Genome-wide Association Study of Plasma Proteins Identifies Putatively Causal Genes, Proteins, and Pathways for Cardiovascular Disease

Chen Yao[1,2*], George Chen[1,2*], Ci Song[1-4*], Michael Mendelson[1,2,5], Tianxiao Huan[1,2], Annika Laser[6,7], Hongsheng Wu[8], Jennifer E. Ho[9], Paul Courchesne[1,2], Asya Lyass[1,10], Martin G. Larson[1,11], Christian Gieger[6,7,12], Johannes Graumann[13], Andrew D. Johnson[1,2], Shih-Jen Hwang[1,2], Chunyu Liu[1,2], Karsten Suhre[14^], Daniel Levy[1,2^].

* and ^ denote equal contributions to this manuscript.

Affiliations: (1) Framingham Heart Study, Framingham, MA; (2) Population Sciences Branch, Division of Intramural Research, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD; (3) Department of Medical Sciences, Uppsala University, 75105 Uppsala, Sweden; (4) Department of Immunology, Genetics and Pathology, Uppsala University, 75105 Uppsala, Sweden; (5) Department of Cardiology, Boston Children's Hospital, Boston, MA; (6) Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany; (7) Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany; (8) Computer Science and Networking, Wentworth Institute of Technology, Boston, MA; (9) Cardiovascular Research Center and Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Boston, MA; (10) Department of Mathematics and Statistics, Boston University, Boston, MA; (11) Department of Biostatistics, Boston University School of Public Health, Boston, MA; (12) German Center for Diabetes Research (DZD), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany; (13) Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Ludwigstr. 43, D—61231 Bad Nauheim, Germany; (14) Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, PO 24144, Doha, Qatar.

Correspondence:
Daniel Levy, MD (email levyd@nih.gov; phone 508-935-3458), Framingham Heart Study, 73 Mount Wayte Avenue, Framingham, MA 01702.

Summary:

Identifying genetic variants associated with circulating protein concentrations (pQTLs) and integrating them with variants from genome-wide association studies (GWAS) may illuminate the proteome's causal role in disease and bridge a GWAS knowledge gap for hitherto unexplained SNP-disease associations. We conducted GWAS of 71 high-value proteins for cardiovascular disease in 6,861 Framingham Heart Study participants followed by external replication. We comprehensively mapped thousands of pQTLs, including functional annotations and clinical-trait associations, and created an integrated plasma-protein-QTL searchable database. We next identified 15 proteins with pQTLs coinciding with coronary heart disease (CHD)-related variants from GWAS or tested causal for CHD by Mendelian randomization; most of these proteins were associated with new-onset cardiovascular disease events in Framingham participants with long-term follow-up. Identifying pQTLs and integrating them with GWAS results yields insights into genes, proteins, and pathways that may be causally associated with disease and can serve as therapeutic targets for treatment and prevention.

Key words: pQTL, GWAS, proteomics, cardiovascular disease, Mendelian randomization

Introduction

Considerable progress has been made in identifying genetic underpinnings of coronary heart disease (CHD),[1-4] which remains the leading cause of death worldwide.[5] Proteins are the functional products of the genome and serve as critical factors for biological processes involved in health and disease as well as primary drug targets. Numerous proteins have been reported to be associated with CHD; it is often difficult, however, to establish with certainty whether CHD-associated proteins are causally related to risk or simply represent downstream markers of disease-related processes. Identifying genetic variants associated with protein levels (protein quantitative trait loci; pQTLs), characterizing pQTLs that also are associated with CHD from genome-wide association studies (GWAS), and inferring causality may provide novel insights into the roles of genetic variants, genes, and the proteins they code in the pathogenesis of CHD. To date, most pQTL studies[6-15] have been based on small sample sizes or did not conduct prospective testing of associations between protein levels and clinical disease.

To address a GWAS knowledge gap for genetic variants of unknown relevance to CHD, we conducted a multistage study (Figure 1) consisting of GWAS of high-value cardiovascular disease (CVD) plasma proteins that were measured in Framingham Heart Study (FHS) participants, followed by external replication in participants from the Cooperative Health Research in the Region of Augsburg (KORA) F4 study[12] and from other protein GWAS. We integrated pQTLs with genetic variants from CHD GWAS databases[1-4] and employed Mendelian randomization (MR)[16] to reveal proteins with potentially causal effects on CHD. Last, we tested proteins for association with new-onset CHD events in FHS participants with long-term follow-up. We hypothesized that a

3

strategy of protein GWAS followed by causal testing and prospective association with CHD outcomes would identify putatively causal genes, proteins, and pathways for CHD and highlight novel targets for its prevention and treatment.

Findings

<u>Discovery Set</u>: Seventy-one proteins, selected *a priori* based on prior evidence of association with CVD, were measured in 7,333 FHS participants (Table S1). The sample size available for GWAS was up to 6,861 participants (mean age 50 years, 53% women); clinical characteristics of the discovery sample are summarized in Table S2.

<u>pQTL Mapping</u>: With a GWAS sample size of ~6,800 participants and a significance threshold of $p<5x10^{-8}$, our study had 80% power to detect a pQTL that explained $\geq 0.6\%$ of variance in protein levels (Table S3). We identified 1,793 insertion/deletion polymorphisms for 57 proteins (Table S4) and 20,495 pQTLs with Reference SNP cluster IDs for 60 proteins (Table S5), including 11,974 cis-pQTLs representing 39 sentinel *cis*-pQTL loci for 39 proteins (Figure 2a; Table S6) and 8,521 trans-pQTLs representing 91 sentinel *trans*-pQTL loci for 48 proteins (Figure 2b; Table S6). Pruning the 1000 Genomes Project (1000G) reference panel[17] GWAS pQTLs (linkage disequilibrium $r^2<0.2$) yielded 4,588 non-redundant variants (Table S7). Thirty-four pQTLs were rare variants (minor allele frequency<1%) associated with 18 proteins (Table S8).

The effect sizes and the proportion of inter-individual variation explained by some pQTLs were large. For example, *cis*-pQTL rs941590, is a missense variant that explained 32% of inter-individual variation in SERPINA10 levels (Figure S1) and was previously reported to be associated with family history of venous thrombosis.[18] Three

4

proteins (PON1, GRN, and LPA) had pQTLs that explained 10-30% of variation in protein levels. Minor allele frequency was inversely correlated with effect size, but not with proportion of variance explained. In general, *cis*-pQTLs and missense variants had larger effect sizes and explained a greater proportion of the variation in protein levels than did *trans*-pQTLs and non-coding variants, respectively (Figure 3).

External Replication: Among our 60 proteins linked to 130 sentinel pQTLs, 46 proteins (associated with 105 sentinel pQTLs) were measured in the KORA F4 study[12] or in prior GWAS. For each pQTL locus identified in the FHS, the pQTL with the lowest protein-association p-value was selected as the sentinel pQTL for external replication. Thirty-six proteins, encompassing 82 sentinel pQTLs, were measured in KORA (21 *cis*-pQTLs and 61 *trans*-pQTLs); 23 additional loci were evaluated for replication in other GWAS. Based on 1,000 re-samplings of 1,000 unrelated FHS participants in the discovery sample, 34 pQTL-protein associations yielded $p<4.8\times10^{-4}$ (alpha level of 0.05 after Bonferroni correction for 105 tests; 0.05/105) in $\geq80\%$ of samplings and thus were considered likely to replicate in a GWAS sample size of 1,000 (Table S9). Of the 21 qualifying sentinel *cis*-pQTLs from FHS, 10 replicated in KORA ($p<4.8\times10^{-4}$; Table S10). Of the 61 sentinel *trans*-pQTLs (31 proteins) from FHS, 17 (14 proteins) replicated in KORA ($p<4.8\times10^{-4}$; Table S10). Next, replication from other external protein GWAS was conducted. Overall, among sentinel pQTLs from FHS for which replication was possible, all of the ten most significant sentinel *cis*-pQTLs and seven of the ten most significant sentinel *trans*-pQTLs replicated. In total, 19 of 31 (61%) *cis*-pQTL loci and 22 of 74 (30%) *trans*-pQTL loci replicated at $p<4.8\times10^{-4}$ (Table S10).

pQTL Functional, Regulatory, and Clinical Annotation: Among the entire set of pQTLs, 334 are missense variants associated with 43 proteins and 8,217 are intronic variants (Table S11; Figure S2).[19] Pathway enrichment analysis of all pQTLs using DEPICT[20] identified 728 interrelated gene sets (false discovery rate [FDR]<0.05, p<0.0034); many are associated with lipids, metabolic processes, or inflammatory response (Table S12). Tissue enrichment analysis revealed that pQTL-mapped genes are highly expressed in monocytes ($p=2.38 \times 10^{-4}$) and hepatocytes ($p=4.55 \times 10^{-5}$). We employed Functional Mapping and Annotation[21] of GWAS (FUMA; http://fuma.ctglab.nl) to generate detailed annotations of pQTLs for each protein (regional plot of each pQTL locus, functional categorization of pQTL SNPs, gene mapping, and pathway enrichment analyses) that are provided in Figure S3. These annotations revealed that pQTLs often reside in active regulatory regions and are frequently located in intergenic and intronic regions. Protein-specific Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses of the corresponding pQTLs revealed a preponderance of pathways concordant with the function of the studied protein.

