Identification of seven novel loci associated with amino acid levels using single variant and gene-based tests in 8,545 Finnish men from the METSIM study

Tanya M. Teslovich,[1†] Daniel Seung Kim,[1†] Xianyong Yin,[1†] Alena Stančáková,[2] Anne U. Jackson,[1] Matthias Wielscher,[3] Adam Naj,[4,5] John R.B. Perry,[6] Jeroen R. Huyghe,[1] Heather M. Stringham,[1] James P. Davis,[7] Chelsea K. Raulerson,[7] Ryan P. Welch,[1] Christian Fuchsberger,[1] Adam E. Locke,[1] Xueling Sim,[1] Peter S. Chines,[8] Narisu Narisu,[8] Antti J. Kangas,[9] Pasi Soininen,[9,10] Genetics of Obesity-related Liver Disease Consortium (GOLD), The Alzheimer's Disease Genetics Consortium (ADGC) , The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM), Mika Ala-Korpela,[9-14] Vilmundur Gudnason,[15] Solomon K. Musani,[16] Marjo-Riitta Jarvelin,[3,17-19] Gerard D. Schellenberg,[4] Elizabeth K. Speliotes,[20,21] Johanna Kuusisto,[2] Francis S. Collins,[8] Michael Boehnke,[1*§] Markku Laakso,[2*§] Karen L. Mohlke[7*§]

[*] **Corresponding Authors:**
Michael Boehnke, Ph.D.
boehnke@umich.edu
+1-734-936-1001 / (Fax): +1-734-615-8322

Markku Laakso, M.D., Ph.D.
markku.laakso@uef.fi
+358-02-9445-4046

Karen L. Mohlke, Ph.D.
mohlke@med.unc.edu
+1-919-966-2913 / (Fax): +1-919-843-0291

[†] These authors contributed equally to this work (co-first authors).
[§] These authors jointly supervised this work (co-last authors).

This is a pre-copyedited, author-produced version of an article accepted for publication in [insert journal title] following peer review. The version of record Tanya M Teslovich, Daniel Seung Kim, Xianyong Yin, Alena Stančáková, Anne U Jackson, Matthias Wielscher, Adam Naj, John R B Perry, Jeroen R Huyghe, Heather M Stringham, James P Davis, Chelsea K Raulerson, Ryan P Welch, Christian Fuchsberger, Adam E Locke, Xueling Sim, Peter S Chines, Narisu Narisu, Antti J Kangas, Pasi Soininen, Genetics of Obesity-Related Liver Disease Consortium (GOLD), The Alzheimer's Disease Genetics Consortium (ADGC), The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM), Mika Ala-Korpela, Vilmundur Gudnason, Solomon K Musani, Marjo-Riitta Jarvelin, Gerard D Schellenberg, Elizabeth K Speliotes, Johanna Kuusisto, Francis S Collins, Michael Boehnke, Markku Laakso, Karen L Mohlke, Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study, *Human Molecular Genetics*, Volume 27, Issue 9, 01 May 2018, Pages 1664–1674, is available online at: https://doi.org/10.1093/hmg/ddy067.

1) Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.
2) Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland.
3) Department of Epidemiology and Biostatistics, MRC–PHE Centre for Environment & Health, School of Public Health, Imperial College London, London, UK.
4) Department of Pathology and Laboratory Medicine, Penn Neurodegeneration Genomics Center, University of Pennsylvania, Philadelphia, USA.
5) Departments of Biostatistics, and Epidemiology (DBE) and Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, USA.
6) MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK.
7) Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, USA.
8) National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.
9) Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland.
10) NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland.
11) Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK
12) Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
13) Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia
14) Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia
15) Icelandic Heart Association, Kopavogur IS-201, Iceland
16) University of Mississippi Medical Center, Jackson, Mississippi, 39213, USA
17) Center for Life Course Health Research, Faculty of Medicine, University of Oulu, 90014 Oulu, Finland.
18) Biocenter Oulu, University of Oulu, Finland.
19) Unit of Primary Care, Oulu University Hospital, Oulu, Finland
20) Division of Gastroenterology, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA.
21) Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA.

ABSTRACT

Comprehensive metabolite profiling captures many highly heritable traits, including amino acid levels, which are potentially sensitive biomarkers for disease pathogenesis. To better understand the contribution of genetic variation to amino acid levels, we performed single variant and gene-based tests of association between nine serum amino acids (alanine, glutamine, glycine, histidine, isoleucine, leucine, phenylalanine, tyrosine, and valine) and 16.6 million genotyped and imputed variants in 8,545 non-diabetic Finnish men from the METabolic Syndrome In Men (METSIM) study. We identified five novel loci associated with amino acid levels ($P<5\times10^{-8}$): *LOC157273*/*PPP1R3B* with glycine (rs9987289, $P=2.3\times10^{-26}$); *ZFHX3* (chr16:73326579, minor allele frequency (MAF)=0.42%, $P=3.6\times10^{-9}$), *LIPC* (rs10468017, $P=1.5\times10^{-8}$), and *WWOX* (rs9937914, $P=3.8\times10^{-8}$) with alanine; and *TRIB1* with tyrosine (rs28601761, $P=8.8\times10^{-9}$). Gene-based tests identified two novel genes harboring missense variants of MAF<1% that show aggregate association with amino acid levels: *PYCR1* with glycine ($P_{gene}=1.5\times10^{-6}$) and *BCAT2* with valine ($P_{gene}=7.4\times10^{-7}$); neither gene was implicated by single variant association tests. These findings are among the first applications of gene-based tests to identify new loci for amino acid levels. In addition to the seven novel gene associations, we identified five independent signals at established amino acid loci, including two rare variant signals at *GLDC* (rs138640017, MAF=0.95%, $P_{conditional}=5.8\times10^{-40}$) with glycine levels and *HAL* (rs141635447, MAF=0.46%, $P_{conditional}=9.4\times10^{-11}$) with histidine levels. Examination of all single variant association results in our data revealed a strong inverse relationship between effect size and MAF ($P_{trend}<0.001$). These novel signals provide further insight

into the molecular mechanisms of amino acid metabolism and potentially, their perturbations in disease.

Abstract word count: 249/250

INTRODUCTION

Amino acid levels are highly heritable biomarkers of human disease (1) that have been implicated in a range of clinical syndromes including type 2 diabetes/insulin resistance (2–4), liver disease (5), and Alzheimer's disease (6). Previous studies have together identified >200 common variant signals associated with amino acid levels (7–20). However, the contribution of genetic variation to amino acid level trait variance, and the role of rare genetic variation in particular, is not fully understood.

One method of assessing rare variant associations is through aggregation of multiple rare variants into a single test (21). One such approach groups rare, protein-altering variants into one test for association for each gene (21). This method has been used successfully to identify gene-based associations with *HAL* for histidine levels and with *PAH* for phenylalanine levels (17). Notably, this result occurred in the absence of any single variant reaching genome-wide significance in either *HAL* or *PAH*, highlighting the importance of gene-based tests in identifying novel genetic loci for complex traits.

In this study, we performed genome-wide single variant and gene-based association analysis in 8,545 non-diabetic Finnish men from the METabolic Syndrome In Men (METSIM) study to identify genetic associations with serum amino acid levels. We identified seven novel amino acid loci – five from single variant tests (of which two signals replicated in the Northern Finnish Birth Cohort 1966 (NFBC1966) dataset) and two from gene-based associations. We also performed analyses conditioned on all previously known amino acid genome-wide association studies (GWAS) signals and identified five

additional novel and independent signals in known amino acid loci, of which three replicated in the NFBC1966 data. In total, we identified five novel and replicated loci-amino acid associations, and two novel gene-based associations. These results help clarify the role of the specific variants and genes in amino acid homeostasis.

