









A genome-wide association study of serum proteins reveals shared loci with common diseases

Alexander Gudjonsson^{1,6}, Valborg Gudmundsdottir ^{1,2,6}, Gisli T. Axelsson ^{1,2}, Elias F. Gudmundsson ¹, Brynjolfur G. Jonsson¹, Lenore J. Launer ³, John R. Lamb⁴, Lori L. Jennings ⁵, Thor Aspelund ^{1,2}, Valur Emilsson ^{1,2,7} & Vilmundur Gudnason ^{1,2,7} ✉

With the growing number of genetic association studies, the genotype-phenotype atlas has become increasingly more complex, yet the functional consequences of most disease associated alleles is not understood. The measurement of protein level variation in solid tissues and biofluids integrated with genetic variants offers a path to deeper functional insights. Here we present a large-scale proteogenomic study in 5,368 individuals, revealing 4,035 independent associations between genetic variants and 2,091 serum proteins, of which 36% are previously unreported. The majority of both *cis*- and *trans*-acting genetic signals are unique for a single protein, although our results also highlight numerous highly pleiotropic genetic effects on protein levels and demonstrate that a protein's genetic association profile reflects certain characteristics of the protein, including its location in protein networks, tissue specificity and intolerance to loss of function mutations. Integrating protein measurements with deep phenotyping of the cohort, we observe substantial enrichment of phenotype associations for serum proteins regulated by established GWAS loci, and offer new insights into the interplay between genetics, serum protein levels and complex disease.

¹Icelandic Heart Association, Holtasmari 1, 201 Kopavogur, Iceland. ²Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland. ³Laboratory of Epidemiology and Population Sciences, Intramural Research Program, National Institute on Aging, Bethesda, MD 20892-9205, USA. ⁴GNF Novartis, 10675 John Jay Hopkins Drive, San Diego, CA 92121, USA. ⁵Novartis Institutes for Biomedical Research, 22 Windsor Street, Cambridge, MA 02139, USA. ⁶These authors contributed equally: Alexander Gudjonsson, Valborg Gudmundsdottir. ⁷These authors jointly supervised this work: Valur Emilsson, Vilmundur Gudnason. ✉email: v.gudnason@hjarta.is

The identification of causal genes underlying common diseases has the potential to reveal novel therapeutic targets and provide readouts to monitor disease risk. Genome-wide association studies (GWAS) have identified thousands of genetic variants conferring risk of disease, however, the highly polygenic architecture of most common disorders¹ implies that the genetic component of common diseases is largely mediated through complex biological networks^{2,3}. Identifying the causal mediators of mapped phenotype-associated genetic variation remains a largely unresolved challenge as majority of such variants reside in non-coding regulatory regions of the genome⁴. In fact, disease risk loci are enriched in regions of active chromatin involved in gene regulation^{5,6}. Thus, the integration of intermediate molecular traits like mRNA⁷ or proteins^{8–12} with genetics and phenotypic information may aid the identification of causal candidates and functional consequences. Furthermore, the phenotypic pleiotropy observed at many loci¹³ calls for a better understanding of the chain of events that are introduced by disease-associated variants. Genetic perturbations may for instance drive molecular cascades through regulatory networks⁸, most of which have not yet been fully mapped, or as a consequence of their phenotypic effects. Such downstream effects of genetic variants can be reflected in the molecular pleiotropy observed at some genetic loci, which could have significant ramifications for drug development, including assessing potential adverse effects¹⁴. For instance, many GWAS risk loci for complex diseases regulate multiple proteins in *cis* and *trans*, which often cluster in the same co-regulatory network modules⁸. Through the serum proteome we can gain a broad and well-defined description of the downstream effects of genetic variants, and their complex relationship with disease-relevant traits.

The human plasma proteome consists of proteins that are secreted or shed into the circulation, either to carry out their function there or to mediate cross-tissue communications¹⁵. Proteins may also leak from tissues, for example as a result of tissue damage¹⁵. It has been noted that a large subset of *cis*-to-*trans* serum protein pairs (i.e. proteins that are regulated by the same genetic variant in *cis* or *trans*, respectively) have tissue-specific expression but often involving distinct organ systems⁸, indicating that proteins in circulation may originate from virtually any tissue in the body. This suggests that system-level coordination is facilitated to a considerable degree by proteins in blood, which if perturbed may mediate common disease¹⁶. These observations, together with the accessibility of blood compared to other tissues, make circulating proteins an attractive source for identifying molecular signatures of disease in large cohorts.

Recent technological advances now allow for high-throughput quantification of circulating proteins, which has resulted in the first large-scale studies^{8–12} of protein quantitative trait loci (pQTLs) as recently reviewed¹⁷. Here, we present a large-scale proteogenomic study revealing thousands of independent genetic loci affecting a substantial proportion of the serum proteome, highlighting widespread pleiotropic effects of disease-associated genetic variation on serum protein levels. While our previous work reported associations to a restricted set of loci⁸, this is the first comprehensive GWAS for this number of serum proteins. A systematic integrative analysis furthermore demonstrates extensive associations between serum proteins and phenotypes that are regulated by the same genetic signals, adding further support to the therapeutic target and biomarker potential among proteins regulated by established GWAS risk variants.

Results

Identification of *cis* and *trans* acting protein quantitative trait loci (pQTLs). We performed a GWAS of 4782 serum proteins encoded by 4135 unique human genes in the population-based

AGES cohort of elderly Icelanders ($n = 5368$, Supplementary Data 1), measured by the slow-off rate modified aptamer (SOMAmer) platform as previously described^{8,18}. On average the genomic inflation factor was low (mean $\lambda = 1.045$, $sd = 0.033$) and of the 7,506,463 genetic variants included in the analysis (Supplementary Fig. 1), 269,637 variants exhibited study-wide significant associations ($P < 5 \times 10^{-8}$ /4,782 SOMAmers = 1.046×10^{-11}) with 2112 unique proteins, dubbed protein quantitative trait loci (pQTLs). In a conditional analysis using GCTA-COJO^{19,20} and validated using individual-level data (Supplementary Fig. 2), we identified 4035 study-wide significant associations between 2024 independent genetic signals in 772 loci (defined as genetic signals within 300 kb of each other) and 2091 unique proteins (Fig. 1a–c and Supplementary Data 2–4). Here we defined a genetic signal as a set of genetic variants in linkage disequilibrium (LD) that were associated with one or more proteins. For each associated protein, a genetic signal has a lead variant, defined as the genetic variant that is most confidently associated with the protein and with the lowest P -value (see Methods section for details). Among the 4035 independent associations, those in *cis* (signal lead variant within 300 kb of the protein-encoding gene boundaries, $n = 1415$) tended to have larger effect sizes than those in *trans* (signal lead variant >300 kb from the protein-encoding gene boundaries, $n = 2620$) (Supplementary Fig. 3A). Protein-altering variants (PAVs) in the gene encoding the protein target have the potential to alter the binding affinity of any targeted assay. For 336 (23.7%) of the 1415 *cis*-associations, the lead variant was, or was in LD ($r^2 > 0.8$) with, a PAV affecting the corresponding gene, thus potentially representing epitope effects (Supplementary Data 3). We found that almost half ($966/2091 = 46\%$) of all proteins with any independent genetic associations had more than one signal (Fig. 1b). Of those, 576 proteins (60%) had more than one independent signal within the same locus (Supplementary Fig. 3B) and 679 proteins (71%) had signals in distinct locations in the genome. The proteins with the largest number of associated loci were NOG (9 loci), TMCC3 (7 loci), and GRAMD1C, MANF and MMP7 (6 loci each).

The majority of genetic signals were only associated with a single protein (Fig. 1c), or 98% of *cis* signals and 73% of *trans* signals, and can as such be considered specific for the given protein based on a recently proposed classification of *trans*-pQTLs¹¹. Furthermore, we have previously shown that proteins regulated in *trans* by the same genetic variant often cluster in the same co-regulatory networks, sharing functionality and a disease relationship, although they may often differ in tissue origin⁸. However, as in previous studies^{8–11}, we identified numerous hotspots of *trans* protein associations, or more specifically 35 independent signals that were associated with 10 or more proteins each at a study-wide significant threshold (Fig. 1a, c). The largest of these *trans* hotspots represents the variant rs704, a missense variant within the Vitronectin (VTN) gene, which was associated with 595 proteins. Many of these *trans* hotspots are well established as such, including the VTN, ABO, APOE, CFH, and BCHE loci^{8–11}. Other notable *trans* hotspots included for instance variants in or near the Lipopolysaccharide Binding Protein (LBP) and Metastasis-Associated 1 (MTA1) genes. LBP is involved in the innate immune response to bacterial infections and MTA1 encodes a transcriptional coregulator upregulated in numerous cancer types and associated with cancer progression²¹. Of the 35 *trans* hotspots, 14 also affected protein levels encoded by proximal genes, thus acting in *cis* as well (Supplementary Data 3).

