medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Efficient candidate drug target discovery through proteogenomics in a Scottish cohort

Jurgis Kuliesius[1], Paul R.H.J. Timmers[1, 2], Pau Navarro[2,3], Lucija Klaric[2*], James F. Wilson[1,2*]✉

[1] Centre for Global Health Research, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland, UK. [2] MRC Human Genetics Unit, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, Scotland, UK. [3]The Roslin Institute, University of Edinburgh, Easter Bush Campus, Midlothian EH25 9RG, UK. ✉ jim.wilson@ed.ac.uk

Understanding the genomic basis of human proteomic variability provides powerful tools to probe potential causal relationships of proteins and disease risk, and thus to prioritise candidate drug targets. Here, we investigated 6432 plasma proteins (1533 previously unstudied in large-scale proteomic GWAS) using the SomaLogic (v4.1) aptamer-based technology in a Scottish population from the Viking Genes study. A total of 505 significant independent protein quantitative trait loci (pQTL) were found for 455 proteins in blood plasma: 382 *cis*- ($P < 5x10^{-8}$) and 123 *trans*- ($P < 6.6x10^{-12}$). Of these, 31 *cis*-pQTL were for proteins with no previous GWAS. We leveraged these pQTL to perform causal inference using bidirectional Mendelian randomisation and colocalisation against complex traits of biomedical importance. We discovered 42 colocalising associations (with a posterior probability >80% that pQTL and complex traits share a causal variant), pointing to plausible causal roles for the proteins. These findings include hitherto undiscovered causal links of leukocyte receptor tyrosine kinase (LTK) to type-2 diabetes and beta-1,3-glucuronyltransferase (B3GAT1) to prostate cancer. These new connections will help guide the search for new or repurposed therapies. Our findings provide strong support for continuing to increase the number of proteins studied using GWAS.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Genome-wide association studies (GWAS) are set to significantly impact the rapidly evolving domain of personalized medicine. This specialized area is dedicated to recognising the genetic and other variation among individuals, guiding the way towards precise risk evaluations and subsequent therapies tailored to distinct genetic (and other) profiles[1].

The field of proteomics adds another dimension to our understanding. The proteome participates in virtually every biological process, playing a critical role in both health and disease, with proteins serving as structural components, enzymes, signalling molecules, and more. Elucidating the genetic determinants of protein abundance is essential for understanding the complex interplay of genes, proteins, and their downstream effects on human physiology and disease susceptibility. This paves the way for the development of new prognostic markers[2], drug repurposing[3], and Precision Medicine[4] approaches.

Recent advances in technology have dramatically increased the number of proteins that can be quantified to over 10,000[5], with single proteomics studies exceeding sample sizes of 50,000[6]. The competing aptamer, antibody and mass spectrometry technologies differ in their mode of action, throughput, and the number of protein targets. Enabled by these innovations, GWAS, Mendelian Randomisation (MR) and comparison of local genetic architectures (colocalisation) are employed to unravel the complex relationships between circulating plasma protein levels and phenotypes such as disease risk. Leveraging naturally occurring genetic variants as instruments allows the assessment of the effects of lifelong exposure to altered protein levels on disease susceptibility, in a conceptually comparable way to performing a randomized control trial.

In this study, we present GWAS of 6432 proteins, representing one of the most comprehensive protein-centric association analyses to date[7]. We then use the resulting 31 *cis* associations from a little-studied 1533-protein subset and explore connections with medically relevant traits and diseases. Post-GWAS analyses resulted in 43 promising links between protein abundance and phenotype, 7 of which we highlight due to their potential therapeutic relevance for future in-depth follow-up. Our primary aim is to identify novel genetic loci associated with protein abundance, thereby uncovering new regulatory mechanisms, and shedding light on the interplay between genetic variants and disease, mediated by the human proteome.

Results

Discovery of pQTLs. We performed genome-wide association analysis of over 10.5 million imputed autosomal single nucleotide polymorphisms (SNPs) in 200 individuals using 7595 aptamers targeting 6432 blood plasma proteins measured with the SomaLogic v4.1 assay. Two different genome-wide significance thresholds were used: $p<5x10^{-8}$ *for cis* associations, defined as being within 1 Mb from the gene encoding the targeted protein) and $p<6.58x10^{-12}$ for *trans* associations, defined as all non-*cis* associations. After pruning SNPs with low allele frequency (MAF<0.05, due to low sample size), we identified a total of 1478 significant associations for the levels of 455 proteins**.** This corresponded to 505 independent "sentinel" SNPs, as determined by clumping (Fig. 1). 76% (382/505) of the sentinel SNPs were *cis* associations (Supplementary Table 1). In total, 333 proteins had only *cis* associations, 117 only *trans*, with 5 proteins having at least 1 *cis* and *trans* signal. The level of genomic inflation was well controlled for all 7595 aptamers, with the median λ value of 1.005, standard error 0.015.

The majority, 82% (412/505), of the independent, sentinel SNPs were associated with a single protein. 6 genomic regions were associated with 5 or more protein measurements (Fig. 2, vertical lines). These regions contained the *CFH*, *HRG*, *BCHE*, *ABO*, *VTN* and *APOE* genes, which have already been discovered as pleiotropic hubs or hotspots in previous proteomics studies[8, 9, 10].

Our analysis reveals 31 novel cis-pQTL, such as those for B3GAT1 (beta-1,3-glucuronyltransferase 1), DCC (Deleted in Colorectal Cancer netrin 1 receptor) and LTK (leukocyte receptor tyrosine kinase), allowing instrumentation of these proteins in Mendelian randomisation analyses.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

