**Supplementary Materials**

**Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases**

Youwen Qin1,2, Guillaume Méric1,3, Tao Long4,5, Jeramie D. Watrous5, Stephen Burgess6,7, Aki S. Havulinna8, Scott C. Ritchie1,9-11, Marta Brożyńska1, Pekka Jousilahti8, Markus Perola8, Leo Lahti12, Teemu Niiranen8,12, Susan Cheng13,14, Veikko Salomaa8, Mohit Jain5, Michael Inouye1,2,7,9-11,15,16

1Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

2School of BioSciences, University of Melbourne, Melbourne, Victoria, Australia

3Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia

4Bioinformatics and Structural Biology Program, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California, USA

5Departments of Medicine and Pharmacology, University of California, San Diego, CA, USA

6MRC Biostatistics Unit, Institute of Public Health, University of Cambridge, Cambridge

7British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

8Department of Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland

9Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

10British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

11National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK

12Department of Future Technologies, University of Turku, Turku, Finland

13Division of Cardiology, Brigham and Women’s Hospital, Boston, USA

14Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA

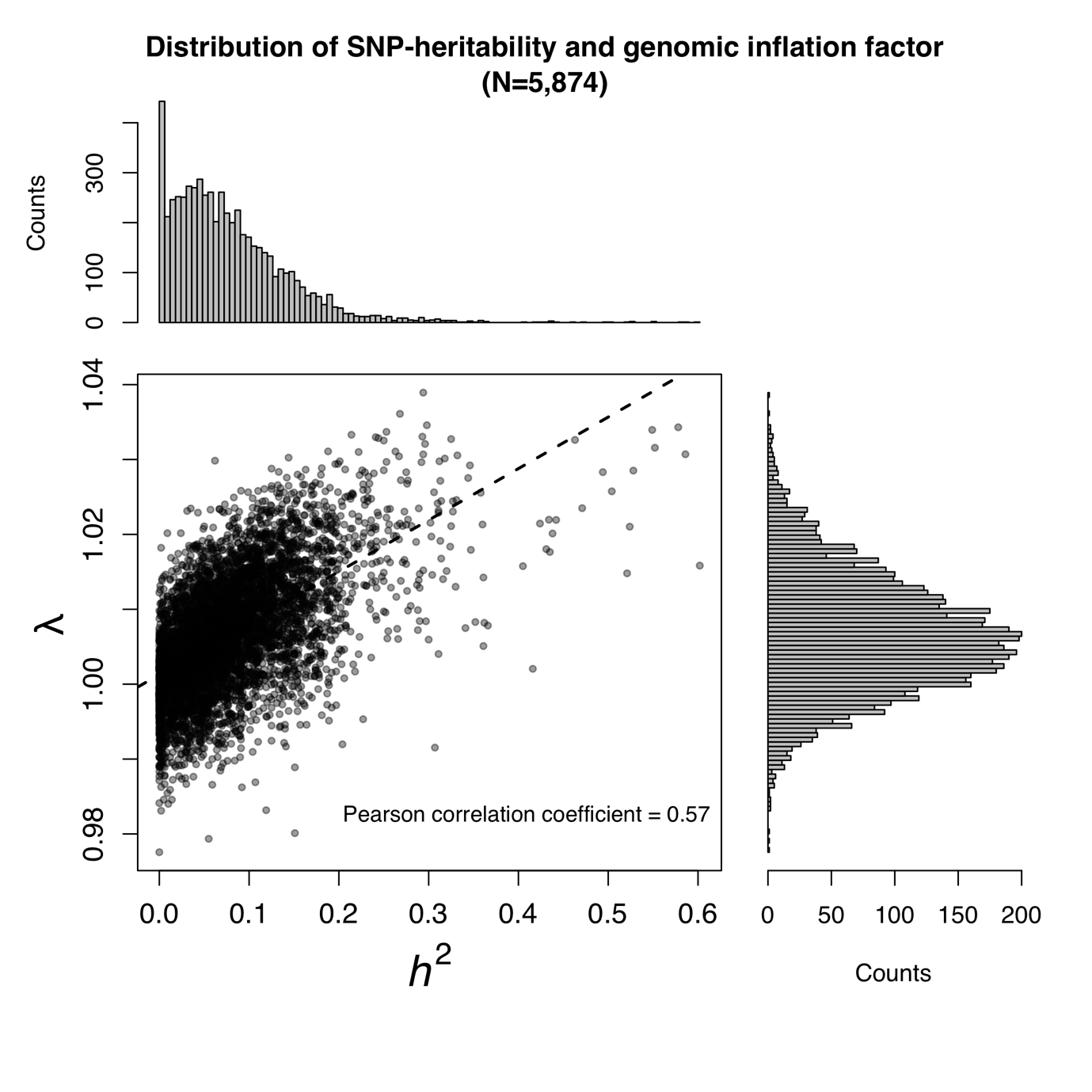
15Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

16The Alan Turing Institute, London, UK.

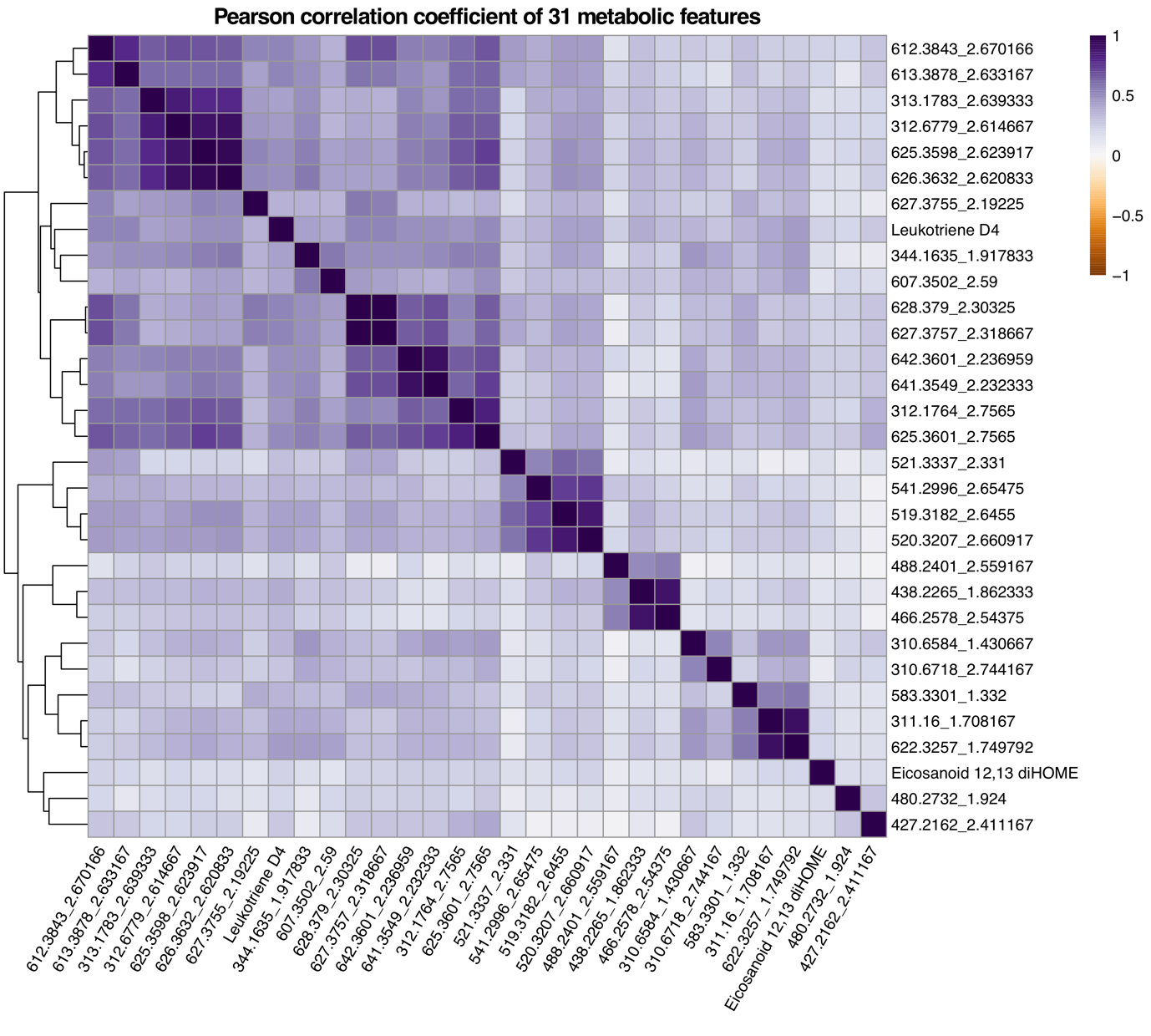
\* Corresponding author: MI ([minouye@baker.edu.au](mailto:minouye@baker.edu.au) or [mi336@medschl.cam.ac.uk](mailto:mi336@medschl.cam.ac.uk))

**Supplementary Figures**

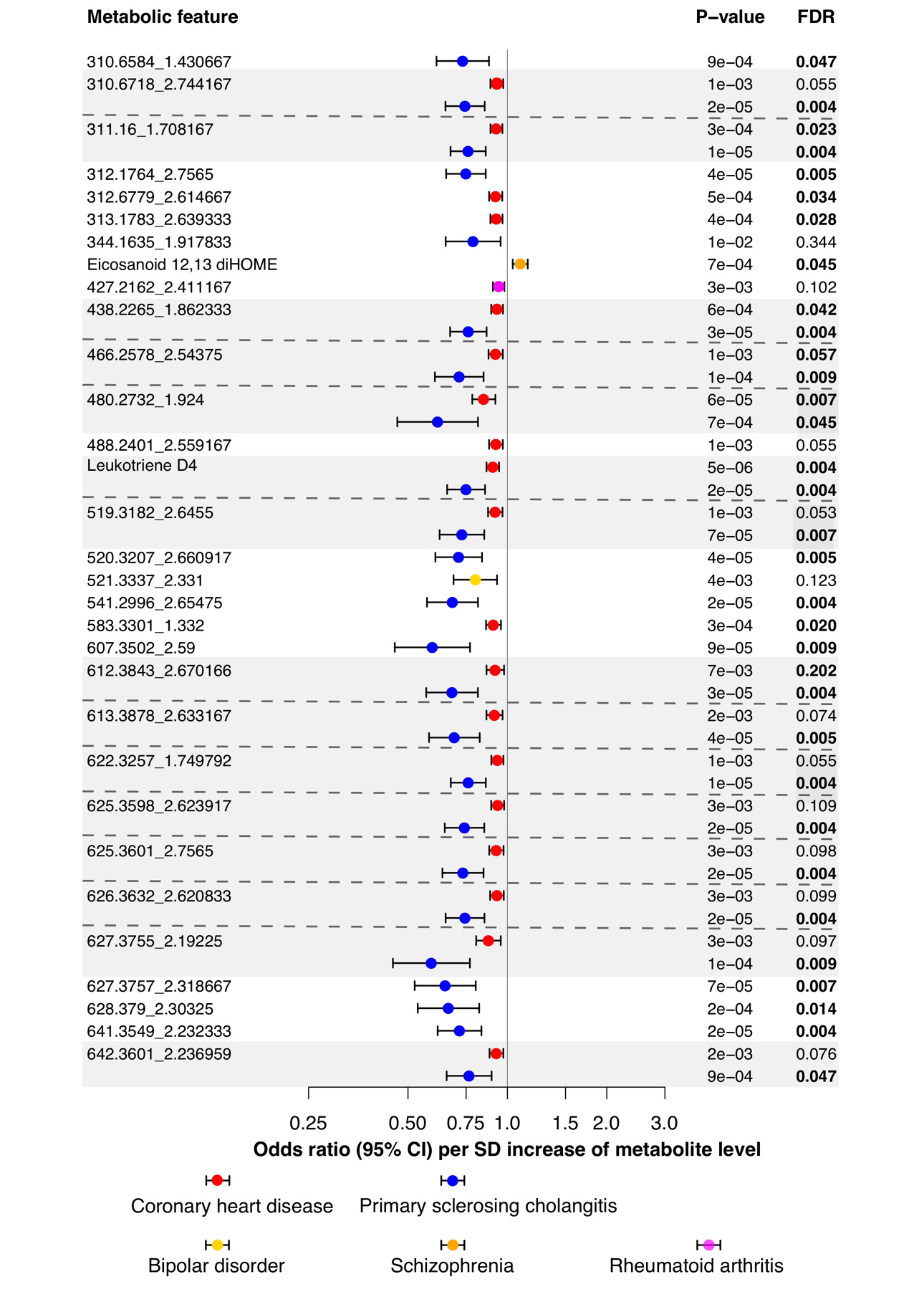
**Figure S1: Distribution of the SNP-heritability *h2* and genomic inflation factor λ for 5,874 circulating metabolite feature levels**. Levels of these metabolite features were associated with genetic variants in FINRISK02 cohort at genome-wide significant threshold (*p*<5×10-8). The SNP-heritability *h****2*** was estimated using 106K pruned SNPs. More than 400 metabolite features (first bar) had small but non-zero SNP-heritability. The inflation factor λ was calculated by dividing the median chi-square statistics of 106K pruned SNPs to 0.4549364. An expected positive association (Pearson correlation coefficient = 0.57) between genomic inflation factor and SNP-heritability was observed.