In contrast to the *trans* acting hotspots, we also observed genetic regions with high density of independent signals, each of which was not necessarily associated with many proteins. One such region stood out on chromosome 3 (Fig. 1a), where 29 independent signals were observed for a total of 54 proteins within a 300 kb window (Supplementary Fig. 4A), of which six

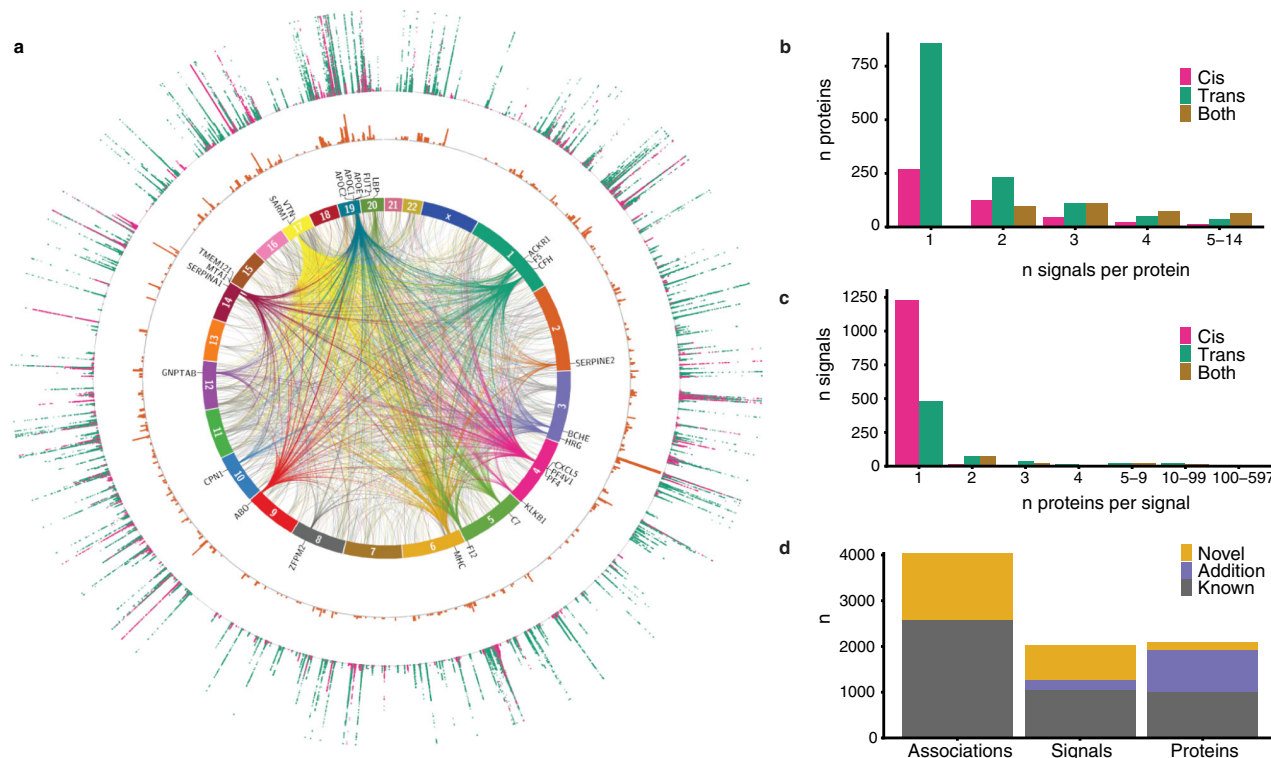


Fig. 1 A summary of the findings for genetic associations to 4782 proteins in serum. **a** Circos plot showing every study-wide significant variant-protein association from the protein GWAS (linear regression, $n = 5368$). The innermost layer shows links between independent signals (conditional and joint analysis, GCTA-COJO)^{19,20} and *trans* gene locations of associated proteins. *Trans* hotspots are colored by the chromosome they originate from. The second layer states the nearest genes to these *trans* hotspots. The third layer is a histogram of the distribution of the independent signals, where each bar represents the number of independent signals within 300 kb from each other, values ranging from 1 to 38. The outermost layer is a Manhattan plot for all proteins, P -values ranging from 1×10^{-11} to 1×10^{-300} (capped), colored by *cis* (pink), or *trans* (green). **b** Barplot showing number of proteins, binned by the number of associated independent signals, colored by *cis* (pink), *trans* (green) or both (mustard). **c** Barplot showing number of independent signals, binned by the number of associated proteins, colored by *cis* (pink), *trans* (green), or both (mustard). **d** Barplot showing the number of novel associations compared to similar large-scale genotype-protein association studies.

proteins (ADIPOQ, AHSB, DNAJB11, FETUB, HRG, and KNG1) were regulated in *cis*. Further analysis of this region demonstrated a sparse LD structure (Supplementary Fig. 4A), allowing for this high density of independent signals, and revealing a subcluster of 15 genetic signals affecting 32 proteins in various constellations (Supplementary Fig. 4B), that were enriched for Toll Like Receptor 7/8 cascade ($FDR = 4.8 \times 10^{-3}$) and MAP kinase activation ($FDR = 4.8 \times 10^{-3}$).

To define what proportion of the pQTLs identified in the present study can be considered novel, we compared all study-wide significant pQTLs with previously reported pQTL studies (Supplementary Data 5), including the recent exome-array analysis of the AGES cohort²². Of the 4035 independent associations detected in the current study, 1452 (36%) are considered novel based on this comparison (Supplementary Note 1, Fig. 1e, and Supplementary Fig. 5). Of the 2,024 independent genetic signals, 760 (38%) are novel, in the sense that they have not been reported to associate with any protein, and we find new protein associations for 204 known signals. Out of the 2091 proteins, 169 (8%) had no previously reported genetic associations in the comparison and we identified new genetic associations for additional 907 proteins.

We evaluated how well independent pQTLs reported by the INTERVAL study⁹ ($n = 3301$) replicated in our results and found 75.6% to be both directionally consistent and nominally significant ($P < 0.05$) (Supplementary Note 2 and Supplementary Figs. 6 and 7). This proportion furthermore increased to 93.9% when the *NLRP12* locus was excluded, a reported *trans* hotspot

that did not replicate in the AGES cohort (Supplementary Note 2 and Supplementary Figs. 6 and 7). This locus has in fact recently been identified as platform specific²³ and was suggested to be related to white blood cell lysis during sample handling. We similarly performed a lookup of the independent study-wide significant associations identified in our current study in the INTERVAL study summary statistics (Supplementary Note 2 and Supplementary Fig. 8). Of 2690 associations with information in the INTERVAL study we find that 94% are directionally consistent and 83% were both directionally consistent and nominally significant ($P < 0.05$). Of 645 associations defined as novel in our study (Supplementary Note 1) and with information available in the INTERVAL study, we again find a very high directional consistency between the two studies, or 90% of associations, and 64% are both directionally consistent and nominally significant ($P < 0.05$) in the smaller INTERVAL study.

Finally, with more individuals genotyped we revisited the GWAS of the serum protein co-regulatory network⁸, now represented by the first two eigenproteins of each protein module, and find that almost all the network modules are under strong genetic control (Supplementary Note 3).

Characterization of proteins by genetic association profiles.

Taking advantage of the broad coverage of the protein measurements in our study, to determine which protein characteristics can provide additional insights into the observed differences in genetic profiles for the measured proteins we compared

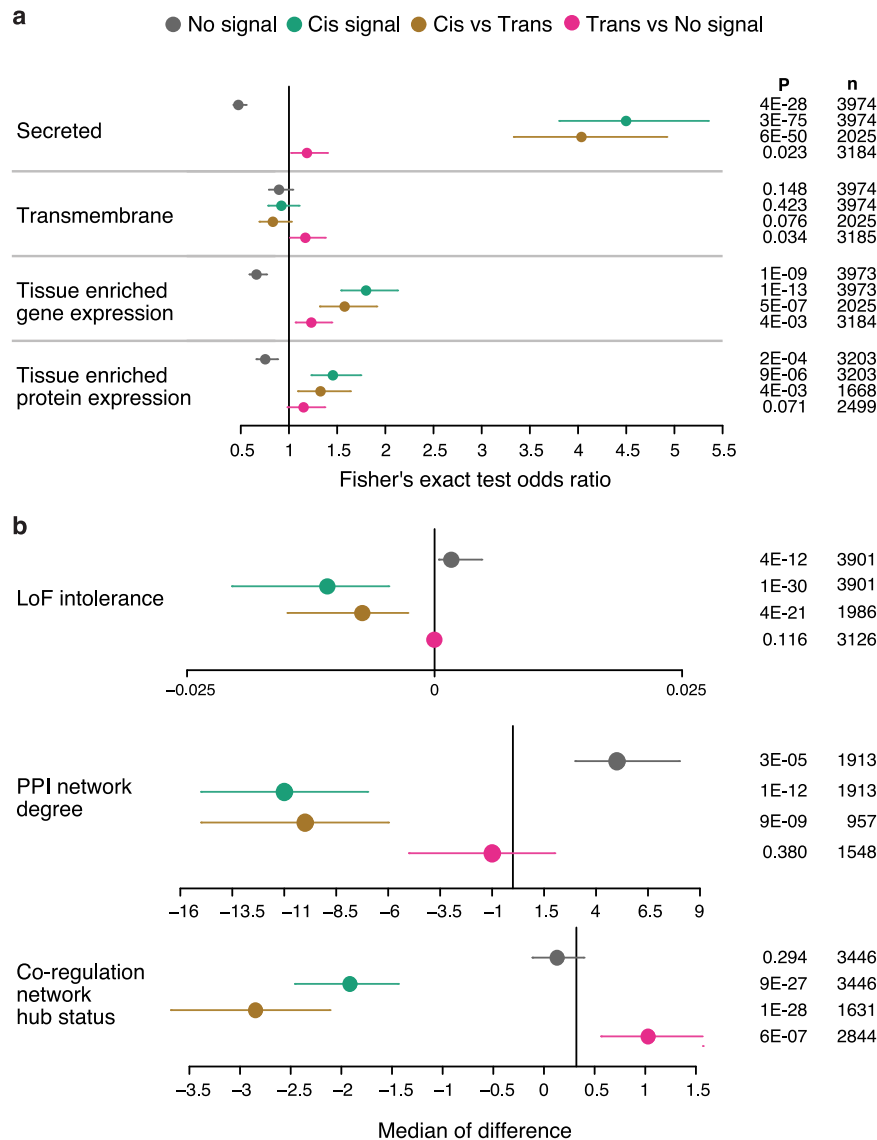


Fig. 2 Enrichment analysis comparing characteristics between proteins classified by types of genetic association signals. See Methods for definitions. **a** Fisher's exact test (two-sided) for comparing two classifications. Odds ratio estimates are presented with 95% confidence intervals. **b** Wilcoxon's rank-sum test (two-sided) for comparing classifications with continuous traits. Estimates of the median of the difference between values from the two classes are presented with 95% confidence intervals. *P*-values (two-sided) for significant enrichment of protein-phenotype associations are provided to the right.

