High-throughput multivariable Mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease

Verena Zuber[1,2], Dipender Gill[1], Mika Ala-Korpela[3,4], Claudia Langenberg[5], Adam Butterworth[6,7,8,9,10,11], Leonardo Bottolo[12,13,2], and Stephen Burgess[2,6]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[2]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK
[3]Computational Medicine, Faculty of Medicine, University of Oulu & Biocenter Oulu, Oulu, Finland
[4]NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland
[5]MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK
[6]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
[7]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK
[8]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK
[9]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK
[10]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
[11]Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK
[12]Department of Medical Genetics, School of Clinical Medicine, University of Cambridge, Cambridge, UK
[13]The Alan Turing Institute, London, UK

February 10, 2020

Background: Genetic variants can be used to prioritize risk factors as potential therapeutic targets via Mendelian randomization (MR). An agnostic statistical framework using Bayesian model averaging (MR-BMA) can disentangle the causal role of correlated risk factors with shared genetic predictors. Here, our objective is to identify lipoprotein measures as mediators between lipid-associated genetic variants and coronary artery disease (CAD) for the purpose of detecting therapeutic targets for CAD.

Methods: As risk factors we consider 30 lipoprotein measures and metabolites derived from a high-throughput metabolomics study including 24,925 participants. We fit multivariable MR models of genetic associations with CAD estimated in 453,595 participants (including 113,937 cases) regressed on genetic associations with the risk factors. MR-BMA assigns to each combination of risk factors a model score quantifying how well the genetic associations with CAD are explained. Risk factors are ranked by their marginal score and selected using false discovery rate (FDR) criteria. We perform sensitivity and replication analyses varying the dataset for genetic associations with CAD.

Results: In the main analysis, the top combination of risk factors ranked by the model score contains apolipoprotein B (ApoB) only. ApoB is also the highest ranked risk factor with respect to the marginal score (FDR$<$ 0.005). Additionally, ApoB is selected in all replication analyses. No other measure of cholesterol or triglyceride is consistently selected otherwise.

Conclusions: Our agnostic genetic investigation prioritizes ApoB across all datasets considered, suggesting that ApoB, representing the total number of hepatic-derived lipoprotein particles, is the primary lipid determinant of CAD.

Wordcount: 249/250

Key messages

- It is a common consensus that lipoproteins increase cardiovascular disease risk, yet little is known about the exact mechanisms.
- We use genetic associations with high-throughput metabolomics features to draw a detailed picture of lipid traits and characteristics allowing for an unprecedented resolution when considering lipids as risk factors for cardiovascular disease.
- This study integrates genetic data from a large scale metabolomics study including 25,000 samples and the largest study on cardiovascular disease risk including 113,937 cases and 339,658 controls.
- MR-BMA, a novel algorithm for multivariable MR (Zuber and Burgess, Nature Communications 2020; 11(1):29) is used to identify the most likely causal lipid determinants

of cardiovascular disease from a large set of candidate risk factors with shared genetic predictors.

- Our agnostic genetic investigation prioritizes apolipoprotein B across all datasets considered, suggesting that apolipoprotein B, representing the total number of hepatic-derived lipoprotein particles, is the primary lipid determinant of cardiovascular disease risk.

Introduction

Genetic variants have the potential to contribute greatly to our understanding of mechanisms underlying disease processes, and to guide target validation for pharmacological and clinical interventions that reduce disease risk [1]. Coronary artery disease (CAD) is the most common cause of death globally. While it has been shown that genetic variants predisposing individuals to higher levels of low-density lipoproteins (LDL)-cholesterol also associate with increased CAD risk [2], genetic variants predisposing individuals to higher levels of high-density lipoproteins (HDL)-cholesterol are not associated with CAD risk [3] after accounting for other lipid traits. These genetic analyses may suggest that LDL-cholesterol is a causal risk factor for CAD risk, but HDL-cholesterol is not − as has generally been observed in clinical trials of lipid-altering therapies [4, 5, 6]. Genetic studies have also suggested that triglyceride levels are an independent risk factor for CAD risk [7]. Triglycerides are another component of body fat which are transported by lipoprotein particles, and in particular by very low density lipoproteins (VLDL). However, two recent studies showed that genetic associations with CAD risk are proportional to the change in apolipoprotein B (ApoB), the primary protein component of VLDL, LDL, and intermediate-density lipoprotein (IDL) particles, and that LDL-cholesterol and triglycerides do not appear to be independent risk factors for CAD after accounting for ApoB [8, 9].

Genome-wide association studies (GWAS) are increasingly used to combine genomic profiling with high-throughput molecular measures on a large scale, including tens of thousands of samples, to explore the genetic regulation of molecular processes. For example, Kettunen et al. have combined high-throughput metabolomics with genomic profiling on nearly 25 000 individuals [10]. Given the large sample size, these studies are well powered to explore the causal role of molecular mechanisms. The metabolomics study by Kettunen et al. was conducted using nuclear magnetic resonance (NMR) spectroscopy to provide a detailed characterization of lipid-related traits, including 14 size categories of lipoprotein particles ranging from small HDL to extra-extra-large VLDL. For each lipoprotein category, measurements are available of cholesterol, triglycerides, cholesterol ester, and phospholipid content. Additional mean diameter of the lipoprotein particles is measured for some lipoprotein size categories. Measurements also include apolipoprotein A1 and ApoB, sphingomyelins, fatty acids, and phosphoglycerides (Supplementary Table A1).

Previous MR studies on lipid determinants for CAD risk have included only a few curated lipid traits at a time [8, 9]. In this study, we build on a high-throughput metabolomics data resource [10] to investigate a much wider set of lipoprotein measurements as candidate risk factors for CAD. We use a recently published algorithm called Mendelian randomization with

Bayesian model averaging (MR-BMA) [6] that applies principles from high-dimensional data analysis and machine learning to detect causal risk factors from a large set of candidate risk factors. Our goal is to select the lipoprotein measures that are the most likely causal risk factors for CAD.

Methods

Variable selection method for finding likely causal risk factors

We provide a brief outline of the MR-BMA method here. More details are given in the Supplementary Material, and a diagram illustrating the method is shown in Figure 1.

We consider each set of risk factors in turn: all single risk factors, all pairs of risk factors, all triples, and so on. For each set of risk factors, we undertake a multivariable MR analysis using weighted regression based on summarized genetic data. We assess goodness-of-fit in the regression model, and assign a score to the risk factor set that is the model posterior probability of that set being the true causal risk factors. We repeat this to get a posterior probability for all models (i.e. all sets of risk factors). Then, for each of the candidate risk factors, we sum up the posterior probability over models including that risk factor to compute the marginal inclusion probability for the risk factor, representing the probability of that risk factor being a causal determinant of disease risk. We also calculate the model-averaged causal effect, representing the average causal effect across models including that risk factor. P-values are calculated for each risk factor using a permutation method, with adjustment for multiple testing via the Benjamini and Hochberg false discovery rate (FDR) procedure [10].

Study design

A summary of our study design is given in Figure 2. The three key steps in designing a two-sample multivariable MR study are instrument selection, risk factor selection, and the choice of outcome data, including main and replication analysis.

Selecting lipid-associated variants as instrumental variables

We took an initial list of 185 variants associated with blood lipids (LDL-cholesterol, HDL-cholesterol or triglycerides) in the Global Lipid Genetics Consortium at a genome-wide level of significance ($p < 5 \times 10^{-8}$) [7] which was pruned at a linkage disequilibrium threshold of $r^2 < 0.05$, and further refined by genomic distance, excluding variants that are less than 1 megabase pair apart, to provide a list of $n = 150$ genetic variants. We selected these lipid-associated genetic variants as instrumental variables because we wanted to investigate lipid determinants of CAD risk. This is important to keep in mind when interpreting the results as the prioritization of risk factors by MR-BMA is conditional on the genetic variants selected as instrumental variables. There are two direct consequences of this choice. Firstly, this choice of instrumental variables will downweight non-lipid risk factors, and so results should not be interpreted as evidence that those risk factors are not on the causal path to CAD. Secondly, basing the selection of instrumental variables on an external dataset (e.g. the Global Lipid Genetics Consortium) reduces the risk of winner's course [13].

Lipoprotein measures as risk factors

Genetic associations with lipoprotein measures and metabolites are taken from Kettunen et al. [10] who measured 118 variables on 24,925 European individuals using the high-throughput Nightingale NMR platform. Estimates were obtained by linear regression of each NMR measurement on each of the genetic variants in turn, with adjustment for age, sex, time from last meal (if available), and ten genomic principal components. NMR measurements were inverse rank-based normal transformed, so that association estimates are presented in standard deviation units for the relevant risk factor throughout.

Several measurements from the Nightingale platform were highly correlated, judged by the correlation between the genetic associations for the 150 genetic variants. While MR-BMA was able to identify the causal risk factors reliably in a simulation study when risk factors were highly correlated (up to $|r| = 0.99$) [6], several risk factors were more highly correlated than this. We therefore pruned the set of risk factors to avoid inaccurate results due to collinearity. For each lipoprotein diameter category representing the size of lipoproteins, we retained only the measurement of cholesterol and/or triglyceride content, and mean particle diameter where available. We also included only total fatty acid content and not other fatty acid measurements, as genetic predictors that were able to distinguish reliably between these risk factors were not included as instruments. Other non-lipoprotein metabolite measurements were retained in the analysis as they had substantially weaker correlations with lipoprotein measurements, and so would only be selected by MR-BMA if they mediated CAD risk from the genetic predictors included in the model. No pair of risk factors included in the final analysis were more highly correlated than $|r| = 0.99$ (see correlation heatmap in Supplementary Figure A1). Finally, we only included risk factors into the MR analysis that had at least one genetic variant that was a strong predictor (genome-wide significant). The final list of 30 lipoprotein measures and metabolites included in the analysis is provided in Supplementary Table A1.

