Dual site proteomic analyses reveal potential drug targets for cardiovascular disease

Christopher Aldous Oldnall[1,2], Julian Ng-Kee-Kwong[2], Jimi Wills[3], Anne Richmond[2], Tim Regan[4], Sara Clohisey Hendry[4,5], Archie Campbell[6,7], J. Kenneth Baillie[2,4,5], Alex von Kriegsheim[3], Chris Haley[2,4], Ava Khamseh[2,8], Sjoerd Viktor Beentjes[1,2*], Andrew D. Bretherick[2,9*]

[1]School of Mathematics, The University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Rd, Edinburgh, EH9 3FD, UK.
[2]MRC Human Genetics Unit, The Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital Campus, Crewe Road South, Edinburgh, EH4 2XU, UK.
[3]CRUK Scotland Centre, The Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital Campus, Crewe Road South, Edinburgh, EH4 2XU, UK.
[4]The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, UK.
[5]Baillie Gifford Pandemic Science Hub, Centre for Inflammation Research, The Queen's Medical Research Institute, The University of Edinburgh, Edinburgh, EH16 4UU, UK.
[6]Centre for Genomic and Experimental Medicine, The Institute of Genetics and Cancer, The University of Edinburgh, Western General Hospital Campus, Crewe Road South, Edinburgh, EH4 2XU, UK.
[7]Centre for Medical Informatics, Usher Institute, The University of Edinburgh, Edinburgh Bioquarter, 5-7 Little France Road, Edinburgh, EH16 4UX, UK.
[8]School of Informatics, The University of Edinburgh, Informatics Forum, 10 Crichton St, Edinburgh, EH8 9AB, UK.
[9]Pain Service, NHS Tayside, Ninewells Hospital, Dundee, DD1 9SY, UK.

*Corresponding author(s). E-mail(s): sjoerd.beentjes@ed.ac.uk; a.bretherick@ed.ac.uk;

Abstract

Background:
While genome-wide association studies (GWAS) hold great promise for unravelling disease pathophysiology, the translation of disease-associated genetic loci into clinically actionable information remains a challenge. Mendelian randomisation (MR), using expressed proteins as exposures and disease as an outcome, stands as a powerful analytical approach for leveraging GWAS data to identify potential drug-targets—at scale—in a data-driven manner. Cardiovascular disease (CVD) is a major health burden worldwide, and therefore is an important outcome for which to establish and prioritise potential therapeutic targets.

Methods:
In this study, we utilised generalised summary-data-based MR (GSMR) with novel mass-spectrometry-based isoform-specific protein groups measured from peripheral-blood mononuclear cell (PBMC) obtained from Generation Scotland and antibody-based plasma protein measures from UK Biobank as exposures, and two CVD and three CVD-related risk-factors from UK Biobank as outcomes. Further, we used colocalisation to assess support for a shared causal variant between the proteins and the disease outcomes providing further evidence supporting a causal link.

Results:
We evaluate expression of 5,114 isoform-specific protein groups in PBMCs from 862 individuals. GSMR analysis, using this data, found 16 putative causal proteins across three of the CVD/CVD-related risk-factors with seven supported by colocalisation analysis. Within the plasma GSMR analysis, 761 putative causal proteins were identified, of which 145 were supported by colocalisation. In addition, we go on to examine enrichment amongst the results and find enrichment of pathways which relate to cholesterol metabolism and platelet function. There was an overlap of three proteins between significant GSMR results in PBMCs and plasma, with two proteins (COMT and SWAP70) identifying opposite directions of effect of the relevant outcome, and one identifying a concordant direction of effect (HLA-DRA).

Discussion:
This study identifies a number of proteins and pathways that may be involved in CVD pathogenesis. It also demonstrates the importance of the location of protein measurement and the methods by which it is quantified. Our research contributes to ongoing efforts to bridge the gap between genotype and phenotype.

1 Introduction

Cardiovascular disease (CVD) is of clear global importance [1–3]. With respect to its genetics, it is a complex-trait, underpinned by multiple genetic variants. Genome-wide association studies (GWAS) have emerged as a powerful tool for uncovering

the intricate genetic architecture of many diseases, including CVD. However, a substantial portion of disease-associated variation resides outside the coding sequence of genes and the translation of GWAS findings into actionable clinical insight remains a significant challenge [4].

Here we attempt to identify proteins that are on the causal pathway to CVD, and its risk-factors. By combining a novel data set of protein-quantitative trait loci (pQTLs) identified by mass-spectrometry in peripheral-blood mononuclear cells (PBMCs) from Generation Scotland (GS), pQTLs in plasma from the UK Biobank (UKB) [5], and GWAS of CVD and CVD-related risk-factors, we identify 774 of unique putative drug-targets by employing a two-sample Mendelian randomisation (MR) approach. MR identifies naturally occurring subgroups with genetically determined high, or low, values of an exposure (e.g., a protein) and tests for an association between these groups and outcome (e.g., CVD). Since the set of genetic variants inherited by an individual is (approximately) random, MR is akin to a 'natural' randomised controlled trial.

Under such circumstances, the finding that a genetic variant is significantly associated with the risk-factor under investigation, which may be a particular protein, while also being associated with the disease, suggests a potential causal role for that protein in that disease. It should be noted that the following three assumptions must hold to ensure the validity of such an analysis: (i) the genetic variant must be strongly associated with the protein of interest, (ii) it must be independent of any confounding of the protein and phenotype, and (iii) it must influence the disease solely through its effect on the protein in question, i.e., there is no horizontal pleiotropy [6]. When a genetic variant satisfies these three assumptions it is called a valid instrumental variable (IV).

Summary data-based Mendelian Randomization (SMR) and generalised SMR (GSMR) have emerged as MR techniques that integrate data from independent studies [7, 8]. These methods harness GWAS summary-level data, thereby allowing analyses to be completed where the exposure and outcome have been measured in two different cohorts sampled from the same underlying population. Additionally, the GSMR pipeline incorporates outlier analysis (termed HEIDI), which attempts to identify and remove potential pleiotropic effects within the single nucleotide polymorphism (SNP) set. This is done to ensure that all included SNPs are valid IVs. Colocalisation of the exposure and outcome is also used here, in conjunction with the MR analyses, in order to provide further support for potential therapeutic targets [9]. Bayesian testing for colocalisation using the software 'coloc' allows for intuitive interpretation of posterior probabilities aligning with different colocalisation possibilities [10].

Previous studies have employed RNA expression (eQTLs) and pQTLs in MR analyses with great success, identifying putative drug-targets (e.g., see [11–13]). While substantial literature exists on statistical methodology related to IV analysis [14], using MR analysis to assess the functional significance of genes at disease-associated loci presents its own set of challenges. Of particular relevance here, for practical reasons, it is often difficult to measure RNA/protein abundance in a disease-relevant

tissue/cell-type. Indeed, most previous pQTL studies have been conducted using proteins measured in plasma [5, 15, 16]. Given the lack of available cellular pQTLs, there has been limited exploration of MR using cell-based measures of proteins abundance to date.

In this study, we present the results of our investigations using GSMR and colocalisation techniques to establish causal links between proteomic data, from both novel cellular pQTL data from PBMCs (GS) and plasma (UKB), and CVD/CVD-related risk-factors from UKB. Here we compare results simultaneously for proteomic associations with CVD/CVD-related risk-factors in both sets of data.

2 Results

2.1 Cellular pQTL analysis reveals 127 unique significant SNP-protein expression associations

The majority of previous pQTL data has been measured in plasma. We performed genome-wide association of the cellular protein abundance for 5,114 isoform-specific protein groups in PBMCs from GS. After applying a Bonferroni correction, of the number of protein groups tested, 127 protein groups were identified which had at least one significantly associated SNP ($p < 9.78 \times 10^{-12}$).

At an uncorrected genome-wide significance threshold ($p < 5 \times 10^{-8}$), 663 protein groups were identified to have at least one SNP, associated with protein expression. Subsequent analyses focused on proteins with at least one cis-pQTL ($\pm$1Mb from the gene encoding the protein) significant at a genome-wide significance threshold ($p < 5 \times 10^{-8}$) [17–19]. This approach narrowed our focus to 165 proteins for GSMR and colocalisation analysis (Fig. 2A). Fig. 1 displays, as exemplars, Manhattan plots for the GWAS results of three proteins which are of subsequent interest in this paper: COMT, HLA-DRA, and SWAP70.

Mass-spectrometry measures the abundance of peptide fragments of the sequence of a protein, and its isoforms, as included in UniProt. If the sequence changes, for example due to missense variation, the detection of the peptide containing the changed residue will be impacted. We identified 570 modification-specific peptides with at least one SNP that was associated with expression at $p < 5 \times 10^{-8}$/5,114, and 541 at $p < 5 \times 10^{-8}$/29,639, corresponding to the number of isoform-specific protein group and modification-specific peptide GWAS considered, respectively; of these, 39% (218/556) and 40% (210/527) contained a missense variant. Peptides lacking a common missense variant in their sequence are more likely to reflect true expression differences, suggesting the associated SNP has a pronounced effect on protein expression.

