

# Correlation guided Network Integration (CoNI) reveals novel genes affecting hepatic metabolism



Valentina S. Klaus<sup>1,2,3,4,17</sup>, Sonja C. Schriever<sup>3,4,5,17</sup>, José Manuel Monroy Kuhn<sup>1,3,4,17</sup>, Andreas Peter<sup>3,6,7</sup>, Martin Irmeler<sup>8</sup>, Janina Tokarz<sup>9</sup>, Cornelia Prehn<sup>9</sup>, Gabi Kastenmüller<sup>3,10</sup>, Johannes Beckers<sup>3,8,13</sup>, Jerzy Adamski<sup>9,11,13</sup>, Alfred Königsrainer<sup>14</sup>, Timo D. Müller<sup>3,4,15</sup>, Martin Heni<sup>3,6,12</sup>, Matthias H. Tschöp<sup>2,3,4,16</sup>, Paul T. Pfluger<sup>2,3,4,5</sup>, Dominik Lutter<sup>1,3,4,\*</sup>

## ABSTRACT

**Objective:** Technological advances have brought a steady increase in the availability of various types of omics data, from genomics to metabolomics. Integrating these multi-omics data is a chance and challenge for systems biology; yet, tools to fully tap their potential remain scarce.

**Methods:** We present here a fully unsupervised and versatile correlation-based method — termed Correlation guided Network Integration (CoNI) — to integrate multi-omics data into a hypergraph structure that allows for the identification of effective modulators of metabolism. Our approach yields single transcripts of potential relevance that map to specific, densely connected, metabolic subgraphs or pathways.

**Results:** By applying our method on transcriptomics and metabolomics data from murine livers under standard Chow or high-fat diet, we identified eleven genes with potential regulatory effects on hepatic metabolism. Five candidates, including the hepatokine *INHBE*, were validated in human liver biopsies to correlate with diabetes-related traits such as overweight, hepatic fat content, and insulin resistance (HOMA-IR).

**Conclusion:** Our method's successful application to an independent omics dataset confirmed that the novel CoNI framework is a transferable, entirely data-driven, flexible, and versatile tool for multiple omics data integration and interpretation.

© 2021 The Authors. Published by Elsevier GmbH. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Keywords** Data integration; Hepatic steatosis; Multi-omics; Systems biology

## 1. INTRODUCTION

In the era of systems biology and high throughput multi-omics data generation, there is an unmet need for effective tools and approaches to compare and integrate these complex data sets [1,2]. Such tools are particularly required for capturing genetic mechanisms associated with metabolic disorders, which typically affect multiple layers of biological regulations and different types of biomolecules.

Present integration approaches to interrogate complex metabolic networks [3] are built mainly on the integration of genetic information through associative approaches or using prior additional knowledge. Integrating multi-omics data on metabolic profiles and genetic information was either attempted by directly correlating genetic variants or transcripts with metabolites [4,5], by using prior knowledge to map genes and enzymes on metabolic pathways [6–8], or by creating deterministic models that abstract enzymatic reactions in metabolic

<sup>1</sup>Computational Discovery Research Unit, Institute for Diabetes and Obesity, Helmholtz Zentrum München, Neuherberg, Germany <sup>2</sup>TUM School of Medicine, Neurobiology of Diabetes, Technical University Munich, Germany <sup>3</sup>German Center for Diabetes Research (DZD), Neuherberg, Germany <sup>4</sup>Institute for Diabetes and Obesity, Helmholtz Diabetes Center at Helmholtz Zentrum München, Germany <sup>5</sup>Research Unit Neurobiology of Diabetes, Helmholtz Zentrum München, Neuherberg, Germany <sup>6</sup>Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Center Munich at the University of Tübingen, Tübingen, Germany <sup>7</sup>Institute for Clinical Chemistry and Pathobiochemistry, University Hospital Tübingen, Germany <sup>8</sup>Institute of Experimental Genetics, Helmholtz Zentrum München, Neuherberg, Germany <sup>9</sup>Research Unit Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, Neuherberg, Germany <sup>10</sup>Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany <sup>11</sup>Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore <sup>12</sup>Department of Internal Medicine, Division of Endocrinology, Diabetology, and Nephrology, Eberhard Karls University Tübingen, Tübingen, Germany <sup>13</sup>Chair of Experimental Genetics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Neuherberg, Germany <sup>14</sup>Department of General, Visceral and Transplant Surgery, University Hospital Tübingen, Germany <sup>15</sup>Department of Pharmacology and Experimental Therapy, Institute of Experimental and Clinical Pharmacology and Toxicology, Eberhard Karls University Hospitals and Clinics, Tübingen, Germany <sup>16</sup>Division of Metabolic Diseases, Department of Medicine, Technical University Munich, Munich, Germany

<sup>17</sup> These authors contributed equally to this work.

\*Corresponding author. Computational Discovery Research, Institute for Diabetes and Obesity, Helmholtz Zentrum München Parking 13 85748 Garching, Germany. E-mail: [dominik.lutter@helmholtz-muenchen.de](mailto:dominik.lutter@helmholtz-muenchen.de) (D. Lutter).

Received February 9, 2021 • Revision received June 24, 2021 • Accepted July 9, 2021 • Available online 13 July 2021

<https://doi.org/10.1016/j.molmet.2021.101295>

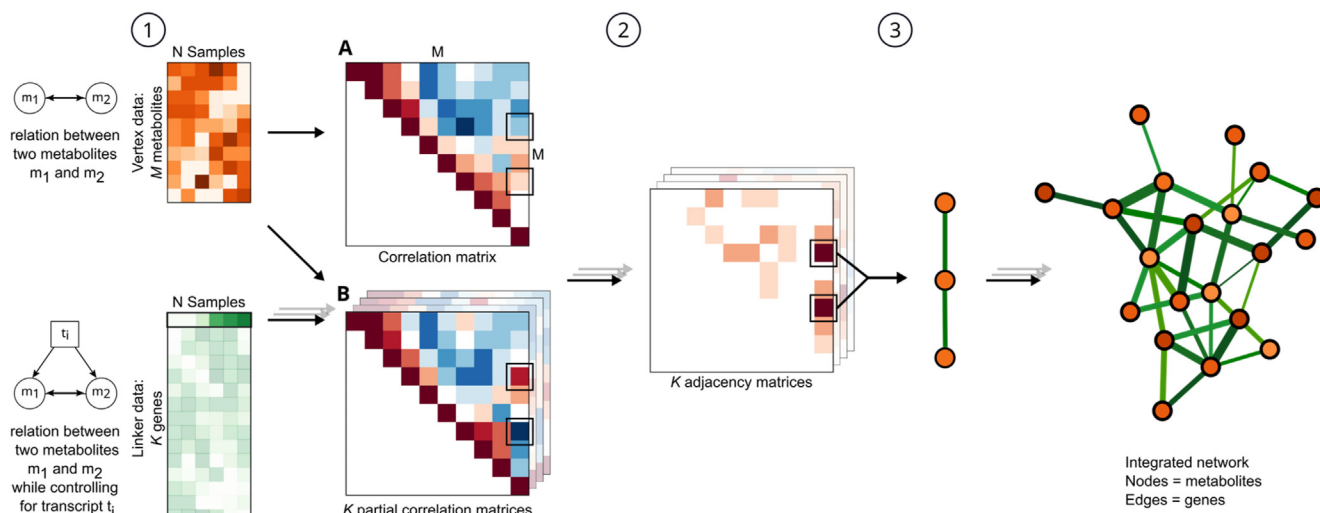
pathways using gene or enzyme levels as rate-limiting denominators [9–11]. However, all these approaches have limitations. Pathway mapping approaches and deterministic models depend either on prior knowledge or on multiple parameters, thus having high degrees of freedom. Correlation-based models typically map direct one-to-one relationships between molecules, but lack a suitable model that reflects biochemical pathways where enzymes or regulatory genes affect the ratio between educts and products in biochemical reactions. A further limiting factor is that metabolomics – targeted or untargeted – typically only captures a small fraction of the metabolome (a few hundred to thousands of the more than 40,000 estimated metabolites) [12], which dramatically limits the use of prior knowledge. Accordingly, novel approaches to reveal distinct regulatory genes are warranted. One area where sophisticated multi-omics data integration could be advantageous is in the study of hepatic steatosis; pathological comorbidity of high body adiposity that is characterized by excess fat accumulation caused by dysfunctional lipid metabolism [13]. Obesity-induced ectopic fat disposition in the liver is a major risk factor in the pathogenesis of type 2 diabetes by locally driving hepatic insulin resistance [14]. However, exact molecular mechanisms that link both pathophysiological conditions remain only partially understood. For the development and implementation of efficient prevention and treatment strategies against fatty liver disease and comorbid hepatic insulin resistance, it is thus crucial to better understand the initial pathogenesis and the exact mechanisms that link both comorbidities. In this study, we present a novel statistical method for correlation-based network integration (CoNI) of generic character and conceivable for multiple approaches. In contrast to common correlation-based methods, that typically estimate one-to-one relationships, we here combine Pearson's correlation with partial correlation to infer transcriptional impact on metabolite pair correlations. We tested CoNI on publicly available proteomics and lipidomics data (SI) and subsequently applied CoNI to murine liver metabolome and transcriptome data sets, and unraveled previously hidden gene–metabolite interactions – that exert major changes to hepatic metabolite levels under normal dietary conditions and obesogenic stress with established fatty liver disease

and hepatic insulin resistance. Validation experiments in human liver biopsy samples revealed that the expression of five selected candidates associated with hepatic triglyceride levels, BMI, or insulin resistance in humans was significant. *In vitro* knockdown experiments confirmed that modulation of selected candidate genes affects metabolite correlations. Overall, our new tool helped to unravel known and potential new genes involved in the regulation of liver metabolism and fatty liver disease along with hepatic insulin resistance under excess energy supply.

## 2. RESULTS

### 2.1. CoNI: correlation guided Network Integration

The CoNI framework uses correlations and partial correlations to combine two types of omics data (**linker data**, **vertex data**), thereby generating a graph where the linker data form the edges and the vertex data form the vertices or nodes. The linker data specifies the impact on the interaction of the vertex data. The general concept of CoNI (Figure 1) is to identify potential confounding variables (transcripts) by estimating the effect of a controlling variable  $t$  (transcript) on the correlation of two random dependent variables  $m_1$   $m_2$  (metabolites). Therefore, Pearson correlation coefficients  $\rho_{m_1 m_2}$  are calculated for each pair of metabolites. Subsequently, each gene's linear effect is estimated by comparing the partial correlation coefficient  $\rho_{m_1 m_2 \cdot t}$  with  $\rho_{m_1 m_2}$ . For  $K$  transcripts and  $M$  metabolites, we thus generated one correlation matrix  $M \times M$  and  $K$   $M \times M$  matrices that contained the partial correlation coefficients. Next, an adapted Steiger test [15] is applied to estimate a significant effect of a transcript on the metabolite pair correlation ( $p < 0.05$ ), thereby generating  $K$  adjacency matrices. These adjacency matrices are then combined to form an integrated graph where the nodes to the metabolite pairs and the edges refer to the controlling genes. A gene can thereby be mapped to multiple edges and edges may consist of multiple genes. Finally, this gene – metabolite pair network assembly is used to identify local controlling genes (LCGs), i.e., genes locally enriched in a densely-connected subgraph.