Coronary artery disease as outcome

Our primary analysis was based on genetic associations with CAD risk taken from the 2017 CARDIoGRAMplusC4D data release meta-analysed together with UK Biobank [14] including 453,595 individuals mostly of European descent, of whom 113,937 had a CAD event. Genetic association estimates with CAD risk were obtained in each study of the CARDIoGRAMplusC4D consortium by logistic regression with adjustment for at least five genomic principal components, and then meta-analysed across studies. There was one rare genetic variant (rs1998013, effect allele frequency 0.8%, in the *PCSK9* gene region) and one common intergenic genetic variant (rs894210, effect allele frequency 43.5%) for which there was no association estimate with CAD available. After excluding the missing genetic variants, we performed MR-BMA with 148 variants and 30 risk factors.

As sensitivity analyses, we repeated the same analysis steps on the 2017 CARDIoGRAMplusC4D data release except: 1) we omitted the variant in the *APOB* gene region from the analysis, to assess whether this variant was overly influential in determining the top ranked models and 2) we omitted the ApoB measurement from the list of risk factors to see if any other risk factor reached a similar level of evidential support. If it is the case that ApoB was selected as representative for a group of highly correlated traits, then upon removal of ApoB another risk factor of this group should be selected as representative instead.

As replication, we considered 1) an earlier release of CARDIoGRAMplusC4D consortium [15] including 60,801 CAD cases and 123,504 controls of European descent, but not including UK Biobank participants and 2) a UK Biobank GWAS on CAD outcomes which includes 29,278 cases and 338,425 controls of European descent (defined by self-report and genomic principal components). Quality control procedures were performed and related individuals were excluded from the analysis as described previously [16]. For the main and the two replication analyses we report the results including all variants and after excluding genetic variants that are influential points and outliers.

Results

Main analysis using outcome data from CARDIoGRAMplusC4D and UK Biobank

Results are provided in Table 1. We show the top 10 models (i.e. sets of risk factors) ranked according to their model posterior probability, and the top 10 risk factors according to their marginal inclusion probability. We also present the model-averaged causal effect estimate for each risk factor. The top-ranked model contains ApoB and no additional risk factors (model posterior probability 0.464). ApoB is also the risk factor with the strongest overall evidence (marginal inclusion probability 0.868, FDR$<$ 0.005). A diagnostic scatterplot of the genetic associations with the outcome against the genetic associations with ApoB is given in Figure 3. Our primary analysis was performed after model diagnostics, which removed influential and outlying genetic variants from the analysis. Similar results were obtained including all variants in the analysis (Supplementary Table A2).

Sensitivity analysis

As sensitivity analyses, we first repeated the primary analysis excluding the genetic variant in the *APOB* gene region, to ensure that this variant was not driving the selection of ApoB as a risk factor. This exclusion did not impact the results (Supplementary Table A3) – ApoB remained the highest ranking individual model (model posterior probability 0.455) and the risk factor with the strongest marginal evidence (marginal inclusion probability 0.862). Secondly, we repeated the primary analysis excluding ApoB from the list of risk factors. No alternative risk factor had similar strength of evidence, suggesting that ApoB is indeed the most important risk factor and not just a representative of a group of highly correlated lipoprotein measures with similar evidence. On exclusion of ApoB, the top risk factors were triglycerides content in small HDL particles (marginal inclusion probability 0.461, FDR$<$ 0.05) and LDL cholesterol (marginal inclusion probability 0.417, FDR$<$ 0.05). Yet, the evidence for these two lipoprotein measures is much weaker compared to the evidence for ApoB in the main analysis.

Replication analysis

As replication, we used genetic associations with CAD risk from two alternative datasets. Results are shown in Table A5. For the earlier release of CARDIoGRAMplusC4D [15], the top ranked model includes ApoB alone (model posterior probability 0.455), and ApoB is the top ranked risk factor overall (marginal inclusion probability 0.673, FDR$<$ 0.005). For UK Biobank, ApoB (marginal inclusion probability 0.325, FDR$<$ 0.05) was ranked second after triglycerides in very small VLDL-cholesterol (marginal inclusion probability 0.456, FDR$<$ 0.01). When looking at the individual models, triglycerides content in very small VLDL-cholesterol particles is ranked first followed by models including both ApoB and a measure of triglycerides content, suggesting an additional causal pathway via triglycerides when deriving genetic associations from UK Biobank analysis.

Discussion

Our results add to the growing evidence that ApoB is the primary causal determinant for CAD risk amongst lipoprotein measurements [17, 18, 19]. These results do not invalidate LDL-cholesterol as a causal risk factor for CAD risk. Indeed, LDL particles contain an apolipoprotein B molecule, as do IDL and VLDL particles. ApoB (in particular ApoB-100) represents the total number of hepatic-derived lipoprotein particles [20]. However, this investigation suggests clinical benefit of lowering triglyceride and LDL-C levels is proportional to the absolute change in ApoB. ApoB measurements are independent of particle density, and are not affected by heterogeneity of particle cholesterol content [21]. This is particularly important for accurately capturing the number of small dense LDL particles, which are believed to be associated with atherosclerosis. ApoB has been shown to be a superior measure to LDL-cholesterol in the prediction of CAD risk [22], and in prediction of coronary artery calcification [23]. From a clinical perspective, statins target LDL-cholesterol levels rather than ApoB, suggesting that greater benefit might be obtained from lipid-lowering drugs that target lipoprotein particle number [24]. When analysing data from UK Biobank only, there was also some evidence for triglyceride content measures as an additional risk factor. This was not evident in the main analysis or the replication analysis including data from the earlier CARDIoGRAMplusC4D release. This finding should therefore be interpreted with some caution.

There are some caveats to the interpretation of the results of this study. Although we were able to distinguish between measures of cholesterol content and triglyceride content for some categories of lipoprotein particles, we were not able to distinguish between other lipoprotein measures, such as cholesterol ester and phospholipid content, which correlated almost perfectly with cholesterol content.

In conclusion, our agnostic investigation to identify risk factors for CAD strongly prioritized ApoB, suggesting that ApoB, representing the number of hepatic-derived lipoprotein particles, is the key determinant of CAD risk amongst lipid-related measurements. This analysis demonstrates the potential of publicly-available genetic association data from high-throughput experiments combined with modern data-adaptive statistical learning techniques for obtaining biological insights into disease aetiology.

Acknowledgment

This work was supported by the UK Medical Research Council (MC_UU_00002/7). S.B. and V.Z. are supported by Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 204623/Z/16/Z). M.A.K is supported by a research grant from the Sigrid Juselius Foundation, Finland. This work was supported by core funding from: the UK Medical Research Council (MR/L003120/1), the British Heart Foundation (RG/13/13/30194; RG/18/13/33946), the National Institute for Health Research [Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust]* and Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

* The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

The research has been conducted using the UK Biobank Resource under Application Number 26865. Lastly, this study would not have been possible without the access to publicly available summary data. We would like to thank the authors of the NMR-GWAS (http://www.computationalmedicine.fi/data) and the CARDIOGRAMplus4CD (http://www.cardiogramplusc4d.org/data-downloads/).

Conflicts of Interest

A.S.B. has received grants outside of this work from AstraZeneca, Biogen, Bioverativ, Merck, Novartis and Sanofi, and personal fees from Novartis. All other authors declare no competing interests.

References

[1] Robert Plenge, Edward Scolnick, and David Altshuler. Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*, 12(8):581–594, 2013. doi: 10.1038/nrd4051.

[2] P. Linsel-Nitschke, A. Götz, J. Erdmann, I. Braenne, P. Braund, C. Hengstenberg, K. Stark, M. Fischer, S. Schreiber, N.E. El Mokhtari, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of

coronary artery disease — a Mendelian randomisation study. *PLoS One*, 3(8):e2986, 2008. doi: 10.1371/journal.pone.0002986.

[3] B.F. Voight, G.M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M.K. Jensen, G. Hindy, H. Hólm, E.L. Ding, T. Johnson, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841): 572–580, 2012. doi: 10.1016/S0140-6736(12)60312-2.

[4] B.M.Y. Cheung, I.J. Lauder, C.P. Lau, and C.R. Kumana. Meta-analysis of large randomized controlled trials to evaluate the impact of statins on cardiovascular outcomes. *British Journal of Clinical Pharmacology*, 57(5):640–651, 2004. doi: 10.1111/j.1365-2125.2003.02060.x.

[5] Gregory G Schwartz, Anders G Olsson, Markus Abt, Christie M Ballantyne, Philip J Barter, Jochen Brumm, Bernard R Chaitman, Ingar M Holme, David Kallend, Lawrence A Leiter, et al. Effects of dalcetrapib in patients with a recent acute coronary syndrome. *New England Journal of Medicine*, 367(22):2089–2099, 2012. doi: 10.1056/nejmoa1206797.

[6] A Michael Lincoff, Stephen J Nicholls, Jeffrey S Riesmeyer, Philip J Barter, H Bryan Brewer, Keith AA Fox, C Michael Gibson, Christopher Granger, Venu Menon, Gilles Montalescot, et al. Evacetrapib and cardiovascular outcomes in high-risk vascular disease. *New England Journal of Medicine*, 376(20):1933–1942, 2017.

[7] Ron Do, Cristen J Willer, Ellen M Schmidt, Sebanti Sengupta, Chi Gao, Gina M Peloso, Stefan Gustafsson, Stavroula Kanoni, Andrea Ganna, Jin Chen, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nature Genetics*, 45:1345–1352, 2013. doi: 10.1038/ng.2795.

