Yazdani et al., Metabolomics (Los Angels) 2018, 8:2
DOI: 10.4172/2153-0769.1000199

Metabolomics : Open Access

Research Article Open Access

Integrated Systems Approach Identifies Pathways from the Genome to Triglycerides through a Metabolomic Causal Network

Azam Yazdani[1*], Akram Yazdani[2], Philip L. Lorenzi[3] and Ahmad Samiei[4]

[1]*Department of Epidemiology, Human Genetics Center, 1200 Pressler Street, Suite E-523, Houston 77030, Texas, USA*

[2]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, 10029, USA*

[3]*Department of Bioinformatics and Computational Biology, The Proteomics and Metabolomics Core Facility, The University of Texas, Houston 77054, Texas, USA*

[4]*Hasso Plattner Institute, 14482 Potsdam, Germany*

[*]**Corresponding author:** Azam Yazdani, Department of Epidemiology, School of Public Health, University of Texas, 1200 Pressler Street, Suite E-523, Houston, Texas, 77030, USA, Tel: 713-500-9808; E-mail: Azam.Yazdani@uth.tmc.edu

Received date: 21 Feb 2018; **Accepted date:** 20 March 2018; **Published date:** 27 March 2018

Abstract

To leverage functionality and clinical relevance into understanding systems biology, one needs to understand the pathway of the genetic effects on risk factors/disease through intermediate molecular level. Systems approaches integrate multi-omics information to find pathways to disease endpoints and make optimal inference decisions. Here, we introduce a multi-stage approach to integrate causal networks and GWAS to facilitate mechanistic understanding through identification of pathways from the genome to risk factors/disease via metabolomics, as an intermediate molecular level. The pathways in causal networks reveal the underlying relationships behind observations to facilitate mechanistic understanding, which do not play a significant role in more traditional correlative analyses, where one variable at a time is considered.

We identified a causal network over the metabolomic level to systematically assess whether variations in the genome lead to variations in triglyceride levels as a risk factor of cardiovascular disease. We found LRRC46 and LRRC69 harbouring loss-of-function mutations have significant effect on two metabolites with direct effects on triglyceride levels. We also found pathways of FAM198B and C6orf25 to triglycerides through indirect paths.

Integrating causal networks with GWAS facilitates mechanistic understanding in comparison to one-phenotype-at-a-time approaches due to accounting for relationships among phenotypes at intermediate molecular levels.

Keywords: Integrated systems approach; Metabolomic causal network; Bayesian network; Mendelian randomization; Triglycerides; Loss of function mutation

Introduction

Genome-wide association studies (GWAS) and recently whole genome sequence (WGS) studies have been widely conducted in humans with the goal of identifying genetic factors predictive of disease. Despite the extensive discovery of those studies, much of the genetic contributions to complex phenotypes remain unexplained. Furthermore, conclusions of noticeable numbers of genetic studies have not been in agreement with clinical presentation in an individual patient [1]. A key attribute for increasing confidence in potential clinical validity of gene variation with risk factors and disease end-points will be the development of assays with more direct mechanistic link. It is biologically meaningful that with a chronic systemic disease, molecular signals more proximal to the disease process may serve as strong biomarkers [2] and as a result, identify more stable pathways from the genome to disease risk factors and end-points. Therefore, integration of information in orthogonal data from different omics provides mechanistic understanding and has attracted attentions [3].

The metabolome is the end product of gene-environment interactions, Figure 1, and may be risk factors for future disease or biomarkers of current disease processes [4-6]. Metabolites can serve as intermediate phenotypes for genomic studies to illuminate mechanisms underlying of a specific SNP/gene, identify biological pathways linking the genome to disease, and discover valuable clinical biomarkers [2,3,7,8]. Therefore, integration of genetics and metabolomics holds potential for elucidating mechanisms for deciphering chronic disease.

Integration of data at different omics, such as genomics, metabolomics, and risk factors/disease endpoints, is challenging in modern biomedical research due to having largescale datasets and association between components. Most of the attempts at large scales are based on one component at a time that do not take into account the underlying relationships among components and lead to association studies. Identification of causal networks based on Mendelian randomization and Bayesian graphical modeling is an established approach to discover relations among components of interest and reduce the risk of false positive discovery. Furthermore, this approach controls for unmeasured confounders at the intermediate level, and presents a one-to-one cause and effect relationship between each two components for the purpose of facilitating mechanistic understanding [9-12]. Moreover, it makes hypotheses for efficacious targets for further experimental studies, which is necessary in the age of big data sets [8,13].

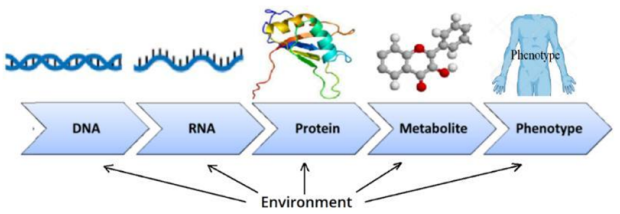

Figure 1: Metabolomics an integrated readout of many biological processes, such as genetic, transcriptomic, and proteomic variation, can characterize and recognize metabolic signatures of common chronic diseases.

In this study, we introduce an integrative approach to identify pathways from the genome to risk factors via metabolomics. To provide insights into underlying relationships among metabolites and represent how their effects spread across metabolomic system, a metabolomic causal network is identified by leveraging near-complete information from the genome based on Mendelian principles. After that, we analyze relation between metabolomics and the risk factor of interest using identified metabolomic network. This narrows metabolomic search space and allows focusing on a subset of metabolites with high impact on the risk factor. To complete the pathway from the genome to the risk factor, we focus on GWAS-metabolite studies. Using this approach, we determine how genome variation leads to variability in risk factor levels through metabolomics. This approach reduces the spurious identification in comparison to one-metabolite-at-a-time approaches due to revealing relation among metabolites and identifying confounders at metabolomic level. To the best of our knowledge so far, no one has systematically analyzed metabolomic data and introduced a systems approach to identify pathways from the genome to risk factor/disease via metabolomic networks.

