Constructing an atlas of associations between polygenic scores from across the human phenome and circulating metabolic biomarkers

Si Fang*, Michael V Holmes, Tom R Gaunt, George Davey Smith, Tom G Richardson*

MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom

Abstract

Background: Polygenic scores (PGS) are becoming an increasingly popular approach to predict complex disease risk, although they also hold the potential to develop insight into the molecular profiles of patients with an elevated genetic predisposition to disease.
Methods: We sought to construct an atlas of associations between 125 different PGS derived using results from genome-wide association studies and 249 circulating metabolites in up to 83,004 participants from the UK Biobank.
Results: As an exemplar to demonstrate the value of this atlas, we conducted a hypothesis-free evaluation of all associations with glycoprotein acetyls (GlycA), an inflammatory biomarker. Using bidirectional Mendelian randomization, we find that the associations highlighted likely reflect the effect of risk factors, such as adiposity or liability towards smoking, on systemic inflammation as opposed to the converse direction. Moreover, we repeated all analyses in our atlas within age strata to investigate potential sources of collider bias, such as medication usage. This was exemplified by comparing associations between lipoprotein lipid profiles and the coronary artery disease PGS in the youngest and oldest age strata, which had differing proportions of individuals undergoing statin therapy. Lastly, we generated all PGS–metabolite associations stratified by sex and separately after excluding 13 established lipid-associated loci to further evaluate the robustness of findings.
Conclusions: We envisage that the atlas of results constructed in our study will motivate future hypothesis generation and help prioritize and deprioritize circulating metabolic traits for in-depth investigations. All results can be visualized and downloaded at http://mrcieu.mrsoftware.org/metabolites_PRS_atlas.
Funding: This work is supported by funding from the Wellcome Trust, the British Heart Foundation, and the Medical Research Council Integrative Epidemiology Unit.

Editor's evaluation

The authors describe their work on an atlas of associations between polygenic scores for 125 different traits representing a variety of quantitative phenotypes and diseases, and a large set of metabolites measured in up to 83,000 participants in the UK Biobank. These associations are all available via a public browser, and may be used to identify candidate intermediate phenotypes, as well as potential biomarkers of disease.

***For correspondence:** si.fang@bristol.ac.uk (SF); Tom.G.Richardson@bristol.ac.uk (TGR)

Competing interest: See page 12

Funding: See page 12

Preprinted: 25 October 2021
Received: 16 September 2021
Accepted: 12 September 2022
Published: 11 October 2022

Reviewing Editor: Jonathan K Pritchard, Stanford University, United States

Introduction

Complex traits and disease have a polygenic architecture meaning that they are influenced by many genetic variants scattered throughout the human genome (***Boyle et al., 2017***). An increasingly popular approach to predict disease risk in a population is to derive weighted scores by summing the number of risk increasing variants that participants harbour. These are typically referred to as 'polygenic scores' (PGS) (***Torkamani et al., 2018***; ***Lewis and Vassos, 2020***). In the last decade, PGS have emerged as powerful tools for predicting lifelong risk of disease, which is predominantly due to the dramatic increase in sample sizes of genome-wide association studies (GWAS) and their continued success in uncovering trait-associated genetic variants across the genome (***Visscher et al., 2017***). Additionally, PGS have utility in a causal inference setting to establish causal effects between risk factors and disease outcomes, as well as to help elucidate putative diagnostic and prognostic biomarkers for disease incidence (***Richardson et al., 2019b***, ***Holmes and Davey Smith, 2019***; ***Ritchie et al., 2021a***).

The human metabolome consists of over 100,000 small molecules and is a rich source of potential risk factors and biomarkers, as well as therapeutic targets (***Holmes et al., 2021***), for complex traits and disease (***Gallois et al., 2019***). Many circulating metabolic traits studied to date have a large heritable component as demonstrated by GWAS endeavours (***Suhre et al., 2011***; ***Shin et al., 2014***; ***MacTel Consortium et al., 2021***), suggesting that they have a polygenic architecture. In-depth molecular profiling has recently been undertaken in the UK Biobank (UKB) study using nuclear magnetic resonance (NMR) to capture measures of 249 circulating metabolites in approximately 120,000 participants who also have genotype data (***Julkunen et al., 2021***; ***Sudlow et al., 2015***). The 249 metabolite measurements include the particle concentration, size, and composition of 14 lipoprotein subclasses, as well as the levels of phospholipids, fatty acids, amino acids, ketone bodies, and other biomarkers as discussed in a recent review (***Ala-Korpela et al., 2022***). This resource therefore provides an unprecedented opportunity to characterize metabolic profiles for disease risk by leveraging genome-wide variation captured by PGS. There are multiple advantages to this approach over conventional observational associations between metabolites and complex traits or endpoints. For example, as UKB is a prospective cohort study many diseases have low prevalence, such as Alzheimer's disease which typically has a late onset. In contrast, evaluations using PGS will likely yield higher statistical power given that a continuous genetic score will be analysed for all participants in UKB based on their liability to disease.

In this study, we sought to construct an atlas of associations between 125 PGS and the 249 circulating metabolic traits in the UKB study (***Figure 1***). We demonstrate the usefulness of this atlas in terms of highlighting putative risk factors and biomarkers for disease risk and advocate the use of an approach known as Mendelian randomization (MR) to investigate whether a causal relationship may underlie findings (Supplementary Note 2) (***Davey Smith and Ebrahim, 2003***, ***Davey Smith and Hemani, 2014***; ***Sanderson et al., 2022***). As an exemplar, we apply MR systematically to investigate all PGS associations highlighted from a hypothesis-free scan of the inflammatory marker glycoprotein acetyls (GlycA). Furthermore, all PGS analyses were initially conducted in the full UKB sample, as well as in sex-stratified samples and age tertiles as proposed previously to evaluate the influence of medication use on findings (***Bell et al., 2022***). As the age of individuals in UKB is unlikely to induce sources of biases into analyses (e.g. collider bias), age-dependent stratification allows comparisons between the youngest and oldest tertiles in UKB where the level of medication use is likely to vary between groups. Stratification on medication use, by contrast, would introduce collider bias. Together our findings provide valuable insights into the effects of PGS on metabolic markers which may influence hypothesis generation and facilitate similar analyses to those presented in this paper.

Results

Constructing an atlas of polygenic score associations across the human metabolome

We obtained genome-wide summary statistics for 125 different complex traits and diseases from large-scale GWAS and constructed PGS for each of these in the UKB study. The majority of these summary statistics were obtained from the OpenGWAS platform and encompassed traits and disease outcomes from across the human phenome (***Elsworth et al., 2020***). GWAS were identified based on those conducted in populations of European descent given that our analysis in UKB was based on the

