
Publisher's PDF, also known as Version of record

License (if available):
CC BY-NC

Link to published version (if available):
10.1161/JAHA.119.011922

Link to publication record in Explore Bristol Research
PDF-document

This is the final published version of the article (version of record). It first appeared online via the American Heart Association at https://doi.org/10.1161/JAHA.119.011922. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
Genetic Determinants of Circulating Glycine Levels and Risk of Coronary Artery Disease

Qiong Jia, MS; Yi Han, PhD; Pin Huang, BS; Nicholas C. Woodward, PhD; Janet Gukasyan, BS; Johannes Kettunen, PhD; Mika Ala-Korpela, PhD; Olga Anufrieva, MSc; Qin Wang, PhD; Markus Perola, MD, PhD; Olli Raitakari, MD, PhD; Terho Lehtimäki, MD, PhD; Jorma Vilkari, MD, PhD; Marjo-Riitta Järvelin, MD, PhD; Michael Boehnke, PhD; Markku Laakso, MD, PhD; Karen L. Mohlke, PhD; Oliver Fiehn, PhD; Zeneng Wang, PhD; W.H. Wilson Tang, MD; Stanley L. Hazen, MD, PhD; Jaana A. Hartiala, PhD; Hooman Allayee, PhD

Background—Recent studies have revealed sexually dimorphic associations between the carbamoyl-phosphate synthase 1 locus, intermediates of the metabolic pathway leading from choline to urea, and risk of coronary artery disease (CAD) in women. Based on evidence from the literature, the atheroprotective association with carbamoyl-phosphate synthase 1 could be mediated by the strong genetic effect of this locus on increased circulating glycine levels.

Methods and Results—We sought to identify additional genetic determinants of circulating glycine levels by carrying out a meta-analysis of genome-wide association study data in up to 30,118 subjects of European ancestry. Mendelian randomization and other analytical approaches were used to determine whether glycine-associated variants were associated with CAD and traditional risk factors. Twelve loci were significantly associated with circulating glycine levels, 7 of which were not previously known to be involved in glycine metabolism (ACADM, PHGDH, COX18-ADAMTS3, PSHP, TRIB1, PTPRD, and ABO). Glycine-raising alleles at several loci individually exhibited directionally consistent associations with decreased risk of CAD. However, these effects could not be attributed directly to glycine because of associations with other CAD-related traits. By comparison, genetic models that only included the 2 variants directly involved in glycine degradation and for which there were no other pleiotropic associations were not associated with risk of CAD or blood pressure, lipid levels, and obesity-related traits.

Conclusions—These results provide additional insight into the genetic architecture of glycine metabolism, but do not yield conclusive evidence for a causal relationship between circulating levels of this amino acid and risk of CAD in humans. (J Am Heart Assoc. 2019;8:e011922. DOI: 10.1161/JAHA.119.011922.)

Key Words: causality • coronary artery disease • genome-wide association study • glycine • Mendelian randomization • meta-analysis

Metabolites derived from gut microbiome and hepatic-mediated metabolism of dietary choline and L-carnitine, such as trimethylamine N-oxide and betaine, have recently been shown to be proatherogenic in mice and novel biomarkers of coronary artery disease (CAD) risk in humans.1–3 In searching for genetic determinants of these metabolites, we identified sexually dimorphic associations between the carbamoyl-phosphate synthase 1 (CPS1) locus and not only plasma trimethylamine N-oxide and betaine levels, but also other intermediates in the metabolic pathway leading from choline to urea.4 We and others further noted that, of the various other biomarkers/metabolites that had previously been linked to CAD,5–14 the strongest effect size and most significant association was with circulating glycine levels in women.4,15–17 Most important, the lead CPS1 variant also exhibited a strikingly significant female-specific association with decreased risk of CAD.4 However, the direction of the associations between CPS1 and the various biomarkers and metabolites was opposite to what would be expected for a variant that decreased risk of CAD.

One explanation for the protective association of CPS1 with CAD could be the strong genetic effect of this locus on increased circulating glycine levels.4 For example, previous in vitro and in vivo studies have shown that glycine reduces inflammation and oxidative stress in endothelial cells, activated macrophages, and other leukocytes.18–22 Furthermore, platelet...
aggregation of both human and rodent platelets can be prevented by glycine in a dose-dependent manner through mechanisms involving the glycine receptor. Interestingly, the same glycine-raising CPS1 variant has been associated with reduced platelet counts. Alternatively, glycine has been reported to have antihypertensive effects in mice and humans. A recent epidemiological study also demonstrated an inverse relationship between plasma glycine levels and risk of an acute myocardial infarction. Taken together, these observations support the concept that glycine could have atheroprotective properties, but direct evidence for a causal relationship between this amino acid and risk of CAD is lacking.

In the present study, we used a meta-analysis approach with genome-wide association study (GWAS) data to identify additional genetics determinants of circulating glycine levels. The identified loci were then used to investigate the possible causal association between circulating glycine levels and risk of CAD and traditional risk factors. In total, 12 loci were identified for circulating glycine levels, 7 of which were novel and not previously known to be involved in glycine metabolism. However, various analytical approaches with glycine-raising alleles at these loci did not provide conclusive evidence for a causal relationship between circulating glycine and risk of CAD in humans.