**Figure 1: CoNI workflow.** 1) Calculation of a full pairwise correlation matrix (A) and partial correlation analysis combining the metabolite concentrations with the transcript expression profiles; for each pair of the metabolites, partial correlation scores are computed by subtracting an estimated regulatory effect for each gene  $k$ . This results in  $K$  partial correlation matrices (B). 2) Calculation of  $K$  adjacency matrices by selecting metabolite pairs significantly altered by individual genes. 3) Selection of significant triplets (metabolite pair plus gene) and construction of an undirected, weighted graph with correlated metabolite pairs as nodes, and influencing genes setting up the edges. The number of genes connecting the two respective metabolite nodes determines the edge weight, indicated by the thickening of the line.

## 2.2. Application and versatility of CoNI

The method presented here was applied to two independent data sets. Besides the presented approach, where we applied CoNI to integrate and compare hepatic metabolite and gene expression, we also applied it to publicly available proteomics and lipidomics data [16,17]. We identified local controlling proteins that — as confounders for lipid correlation — changes in the lung lipidome of C57BL/6 mice under fresh air and smoking conditions (see SI). These findings suggest that we developed an entirely data-driven, flexible, and versatile tool for multiple omics data integration and interpretation.

## 2.3. Transcriptional and metabolic profiling of livers from chow and HFD-fed mice

To investigate the effects of diet-induced obesity (DIO) on the liver transcriptome and metabolome, we applied CoNI on data generated from male C57BL/6J mice either exposed to standard Chow diet or 58% high-fat diet (HFD) for 22 weeks. Exposure to the high-fat diet resulted in significantly higher body weight (BW, Chow 33.5 g  $\pm$  1.6 g; HFD 49.2 g  $\pm$  4.5 g,  $p < 0.0001$ , mean  $\pm$  SD) (Figure 2A). HFD-fed obese mice showed increased plasma triglyceride and cholesterol levels (Figure 2B,C), whereas hyperinsulinemic mice (Figure 2D) had increased hepatic triglyceride (TAG) stores (Figure 2E, Table 1) compared to chow-fed lean controls.

Hepatic metabolism was analyzed by transcriptional and metabolic profiling using Affymetrix microarrays (Figure 2F) and the targeted metabolomics AbsoluteIDQ™ p180 kit (Figure 2G, Table S1). The differential expression analysis of hepatic tissue revealed 989 significantly and differentially expressed genes between Chow and HFD mice (Figure S1A, Table S2). Functional enrichment analyses based on gene ontology (GO) of the up- and down-regulated genes revealed numerous metabolic processes; lipid-related processes in particular (Figure S1B, Table S2). We further identified 91 significantly altered metabolites in HFD livers compared to Chow controls (Figure S1C, Table S1). The most prominently regulated metabolite classes were the sphingomyelins (SM, 67 % regulated), followed by phosphatidylcholines (PC, 64 %) and the acylcarnitines (AC, 45 %). With principal component analyses (PCA), we could reveal that the administered diet was the main contributor to explain variance in gene expression (Figure 2H) and the main driver of metabolite variance (Figure 2I) in murine livers.

## 2.4. Correlation maps of diet-altered metabolites

To further investigate diet-induced changes in liver metabolism, we generated a correlation map for each diet by calculating all metabolites' pairwise Pearson correlation coefficients (Figure 3A). We observed a slight negative skew of the correlation coefficient distribution for both diets (Figure S2A), as previously observed by Bartel et al. [18]. We identified 2,488 significantly correlated metabolite pairs for Chow and 2,322 pairs for HFD (non-adjusted  $p < 0.05$ ), whereas 923 metabolite pairs were identified in both diets (Figure 3B). Furthermore, 1,023 metabolite pairs showed a significant change in their correlation from Chow to HFD, indicating that the administered diet substantially alters the hepatic metabolism. We then deepened our investigations and analyzed the class composition of the correlated metabolite pairs (Figure 3C). The metabolite class with the maximum change in correlated pairs was the sphingomyelins (Jaccard Index = 0.12). In contrast, amino acids (AA) mostly maintained their correlations (Jaccard Index = 0.76), indicating that they interact independently of dietary conditions. Between classes, we generally observed a substantial change in correlations with Jaccard indices between 0.01 (PC — BA (biogenic amines)) and 0.71 (H (hexoses) —

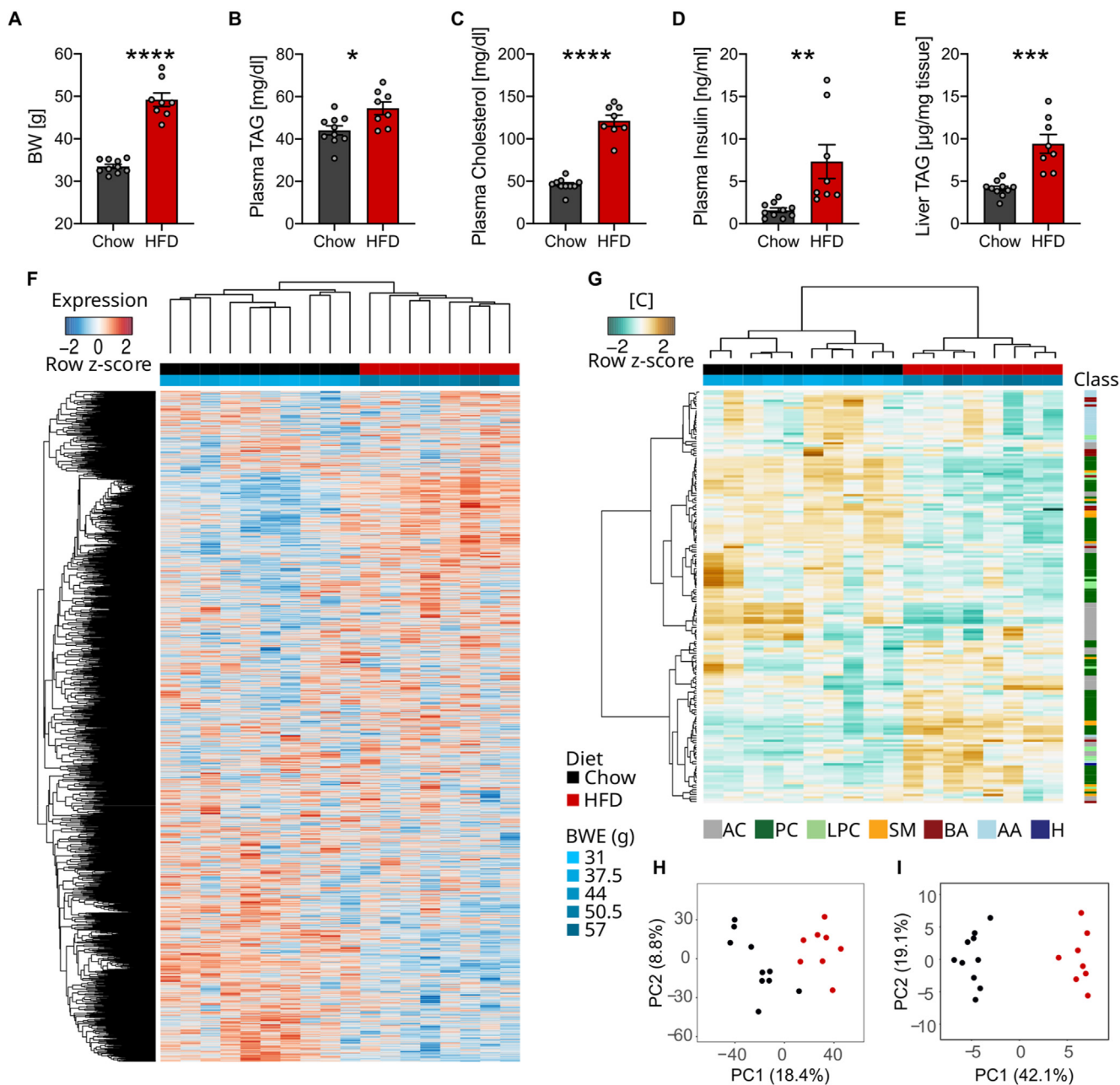
AC) when Chow and HFD livers were compared. A striking difference was observed in the correlated metabolite class pairs for AC and PC with the highest absolute numbers of changes driven by the shift from positive correlations under Chow diet to negative ones under HFD (Figure 3A,C, and S2B), indicating massive changes in these metabolite classes under obesogenic conditions. In summary, the correlation analyses revealed substantial diet-dependent changes in metabolite regulation, which is in accordance with previously reported findings of HFD-induced changes in metabolite concentrations in a circadian manner over multiple tissues [19].

## 2.5. Estimating genetic impact on metabolic networks

Subsequently, we applied CoNI to assess the genetic impact on metabolite correlations under both dietary conditions. Hence, gene expression was integrated into pairwise metabolite correlations to form two independent graphs (Figure 3D,E). The Chow graph was constructed of 485 triplets (gene and metabolite pairs) and the HFD graph of 1,058 triplets (Table S3). Of the 175 metabolites used for the analysis, more metabolites were connected in the CoNI network of HFD-fed mice than the Chow controls (Chow  $n = 133$ ; HFD  $n = 164$  metabolites; Figure 3F), with an overlap of 127 metabolites. Of all connected metabolite pairs (Chow  $n = 407$ ; HFD  $n = 722$ ), 67 were identical in both diets revealing an extensive rewiring of hepatic metabolism under HFD (Figure 3G). This alteration in hepatic metabolism by HFD-feeding was also evident in the node degree distribution, the number of edges per node. The HFD network showed consistently higher degrees than Chow (Figure S3A), which was also observed when comparing the node degrees for the specific metabolite classes (Figure S3B). The Chow network showed a trend towards increased node degrees for PC and LPC compared to the other metabolite classes, which was absent in the overall elevated distribution of node degrees within the HFD network. A striking characteristic of the inferred networks is that both tend to be organized in communities (Figure S4A, B) or densely connected subnetworks, which mainly reflect metabolite classes, but are partly reorganized on dietary change (Figure S4C). This reorganization was also observed in other network characteristics, such as the shortest path length (Figure S5).

Analogous to the reorganization of metabolite interactions, the genes forming the edges in both CoNI constructed graphs substantially changed (Chow  $n = 166$ ; HFD  $n = 319$ ) with only five genes shared between the networks (Figure 3H, S6A, B): *Gm4553*, *Hnmpm*, *Tap1*, *Xpo7*, and *Eya3*, from which *Tap1* was the only one differentially expressed between Chow and HFD. Furthermore, the comparison of the number of genes that map to single edges revealed that most edges consist of a single gene (85.75 % in Chow, 65.93 % in HFD), with a maximum number of six genes per edge in Chow and five in HFD (Figure 3I, Table S3). The distribution of individual genes over the edges showed that most genes appeared in five or fewer edges (Figure 3J). These results highlight the specificity of the gene—metabolite interactions and the substantial metabolic changes in the liver upon HFD-induced obesity.