RESULTS

GWAS for nine amino acids levels

To identify genetic variants associated with the nine amino acid traits measured in the METSIM study (alanine, glutamine, glycine, histidine, isoleucine, leucine, phenylalanine, tyrosine, and valine; see **Supplemental Figure 1-3** and **Supplemental Table 1**), we analyzed 16.6M genotyped and imputed variants in 8,545 non-diabetic Finnish men of mean age 57 years and mean BMI 27 kg/m^2 (see **Supplemental Table 1**).

We identified 2,428 unique variants associated with at least one amino acid trait ($P < 5\times10^{-8}$), and a total of 2,580 variant-trait associations (see **Supplemental Table 2, Supplemental Figures 4** and **5**). Of the 2,580 variant-trait associations, the majority were with glycine (1,403 variants), followed by tyrosine (560), glutamine (164), alanine (95), leucine (89), isoleucine (87), phenylalanine (67), valine (62), and histidine (53). We present a summary of the variants and their distributions in independent loci in **Supplemental Table 3**.

We estimated for each amino acid trait the phenotypic variation that genetic variants explained from 10.4% (histidine) to 28.5% (glycine) of variation (see **Supplemental Table 4)**. Restricting analysis to genome-wide significant variant-trait associations ($P < 5\times10^{-8}$, see **Supplemental Table 2**), the proportion of phenotypic variation explained by significantly associated variants ranged from 1.3% (leucine) to 18.3% (glycine) (see **Supplemental Table 4**).

We attempted to validate genotypes at three rare imputed trait-associated variants with MAF<0.5% (see **Supplemental Table 5**). We confirmed two variants with no discordance between imputed and sequenced genotypes: rs141635447 (0/74 discordant) and chr16:73326579 (0/67). Variant chr3:125173967 showed a discordance rate of 39% (24/61), and was thus removed from subsequent analyses.

Single variant analysis identifies novel associations at *LOC157273/PPP1R3B*, *WWOX*, *LIPC, TRIB1, and ZFHX3*

Genome-wide single variant analyses identified five novel amino acid-associated loci at least 1 Mb away from the nearest known GWAS variant (see **Table 1** and **Supplemental Figure 6a-e**). Of the five novel loci (see **Table 1**), two were located in the introns of *LOC157273* (near *PPP1R3B*) *and WWOX.* At the *LOC157273/PPP1R3B* locus, intronic variant rs9987289-A was associated with decreased glycine levels (MAF=17.0%, β=-0.22, P=2.3×10^{-26}, **Supplemental Figure 6a**). This variant was replicated in the NFBC1966 cohort (β=-0.15, P=7.3×10^{-4}, see **Table 1**), and was associated with the risk of type 2 diabetes (Odds Ratio(OR)=1.05, P=0.02) and liver disease (OR=1.33, P=4.7×10^{-18}, see **Supplemental Table 6**). Within the *WWOX* region*,* intronic variant rs9937914-G was associated with increased alanine levels (MAF=1.47%, β=0.36, P=3.8×10^{-8}, **Supplemental Figure 6b**).

We identified two additional novel loci in regions previously highlighted by genome-wide association studies: first, in the region upstream of *LIPC,* a gene implicated in numerous lipid traits including high-density lipoprotein (HDL) cholesterol (22), phospholipids (20),

and the ratio of isoleucine and serum total cholesterol (serum-c) (10), rs10468017-T was associated with increased alanine levels (MAF=33.2%, β=0.09, P=1.5×10^{-8}, **Supplemental Figure 6c**), and is in strong linkage disequilibrium (LD) with rs1532085 (LD r^2=0.66), a *LIPC* GWAS locus for HDL (23) and ratio of isoleucine and serum-c (10). Its association with increased alanine levels was confirmed in the NFBC1966 cohort (β=0.08, P=7.7×10^{-3}, see **Table 1**). Second, rs28601761-G, for which we report an association with decreased tyrosine levels (MAF=42.2%, β=-0.09, P=8.8×10^{-9}, **Supplemental Figure 6d**), is in strong LD with rs2954029 (LD r^2=0.71), a *TRIB1* GWAS variant for low-density lipoprotein (LDL) cholesterol, triglycerides, and high-density lipoprotein (HDL) cholesterol levels (23, 24).

At the remaining novel locus, rare variant 16:73326579 was associated with increased alanine levels (MAF=0.42%, β=0.76, P=3.6×10^{-9}). 16:73326579 is located within 300 kb of both *HCCAT5* and *ZFHX3,* but is not in strong LD (LD r^2<0.60) with any coding variant observed in the GoT2D study (see **Supplemental Figure 6e**). We computed P values (P_{ACT}) for the novel variants after correcting for the nine correlated amino acid traits (25). All of these five novel variants remained genome-wide significantly associated even after correcting for the nine amino acid traits (P_{ACT} < 5.0 ×10^{-8}).

Conditional analyses identify independent signals at five known amino acid loci: *GLDC, HAL, ALDH1L1, ADAMTS3,* and *GCSH*

We curated 1,519 unique known amino acid associated variants, and then used them as covariates in the genome-wide conditional analyses (see **Supplemental Table 7,**

206 **Materials and Methods**). After conditional analyses, we observed 227 unique variant-
207 trait associations ($P_{conditional}<5\times10^{-8}$) (see **Supplemental Table 8**), whose distributions in
208 genes are presented in **Supplemental Table 3.** Among these, we identified five novel
209 signals at established amino acid loci distinct from the previously published GWAS
210 variants (see **Table 1**): *GLDC* p.Q996H, associated with increased glycine (rs138640017-
211 G, MAF=0.95%, β=1.35, $P_{conditional}=5.8\times10^{-40}$); *HAL* p.G283V, associated with increased
212 histidine levels (rs141635447-A, MAF=0.46%, β=0.85, $P_{conditional}=9.4\times10^{-11}$); rs6564825-
213 G, in an intron of *PKD1L2* and 38 kb downstream of *GCSH,* was associated with
214 increased glycine levels (MAF=11.7%, β=0.17, $P_{conditional}=2.0\times10^{-10}$) and nominally but not
215 coincidently associated with expression level of *PKD1L2* (see **Supplemental Table 9**);
216 rs190671241-G, an intergenic variant near *ADAMTS3*, associated with increased
217 phenylalanine levels (MAF=1.70%, β=0.36, $P_{conditional}=2.2\times10^{-9}$); and rs112981908-G, an
218 intronic variant of the *ALDH1L1* gene associated with decreased glycine levels
219 (MAF=11.9%, β=-0.14, $P_{conditional}=3.1\times10^{-10}$). At each locus, we observed low pairwise LD
220 (LD $r^2<0.10$) between the novel variant identified in the METSIM data and the previously
221 published GWAS variant(s). Notably, of the five novel signals at established amino acid
222 loci, three replicated in the NFBC1966 data: *GLDC* p.Q996H, associated with increased
223 glycine (rs138640017-G, β=0.94, $P=2.7\times10^{-10}$); *HAL* p.G283V, associated with increased
224 histidine levels (rs141635447-A, β=1.04, $P=1.7\times10^{-5}$); and *ADAMTS3* upstream variant
225 rs190671241-G, associated with increased phenylalanine levels (β=0.39, $P=1.6\times10^{-4}$).
226 Functional work is necessary to determine whether the novel signals represent additional

functional variants in genes known to play a role in amino acid metabolism (e.g. *GLDC* and *HAL*), or whether they point to novel mechanisms.