**Methods**

The statistical methods used in this study will be made available to other researchers for purposes of reproducing the results or replicating the analyses. The summary statistics of the meta-analysis for circulating glycine levels will be made available through the NHGRI-EBI Catalog of published GWASs (https://www.ebi.ac.uk/gwas/downloadstats/summary-statistics/).

**Study Populations**

The present analyses included 30,118 subjects of European ancestry from the GeneBank (GB), FINRISK 1997 and 2007 (FR97 and FR07), YFS (Cardiovascular Risk in Young Finns Study), NFBC1966 and NFBC1986 (Northern Finland Birth Cohort), and METSIM (Metabolic Syndrome in Men) studies. Details of subject recruitment and genotyping methodology for each cohort are provided in Data S1. For each cohort, written informed consent was obtained from all participants before being enrolled, and the studies were approved by the institutional review boards of the participating institutions. The present analysis was approved by the institutional review board of USC Keck School of Medicine.

**Measurement of Circulating Glycine Levels**

Glycine levels were quantified using stable isotope dilution high-performance liquid chromatography with online electrospray ionization tandem mass spectrometry in the GB study and by quantitative high-throughput NMR in the FR97, FR07, YFS NFBC66, NFBC86, and METSIM cohorts.

**Data Harmonization and GWAS Analyses**

Circulating glycine levels were first regressed on study-specific covariates chosen by the investigators of each cohort. These included age and sex in GeneBank; age, sex, and time from last meal in FR97, FR07, YFS NFBC66, and NFBC86; and age, age, and body mass index in METSIM. Inverse rank-based normal transformations were carried out on the residuals after adjustment for covariates and used as the outcome in GWAS analyses by linear regression in each study.

**Meta-Analysis for Circulating Glycine Levels**

We performed a fixed-effects meta-analysis for circulating glycine levels with 7,487,927 SNPs that were imputed using 1000 Genomes Project data and that were common to all data sets. This analysis was carried out assuming an additive model and after controlling for population structure within each study, as implemented in GWAMA (Genome-Wide Association Meta-Analysis) software. In addition to a combined meta-analysis with all subjects, we also carried out a sex-stratified fixed-effects meta-analysis. The genome-wide threshold for significant association was set at $P=5.0 \times 10^{-8}$. A locus was defined as novel if the lead single-nucleotide polymorphism (SNP) was in weak or no linkage disequilibrium ($r^2 \leq 0.1$) with variants at genome-wide significant loci previously reported for circulating glycine levels. Manhattan and quantile-quantile plots were constructed using the “qqman” package in R (R Foundation for Statistical Computing, Vienna, Austria). To examine whether...
all novel loci identified in our meta-analysis were also significantly associated with other traits (phenome-wide association studies), we used publicly available databases, such as PhenoScanner,\textsuperscript{35} the UCSC Genome Browser (https://genome.ucsc.edu/), and the GWAS Catalog (https://www.ebi.ac.uk/gwas/home). The significance threshold for phenome-wide association studies analyses was set to $P=5.0 \times 10^{-8}$ with a linkage disequilibrium cut off of $r^2 \geq 0.8$ for proxy SNPs.

Proportion of Phenotypic Variance Explained

The proportion of variation in glycine levels explained by the identified variants was estimated using SumHer software.\textsuperscript{36} SNP heritability was calculated using a weighted linkage disequilibrium adjusted kinships model with the 12 glycine-associated SNPs. 1000 Genomes Project–based imputed genotypes in $\approx 4500$ subjects of European ancestry from the GB cohort were used as a reference panel for linkage disequilibrium ($r^2$) for these estimates.

Analysis of Variants With Risk of CAD and Traditional Risk Factors

Publicly available summary results from large-scale GWAS in subjects of European ancestry\textsuperscript{37–39} were used to determine whether glycine-associated variants were associated with risk of CAD and various lipid-, metabolic-, and blood-pressure–related risk factors. Specifically, we tested associations using 3 analytical strategies with 4 genetic models that were based on various nested combinations of the 12 identified variants. Genetic model 1 included all 12 loci identified for glycine; model 2 was designed to specifically test only the 7 novel loci (ACADM, PHGDH, COX18-ADAMTS3, PSPH, TRIB1, PTPRD, and ABO); and model 3 included only the 4 loci known to be related to glycine metabolism (PSPH, PHGDH, GLDC, and GSH). Model 4 was the most restrictive and included only the 2 glycine-associated loci that are known to be directly involved in the catabolism of glycine through the glycine cleavage system (GLDC and GCSH) and that did not exhibit pleiotropic effects with other traits or metabolites. In the first analytical approach, the average/overall association of CAD and its risk factors with glycine-raising alleles in the 4 genetic models were evaluated by meta-analysis, as implemented in the “meta” R package (https://cran.r-project.org/web/packages/meta/index.html). In the second approach, we generated genetic risk scores (GRS) with the identified variants for the same 4 genetic models to evaluate the cumulative joint effects of glycine-raising alleles. Additive multi-SNP GRS associations were estimated using the grr.summary function of the “gtx: Genetic ToolboX” R package (https://cran.r-project.org/web/packages/gtx). This approach approximates the regression of an intermediate trait or biomarker onto a GRS, which is based on the weighted sum of the single SNP coefficients derived from the association summary statistics.\textsuperscript{40} For the third strategy, we carried out weighted median and inverse variance weighted Mendelian randomization (MR) analyses with the 4 genetic models, as implemented in the “TwoSampleMR” R package.\textsuperscript{41} Because the weighted median MR method requires 3 or more variants, only the inverse variance weighted MR test was used for determining association of the 2 SNPs in model 4 with CAD and traditional risk factors.