To further classify the genes affecting the correlations between metabolite pairs in the independent networks, we performed a functional enrichment analysis using KEGG pathways and GO biological processes (Table S3). However, no informative categories were identified for the chow network's genes; the HFD network's genes were enriched in the KEGG categories 'glycerolipid metabolism' and 'nonalcoholic fatty liver disease' (NAFLD). Strikingly, these two categories were not significantly enriched for the 989 DEGs (Table S2). Thus, CoNI could provide an improved reflection of the metabolic



**Figure 2: Transcriptional and metabolic profiling of murine livers under normal conditions and obesogenic stress.** Barplots comparing mice after 22 weeks of Chow ( $n = 10$ ) or HFD ( $n = 8$ ) for (A) body weight, (B) plasma triacylglyceride (TAG), plasma cholesterol (C), plasma insulin (D), and hepatic TAG levels (E). Asterisks indicate the significance of the differences between the factors (\* $p \leq 0.05$ ; \*\* $p \leq 0.01$ ; \*\*\* $p \leq 0.001$ ). Error bars show standard error of the mean (SEM) (F) Heatmap with 10,159 hepatic mRNAs transcripts detected in chow-fed (black color, upper bar) and HFD-fed (red color, upper bar) mice. The lower color bar indicates individual body weights measured at the end of the study (BWE). (G) Heatmap with concentrations of 175 detected metabolites. Metabolite classes are indicated in the right color bar: 40 acylcarnitines (AC), 76 phosphatidylcholines (PC), 14 lysophosphatidylcholines (LPC), 12 sphingomyelins (SM), 12 biogenic amines (BA), 20 amino acids (AA), and 1 hexose (H). PCA plot of transcript expression (H) and metabolite concentrations (I) for Chow (black) and HFD (red). The amounts of variance explaining the differences are given in brackets.

phenotype induced by HFD (Figure 2A–E) that showed all hallmarks of fatty liver disease.

## 2.6. Effective network genes

We next aimed to identify genes within our CoNI networks that locally drive the observed changes in hepatic metabolism under obese conditions. First, we defined local controlling genes (LCGs) as genes significantly enriched within a local subgraph of correlating metabolite pairs (see methods). With this approach, we could identify 20 LCGs in

the Chow network and 59 LCGs in the HFD network (Table S3) with no overlap. Combining these network characteristics with differential gene expression, we found one LCG in the chow network (*Ddx3x*) and seven LCGs from the HFD network (*Myc*, *Arhgap24*, *Smim13*, *Rapgef4*, *Cd82*, *Inhbe*, and *Gk*) to be differentially expressed between the two diet groups (Figure 4A–H, S6A, B, S7, Table 2, S3).

Subsequently, to uncover functional links to obesity and obesity-associated diseases of the identified 79 LCGs in humans, we queried the Type 2 diabetes knowledge portal [20]. For three LCGs of



**Table 1** — Characteristics of the mouse cohorts (mean ± SD).

	Chow (n = 10)	HFD (n = 8)	p-value
Body weight (end) (g)	33.49 ± 1.56	49.2 ± 4.49	<0.0001
Liver triglyceride levels (μg/mg tissue)	4.13 ± 0.9	9.4 ± 3.11	0.0001
Plasma triglyceride levels (mg/dl)	44.09 ± 6.84	54.49 ± 8.43	0.0105
Plasma cholesterol levels (mg/dl)	45.74 ± 6.84	121.24 ± 18.78	<0.0001
Plasma insulin levels (ng/ml)	1.58 ± 0.85	7.32 ± 5.65	0.0057

the Chow network and 17 LCGs of the HFD network, we found single nucleotide polymorphisms (SNPs) associated with obesity-related traits and disease markers (Table S4, S1). Among the genes that showed the most prominent associations with obesity-related SNPs were *App12* (Figure 4I), which mediates insulin signaling, endosomal trafficking, adiponectin, other signaling pathways [21], and *Cobll1* (Figure 4J), which was strongly associated with obesity- and type 2 diabetes-related markers [22,23]. Additionally, we found the differentially expressed LCG *Inhbe*, a hepatokine, that had recently been linked to insulin resistance in human livers [24].

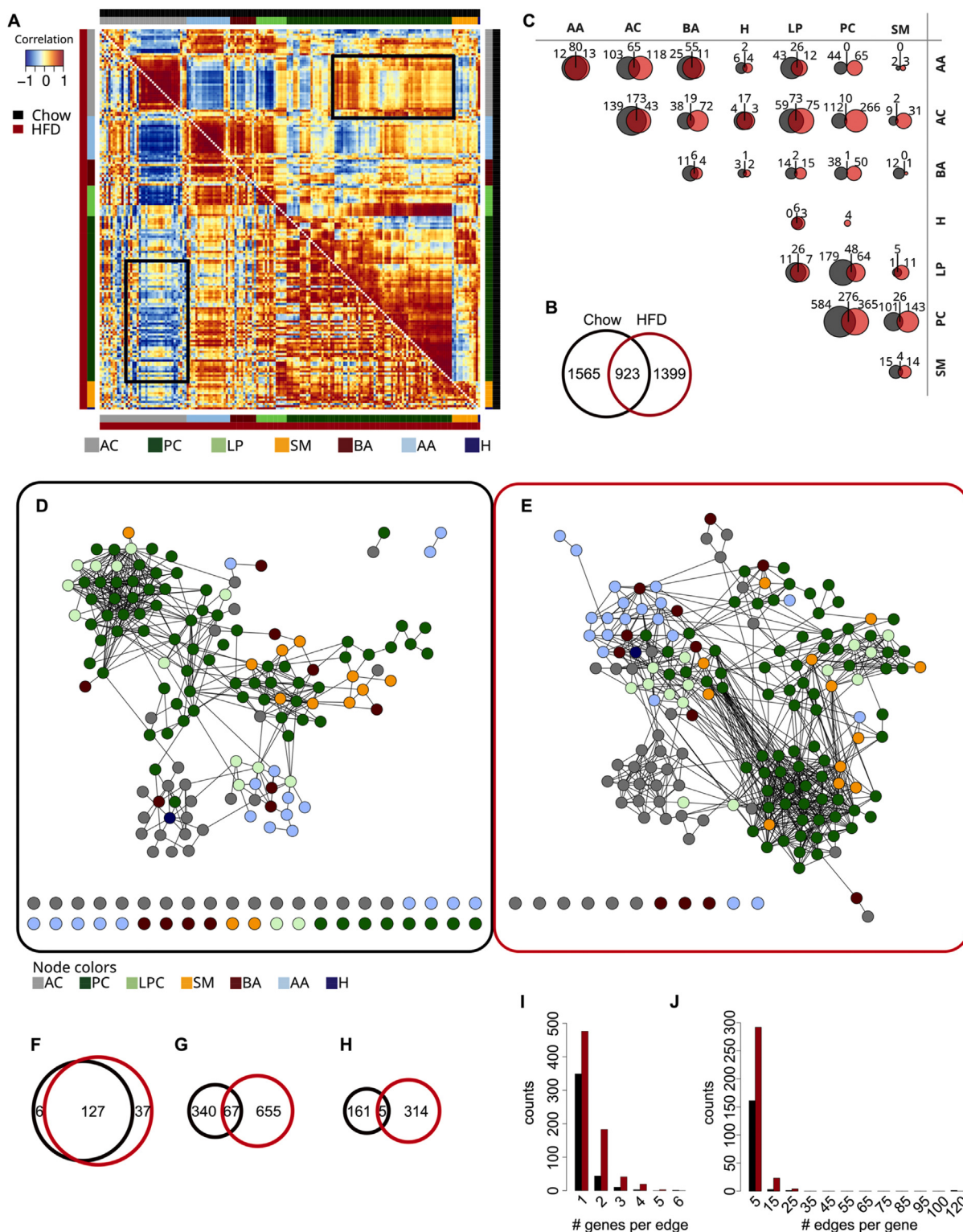
We combined these eight differentially expressed LCGs with two genes that had the strongest associations with diabetic relevant traits (*Cobll1* & *App12*) to a candidate list of 10 genes. When we compared metabolite concentrations of the ten sub-networks controlled by these selected LCGs, we found substantial differences in the metabolite levels between Chow and HFD (Figure 4A–J). To test for the interrelation between the selected LCGs, we combined the isolated metabolite-gene sub-networks and obtained three interconnected subgraphs for HFD (Figure 4K). The metabolite sub-networks of the Chow-derived LCG *Ddx3x* did not overlap with any of the metabolite subgraphs of the HFD network-derived LCGs. The largest interconnected subgraph of the HFD network contained mostly PCs (24/27) and was controlled by six genes, which is in accordance with our previous observation that this metabolite class is highly regulated under obesogenic conditions (Figure 3C). The importance of the highly abundant PCs that had been previously associated with NAFLD [25] was further supported by our KEGG enrichment where we found the PC regulatory phospholipase D signaling pathway enriched in the HFD network genes (Table S3). Importantly, this pathway was not significantly enriched ( $p = 0.32$ ) for the 989 DEGs, although it was present with 11 genes, of which four were also found in the HFD network. In contrast, the metabolite sub-networks of the HFD-derived LCGs *Smim13* and *Rapgef4* not only included BA, PC, LPC, and AC metabolites, but also their subgraphs were interconnected by an overlapping mixture of AA. An impaired AA metabolism was previously linked to NAFLD and liver fibrosis [26,27]. The third HFD sub-network was under the control of the LCG *Inhbe* and contained only AC which is known to be involved in fat metabolism, especially in the carnitine shuttle transporting long-chain fatty acids into mitochondria [28]. We found *Cpt1a*, *Cpt1b*, and *Cpt2* as known regulators of the carnitine shuttle that differentially expressed between chow and HFD, but not as part of the CoNI predicted networks.

We next investigated whether the hepatic expression of the selected target genes is associated with the three metabolic parameters, serum insulin, liver TAG, and body weight in Chow and HFD-fed mice. In addition to the ten LCGs with significant differential expression and/or an association with human obesity-related SNPs (Figure 4), we further selected *Tap1* for validation as it was the only gene present in both CoNI-derived graphs that were differentially expressed between Chow and HFD-fed mice (Table 2, S3). We generated and tested a linear regression model for each candidate mRNA and the metabolic