characteristics such as tissue-enhanced gene²⁴ and protein²⁵ expression and protein localization²⁴ for proteins with genetic signals to those without any detected genetic effect. Moreover, we analyzed loss-of-function (LoF) intolerance²⁶ and hub status in two types of protein networks, i.e. the InWeb protein-protein interaction (PPI) network²⁷ and the serum protein co-regulatory network⁸, but pathogenicity of DNA sequence variation and hub status of proteins in biological networks are well-known features used to study the extent of selection pressure in molecular evolution^{28,29}. We find that proteins with study-wide significant genetic associations, especially those acting in *cis*, are generally more likely to have tissue-enriched gene and protein expression and are more often secreted compared to those with no detected genetic signals (Fig. 2a and Supplementary Data 6 and 7). These effects were slightly attenuated for *cis*-pQTLs tagging PAVs affecting the protein target, although the enrichment of secreted proteins and tissue-enriched gene expression remained significant (Supplementary Fig. 9). These findings suggest that serum *cis*-pQTLs, in part, mirror the regulation of protein secretion from

solid tissues, particularly those that do not affect protein structure, whereas serum levels of proteins without *cis*-pQTLs may be influenced primarily by other mechanisms. By contrast, proteins with *trans* only signals are enriched among transmembrane proteins (Fig. 2a and Supplementary Data 6 and 7). Furthermore, we find that proteins with *cis* signals, and especially those tagging PAVs (Supplementary Fig. 9), generally have lower LoF intolerance, that is they are more tolerant to deleterious mutations, and they tend to have lower hub status in both PPI and co-regulatory networks, indicating a more peripheral position of *cis* regulated proteins in protein networks (Fig. 2b and Supplementary Data 6 and 7). Similarly, larger genetic effects on protein levels are negatively correlated with LoF intolerance and hub status in both the PPI and co-regulatory networks (Supplementary Fig. 10). This suggests that selective pressure may to some extent explain the lack of pQTLs for proteins that are encoded by housekeeping genes, are network hubs and are intolerant to LoF mutations.

Proteins with *trans* acting signals had higher hub status in the co-regulatory network compared to those proteins having no

genetic signals (Fig. 2b). However, *trans* signals were not associated with hub status in the PPI network or influenced by LoF intolerance (Fig. 2b). Complementing this observation, we find that hub proteins in co-regulatory networks are generally connected to more proteins through the same genetic variants (Supplementary Fig. 10). As the co-regulatory network is derived from protein correlations, these results highlight how its structure is to some extent shaped by genetic variants affecting multiple proteins, the majority of which are *trans* regulated⁸ (Supplementary Note 3). These results elucidate key differences between the PPI and the serum protein co-regulatory networks, i.e., while hubs in both types of networks are depleted for *cis*-pQTLs, only those in the co-regulatory network were more likely *trans*-regulated proteins.

Colocalization of pQTLs with GWAS risk loci. Genetic effects on serum proteins may offer novel insights into mechanisms underlying the genetics of common disease and relevant traits. Therefore, we examined the overlap between pQTLs and GWAS loci. We obtained GWAS summary statistics for 81 diseases and clinical traits (Supplementary Data 8) and identified all genome-wide significant ($P < 5 \times 10^{-8}$) GWAS loci overlapping with a study-wide significant pQTL from our results. Of note, the number of significant loci for each of the tested phenotypes is highly dependent on the original study size (Supplementary Fig. 11). GWAS signals for different phenotypes were considered to belong to the same locus if the lead variants were within 300 kb of each other. By this criteria, 1335 GWAS loci for 76 phenotypes were found to be in the vicinity of a study-wide significant pQTL and were tested for colocalization. Of those, 218 GWAS loci (associated with 69 phenotypes) had high support ($PP4 > 0.8$) for colocalization with 1045 proteins (Fig. 3 and Supplementary Data 9 and 10). Additionally, medium support ($0.5 < PP4 \leq 0.8$) was found for colocalization between 171 proteins and 84 loci associated with 49 phenotypes (Fig. 3, Supplementary Data 9 and 10). In a secondary analysis (see Methods), we found that 84% of the protein-phenotype pairs with high support ($PP4 > 0.8$) for colocalization remained so with a more stringent coloc prior selection (Supplementary Data 10 and Supplementary Fig. 12). Of the 772 loci associated with protein levels, 206 (27%) colocalized with at least one GWAS phenotype and the same was true for 1083 (51%) of the 2112 proteins with a study-wide significant pQTL. We found almost all (69/76 or 91%) of the phenotypes tested to have a genetic signal colocalizing with at least one protein, with an average of 9 (11%) colocalized loci per trait (Supplementary Fig. 13). GWAS loci with *cis*-pQTLs were more likely to colocalize (medium or high support) with any protein than those without (22.3% vs 10.4%, Fisher's exact test $P = 7.5 \times 10^{-8}$). For a given phenotype, we observed that its associated loci involved a median of 17 serum proteins (Supplementary Fig. 14). Thus, even a limited proportion of associated loci for a given phenotype generally associates with numerous proteins in serum and consequently implicate multiple affected molecular pathways. To account for multiple independent signals within a given locus, we additionally ran a conditional colocalization analysis for loci that had more than one independent signal per protein, thus including 549 GWAS loci that overlapped with pQTLs for 546 proteins. Here we observed 178 instances of colocalization with medium or high support, of which 51 (involving 19 loci, 14 phenotypes, and 40 proteins) were not captured in the initial colocalization analysis (Supplementary Data 11 and 12).

Colocalized *cis*-acting pQTLs can point to causal genes at GWAS loci. We found 237 and 49 trait-locus-*cis*-protein combinations with high or medium support, respectively. For

102 of 203 (50.2%) unique pairs of GWAS lead variants and colocalized *cis*-pQTLs, the protein was different than that encoded by the nearest gene to the GWAS lead variant (Supplementary Data 10). For example, a GWAS signal for waist-to-hip ratio in the gene *LRRC36*, colocalizes with a pQTL for the serum levels of Agouti-related protein encoded by a nearby gene, *AGRP* (Supplementary Fig. 15), a neuropeptide that increases appetite and decreases metabolism³⁰. A related example involves two loci associated with BMI, located 5 Mb apart on chromosome 20, both of which colocalize with serum levels of the Agouti signaling protein (ASIP) (Supplementary Fig. 16), known to promote obesity via the melanocortin receptor (MC4R)³¹. These two associations are 2.2 Mb and 7.6 Mb upstream of the *ASIP* gene, respectively, however, the colocalization with serum levels of ASIP suggests this may in fact be the causal candidate mediating their effects. Among neurological phenotypes, colocalized *cis*-pQTL examples include a GWAS signal for bipolar disorder on chromosome 2, which colocalizes with the serum levels of the protein encoded by *LMAN2L* (Supplementary Fig. 17A), and a signal for major depression disorder on chromosome 7 colocalizing with *TMEM106B* (Supplementary Fig. 17B), adding support for these being the causal genes at these loci, both of which are also the nearest gene to the GWAS lead variant.

We observed several highly pleiotropic loci, where multiple phenotype signals colocalized with multiple protein signals (Fig. 4a). In fact, among the high ($PP4 > 0.8$) and medium confidence ($PP4 > 0.5$) colocalization results, the number of associated proteins per GWAS locus was positively correlated with the number of associated phenotypes (Spearman's $\rho = 0.50$, $P = 9.9 \times 10^{-17}$). These pleiotropic loci included for example the *ABO* locus, best known for its role in determining the *ABO* blood groups, which was found to harbor eight independent protein signals within a 28 kb region (chr 9, 136,127,268–136,155,127) (Supplementary Data 4), where pQTLs for 63 proteins colocalized with 17 phenotypes, predominantly cardiometabolic and hematopoietic (Fig. 4a and Supplementary Data 10). The complex genetic architecture at this locus gives rise to a wide range of downstream consequences, as indicated by the distinct sets of proteins associated with each independent genetic signal defined here and consistent with previous reports¹⁰, and most traits associated with the locus are affected by more than one of those signals. The 63 proteins in the *ABO* locus were enriched for gene ontology terms and pathways such as “transmembrane signaling receptor activity” ($FDR = 2.7 \times 10^{-6}$), “regulation of cell migration” ($FDR = 2.5 \times 10^{-4}$), and “Hippo-Merlin signaling dysregulation” ($FDR = 1.2 \times 10^{-3}$). Another example of a pleiotropic locus is a 46 kb window (chr 19, 49,206,108–49,252,151), harboring variants adjacent to or within *FUT2* that are associated with diverse traits (Fig. 4b and Supplementary Data 10), including immune (Crohn's disease and type 1 diabetes), anthropometric (waist-to-hip ratio and offspring birth weight), cardiometabolic (blood pressure, LDL, and total cholesterol) and renal (BUN and UACR). *FUT2* encodes for fucosyltransferase-2 that synthesizes the H antigen in body fluids and the intestinal mucosa, while a nearby gene, *FGF21*, is an important metabolic regulator³², acting for example through its effects on sugar intake³³. We find that the genetic signals for 10 phenotypes in this region colocalize with 19 proteins that are collectively enriched for elevated gene expression²⁴ in the intestine ($FDR = 1.4 \times 10^{-6}$), salivary gland ($FDR = 1.7 \times 10^{-6}$), and stomach ($FDR = 8.9 \times 10^{-3}$) (Fig. 4b, c) and include proteins involved in carbohydrate digestion (LCT), taste perception (LPO, PIP) or humoral immunity (CCL25). The proteins regulated by this locus thus suggest downstream effects across different parts of the gastrointestinal tract. Finally, the shared genetic