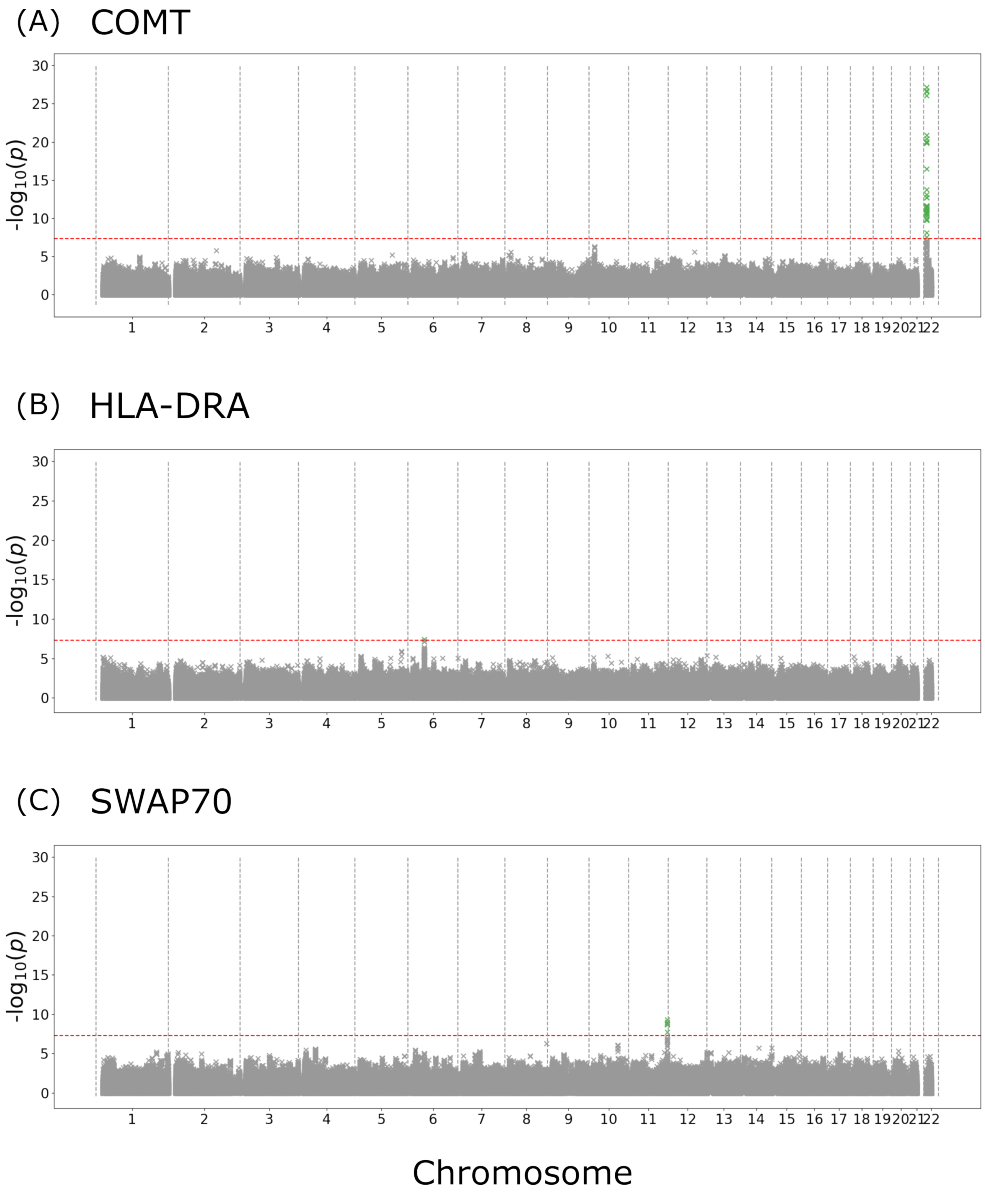


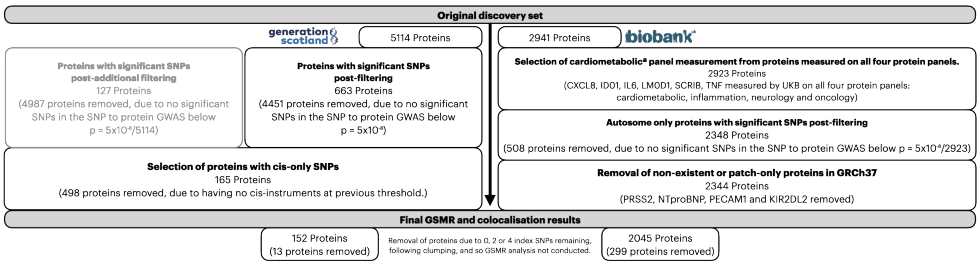
Fig. 1: **Genome-wide SNP associations with protein expression.** Manhattan plots depicting $-\log_{10}(p)$ of SNPs across all chromosomes for **(A)** COMT, **(B)** HLA-DRA, and **(C)** SWAP70 proteins. SNPs associated with the protein below the genome-wide significance threshold ($p < 5 \times 10^{-8}$) are denoted in green. The red dashed line is at $-\log_{10}(5 \times 10^{-8})$) corresponding to the genome-wide significance level.

2.2 Comparison of cellular and plasma pQTLs shows 50 mutual (non-isoform specific) instrumented proteins

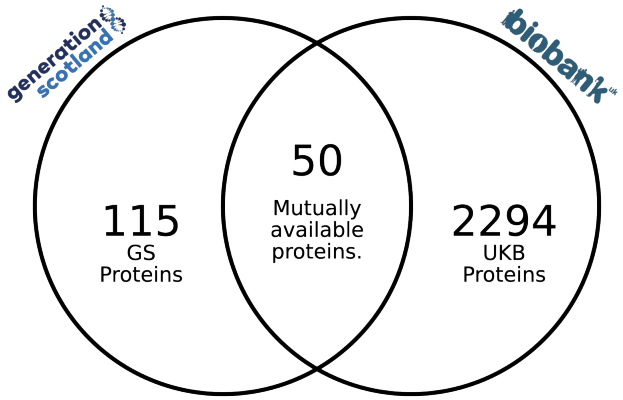
The UK Biobank pharma-proteomics-project (UKB-PPP) provides a set of 2,941 unique proteins [5]. The dataset underwent a filtering process, as outlined in Fig. 2A, and elaborated in the Methods section 4.2. Post-filtering, which included the removal of duplicate measurements, imposing the requirement for proteins to have at least one SNP linearly associated with expression after Bonferroni correction ($p < 5 \times 10^{-8}/2,923$), and the reconciliation of SNPs across datasets, we retained 2,045 proteins for GSMR and colocalisation analyses. Out of the initial 165 GS proteins with at least one cis-SNP associated with cellular expression ($p < 5 \times 10^{-8}$), considering only the non-isoform specific form of the protein, 50 were also found in plasma and had SNPs meeting the Bonferroni corrected threshold for plasma expression, as displayed in Fig. 2B.

We assessed the allele frequencies of the 50 lead SNPs (the SNP most significantly associated with protein expression) of the proteins in GS that were also present in the UK Biobank data. The two populations are similar in allele frequencies, see Fig. 2C, giving confidence in the use of GS SNP-protein data in conjunction with UKB SNP-outcome data for MR and colocalisation analyses. Furthermore, we examined the effect-to-standard error ratios of these SNPs on expression levels in plasma and PBMCs. A strong linear correlation ($R = 0.759, p = 1.757 \times 10^{-10}$) was observed, see Fig. 2D. Together, these findings suggest that any subsequent non-concordance between cellular and plasma MR analysis results may be due to expression difference or measurement technique used.

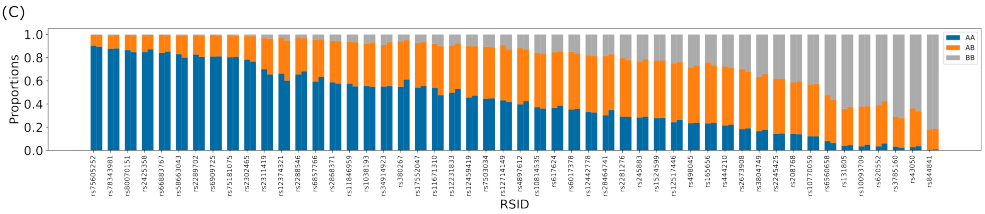
(A)



(B)



(C)



(D)

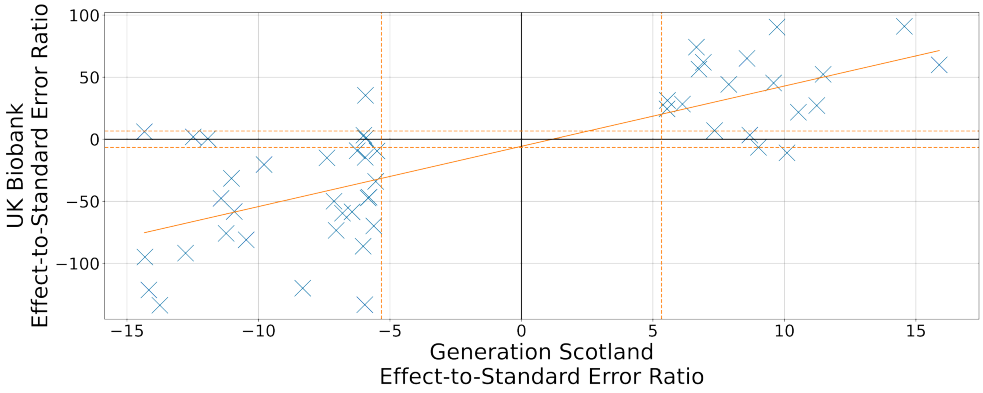


Fig. 2: **Filtering and comparison of the PBMC and plasma proteomic data and populations.** **(A)** Filtering process of both PBMC and plasma measured proteins. Mass-spectrometry protein groups are referred to simply as proteins in the figure panel. Only those proteins which had at least one associated SNP with the expression (at either $p < 5 \times 10^{-8}$ for the protein groups in PBMCs, or the corresponding Bonferonni corrected significance level for the proteins in plasma) were retained. Additional filtering included use of cis-only SNPs for the PBMC set, removal of non-autosomal proteins, and those which did not map to GRCh37 (plasma proteins only). Proteins with either 0, 2 or 4 index SNPs remaining, following clumping, did not have GSMR analysis conducted. **(B)** Venn diagram illustrating the overlap between protein sets in GS and UKB. Proteins included in the diagram are non-isoform specific and 1) in GS, had at least one cis-SNP associated with their expression ($p < 5 \times 10^{-8}$); or 2) in UKB, had at least one SNP associated with their expression following Bonferroni correction ($p < 5 \times 10^{-8}/2923$). **(C)** Comparison of the allele frequencies of the 50 lead SNPs in the PBMC data, where the protein is measured in both PBMC and plasma data, and whose expression is associated with a cis-SNP at genome-wide significance level $p < 5 \times 10^{-8}$ in the PBMCs. For each of the 50 proteins, the lead SNP with respect to expression is selected and its allele frequency reported. The rsID of the lead SNP is denoted on the x-axis, whilst the y-axis displays the proportions of alleles (AA, AB, or BB) for GS and UKB next to one another, i.e., one bar has both the allele frequencies of the same SNP from GS and UKB next to each other. As can be seen, the relative frequencies of alleles are similar between the two cohorts. **(D)** Comparison of effect-to-standard error ratio in GS and UKB, of the lead SNP with respect to expression in GS, for the 50 proteins present in both GS and the UKB. On the x-axis are the effect-to-standard error ratios for the lead SNP to expression association in GS, and on the y-axis are the effect-to-standard error ratios in UKB. The linear correlation is $R = 0.759$ ($p = 1.757 \times 10^{-10}$). Two SNPs are significantly associated with the expression of the same protein in both GS and UKB, but with the opposite direction of effect—these can be found in the upper-left and bottom-right quadrants. This provides evidence that there is a difference between cellular and plasma expression and/or the differing measurement modalities.

2.3 Protein-to-disease

2.3.1 Analyses with PBMC measured proteins identifies 16 protein targets for CVD/CVD-related risk-factors

There were 16 unique protein groups involved in 24 significant protein-to-outcome relationships (FDR $< 5\%$, Benjamini–Hochberg [20]), identified using GSMR for the CVD/CVD related-risk outcomes: diabetes (five protein), high cholesterol (eight protein) and hypertension (11 protein). No significant relationships were identified for the angina or heart attack outcomes. Colocalisation analyses showed support for a range of these same proteins – the probability, H4, that the protein and outcome share a single causal variant. The significant results after multiple hypothesis correction (FDR < 0.05, Benjamini–Hochberg) are presented in Fig. 3 and Table 1. Full PBMC results are provided in Supplementary Table 9.1.

ERLIN1 is causally implicated in diabetes using GSMR, with high support (H4 > 0.8) from colocalisation. MARC1 and HP have high colocalisation H4 probabilities, further supporting the GSMR analysis, prioritising them as putative causal targets for high cholesterol. Additionally, we see SWAP70 linked to hypertension with high colocalisation support. Some proteins are significant (FDR < 0.05, Benjamini–Hochberg) across multiple cardiovascular disease risk-factors, namely, HLA-B, HLA-DRA, HLA-DRB1, HLA-C, and COMT. HLA-DRA and COMT were identified as putative proteins which are causally linked with both high cholesterol and hypertension.

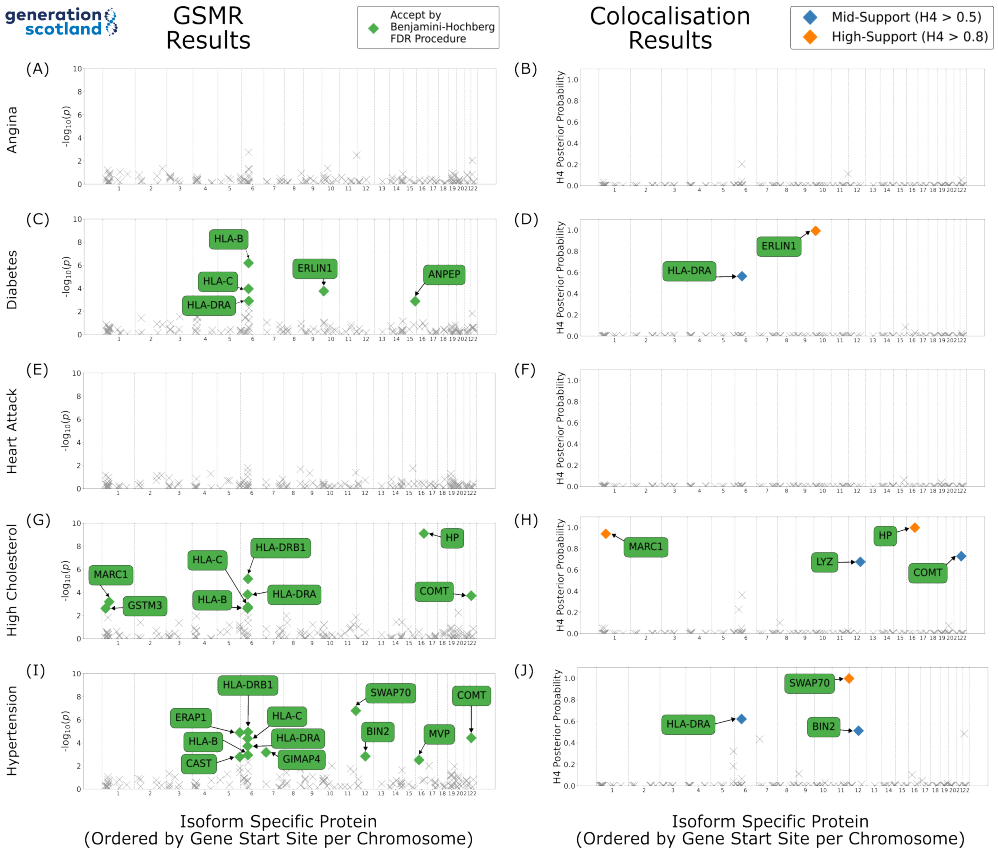


Fig. 3: **PBMC GSMR and colocalisation results. Left:** GSMR results obtained by combining the SNP-protein associations from GS and the SNP-outcome association from Gene Atlas [21]. For each outcome, GSMR results are labeled when they are significant after multiple hypothesis correction (FDR < 0.05, Benjamini–Hochberg). **Right:** Corresponding colocalisation results estimating posterior probability support for the H_4 hypothesis that both the exposure and outcome share a common causal SNP for the corresponding trait. The posterior probability for H_4 is coloured coded: $0.5 - 0.8$ (blue), or > 0.8 (orange); a greater number indicating greater support for the hypothesis. Proteins are ordered by the genomic position of their transcription start site. Traits are as follows: **(A-B)** Angina, **(C-D)** Diabetes, **(E-F)** Heart Attack, **(G-H)** High Cholesterol, **(I-J)** Hypertension.