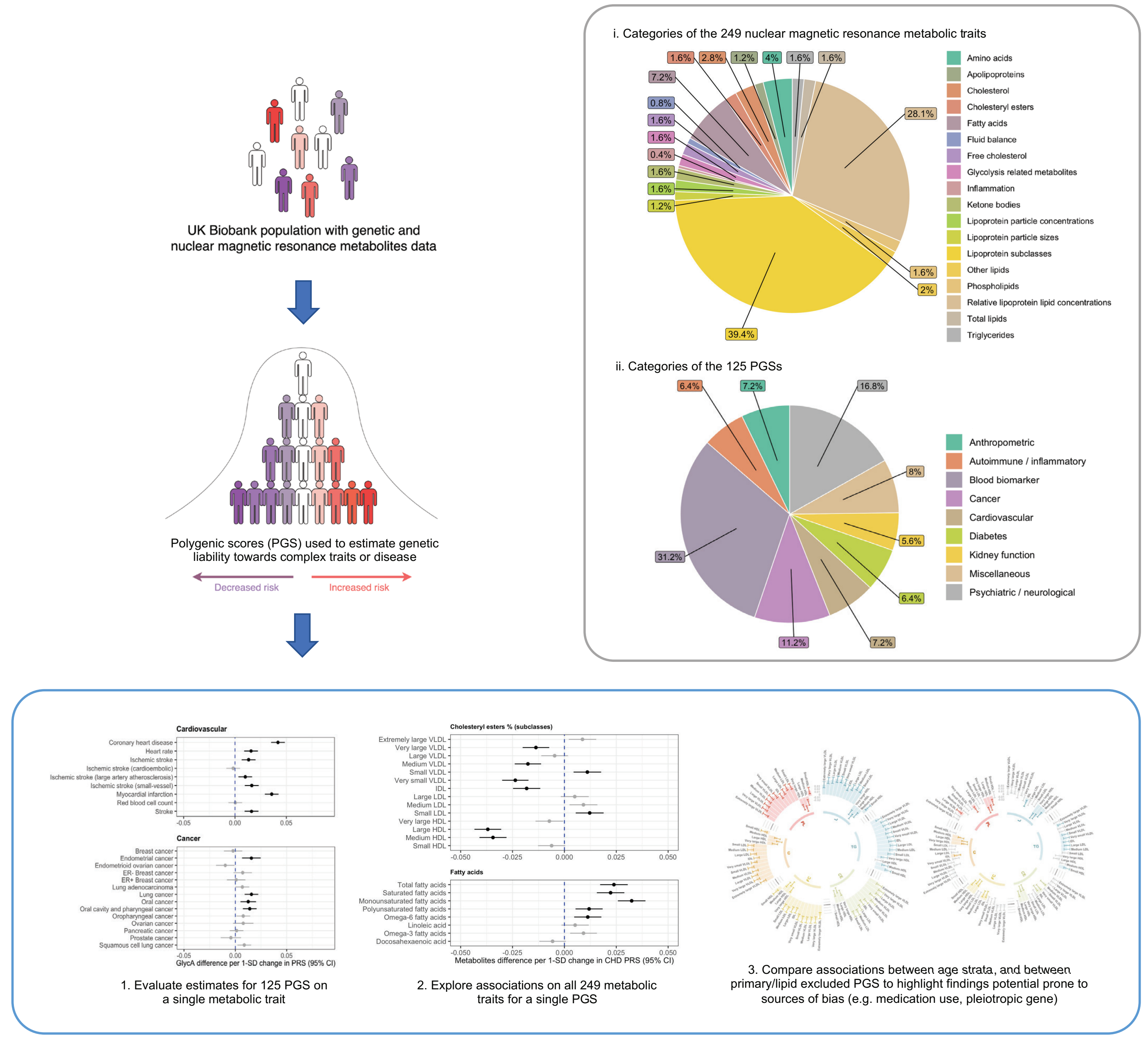
UK Biobank population with genetic and nuclear magnetic resonance metabolites data
Polygenic scores (PGS) used to estimate genetic liability towards complex traits or disease
Decreased risk
Increased risk
i. Categories of the 249 nuclear magnetic resonance metabolic traits
1.6%
2.8%
1.2%
4%
1.6%
1.6%
7.2%
0.8%
1.6%
1.6%
0.4%
1.6%
1.6%
1.2%
28.1%
1.6%
2%
39.4%
Amino acids
Apolipoproteins
Cholesterol
Cholesteryl esters
Fatty acids
Fluid balance
Free cholesterol
Glycolysis related metabolites
Inflammation
Ketone bodies
Lipoprotein particle concentrations
Lipoprotein particle sizes
Lipoprotein subclasses
Other lipids
Phospholipids
Relative lipoprotein lipid concentrations
Total lipids
Triglycerides
ii. Categories of the 125 PGSs
6.4%
7.2%
16.8%
8%
5.6%
6.4%
7.2%
11.2%
31.2%
Anthropometric
Autoimmune / inflammatory
Blood biomarker
Cancer
Cardiovascular
Diabetes
Kidney function
Miscellaneous
Psychiatric / neurological
Cardiovascular
Coronary heart disease
Heart rate
Ischemic stroke
Ischemic stroke (cardioembolic)
Ischemic stroke (large artery atherosclerosis)
Ischemic stroke (small-vessel)
Myocardial infarction
Red blood cell count
Stroke
-0.05
0.00
0.05
Cancer
Breast cancer
Endometrial cancer
Endometrioid ovarian cancer
ER- Breast cancer
ER+ Breast cancer
Lung adenocarcinoma
Lung cancer
Oral cancer
Oral cavity and pharyngeal cancer
Oropharyngeal cancer
Ovarian cancer
Pancreatic cancer
Prostate cancer
Squamous cell lung cancer
-0.05
0.00
0.05
GlycA difference per 1-SD change in PRS (95% CI)
Cholesteryl esters % (subclasses)
Extremely large VLDL
Very large VLDL
Large VLDL
Medium VLDL
Small VLDL
Very small VLDL
IDL
Large LDL
Medium LDL
Small LDL
Very large HDL
Large HDL
Medium HDL
Small HDL
-0.050
-0.025
0.000
0.025
0.050
Fatty acids
Total fatty acids
Saturated fatty acids
Monounsaturated fatty acids
Polyunsaturated fatty acids
Omega-6 fatty acids
Linoleic acid
Omega-3 fatty acids
Docosahexaenoic acid
-0.050
-0.025
0.000
0.025
0.050
Metabolites difference per 1-SD change in CHD PRS (95% CI)
1. Evaluate estimates for 125 PGS on a single metabolic trait
2. Explore associations on all 249 metabolic traits for a single PGS
3. Compare associations between age strata, and between primary/lipid excluded PGS to highlight findings potential prone to sources of bias (e.g. medication use, pleiotropic gene)

Figure 1. A schematic diagram depicting the data composition and analytical approach undertaken in this study.

European subset with NMR data. Furthermore, we identified studies which did not include the UKB in their study to avoid overlapping samples between PGS construction and analysis with metabolic traits. Full details for all GWAS can be found in ***Supplementary file 1a***. Two versions of each PGS were built using different thresholds for variant-trait associations (P) and linkage disequilibrium (LD; r^2). These were (1) 'lenient' thresholds of $p < 0.05$ and $r^2 < 0.1$ and (2) a 'stringent' threshold of $p < 5 \times 10^{-8}$ and $r^2 < 0.001$. PGS were generated using the software PLINK (***Chang et al., 2015***) with LD being calculated using a reference panel of 10,000 randomly selected unrelated UKB individuals of European descent (***Kibinge et al., 2020***). The specific weights for clumped variants used in all PGS can be found at https://tinyurl.com/PRSweights.

We investigated the association between each PGS in turn with 249 circulating metabolites measured using targeted high-throughput NMR metabolomics from Nightingale Health Ltd (biomarker quantification version 2020) (***Supplementary file 1b***; ***Julkunen et al., 2021***). Our final sample size of *n* = 83,004 was determined based on individuals with both genotype and circulating metabolites data after removing participants with withdrawn consent, evidence of genetic relatedness or who were not of 'white European ancestry' based on a *K*-means clustering (*K* = 4). All PGS were standardized to have a mean of 0 and standard deviation of 1 and similarly all metabolites were subject to inverse rank normalization transformations prior to analysis allowing cross-PGS/metabolite comparisons to be made. Analyses were conducted using linear regression adjusting for age, sex, and the top 10 principal components (PCs).

To disseminate all findings from this large-scale analysis we have developed a web application (http://mrcieu.mrsoftware.org/metabolites_PRS_atlas/) to query and visualize metabolic signatures for a given PGS. In this paper, we have discussed findings using PGS that were derived using the more lenient criteria (i.e. $p < 0.05$ and $r^2 < 0.1$), although all findings based on both thresholds can be found in the web atlas. PC analysis suggested that the first 19 PCs captured 95% of the variance in the NMR metabolites data (whereas the first ten PCs captured 90% and the first 41 PCs captured 99% of the variance) (***Supplementary file 1c***). We therefore have applied a heuristic of $p < 0.05/19$ in this manuscript to account for multiple testing of the associations between any single PGS and the NMR metabolic traits for downstream analyses, although users are able to download the full results to apply whatever correction they see fit. For all other analyses (e.g. associations between metabolic traits and all PGS), we apply a false discovery rate (FDR) of less than 0.05 calculated from the Benjamini–Hochberg procedure to correct for multiple testing. Based on the FDR threshold of 0.05, there were a total of 5445 associations between PGS derived (derived based on 'lenient criteria' with $p < 0.05$ variants) in the full sample and NMR-assessed circulating metabolic traits. Heatmaps depicting the *Z* scores of all PGS–metabolic trait associations can be found in ***Figure 2—figure supplements 1 and 2***. The PGS with the largest number of associations robust to multiple testing corrections was body mass index (BMI) (*n* = 217) (***Supplementary file 1d***). Our atlas also includes sex-stratified estimates for PGS weighted by GWAS undertaken in female only (such as breast cancer and age at menarche) and male only (e.g. prostate cancer) populations, as well as sex-stratified estimates in both females and males separately for all other PGS–metabolite associations. We encourage users interested in these sex-stratified estimates to interpret them with caution however, given the widespread sex-differential participation bias in UKB (***Pirastu et al., 2021***).

In this paper, we provide several examples of how results from this atlas can be used to generate hypotheses and pave the way for in-depth and careful evaluations of associations between PGS and circulating traits. Specifically, we believe our findings can facilitate a 'reverse gear Mendelian randomization' approach to disentangle whether associations likely reflect metabolic traits acting as a cause or consequence of disease risk (***Holmes and Davey Smith, 2019***) as illustrated using triglyceride-rich very-low-density lipoprotein (VLDL) particles in the next section. Furthermore, in-depth evaluations allow careful consideration of appropriate instrumental variables for circulating metabolites which can be a challenging task as highlighted in our exemplar analysis of GlycA. Finally, we provide examples of how the plethora of sensitivity analyses within our atlas can help users further investigate the robustness of findings.