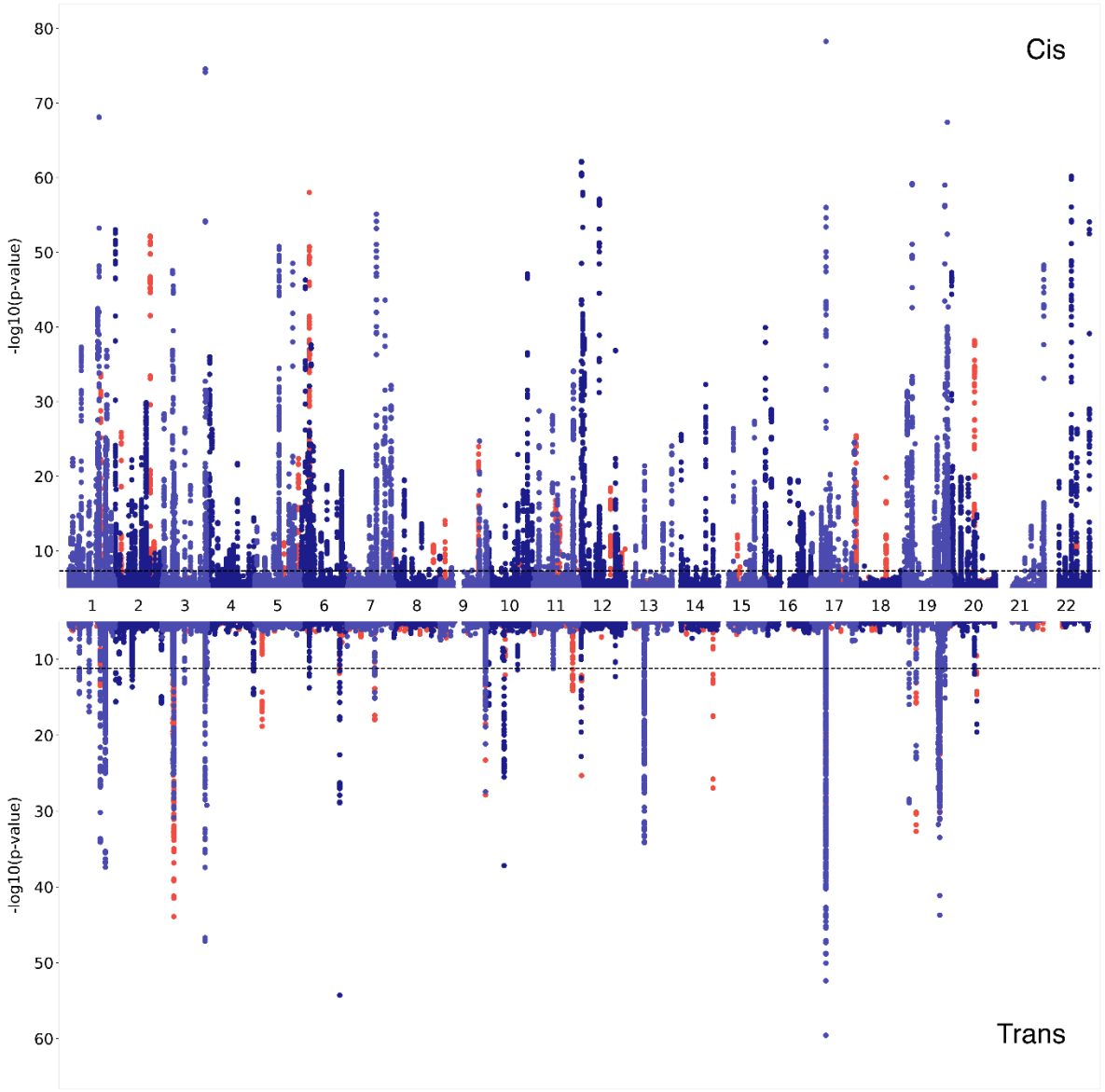

Fig. 1 Miami plot of the 455 proteins with genome-wide significant associations. In two shades of blue (denoting odd and even chromosomes) are the associations of proteins that have previously been reported in large-scale proteomics papers[10, 11, 12, 13, 14, 15] (Supplementary tables 2, 3). Proteins that were not found in aforementioned studies are coloured in red. *Cis* associations (upper part of the graph) are defined as being within 1 Mb of the transcription start site of the targeted protein, meanwhile all other associations are labelled as *trans* (lower part of the graph). Dashed lines represent the multiple-testing adjusted genome-wide significance thresholds, $p \leq 5x10^{-8}$ for *cis* and $p \leq 6.6x10^{-12}$ for *trans* associations.

Notably, some of the genome-wide significant associations were likely due to a degree of amino-acid sequence homology between the aptamer-targeted protein of interest and its paralogues. The strongest non-hub *trans*-pQTL detected in this study was on chromosome 6 (rs11155297, $p=5.4x10^{-55}$, Fig. 2, Supplementary table 1) was associated with FUCA1 (alpha-L-fucosidase 1) protein levels. However, the pQTL maps within the genomic region of *FUCA2* (alpha-L-fucosidase 2) on chromosome 1, the gene product of which shares 55% amino-acid sequence homology with the measured FUCA1 protein, when analysed with Clustal-O[16]. Meanwhile, there was no suggestive association detected within the *FUCA1 cis* genomic region ($p > 1x10^{-5}$). Hence, we conclude that in these examples

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

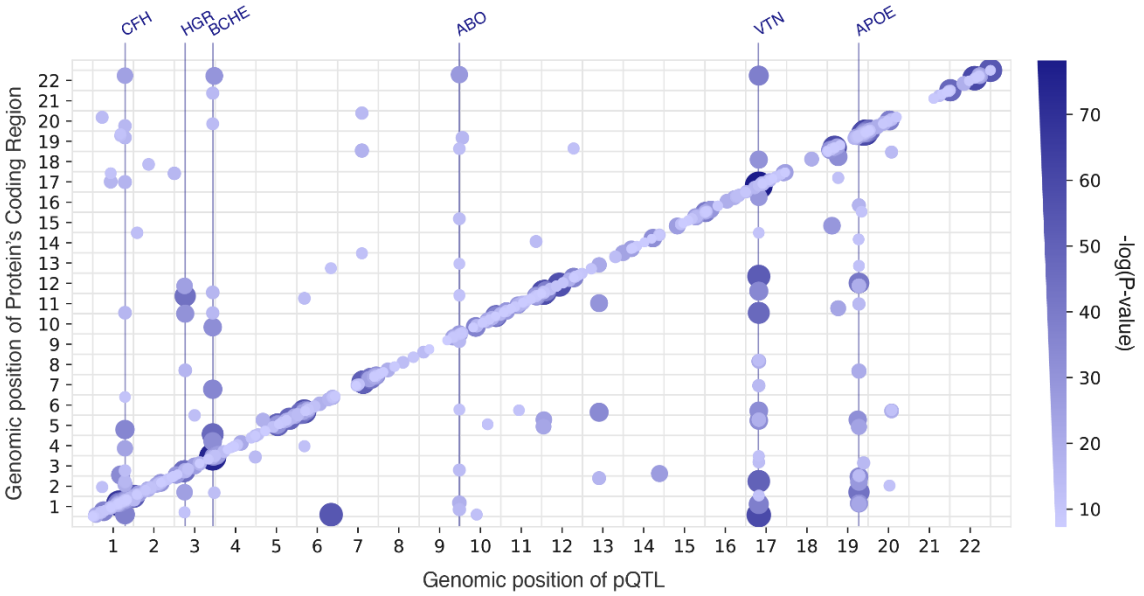

Fig. 2 pQTL distribution across the genome. 505 pQTL are plotted against the locations of the genes encoding those protein targets. Both the size and the colour of the points show the significance of the genetic association. Pleiotropic regions (hubs) are marked with vertical lines and the genes underlying the associated regions are named at the top. Only the pleiotropic regions with at least 5 *trans* associated proteins are shown as hubs. pQTL – protein Quantitative Trait Locus

the aptamer might not be able to distinguish between the two paralogous proteins and therefore the observed strong *trans* association might in fact be a *cis* association of a mislabelled protein.

Furthermore, two different aptamers targeting FCGR2B (Immunoglobulin G Fc Gamma receptor IIb) also had significant *cis*-associations in the vicinity of the nearby *FCGR2A* (Fc gamma receptor IIa) and *FCGR2C* (Fc gamma receptor IIc) coding regions, while exhibiting 73% and 89% amino-acid sequence homology with the encoded proteins, respectively (Supplementary Fig. 3). The connection between *FCGR2B* and *FCGR2C* may also be due to linkage disequilibrium (LD), with top SNPs, rs17413015 and rs61801833 (73 kb apart), exhibiting an LD r^2 value of 0.56, as evaluated with LDproxy[17]. In contrast, the sentinel SNPs, rs17413015 and rs4657041, in *FCGR2B* and *FCGR2A*, respectively, show a lower LD r^2 of 0.06, across the 166 kb between them.

As in other studies[9, 18], an inverse relationship between the minor allele frequency and the absolute effect size was observed for both *cis* and *trans* associations. Overall, *trans* associations displayed both smaller effect sizes and were less detectable at lower allele frequencies (Fig. 3A). Moreover, there was a strong influence of the distance from the transcription start site on the effect size of the *cis*-pQTL, with both the number of associations and their effect size rapidly decreasing outside the 0.15 Mb range (Fig. 3B).

