HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 September 22.

Published in final edited form as:

Genet Epidemiol. 2016 September ; 40(6): 486–491. doi:10.1002/gepi.21980.

Identification of Rare Variants in Metabolites of the Carnitine Pathway by whole genome sequencing analysis

Akram Yazdani[1], **Azam Yazdani**[1], **Xiaoming Liu**[1], and **Eric Boerwinkle**[1,2]

[1]Human Genetics Center, The University of Texas Health Science Center at Houston, USA

[2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

Abstract

We use whole genome sequence data and rare variant analysis methods to investigate a subset of the human serum metabolome, including 16 carnitine-related metabolites that are important components of mammalian energy metabolism. Medium pass sequence data consisting of 12,820,347 rare variants and serum metabolomics data were available on 1,456 individuals. By applying a penalization method, we identified two genes *FGF8* and *MDGA2* with significant effects on *lysine* and *cis-4-decenoylcarnitine*, respectively using Delta-AIC and LRT statistics. Single variant analyses in these regions did not identify a single low frequency variant (MAC>3) responsible for the underlying signal. The results demonstrate the utility of whole genome sequence and innovative analyses for identifying candidate regions influencing complex phenotypes.

Keywords

linkage disequilibrium; carnitine; metabolomics; penalization; rare variants

Introduction

The metabolome is a collection of small molecules resulting from a multitude of cellular and physiologic processes (Pearson, 2007). Measurement and analysis of the metabolome in well characterized sample sets can provide insight into the underlying mechanisms of genomic and environmental actions and interactions on metabolism. In addition, investigations of the metabolome are likely to prove more productive for novel gene and locus discovery because the metabolome is functionally closer to the DNA level, where effect sizes will be larger than similar analyses of distant risk factor levels or disease endpoints (Yu et al., 2014). Unlike proteomics and genomics, which have enjoyed considerable concentrated efforts, the metabolome is relatively under-studied, partly because of the difficulty of measuring and annotating such a large number of compounds, and the identity of many of the analytes in the metabolome remain unknown. Indeed, genetic studies of the metabolome can lead to

Address correspondence to: Eric Boerwinkle, PhD, Human Genetics Center, 1200 Pressler St., Suite E-447, Houston, TX 77030, 713-500-9816, Eric.Boerwinkle@uth.tmc.edu.

Competing interests

The authors declare that they have no competing interests.

insight into the identity of these unknown compounds (Zheng et al., 2013). Multiple genomic studies have been done to identify common variants influencing subsets of the metabolome in plasma (Yu et al., 2014; Rhee et al. 2013) and urine (Suhre et al., 2011), and such studies have been valuable tools in agriculture research (Wen et al., 2014). GWAS studies, however, capture the influence of common variants and exome sequencing is limited to the small protein-encoding portion of the genome (Demirkan et al., 2015).

Whole genome sequencing in large sample sizes is now a reality (Morrison et al., 2013). Analysis of whole genome sequence offers the advantage of ascertaining genome variation outside of the protein-encoding region and the ability to investigate the role of rare variants on phenotypic variation. Studies using whole genome sequence to analyze phenotypes also encounter a number of challenges that must be overcome. First, there are a very large number of rare variants and many of these variants are limited to a single individual. Therefore, usual single site methods comparing mean values or frequencies do not have sufficient statistical power and are often not appropriate or feasible (Lee et al., 2014, Yazdani et al., 2015). Second, outside of genes and a few well-characterized regulatory elements, the role of much of the genome in normal development, metabolism and physiology remains unknown. Despite impressive advances from studies such as ENCODE (Consortium et al., 2011), predicting the consequences of sequence variation outside of a protein encoding gene is difficult.

We use whole genome sequence data and rare variant analysis methods introduced by Yazdani et al. (2015) called CCRS to investigate a model subset of the human serum metabolome, specifically the carnitine-related metabolites. Carnitine is synthesized in the liver and kidney and is required for the transport of fatty acids within mitochondria for energy production (Steiber et al. 2004). A major source of carnitine in the body, however, is dietary, with red meat having the highest levels (Flanagan et al. 2010). There are number of genetic forms of carnitine deficiency that influence the cycling of carnitine or acylcarnitine from the cytosol to the mitochondrial matrix (Olpin, 2004). Primary carnitine deficiency is a rare autosomal recessive disorder of fatty acid transport caused by mutations in SLC22A5 (Li et al. 2010). However, the complete picture of the genetic architecture of carnitine is unclear. In this study, we use whole genome sequence data containing 12,820,347 rare variants (i.e. minor allele frequency, MAF, less than 5%) to analyze 16 metabolites involved in carnitine synthesis and metabolism measured on 1,456 European-Americans from the Atherosclerosis Risk in Communities (ARIC) study. For this analysis, we implemented the CCRS rare variant analysis and used sliding windows across the genome in order to identify two loci spanning the *FGF8* and *MDGA2* genes associated with *lysine* and cis-4-decenoylcarnitine levels, respectively.