Orienting the direction of effect between putative causal relationships using Mendelian randomization

Many top associations across PGS were consistent with the known underlying biology of their corresponding diseases, as well as various proof of concepts that associations between PGS and metabolic traits may reflect both causes of disease and consequences of genetic liability towards disease. For example, we applied MR to further evaluate associations highlighted in our atlas with VLDL particles, where both VLDL particle average diameter size and concentration were associated with the PGS for BMI (Beta = 0.04, 95% CI = 0.033 to 0.046, $p = 3.53 \times 10^{-35}$ and Beta = 0.012, 95% CI = 0.006 to 0.019, $p = 2.7 \times 10^{-4}$, respectively) and also coronary heart disease (CHD) liability (Beta = 0.026, 95% CI = 0.019 to 0.032, $p = 2.12 \times 10^{-15}$ and Beta = 0.035, 95% CI = 0.028 to 0.042, $p = 2.73 \times 10^{-24}$, respectively). Conducting bidirectional MR suggested that the associations with average diameter of VLDL particles are likely attributed to a consequence of BMI and CHD liability as opposed to the size of VLDL particles having a causal influence on these outcomes (***Supplementary file 1e***). In contrast, MR analyses suggested that the concentration of VLDL particles increases risk of CHD (Beta = 1.28 per

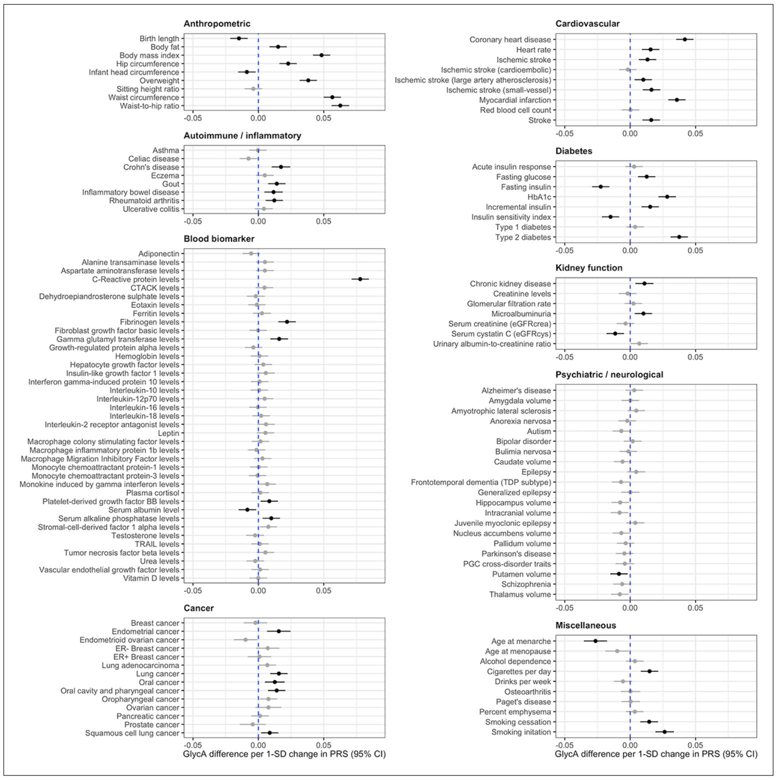

Figure 2. Forest plots depicting results from a systematic evaluation of 125 polygenic scores and their associations with circulating glycoprotein acetyls (GlycA). Associations were assessed by linear regression on up to 83,004 individuals in the UK Biobank. Error bars represent the 95% confidence intervals for the effect estimates. Results coloured in grey are associations which did not surpass a false discovery date of less than 0.05 to account for multiple testing.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Heatmap showing the *Z* scores of associations between metabolites and PGS that were derived using the lenient criteria (i.e. variant-trait associations with $p < 0.05$ and linkage disequilibrium (LD) $r^2 < 0.1$).

Figure supplement 2. Heatmap showing the *Z* scores of associations between metabolites and PGS that were derived using the more stringent criteria (i.e. variant-trait associations with $p < 5 \times 10^{-8}$ and linkage disequilibrium (LD) $r^2 < 0.001$).

1-SD change in VLDL particle concentration, 95% CI = 1.25 to 1.65, $p = 2.8 \times 10^{-7}$) which may explain associations between the CHD PGS and this metabolic trait within our atlas. Similar MR analyses to investigate findings from our atlas can be conducted using the full GWAS summary statistics for all 249 circulating metabolic traits available via the GWAS catalog (https://www.ebi.ac.uk/gwas/) under accession IDs GCST90092803 to GCST90093051 (***Richardson et al., 2022***).

Along with comparing metabolic signatures for a given PGS, our atlas facilitates hypothesis-free evaluations to inspect all PGS associations for a given metabolic trait. As an example of this, we have undertaken such an analysis based on the associations between all 125 PGS in our atlas with circulating GlycA. GlycA is a biomarker of chronic inflammation and has been found to predict various endpoints, including types of cardiovascular disease, cancer, and all-cause mortality (***Lawler et al., 2016***; ***Connelly et al., 2017***). Although previous studies of genetically predicted GlycA have been conducted for hypotheses regarding single endpoints (***Lord et al., 2021***), whether or not circulating GlycA has a causal effect on outcomes from across the disease spectrum has yet to be comprehensively investigated. The role of GlycA is important to establish given the emerging role of inflammation as a pharmacologically modifiable pathway for the prevention and treatment of cardiovascular disease.

There were 44 PGS associations with GlycA which were robust to an FDR <5%, used as a heuristic to determine which results to investigate in further detail (***Figure 2*** and ***Supplementary file 1g***). We firstly applied the inverse variance weighted (IVW) MR method to systematically assess whether genetic liability to any of these disease endpoints or complex traits provided evidence of an effect on GlycA levels. Of the 44 PGS, 36 contain one or more genetic variants that reached genome-wide significance which can be used as instrumental variables for MR. In total, eight of these exposures provided evidence of a genetically predicted effect from MR analyses based on FDR <5% (***Supplementary file 1g***), which included anthropometric traits such as BMI (Beta = 0.16 SD increase in GlycA levels per 1 SD increase in BMI, 95% CI = 0.11 to 0.21, FDR = 1.59 × 10^{-8}) and genetic liability to cigarettes smoked per day (Beta = 0.27 SD change GlycA levels per 1 SD increase in cigarettes per day, 95% CI = 0.20 to 0.34, FDR = 2.84 × 10^{-12}). Estimates based on the IVW method were typically supported by the weighted median approach, although only cigarettes smoked per day were supported by both the weighed median (Beta = 0.24, 95% CI = −0.16 to 0.33, p = 2.08 × 10^{-8}) and MR-Egger (Beta = 0.22, 95% CI = 0.07 to 0.37, p = 0.02) methods (***Supplementary file 1h***).

Next, we investigated the converse direction of effect using MR to assess whether genetically predicted GlycA may influence any of the 44 complex traits or disease endpoints highlighted by our atlas of results. Undertaking a GWAS of GlycA in the UKB identified 59 independent genetic variants which were harnessed as instrumental variables (mean F = 100.1) (***Supplementary file 1i***). In contrast to the previous analysis, we identified very weak evidence using the IVW method that genetically predicted GlycA has an effect on any of the 44 traits or diseases assessed based on FDR <5% (***Supplementary file 1j***). We also conducted further sensitivity analyses given that the NMR signal of GlycA is a composite signal contributed by the glycan *N*-acetylglucosamine residues on five acute-phase proteins, including alpha1-acid glycoprotein, haptoglobin, alpha1-antitrypsin, alpha1-antichymotrypsin, and transferrin (***Otvos et al., 2015***). Using cis-acting plasma protein (where possible) and expression quantitative trait loci (pQTLs and eQTLs respectively) (*F*-stats range from 43.9 to 4468.0) as instrumental variables for these proteins (***Supplementary file 1k***) did not provide convincing evidence that they play a role in disease risk for associations between PGS and GlycA (***Supplementary file 1l***). The only effect estimate robust to multiple testing was found for higher genetically predicted alpha1-antitrypsin levels on gamma glutamyl transferase (GGT) levels (Beta = 0.05 SD change in GGT per 1 SD increase in protein levels, 95% CI = 0.03 to 0.07, FDR = 3.6 × 10^{-3}), although this was not replicated when using estimates of genetic associations with GGT levels from a larger GWAS conducted in the UKB data (Beta = 1.6 × 10^{-3}, 95% CI = −6.9 × 10^{-3} to 0.01, p = 0.71). For details of pleiotropy robust analysis and replication results see ***Supplementary file 1m***.