We next annotated our sentinel pQTLs with the functional consequence information by considering the most severe consequence of any variant that is in r^2>0.8 with our sentinel variants. 32 out of 505 variants in this study have a high impact (loss-of-function) on the protein structure (e.g. stop/start gain/lost, frameshift). These were not distinguishable in their protein-level variances explained from the 229 protein-altering variant group of moderate impact (in-frame insertion/deletion,

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

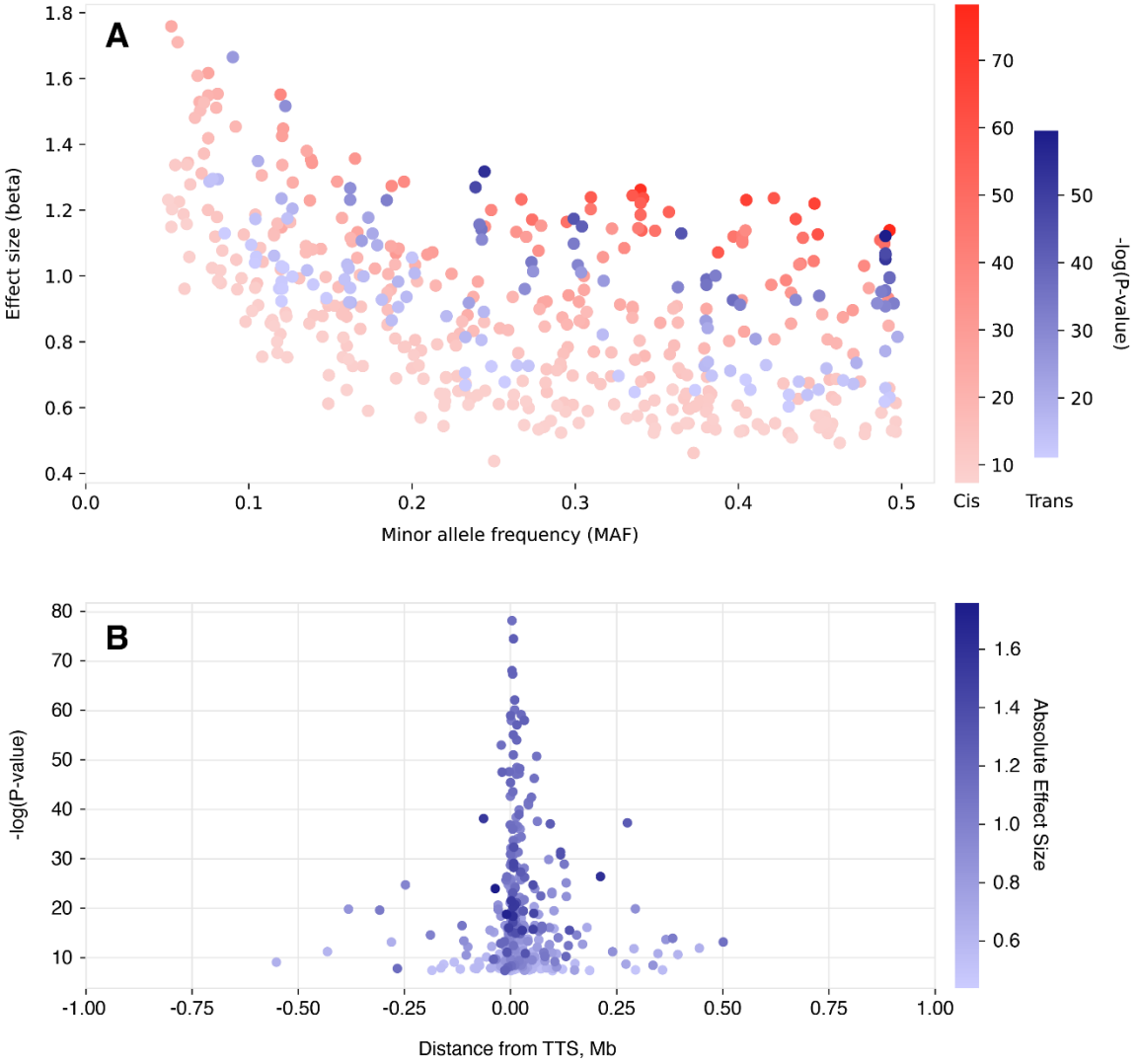

Fig. 3A SNPs with lower minor allele frequency tend to have larger absolute effect sizes. *Cis* (in red) and *trans* (in blue) associations are coloured separately to show the same L-shaped distribution. *Cis*-pQTL are both more detectable and have higher effect size at lower allele frequencies than *trans*-pQTL. Variants with minor allele frequency <0.05 were excluded from this analysis. Effect sizes (beta) were calculated on rank-based inverse normal transformed protein level data. **B. Distribution of *cis*-pQTL association p-values by distance to the transcription start site of the gene encoding the protein.** Closer to the TSS, pQTLs tend to be more significant, with a corresponding increase in effect size. TSS, Transcription Start Site, pQTL – protein Quantitative Trait Locus

missense), p=0.731. High and moderate impact genetic variants showed a significantly stronger effect on protein levels compared to 235 low-impact variants, with p-values of 2.21×10^{-3} and 2.98×10^{-8}, respectively. This effect was observed for both *cis* and *trans* associations (Supplementary Fig. 2).

Protein – disease links. To assess possible causal connections between plasma protein levels and disease outcomes or risk, we next performed bidirectional two-sample Mendelian Randomisation (MR). We focused on the proteins that have not yet been reported in large-scale proteomic MR studies, by cross-referencing the proteins targeted by the SomaLogic v4.1 assay with those measured with the SomaLogic v4.0 and Olink Explore 1536[15] assays. 1533 of the 6432 proteins quantified with the SomaLogic v4.1 assay were not measured in previous large-scale proteomics studies[10, 11, 12, 13, 14, 15]

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

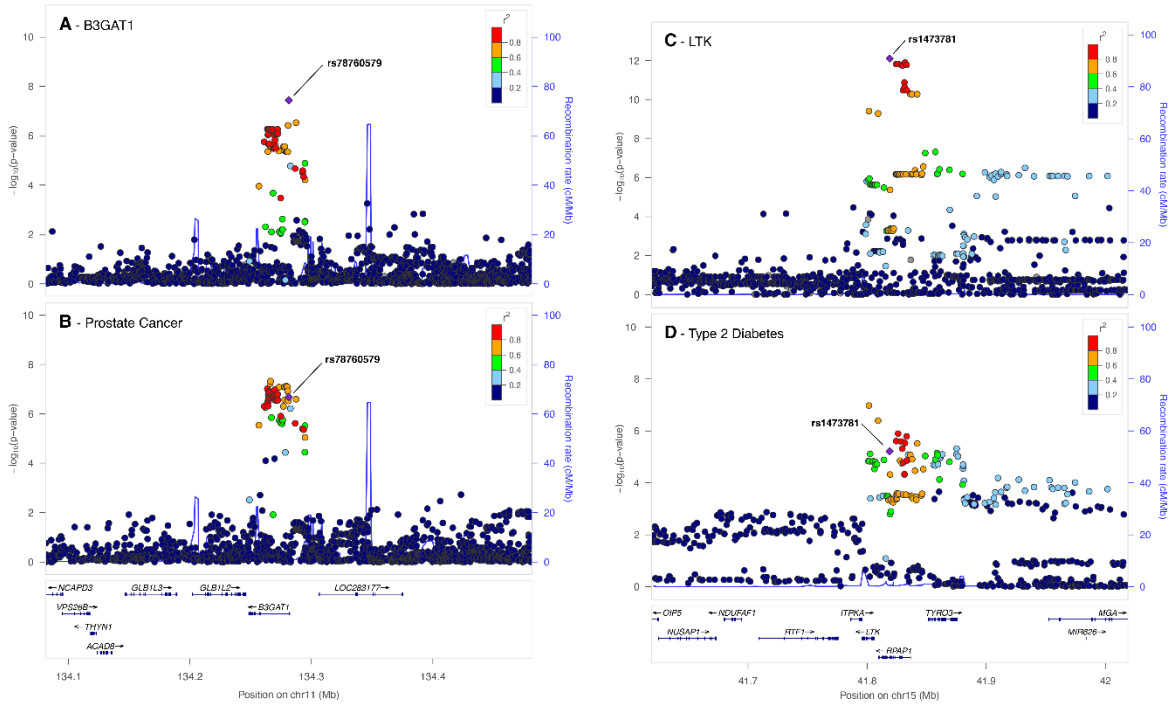

Fig 4. LocusZoom plots showcasing notable potentially causal protein level and disease associations. A, B - B3GAT1 and prostate cancer. LocusZoom plots show a visual colocalization comparison between the local association architecture for circulating B3GAT1 protein levels in our study (A) and that for ebi-a-GCST006085, a case-control prostate cancer study (B). The two studies colocalise in this locus with posterior probability of colocalisation (H4) = 0.91. **C, D - LTK and type 2 diabetes.** Colocalization comparison between the local association architecture for circulating LTK protein levels in our study (C) and that for ebi-a-GCST006867, a case-control type 2 diabetes study (D). The two studies colocalise in this locus with H4 = 0.95. The legends indicate Linkage Disequilibrium (LD) patterns between the sentinel SNP and others in the region.