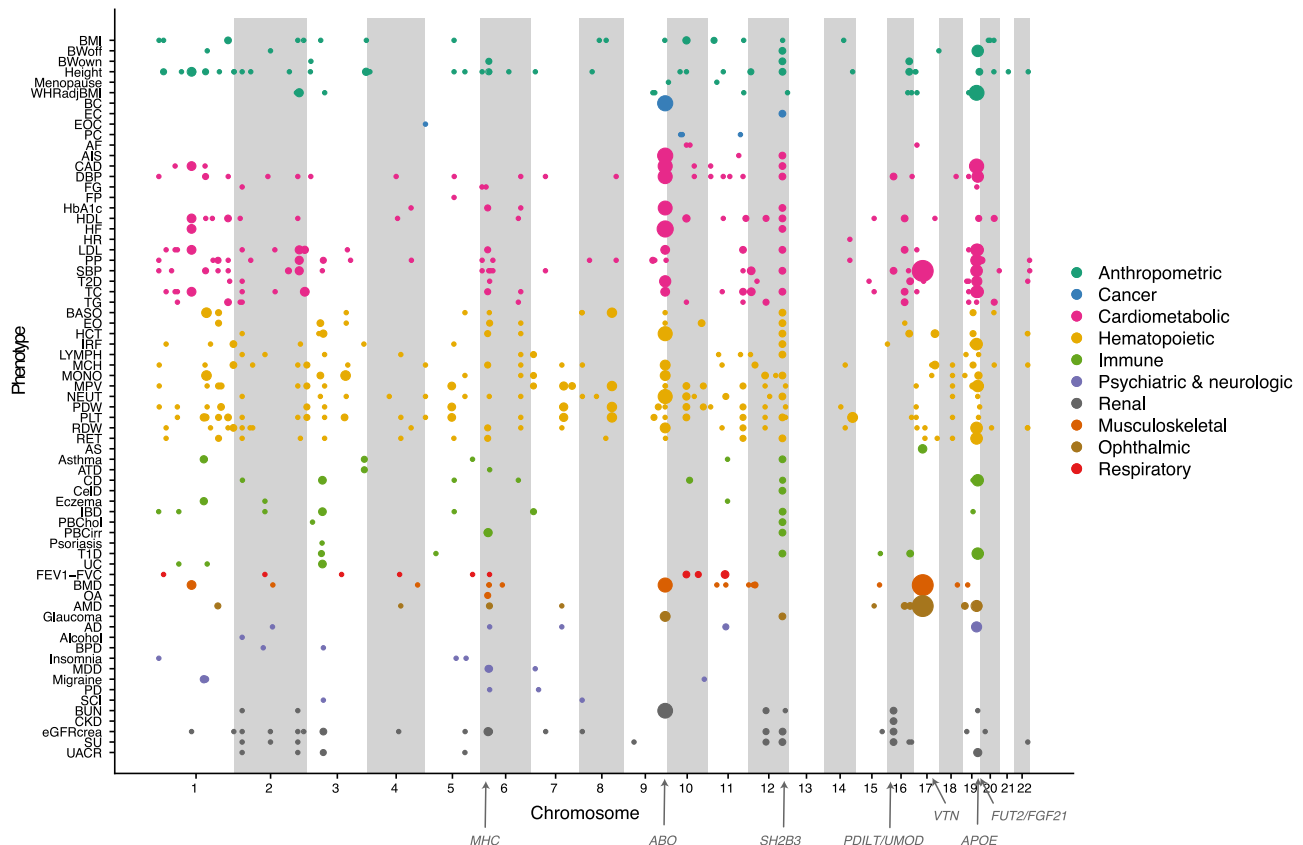


Fig. 3 Overview of colocalization between protein and phenotype associations across the genome. Each dot represents a genetic locus (genomic location on x-axis) that is associated with a phenotype (y-axis), where the size of the dots indicates the number of colocalized proteins (color $PP > 0.5$). Phenotype abbreviations are available from Supplementary Data 8.

architecture of immune disorders has been well documented in the literature and is mirrored in multiple colocalized pQTLs shared between various immune diseases (Supplementary Fig. 18). In particular, the *SH2B3* locus on chromosome 12 stands out in this regard, with GWAS signals for seven immune disorders colocalizing with three *trans*-regulated proteins (THPO, ICAM2, CXCL11), all involved in positive regulation of immune system processes (GO:0002684).

In some cases, we observed more than one colocalized *trans*-pQTLs converging on the same protein for a given phenotype. For example, HDL-associations in the *LIPC* (chromosome 15) and *APOB* (chromosome 2) loci both colocalized with the serum levels of the sodium-coupled transporter *SLC5A8* (Supplementary Fig. 19), involved in the transport of monocarboxylates such as lactate and short-chain fatty acids. Similarly, variants in the *GALNT2* (chromosome 1) and *GCKR* loci (chromosome 2) both regulate the serum levels of NRP1, colocalizing with GWAS signals for triglyceride levels (Supplementary Fig. 20). A more extreme example is a network of 12 loci with GWAS signals for platelet counts that colocalize with serum levels of 24 proteins (Supplementary Fig. 21). These proteins include noggin (NOG) and cochlin (COCH), colocalizing with platelet count signals in five and four loci, respectively.

Associations of proteins with phenotypes in the AGES cohort.

Taking advantage of the deep phenotyping of the AGES cohort, we examined direct associations between colocalized proteins and 37 phenotypes that were measured in the AGES cohort (Supplementary Data 13). For a quarter (10/37) of the phenotypes tested we observed a significant enrichment of phenotype associations among the sets of colocalized proteins compared to

randomly sampled proteins (Fig. 5, Supplementary Fig. 22, and Supplementary Data 14), demonstrating more generally that GWAS loci for complex phenotypes regulate serum proteins that themselves are often directly associated to the phenotype itself. At a more relaxed genome-wide significant ($P < 5 \times 10^{-8}$) threshold for pQTLs, the proportion of phenotypes with significant enrichment of protein associations increased to 45% (18/40 phenotypes, Supplementary Fig. 23), likely due to an increase in statistical power with more colocalized proteins per phenotype at this threshold and indicating that more associations between proteins regulated by GWAS-loci and the respective phenotypes can be expected to be identified as sample sizes for proteogenomic studies increase. Among the diseases and clinical traits with the strongest enrichment for direct protein-trait associations, we found age-related macular degeneration (AMD) (14% of colocalized proteins associated compared to an average of 7% for random proteins, $P < 0.001$), total cholesterol (67% vs 35% for random, $P < 0.001$), Alzheimer's disease (21% vs 1% for random, $P = 0.001$), and type 2 diabetes (60% vs 40% for random, $P = 0.017$). In some cases, this enrichment was driven by proteins regulated from a few *trans* loci, as evident by the loss of significance when the analysis was repeated without pleiotropic loci regulating five or more proteins, leaving on average 17 proteins per trait (Fig. 5 and Supplementary Data 14). This was particularly evident for Alzheimer's disease, where the enrichment was entirely driven by the associations of proteins regulated by the *APOE* locus (Supplementary Data 13). In other cases, the removal of proteins regulated by pleiotropic loci resulted in an enhanced enrichment of phenotype associations, such as for HbA1c, mean platelet volume and diastolic blood pressure (Supplementary Fig. 22 and Supplementary Data 14).