Results

GWAS for Circulating Glycine Levels

To identify novel loci for circulating glycine levels, we carried out a meta-analysis of GWAS summary-level data with 7 487 927 genotyped and imputed SNPs in 30 118 subjects of European ancestry. Table 1 shows the characteristics of the study cohorts and data sets used for these analyses. A GWAS was carried out for circulating glycine levels in each cohort, followed by a fixed-effects meta-analysis. The genomic control factor (lambda, $\lambda$) in GB I (0.995), GB II (0.989), and the combination of the FR97, FR07, YFS NFBC66, and NFBC86 cohorts (1.039), and METSIM (1.014) were small or modest, thus decreasing the likelihood of identifying spurious associations attributed to population stratification (Figure S1). To further account for this potential confounder, we also applied genomic control to each study before the meta-analysis. In total, 4934 variants distributed across 12 loci were associated with circulating glycine levels at the genomewide significance threshold ($P=5.0 \times 10^{-8}$; Figure 1, Table 2, and Table S1). Seven of these loci (ACADM, PHGDH, COX18-ADAMTS3, PSPH, TRIB1, PTPRD, and ABO) were novel and identified as being associated with circulating glycine levels for the first time herein (Figure 1, Table 2, and Figure S2). The other 5 loci (CPS1, ALDH1L1, PPP1R3B-LOC157273, GLDC, and GCSH) have previously been reported for circulating glycine levels, but the association signals became more significant in our meta-analysis because of increased sample size (Figure 1, Table 2, and Figure S2). Overall, the 12 identified loci explained $\approx 15\%$ of the variation in circulating glycine levels.

Based on previous observations that the CPS1 locus exhibited a pattern of sexually dimorphic associations with glycine, various other metabolites, and risk of CAD,\textsuperscript{4,17} we also carried out meta-analyses in men and women separately. Five and 9 regions were significantly associated with circulating glycine levels in females and males, respectively (Figures S3 and S4), all of which were also observed in the combined GWAS analysis with all subjects (Figure 1). With the exception of the previously observed stronger association
signal for glycine levels at the CPS1 locus in women ($\beta=0.572; P=1.0 \times 10^{-300}$) compared with men ($\beta=0.322; P=5.9 \times 10^{-189}$), the effect sizes at the remaining 11 loci were similar in males and females with no significant evidence for heterogeneity (Table S2 and Figure S5). We next carried out a phenome-wide association studies analysis based on publicly available data to determine whether any of the loci for glycine were associated with other traits. Six of the 12 loci (ACADM, CPS1, ALDH1L1, PPP1R3B-LOC157273, TRIB1, and ABO) exhibited pleiotropic associations with blood cell counts or lipid levels, some of which were even more significant than the association signals for glycine (Table S3). Two other loci (PSPH and PHGDH) had also been associated with serine and homocysteine levels, which are metabolites related to glycine metabolism (Table S3). However, no genome-wide significant associations have previously been reported for the 4 remaining loci (COX18-ADAMTS3, GLDC, PTPRD, and GCSH).

**Association of Loci for Circulating Glycine Levels With CAD and Traditional Risk Factors**

We next sought to evaluate association of loci for glycine levels with risk of CAD and traditional risk factors. Of the 12 regions identified, glycine-raising alleles of the lead variants at the CPS1, PSPH, TRIB1, and ABO loci individually yielded directionally consistent associations with decreased risk of CAD at the Bonferroni-corrected threshold of $P=4.2 \times 10^{-3}$ for testing 12 loci ($0.05/12$; Table S4). We next tested 4 genetic models based on various nested combinations of the 12 glycine loci for association with risk of CAD using 3 analytical strategies (details provided in Methods). Consistent with the individual SNP results, meta-analysis or GRS-based joint SNP

**Table 1. Description of Cohorts Used in Meta-Analysis for Circulating Glycine Levels**

<table>
<thead>
<tr>
<th>Cohort</th>
<th>No. of SNPs</th>
<th>N (Male/Female)</th>
<th>Metabolomics Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB I</td>
<td>8 986 545</td>
<td>391 (195/196)</td>
<td>HPLC-MS</td>
</tr>
<tr>
<td>GB II</td>
<td>8 986 545</td>
<td>885 (602/283)</td>
<td>HILIC-MS</td>
</tr>
<tr>
<td>FR97</td>
<td>11 512 433</td>
<td>6631 (3198/3433)</td>
<td>NMR</td>
</tr>
<tr>
<td>FR07</td>
<td>11 512 433</td>
<td>4124 (1860/2264)</td>
<td>NMR</td>
</tr>
<tr>
<td>YFS</td>
<td>11 512 433</td>
<td>1947 (1052/895)</td>
<td>NMR</td>
</tr>
<tr>
<td>NFBC66</td>
<td>11 512 433</td>
<td>4483 (2152/2331)</td>
<td>NMR</td>
</tr>
<tr>
<td>NFBC86</td>
<td>11 512 433</td>
<td>3112 (1508/1604)</td>
<td>NMR</td>
</tr>
<tr>
<td>METSIM</td>
<td>16 888 882</td>
<td>8545 (8545/0)</td>
<td>NMR</td>
</tr>
</tbody>
</table>

