlation between the lipids l_1 and l_2 is the correlation of the residuals that result from linearly regressing l_1 and l_2 against the remaining lipids $(l_3, \ldots, l_p)$.[6] In out study, the partial correlation ζ_{ij} provides information on the coregulation of two lipids l_i, l_j.

To generate a GGM, the number of samples with respect to the number of variables determine the approach used for the calculation. If the number of samples n exceeds the number of variables p, full-order partial correlations $Z = (\zeta_{ij})$ can be calculated in a straight-forward manner from the inverse of the covariance matrix P as

$$\Omega = (\omega_{ij}) = P^{-1}$$

$$Z = (\zeta_{ij}) = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}.$$

Statistical tests are next applied to determine whether a partial correlation ζ_{ij} is significantly different from zero ζ^*_{ij} (we mark a significant partial correlation with an asterisk) resulting in the GGM Z^*. Of the partial correlation matrix Z we construct Z^* as

$$Z^* = (\zeta^*_{ij}) = \begin{cases} \zeta_{ij} & \text{if } \zeta_{ij} \text{ is significant} \\ 0 & \text{else} \end{cases}$$

and we denote $\exists\zeta^*_{ij}$ for $\zeta^*_{ij} > 0$. A GGM is an undirected graph obtained by partial correlation calculation with subsequent statistical testing for edge significance (Fig. 2a). The graph nodes represent the measured variables whereas the edge weights correspond to significant partial correlation coefficients. If the number of samples is smaller than the number of variables ($n < p$) the straight-forward GGM calculation cannot be applied but a regularization and a likelihood estimation step have to be included. For $n < p$ the covariance matrix is rank-deficient,[7–9] as a consequence the covariance matrix is not positive definite and can, thus, not be inverted. In the case of the present lipidomics data, we indeed have $n < p$ with $p = 157$ lipids and $n = 8^*$ samples. To estimate the GGM for $n < p$, Strimmer and colleagues[10] introduced an all-in-one approach. One estimation step is a shrinkage approach and is applied[9,10] to obtain the true correlation matrix $\hat{P}$. The other estimation step distinguishes actually existing edges from "null" edges in the GGM by fitting a statistical model assuming these two population of edges. The GGM is finally build by adjusting for local false-discovery rates (FDR).[9,10] This method of regularized GGMs was already applied to transcriptomics datasets[6,11] and will here be applied to our lipidomics dataset.

*Eight samples were measured with two technical replicates. Analyses were performed on the raw data including the replicates.

When calculating the GGM, all samples are assumed to be independent,[7] but inspection of the present lipidome dataset showed a strong correlation between all samples. Although correlations between the technical replicates were higher than between perturbations, the overall correlation of disease and control samples was very high ($> .95$). In case of dependent samples the covariance estimates are no longer optimal: its standard deviation monotonically increases with larger correlation coefficients of samples.[8] Note that the result of the strong correlation between all samples already indicates that the successful perturbation of cells transfected with wt p53 prior to SN-38 chemotherapy has strong effects only on few lipids and not the lipid levels in general. To account for the high dependencies between samples, we calculated the GGM mimicking that all samples are replicates of one another. Since seven of the eight samples are only measured as controls (which were introduced as control replicates with respect to the one perturbation of interest), this approach is reasonable for our study.