(Supplementary tables 2, 3) and did not have associations uncovered through GWAS that could be used as instrumental variables.

We further restricted our selection of instrumental variables to *cis* sentinel pQTL, which are near the genes encoding their respective proteins. This approach was intended to mitigate the impact of pleiotropy on our findings, as a genetic locus in *cis* is less likely to influence multiple unrelated phenotypic traits, thereby simplifying the interpretation of resulting causal relationships. In addition, the focus on *cis* pQTL effectively reduced the multiple testing burden.

31 of the 1533 proteins had *cis* associations and were used as exposures in forward MR, while a curated list of diseases and risk factors (see Methods) from the MRC IEU OpenGWAS database[19] were used as outcomes. We found statistically significant associations (FDR < 0.01) for 17 (out of 31) proteins and 95 outcomes (149 protein-outcome pairs). Next, we assessed the possibility of reverse causality by running reverse MR, with the outcomes as exposures and the proteins as outcomes. There was no evidence of reverse causality (reverse MR p-value > 0.01) for any of the 149 significant protein-outcome pairs (Supplementary table 6).

Given that only a single instrumental variable was used for each of these proteins (their *cis*-pQTL), to further validate our findings we next performed a colocalisation analysis, using the "coloc" R package[20]. Colocalisation compares the local architecture of association for each trait in a Bayesian framework to assess whether the same underlying causal variant is responsible for the association with protein levels

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

cis pQTL	Gene	Colocalising MR Outcome	Coloc PP.H4	GWAS -log(p)	GWAS Effect size	MR -log(p)	MR Effect size
rs78760579	*B3GAT1*	Prostate cancer	0.91	7.4	-0.80	6.7	-0.080
rs1473781	*LTK*	Type 2 diabetes	0.95	12.1	0.69	5.2	0.054
rs10931931	*NIF3L1*	Early age-related macular degeneration	0.81	11.2	0.88	5.7	-0.11
rs13258747	*NTAQ1*	Total Testosterone	0.88	10.8	-0.62	5.5	-0.02
rs72941336	*AAMDC*	Intrinsic epigenetic age acceleration	0.90	14.6	-0.99	5.0	0.22
rs1169084	*BCL7A*	Systolic blood pressure	0.85	9.7	0.68	5.5	0.23
rs56953556	*COMMD10*	Parental longevity (mother's attained age)	0.90	11.1	-1.34	5.4	-0.016

Table 1. Noteworthy cis-pQTL associations with colocalising Mendelian Randomisation outcomes. The table showcases causal associations identified in this study between pQTL, protein levels and diseases or health outcomes that may be of medical interest. pQTL – protein Quantitative Trait Locus, MR – Mendelian Randomisation. Coloc PP.H4, posterior probability of colocalisation. Proteins represented: B3GAT1 – Beta-1,3-Glucuronyltransferase 1, LTK – Leukocyte Receptor Tyrosine Kinase, NIF3L1 – NGG1 Interacting Factor 3 Like 1, NTAQ1 – N-Terminal Glutamine Amidase 1, AAMDC – Adipogenesis Associated Mth938 Domain Containing protein, BCL7A – BAF Chromatin Remodelling Complex Subunit BCL7A, COMMD10 – COMM Domain-Containing Protein 10.

and the association with the outcome (disease risk). Of the 149 exposure (pQTL) – outcome (disease) pairs, 43 showed strong evidence of colocalisation (PPH4 > 0.8), suggesting direct genetic influences on disease via specific proteins, highlighting targets for future therapeutic intervention. 14 out of the 17 proteins with statistically significant forward MR associations had at least one association passing this sensitivity test (Supplementary Table 7). Among the 43 colocalising protein-disease outcome pairs, a few of the most interesting associations will be discussed in greater detail (Table 1).

The genes harbouring each of the pQTLs passing MR and sensitivity tests were checked in the genebass database of aggregate associations of rare variants, but no significant aggregate associations with any medical phenotype in the UK Biobank were found[21].

B3GAT1 and prostate cancer. We found that genetically decreased levels of B3GAT1 (CD57; beta-1.3-glucoronyltransferase 1) are associated with increasing risk for prostate cancer (MR effect size = -0.080, MR p = 1.9×10^{-7}). The reverse MR is not significant and the coloc posterior probability H4 is 0.91 (Fig. 4A, 4B). Notably, the *cis*-associated rs78760579 (effect allele G, effect size = -0.80, p = 3.7×10^{-8}) is in LD ($r^2 > 0.8$) with a recently reported variant, rs878987 (p = 2.7×10^{-6} in this study), detected as the lead variant in large case-control prostate cancer GWAS[22, 23] (p = 4.8×10^{-8}).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

LTK and diabetes. We have shown that rs1473781 is a *cis*-pQTL (effect allele A, effect size = 0.69, p = 7.75×10^{-13}) for LTK (leukocyte receptor tyrosine kinase). This SNP was shown to causally affect type-2 diabetes risk, mediated by LTK (MR effect size 0.054, MR p = 6.4×10^{-6}), there is no effect in reverse MR, and the association passed the colocalisation sensitivity test with a posterior probability of colocalisation (H4) of 0.95 (Fig. 4C, 4D).

NIF3L1 and Macular Degeneration. In the NIF3L1 (transcriptional activator NGG1 Interacting Factor 3 Like 1) GWAS, the *cis* sentinel SNP, rs10931931, (effect allele T, effect size = 0.87, p = 6.3×10^{-12}) was found to be causally linked to macular degeneration and decrease its risk (MR effect size -0.11, MR p = 2.0×10^{-6}), with no effect in reverse MR. This association passed the colocalisation test (H4 = 0.81).

Other notable associations. A *cis* sentinel SNP, rs13258747, (effect allele T, effect size = -0.62, p = 1.7×10^{-11}) for *NTAQ1* (N-terminal glutamine amidase 1; also known as WDYHV1), showed an effect on the levels of testosterone (MR effect size 0.026, MR p = 4.6×10^{-7}). Finally, we found a *cis* SNP, rs72941336, (effect allele T, effect size = 0.13, p = 2.7×10^{-15}) for AAMDC (Adipogenesis Associated Mth938 Domain-Containing protein) to be involved with intrinsic epigenetic age acceleration related to DNA methylation (MR effect size 0.22, MR p = 1.0×10^{-5}). Neither of these showed significant evidence for reverse causality and both colocalised (PPH4>0.8).

Discussion

The application of broad-capture proteomic profiling and linking that to genomics holds great potential to increase our understanding of biology and the mechanisms underlying various diseases. In this study we present the results of one of the most comprehensive proteomic GWAS, encompassing 6432 blood plasma proteins of the SomaScan v4.1 assay, of which 1533 have not been measured in any large-scale proteogenomic study to date. A total of 505 pQTL were identified for 455 proteins, 76% of which were in *cis* (within 1Mb of the gene encoding that protein). These results include unexplored associations with 58 proteins, 31 (53%) of which were categorised as *cis*. As for *trans* associations, we observed that 49% of them (60 out of 123) are linked to one of the *trans CFH*, *HRG*, *BCHE*, *ABO*, *VTN* or *APOE* hubs (Figure 2, Supplementary Table 1). Notably, this group of 6 genes has a marked enrichment in the regulation of coagulation and proteoglycan binding (Supplementary Table 8 GOEnrichment). A comparable enrichment pattern was also observed in another plasma proteomics study, utilising a different proteomics assay[24], though not noted in other tissues[25]. These findings suggest that these hubs may represent true biological effects or, alternatively, be artifacts related to the plasma sample preparation process.