Enrichment of pQTLs with eQTLs: More than 90% pQTLs were annotated with chromatin marks or DNase hypersensitive sites by HaploReg[22] (Table S13), suggesting that they play an important role in gene regulation. From the 1000G GWAS, we identified 8,542 pQTLs (46% of total discovery) for 20 proteins that also are whole blood eQTLs (genetic variants associated with gene expression levels in 5,257 FHS participants at FDR<0.05),[23] including 8,532 *cis*-eQTLs and 596 *trans*-eQTLs (Fisher's exact test for enrichment $p<1 \times 10^{-8}$; Table S14). Among the 130 sentinel pQTLs, 72 (55%) are eQTLs. Moreover, we identified pQTLs associated with expression of the corresponding protein-

coding genes for 15 proteins, suggesting that many pQTLs affect circulating protein levels by regulating blood cell gene expression (Table S15).

Clinical Annotation: We found that 58 missense pQTLs (from the Exome Chip) were linked to clinical disorders in the NCBI ClinVar[24] database (Table S16). We identified examples where the missense pQTL and its associated protein are both linked to CHD-related traits. For example, rs2228671, a missense variant in the LDL-receptor gene (*LDLR*), was associated in our pQTL analysis with circulating APOB levels, the major lipoprotein of LDL particles, and was previously reported to be pathogenic for familial hypercholesterolemia (FH), a monogenic disorder of LDL cholesterol.[25] Additionally, for several variants reported to be benign in ClinVar, we demonstrated associations with the disease-relevant protein, suggesting that they have clinical consequences.

Integrating pQTLs with CHD-associated SNPs: We integrated our pQTLs with 2,738 CHD-related SNPs from the CARDIoGRAMplusC4D Consortium[1] and other CHD GWAS.[2-4] A total of 201 pQTLs (19 independent pQTLs with linkage disequilibrium $r^2<0.2$, representing 14 proteins) matched SNPs associated with CHD in GWAS ($p<5\times10^{-8}$; Table S17). Table 1 displays the sentinel pQTL, based on lowest protein-association p-value that coincided with a CHD-related GWAS SNP, and the corresponding protein at each genetic locus. The proteins identified by this approach included GRN, APOB, ANGTL3, CRP, B2M, GMP140, sICAM1, REG1A, MCAM, LPA, sGP130, BCHE, sRAGE, and CXCL16.

We found the ABO locus to have links to CHD through five circulating proteins (MCAM, sICAM1, GMP140, sGP130, REG1A). ABO blood type has long been linked

to CVD risk, including in the FHS,[26] additional reports have linked the ABO locus to CVD via coagulation pathways.[27,28] ABO locus-related proteins, identified in our study, are involved in inflammatory pathways, including interleukin and interferon signaling (Figure 2). The multi-protein association of this locus may be driven by the general function of ABO as a glycosyltransferase.

Some of the genes coding for CHD-related proteins have been linked to known CHD risk pathways in previous GWAS of lipids (APOB, LPA, ANGPTL3), coagulation (GMP140), and systemic inflammation (sGP130, sICAM1) (Figure 2). Many of the proteins that share genetic underpinnings with CHD are known drug targets (per the DrugBank database),[29] or currently under development as such (*e.g.* ANGPTL3, LPA, sICAM1, GMP140). Several proteins with pQTLs linked to CHD, however, are not known drug targets, especially those from gene loci not previously linked to CHD risk pathways (e.g. BCHE, CXCL16, GRN, MCAM, and sRAGE).

Causal Testing: We applied MR testing to infer the causal association between protein levels and CHD for all proteins having *cis*-pQTLs and those with at least four *cis*- or *trans*-pQTL loci that coincided with CHD-associated SNPs from GWAS.[1] MR causally implicated LPA, REG1A, MCAM, and SAA1 via *cis*-pQTLs as instrument variables (p<0.05; Table S18). For 11 proteins with pQTLs that coincided with SNPs from CHD GWAS and had at least four non-redundant *cis*- or *trans*-pQTLs, we conducted MR analyses using a multi-SNP approach, implemented in MRbase,[30] which revealed causal CHD associations for APOB (p=0.0005) and GRN (p=7.0x10^{-5}).

Protein Associations with Clinical Outcomes: For 15 proteins (Figure 4) with pQTLs that coincided with CHD GWAS SNPs or tested positive in MR analyses at p<0.05 we tested

8

the associations of protein levels with a) major CHD (recognized myocardial infarction or CHD death; n=213 events) and b) CVD death (fatal CHD or death due to stroke, peripheral arterial disease, heart failure, or other CVD causes; n=199 events) occurring during a median follow-up of 14.3 years (25th percentile 11.4, 75th percentile 15.2 years) among 3,520 FHS participants age ≥50 years. Twelve of the 14 proteins with pQTLs that coincided with CHD GWAS SNPs were associated (nominal p<0.05) with incident CHD or CVD death (Table 2). After adjusting for multiple testing of 15 proteins (p<0.05/15 = p<3.3×10^{-3}), nine proteins remained associated with incident events. Four (REG1A, SAA1, APOB, and GRN) of the six proteins that tested causal for CHD by MR (at p<0.05) were associated with CHD/CVD outcomes (at p<0.05). Two proteins (AGP1 and HPX) that tested marginally positive for CHD risk in MR analysis (0.05<p<0.10) were associated with new-onset CHD events (p=3.5×10^{-8} for AGP1 and p=2.0×10^{-5} for HPX). The protein effect sizes on CHD predicted from MR were consistent with the observed prospective protein-CHD associations (Figure 5).

Novel Proteins and Pathways Implicated in CHD: Whereas LPA and APOB can be viewed as positive controls because prior studies identified them as causal for CHD,[31-34] four proteins that were causal for CHD in MR are novel, including REG1A, MCAM, SAA1, and GRN (Table S18). REG1A is a protein secreted by the pancreas and may be related to islet cell regeneration and diabetogenesis, potentially contributing to increased atherogenic risk of diabetes.[35] REG1A levels were positively associated with CHD and CVD outcomes in our protein-trait association analyses (Table 2), and previous studies have shown that levels of REG1A are elevated in individuals with CHD and type 2 diabetes.[35,36] MCAM, also known as CD146, a trans-membrane glycoprotein, is highly

9

expressed in vascular cells and plays a role in cell adhesion. MCAM/CD146 levels are a biomarker of endothelial activation/injury and are associated with carotid intima thickness[37,38] and risk for acute CHD events.[39-41] Our MR and integrative GWAS analyses suggest a protective role of MCAM/CD146 on CHD risk. Protein-trait association analysis for MCAM/CD146 in the FHS did not find evidence of association with CHD or CVD events, however, a previous FHS case-control analysis reported an inverse association of MCAM/CD146 with myocardial infarction,[36] concordant with our MR results.

SAA1, a precursor to amyloid A, is an inflammatory apolipoprotein inversely associated with HDL-cholesterol.[42] Serum SAA1 levels are elevated in patients with CHD.[43] In addition, SNPs located in the *SAA1* gene are associated with carotid intima media thickness and HDL-cholesterol levels.[44,45] Our follow-up analyses revealed association of circulation SAA1 levels with CHD and CVD risk. MR results for GRN were positive, largely by virtue of sentinel *trans*-pQTLs. One of our sentinel *trans*-pQTLs for GRN, rs12740374, is located at the *CELSR2/SORT1* locus on Chromosome 1; it explained 15% of variance in GRN levels and was associated with CHD at $p<4.6 \times 10^{-23}$ in prior GWAS.[1] Previous studies reported that rs12740374 affects expression levels of the *SORT1* gene in human hepatocytes, which in turn regulate LDL-cholesterol levels.[46,47] Our longitudinal analyses revealed association of GRN with CVD death in FHS participant (p=0.001; Table 2).

Molecular QTL browser: Our pQTL resource is accessible through the NCBI Molecular QTL Browser (ftp://ftp.ncbi.nlm.nih.gov/eqtl/original_submissions/FHS_pQTLs/; a link to the browser will be sent to reviewers under separate cover), which serves as a data

resource for associations between genetic variants and molecular phenotypes. The browser links our pQTL results to eQTLs and other molecular resources via a user-friendly interface. Users can browse and search results and specify p-value cutoffs and other data filters. The Molecular QTL Browser also permits users to conduct targeted studies of specific genes based on prior evidence. The integrated data resource enables searches across datasets and filtering by functional annotation and genomic position.

Discussion

Using a multistage strategy, we discovered thousands of pQTLs associated with scores of proteins that were selected *a priori* as high-value plasma proteins for CVD. Integration of pQTLs with CHD GWAS revealed 14 proteins with pQTLs that coincide with CHD SNPs (Table 1) and MR analyses identified six proteins with evidence for causal effects on CHD (Table S18), four of which were novel (REG1A, MCAM, SAA1, and GRN). Furthermore, most of these proteins were associated with new-onset CHD or CVD events in FHS participants with long-term follow-up (Table 2). Our strategy connected pQTLs, genes, the proteins they code, and CHD risk (Figure 4) and highlighted a comprehensive approach to bridge the GWAS knowledge gap for genetic variants that have no links to disease via known mechanisms.

We acknowledge several limitations of our study. Participants were of European ancestry; consequently, the results may not be directly generalizable to populations with different genetic backgrounds. Although our sample size for GWAS was large, our ability to detect pQTLs and to test them for causality using MR was limited by power. Protein

11

levels were measured in whole blood and may not reflect tissue-specific patterns of expression.