4. DIFFERENTIAL GGM

To identify those partial correlations of lipids only resulting from the biologically relevant perturbation and not from side effects of one perturbation, we implemented the following concept of disease-specificity. For simplicity, we name the biologically relevant perturbation "disease" in contrast to the "controls" in the following, although this combination of perturbation is the one inhibiting tumor cell growth. Let S be the set of n samples composed of control and one disease sample $S = (s_1, \ldots, s_n) = (s_D, s_{C_1}, \ldots, s_{C_{n-1}}) = (s_D, s_{C_\cdot})$ with the disease sample s_D and the union of all control samples $s_{C_\cdot}$. Imagine $\zeta^*(S)$ to be a significant correlation on the entire dataset S. It may then be a result of a perfect correlation of controls not substantially affected by the disease samples or be a result where primarily the disease samples induce a correlation on the entire dataset (controls alone are not correlated). In other words: if a correlation has no specific relevance to the disease, we would still detect a correlation when using a truncated dataset with solely control samples. These correlations, which are mostly a result of strong control sample correlation, can be considered "false positive" with respect to true disease relevance. In order to gather all truly disease relevant correlation, we also have to account for the reverse case, equivalently the "false negatives" with respect to disease relevance. If a correlation exists on the control samples $s_{C_\cdot}$ but is suppressed on the entire dataset S, the disease samples do not follow the correlation of the controls, wherein the correlation is again relevant with respect to the disease. This reverse case corresponds to the concept of suppressed variables, which denotes a variable to be a suppressor if it suppresses the correlation between some other variable to the remaining variables.[12,13]

All disease relevant partial correlations were assessed in an approach inspired by jackknife resampling.[14] GGMs are calculated by leaving out one sample from the dataset ($Z^*_{S\setminus s_i}$) during each iteration, resulting in a set of partial correlation coefficients for each lipid pair (l_i, l_j) of $\{\zeta^*(S), \zeta^*(S\setminus s_D), \zeta^*(S\setminus s_{C_1}), \ldots, \zeta^*(S\setminus s_{C_{n-1}})\}$ for all existing significant partial correlations. Figure 2b illustrates the approach to build a differential GGM by evaluating the set of leave-one-out GGMs with respect to the criterion of disease-specificity. A pseudo-code formalizes the differential GGM approach:

```
ggm <- empty set of GGMs
ggm(0) <- result of GGM with S
for (i = 1:n){
     ggm(i) <- result of GGM with S\Si
}

dGGM <- empty set of differential GGM edges
for (e=(li,lj) : all possible edges){
     if (e fulfills IAij w.r.t. ggm) {
       dGGM -> add e between nodes li and lj
     }
 }
 return dGGM
```