Overall, we observed a higher proportion of *cis* signals compared to 10-31% in other large-scale proteomics studies[10, 12]. This is in accord with the idea that the smaller cohort size of this study (n=200) only allows for detection of proteomic GWAS signal with higher effect size, which are more likely to be in *cis*[9]. Furthermore, *cis* associations reach a plateau with increasing sample size, while the number of *trans* associations continues to rise[6].

Consistent with patterns noted in both molecular and whole-organism characteristics[26], we have identified an inverse correlation between allele frequency and effect size (Fig. 3A). Remarkably, this trend holds even though we omitted rare variants (MAF < 0.05) from our analysis, due to the constraints of a smaller sample size. Similarly, an inverse correlation between the *cis*-pQTL association strength (p-value/effect size) and distance to the transcription start site is observed (Fig. 3B), as in previous proteomics studies[9, 24]. Notably, only 2 out of 382 *cis*-pQTL identified herein fell outside the 500 kb range of the transcription start site for that protein. This is in agreement with previous theoretical research[27] and suggests interpreting pQTL outside of this range as *cis* with caution. Finally, we have observed that when the pQTL were categorized based on their predicted protein-altering properties, there were significant differences in the explained variance of protein abundances. Specifically, the variants predicted to have a high or moderate impact played a more significant role in contributing to changes in protein abundance, consistent in both *cis* and *trans* associations (Supplementary Graph 2A, B).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

In our analysis, we have identified notable cases with FCGR2B and FUCA1 where other proteins (FCGR2A, FCGR2C, and FUCA2, respectively) exhibit extensive similarity in their amino acid sequences. This high level of homology can lead aptamers to bind incorrectly to proteins they were not meant to target, creating false or misleading associations. With 70.5% of human genes having at least one paralog, 71.4% of which are located within 1 Mb[28], future research in proteomics should exercise caution when interpreting results due to misidentification and amino-acid sequence similarity for both *cis* and *trans* associations[29].

We then followed up the 31 *cis*-pQTL for the hitherto unexplored protein group with Mendelian Randomisation, incorporating a reverse MR filter to address reverse causation. This enabled us to assess the potential causal role of the levels of these proteins in disease, uncover new biology and potential drug targets. After extracting traits categorised as medically relevant (see methods, Supplementary Tables 4, 5) from the OpenGWAS[19] database, we have identified 149 significant protein-outcome associations passing both forward and reverse MR tests for 16 distinct proteins. Of these, a total of 43 colocalising protein-outcome pairs for 14 proteins and 39 medically relevant outcomes were observed. We highlight 7 protein-disease associations with newly discovered pQTL. Of these, B3GAT1, LTK and NIF3L1 have been researched more thoroughly in pre-existing literature and are discussed in more detail here.

The B3GAT1 protein is an enzyme that participates in the biosynthesis of glycosaminoglycans - long, unbranched polysaccharides found on the cell surface and in the extracellular matrix that play a part in cell signalling[30] and adhesion[31]. B3GAT1 knockdown in human tissues and mice experiments have been previously shown to moderate glycosaminoglycan structure, inhibiting spreading of tumour cells and increasing the survival of the animals[30]. Other studies have shown that human prostate luminal cell tumours continue expressing B3GAT1 (CD57) upon turning malignant and that this differentiation is among the most common prostate cancer phenotypes[32, 33]. Taken together, these results show that while B3GAT1 was shown to be involved in prostate cancer, this is the first time the causal role of the protein in the disease has been suggested.

The LTK (Leukocyte Tyrosine Kinase) protein is a receptor tyrosine kinase that belongs to the insulin receptor superfamily. Its specific function is not fully understood, but it is believed to play a role in neuronal development, immune response and cancer[34]. LTK variants also have evidence of being involved in lupus erythematosus, an autoimmune disease[35, 36]. Despite its name, LTK is primarily internally expressed in adipocytes[37]. Annotation of LTK via STRING[38] for protein-protein interactions has elucidated two potential pathways through which LTK may be participating in Type 2 diabetes. PIK3R1 (phosphoinositide-3-kinase regulatory subunit 1), a protein facilitating insulin signal transduction, mutations of which have been shown to trigger insulin resistance[39], has been shown to be essential in for LTK signal transduction through co-immunoprecipitation[40]. In contrast, IRS1 (insulin receptor substrate 1), a protein in which mutations are also well-documented to lead to type 2 diabetes[41], does not have biochemical data on interactions with LTK in humans. Instead, they are linked through text mining[36]. Finally, LTK was found to be associated with asthma, dermatitis, and cardiac arrhythmia in the FinnGen database, but not with type 2 diabetes[42].

NIF3L1 (NGG1 Interacting Factor 3 Like 1) is a little-studied protein that is broadly expressed in most tissues, including the retina[43], and involved in transcriptional regulation[37]. A recent study utilizing single-cell RNA sequencing has highlighted *NIF3L1* as actively transcribed across multiple retinal cell types[44]. However, their exclusive investigation into the non-coding region was unsuccessful in identifying a causal variant responsible for risk of age-related macular degeneration or any other eye disease tested. The complex Linkage Disequilibrium pattern observed (Supplementary Fig. 4) encompasses 7 other genes and further research is necessary to see if the sentinel variant rs10931931 identified in this study is indeed driving the causal relationship between NIF3L1 and age-related macular degeneration.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

As for drug repurposing, 8 out of 14 proteins identified to be linked to a medical trait in this study have approved or investigational drugs in DrugBank[45]. This illustrates the utility of pQTL in both discovering potential new drug targets and reimagining existing targets for different diseases.

Only a single other GWAS study has used the SomaLogic v4.1 panel, to date. The study focused on 466 chronic kidney disease patients of African American descent, performing Mendelian Randomisation on estimated Glomerular Filtration Rate and colocalisation with 778 phenotypes from the UK Biobank[8]. In contrast to the previous study, we have analysed a healthy European population and have broadened our research to encompass a comprehensive range of 3772 medically relevant traits available in the OpenGWAS database.

The main limitation of this study is its relatively low sample size, with proteomic profiling being performed in 200 individuals. While the sample size falls short of the standard typically required for GWAS with effect sizes of organismal-level traits[26], it has proven sufficient for analysing molecular traits. Specifically, we detect pQTL with large effect sizes, consistent with earlier molecular phenotype studies with maximum trait variance explained by the QTL reaching 22 – 39%[26, 46, 47]. Similarly, other proteomics studies using SomaLogic technology, albeit with a much larger sample size, were able to uncover pQTL that explained up to 75% of the variance in the observed protein levels[12].

Clearly, an increased sample size would enable discovery of variants with smaller effects. For instance, in this study only 4.6% of the proteins had a genome-wide significant *cis* association, while the largest proteomic studies report >90%[18]. Increasing the sample size would help not only with identifying smaller *cis* associations, but would also allow detection and pathway analysis of more *trans* associations, as their number increases and exceeds that of *cis* associations as the study power grows[15]. A better powered proteomics association study would in turn increase power in the downstream MR, allowing further causal protein-disease connection discovery. Nevertheless, studies of this size (n=200) are manifestly able to identify the strongest proteomics MR instruments. This approach to proteomics may serve as an alternative, yet complementary strategy to large-scale studies toward cost-effective instrumentation of more human proteins.

The other limitation of the study is that we focused exclusively on the common variation in the genome (MAF > 0.05), again due to sample size. As shown in Figure 3a, rarer variants are more likely to have a larger effect size and therefore a potentially increased risk for disease.

Finally, the sample used in this study is of exclusively European heritage, representing the predominant ancestry of populations where the majority of current discoveries have been made. As is the case with disease studies[48], deploying proteomic GWAS to populations of diverse continental ancestries will reveal further components of the genetic architecture, due to the different variants segregating. Indeed, our analysis suggests that proteomic GWAS of even relatively modest sample sizes from diverse populations may be a fruitful strategy to increase the number of proteins that can be used as instruments for MR.