2.3.2 UKB plasma protein analysis provides support for 761 targets

GSMR analysis was run on the 2,045 plasma proteins in UKB, post-filtering, and the same 5 CVD/CVD risk-related traits as the PBMCs. There were 761 unique proteins identified as significant after multiple testing correction across the different outcomes

Protein Name	UniProtID(s)	GSMR p-value	H_3 PP	H_4 PP
Diabetes				
HLA-B	P01889	6.44×10^{-7}	0.999	4.07×10^{-11}
HLA-C	P10321, P10321-2	2.73×10^{-45}	0.999	4.90×10^{-19}
ERLIN1	O75477	1.76×10^{-4}	1.93×10^{-3}	0.993
HLA-DRA	P01903	1.26×10^{-3}	1.88×10^{-5}	0.564
ANPEP	P15144	1.30×10^{-3}	0.883	8.38×10^{-2}
High Cholesterol				
HP	P00738	8.04×10^{-10}	1.94×10^{-3}	0.998
HLA-B	P01889	2.38×10^{-3}	0.999	3.46×10^{-6}
HLA-DRA	P01903	1.79×10^{-3}	0.0	0.363
HLA-DRB1	P01911	6.63×10^{-6}	0.999	1.06×10^{-5}
HLA-C	P10321, P10321-2	5.28×10^{-38}	0.999	1.44×10^{-9}
GSTM3	P21266	2.43×10^{-3}	2.00×10^{-3}	0.051
COMT	P21964, P21964-2	2.03×10^{-20}	2.68×10^{-3}	0.729
MARC1	Q5VT66, Q5VT66-2, Q5VT66-3	6.43×10^{-4}	4.22×10^{-3}	0.940
Hypertension				
HLA-B	P01889	1.95×10^{-4}	0.999	1.18×10^{-7}
HLA-DRA	P01903	1.16×10^{-3}	0.0	0.621
HLA-DRB1	P01911	1.14×10^{-8}	0.999	3.57×10^{-8}
HLA-C	P10321, P10321-2	4.23×10^{-5}	0.999	3.38×10^{-9}
CAST	P20810-2, P20810-3, P20810-5	1.65×10^{-3}	0.116	0.183
COMT	P21964, P21964-2	3.78×10^{-5}	0.497	0.485
MVP	Q14764	3.04×10^{-3}	2.75×10^{-3}	0.102
GIMAP4	Q9NUV9	6.76×10^{-4}	0.131	0.435
ERAP1	Q9NZ08, Q9NZ08-2	1.29×10^{-5}	0.665	0.322
BIN2	Q9UBW5, Q9UBW5-2, Q9UBW5-3	1.45×10^{-3}	0.122	0.512
SWAP70	Q9UH65	1.65×10^{-7}	2.88×10^{-3}	0.997

Table 1: **GS protein analysis.** The table is divided into sections by the CVD/CVD risk-related trait. Protein name along with the isoform-specific UniProtID(s) are provided in the first two columns. Additionally, the p-value from the GSMR procedure and the H_3 and H_4 posterior probabilities from the colocalisation are given. H_3 corresponds to the hypothesis that both the exposure and outcome have an associated SNP, but a different SNP in each case. Results are highlighted by H4 colocalisation posterior probability: the probability that both the exposure and outcome share a common associated SNP. Blue corresponds to mid-support, $0.5 < H4 < 0.8$, and orange corresponds to high-support, H4 > 0.8).

(FDR $<$ 0.05, Benjamini–Hochberg): angina (78 proteins), diabetes (152 proteins), heart attack (187 proteins), high cholesterol (279 proteins) and hypertension (490 proteins). Similar to the GS PBMC proteins, colocalisation analyses showed high support for a range of proteins to be colocalised with all five outcomes. Proteins APOE, TP53BP1, DEFB4A/DEFB4B, CD22, GZMA, LTB and PLXDC2 were noted as significant (FDR $<$ 0.05, Benjamini-Hochberg) in the GSMR results for all 5 CVD/CVD risk-related traits. Results are shown in Fig. 4, annotated by proteins which were also significant in the PBMC analysis. The full results are provided in Supplementary Table 9.2.

Using StringDB [22], we conducted functional enrichment analysis, for the sets of proteins associated with each trait, employing an FDR threshold of 0.05 (Methods 4.7). The final networks of proteins identified by StringDB were of sizes 69, 144, 181, 271, 476 for the traits angina, diabetes, heart attack, high cholesterol and hypertension respectively. The network for angina is provided in Fig. 5, with results highlighted by colocalisation support, and additional network images are provided in Supplementary Material 9.3.

Functional enrichment for cholesterol metabolism was found for angina (6/21 proteins, KEGG) and high cholesterol (13/21 proteins, KEGG). For heart attack, multiple sets relating to platelet functions were found to be enriched: platelet degranulation (17/64 proteins, Reactome), platelet alpha granule lumen (12/30 proteins, Gene Ontology Component), platelet activation, signalling and aggregation (23/86 proteins, Reactome). No functional enrichment was found for diabetes and hypertension. Full functional enrichment analyses from StringDB for angina, high cholesterol and heart attack are provided in Supplementary Table 9.4.

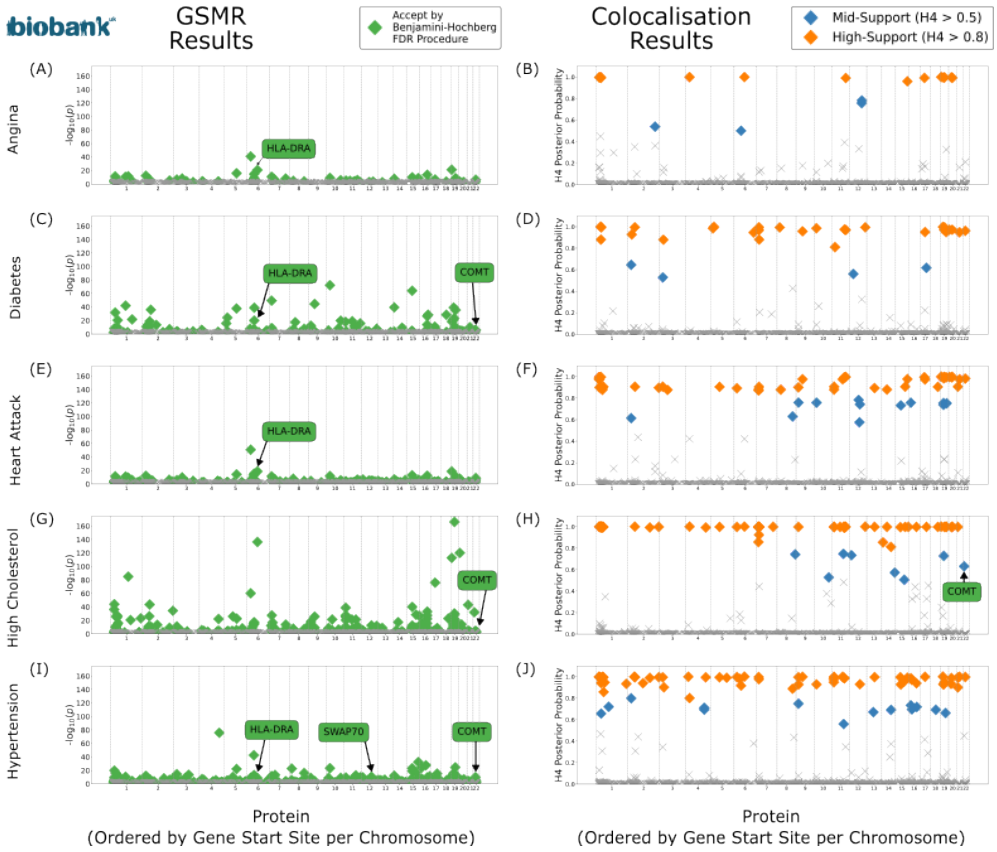


Fig. 4: **Plasma GSMR and colocalisation results. Left:** GSMR results obtained by combining the SNP-protein associations from UKB and the SNP-outcome association from Gene Atlas [21]. **Right:** Corresponding colocalisation results estimating posterior probability support for the H4 hypothesis that both the exposure and outcome share a common causal SNP for the corresponding trait. The posterior probability for H4 is coloured when between 0.5 and 0.8 (blue) or larger than 0.8 (orange), where a greater number indicates greater support for the hypothesis. Proteins are ordered by the genomic position of their transcription start site. For each trait, GSMR and colocalisation results are labeled when they were significant in the PBMC results (see Fig 3) and were also significant after multiple hypothesis correction (FDR < 0.05, Benjamini–Hochberg) or had medium/high colocalisation support in the plasma results. Traits are as follows: **(A-B)** Angina, **(C-D)** Diabetes, **(E-F)** Heart Attack, **(G-H)** High Cholesterol, **(I-J)** Hypertension.

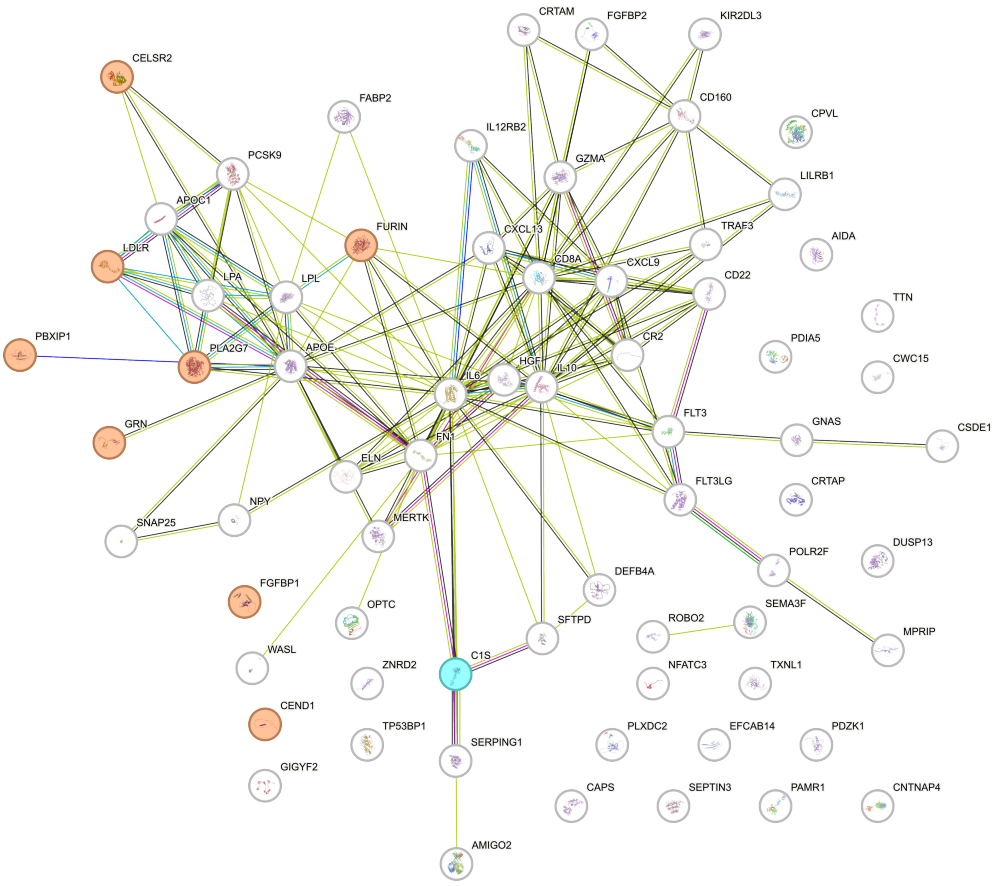


Fig. 5: **UKB GSMR and colocalisation results network for angina.** The String-DB generated network is presented for all non-MHC region proteins (those with a gene starting location outside of chr6:28477797-chr6:33448354) which were found by GSMR to have a significant relationship between the protein and angina (FDR < 0.05, Benjamini–Hochberg). Proteins, each represented by a node on the network, are shaded when the H4 colocalisation posterior probability, the hypothesis that the association between the set of SNPs provided with both the exposure and outcome through one shared SNP, is above 0.5. The posterior probability for H4 is coloured when between 0.5 and 0.8 (blue) or larger than 0.8 (orange), where a greater number indicates greater support for the hypothesis. Lines between nodes represent different forms of interaction evidence through: co-occurrence, co-expression, databases, experiments, gene fusion, neighbourhood and text-mining.