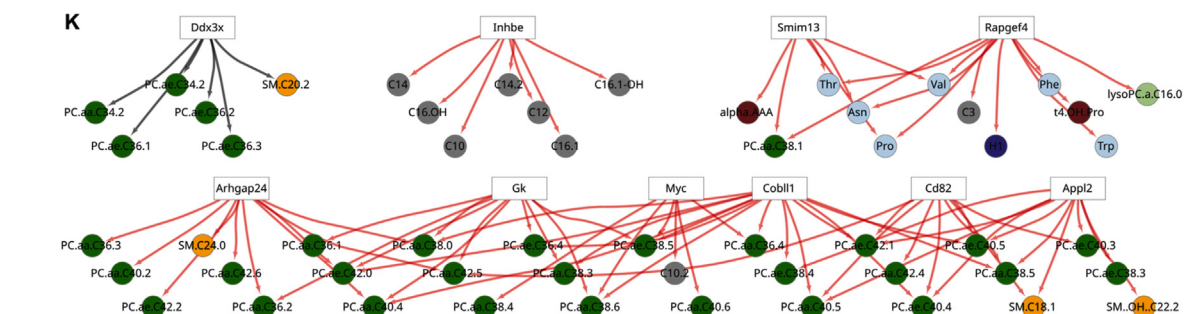
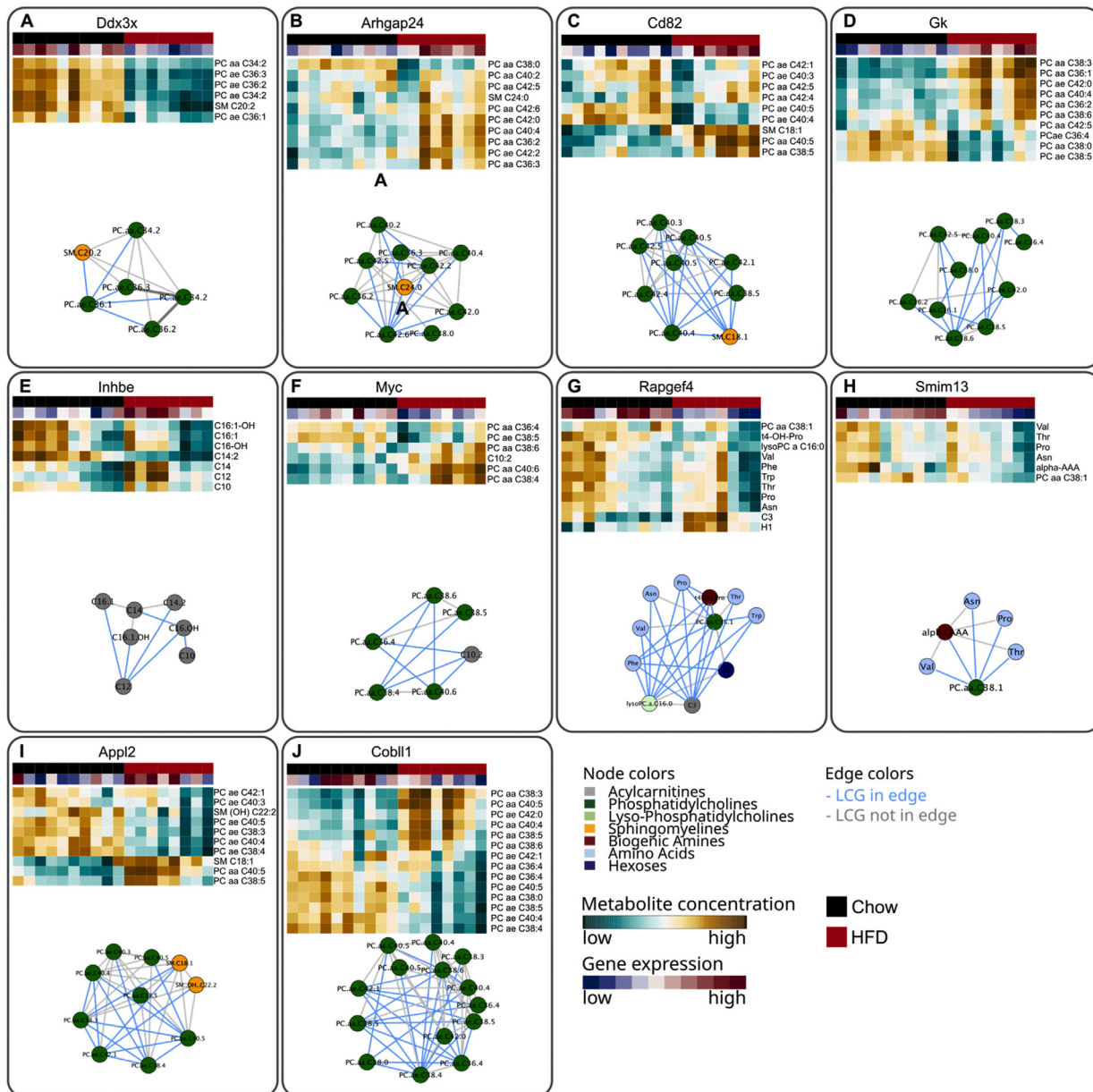
parameters including diet as a cofactor and an interaction term including diet-mRNA interactions and two diet-specific regression models (Figure S8–S10). All linear models significantly predicted the metabolic parameter (Table S5A–C), which could mostly be explained by the diet. Accordingly, we found for the interaction model and the diet-specific model, a diet-dependent effect on the metabolic parameter for almost all genes except *Myc*. To test whether this frequency of significant predictions would also be expected with a randomly selected gene list, we performed an enrichment analysis for all 989 DEGs (Figure S1A, Table S2) and all genes mapped to one of the diet-specific networks (Figure S11). We found that the Chow network genes were highly enriched to predict BW, whereas the HFD network genes significantly predicted plasma insulin and liver TAG levels. Although the eleven genes from the candidate list were less enriched than the full network genes, they still outperformed the differentially expressed genes in the HFD condition. The candidate gene list was (except *Ddx3x* and *Tap1*) exclusively selected from the HFD network. Taken together, our data support the diet-dependent differences in our omics datasets and point toward specific roles of the selected LCGs in hepatic metabolism.

### 2.7. Network genes are associated with human hepatic metabolism

To assess the translational relevance of the selected eleven candidate genes identified with the CoNI approach presented here, we validated them in human-derived liver biopsies. The expression of the eleven candidate genes was quantified by qPCR analyses in human liver biopsies of 170 patients. Anthropometrics and metabolic characteristics of these participants, which covered a wide range of hepatic triglyceride content, are shown in Tables S6A, B. Hepatic gene expression levels were then correlated with liver fat content and BMI. Associations of hepatic mRNA levels with insulin resistance (HOMA-IR) were additionally analyzed for a subgroup of 77 subjects. Significant associations of gene expression and metabolic traits were found for five of the eleven genes (Figure 5, Table 3). Expression of *GK*, *INHBE*, and *TAP1* in human livers was positively associated with BMI (Figure 5), which was in accordance with their increased expression in the livers of HFD-fed mice (Table S2). In contrast to its reduced hepatic expression in obese mice, we found that *MYC*'s expression increased with BMI in human livers (Figure 5). The expression of *SMIM13* correlated significantly with liver fat content (Figure 5). Strikingly, the expression of the LCG *INHBE* in the human livers was not only significantly associated with BMI, but also with liver TAG content and HOMA-IR (Figure 5); thus showing the strongest impact on cellular metabolism of all LCGs selected for validation. These findings show that five out of eleven murine-derived LCGs were associated with metabolic traits in humans. Finally, to test whether the expression of our candidate genes had an impact on cellular metabolite levels, we selected five representative genes to perform siRNA-mediated knockdown (KD) experiments followed by metabolic profiling using the Absolute*IDQ*<sup>TM</sup> p180 Kit in HepG2 cells. All five specific siRNAs significantly reduced target mRNA levels compared to a nontarget-siRNA (Figure S12A). After processing and filtering out metabolites under limits of detection (LOD), 107 metabolites were analyzed in total (Figure S12B). We observed a batch effect caused by performing the experiments on two consecutive days with  $n = 3$  replicates each, which we tried to minimize using ComBat [29]. Differentially regulated metabolites were estimated by ANOVA. After FDR correction, only the KD of *Rapgef4* and *Gk* caused significant effects on two and six metabolites, respectively (Table S7), but none of those were part of the predicted LCG sub-networks (Figure S12C). For *Inhbe*, none of the predicted ACs passed the LOD filter. Since the Chow



**Figure 3: Hepatic metabolite correlations under Chow and HFD feeding. (A)** Pairwise metabolite correlation matrix showing the correlation coefficients obtained for Chow (black) in the upper right and for HFD (red) in the lower left triangle. The metabolite classes are indicated by the respective color bars. Black boxes mark correlations between AC and PC in Chow and HFD. **(B)** Venn diagram showing significantly correlating metabolite pairs that differ and overlap between Chow and HFD. The sizes of the Venn diagrams correspond to the number of metabolite pairs. **(D, E)** Integrated graphs generated using CoNI for Chow **(D)** and HFD livers **(E)**. The nodes of the graphs refer to metabolites and edges to genes that significantly affect metabolite correlations. Node colors reflect the metabolite class. Unconnected nodes are displayed at the bottom. **(F)** Comparison of connected metabolite nodes between Chow (black) and HFD (red). **(G)** Comparison of significantly correlated metabolite pairs between the two graphs. **(H)** Comparison of genes contained in edges of both graphs. **(I)** The number of genes per edge and **(J)** number of edges per gene in both graphs.



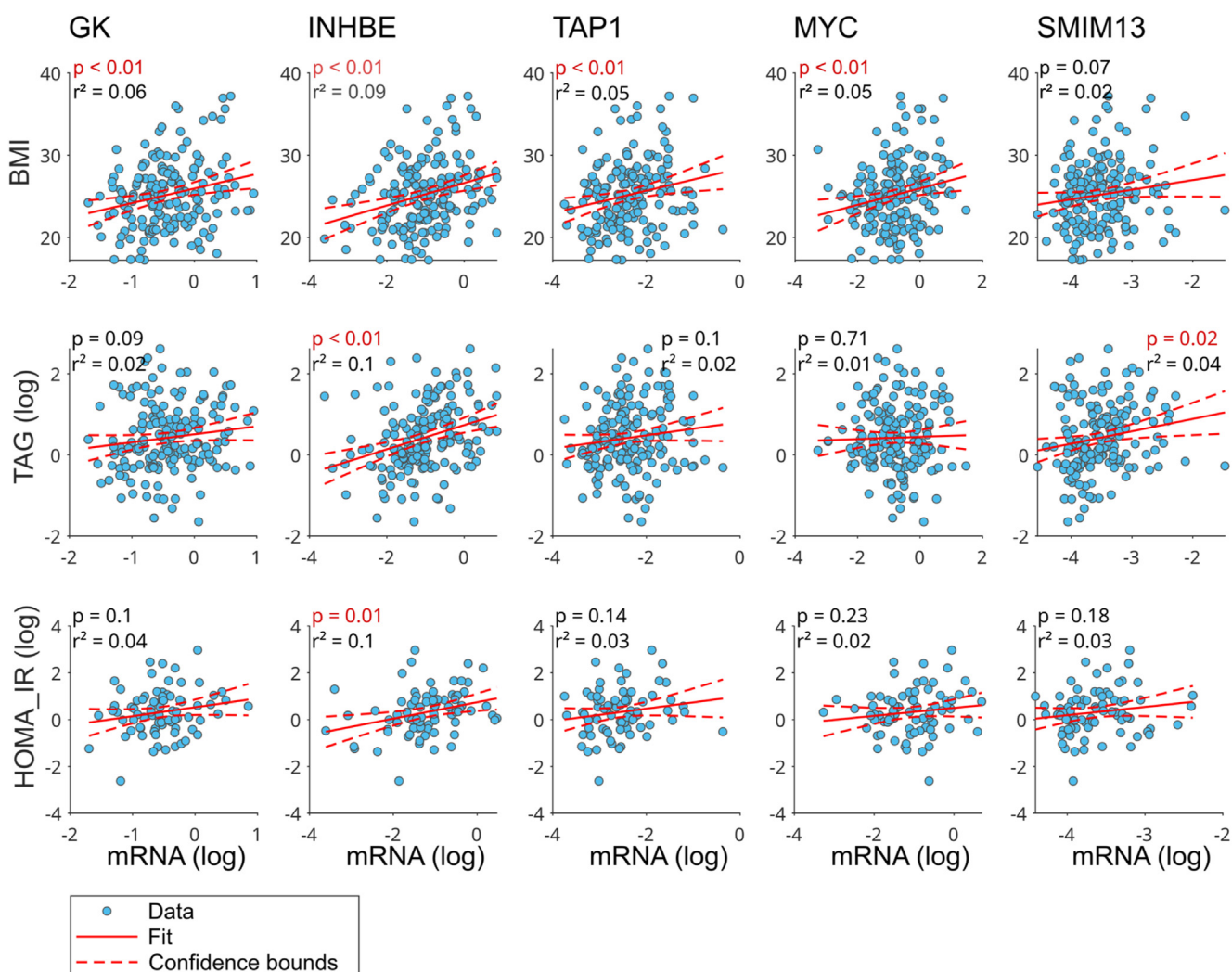
**Figure 4: Local controlling genes (LCGs) and associated metabolite subgraphs. (A–J)** Metabolite subgraphs for the selected LCGs in murine livers. Node colors refer to the different metabolite classes; the respective heatmaps show the concentrations of node metabolites. Gene expression levels of the respective LCG are displayed in the upper color bar for Chow (black) and HFD (red). Blue edges denote edges that contain the LCG in the sub-graph, grey edges are void of the LCG. LCGs were identified in the Chow (A) and HFD (B–J) network from which eight were differentially expressed depending on the diet (A–H) and for three genes SNPs associated with obesity and related disease markers could be identified (E,I,J). (K) Combined LCG network. Edges display direct gene – metabolite connections. Edge colors refer to diet, i.e., Chow (black) or HFD (red).