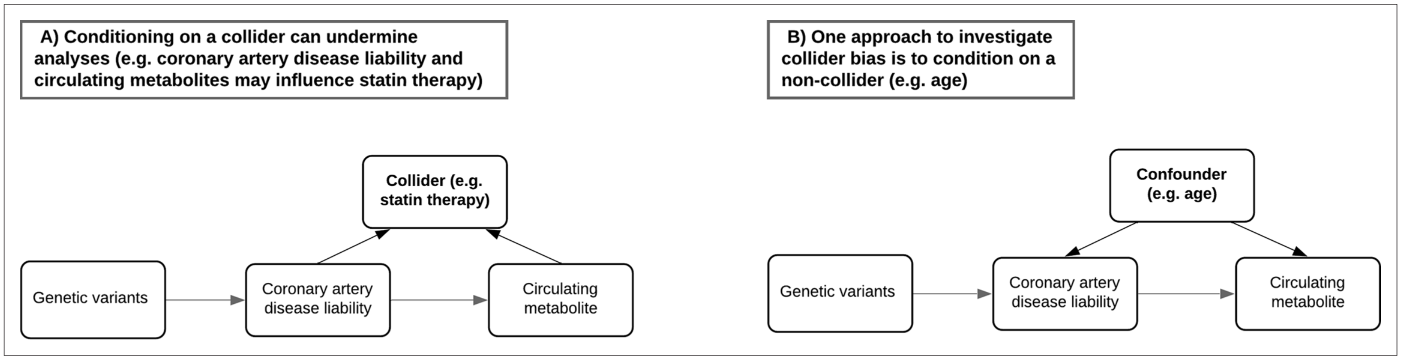
A) Conditioning on a collider can undermine analyses (e.g. coronary artery disease liability and circulating metabolites may influence statin therapy)
Collider (e.g. statin therapy)
Genetic variants
Coronary artery disease liability
Circulating metabolite
B) One approach to investigate collider bias is to condition on a non-collider (e.g. age)
Confounder (e.g. age)
Genetic variants
Coronary artery disease liability
Circulating metabolite

Figure 3. Directed acyclic graphs illustrating the potential collider bias involved in the causal relationship between the coronary artery disease polygenic score and circulating metabolites. (**A**) The likelihood of participants in UK Biobank taking medication such as statins is influenced by having a higher genetic predisposition to coronary artery disease but may also be influenced by certain metabolic traits measured on the nuclear magnetic resonance (NMR) panel (e.g. having elevated low-density lipoprotein cholesterol levels). Either stratifying or adjusting for statin use in regression models may therefore induce collider bias into the association between disease liability and metabolic traits. (**B**) Age is commonly adjusted for in association analyses due to its role as a confounder and cannot be a collider (i.e. exposures and outcomes cannot influence the age of participants). Stratifying samples by age therefore enables the analysis of exposure–outcome associations in a group of participants with relatively consistent confounding effect from age, leading to more robust association estimates in the lower age tertile where the percentage of participants who are regularly taking medication is low. Furthermore, comparisons with participants in the highest age tertile can help highlight associations between polygenic scores and metabolic traits most likely distorted by potential colliders such as statins in the full sample.

Stratifying analyses by age to investigate potential sources of bias induced by medication use

A critical challenge when analysing the NMR metabolites data in UKB concerns the most appropriate manner to account for participants taking medications which may undermine inference (**Bell et al., 2022**). For example, UKB participants undergoing statin therapy will likely have altered levels of lipoprotein lipid metabolites compared to others. However, adjusting for statin therapy as a covariate or by stratification can induce collider bias, which may be encountered when investigating the relationship between two factors (such as genetic liability towards CHD and a lipoprotein lipid metabolite) when both influence a third factor (e.g. statin therapy) (***Figure 3A***). In particular due to the large sample sizes provided, collider bias in the UKB study has been shown to distort findings (**Griffith et al., 2020**) and in extreme cases can even result in opposite conclusions being drawn (**Richardson et al., 2019a**). Therefore, to investigate the influence of medication use on the results within our atlas, we repeated all analyses stratified by age tertiles as proposed previously (**Bell et al., 2022**), given that age is very unlikely to act as a collider between PGS and circulating metabolites (***Figure 3B***), and medication use is lower in the younger tertiles. Comparisons between the youngest and oldest tertiles in UKB can be systematically investigated and visualized using our web application to evaluate how medications may bias findings.

As an example of this, in the full UKB sample there were 193 circulating metabolites associated with the PGS for CHD (constructed using genetic variants with $p < 0.05$ and $r^2 < 0.1$) under a p value <0.05/19 for multiple testing correction (***Supplementary file 1n***). The vast majority of these were lipoprotein lipid traits, which are likely capturing causal risk factors for CHD. Amongst the top associations for this PGS was apolipoprotein B (apo B) (Beta = 0.027, 95% CI = 0.020 to 0.033, $p = 7.2 \times 10^{-15}$), which acts as an index of the number of circulating atherogenic lipoprotein particles and has been postulated previously to be the predominating lipoprotein lipid trait indexing CHD risk (**Ference et al., 2019**; **Sniderman et al., 2019**; **Richardson et al., 2020b**, **Richardson et al., 2021**).

Evaluating this association between age tertiles allowed us to investigate whether it may be influenced by medications in UKB, such as the impact of statin therapy on lowering low-density lipoprotein (LDL) cholesterol, which apo B particles carry. In the youngest tertile (mean age = 47.3 years, 5%

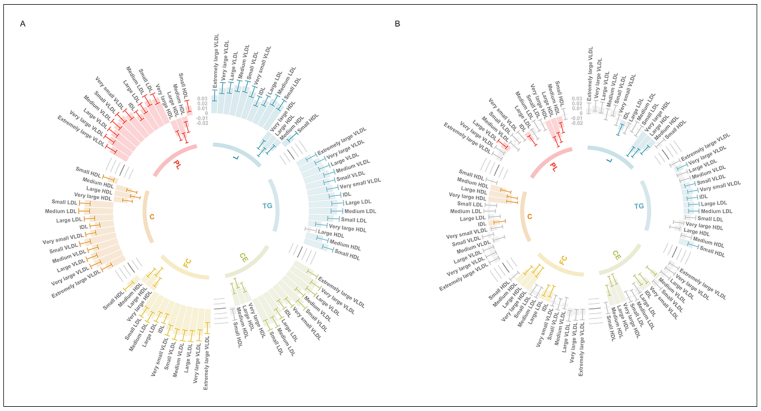

Figure 4. Circos plots illustrating the utility of age-stratified analyses in UK Biobank to investigate potential sources of bias when evaluating associations between polygenic scores (PGS) and circulating metabolites. Associations were assessed by linear regression on up to 83,004 individuals in the UK Biobank. Error bars represent confidence intervals of the effect estimates between the coronary heart disease (CHD) PGS and traits from the six subclasses: L = total lipids, TG = triglycerides, CE = cholesteryl esters, FC = free cholesterol, C = cholesterol, PL = phospholipids. Grey bars represent associations not robust to multiple testing based on p > 0.05/19. These barcharts are oriented such that those extending to the outer rim reflect a positive association between the CHD PGS and metabolic traits whereas those extending inwards indicate inverse associations. (**A**) Analyses undertaken for participants in the lowest age tertile (mean age = 47.3 years, 5% statin users) and (**B**) the corresponding results for the oldest age tertile (mean age = 65.3 years, 29% statin users).

The online version of this article includes the following figure supplement(s) for figure 4:

Figure supplement 1. A forest plot illustrating the associations between the polygenic score (PGS) for coronary heart disease (CHD) with 249 circulating metabolic traits in UK Biobank.

statin users), the association between the CHD PGS with apo B was markedly stronger than in the total sample (Beta = 0.059, 95% CI = 0.048 to 0.070, $p < 1.6 \times 10^{-26}$). In stark contrast, there was weak evidence of an inverse association between apo B and the CHD PGS in the oldest tertile (mean age = 65.3 years, 29% statin users) (Beta = −0.007, 95% CI = −0.019 to 0.004, p = 0.223), which is likely attributed to the higher proportion of participants undergoing statin therapy in this sample. Similarly, concentrations of VLDL, LDL, and IDL provided evidence of a positive association with the CHD PGS in the youngest tertile (***Figure 4A***), whereas the corresponding associations in the oldest tertile provided weak evidence of association (and in some cases reversed direction entirely) (***Figure 4***). A comparison of all 249 associations with the CHD PGS derived in the youngest and oldest age tertiles can be found in ***Supplementary file 1o***.

Elucidating polygenic associations with metabolic traits by excluding major regulators of NMR lipoprotein lipids

The polygenic nature of complex traits means that the inclusion of highly weighted pleiotropic genetic variants in PGS may introduce bias into genetic associations within our atlas. To provide insight into this issue, we constructed PGS excluding variants within the regions of the genome which encode the genes for 13 major regulators of NMR lipoprotein lipids signals which captured 75% of the gene–metabolite associations in the Finnish Metabolic Syndrome In Men (METSIM) cohort (***Gallois et al., 2019***). For details of these genes see ***Supplementary file 1p***.

For PGS with these lipid loci excluded, anthropometric traits such as waist-to-hip ratio (N = 209), waist circumference (N = 206), and BMI (N = 205) still provided strong evidence of association with the majority of metabolic measurements on the NMR panel based on multiple testing corrections. Elsewhere however, the Alzheimer's disease PGS, which was associated with 60 metabolic traits robust to $p < 0.05/19$ in the initial analysis including these lipid loci (***Supplementary file 1q***), provided no convincing evidence of association with the 249 circulating metabolites after excluding the lipid loci based on the same multiple testing threshold (***Supplementary file 1r***). Further inspection suggested that the likely explanation for this attenuation of evidence were due to variants located within the *APOE* locus (near one of the major lipid regulators *APOC1*) which are recognized to exert their influence on phenotypic traits via horizontally pleiotropic pathways (***Ferguson et al., 2020***).