Study Sample

Metabolomic and genomic data were available on a subset of the Atherosclerosis Risk in Communities (ARIC) study, a biracial longitudinal cohort study of 15,792 middle-aged individuals who were randomly sampled from four US communities and have been measured for multiple risk factor phenotypes related to health and chronic disease. A detailed description of the ARIC study design and methods has been published elsewhere

(The ARIC investigator, 1989). The data presented here includes 1,456 European-American individuals having metabolomic and whole genome sequence data. Written informed consent was obtained from each study participant, including that for broad data sharing. The dbGAP accession number for ARIC study data, including genomic data, is phs000668.v1.p1.

Laboratory Methods

Metabolic profiling was carried out on fasting serum samples from the baseline examination and stored at −80oC. Metabolites were measured using untargeted gas chromatography-mass spectrometry and liquid chromatography-mass spectrometry (GC-MS and LC-MS)-based quantification protocols. Details of these procedures have been previously published (Zheng et al., 2014). The analyses presented here focus on 16 metabolites in the pathway of carnitine synthesis and metabolism, which were near normally distributed or could be transformed to such and had a missing data rate less than 25%.

Missing data were imputed using the k nearest neighbor algorithm of (Verboven et al., 2007). This algorithm starts from a complete subset of the data set X_c and sequentially estimates the missing values for an incomplete observation, x^*, by minimizing the determinant of the covariance of the augmented data matrix $X^* = [X_c; x']$. Then the observation x^* was added to the complete data matrix and the algorithm continues with the next observation.

Details of the sequencing methods are in the supplementary materials of Morrison (2013). Briefly, automated Illumina PE libraries were barcoded with Illumina's multiplex adapters and pooled for sequencing in sets of three samples to generate an average of 6-fold sequence coverage per sample. Methods for WGS sequencing followed standard Illumina PairEnd library protocols with minor modifications. Variants were called using SNPTools (http://www.hgsc.bcm.tmc.edu/cascade-tech-software_snp_tools-ti.hgsc), which consists of variant discovery, genotype likelihood estimation, and genotype inference via imputation. Quality control steps included: low coverage, low variant quality, high degree of relatedness, and evidence of hidden population substructure excluding variants and individuals. To identify population structure, we calculated principal components, PCs, over individuals using common variants and selected first ten PCs. We then adjusted metabolites using those PCs in addition to age and sex as covariates by fitting a linear regression and using its residuals as outcome for the main analysis.

Statistical Analysis

To address the multitude of analytic challenges confronted when analyzing rare variants across the genome, we applied a new statistical approach based on penalization methods by Yazdani et al. (2015). This approach follows a two-stage analysis. In the first stage, it applies principal component analysis with convex penalization (Witten et al., 2009) to avoid over fitting due to linkage disequilibrium (LD) among variants (Talluri et al., 2013; Feng et al., 2014) in a multiple regression model. In the second stage, it uses concave penalization in a multiple regression model to obtain sparse estimates of the coefficient for the principal

components, in the sense that some of the coefficients may be explicitly shrunk to zero (Frank et al., 1993; Fu, 1998). Although this method is a multi-stage analysis, it does not increase the false discovery rate since the first stage is an unsupervised analysis approach. Instead, it avoids missing important region through multiple hypotheses testing steps by incorporate local LD structure, and maintains power while keeping the false discovery rate low by considering sparsity in the data. This method is called CCRS due to its convex and concave penalization for rare variants analysis.

We implemented the CCRS method in sliding windows across the genome. Each window included 50 variants with a step size of 25 variants to the next window. We selected the most promising window associated with the metabolites using Δ-AIC (Burnham and Anderson, 2002) and Likelihood Ratio test (LRT). The Δ-AIC is a model selection criterion that comparers the Akaike's (1973) information criterion (AIC) of all the models with model with min AIC, i.e. $\Delta\text{-AIC}_i = \text{AIC}_i - \min(\text{AIC})$. $\Delta\text{-AIC}_i < 2$ suggests substantial evidence for the *i*th model, values between 3 and 7 indicate that the model has considerably less support, whereas an $\text{AIC}_i > 10$ indicates that the model is very unlikely (Burnham and Anderson, 2002).