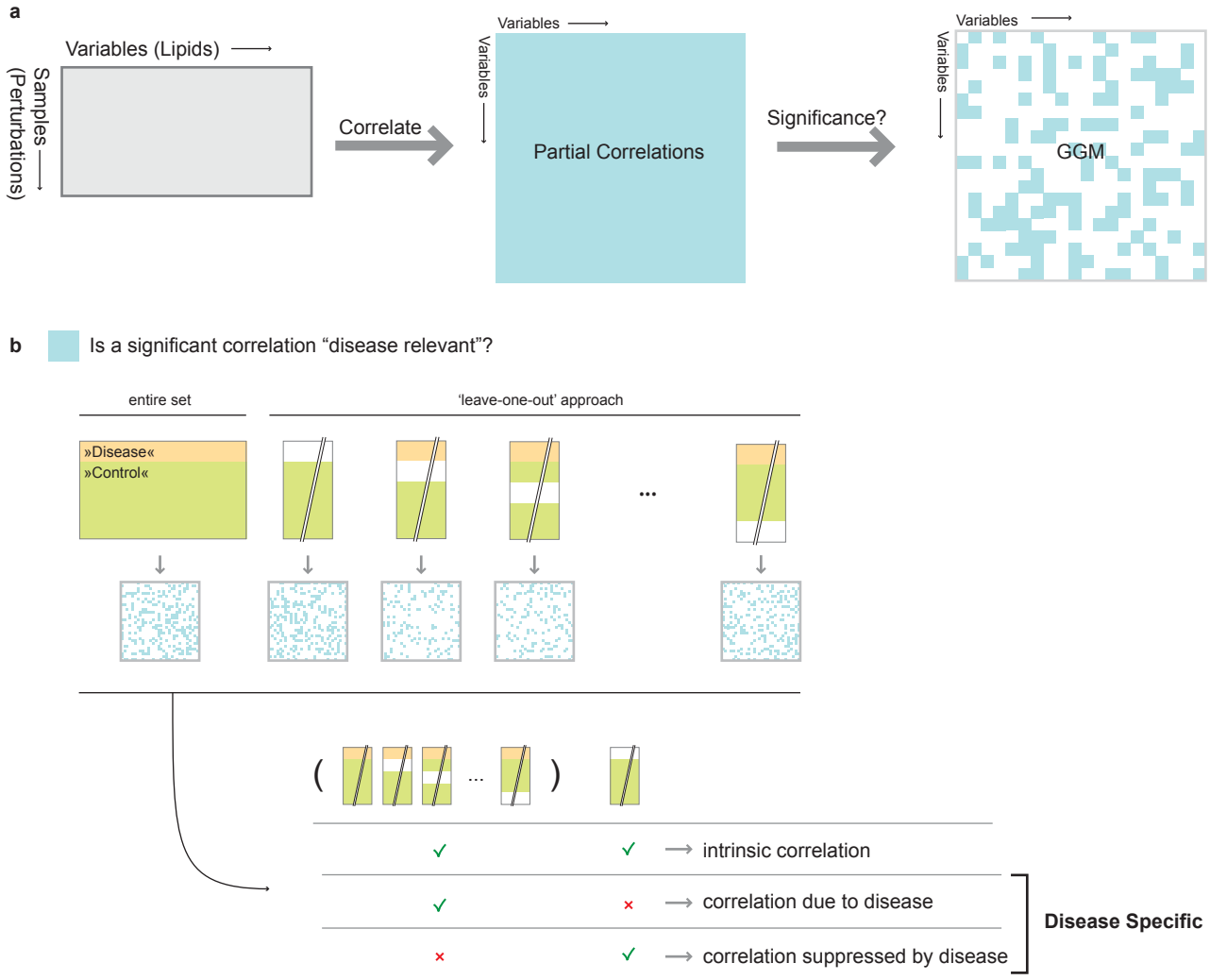


Figure 2. **Raw data transformed to disease specific correlations. a.** The lipidome raw data is a matrix of samples over variables. The samples are the individual perturbations which are grouped into control samples and the sample(s) of interest to the study, here simply called ‘disease’. Partial correlations of all variables are obtained and later evaluated with respect to statistical significance. **b.** To investigate whether a significant partial correlation is specific for the disease sample, partial correlations (as in **a**) were calculated for the entire dataset as well as for datasets where each one sample was left out. Unless a correlation is significant in all GGMs, it is considered disease-specific.

In detail, we extract those interactions IA_{ij} of (l_i, l_j) which fulfill the criterion to be disease relevant by comparing all GGMs with respect to the disease sample s_D as

$$\mathrm{IA}_{ij} = [\neg\exists\zeta^*(S\backslash s_D) \wedge \forall_{s_i\in\{S,S\backslash s_{C_\cdot}\}}\exists\zeta^*(s_i)] \quad \vee \quad [\exists\zeta^*(S\backslash s_D) \wedge \forall_{s_i\in\{S,S\backslash s_{C_\cdot}\}}\neg\exists\zeta^*(s_i)].$$

In other words, we consider an edge disease-specific if it fulfills either one of two criteria: (1) The edge is not significant in the GGM of $S\backslash s_D$, the dataset S without the disease sample s_D, while it is significant in the GGM constructed from the entire dataset S as well as in all GGMs of $S\backslash s_{C_\cdot}$ where each one control sample was left out for the calculation. (2) The reverse case holds if the edge is significant on the dataset without the disease sample $(S\backslash s_D)$ – equivalent to a correlation of control samples – while the edge is not significant if the disease sample is present in the dataset (that are the datasets of S and any $S\backslash s_{C_\cdot}$). As a result, we obtained one differential GGM of only direct lipid-lipid correlations resulting from the combination of wt p53 transfection prior to SN-38 chemotherapy for the Glioblastoma lipidome.