Overview

To understand how components and their interactions give rise to emergent properties of a system, causal networks are employed [14-16]. Causal networks illustrate underlying relationship among components in observational studies and can be identified through application of Mendelian randomization and Bayesian graphical models [10,17]. In this section, we briefly review Mendelian randomization approach and introduce a systems approach for the case study.

Mendelian randomization/instrumental variable approaches: Genetic markers have been employed in multiple studies to prevent the analysis being confounded [18]. This is called Mendelian randomization or instrumental variable (IV) to estimate causal relationships. IV approach as established approach for causal inference utilizes variation in the system that is free of unmeasured confounders. Therefore, relative to regression, matching, and propensity score methods, the IV approach seeks to find a randomized experiment embedded in an observational study and estimates causal effects rather than associations [9,15]. A variable must satisfy three key assumptions discussed in multiple studies to qualify as an IV [9,10]. To review those assumptions briefly, assume we aim to measure causal effect of a component (T) on a component (Y) using IV approach. Therefore, we need to find/generate a genetic variant (G) such that it fulfils the following three assumptions graphically represented in Figure 2: G has a significant association with changes in T represented by an arrow from G to T, Any direct path from G to Y is only through T represented by a missing arrow from G to Y, There is no unmeasured factors that confound G and Y relationship. It is represented as missing arrow from G to U.

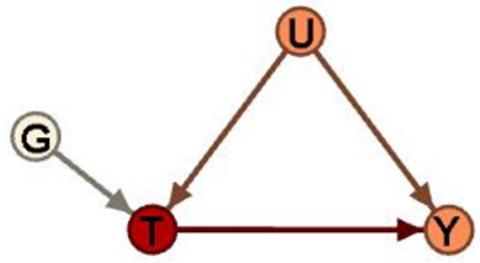

Figure 2: A graphical explanation of Mendelian randomization/instrumental variable assumptions. Note that lack of a link corresponds to lack of relationship.

The IV assumptions are strong and cannot be fully empirically verified. However, attentions need to be paid to hold them. The assumptions are violated through application of weak and invalid IVs. If an IV does not explain sufficient variation of a component of interest (T in Figure 2), it is called a weak IV. Application of weak IVs results in bias and unstable directionality even with large samples. Invalid IVs are those that the effect of IV reaches to outcome (Y) through other paths and not only through the component of interest (T in Figure 2). Application of invalid IVs may happen in different ways, such as pleiotropic effect of a genetic variant, linkage disequilibrium between genetic variants, and genetic interactions. Since the genome includes millions of variants, identifying genetic variants to satisfy IV assumptions is a fundamental question in Mendelian randomization approaches. To hold IV assumptions, the common approach is to find a genetic variant strongly associated with the variable of interest [18,19]. However, not being pragmatic at largescale metabolomic studies, see for example [20], is a major limitation of this approach.

To overcome this challenge and be able to identify a metabolomic causal network in large scales and try to hold the IV assumptions, we extract near-complete information across the genome to create IVs. This IV approach is combined with Bayesian graphical modeling [21] and implemented in a constraint-based algorithm called genome directed acyclic graph (G-DAG) [11]. The G-DAG algorithm has been discussed in multiple publications [8,22,23], for the readers' convenience, the algorithm is provided in details in Supplementary, Section 2.

The G-DAG algorithm extracts information across genome and creates several hundred instrumental variables. Since the information in multiple SNPs/genes are combined in each IV, it is stronger than a single SNP/gene. Below are features of the G-DAG algorithm to hold the IV assumptions: Creating and employing strong instrumental variables through extracting information from multiple variants; Independent instrumental variables toward holding the assumption of

validity; Multiple independent instrumental variables for each metabolite, to make overall IVs even stronger.

Note that the third feature is possible due to generating and applying independent IVs in the G-DAG algorithm. Otherwise, the underlying assumptions would be violated.

A systems approach to multi-omic integration for pathway identification: In a multi-omic approach different molecular levels provide measures from different biological inputs to disease and as a result increase predictive and discriminative ability [24,25]. Here, we introduce a systems approach to integrate data at three biological levels, genomics, metabolomics, and risk factor/disease with the aim of pathway identification, Figure 3. This approach is explained below in four steps.

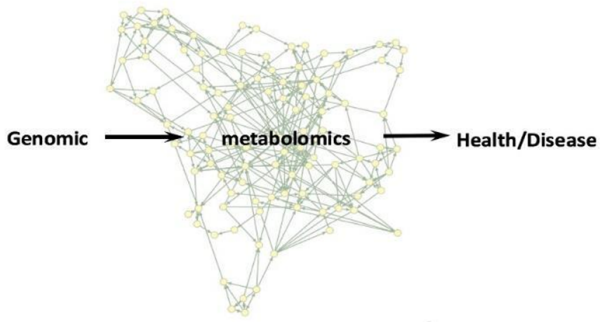

Figure 3: Overall picture of the strategy for identifying pathways from the genome to risk factors or clinical end points via metabolomics. The network is a causal network across metabolomics where nodes stand for metabolites and edges represent the direction of effect identified based on Mendelian principles.

Underlying relationships among metabolites: Instead of analyzing an individual metabolite at a time, we identify a metabolomic causal network (Figure 3), to infer underlying relationships among metabolites. In the network, nodes represent metabolites linked by directed edges. A missing link between two metabolites means no relationship. A link between two metabolites represents the relationship after excluding the effect of other metabolites in the analysis. Directions represent cause and effect relationships and are identified based on Mendelian principles, using variation in the system that is free of confounding.

Further analysis of the metabolomic causal network leads to understanding the principles governing at metabolomics, such as identification of modules, identification of the role of each metabolite, measuring effects of actual or hypothetical manipulations, distinguishing intervention targets from disease predictors, identification of more efficacious intervention targets, and inference of pathways [12,26].