We also calculated a likelihood ratio test (LRT) statistic to test the null hypothesis that considers no genotype effect within a window. The LRT approximately follows a chi-square distribution with degrees of freedom equal to a complex function of the design matrix of the model and the penalization (Hastie et al., 2005). A permutation test was carried out to estimate the empirical distribution of *p*-values under the null hypothesis and to calculate a quantile threshold defining statistical significance for each trait (Box 1).

Result

There were 16 analyses in the carnitine synthesis and metabolism pathways that were represented in this metabolomics dataset. Their names, averages and other descriptive statistics for the study sample are shown in Table I and Table II. By design, the sample consisted of only European-Americans. Consistent with enrollment and participation, there were more females than males in the study sample, and the average BMI was 27.3. In general, the individuals included in this metabolomics substudy were similar to participants in the entire ARIC study.

The QQ plots for all 16 metabolites are shown in the online Supplement. Two genomic regions in the analysis of *lysine* and *cis-4-decenoylcarnitine* yielded minimum AIC such that $\Delta\text{-AIC}_i$ for all other models are greater than 4.2 and 2.1, respectively, in addition to having very small *p*-values relative to other calculated *p*-values using LRT. These regions advanced to the permutation algorithm described in Box 1 to obtain a significance threshold for these two metabolites. By permuting each metabolite n=100,000 times and calculating the LRT statistic and corresponding *p*-value in m=1000 randomly selected windows, we defined the significance threshold to be 10^{-8} and 10^{-6} at the level of type I error equal to α=0.001 for *lysine* and c*is-4-decenoylcarnitine*, respectively. Using these two thresholds and $\Delta\text{-AIC}_i$, we identified two significant regions listed in Table III. Among the sites having a MAF<0.05 included in this analysis, the median minor allele frequency within the significant windows

was 0.0034 and 0.0014 for *lysine* and c*is-4-decenoylcarnitine*, respectively. To observe the influence of single variants in significant regions on the traits of interest, we used simple regression to estimate the effects of each variant with minor allele count, MAC, greater than 3. The single variants having a p-value < 0.05 are shown in Table IV.

Figure 1 and Figure 2 show regional plots of 201 windows centered on the significant regions for lysine and c*is-4-decenoylcarnitin*, respectively. The differences between the very small *p*-value in the selected regions and other surrounding regions can be readily observed. There appears to be a background null distribution of *p*-values > 10^{-4}, and the two regions of interest rise notably above this background. To help fine map the possible signal in the region significantly associated with *lysine* included two genes, we serially excluded variants in *FGF8* or *NPM3* and computed the *p*-values using CCRS for each subset.

After excluding 11 variants in *FGF8*, the signal for the metabolite *lysine* is attenuated; the *p*-value of the model is increased from 1.75e-10 to 2.63e-06. In contrast, the signal was not changed appreciably after excluding 7 variants in the *NPM3*; the *p*-value of the model is increased to 4.01e-09. Figure 3 shows a regional plot of individual variant effects with MAC>3 based on a single-based regression approach. The figure shows multiple variants with small P-values within FGF8 gene. The SNP on the right hand side of figure 3 is outside of the boundaries of the NPM3 gene.

In addition to the CCRS method, we applied other common methods for rare variants analysis such as CAST (Morgenthaler et al., 2007) and SKAT-O (Lee et al., 2012). The results of these analyses, identified one significant region using SKAT-O and this region is on chromosome 14 and is the same region associated with c*is-4-decenoylcarnitine* in our analysis using CCRS.

Discussion

Whole genome sequence analysis of complex traits is in its infancy, and to date there has been only one other foray into the area (Morrison et al., 2013). Success depends on the bringing together of multiple complementary lines of investigation. In this study, we combined whole genome sequencing in 1,456 individuals, detailed metabolomics measures of 16 analytes involved in carnitine synthesis and metabolism, and innovative analytic methods that take into account the very large number of rare genomic variants while maintaining statistical power and controlling the false discovery rate in order to identify that the FGF8 and MDGA2 genes were associated with inter-individuals variation in lysine and cis-4-decenoylcarnitine, respectively. In the absence of replication or detailed mechanistic studies, these findings should be viewed as preliminary, and FGF8 and MDGA2 should be viewed as candidate genes influencing lysine and cis-4-decenoylcarnitine, respectively.