**Table 2** — Characteristics and selection criteria for the eleven genes further subjected to validation experiments: identification as LCG, present in a specific network; identification as obesity and type 2 diabetes-related genetic variant (SNP) and differential hepatic expression between chow and HFD mice (log2FC and p).

Gene	LCG	Network	SNP	log2FC	p adjusted
<i>Arhgap24</i>	yes	HFD	no	0.99	0.02
<i>Cd82</i>	yes	HFD	no	0.5	0.03
<i>Gk</i>	yes	HFD	no	0.73	0.01
<i>Inhbe</i>	yes	HFD	yes	1.2	0.02
<i>Myc</i>	yes	HFD	no	-1.31	0.04
<i>Rapgef4</i>	yes	HFD	no	-1.23	<0.00
<i>Smim13</i>	yes	HFD	no	-0.89	0.01
<i>Ddx3x</i>	yes	Chow	no	-0.6	0.01
<i>Cobll1</i>	yes	HFD	yes	-0.33	0.24
<i>Appl2</i>	yes	HFD	yes	0.63	0.15
<i>Tap1</i>	no	Chow/HFD	no	1.04	<0.00

and HFD CoNI networks were predicted as genetically controlled metabolite correlations, we next compared metabolite correlations between control and siRNA KD data. For all five genes tested, we observed a loss of metabolite correlations after KD in cells (Figure S12D). Comparison of correlation graphs from all significant correlations in the control group with the correlations affected by the respective KD (Fisher's z transformation) revealed that a minimum of 14 (*Gk*) up to 228 (*Appl2*) edges were affected by the KD (Figure 6A). We then extracted the correlation sub-networks build by the metabolites from the initial LCG's sub-networks (Figure 5) and compared their edges to the ones found in the KD sub-networks. For all four sub-networks with detectable metabolites, we found a significant reduction in correlated metabolites (Figure 6B). For *Appl2* and *Rapgef4*, no edge remained in the KD networks, for *Cobll1* and *Gk* only 3 of 25 and 3 of 15, respectively, could be found after KD. Finally, we compared all metabolite pairs predicted with CoNI between control and KD conditions (Figure 6C). For each gene, we found that between two to nine



**Figure 5:** CoNI identified genes correlate with hepatic lipid metabolism in humans. Pearson correlation analysis of human hepatic gene expression (log mRNA) for five selected genes: *GK*, *INHBE*, *MYC*, *SMIM13*, and *TAP1* compared to BMI (log) (upper row) and hepatic TAG content in mg/100 mg tissue (log) (middle row),  $n = 170$ . Human hepatic mRNA expression of *INHBE* correlated with HOMA-IR,  $n = 77$  (bottom row). Confidence bounds for the selected models are denoted by dashed lines. Significant p-values are colored red.



**Table 3** — Associations between hepatic biopsy gene expression levels and metabolic traits BMI (N = 170), hepatic TAG (N = 170), and HOMA-IR (N = 77) in human liver samples. Significant p Values (< 0.05) are highlighted in bold.

Genes	BMI p	BMI r2	HOMA-IR p (log)	HOMA-IR r2 (log)	TAG p (log)	TAG r2 (log)
<i>RAPGEF4</i>	0.75138	0.0006	0.23293	0.01892	0.70119	0.00088
<i>ARHGAP24</i>	0.09284	0.01672	0.86126	0.00041	0.08886	0.01713
<i>GK</i>	<b>0.00222</b>	0.05435	0.10077	0.0355	0.08875	0.01715
<i>COBLL1</i>	0.86194	0.00018	0.73025	0.00159	0.15734	0.01187
<i>INHBE</i>	<b>0.00014</b>	0.08261	<b>0.00781</b>	0.09061	<b>3E-05</b>	0.09904
<i>CD82</i>	0.89931	0.0001	0.64076	0.00292	0.10298	0.01575
<i>TAP1</i>	<b>0.00656</b>	0.04316	0.1444	0.02819	0.10229	0.01581
<i>SMIM13</i>	0.06884	0.01957	0.18164	0.02366	<b>0.01693</b>	0.03348
<i>DDX3X</i>	0.54836	0.00215	0.64243	0.00289	0.96771	1E-05
<i>MYC</i>	<b>0.00636</b>	0.04347	0.22892	0.01924	0.7102	0.00082
<i>APPL2</i>	0.30545	0.00625	0.97038	2E-05	0.61441	0.00151

metabolite pairs significantly changed their correlation. Furthermore, for *Appl2*, *Cobll1*, and *Gk*, we also saw a general reduction of metabolite correlations between control and KD cells.

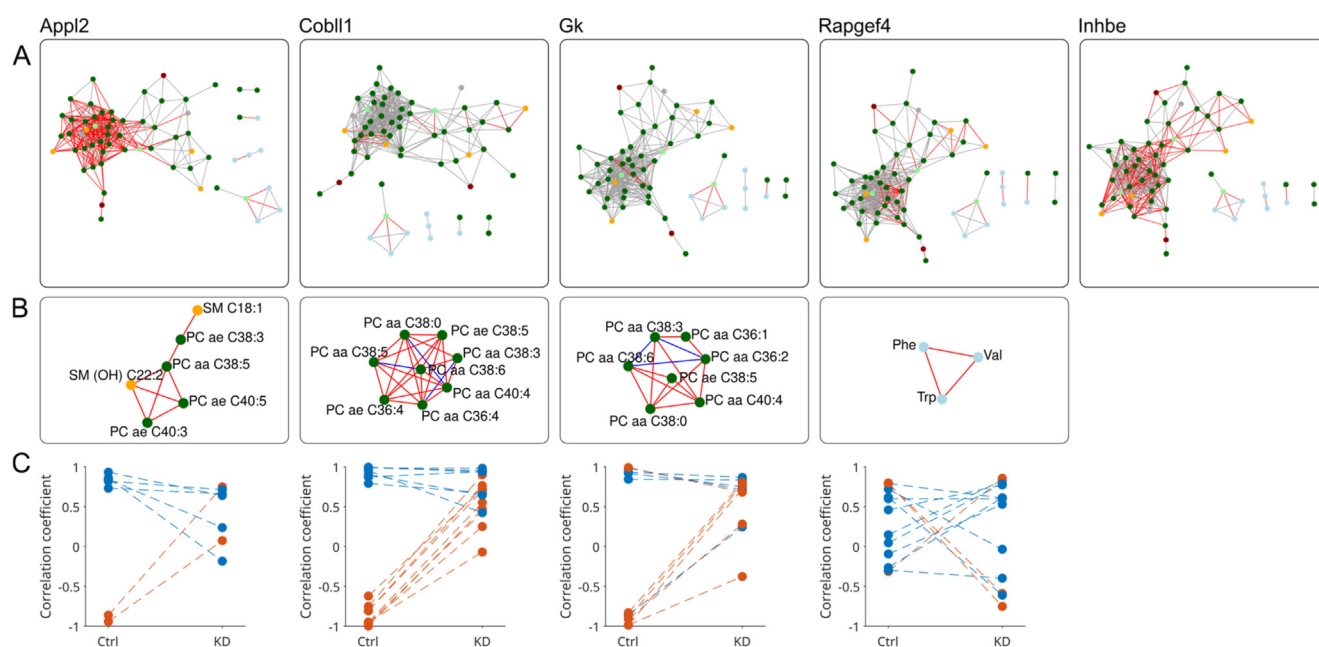
In summary, for eight of the eleven selected candidate genes, we could validate functional associations with hepatic metabolism by showing correlations of human hepatic transcript levels with hallmarks of clinical obesity or insulin resistance to changes in cellular metabolite correlations after siRNA KD *in vitro*. Of the three LCGs harboring SNPs that are associated with obesity or type 2 diabetes, *COBLL1* and *INHBE* could be confirmed with at least one performed validation experiment.

### 3. DISCUSSION

We here present with our CoNI approach a novel, fully unsupervised, and data-driven method that allows for the integration of different

omics data types based on a combined correlation approach followed by the construction of integrated hypergraphs. Our method is the first to introduce a new paradigm in correlation-based data integration. Molecular interrelationships are not inferred as paired interactions, but as triplets where one interactor is estimated as a controlling variable of the other two. This allows for a more natural reconstruction of metabolic networks where we estimate genes affecting correlations between metabolites. With the CoNI-driven integration of transcriptomics and metabolomics data from murine livers, we identified and unraveled previously hidden local controlling genes (LCG) that exerted major changes to hepatic metabolite correlation levels. In a well-characterized human cohort, we confirmed the translational relevance of these LCGs for human liver metabolism in obesity. Subsequently, *in vitro* experiments showed that the siRNA-mediated KD of selected LCGs perturbs correlation networks. In addition to the liver dataset, the CoNI-derived reconstruction of known and potentially novel protein-lipid interactions in murine lungs under clean air and smoking conditions (SI, Figures S11–13) demonstrated that our method is a versatile framework that can be applied to various data integration problems. We also postulate that CoNI can be applied not only to various kinds of omics data, but also to other nonbiological data where the aim is to investigate factors affecting network interactions. Our approach studies presented here are completely data-driven, which makes them applicable to various kinds of paired multivariate datasets from the same samples.

The requirement of paired samples is a conceptual limitation of CoNI, but in turn, allows the application on datasets without different conditions or temporal resolution like large cohort data. A further limitation is that Pearson's and partial correlations are prone to outliers which may increase the false positive predictions. This effect can be reduced, with increasing sample sizes. Additionally, although we here use partial correlation to uncover genes as confounding factors for



**Figure 6: siRNA KD of LCGs perturbs metabolite correlation networks.** (A) Correlation networks generated from metabolite correlations in HepG2 cells. Edges refer to significant correlations in control experiments. Red edges denote correlations significantly altered by siRNA-mediated knockdown of an LCG. (B) Correlation subgraphs of metabolites mapped to LCG subgraphs from CoNI applied to murine data. Red edges denote edges only present in the control; blue edges denote significant correlations in KD conditions. (C) Correlation coefficients of predicted metabolite pairs from murine CoNI networks in Control and KD condition. Red pairs denote correlations that significantly change upon LCG knockdown.

metabolite correlations, our method is unsuitable to infer any causality; thus interactions estimated with CoNI do not necessarily reflect directed regulations. The high computational costs become more and more insignificant with the increase of available high throughput computing. However, the major limitation that we perceive lies in the limitations of available datasets. In particular, for the prediction of genetic and metabolic interactions, the lack of detectable metabolites — either in targeted or untargeted metabolomics approaches — limits the capabilities to recover real molecular interaction, thus creating abstractions of molecular interaction networks. In contrast, prior knowledge-based approaches, that are successful in modeling specific pathways or biochemical reactions [30] face their limitations for the identification of novel or indirect interactions; in particular, when available data are incomplete. We here applied CoNI to a very unbalanced set of data — 180 targeted metabolites and thousands of transcripts — but successfully integrated these.