Discussion

In this study, we have developed an atlas of polygenic risk score associations with circulating metabolic traits in an unprecedented sample size compared to previous studies. Our results can be used to help prioritize findings worthy of follow-up, using techniques such as MR, as a means of disentangling putative causal and non-causal relationships underlying associations between PGS and circulating biomarkers. Furthermore, conducting all analyses within age tertiles illustrates the potential of medication use within the UKB population to bias relationships within our atlas of results. These results should help highlight disease–metabolic trait relationships where researchers should exercise caution when interpreting findings from their own analyses of the recently generated NMR metabolites data in UKB, which are due to be available in all ~500,000 participants in the forthcoming years.

Amongst the thousands of PGS associations identified in this study with a p value <0.05/19, we observed an enrichment of scores derived using GWAS of anthropometric traits. This was exemplified by the waist circumference PGS which yielded the largest number of associations in our atlas (n = 212). Previous studies in the field have demonstrated the strong influence that adiposity has on circulating traits from across the metabolome (***Würtz et al., 2014***), and indeed across the proteome (***Folkersen et al., 2020***). Furthermore, as shown previously by an MR study (***Bell et al., 2022***), certain associations with the BMI PGS may be due to the influence of medication use in the UKB sample, for example those related to LDL (e.g. BMI PGS on total lipids in LDL: Beta = −0.020, 95% CI = 0.027 to −0.013, $p < 6.17 \times 10^{-9}$). In our atlas, evidence of these associations strongly attenuated in the youngest age tertile, where the influence of such factors in the UKB population may be weakest (e.g. total lipids in LDL from youngest age tertile: Beta = 0.005, 95% CI = −0.006 to 0.016, p = 0.35). In addition to the striking difference highlighted in our study between the CHD PGS and apolipoprotein B across age tertiles, findings such as this further emphasize the importance of evaluating results from the full sample analysis together with those derived in age-stratified subsamples. Moreover, we suggest that users interested in the sex-stratified estimates within our atlas should interpret them in conjunction with estimates derived across age tertiles as in this example, given that the proportions of males and females in UKB taking certain medications may differ (e.g. statins).

As an exemplar, we conducted a hypothesis-free evaluation of one of the metabolic traits on the UKB NMR panel, GlycA, as a means of demonstrating how findings from our atlas may help generate hypotheses and follow-up analyses. Whilst our MR results indicated that modifiable risk factors such as BMI and cigarette smoking may increase levels of this circulating inflammatory biomarker, they suggest that targeting GlycA itself is unlikely to yield a beneficial therapeutic effect on the complex traits and disease endpoints evaluated in this study. This highlights the value of findings from our atlas, complemented by approaches such as MR, to help both prioritize and deprioritize circulating metabolic traits for further evaluation. Similar hypothesis-free evaluations on the other 248 metabolic traits can be routinely undertaken using our web tool, in addition to evaluations using the more stringent PGS construction criteria of $p < 5 \times 10^{-8}$ and $r^2 < 0.001$. We reiterate the importance of using approaches such as MR (including sensitivity analyses, which are at least partially robust to various forms of pleiotropy) to formally assess putative causal relationships which may underlie findings in our atlas however, as well as to help orient their directionality. This is particularly important given that PGS may be more prone to recapitulating sources of bias commonly encountered in observational studies in comparison to formal MR analyses (**Richardson et al., 2019b**, **Ritchie et al., 2021a**). We likewise conducted bidirectional MR to demonstrate that associations between the CHD PGS and VLDL particle size likely reflect an effect of CHD liability on this metabolic trait. In contrast, the association between the CHD PGS and VLDL concentrations are likely attributed to the causal influence of this metabolic trait on CHD risk, suggesting that it is the concentration of these triglyceride-rich particles that are important in terms of the aetiology of CHD risk as opposed to their actual size. We believe that findings from our atlas, as well as other ongoing efforts which leverage the large-scale NMR data within UKB, should facilitate further granular insight into lipoprotein lipid biology.

In terms of study limitations, we note that the NMR panel is predominantly focussed on lipoprotein lipids and as such our atlas does not facilitate analyses across the entire metabolome. Availability of metabolomics quantified by other platforms (e.g. mass spectroscopy) in large numbers with GWAS genotyping will aid in this effort (**MacTel Consortium et al., 2021**). Furthermore, whilst these data provide an unparalleled sample size compared to predecessors, findings are based on traits derived from whole blood and may therefore not be reflective of molecular signatures identified in other tissue types (**Richardson et al., 2020a**). In terms of interpretation, we emphasize that PGS can capture an estimate of an individual's lifelong disease risk, and as such results based on the UKB NMR metabolites dataset, measured at a midlife timepoint in the lifecourse in predominantly healthy participants, may differ substantially to metabolomic profiles of patients with a disease. Conversely, findings may hold the potential to highlight biomarkers useful for disease prediction before clinical manifestation, therefore indicating a potential window of opportunity for early detection and/or intervention. Lastly, in this study we leveraged data from the European subset of the UKB study, which may therefore not be representative of individuals from other ancestries (**Duncan et al., 2019**). Larger sample sizes of non-European individuals with metabolomics data will facilitate analyses in other ancestries once available, in addition to findings from future large-scale GWAS which have been principally confined to individuals of European descent to date (**Sirugo et al., 2019**).

We envisage that findings from our atlas will motivate future study hypotheses and help prioritize (and deprioritize) circulating metabolic traits for further in-depth research. Although we highlight several key findings in this manuscript, all our findings can be queried using our web application which provides a platform to inform researchers in the field planning similar analyses. Similar evaluations to those conducted in this manuscript should help develop a deeper understanding into how circulating metabolic traits contribute towards complex trait variation and assess their putative mediatory roles along the causal pathways between modifiable lifestyle risk factors and disease endpoints.

Methods

Data sources

The UKB study

Metabolic profiling was undertaken on a random subset of individuals from the UKB study (***Sudlow et al., 2015***) (range between 116,353 and 121,695). Full details on genotyping quality control (QC), phasing and imputation in UKB have been described previously (***Bycroft et al., 2018***). In brief, samples were restricted to individuals of white British ancestry who self-report as 'White British' and who have very similar ancestral backgrounds according to PC analyses performed by Bycroft et al ($n = 409,703$). In total, 107,162 pairs of related individuals were removed based on estimated kingship coefficients derived using the KING toolset. An in-house algorithm was then applied to preferentially remove individuals related to the greatest number of other individuals until no related pairs were left (removing $n = 79,448$ in total). A further 2 individuals were removed as they were related to over 200 other individuals. There were $n = 814$ individuals with sex-mismatch (derived by comparing genetic sex and reported sex) or individuals with sex chromosome aneuploidy excluded from analyses.

In total, 249 metabolic biomarkers were generated using non-fasting plasma samples (aliquot 3) taken from UKB participants at initial or subsequent clinical visits. Targeted high-throughput NMR metabolomics from Nightingale Health Ltd (biomarker quantification version 2020) were used to generate data on each of the 249 measures. These included biomarkers on lipoprotein lipid traits, their concentrations and subclasses, fatty acids, ketone bodies, glycolysis metabolites, and amino acids. Further details are described elsewhere (***Julkunen et al., 2021***). For QC, data of the 249 NMR metabolomics traits were processed using the 'ukbnmr' R package to remove variation due to technical factors caused by differences in sample handing and measurement (***Ritchie et al., 2021b***). Statin users in UKB were identified based on medication codes as defined previously (***Sinnott-Armstrong et al., 2021***). A full list of these metabolic biomarkers and their summary characteristics can be found in ***Supplementary file 1b***.

Ethical approval for this study was obtained from the Research Ethics Committee (REC; approval number: 11/NW/0382) and informed consent was collected from all participants enrolled in UKB. Data were accessed under UKB application #15825 and #81499.

GWAS summary statistics

Publicly available GWAS summary statistics were extracted from the OpenGWAS platform (https://gwas.mrcieu.ac.uk/) and publicly available repositories (***Elsworth et al., 2020***). We identified GWAS for 125 different complex traits and diseases which were selected to encompass a broad range of human phenotypes for which genome-wide data were available allowing us to construct PGS based on all variants available with $p < 0.05$. Furthermore, we identified GWAS based on study populations with participants of European descent, as our study was based on the unrelated European participants of UKB with NMR measures, as well as studies which had not analysed the UKB study population to avoid overlapping samples which can lead to overfitting bias in results (***Fang et al., 2022***). All details of these GWAS can be found in ***Supplementary file 1a***.

PGS construction

We built two versions of each PGS in this study using the following criteria. Firstly, scores were developed with independent variants (i.e. $r^2 < 0.001$) which were robustly associated with their traits or disease based on conventional genome-wide corrections (i.e. $p < 5 \times 10^{-8}$). The second versions of scores were derived using more lenient thresholds which were $r^2 < 0.1$ and $p < 0.05$. LD to estimate correlation between variants was based on a previously constructed reference panel of 10,000 randomly selected unrelated UKB individuals of European descent (***Kibinge et al., 2020***). PGS were derived for all participants with both genotype and NMR metabolites data after firstly excluding individuals with withdrawn consent, evidence of genetic relatedness or who were not of 'white European ancestry' based on a *K*-means clustering ($K = 4$). These scores were built by summing trait/disease risk increasing alleles which participants harboured weighted by their effect size reported by GWAS using genotype data from hard call dosages files (plink binaries bed/bim/fam) and the software PLINK v2.0 (***Chang et al., 2015***).