We selected carnitine biosynthesis and metabolism for this whole genome analysis because of its moderate size (i.e. 16 metabolites were represented in the serum metabolomics measurements) and its role in health and disease, primarily through fatty acid metabolism (Hoppel C. et al., 2003). In addition to overt carnitine deficiency, carnitine supplementation has been suggested as a therapeutic for multiple heart- and muscle-related disorders (Kelly

G. S. et al., 1998). Biosynthesis of carnitine occurs primarily in the liver and kidney from the amino acids lysine and methionine (Bremer J., 1983). Lysine is an essential amino acid and is abundant in foods high in protein. In addition to lysine, another metabolite having significant effects from this whole genome analysis was cis-4-decenoylcarnitine. Unlike its cousin octanoylcarnitine, there is little literature on the function and metabolism of decenoylcarnitine. Decenoylcarnitine is elevated with carnitine treatment (Vernez et al., 2006). Bhuiyan et al. (1992) report that decenoylcarnitine can be found in the urine of patients suspected of having defects in mitochrondrial beta-oxidation.

The mechanisms of the association of FGF8 and MDGA2 with lysine and cis-4-decenoylcarnitine, respectively, is unknown. Although carnitine is synthesized from the essential amino acid lysine, we did not observe evidence that reduced lysine leads to reduced carnitine. In the data set analyzed here, the correlation between lysine and carnitine was only 0.02. In addition, individuals with rare variants in FGF8 associated with reduced lysine did not show deficiency in the other 15 metabolites selected from the pathway of carnitine synthesis and metabolism (data not shown). This most likely is due to only a small percentage of lysine going to carnitine biosynthesis. FGF8 is a member of the fibroblast growth factor family expressed at high levels in the testes and ovaries (www.genecards.org), and the protein is necessary for normal brain development and maintenance, specifically the borders among certain segments (Crossley and Martin, 1995). Carnitine metabolism is involved in both male fertility (Ahmed S. D. et al., 2011) and a variety of brain processes (Nałecz K. A. et al., 2004). Furthermore, relative lysine deficiency may be related to multiple cognitive and behavioral-related disorders by acting as precursor of glutamate in the central nervous system (Papes F. et al., 2001) and its relationship with cortisol (Smriga M. et al., 2007).

Using the CCRS test, MDGA2 was significantly associated with cis-4-decenoylcarnitine. In follow-up analyses, there were three rare synonymous variants that were individually significantly related to the levels of this metabolite. As can be seen in Table III, one of the variants that is in the beginning of the gene increased levels, while the other two toward the end of the gene decreased levels. MDGA2 is expressed in the central and peripheral nervous system in the rat (http://rgd.mcw.edu/). Previously, MDGA2 was known as MAMDC1. MDGA2 gene variation has previously been associated with multiple neurologic and behavioral conditions (https://www.ebi.ac.uk/gwas/). Similarly, carnitine and carnitine metabolites have also been related to multiple neurologic and behavioral conditions, especially autism (e.g. Filipek P. A. et al., 2004).

The results reported here document the feasibility of whole genome sequence analysis of endophenotypes close to the gene level. They also document the benefits of applying multiple methods of analysis, in this case a burden test, SKAT and CCRS, to promote discovery because each method has relative strengths and weaknesses. In the future, sample sizes for whole genome sequence analysis of metabolomics data will increase, partly fueled by declining costs. Likewise, projects such as ENCODE and GTEX will permit better annotation of non-genic coding regions. In addition to the domains of genetic epidemiology, adequate sample sets, quality phenotyping and whole genome sequencing, future studies will

need to directly incorporate high through put functional assays as an integral part of discovery.

Acknowledgments

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The metabolome measurement work obtained through support from the National Genome Research Institute (HG004402) and the generous support of the University of Texas Health Science Center at Houston. The DNA sequence data work obtained through support from the National Heart Lung and Blood Institute (HL102419) and National Human Genome Research Institute (HG003273 and HG006542) of the National Institute of Health.

References

Ahmed SDH, Karira KA, Ahsan S. Role of L-carnitine in male infertility. J. Pak. Med. Assoc. 2011; 61(8):732. [PubMed: 22355991]

Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, BN., Csaki, BF., editors. Second International Symposium on Information Theory. Academiai Kiado; Budapest: 1973. p. 267-281.

Bhuiyan AKMJ, Jackson S, Turnbull DM, Aynsley-Green A, Leonard JV, Bartlett K. The measurement of carnitine and acyl-carnitines: application to the investigation of patients with suspected inherited disorders of mitochondrial fatty acid oxidation. Clin. Chim. Acta. 1992; 207(3):185–204. [PubMed: 1327583]

Bremer J. Carnitine-metabolism and functions. Physiol Rev. 1983; 63(4):1420–1480. [PubMed: 6361812]

Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media. 2002

Consortium EP, et al. A user's guide to the encyclopedia of DNA elements (encode). PLoS Biol. 2011; 9(4):e1001046. [PubMed: 21526222]

Crossley PH, Martin GR. The mouse Fgf8 gene encodes a family of polypeptides and is expressed in regions that direct outgrowth and patterning in the developing embryo. Development. 1995; 121(2): 439–451. [PubMed: 7768185]