[8] Brian A. Ference, John J. P. Kastelein, Kausik K. Ray, Henry N. Ginsberg, M. John Chapman, Chris J. Packard, Ulrich Laufs, Clare Oliver-Williams, Angela M. Wood, Adam S. Butterworth, Emanuele Di Angelantonio, John Danesh, Stephen J. Nicholls, Deepak L. Bhatt, Marc S. Sabatine, and Alberico L. Catapano. Association of triglyceride-lowering LPL variants and LDL-C–lowering LDLR variants with risk of coronary heart disease. *JAMA*, 321(4):364–373, 10/3/2019 2019. doi: 10.1001/jama.2018.20045.

[9] Tom G Richardson, Eleanor Sanderson, Tom M Palmer, Mika Ala-Korpela, Brian A Ference, George Davey Smith, and Michael V Holmes. Apolipoprotein B underlies the causal relationship of circulating blood lipids with coronary heart disease. *medRxiv*, 2019. doi: 10.1101/19004895. URL `https://www.medrxiv.org/content/early/2019/08/29/19004895`.

[10] Johannes Kettunen, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikäinen, Matti Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of *LPA*. *Nature Communications*, 7:11122, 2016.

[6] Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nature Communications*, 11(1):29, 2020. doi: 10.1038/s41467-019-13870-3.

[10] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[13] Philip C Haycock, Stephen Burgess, Kaitlin H Wade, Jack Bowden, Caroline Relton, and George Davey Smith. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition*, 103(4):965–978, 2016. doi: 10.3945/?ajcn.115.118216.

[14] Christopher P Nelson, Anuj Goel, Adam S Butterworth, Stavroula Kanoni, Tom R Webb, Eirini Marouli, Lingyao Zeng, Ioanna Ntalla, Florence Y Lai, Jemma C Hopewell, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nature Genetics*, 49(9):1385–1391, 2017. doi: 10.1038/ng.3913.

[15] CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47:1121–1130, 2015. doi: 10.1038/ng.3396.

[16] Elias Allara, Gabriele Morani, Paul Carter, Apostolos Gkatzionis, Verena Zuber, Christopher N Foley, Jessica M B Rees, Amy M Mason, Steven Bell, Dipender Gill, Sara Lindström, Adam S Butterworth, Emanuele Di Angelantonio, James Peters, Stephen Burgess, and INVENT consortium. Genetic determinants of lipids and cardiovascular disease outcomes: A wide-angled mendelian randomization investigation. *Circulation. Genomic and precision medicine*, 12(12):e002711–e002711, 12 2019. doi: 10.1161/CIRCGEN.119.002711. URL `https://www.ncbi.nlm.nih.gov/pubmed/31756303`.

[17] PJ Barter, CM Ballantyne, R Carmena, M Castro Cabezas, M John Chapman, P Couture, J De Graaf, PN Durrington, O Faergeman, J Frohlich, et al. Apo B versus cholesterol in estimating cardiovascular risk and in guiding therapy: report of the thirty-person/ten-country panel. *Journal of Internal Medicine*, 259(3):247–258, 2006.

[18] Mika Ala-Korpela. The culprit is the carrier, not the loads: cholesterol, triglycerides and apolipoprotein b in atherosclerosis and coronary heart disease. *International Journal of Epidemiology*, 48(5):1389–1392, 1/22/2020 2019. doi: 10.1093/ije/dyz068. URL `https://doi.org/10.1093/ije/dyz068`.

[19] Allan D. Sniderman, George Thanassoulis, Tamara Glavinovic, Ann Marie Navar, Michael Pencina, Alberico Catapano, and Brian A. Ference. Apolipoprotein b particles and cardiovascular disease: A narrative review. *JAMA Cardiology*, 4(12):1287–1295, 1/22/2020 2019. doi: 10.1001/jamacardio.2019.3780. URL `https://doi.org/10.1001/jamacardio.2019.3780`.

[20] T Ji Knott, RJ Pease, LM Powell, SC Wallis, SC Rall Jr, TL Innerarity, B Blackhart, WH Taylor, Y Marcel, R Milne, et al. Complete protein sequence and identification of structural domains of human apolipoprotein B. *Nature*, 323(6090):734, 1986.

[21] John H Contois, G Russell Warnick, and Allan D Sniderman. Reliability of low-density lipoprotein cholesterol, non-high-density lipoprotein cholesterol, and apolipoprotein B measurement. *Journal of clinical lipidology*, 5(4):264–272, 2011.

[22] Carl E Orringer. Non-HDL cholesterol, ApoB and LDL particle concentration in coronary heart disease risk prediction and treatment. *Clinical Lipidology*, 8(1):69–79, 2013.

[23] John T Wilkins, Ron C Li, Allan Sniderman, Cheeling Chan, and Donald M Lloyd-Jones. Discordance between apolipoprotein B and LDL-cholesterol in young adults predicts coronary artery calcification: the CARDIA study. *Journal of the American College of Cardiology*, 67(2):193–201, 2016.

[24] Terry A Jacobson. Opening a new lipid "apo-thecary": incorporating apolipoproteins as potential risk factors and treatment targets to reduce cardiovascular risk. *Mayo Clinic Proceedings*, 86(8):762–780, 2011.

Tables and Figures

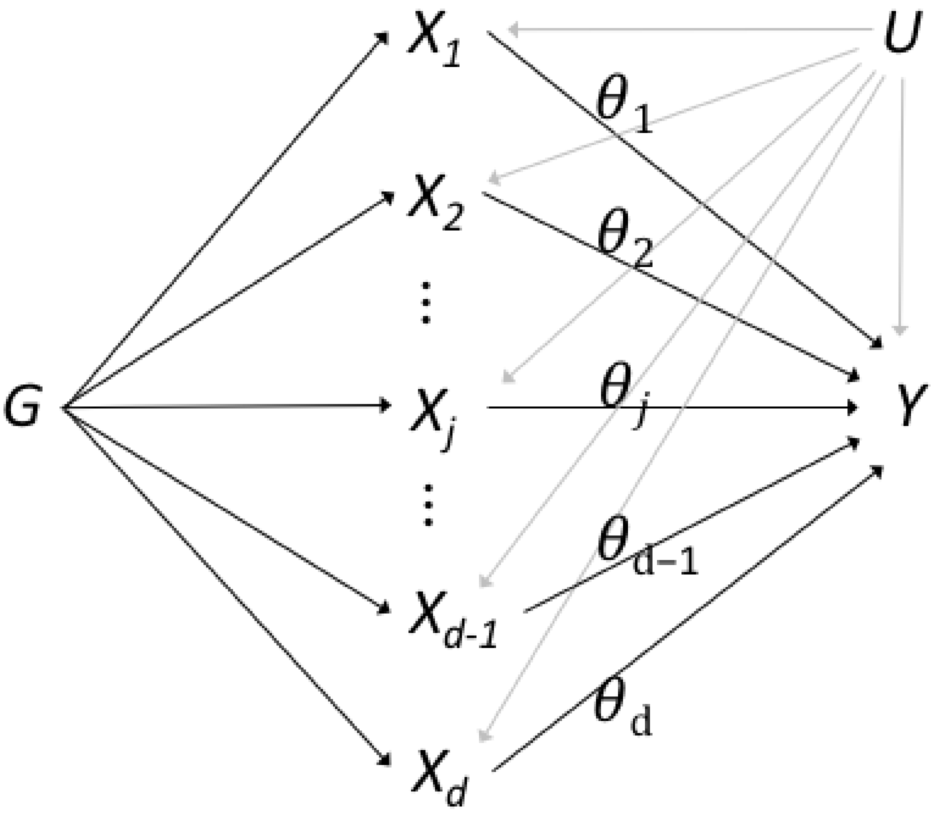
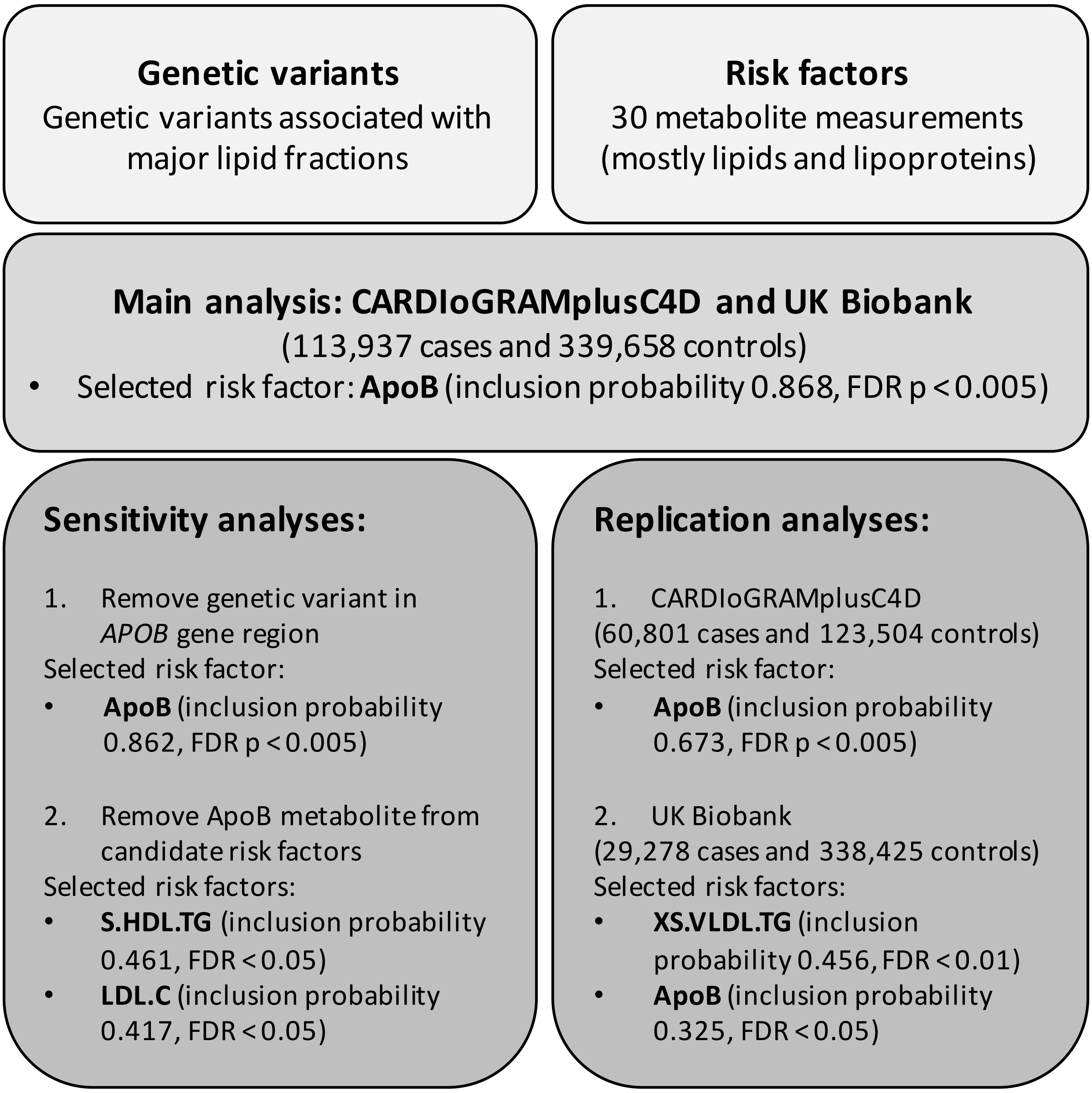