2.3.3 Comparison of cellular and plasma results highlights opposing directions of effect for 2 proteins.

Comparing GSMR results obtained using the proteomic data from PBMCs measured by mass-spectrometry, and those measured in plasma using the Olink Explore 3072 PEA platform [5], only proteins HLA-DRA, COMT and SWAP70 were significant (FDR < 0.05, Benjamini-Hochberg) in both PBMCs and plasma. An overview of results from both biobanks is presented in Fig. 6.

HLA-DRA is significantly associated with diabetes and hypertension in both the PBMC and plasma data (FDR < 0.05, Benjamini–Hochberg), showing a concordant direction of effect in plasma, as assessed by antibody affinity, and in PBMCs, as measured by mass spectrometry. Furthermore, HLA-DRA is implicated in angina and heart attack within plasma, and high cholesterol within PBMCs (FDR < 0.05, Benjamini–Hochberg; Table 2). The PBMC dataset utilised a single SNP in the GSMR analysis for HLA-DRA, which may have resulted in lower statistical power compared to the plasma data – where multiple SNPs were used in the GSMR procedure, potentially explaining the lack of significant association with angina and heart attack in this context.

COMT has opposing estimated direction-of-effect between the two protein measures for high cholesterol and hypertension. In the cellular GSMR results, a single SNP was used for COMT; rs165656, a SNP that is in complete LD with rs4680 ($R^2 = 1.0$, within the 1000 genomes, phase 3 GBR population [23–25]), a missense SNP within COMT itself. Despite not being the lead-variant, rs4680 is also strongly associated ($p = 1.84 \times 10^{-27}$ compared with rs165656 $p = 6.70 \times 10^{-28}$) with expression of COMT in PBMCs. A base change of G to A at rs4680 is associated with a V to M amino acid change. The observed direction-of-effect of a G to A substitution at this SNP is consistent with an observed reduction in the expression of COMT with the unmodified sequence, as one would expect from proteins measured by mass-spectrometry.

Given the nature of mass spectrometry proteomics, it is possible to assess the abundance of the measured peptide fragments that have contributed to evaluating protein abundance separately. From COMT (UniProt ID: P21964), there were 17 peptides detected. Of these, 10 passed quality controls (Methods 4.1.8) and were assessed. However when applying a Bonferroni correction for the number of peptide GWAS ($p < 5 \times 10^{-8}/29,639$), significant association was detected for one only. The sequence of that peptide – MVDFAGVK – contains the expected amino-acid change due to rs4680. This is consistent with this association being driven by the reduction in the abundance of the unmodified sequence of COMT, and does not speak to its overall abundance (with and without the amino acid change caused by the missense variant). In plasma, there were 32 and 33 SNPs used in the GSMR of high cholesterol and hypertension, respectively. When only the lead SNP from PBMCs was compared (that most significantly associated with the expression of COMT), the PBMC data had a consistent direction-of-effect as that observed in the plasma data (see Table 3). Therefore the conflicting direction of effect is not driven by altered direction-of-effect

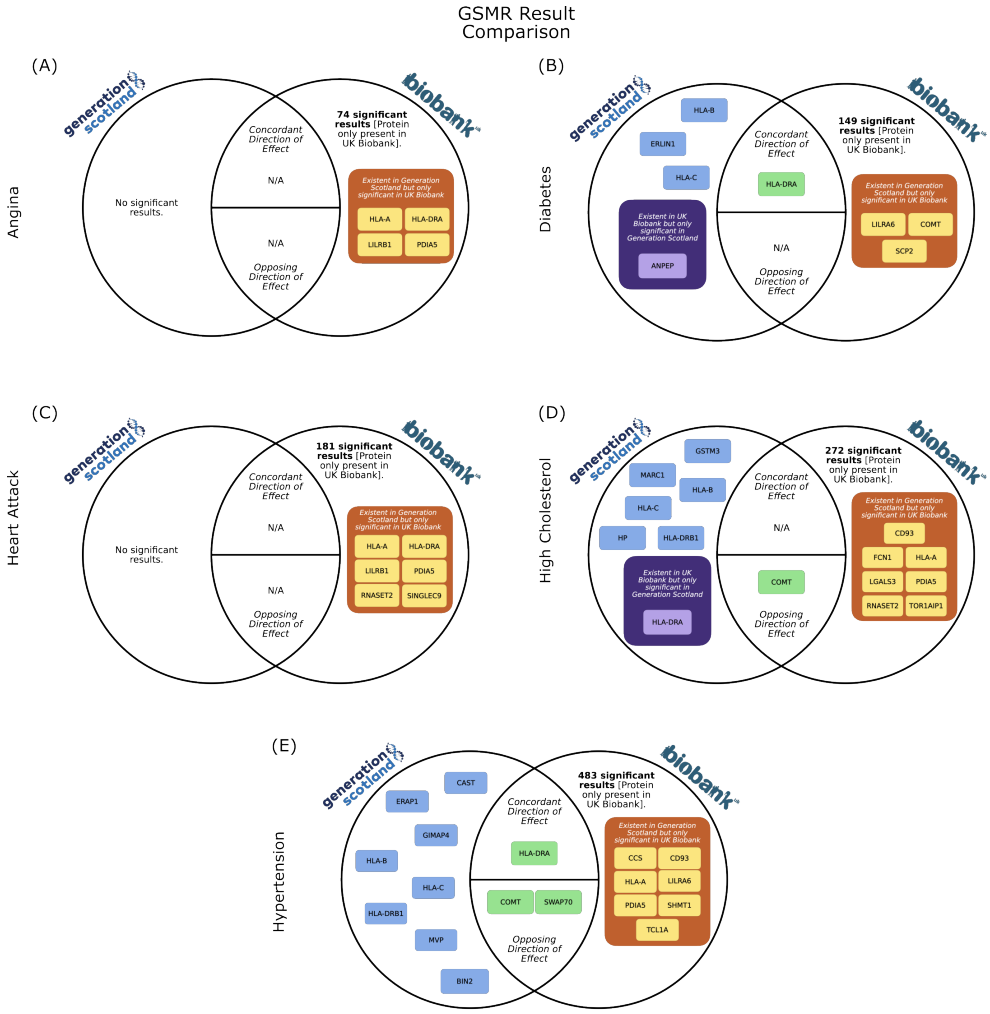


Fig. 6: **GS/UKB overlap and replication of UKB results in GS.** Venn diagrams illustrating the intersection between the PBMC (GS) and plasma (UKB) GSMR significant results (FDR < 0.05, Benjamini-Hochberg) in either site, and their overlap. In these diagrams, purple signifies results significant in GS but not in UKB, whereas yellow indicates significance in UKB but not in GS. Blue denotes proteins significant in GS but unmeasured in UKB. Green represents proteins significant in both biobanks, further categorised by either concordant or opposing direction of effect. The proteins are displayed in relation to various health outcomes: **(A)** angina, **(B)** diabetes, **(C)** heart attack, **(D)** high cholesterol, **(E)** hypertension. The results highlight instances of singular significance in one tissue and shared significance with varying effects between the tissues.

of the single SNP associated with the missense variant.

SWAP70 was also found to have an opposing direction-of-effect when assessed in PBMCs and plasma for hypertension. In the PBMC GSMR analysis, a single SNP was used for SWAP70; rs10770059, a SNP that is in complete LD with rs415895 (R^2 = 1.0, within the 1000 genomes, phase 3 GBR population [23–25]), a missense SNP within SWAP70 itself. Despite not being the lead-variant, rs415895 is also strongly associated ($p = 1.08^{-9}$; cf. rs10770059 $p = 4.74 \times 10^{-10}$) with expression of SWAP70 in PBMCs. A base change of C to G at rs415895 results in a Q to E amino acid change. The observed effect of a C to G substitution at this SNP is consistent with a reduction in the expression of SWAP70 with the Q containing sequence, as one would expect from proteins measured by mass-spectrometry. Again, for SWAP70, when the measured peptide fragments are assessed separately, only one – EQALQEAMEQLE-QLELERK – was found to contain a significant association after Bonferroni correction for the number of peptide GWAS (p-value $< 5 \times 10^{-8}/29,639$). The genetic sequence encoding this peptide contains rs415895, which is consistent with the association being driven by the reduction in the abundance of the unmodified sequence.

Fig. 7 contains locus zoom plots for the two non-MHC region proteins which were significant (FDR < 0.05, Benjamini–Hochberg) in the PBMC GSMR analysis, which were also significant within the plasma GSMR analysis. Table 3 provides details from the corresponding GWAS for the lead SNPs for the PBMC data.