FR97 and FR07 indicates FINRISK; GB, GeneBank; HPLC-MS, high-performance liquid chromatography with mass spectrometry; METSIM, METabolic Syndrome In Men Study; NFBC, Northern Finland Birth Cohort; NMR, nuclear magnetic resonance; SNP, single-nucleotide polymorphism; YFS, Cardiovascular Risk in Young Finns.

**Figure 1.** Results of GWAS meta-analysis for circulating glycine levels. The Manhattan plot shows 7 novel significantly associated loci for circulating glycine levels (red dots) identified through meta-analyses of GWAS data from 30 118 subjects in the GeneBank, FR97, FR07, YFS, NFBC66, NFBC86, and METSIM cohorts. The 5 previously known loci are indicated by blue dots and all increased in significance in the meta-analysis. Genome-wide thresholds for significant ($P=5.0 \times 10^{-8}$) and suggestive ($P=5.0 \times 10^{-6}$) association are indicated by the horizontal red and dark blue lines, respectively. $P$ values are truncated at $-\log_{10} (P)=40$. FR97 and FR07 indicates FINRISK; GWAS, genome-wide association study; METSIM, METabolic Syndrome In Men Study; NFBC, Northern Finland Birth Cohort; YFS, Cardiovascular Risk in Young Finns.

DOI: 10.1161/JAHA.119.011922

Journal of the American Heart Association
effects analysis of glycine-raising alleles in all 4 genetic models yielded modest, but significant, associations (odds ratios, \( \approx 0.98 \)) with decreased risk of CAD (Figure 2). By comparison, weighted median and inverse variance weighted MR tests yielded much weaker or no evidence for a protective association of glycine-raising alleles with CAD, including the most restrictive model constructed with only variants at the 2 glycine cleavage system loci (Figure 2). We next evaluated whether loci for glycine were associated with blood pressure, lipid levels, and obesity-related traits using the same analytical strategies. Glycine-raising alleles at several loci (CPS1, PPP1R3B-LOC157273, TRIB1, and ABO) individually exhibited highly significant associations with decreased blood pressure and lipid levels (Table S5). The meta-analysis and GRS-based joint SNP effects analysis also provided evidence for similar associations with blood pressure and lipid levels, although these were only observed for the genetic models that included either all 12 glycine-associated loci or the 7 novel loci. However, the 2 MR analyses provided no evidence that glycine-raising alleles were causally associated with any of the selected traditional risk factors (Table S5).

**Discussion**

In the present study, we used a meta-analysis approach to identify 7 novel genomic regions associated with circulating glycine levels and strengthen the association signals at 5 previously known loci. Among all 12 loci, CPS1 and GLDC were the most strongly associated with glycine levels, with variants at the remaining 10 loci having anywhere between \( \approx 60\% \) and 90\% lower effect sizes. Furthermore, sex-stratified analyses confirmed the strong effect of CPS1 on glycine levels in women compared with men, but did not reveal sexually dimorphic associations with any of the remaining 11 loci. Follow-up analyses with the identified loci also yielded evidence that glycine could be causally associated with risk of CAD, although the biological mechanism(s) through which this effect occurs remains to be determined.

Based on what is known about amino acid metabolism, plausible biological links could be inferred between several of the newly identified loci and glycine levels. For example, PHGDH and PSPH encode phosphoglycerate dehydrogenase and phosphoserine phosphatase, which catalyze the first and last reactions, respectively, in the 3-step process leading to the synthesis of serine from 3-phosphoglycerate.47 Although the PHGDH and PSPH loci have both been strongly associated with circulating serine or homocysteine levels,8,15,43–47 they were not known to be associated with glycine levels before the results of our meta-analysis. Interestingly, serine can serve as a substrate for the synthesis of glycine in a reversible reaction catalyzed by SHMT,48 and glycine levels have been reported to be lower in humans deficient for PHGDH or PSPH.49–52 With respect to our results, the lead variant at PHDGH has yielded several highly significant \( P \) values ranging from \( \approx 10^{-10} \) to \( 10^{-34} \) cis expression quantitative trait loci where the glycine-raising allele of rs478093 (G) increases PHGDH mRNA levels.53 This would presumably lead to increased production of serine and,
by extension, glycine, thus providing a directionally consistent molecular mechanism for the observed association of the 
PHGDH locus with circulating glycine levels. However, even when taking into account previously identified associations at 
loci harboring enzymes involved in either glycine catabolism (GLDC, GCSH)\(^{53,54}\) or downstream detoxification through the urea cycle (CPS1),\(^{55,56}\) biological mechanisms for half of the loci associated with circulating glycine levels are not directly evident.