5. RESULTS

We generated GGMs for all perturbation combinations of the Glioblastoma lipidome according to out jackknife-inspired approach. The FDR cutoff value was set to $q = 0.01$. Compared to conventional GGM applications (analysis only of the entire dataset), we can break down each significant correlation with respect to the contribution of each sample. If we examine the lipidome solely from the perspective of conventional GGM calculations, we would obtain 256 significant lipid-lipid correlations. Thereof 25 correlations are disease relevant with respect to the perturbation of p53 gene therapy prior to SN-38 chemotherapy (Fig. 3a). Surprisingly, less than 10% of all significant interactions of a GGM from the entire dataset were actually disease-specific, or figuratively speaking true positive. Drawing any biological conclusions from correlations on the entire data set may therefore be misleading. In addition, we identified 9 significant lipid-lipid correlations which are suppressed by the disease relevant sample. Subsequently, we were able to identify 34 lipid-lipid interactions on the Glioblastoma lipidome which are significantly correlated upon p53 gene therapy prior to SN-38 chemotherapy.

The resulting disease-specific, differential GGM is depicted in Figure 3b. Since we obtained correlations across all six lipid species, our results are more comprehensive than the results of previous analyses[2,5] where lipid species were always handled separately. Closer inspection of the 45 lipids involved in the disease relevant differential