Figure 1: Diagram illustrating multivariable Mendelian randomization for selecting causal risk factors from a large number of candidate risk factors. Legend: G = genetic variants, $X_1, ..., X_d$ = risk factors, Y = outcome, U = confounders, θ_j = causal effect of risk factor j on the outcome.



Genetic variants
Genetic variants associated with major lipid fractions

Risk factors
30 metabolite measurements (mostly lipids and lipoproteins)

Main analysis: CARDIoGRAMplusC4D and UK Biobank
(113,937 cases and 339,658 controls)

- Selected risk factor: **ApoB** (inclusion probability 0.868, FDR p < 0.005)

Sensitivity analyses:

1. Remove genetic variant in *APOB* gene region

Selected risk factor:

- **ApoB** (inclusion probability 0.862, FDR p < 0.005)

2. Remove ApoB metabolite from candidate risk factors

Selected risk factors:

- **S.HDL.TG** (inclusion probability 0.461, FDR < 0.05)
- **LDL.C** (inclusion probability 0.417, FDR < 0.05)

Replication analyses:

1. CARDIoGRAMplusC4D (60,801 cases and 123,504 controls)

Selected risk factor:

- **ApoB** (inclusion probability 0.673, FDR p < 0.005)

2. UK Biobank (29,278 cases and 338,425 controls)

Selected risk factors:

- **XS.VLDL.TG** (inclusion probability 0.456, FDR < 0.01)
- **ApoB** (inclusion probability 0.325, FDR < 0.05)

Figure 2: Schematic diagram of the study design and results for the main, sensitivity, and replication analyses. Selected risk factors are those which had a empirical p-value of less than 0.05 after correction for multiple testing.

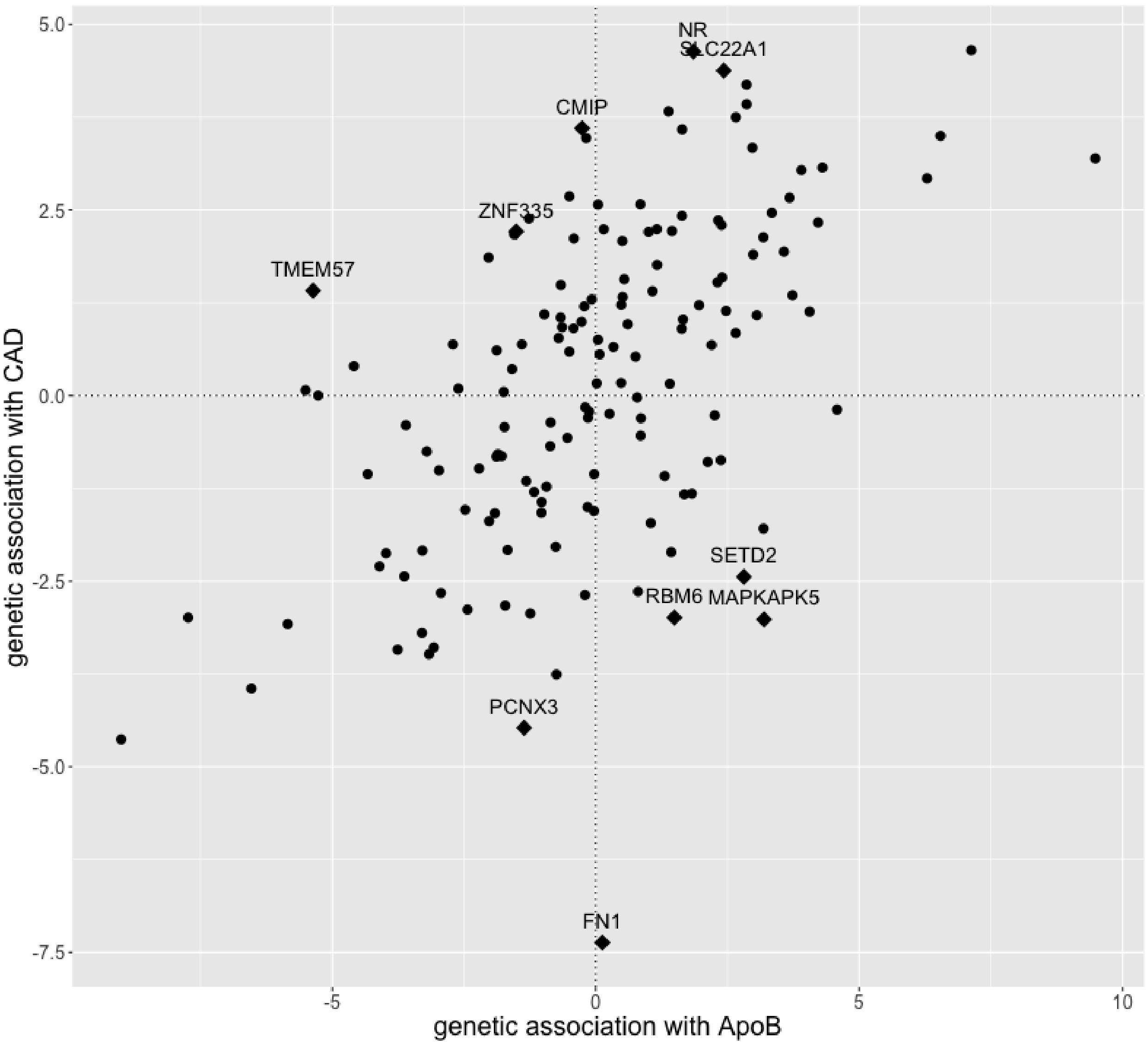


Figure 3: Estimates of genetic associations with coronary artery disease (CAD) risk (y-axis) against genetic associations with apolipoprotein B (x-axis) for each genetic variant from the primary analysis using CARDIoGRAMplusC4D and UK Biobank. Outliers removed from the analysis are highlighted as diamonds (◆) and their annotated gene-region is displayed.

CARDIoGRAMplusC4D and UK Biobank								
		Posterior	Causal		Marginal inclusion	Model-averaged	Empirical	
	Model	probability	effect	Risk factor	probability	causal effect	$p-$value	FDR
1	ApoB	0.480	0.464	ApoB	0.868	0.392	0.0001	0.003
2	ApoB,S.HDL.TG	0.043	0.349,0.175	S.HDL.TG	0.136	0.033	0.0165	0.247
3	LDL.C,S.HDL.TG	0.021	0.276,0.301	LDL.C	0.075	0.015	0.0882	0.882
4	ApoB,M.HDL.C	0.020	0.437,-0.111	XXL.VLDL.TG	0.047	0.010	0.4823	0.995
5	ApoB,S.LDL.C	0.014	0.570,-0.121	Serum.C	0.045	0.011	0.2295	0.995
6	ApoB,XXL.VLDL.TG	0.014	0.419,0.112	IDL.C	0.042	0.008	0.2401	0.995
7	ApoB,XS.VLDL.TG	0.011	0.375,0.099	S.LDL.C	0.040	0.001	0.3745	0.995
8	ApoB,S.VLDL.C	0.011	0.480,-0.017	M.HDL.C	0.038	-0.005	0.2885	0.995
9	ApoB,LDL.C	0.011	0.522,-0.062	HDL.C	0.036	-0.006	0.2266	0.995
10	ApoB,HDL.C	0.011	0.453,-0.073	Serum.TG	0.035	0.006	0.7583	0.995

Table 1: Main analysis: Top 10 models (combination of risk factors) ranked by the model posterior probability and top 10 risk factors ranked by the marginal inclusion probability in the primary analysis based on $n = 138$ genetic variants after model diagnostics. Causal effects are log odds ratios for coronary artery disease per 1 standard deviation increase in the risk factor. Empirical p-values are computed using $1,000$ permutations and adjusted for multiple testing using False Discovery Rate (FDR) procedure.