To our knowledge, this is the largest sample size pQTL study with well-powered discovery, independent external replication, CHD causal testing, and confirmatory prospective protein-CHD outcome findings. We provide a large and comprehensive compilation of pQTLs as a resource for other researchers (via the NCBI Molecular QTL Browser) and provide evidence that an integrated genomic approach can identify proteins with putatively causal effects on CHD. Although some of our causally-implicated proteins may act through classic CHD risk factors and known pathways, many do not and thus represent attractive candidate targets for drug development. Additional studies are needed to elucidate the mechanisms by which such proteins alter CHD risk as well as trials to confirm our MR prediction that perturbing these pathways can prevent CHD events. Taken together, the genetic variants associated with circulating protein levels in this study shed new light on genes, proteins, and pathways contributing to the pathogenesis of CHD, which could have profound implications for the treatment and prevention of the leading cause of death worldwide.

Acknowledgements

References

1. CARDIoGRAMplusC4D Consortium, M Nikpey, A Goel, H Won, LM Hall C. Willenborg, S Kanoni, D Saleheen et al. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. Nat Genet 2015 47:1121-1130.

2. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009;106(23):9362-7.

3. Leslie R, O'Donnell CJ, Johnson AD (2014) GRASP: analysis of genotype-phenotype results from 1,390 genome-wide association studies and corresponding open access database. Bioinformatics 30(12), i185-94. GRASP Build 2.0.0.0.

4. Eicher JD, Landowski C, Stackhouse B, et al. GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. Nucleic Acids Res. 2015;43(Database issue):D799-804.

5. Gaziano TA, Bitton A, Anand S, Abrahams-gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. Curr Probl Cardiol. 2010;35(2):72-115.

6. Melzer D, et al. A genome-wide association study identifies protein quantitative trait loci (pQTLs). PLoS Genet. 2008;4:e1000072.

7. Chick JM, Munger SC, Simecek P, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016;534(7608):500-5.

8. Lourdusamy A, Newhouse S, Lunnon K, et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum Mol Genet. 2012;21(16):3719-26.

9. Enroth S, Johansson A, Enroth SB, Gyllensten U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. Nat Commun. 2014;5:4684.

10. Liu Y, Buil A, Collins BC, et al. Quantitative variability of 342 plasma proteins in a human twin population. Mol Syst Biol. 2015;11(1):786.

11. Solomon T, Smith EN, Matsui H, et al. Associations Between Common and Rare Exonic Genetic Variants and Serum Levels of 20 Cardiovascular-Related Proteins: The Tromsø Study. Circ Cardiovasc Genet. 2016;9(4):375-83.

12. Suhre K, Arnold M, Bhagwat AM, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357.

13. Hause RJ, Stark AL, Antao NN, et al. Identification and validation of genetic variants that influence transcription factor and cell signaling protein levels. Am J Hum Genet. 2014;95(2):194-208.

14. Johansson Å, Enroth S, Palmblad M, Deelder AM, Bergquist J, Gyllensten U. Identification of genetic variants influencing the human plasma proteome. Proc Natl Acad Sci USA. 2013;110(12):4673-8.

15. Folkersen L, Fauman E, Sabater-Lleal M, et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. doi: http://dx.doi.org/10.1371/journal.pgen.1006706.

16. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?. Int J Epidemiol. 2003;32(1):1-22.

17. Peloso GM, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014;94:223-32.

18. Corral J, González-conejero R, Soria JM, et al. A nonsense polymorphism in the protein Z-dependent protease inhibitor increases the risk for venous thrombosis. Blood. 2006;108(1):177-83.

19. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015;16:56.

20. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun. 2015;6:5890.

21. FUMA: Functional mapping and annotation of genetic associations. Kyoko Watanabe, Erdogan Taskesen, Arjen van Bochoven, Danielle Posthuma. bioRxiv 110023; doi: https://doi.org/10.1101/110023.

22. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. Nucleic Acids Res. 2016;44(D1):D877-81.

23. Joehanes R, Zhang X, Huan T, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. Genome Biol. 2017;18(1):16.

24. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980-5

25. Brown MS, Goldstein JL. Receptor-mediated control of cholesterol metabolism. Science. 1976;191(4223):150-4.

26. Garrison RJ, Havlik RJ, Harris RB, Feinleib M, Kannel WB, Padgett SJ. ABO blood group and cardiovacular disease: the Framingham study. Atherosclerosis. 1976;25(2-3):311-8.

27. Song J, Chen F, Campos M, et al. Quantitative Influence of ABO Blood Groups on Factor VIII and Its Ratio to von Willebrand Factor, Novel Observations from an ARIC Study of 11,673 Subjects. PLoS ONE. 2015;10(8):e0132626.

28. Larson NB, Bell EJ, Decker PA, et al. ABO blood group associations with markers of endothelial dysfunction in the Multi-Ethnic Study of Atherosclerosis. Atherosclerosis. 2016;251:422-9.

29. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014;42(Database issue):D1091-7.

30. Gibran Hemani, Jie Zheng, Kaitlin H Wade, Charles Laurin, Benjamin Elsworth, Stephen Burgess, Jack Bowden, Ryan Langdon, Vanessa Tan, James Yarmolinsky, Hashem A. Shihab, Nicholas Timpson, David M Evans, Caroline Relton, Richard M Martin, George Davey Smith, Tom R Gaunt, Philip C Haycock, The MR-Base Collaboration. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. bioRxiv. doi: https://doi.org/10.1101/078972

31. Kamstrup PR, Tybjærg-hansen A, Nordestgaard BG. Lipoprotein(a) and risk of myocardial infarction--genetic epidemiologic evidence of causality. Scand J Clin Lab Invest. 2011;71(2):87-93.

32. Lamon-fava S, Marcovina SM, Albers JJ, et al. Lipoprotein(a) levels, apo(a) isoform size, and coronary heart disease risk in the Framingham Offspring Study. J Lipid Res. 2011;52(6):1181-7.

33. Tybjaerg-hansen A, Steffensen R, Meinertz H, Schnohr P, Nordestgaard BG. Association of mutations in the apolipoprotein B gene with hypercholesterolemia and the risk of ischemic heart disease. N Engl J Med. 1998;338(22):1577-84.

34. Smith GD, Ebrahim S. Mendelian randomization: prospects, potentials, and limitations. Int J Epidemiol. 2004;33(1):30-42.

35. Yang J, Li L, Raptis D, et al. Pancreatic stone protein/regenerating protein (PSP/reg): a novel secreted protein up-regulated in type 2 diabetes mellitus. Endocrine. 2015;48(3):856-62.

36. Yin X, Subramanian S, Hwang SJ, O'Donnell CJ, Fox CS, Courchesne P, et al. Protein biomarkers of new-onset cardiovascular disease: prospective study from the systems approach to biomarker research in cardiovascular disease initiative. Arterioscler Thromb Vasc Biol. 2014;34:939-945.

37. Qian YN, Luo YT, Duan HX, et al. Adhesion molecule CD146 and its soluble form correlate well with carotid atherosclerosis and plaque instability. CNS Neurosci Ther. 2014;20(5):438-45.

38. Ciftci dogansen S, Helvaci A, Adas M, Deniz onal S. The relationship between early atherosclerosis and endothelial dysfunction in type 1 diabetic patients as evidenced by measurement of carotid intima-media thickness and soluble CD146 levels: a cross sectional study. Cardiovasc Diabetol. 2013;12:153.

39. Kubena P, Arrigo M, Parenica J, et al. Plasma Levels of Soluble CD146 Reflect the Severity of Pulmonary Congestion Better Than Brain Natriuretic Peptide in Acute Coronary Syndrome. Ann Lab Med. 2016;36(4):300-5.

40. Gayat E, Caillard A, Laribi S, et al. Soluble CD146, a new endothelial biomarker of acutely decompensated heart failure. Int J Cardiol. 2015;199:241-7.

41. Arrigo M, Truong QA, Onat D, et al. Soluble CD146 Is a Novel Marker of Systemic Congestion in Heart Failure Patients: An Experimental Mechanistic and Transcardiac Clinical Study. Clin Chem. 2017;63(1):386-393.

42. Benditt EP, Eriksen N. Amyloid protein SAA is associated with high density lipoprotein from human serum. Proc Natl Acad Sci USA. 1977;74(9):4025-8.

43. Fyfe AI, Rothenberg LS, Debeer FC, Cantor RM, Rotter JI, Lusis AJ. Association between serum amyloid A proteins and coronary artery disease: evidence from two distinct arteriosclerotic processes. Circulation. 1997;96(9):2914-9.

44. Xie X, Ma YT, Yang YN, et al. Polymorphisms in the SAA1/2 gene are associated with carotid intima media thickness in healthy Han Chinese subjects: the Cardiovascular Risk Survey. PLoS ONE. 2010;5(11):e13997.

45. Carty CL, Heagerty P, Heckbert SR, et al. Association of genetic variation in serum amyloid-A with cardiovascular disease and interactions with IL6, IL1RN, IL1beta and TNF genes in the Cardiovascular Health Study. J Atheroscler Thromb. 2009;16(4):419-30.

46. Musunuru K, Strong A, Frank-kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010;466(7307):714-9.

47. Kjolby M, Andersen OM, Breiderhoff T, et al. Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export. Cell Metab. 2010;12(3):213-23.