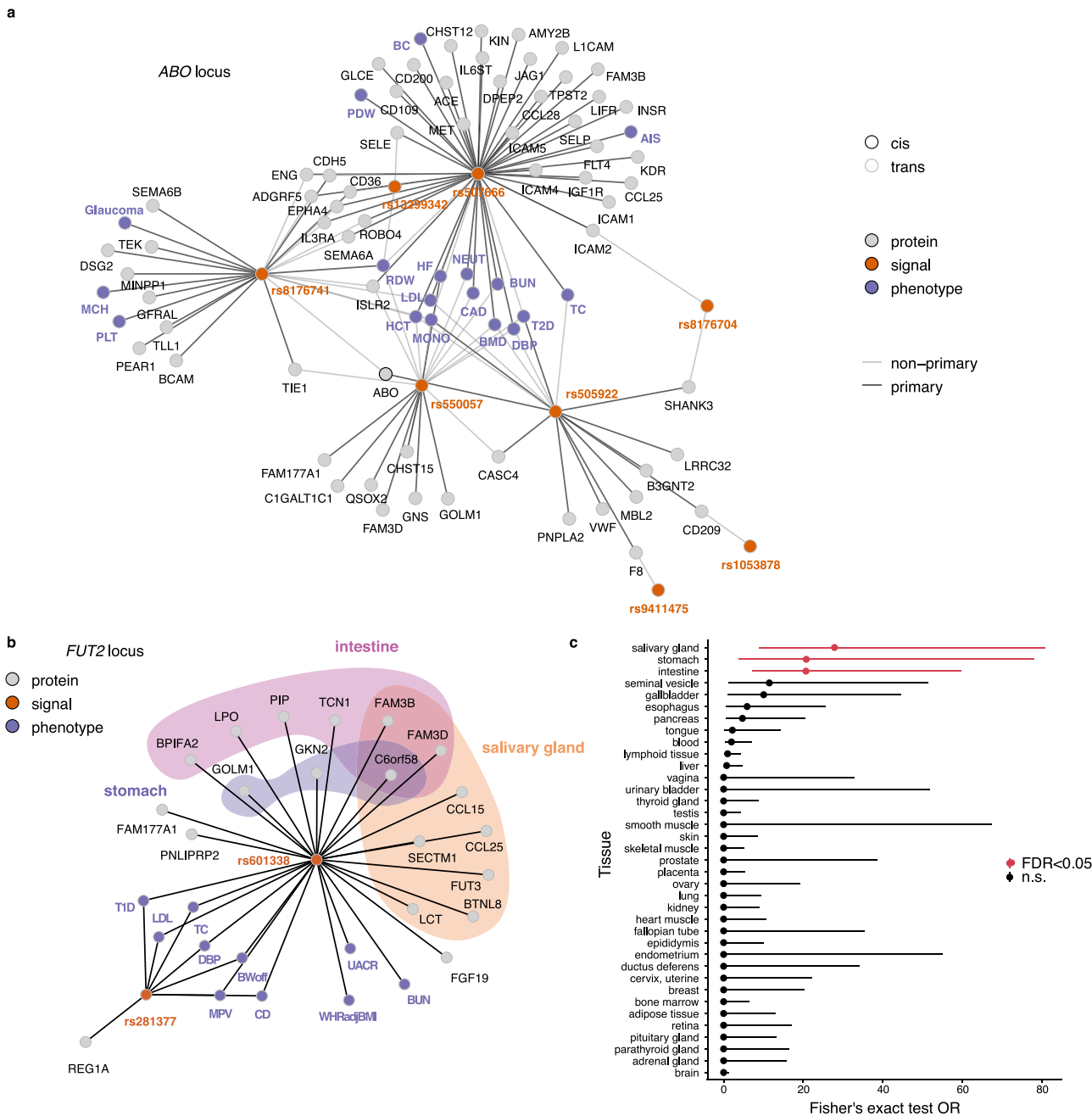


Fig. 4 An overview of independent genome-wide significant genetic signals. **a** Genetic signals (orange nodes), using conditional and joint analysis (GCTA-COJO)^{19,20}, annotated by the SNP with the strongest protein association, at the *ABO* locus (chr 9, 136,127,268–136,155,127) and their links to proteins (gray nodes) and phenotypes (purple nodes). Edges between genetic signals and proteins indicate primary (dark edges) and secondary (light edges) independent signals from the conditional analysis. Edges between genetic signals and traits indicate that any of the lead pQTL SNPs within that signal reaches $P < 5 \times 10^{-8}$ (two-sided) in GWAS summary statistics for the given trait, and the primary signal is assigned for the trait based on the lowest *P*-value. **b** An overview of the independent genome-wide significant genetic signals (orange nodes), annotated by the SNP with the strongest protein association, at the *FUT2* locus (chr 19, 49,206,108–49,252,151) and their links to proteins (gray nodes) and the phenotypes they colocalize with (purple nodes). The background color indicates tissue-elevated expression in the salivary gland, intestine or stomach. **c** Enrichment (Fisher's exact test, two-sided) of tissue-elevated expression among the 19 proteins regulated by the *FUT2* locus where Benjamini–Hochberg FDR < 0.05 is considered significant (red). Here 4016 proteins with available data in the Human Protein Atlas were included. Odds ratio estimates are presented with 95% confidence intervals. Phenotype abbreviations are available from Supplementary Data 8.

By evaluating each individual locus separately, we identified six loci with significant phenotype-association enrichment among its linked proteins that colocalized with GWAS signals for the respective phenotype, thus demonstrating specific examples of genetic variants whose molecular and phenotypic consequences

are linked within the same cohort (Supplementary Data 15). Here the *APOE* locus stood out in terms of number of enriched phenotypes, with its regulated proteins being enriched for associations with Alzheimer's disease, AMD, and numerous cardiometabolic traits including coronary artery disease. The 641

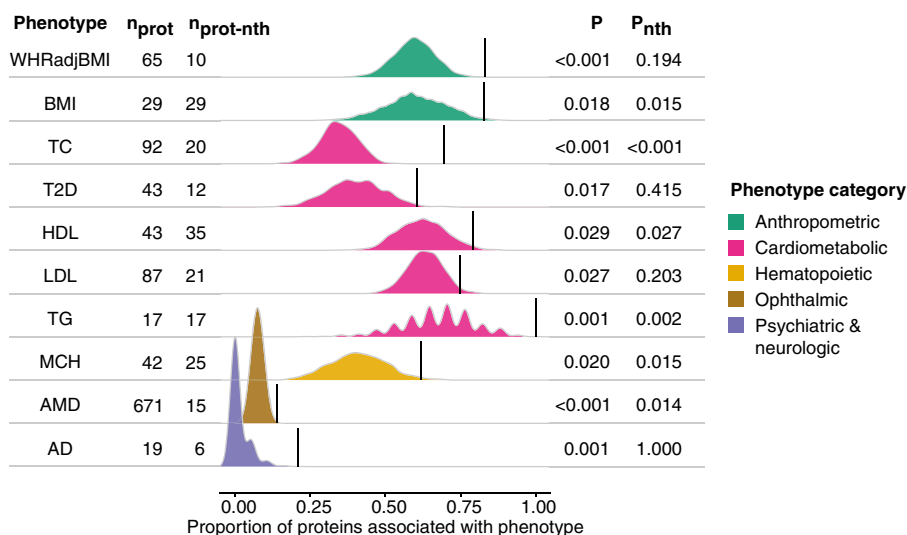


Fig. 5 Enrichment of phenotype associations among sets of colocalized proteins. The ridgeline plot illustrates for each GWAS phenotype the proportion of colocalized proteins that were significantly associated with the same trait in AGES (linear regression, FDR < 0.05, $n = 5457$) (black lines) compared to 1000 randomly sampled sets of proteins of the same size (density curves), here showing only those with empirical $P < 0.05$. See full results in Supplementary Fig. 22. The number of colocalized proteins for each trait are provided on the left-hand side, along with the number of proteins remaining after the removal of proteins originating from loci with 5 or more colocalized proteins from the analysis, annotated as no *trans* hotspots (nth). Empirical P-values for significant enrichment of trait-associations are shown to the right. WHRadjBMI waist-to-hip ratio adjusted for BMI, TC total cholesterol, T2D type 2 diabetes, HDL high-density lipoprotein cholesterol, LDL low-density lipoprotein cholesterol, TG triglycerides, MCH mean corpuscular hemoglobin, AMD age-related macular degeneration, AD Alzheimer's disease.

proteins regulated by all seven independent signals in the *VTN* locus on chromosome 17 were also enriched for associations with AMD. The *PSRC1-CELSR2-SORT1* locus, best known for its associations with coronary artery disease and cholesterol levels, showed enrichment for protein associations with bone mineral density. Proteins regulated by the *ABO* locus on chromosome 9 and the *UGT* gene family cluster on chromosome 8 were enriched for associations with total cholesterol and finally the proteins regulated by the *ZFPM2* locus on chromosome 8 were enriched for associations with basophil counts.

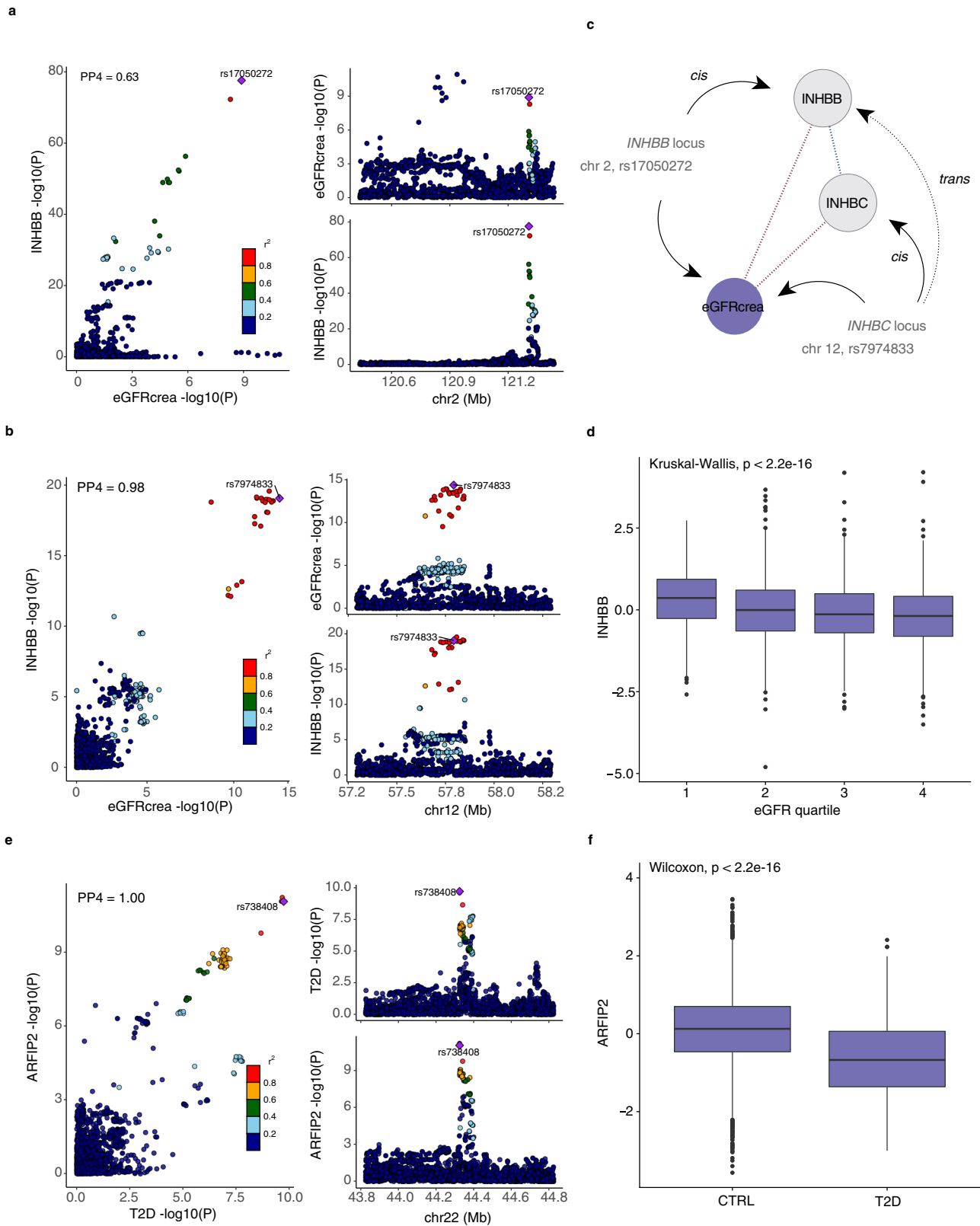
Other examples of colocalized proteins showing significant associations with the respective phenotype include the inhibin beta subunit B (*INHBB*) protein, which has a *cis*-pQTL on chromosome 2 and a *trans*-signal on chromosome 12, near the *INHBC* gene that encodes another subunit of the same protein complex, both of which colocalize with GWAS signals for estimated glomerular filtration rate (eGFR), a marker of renal function (Fig. 6a–c). The *INHBB* protein itself is associated with eGFR in the AGES cohort in a directionally consistent manner (Fig. 6c, d). Thus, the associations of these genetic variants affecting different components of the same protein complex together with the consistent association between the protein itself and eGFR indicate a possible role for the inhibin/activin proteins in renal function. Another example is the colocalization between a GWAS signal for type 2 diabetes with the missense lead variant rs738409 in the *PNPLA3* gene, a well-established locus for non-alcoholic fatty liver disease³⁴, and a *trans*-pQTL for ADP Ribosylation Factor Interacting Protein 2 (*ARFIP2*) (Fig. 6e), which is strongly downregulated in type 2 diabetes patients in AGES (Fig. 6f)¹⁸. These observations raise several questions, for example how a missense variant in *PNPLA3* leads to a change in the circulating levels of *ARFIP2*, if *ARFIP2* provides some sort of readout of *PNPLA3* function and finally how *ARFIP2* relates to type 2 diabetes, i.e., if it mediates any of the risk associated with this locus or if it is merely a bystander. Thus, the links discovered here between genetic loci, proteins, and disease risk can be used to derive new hypotheses for future research.