Supplementary Material

Abbreviation	Lipoprotein and metabolite measurements included
XXL.VLDL.TG	Triglyceride content in chylomicrons and extra-extra large VLDL
XL.VLDL.TG	Triglyceride content in extra-large VLDL
L.VLDL.TG	Triglyceride content in large VLDL
M.VLDL.TG	Triglyceride content in medium VLDL
S.VLDL.TG	Triglyceride content in small VLDL
XS.VLDL.TG	Triglyceride content in extra-small VLDL
IDL.TG	Triglyceride content in IDL
XL.HDL.TG	Triglyceride content in extra-large HDL
S.HDL.TG	Triglyceride content in small HDL
Serum.TG	Serum total triglycerides
L.VLDL.C	Cholesterol content in large VLDL
M.VLDL.C	Cholesterol content in medium VLDL
S.VLDL.C	Cholesterol content in small VLDL
LDL.C	Cholesterol content in LDL
S.LDL.C	Cholesterol content in small LDL
IDL.C	Cholesterol content in IDL
XL.HDL.C	Cholesterol content in extra-large HDL
L.HDL.C	Cholesterol content in large HDL
M.HDL.C	Cholesterol content in medium HDL
HDL.C	Cholesterol content in HDL
Est.C	Esterified cholesterol
Serum.C	Serum total cholesterol
VLDL.D	VLDL diameter
LDL.D	LDL diameter
HDL.D	HDL diameter
ApoA1	Apolipoprotein A1
ApoB	Apolipoprotein B
SM	Sphingomyelins
Tot.FA	Total fatty acids
Tot.PG	Total phosphoglycerides

Supplementary Table A1: List of lipoprotein and metabolite measurements included in the analyses.

CARDIoGRAMplusC4D and UK Biobank

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect
1	ApoB	0.347	0.432	ApoB	0.706	0.298
2	ApoB,M.HDL.C	0.048	0.392,-0.17	M.HDL.C	0.124	-0.024
3	XS.VLDL.TG	0.039	0.411	XS.VLDL.TG	0.103	0.032
4	ApoB,S.LDL.C	0.015	0.613,-0.208	IDL.TG	0.079	0.021
5	ApoB,SM	0.014	0.501,-0.139	XXL.VLDL.TG	0.076	0.02
6	IDL.TG	0.014	0.38	IDL.C	0.074	0.018
7	ApoB,S.HDL.TG	0.014	0.334,0.151	LDL.C	0.052	0.005
8	ApoB,XS.VLDL.TG	0.014	0.287,0.163	Serum.TG	0.049	0.014
9	ApoB,XXL.VLDL.TG	0.013	0.37,0.156	Serum.C	0.048	0.009

CARDIoGRAMplusC4D only

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect
1	ApoB	0.24	0.438	ApoB	0.488	0.197
2	XS.VLDL.TG	0.058	0.42	IDL.TG	0.159	0.048
3	IDL.TG	0.033	0.395	XS.VLDL.TG	0.153	0.05
4	S.VLDL.C	0.015	0.447	Serum.TG	0.095	0.036
5	ApoB,XS.VLDL.TG	0.014	0.272,0.186	Tot.FA	0.088	0.026
6	ApoB,S.HDL.TG	0.012	0.331,0.163	IDL.C	0.076	0.016
7	ApoB,IDL.TG	0.012	0.283,0.167	S.HDL.TG	0.07	0.016
8	IDL.TG,XXL.VLDL.	0.012	0.319,0.256	XXL.VLDL.TG	0.067	0.016
9	ApoB,M.HDL.C	0.01	0.407,-0.127	Serum.C	0.065	0.016
10	ApoB,Serum.TG	0.01	0.318,0.16	S.LDL.C	0.064	0.011

UK Biobank only

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect
1	XS.VLDL.TG	0.205	0.459	XS.VLDL.TG	0.388	0.161
2	S.VLDL.C	0.032	0.488	Tot.FA	0.321	0.139
3	HDL.C,Tot.FA	0.03	-0.255,0.475	ApoB	0.147	0.047
4	ApoB	0.023	0.452	IDL.TG	0.145	0.045
5	IDL.TG	0.019	0.425	HDL.C	0.103	-0.023
6	ApoB,XS.VLDL.TG	0.014	0.191,0.294	S.VLDL.C	0.099	0.033
7	L.HDL.C,Tot.FA	0.013	-0.221,0.448	S.HDL.TG	0.097	0.026
8	S.HDL.TG,Tot.FA	0.011	0.329,0.259	TotPG	0.089	-0.032
9	Tot.FA,TotPG	0.01	0.883,-0.504	IDL.C	0.073	0.015
10	LDL.C,XS.VLDL.TG	0.009	0.129,0.369	Serum.TG	0.072	0.026

Supplementary Table A2: Analysis including all genetic variants: Top 10 models ranked by the model posterior probability and top 10 risk factors ranked by the marginal inclusion probability including all genetic variants before removing influential genetic variants and outliers. Causal effects are log odds ratios for coronary artery disease per 1 standard deviation increase in the risk factor.

CARDIoGRAMplusC4D and UK Biobank						
		Posterior	Causal		Marginal inclusion	Model-averaged
	Model	probability	effect	Risk factor	probability	causal effect
1	ApoB	0.472	0.46	ApoB	0.862	0.385
2	ApoB,S.HDL.TG	0.043	0.343,0.177	S.HDL.TG	0.136	0.033
3	LDL.C,S.HDL.TG	0.02	0.272,0.301	LDL.C	0.076	0.015
4	ApoB,M.HDL.C	0.019	0.435,-0.11	XXL.VLDL.TG	0.05	0.011
5	ApoB,XXL.VLDL.TG	0.015	0.408,0.123	Serum.C	0.045	0.01
6	ApoB,S.LDL.C	0.015	0.571,-0.127	IDL.C	0.043	0.008
7	ApoB,XS.VLDL.TG	0.012	0.367,0.102	S.LDL.C	0.041	0.001
8	ApoB,Serum.TG	0.011	0.385,0.098	Serum.TG	0.038	0.007
9	ApoB,LDL.C	0.011	0.525,-0.071	M.HDL.C	0.037	-0.004
10	ApoB,S.VLDL.C	0.011	0.474,-0.015	HDL.C	0.035	-0.005

Supplementary Table A3: Sensitivity analysis 1: After excluding the genetic variant in the *APOB* gene region, these are the top 10 models judged by posterior probability and top 10 risk factors judged by marginal inclusion probability in the primary analysis based on $n = 137$ genetic variants. Causal effects are log odds ratios for coronary artery disease per 1 standard deviation increase in the risk factor.

CARDIoGRAMplusC4D and UK Biobank including all genetic variants (n = 148)

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect
1	XS.VLDL.TG	0.127	0.411	XS.VLDL.TG	0.263	0.094
2	IDL.TG	0.046	0.38	IDL.C	0.213	0.059
3	S.VLDL.C	0.041	0.44	IDL.TG	0.204	0.063
4	IDL.C,XXL.VLDL.TG	0.03	0.299,0.347	XXL.VLDL.TG	0.168	0.05
5	IDL.TG,XXL.VLDL.TG	0.022	0.304,0.267	M.HDL.C	0.162	-0.037
6	M.HDL.C,Serum.C	0.015	-0.317,0.367	Serum.C	0.116	0.034
7	LDL.C,XS.VLDL.TG	0.015	0.178,0.286	LDL.C	0.114	0.026
8	IDL.C,S.HDL.TG	0.012	0.241,0.282	Serum.TG	0.107	0.038
9	IDL.C,Serum.TG	0.011	0.219,0.287	S.VLDL.C	0.096	0.031
10	S.LDL.C,XS.VLDL.TG	0.011	0.175,0.294	S.HDL.TG	0.079	0.019

CARDIoGRAMplusC4D and UK Biobank after model diagnostics (n = 138)

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect	Empirical $p-$value	FDR
1	LDL.C,S.HDL.TG	0.156	0.261,0.3	S.HDL.TG	0.461	0.144	0.0021	0.025
2	IDL.TG	0.063	0.436	LDL.C	0.417	0.119	0.0013	0.025
3	S.HDL.TG,S.LDL.C	0.049	0.294,0.266	Serum.C	0.17	0.056	0.0143	0.087
4	IDL.C,S.HDL.TG	0.048	0.237,0.325	IDL.TG	0.159	0.055	0.0151	0.087
5	L.HDL.C,Serum.C	0.032	-0.272,0.381	S.LDL.C	0.156	0.038	0.0274	0.101
6	HDL.C,Serum.C	0.027	-0.277,0.441	IDL.C	0.128	0.029	0.0296	0.101
7	S.HDL.TG,Serum.C	0.021	0.354,0.23	L.HDL.C	0.118	-0.026	0.0181	0.087
8	LDL.C,XS.VLDL.TG	0.016	0.233,0.249	HDL.C	0.095	-0.019	0.0384	0.115
9	Est.C,S.HDL.TG	0.014	0.197,0.393	XS.VLDL.TG	0.076	0.017	0.0682	0.182
10	LDL.C,XXL.VLDL.TG	0.012	0.337,0.273	XXL.VLDL.TG	0.073	0.016	0.2636	0.575

Supplementary Table A4: Sensitivity analysis 2: After excluding the ApoB measurement as risk factor from the set of candidate risk factors these are the top 10 models ranked by the posterior probability and top 10 risk factors ranked by the marginal inclusion probability in the primary analysis based on all available genetic variants (n = 148) and after model diagnostics (n = 138). Causal effects are log odds ratios for coronary artery disease per 1 standard deviation increase in the risk factor.