Our findings provide strong support for continuing to increase the number of proteins under study in genome-wide association, so that many hitherto unstudied proteins will have genetic evidence available in drug development pathways. We further show that studies of modest sample sizes can reveal highly significant, novel pQTL. Deployment of this approach across multiple ancestries may be a cost-effective way to maximise the number of proteins for which genetic instruments are available. Finally, we identify new connections between proteins and disease risk which illuminate mechanisms and will help pave the way for new or repurposed therapies.

Methods

Study participants. The Viking Health Study – Shetland (hereafter VIKING1) is a geographically defined cohort with grandparents from the Shetland Isles, north of Scotland, which seeks to identify genetic factors influencing cardiovascular and other disease risk[49]. High levels of historical endogamy are reflected in the distinct gene pool of the VIKING1 cohort, as indicated by both common and rare genetic variants that set it apart from the rest of the British Isles and Europe[50, 51]. Recruitment of 2105

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

volunteers took place between 2013 – 2015. Each participant completed a health survey questionnaire and attended a 2-hour measurement clinic. Following that, overnight fasting blood samples were collected and frozen for downstream analyses**.** All participants gave informed consent, and the study was approved by the Southeast Scotland Research Ethics Committee, NHS Lothian (reference: 12/SS/0151).

A subsample of 200 participants was chosen for this study, with all 4 grandparents originating from the Shetland Isles and with minimal kinship to the rest of the n=200 sub-cohort. The highest genomic relatedness in this sub-cohort was 6%. Ages ranged from 19 – 91 (mean 52.6, s.e. 16.0), with females composing 53.5% of the sub-cohort.

Plasma samples and protein measurement. Following standard processing protocols (clotting, centrifugation and aliquoting), EDTA-treated fasting blood plasma samples were immediately frozen at -40°C and thereafter kept at -70°C for long-term storage. Frozen aliquots (500 µl) were shipped on dry ice to SomaLogic Inc. (Boulder, Colorado, USA) for proteomic analysis. All 200 blood plasma samples were measured with the SomaScan assay, version 4.1 (SomaLogic Inc.). The assay is specifically designed for human plasma analysis and measures protein levels using 7596 aptamers, covering 6432 unique human protein targets. Protein concentrations are measured in relative fluorescent units that are proportional to the actual amount of target protein in the plasma sample across a large dynamic range spanning 10 orders of magnitude in concentration[52].

Data quality control. Quality control of the assay results was performed both by SomaLogic and using in-house methods. In brief, SomaLogic quality control used hybridisation controls, median signal normalisation, and calibrator samples to account for the variability in target-aptamer hybridization and to allow a between-run comparison. This process altogether marked 289 aptamer measurements as inconsistent.

We performed further quality control based on the overlap (>5%) of observed signal data points between calibrator (plasma-free) and the actual samples. This may indicate a lack of sensitivity or specificity in the aptamer binding, considering that most (7526 out of 7596) aptamer signals do not overlap with the calibrator signal. Aptamers were also flagged for targeting non-human proteins or having no specified target at all. A total of 595 aptamers were marked this way, with an overlap of 33 with the SomaLogic quality control. The flagging system was only used descriptively to track the robustness of downstream analysis.
Protein abundances were then filtered by removing outliers outside the three interquartile range from the median raw measurement of each protein level. This resulted in GWAS having differing sample sizes with a median of 198 (min 174, max 200).

Protein-phenotype and technical covariate associations. Confounding factors are variables that can influence measured protein levels in the samples. These can be either inherent (e.g. sex, age), or technical (e.g. blood plasma sample storage time or co-ordinates on measurement plate). Technical artifacts, such as batch effects and some confounding factors can be accounted for to improve power and decrease the risk of false associations[53, 54]. For each aptamer, we performed multiple regression with the following covariates: biological sex, age, sample storage time in the freezer, season of the year when the plasma sample was taken, 96-well measurement plate number, row, and column. Forward stepwise selection with a likelihood ratio test (scipy 1.9.1, python 3) was performed to determine the influence of these covariates on each of the measured aptamer levels. All previously described covariates, except for sampling season had a statistically significant ($P<0.05/7596 = 6.58 \times 10^{-6}$, Bonferroni corrected for 7596 aptamers) effect on at least one of the measured protein levels. Therefore, all covariates except for sampling season were included in the GWAS model as fixed effects.

Similarly, principal components (PC) of the Genomic Relatedness Matrix (GRM) were used to correct for population stratification[55]. The GRM was created, and PC1-20 extracted for the whole 2005 individuals in the VIKING1 cohort using PLINK[56]. PCs were computed using parameters --maf 0.0025 and --nonfounders for autosomal variants.

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

For the 200 individual sub-cohort with proteomic data, the PC1-20 were analysed using multivariate regression one at a time via nested models for each aptamer. No principal components were significant ($P<0.05/(7596*20)=3.29x10^{-7}$, Bonferroni corrected for 7596 aptamers and 20 PCs) for all aptamer measurements and most of the aptamers exhibited a unique combination of significant PCs. However, due to the small sample size and limited degrees of freedom, in addition to the Scree plot having a clear Inflection Point (Supplementary Fig. 1), we decided to include only PC1-3 for all aptamers. As further post-GWAS analysis showed, this controlled for population stratification with a median genomic inflation control factor λ = 1.005, s.e. 0.015 (min 0.945, max 1.127) across all GWAS. λ of 44 out of 7596 GWAS performed in this study fell outside the accepted 0.95 - 1.05 range, with 2 yielding genome-wide significant pQTL. These pQTL were not considered for further post-GWAS analysis.

Genotyping and imputation. Individuals were genotyped using the HumanOmniExpressExome8 v1-2_A (Illumina) platform. Data was called with Beadstudio-Gencall v3.0 (Illumina). SNP genotype quality control (QC) was carried out using PLINK 1.9[56]. Only high-quality variants were selected: those with Hardy-Weinberg Equilibrium test $P>1x10^{-6}$, SNP call rate >98%, individual call rate >97%. In addition, we detected and removed Mendelian errors by using cohort pedigree information and removed monomorphic SNPs. After initial Quality Control, 611,836 autosomal SNPs remained in the dataset. We then imputed SNPs to the Haplotype Reference Consortium (HRC) panel v1.1 using the Sanger Imputation Service[57]. Imputed variants with low imputation quality scores (INFO < 0.4) were removed prior to downstream analysis.

Genome-wide association study. Due to skewness in the distributions, the relative protein detection levels of all 7596 aptamers were independently rank-based inverse normal transformed prior to GWAS. GRAMMAR residuals were computed by first regressing out the fixed effect covariates: sex, age, the previously described technical covariates and PC1-3 of the genetic relationship matrix (Supplementary Fig. 1), before modelling the relationship matrix as a random effect (GenABEL[58]). The resulting GRAMMAR residuals were then divided by the Gamma factor (GRAMMAR-Gamma)[citing https://www.nature.com/articles/ng.2410] and tested against genotypes using RegScan v0.5[59].

Following the GWAS, due to low sample size we removed variants with minor allele frequency (MAF) < 0.05.

Gene enrichment analysis. Gene enrichment analysis was conducted using the PANTHER 18.0 Overrepresentation Test[60, 61] with the GO Ontology database, version 2023-05-10[62, 63]. The analysed dataset contained *trans* hub genes identified in this study (*CFH*, *HRG*, *BCHE*, *ABO*, *VTN* and *APOE*), compared against a reference list of 20592 human genes. Fisher's Exact Test was utilized to identify significant overrepresentations, with False Discovery Rate (FDR) correction applied for multiple testing adjustments (Supplementary Table GOEnrichment).