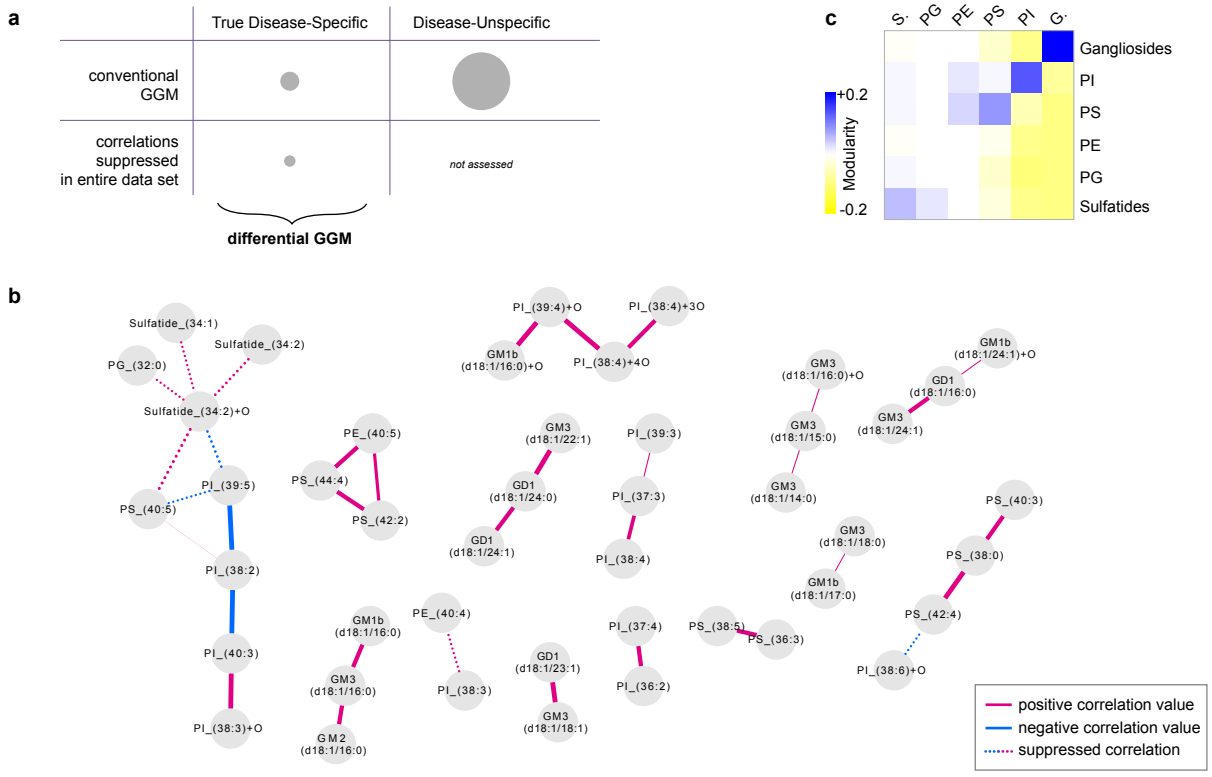


Figure 3. **Lipids specifically regulated when Glioblastoma cell lines were effectively perturbed. a.** Relative number of disease specific and unspecific lipid-lipid partial correlations in the GGM. Analysis of the entire dataset is named "conventional" GGM with respect to disease specificity. **b.** Disease relevant GGM which is associated with the combined perturbation of p53 adenoviral transfection prior to SN-38 chemotherapy in U87 Glioblastoma cell lines. Edges between lipid nodes are drawn if a significant correlation exists. Positive and negative correlations were color-coded in pink and blue, respectively; Suppressed correlations drawn with dotted lines. Edge line widths indicate degree of dependencies (absolute partial correlation value). The numbers C:D indicates the number of carbon atoms (C) and double bonds (D) of the fatty acid side chain(s). **c.** Modularity matrix was calculated by using lipid classes as cluster label for the GGM shown in **b**. Modularity values were color-coded between -0.2 and $+0.2$ from yellow to blue, respectively. Modularity values close to 1 indicate strong inner-cluster connectivity and little links outside its cluster.

GGM revealed an overrepresentation of specific lipid classes. Sulfatides are glycosphingolipids with two variable ceramide tails. Out of five measured sulfatides, three (60%) were differentially correlated. The three C31:1, C34:2 and C34:2+O are all short chain ceramides with increased levels for the p53 plus SN-38 perturbation.[2] We can assign the C34:2+O sulfatide a more important role with respect to the disease, as it has a prominent role in the differential GGM with five edges. Note, that we revealed the sulfatide regulation only by inspecting the suppressed correlations, which would have been overlooked by conventional GGM analysis. Gangliosides are glycosphingolipids where one of the two side chains is fixed to a C18:1 fatty acid. They additionally vary in their number of salic acid residues (mono, di or tri). In general, 17 out of 32 (53%) measured gangliosides were coregulated in the disease-specific GGM. Of the the major gangliosides found in adult brain (GM3/GD3),[15] only one was measured by MS. Interestingly, the GM3 was found to be overrepresented with 61% in the GGM (8 out of 13 measured). As previously shown to have decreased level for the p53 plus SN-38 perturbation,[2] the long chain gangliosides GD1 and GM1b were also found to be overrepresented in the GGM by 50% (4 of 8) and 66% (4 of 6), respectively. Besides the two lipid classes which are overrepresented by more than a half of the measured lipids, another interesting lipid class were Phosphoinositols (PIs). PIs are phospholipids with two esterified fatty acyl residues and inositol as the polar head group. One fourth of the PI were found to be enriched in the GGM (14 of 55). In the original study, the phosphatidylglycerols (PGs) were used as a generic example to show the increased levels of all four phospholipids subclasses.[2] Nevertheless, we detected an overrepresentation of PIs. A more detailed biological analysis of the PI may reveal the affected mechanisms.

Finally, we aimed to analyze the extend to which the lipid classes were interlinked with each other in the disease-specific GGM. We calculated the modularity † by considering each lipid class as the node class label (Fig. 3c). We assume the lipid classes with little or no links to other classes to have a disease relevant regulation based on their molecular characteristics and not due to e.g. fatty acid remodeling. The sulfatides show the most prominent inner-group linkage, indicating that this class was specifically affected by the p53 plus SN-38 perturbation. The gangliosides and all four phospholipids classes were generally interlinked, indicating that a disease relevant mechanism is rather linked to common fatty acid side chains than their unique characteristic head groups.

6. CONCLUSION

We have developed a biologically driven technique to analyze high-throughput measurements. The novel method of a differential GGM is inspired by the experimental design of the biological study to reveal disease relevant information. The differential GGM was applied to the influence of p53 gene therapy prior to SN-38 chemotherapy on U87 Glioblastoma cell lines. We identified only those lipid correlations which are solely induced by the combined perturbation and not just by a single perturbation. Beyond prior studies of quantification histograms and lipid profiles on single lipid classes, we succeeded in analyzing lipids across their classes for the Glioblastoma lipidome which is easily comprehensibly. The disease-specific correlations will advance the understanding of primary brain tumors and their mechanism to immortality.