Demirkan A, Henneman P, Verhoeven A, Dharuri H, Amin N, van Klinken JB, Karssen LC, de Vries B, Meissner A, Göraler S, et al. Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. PLoS Genet. 2015; 11(1):e1004835. [PubMed: 25569235]

Feng T, Zhu X. Whole genome sequencing data from pedigrees suggests linkage disequilibrium among rare variants created by population admixture. BMC Proc. 2014; 8(Supp 1):S44. [PubMed: 25519326]

Filipek PA, Juranek J, Nguyen MT, Cummings C, Gargus JJ. Relative carnitine deficiency in autism. J. Autism Dev. Disord. 2004; 34(6):615–623. [PubMed: 15679182]

Flanagan JL, Simmons PA, Vehige J, Willcox M, Garrett Q. Review role of carnitine in disease. Nutr.Metab. 2010; 7:30.

Frank LE, Friedman JH. A statistical view of some chemometrics regression tools. Technometrics. 1993; 35(2):109–135.

Fu WJ. Penalized regressions: the bridge versus the lasso. J. Comp. Graph. Stat. 1998; 7(3):379–416.

Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005; 27(2):83–85.

Hoppel C. The role of carnitine in normal and altered fatty acid metabolism. Am. J. Kidney Dis. 2003; 41:S4–S12. [PubMed: 12751049]

Kelly GS. L-Carnitine: therapeutic applications of a conditionally-essential amino acid. Alternative medicine review: a journal of clinical therapeutic. 1998; 3(5):345–360. [PubMed: 9804680]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team ELP, Christiani DC, Wurfel MM, Lin X, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The Am. J. Hum. Genet. 2012; 91(2):224–237. [PubMed: 22863193]

Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The Am. J. Hum. Genet. 2014; 95(1):5–23. [PubMed: 24995866]

Li F-Y, El-Hattab AW, Bawle EV, Boles RG, Schmitt ES, Scaglia F, Wong LJ. Molecular spectrum of slc22a5 (octn2) gene mutations detected in 143 subjects evaluated for systemic carnitine deficiency. Hum. Mutat. 2010; 31(8):E1632–E1651. [PubMed: 20574985]

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2007; 615(1):28–56. [PubMed: 17101154]

Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, Li A, Muzny D, Yu F, Rice K, Zhu C, et al. Whole genome sequence-based analysis of a model complex trait, high density lipoprotein cholesterol. Nat. Genet. 2013; 45(8):899. [PubMed: 23770607]

Nałęcz KA, Miecz D, Berezowski V, Cecchelli R. Carnitine: transport and physiological functions in the brain. Molecular aspects of medicine. 2004; 25(5):551–567. [PubMed: 15363641]

Olpin SE. Fatty acid oxidation defects as a cause of neuromyopathic disease in infants and adults. Clin. Lab. 2004; 51(5–6):289–306.

Pearson H. Meet the human metabolome. Nature. 2007; 446(7131):8. [PubMed: 17330009]

Papes F, Surpili MJ, Langone F, Trigo JR, Arruda P. The essential amino acid lysine acts as precursor of glutamate in the mammalian central nervous system. FEBS letters. 2001; 488(1):34–38. [PubMed: 11163791]

Rhee EP, Ho JE, Chen MH, Shen D, Cheng S, Larson MG, Ghorbani A, Shi X, Helenius IT, O'Donnell CJ, et al. A genome-wide association study of the human metabolome in a community-based cohort. Cell Metab. 2013; 18(1):130–143. [PubMed: 23823483]

Smriga M, Ando T, Akutsu M, Furukawa Y, Miwa K, Morinaga Y. Oral treatment with L-lysine and L-arginine reduces anxiety and basal cortisol levels in healthy humans. Biomedical Research. 2007; 28(2):85–90. [PubMed: 17510493]

Steiber A, Kerner J, Hoppel CL. Carnitine: a nutritional, biosynthetic, and functional perspective. Molecular aspects of medicine. 2004; 25(5):455–473. [PubMed: 15363636]

Suhre K, Wallaschofski H, Raffler J, Friedrich N, Haring R, Michael K, Wasner C, Krebs A, Kronenberg F, Chang D, et al. A genome-wide association study of metabolic traits in human urine. Nat. Genet. 2011; 43(6):565–569. [PubMed: 21572414]

Talluri R, Shete S. A linkage disequilibrium-based approach to selecting disease-associated rare variants. PloS one. 2013; 8(7):e69226. [PubMed: 23874919]

The ARIC investigator. The atherosclerosis risk in communities (aric) study: design and objectives. Am. J. Epidemiol. 1989; 129:687–702. [PubMed: 2646917]

Verboven S, Branden KV, Goos P. Sequential imputation for missing values. Comp. Biol. Chem. 2007; 31(5):320–327.