Metabolites with direct effect on risk factors/disease: Association analyses of metabolites with disease endpoints may lead to spurious detections or findings with low impact on the endpoints. Considering underlying relationships among metabolites, we take into account confounders at metabolomics and identify metabolites with direct effect on the endpoints using structural equation modeling [8] and distinguish them from those with indirect effects or spurious effects. Given the metabolites with direct effect on the endpoints, the rest of metabolites in the study do not have a significant effect on the endpoints. Therefore, for the further analyses and interpretations, a focus on the set of metabolites with the direct effect is sufficient.

Gene-metabolite relationship through GWAS: Metabolites can serve as intermediate phenotypes for detecting novel genes with variants of functional effect and bridge gene effects to clinical end points [2,26]. Focusing on genetic variations that exert their function through metabolomic mechanisms prevents reducing the signal to noise through long pathway from genetics to risk factors/disease.

Pathways from the genome to disease via metabolomics: We integrate results of the three aforementioned steps to identify pathways. The path is identified, if significant genes from the genome analysis (step 3) has a significant relationship with one of the metabolites with a direct/indirect path to a risk factor/disease (step 2). These pathways are identified after overcoming confounders at metabolomic level (step 1) and visualizing underlying relationships of this intermediate level. Therefore, they facilitate mechanistic understanding and generate more efficacious hypotheses for clinical experiment.

Case Study

To identify pathways from the genome to risk factor through metabolomics, we focused on plasma triglycerides as a risk factor of cardiovascular disease. For genotype-phenotype relationships, we focused on loss-of-function (LoF) mutations. LoF mutations are defined as sequence changes caused by single nucleotide variants or small insertions and deletions, which are predicted to result in a non-viable transcript or greatly truncated protein product [27]. A typical human exome harbors dozens of LoF variants predicted to severely disrupt or abolish gene function. Regarding metabolites, we focused on African-American individuals to overcome environmental confounders, such as population-to-population and regional dietary variations in the metabolome.

Study sample and data preparation: Genomic data and serum metabolites were available on a subset of the Atherosclerosis Risk in Communities (ARIC) study [28], 2,479 African-American (range in age from 45-64 years) who were randomly sampled from Jackson, Mississippi field center. In addition to metabolites and dense genetic marker data, multiple risk factor phenotypes related to health and chronic diseases including plasma triglycerides were measured.

Metabolic profiling was completed in June 2010 carried out on fasting serum samples stored at -80 degrees centigrade since collection at baseline in 1986–1987. A total of 602 metabolites were detected and semiquantified by Metabolon Inc. (Durham, North Carolina) using an untargeted gas chromatography-mass spectrometry and liquid chromatography-mass spectrometry-based quantification protocol [29]. Metabolites were excluded on the basis of 3 criteria. First, more than 50% of the samples had missing values. Second, they had unknown chemical structures. Third, the metabolites or any transformation of them did not follow normal distribution. After this assessment, a total of 122 named metabolites were included in the study. We carried out some preliminary assessments and found KNN algorithm the best approach for the missing value imputation in our data set. Therefore, the metabolites were imputed by KNN algorithm, which also is identified as the best approach for imputation in some other metabolomics studies, such as [30].

Common single nucleotide polymorphisms (SNPs) were genotyped using the Affymetrix platform (version 6.0) consisting of 1,034,945 common variants spread across the genome. Variation across this data set was extracted and used to identify a metabolomic causal network.

Sequencing data of the protein-coding regions of the genome were also available; the annotations were captured by NimbleGen's VCRome2.1 (Roche NimbleGen), and the captured exons were sequenced using Illumina HiSeq 2000. The Burrows-Wheeler Aligner was used to align sequences to the hg19 reference genome [31]. Allele calling and variant call file construction were performed using the Atlas2 suite [32] (Atlas-SNP and Atlas-Indel). Variants were annotated using ANNOVAR [33] according to the reference genome GRCh37 and National Center for Biotechnology Information Reference Sequence. More details of the study sample and measurements are provided in Supplementary, Section 5.

Identification of the metabolomic network: The identification and analysis of a metabolomic network was carried out using the G-DAG algorithm [11]. The G-DAG algorithm first utilizes hierarchical clustering to measure linkage disequilibrium using square of correlation [34] and determine proxies from SNPs that are nearly perfectly correlated (>0.80) with others. Assuming that the genome inherited variation is a causal factor of metabolomic changes and not the other way around, the G-DAG algorithm extracted information from 1,034,945 SNPs scattered across the genome to generate multiple IVs. Around 80% of variation in the genome was driven by 788 IVs. Then, the G-DAG algorithm, which is a constraint-based algorithm, found 353 IVs significantly correlated with 122 metabolites. These IVs provided possibility to identify directions (cause and effect relationships) over metabolomic topology (undirected network) based on Mendelian principles, Figure 4. More details of the G-DAG algorithm are provided in Supplementary, Section 2. The analysis was carried out at statistical significance level 0.001 determined by structural Hamming distance [35], a well-established assessment for the quality of fit in networks, e.g. see [36,37].