Table 1. Proteins with pQTLs that Coincide with Coronary Heart Disease-associated SNPs from Genome-wide Association Studies

Protein	Protein-coding Gene	Location of Protein-coding Gene	pQTL*	pQTL Annotated Gene Locus	pQTL Location	pQTL Function	pQTL-protein P-value	CHD GWAS P-value†
ANGPTL3	*ANGPTL3*	7:63.1M	rs964184	*ZNF259*	11:116.6Mb	UTR3	1.1×10^{-14}	8.0×10^{-10}
APOB	*APOB*	2:21.2M	rs12740374	*CELSR2*	1:109.8Mb	UTR3	3.2×10^{-15}	3.3×10^{-18}
APOB	*APOB*	2:21.2M	rs964184	*ZNF259*	11:116.6Mb	UTR3	9.5×10^{-9}	8.0×10^{-10}
APOB	*APOB*	2:21.2M	rs445925	*APOC1*	19:45.4Mb	Upstream	7.0×10^{-27}	9.4×10^{-11}
B2M	*B2M*	15:45.0M	rs2508015	*HLA-C*	6:31.0Mb	Intergenic	5.2×10^{-10}	1.5×10^{-9}
B2M	*B2M*	15:45.0M	rs10774625	*ATXN2*	12:119.1Mb	Intronic	8.7×10^{-11}	8.0×10^{-9}
BCHE	*BCHE*	3:165.5M	rs35071165	*ABI2*	2:204.2Mb	Intronic	3.4×10^{-8}	5.2×10^{-10}
CRP	*CRP*	1:159.7M	rs12721051	*APOC1*	19:45.4Mb	UTR3	2.4×10^{-17}	2.0×10^{-10}
CXCL16	*CXCL16*	17:4.6M	rs11065987	*CUX2*	12:112.1Mb	Intergenic	1.1×10^{-8}	2.5×10^{-10}
GMP140	*SELP*	1:169.6M	rs2519093††	*ABO*	9:136.1Mb	Intronic	1.2×10^{-78}	1.2×10^{-11}
GRN	*GRN*	17:42.4M	rs12740374	*CELSR2*	1:109.8Mb	UTR3	2.7×10^{-268}	4.6×10^{-23}
LPA	*LPA*	6:161.0M	rs55730499**	*LPA*	6:161.0Mb	Intronic	3.8×10^{-167}	5.4×10^{-39}
MCAM	*MCAM*	11:119.2M	rs550057	*ABO*	9:136.1Mb	Intronic	1.4×10^{-11}	4.2×10^{-9}
REG1A	*REG1A*	2:79.3M	rs687289	*ABO*	9:136.1Mb	Intronic	2.6×10^{-12}	7.7×10^{-9}
sGP130	*IL6ST*	5:55.2M	rs507666††	*ABO*	9:136.1Mb	Intronic	8.2×10^{-19}	1.6×10^{-11}
sICAM1	*ICAM1*	19:10.4M	rs532436††	*ABO*	9:136.1Mb	Intronic	7.7×10^{-41}	1.6×10^{-11}
sRAGE	*AGER*	6:32.1M	rs2523535**	*MICA*	6:31.3Mb	Upstream	1.3×10^{-9}	8.1×10^{-10}

*For proteins with multiple pQTLs that coincided with coronary heart disease GWAS SNPs, the pQTL with the lowest p-value of association with its corresponding protein level is shown.

18

**Indicates *cis*-pQTL. All other pQTLs shown in this table are *trans*-pQTLs.

†P-value of associations with coronary heart disease risk in GWAS reported in the NHGRI GWAS Catalog, GRASP Database, or CARDIOGRAMplusC4D Consortium.

††Indicates pQTLs located in the *ABO* gene that are in high LD ($r^2 > 0.8$) with each other.

Abbreviations: CHD = coronary heart disease; GWAS = genome-wide association study; pQTL = protein quantitative trait locus (i.e. genetic variant associated with protein level)

19

Table 2. Protein Associations with Coronary Heart Disease Events and Cardiovascular Disease Death in Framingham Heart Study Participants with Long-term Follow-up

Protein	Link to CHD in GWAS, MR, or Both	Protein Description	KEGG Pathways for Which pQTLs of Each Protein Are Enriched for†	Association with CHD Events*		Association with CVD Death*	
				Hazards Ratio (95% CI)	P-value††	Hazards Ratio (95% CI)	P-value††
ANGPTL3	CHD GWAS	A member of a family of secreted proteins that function in angiogenesis	PPAR signaling pathway	1.18 (1.01-1.36)	0.03	0.99 (0.85-1.16)	0.94
APOB	Both	The main apolipoprotein of chylomicrons and low density lipoproteins	Endocytosis	1.44 (1.24-1.67)	**$1.8x10^{-6}$**	1.07 (0.91-1.26)	0.41
B2M	CHD GWAS	A serum protein found in association with the major histocompatibility complex (MHC) class I heavy chain on the surface of nearly all nucleated cells	Type I diabetes mellitus Antigen processing and presentation Allograft rejection Graft versus host disease Autoimmune thyroid disease	1.47 (1.24-1.75)	**$9.0x10^{-6}$**	1.97 (1.63-2.38)	**$2.3x10^{-12}$**
BCHE	CHD GWAS	A cholinesterase enzyme and member of the type-B carboxylesterase/lipase family of proteins	None	1.09 (0.94-1.26)	0.3	0.82 (0.70-0.96)	0.01
CRP	CHD GWAS	A member of the pentaxin family	None	1.4 (1.20-1.62)	**$1.4x10^{-5}$**	1.43 (1.23-1.68)	**$5.6x10^{-6}$**
CXCL16	CHD GWAS	A scavenger receptor on macrophages	Chemokine signaling pathway Cytokine-cytokine receptor interaction	1.13 (0.97-1.31)	0.1	1.17 (1.00-1.37)	0.04

			Intestinal immune network for IgA production Neurotrophin signaling pathway				
GMP140	CHD GWAS	Ca(2+)-dependent receptor for myeloid cells that binds to carbohydrates on neutrophils and monocytes	Cell adhesion molecules	1.23 (1.06-1.42)	7.1×10^{-3}	1.25 (1.06-1.46)	6.0×10^{-3}
GRN	Both	A family of secreted, glycosylated peptides	None	1.14 (0.98-1.32)	0.08	1.29 (1.11-1.51)	$\mathbf{1.2 \times 10^{-3}}$
LPA	Both	A serine proteinase that inhibits the activity of tissue-type plasminogen activator I	None	1.09 (0.94-1.26)	0.2	1.09 (0.93-1.27)	0.29
MCAM	Both	Plays a role in cell adhesion, and in cohesion of the endothelial monolayer at intercellular junctions in vascular tissue	None	0.88 (0.76-1.03)	0.1	1.11 (0.95-1.30)	0.19
REG1A	Both	A type I subclass member of the Reg gene family	Glycosphingolipid biosynthesis - lacto and neolacto series Glycosphingolipid biosynthesis - globo series	1.28 (1.10-1.48)	$\mathbf{1.2 \times 10^{-3}}$	1.47 (1.25-1.73)	$\mathbf{2.6 \times 10^{-6}}$
SAA1	MR	A member of the serum amyloid A family of apolipoproteins	Propanoate metabolism Cysteine and methionine metabolism Pyruvate metabolism	1.28 (1.10-1.48)	$\mathbf{1.0 \times 10^{-3}}$	1.37 (1.17-1.59)	$\mathbf{6.2 \times 10^{-5}}$
sGP130	CHD GWAS	A signal transducer shared by many cytokines, including Interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF), and oncostatin M	None	1.02 (0.88-1.18)	0.8	1.34 (1.14-1.56)	$\mathbf{2.7 \times 10^{-4}}$

21

		(OSM)					
sICAM1	CHD GWAS	A cell surface glycoprotein	Complement and coagulation cascades	1.23 (1.06-1.42)	5.1×10^{-3}	1.33 (1.14-1.55)	**2.1×10^{-4}**
sRAGE	CHD GWAS	A member of the immunoglobulin superfamily of cell surface receptors	Antigen processing and presentation Type I diabetes mellitus Allograft rejection Graft versus host disease Autoimmune thyroid disease	1.03 (0.89-1.20)	0.7	1.18 (1.01-1.37)	0.04

*CHD events (n=213) included recognized myocardial infarction or death from CHD, and CVD death (n=199) included fatal CHD or death due to stroke, peripheral arterial disease, heart failure, or other CVD causes occurring during a median follow-up of 14.3 years (25th percentile 11.4, 75th percentile 15.2 years) among 3,520 Framingham Heart Study participants age ≥50 years.

**Proteins that tested positive in Mendelian randomization analysis for coronary heart disease risk at 0.05<p<0.10.

†For proteins with pQTLs that are enriched for more than five KEGG pathways, the top five most significant pathways based on enrichment p-value are shown.

††The p-value threshold for significance ($p < 3.3 \times 10^{-3}$) was determined by the Bonferroni method (0.05/15 proteins tested). Significant p-values are shown in **bold**.