The majority of PGS were constructed in all eligible participants, with the exception of those based on GWAS in sex-stratified populations. These were breast cancer (including ER+ and ER− PGS), endometrial cancer, ovarian cancer, endometrioid ovarian cancer, age at menarche, age at menopause, and bulimia nervosa, which were derived in females only, as well as the prostate cancer PGS derived in males only. Additionally, we also built two versions of PGS for all complex traits excluding variants at 13 lipid-associated gene loci, which were *DOCK7*, *CELSR2*, *GALNT2*, *PCSK9*, *GCKR*, *TRIB1*, *LPL*, *APOA5*, *FADS2*, *LIPC*, *CETP*, *LDLR*, and *APOC1* (consisting of the encoding gene region itself as well as a 1 Mb window either side). Details of the 13 genes are presented in ***Supplementary file 1p***.

Statistical analysis

PC analysis of the metabolites

PC analysis was performed on the post-QC metabolite data to identify the number of independent traits among the 249 highly correlated metabolites. The analysis was conducted using the *princomp* function from the 'stats' R package.

PGS analysis

To allow us to draw comparisons between PGS–metabolite associations, we standardized all PGS to have a mean of 0 and standard deviation of 1 and additionally applied inverse rank normalization transformations to all metabolic traits prior to analysis. Associations between PGS and normalized metabolites were determined by linear regression with adjustment for age, sex (where appropriate), genotyping chip, the top 10 PCs, and fasting time. Each analysis was conducted initially in the full sample, followed by analyses after stratification into age tertiles to investigate the influence of medication use on findings. Sex-stratified association analyses in the full sample were also conducted whereby metabolic traits were transformed separately among males and females before applying linear regression. All analyses were undertaken using both versions of each PGS as long as their corresponding GWAS had at least one variant with $p < 5 \times 10^{-8}$ necessary for the more stringent criteria. To account for multiple testing in this study, we applied a p value threshold of 0.05/19 (accounting for 19 independent variables captured 95% variances in the metabolites from PC analysis) to highlight findings worthwhile evaluating in further detail. However, all results from our analyses are available in the web application should users decide to apply a more stringent (or lenient) heuristic to prioritize findings for in-depth analyses.

Instrument selection for GlycA analysis

Genetic instruments for all PGS traits/disease points evaluated in this study using MR were obtained from the 'TwoSampleMR' v0.5.6 R package (***Hemani et al., 2018***), or by manually uploading GWAS summary statistics and using the *clump_data* function from this package to identify them. Instruments for GlycA and particle size of very low-density lipoprotein (VLDL) were identified by conducting a GWAS of this trait in the UKB study using the BOLT-LMM (linear mixed model) software to control for population structure (***Loh et al., 2015***). Analyses were undertaken after excluding individuals of non-European descent (based on *K*-mean clustering of *K* = 4) and standard exclusions, including withdrawn consent, mismatch between genetic and reported sex, and putative sex chromosome aneuploidy. Analyses were adjusted for age, sex, fasting status, and a binary variable denoting the genotyping chip used in individuals (the UKBB Axiom array or the UK BiLEVE array). Genetic instruments were defined as variants with $p < 5 \times 10^{-8}$ after removing those in LD using the *clump_data* function as above. The full GWAS summary statistics for GlycA as well as the other 248 circulating metabolic traits are available via the GWAS catalog (https://www.ebi.ac.uk/gwas/) under accession IDs GCST90092803 to GCST90093051 (***Richardson et al., 2022***).

The NMR signal of serum GlycA is contributed by five acute-phase proteins (alpha1-acid glycoprotein, haptoglobin, alpha1-antitrypsin, alpha1-antichymotrypsin, and transferrin) (***Otvos et al., 2015***). Thus, another set of genetic instruments for GlycA were selected among variants strongly associated with these five proteins for further evaluation of the genetically predicted effect of GlycA. Instrumental variables for alpha1-acid glycoprotein were identified usinge eQTLs ($p < 5 \times 10^{-8}$, $r^2 < 0.001$) for *ORM1* from the eQTLGen Consortium (***Võsa et al., 2021***) due to a lack of available protein data. Genetic instruments for haptoglobin, alpha1-antitrypsin, alpha1-antichymotrypsin, and transferrin were identified using pQTLs ($p < 5 \times 10^{-8}$, $r^2 < 0.001$) for these proteins identified from 35,559 Icelanders

(***Ferkingstad et al., 2021***). To identify cis-acting instruments for these five proteins, we restricted e/pQTL associations to variants located within 1 Mb around their encoding genes: *ORM1* (encoding alpha1-acid glycoprotein; Ensembl ID: ENSG00000229314), *HP* (encoding haptoglobin; Ensembl ID: ENSG00000257017), *SERPINA1* (encoding alpha1-antitrypsin; Ensembl ID: ENSG00000197249), *SERPINA3* (encoding alpha1-antichymotrypsin; Ensembl ID: ENSG00000196136), or *TF* (encoding transferrin; Ensembl ID: ENSG00000196136). Independent instruments were identified using the same protocol as above (***Kibinge et al., 2020***).

MR analyses

MR analyses were undertaken using the 'TwoSampleMR' package (***Hemani et al., 2018***) to estimate the bidirectional effects between PGS traits/disease endpoints and metabolic traits, including GlycA and VLDL particle size. This was firstly estimated using the IVW method, which takes the SNP-outcome estimates and regresses them on those for the SNP-exposure associations (***Burgess et al., 2013***), followed by the weighted median and MR-Egger methods which are considered to be more robust to horizontal pleiotropy than the IVW approach (***Bowden et al., 2015***; ***Bowden et al., 2016***). If only one SNP is available as genetic instrument, Wald ratio estimates were calculated by dividing the SNP-outcome estimates by the SNP-exposure estimates (***Burgess et al., 2017***). *F*-Statistics were calculated to assess weak instrument bias (***Burgess and Thompson, 2013***). Benjamini–Hochberg FDR threshold of less than 5% was applied as a heuristic to account for multiple testing in the results.

Forest and circos plots in this study were generated using the R package 'ggplot' v3.3.3 (***Ginestet, 2011***). Heatmaps were generated using the R package 'pheatmap' v1.0.12 (***Kolde, 2015***). The web application was developed using the R package 'shiny' v1.0.4.2 (***Chang et al., 2020***). All analyses were undertaken using R (version 3.5.1).

Acknowledgements

We would like to thank all the consortia for making their summary statistics available for the benefit of this work. We also thank the participants of the UK Biobank study. Data were accessed under UKB application #15825 and #81499. All authors work at the MRC Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/1, MC_UU_00011/4). TRG and GDS conduct research at the NIHR Biomedical Research Centre at the University Hospitals Bristol NHS Foundation Trust and the University of Bristol. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health. SF is supported by a Wellcome Trust PhD studentship in Molecular, Genetic and Lifecourse Epidemiology [108902/Z/15/Z]. MVH works in a unit that receives funding from the UK Medical Research Council and is supported by a British Heart Foundation Intermediate Clinical Research Fellowship (FS/18/23/33512) and the National Institute for Health Research Oxford Biomedical Research Centre.

Additional information

Competing interests

Michael V Holmes: MVH has consulted for Boehringer Ingelheim, and in adherence to the University of Oxford's Clinical Trial Service Unit & Epidemiological Studies Unit (CSTU) staff policy, did not accept personal honoraria or other payments from pharmaceutical companies. Tom R Gaunt: TRG receives funding from Biogen for unrelated research. Tom G Richardson: TGR is employed part-time by Novo Nordisk outside of this work. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Medical Research Council	MC_UU_00011/1	George Davey Smith
Medical Research Council	MC_UU_00011/4	Tom R Gaunt
British Heart Foundation	FS/18/23/33512	Michael V Holmes

Funder	Grant reference number	Author
Wellcome Trust	108902/Z/15/Z	Si Fang

The funders had no role in study design, data collection, and interpretation, or the decision to submit the work for publication. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Author contributions

Si Fang, Formal analysis, Investigation, Visualization, Methodology, Writing – review and editing; Michael V Holmes, Tom R Gaunt, George Davey Smith, Methodology, Writing – review and editing; Tom G Richardson, Conceptualization, Resources, Data curation, Formal analysis, Supervision, Funding acquisition, Validation, Investigation, Visualization, Methodology, Writing – original draft, Writing – review and editing

Author ORCIDs

Si Fang http://orcid.org/0000-0003-4934-1212
Michael V Holmes http://orcid.org/0000-0001-6617-0879
Tom R Gaunt http://orcid.org/0000-0003-0924-3247
George Davey Smith http://orcid.org/0000-0002-1407-8314
Tom G Richardson http://orcid.org/0000-0002-7918-2040

Ethics

Our study involves previously collected data (genomic sequencing data and metabolites data) of human participants in the UK Biobank (UKB) cohort study. Ethical approval for the UKB was obtained from the Research Ethics Committee (REC; approval number: 11/NW/0382) and informed consent was collected from all participants enrolled in UKB.