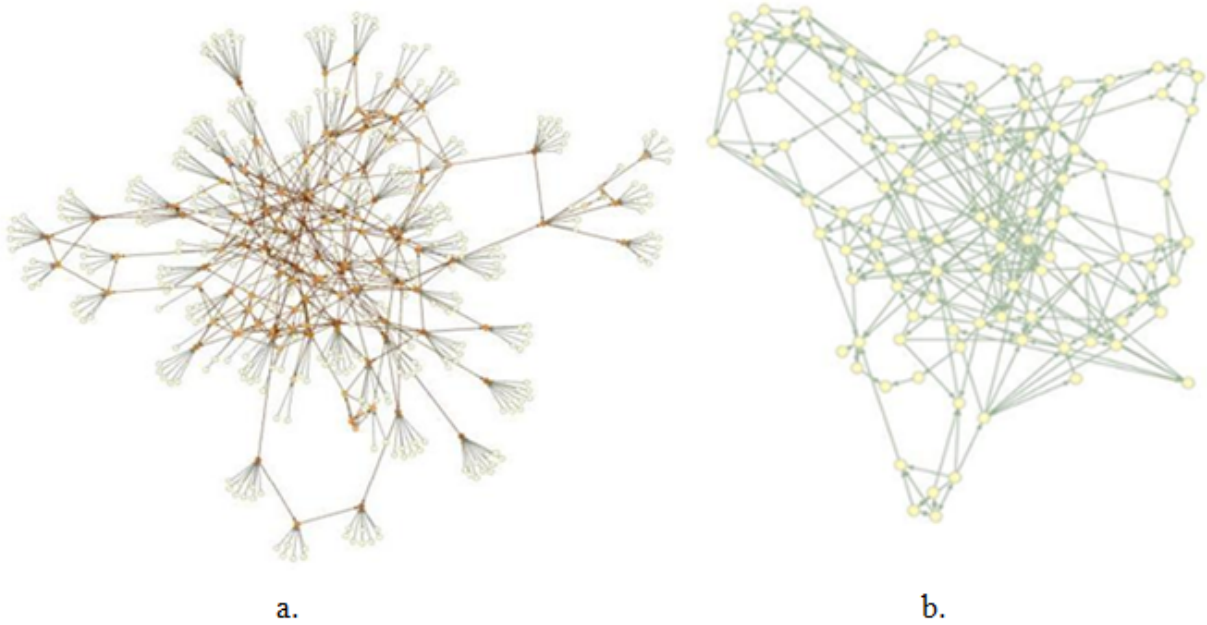

Figure 4: a. Identified metabolomic causal network using the G-DAG algorithm established in Mendelian randomization and Bayesian network modeling. Pale nodes represent genome IVs which explained even up to 96% variation in metabolites. Orange nodes represent metabolites. b. The metabolite relationships from Figure 4a without depicting the IVs. Each link represents relationship between two corresponding metabolites when effects from other metabolites are excluded. Directions are identified based on Mendelian principles using genome IVs.

Note that the genome IVs were employed as a tool to identify the metabolomic causal network. Therefore, after building the network, we removed the IVs from the network to focus on the metabolomic relationships, Figure 4b. In the metabolomic network, directions identified robustly using Mendelian principles and represent cause and effect relationships.

Metabolite-triglyceride pathways: An extension of the G-DAG algorithm [8] was conducted to identify metabolites with direct effects on triglyceride levels and distinguish them from those with indirect effect, Figure 5. Using the underlying relationship between metabolites and triglycerides, we employed structural equation modeling and measured the causal effects, for details on the model see Supplementary, section 4. Nine metabolites out of 122 metabolites under study were identified with direct effect on triglycerides at significance level 0.001; the estimated causal effects are presented in Table 1. The effect of the other metabolites to triglyceride levels is through the set of metabolites with direct effect on triglycerides. For the analysis, log transformation of triglyceride levels was adjusted for covariates including age and principle components for population stratification applying a linear regression. The analysis after including body mass index (BMI) in the set of covariates did not show significant effect of glutamate, glycine, and deoxycarnitine on triglyceride levels, which are noted with superscript "b" in Table 1.

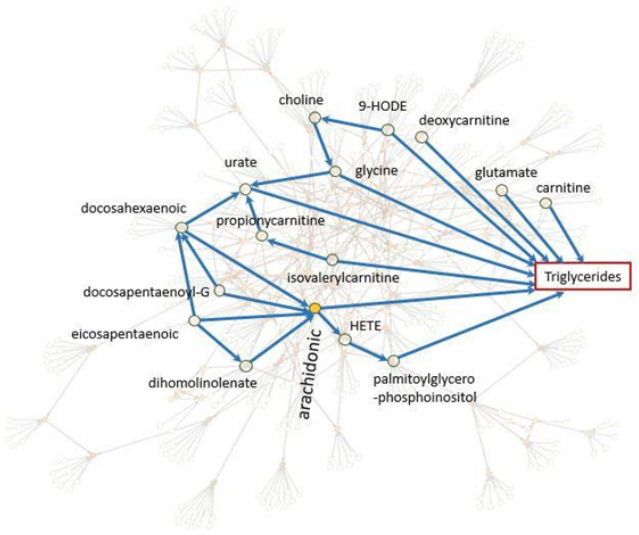

Figure 5: Relationship among nine metabolites with direct effect on triglyceride levels. Arachidonic acid has the largest effect on triglycerides. No feed-back loop (i.e. cycles) was identified between these nine metabolites and triglyceride levels. The relationships among metabolites are from the genomic-metabolomic network depicted in the background to emphasize that the directions are identifies based on Mendelian principles.

The systems approach applied here not only reduces the false discovery of identification but also visualizes the underlying relationships among metabolites and leads to understand mechanistic of metabolite-triglyceride relationship [12]. For instance, it facilitates distinguishing between direct and indirect metabolite-triglyceride pathways. While association study found 21 metabolites out of the 122 metabolites with significant effect on triglyceride levels, applying the systems approach identified that only 6 of them (after adjustment for BMI) have direct effect on triglyceride levels. Therefore, the systems approach applied here provides efficacious targets for intervention compared to association studies. (For the results of the association study, see Supplementary Tables S1-S3). Figure 5 shows that four metabolites eicosapentaenoic acid (EPA), docosapentaenoyl-glycerophosphocholine (DPA-G), docosahexaenoic acid (DHA), and dihomolinolenate influence levels of arachidonic acid which has a direct effect on triglycerides. Arachidonic acid has a positive and the largest effect on triglyceride levels, see Table 1. The association study showed significant relationship between these metabolites with triglycerides due to their relationship with arachidonic acid. The largest and positive effect of arachidonic acid on triglyceride levels among the other associated metabolites is already validated clinically [38].