Abbreviations: CHD = coronary heart disease; CI = confidence interval; CVD = cardiovascular disease; GWAS = genome-wide association study; KEGG = Kyoto Encyclopedia of Genes and Genomes; MR = Mendelian randomization

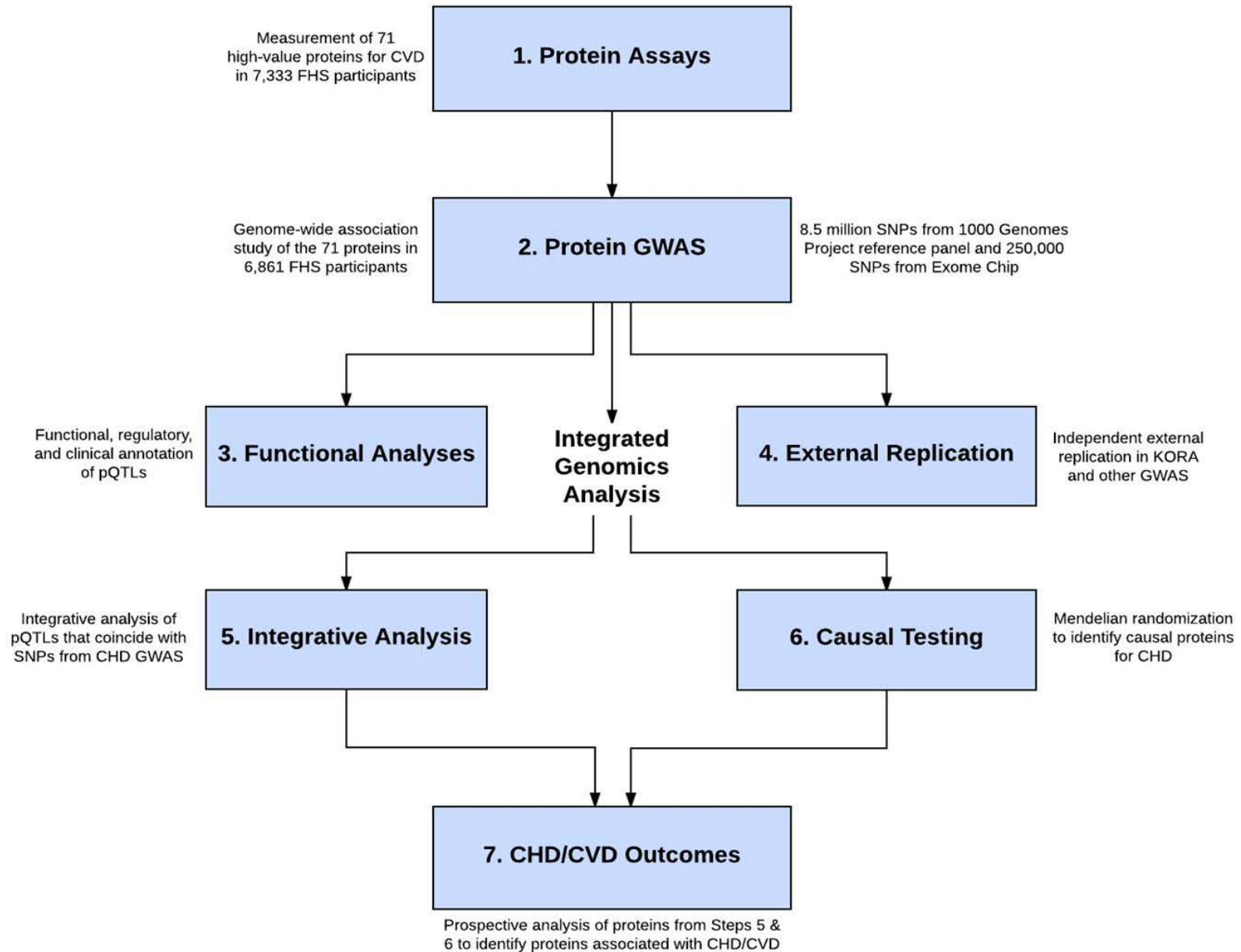**Figure 1. Flowchart of Study Design**

Figure 1. Flowchart of Study Design

The study consisted of seven steps: 1) selection and measurement of 71 high-value plasma proteins for atherosclerotic CVD via multiplex immunoassays in 7,333 FHS participants, 2) GWAS of the 71 proteins in 6,861 FHS participants to identify genome-wide significant pQTLs, 3) functional enrichment analyses of the identified pQTLs, 4) independent external replication of the sentinel pQTLs in KORA and other previous GWAS, 5) integrated analysis to pQTLs that coincide with CHD SNPs from GWAS, 6) identification of causal proteins for CHD using a Mendelian randomization approach, 7) association analysis of proteins from steps 5 and 6 with risk for incident CHD death and CVD death in 3,520 FHS participants age 50 years or older with available long-term follow-up.

Abbreviations: CHD = coronary heart disease; CVD = cardiovascular disease; FHS = Framingham Heart Study; GWAS = genome-wide association study; KORA = Cooperative Health Research in the Region of Augsburg Study; pQTL = protein quantitative trait locus (i.e. genetic variant associated with protein level); SNP = single nucleotide polymorphism

24

Figure 2. Sentinel *cis*- and *trans*-pQTLs and the Corresponding Proteins
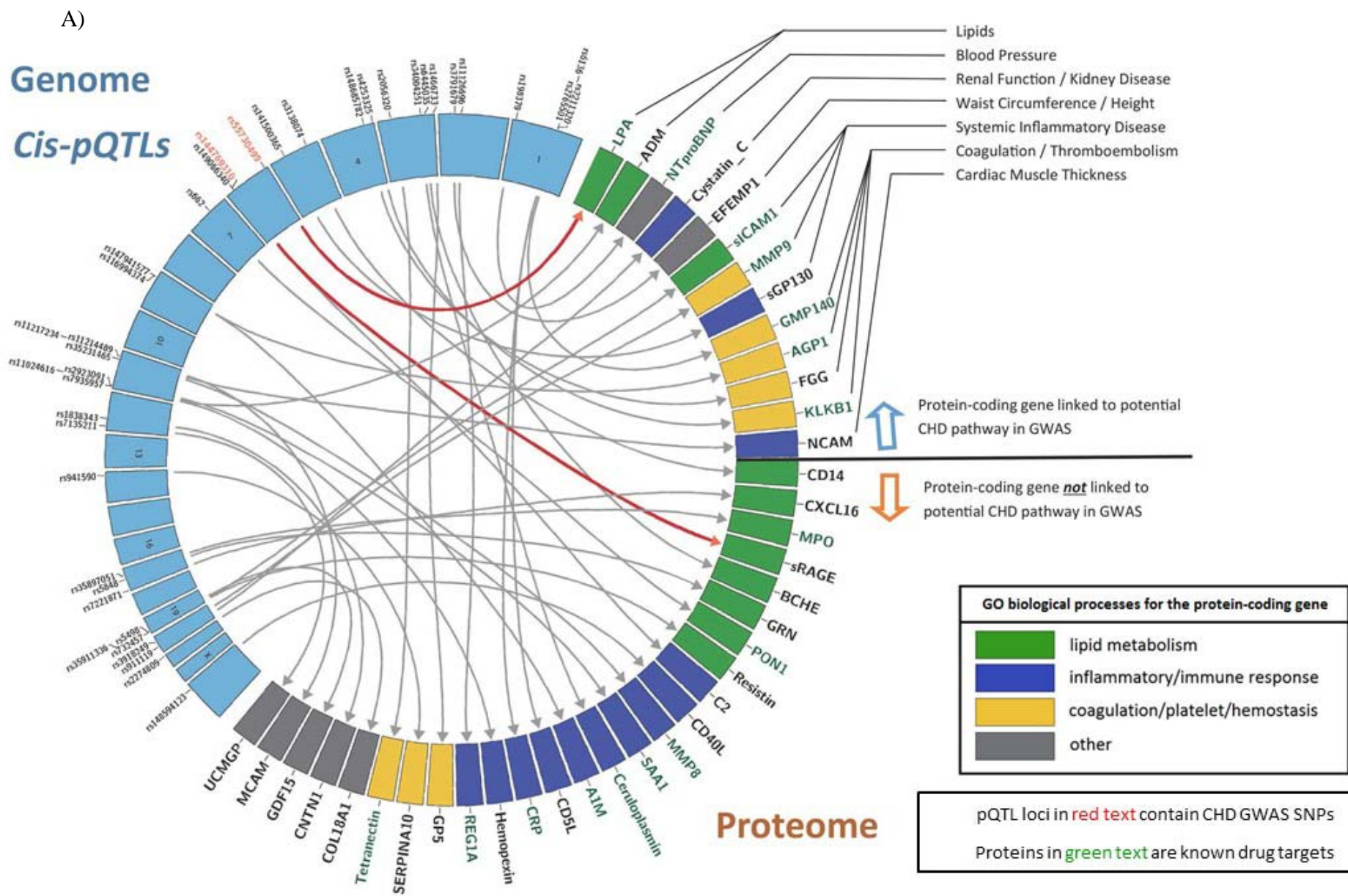

A)

Figure 2. Sentinel *cis*- and *trans*-pQTLs and the Corresponding Proteins.

Circos plots of sentinel *cis*- (Panel A) and *trans*-pQTLs (Panel B) and the plasma protein levels with which they are associated. Sentinel pQTLs are listed in order of chromosomal locations (blue boxes in the left semicircle). Loci containing pQTLs previously identified in GWAS to be associated with CHD are written in red text. Proteins with genome-wide significant pQTLs are listed in the right semicircle. The following three conditions are summarized for each protein: 1) The corresponding protein-coding gene is linked to a potential CHD pathway in previous GWAS (above the black line). 2) The corresponding protein-coding gene is a known drug target (green text). 3) GO biological processes for the protein-coding gene (green box denotes lipid metabolism pathways, blue box denotes inflammatory/immune response pathways, yellow box denotes coagulation/platelet/hemostasis pathways, and gray box denotes other pathways not included in the three most common, previously listed pathways). A single primary GO process was chosen when the protein-coding gene was included in multiple pathways.