Vernez L, Dickenmann M, Steiger J, Wenk M, Krähenbühl S. Effect of L-carnitine on the kinetics of carnitine, acylcarnitines and butyrobetaine in long-term haemodialysis. Nephrol. Dial. Transplant. 2006; 21(2):450–458. [PubMed: 16286428]

Wen W, Li D, Li X, Gao Y, Li W, Li H, Liu J, Liu H, Chen W, Luo J, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat. Commun. 2014; 5

Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostat. 2009; doi: 10.1093/biostatistics/kxp008

Yazdani A, Yazdani A, Boerwinkle E. Rare Variants Selection using Penalization Methods for the Analysis of Whole Genome Sequence Data. BMC Bioinform. 2015; 16(1):405.doi: 10.1186/s12859-015-0825-4

Yu B, Zheng Y, Alexander D, Morrison AC, Coresh J, Boerwinkle E. Genetic determinants inuencing human serum metabolome among african americans. PLoS Genet. 2014; 10:e1004212. [PubMed: 24625756]

Zheng Y, Yu B, Alexander D, Manolio TA, Aguilar D, Coresh J, Heiss G, Boerwinkle E, Nettleton JA. Associations between metabolomic compounds and incident heart failure among african americans: the aric study. Am. J. Epidemiol. 2013 page kwt004.

Zheng Y, Yu B, Alexander D, Couper DJ, Boerwinkle E. Medium-term variability of the human serum metabolome in the atherosclerosis risk in communities (aric) study. omics. A Journal of Integrative Biology. 2014; 18(6):364–373. [PubMed: 24910946]

Box 1: Permutation Algorithm to obtain Significance Threshold

- Randomly selecting *m* windows.
- Permute the observed phenotypes *n* times and obtain a *p*-value for likelihood ratio test for each permutation.
- Estimate the empirical distribution of *p*-values and calculate the quantile for α in each window.
- Define the significant threshold for the observed *p*-value as the minimum value in the set of m calculated quantiles in Step 3. *p*-values less than this threshold are deemed significant and the specific window of the genome and metabolite deserve further follow-up.

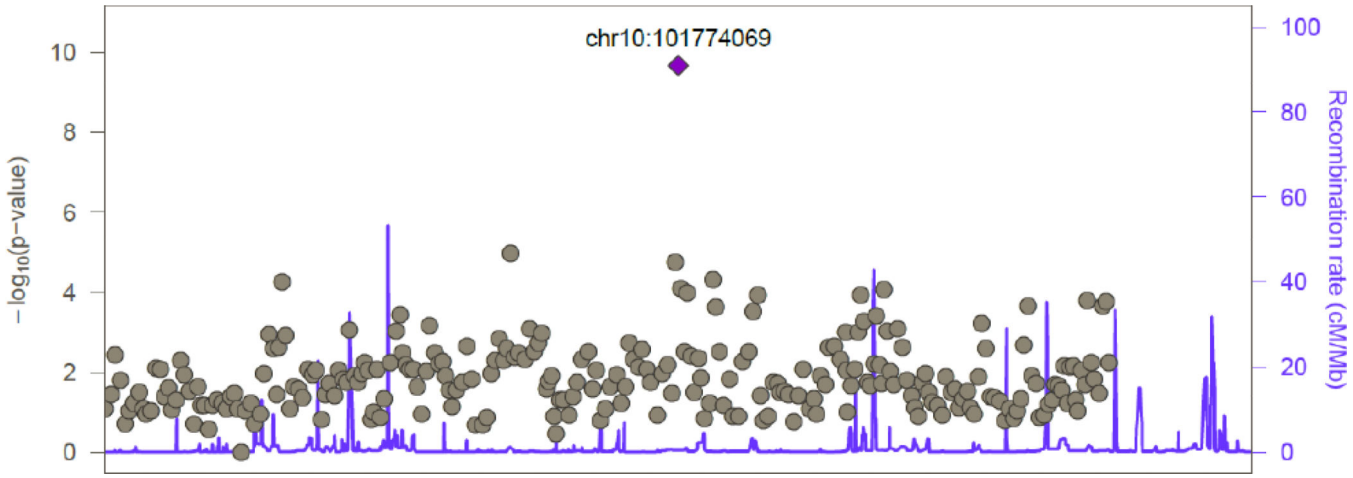

Figure 1.
Regional plot of 201 windows centered on associated region with *lysine*