More information that can be extracted from pathway visualization is causal parameters [12]. In Table 1, in addition to p-value and effect size of metabolites with direct effect on triglyceride levels, three causal network parameters Out-degree, In-degree, and Strength are presented to investigate potential roles of the metabolites at metabolomics. In-degree represents the number of metabolites that influence a particular metabolite. At the metabolomic causal network the In-degree parameter has a range from 0 to 9. Metabolites with higher In-degree capture features of higher number of metabolites in the metabolomic network.

Metabolite	Out-degree	In-degree	strength	Pathway	P-value	Effect Sizes (SE)
Arachidonic	1	4	50	Lipid	2.3e-17	0.17 (0.03)
Carnitine	1	1	7	Lipid	1.4e -11	0.15 (0.04)
9-HODE	1	1	50	Lipid	1.4e -7	0.12 (0.03)
Palmitoylglycero-phosphoinositol	1	5	8	Lipid	1.6e -6	0.1 (0.01)
Urate	0	5	11	Nucleotide	2.2e -5	0.09 (0.01)
Isovalerylcarnitine	2	0	20	Amino acid	2.0e -4	0.09 (0.02)
Glycineb	4	2	20	Amino acid	4.0e -3	-0.09 (0.02)
Deoxycarnitine b	0	6	11	Lipid	1.0e -3	-0.08 (0.03)
Glutamate b	0	0	0	Amino acid	1.0e -3	-0.07 (0.02)

Table 1: The causal network parameters and the effect sizes of nine metabolites with direct effect on plasma triglycerides. a. Effect sizes measured in standard deviation units to facilitate comparison, b. Metabolite with no significant effect at level 0.0001 after adjustment for BMI

Out-degree represents the number of metabolites influenced by a particular metabolite. At the metabolomic causal network the Out-degree parameter has a range from 0 to 8 at the metabolomic causal network. Metabolites with higher Out-degree influence higher number of metabolite in the network.

Strength represents how strong a metabolite is connected to the network, ranges from 0 to 50 at the metabolomic causal network. The Strength zero (the lowest strength) means the metabolite is not connected to the network.

In comparison to the range of Out-degree in the metabolomic network (0 to 8), the study reveals that the metabolites with direct effect on triglycerides have very low Out-degree, which is corresponding to their low influence in metabolomic network.

Pathways from LoF mutations to metabolites through GWAS: LoF variants included in this study were defined as premature stop codons occurring in the exon, essential splice site disrupting, and indels predicted to disrupt the downstream reading frame. A total of 7038 genes harboring 12,522 LoF variants were identified. Single-variant tests for the single variants and gene-based burden tests for the genes were utilized to investigate the relationship with individual metabolites [39]. Seven genes with significant effect (P_val < 1.3×10^{-7}) on metabolites in our analysis are presented in Table 2. Potential roles of these metabolites in the metabolomic system are investigated through causal network parameters.

The metabolites influenced by LoF mutations tabulated in Table 2 do not have important roles at the metabolomic network due to low number of Out-degrees and relatively higher number of In-degree. Later, we see both metabolites urate and deoxycarnitine are in direct paths to triglycerides, Figures 6 and 7. Interestingly, both of them have a high In-degree and zero Out-degree. We conclude the metabolites influenced by LoF mutations are mostly influenced by metabolomic system rather than influence the system. This may be interpreted as below: metabolites influences by LoF mutations cannot be so critical at metabolomics. Otherwise they will lead to debilitating disease or be inconsistent with life. This point can be considered and discussed for functional understanding although it may require further assessments.

Gene and Variant	Metabolite	Out-degree	In-degree	Strength	Pathway	P-value
LRRC69 8:92213022:T:A & 8:92212839:A:G	Deoxycarnitine	0	6	11	Lipid	9.00E-16
LRRC46 17:45913719:TG:T&17.45911791:GT:G	Urate	0	5	11	Nucleotide	le-7
FAM198B 4:159091864:G:A & 4:159091422:T:A	HETE	2	5	17	Lipid	5e-9
CD36 7:80300449:T:G	Oetanoylcamitine	1	1	50	Lipid	4e-8
PCSK9 1:55529215:C:A & 1:55512222:C:G 1:55524293:TG:T	Cholesterol	0	3	29	Lipid	5e-9
TEX15 8:30700833:TTC:T & 8:30694729:C:CA 8:30701535:C:A &8:30705099:CTCTA:C 8:30702839:CMCA:Cc	Mannose	I	1	14	Carbohydrate	9e-9
C6orf25 6:31692558:C:T	Methionine sulfoxide	3	1	13	Amino acid	3e-8

Table 2: Genes harboring LoF mutation with significant effect on individual metabolites under study. The causal network parameters are measured from the metabolomic causal network. Variant includes chromosome:position:reference allele:alternative allele.

Identified genome-metabolite-triglyceride pathways: Through combining the results of the three steps above, we identify pathways from the genome to plasma triglycerides via metabolomic network. By inspecting the metabolites with direct effect on triglyceride levels and those influenced by LoF mutations, we identified two direct pathways linking the genome to plasma triglycerides: One path from LRRC46 through metabolite urate (LRRC46 UrateTriglycerides), and another path from LRRC69 through metabolite deoxycarnitine (LRRC69 Deoxycarnitine Triglycerides). These pathways are visualized in Figure 6.