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.73951.sa1
Author response https://doi.org/10.7554/eLife.73951.sa2

Additional files

Supplementary files

- Supplementary file 1. Supplementary tables. (a) Genome-wide association studies used to derive weights for polygenic risk scores in this study. (b) Metabolic traits analyses in this study from the UK Biobank nuclear magnetic resonance (NMR) panel. (c) Principal component analysis of the NMR metabolites data. (d) Associations with the body mass index polygenic risk score. (e) Mendelian randomization results with very low-density lipoprotein (VLDL) particle size as the exposure and complex traits as the outcome. (f) Polygenic risk scores association for 125 complex traits with glycoprotein acetyls levels. (g) Mendelian randomization results for complex traits and disease liability with glycoprotein acetyls as an outcome. (h) Mendelian randomization results using weighted median and MR-Egger for complex traits and disease liability with glycoprotein acetyls as an outcome. (i) Genetic instruments for glycoprotein acetyls. (j) Mendelian randomization results with glycoprotein acetyls as our exposure and complex traits/diseases as our outcome. (k) Genetic instruments for five proteins contributing to NMR signal of glycoprotein acetyls. (l) Mendelian randomization results with each of the five acute-phase proteins as the exposure and complex traits/diseases as the outcome. (m) Mendelian randomization results using alpha1-antitrypsin as the exposure and gamma glutamyl transferase as the outcome. (n) Associations with the coronary artery disease polygenic score in the full sample. (o) Associations with the coronary artery disease polygenic risk score in the youngest and oldest age tertiles. (p) List of lipid-associated genes removed from polygenic scores in sensitivity analysis. (q) Associations with the Alzheimer's disease polygenic risk score. (r) Associations with the Alzheimer's disease polygenic risk score (excluding lipid loci).
- Transparent reporting form

Data availability

All data generated in this study can be downloaded from the web application of our metabolites-PGS atlas: http://mrcieu.mrsoftware.org/metabolites_PRS_atlas/.

References

Ala-Korpela M, Zhao S, Järvelin M-R, Mäkinen V-P, Ohukainen P. 2022. Apt interpretation of comprehensive lipoprotein data in large-scale epidemiology: disclosure of fundamental structural and metabolic relationships. *International Journal of Epidemiology* **51**:996–1011. DOI: https://doi.org/10.1093/ije/dyab156, PMID: 34405869

Bell JA, Richardson TG, Wang Q, Sanderson E, Palmer T, Walker V, O'Keeffe LM, Timpson NJ, Cichonska A, Julkunen H, Würtz P, Holmes MV, Davey Smith G. 2022. Effects of general and central adiposity on circulating lipoprotein, lipid, and metabolite levels in UK Biobank: a multivariable Mendelian randomization study. *The Lancet Regional Health. Europe* **21**:100457. DOI: https://doi.org/10.1016/j.lanepe.2022.100457, PMID: 35832062

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. DOI: https://doi.org/10.1093/ije/dyv080, PMID: 26050253

Bowden J, Davey Smith G, Haycock PC, Burgess S. 2016. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40**:304–314. DOI: https://doi.org/10.1002/gepi.21965, PMID: 27061298

Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**:1177–1186. DOI: https://doi.org/10.1016/j.cell.2017.05.038, PMID: 28622505

Burgess S, Butterworth A, Thompson SG. 2013. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37**:658–665. DOI: https://doi.org/10.1002/gepi.21758, PMID: 24114802

Burgess S, Thompson SG. 2013. Use of allele scores as instrumental variables for Mendelian randomization. *International Journal of Epidemiology* **42**:1134–1144. DOI: https://doi.org/10.1093/ije/dyt093, PMID: 24062299

Burgess S, Small DS, Thompson SG. 2017. A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research* **26**:2333–2355. DOI: https://doi.org/10.1177/0962280215597579, PMID: 26282889

Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, Motyer A, Vukcevic D, Delaneau O, O'Connell J, Cortes A, Welsh S, Young A, Effingham M, McVean G, Leslie S, Allen N, Donnelly P, Marchini J. 2018. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**:203–209. DOI: https://doi.org/10.1038/s41586-018-0579-z, PMID: 30305743

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-Generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**:7. DOI: https://doi.org/10.1186/s13742-015-0047-8, PMID: 25722852

Chang W, Cheng J, Allaire JJ, Xie Y, Mcpherson J. 2020. Shiny: web application framework for R. 1.4.0.2. R Package. https://CRAN.R-project.org/package=shiny

Connelly MA, Otvos JD, Shalaurova I, Playford MP, Mehta NN. 2017. GlycA, a novel biomarker of systemic inflammation and cardiovascular disease risk. *Journal of Translational Medicine* **15**:219. DOI: https://doi.org/10.1186/s12967-017-1321-6, PMID: 29078787

Davey Smith G, Ebrahim S. 2003. " Mendelian randomization ": can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**:1–22. DOI: https://doi.org/10.1093/ije/dyg070, PMID: 12689998

Davey Smith G, Hemani G. 2014. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23**:R89–R98. DOI: https://doi.org/10.1093/hmg/ddu328, PMID: 25064373

Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* **10**:3328. DOI: https://doi.org/10.1038/s41467-019-11112-0, PMID: 31346163

Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, Bates P, Palmer T. 2020. The MRC IEU OpenGWAS Data Infrastructure. *bioRxiv*. DOI: https://doi.org/10.1101/2020.08.10.244293

Fang S, Hemani G, Richardson TG, Gaunt TR, Smith GD. 2022. Evaluating and implementing block jackknife resampling mendelian randomization to mitigate bias induced by overlapping samples. *Human Molecular Genetics* **6**:ddac186. DOI: https://doi.org/10.1093/hmg/ddac186, PMID: 35932451

Ference BA, Kastelein JJP, Ray KK, Ginsberg HN, Chapman MJ, Packard CJ, Laufs U, Oliver-Williams C, Wood AM, Butterworth AS, Di Angelantonio E, Danesh J, Nicholls SJ, Bhatt DL, Sabatine MS, Catapano AL. 2019. Association of triglyceride-lowering LPL variants and LDL-C-lowering LDLR variants with risk of coronary heart disease. *JAMA* **321**:364–373. DOI: https://doi.org/10.1001/jama.2018.20045, PMID: 30694319

Ferguson AC, Tank R, Lyall LM, Ward J, Celis-Morales C, Strawbridge R, Ho F, Whelan CD, Gill J, Welsh P, Anderson JJ, Mark PB, Mackay DF, Smith DJ, Pell JP, Cavanagh J, Sattar N, Lyall DM. 2020. Alzheimer's disease susceptibility gene apolipoprotein E (apoe) and blood biomarkers in UK biobank (N = 395,769). *Journal of Alzheimer's Disease* **76**:1541–1551. DOI: https://doi.org/10.3233/JAD-200338, PMID: 32651323

Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, Gunnarsdottir K, Helgason A, Oddsson A, Halldorsson BV, Jensson BO, Zink F, Halldorsson GH, Masson G, Arnadottir GA, Katrinardottir H, Juliusson K, Magnusson MK, Magnusson OT, Fridriksdottir R, et al. 2021. Large-Scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53**:1712–1721. DOI: https://doi.org/10.1038/s41588-021-00978-w, PMID: 34857953

Folkersen L, Gustafsson S, Wang Q, Hansen DH, Hedman ÅK, Schork A, Page K, Zhernakova DV, Wu Y, Peters J, Eriksson N, Bergen SE, Boutin TS, Bretherick AD, Enroth S, Kalnapenkis A, Gådin JR, Suur BE, Chen Y, Matic L, et al. 2020. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nature Metabolism* **2**:1135–1148. DOI: https://doi.org/10.1038/s42255-020-00287-2, PMID: 33067605

Gallois A, Mefford J, Ko A, Vaysse A, Julienne H, Ala-Korpela M, Laakso M, Zaitlen N, Pajukanta P, Aschard H. 2019. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nature Communications* **10**:4788. DOI: https://doi.org/10.1038/s41467-019-12703-7, PMID: 31636271

Ginestet C. 2011. Ggplot2: elegant graphics for data analysis. *Journal of the Royal Statistical Society* **174**:245–246. DOI: https://doi.org/10.1111/j.1467-985X.2010.00676_9.x

Griffith GJ, Morris TT, Tudball MJ, Herbert A, Mancano G, Pike L, Sharp GC, Sterne J, Palmer TM, Davey Smith G, Tilling K, Zuccolo L, Davies NM, Hemani G. 2020. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications* **11**:5749. DOI: https://doi.org/10.1038/s41467-020-19478-2, PMID: 33184277

Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, Laurin C, Burgess S, Bowden J, Langdon R, Tan VY, Yarmolinsky J, Shihab HA, Timpson NJ, Evans DM, Relton C, Martin RM, Davey Smith G, Gaunt TR, Haycock PC. 2018. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**:e34408. DOI: https://doi.org/10.7554/eLife.34408, PMID: 29846171

Holmes MV, Davey Smith G. 2019. Can Mendelian randomization shift into reverse GEAR? *Clinical Chemistry* **65**:363–366. DOI: https://doi.org/10.1373/clinchem.2018.296806, PMID: 30692117