The CoNI networks allowed us to define LCGs that presumably play a role in the development of liver steatosis and hepatic insulin resistance in diet-induced obesity. None of the Chow or HFD LCGs was present in both networks, which is in accordance with recent findings that show the genetic control of metabolic networks widely being altered by diet [19]. Furthermore, for some of the LCGs, an involvement in lipid metabolism or diabetes-related traits had already been discussed. The membrane-associated protein *Tap1* was recently linked to the initiation and propagation of liver inflammation and insulin resistance in mice [31], which is in accordance with our finding of an association between hepatic TAP1 expression and BMI in humans. We found that the expression of the phosphotransferase enzyme glycerol kinase (*Gk*) — a gene that had been proposed as a regulator for several lipids [32] — was increased in the livers of obese mice and humans, thus suggesting an adaptive mechanism to handle the increased hepatic lipid load. This hypothesis is supported by the finding that overexpression of *Gk* favors recycling of free fatty acids leading to increased fat storage in rat hepatoma cells [33–35]. Also, *Myc* seems to be involved in the regulation of hepatic glycolysis [36,37]. Under HFD exposure, *Myc* overexpression in transgenic mice normalized glycemia, insulinemia, and the expression of genes involved in hepatic metabolism [38]. The potential regulatory role for the putative hepatokine *Inhbe* was confirmed by the strong associations of its hepatic expression with metabolic traits in humans. This is in accordance with previous reports identifying *Inhbe* as a diet-responsive gene in the rodent liver, regulated by HFD feeding, fasting, or refeeding [39–41]. Recently, Sugiyama et al. [24] demonstrated that the siRNA-mediated knockdown of *Inhbe* in obese insulin-resistant *Lep<sup>db</sup>* mice decreased fat mass and respiratory quotient, thus suggesting enhanced whole body fat usage. We here link *Inhbe* to the regulation of AC, which is also known to interfere with hepatic insulin sensitivity [28,42].

Only one of the genes selected for validation was among the top 100 differentially expressed genes between Chow and HFD. Here, we could not only identify genes involved in the metabolic rearrangement of the liver under HFD feeding that would have been remained undiscovered by traditional approaches, but also link them to metabolic pathways. Future studies that clarify the impact of the LCG expression changes on hepatic metabolite levels in humans are warranted, but are beyond the scope of this study.

Overall, our results indicate that the data integration approach, CoNI, is a useful method— to successfully integrate transcriptional data into metabolic networks to ultimately facilitate the identification of gene candidates involved in hepatic steatosis and comorbid hepatic insulin resistance in mice and humans. CoNI can be used to integrate various types of multidimensional omics data, and it can make them available

for useful holistic analyses in various fields of health research and beyond.

## 4. METHODS

### 4.1. Ethics statement

*In vivo* experiments were performed without blinding of the investigators. All studies were based on power analyses to assure adequate sample sizes and performed with approval by the State of Bavaria, Germany, under the following protocol numbers: 55.2.1-54-2532-75-13. The clinical study has been approved by the Ethical Review Board of the University Hospital Tübingen, and all human participants provided informed written consent.

### 4.2. Animals

All experiments were performed in 20 adult male C57BL/6J mice purchased from Janvier Labs (Saint-Berthevin, Cedex, France). Mice were maintained on a 12-h light–dark cycle with free access to water and a standard Chow diet (Altromin, #1314). To promote diet-induced obesity (DIO), mice were ad-libitum fed with a 58 % high-fat diet (HFD) (Research Diets, D12331) for 22 weeks. Two mice from the HFD cohort with body weights comparably lower to those of the other HFD-fed mice (BW at week 22: 37.5 g and 33.3 g) were excluded from all further analyses. Mice were fasted for 5 h and then sacrificed by cervical dislocation for organ withdrawal. Livers were removed immediately, flash-frozen in liquid nitrogen, and stored at  $-80^{\circ}\text{C}$  until further analysis. Two animals were excluded from the study cohort owing to their comparatively lower body weight gain on HFD.

#### 4.2.1. Plasma analysis

Blood was collected in tubes containing 50  $\mu\text{L}$  EDTA and then centrifuged at  $2000\times g$  and  $4^{\circ}\text{C}$  for 10 min. Plasma was collected and stored at  $-80^{\circ}\text{C}$  until further testing. Plasma triglycerides (TAG), cholesterol, and nonesterified fatty acids (NEFA) were measured by commercial enzymatic assay kits (WAKO Chemicals, Neuss, Germany). Insulin was measured by the ultrasensitive murine insulin ELISA kit (Merck Millipore, Darmstadt, Germany).

#### 4.2.2. Hepatic triglyceride content measurements

Hepatic triglyceride content was determined after chloroform/methanol (2:1) extraction by using the triglyceride assay kit according to the manufacturer's protocol (Wako Chemicals).

### 4.3. Metabolomics

#### 4.3.1. Tissue homogenization and metabolite extraction

Frozen murine liver samples were weighed, and metabolites were extracted as previously described in ice-cold extraction solvent, an 85/15 (v/v) ethanol/10 mM phosphate buffer pH 7.5 mixture at a ratio of 3  $\mu\text{L}$  solvent per 1 mg tissue weight [43]. The liver samples were homogenized using a Precellys24 homogenizer (PeqLab Biotechnology, Erlangen, Germany) thrice for 20 s at 5,500 rpm and  $4^{\circ}\text{C}$ , with 30 s pause intervals to ensure constant temperature, followed by centrifugation at  $4^{\circ}\text{C}$  and  $10,000\times g$  for 5 min. Subsequently, 10  $\mu\text{L}$  of the supernatants were used for metabolite quantification.

#### 4.3.2. siRNA knockdown and metabolite extraction in HepG2 cells

Cells were cultured in DMEM supplemented with 10% fetal bovine serum and antibiotics (penicillin 100 IU/ml and streptomycin 100  $\mu\text{g}/\text{ml}$ ) in 5%  $\text{CO}_2$  at  $37^{\circ}\text{C}$ . At 70–80% confluence, cells were transfected with five different human SMARTpool On Target plus siRNA clones (L-

006727-00, L-021435-02, L-009511-00, L-020477-02, L-016272-01; Dharmacon, Lafayette, USA) or ON-TARGETplus Nontargeting Pool (D-001810-10) using DharmafECT #4 (Dharmacon) for 48 h and subjected to metabolite extraction in an ice-cold 80/20 (v/v) methanol/water mixture as described previously [44]. Subsequently, 20  $\mu$ L of the supernatants were used for the metabolite quantification. Each target-specific siRNA and the nontarget control was transfected and measured in six biological replicates.

#### 4.3.3. RNA extraction and qPCR in HepG2 cells

RNA was extracted from HepG2 cells after siRNA KD using the NucleoSpin RNA isolation kit (Macherey–Nagel, Düren, Germany). Equal amounts of RNA were reverse transcribed to cDNA using the QuantiTect Reverse Transcription kit (Qiagen, Hilden, Germany). Gene expression was analyzed using TaqMan probes for *APPL2* (Hs01565861\_m1), *COBLL1* (Hs01117513\_m1), *GK* (Hs02340007\_g1), *INHBE* (Hs00368884\_g1), *RAPGEF4* (Hs00199754\_m1), and *HPRT* (Hs02800695\_m1) as the housekeeping gene with the respective TaqMan mastermix (Thermo Fischer Scientific, Inc., Rockford, IL USA). qPCRs were carried out using a Quantstudio 6 real-time PCR system (Applied Biosystems). Gene expression was evaluated using the  $\Delta$ - $\Delta$  Ct method.

#### 4.3.4. Fluorescence-based DNA quantification in cell homogenates

To normalize the obtained metabolomics data from cell homogenates for differences in cell numbers, the DNA content was determined using fluorochrome Hoechst 33342 (ThermoFisher Scientific, Schwerte, Germany) and a GloMax Multi Detection System (Promega, Mannheim, Germany) from a small aliquot taken before the final centrifugation step, as previously described [44].

#### 4.3.5. Metabolite quantification by Absolute/*DQ*<sup>TM</sup> p180 kit

The targeted metabolomics approach was based on liquid chromatography-electrospray ionization-tandem mass spectrometry (LC-ESI-MS/MS) and flow injection-electrospray ionization-tandem mass spectrometry (FIA-ESI-MS/MS) measurements using the Absolute/*DQ*<sup>TM</sup> p180 kit (BIOCRATES Life Sciences AG, Innsbruck, Austria). The assay allows simultaneous quantification of 188 metabolites out of 10  $\mu$ L tissue lysate or 20  $\mu$ L cell lysate, and includes free carnitine, 39 acylcarnitines (Cx:y), 21 amino acids (19 proteinogenic + citrulline + ornithine), 21 biogenic amines, hexoses (sum of hexoses — about 90–95% glucose), 90 glycerophospholipids (14 lysophosphatidylcholines (LPC) and 76 phosphatidylcholines (PC)), and 15 sphingolipids (SMx:y). The abbreviations Cx:y are used to describe the total number of carbons and double bonds of all chains, respectively [43,45]. For the LC-part, compound identification and quantification were based on scheduled multiple reaction monitoring measurements (sMRM). The method of Absolute/*DQ*<sup>TM</sup> p180 kit has been proven to be in conformance with the EMEA-Guideline “Guideline on bioanalytical method validation (July 21st, 2011)” [46], which implies proof of reproducibility within a given error range. Sample preparation and LC-MS/MS measurements were performed as described in the manufacturer manual UM-P180. The limits of detection (LOD) were set to three times the values of the zero samples (PBS). The lower and upper limits of quantification (LLOQ and ULOQ) were determined experimentally by Biocrates. The assay procedures of the Absolute/*DQ*<sup>TM</sup> p180 kit and the metabolite nomenclature have been described in detail previously [47]. Metabolite concentrations were calculated using internal standards and reported in  $\mu$ M.

## 4.4. Transcriptomics

### 4.4.1. RNA preparation and microarray analysis

Microarray data were obtained from liver samples of 10 chow and 8 HFD-fed mice. Total RNA was isolated from tissues employing a commercially available kit (NucleoSpin RNA, #740955, Macherey–Nagel, Düren, Germany). Total RNA (150 ng, RIN > 7) was amplified using the WT PLUS Reagent Kit (Affymetrix, Santa Clara, US). Amplified cDNA was hybridized on Mouse Clariom S arrays (Affymetrix). Staining and scanning were performed according to the Affymetrix expression protocol. Expression console (v.1.4.1.46, Affymetrix) was used for quality control and to obtain annotated normalized RMA gene level data (Gene Level - SST-RMA). Genes with low expression levels (probe intensity < 40 in 5 out of 18 samples) were removed from the data set. For probe sets with identical values across all samples, only one probe set was kept in the final gene sets. Before calculating the partial correlation coefficients, genes with high within-group variance (variance > 0.5) were excluded from the downstream analysis to reduce the number of identified false positives because of noisy expression patterns. This resulted in 10,159 gene expression profiles that were used in the downstream analysis.