Discussion

To the best of our knowledge, this is the largest genome-wide association study of serum protein levels in terms of protein coverage to date, and it demonstrates a significant increase in existing knowledge in terms of the number of significant genetic associations to proteins in circulation. We furthermore provide a systematic evaluation of protein-phenotype associations in the context of established risk loci for numerous diseases and clinical traits.

The current study expands on our previous work⁸ by increasing the number of genetic variants included in the analysis (from *cis*-regions only to a genome-wide analysis), thus increasing the search space, but also enhancing statistical power for identifying genetic associations by increasing the sample size in genetic analyses from 3219 previously to 5368 participants in the current study. Here, we identified study-wide significant genetic signals for half of the measured proteins and up to 13 independent genetic signals for a given protein. Thus, as for any other traits, the expected number of genetic associations for serum proteins can only be expected to increase with larger sample sizes, as has been demonstrated for CRP³⁵. Large-scale meta-analyses across cohorts and biobanks will with time provide a more complete understanding of the genetic regulation of individual circulating proteins and their networks, including the effect of variability between different tissues on serum protein levels. The majority of *cis* and *trans* acting pQTLs detected in serum and plasma can be readily replicated across different populations, as shown in the current study, and different proteomic platforms^{8,9,17,23}. However, a recent cross-platform comparison has shown that a subset of pQTLs are platform-specific and may in some cases represent epitope effects or other technical factors²³. Thus, meta-analyses across platforms will still need to consider differences in analytical approaches and in cases where protein quantifications obtained by orthogonal methods differ, *cis*-pQTLs and mass spectrometry validation of probe targets may be good indicators of platform specificity³⁶.

We demonstrate that proteins that are secreted, tissue-enriched, more tolerant to LoF variants and with few connections in



protein networks were most likely to be genetically controlled. This pattern was mainly driven by *cis* acting signals and not as apparent for the *trans* effects on protein levels, illustrating that *cis*- and *trans*-signals for serum proteins arose by different means and may differ in evolutionary properties. Our results are consistent with the notion that evolutionary important, and likely disease-relevant, genes undergo a negative selection against

genetic *cis*-variants, which has been proposed as an explanation of the extreme polygenicity of complex traits³⁷. The observed depletion of *cis*-variants among network hubs in our study are furthermore in line with the recently proposed omnigenic model², which suggests that core disease genes are rarely affected directly by GWAS variants but rather through a multitude of smaller effects mediated through *cis*-regulation of peripheral

Fig. 6 Colocalization between GWAS signals for eGFR and *INHBB* and *INHBC*. **a** Colocalization between GWAS signals (linear regression) at the *INHBB* locus on chromosome 2 and **b** the *INHBC* locus on chromosome 12 and eGFR. The PP4 value indicates the posterior probability for colocalization obtained from colocalization analysis. **c** A schematic diagram showing the convergence of genetic effects on serum levels of *INHBB* at the *INHBB* locus in *cis* and *INHBC* locus in *trans*. Variants in the *INHBC* locus furthermore affect *INHBC* serum levels in *cis*, albeit not reaching study-wide significance ($P = 8.5 \times 10^{-8}$, two-sided). Serum levels of *INHBB* and *INHBC* are positively correlated (Pearson's $r = 0.32$, $P = 3.4 \times 10^{-130}$, two-sided), while both are negatively associated (linear regression) with eGFR (beta = -4.52 , SE = 0.23, $P = 1.3 \times 10^{-82}$, two-sided, and beta = -2.62 , SE = 0.22, $P = 5.4 \times 10^{-32}$, two-sided, respectively). **d** Boxplot showing *INHBB* serum levels in the AGES cohort ($n = 5457$) by eGFR quartiles. **e** Colocalization between a GWAS signals for T2D and a *trans* signal for ARFIP2 at the *PNPLA3* locus on chromosome 22. **f** Boxplot showing ARFIP2 serum levels in the AGES cohort by T2D status ($n_{T2D} = 658$, $n_{CTRL} = 4799$). Boxplots in **d** and **f** indicate median value, 25th and 75th percentiles. Whiskers extend to smallest/largest value no further than $1.5 \times$ interquartile range. Outliers are shown.

genes in regulatory networks. Thus, while our results provide a map of *cis*-regulatory effects for 812 proteins, linking many of these to disease signals from GWAS studies, those without *cis*-effects may be even more important in the context of disease and should be studied further by other means. While hubs in the PPI network were depleted for any genetic signal, *trans*-affected proteins showed higher degree of connectivity in the co-regulatory network compared to those with no detectable genetic signal. These findings demonstrate that the structure of the co-regulatory network is to some extent driven by genetic variants affecting multiple proteins. We also note that unlike PPI networks constructed in solid tissues, the serum protein networks are composed of protein members synthesized across different tissues of the body and as such may reflect cross-tissue regulation⁸ or factors that affect the levels of circulating proteins independently of their origin.

Among proteins with genetic associations, we find that many have multiple genetic signals, both across different loci throughout the genome but also within a given locus as revealed by conditional analysis, indicating that allelic heterogeneity is common in loci regulating serum protein levels. Widespread allelic heterogeneity has been described for gene expression³⁸ and complex traits in general³⁹. For serum proteins, this may reflect the complex regulation and diverse origin of proteins in circulation, as these proteins may arise from almost any tissue of the body. Furthermore, *cis*-pQTLs show a roughly 40% overlap with gene expression QTLs^{8,9}, suggesting that a large fraction of the genetic effect is mediated through any of the many post-transcriptional steps involved in protein maturation.

The integration of well-established genetic associations for 81 diseases and disease-related traits revealed a profound overlap with the genetic signals affecting protein levels in our study, where a third of the identified loci regulating serum protein levels colocalized with at least one GWAS phenotype. We identify examples of disease-associated loci colocalizing with many proteins, especially loci that also exhibit pleiotropic phenotype associations. Thus, it seems likely that the more complex the molecular consequences of a variant, the more likely it is to be associated with many different phenotypes, which has also been observed at the transcriptomic level⁴⁰. The serum protein changes associated with any given disease signal can shed new light on the underlying pathways that are affected either before or after the onset of disease. The deep phenotyping of the AGES cohort allowed for an integrative analysis of genetic variants, serum protein measurements and phenotypes within the same population. For proteins regulated by loci linked to a given disease-relevant phenotype, we observed an enrichment for associations to the same phenotype measures in our cohort, thus pointing to many novel candidate proteins that may play a role in regulating or responding to these phenotypes. However, it should be noted that while a pQTL that colocalizes with a signal for a disease or clinical trait may implicate causal candidates for mediating the genetic risk, it may just as well indicate

downstream events or even unrelated parallel effects of a pleiotropic variant. Furthermore, the plasma proteome has been shown to change in waves throughout the human lifespan⁴¹, with a large proportion of proteins changing in old age. Thus, some of the associations observed in the elderly AGES cohort may not be directly transferable to a younger population but may at the same time shed light on the physiological relevance of circulating proteins in the aging process. Our study provides genetic instruments for further studies of causal relationships for specific examples, however mechanistic and experimental studies are warranted for determining the underlying chains of events behind these complex associations. Our results offer an in-depth inventory of information regarding the interconnections between genetic variants, serum proteins, and disease-relevant traits, which may encourage discoveries of therapeutic targets and fluid biomarkers, providing a robust framework for understanding the pathobiology of complex disease.

Methods

The AGES cohort. Cohort participants aged 66 through 96 were included from the AGES-Reykjavik Study⁴², a prospective study of deeply phenotyped individuals of Northern European ancestry (Supplementary Data 1). Blood samples were collected at the baseline visit after overnight fasting and serum lipids, glucose, HbA1c, insulin, uric acid, and urea were measured using standard protocols. LDL and total cholesterol levels were adjusted for statin use, with an approach similar to what has previously been described⁴³. Hypertension medication use was accounted for by adding 15 mmHG to systolic blood pressure and 10 mmHG to diastolic blood pressure⁴⁴. Serum creatinine was measured with the Roche Hitachi 912 instrument and estimated glomerular filtration rate (eGFR) derived with the four-variable MDRD Study equation⁴⁵. Type 2 diabetes was defined from self-reported diabetes, diabetes medication uses or fasting plasma glucose ≥ 7 mmol/L. Type 2 diabetes patients were excluded from all analyses for fasting glucose, fasting insulin, and HbA1c. The presence of coronary artery disease was determined using hospital records and/or data from the cause of death registry. A coronary artery disease event was any occurrence of myocardial infarction, ICD-10 codes: I21–I25, coronary revascularization (either CABG surgery or percutaneous coronary intervention (PCI)) or death from CHD according to a complete adjudicated registry of deaths available from the national mortality register of Iceland (ICD-10 codes I21–I25). The prostate cancer diagnosis was obtained from medical records (ICD-10 code C61). Information on migraine, Parkinson's disease, eczema, and thyroid disease were obtained from questionnaires. Alzheimer's disease was determined with a consensus diagnosis based on international guidelines was made by a panel that includes a geriatrician, neurologist, neuropsychologist, and neuroradiologist and defined according to the criteria of the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA), as previously described⁴⁶. Hospital- and mortality data were also used to identify cases according to the ICD-10 code F00. Age-related macular degeneration (AMD) in the AGES-Reykjavik study has been previously described⁴⁷, but in short was defined by the presence of any soft drusen and pigmentary abnormalities (increased or decreased retinal pigment) or the presence of large soft drusen ≥ 125 μm in diameter with a large drusen area >500 μm in diameter or large ≥ 125 μm indistinct soft drusen in the absence of signs of late AMD. The maximum grip strength of the dominant hand was measured by a computerized dynamometer, as previously described⁴⁸. Bone mineral density was estimated from a CT scan of the femur⁴⁹. The AGES-Reykjavik study was approved by the NBC in Iceland (approval number VSN-00-063), and the National Institute on Aging Intramural Institutional Review Board, and the Data Protection Authority in Iceland. All participants provided informed consent.