Gene	UniProtID GS	$\beta_{\text{GSMR-GS}}$	$P_{\text{GSMR-GS}}$	$\beta_{\text{GSMR-UKB}}$	$P_{\text{GSMR-UKB}}$	N_{GS}	N_{UKB}
Angina							
HLA-A	P04439, P04439-2	3.75×10^{-4}	7.27×10^{-1}	1.34×10^{-3}	8.26×10^{-8}	1	285
HLA-DRA	P01903	-4.11×10^{-3}	1.62×10^{-1}	-9.98×10^{-4}	7.78×10^{-7}	1	204
LILRB1	Q8NHL6, Q8NHL6-2, Q8NHL6-3, Q8NHL6-4, Q8NHL6-5	2.25×10^{-4}	8.50×10^{-1}	-1.67×10^{-3}	3.70×10^{-5}	1	190
PDIA5	Q14554, Q14554-2	-2.11×10^{-3}	2.25×10^{-1}	4.58×10^{-3}	2.17×10^{-7}	1	47
Diabetes							
ANPEP	P15144	5.63×10^{-3}	1.30×10^{-3}	-1.97×10^{-3}	1.91×10^{-2}	1	49
HLA-DRA	P01903	-1.31×10^{-2}	1.26×10^{-3}	-3.71×10^{-3}	1.08×10^{-39}	1	151
LILRA6	Q6PI73	7.57×10^{-4}	5.14×10^{-1}	-1.44×10^{-3}	4.65×10^{-4}	1	101
COMT	P21964, P21964-2	-4.57×10^{-3}	1.52×10^{-2}	3.21×10^{-3}	2.19×10^{-3}	1	31
SCP2	P22307	-5.68×10^{-4}	6.55×10^{-1}	2.86×10^{-2}	1.75×10^{-3}	1	1
Heart Attack							
HLA-A	P04439, P04439-2	3.46×10^{-4}	7.43×10^{-1}	1.69×10^{-3}	6.17×10^{-14}	1	286
HLA-DRA	P01903	-5.57×10^{-3}	3.84×10^{-2}	-6.87×10^{-4}	1.17×10^{-4}	1	205
LILRB1	Q8NHL6, Q8NHL6-2, Q8NHL6-3, Q8NHL6-4, Q8NHL6-5	5.42×10^{-4}	6.36×10^{-1}	-1.17×10^{-3}	9.44×10^{-4}	1	189
PDIA5	Q14554, Q14554-2	7.94×10^{-4}	6.13×10^{-1}	3.68×10^{-3}	2.29×10^{-6}	1	48
RNASET2	O00584, O00584-2	-1.13×10^{-3}	4.40×10^{-1}	1.78×10^{-3}	6.91×10^{-4}	1	77
SIGLEC9	Q9Y336, Q9Y336-2	-2.11×10^{-3}	2.05×10^{-1}	9.63×10^{-4}	7.24×10^{-2}	1	59
High Cholesterol							
HLA-DRA	P01903	-1.84×10^{-2}	1.79×10^{-3}	-7.32×10^{-4}	2.45×10^{-2}	1	181
COMT	P21964, P21964-2	-1.01×10^{-2}	1.91×10^{-4}	4.85×10^{-3}	1.30×10^{-3}	1	31
CD93	Q9NPY3	-7.97×10^{-4}	7.39×10^{-1}	2.15×10^{-2}	6.90×10^{-4}	1	3
FCN1	O00602	-2.19×10^{-4}	9.08×10^{-1}	3.19×10^{-3}	1.27×10^{-3}	1	62
HLA-A	P04439, P04439-2	-5.42×10^{-4}	6.55×10^{-1}	1.34×10^{-3}	1.40×10^{-3}	1	274
LGALS3	P17931	9.54×10^{-3}	3.74×10^{-2}	-2.66×10^{-3}	5.17×10^{-3}	1	74
PDIA5	Q14554, Q14554-2	1.02×10^{-3}	7.06×10^{-1}	7.73×10^{-3}	2.01×10^{-5}	1	41
RNASET2	O00584, O00584-2	-3.06×10^{-3}	2.26×10^{-1}	3.32×10^{-3}	1.47×10^{-3}	1	74
TOR1AIP1	Q5JTV8	4.04×10^{-3}	1.51×10^{-2}	-4.35×10^{-3}	1.12×10^{-5}	1	71
Hypertension							
HLA-DRA	P01903	-2.45×10^{-2}	1.16×10^{-3}	-2.34×10^{-3}	4.56×10^{-8}	1	164
COMT	P21964, P21964-2	-1.38×10^{-2}	3.78×10^{-5}	8.45×10^{-3}	6.60×10^{-6}	1	30
SWAP70	Q9UH65	-5.53×10^{-2}	1.65×10^{-7}	3.99×10^{-2}	8.53×10^{-11}	1	4
CD93	Q9NPY3	-1.43×10^{-3}	6.24×10^{-1}	3.25×10^{-2}	5.23×10^{-4}	1	2
CCS	O14618	8.78×10^{-3}	1.76×10^{-2}	-5.72×10^{-3}	1.20×10^{-3}	1	38
CPVL	Q9H3G5	8.78×10^{-3}	1.77×10^{-2}	5.67×10^{-3}	5.45×10^{-14}	1	196
HLA-A	P04439, P04439-2	1.77×10^{-3}	1.83×10^{-1}	2.57×10^{-3}	1.89×10^{-6}	1	246
LILRA6	Q6PI73	8.22×10^{-4}	5.84×10^{-1}	-2.51×10^{-3}	4.41×10^{-4}	1	101
PDIA5	Q14554, Q14554-2	-1.71×10^{-4}	9.59×10^{-1}	6.81×10^{-3}	7.21×10^{-4}	1	40
TCL1A	P56279	9.74×10^{-3}	2.28×10^{-2}	-5.68×10^{-3}	1.77×10^{-3}	1	34
SHMT1	P34896	9.31×10^{-3}	9.58×10^{-3}	-3.66×10^{-3}	5.58×10^{-4}	1	81

Table 2: **Proteins instrumented in both biobanks, and significant in at least one**. GSMR analysis was conducted on both the PBMC and plasma data sets independently. Shown in the table are the p-values for the GSMR analysis for each gene that was instrumented in both sets of data, and then found to be significant in one of the data sets (FDR < 0.05, Benjamini–Hochberg). All genes are accompanied by the UniProtIDs for the isoforms considered in the PBMC data set. The number of instruments used in the GSMR processes are present at the end of the table for each set of data, with a subscript indicating the biobank which the data was obtained from.

Gene	Chr	Pos	Effect Allele	Other Allele	β_{GS}	SE_{GS}	p-value$_{GS}$	β_{UKB}	SE_{UKB}	p-value$_{UKB}$
COMT	22	19948863	C	G	-0.273	2.500×10^{-2}	6.700×10^{-28}	-0.445	7.643×10^{-3}	0.000
SWAP70	11	9770910	C	T	-0.148	2.370×10^{-2}	4.740×10^{-10}	-0.074	7.964×10^{-3}	1.709×10^{-20}

Table 3: Lead SNPs for significant proteins within the GS based GSMR which also existed within the UKB. Here we see two proteins which were significant in the PBMC based GSMR, which were also tested in the plasma. The SNP shown per instrument is the one which was selected as the lead SNP in the PBMC GSMR (smallest p-value). It is observed that, for both SNPs, the direction of effect, as measured by the respective GWAS, on the protein was the same.

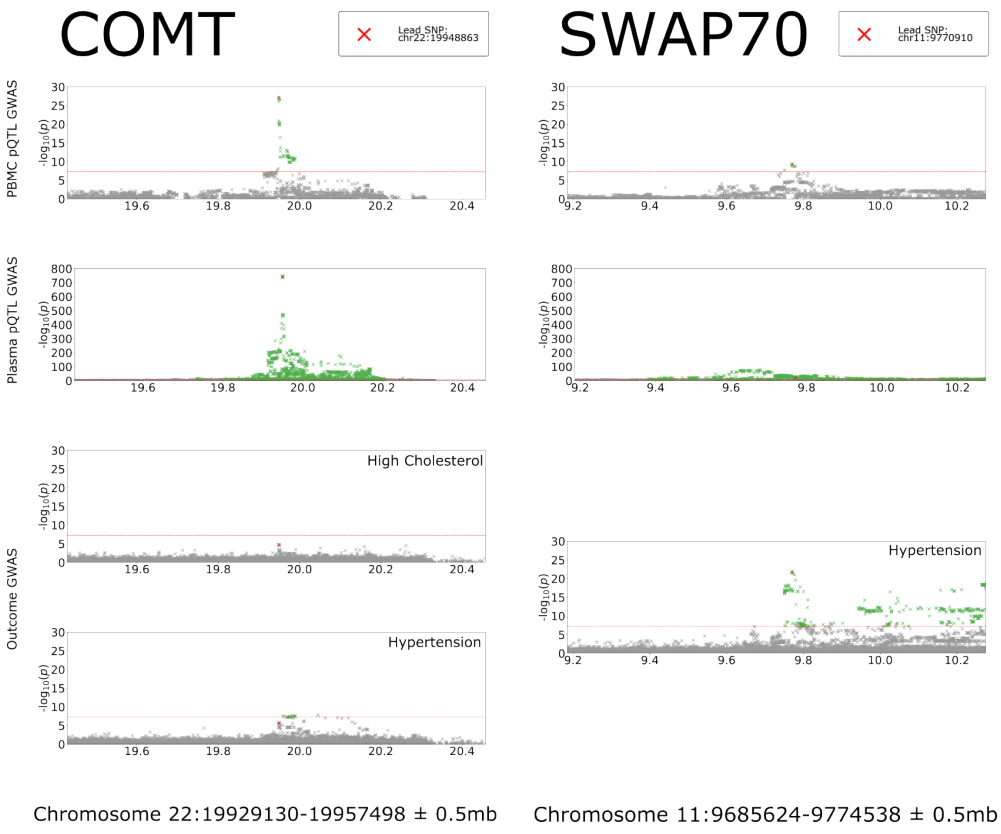


Fig. 7: **Locus Zoom of PMBC measured proteins significantly (FDR $< 5\%$) associated with an outcome, that were also significant in plasma.** For each protein we plot the gene encoding it ± 0.5Mb. Top row: peripheral-blood mononuclear cell pQTL results; middle row plasma pQTL results; bottom row outcome GWAS results. The most significant SNP in the peripheral-blood mononuclear cell protein GWAS is marked as a red cross. The red line is at genome-wide significance ($p = 5 \times 10^{-8}$).

3 Discussion

Cardiovascular disease (CVD) is a major worldwide cause of mortality and morbidity. In this study we have presented and compared cellular (PBMCs; GS) and plasma (UKB) protein links to CVD/CVD-related risk-factors (angina, diabetes, heart attack, high cholesterol and hypertension). We applied generalised summary-data-based MR (GSMR) and Bayesian colocalisation ('coloc') to provide a robust list of potential drug-targets for CVD and its risk factors. We examined the causal contribution of 152 proteins in peripheral-blood mononuclear cells (PBMCs) and 2,045 proteins in plasma on cardiovascular disease (heart attack and angina) and CVD-related risk-factors (hypertension, diabetes, and high cholesterol). By measuring proteins in PBMCs from 736 unique individuals – novel data reported here – we are able to comment on protein abundance in a cellular context.

Despite practical constraints limiting sample size, we have revealed significant associations for 16 isoform-specific protein groups identified in PBMCs. Among these, three proteins – HLA-DRA, COMT, and SWAP70 – were also significant (FDR < 0.05, Benjamini-Hochberg) in the plasma data. In plasma, 761 proteins showed significant associations with one or more CVD-related risk-factors. It is testament to the power of modern GWA studies that approximately 37% (761/2,045) of the proteins tested in plasma were linked to CVD, or its risk-factors. Notably, seven proteins – APOE, TP53BP1, DEFB4A/DEFB4B, CD22, GZMA, LTB, and PLXDC2 – were consistently associated across all five outcomes in plasma, highlighting their potential importance in CVD pathogenesis.

As with all MR studies, the validity of the conclusions drawn depends on the validity of the underlying assumptions (see Introduction). In order to ensure that the SNP (instrument) affects the protein (exposure), we filtered potential instruments based upon the observed strength of this association as part of our analysis pipeline and, in order to minimise the risk of horizontal pleiotropy, applied HEIDI outlier analysis as part of the GSMR analysis. However, we acknowledge that assumptions still remain.

Li *et al.* have previously examined the relationship between the UKB-PPP plasma protein data and CVD/CVD risk-related factors, identifying 221 putative proteins [3]. 132 proteins overlapped with our findings, including APOE, TP53BP1, CD22, GZMA, and LTB. The additional 629 targets that, to our knowledge, have not been previously identified in the plasma data set we identified – including DEFB4A/DEFB4B and PLXDC2 – are likely due to the different CVD/CVD risk-related outcome traits selected and differing MR analysis methods.

As proteins rarely act in isolation, we sought to identify biological processes common to those that were linked to CVD, and CVD-related risk-factors, by GSMR in plasma. Multiple protein sets were found to be enriched; including those related to current primary and secondary prevention treatment modalities, for example, statins ('HMG CoA reductase inhibitor use measurement'; Monarch; observed gene count 6;

background gene count 17; FDR 0.0285) in angina, and antiplatelets ('Platelet activation, signaling and aggregation'; Reactome; observed gene count 23; background gene count 86; FDR 0.0013) in 'heart attack'. Indeed, the mortality burden of CVD is also revealed here: proteins linked to angina were enriched amongst those linked to longevity (Monarch; observed gene count 9; background gene count 46; FDR 0.0187).

Of the proteins identified as significantly associated with the same outcome in both PBMCs and plasma, two exhibited opposite directions-of-effect (COMT and SWAP70). This may be due to statistical uncertainty, true biological differences, or a technical artefact. For example, interpretation of protein abundance in the context of protein altering genetic variation is exceptionally difficult, especially for affinity-based assays, and may be a wide-spread problem [26]. Regardless, protein location and activity are likely more important for phenotypic variation than abundance; information that needs to come from careful additional research.