Abbreviations: CHD = coronary heart disease; GO = Gene Ontology; GWAS = genome-wide association study; pQTL = protein quantitative trait locus (i.e. genetic variant associated with protein level); SNP = single nucleotide polymorphism

27

Figure 3. pQTL Minor Allele Frequency vs. Effect Size and Proportion of Variance Explained
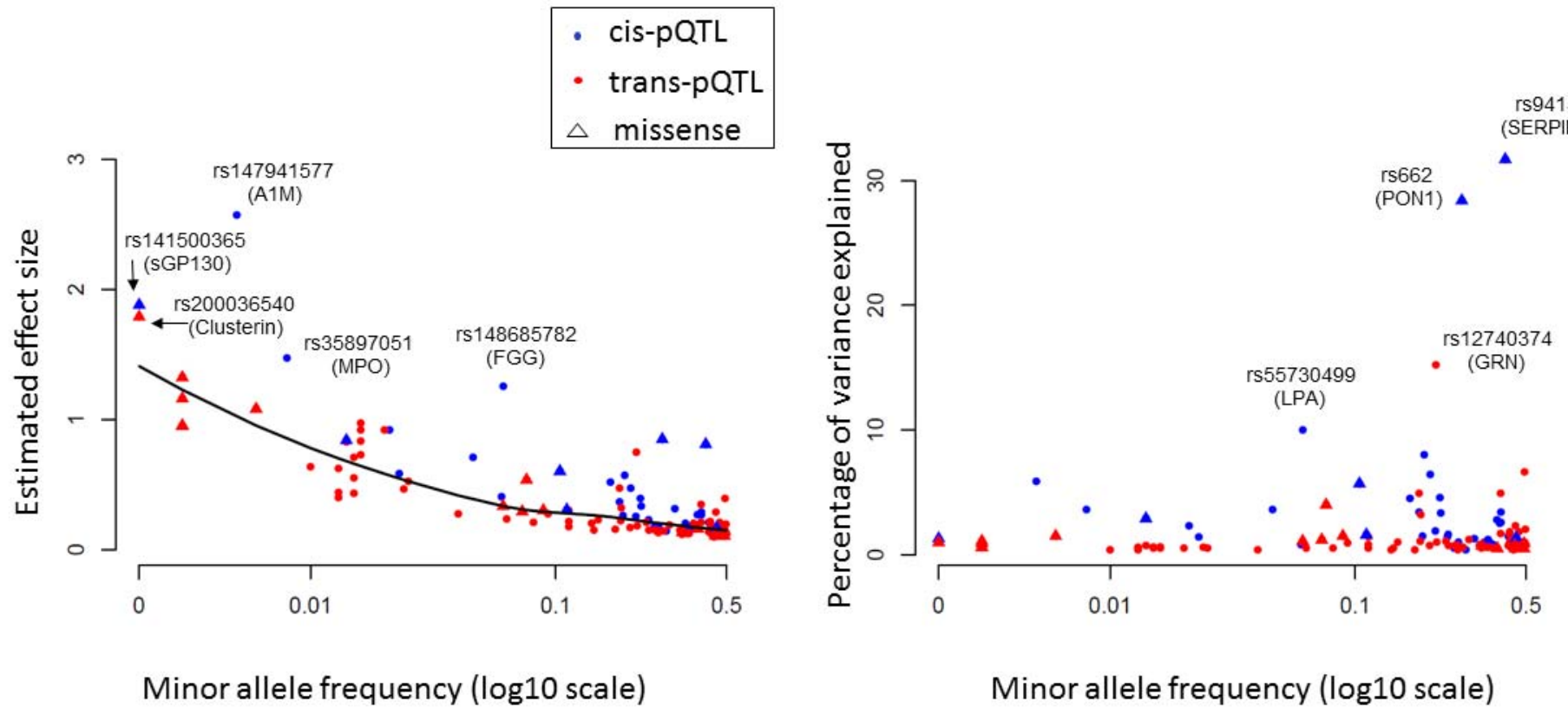
Figure 3. pQTL Minor Allele Frequency vs. Effect Size and Proportion of Variance Explained
Minor allele frequency of pQTLs (X-axis) vs. effect size of variants on proteins (left panel) and proportion of protein variance explained (right panel) for each sentinel pQTL locus.

Abbreviations: pQTL = protein quantitative trait locus (i.e. genetic variant associated with protein level)

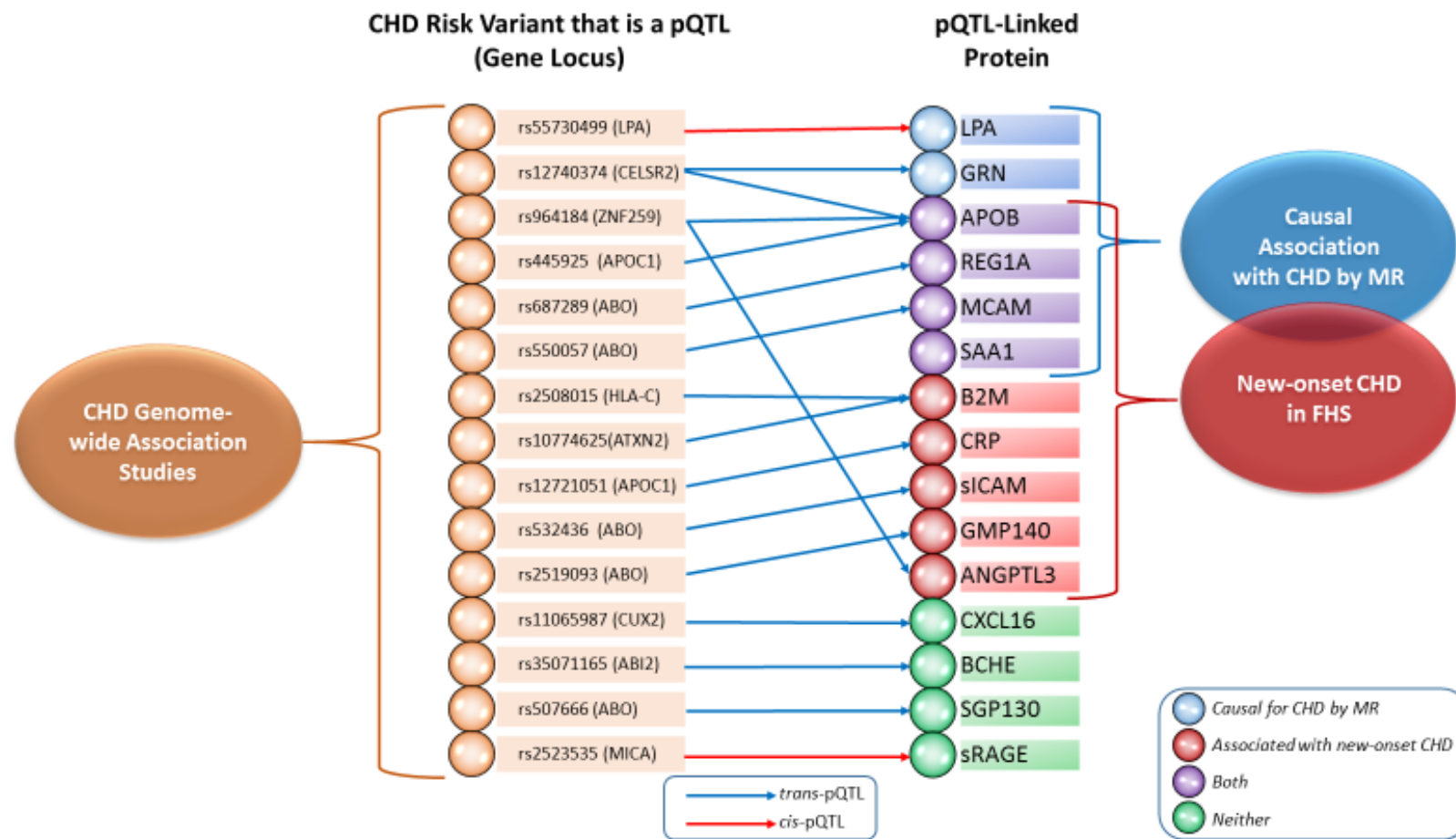**Figure 4. pQTL-Protein-Coronary Heart Disease Network**

30

Figure 4. pQTL-Protein-Coronary Heart Disease Network

Network of proteins and significant pQTLs with annotated gene loci for the pQTLs that are also GWAS risk SNPs for CHD (see Table 1). For proteins with multiple pQTLs that coincide with coronary heart disease GWAS SNPs, the pQTL with the lowest p-value of association with its corresponding protein level is shown. The following two conditions are summarized: 1) Proteins that tested causal for CHD in Mendelian randomization (p<0.05). 2) Proteins associated with new-onset CHD (p<0.05) in 3,520 Framingham Heart Study participants age 50 years or older with long-term follow-up. Proteins in green fulfill neither condition 1 nor 2; proteins in blue fulfill condition 1; proteins in red fulfill condition 2; proteins in purple fulfill conditions 1 and 2. The pQTL rs2523535 for sRAGE was reported to be associated with CHD (p=8.1x10-10) in a Japanese GWAS (PMID 21971053).