## 4.5. Human data

### 4.5.1. Patients with liver tissue samples

For the analysis of gene expression in human liver tissue samples, a cohort of 170 men and women of European descent undergoing liver surgery at the Department of General, Visceral, and Transplant Surgery at the University Hospital of Tübingen (Tübingen, Germany) was included in the present study. Participants fasted overnight before collection of the liver biopsies, and in a subgroup of 77 individuals, fasting plasma samples for the calculation of the homeostasis model assessment of insulin resistance (HOMA-IR) were also obtained as proposed by Matthews et al. [48]. Characteristics are shown in Table S5A for the whole group and in Table S5B for the subgroup with fasting plasma samples. All patients tested negative for viral hepatitis and had no liver cirrhosis. Only samples from normal, nondiseased tissue, judged by an experienced pathologist, were used. Informed, written consent was obtained from all participants, and the Ethics Committee of the University of Tübingen approved the protocol (239/2013B01) according to the Declaration of Helsinki. Liver samples were taken from normal, nondiseased tissue during surgery, immediately frozen in liquid nitrogen, and stored at  $-80$  °C.

### 4.5.2. Determination of liver tissue triglyceride content

Liver tissue samples were homogenized in phosphate-buffered saline containing 1% Triton X-100 with a TissueLyser (Qiagen, Hilden, Germany). To determine the liver fat content, triglyceride concentrations in the homogenate were quantified using an ADVIA XPT clinical chemistry analyzer (Siemens Healthineers, Eschborn, Germany), and the results were calculated as TAG(mg)/100 mg tissue weight.

### 4.5.3. Real-time PCR

For real-time (RT)-PCR and quantitative RT-PCR analyses of hepatic mRNA expression in liver biopsies, frozen tissue was homogenized in a TissueLyser (Qiagen), and RNA was extracted with the RNeasy Tissue kit (Qiagen) according to the manufacturer's instructions. Total RNA treated with RNase-free DNase I was transcribed into cDNA using a



first-strand cDNA kit, and PCRs were performed in duplicates on a LightCycler480 (Roche Diagnostics, Mannheim, Germany). The human primer sequences that were used are shown in Table S8. Data are presented relative to the housekeeping gene Rps13 using the  $\Delta\text{-}\Delta$  Ct method.

#### 4.5.4. Quantification of blood parameters

Plasma insulin was determined on the ADVIA Centaur XPT chemiluminometric immunoassay system. Fasting plasma glucose concentrations were measured using the ADVIA XPT Clinical chemistry analyzer (both from Siemens Healthineers, Eschborn, Germany).

#### 4.6. Data availability

The microarray data have been submitted to the GEO database at NCBI (GSE137923: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc>). All other data generated or analyzed during this study are available within the study and its supplementary information file. Human data are not available at the patient level because of data protection regulation.

#### 4.7. Statistics

##### 4.7.1. Metabolite data analysis

After removal of metabolites with nonimputable concentration levels (>5 % missing values), TwoGroup from the R metabolomics package [49] was used to compare the log<sub>2</sub>-normalized concentration levels between the Chow and HFD-fed mice. The siRNA KD altered metabolites in HepG2 cells were compared using a two-way ANOVA with experiment day as an additional covariate. Unless otherwise stated, all metabolites showing a Benjamini–Hochberg [50] corrected p-value less than 0.05 were defined to differ significantly with respect to their concentration. For the downstream analysis, the log<sub>2</sub>-transformed metabolite concentrations were scaled using the square root of the standard deviation as a scaling factor (Pareto scaling) [51].

##### 4.7.2. Differential gene expression analysis of murine liver samples

After log<sub>2</sub>-transformation, the R package limma [52] (linear model for microarray data) was applied to infer differential expression between the two diet groups. We defined all genes with Benjamini–Hochberg corrected p-value less than 0.05 to be significantly deregulated.

##### 4.7.3. Human data analyses

Data that were not normally distributed (Shapiro–Wilk W-test) were logarithmically transformed. Univariate associations between parameters were tested using Pearson correlation analyses. To adjust the effects of covariates and identify independent relationships, multivariate linear regression analyses were used. The statistical software package JMP 14.0 (SAS Institute, Cary, NC) was used.

##### 4.7.4. Correlation analyses

For each metabolite pair,  $M_i$  and  $M_j$  given,  $M, i = 1, \dots, n, j = 1, \dots, n, i \neq j$ , with  $n$  metabolites, the Pearson correlation coefficients were obtained with R package ‘Hmisc’. Significant differences in metabolite correlations under different dietary conditions were tested using Steiger’s test [15] function of R’s cocor package [53].

##### 4.7.5. Identification of main influencing factor

Principal component analysis (PCA) was performed on each set to find the main factor separating the samples. Tested factors were as follows: Bodyweight measured at the end of the study (BWE), liver

triglyceride level (TAG), and administered diet. To assess whether the factors differ between Chow and HFD, we applied the Wilcoxon signed rank test.

##### 4.7.6. Pathway enrichment analysis of differentially expressed genes

Transcriptional enrichments were calculated using the R package ClusterProfiler [54] to test for overrepresented GO [55] biological process terms. We summarized terms that were completely contained in another term with respect to the enriched gene list.

##### 4.8. Regression models predicting body weight, plasma insulin, and liver TAG

Models were fitted and analyzed using MATLAB 2020b. The significance of enriched models for each gene group was estimated using Fisher’s exact test.

##### 4.8.1. Partial correlation based network integration (CoNI)

The framework (Figure 1) includes three steps carried out for each treatment group independently: 1) Performing pairwise correlation analysis on metabolite data set; 2) Partial correlation analysis combining the metabolite concentrations with the gene expression profiles; and 3) Construction of undirected, weighted graph.

**1) Correlation analysis.** First, for  $M$  metabolites, the  $M \times M$  correlation matrix was calculated. Here, for subsequent analyses, only metabolite pairs showing a Pearson correlation p-value < 0.05 were selected.

**2) Partial correlation based gene extension.** For each pair of selected metabolites,  $M_i$  and  $M_j$  given,  $M, i = 1, \dots, M, j = 1, \dots, M, i \neq j$  and each gene  $G_k$  with  $k = 1, \dots, K$  the partial correlation  $\rho(M_i M_j^* G_k)$  reflecting the correlation between  $M_i$  and  $M_j$  after removing the linear effects of gene  $G_k$  were calculated using R’s package ppcor [56]. Such a combination was denoted as a triplet. Steiger’s test was adapted to select triplets, where the partial correlation coefficient differed significantly from the correlation coefficient of the respective metabolite pair. The original test assesses the significance for the difference between two correlation coefficients that have one variable in common. The significance depends on the intercorrelation between the two variables that are not shared, which has to be provided as an additional parameter. To use this test and compare a partial correlation coefficient and a correlation coefficient, the test was applied twice. The provided, additional parameter in the first test was the correlation between  $M_i$  and  $G_k$ , and in the second test, the correlation between  $M_j$  and  $G_k$ . To be selected, the triplet had to significantly reject the null hypothesis (Bonferroni adjusted p-value < 0.05), stating that the correlation coefficients did not differ in both tests. The method `cocor.dep.groups.overlap` of R’s cocor package [53] was used to perform the testing.

**3) Undirected graph construction and clustering.** Next, an undirected and weighted graph was generated where nodes are formed by metabolites and genes set up the edges. Edges were drawn if a metabolite pair correlated and this correlation was significantly influenced by at least one gene. Several genes can connect more than two correlating metabolites and a pair of metabolites might also be connected by more than one gene. The number of genes connecting the two respective metabolite nodes determines the edge weight.

#### 4.8.2. Local controlling genes (LCGs)

Starting from each node in the network, the appearance of each gene in all edges connecting nodes with a distance  $\leq$  two was counted. Statistical significance was estimated using a binomial distribution test. P-values were Bonferroni corrected for multiple testing and genes with adjusted p-value  $< 0.05$  were defined as LCGs.

#### 4.8.3. Communities

To find densely connected subgraphs in the graph, the fast greedy modularity optimization algorithm implemented in the igraph package [57] was applied.

### AUTHOR CONTRIBUTIONS

V.S.K. and D.L. conceived the method. V.S.K. and J.M.M-K. wrote the code and analyzed the data. S.C.S. performed the mouse study and *ex vivo* and *in vivo* experiments. M.I. and J.B. performed transcriptomics microarray analysis. J.T., C.P., G.K., and J.A. performed metabolomics. A.P., A.K., and M.H. contributed the human liver data. M.H.T. and T.D.M. contributed reagents, materials, and critical input. P.T.P. and D.L. conceptualized the project. V.S.K., S.C.S., P.T.P., and D.L. interpreted the data. V.S.K., S.C.S., and D.L. wrote the article. All authors critically reviewed and edited the manuscript and approved the final version.

### FUNDING

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.), by the German Research Foundation (DFG) grants TRR152 and TRR-296, by the Helmholtz Portfolio Program “Metabolic Dysfunction” (MHT), by the Alexander von Humboldt Foundation (MHT), by DZD tandem grant funds (SCS, PP, MH), by the Helmholtz-Israel-Cooperation in Personalized Medicine (PP), by the Helmholtz Initiative for Personalized Medicine (iMed; MHT), by Helmholtz Alliance Aging and Metabolic Programming (AMPPro), and through the Initiative and Networking Fund of the Helmholtz Association.

### ACKNOWLEDGMENTS

We gratefully acknowledge Alke Guirguis and Ann-Kathrin Horlacher (Institute for Clinical Chemistry and Pathobiochemistry, University Hospital Tübingen, Germany) and Silke Becker (Research Unit Molecular Endocrinology and Metabolism, HMGU) for their technical assistance.

### CONFLICT OF INTEREST

Dr. Matthias Tschöp is a member of the scientific advisory board of ERX Pharmaceuticals, Inc., Cambridge, MA. He was a member of the Research Cluster Advisory Panel (ReCAP) of the Novo Nordisk Foundation between 2017 and 2019. He attended a scientific advisory board meeting of the Novo Nordisk Foundation Center for Basic Metabolic Research in 2016. He received funding for his research projects from Novo Nordisk (2016–2020) and Sanofi-Aventis (2012–2019). He was a consultant for Bionorica SE (2013–2017), Menarini Ricerche S.p.A. (2016), and Bayer Pharma AG Berlin (2016).

As former Director of the Helmholtz Diabetes Center and the Institute for Diabetes and Obesity at Helmholtz Zentrum München (2011–2018) and since 2018 as CEO of Helmholtz Zentrum München, he has been responsible for collaborations with a multitude of companies and institutions, worldwide. In this capacity, he discussed potential projects with and has signed/signs contracts for his institute(s) and the staff for research funding and/or collaborations with industry and academia, worldwide;

including, but not limited to pharmaceutical corporations such as Boehringer Ingelheim, Eli Lilly, Novo Nordisk, Medigene, Arbormed, BioSyngen, and others. In this role, he was/is further responsible for commercial technology transfer activities of his institute(s), including diabetes-related patent portfolios of Helmholtz Zentrum München, e.g. WO/2016/188932 A2 or WO/2017/194499 A1. Dr. Tschöp confirms that to the best of his knowledge, none of the above funding sources was involved in the preparation of this study.