A primary goal of our study was to test whether glycine is a causal and protective biomarker of CAD risk. To address this question, we used the results of large GWAS meta-analyses to determine whether loci identified for glycine levels were associated with CAD and traditional risk factors. Glycine-raising alleles at 3 of the 7 novel loci (PSPH, TRIB1, and ABO) were individually associated with decreased risk of CAD at the Bonferroni-corrected significance threshold ($P=4.2 \times 10^{-3}$), of which TRIB1 and ABO had been identified as CAD susceptibility loci in previous GWASs.\(^{37,57,58}\) Rather than glycine levels, it is likely that association of TRIB1 and ABO with CAD is attributed to their stronger effect sizes on lipid levels and hematological parameters\(^{12,59,60}\) and, in the case of ABO, numerous other CAD-relevant traits.\(^{35}\) When all 12 loci or only the 7 novel loci were considered in combination, the meta-analyses and joint SNP effects analyses also revealed association of glycine-raising alleles with decreased risk of CAD. Because several of the loci included in these analyses (CPS1, PSPH, PPP1RB-LOC157273, TRIB1, and ABO) exhibited associations with other CAD-related traits, either individually or in various combinations, it was not possible based on these results alone to conclude that glycine is the causal biomarker driving the association of these loci with CAD. Therefore, we assessed causality more directly with 2 different MR tests, which provided little to no evidence that

---

**Figure 2.** Association of loci identified for circulating glycine levels with risk of CAD. Individual associations between glycine-raising alleles at each locus and risk of CAD are shown by blue squares in the forest plots. Purple diamonds indicate combined associations based on meta-analysis, joint SNP effects with a genetic risk score (GRS), and weighted median (WM) or inverse variance weighted (IVW) Mendelian randomization (MR) test. Model 1 included all 12 glycine-associated loci (A), model 2 included the 7 novel loci for glycine in this study (B), model 3 included the 4 loci known to be involved in glycine metabolism (C), and model 4 was constructed with only the 2 loci directly involved in the catabolism of glycine through the glycine cleavage complex (D). CAS indicates coronary artery disease; OR, odds ratio; SNP, single-nucleotide polymorphism.

---

DOI: 10.1161/JAHA.119.011922
glycine-raising alleles were associated with risk of CAD or lipid levels, blood pressure, and obesity-related traits. In this regard, the results of MR tests with the most restrictive genetic model that included only the 2 loci directly involved in glycine degradation (GLDC and GCSH) are particularly relevant. For example, no CAD-related traits, aside from glycine levels, are known to be associated with the GLDC and GCSH loci, thus satisfying the lack of pleiotropy as 1 of the major assumptions in MR analysis. Moreover, the glycine-raising alleles of rs71503800 at the GLDC locus and rs1047891 at the CPS1 locus have nearly equivalent effect sizes on circulating glycine levels. However, none of the analyses with rs71503800 at the GLDC locus yielded evidence for association of this variant with risk of CAD or traditional risk factors. Taken together, we conclude that evidence for a causal relationship between circulating glycine and risk of CAD is relatively weak and requires additional studies.

Whereas the present results have revealed novel genetic determinants of circulating glycine levels, our study should also be taken in the context of certain limitations. First, depending on the cohort, metabolomic analysis was carried out using different platforms and glycine was measured in either serum or plasma, some of which were not fasting samples. Although this may have led to identifying fewer significant associations for circulating glycine levels, our relatively large sample size in the meta-analysis still provided sufficient power to detect robust associations at several previously known loci and 7 novel genomic regions. Second, the sex-stratified analyses had approximately half the number of females than males, which likely decreased power to identify loci for circulating glycine levels that were either specific to, or more strongly associated in, 1 sex or the other. Third, all study subjects in our study were of European ancestry, and it is possible that the genetic association results for either circulating plasma glycine levels may not be generalizable to other populations. Last, our evaluation of the causal relationship between glycine and risk of CAD or traditional risk factors may have resulted in biased estimates because of pleiotropic effects, especially in models that included all 12 loci or the 7 newly identified SNPs, or because of weak instruments in nested models that included only the 4 or 2 loci directly involved in glycine metabolism.

In summary, the results of our study provide additional insight into the genetic architecture of glycine metabolism, but a more-complete understanding of the mechanisms through which some of these loci influence circulating levels remains to be determined. Despite these genetic findings, we did not obtain conclusive evidence for a causal relationship between glycine and risk of CAD, raising the possibility that another unknown metabolite or biological pathway is driving the protective association of glycine-raising alleles at the CPS1 locus with risk of CAD.

Acknowledgments
We gratefully acknowledge the subjects in each of the cohorts for their participation in these studies.

Sources of Funding
This work was supported, in part, by NIH grants R01HL133169, R01ES021801, R01ES025786, R01HL103866, P20HL113452, R01DK062370, and S100D016346 and Transatlantic Networks of Excellence Awards from Foundation Leducq. The GeneBank study was supported, in part, by NIH grants P01HL098055, P01HL076491, and R01HL103931. Mass Spectrometry instrumentation used for the GeneBank study was housed in a facility supported, in part, through a Shimadzu Center of Excellence Award. Kettunen is supported through funds from the Academy of Finland (Grant Nos. 297338 and 307247) and Novo Nordisk Foundation (Grant No. NNF17OC0026062). Ala-Korpela is supported by a Senior Research Fellowship from the National Health and Medical Research Council (NHMRC) of Australia (APP1158958) and also works in a unit that is supported by the University of Bristol and the UK Medical Research Council (MC_UU_12013/1). The Baker Institute is supported, in part, by the Victorian Government’s Operational Infrastructure Support Program.