Abbreviations: CHD = coronary heart disease; FHS = Framingham Heart Study; MR = Mendelian randomization; pQTL = protein quantitative trait locus (i.e. genetic variant associated with protein level)

31

Figure 5. Comparison of Protein Effects on Coronary Heart Disease from Mendelian Randomization Estimate vs. Observed Hazards
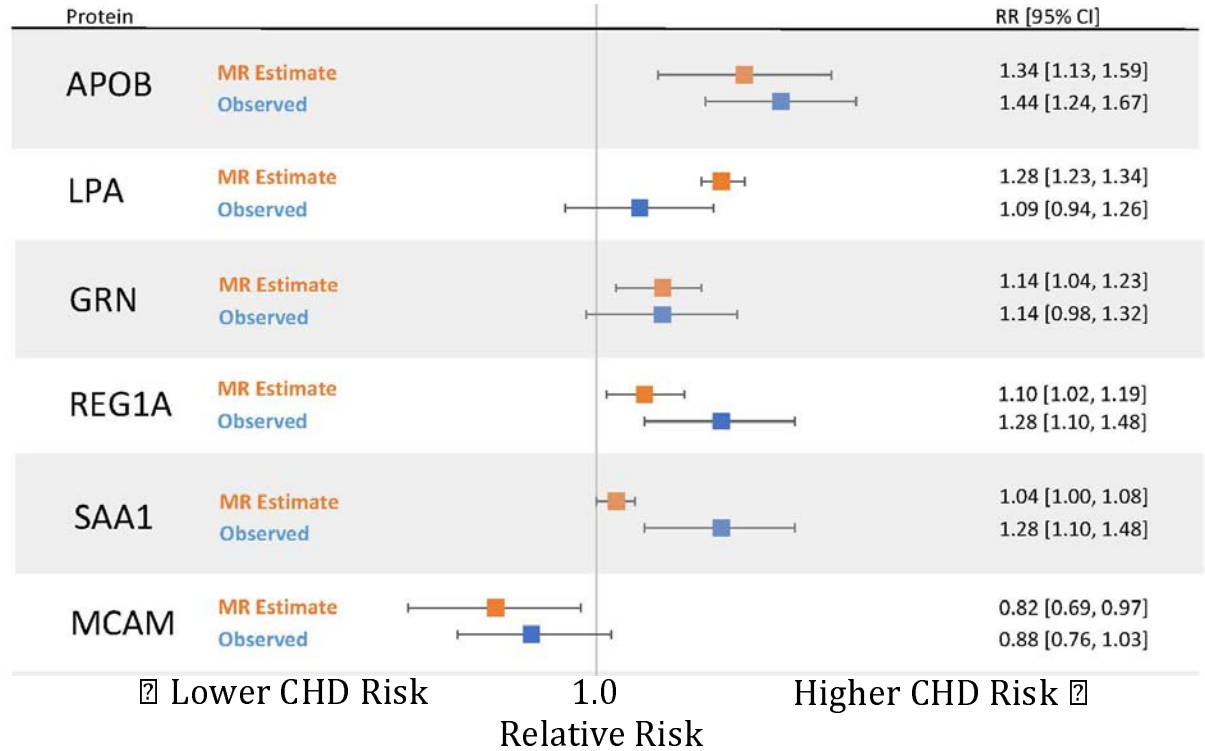
Figure 5. Comparison of Protein Effects on Coronary Heart Disease from Mendelian Randomization Estimate vs. Observed Hazards
Comparison of protein effects on CHD estimated from Mendelian randomization versus the observed hazards in 3,520 Framingham Heart Study participants with long-term follow-up.

Abbreviations: CHD = coronary heart disease; CI = confidence interval; MR = Mendelian randomization; RR = relative risk

Methods

Study Design: The study consisted of seven steps (Figure 1): 1) selection and measurement of 71 high-value plasma proteins for atherosclerotic CVD via multiplex immunoassays in 7,333 FHS participants, 2) genome-wide association study of the 71 proteins in 6,861 FHS participants to identify genome-wide significant pQTLs, 3) functional enrichment analyses of the identified pQTLs, 4) independent external replication of the sentinel pQTLs in KORA, 5) integrative analysis to pQTLs that coincide with CHD SNPs from GWAS, 6) identification of causal proteins for CHD using a Mendelian randomization approach, 7) association analysis of proteins from steps 5 and 6 with risk for incident CHD death and CVD death in 3,520 FHS participants age 50 years or older with available long-term follow-up.

Discovery Study Sample: The FHS is a community-based prospective study of CVD and its risk factors that recruited three generations of participants within families in 1948, 1971, and 2002, respectively.[48-50] The study samples for this investigation were collected from 7,333 participants from the FHS Offspring (Exam 7; 1998-2001) and Third Generation (Exam 1; 2002-2005) cohorts. The final sample for GWAS was composed of 6,861 participants with complete imputed dosage data based on the 1000 Genomes Project reference panel (1000G).[17] For association analyses using Exome Chip genotypes (see Genotyping for details), the sample size was 6,763. Genome-wide analysis of SNPs associated with gene expression levels (eQTLs) was performed on 5,257 FHS participants in whom both genotype and gene expression data were available.[23]

Replication Study Sample: The KORA F4 study is a prospective population-based cohort study consisting of 3,080 participants living in Augsburg, Southern Germany.[12,51] A total

34

of 1,000 participants who also participated in a metabolomic study with follow-up information for aging-related diseases composed the study population for replication. After excluding participants with missing genotype or protein data (n=3), the final KORA sample included 997 individuals.

Power Calculation: For power in the discovery stage with n=6,800, we assumed an additive genetic model with no interaction and a population mean=0 and standard deviation=1 for all rank-normalized protein levels. At $\alpha=5 \times 10^{-8}$ for a two-sided test, power was estimated for MAF=0.002, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5 with QUANTO.[52] For empirical power in the replication stage, we performed pQTL analysis with 1,000 resamplings of 1,000 unrelated FHS participants. We counted the number of tests with $p<0.05/n$ in the 1,000 resamplings, where n is the number of pQTLs that tested for replication in KORA.

Clinical Measures: All FHS participants underwent periodic clinical examinations with standard protocols as described previously.[50] A three-physician panel was formed to perform medical chart review weekly. The review panel jointly assigned CVD diagnoses and causes of death. All suspected CVD events were adjudicated by the physician-panel after reviewing all available medical evidence including hospital records, personal physician records, and interviews with next of kin in the event of an out-of-hospital death. Recognized myocardial infarction (MI) was diagnosed when two of three of the following conditions were present: prolonged chest discomfort or symptoms of coronary ischemia, elevated biomarkers of myocardial necrosis (e.g. CK-MB or troponin), and the development of new diagnostic Q-waves on the ECG. Fatal CHD events included fatal MI and other deaths due to CHD as an underlying cause in the absence of evidence of

recent infarction. Fatal CVD events additionally included deaths due to stroke, peripheral arterial disease, heart failure, or other cardiovascular causes.

Protein Quantification: FHS fasting blood plasma samples were collected and stored at -80°C. Candidate protein biomarkers were selected *a priori* based on previous evidence of association with atherosclerotic CVD or its risk factors using the following complementary approaches: a) comprehensive literature search,[53] b) proteomics discovery via mass spectrometry in the FHS or elsewhere,[36,54] and c) targeting proteins coded by genes identified via gene expression profiling studies[55,56] or GWAS[57] of atherosclerotic CVD and its risk factors. A total of 85 plasma protein biomarkers were assayed using a modified enzyme-linked immunosorbent assay sandwich method, multiplexed on a Luminex xMAP platform (Luminex, Inc., Austin, TX). All targets were first developed as singleton assays before compatible targets were pooled to create multiplex panels. Standard Luminex assays with previously published methods were used.[58,59] Measurements were calibrated using a seven-point calibration curve (in triplicate) and tested for recovery at both ends of the quantitation scale. The 'High' and 'Low' spike controls (QC1 and QC2 respectively) were used to calculate intra- and inter-assay coefficients of variation (CV) for each protein. A total of 14 proteins had low call-rate (<90%) mainly due to values falling below the lower detection limit that were excluded for the current study. A list of the 71 proteins and their coefficients of variation and selection criteria were shown in Table S1.

For the KORA study, plasma levels of 1,129 proteins in 1,000 blood samples were measured using the SOMAscan platform (SomaLogic Inc., Boulder, Colorado), a multiplexed aptamer-based affinity proteomics platform; 1,124 proteins passed quality

control. Protein measurement protocol, normalization of protein values, and data quality are described elsewhere.[12]

Genotyping: Genotyping and QC methods in the FHS have previously been described.[17] In brief, genome-wide genotyping was conducted using the Affymetrix 500K mapping arrays and 50K supplemental Human Gene Focused arrays (Affymetrix, Inc., Santa Clara, CA) as well as the Illumina Human Exome BeadChip v.1.0 (Exome Chip; Illumina, Inc., San Diego, CA). Genotypes from the Affymetrix arrays were used in conjunction with the 1000G reference panel[17] to generate an imputed set of ~30 million variants using MACH.[60] SNPs with imputation quality ratio <0.3 (imputation quality ratio is calculated by the ratio of the variances of the observed and the estimated allele counts) or minor allele frequency (MAF) <0.01 were excluded, leaving a final set of 8,509,364 SNPs for 1000 genomes imputed GWAS.

The Exome Chip includes rare coding variants not covered by previous genotyping arrays.[61] More than 90% of the SNPs included in the Exome Chip are non-synonymous variants, splice variants, or stop codon altering variants. Common variants on the Exome Chip include 5,542 SNPs that were selected based on their associations with disease traits reported in the NHGRI GWAS Catalog.[1] Rare variants with $MAF < 1 \times 10^{-4}$ were excluded from analysis.