Single variant associations exhibit an inverse relationship between allele frequency and effect size

To visualize the relationship between allele frequency and effect size of amino acid-associated variants, we plotted the absolute value of effect size estimates vs. MAF for all loci associated with amino acid traits in the METSIM study ($P<5\times10^{-8}$) (see **Table 1**) and fitted a fractional polynomial spline to the data (see **Figure 1**). These results demonstrated a strong, inverse relationship between MAF and effect size ($P_{trend}<0.001$), consistent with past findings for other traits (e.g. (26)). This relationship is largely driven by variants with MAF<5%, five of which were newly identified in this study.

Variant associations with amino acid ratios

Prior studies found more genome-wide significant variant associations with amino acid ratios as compared to amino acid traits alone (8, 12). We identified 15,220 significant variant-ratio associations (3,822 unique variants) from unconditional analyses of the 36 possible ratios among the nine amino acids measured in the METSIM study. These results are presented in **Supplemental Table 10** as a reference for other investigators.

Gene-based tests identify novel gene associations with *BCAT2* and *PYCR1*

To determine the joint contribution of the protein-truncating and missense variants of MAF<1% on amino acid traits, we performed gene-based tests (see **Supplemental Table**

11 and Supplemental Figure 7) and applied a significance threshold based on the number of genes tested (~20,000). Given the high correlation among amino acid traits levels (see **Supplemental Figure 2**), we did not correct for the number of amino acid traits as a Bonferroni-corrected significance threshold would be overly strict. We identified six gene-trait associations ($P_{gene} < 2.5\times10^{-6}$), including four genes previously identified through single variant association tests: *ALDH1L1* (8) and *GLDC* (16) associated with glycine levels, *HAL* with histidine levels (16) (see **Supplemental Table 12**), and *DHODH* (previously associated with alanine-to-tyrosine ratio) with alanine levels, as well as two novel associations between *PYCR1* and glycine levels ($P_{gene}=1.5\times10^{-6}$), and *BCAT2* and valine levels ($P_{gene}=7.4\times10^{-7}$, see **Table 2**).

No missense variants within either *PYCR1* or *BCAT2* achieved genome-wide significance with any amino acid trait in the single variant association tests, highlighting the utility of gene-based test in novel gene discovery (see **Table 2**). Despite only suggestive association evidence, the effect of these variants on amino acid trait variance was considerable (the range of absolute β: 0.36-1.02): the carriers of the two missense variants within the gene *PCYR1* exhibited lower mean glycine levels, while the carriers of the three variants within the *BCAT2* gene showed higher mean valine levels, suggesting altering their protein sequences would affect the serum glycine and valine levels, respectively (see **Figure 2**).

DISCUSSION

Amino acids are highly heritable traits whose levels have been implicated in the pathogenesis of human complex diseases such as type 2 diabetes (2–4) and Alzheimer's disease (6). Here, we leveraged dense, experimentally-determined and imputed genotypes and report seven novel amino acid associations in the METSIM study that replicated in NFBC1966. Of these associations, two were identified from single variant testing in the METSIM study and replicated in the NFBC1966 data, and two other loci were identified from gene-based analyses in the METSIM study alone. One of these newly identified variants from single-variant analyses, *LOC157273/PPP1R3B* variant rs9987289-A, also confers increased risk of type 2 diabetes and liver disease. In addition, we fine-mapped known amino acid loci and identified and replicated distinct association signals at three of these loci. Phenotypic variance explained for these nine amino acids by known and novel associations ranged from 10.4% for histidine levels to 28.5% for glycine levels in our data. These results further elucidate the potential mechanisms through which amino acid levels are perturbed, and their potential relationship to disease.

Novel loci highlight a potential role for genes implicated in lipid metabolism and human diseases

Of the five novel amino acid loci highlighted in this study through single variant analyses, four have previously been implicated in lipid metabolism. First, the rs9987289-A signal near *LOC157273*/*PPP1R3B*, for which we report an association with decreased glycine levels, has previously been associated with decreased high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and total cholesterol

levels (23), as well as increased *PPP1R3B* expression levels in human liver tissue (23). In addition, overexpression of *Ppp1r3b* led to a significant decrease in HDL-C and total cholesterol in a mouse model (23). This variant is also associated with increased risk of type 2 diabetes (27) and liver disease (28). Second, the rs10468017-T signal upstream of *LIPC* associated with increased alanine levels has previously been associated with increased HDL-C (22), altered levels of several circulating phospholipids (20), and ratio of isoleucine and serum-c (10). Third, the signal near *TRIB1* associated with decreased tyrosine levels has previously been associated with decreased total cholesterol, LDL-C, and triglyceride levels (23). Finally, we reported an intronic *WWOX* variant, rs9937914-G, associated with increased alanine levels, and human carriers of predicted loss-of-function variants in *WWOX* were reported to have lower HDL-C (31) levels; in addition, mice lacking *Wwox* exhibit decreased fasting cholesterol, triglyceride and glucose levels (32). These variant associations may represent a secondary effect of altered lipid levels on amino acid metabolism, as previously demonstrated in the setting of obesity (33) and insulin resistance (2). Further work will be required to determine the mechanisms through which these lipid-related loci affect amino acid levels and human complex diseases.

Gene-based tests highlight two novel loci not identified from single variant testing

PYCR1, which we identified as a gene implicated in glycine levels, encodes a mitochondrial protein involved in biosynthesis of proline and generation of oxidative potential through NADP+ production (34). *PYCR1* was recently identified as the genetic cause of autosomal recessive cutis laxa type 2, highlighting the importance of normal *PYCR1* function in neurodevelopment (35). Functional studies of fibroblasts from affected

316 individuals found increased sensitivity to oxidative stress (36). As redox reactions are
317 critical to amino acid biosynthesis (37), our finding that *PYCR1* missense variants result
318 in decreased glycine levels may suggest reduced oxidative potential *in vivo*.
319
320 We also identified an association between *BCAT2* variants and valine levels through
321 gene-based testing. *BCAT2* encodes a mitochondrial enzyme responsible for the first
322 steps in the breakdown of branched chain amino acids (isoleucine, leucine, and valine)
323 (37); thus the relationship of *BCAT2* with valine levels is clear. Prior literature reports that
324 the deletion of exon 2 in the mouse homologue of *BCAT2* resulted in a phenotype similar
325 to Maple Syrup Urine Disease (38), an autosomal recessive human inborn error of
326 metabolism characterized by high levels of branched chain amino acids and resulting
327 neurologic symptoms due to the inability to catabolize dietary branched chain amino acids.
328 A human case study of an adult male with mild neurologic symptoms (headaches and
329 mild memory loss) has been reported similar findings, with R170Q and E264K missense
330 variants in *BCAT2* resulting in higher-than-expected levels of leucine, isoleucine, and
331 valine (39). Therefore, our finding of *BCAT2* missense variants resulting in increased
332 valine levels is supported by prior literature.
333
334 **Limitations**
335 Some limitations of this study should be considered. First, our analyses were limited by
336 statistical power secondary to sample size. This was likely one of the contributing factors
337 to our lack of replication of single variant signals in the NFBC1966 data, as the available
338 NFBC1966 replication sample size was modest. Future meta-analyses of amino acid

associations are likely to clarify true signals from false positives. Second, our discovery and replication populations of Finnish men (and women for the replication study) limit the generalizability of our findings. However, this study design also provided the genetic homogeneity needed to identify Finnish-ancestry-specific rare variant associations with amino acid traits. Third, our results infer association between genetic markers and amino acid traits; however, elucidating the causal mechanism through which these variants affect amino acid levels will require further functional work.

Summary

These GWAS of nine amino acid traits in 8,545 participants of the METSIM study identified five novel single variant associations, including variant-trait associations near *LOC157273/PPP1R3B*, *WWOX*, *TRIB1, LIPC*, and *ZFHX3,* of which two were replicated in the NFBC1966 cohort and one (*LOC157273/PPP1R3B*) was also associated with the risk of type 2 diabetes and liver disease. In addition, we identified two novel gene-based signals driven by two and three potentially functional missense variants at *PYCR1* and *BCAT2,* respectively. In *BCAT2,* we validated the association of one rare missense variant in the NFBC1966 study. Our use of a dense reference panel yielded 16.6M genotyped and imputed variants, allowing for high-resolution analyses and fine-mapping of independent genetic signals at *GLDC, GCSH, ALDH1L1*, *ADAMTS3*, and *HAL*; of which the signals at *GLDC, HAL,* and *ADAMTS3* were replicated in the NFBC1966 data. Further work is needed to determine which of the variants identified in this study may affect gene function and the precise roles of the identified genes in amino acid metabolism. These analyses provide further insight into the molecular mechanisms of amino acid

362 metabolism, and, given the importance of amino acid level perturbation in the
363 pathogenesis of numerous human diseases, may yield insights into a wide spectrum of
364 human complex disease.

MATERIALS AND METHODS

Study Participants

Of the 10,197 participants in the METSIM study, we analyzed the subset of 8,545 non-diabetic men of mean age 57.3±7.1 years and BMI 27.0±4.0 kg/m^2 with NMR amino acid trait measurements (see **Supplemental Table 1**). Institutional review boards at the University of Kuopio and Kuopio University approved the METSIM study. Written informed consent was obtained from each participant.

Amino Acid Trait Measurement

Blood samples from METSIM participants were obtained and stored in liquid nitrogen until measurement by NMR, as previously described (40). In brief, fasting serum samples collected at enrollment were stored at -80^{o}C and thawed overnight in a refrigerator before sample preparation. A high-throughput serum NMR metabolomics platform was then used to quantify the levels of individual metabolites using a low-molecular weight metabolite data window (1^H NMR spectra) used to identify amino acids (41). We then used iterative lineshape fitting with known chemical shifts to identify and quantify each specific metabolite (42).

We measured nine amino acid levels by NMR spectroscopy: alanine, glutamine, glycine, histidine, isoleucine, leucine, phenylalanine, tyrosine, and valine (see **Supplemental Table 1**). Visualization of the Pearson pairwise correlation matrix between the nine measured amino acid traits was generated using the corrplot package (https://cran.r-project.org/web/packages/corrplot/index.html) within R (see **Supplemental Figure 2**).

388

389 **Genotyping and Imputation**

390 METSIM participant samples were genotyped on the HumanOmniExpresss-12v1_C
391 BeadChip (OmniExpress) and Infinium HumanExome-12 v1.0 BeadChip (Exome Chip)
392 platforms. Quality controls included sample-level controls for sex and relatedness
393 confirmation, sample duplication, and detection of sample genetic ancestry outliers using
394 principal component analysis. Based on these quality control measures, we removed 14
395 samples with sex chromosome anomalies, 18 with evidence of participant duplication, 12
396 population outliers, and 9 samples with non-Mendelian inheritance inconsistencies. In
397 addition, we removed one individual from each of seven monozygotic twin pairs.

398

399 We filtered variants with low mapping quality of probes to genome build GRCh37, low
400 genotype completeness (<95% and <98% for the OmniExpress and ExomeChip,
401 respectively), or Hardy-Weinberg equilibrium $P<10^{-6}$.

402

403 We phased OmniExpress variants passing quality control with SHAPEIT v2 (43) and
404 imputed them using minimac v2 (44). For imputation, we used a reference panel of 20.9M
405 variants from the GoT2D study (including SNVs, indels, and large deletions) based on the
406 whole genome sequence of 2,874 Europeans, including 1,004 Finnish individuals – the
407 largest panel of Finnish genomes available (45). Following imputation, variants directly
408 genotyped on the ExomeChip were added. In cases of common markers between
409 imputed and genotyped variants, we used the directly genotyped call from the ExomeChip.
410 The distribution of imputation quality and MAF for each imputed variants are presented in

411 **Supplemental Figure 3**. We carried forward 16,607,533 variants with high imputation
412 quality (i.e. minimac RSQ ≥ 0.3) for further single variant association testing.
413
414 **Single Variant Analyses**
415 We performed single variant association tests on imputed genotype dosages for all
416 variants with a minor allele count ≥ 3. Association tests assumed an additive genotype
417 model and accounted for cryptic relatedness among the Finnish population using the
418 EMMAX linear mixed model approach (46), as implemented in EPACTS
419 (http://genome.sph.umich.edu/wiki/EPACTS). We adjusted amino acid traits for age, age^2,
420 and BMI, and then inverse normalized the residuals. We applied normalization of trait
421 levels to control for type-I error caused by skewed distributions, although this
422 normalization may reduce power to discover associated variants. We created association
423 plots for the novel variants using LocusZoom
424 (http://locuszoom.sph.umich.edu/locuszoom/). In addition, we computed *P* values (P_{ACT})
425 for the novel variants after correcting for the nine correlated amino acid traits (25). We
426 used a conventional significance threshold of $P < 5\times10^{-8}$ in single variant association
427 testing.
428
429 **Replication in the Northern Finnish Birth Cohort**
430 The associations in the novel regions were replicated *in silico* in the Northern part of
431 Finland: The 1966 cohort (the “Northern Finnish Birth Cohort”, or NFBC1966) (47).
432 NFBC1966 is a prospective follow-up study of children from the two northernmost
433 provinces of Finland born in 1966. All individuals still living in northern Finland or the

Helsinki area (n=8,463) were contacted and invited for clinical examination. A total of 6,007 participants attended the clinical examination at the participants' age of 31 years. Among them, 5,402 samples were genotyped on Illumina HumanCNV370DUO Analysis BeadChip (48), and were then imputed to the Haplotype Reference Consortium (HRC) reference (49) and 1000 Genomes Project Phase 3 (50) on the Michigan Imputation Server (https://imputationserver.sph.umich.edu/index.html). The association for the novel region variants and rare variants were looked up in 2,591 samples. Given our focused hypothesis, we set a threshold for significance in replication as $P \leq 0.05$.

Associations of Novel Amino Acid SNVs with End-organ Phenotypes

We investigated the association of the novel amino acid regions variants with the risk of type 2 diabetes, Alzheimer's diseases, and liver disease. For type 2 diabetes, we used publically available data in large-scale Europeans (N = 159,208) from the DIAGRAM consortium (http://www.diagram-consortium.org) (27). For Alzheimer's disease, we examined associations in the ADGC consortium (N = 54,162) (51). Finally, for liver disease, we used association summary statistics data the GOLD consortium (N = 7,176) (28). We used proxy single nucleotide polymorphisms (SNPs) tightly linking with the novel variant if our variant was not available.

Analysis of Amino Acid Trait Variance

We estimated the phenotypic variance explained by genetic variants for inverse normalized amino acid traits as previously described through GCTA v1.26 (see **Supplemental Figure 1**) (52). We removed 1,153 close relatives through kinship cutoff

of 0.0075 in KING 1.4 (53), and then estimated the phenotypic variance in 7,392 unrelated samples (54). To account for the effect of population structure, we used the top ten principal components as covariates. In brief, we carried out a primary analysis that consisted of a simultaneous analysis of all 16.6M variants, and a secondary analysis considering only the 2,580 variants determined to be genome-wide significant for at least one amino acid trait (see **Supplemental Table 2**).

Validation of Imputed Rare Variants

We used TaqMan SNP genotyping (Thermo Fisher Scientific) or Sanger sequencing to validate genotypes at three trait-associated ($P<5\times10^{-8}$) and rare imputed variants (MAF<0.5%) (see **Supplemental Table 5**). We examined all individuals predicted (on the basis of imputation) to be heterozygous carriers at any of the three sites, as well as additional non-carriers.

Genome-Wide Conditional Analyses

To identify additional independent genetic signals for amino acid traits at known GWAS loci, we conducted a comprehensive genome-wide conditional association analysis. We curated a database of prior published studies of genetic associations with amino acid and related traits to identify distinct variant associations in conditional analyses. To identify published studies, we screened a GWAS catalogue (http://www.ebi.ac.uk/gwas/), used SNIPPER (https://csg.sph.umich.edu/boehnke/snipper/) to query publicly available databases for published variants and loci, and performed literature review using PubMed (https://www.ncbi.nlm.nih.gov/pubmed/), and Google Scholar

480 (https://scholar.google.com). We focused on proteinogenic amino acid and related traits
481 (e.g., citrate) in European populations (see **Supplemental Table 7**).
482
483 The curated list contained 2,615 variants (of which 1,519 were unique, with several
484 variants having multiple trait associations) spanning >100 loci from 14 studies (see
485 **Supplemental Table 7**). These associated variants were then filtered for pairwise LD
486 (r^2>0.95) to 408 variants (see **Supplemental Table 7**). For the 2,580 amino acid
487 associated variants identified in discovery analyses (see **Supplemental Table 2**), we
488 performed a secondary analysis conditioning on the LD-pruned list of 408 independent
489 genetic variants. A variant with *P* value < 5×10^{-8} was considered to be a novel secondary
490 signal within known amino acid traits loci after conditioning on these 408 independent
491 genetic variants.
492
493 **Amino Acid Ratio Tests of Association**
494 Prior investigations of genetic associations with amino acid trait variation have reported
495 extensive findings with amino acid ratios (8, 12). While amino acid ratios were not the
496 focus of our investigation, we included discovery analyses with 36 amino acid ratios listed
497 in Supplemental Table 10.
498
499 **Gene-Based Tests of Association**
500 We performed gene-based tests of association using SKAT-O (21) with EMMAX (46) to
501 determine the joint contribution of protein-truncating (i.e. nonsense, frameshift, and
502 essential splice variants) and missense variants with MAF<1% on amino acid traits, as

described in our previous study (55). For these analyses, we included only coding variants directly genotyped on either the OmniExpress or Exome array. Missing genotype data (proportion < 2%) were imputed with variant-specific mean genotype since SKAT-O requires complete data (21). A total of 51,898 protein-truncating or missense variants in 13,996 genes met these criteria (see **Supplemental Table 11** for a summary of variant distributions within genes). We considered a gene-based result exome-wide significant at a p-value threshold of 2.5×10^{-6} (0.05/20,000) to account for the number of genes in these gene-based analyses.

Conflicts of Interest: A.J.K. and P.S. are shareholders of Brainshake Ltd., a company offering NMR-based metabolite profiling. A.J.K. and P.S. report employment relation for Brainshake Ltd. All other authors report no conflicts of interest.

Acknowledgements: We thank the participants of the METSIM study. We thank Seunggeun Lee and Hyun-Min Kang for their expertise and consultation. Data on type 2 diabetes have been contributed by DIAGRAM investigators and have been downloaded from diagram-consortium.org. We acknowledge the contribution of Alzheimer's Disease Genetic Consortium (ADGC) and Genetics of Obesity-related Liver Disease Consortium (GOLD) to the summary statistics for Alzheimer's diseases and Nonalcoholic fatty liver disease, respectively. This study was supported by Academy of Finland grants 77299, 124243, and 141226 (M.L.); the Finnish Heart Foundation (M.L.); the Finnish Diabetes Foundation (M.L.); the Juselius Foundation (ML); the Commission of the European Community HEALTH-F2-2007-201681 (M.L.); National Institutes of Health grants R01DK093757 (K.L.M.), R01DK072193 (K.L.M.), U01DK105561 (K.L.M.), R01DK062370 (M.B.), T32 HL129982 (J.P.D.), and T32 GM067553 (C.K.R.); National Human Genome Research Institute Division of Intramural Research project number Z01HG000024 (F.S.C.) and American Heart Association 16POST27250048 (D.S.K.). M.A.K. has been supported by the Sigrid Juselius Foundation and the Strategic Research Funding from the University of Oulu. M.A.K. works in a Unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_1201/1). NFBC1966 received financial support from the Academy of Finland (project grants 104781, 120315, 129269, 1114194, 24300796, Center of Excellence in Complex Disease Genetics and SALVE), University

Hospital Oulu, Biocenter, University of Oulu, Finland (75617), NHLBI grant 5R01HL087679-02 through the STAMPEED program (1RL1MH083268-01), NIH/NIMH (5R01MH63706:02), ENGAGE project and grant agreement HEALTH-F4-2007-201413, EU FP7 EurHEALTHAgeing -277849, the Medical Research Council, UK (G0500539, G0600705, G1002319, PrevMetSyn/SALVE) and the MRC, Centenary Early Career Award. The program is currently being funded by the H2020-633595 DynaHEALTH action, academy of Finland EGEA-project (285547) and EU H2020 ALEC project (Grant Agreement 633212).

REFERENCES

1. McBride,K.L., Belmont,J.W., O'Brien,W.E., Amin,T.J., Carter,S. and Lee,B.H. (2007) Heritability of plasma amino acid levels in different nutritional states. *Mol. Genet. Metab.*, **90**, 217–220.

2. Stančáková,A., Civelek,M., Saleem,N.K., Soininen,P., Kangas,A.J., Cederberg,H., Paananen,J., Pihlajamäki,J., Bonnycastle,L.L., Morken,M.A., *et al.* (2012) Hyperglycemia and a common variant of GCKR are associated with the levels of eight amino acids in 9,369 Finnish men. *Diabetes*, **61**, 1895–1902.

3. Würtz,P., Mäkinen,V.-P., Soininen,P., Kangas,A.J., Tukiainen,T., Kettunen,J., Savolainen,M.J., Tammelin,T., Viikari,J.S., Rönnemaa,T., *et al.* (2012) Metabolic signatures of insulin resistance in 7,098 young adults. *Diabetes*, **61**, 1372–1380.

4. Würtz,P., Soininen,P., Kangas,A.J., Rönnemaa,T., Lehtimäki,T., Kähönen,M., Viikari,J.S., Raitakari,O.T. and Ala-Korpela,M. (2013) Branched-chain and aromatic amino acids are predictors of insulin resistance in young adults. *Diabetes Care*, **36**, 648–655.

5. Tajiri,K. and Shimizu,Y. (2013) Branched-chain amino acids in liver diseases. *World J. Gastroenterol.*, **19**, 7620–7629.

6. Kan,M.J., Lee,J.E., Wilson,J.G., Everhart,A.L., Brown,C.M., Hoofnagle,A.N., Jansen,M., Vitek,M.P., Gunn,M.D. and Colton,C.A. (2015) Arginine deprivation and immune suppression in a mouse model of Alzheimer's disease. *J. Neurosci. Off. J. Soc. Neurosci.*, **35**, 5969–5982.

7. Suhre,K., Shin,S.-Y., Petersen,A.-K., Mohney,R.P., Meredith,D., Wägele,B., Altmaier,E., CARDIoGRAM, Deloukas,P., Erdmann,J., *et al.* (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature*, **477**, 54–60.

8. Kettunen,J., Tukiainen,T., Sarin,A.-P., Ortega-Alonso,A., Tikkanen,E., Lyytikäinen,L.-P., Kangas,A.J., Soininen,P., Würtz,P., Silander,K., *et al.* (2012) Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.*, **44**, 269–276.

9. Krumsiek,J., Suhre,K., Evans,A.M., Mitchell,M.W., Mohney,R.P., Milburn,M.V., Wägele,B., Römisch-Margl,W., Illig,T., Adamski,J., *et al.* (2012) Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet.*, **8**, e1003005.

10. Tukiainen,T., Kettunen,J., Kangas,A.J., Lyytikäinen,L.-P., Soininen,P., Sarin,A.-P., Tikkanen,E., O'Reilly,P.F., Savolainen,M.J., Kaski,K., *et al.* (2012) Detailed metabolic and genetic characterization reveals new associations for 30 known lipid loci. *Hum. Mol. Genet.*, **21**, 1444–1455.

11. Rhee,E.P., Ho,J.E., Chen,M.-H., Shen,D., Cheng,S., Larson,M.G., Ghorbani,A., Shi,X., Helenius,I.T., O'Donnell,C.J., *et al.* (2013) A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.*, **18**, 130–143.

12. Shin,S.-Y., Fauman,E.B., Petersen,A.-K., Krumsiek,J., Santos,R., Huang,J., Arnold,M., Erte,I., Forgetta,V., Yang,T.-P., *et al.* (2014) An atlas of genetic influences on human blood metabolites. *Nat. Genet.*, **46**, 543–550.

13. Demirkan,A., Henneman,P., Verhoeven,A., Dharuri,H., Amin,N., van Klinken,J.B., Karssen,L.C., de Vries,B., Meissner,A., Göraler,S., *et al.* (2015) Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet.*, **11**, e1004835.

14. Draisma,H.H.M., Pool,R., Kobl,M., Jansen,R., Petersen,A.-K., Vaarhorst,A.A.M., Yet,I., Haller,T., Demirkan,A., Esko,T., *et al.* (2015) Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels. *Nat. Commun.*, **6**, 7208.

15. Raffler,J., Friedrich,N., Arnold,M., Kacprowski,T., Rueedi,R., Altmaier,E., Bergmann,S., Budde,K., Gieger,C., Homuth,G., *et al.* (2015) Genome-Wide Association Study with Targeted and Non-targeted NMR Metabolomics Identifies 15 Novel Loci of Urinary Human Metabolic Individuality. *PLoS Genet.*, **11**, e1005487.

16. Kettunen,J., Demirkan,A., Würtz,P., Draisma,H.H.M., Haller,T., Rawal,R., Vaarhorst,A., Kangas,A.J., Lyytikäinen,L.-P., Pirinen,M., *et al.* (2016) Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat. Commun.*, **7**, 11122.

17. Rhee,E.P., Yang,Q., Yu,B., Liu,X., Cheng,S., Deik,A., Pierce,K.A., Bullock,K., Ho,J.E., Levy,D., *et al.* (2016) An exome array study of the plasma metabolome. *Nat. Commun.*, **7**, 12360.

18. Yet,I., Menni,C., Shin,S.-Y., Mangino,M., Soranzo,N., Adamski,J., Suhre,K., Spector,T.D., Kastenmüller,G. and Bell,J.T. (2016) Genetic Influences on Metabolite Levels: A Comparison across Metabolomic Platforms. *PloS One*, **11**, e0153672.

19. Long,T., Hicks,M., Yu,H.-C., Biggs,W.H., Kirkness,E.F., Menni,C., Zierer,J., Small,K.S., Mangino,M., Messier,H., *et al.* (2017) Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.*, **49**, 568–578.

20. Demirkan,A., van Duijn,C.M., Ugocsai,P., Isaacs,A., Pramstaller,P.P., Liebisch,G., Wilson,J.F., Johansson,A., Rudan,I., Aulchenko,Y.S., *et al.* (2012) Genome-Wide Association Study Identifies Novel Loci Associated with Circulating Phospho- and Sphingolipid Concentrations. *PLoS Genet.*, **8**, e1002490-14.

21. Lee,S., Abecasis,G.R., Boehnke,M. and Lin,X. (2014) Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, **95**, 5–23.

22. Kathiresan,S., Willer,C.J., Peloso,G.M., Demissie,S., Musunuru,K., Schadt,E.E., Kaplan,L., Bennett,D., Li,Y., Tanaka,T., *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56–65.

23. Teslovich,T.M., Musunuru,K., Smith,A.V., Edmondson,A.C., Stylianou,I.M., Koseki,M., Pirruccello,J.P., Ripatti,S., Chasman,D.I., Willer,C.J., *et al.* (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.

24. Consortium,G.L.G., Willer,C.J., Schmidt,E.M., Sengupta,S., Peloso,G.M., Gustafsson,S., Kanoni,S., Ganna,A., Chen,J., Buchkovich,M.L., *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.

25. Conneely,K.N. and Boehnke,M. (2007) So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am. J. Hum. Genet.*, **81**, 1158–1168.

26. Lange,L.A., Hu,Y., Zhang,H., Xue,C., Schmidt,E.M., Tang,Z.-Z., Bizon,C., Lange,E.M., Smith,J.D., Turner,E.H., *et al.* (2014) Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.*, **94**, 233–245.

27. Scott,R.A., Scott,L.J., Mägi,R., Marullo,L., Gaulton,K.J., Kaakinen,M., Pervjakova,N., Pers,T.H., Johnson,A.D., Eicher,J.D., *et al.* (2017) An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. *Diabetes*, **66**, 2888–2902.

28. Speliotes,E.K., Yerges-Armstrong,L.M., Wu,J., Hernaez,R., Kim,L.J., Palmer,C.D., Gudnason,V., Eiriksdottir,G., Garcia,M.E., Launer,L.J., *et al.* (2011) Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.*, **7**, e1001324.

29. McCarty,M.F. and DiNicolantonio,J.J. (2014) The cardiometabolic benefits of glycine: Is glycine an ‘antidote’ to dietary fructose? *Open Heart*, **1**, e000103.

30. Kim,D.S., Jackson,A.U., Li,Y.K., Stringham,H.M., FinMetSeq Investigators, Kuusisto,J., Kangas,A.J., Soininen,P., Ala-Korpela,M., Burant,C.F., *et al.* (2017) Novel association of TM6SF2 rs58542926 genotype with increased serum tyrosine levels and decreased apoB-100 particles in Finns. *J. Lipid Res.*, **58**, 1471–1481.

31. Iatan,I., Choi,H.Y., Ruel,I., Reddy,M.V.P.L., Kil,H., Lee,J., Odeh,M.A., Salah,Z., Abu-Remaileh,M., Weissglas-Volkov,D., *et al.* (2014) The WWOX gene

modulates high-density lipoprotein and lipid metabolism. *Circ. Cardiovasc. Genet.*, **7**, 491–504.

32. Aqeilan,R.I., Hassan,M.Q., de Bruin,A., Hagan,J.P., Volinia,S., Palumbo,T., Hussain,S., Lee,S.-H., Gaur,T., Stein,G.S., *et al.* (2008) The WWOX tumor suppressor is essential for postnatal survival and normal bone metabolism. *J. Biol. Chem.*, **283**, 21629–21639.

33. Felig,P., Marliss,E. and Cahill,G.F.J. (1969) Plasma Amino Acid Levels and Insulin Secretion in Obesity. *N. Engl. J. Med.*, **281**, 811–816.

34. Yeh,G.C., Harris,S.C. and Phang,J.M. (1981) Pyrroline-5-carboxylate reductase in human erythrocytes. *J. Clin. Invest.*, **67**, 1042–1046.

35. Guernsey,D.L., Jiang,H., Evans,S.C., Ferguson,M., Matsuoka,M., Nightingale,M., Rideout,A.L., Provost,S., Bedard,K., Orr,A., *et al.* (2009) Mutation in pyrroline-5-carboxylate reductase 1 gene in families with cutis laxa type 2. *Am. J. Hum. Genet.*, **85**, 120–129.

36. Reversade,B., Escande-Beillard,N., Dimopoulou,A., Fischer,B., Chng,S.C., Li,Y., Shboul,M., Tham,P.-Y., Kayserili,H., Al-Gazali,L., *et al.* (2009) Mutations in PYCR1 cause cutis laxa with progeroid features. *Nat. Genet.*, **41**, 1016–1021.

37. Berg,J.M., Tymoczko,J.L. and Stryer,L. (2002) Biochemistry 5th edition. W H Freeman, New York.

38. Wu,J.-Y., Kao,H.-J., Li,S.-C., Stevens,R., Hillman,S., Millington,D. and Chen,Y.-T. (2004) ENU mutagenesis identifies mice with mitochondrial branched-chain aminotransferase deficiency resembling human maple syrup urine disease. *J. Clin. Invest.*, **113**, 434–440.

39. Wang,X.L., Li,C.J., Xing,Y., Yang,Y.H. and Jia,J.P. (2015) Hypervalinemia and hyperleucine-isoleucinemia caused by mutations in the branched-chain-amino-acid aminotransferase gene. *J. Inherit. Metab. Dis.*, **38**, 855–861.

40. Soininen,P., Kangas,A.J., Würtz,P., Tukiainen,T., Tynkkynen,T., Laatikainen,R., Jarvelin,M.-R., Kähönen,M., Lehtimäki,T., Viikari,J., *et al.* (2009) High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *The Analyst*, **134**, 1781–1785.

41. Soininen,P., Kangas,A.J., Würtz,P., Suna,T. and Ala-Korpela,M. (2015) Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.*, **8**, 192–206.

42. Soininen,P., Haarala,J., Vepsäläinen,J., Niemitz,M. and Laatikainen,R. (2005) Strategies for organic impurity quantification by 1H NMR spectroscopy: Constrained total-line-shape fitting. *Anal. Chim. Acta*, **542**, 178–185.

693 43. Delaneau,O., Zagury,J.-F. and Marchini,J. (2013) Improved whole-chromosome
694 phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5–6.

695 44. Fuchsberger,C., Abecasis,G.R. and Hinds,D.A. (2015) minimac2: faster genotype
696 imputation. *Bioinformatics*, **31**, 782–784.

697 45. Fuchsberger,C., Flannick,J., Teslovich,T.M., Mahajan,A., Agarwala,V., Gaulton,K.J.,
698 Ma,C., Fontanillas,P., Moutsianas,L., McCarthy,D.J., *et al.* (2016) The genetic
699 architecture of type 2 diabetes. *Nature*, **536**, 41–47.

700 46. Kang,H.M., Sul,J.H., Service,S.K., Zaitlen,N.A., Kong,S., Freimer,N.B., Sabatti,C.
701 and Eskin,E. (2010) Variance component model to account for sample structure
702 in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.

703 47. Rantakallio,P. (1988) The longitudinal study of the northern Finland birth cohort of
704 1966. *Paediatr. Perinat. Epidemiol.*, **2**, 59–88.

705 48. Sovio,U., Bennett,A.J., Millwood,I.Y., Molitor,J., O'Reilly,P.F., Timpson,N.J.,
706 Kaakinen,M., Laitinen,J., Haukka,J., Pillas,D., *et al.* (2009) Genetic determinants
707 of height growth assessed longitudinally from infancy to adulthood in the northern
708 Finland birth cohort 1966. *PLoS Genet.*, **5**, e1000409.

709 49. McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A.,
710 Kang,H.M., Fuchsberger,C., Danecek,P., Sharp,K., *et al.* (2016) A reference
711 panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.

712 50. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J.,
713 Zhang,Y., Ye,K., Jun,G., Fritz,M.H.-Y., *et al.* (2015) An integrated map of
714 structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.

715 51. Lambert,J.C., Ibrahim-Verbaas,C.A., Harold,D., Naj,A.C., Sims,R., Bellenguez,C.,
716 DeStafano,A.L., Bis,J.C., Beecham,G.W., Grenier-Boley,B., *et al.* (2013) Meta-
717 analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's
718 disease. *Nat. Genet.*, **45**, 1452–1458.

719 52. Yang,J., Benyamin,B., McEvoy,B.P., Gordon,S., Henders,A.K., Nyholt,D.R.,
720 Madden,P.A., Heath,A.C., Martin,N.G., Montgomery,G.W., *et al.* (2010) Common
721 SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*,
722 **42**, 565–569.

723 53. Manichaikul,A., Mychaleckyj,J.C., Rich,S.S., Daly,K., Sale,M. and Chen,W.-M.
724 (2010) Robust relationship inference in genome-wide association studies.
725 *Bioinforma. Oxf. Engl.*, **26**, 2867–2873.

726 54. Manichaikul,A., Mychaleckyj,J.C., Rich,S.S., Daly,K., Sale,M. and Chen,W.-M.
727 (2010) Robust relationship inference in genome-wide association studies.
728 *Bioinforma. Oxf. Engl.*, **26**, 2867–2873.

729 55. Davis,J.P., Huyghe,J.R., Locke,A.E., Jackson,A.U., Sim,X., Stringham,H.M.,
730 Teslovich,T.M., Welch,R.P., Fuchsberger,C., Narisu,N., *et al.* (2017) Common,
731 low-frequency, and rare genetic variants associated with lipoprotein subclasses
732 and triglyceride measures in Finnish men from the METSIM study. *PLoS Genet.*,
733 **13**, e1007079.

734

FIGURE LEGENDS

Figure 1. Relationship between minor allele frequency and estimated beta coefficient (β) for loci associated with amino acid levels in the METSIM data. All novel amino acid loci (triangles in pink) are highlighted, in addition to novel signals at known amino acid loci (squares in green) identified through analyses conditioned on all known amino acid GWAS variants. The known amino acid signals are represented with blue circles. Gray dashed line represents a fractional polynomial spline fitted to the data points (P<0.001). β, on the y-axis, is the absolute value of the estimated regression coefficient for a given variant-trait association.