Holmes MV, Richardson TG, Ference BA, Davies NM, Davey Smith G. 2021. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nature Reviews. Cardiology* **18**:435–453. DOI: https://doi.org/10.1038/s41569-020-00493-1, PMID: 33707768

Julkunen H, Cichońska A, Slagboom PE, Würtz P, Nightingale Health UK Biobank Initiative. 2021. Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *eLife* **10**:e63033. DOI: https://doi.org/10.7554/eLife.63033, PMID: 33942721

Kibinge NK, Relton CL, Gaunt TR, Richardson TG. 2020. Characterizing the causal pathway for genetic variants associated with neurological phenotypes using human brain-derived proteome data. *American Journal of Human Genetics* **106**:885–892. DOI: https://doi.org/10.1016/j.ajhg.2020.04.007, PMID: 32413284

Kolde R. 2015. Pheatmap: pretty heatmaps. 1.0.12. CRAN. https://CRAN.R-project.org/package=pheatmap

Lawler PR, Akinkuolie AO, Chandler PD, Moorthy MV, Vandenburgh MJ, Schaumberg DA, Lee I-M, Glynn RJ, Ridker PM, Buring JE, Mora S. 2016. Circulating N-linked glycoprotein acetyls and longitudinal mortality risk. *Circulation Research* **118**:1106–1115. DOI: https://doi.org/10.1161/CIRCRESAHA.115.308078, PMID: 26951635

Lewis CM, Vassos E. 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome Medicine* **12**:44. DOI: https://doi.org/10.1186/s13073-020-00742-5, PMID: 32423490

Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, Price AL. 2015. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics* **47**:284–290. DOI: https://doi.org/10.1038/ng.3190, PMID: 25642633

Lord J, Jermy B, Green R, Wong A, Xu J, Legido-Quigley C, Dobson R, Richards M, Proitsi P. 2021. Mendelian randomization identifies blood metabolites previously linked to midlife cognition as causal candidates in Alzheimer ' S disease. *PNAS* **118**:e2009808118. DOI: https://doi.org/10.1073/pnas.2009808118, PMID: 33879569

MacTel Consortium, Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, Raffler J, Biggs EK, Oliver-Williams C, Auyeung VPW, Luan J, Wheeler E, Paige E, Surendran P, Michelotti GA, Scott RA, Burgess S, Zuber V, Sanderson E, et al. 2021. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics* **53**:54–64. DOI: https://doi.org/10.1038/s41588-020-00751-5, PMID: 33414548

Otvos JD, Shalaurova I, Wolak-Dinsmore J, Connelly MA, Mackey RH, Stein JH, Tracy RP. 2015. GlycA: a composite nuclear magnetic resonance biomarker of systemic inflammation. *Clinical Chemistry* **61**:714–723. DOI: https://doi.org/10.1373/clinchem.2014.232918, PMID: 25779987

Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, Kanai M, Rajagopal VM, Parolo PDB, Baya N, Carey CE, Karjalainen J, Als TD, Van der Zee MD, Day FR, Ong KK, FinnGen Study, 23andMe Research Team, Agee M, Aslibekyan S, et al. 2021. Genetic analyses identify widespread sex-differential participation bias. *Nature Genetics* **53**:663–671. DOI: https://doi.org/10.1038/s41588-021-00846-7, PMID: 33888908

Richardson TG, Davey Smith G, Munafò MR. 2019a. Support a health-protective effect of neuroticism in population subgroups? *Psychological Science* **30**:629–632.

Richardson TG, Harrison S, Hemani G, Davey Smith G. 2019b. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8**:e43657. DOI: https://doi.org/10.7554/eLife.43657, PMID: 30835202

Richardson TG, Hemani G, Gaunt TR, Relton CL, Davey Smith G. 2020a. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nature Communications* **11**:185. DOI: https://doi.org/10.1038/s41467-019-13921-9, PMID: 31924771

Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, Holmes MV. 2020b. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLOS Medicine* **17**:e1003062. DOI: https://doi.org/10.1371/journal.pmed.1003062, PMID: 32203549

Richardson TG, Wang Q, Sanderson E, Mahajan A, McCarthy MI, Frayling TM, Ala-Korpela M, Sniderman A, Davey Smith G, Holmes MV. 2021. Effects of apolipoprotein B on lifespan and risks of major diseases including type 2 diabetes: a Mendelian randomisation analysis using outcomes in first-degree relatives. *The Lancet. Healthy Longevity* **2**:e317–e326. DOI: https://doi.org/10.1016/S2666-7568(21)00086-6, PMID: 34729547

Richardson TG, Leyden GM, Wang Q, Bell JA, Elsworth B, Davey Smith G, Holmes MV. 2022. Characterising metabolomic signatures of lipid-modifying therapies through drug target Mendelian randomisation. *PLOS Biology* **20**:e3001547. DOI: https://doi.org/10.1371/journal.pbio.3001547, PMID: 35213538

Ritchie SC, Lambert SA, Arnold M, Teo SM, Lim S, Scepanovic P, Marten J, Zahid S, Chaffin M, Liu Y, Abraham G, Ouwehand WH, Roberts DJ, Watkins NA, Drew BG, Calkin AC, Di Angelantonio E, Soranzo N, Burgess S, Chapman M, et al. 2021a. Integrative analysis of the plasma proteome and polygenic risk of cardiometabolic diseases. *Nature Metabolism* **3**:1476–1483. DOI: https://doi.org/10.1038/s42255-021-00478-5, PMID: 34750571

Ritchie SC, Surendran P, Karthikeyan S, Lambert SA, Bolton T, Pennells L, Danesh J. 2021b. Quality Control and Removal of Technical Variation of NMR Metabolic Biomarker Data in ~120,000 UK Biobank Participants. [medRxiv]. DOI: https://doi.org/10.1101/2021.09.24.21264079

Sanderson E, Glymour MM, Holmes MV, Kang H, Morrison J, Munafò MR, Palmer T, Schooling CM, Wallace C, Zhao Q, Davey Smith G. 2022. Mendelian randomization. *Nature Reviews Methods Primers* **2**:6. DOI: https://doi.org/10.1038/s43586-021-00092-5

Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte I, Forgetta V, Yang TP, Walter K, Menni C, Chen L, Vasquez L, Valdes AM, Hyde CL, Wang V, Ziemek D, Roberts P, Xi L, et al. 2014. An atlas of genetic influences on human blood metabolites. *Nature Genetics* **46**:543–550. DOI: https://doi.org/10.1038/ng.2982, PMID: 24816252

Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, Venkataraman GR, Wainberg M, Ollila HM, Kiiskinen T, Havulinna AS, Pirruccello JP, Qian J, Shcherbina A, FinnGen, Rodriguez F, Assimes TL, Agarwala V, Tibshirani R, Hastie T, et al. 2021. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics* **53**:185–194. DOI: https://doi.org/10.1038/s41588-020-00757-z, PMID: 33462484

Sirugo G, Williams SM, Tishkoff SA. 2019. The missing diversity in human genetic studies. *Cell* **177**:1080. DOI: https://doi.org/10.1016/j.cell.2019.04.032, PMID: 31051100

Sniderman AD, Thanassoulis G, Glavinovic T, Navar AM, Pencina M, Catapano A, Ference BA. 2019. Apolipoprotein B particles and cardiovascular disease: a narrative review. *JAMA Cardiology* **4**:1287–1295. DOI: https://doi.org/10.1001/jamacardio.2019.3780, PMID: 31642874

Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. 2015. Uk Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* **12**:e1001779. DOI: https://doi.org/10.1371/journal.pmed.1001779, PMID: 25826379

Suhre K, Shin S-Y, Petersen A-K, Mohney RP, Meredith D, Wägele B, Altmaier E, CARDIoGRAM, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmüller G, Köttgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes H-W, et al. 2011. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**:54–60. DOI: https://doi.org/10.1038/nature10354, PMID: 21886157

Torkamani A, Wineinger NE, Topol EJ. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews. Genetics* **19**:581–590. DOI: https://doi.org/10.1038/s41576-018-0018-x, PMID: 29789686

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics* **101**:5–22. DOI: https://doi.org/10.1016/j.ajhg.2017.06.005, PMID: 28686856

Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, Brugge H, Oelen R, de Vries DH, van der Wijst MGP, Kasela S, Pervjakova N, Alves I, Favé M-J, Agbessi M, Christiansen MW, et al. 2021. Large-Scale cis- and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**:1300–1310. DOI: https://doi.org/10.1038/s41588-021-00913-z, PMID: 34475573

Würtz P, Wang Q, Kangas AJ, Richmond RC, Skarp J, Tiainen M, Tynkkynen T, Soininen P, Havulinna AS, Kaakinen M, Viikari JS, Savolainen MJ, Kähönen M, Lehtimäki T, Männistö S, Blankenberg S, Zeller T, Laitinen J, Pouta A, Mäntyselkä P, et al. 2014. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLOS Medicine* **11**:e1001765. DOI: https://doi.org/10.1371/journal.pmed.1001765, PMID: 25490400