REFERENCES

[1] Puchades, M., Nilsson, C. L., Emmett, M. R., Aldape, K. D., Ji, Y., Lang, F. F., Liu, T.-J., and Conrad, C. A., "Proteomic investigation of glioblastoma cell lines treated with wild-type p53 and cytotoxic chemotherapy demonstrates an association between galectin-1 and p53 expression.," *J Proteome Res* **6**, 869–875 (Feb 2007).

[2] He, H., Nilsson, C. L., Emmett, M. R., Ji, Y., Marshall, A. G., Kroes, R. A., Moskal, J. R., Colman, H., Lang, F. F., and Conrad, C. A., "Polar lipid remodeling and increased sulfatide expression are associated with the glioma therapeutic candidates, wild type p53 elevation and the topoisomerase-1 inhibitor, irinotecan.," *Glycoconj J* **27**, 27–38 (Jan 2010).

†Modularity was introduced as a measure of community structure.[16] It measures how modular a set of nodes is compared to a random network model with identical connectivity. The fraction of edges between cluster i and cluster j is corrected for the connectivity of cluster i. The fraction $e_{ij} = \frac{l_{ij}}{L}$ is determined by the number of edges l_{ij} between vertices of i and j divided by the number of edges in the graph L. The connectivity of the cluster a_i is defined by the sum of degrees d_i of vertices in i, or $a_i = \sum_j e_{ij}$. Modularity is then $M_{ij} = (e_{ij} - a_i^2)$

[3] He, H., Conrad, C. A., Nilsson, C. L., Ji, Y., Schaub, T. M., Marshall, A. G., and Emmett, M. R., "Method for lipidomic analysis: p53 expression modulation of sulfatide, ganglioside, and phospholipid composition of u87 mg glioblastoma cells.," *Anal Chem* **79**, 8423–8430 (Nov 2007).

[4] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., and Theis, F. J., "Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data.," *BMC Syst Biol* **5**, 21 (Jan 2011).

[5] Görke, R., Meyer-Bäse, A., Wagner, D., He, H., Emmett, M. R., and Conrad, C. A., "Determining and interpreting correlations in lipidomic networks found in glioblastoma cells.," *BMC Syst Biol* **4**, 126 (2010).

[6] Schäfer, J. and Strimmer, K., "Learning large-scale graphical gaussian models from genomic data," in [*In Science of Complex Networks: From Biology to the Internet and WWW*], (2005).

[7] Opgen-Rhein, R. and Strimmer, K., "Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data.," *4th International Workshop on Computational Systems Biology, WCSB 2006* , 73–76 (June 2006).

[8] Monakov, A. A., "Estimation of the covariance matrix for dependent signal samples: polarization diversity systems," **30**(2), 484–492 (1994).

[9] Schäfer, J. and Strimmer, K., "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics.," *Stat Appl Genet Mol Biol* **4**, Article32 (2005).

[10] Opgen-Rhein, R. and Strimmer, K., "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.," *BMC Syst Biol* **1**, 37 (2007).

[11] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P., "Discovery of meaningful associations in genomic data using partial correlation coefficients.," *Bioinformatics* **20**, 3565–3574 (Dec 2004).

[12] Velicer, W. F., "Suppressor Variables and the Semipartial Correlation Coefficient," *Educational and Psychological Measurement* **38**(4), 953–958 (1978).

[13] Das, A. and Kempe, D., "Algorithms for subset selection in linear regression," in [*Proceedings of the 40th annual ACM symposium on Theory of computing*], *STOC '08*, 45–54, ACM, New York, NY, USA (2008).

[14] Miller, R. G., "The jackknife-a review," *Biometrika* **61**(1), 1–15 (1974).

[15] Ando, S. and Yu, R. K., "Fatty acid and long-chain base composition of gangliosides isolated from adult human brain.," *J Neurosci Res* **12**(2-3), 205–211 (1984).

[16] Newman, M. E. J. and Girvan, M., "Finding and evaluating community structure in networks.," *Phys Rev E Stat Nonlin Soft Matter Phys* **69**, 026113 (Feb 2004).