In Figure 7, we see the connectivity of the metabolites urate and deoxycarnitine at the metabolomic network. Blue pathways depict metabolites that influence urate and deoxycarnitine and the effect of these individual metabolites on triglyceride levels is not significant conditionally. In the left panel, propionylcarnitine and isovalerate are lipids involved in fatty acid metabolism; docosahexaenoate is an essential fatty acid; glycine is an amino acid involved in glycine, serine, and threonine metabolism; and finally gamma-glutamylthreonine is a peptide in gamma-glutamyl metabolism. In the right panel of Figure 7, 3-carboxy-4-methyl-5-propyl-2-furanpropanoate, azelate, and laurate are fatty acids; citrate is a component of the tricarboxylic acid cycle that is central to energy metabolism; and trans-4-hydroxyproline is a modified amino acid associated with the urea cycle and is thought to be associated with oxidative stress.

In addition to the above direct pathways from the genome to triglycerides via metabolites urate and deoxycarnitine, another pathway is identified from FAM198B and C6orf25 to metabolite HETE with indirect effect on triglyceride levels. These pathways are through palmitoylglycerophosphoinositol (Figure 8).

Figure 8 shows that the identified path to triglycerides is not through arachidonic acid. Rather, arachidonic acid influences this path through HETE. These results are identified using the metabolomic causal network. The identified pathways through the systems approach introduced here reduce false discovery and may facilitate understanding underlying mechanism and generate hypotheses for further experimental studies. Through a standard GWAS analysis, we investigated genome-triglyceride relationship and no variant with significant effect on triglyceride levels was identified. The results are provided in Supplementary Table S2.

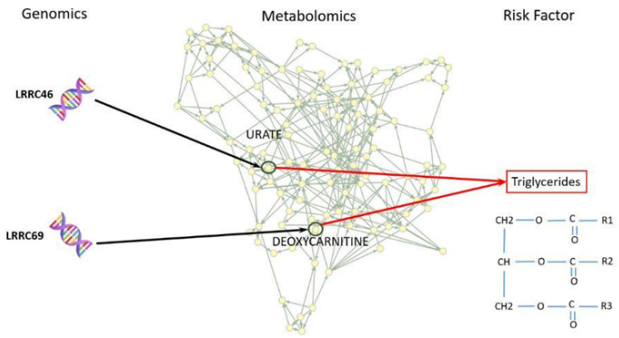

Figure 6: Two identified direct pathways from LRRC46 and LRRC96 to plasma triglycerides via metabolites urate and deoxycarnitine respectively.

Discussion

The genetics of complex diseases, such as cardiovascular disease, produce alterations in the molecular interactions of intermediate phenotypes, such as metabolites, which contribute to early disease-related changes [40,41]. The collective effect in cellular pathways may become clear through integrated approaches to identify underlying mechanisms. Powerful and advanced analytic strategies are required to integrate largescale data in systems biology for the elucidation of pathways across human biological levels. We introduced a systems approach for pathway identification by integrating data at three different biological levels using causal networks. Causal networks compatible with structural equation modeling improve the power of discovery by reducing the influence of phenotype-phenotype associations to illustrate underlying relationships [8,42].

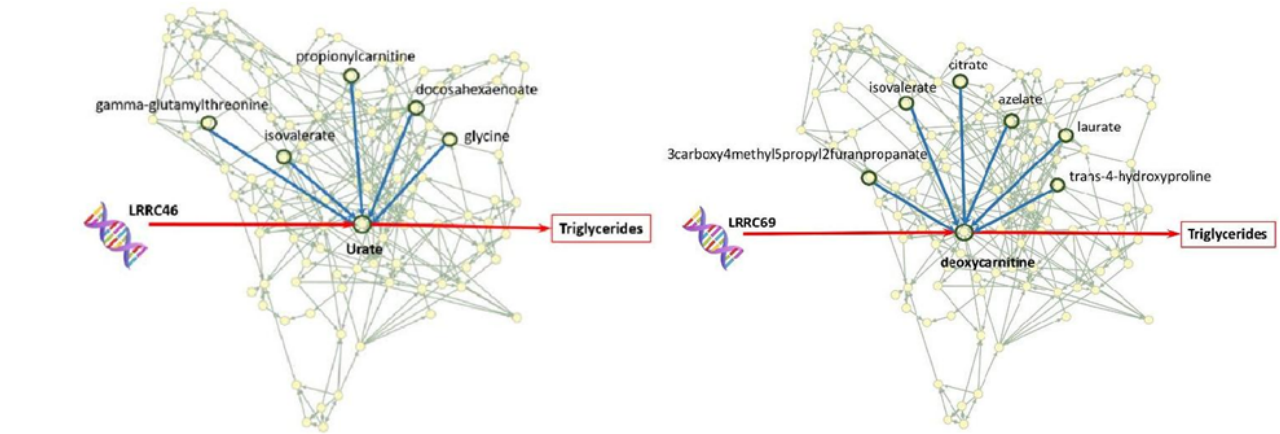

Figure 7: Red pathways: Direct pathways from the genome to triglycerides via metabolites. Blue pathways: Metabolites that influence urate and deoxycarnitine at metabolomic network.

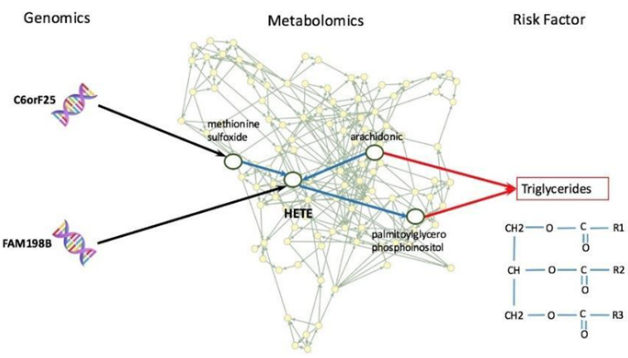

Figure 8: Indirect pathways from FAM198B and C6orf25 to triglycerides through metabolite HETE and palmitoylglycerophosphoinositol.