***Cis*- and *trans*-associations.** An association was defined as *cis* if the associated SNP was within 1 Megabase (Mb) of the transcription start site of the gene encoding the protein that was targeted by that aptamer. Conversely, associations found outside this region or on another chromosome were defined as *trans* associations. To assess the proximity of a particular pQTL to the gene encoding the protein, we extracted the transcription start sites for the aptamer protein targets using Ensembl Biomart (build GRCh37), accessed in July 2022[64]. We used two different thresholds to define significance of the association: $5x10^{-8}$ for *cis*-pQTL, where there is prior expectation of an association, and $6.58x10^{-12}$ ($5x10^{-8}$/7596, number of aptamers) for *trans* associations.

Independent associations and LD proxies. To identify independent association signals (sentinel SNPs) we used clumping, as implemented in PLINK 1.9[65], with a window of ±250kb around the significant variants, and LD $r^2<0.01$ against a reference panel of a random subset of 10,000 unrelated genomically British individuals from UK Biobank[66, 67]**.** PLINK options used were –clump-kb 250 --clump-r2 0.01 --clump-p1 0.00000005 --clump-p2 0.0000025.

For proteins with multiple genome-wide significant *cis* associations, further filtering was performed since some genomic regions have long-range Linkage Disequilibrium (LD) patterns. Such associations

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

were not considered independent if their clumping windows overlapped and only the SNP with the lowest p-value was retained.

The LDproxy[17] API was then used to define LD proxies in the regions of the genome-wide significant results. Proxies were selected if they were in LD with associated variants, using European 1000 Genomes Project populations (CEU, TSI, GBR, IBS), with $r^2 > 0.8$ within a 1 Mb window. LDproxy and its associated databases were accessed in August 2022.

pQTL and their linked proxies were annotated using Variant Effect Predictor (VEP) for their consequences on the protein structure[64]. The consequences were divided into three categories – High, Moderate, and Low-Modifier. Only the most severe consequence(s) for each SNP was retrieved from the database and only the most severe consequence category was retained for the SNPs in LD with the sentinel variants in Supplementary Table 1.

The genome-wide significant results were then assessed for novelty. Proteins targeted in the most comprehensive proteomics studies using the SomaLogic v4.0 protein assay were not considered novel. Proteins from the Olink Explore 1536 panel were also treated as non-novel[15].

Subsequent analyses were only performed on the proteins that were not reported in the largest published proteomics study using the SomaLogic v4.0 assay[12], were not present in the Olink Explore 1536 panel and had at least one genome-wide significant *cis* signal in our study (Supplementary table 3).

Mendelian randomisation. Bidirectional two-sample Mendelian randomisation (MR) was performed to assess potentially causal associations between proteins (using *cis* sentinel SNPs as instrumental variables) and diseases and risk factors from the OpenGWAS[19, 68] database. The MR was performed using the TwoSampleMR (0.5.6) R package[68], with the Wald ratio method. TwoSampleMR proxy search was enabled if the sentinel SNP in our study was absent in the referenced studies with default parameters (1000 Genomes reference, rsq = 0.8, palindromes = yes, maf_threshold = 0.3). In the two cases when the TwoSampleMR integrated proxy search failed to find a proxy for the sentinel SNP, the next strongest association within LD>0.8 from our GWAS was supplied to the pipeline instead. Because of their pleiotropy and complex LD structure, we excluded SNPs that fall within the ABO (build GRCh37 chr9: 136.1311 – 136.1506 Mb) and HLA (build GRCh37 chr6: 2.9645 – 3.3365 Mb) regions. A subset of OpenGWAS[19, 68] datasets was used: ieu-a, ieu-b, ebi-a, ukb-b (database accessed April 2023). These datasets were further filtered for medically relevant traits and outcomes by employing a large language model (LLM), ChatGPT 4 (version July 20). The LLM was asked to categorise each outcome in the OpenGWAS database on whether they are of medical importance. After five repeats, the resulting categorisation (yes/no/ambiguous) for each trait was assigned a numeric value (1, 0 and 0.5, respectively). If the five-round sum was 2.5 or higher, the corresponding trait was designated as exhibiting medical relevance and was selected for further study. The traits with resulting scores of less than 2.5 were manually curated for medical relevance before being discarded from the study (Supplementary Tables 4, 5).

First, we performed forward MR, with protein levels being used as exposures and disease and risk factors as outcomes. To distinguish causal effects from reverse causality, the significantly associated (FDR < 0.01; equalling $p < 1.67 \times 10^{-5}$ in this study) 231 protein-disease pairs from the forward MR were used for reverse MR[69], where disease/risk factors were used as exposures and protein levels as outcomes using the "extract_instruments" function of TwoSampleMR[68]. We considered there to be no evidence for reverse causality if the reverse MR association was non-significant ($p > 0.01$).

Colocalisation. Robust associations passing the bidirectional MR sensitivity test were then tested for colocalisation using the R coloc 5.1.0 package[20, 70]. This method is used for analysis of two potentially related traits or diseases to investigate whether they share common underlying causal genetic variant(s), based on shared local genetic architectures of association. It involves testing five hypotheses: H_0 (no causal variants for either trait), H_1 and H_2 (causal variant for one trait only), H_3 (two independent causal variants, one for each trait), and H_4 (a single shared causal variant influencing both traits).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

A 300 kb window around each sentinel SNP was selected for the test with default priors using the "coloc.abf" function. Default package prior probabilities (priors) were used, with Hypotheses 1 and 2 being assigned $1x10^{-4}$ and Hypothesis 4 $1x10^{-5}$. MAF unfiltered summary statistics were used for the colocalisation tests. Colocalisation was declared for tests for which the posterior probability of colocalisation (H_4) > 0.8.

Association annotation. Clinically important associations passing all sensitivity tests described previously were manually assessed for recapturing known biology. The databases and tools used include the text-mining DISEASES platform[36], The Human Protein Atlas proteinatlas.org[71], GWAS and functional genomics database Open Targets Genetics[72, 73], drug database DrugBank[45], and the protein-protein interaction network database STRING[38].

Protein sequence alignment was investigated using clustalo v1.2.4[16].

Cumulative impact of multiple rare genetic variants for the protein and a linked disease outcome was investigated using genebass (gene-based association summary statistics)[21]. Burden (overall burden of multiple variants) and SKAT-O (incorporates weighting for rare variants) test outcomes were checked for phenotype-wide significant associations in all Burden sets – putative loss of function, missense and synonymous.

Data availability. The summary association statistics for all proteomic GWAS in this study have been deposited in the DataShare repository (available under https://datashare.ed.ac.uk/handle/10283/705) There is neither Research Ethics Committee approval, nor consent from individual participants, to permit open release of the individual-level research data underlying this study. The datasets generated and analysed during the current study are therefore not publicly available. Instead, the research data and/or DNA samples are available by managed access from accessQTL@ed.ac.uk on reasonable request, following approval by the QTL Data Access Committee and in line with the consent given by participants. Each approved project is subject to a data or materials transfer agreement (D/MTA) or commercial contract. The UK Biobank genotypic data used in this study as a LD reference panel were approved under application 19655 and are available to qualified researchers via the UK Biobank data access process.

Code availability. All analyses were conducted using publicly accessible software tools, which are detailed both in the main text and within the Methods section.

Data handling was done in Python 3. Main modules used include *pandas (v1.4), scipy (v1.4), numpy (v1.20)* for data transformation and statistical analysis, *requests (v2.22)* for data download via API, and *matplotlib (v3.2)* for creating graphs. Scripts will be made available upon request.

Acknowledgements. The Viking Health Study – Shetland (VIKING1) was supported by the MRC Human Genetics Unit quinquennial programme grant "QTL in Health and Disease" (U. MC_UU_00007/10). DNA extractions and genotyping were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. J.K. acknowledges the MRC Doctoral Training Programme in Precision Medicine (MR/N013166/1). L.K. was supported by an RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1). PN was supported by UKRI's Medical Research Council (MC_PC_U127592696, MC_PC_U127561128 and MC_UU_00007/10) and the BBSRC (BBS/E/RL/230001A). We would like to acknowledge the invaluable contributions of the research nurses in Shetland, the administrative team in Edinburgh and the people of Shetland. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Author contributions. Conception and design J.F.W, L.C.; Data curation and analysis J.K.; Scripting J.K., P.T.; Writing J.K.; Review and editing J.K., J.F.W., L.C., P.N., P.T.