Disclosures
Z. Wang and Hazen are named as co-inventors on pending and issued patents held by the Cleveland Clinic relating to cardiovascular diagnostics and therapeutics and have the right to receive royalty payment for inventions or discoveries related to cardiovascular diagnostics or therapeutics from Cleveland Heart Lab, Quest Diagnostics, and Procter & Gamble Company. Hazen also reports having been paid as a consultant from Procter & Gamble Company and having received research funds from Procter & Gamble Company and Roche. Kettunen reports owning a modest amount of stock options for Nightingale Health Ltd, a company offering metabolic profiling. The remaining authors have no disclosures to report.

Author’s Affiliations
From the Departments of Preventive Medicine (Q.J., Y.H., P.H., N.C.W., J.G., J.A.H., H.A.) and Biochemistry & Molecular Medicine (Q.J., Y.H., P.H., N.C.W., J.G., J.A.H., H.A.), Keck School of Medicine, University of Southern California, Los Angeles, CA; Xiangya School of Medicine, Central South University, Hunan, China (P.H.); Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland (J.K., M.A.-K., O.A., Q.W., M.-R.J.); National...


Genetics of Glycine and Coronary Artery Disease

13. Danik JS, Pare G, Chasman DI, Zee RY, Kwiatkowski DJ, Parker A, Miletich JP, DOI: 10.1161/JAHA.119.011922

Journal of the American Heart Association


19. Wheeler M, Stachlewitz RF, Yamashina S, Ikejima K, Morrow AL, Thurman RG. 2014;126:19


Supplemental Material
Data S1.

Detailed Description of Cohorts

GeneBank Study: The Cleveland Clinic GeneBank study is a single site sample repository generated from consecutive patients undergoing elective diagnostic coronary angiography or elective cardiac computed tomographic angiography with extensive clinical and laboratory characterization and longitudinal observation. Subject recruitment occurred between 2001 and 2006. Ethnicity was self-reported and information regarding demographics, medical history, and medication use was obtained by patient interviews and confirmed by chart reviews. All clinical outcome data were verified by source documentation. Coronary artery disease (CAD) was defined as adjudicated diagnoses of stable or unstable angina, myocardial infarction (MI) (adjudicated definition based on defined electrocardiographic changes or elevated cardiac enzymes), angiographic evidence of \( \geq 50\% \) stenosis of one or more major epicardial vessel, and/or a history of known CAD (documented MI, CAD, or history of revascularization). The GeneBank Study has been used previously for discovery and replication of novel genes and risk factors for atherosclerotic disease\(^1\text{-}\text{4}\). Plasma glycine levels were measured in blood samples obtained upon entry into GeneBank. Genome-wide genotyping was carried out on 3031 GeneBank subjects of European ancestry using the Affymetrix Genome-Wide Human Array 6.0 SNP chip. After conversion of genomic coordinates to GRCh37/hg19, exclusion of SNPs with duplicates, call rates <97%, minor allele frequencies (MAFs) <1%, and without chromosome and base pair position, and exclusion of 44 subjects with genotype call rates <90%, 642,766 were available for imputation in 2972 participants. Imputation was carried out on the forward (+)
strand using the University of Michigan Imputation Server (https://imputationserver.sph.umich.edu) and data from the 1000 Genomes Project (Phase 3, Version 5). Application of the same quality control filters described above to the 46,180,700 imputed SNPs, with the addition of excluding SNPs with Hardy-Weinberg equilibrium p-values <0.0001 and imputation Rsq scores <0.3, resulted in 8,986,545 autosomal SNPs that were available for analysis in 1276 GeneBank subjects for whom plasma glycine levels were also available. All patients provided written informed consent prior to being enrolled in GeneBank and the study was approved by the Institutional Review Board of the Cleveland Clinic.

**FINRISK:** FINRISK (FR) surveys are cross-sectional, population-based studies conducted every five years since 1972 to monitor risk of chronic diseases. For each survey, a representative random sample was selected from 25- to 74-year-old inhabitants of different regions in Finland. The survey included a questionnaire and a clinical examination, at which a blood sample was drawn, with linkage to national registries of cardiovascular disease and other health outcomes. The study protocol has been described elsewhere. Study participants were followed up through December 31, 2012. Eligible individuals from FINRISK surveys conducted in 1992, 1997, 2002, and 2007 (total n=27,838) were genotyped in three separate batches and analyzed separately to avoid batch effects, followed by a meta-analysis for glycine levels as described previously.