Recent technological and logistical advances have made population-level plasma proteome analysis possible, to this we have added the cellular proteome of PBMCs. Instrumental variable analyses, like MR, offer a practical alternative to randomised control trials for screening large numbers of proteins in the search for potential therapeutic protein targets; positioning, with careful interpretation of the results, proteome-wide analysis as a cornerstone of future drug-discovery efforts, helping to prioritise targets for further investigation.

4 Methods

4.1 Generation Scotland PBMC preparation

Generation Scotland: Scottish Family Health Study (GS:SFHS) is a population and family-based cohort, recruited from the Scottish population, between 2006 and 2011 [27]. Peripheral blood mononuclear cell samples (PBMCs) were available from Generation Scotland for 862 individuals. PBMC samples were available for a sample of those recruited in Glasgow or Dundee with parents born in Scotland, no further selection criteria were applied.

4.1.1 Protein measurement

PBMCs were isolated from approximately 5ml whole blood that had been collected in acid-citrate-dextrose. Separation was performed using density gradient separation (Histopaque-1077; Sigma-Aldrich) by the European Collection of Authenticated Cell Cultures (ECACC) using a standardised protocol. PBMCs were then suspended in foetal-calf serum with 10% DMSO, frozen in a rate-controlled manner, and stored in liquid nitrogen until withdrawn for this study. Sample preparation and mass-spectrometry is described elsewhere [28], and summarised here for clarity.

Frozen PBMC cell pellets were retrieved from liquid nitrogen storage, thawed and washed in phosphate-buffered saline prior to lysis in 40μL of 6M guanidine hydrochloride with 100mM tris(hydroxymethyl)aminomethane ('lysis buffer') and sonicated. 2μl of sample was taken for a protein assay (Pierce BCA protein assay, Thermo-Fisher; Catalog number 23227). Protein concentration, per sample, was then standardised to a ceiling of 15μg of protein in 20μL lysis buffer. Samples were reduced and alkylated (1μL of 100mM tris-carboxyethylphosphine, 1μL of 200mM chloroacetamide, per sample) and heated for 5 minutes at 90-95°C.

Proteins were digested with LysC (300ng per sample, overnight digest, 37°C; Wako) and trypsin (150ng per sample, 4 hour digest, 37°C; Pierce, Thermo Fisher). Digestion was stopped by the addition of 16μL 10% trifluoroacetic acid (TFA), per sample.

Peptides were desalted on C18 stage tips. Stage tips were activated with 15μL methanol, washed with 50μL 0.1% TFA (pre- and post- sample loading), and eluted with 40μL 80% acetonitrile + 0.1% TFA. Following elution, samples were dried and resuspended in 14μL mass-spectrometry grade water. 9μL was acidified with 1μL 1% TFA and stored frozen prior to mass-spectrometry analysis.

Samples were prepared for mass-spectrometry in 20 batches (19 batches of 43 samples, and one batch of 45), and (where possible) one batch of sample repeats where the original preparation had failed (6 samples). Each batch included a pooled standard as well as within and between batch repeats. Peptide preparation was successful for 861 samples.

4.1.2 Mass-spectrometry

LC-MS/MS was performed on a Thermo Ultimate 3000 RSLC Nano UPLC coupled to a Thermo Fisher Q Exactive plus mass-spectrometer (Thermo Fisher). Samples were directly injected from a 96-well plate onto an Aurora UHPLC column from IonOpticks (Ion Opticks Pty Ltd). A Proxeon nano-spray ionisation source (Proxeon Biosystems) with a capillary temperature of 250°C and an optimised voltage of 1.4-1.7kV was used. A 120 minute gradient (2%-30% B in 110 min, 30%-45% B in the next 10 min; A=2% acetonitrile, B=80% acetonitrile, 0.5% acetic acid throughout; the composition was raised to 100% B in 7 minutes after the analytical gradient to wash the column, and total equilibration time was 20 minutes). Data-dependent acquisition (DDA), was run with a scan range of 350 to 1400 m/z using the Orbitrap at a resolution of 70,000 in profile mode. The top 24 parent peaks were selected for fragmentation. HCD fragmentation was performed with a normalised collision energy of 26 and spectra were acquired in centroid mode at a resolution of 17,500. Charge states accepted for MS2 were 2-5, peptide match was preferred and dynamic exclusion was 30 seconds.

A single injection was performed for all batches, and a second injection on a separate occasion for 75% (15/20) of them.

4.1.3 Data searching and annotation

Searches were conducted per batch. Data were processed using MaxQuant(v1.6.5.0) [29], matching against UniProt human (9606) reference proteome (5640) release 2019_11, canonical sequence and isoforms [30].

The search was performed for trypsin-digested peptides with up to two permitted missed cleavages. The fixed modification carbamidomethyl (C), and the variable modifications oxidation (M), acetylation (protein N-terminus), and methyl (KR;E) were included. Match-between-runs was allowed. MaxQuant built-in label-free quantification was not used. A false discovery rate of the peptide-spectrum matches of 0.01 was used for the search.

4.1.4 Imputation

Missing data were imputed, per batch, using IceR (version 0.9.12) [31]. Parameter settings used can be found in supplementary section 9.5. Once imputation was complete, quality control was performed. Batches with any of the following were removed: 1) an inconsistent gradient length (1 batch run) or low quality chromatography (1 batch run); 2) a median RT-deviation across the batch of > 1 min (4 batches runs); or 3) an FDR of $> 5\%$ of IceR quantification (when compared to 500 randomly chosen MaxQuant quantifications) (5 batch runs).

Twenty-four batch runs passed QC (17 unique batches: 10 with a single injection, 7 with two). A single injection of the repeats plate (6 samples) was included.

4.1.5 Normalisation

Intensity measurements were normalised using variance stabilisation normalisation in the R package 'vsn' (version 3.62.0) [32, 33]. The fit model ('vsn2') was computed using non-imputed intensities only (that is, prior to imputation with IceR) and applied ('predict') to both imputed and non-imputed data combined.

Label-free quantification (LFQ; Section: 'Extraction of Maximum Peptide Ratio Information') [34], was performed at the modification-specific peptide level and at the isoform-specific protein group level. An isoform-specific protein group was defined as a set of isoform-specific SwissProt proteins. Subsets not merged into supersets. The peptides contributing to an isoform-specific protein group were all those that mapped specifically to it.

In total, 1,068 mass-spectrometry runs, from 736 unique individuals, were included in the GWAS analyses.

4.1.6 pQTL assessment (GWAS Model)

A mixed-linear model was performed using the GENESIS (v2.22) R package [35]. The following were included as fixed effects 1) age, 2) sex, 3) age squared, and 4) 20 genetic principal components, and as random effects: 1) genetic relationship matrix (GRM), 2) peptide digest batch, 3) mass-spectrometry batch, and 4) biological replicates.

Genomic PCs and the GRM were created using the 'pcair' and 'pcrelate' functions from the GENESIS (v2.22) package and the 'snpgdsIBDKING' and 'snpgdsLDpruning' functions from the SNPRelate (v1.26) package [36]. Where the matrix denoting a random-effect was invariant it was removed from the model.

4.1.7 Quality control: Genotyping

The HRC v1.1 imputation was used [37], as described for Generation Scotland [38]. SNPs that had an info score of > 0.9 and a minor-allele frequency > 0.01 in the 861 sample for which peptide was successfully prepared were considered in the GWAS: in total, 6,804,554 variants.

4.1.8 Quality control: GWAS

Isoform-specific proteins/modification specific peptides that had been measured in 400, or more, unique individuals, and had a GWAS genomic inflation (lambda) of $<$ 1.1 were considered for further analysis. In total, 5,114 isoform-specific protein groups, and 29,639 modification-specific peptide GWAS were considered.

4.1.9 Missense variation mapping

Measured peptide fragments were first mapped to their corresponding Ensembl transcript ID by using BLAT v.37x1 [39] against the set of all peptides translated from Ensembl genes (available at: https://ftp.ensembl.org/pub/grch37/release-108/

fasta/homo_sapiens/pep/Homo_sapiens.GRCh37.pep.all.fa.gz; accessed: 26 October 2022). They were then mapped based on Ensembl transcript ID to their corresponding genomic coordinates, which were extracted from Ensembl using BioMart [23] (accessed: 26 October 2022).

Annotated SNPs catalogued by dbSNP [40] (dbSNP build 155; available at: https://ftp.ncbi.nlm.nih.gov/snp/latest_release/VCF/GCF_000001405.25.gz; accessed: 26 October 2022) were filtered for missense variants using BCFtools v.1.13 [41]. Only SNPs with MAF above 0.05 in the 1,000 Genomes GBR subpopulation were retained [24]. Missense variants located within the mapped genomic regions were obtained using the 'intersect' command from BedTools v.2.30.0 [42]. Unmapped peptides and peptides mapped to multiple genomic locations were removed.

4.2 UK Biobank plasma proteins

The UK Biobank is a large-scale cohort study based in the United Kingdom, with a range of different measurements and data for approximately 500,000 individuals. The UK Biobank Pharma Protemoics Project (UKB-PPP) is a pre-competitive biopharmaceutical consortium formed in order to produce the plasma proteomic profiles of 54,219 individuals from the UKB. They identified 2,941 unique proteins in their data. Data preparation, quality control and GWAS summary creation can be found elsewhere [5]. GWAS summary files for all 2,941 proteins were downloaded from the Synapse data storage platform [5, 43]. These files were used in the GSMR procedure.

As 6 proteins (CXCL8, IDO1, IL6, LMOD1, SCRIB and TNF) were measured on multiple protein panels, we only retained the cardiometabolic panels to ensure no proteins were duplicated, with the exception of SCRIB for which no cardiometabolic panel existed – the oncology panel was retained. This filtering gave 2,923 proteins. Following this we only retained SNPs for each protein which were significant below a Bonferroni corrected genome wide significance threshold ($p < 5 \times 10^{-8}/2923$). This left 2,415 proteins which had at least one instrument. Additionally 3 proteins were removed due to inability to map between GRCh38 and GRCh37 (PRSS2, PECAM1 and KIR2DL2). NTproBNP was removed due to its complexity when attempting to map. We considered proteins mapped to a single gene on an autosome only. This resulted in 2,344 proteins remaining for GSMR and colocalisation analyses.

4.3 LD reference

An unrelated set of individuals was selected from Generation Scotland [38] for use as an LD reference and reporting allele frequencies. The '–make-king-table' flag in PLINK was used to determine the relatedness of each of the samples. For sample pairs with a relatedness score of 0.025, or greater, one of the samples was removed to leave an unrelated subset of 6,862 samples. SNPs reported as significant and used in subsequent analyses from the PBMC GWAS were filtered to those with a minor allele frequency > 0.05 within this unrelated set of individuals.

Outcome	No. of Cases
Angina	14,399
Diabetes	21,105
Heart Attack	10,356
High Cholesterol	55,265
Hypertension	120,333

Table 4: Table presenting the various traits used in this study. All traits here are self-reported. UKB data field 20002.

This LD reference was used for the GSMR runs with both the PBMC and plasma data.

4.4 Outcome GWAS data

Summary-level data, from Gene Atlas [21], was used for the self-reported traits in Table 4 from UKB field 20002.

4.5 GSMR

GSMR results were generated using GCTA 1.94 [8, 44]. The pQTLs (PBMC or plasma) were used as the exposure, and each of the selected UK Biobank CVD, and CVD-related risk-factors as the outcome. The unrelated set of individuals from Generation Scotland was used as the LD reference.

In the GCTA software three items are specified, namely (i) direction of GSMR, we specify '0' for a forward direction assuming the exposure is directed to the outcome; (ii) maximum difference in allele frequency of each SNP between the data inputs, which we set to '1.0' to allow for any difference between the GWAS and the LD reference sample; (iii) minimum number of genome-wide significant and independent SNPs, set to '1' in our case to allow for any number of SNPs to be used. Explicitly, the following settings were used in GCTA: –gsmr-direction 0 –diff-freq 1.0 –gsmr-snp-min 1. The built-in HEIDI procedure was run before the GSMR procedure to identify and remove potentially pleiotropic SNPs. We adjust for multiple testing using the Benjamini–Hochberg procedure to bound FDR at 0.05 [20].