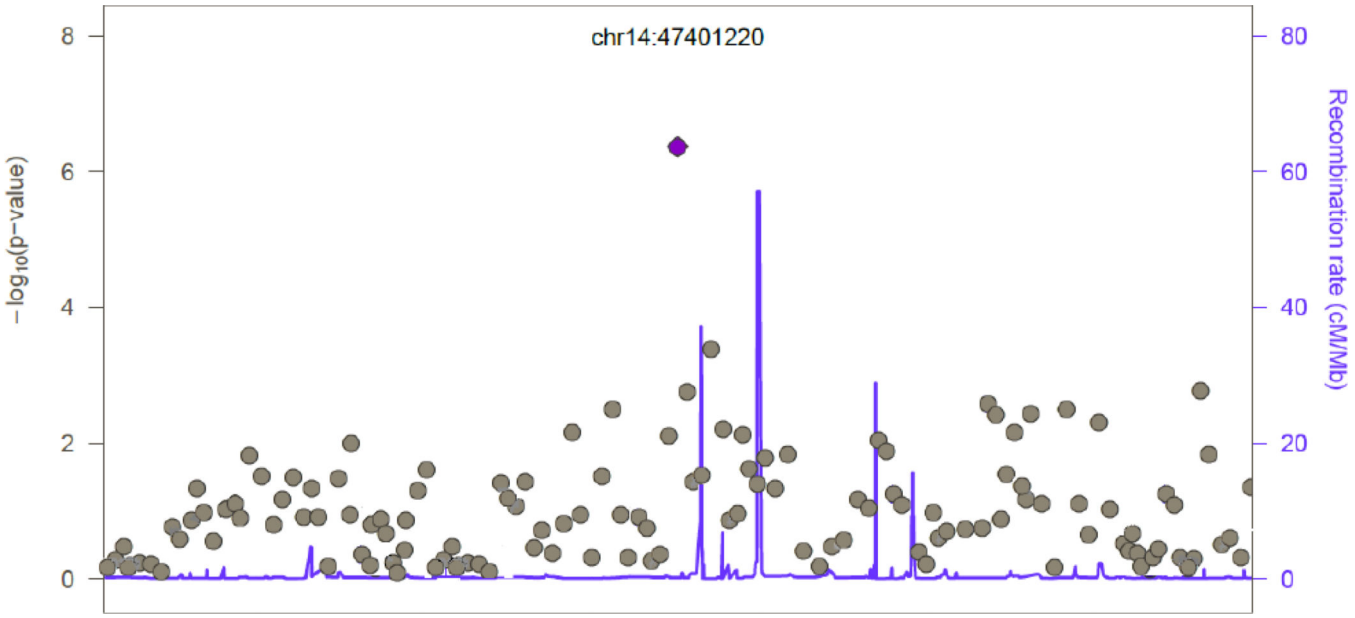

Figure 2.
Regional plot of 201 windows centered on associated region with *Cis-4-decenoylcarnitine*