The introduced approach is toward mechanistic understanding through pathway identification linking the genome to health/disease. The pathways reveals the underlying relationships behind observations [21,43,44], which do not play a significant role in more traditional correlative analyses. We first construct a causal network over metabolomics using instrumental variables/Mendelian randomization [8]. Through a metabolomic causal network, we not only account for association between metabolites but also confounders at metabolomics. Second, we take an improvement in understanding the role of metabolites in quantitative risk factors, here triglycerides [8]. This step filters the number of metabolites to a subset that impact triglyceride levels. In addition, the visualization of underlying relationships reveals even more information.

Through integration of the metabolomic causal network and genome association study, two pathways were identified, from the genome—leucine rich repeat containing 46, LRRC46, and leucine rich repeat containing 69, LRRC69—linked to plasma triglycerides via the metabolites urate and deoxycarnitine, respectively. Those metabolite-triglyceride connections are consistent with known biochemical

information. For example, urate is biochemically linked to the thioredoxin system, which mediates cellular redox balance and has been associated with triglyceride levels [45]. Similarly, deoxycarnitine lies at a biochemical crossroad between triglyceride metabolism and fatty acid oxidation, and carnitine metabolism has been implicated in the regulation of triglycerides [46]. Further, carnitine is essential for β-oxidation of long-chain fatty acids, and metabolic enzymes involved in carnitine biosynthesis mediate a decrease in fatty acid oxidation and increase in glycolysis in heart failure progression [47].

The genetic components LRRC46 and LRRC69 provide new mechanistic insights into the regulation of triglyceride levels. Experimental validation will need to be conducted to assess the contributions of LRRC46 and LRRC69 to the modulation of triglyceride levels, but our new approach allows such validation experiments to be focused. For instance, instead of measuring only triglyceride levels as an endpoint in a model biological system, we can additionally measure levels of uric acid, deoxycarnitine, and other metabolites known to be associated with each of them, such as thioredoxins. For example, since thioredoxins have been found to mediate cardioprotection, it may be possible to implicate LRRC46 in cardioprotection upstream of thioredoxins regulation.

The leucine-rich repeat (LRR) structural motif is characterized by the α/β horseshoe fold, composed of 20-30 hydrophobic amino acid stretches of leucine. LRRs mediate protein-ligand interactions and in the case of cascade interaction model in fatty-acid-uremic toxin-drug system, in which long-chain fatty acids concentrations are increased, cascade displacement of bound drugs occurs by a competitive inhibitor, such as CMPF and uremic toxins containing an indole ring [48]. Previous studies reveal a relationship between leucine-rich repeat (LRR) and cardiovascular disease and triglycerides [49,50]. The pathways identified here strengthen those associations and provide a plausible mechanism for them.

Conclusion

The approach presented here provides a step toward addressing challenges in modern biomedical research, such as large scale data sets, highly correlated phenotypes, and integrating information at different biological levels. Additionally, future method development will include genome variation effects on multiple metabolites, hypothesized pleiotropy in GWAS/WGS, to provide further insights into the mechanistic underpinnings of chronic diseases.

Acknowledgment

Thanks go to Drs. Eric Boerwinkle and Christie Ballantyne for valuable comments which clarified the message of this work. Thanks also go to the staff and participants of the Atherosclerosis Risk in Communities (ARIC) Study for their important contributions.

Conflict of Interest

There is no conflict of interest among authors.

References

1. Reitz C, Brayne C, Mayeux R (2011) Epidemiology of Alzheimer disease. Nature Reviews Neurology.
2. Suhre K (2012) Genetics meets metabolomics: From experiment to systems biology. Genetics Meets Metabolomics: From Experiment to Systems Biology (Vol. 9781461416).
3. Shah SH, Newgard CB (2015) Integrated Metabolomics and Genomics: Systems Approaches to Biomarkers and Mechanisms of Cardiovascular Disease. Circulation: Cardiovascular Genetics 8: 410- 419.
4. Lewis GD, Wei R, Liu E, Yang E, Shi X, et al. (2008) Metabolite profiling of blood from individuals undergoing planned myocardial infarction reveals early markers of myocardial injury. J Clin Invest 118: 3503-3512.
5. Blasco H, Nadal-Desbarats L, Pradat PF, Gordon PH, Madji Hounoum B, et al. (2016) Biomarkers in amyotrophic lateral sclerosis: Combining metabolomic and clinical parameters to define disease progression. Eur J Neurol 23: 346-353.
6. Nicholson JK, Wilson ID (2003) Opinion: Understanding "Global" Systems Biology: Metabonomics and the Continuum of Metabolism. Nat Rev Drug Discov 2: 668-676.
7. Yazdani A, Yazdani A, Boerwinkle E (2015) Rare variants analysis using penalization methods for whole genome sequence data. BMC Bioinformatics 16: 405.
8. Yazdani A, Yazdani A, Samiei A, Boerwinkle E (2016a) Erratum to: A causal network analysis in an observational study identifies metabolomics pathways influencing plasma triglyceride levels. J Biomed Inform 63: 337-343.
9. Dawid AP (2007) Fundamentals of statistical causality. RSS/EPSRC Graduate Training Program 279: 1-94.
10. Burgess S, Thompson SG (2013) Use of allele scores as instrumental variables for Mendelian randomization. Int J Epidemiol 42: 1134-1144.
11. Yazdani A, Samiei A, Boerwinkle E (2016b) Generating a robust statistical causal structure over 13 cardiovascular disease risk factors using genomics data. J Biomed Inform 60: 114-119.
12. Yazdani A, Samiei A, Boerwinkle E (2016c) Identification, analysis, and interpretation of a human serum metabolomics causal network in an observational study. J Biomed Inform 63: 337-343.
13. Yazdani A, Yazdani A, Boerwinkle E (2016) Conceptual aspects of causal networks in an applied context. J Data Mining Genomics Proteomics 7: 602-2153.
14. Pearl J (2010) The International Journal of Biostatistics An Introduction to Causal Inference An Introduction to Causal Inference. Int J Biostat 6: 7.
15. Berzuini C, Dawid P, Bernardinell L, VanderWeele TJ, Hernán MA (2012) Causality: Statistical Perspectives and Applications. Causality: Statistical Perspectives and Applications.
16. Yazdani A, Yazdani A, Boerwinkle E (2016) A Causal Network Analysis of the Fatty Acid Metabolome in African-Americans Reveals a Critical Role for Palmitoleate and Margarate. OMICS: A J Integrative Biology 20: 480-484.
17. Baiocchi M, Cheng J, Small DS (2014) Instrumental variable methods for causal inference. Stat Med 33: 2297-2340.
18. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Smith GD (2008) Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. Stat Med 27: 1133-1163.
19. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet, 37: 710-717.
20. Inouye M, Kettunen J, Soininen P, Silander, K, Ripatti S, et al. (2010) Metabonomic, transcriptomic, and genomic variation of a population cohort. Molecular Systems Biology 6.
21. Pearl J (2000) Causality. New York: Cambridge.
22. Ainsworth HF, Shin SY, Cordell HJ (2017) A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements. Genet Epidemiol 41: 577-586.
23. Aksam VK, Chandrasekaran VM, Pandurangan S (2017) Identification of cluster of proteins in the network of MAPK pathways as cancer drug targets. Informatics in Medicine Unlocked 9: 86-92.
24. Chen Q, Park HC, Goligorsky MS, Chander P, Fischer SM, et al. (2012) Untargeted plasma metabolite profiling reveals the broad systemic consequences of xanthine oxidoreductase inactivation in mice. PLoS ONE 7.