CARDIoGRAMplusC4D only

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect	Empirical $p-$value	FDR
1	ApoB	0.394	0.455	ApoB	0.673	0.293	0.0001	0.003
2	ApoB,M.HDL.C	0.018	0.425,-0.121	LDL.C	0.107	0.027	0.0544	0.461
3	S.VLDL.C	0.018	0.464	S.LDL.C	0.097	0.027	0.0709	0.461
4	IDL.TG	0.014	0.444	Serum.TG	0.084	0.028	0.0599	0.461
5	HDL.C,Serum.C	0.014	-0.263,0.464	Serum.C	0.072	0.021	0.1176	0.510
6	LDL.C,Serum.TG	0.012	0.276,0.263	HDL.C	0.062	-0.012	0.0974	0.506
7	ApoB,Serum.TG	0.011	0.369,0.115	S.VLDL.C	0.059	0.015	0.1667	0.542
8	ApoB,IDL.TG	0.011	0.358,0.109	IDL.TG	0.056	0.015	0.1539	0.542
9	S.LDL.C	0.010	0.461	M.HDL.C	0.055	-0.008	0.1889	0.546
10	ApoB,S.VLDL.C	0.010	0.402,0.06	IDL.C	0.052	0.010	0.2423	0.630

UK Biobank only

	Model	Posterior probability	Causal effect	Risk factor	Marginal inclusion probability	Model-averaged causal effect	Empirical $p-$value	FDR
1	XS.VLDL.TG	0.195	0.435	XS.VLDL.TG	0.456	0.169	0.0002	0.006
2	ApoB,S.HDL.TG	0.056	0.281,0.233	ApoB	0.325	0.102	0.0010	0.015
3	ApoB,XS.VLDL.TG	0.043	0.207,0.258	S.HDL.TG	0.222	0.060	0.0061	0.061
4	ApoB	0.039	0.437	IDL.TG	0.109	0.027	0.0157	0.103
5	LDL.C,XS.VLDL.TG	0.024	0.14,0.338	LDL.C	0.108	0.018	0.0446	0.191
6	S.VLDL.C	0.024	0.467	Serum.TG	0.104	0.032	0.0171	0.103
7	LDL.C,S.HDL.TG	0.015	0.216,0.334	S.VLDL.C	0.086	0.024	0.0444	0.191
8	S.LDL.C,XS.VLDL.TG	0.015	0.133,0.346	Tot.FA	0.079	0.018	0.0677	0.254
9	IDL.C,S.HDL.TG	0.012	0.201,0.345	IDL.C	0.063	0.009	0.0994	0.331
10	ApoB,Serum.TG	0.012	0.273,0.218	S.LDL.C	0.059	0.003	0.1739	0.522

Supplementary Table A5: Replication analysis: Top 10 models ranked by the posterior probability and top 10 risk factors ranked by the marginal inclusion probability after model diagnostics (including $n = 144$ genetic variants for CARDIoGRAMplusC4D and $n = 141$ for UK Biobank). Causal effects are log odds ratios for coronary artery disease per 1 standard deviation increase in the risk factor.

CARDIoGRAMplusC4D and UK Biobank

	rs	gene region	$Cd1$	$Cd2$	$Cd3$	max Cd
1	rs10903129	TMEM57	0.108	0.054	0.018	0.108
2	rs2923084	AMPD3	0.049	0.069	0.068	0.069
3	rs6489818	MAPKAPK5	0.051	0.029	0.004	0.051
4	rs1515110	NR	0.013	0.042	0.027	0.042
5	rs515135	APOB	0.013	0.003	0.041	0.041
6	rs6859	APOE	0.035	0.018	0.039	0.039
7	rs2326077	intergenic	0.039	0.027	0.015	0.039
8	rs5880	CETP	0.001	0.038	0.023	0.038
9	rs799160	intergenic	0.004	0.002	0.037	0.037
10	rs4465830	ZNF335	0.005	0.037	0.037	0.037
		threshold	0.457	0.696	0.457	

CARDIoGRAMplusC4D only

	rs	region	$Cd1$	$Cd2$	$Cd3$	max Cd
1	rs261342	LIPC	0.008	0.024	0.911	**0.911**
2	rs5880	CETP	0.006	0.164	0.057	0.164
3	rs515135	APOB	0.116	0.129	0.125	0.129
4	rs2923084	AMPD3	0.081	0.109	0.096	0.109
5	rs10903129	TMEM57	0.078	0.012	0.025	0.078
6	rs4530754	CSNK1G3	0.076	0.001	0.001	0.076
7	rs6489818	MAPKAPK5	0.062	0.005	0.009	0.062
8	rs2326077	intergenic	0.039	0.016	0.015	0.039
9	rs12133576	DR1	0.036	0.006	0.004	0.036
10	rs4465830	ZNF335	0.005	0.036	0.000	0.036
		threshold	0.457	0.457	0.457	

UK Biobank only

	rs	region	$Cd1$	$Cd2$	$Cd3$	$Cd4$	max Cd
1	rs10401969	SUGP1	0.302	0.224	0.248	0.096	0.302
2	rs2923084	AMPD3	0.124	0.064	0.026	0.079	0.124
3	rs5880	CETP	0.107	0.033	0.025	0	0.107
4	rs2297374	SLC22A1	0.024	0.054	0.091	0.057	0.091
5	rs10903129	TMEM57	0.012	0.051	0.005	0.071	0.071
6	rs7703051	HMGCR	0.006	0.053	0.009	0.065	0.065
7	rs6489818	MAPKAPK5	0.005	0.032	0.001	0.055	0.055
8	rs894210	intergenic	0.05	0.051	0.019	0.039	0.051
9	rs687339	intergenic	0.038	0.044	0.039	0.045	0.045
10	rs998584	VEGFA	0.041	0.039	0.037	0.036	0.041
		threshold	0.457	0.457	0.696	0.457	

Supplementary Table A6: Influential genetic variants: This table displays for each study the 10 variants with the largest Cook's distance (Cd) and the annotated genomic region based on the best individual models (model posterior probability > 0.02). The maximum Cd of each variant in all models is used for diagnostics. The final row gives the suggested cut-off for Cook's distance and genetic variants with Cd above the threshold are marked in bold.

CARDIoGRAMplusC4D and UK Biobank

	rs	gene region	$q1$	$q2$	$q3$	max q
1	rs1250229	FN1	55.077	54.867	57.211	**57.211**
2	rs6489818	MAPKAPK5	19.308	20.150	12.288	**20.150**
3	rs12801636	PCNX3	15.124	14.625	15.845	**15.845**
4	rs1515110	NR	14.697	10.106	11.196	**14.697**
5	rs2290547	SETD2	13.361	14.316	8.53	**14.316**
6	rs2297374	SLC22A1	11.075	11.676	14.204	**14.204**
7	rs10903129	TMEM57	13.910	12.503	7.369	**13.910**
8	rs2925979	CMIP	13.787	11.646	10.338	**13.787**
9	rs2240327	RBM6	13.213	11.505	11.223	**13.213**
10	rs4465830	ZNF335	8.194	2.964	12.962	**12.962**
11	rs6450176	ARL15	8.271	6.936	12.705	12.705
12	rs731839	PEPD	12.596	10.504	10.14	12.596
13	rs4148218	ABCG8	11.789	12.03	12.032	12.032
14	rs2247056	HLA	8.710	9.897	11.563	11.563
15	rs9930333	FTO	7.213	6.599	11.191	11.191
		threshold				12.84801

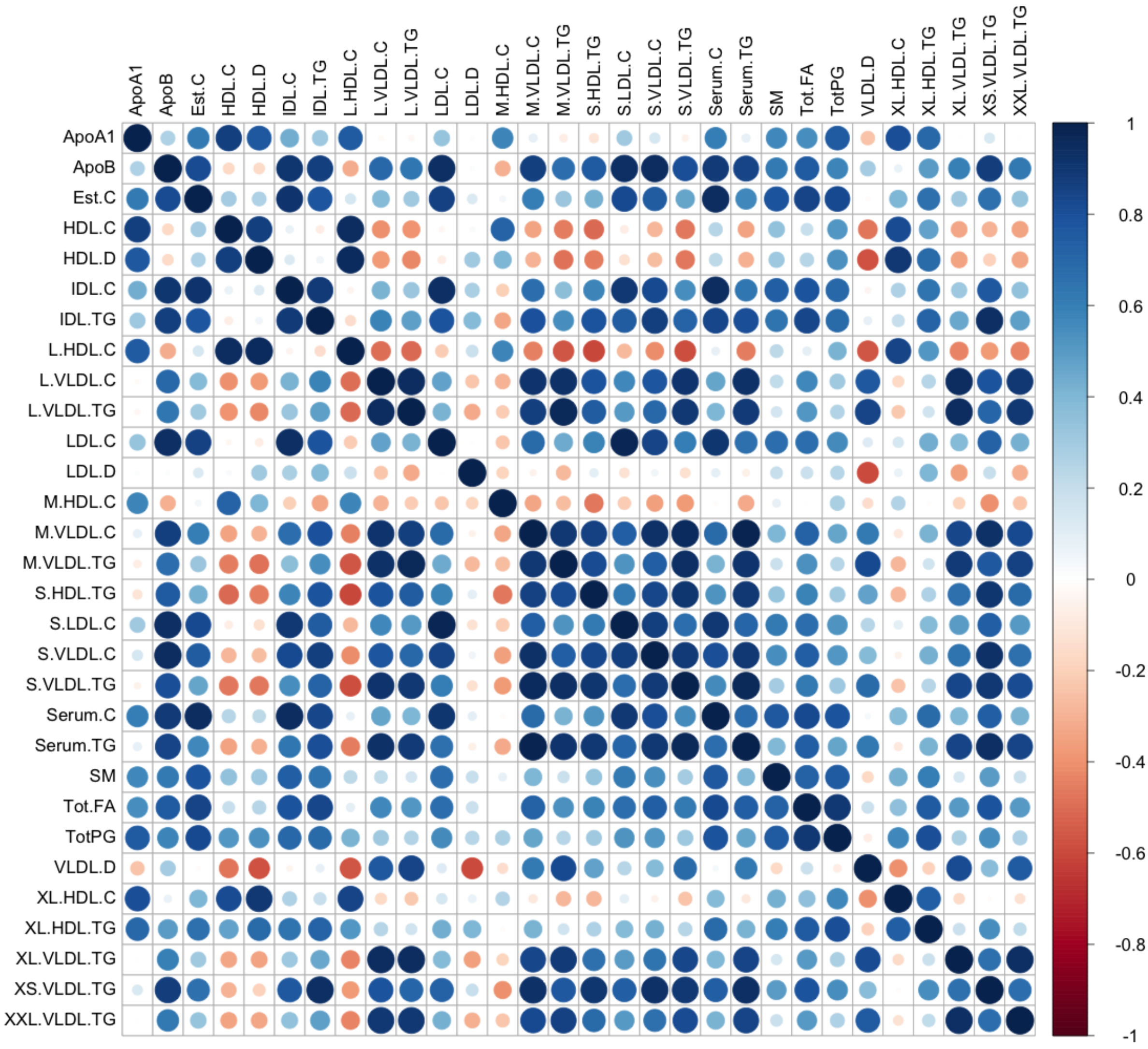
CARDIoGRAMplusC4D only

	rs	gene region	$q1$	$q2$	$q3$	max q
1	rs4530754	CSNK1G3	24.505	15.468	15.292	**24.505**
2	rs6489818	MAPKAPK5	19.513	14.598	13.255	**19.513**
3	rs12801636	PCNX3	16.290	16.800	16.810	**16.810**
4	rs4148218	ABCG8	14.936	14.107	15.098	**15.098**
5	rs1250229	FN1	9.932	12.776	10.769	12.776
6	rs952044	AC090771.2	10.333	12.468	11.714	12.468
7	rs2297374	SLC22A1	9.125	9.187	11.492	11.492
8	rs4465830	ZNF335	7.196	5.401	11.390	11.390
9	rs998584	VEGFA	8.781	11.195	7.745	11.195
10	rs2923084	AMPD3	8.404	10.802	9.845	10.802
		threshold				12.84801