Protein measurements. Serum levels of 4135 human proteins, targeted by 4782 SOMAmers⁵⁰, were determined at SomaLogic Inc. (Boulder, US) in samples from 5457 AGES-Reykjavik participants as previously described⁸. A few SOMAmers are annotated to more than one gene, for example when the target is a protein complex, thus the 4782 SOMAmers are annotated to a total of 4118 unique targets (annotated as one or more Entrez gene symbols) in the most up to date inhouse annotation database, which was used in all analyses. Sample collection and processing for protein measurements were randomized and all samples were run as a single set. The SOMAmers that passed quality control had median intra-assay and inter-assay coefficient of variation (CV) < 5% similar to that reported on variability in the SOMAscan assays⁵¹. In addition to multiple types of inferential support for SOMAmer specificity towards target proteins including cross-platform validation and detection of *cis*-acting genetic effects⁸, direct measures of the SOMAmer specificity for 779 of the SOMAmers in complex biological samples was performed using tandem mass spectrometry⁸. Previous studies have shown that pQTLs replicate well across proteomics platforms^{8,9}. While a recent comparison of protein measurements across different platforms showed a wide range of correlations^{23,36}, *cis* pQTLs and validation by mass spectrometry were predictive of a strong correlation across platforms and are likely good indicators of platform specificity when protein concentrations obtained by orthogonal methods differ³⁶. Hybridization controls were used to correct for systematic variability in detection and calibrator samples of three dilution sets (40%, 1%, and 0.005%) were included so that the degree of fluorescence was a quantitative reflection of protein concentration. In the main text the results are described at a protein level instead of SOMAmer level, to avoid overcounting as some proteins are targeted by more than one SOMAmer that were selected to different forms or domains of the same protein. Thus, when we refer to a protein having a genetic signal, this indicates that any of the protein's SOMAmers are associated with that genetic signal.

Genotyping and imputation. Within the AGES cohort, 3219 individuals were genotyped with the Illumina hu370CNV array, and 2,705 individuals genotyped with the Illumina Infinium Global Screening Array. Data from both genotype arrays underwent quality control procedure, separately, removing variants with call rate < 95% and HWE P -value < 1×10^{-6} . Both arrays were imputed against the Haplotype Reference Consortium imputation panel r1.1 with the Minimac3 software⁵². Post-imputation quality control consisted of filtering out variants with imputation quality $R^2 < 0.7$, MAF < 0.01, as well as monomorphic and multiallelic variants for each platform separately. Genotypes for remaining variants, with matching location and alleles between platforms, were merged to create a dataset with 7,506,463 variants for 5656 individuals (268 individuals were genotyped on both platforms, with a 99% match of genotypes for the final set of variants between platforms). The quality control procedure was performed using bcftools (v1.9)⁵³ and PLINK 1.9⁵⁴. All positions are based on genome assembly GRCh37.

GWAS and conditional analysis. Data processing and statistical analysis were performed using R (v3.5.1 & 4.0.1) and Rstudio (v1.1.456), unless otherwise specified. Box-Cox transformation was applied on the protein data⁵⁵ and extreme outlier values were excluded, defined as values above the 99.5th percentile of the distribution of 99th percentile cutoffs across all proteins after scaling, resulting in the removal of an average 11 samples per SOMAmer, as previously described¹⁸. Within the AGES cohort, 5368 individuals had both genetic data and protein measurements. With that sample set, 7,506,463 variants were tested for association with each of the 4782 SOMAmers separately, in a linear regression model with age, sex, 5 genetic principal components, and genotyping platform as covariates using PLINK 2.0. To obtain independent genetic signals, we performed a stepwise conditional association analysis for each SOMAmer separately with the GCTA-COJO software^{19,20}. We conditioned on the current lead variant, defined as the variant with the lowest P -value, and then kept track of any new lead variants with study-wide-significant associations. Variants in strong LD ($r^2 > 0.9$) with previously chosen lead variants were not considered for joint analysis to avoid multicollinearity. The independent signals defined by GCTA-COJO were subsequently subjected to a validation analysis where the joint models were tested using individual-level data in AGES and those remaining study-wide significant retained. Associations with independent lead variants within 300 kb window of the gene boundaries of the protein-coding gene were defined as *cis*-signals, and otherwise in *trans*. To compare independent signals between SOMAmers, we define any signals with lead variants in strong LD ($r^2 > 0.9$) as the same signal. Due to the complex LD structure and high pleiotropy of the MHC region⁵⁶ (chr.6, 28.47–34.45 Mb) we collapsed all signals within that region to a single signal. To define loci harboring independent signals, we defined a 300 kb window around each independent signal (150 kb up- and downstream of lead variants) and collapsed all such intersecting windows. Therefore, the definition of loci is solely based on physical distances while the definition of independent signals is solely based on LD structure. Variants were annotated using the Ensembl Variant Effect Predictor⁵⁷ (v104, “per_gene” option), where PAVs affecting the corresponding protein target were defined as those with the following consequences: splice acceptor variant, splice donor variant, splice region variant, stop gained, stop lost, start lost, frameshift variant, missense variant or frameshift variant. The GWAS results were visualized using Circos⁵⁸. Pathway enrichment was performed using gProfiler⁵⁹, using the full set of

measured proteins as background and considering Benjamini–Hochberg FDR < 0.05 as statistically significant. Enrichment of tissue-elevated gene expression was performed using data from the Human Protein Atlas²⁴ with a Fisher's exact test, considering Benjamini–Hochberg FDR < 0.05 as statistically significant.

Comparison with previous proteogenomic studies. To evaluate the novelty of the genetic associations identified in the current study, we compared our results to 20 previously published proteogenomic studies (Supplementary Data 5), including the protein GWAS in the INTERVAL study⁹, our previously reported genetic analysis of 3,219 AGES cohort participants⁸, and a recent Illumina exome-array analysis in 5,343 AGES participants²². In a previous proteogenomic analysis of AGES participants⁸, one *cis* variant was reported per protein using a locus-wide significance threshold, as well as *cis*-to-*trans* variants at a Bonferroni corrected significance threshold, whereas the more recent exome-array analysis²² reported results at a study-wide significant threshold ($P < 1 \times 10^{-10}$). Due to these differences in reporting criteria, we only considered the associations in previous AGES results that met the current study-wide P -value threshold ($P < 1.046 \times 10^{-11}$). For all other studies, we retained the pQTLs at the reported significance threshold. In addition, we performed a lookup of all independent pQTLs from the current study available in summary statistics from the INTERVAL study, considering them known if they reached a study-wide significance in their data. We calculated the LD structure between the reported significant variants for all studies, using 1000 Genomes v3 EUR samples, but using AGES data when comparing to previously reported AGES results. We considered variants in LD ($r^2 > 0.9$ for consistency for defining signals across SOMAmers described above but results for $r^2 > 0.5$ are additionally shown in Supplementary Note 1) to represent the same signal across studies. Comparison was performed on protein level, by matching the reported Entrez gene symbol from each study.

Enrichment analysis. We grouped the proteins into three categories derived from our GWAS results; (a) proteins with at least one *cis* signal, (b) proteins with no *cis* signals and at least one *trans* signal, and (c) proteins with no genetic signal. From our data we also derived three continuous traits for a given protein; (a) a number of associated independent signals, (b) highest absolute beta coefficient of all associated signals, and (c) the number of proteins that share genetic signals with the given protein, which is essentially a quantitative representation of whether a protein is a part of a *trans* hotspot. We fetched publicly available data regarding; (a) tissue-elevated gene expression, where “tissue-enriched” in our analyses refers to the “Tissue Enriched”, “Tissue Enhanced” or “Group Enriched” categories defined by Uhlen et al.²⁴, (b) tissue-elevated protein expression, where “tissue-enriched” in our analyses refers to the “Tissue Enriched”, “Tissue Enhanced” or “Group Enriched” categories defined by Wang et al.²⁵, (c) annotation of secreted and transmembrane proteins, classifying proteins as secreted or transmembrane if it was predicted so by at least one method or one segment, respectively²⁴, (d) gene-level loss-of-function intolerance²⁶, and (e) network degree in the InWeb protein-protein interaction network²⁷. Furthermore, we estimated hub status of proteins within the serum protein co-regulation network derived from the AGES cohort⁸. Protein classifications were compared using a Fisher's exact test, where the estimate is the odds ratio. Continuous parameters were compared between protein classes using the Wilcoxon rank-sum test and for the estimate we calculated the median of the difference between values from the two classes, so the size of the estimate is dependent on the scale of the values. For comparing two continuous traits we used Spearman's Rho correlation. We report 95% confidence intervals of all estimates.