25. Then C, Wahl S, Kirchhofer A, Grallert H, Krug S, et al. (2013) Plasma Metabolomics Reveal Alterations of Sphingo- and Glycerophospholipid Levels in Non-Diabetic Carriers of the Transcription Factor 7-Like 2 Polymorphism rs7903146. PLoS ONE 8.
26. Yazdani A, Yazdani A, Liu X, Boerwinkle E (2016) Identification of Rare Variants in Metabolites of the Carnitine Pathway by Whole Genome Sequencing Analysis. Genet Epidemiol 40: 486-491.
27. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science 335: 823-828.
28. The ARIC Investigators (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. American Journal of Epidemiology, 129: 687-702.
29. Perkel J (2011) Metabolomics: where seeing is believing. BioTechniques.
30. Hrydziuszko O, Viant MR (2012) Missing values in mass spectrometry based metabolomics: An undervalued step in the data processing pipeline. Metabolomics 8: 161-174.
31. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
32. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, et al. (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13.
33. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38.
34. Yazdani A, Dunson DB (2015) A hybrid Bayesian approach for genome-wide association studies on related individuals. Bioinformatics 31: 3890-3896.
35. Tsamardinos I, Brown LE, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65: 31-78.
36. Norouzi M, Fleet DJDDJ, Salakhutdinov R, Blei DM (2012). Hamming distance metric learning. Advances in Neural Information Processing Systems, 1-9.
37. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. Proc Biol Sci 255: 279-284.
38. Maki KC, Dicklin MR, Schild AL, Kling D, Davidson MH (2014) Plasma Fatty Acids as Predictors of Triglyceride and Non-HDL Cholesterol Responses to Omega-3 Free Fatty Acid Therapy in Hypertriglyceridemia. J Clin Lipidol 8: 341-342.
39. Yu B, Li AH, Metcalf GA, Muzny DM, Morrison AC, et al. (2016) Loss-of-function variants influence the human serum metabolome. Science Advances 2: e1600800-e1600800.
40. Hatano T, Saiki S, Okuzumi A, Mohney RP, Hattori N (2016) Identification of novel biomarkers for Parkinson's disease by Metabolomic technologies. J Neurol Neurosurg Psychiatry 87: 295-301.
41. Blasco H, Patin F, Madji Hounoum B, Gordon PH, Vourc'h P, et al. (2016) Metabolomics in amyotrophic lateral sclerosis: How far can it take us? Eur J Neurol.
42. Gowda GAN, Zhang S, Gu H, Asiago V, Shanaiah N, et al. (2008) Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn 8: 617-633.
43. Rubin DB (2005) Causal Inference Using Potential Outcomes. J Am Stat Assoc 100: 322-331.
44. Yazdani A, Boerwinkle E (2014) Causal inference at the population level. Int J Res Med Sci 2: 1368.
45. Van Greevenbroek Mm, Vermeulen Vm, Feskens Ej, Evelo Ct, Kruijshoop M, et al. (2007) Genetic Variation In Thioredoxin Interacting Protein (Txnip) Is Associated With Hypertriglyceridaemia And Blood Pressure In Diabetes Mellitus. Diabet Med 24: 498-504.
46. Stefanovic-Racic M, Perdomo G, Mantell BS, Sipula IJ, Brown NF, et al. (2008) A moderate increase in carnitine palmitoyltransferase 1a activity is sufficient to substantially reduce hepatic triglyceride levels. Am J Physiol Endocrinol Metab 294: E969-E977.
47. West JA, Beqqali A, Ament Z, Elliott P, Pinto YM, et al. (2016) A targeted metabolomics assay for cardiac metabolism and demonstration using a mouse model of dilated cardiomyopathy. Metabolomics 12.
48. Takamura N, Maruyama T, Otagiri M (1997) Effects of urernic toxins and fatty acids on serum protein binding of furosemide: Possible mechanism of the binding defect in uremia. Clinical Chem 43: 2274-2280.
49. Hultgårdh-Nilsson A, Borén J, Chakravarti S (2015) The small leucine-rich repeat proteoglycans in tissue repair and atherosclerosis. J Intern Med.
50. Will RD, Eden M, Just S, Hansen A, Eder A, et al. (2010) Myomasp/LRRC39, a heart-and muscle-specific protein, is a novel component of the sarcomeric m-band and is involved in stretch sensing. Circ Res 107: 1253-1264.