### APPENDIX A. SUPPLEMENTARY DATA

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.molmet.2021.101295>.

### REFERENCES

- [1] Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G., et al., 2016. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics*, 15, 17 Suppl 2.
- [2] Hawe, J.S., Theis, F.J., Heinig, M., 2019. Inferring interaction networks from multi-omics data. *Frontiers in Genetics* 10:535.
- [3] Krumsiek, J., Suhre, K., Illig, T., Adamski, J., Theis, F.J., 2011. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology* 5:21.
- [4] Parker, B.L., Calkin, A.C., Seldin, M.M., Keating, M.F., Tarling, E.J., Yang, P., et al., 2019. An integrative systems genetic analysis of mammalian lipid metabolism. *Nature* 567(7747):187–193.
- [5] Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., et al., 2014. An atlas of genetic influences on human blood metabolites. *Nature Genetics* 46(6):543–550.
- [6] Fahrman, J.F., Grapov, D.D., Wanichthanarak, K., DeFelice, B.C., Salemi, M.R., Rom, W.N., et al., 2017. Integrated metabolomics and proteomics highlight altered nicotinamide- and polyamine pathways in lung adenocarcinoma. *Carcinogenesis*.
- [7] Kindt, A., Liebisch, G., Clavel, T., Haller, D., Hormannspurger, G., Yoon, H., et al., 2018. The gut microbiota promotes hepatic fatty acid desaturation and elongation in mice. *Nature Communications* 9(1):3760.
- [8] Kresnowati, M.T., Van Winden, W.A., Almering, M.J., Ten Pierick, A., Ras, C., Knijnenburg, T.A., et al., 2006. When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Molecular Systems Biology* 2:49.
- [9] Blazier, A.S., Papin, J.A., 2012. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology* 3:299.
- [10] Moreno-Sanchez, R., et al., 2008. Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways. *Journal of Biomedicine and Biotechnology* 2008:597913.
- [11] ter Kuile, B.H., Westerhoff, H.V., 2001. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters* 500(3):169–171.
- [12] Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vazquez-Fresno, R., et al., 2018. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Research* 46(D1):D608–D617.
- [13] Stefan, N., Haring, H.U., Cusi, K., 2019. Non-alcoholic fatty liver disease: causes, diagnosis, cardiometabolic consequences, and treatment strategies. *Lancet Diabetes Endocrinol* 7(4):313–324.
- [14] Stefan, N., Fritsche, A., Schick, F., Haring, H.U., 2016. Phenotypes of prediabetes and stratification of cardiometabolic risk. *Lancet Diabetes Endocrinol* 4(9):789–798.
- [15] Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin* 87(2):245.
- [16] Phillips, B., Veljkovic, E., Peck, M.J., Buettner, A., Elamin, A., Guedj, E., et al., 2015. A 7-month cigarette smoke inhalation study in C57BL/6 mice

- demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product. *Food and Chemical Toxicology* 80:328–345.
- [17] Titz, B., Boué, S., Phillips, B., Talikka, M., Vihervaara, T., Schneider, T., et al., 2015. Effects of cigarette smoke, cessation, and switching to two heat-not-burn tobacco products on lung lipid metabolism in C57BL/6 and Apoe<sup>-/-</sup> Mice—an integrative systems toxicology analysis. *Toxicological Sciences* 149(2):441–457.
- [18] Bartel, J., Krumsiek, J., Schramm, K., Adamski, J., Gieger, C., Herder, C., et al., 2015. The human blood metabolome-transcriptome interface. *PLoS Genetics* 11(6):e1005274.
- [19] Dyar, K.A., Lutter, D., Artati, A., Ceglia, N.J., Liu, Y., Armenta, D., et al., 2018. Atlas of circadian metabolism reveals system-wide coordination and communication between clocks. *Cell* 174(6):1571–1585 e11.
- [20] T2Dkp, 2018. Type 2 diabetes knowledge portal. Nov 15th. <http://www.type2diabetesgenetics.org>.
- [21] Liu, Z., Xiao, T., Peng, X., Li, G., Hu, F., 2017. APPLs: More than just adiponectin receptor binding proteins. *Cellular Signalling* 32:76–84.
- [22] Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., et al., 2018. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics* 50(11):1505–1513.
- [23] Mahajan, A., Wessel, J., Willems, S.M., Zhao, W., Robertson, N.R., Chu, A.Y., et al., 2018. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nature Genetics* 50(4):559–571.
- [24] Sugiyama, M., Kikuchi, A., Misu, H., Igawa, H., Ashihara, M., Kushima, Y., et al., 2018. Inhibin  $\beta$ E (INHBE) is a possible insulin resistance-associated hepatokine identified by comprehensive gene expression analysis in human liver biopsy samples. *PLoS One* 13(3):e0194798.
- [25] van der Veen, J.N., Kennelly, J.P., Wan, S., Vance, J.E., Vance, D.E., Jacobs, R.L., 2017. The critical role of phosphatidylcholine and phosphatidylethanolamine metabolism in health and disease. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1859(9 Pt B):1558–1572.
- [26] Hasegawa, T., Iino, C., Endo, T., Mikami, K., Kimura, M., Sawada, N., et al., 2020. Changed amino acids in NAFLD and liver fibrosis: a large cross-sectional study without influence of insulin resistance. *Nutrients*(5):12.
- [27] Jin, R., Banton, S., Tran, V.T., Konomi, J.V., Li, S., Jones, D.P., et al., 2016. Amino acid metabolism is altered in adolescents with nonalcoholic fatty liver disease—an untargeted, high resolution metabolomics study. *The Journal of Pediatrics* 172:14–19 e5.
- [28] Schooneman, M.G., Vaz, F.M., Houten, S.M., Soeters, M.R., 2013. Acylcarnitines: reflecting or inflicting insulin resistance? *Diabetes* 62(1):1–8.
- [29] Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118–127.
- [30] Gille, C., Bolling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hubner, K., et al., 2010. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular Systems Biology* 6: 411.
- [31] Arindkar, S., Bhattacharjee, J., Kumar, J.M., Das, B., Upadhyay, P., Asif, S., et al., 2013. Antigen peptide transporter 1 is involved in the development of fructose-induced hepatic steatosis in mice. *Journal of Gastroenterology and Hepatology* 28(8):1403–1409.
- [32] Hara-Chikuma, M., Sohara, E., Rai, T., Ikawa, M., Okabe, M., Sasaki, S., et al., 2005. Progressive adipocyte hypertrophy in aquaporin-7-deficient mice adipocyte glycerol permeability as a novel regulator of fat accumulation. *Journal of Biological Chemistry* 280(16):15493–15496.
- [33] Hibuse, T., Maeda, N., Funahashi, T., Yamamoto, K., Nagasawa, A., Mizunoya, W., et al., 2005. Aquaporin 7 deficiency is associated with development of obesity through activation of adipose glycerol kinase. *Proceedings of the National Academy of Sciences* 102(31):10993–10998.
- [34] Maeda, N., Funahashi, T., Hibuse, T., Nagasawa, A., Kishida, K., Kuriyama, H., et al., 2004. Adaptation to fasting by glycerol transport through aquaporin 7 in adipose tissue. *Proceedings of the National Academy of Sciences* 101(51): 17801–17806.
- [35] Sriram, G., Parr, L.S., Rahib, L., Liao, J.C., Dipple, K.M., 2010. Moonlighting function of glycerol kinase causes systems-level changes in rat hepatoma cells. *Metabolic Engineering* 12(4):332–340.
- [36] Riu, E., Bosch, F., Valera, A., 1996. Prevention of diabetic alterations in transgenic mice overexpressing Myc in the liver. *Proceedings of the National Academy of Sciences* 93(5):2198–2202.
- [37] Valera, A., Pujol, A., Gregori, X., Riu, E., Visa, J., Bosch, F., 1995. Evidence from transgenic mice that myc regulates hepatic glycolysis. *The FASEB Journal* 9(11):1067–1078.
- [38] Riu, E., Ferre, T., Hidalgo, A., Mas, A., Franckhauser, S., Otaegui, P., et al., 2003. Overexpression of c-myc in the liver prevents obesity and insulin resistance. *The FASEB Journal* 17(12):1715–1717.
- [39] Hashimoto, O., Funaba, M., Sekiyama, K., Doi, S., Shindo, D., Satoh, R., et al., 2018. Activin E controls energy homeostasis in both brown and white adipose tissues as a hepatokine. *Cell Reports* 25(5):1193–1203.
- [40] Hashimoto, O., Sekiyama, K., Matsuo, T., Hasegawa, Y., 2009. Implication of activin E in glucose metabolism: transcriptional regulation of the inhibin/activin  $\beta$ E subunit gene in the liver. *Life Sciences* 85(13–14):534–540.
- [41] Rodgarkia-Dara, C., Vejda, S., Erlach, N., Losert, A., Bursch, W., Berger, W., et al., 2006. The activin axis in liver biology and disease. *Mutation Research: Reviews in Mutation Research* 613(2–3):123–137.
- [42] Magkos, F., Su, X., Bradley, D., Fabbrini, E., Conte, C., Eagon, J.C., et al., 2012. Intrahepatic diacylglycerol content is associated with hepatic insulin resistance in obese subjects. *Gastroenterology* 142(7):1444–1446 e2.
- [43] Zukunft, S., Prehn, C., Röhring, C., Möller, G., de Angelis, M.H., Adamski, J., et al., 2018. High-throughput extraction and quantification method for targeted metabolomics in murine tissues. *Metabolomics* 14(1):18.
- [44] Muschet, C., Möller, G., Prehn, C., de Angelis, M.H., Adamski, J., Tokarz, J., 2016. Removing the bottlenecks of cell culture metabolomics: fast normalization procedure, correlation of metabolites to cell number, and impact of the cell harvesting method. *Metabolomics* 12(10):151.
- [45] Römisch-Margl, W., Prehn, C., Bogumil, R., Röhring, C., Suhre, K., Adamski, J., 2012. Procedure for tissue sample preparation and metabolite extraction for high-throughput targeted metabolomics. *Metabolomics* 8(1): 133–142.
- [46] Agency, E.M., 2011. Guideline on bioanalytical method validation. Committee for medicinal products for human use (EMA/CHMP/EWP/192217/2009).
- [47] Zukunft, S., Sorgenfrei, M., Prehn, C., Möller, G., Adamski, J., 2013. Targeted metabolomics of dried blood spot extracts. *Chromatographia* 76(19–20): 1295–1305.
- [48] Matthews, D., Hosker, J., Rudenski, A., Naylor, B., Treacher, D., Turner, R., 1985. Homeostasis model assessment: insulin resistance and  $\beta$ -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia* 28(7):412–419.
- [49] De Livera, A.M., Olshansky, M., Speed, T.P., 2013. Statistical analysis of metabolomics data. *Methods in Molecular Biology* 1055:291–307.
- [50] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1):289–300.
- [51] van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7:142.



- [52] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., et al., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 43(7):e47 e47.
- [53] Diedenhofen, B., Musch, J., 2015. cocor: a comprehensive solution for the statistical comparison of correlations. *PloS One* 10(4):e0121945.
- [54] Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 16(5):284–287.
- [55] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* 25(1):25–29.
- [56] Kim, S., 2015. Ppcor: an R Package for a fast Calculation to semi-partial correlation coefficients. *Commun Stat Appl Methods* 22(6): 665–674.
- [57] Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. *InterJournal. Complex Systems* 1695(5):1–9.