13 and 299 proteins were removed from PBMCs and plasma, respectively, as they had 0, 2, or 4 index SNPs following clumping and GSMR analysis was not completed.

4.6 Colocalisation analysis

'Coloc' version 5.2.2 [10], jointly tested five hypothesis under a Bayesian framework:

- H_0: No association with either the exposure or outcome.
- H_1: Association with only the exposure, but not the outcome.
- H_2: Association with only the outcome, but not the exposure.
- H_3: Association with both exposure and outcome, through two independent SNPs.
- H_4: Association with both the exposure and outcome, through one shared SNP.

Default setting were used within the 'coloc' R package provided by the author for the setting of the prior probabilities.

4.7 Enrichment Analyses

Enrichment analyses were performed using StringDB [22]. For each of the outcome traits, the protein names of all significant (FDR < 0.05) non-MHC (chr6:28477797-chr6:33448354) plasma proteins were input into StringDB, with the organism selected as 'Homo sapiens'. The standard protein mapping performed by StringDB, between the input set and its database, was manually checked and accepted. A bespoke background of all 2,045 post-filtering instrumented proteins from plasma was used.

The final networks of proteins identified by StringDB were of sizes 69, 144, 181, 271, 476 for the traits angina, diabetes, heart attack, high cholesterol and hypertension respectively.

5 Ethics

Generation Scotland participants provided written informed consent. Generation Scotland was granted Research Tissue Bank status by the East of Scotland Research Ethics Service committee (REC reference 15/ES/0040). This project was approved by Generation Scotland as reference GS18318.

6 Data Availability - Generation Scotland

Summary statistics for the significant (p-value $< 5 \times 10^{-8}/5114$; Bonferroni correction) results from the isoform-specific SwissProt protein group GWAS are available in Supplementary Table 4, and those with a p-value $< 5 \times 10^{-8}$ in the cis region ($\pm$ 1Mb) of the gene to which protein group was mapped are available in Supplementary Table 5.

Full GWAS summary statistics of the isoform-specific SwissProt protein groups and modification-specific peptides are available via Generation Scotland, and individual level data are available to bona fide researchers, subject to approval by the Generation Scotland data access committee: https://genscot.ed.ac.uk/for-researchers/access.

7 Acknowledgements

This study would not have been possible without the participants and staff of all the studies included for which we are very grateful.

8 Funding

CAO was supported by the EPSRC Centre for Doctoral Training in Mathematical Modelling, Analysis and Computation (MAC-MIGS) funded by the UK Engineering

and Physical Sciences Research Council (grant EP/S023291/1), Heriot-Watt University and The University of Edinburgh.

TR would like to acknowledge strategic funding grant BBSRC Institute Strategic Programme grants (BBS/E/D/20002172 and BBS/E/D/20002174).

SCH would like to acknowledge the BBSRC Strategic Programme Grant to the Roslin Institute (BB/P013732/1, BB/P013759/1).

JKB gratefully acknowledges funding support from a Wellcome Trust Senior Research Fellowship (223164/Z/21/Z), UKRI grants (MR/Y030877/1, MC_PC_20004, MC_PC_19025, MC_PC_1905, MRNO2995X/1, and MC_PC_20029), Sepsis Research (Fiona Elizabeth Agnew Trust), a BBSRC Institute Strategic Programme Grant to the Roslin Institute (BB/P013732/1, BB/P013759/1) and support of Baillie Gifford and the Baillie Gifford Science Pandemic Hub at the University of Edinburgh.

AK is supported by a Langmuir Talent Development Fellowship from the Institute of Genetics and Cancer, and a philanthropic donation from Hugh and Josseline Langmuir.

ADB would like to acknowledge funding from the Wellcome PhD training fellowship for clinicians (204979/Z/16/Z), the Edinburgh Clinical Academic Track (ECAT) programme.

CH and AR would like to acknowledge the MRC University Unit Programme Grant to the Human Genetics Unit (MC_UU_00007/10).

Mass-spectrometry proteomics was funded by the Wellcome Trust (204979/Z/16/Z), with supplemental funding from the MRC University Unit Programme Grant to the Human Genetics Unit (MC_UU_00007/10).

Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award "STratifying Resilience and Depression Longitudinally" (STRADL) Reference 104036/Z/14/Z).

9 Supplementary Materials

9.1 GS PBMC GSMR and colocalisation results

The full results for all PBMC measured proteins, for all five CVD/CVD-risk related outcomes are provided in Supplementary Table 1. The included columns are: 'trait', 'Gene name', 'Gene chromosome', 'Gene start (bp)', 'p', 'accept', 'H0', 'H1', 'H2', 'H3' and 'H4'. The trait refers to the outcome, whilst the gene name refers to the protein tested along with its corresponding known chromosome and starting location. The p-value represents the p-value of the GSMR procedure and accept determines whether the protein is accepted at a Benjamini–Hochberg FDR rate of 0.05. Finally H0-H4 relate to the posterior probabilities of the colocalisation tests.

9.2 UKB plasma GSMR and colocalisation results

The full results for all plasma measured proteins, for all five CVD/CVD-risk related outcomes are provided in Supplementary Table 2. The included columns are: 'trait', 'Gene name', 'Gene chromosome', 'Gene start (bp)', 'p', 'accept', 'H0', 'H1', 'H2', 'H3' and 'H4'. The trait refers to the outcome, whilst the gene name refers to the protein tested along with its corresponding known chromosome and starting location. The p-value represents the p-value of the GSMR procedure and accept determines whether the protein is accepted at a Benjamini–Hochberg FDR rate of 0.05. Finally H0-H4 relate to the posterior probabilities of the colocalisation tests.

9.3 UKB plasma GSMR networks

Fig. 8, 9, 10, and 11 show the String-DB network graphs for diabetes, heart attack, high cholesterol, and hypertension respectively, of all non-MHC region proteins identified by GSMR to have a significant relationship with the corresponding outcome, following a multiple testing correction (FDR < 0.05, Benjamini–Hochberg). Results on each network are shaded when the H4 colocalisation posterior probability, the hypothesis that both the exposure and outcome share a common associated SNP, is above 0.5. The posterior probability for H4 is coloured when between 0.5 and 0.8 (blue) or larger than 0.8 (orange). A greater number indicates greater support for the hypothesis. Lines between nodes represent different forms of interaction evidence through: co-occurrence, co-expression, databases, experiments, gene fusion, neighbourhood and text-mining.

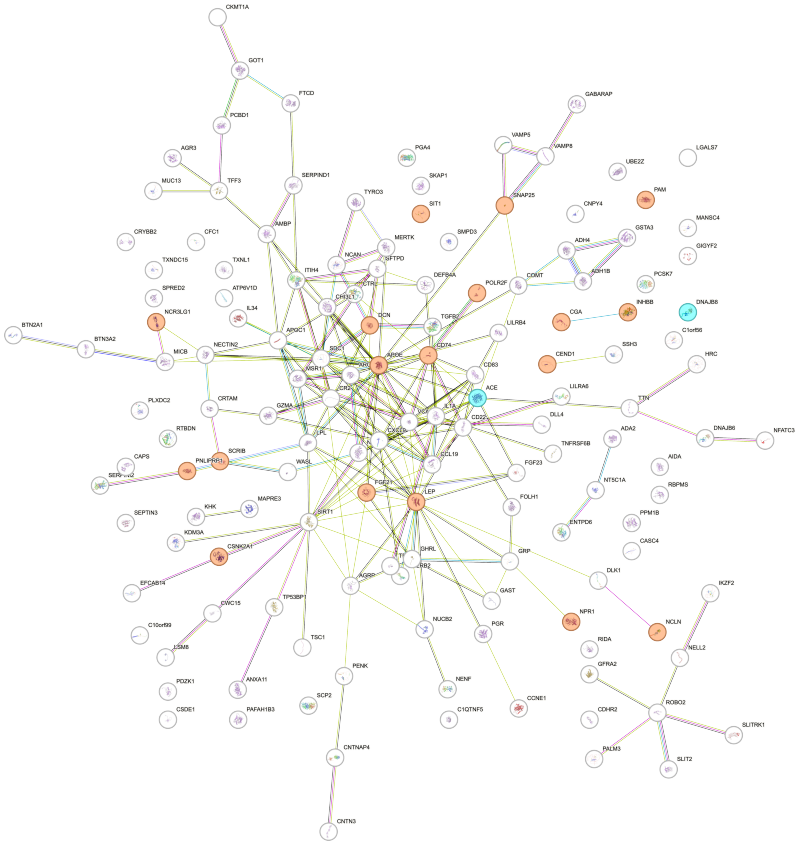


Fig. 8: UKB GSMR and colocalisation results network for the **diabetes** outcome from StringDB.

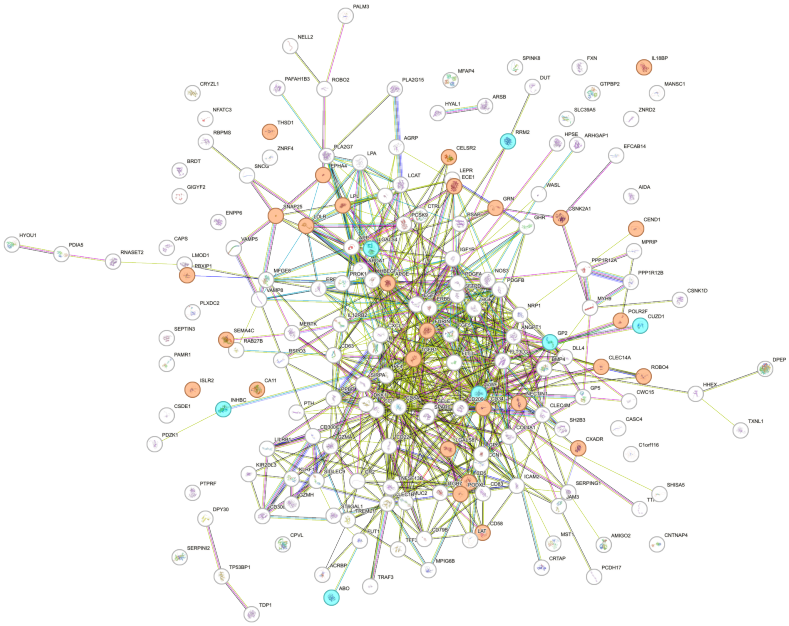


Fig. 9: UKB GSMR and colocalisation results network for the **heart attack** outcome from StringDB.

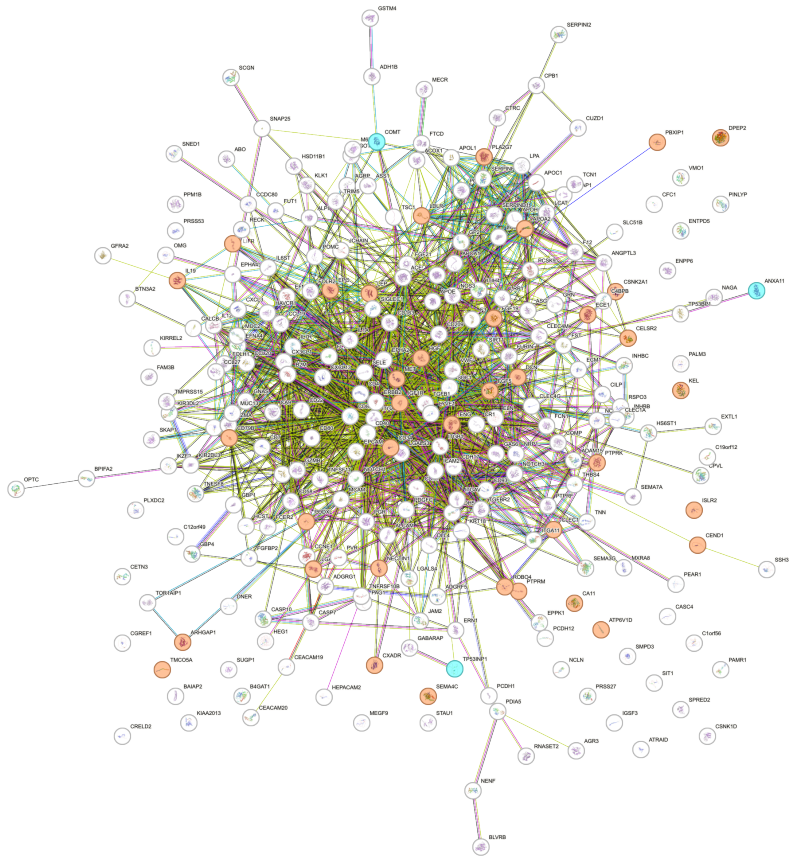


Fig. 10: UKB GSMR and colocalisation results network for the **high cholesterol** outcome from StringDB.