Genome-wide genotyping was carried out on an Illumina core-exome chip. After quality controls, including SNP call rates ≥ 95%, minor allele frequencies (MAFs) ≥ 1%, and sample call rates ≥ 95%, identity-by-descent (IBD) ≤ 0.1, without sex mismatches, duplicates, and heterozygosity outliers by eye from distribution, 273,113 SNPs was available for imputation. IMPUTE2 was used for imputation based on 1000 Genomes Project March 2012 version.
Further exclusions included p for Hardy–Weinberg equilibrium $\leq 1.0 \times 10^{-6}$ and imputation info $\leq 0.4^6$.

**Cardiovascular Risk in Young Finns Study (YFS):** The Cardiovascular Risk in Young Finns Study (YFS) is a population based prospective cohort study. It was conducted at five medical schools in Finland (Turku, Helsinki, Kuopio, Tampere and Oulu) with the aim of studying the levels of cardiovascular risk factors in children and adolescents in different parts of the country. The latest follow-up was conducted in 2007 at which serum samples were used for metabolomics analyses. The study and data collection protocols have been described in detail previously$^7$. Genome-wide SNP data were generated from a custom Illumina BeadChip containing 670,000 SNPs and CNV probes. The custom content on the custom 670K array replaced some poor performing SNPs on the Human610 BeadChip and added more CNV content, and includes 546,677 SNPs passing QC from 594,210 SNPs on the chip. The custom 670K chip shares 562,643 SNPs in common with the Illumina Human610 BeadChip. Genotypes were called using Illumina's clustering algorithm. A total of 2,556 samples were genotyped. After initial clustering, we removed 2 subjects for poor call rates (CR<0.90), and 54 samples failed subsequent QC filters (i.e., duplicated samples, heterozygosity, low call rate, or custom SNP fingerprint genotype discrepancy). The following filters were then applied to the remaining data: MAF 0.01, GENO 0.05, MIND 0.05, and HWE $1 \times 10^{-6}$. Three individuals were removed for low genotyping (MIND>0.05), 11,766 markers were excluded based on HWE test (P$\leq 1 \times 10^{-6}$), 7,746 SNPs failed missingness test (GENO>0.05), 34,596 SNPs failed frequency test (MAF<0.01), and one individual failed gender check. A final list of 546,677 SNPs passed QC and allele frequency filters$^8$. IMPUTE2 was used for imputation based on 1000 Genomes
Further exclusions included $p$ for Hardy–Weinberg equilibrium $\leq 1.0 \times 10^{-6}$ and imputation info $\leq 0.4$.

**Northern Finland Birth Cohort (NFBC):** The Northern Finland Birth Cohorts were initiated 20 years apart in 1966 (NFBC66) and 1986 (NFBC86) to examine risk factors involved in preterm birth and intrauterine growth retardation, and the consequences of these early adverse outcomes on subsequent morbidity and mortality, as described in detail previously. Mothers living in the two northern-most provinces of Finland (Oulu and Lapland) were invited to participate if they had expected delivery dates during 1966 or 1986. Individuals still living in the Helsinki area or Northern Finland were asked at age 31 to participate in a detailed biological and medical examination as well as a questionnaire. GWAS analyses for circulating glycine levels, as measured by NMR, were carried out in 4,483 and 3,112 from the NFBC66 and 1986 NFBC86 studies, respectively. Genomic DNA was extracted from whole blood using standard methods and samples were genotyped on the Illumina Infinium 370cvnDuo array at the Broad Institute Biological Sample Repository. All individuals in the study were genotyped with call rates $>95\%$. Individuals with discrepancy between their reported sex and the sex determined from the X chromosome were excluded from analysis. The identity-by-descent (IBD) analysis option of PLINK45 was used to determine possible relatedness among sample subjects and identify sample duplications and sample contamination (the latter identified as individuals who seemed to be related to nearly everyone in the sample). If the sample duplication issue could not be resolved by external means, both samples were excluded. All apparently contaminated samples were also excluded. For pairs of individuals identified to be related at the level of half-sibs or closer in the IBD analysis, the subject with less complete genotyping was excluded. Variants were excluded from the analysis if the call rate in the final sample was $<95\%$, if the $P$ value from a test of
Hardy-Weinberg Equilibrium (HWE) was <0.0001, or if the MAF was <1%\textsuperscript{10}. This resulted in 335,118 SNPs that were available for imputation. IMPUTE2 was used for imputation based on 1000 Genomes Project March 2012 version, with further exclusions for p for Hardy–Weinberg equilibrium ≤ 1.0×10\textsuperscript{-6} and imputation info ≤ 0.4\textsuperscript{6}.