GWAS colocalization analysis. We included 81 phenotypic traits including major disease classes in the colocalization analysis, for which GWAS summary statistics were publicly available from consortium websites and the GWAS catalog⁶⁰. We restricted the study selection to those with study sample sizes of $n > 10$ K, of primarily European Ancestry (to match the AGES cohort's LD structure), having at least one genome-wide significant association ($P < 5 \times 10^{-8}$) and selecting one study per phenotype (Supplementary Data 8). For each trait, significant loci were defined by identifying all genome-wide variants ($P < 5 \times 10^{-8}$) at least 500 kb apart, defining a flanking region of 1 Mb around each lead variant, and finally merging overlapping regions. For each GWAS locus, all SOMAmers with a study-wide significant association (*cis* or *trans*) within the given region were tested for colocalization, if at least 50 SNPs in the region had complete information from both trait and protein GWAS and the overlapping set of SNPs included at least one SNP with a genome-wide significant ($P < 5 \times 10^{-8}$) phenotype association and at least one SNP with a study-wide significant ($P < 1.046 \times 10^{-11}$) protein association. When the MAF was not available for a given GWAS, the 1000 Genomes EUR MAF was used instead. Colocalization analysis was performed with coloc (v.3.2-1)⁶¹, using the coloc.abf function with default priors. In a secondary analysis we repeated the analysis with a more stringent prior selection, $p12 = 5 \times 10^{-5}$, as recently proposed⁶². High and medium colocalization support was defined as PP.H4 > 0.8 and PP.H4 > 0.5, respectively. Conditional colocalization analysis was performed using coloc 4.0–4⁶², using the “allbutone” option and restricted to loci harboring more than one independent signal per protein. Unlike the primary coloc analysis, the conditional analysis requires the GWAS effect size to be included, thus the phenotypes AMD, ATD, and PD were excluded from this analysis which did not

have this information available in the GWAS summary statistics. Results were visualized with LocusCompare⁶³.

Phenotype associations. For each GWAS phenotype with a corresponding measurement in AGES and well represented at the population level (Supplementary Data 8), the colocalized proteins were tested for association with the phenotype in all AGES participants with protein data available ($n = 5,457$, see n missing per phenotype in Supplementary Data 1), in a linear or logistic regression model adjusted for age and sex. The SOMAmer with the lowest P -value was chosen for each protein, and P -values were subsequently adjusted for the number of proteins tested for each trait by Benjamini-Hochberg FDR. For each phenotype with at least five colocalized proteins, the proportion of significantly associated proteins (FDR < 0.05) was compared to that obtained by 1000 randomly sampled protein sets of the same size, again choosing the SOMAmer with the lowest P -value per protein, and an empirical P -value calculated. The analysis was repeated by excluding proteins originating from loci where five or more proteins colocalized with the same phenotype. The same enrichment analysis was additionally performed for each individual locus where five or more proteins colocalized with the same phenotype.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The custom-design Novartis SOMAscan is available through a collaboration agreement with the Novartis Institutes for BioMedical Research (lori.jennings@novartis.com). Data from the AGES-Reykjavik study are available through collaboration (AGES_data_request@hjarta.is) under a data usage agreement with the IHA. All access to data is controlled via the use of a subject-signed informed consent authorization. The time it takes to respond to requests varies depending on the nature and circumstances of the request, but it will not exceed 14 working days. The protein GWAS summary statistics data from this study were deposited in the GWAS catalog database with accession IDs for each summary statistics dataset based on unique SOMAmers, as listed in Supplementary Data 16. SNP correlations at protein-associated loci from the AGES cohort are available from zenodo.org (<https://doi.org/10.5281/zenodo.5711426>). All other data supporting the conclusions of the paper are presented in the main text and freely available as a supplement to this manuscript (Supplementary Information and Supplementary Data).

Received: 29 June 2021; Accepted: 15 December 2021;

Published online: 25 January 2022

References

- Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
- Schadt, E. E. Molecular networks as sensors and drivers of common human diseases. *Nature* **461**, 218–223 (2009).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Farh, K. K. H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
- Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
- Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).
- Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
- Suhre, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
- Pietzner, M. et al. Genetic architecture of host proteins involved in SARS-CoV-2 infection. *Nat. Commun.* **11**, 1–14 (2020).
- Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
- Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
- Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat. Commun.* **10**, 1579 (2019).
- Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteom.* **1**, 845–867 (2002).
- Lamb, J. R., Jennings, L. L., Gudmundsdottir, V., Gudnason, V. & Emilsson, V. It's in our blood: a glimpse of personalized medicine. *Trends Mol. Med.* **27**, 20–30 (2021).
- Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
- Gudmundsdottir, V. et al. Circulating protein signatures and causal candidates for type 2 diabetes. *Diabetes* **69**, 1843–1853 (2020).
- 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).
- Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
- Sen, N., Gui, B. & Kumar, R. Role of MTA1 in cancer progression and metastasis. *Cancer Metastasis Rev.* **33**, 879–889 (2014).
- Emilsson, V. et al. Coding and regulatory variants are associated with serum protein levels and disease. *Nat. Commun.* <https://doi.org/10.1038/s41467-022-28081-6> (2022).
- Pietzner, M. et al. Cross-platform proteomics to advance genetic prioritisation strategies. *bioRxiv* <https://doi.org/10.1101/2021.03.18.435919> (2021).
- Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
- Wang, D. et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, 1–16 (2019).
- Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
- Li, T. et al. A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64 (2017).
- Cvijović, I., Good, B. H. & Desai, M. M. The effect of strong purifying selection on genetic diversity. *Genetics* **209**, 1235–1278 (2018).
- Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
- Keen-Rhinehart, E., Ondek, K. & Schneider, J. E. Neuroendocrine regulation of appetitive ingestive behavior. *Front. Neurosci.* **7**, 213 (2013).
- Adan, R. A. H. et al. The MC4 receptor and control of appetite. *Br. J. Pharmacol.* **149**, 815–827 (2006).
- Bookout, A. L. et al. FGF21 regulates metabolism and circadian behavior by acting on the nervous system. *Nat. Med.* **19**, 1147–1152 (2013).
- Von Holstein-Rathlou, S. et al. FGF21 mediates endocrine control of simple sugar intake and sweet taste preference by the liver. *Cell Metab.* **23**, 335–343 (2016).
- Speliotes, E. K. et al. Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.* **7**, 1001324 (2011).
- Ligthart, S. et al. Genome analyses of >200,000 individuals identify 58 loci for chronic inflammation and highlight pathways that link inflammation and complex disorders. *Am. J. Hum. Genet.* **103**, 691–706 (2018).
- Raffield, L. M. et al. Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* **20**, e1900278 (2020).
- O'Connor, L. J. et al. Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
- Jansen, R. et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum. Mol. Genet.* **26**, 1444–1451 (2017).
- Hormozdiari, F. et al. Widespread allelic heterogeneity in complex traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).
- The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
- Harris, T. B. et al. Age, gene/environment susceptibility-Reykjavik study: Multidisciplinary applied phenomics. *Am. J. Epidemiol.* **165**, 1076–1087 (2007).
- Peloso, G. M. et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 (2014).
- Evangelou, E. et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).
- Levey, A. S., Greene, T., Kusek, J. & Beck, G. A simplified equation to predict glomerular filtration rate from serum creatinine. *J. Am. Soc. Nephrol.* **11**, A0828 (2000).
- Qiu, C. et al. Cerebral microbleeds, retinopathy, and dementia: The AGES-Reykjavik Study. *Neurology* **75**, 2221–2228 (2010).
- Jonasson, F. et al. Five-year incidence, progression, and risk factors for age-related macular degeneration: the age, gene/environment susceptibility study. *Ophthalmology* **121**, 1766–1772 (2014).
- Mijnarends, D. M. et al. Physical activity and incidence of sarcopenia: The population-based AGES-Reykjavik Study. *Age Ageing* **45**, 614–621 (2016).

49. Steingrimsdottir, L. et al. Hip fractures and bone mineral density in the elderly —importance of serum 25-hydroxyvitamin D. *PLoS ONE* **9**, e91122 (2014).
50. Gold, L. et al. Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS ONE* **5**, e15004 (2010).
51. Hathout, Y. et al. Large-scale serum protein biomarker discovery in Duchenne muscular dystrophy. *Proc. Natl Acad. Sci. USA* **112**, 7153–7158 (2015).
52. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
53. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
54. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047-8 (2015).
55. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (Springer-Verlag, New York, 2013).
56. Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu. Rev. Genomics Hum. Genet.* **14**, 301–323 (2013).
57. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).
58. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
59. Reimand, J. et al. g:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44**, W83–W89 (2016).
60. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
61. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
62. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, 1–20 (2020).
63. Liu, B., Gludemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).

Acknowledgements

The authors acknowledge the contribution of the Icelandic Heart Association (IHA) staff to AGES-Reykjavik, as well as the involvement of all study participants. The National Institute on Aging (NIA) contracts N01-AG-12100 and HHSN271201200022C for V.G. financed the study. V.G. received funding from the NIA (1R01AG065596-01A1), and IHA received a grant from Althingi (the Icelandic Parliament). The Icelandic Research Fund (IRF) funded V.E. and Va.G. with grants 195761-051, 184845-053, and 206692-051, while Va.G. received a postdoctoral research grant from the University of Iceland Research Fund.

Author contributions

A.G., Va.G., V.E., and Vi.G. designed the study. A.G., Va.G., G.T.A., E.F.G., B.G.J., L.J.L., and T.A. performed data analysis. J.R.L. and L.L.J. provided expertise on proteomics data and contributed to discussion. Vi.G. and V.E. supervised the project. A.G. and Va.G. wrote the first draft of the manuscript, with all coauthors contributing to data interpretation, manuscript editing, and revision.

Competing interests

The study was supported by the Novartis Institute for Biomedical Research, and protein measurements for the AGES-Reykjavik cohort were performed at SomaLogic. J.R.L. and L.L.J. are employees and stockholders of Novartis. All other authors have no conflict of interests to declare.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-27850-z>.

Correspondence and requests for materials should be addressed to Vilmundur Gudnason.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022