****

**Figure S2: Pearson correlation heatmap of 31 metabolite feature levels in 7,842 individuals from the FINRISK02 cohort.** The chosen metabolite features correspond to distinct features predicted to have causal effects on disease outcomes by MR (**Figure 2A)**. Clustering was based on the Pearson correlation coefficients using complete linkage method.

****

**Figure S3: Forest plot of MR-based causal effects for metabolite features on various common diseases.** Metabolites are sorted according to increasing mass charge ratio. Effects on coronary heart disease (CHD) are in red, primary sclerosing cholangitis (PSC) in blue, bipolar disorder in gold, schizophrenia in orange, and rheumatoid arthritis in purple. Grey background indicates metabolite features with predicted causal effects on both CHD and PSC. Given causal effects are from the weighted median method with whiskers indicating 95% conference intervals.

****

**Supplementary Table Legends**

**Table S1. Physical characteristics of 11,067 high quality metabolite features obtained using untargeted LC-MS from 7,842 individuals (FINRISK02) after quality control**. Compound identification is a putative identification of the feature ion based on matching with our in-house library of chemical standards on the basis of mass/charge ration and retention time only. Adduct information provides feature ions that have been identified as likely measurements of the same compound based on retention time and measurement correlation. Criteria for matching were a retention time window of 0.01 minutes and a spearman correlation coefficient of at least 0.95. Where possible, the difference in mass between the matched ions was compared with mass differences of common adducts to identify likely adducts and the neutral mass of the measured compound.

**Table S2. Details of all 3,602 study-wide significant associations in FINRISK02 and validated in FHS cohort.** These were independent associations between 609 unique SNPs and 2,319 unique metabolite feature levels from joint model analysis in GCTA-COJO. Genomic coordinates are human genome build GRCh38/hg38.

**Table S3. Details of 1,197 genome-wide significant associations in FINRISK02 and validated in FHS cohort but not passing the study-wide threshold.** These were independent associations after adjustment of top SNPs with GCTA-COJO. Genomic coordinates are human genome build GRCh38/hg38.

**Table S4. Summary of GWAS results for 118 loci and their associations with plasma metabolite levels.** Association metrics (in discovery cohort FINRISK02, validation cohort FHS and meta-analysis of both cohorts) are reported for the top SNP-metabolite association in the locus. Previously reported associations in GWAS Catalog are indicated.

**Table S5. Levels of chemically identified and unidentified metabolite features associated with previously reported or novel loci.** Only associations passing study-wide significant threshold (*p*<3.7×10-11) were counted. Shaded lines denote newly discovered loci being associated with metabolite features.

**Table S6. Metabolite features associated with previously reported and novel loci.** Only associations passing study-wide significant threshold (*p*<3.7×10-11) in meta-analysis were shown.

**Table S7. Details of 45 disease GWASs included in the two-sample Mendelian randomization analysis.** Six diseases causally associated with metabolite feature levels were highlighted in bold.The data was extracted using MRbase R package.

**Table S8. Details of instrumental variables used for the two-sample Mendelian randomization.** Only the summary results for metabolite-disease pairs with significant predicted causal effect were shown.

**Table S9. Causal effects of metabolite levels to diseases inferred by two-sample Mendelian randomization.** Red color denotes significant test of causal effect at *p*<0.05 threshold.

**Table S10. Hazard ratios of 19 metabolite feature levels predicted to have causal effects on CHD.** Hazard ratios on CHD were calculated in FINRISK02 and FHS cohorts independently. Significant results (*p*<0.05) in meta-analysis of two cohorts were highlighted in bold.