**The Metabolic Syndrome in Men (METSIM) Study.** METSIM is a population-based study that recruited 10,197 Finnish men from the city of Kuopio in Eastern Finland between 2005-2010. The aims of METSIM are to investigate nongenetic and genetic factors associated with the risk of type 2 diabetes and cardiovascular disease, and with cardiovascular risk factors\textsuperscript{11}. The protocol included a detailed phenotyping of the participants, an oral glucose tolerance test, fasting laboratory measurements, including proton NMR measurements, mass spectrometry metabolomics, as well as adipose tissue biopsies and stool samples in a subset of participants. Participants were genotyped on the Human OmniExpresss-12v1_C BeadChip (OmniExpress) and Infinium HumanExome-12 v1.0 BeadChip (Exome Chip) platforms. Quality controls included sample-level controls for sex and relatedness confirmation, sample duplication, and detection of sample genetic ancestry outliers using principal component analysis. Based on these quality control measures, 14 samples with sex chromosome anomalies, 18 with evidence of participant duplication, 12 population outliers, and 9 samples with non-Mendelian inheritance inconsistencies were removed. In addition, one individual from each of seven monozygotic twin pairs was removed. Variants with low mapping quality of probes to genome build GRCh37, low genotype completeness (<95% and <98% for the OmniExpress and ExomeChip, respectively), or Hardy-Weinberg equilibrium P<10\textsuperscript{-6} were also filtered out. OmniExpress variants passing quality control with SHAPEIT v2 were phased and imputed using minimac v2. For imputation, a
reference panel of 20.9M variants from the GoT2D study (including SNVs, indels and large deletions) based on the whole genome sequence of 2874 Europeans, including 1004 Finnish individual, was used. Following imputation, variants directly genotyped on the ExomeChip were added. In cases of common markers between imputed and genotyped variants, the directly genotyped calls from the ExomeChip were used. Subsequently, 16,607,533 variants with high imputation quality (i.e. minimac RSQ0.3) were carried forward for single-variant association testing. GWAS analyses for circulating glycine levels, as measured by NMR, were carried out in a subset of 8545 non-diabetic men as described previously\textsuperscript{12}. The institutional review boards of the University of Kuopio and Kuopio University approved the METSIM study. Written informed consent was obtained from each participant.
Supplemental Table Legends (see Excel file):

**Table S1.** Results of 12 Loci Significantly Associated with Circulating Glycine Levels Stratified by Metabolomics Platform.

**Table S2.** Results of 12 Loci Significantly Associated with Circulating Glycine Levels Stratified by Sex.

**Table S3.** PheWAS Results for 12 Loci Significantly Associated with Circulating Glycine Levels.

**Table S4.** Association of 12 Glycine-associated Loci with CAD in CARDIoGRAM+C4D and UK Biobank.

**Table S5.** Individual and Joint SNP Effect Associations and Mendelian randomization analysis of Glycine-associated Loci with Traditionally CAD Risk Factors.
Figure S1. Quantile-quantile (Q-Q) plot of GWAS meta-analysis results for circulating glycine levels in 30,118 subjects. The observed versus the expected p-values from the meta-analyses for glycine levels are shown in the Q-Q plot. These analyses yielded a genomic inflation factor ($\lambda$) of 1.035, indicating that the GWAS meta-analyses were not confounded by underlying population stratification.
Figure S2. Twelve loci identified for circulating glycine associated levels. Regional plots for the *ACADM, PHGDH, CPS1, ALDH1L1, COX18-ADAMTS3, PSPH, PPP1R3B-LOC157273, TRIB1, GLDC, PTPRD, ABO*, and GCSH loci are shown in panels A-L. Each region is centered on the lead SNP (purple diamond) and the genes in the interval are indicated in the bottom panel. The degree of linkage disequilibrium (LD) between the lead SNP and other variants is shown as $r^2$ values according to the color-coded legend in the box.
Figure S3. Results of GWAS meta-analysis for circulating glycine levels in women. (A) The Manhattan plot shows five previously identified loci significantly associated with circulating glycine levels (blue dots) in a stratified GWAS analysis with 10,886 women. Red dots indicate association signals for the seven novel identified in our meta-analysis with all 30,118 subjects, all of which were only suggestively associated in women. Genome-wide thresholds for significant (\(P=5.0 \times 10^{-8}\)) and suggestive (\(P=5.0 \times 10^{-6}\)) association are indicated by the horizontal red and dark blue lines, respectively. P-values are truncated at \(-\log_{10}(P)=40\). (B) The Q-Q plot shows the observed versus the expected p-values from the meta-analyses for glycine levels in women. These analyses yielded a genomic inflation factor (\(\lambda\)) of 1.002, indicating that the GWAS meta-analyses were not confounded by underlying population stratification.
Figure S4. Results of GWAS meta-analysis for circulating glycine levels in men. (A) The Manhattan plot shows nine loci significantly associated with circulating glycine levels in a stratified GWAS analysis with 19,004 men. The five loci identified in previous studies are indicated by blue dots. The red dots indicate association signals at the seven novel identified by our meta-analysis with all 30,118 subjects, of which four were also significant in only men. Genome-wide thresholds for significant ($P=5.0\times10^{-8}$) and suggestive ($P=5.0\times10^{-6}$) association are indicated by the horizontal red and dark blue lines, respectively. P-values are truncated at $-\log_{10}(P)=40$. (B) The Q-Q plot shows the observed versus the expected p-values from the meta-analyses for glycine levels in men. These analyses yielded a genomic inflation factor ($\lambda$) of 1.035, indicating that the GWAS meta-analyses were not confounded by underlying population stratification.
Figure S5. Sex-stratified results for 12 loci identified for circulating glycine levels. Effect sizes for the lead SNPs at the 12 loci identified for circulating glycine levels are shown in men (blue) and women (red) separately. With the exception of CPS1, which is associated with approximately two-fold higher glycine levels in women compared to men, effect sizes at the 11 other loci were similar in males and females. EA, effect allele; OA, other allele.
Supplemental References:


