Metabolomics (2022) 18:1
https://doi.org/10.1007/s11306-021-01856-6

ORIGINAL ARTICLE

Systems epidemiology of metabolomics measures reveals new relationships between lipoproteins and other small molecules

Fotios Drenos[1,2]

Received: 11 January 2021 / Accepted: 20 November 2021 / Published online: 16 December 2021

Abstract

Introduction The study of lipoprotein metabolism at the population level can provide valuable information for the organisation of lipoprotein related processes in the body. To use this information towards interventional hypotheses generation and testing, we need to be able to identify the mechanistic connections among the large number of observed correlations between the measured components of the system.

Objectives To use population level metabolomics information to gain insight on their biochemical networks and metabolism.

Methods Genetic and metabolomics information for 230 metabolic measures, predominately lipoprotein related, from a targeted nuclear magnetic resonance approach, in two samples of an established European cohort, totalling more than 9400 individuals analysed using phenotypic and genetic correlations, as well as Mendelian Randomisation.

Results More than 20,500 phenotypic correlations were identified in the data, with almost 2000 also showing evidence of strong genetic correlation. Mendelian randomisation, provided evidence for a causal effect between 9496 pairs of metabolic measures, mainly between lipoprotein traits. The results provide insights on the organisation of lipoproteins in three distinct classes, the heterogeneity between HDL particles, and the association, or lack of, between CLA, glycolysis markers, such as glucose and citrate, and glycoproteins with lipids subfractions. Two examples for the use of the approach in systems biology of lipoproteins are presented.

Conclusions Genetic variation can be used to infer the underlying mechanisms for the associations between lipoproteins for hypothesis generation and confirmation, and, together with biological information, to map complex biological processes.

Keywords ALSPAC · Metabolomics · Genetic correlation · Causality · Mendelian randomization · Systems epidemiology

1 Introduction

Metabolomics is the study of the quantitative complement of small molecules in biological systems. The metabolic measures obtained are mostly organic compounds involved in the biochemical reactions of the organism and represent the final stage of the flow of information from the genome to the biological phenotype (Dunn et al., 2011a). One of the main characteristics of metabolomics approaches is their ability to obtain multiple measures from a biological pathway, representing its intermediate steps and chemical compounds involved. As expected, these measures are usually tightly correlated with each other and as metabolic pathways intersect, these correlations can be extensive among metabolomics data (Camacho et al., 2005). For lipoproteins measures the interactions of the particles during lipids metabolism give rise to a large number of correlations that complicate our understanding of their impact on health (Holmes et al., 2015). Understanding the nature of these correlations and the flow of information between them leading to the observed phenotype is a challenging problem and the focus of systems studies (Dunn et al., 2011b). Distinguishing between lipoprotein measures belonging in the same process and those that are correlated due to other reasons has implication on our understanding of the underlying causes for disease and the identification of relevant interventional strategies (Steuer, 2006).

✉ Fotios Drenos
fotios.drenos@brunel.ac.uk

[1] Department of Life Sciences, College of Health, Medicine and Life Sciences, Brunel University London, Uxbridge UB8 3PH, UK

[2] Institute of Cardiovascular Sciences, UCL, London WC1E 6JF, UK

In contrast to phenotypic correlations between biological measures, genetic correlations suggest the presence of a partially shared underlying genetic mechanism (Bulik-Sullivan et al., 2015a; Lee et al., 2012). Genetically correlated metabolic measures can be considered as taking part in the same biological process, though each one might also affect other pathways and processes in the system. In addition to the existence of common mechanisms between metabolic measures, we are also interested in the flow of information between them. Mendelian randomization (MR) is a popular method able to assess the direction of effect between two correlated traits (Davey Smith & Hemani, 2014). In this case, genetic polymorphisms are used as instruments to estimate the causal effect between the two measures and identify pairs of metabolic traits that are part of the same chain of effects from those correlated through other mechanisms.

Here, using a well characterised European population cohort, of children and their mothers, and metabolomics measures from a targeted NMR approach focusing mainly on lipoproteins, fatty acids, and amino acids, I aimed to elucidate the relationships between the available metabolic measures. Understanding of the relationships between the metabolic measures can then be used to test hypotheses.

2 Methods

2.1 Study population

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a population based, prospective birth cohort (www.bris.ac.uk/alspac). The study recruited 14,541 pregnancies and has since followed participants in a number of phases during development and maturity. Full details of the study have been published previously (Boyd et al., 2013; Fraser et al., 2013). Here we use the unrelated offspring of this study at age 7, as the discovery sample, and mothers from the first focus on mothers sample collection, as replication. The study website contains details of all the data that is available through a fully searchable data dictionary http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and from the UK NHS National Health Service Local Research Ethics Committees. Participants have provided informed consent for the use of the data.

2.2 Serum NMR metabolomics

A high-throughput serum nuclear magnetic resonance (NMR) metabolomics platform was used to quantify 230 metabolic measures representing a broad molecular signature of systemic metabolism (Soininen et al., 2015). The measured set covers multiple metabolic pathways, including lipoprotein lipids and subclasses, fatty acids and fatty acid composition, as well as amino acids and glycolysis precursors. This applied NMR-based metabolic profiling platform has recently been used in various epidemiological and genetic studies (Beaney et al., 2016; Drenos et al., 2016; Würtz et al., 2016a). Applications of this high-throughput metabolomics platform has been reviewed (Soininen et al., 2015) and details of the experimentation have been described elsewhere (Soininen et al., 2009). Previous work (Würtz et al., 2017) has shown excellent correlation between the NMR determined measures and their respective clinically assessed values.

For 5645 of the ALSPAC young participants, 48.5% females, metabolic measures were obtained from serum under non-fasting conditions. For 4530 ALSPAC mothers, at a median age of 48 years (IQR 45-51), the measures were obtained from overnight, or at least 6 h, fasted samples.

2.3 Genotyping

The ALSPAC children were genotyped using the Illumina HumanHap610 array. The ALSAPC mothers were genotyped with the Illumina HumanHap550 array. Standard metrics were employed to assess the quality of these data: individual call rate > 97%; heterozygosity threshold: 0.34; minor allele frequency of < 0.005%; SNP call rate of > 97%; and Hardy–Weinberg equilibrium (HWE) ($p < 5 \times 10^{-7}$) (Bønnelykke et al., 2013).

2.4 Statistical analysis

The metabolomics measures were transformed using a rank-based inverse normal transformation (Blom, 1958). The SNPs effects on the metabolites were obtained through PLINK (Purcell et al., 2007) using a linear model adjusted for age or sex, as appropriate. Independent SNPs were obtained using a pairwise linkage disequilibrium of $r^2 < 0.01$ per chromosome through PLINK. When a pair of siblings was present in the data, one of them was randomly removed from the sample. The phenotypic correlation between metabolic measures was represented by the Pearson's correlation coefficient of their transformed values. Pearson's correlation coefficients and their respective p-values were estimated through the cor.test in R (Team, 2008). The genetic correlation was estimated through (1) the LD score regression (Bulik-Sullivan et al., 2015b), as found in LD Hub (Zheng et al., 2017), for measures with previous genome wide association study (GWAS) results available, (2) Bivariate REML on individual level data, for measures with no published GWAS (Lee et al., 2012), and (3) approximated by the inverse variance weighted regression between the normalised beta coefficients of the independent SNP-trait associations, for a small number of pairs where both LD score and GCTA algorithms failed to converge. Principal component analysis (PCA) was performed through the caret package in R (Kuhn, 2008). Only independent SNPs with an effect > 3 standard deviations from zero were used for Mendelian randomisation analysis. We used individual SNPs estimates in the weighted inverse variance and MR-Egger methods, as described elsewhere (Bowden et al., 2015), to obtain the causal effects between the traits and test for pleiotropy. The tree and network of lipoproteins characteristics presented were constructed through igraph in R (Csardi & Nepusz, 2006). The rooted tree was based on an algorithm starting with the largest particle and selecting the next particle based on the -log of the association p-value. P-values higher than the threshold were considered as equal to 1. All plots were constructed in R using either the corrplot (Simko, 2016) or ggplot2 packages (Wickham, 2016).

3 Results

3.1 Samples characteristics

In the Avon Longitudinal Study of Parents and Children (ALSPAC), excluding all subjects with missing values, resulted in 5353 unrelated children at age 7 and 4120 mothers with available metabolomics measurements. In the ALSPAC offspring, principal component analysis showed that 44 principal components (PCs) accounted for 0.99 of the metabolic measures variance. The corresponding figure in the mothers sample was 0.98. Based on this, our Bonferroni adjusted p-value threshold for the pairwise correlations was set at 5.28×10^{-5} for correlation tests and 2.64×10^{-5} for regression based tests. In total 465,740 genotyped SNPs passed the quality control criteria in the two samples, with 12,516 SNPs found to be independent (LD $r^2 < 0.01$) and used to estimate the correlation of SNP effects between the traits and in selecting MR instruments.

3.2 Phenotypic correlations

Of the 26,335 Pearson's correlation coefficients tested between the metabolic measures, 23,032 showed evidence of phenotypic association. A table showing all correlation coefficients and their respective *p*-values can be found in the Supplementary material (Table S1) and plotted in Fig. S1. When the lipoprotein measures were ordered by size, as the current understanding links size to function, four clusters of high correlation were evident in their concentration measures, two major and two minor. The first major cluster included Chylomicrons and extra-large very low density lipoprotein (VLDL) and extended to very-large, large, medium and small VLDL particles. The second major cluster included very-small VLDL, intermediate density lipoprotein (IDL) and the various sizes of low density lipoprotein (LDL). The majority of measures were associated across the two clusters except triglycerides and free cholesterol measures. Both of the minor clusters of strong correlations were in the high-density lipoprotein (HDL) measures, with one cluster including particles of very-large and large size and the second measures of very-small HDL. Medium size HDL measures were associated with both. Again, cross cluster correlations were present throughout. The other prominent feature of the phenotypic correlations matrix was the complex correlations pattern between glycolysis, amino acids, ketone bodies, fluid balance, and inflammation markers with lipids, although most were consistently associated with the larger VLDL particles measures. Results from the mothers were similar, replicating 20,758 of the associations, and showed the same major features. The results can be seen in Table S2 and Fig. S2.

3.3 Genetic correlations

Of the 26,335 possible pairs of metabolic measures, 5050 were tested through LD score regression (Bulik-Sullivan et al., 2015a) in LD Hub (Zheng et al., 2017) from external data (Kettunen et al., 2016) with 1551 showing evidence of genetic correlation, 24,484 were estimated using bivariate REML (Lee et al., 2012) in the children sample with 2330 showing evidence of genetic correlation, while 24,121 had evidence of correlation when the beta coefficients of the independent SNPs were considered. Tables S3 and Fig. S3, show the correlation coefficients and their p-values obtained for all pairs of measures. The Pearson correlation between the estimates of the three methods were: 0.753 (CIs 0.740–0.766) between LD score and bivariate REML, 0.827 (CIs 0.818–0.835) between LD score and the correlation of SNP effects and 0.836 (CIs 0.832–0.839) for bivariate REML and the correlation of SNP effects. The main clusters of high correlations present in the phenotypic level were also evident for the genetic correlations, but in this case, the pairwise associations tended to be confined mostly within the observed clusters. The majority of VLDL measures were associated with each other. IDL and LDL measures, formed another cluster of genetic correlations and they were also correlated to some of the medium and smaller VLDL measures. Very large and large HDL measures had evidence of genetic correlations between them, which were less pronounced for medium HDL and mostly absent for small HDL measures. Measures of large HDL showed evidence of correlation with VLDL measures. Fatty acids were correlated to the measures of IDL, LDL and the larger HDL, while the unsaturated fatty acids were also correlated with VLDL measures. Finally, both isoleucine and glycoproteins acetyls had evidence of correlation with VLDL measures. The main patterns of correlation were replicated in the mothers sample, where bivariate REML identified 5040 pairwise correlations while 24,459 pairs of metabolites had correlated SNP effects (Table S4 and Fig. S4).

In general, the magnitudes of phenotypic correlations and genetic correlations between the measures had similar patterns (Fig. 1), though their statistical significance evidence differed. There were 3579 correlation pairs where both phenotypic and genetic correlations were evident and 19,543 pairs of measures that were correlated in the phenotypic level but not the genetic level. In the mothers sample, 1966 of the 3579 and 15,692 of the 19,543 relationships were replicated. These 1966 pairs can be found in Table S5.

3.4 Mendelian randomisation

To better assess the nature of correlations seen between the metabolomics measures, we performed MR analysis for all pairs in both directions. The number of SNPs used for each of the metabolic measures can be seen in Table S6. Of the 52,670 associations tested 26,909 had evidence of a causal effect. The full list of results can be seen in Tables S7 and their respective p-values in Table S8. From the 26,909 relationships observed in the offspring data, 20,663 were replicated in the mother's sample. Of these, 9297 were bidirectional associations, out of the 12,188 observed in the discovery sample (Table S9), and 199 one-directional causal effects, from the 2533 observed initially (Table S10), with the rest being bidirectional associations replicated in one direction. Figure 2, summarises all the replicated p-values. The three major clusters evident in the genetic correlation results around VLDL, IDL and LDL, and larger HDL particles were again visible as bidirectional associations. The VLDL cluster included measures of the largest lipoprotein particles, such as chylomicrons and very large VLDL particles, to smaller VLDL. The small VLDL measures were

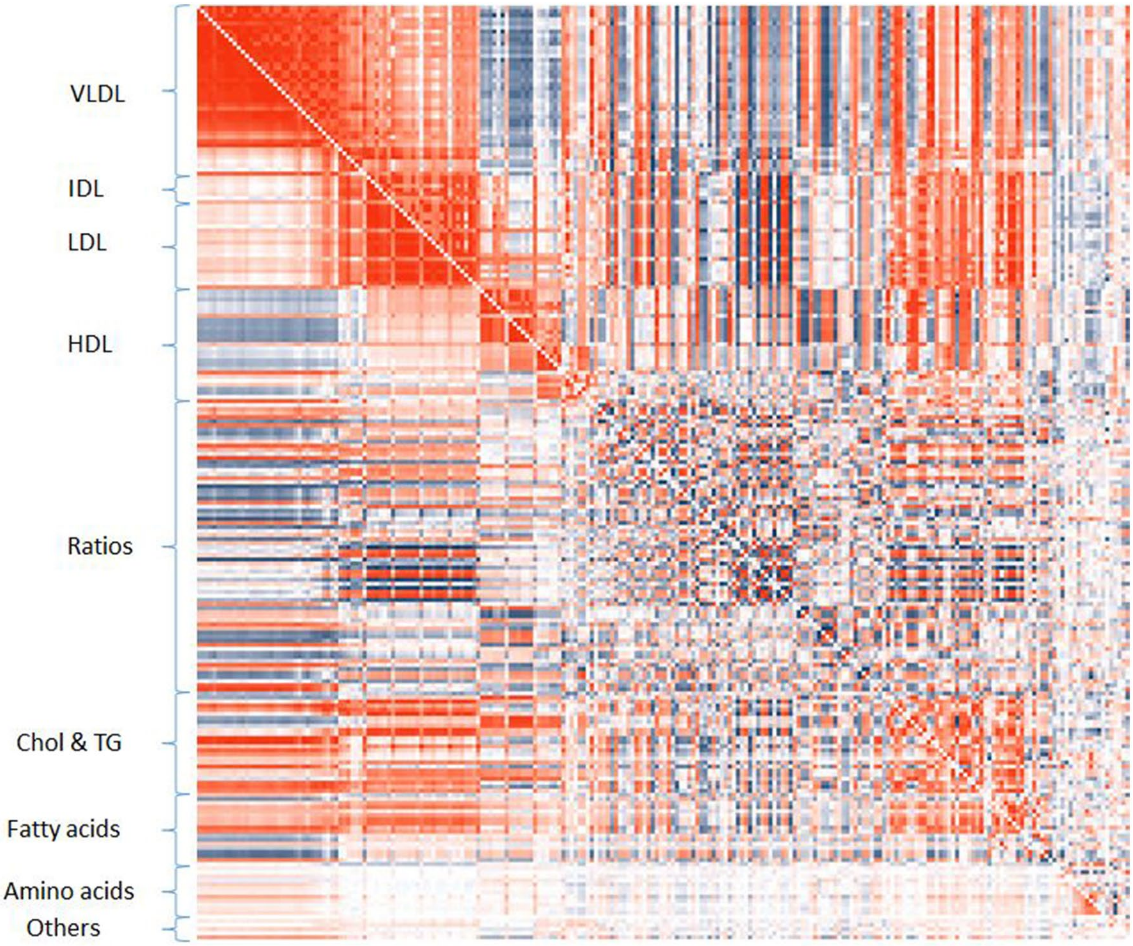

Fig. 1 Correlations between 230 metabolites measured through a targeted metabolomics NMR platform. Red for positive correlation and blue for negative. The lower left part of the square shows the correlation between the levels of the metabolic measures. The upper right part shows the correlations of their genetic effects. High degree of similarity is evident in the two triangles in terms of the sign and level of correlation. Only some of these correlations are statistically significant for both levels of correlation

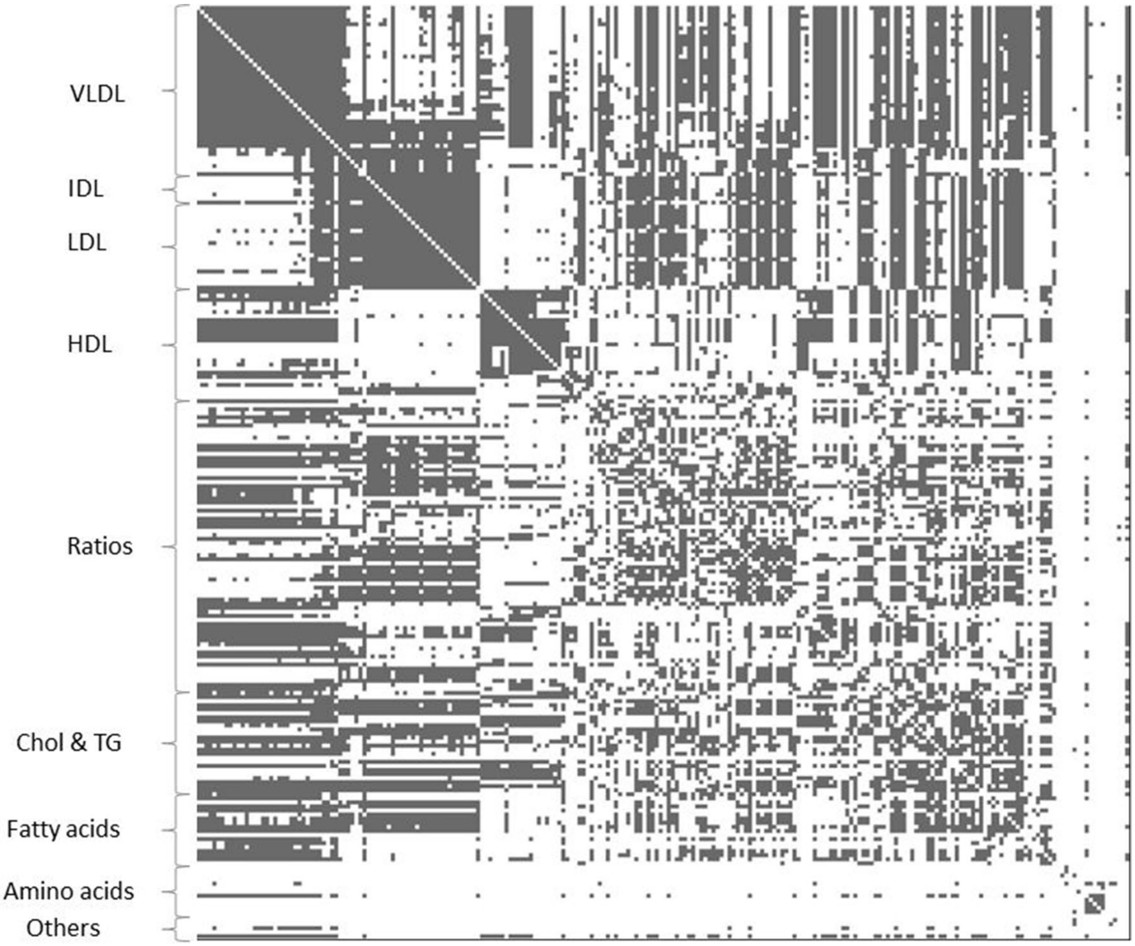

Fig. 2 Evidence of MR effects of the row metabolic measure on the column metabolic measure obtained from a targeted metabolomics NMR platform. Grey for a causal association, white for associations not reaching the pre-specified p-value threshold

associated with both the VLDL cluster and the remnant and LDL cluster of associations. The VLDL cluster was also bidirectionally associated with triglycerides levels in remnant particles and medium and small HDL, but not with triglycerides in large and very large HDL. Phospholipids on LDL, and the large and very large HDL subclasses were associated with the VLDL cluster. Large HDL particle measures also had multiple associations with the VLDL cluster for total, esterified and free cholesterol measures. The remnant lipoprotein and LDL measures cluster were associated with overall esterified and free cholesterol as well as total and esterified cholesterol in small HDL particles. The MR results support a clear distinction between the larger HDL lipoproteins (extra-large and large) and small HDL. Medium HDL measures are more similar to the larger HDL measures, though they also have overlapping effects with total lipids and triglycerides measures of small HDL. Fatty acids associations with both the VLDL and remnant and LDL clusters were observed, though the degree of unsaturation, length of fatty acid chain and ratios of fatty acids were associated only with the VLDL cluster, while omega-3 measures were associated with the remnant and LDL cluster. Of note, isoleucine and alpha-1-acid glycoprotein were linked to the VLDL measures cluster, with the later also showing bidirectional associations with the larger HDL measures.

3.5 Adjusting for pleiotropy

Using the MR-Egger approach, as a sensitivity analysis for the effect of pleiotropy in our estimates, resulted in 7074 associations, in accordance with the lower statistical power of the method, with 2224 pairs being bidirectional associations and 2626 in one direction only (Tables S11 and S12). Of the MR-Egger observed associations 6000 were also observed when the standard approach was used. In the children sample, 1719 associations also had evidence for the presence of pleiotropy based on their intercept (Table S13 and S14). MR-Egger analysis of the ALSPAC mothers revealed 176 pairs of bidirectional associations and 305 in one direction, with 137 of these also showing evidence of pleiotropy (Tables S15–S18). Compared to the standard approach 521 pairs of metabolites also had evidence of association in the MR-Egger analysis. In total, the mothers sample provided replication for 486 associations seen in the children.

3.6 Examples

To demonstrate the potential use of the results in trying to infer the mechanistic relationships between the metabolic measures, we focused on two examples. VLDL is the main transport form of endogenous triglycerides in the body. It is produced in the hepatocytes and released into circulation progressively losing its triglyceride content to give rise to remnant VLDL, IDL and LDL particles of smaller sizes (Marshall et al., 2012). Using the nine measures for the triglycerides to total lipids ratio in VLDL (except chylomicrons and extremely large VLDL which can carry triglycerides from diet), IDL and LDL lipoproteins, a very-large VLDL one directional routed tree (see methods) was constructed, recreating the process of lipoproteins metabolism relatively accurately (Fig. 3). The second example focused on the relationships of the small HDL measures with the rest of the lipoproteins measures (Fig. 4). The triglyceride content of small HDL particles vertex was located within the VLDL measures cluster, while total and esterified cholesterol were closely related with the remnant and LDL cluster, which were the main cholesterol carrying particles. Small HDL total lipids and free cholesterol showed a small number of connections and were situated away from other clusters. The phospholipids in the small HDL vertex had an equal distance from other clusters. Finally, the concentration of small HDL particles was mostly associated and located close to the area of the larger HDL particle measures.

4 Discussion

Using two samples of mothers and young participants from a European population cohort and data from a targeted metabolomics platform, predominately of lipoproteins, I illustrate the use of popular epidemiological approaches to assess the

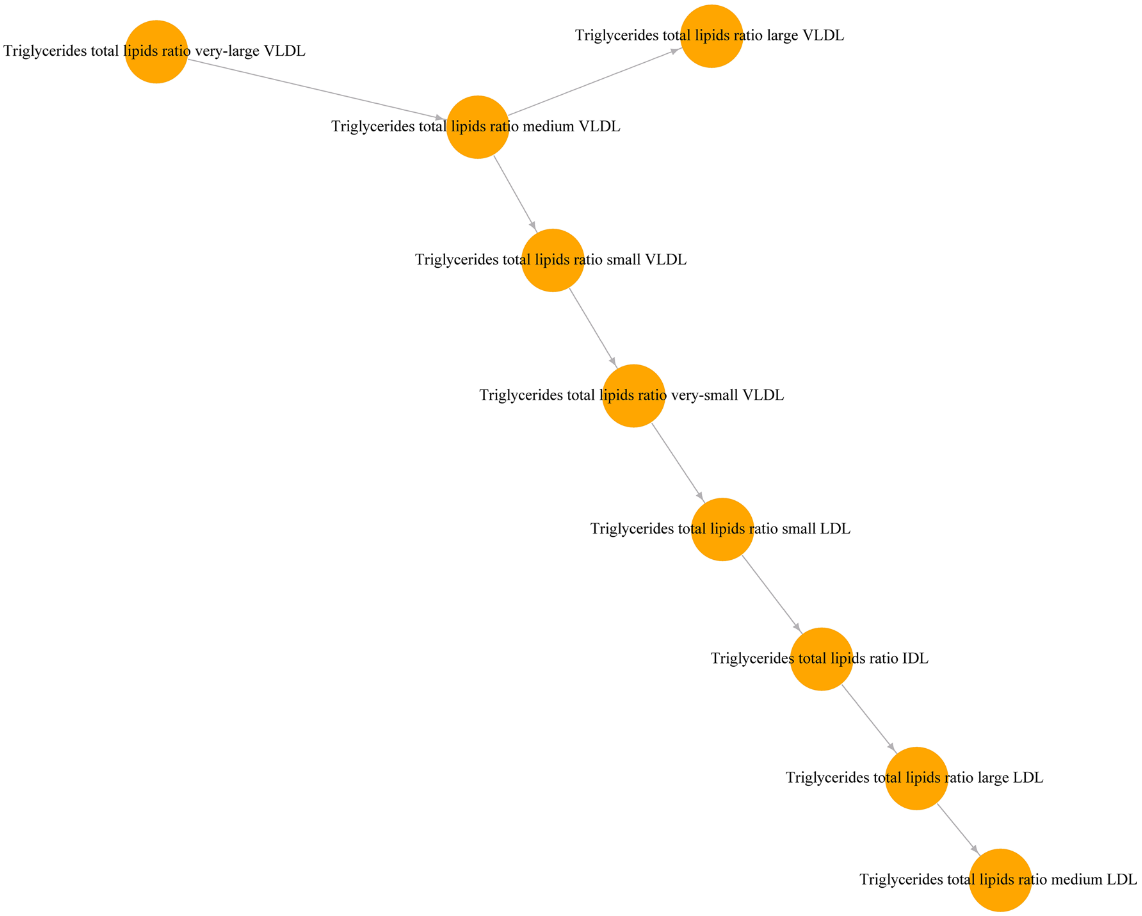
Triglycerides total lipids ratio very-large VLDL
Triglycerides total lipids ratio large VLDL
Triglycerides total lipids ratio medium VLDL
Triglycerides total lipids ratio small VLDL
Triglycerides total lipids ratio very-small VLDL
Triglycerides total lipids ratio small LDL
Triglycerides total lipids ratio IDL
Triglycerides total lipids ratio large LDL
Triglycerides total lipids ratio medium LDL

Fig. 3 A one directional routed tree representing the strongest (smaller *p*-value) associations between the nine measures of triglyceride concentration in lipoprotein particles

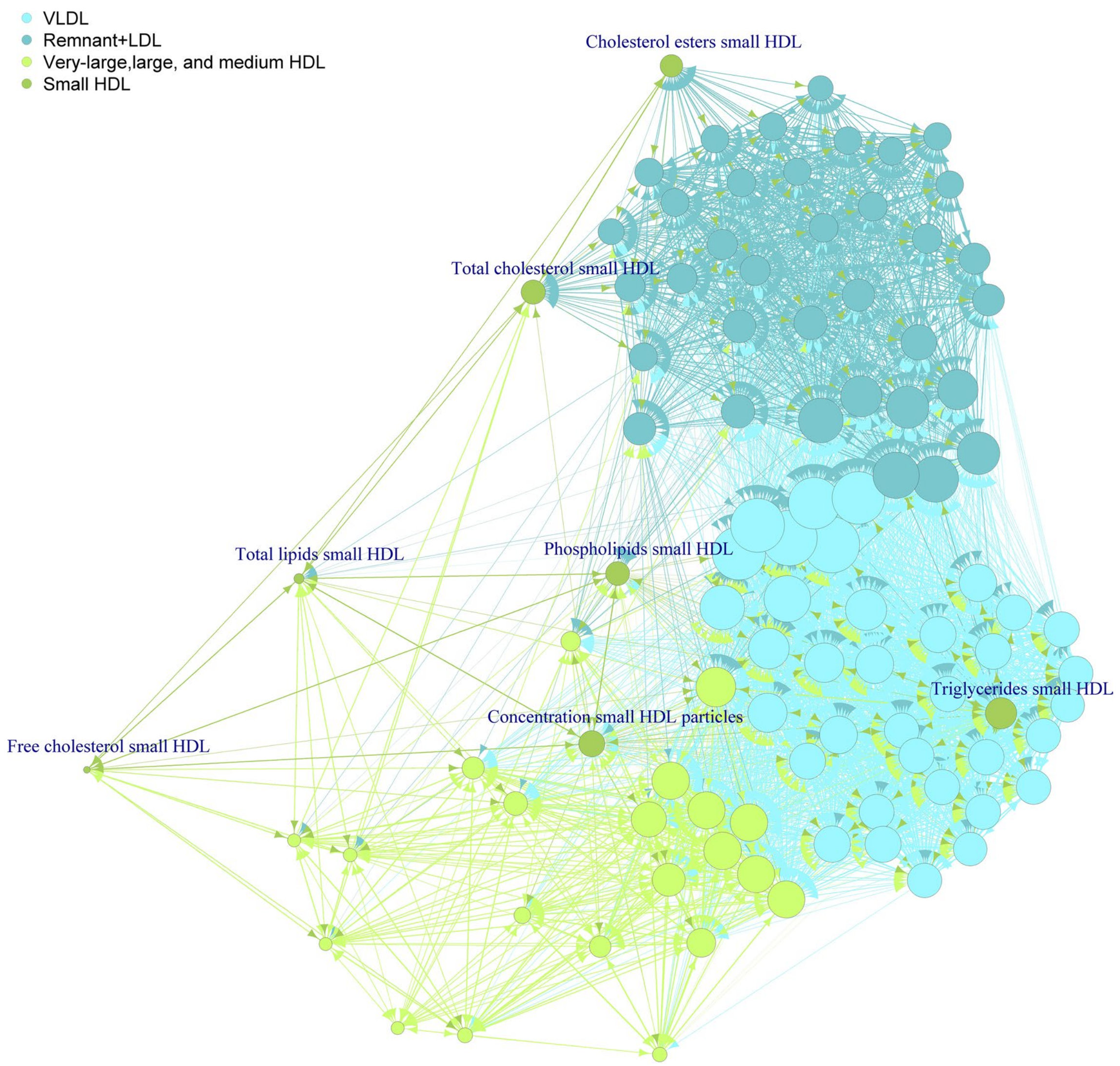
VLDL
Remnant+LDL
Very-large,large, and medium HDL
Small HDL
Cholesterol esters small HDL
Total cholesterol small HDL
Total lipids small HDL
Phospholipids small HDL
Concentration small HDL particles
Free cholesterol small HDL
Triglycerides small HDL

Fig. 4 A network of MR replicated associations highlighting the relationships of small HDL measures with other lipoprotein subfractions measures. The size of the node is proportional to the number of connections to the node. The log *p*-value was used as weights for the edges of the network

relationship between metabolic measures. The results show the presence of strong correlations between the lipoproteins and other measures, at both the phenotypic and genetic levels, and a wide range of causal effects between them, including bidirectional associations. Three main causal clusters of lipoprotein relationships were evident in the results. Heterogeneity of links between HDL lipoprotein particles of different sizes was also observed. Additional insights were obtained for the relationships between CLA, glycolysis markers, and glycoproteins with lipids subfractions.

As the metabolomics platform used is focused mainly on lipoproteins measures, the large number of phenotypic correlations is of no surprise, though similar levels of correlation should be expected within most metabolomics datasets. Except the more obvious reasons of correlations between the metabolic measures, such as the mathematical relationship between them, were ratios are considered, and their proximity in biological processes, more complex mechanisms such as chemical equilibrium, mass conservation, confounding from other internal or external modifiers, and global perturbations due to the conditions of the sample measurement, can also contribute to the correlations observed between metabolites (Camacho et al., 2005; Steuer, 2006).

In the present study, measures of the triglyceride rich VLDL particles made up a distinct group, while the products of their remodelling, triglyceride poor and depleted particles of very small VLDL, IDL and LDL, formed another closely correlated group. These phenotypic correlation features were also present when the genetic correlations were considered, while the MR analysis provided further evidence for both bi-directional and one-directional effects within the two clusters, but very few associations between the clusters. A study on the metabolic profile of statins and the rs12916 *HMGCR* gene polymorphism, identified that disruption of the mevalonate pathway, in addition to the expected LDL lowering effect, also produces large changes on the very-small VLDL and IDL concentration measures and only modest changes on the larger VLDL particles (Würtz et al., 2016b). These results fit very well with the idea of two distinct clusters of associations and the two-way causal effects observed.

The correlations between HDL subfractions were also arranged in two blocks, one for larger and medium and one for small HDL particles. The sub-speciation of HDL to a number of distinct proteins and lipids combinations sharing a similar density has been previously suggested based on proteomics analysis (Davidson et al., 2009) and has been used to explain the large number of properties associated with HDL (Rosenson et al., 2012). The majority of causal associations were between and within the measures of very large and large HDL particles, suggesting that these are quite different from the smaller HDL particles. Interestingly, triglycerides in small and medium, but not large and very large, HDL particles were causally associated with the larger VLDL particles measures and only the smallest HDL particles, corresponding to HDL3, showed evidence of causal effects with the triglyceride depleted, cholesterol rich, low density lipoprotein particles. The observed relationships correspond well to our current understanding of the molecular exchanges taking place during lipoproteins metabolism, with triglycerides moving from VLDL particles to HDL particles, with esterified cholesterol following the opposite direction through the action of the cholesteryl ester transfer protein (Marshall et al., 2016), though now we can provide further information on the size of particles involved.

Other interesting relationships included the associations of conjugated linoleic acid (CLA). CLA is a popular dietary supplement associated with a number of suggested beneficial effects on common diseases and BMI, while recent studies correlated CLA with a decrease of LDL (Derakhshande-Rishehri et al., 2015) and HDL (Kim et al., 2016). We did not find any evidence for a causal effect of CLA on either LDL- or HDL-cholesterol concentrations in the present study. In contrast, we found positive bidirectional associations of CLA with measures of esterified and total cholesterol, as well as triglycerides, in large and very large VLDL particles.

No causal associations between glycolysis markers and lipids were observed. Previously, insulin has been implicated in lipogenesis and VLDL production (Brown & Gibbons, 2001) but a second study looking at the effect of glucose metabolism on the transcriptional regulation of genes involved in VLDL assembly and secretion, did not find any major effects (Morral et al., 2007). The current results do not support the existence of a causal effect between glucose and VLDL concentration and composition measures. Similarly, Citrate, a popular additive to foods and an intermediate product of the Krebs cycle, has been described as a "fundamental precursor" for the endogenous production of cholesterol (Leandro et al., 2016). The observed results do not support any causal associations between plasma measured citrate and lipoprotein measures or glucose.

Evidence for the causal effect of Glycoprotein acetyls, mainly a1-acid glycoprotein (AGP), on large and medium VLDL concentration measures and particle diameter were observed. AGP is an acute-phase protein believed to be involved in a wide range of biological processes, including immuno-modulation, drug compound transport, maintaining capillary function, sphingolipid biosynthesis and glucose and insulin metabolism (Luo et al., 2015). AGP has been suggested as a marker for all-cause mortality (Fischer et al., 2014). The complete function of the protein and how it can interact with VLDL concentration is not known, but our results suggest a role in lipoprotein metabolism that has not previously been identified.

Two examples for the use of the results towards understanding and mapping metabolic networks have been illustrated. The first example looking at the metabolism of triglyceride rich VLDL fits almost perfectly with the current understanding of the process (Marshall et al., 2016). The second example is mapping the characteristics of the small HDL particles in relation to other lipoproteins. According to our results, most of the cholesterol exchange, in the form of esterified cholesterol, is taking place between small HDL and remnant and LDL particles. In contrast, the node of triglyceride concentration of the small HDL particles is embedder within the VLDL cluster of measures. Both features of the network are well established (Marshall et al., 2016). Phospholipids and free cholesterol in small HDL particles are believed to be obtained by the interaction of the early HDL particle with cell membranes (Gurr et al., 2016). This can explain the lack of associations seen with the free cholesterol measure, but the associations with both the VLDL and the remnant and LDL clusters of measures suggests the existence of additional mechanisms.

A number of limitations are evident in this work. The discovery sample was of unfasted children aged 7 with 51.5% of them boys, while the replication sample was of partially related adult fasted women with a mean age of 47.9. This means that the replication sample could confirm the common observed associations, but false positives are indistinguishable from changes due to age and sex. The selection of SNPs to be used as instruments were from the same sample where the MR was performed. This can introduce bias in the analysis resulting in the identification of associations that are not causal (Taylor et al., 2014), but the use of replication and previous biological evidence, suggest that the main associations observed are indeed causal. Finally, we treat each metabolic measure as an independent variable without modelling their potential interactions which will require the use of pathway analysis and latent variables (Burgess et al., 2015), but this will shift the focus of the analysis from hypothesis generating to modelling mediation between a small number of preselected measures.

Focus areas for the use of metabolomics are the understanding and reconstruction of the molecular processes underpinning health and disease and the identification of new risk factors. Genetic variation can be used to provide the direction of effects in a correlation network and disentangle the flow of information in such systems. Here we used measures from a targeted metabolomics platform to explore the correlations between the metabolic measures in the context of common underlying mechanisms with the help of genetic variants. The results of this work provide evidence for, or against, a number of interesting phenotypic associations between the lipoprotein measures and other metabolites and illustrate the challenges and potential uses of this kind of approaches in metabolomics data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11306-021-01856-6.

Acknowledgements I am extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. This publication is the work of the author and Fotios Drenos will serve as guarantors for the contents of this paper. GWAS data was generated by Sample Logistics and Genotyping Facilities at the Wellcome Trust Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe. The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. Grants from the British Heart Foundation (SP/07/008/24066), the UK Medical Research Council (G1001357) and Wellcome Trust (WT092830M and WT088806) funded data collection, and completion of NMR metabolites, from the ALSPAC mothers. Ethical approval for the study was obtained from the Avon Longitudinal Study of Parents and Children (ALSPAC) Ethics and Law Committee and from the UK NHS National Health Service Local Research Ethics Committees. Participants have provided informed consent for the use of the data. Individual level data are available from the ALSPAC study following an application submission and review from the board. Details can be found at http://www.bristol.ac.uk/alspac/researchers/access/.

Funding This work has been partly supported through a BHF-Turing Cardiovascular Data Science Award (BHF-Turing-19/2/1008).

Declarations

Conflict of interest There are no competing interests concerning this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The data are available from the ALSPAC study following request to the study.

References

Beaney, K. E., Cooper, J. A., McLachlan, S., Wannamethee, S. G., Jefferis, B. J., Whincup, P., Ben-Shlomo, Y., Price, J. F., Kumari, M., Wong, A., Ong, K., Hardy, R., Kuh, D., Kivimaki, M., Kangas, A. J., Soininen, P., Ala-Korpela, M., Drenos, F., & Humphries, S. E. (2016). Variant rs10911021 that associates with coronary heart disease in type 2 diabetes, is associated with lower concentrations of circulating HDL cholesterol and large HDL particles but not with amino acids. *Cardiovascular Diabetology, 15*, 115.

Blom, G. (1958). *Statistical estimates and transformed beta-variables*. Wiley.

Bønnelykke, K., Matheson, M. C., Pers, T. H., Granell, R., Strachan, D. P., Alves, A. C., Linneberg, A., Curtin, J. A., Warrington, N. M., Standl, M., Kerkhof, M., Jonsdottir, I., Bukvic, B. K., Kaakinen, M., Sleimann, P., Thorleifsson, G., Thorsteinsdottir, U., Schramm, K., Baltic, S., Kreiner-Møller, E., Simpson, A., St. Pourcain, B., Coin, L., Hui, J., Walters, E. H., Tiesler, C. M. T., Duffy, D. L., Jones, G., Aagc, Ring, S. M., McArdle, W. L., Price, L., Robertson, C. F., Pekkanen, J., Tang, C. S., Thiering, E., Montgomery, G. W., Hartikainen, A. -L., Dharmage, S. C., Husemoen, L. L., Herder, C., Kemp, J. P., Elliot, P., James, A., Waldenberger, M., Abramson, M. J., Fairfax, B. P., Knight, J. C., Gupta, R., Thompson, P. J., Holt, P., Sly, P., Hirschhorn, J. N., Blekic, M., Weidinger, S., Hakonarsson, H., Stefansson, K., Heinrich, J., Postma, D. S., Custovic, A., Pennell, C. E., Jarvelin, M. -R., Koppelman, G. H., Timpson, N., Ferreira, M. A., Bisgaard, H., Henderson, A. J., for the, E.G. and Lifecourse Epidemiology, C. (2013) Meta-analysis of genome-wide association studies identifies 10 loci influencing allergic sensitization. *Nature genetics* 45, 902–906.

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology, 44*, 512–525.

Boyd, A., Golding, J., Macleod, J., Lawlor, D. A., Fraser, A., Henderson, J., Molloy, L., Ness, A., Ring, S., & Davey Smith, G. (2013). Cohort Profile: the 'children of the 90s'—the index offspring of the Avon longitudinal study of parents and children. *International Journal of Epidemiology, 42*, 111–127.

Brown, A.-M., & Gibbons, G. F. (2001). Insulin Inhibits the Maturation Phase of VLDL Assembly via a Phosphoinositide 3-Kinase—Mediated Event. *Arteriosclerosis, Thrombosis, and Vascular Biology, 21*, 1656–1661.

Bulik-Sullivan, B., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. -R., ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control, C., Duncan, L., Perry, J. R. B., Patterson, N., Robinson, E. B., Daly, M. J., Price, A. L. and Neale, B. M. (2015a). An atlas of genetic correlations across human diseases and traits. *Nature Genetics* 47, 1236–1241.

Bulik-Sullivan, B. K., Loh, P. -R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M. J., Price, A. L. and Neale, B. M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics* 47, 291–295.

Burgess, S., Daniel, R. M., Butterworth, A. S., & Thompson, S. G. (2015). Network Mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. *International Journal of Epidemiology, 44*, 484–495.

Camacho, D., de la Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. *Metabolomics, 1*, 53–63.

Csardi, G. and Nepusz, T. (2006) The igraph Software Package for Complex Network Research. *InterJournal* Complex Systems, 1695.

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics, 23*, R89–R98.

Davidson, W. S., Silva, R. A. G. D., Chantepie, S., Lagor, W. R., Chapman, M. J., & Kontush, A. (2009). Proteomic analysis of defined HDL subpopulations reveals particle-specific protein clusters. *Relevance to Antioxidative Function, 29*, 870–876.

Derakhshande-Rishehri, S.-M., Mansourian, M., Kelishadi, R., & Heidari-Beni, M. (2015). Association of foods enriched in conjugated linoleic acid (CLA) and CLA supplements with lipid profile in human studies: A systematic review and meta-analysis. *Public Health Nutrition, 18*, 2041–2054.

Drenos, F., Davey Smith, G., Ala-Korpela, M., Kettunen, J., Würtz, P., Soininen, P., Kangas, A. J., Dale, C., Lawlor, D. A., Gaunt, T. R., Casas, J. -P. and Timpson, N. J. (2016) Metabolic Characterization of a Rare Genetic Variation Within APOC3 and Its Lipoprotein Lipase–Independent Effects. *Circulation: Cardiovascular Genetics* 9, 231–239.

Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R., & Griffin, J. L. (2011). Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chemical Society Reviews, 40*, 387–426.

Fischer, K., Kettunen, J., Würtz, P., Haller, T., Havulinna, A. S., Kangas, A. J., Soininen, P., Esko, T., Tammesoo, M. -L., Mägi, R., Smit, S., Palotie, A., Ripatti, S., Salomaa, V., Ala-Korpela, M., Perola, M., & Metspalu, A. (2014). Biomarker Profiling by Nuclear Magnetic Resonance Spectroscopy for the Prediction of All-Cause Mortality: An Observational Study of 17,345 Persons. *PLOS Medicine* 11, e1001606.

Fraser, A., Macdonald-Wallis, C., Tilling, K., Boyd, A., Golding, J., Davey Smith, G., Henderson, J., Macleod, J., Molloy, L., Ness, A., Ring, S., Nelson, S. M., & Lawlor, D. A. (2013). Cohort Profile: The avon longitudinal study of parents and children: ALSPAC mothers cohort. *International Journal of Epidemiology, 42*, 97–110.

Gurr, M. I., Harwood, J. L., Frayn, K. N., Murphy, D. J., & Mitchell, R. H. (2016). *Lipids: Biochemistry, Biotechnology and Health*, 6 edn. Wiley Blackwell.

Holmes, M. V., Asselbergs, F. W., Palmer, T. M., Drenos, F., Lanktree, M. B., Nelson, C. P., Dale, C. E., Padmanabhan, S., Finan, C., Swerdlow, D. I., Tragante, V., van Iperen, E. P. A., Sivapalaratnam, S., Shah, S., Elbers, C. C., Shah, T., Engmann, J., Giambartolomei, C., White, J., Zabaneh, D., Sofat, R., McLachlan, S., Doevendans, P. A., Balmforth, A. J., Hall, A. S., North, K. E., Almoguera, B., Hoogeveen, R. C., Cushman, M., Fornage, M., Patel, S. R., Redline, S., Siscovick, D. S., Tsai, M. Y., Karczewski, K. J., Hofker, M. H., Verschuren, W. M., Bots, M. L., van der Schouw, Y. T., Melander, O., Dominiczak, A. F., Morris, R., Ben-Shlomo, Y., Price, J., Kumari, M., Baumert, J., Peters, A., Thorand, B., Koenig, W., Gaunt, T. R., Humphries, S. E., Clarke, R., Watkins, H., Farrall, M., Wilson, J. G., Rich, S. S., de Bakker, P. I. W., Lange, L. A., Smith, G. D., Reiner, A. P., Talmud, P. J., Kivimaki, M., Lawlor, D. A., Dudbridge, F., Samani, N. J., Keating, B. J., Hingorani, A. D., Casas, J. P., & Consortium, U. (2015) Mendelian randomization of blood lipids for coronary heart disease. *European Heart Journal* 36, 539-+.

Kettunen, J., Demirkan, A., Würtz, P., Draisma, H. H. M., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A. J., Lyytikäinen, L.-P., Pirinen, M., Pool, R., Sarin, A.-P., Soininen, P., Tukiainen, T., Wang, Q., Tiainen, M., Tynkkynen, T., Amin, N., Zeller, T., … Ala-Korpela, M. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications, 7*, 11122.

Kim, B., Lim, H. R., Lee, H., Lee, H., Kang, W., & Kim, E. (2016). The effects of conjugated linoleic acid (CLA) on metabolic syndrome patients: A systematic review and meta-analysis. *Journal of Functional Foods, 25*, 588–598.

Building Predictive Models in R Using the caret Package. *2008* 28, 26

Leandro, J. G. B., Espindola-Netto, J. M., Vianna, M. C. F., Gomez, L. S., DeMaria, T. M., Marinho-Carvalho, M. M., Zancan, P., Neto, H. A. P., & Sola-Penna, M. (2016). Exogenous citrate impairs glucose tolerance and promotes visceral adipose tissue inflammation in mice. *British Journal of Nutrition, 115*, 967–973.

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., & Wray, N. R. (2012). Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics, 28*, 2540–2542.

Luo, Z., Lei, H., Sun, Y., Liu, X., & Su, D.-F. (2015). Orosomucoid, an acute response protein with multiple modulating activities. *Journal of Physiology and Biochemistry, 71*, 329–340.

Marshall, W. J., Bangert, S. K., & Lapsley, M. (2012). *Clinical chemistry* (7th ed.). Elsevier.

Marshall, W., Lapsley, M., & Day, A. (2016). *Clinical chemistry* (8th ed.). Elsevier.

Morral, N., Edenberg, H. J., Witting, S. R., Altomonte, J., Chu, T., & Brown, M. (2007). Effects of glucose metabolism on the regulation of genes of fatty acid synthesis and triglyceride secretion in the liver. *Journal of Lipid Research, 48*, 1499–1510.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics, 81*, 559–575.

Rosenson, R. S., Brewer, H. B., Davidson, W. S., Fayad, Z. A., Fuster, V., Goldstein, J., Hellerstein, M., Jiang, X.-C., Phillips, M. C., Rader, D. J., Remaley, A. T., Rothblat, G. H., Tall, A. R., & Yvan-Charvet, L. (2012). Cholesterol efflux and atheroprotection. *Advancing the Concept of Reverse Cholesterol Transport, 125*, 1905–1919.

Simko, T.W.a.V. (2016) R package 'corrplot': Visualization of a correlation matrix.

Soininen, P., Kangas, A. J., Wurtz, P., Suna, T., & Ala-Korpela, M. (2015). Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circulation. Cardiovascular Genetics, 8*, 192–206.

Soininen, P., Kangas, A. J., Wurtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Jarvelin, M.-R., Kahonen, M., Lehtimaki, T., Viikari, J., Raitakari, O. T., Savolainen, M. J., & Ala-Korpela, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *The Analyst, 134*, 1781–1785.

Steuer, R. (2006). Review: On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics, 7*, 151–158.

Taylor, A. E., Davies, N. M., Ware, J. J., VanderWeele, T., Smith, G. D., & Munafò, M. R. (2014). Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates(). *Economics and Human Biology, 13*, 99–106.

Team, R.D.C. (2008) R: A language and environment for statistical computing, R Foundation for Statistical Computing.

Wickham, H. (2016) *ggplot2 Elegant Graphics for Data Analysis*. Springer International Publishing.

Würtz, P., Kangas, A. J., Soininen, P., Lawlor, D. A., Davey Smith, G., & Ala-Korpela, M. (2017). Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: A primer on -omic technologies. *American Journal of Epidemiology, 186*, 1084–1096.

Würtz, P., Wang, Q., Niironen, M., Tynkkynen, T., Tiainen, M., Drenos, F., Kangas, A. J., Soininen, P., Skilton, M. R., Heikkilä, K., Pouta, A., Kähönen, M., Lehtimäki, T., Rose, R. J., Kajantie, E., Perola, M., Kaprio, J., Eriksson, J. G., Raitakari, O. T., … Auro, K. (2016a). Metabolic signatures of birthweight in 18 288 adolescents and adults. *International Journal of Epidemiology, 45*, 1539–1550.

Würtz, P., Wang, Q., Soininen, P., Kangas, A. J., Fatemifar, G., Tynkkynen, T., Tiainen, M., Perola, M., Tillin, T., Hughes, A. D., Mäntyselkä, P., Kähönen, M., Lehtimäki, T., Sattar, N., Hingorani, A. D., Casas, J.-P., Salomaa, V., Kivimäki, M., Järvelin, M.-R., … Ala-Korpela, M. (2016b). Metabolomic profiling of statin use and genetic inhibition of HMG-CoA reductase. *Journal of the American College of Cardiology, 67*, 1200–1210.

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., Hemani, G., Tansey, K., Laurin, C., St Pourcain, B., Warrington, N. M., Finucane, H. K., Price, A. L., Bulik-Sullivan, B. K., Anttila, V., Paternoster, L., Gaunt, T. R., Evans, D. M., Neale, B. M., & Early Genetics, L. (2017). LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics, 33*, 272–279.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.