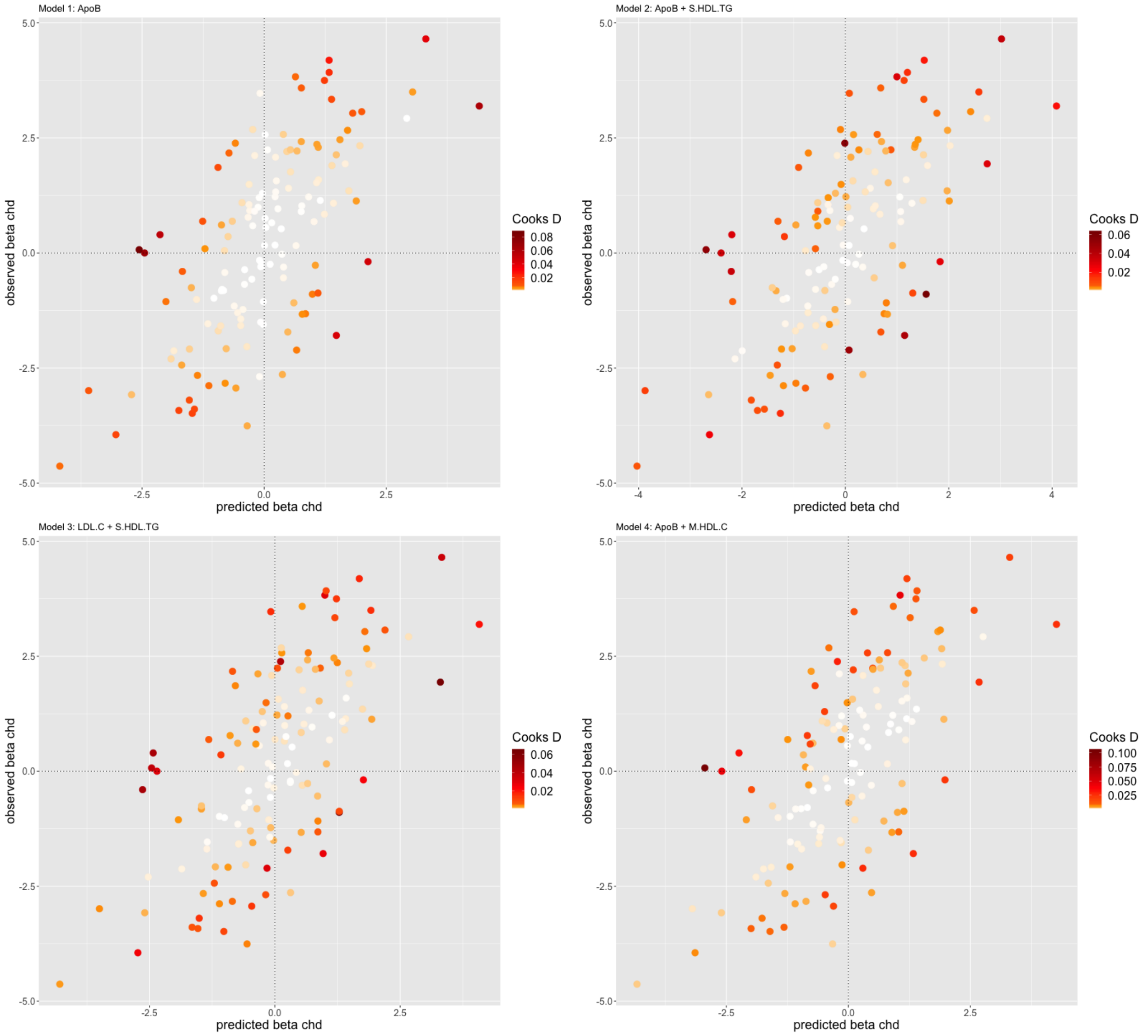
UK Biobank only

	rs	gene region	$q1$	$q2$	$q3$	$q4$	max q
1	rs2297374	SLC22A1	38.863	34.587	27.820	34.345	**38.863**
2	rs1250229	FN1	25.528	22.807	31.310	24.057	**31.310**
3	rs6489818	MAPKAPK5	14.222	18.563	12.616	20.625	**20.625**
4	rs2240327	RBM6	15.063	16.844	16.278	17.034	**17.034**
5	rs687339	intergenic	16.284	15.452	7.003	15.328	**16.284**
6	rs2925979	CMIP	10.424	15.160	9.803	13.903	**15.160**
7	rs4148218	ABCG8	14.250	14.512	11.681	14.137	**14.512**
8	rs4921914	NAT2	13.259	10.640	10.262	11.642	**13.259**
9	rs1186380	HNF1A-AS1	9.758	12.067	11.982	13.168	**13.168**
10	rs2241210	UBE3B	12.630	11.045	10.759	9.015	12.630
		threshold					12.87313

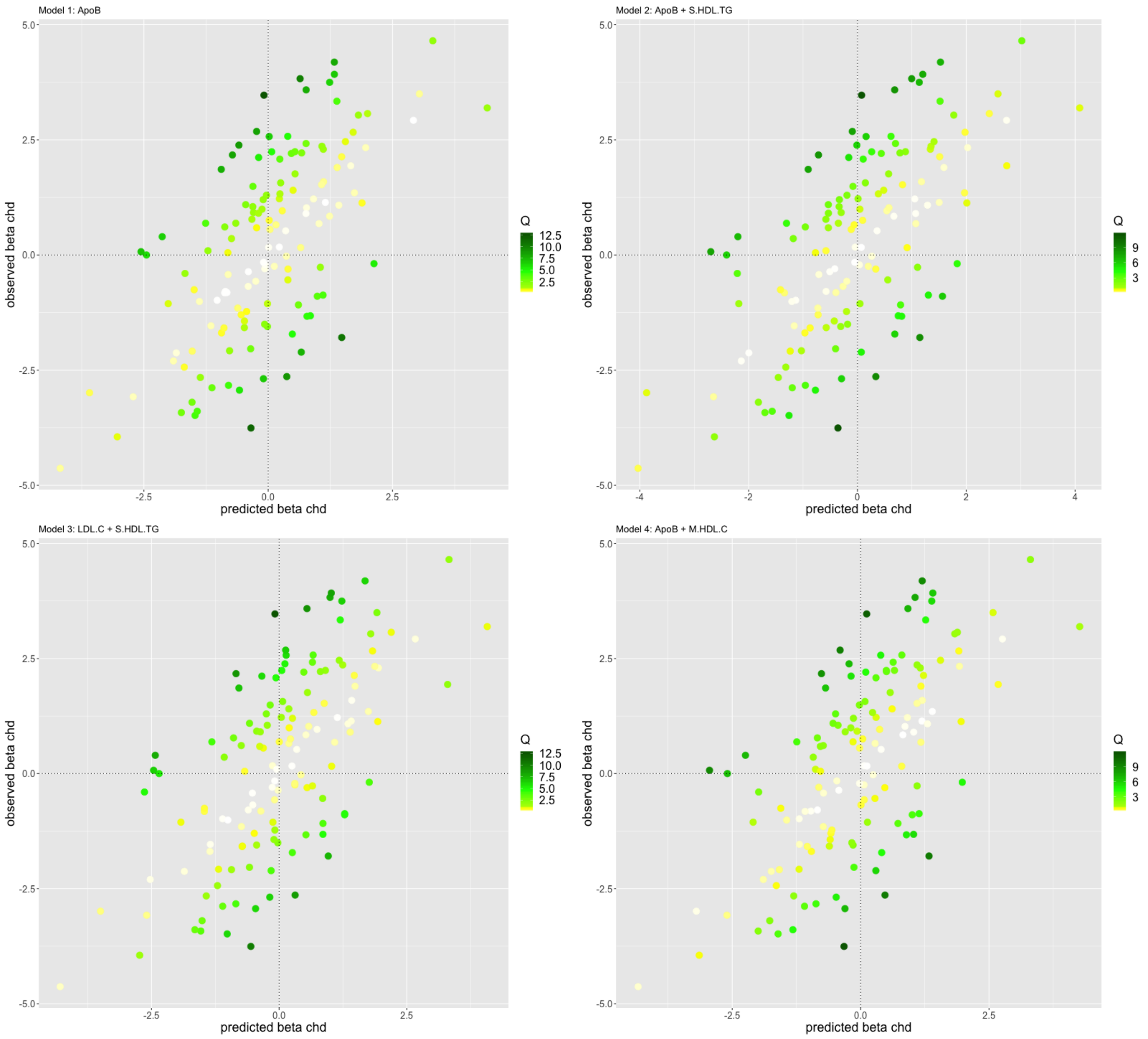
Supplementary Table A7: Outlying genetic variants: This table displays for each study the 10 variants with the largest maximum q and the annotated genomic region based on the best individual models (model posterior probability > 0.02). The maximum q of each variant in all models is used for diagnostics. The final row gives the suggested threshold for the $q-$statistic and variants with q above this threshold are given in bold.



Supplementary Figure A1: Genetic correlation between lipoprotein measures and metabolites based on the $n = 148$ lipid-associated genetic variants as used in the main analysis. Color-code indicates correlation strength (darkblue=strong positive correlation to darkred=strong negative correlation). The size of the square is proportional to the absolute correlation.



Supplementary Figure A2: Diagnostic plots with Cooks distance: Estimates of genetic associations with the outcome against predicted genetic associations with the outcome from the primary analysis based on $n = 138$ genetic variants after exclusion of outliers. Here we show the diagnostics for all four top models with posterior probability > 0.02 as given in Main Table 1. Colour code of points indicates influence, as measured by the variant's Cook's distance.



Supplementary Figure A3: Diagnostic plots with Cooks distance: Estimates of genetic associations with the outcome against predicted genetic associations with the outcome from the primary analysis based on n = 138 genetic variants after exclusion of outliers. Here we show the diagnostics for all four top models with posterior probability > 0.02 as given in Main Table 1. Colour code of points indicates heterogeneity, as measured by the variant's q-statistic.

Supplementary Methods

Mendelian randomization using summarized data

A genetic variant can be used to make causal inferences about the effect of a risk factor on an outcome if it satisfies the three instrumental variable assumptions:

IV1 The variant is associated with the risk factor;

IV2 The variant is not confounded in its associations with the outcome;

IV3 The variant does not influence the outcome directly, only potentially indirectly via its association with the risk factor.

These assumptions imply that a genetic variant behaves analogously to random assignment to a treatment group in a randomized controlled trial, in that it divides the population into subgroups that differ only with respect to their average level of the risk factor [1]. Any difference in the outcome between these groups implies a causal effect of the risk factor on the outcome, analogous to an intention-to-treat effect in a randomized trial [2].

We consider an extension of the Mendelian randomization paradigm known as multivariable Mendelian randomization, in which genetic variants are allowed to influence multiple risk factors, provided that any causal pathway from the genetic variants to the outcome passes via one or more of the measured risk factors [3]. The assumptions for genetic variants to be valid instruments in multivariable Mendelian randomization are:

MV-IV1 Each variant is associated with at least one of the risk factors;

MV-IV2 Variants are not confounded in their associations with the outcome;

MV-IV3 Variants are not associated with the outcome conditional on the risk factors and confounders.

In turn, the assumptions for a risk factor to be included in a multivariable Mendelian randomization model are:

RF1 No risk factor can be linearly explained by any other included risk factor or a combination of multiple risk factors.

RF2 Each risk factor is associated with at least one of the genetic variants.

Assumption RF1 is needed to distinguish between correlated risk factors [4]. RF2 ensures that each risk factor is adequately predicted by the genetic variants selected as instrumental variables in the analysis.

For a particular set of risk factors, causal effects are estimated by weighted linear regression of the genetic associations with the outcome on the genetic associations with the risk factors

$$\beta_Y = \theta_1\beta_{X1} + \theta_2\beta_{X2} + \ldots + \theta_d\beta_{Xd} + \varepsilon, \quad \varepsilon \sim N(0, \text{diag}(\text{se}(\beta_Y)^2)),$$

where β_Y is the vector of genetic associations with the outcome of length n, with n the number of genetic variants used as instrumental variables, $\text{se}(\beta_Y)$ is the vector of standard errors of these associations of length n and diag the diagonal operator. $\beta_{X1}, \beta_{X2}, \ldots, \beta_{Xd}$ are the genetic associations with the d risk factors, and $\theta_1, \theta_2, \ldots, \theta_d$ are the causal effects of the d risk factors on the outcome. If there are causal relationships between the risk factors, then these parameters represent the direct effects of the risk factors, i.e. the effect of changing the target risk factor keeping all other risk factors constant [4, 5].

Variable selection and Bayesian model averaging

The model averaging approach is implemented by considering different sets of risk factors in turn [6]. For each risk factor set, MR-BMA fits the relevant multivariable Mendelian randomization model and assigns a score to the set of risk factors considered that captures the posterior probability that this particular model represents the true causal risk factors for the outcome given the observed genetic association data [6].

When considering many candidate risk factors, the model space (including all possible combinations of risk factors) may be prohibitively large to consider all possible combinations of risk factors. To alleviate this we have implemented a stochastic search algorithm [7] to explore the relevant model space (all models with a non-negligible posterior probability) in an efficient way.

When the number of risk factors considered is large, the evidence for each particular model may be small. Hence, we average over the models visited and for each risk factor compute its marginal inclusion probability, which is the sum of the posterior probabilities for all models visited that include this particular risk factor. Further, we provide the model-averaged causal effect estimate, representing the average causal effect estimate for the given risk factor across models in which it is included. As is common for variable-selection methods, this is a conservative estimates of the true causal effect and underestimates its magnitude, but may be used for the interpretation of effect direction and for comparison among the risk factors.

Resampling to compute empirical $p-$values

Empirical p-values for the marginal inclusion probability of each risk factor are obtained using a permutation procedure, where the risk factor association data are held constant and the outcome associations of the genetic variants are randomly perturbed [8]. The empirical p-value for risk factor j quantifies how extreme the actual observed marginal inclusion probability is with respect to all permuted marginal inclusion probabilities for that particular risk factor. Formally, the empirical p-value is computed by the rank (r_j) of the actual observed marginal inclusion probability for risk factor j among all permuted marginal inclusion probabilities for risk factor j over the total number of permutations (n_{perm} = 1,000). Following [9] we add one to the computation to obtain the probability that under the null hypothesis the observed marginal inclusion probability has the observed or a higher rank

$$p_j = (r_j + 1)/(n_{perm} + 1).$$

Multiple testing adjustment is done using the Benjamini and Hochberg false discovery rate (FDR) procedure [10].

Model diagnostics

Two approaches are considered for model diagnostics. Firstly, to identify influential variants for each visited model with a model posterior probability larger than 0.02, we calculated Cook's distance for each genetic variant [11] and excluded all variants that have in any selected model a Cook's distance which exceeds the median of a central F-distribution with d and $n-d$ degrees of freedom, where d is the number of risk factors and n the number of genetic variants used as instrumental variables.

Secondly, to identify outlying variants, we consider for each visited model with a model posterior probability larger than 0.02 a version of Cochran's Q statistic used to detect heterogeneity in meta-analysis [12]

$$Q = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \mathrm{se}(\beta_{Y_i})^{-2}(\beta_{Y_i} - \hat{\beta}_{Y_i})^2,$$

where i indexes the genetic variants and $\hat{\beta}_{Y_i}$ is the predicted value of the genetic association with the outcome β_{Y_i} based on the relevant multivariable Mendelian randomization model. A genetic variant with a high value of q_i (compared to the $0.05/n$th upper tail of a χ^2 distribution with one degree of freedom representing Bonferroni multiple testing adjustment by the number of variants included) in any of the models visited (with a model posterior probability larger than 0.02) was considered to be an outlying variant.

We then repeated the analyses excluding such variants. The reason for excluding outliers and influential variants is that a single genetic variant can have a strong impact on the models visited and subsequently on variable selection. However, in this case for both main and replication analyses, excluding these variants did not change the headline results.

References

[1] G. Thanassoulis and C.J. O'Donnell. Mendelian randomization: nature's randomized trial in the post-genome era. *Journal of the American Medical Assocation*, 301(22): 2386–2388, 2009. doi: 10.1001/jama.2009.812.

[2] S. Burgess, C. N. Foley, and V. Zuber. Inferring causal relationships between risk factors and outcomes from genome-wide association study data. *Annual Review of Genomics and Human Genetics*, 2018.

[3] S Burgess and Simon G Thompson. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, 181(4):251–260, 2015. doi: 10.1093/aje/kwu283.

[4] Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable Mendelian Randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, pages dyy262–dyy262, 12 2018.

[5] Stephen Burgess, Deborah J Thompson, Jessica MB Rees, Felix R Day, John R Perry, and Ken K Ong. Dissecting causal pathways using mendelian randomization with summarized genetic data: application to age at menarche and risk of breast cancer. *Genetics*, 2017. doi: 10.1534/genetics.117.300191.

[6] Verena Zuber, Johanna Maria Colijn, Caroline Klaver, and Stephen Burgess. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nature Communications*, 11(1):29, 2020. doi: 10.1038/s41467-019-13870-3.

[7] Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

[8] Matthew Stephens and David J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009. doi: 10.1038/nrg2615. URL `https://doi.org/10.1038/nrg2615`.

[9] B V North, D Curtis, and P C Sham. A note on the calculation of empirical p values from monte carlo procedures. *American journal of human genetics*, 71(2):439–441, 08 2002. doi: 10.1086/341527. URL `https://www.ncbi.nlm.nih.gov/pubmed/12111669`.

[10] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[11] Laura J Corbin, Rebecca C Richmond, Kaitlin H Wade, Stephen Burgess, Jack Bowden, George Davey Smith, and Nicholas J Timpson. Body mass index as a modifiable risk factor for type 2 diabetes: Refining and understanding causal estimates using Mendelian randomisation. *Diabetes*, 2016. doi: 10.2337/db16-0418.

[12] MDF Greco, Cosetta Minelli, Nuala A Sheehan, and John R Thompson. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine*, 34(21):2926–2940, 2015. doi: 10.1002/sim.6522.