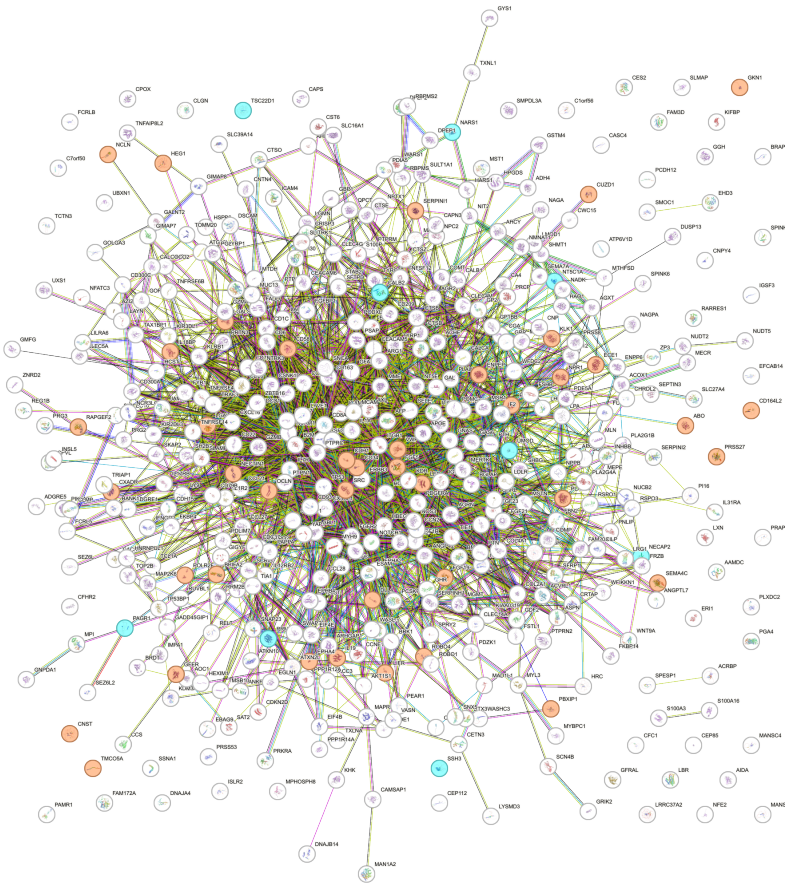


Fig. 11: UKB GSMR and colocalisation results network for the **hypertension** outcome from StringDB.

9.4 UKB plasma enrichment analyses

Full enrichment analyses using StringDB are provided as one table across all traits as Supplementary Table 3. The background for all of these networks were the 2,045 proteins instrumented in the plasma data set. No enrichment pathways were found for the Diabetes and Hypertension traits.

9.5 IceR Parameters

IceR was run with the following parameters: min_mz_window = 0.001; min_RT_window = 1; feature_mass_deviation_collapse = 0.002; only_unmodified_peptides = FALSE; align_unknown = FALSE; use_isotope_peaks = TRUE; peak_detection = TRUE; abundance_estimation_correction = TRUE; alignment_score_cut = 0.05; Quant_pVal_cut = 0.05; RT_correction = TRUE; mz_correction = TRUE; add_PMPs = FALSE; plot_peak_detection = FALSE; calc_protein_LFQ = TRUE; kde_resolution = 50; num_peaks_store = 5; MassSpec_mode = Orbitrap; use_IM_data = FALSE. Feature level output from IceR was used for subsequent analyses.

9.6 PBMC pQTL results

Supplementary Table 4 contains significant (GWA p-value $< 5 \times 10^{-8}/5114$; Bonferroni correction) results for isoform-specific, SwissProt only, protein groups measured in ≥ 400 unique individuals with a GWA genomic inflation (lambda) < 1.1. SNPs are limited to those with a minor allele frequency > 0.05 in a set of unrelated individuals from Generation Scotland (relatedness score of < 0.025 for all pairwise comparisons).

Included columns are as follows: Proteins_allpep_spIso: Isoform-specific SwissProt IDs in the group; N: Number of unique individuals included in the GWA study; lambda: Genomic inflation (lambda) of the GWA study; chr: Chromosome (GRCh37); pos: Position (GRCh37); Score: The value of the score function (as per the GENESIS v2.22 'assocTestSingle' function); Score.SE: The estimated standard error of the Score (as per the GENESIS v2.22 'assocTestSingle' function); Score.pval: The score p-value (as per the GENESIS v2.22 'assocTestSingle' function); Est: An approximation of the effect-size estimate for each additional copy of the effect allele (as per the GENESIS v2.22 'assocTestSingle' function); Est.SE: An approximation of the standard error of the effect-size estimate (as per the GENESIS v2.22 'assocTestSingle' function); PVE: An approximation of the proportion of phenotype variance explained (as per the GENESIS v2.22 'assocTestSingle' function); other.allele: Other allele in the GWA study; effect.allele: Effect allele in the GWA study; and effect.allele.frequency: The frequency of the effect allele from a set of unrelated individuals from Generation Scotland (relatedness score of < 0.025 for all pairwise comparisons).

Considering only SNPs located within the cis region ($\pm$1Mb) of the gene to which protein group was mapped, Supplementary Table 5 contains significant (GWA p-value $< 5 \times 10^{-8}$) results for isoform-specific, SwissProt only, protein groups measured in ≥ 400 unique individuals with a GWA genomic inflation (lambda) < 1.1. SNPs are limited to those with a minor allele frequency > 0.05 in a set of unrelated individuals

from Generation Scotland (relatedness score of < 0.025 for all pairwise comparisions). The results are limited to those protein groups that mapped to a single UniProtKB ID (SwissProt) and that that UniProt ID mapped to a single gene. Mapping was performed using https://grch37.ensembl.org/biomart/martview/ (accessed 25 May 2023).

The columns are as per Supplementary Table 4 with the following additions: Gene stable ID: Ensembl Gene ID of the mapped Gene; Gene start (bp): Gene start location; Gene end (bp): Gene end location; and Gene name: Gene name of the mapped Gene.

References

[1] Bhatnagar, P., Wickramasinghe, K., Wilkins, E. & Townsend, N. Trends in the epidemiology of cardiovascular disease in the UK. *Heart* **102**, 1945–1952 (2016). URL https://doi.org/10.1136/heartjnl-2016-309573.

[2] Tsao, C. W. *et al.* Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. *Circulation* **147**, e93–e621 (2023). URL https://doi.org/10.1161/cir.0000000000001123.

[3] Li, C. *et al.* Proteome-wide Mendelian randomization identifies candidate causal proteins for cardiovascular diseases. *medRxiv* (2023). URL https://doi.org/10.1101/2023.10.16.23297103.

[4] Abdellaoui, A., Yengo, L., Verweij, K. J. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 179–194 (2023). URL https://doi.org/10.1016/j.ajhg.2022.12.011.

[5] Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023). URL https://doi.org/10.1038/s41586-023-06592-6.

[6] Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol* **181**, 251–260 (2015). URL https://doi.org/10.1093/aje/kwu283.

[7] Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016). URL https://doi.org/10.1038/ng.3538.

[8] Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018). URL https://doi.org/10.1038/s41467-017-02317-2.

[9] Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet* **109**, 767–782 (2022). URL https://doi.org/10.1016/j.ajhg.2022.04.001.

[10] Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, 1–15 (2014). URL https://doi.org/10.1371/journal.pgen.1004383.

[11] Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021). URL https://doi.org/10.1038/s41586-020-03065-y.

[12] Kousathanas, A. *et al.* Whole-genome sequencing reveals host factors underlying critical COVID-19. *Nature* **607**, 97–103 (2022). URL https://doi.org/10.1038/s41586-022-04576-6.

[13] Pairo-Castineira, E. *et al.* GWAS and meta-analysis identifies 49 genetic variants underlying critical COVID-19. *Nature* **617**, 764–768 (2023). URL https://doi.org/10.1038/s41586-023-06034-3.

[14] Lousdal, M. L. An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology* **15**, 1 (2018). URL https://doi.org/10.1186/s12982-018-0069-7.

[15] Henry, A. *et al.* Therapeutic Targets for Heart Failure Identified Using Proteomics and Mendelian Randomization. *Circulation* **145**, 1205–1217 (2022). URL https://doi.org/10.1161/CIRCULATIONAHA.121.056663.

[16] Rasooly, D. *et al.* Genome-wide association analysis and Mendelian randomization proteomics identify drug targets for heart failure. *Nature Communications* **14**, 3826 (2023). URL https://doi.org/10.1038/s41467-023-39253-3.

[17] Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nature Communications* **11**, 3255 (2020). URL https://doi.org/10.1038/s41467-020-16969-0.

[18] Gill, D. *et al.* Mendelian randomization for studying the effects of perturbing drug targets [version 2; peer review: 3 approved, 1 approved with reservations]. *Wellcome Open Research* **6** (2021). URL https://doi.org/10.12688%2Fwellcomeopenres.16544.2.

[19] Gkatzionis, A., Burgess, S. & Newcombe, P. J. Statistical methods for cis-Mendelian randomization with two-sample summary-level data. *Genetic Epidemiology* **47**, 3–25 (2023). URL https://doi.org/10.1002/gepi.22506.

[20] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289–300 (1995). URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[21] Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nature Genetics* **50**, 1593–1599 (2018). URL https://doi.org/10.1038/s41588-018-0248-z.

[22] Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**, D638–D646 (2023). URL https://doi.org/10.1093/nar/gkac1000.

[23] EMBL-EBI. Ensembl LD Calculator (2023). URL https://grch37.ensembl.org/. [Online; accessed 05-10-23].

[24] Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL https://doi.org/10.1038/nature15393.

[25] Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015). URL https://doi.org/10.1038/nature15394.

[26] Pietzner, M. *et al.* Cross-platform proteomics to advance genetic prioritisation strategies. *bioRxiv* (2021). URL https://doi.org/10.1101/2021.03.18.435919.

[27] Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *International Journal of Epidemiology* **42**, 689–700 (2012). URL https://doi.org/10.1093/ije/dys084.

[28] Bretherick, A. D. *On the genetics of intermediate phenotypes and their utility.* Ph.D. thesis, Institute of Genetics and Cancer (2020). URL http://dx.doi.org/10.7488/era/866.

[29] Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**, 1367–1372 (2008). URL https://doi.org/10.1038/nbt.1511.

[30] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2022). URL https://doi.org/10.1093/nar/gkac1052.

[31] Kalxdorf, M., Müller, T., Stegle, O. & Krijgsveld, J. IceR improves proteome coverage and data completeness in global and single-cell proteomics. *Nature Communications* **12**, 4787 (2021). URL https://doi.org/10.1038/s41467-021-25077-6.

[32] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96–104 (2002). URL https://doi.org/10.1093/bioinformatics/18.suppl_1.S96.

[33] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A. & Vingron, M. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol* **2** (2003). URL https://doi.org/10.2202/1544-6115.1008.

[34] Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513–2526 (2014). URL https://doi.org/10.1074/mcp.M113.031591.

[35] Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019). URL https://doi.org/10.1093/bioinformatics/btz567.

[36] Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012). URL https://doi.org/10.1093/bioinformatics/bts606.

[37] McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016). URL https://doi.org/10.1038/ng.3643.

[38] Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Medicine* **9**, 23 (2017). URL https://doi.org/10.1186/s13073-017-0414-4.

[39] Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res* **12**, 656–64 (2002). URL https://doi.org/10.1101/gr.229202.

[40] Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (2001). URL https://doi.org/10.1093/nar/29.1.308.

[41] Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021). URL https://doi.org/10.1093/gigascience/giab008.

[42] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). URL https://doi.org/10.1093/bioinformatics/btq033.

[43] Synapse (2023). URL https://doi.org/10.7303/syn51364943.

[44] Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011). URL https://doi.org/10.1016/j.ajhg.2010.11.011.