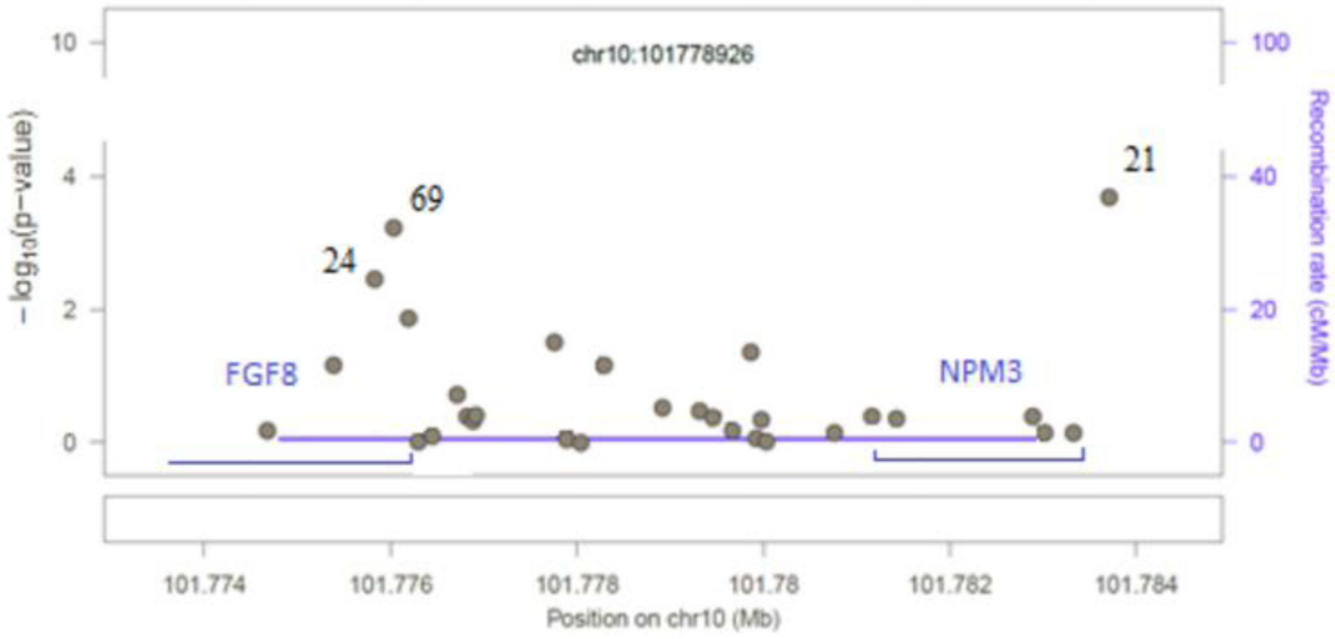

Figure 3.

Regional plot of 30 variants with MAC≥ 3 included in associated region with *lysine* with numbers of mutation for variants with *p*-value < 10^{-2}

Table I

Name and the range of metabolites after log transformation

Name	Average (Range)	Name	Average (Range)
octanoylcarnitine	−0.101 (−0.621, 0.542)	*deoxycarnitine*	−0.066 (−2.074, 1.640)
decanoylcarnitine	−0.100 (−0.786, 0.657)	*carnitine*[1]	−0.042 (−2.130, 2.365)
cis-4-cenoylcarnitine	−0.080 (−1.198, 1.061)	*glutarate*	−0.130 (−1.136, 1.124)
laurylcarnitine	−0.076 (−0.820, 0.642)	*leucine*	−0.092 (−1.133, 0.945)
glutarylcarnitine	−0.080 (−0.943, 0.786)	*lysine*	−0.081 (−0.696, 0.489)
isovalerylcarnitine	−0.129 (−1.388, 1.134)	*N6acetyllysine*	−0.066 (−0.635, 0.620)
isobutyrylcarnitine	−0.088 (−0.799, 0.748)	*citrate*	−0.094 (−1.805, 1.676)
propionylcarnitine	−0.091 (−1.720, 1.243)	*succinate*	−0.061 (−0.891, 0.935)

[1] Variants that are not transformed.

Table II

Baseline characteristics of ARIC European-Americans participants and those with metabolomics and genomic data. The numbers in parentheses represent standard deviations.

ARIC European-Americans		
	Whole data set	**Subset under study**
N	11,478	1,456
Age (years)	54(6)	55(6)
Male (%)	47.3	45.9
Diabetes (%)	9.1	8.0
Current smoker (%)	24.8	25.8
Hypertension (%)	27.3	31.7
Systolic bp (mmHg)	118.5(17.0)	119.4(18.4)
Diastolic bp (mmHg)	71.5(10.0)	71.7(10.8)
Glucose (mg/dL)	105.6(32.1)	105.7(29.8)
BMI (kg/m^2)	27.0(4.9)	27.3(5.0)
HDL (mg/dL)	50.4(16.8)	50.0(16.5)
Total cholesterol (mg/dL)	215.0(40.8)	216.2(40.3)
Triglycerides (mg/dL)	138.1(93.0)	144.5(110.1)
eGFRCKD-EPI (mL/min/1.73 m^2)	0.48(0.08)	0.47(0.09)

Table III

Summary information for two significant windows influencing two metabolites

	Chr	Location	Physical Distance	Genes	*p*-value of LRT	AIC_i - $AIC_{selected}$
lysine	10	101774069-83714	9,646 bp	*FGF8 NPM3*	1.752e-10	>4.2
cis-4-decenoylcarnitine	14	47401220-11509	10,290 bp	*MDGA2*	4.239e-07	>2.4

Table IV

Summary information of promising variants in two significant windows influencing two metabolites

Chromosome No. Position	Associated Metabolite	MAC	Estimated Effect Size	Standard Deviation	*p*-value
chr10.101775830	*lysine*	24	−0.127	0.0434	0.0035
chr10.101776036	*lysine*	69	−0.090	0.026	0.0005
chr10.101783714	*lysine*	21	0.172	0.046	0.0002
chr14.47403473	*cis-4-decenoyl carnitine*	58	0.235	0.053	9.202e-06
chr14.47408133	*cis-4-decenoyl carnitine*	23	−0.284	0.083	6.547e-04
chr14.47411509	*cis-4-decenoyl carnitine*	14	−0.433	0.1065	4.729e-05