Competing interests. P.T. and L.K. are currently employed by and have share options in BioAge Labs. The remaining authors declare no conflicts of interest.

References

1. Ashley EA. Towards precision medicine. *Nature Reviews Genetics* **17**, 507-522 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

2. Thadikkaran L, Siegenthaler MA, Crettaz D, Queloz P-A, Schneider P, Tissot J-D. Recent advances in blood-related proteomics. *PROTEOMICS* **5**, 3019-3034 (2005).

3. Reay WR, Cairns MJ. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet* **22**, 658-671 (2021).

4. Duarte TT, Spencer CT. Personalized Proteomics: The Future of Precision Medicine. *Proteomes* **4**, 29 (2016).

5. Somalogic. SomaScan® 11K Assay v5.0.) (2023).

6. Sun BB*, et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329-338 (2023).

7. Suhre K. A Table of all published GWAS with proteomics.) (2023).

8. Surapaneni A*, et al.* Identification of 969 protein quantitative trait loci in an African American population with kidney disease attributed to hypertension. *Kidney International* **102**, 1167-1177 (2022).

9. Sun BB*, et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

10. Pietzner M*, et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

11. Emilsson V*, et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769-773 (2018).

12. Ferkingstad E*, et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53**, 1712-1721 (2021).

13. Gudjonsson A*, et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nature Communications* **13**, 480 (2022).

14. Pietzner M*, et al.* Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nature Communications* **11**, 6397 (2020).

15. Sun BB*, et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv*, 2022.2006.2017.496443 (2022).

16. Madeira F*, et al.* Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res* **50**, W276-w279 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

17. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555-3557 (2015).

18. Macdonald-Dunlop E*, et al.* Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases. *medRxiv*, 2021.2008.2003.21261494 (2021).

19. Elsworth B*, et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv*, 2020.2008.2010.244293 (2020).

20. Giambartolomei C*, et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* **10**, e1004383 (2014).

21. Karczewski KJ*, et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics* **2**, 100168 (2022).

22. Schumacher FR*, et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936 (2018).

23. Conti DV*, et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nature Genetics* **53**, 65-75 (2021).

24. Repetto L*, et al.* Genetic mechanisms of 184 neuro-related proteins in human plasma. *medRxiv*, (2023).

25. Yang C*, et al.* Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nature Neuroscience* **24**, 1302-1312 (2021).

26. Yengo L*, et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704-712 (2022).

27. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? *Nucleic Acids Res* **44**, 6046-6054 (2016).

28. Ibn-Salem J, Muro EM, Andrade-Navarro MA. Co-regulation of paralog genes in the three-dimensional chromatin architecture. *Nucleic Acids Res* **45**, 81-91 (2017).

29. Hu Y, Ewen-Campen B, Comjean A, Rodiger J, Mohr SE, Perrimon N. Paralog Explorer: A resource for mining information about paralogs in common research organisms. *Comput Struct Biotechnol J* **20**, 6570-6577 (2022).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

30. Clausen TM*, et al.* Oncofetal Chondroitin Sulfate Glycosaminoglycans Are Key Players in Integrin Signaling and Tumor Cell Motility. *Molecular Cancer Research* **14**, 1288-1299 (2016).

31. Jeffries AR*, et al.* β-1,3-Glucuronyltransferase-1 gene implicated as a candidate for a schizophrenia-like psychosis through molecular analysis of a balanced translocation. *Molecular Psychiatry* **8**, 654-663 (2003).

32. Liu AY*, et al.* Analysis and sorting of prostate cancer cell types by flow cytometry. *The Prostate* **40**, 192-199 (1999).

33. Liu AY, Roudier MP, True LD. Heterogeneity in Primary and Metastatic Prostate Cancer as Defined by Cell Surface CD Profile. *The American Journal of Pathology* **165**, 1543-1556 (2004).

34. Zhang H*, et al.* Deorphanization of the human leukocyte tyrosine kinase (LTK) receptor by a signaling screen of the extracellular proteome. *Proceedings of the National Academy of Sciences* **111**, 15741-15745 (2014).

35. Li N*, et al.* Gain-of-function polymorphism in mouse and human Ltk: implications for the pathogenesis of systemic lupus erythematosus. *Human Molecular Genetics* **13**, 171-179 (2004).

36. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **74**, 83-89 (2015).

37. Stelzer G*, et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics* **54**, 1.30.31-31.30.33 (2016).

38. Szklarczyk D*, et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605-d612 (2021).

39. Kwok A*, et al.* Truncation of Pik3r1 causes severe insulin resistance uncoupled from obesity and dyslipidaemia by increased energy expenditure. *Mol Metab* **40**, 101020 (2020).

40. Ueno H*, et al.* The phosphatidylinositol 3' kinase pathway is required for the survival signal of leukocyte tyrosine kinase. *Oncogene* **14**, 3067-3072 (1997).

41. Rung J*, et al.* Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. *Nature Genetics* **41**, 1110-1115 (2009).

42. Kurki MI*, et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

43. Gautam P*, et al.* Multi-species single-cell transcriptomic analysis of ocular compartment regulons. *Nature Communications* **12**, 5675 (2021).

44. Wang SK*, et al.* Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. *Cell Genom* **2**, (2022).

45. Wishart DS*, et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074-d1082 (2018).

46. Bretherick AD*, et al.* Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS Genet* **16**, e1008785 (2020).

47. Klarić L*, et al.* Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci Adv* **6**, eaax0301 (2020).

48. Tsuo K*, et al.* Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *Cell Genomics* **2**, 100212 (2022).

49. Kerr SM*, et al.* An actionable KCNH2 Long QT Syndrome variant detected by sequence and haplotype analysis in a population research cohort. *Sci Rep* **9**, 10964 (2019).

50. Gilbert E*, et al.* The genetic landscape of Scotland and the Isles. *Proc Natl Acad Sci U S A* **116**, 19064-19070 (2019).

51. Halachev M*, et al.* Increased ultra-rare variant load in an isolated Scottish population impacts exonic and regulatory regions. *PLoS Genet* **15**, e1008480 (2019).

52. SomaLogic. SomaScan® Assay v4.1.). SL00000572 Rev 4: 2022-01 edn (2022).

53. Luo J*, et al.* Genetic regulation of human brain proteome reveals proteins implicated in psychiatric disorders.). Research Square (2022).

54. Wang Y*, et al.* The Association Between Glycosylation of Immunoglobulin G and Hypertension: A Multiple Ethnic Cross-Sectional Study. *Medicine* **95**, e3379 (2016).

55. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904-909 (2006).

56. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

57. McCarthy S*, et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016).

58. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294-1296 (2007).

59. Haller T, Kals M, Esko T, Mägi R, Fischer K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Briefings in Bioinformatics* **16**, 39-44 (2013).

60. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L-P, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* **31**, 8-22 (2022).

61. Mi H*, et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols* **14**, 703-721 (2019).

62. Ashburner M*, et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25-29 (2000).

63. Consortium TGO*, et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224**, (2023).

64. Cunningham F*, et al.* Ensembl 2022. *Nucleic Acids Res* **50**, D988-d995 (2022).

65. Shaun Purcell CC. PLINK 1.9.).

66. Bycroft C*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).

67. Landini A*, et al.* Genetic regulation of post-translational modification of two distinct proteins. *Nature Communications* **13**, 1586 (2022).

68. Hemani G*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

69. Zheng J*, et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122-1131 (2020).

70. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics* **16**, e1008720 (2020).

71. Uhlen M*, et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2024.01.29.24301808; this version posted February 6, 2024. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

72. Ghoussaini M*, et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research* **49**, D1311-D1320 (2020).

73. Mountjoy E*, et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics* **53**, 1527-1533 (2021).