Figure 2. Plots show the trait values of rare variant carriers relative to the distribution of amino acid levels in all individuals. The tables in the right panel show gene-based tests of association with amino acid levels for genes *PYCR1* and *BCAT2*. Histograms show the distribution of the inverse normalized residuals of the trait across all participants for the gene-based test of association at (a) *PYCR1* with glycine levels and (b) *BCAT2* with valine. The dashed gray line represents the mean inverse normalized residual of trait level for all individuals. The solid black line in each row represents the mean trait level for carriers of each variant. Triangles represent rare variant carriers. The locations of triangles relative to the distribution across all participants indicate the trait levels of rare variant carriers. No individuals were homozygous for the minor allele of any of the listed variants.

SUPPLEMENTAL DATA

Seven Supplemental Figures are available online as a single PDF file. Twelve Supplemental Tables are available in a single Excel file.

Table 1. Novel and known single variants associations with amino acid traits.

Lead Variant	Trait	Chr:Pos [a]	Variant Annotation	METSIM							NFBC1966		
				MAF (%) (Allele)	β (SE)	Var%	*P*	GWAS SNV [b]	GWAS Trait	$P_{conditional}$ [c]	MAF (%)	β (SE)	*P*
Novel Single Variant Associations with Amino Acid Traits													
rs9987289	Gly	8:9183358	*LOC157273* intronic	17.0 (A)	-0.22 (0.02)	1.31	2.3×10^{-26}	-	-	2.3×10^{-25}	13.5	-0.15 (0.04)	7.3×10^{-4}
16:73326579 [d]	Ala	16:73326579	267 kb upstream of *ZFHX3*	0.42 (T)	0.76 (0.13)	0.41	3.6×10^{-9}	-	-	1.3×10^{-6}	0.7	-0.27 (0.17)	0.10
rs10468017	Ala	15:58678512	45 kb upstream of *LIPC*	33.2 (T)	0.09 (0.02)	0.37	1.5×10^{-8}	-	-	1.2×10^{-3}	33.2	0.08 (0.03)	7.7×10^{-3}
rs9937914	Ala	16:78422354	*WWOX* intronic	1.47 (G)	0.36 (0.07)	0.35	3.8×10^{-8}	-	-	1.4×10^{-7}	1.3	0.14 (0.13)	0.29
rs28601761	Tyr	8:126500031	49 kb downstream of *TRIB1*	42.2 (G)	-0.09 (0.02)	0.39	8.8×10^{-9}	-	-	1.5×10^{-7}	40.0	-0.02 (0.03)	0.52
Novel Single Variant Signals at Established Amino Acid Loci													
rs138640017	Gly	9:6533092	*GLDC* Q996H	0.95 (G)	1.35 (0.08)	3.35	3.5×10^{-65}	rs140348140	Gly (16)	5.8×10^{-40}	1.02	0.94 (0.15)	2.7×10^{-10}
rs141635447 [d]	His	12:96374381	*HAL* G283V	0.46 (A)	0.85 (0.12)	0.62	2.5×10^{-13}	rs7954638	His (16)	9.4×10^{-11}	0.35	1.04 (0.24)	1.7×10^{-5}
rs6564825	Gly	16:81153894	*PKD1L2* intronic	11.7 (G)	0.17 (0.02)	0.57	3.1×10^{-12}	rs74249229	Gly (16)	2.0×10^{-10}	10.4	-0.01 (0.05)	0.87
rs190671241	Phe	4:73574142	139 kb upstream of *ADAMTS3*	1.70 (G)	0.36 (0.06)	0.42	2.4×10^{-9}	4:73542640	His/Phe (8)	2.2×10^{-9}	2.10	0.39 (0.10)	1.6×10^{-4}
rs112981908	Gly	3:125858480	*ALDH1L1* intronic	11.9 (G)	-0.14 (0.02)	0.37	1.8×10^{-8}	rs1107366	Gly (16)	3.1×10^{-10}	e	e	e

Abbreviations: β (SE) – estimated regression coefficient and standard error for the minor allele; Chr/Pos – variant chromosome and position based on hg19 build; MAF% (Allele) - minor allele frequency (in percent) with minor allele in parentheses; SNV – single nucleotide variant; Var% – trait variance explained by variant (in percent)

Amino acid abbreviations: Ala – alanine; Gly – glycine; His – histidine; Phe – phenylalanine; Tyr – tyrosine.

[a] Position based on hg19 build.

[b] Lead GWAS SNV within 1 Mb of the identified lead SNV.

[c] $P_{conditional}$ values result from conditional analyses adjusting for known amino acid signals from previous studies (see **Supplemental Table 7**)

[d] These two variants were directly genotyped for validation and had 100% concordance with the imputed genotype (see text).

[e] This variant was not in the HRC panel used for analyses in the NFBC1966 dataset.

Table 2. Novel gene-based associations with amino acid traits.

Trait	Gene	rsID	Chr:Pos [a]	Annotation	METSIM						NFBC1966		
					MAF (%) (Allele)	Genotype Counts	β	Variant *P*	Gene *P*		MAF (%)	β	Variant *P*
Novel Gene-Based Associations with Amino Acid Traits													
Gly	*PYCR1*								1.5×10^{-6}		-	-	-
		rs3744807	17:79890818	G297R	0.13(T)	8532 / 22 / 0	-1.02	1.62×10^{-6}			0.01	1.2	0.23
		rs142225075	17:79893020	L108V	0.10(C)	8528 / 17 / 0	-0.42	0.08			0.09	-0.5	0.30
Val	*BCAT2*								7.4×10^{-7}		-	-	-
		rs199999090	19:49299714	R331C	0.18(A)	8515 / 30 / 0	0.63	5.36×10^{-4}			-	-	-
		rs117048185	19:49309776	Q60E	0.59(C)	8445 / 100 / 0	0.36	4.12×10^{-4}			0.71	0.6	0.0003
		rs201148940	19:49309937	H6R	0.02(C)	8542 / 3 / 0	0.42	0.46			monomorphic		
Novel Gene-Based Associations with Amino Acid Traits, Established Association with Amino Acid Ratios													
Ala	*DHODH* [b]								1.36×10^{-7}		-	-	-
		rs201970636	16:72055088	A195T	0.41(A)	8475 / 70 / 0	0.64	1.94×10^{-7}			0.76	0.3	0.047
		rs201202896	16:72055187	E228K	0.01(A)	8544 / 1 / 0	-0.80	0.42			-	-	-
		rs199626701	16:72057113	A290V	0.01(T)	8544 / 1 / 0	1.13	0.26			-	-	-
		rs200181357	16:72057134	R297H	0.01(A)	8543 / 2 / 0	-0.22	0.76			monomorphic		
		rs192923495	16:72057193	R317W	0.09(T)	8529 / 16 / 0	0.26	0.30			0.32	0.0	0.92

Abbreviations: β – estimated regression coefficient for the minor allele; Chr:Pos – variant chromosome and position based on hg19 build; MAF(%) – minor allele frequency (in percent) with minor allele in parentheses.

Amino acid abbreviations: Ala – alanine; Gly – glycine; His – histidine; Val – valine.

[a] Position based on hg19 build.

[b] SNPs in *DHODH* previously associated with Alanine-to-Tyrosine ratio by Kettunen *et al.*(8).