For KORA, the Affymetrix Axiom Array (Affymetrix, Inc., Santa Clara, CA) was used to genotype 3,788 study participants.[12,51] Genotypes were then imputed from the 1,000G reference panel[15] and used for lookup of the replication targets.

Functional Annotation of pQTLs: We used HaploReg[22] v4.1 to functionally annotate our pQTL results. Using information from the Roadmap Epigenomics[62] and ENCODE

projects,[63] HaploReg linked SNPs and small insertions/deletions with chromatin state, protein binding annotation, and regulatory motifs. A total of 14,756 pQTLs could be found in the HaploReg database. We used DEPICT[20] to conduct gene prioritization, pathway analysis, and tissue/cell type enrichment analysis. DEPICT used information from co-regulation of gene expression from 77,840 samples, in conjunction with 14,461 reconstituted functional gene sets, to assess pathway enrichment and prioritize genes. In addition, DEPICT utilized a set of 37,427 human microarrays to identify enrichment of highly expressed genes in specific tissue/cell types. We used Functional Mapping and Annotation[21] of GWAS (FUMA; http://fuma.ctglab.nl) to categorize proteins based on known pathways and conduct functional annotation of pQTLs (regional plot of each pQTL locus, functional categorization of pQTL SNPs, gene mapping, and pathway enrichment analyses).

Gene Expression: Gene expression profiling was conducted using the Affymetrix Human Exon 1.0 ST GeneChip platform (Affymetrix Inc., Santa Clara, CA), comprised of >5.5 million probes covering expression of 17,873 mRNA transcripts. Gene expression values were normalized and adjusted for three technical covariates (batch, first principal component, and residual of probeset mean values) as described previously.[23]

Coronary Heart Disease-associated SNPs: The CARDIoGRAMplusC4D Consortium[1] GWAS of CHD yielded 1,892 genome-wide significant SNPs (at $p<5\times10^{-8}$) from 1000G imputation. The National Human Genome Research Institute (NHGRI) GWAS catalog[2] (downloaded in July 2016) and Genome-wide Repository of Associations Between SNPs and Phenotypes (GRASP)[3,4] v.2.0 (downloaded in June 2016) included 846 SNPs associated in GWAS with CHD at genome-wide significance level ($p<5\times10^{-8}$).

<u>Statistical Methods</u>: Statistical analyses in the FHS were performed using R software version 3.1.1[64] or SAS software version 9.4.

Genome-wide association (pQTL) analyses: Linear mixed effects models (the "LMEKIN" function of Kinship Package in R[64]) were used to test associations of inverse-rank normalized protein levels with 1000G or Exome Chip variants in conjunction with an additive genetic model. We applied a p-value threshold of 5×10^{-8} for defining significant pQTLs. A *cis*-pQTL was defined as a SNP residing within 1 megabase (Mb) upstream or downstream of the transcription start site of the corresponding protein-coding gene. A SNP located >1 Mb upstream or downstream of the gene transcript or on a different chromosome from its associated gene was categorized as a *trans*-pQTL.

Linkage disequilibrium (LD) was computed as the square of Pearson's correlation (r^2) between imputed additive dosages of genotypic variants within the same chromosome across 8,481 FHS individuals with genotype data. Independent pQTLs for a given protein were defined as those with LD $r^2 < 0.2$ with other pQTLs at a genomic locus. For a genetic locus with multiple pQTLs in LD (*i.e.,* LD $r^2 > 0.2$), we selected the pQTL with the lowest p-value to represent the sentinel pQTL for that locus.

For KORA, linear regression models were performed on the follow-up SNPs using R version 3.1.3.[64] Associations between inverse-normalized protein levels and imputed dosages were tested using linear additive genetic regression models adjusted for age, sex, and body mass index.[12]

eQTL Mapping: We used linear mixed effects models, accounting for familial relationships using "PEDIGREEMM" in R,[64] to assess associations between ~8.5 million 1000G SNPs that were additively coded and expression levels of 17,873 transcripts.[23]

39

Models were adjusted for age, sex, platelet count, differential white cell count (percentages of lymphocyte, monocyte, eosinophil, and basophil), and for 20 PEER factors[65,66] to reduce confounding due to unmeasured factors. The criteria used to define *cis* and *trans* effects for pQTLs were also applied to eQTLs. A false discovery rate (FDR) threshold of 0.05 was applied separately for *cis*- and *trans*-eQTLs.

Mendelian randomization: We used an MR approach to test for causal associations between protein biomarkers and CHD risk. The sentinel *cis*-pQTL for each protein, based on lowest p-value of association in either 1000G GWAS or Exome Chip analysis, was selected as the instrumental variable (IV) for its perspective protein in MR analysis. Based on the association between the sentinel *cis*-pQTL and CHD in prior GWAS,[1] a putative causal effect of one standard error difference in inverse-rank normalized protein level on CHD was calculated as the per risk allele effect on CHD risk dependent on the per risk allele effect on one standard error difference in inverse-rank normalized protein level (the Wald ratio test).[67] For proteins with suggestive single *cis*-pQTL results ($0.05<p<0.1$) and with additional non-redundant *cis*-pQTLs, we conducted single-locus multi-SNP MR. Low-level correlation (LD $r^2<0.2$) between variants in the genetic risk score was adjusted for in MR analysis using the method developed by Burgess et al.[68] Similarly, for proteins with pQTLs that shared genetic signals with CHD from GWAS, we conducted multi-SNP MR using MRbase[30] when there were at least four non-redundant pQTL loci.

Associations of protein levels with CVD: To analyze associations between plasma protein levels and MI/CHD death and CVD death in FHS participants, protein biomarkers were rank-normalized. Cox proportional hazard models were used to predict MI/CHD death

40

and CVD death for each biomarker, adjusting for age and sex. Participants younger than 50 years of age at baseline were excluded from outcome analyses due to a paucity of events in this age group. In addition, participants with prevalent MI/CHD or CVD at baseline were excluded from analyses of incident events, leaving a final sample size of 3,520 FHS participants.

Independent External Replication: After merging our 1000G and Exome Chip GWAS results, the pQTL with the lowest p-value of association at each genetic locus was selected as the sentinel pQTL. We conducted independent external replication of our sentinel pQTLs in the KORA study[12] and in other protein GWAS. Out of the 60 proteins with pQTL SNPs in the FHS, replication was conducted for 47 proteins from discovery that also were measured in KORA or other studies. The sentinel pQTL at each genetic locus in the FHS was determined to be successfully validated if its corresponding 1000G-imputed genotype or strong proxy (LD $r^2>0.8$) in KORA was also a significant pQTL for the corresponding protein and if directionality of pQTL-protein association was preserved. Statistical significance was defined as a p-value $<0.05/n$ (n was the number of pQTLs that were studied in KORA).

Study Approval: All participants from the FHS and KORA study gave informed consent for participation in this study and for the collection of plasma and DNA for analysis. The KORA study was approved by the Ethics Committee of the Bavarian Medical Association, Germany.

Data Access: All data from the FHS for this study are accessible (dbGaP Study Accession: phs000363.v16.p10). Data for KORA are available upon request from

41

KORA-gen (http://epi.helmholtz-muenchen.de/kora-gen). Requests are submitted online

and are subject to approval by the KORA board.

Methods References

48. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 1951;41:279-81.
49. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. Prev Med 1975;4:518-25.
50. Splansky GL, Corey D, Yang Q et al. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. Am J Epidemiol 2007;165:1328-35.
51. Wichmann HE, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. Gesundheitswesen. 2005;67 Suppl 1:S26-30.
52. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med. 2002;21(1):35-50.
53. Anderson L. Candidate-based proteomics in the search for biomarkers of cardiovascular disease. J Physiol (Lond). 2005;563(Pt 1):23-60.
54. Prentice RL, Paczesny S, Aragaki A, et al. Novel proteins associated with risk for coronary heart disease or stroke among postmenopausal women identified by in-depth plasma proteome profiling. Genome Med.2010;2:48.
55. Joehanes R, Ying S, Huan T, Johnson AD, Raghavachari N, Wang R, Liu P, Woodhouse KA, Sen SK, Tanriverdi K, et al. Gene expression signatures of coronary heart disease. Arterioscler Thromb Vasc Biol. 2013;33:1418-1426.
56. Huan T, Esko T, Peters MJ, Pilling LC, Schramm K, Schurmann C, et al. A meta-analysis of gene expression signatures of blood pressure and hypertension. PLoS Gen. 2015;11(3):e1005035.
57. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011;43(4):333-8.
58. Dupont NC, Wang K, Wadhwa PD, Culhane JF, Nelson EL. Validation and comparison of luminex multiplex cytokine analysis kits with ELISA: determinations of a panel of nine cytokines in clinical sample culture supernatants. J Reprod Immunol. 2005;66(2):175-91.
59. Carson RT, Vignali DA. Simultaneous quantitation of 15 cytokines using a multiplexed flow cytometric assay. J Immunol Methods. 1999;227(1-2):41-52.
60. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
61. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS ONE. 2013;8(7):e68095.
62. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010;28(10):1045-8.
63. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
64. Team RC. R: A language and environment for statistical computing. 2012.

65. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500-7.
66. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6(5):e1000770.
67. Fall T, Hägg S, Mägi R, et al. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. PLoS Med. 2013;10(6):e1001474.
68. Burgess S, Dudbridge F, Thompson SG. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. Stat Med. 2016;35(11):1880-906.