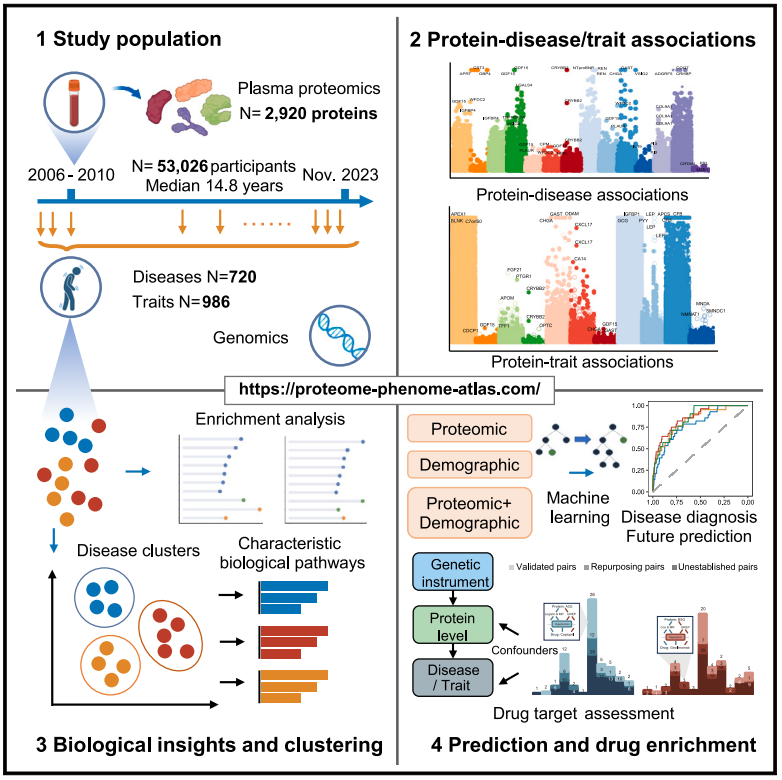
Cell

Resource

Atlas of the plasma proteome in health and disease in 53,026 adults

Graphical abstract



Highlights

- Construct a comprehensive proteomics atlas for 1,706 human diseases and traits
- Machine-learning-based big data uncover promising diagnostic and predictive biomarkers
- Identify 37 drug repurposing prospects and 26 potential targets with good safety
- Provide an open-access proteome-phenome resource to advance precision medicine

Authors

Yue-Ting Deng, Jia You, Yu He, ..., Jian-Feng Feng, Wei Cheng, Jin-Tai Yu

Correspondence

maoying@fudan.edu.cn (Y.M.),
jianfeng64@gmail.com (J.-F.F.),
wcheng@fudan.edu.cn (W.C.),
jintai_yu@fudan.edu.cn (J.-T.Y.)

In brief

A large-scale proteomics study involving 53,026 individuals maps 2,920 plasma proteins to 406 prevalent diseases, 660 incident diseases, and 986 health-related traits, identifying promising biomarkers for disease discrimination and potential therapeutic targets and paving the way for precision medicine.

Deng et al., 2025, Cell *188*, 253–271
January 9, 2025
https://doi.org/10.1016/j.cell.2024.10.045

CellPress

Cell

CellPress
OPEN ACCESS

Resource

Atlas of the plasma proteome in health and disease in 53,026 adults

Yue-Ting Deng,[1,9] Jia You,[1,2,3,9] Yu He,[1,9] Yi Zhang,[1,9] Hai-Yun Li,[1,9] Xin-Rui Wu,[1,9] Ji-Yun Cheng,[1,9] Yu Guo,[1,9] Zi-Wen Long,[4,5,9] Yi-Lin Chen,[1,9] Ze-Yu Li,[1,2,3] Liu Yang,[1] Ya-Ru Zhang,[1] Shi-Dong Chen,[1] Yi-Jun Ge,[1] Yu-Yuan Huang,[1] Le-Ming Shi,[6] Qiang Dong,[1] Ying Mao,[7,*] Jian-Feng Feng,[2,3,8,10,*] Wei Cheng,[1,2,3,*] and Jin-Tai Yu[1,*]

[1]Department of Neurology and National Center for Neurological Disorders, Huashan Hospital, State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Shanghai Medical College, Fudan University, Shanghai, China
[2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[3]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Ministry of Education, Shanghai, China
[4]Department of Gastric Cancer Surgery, Fudan University Shanghai Cancer Center, Shanghai, China
[5]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China
[6]State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China
[7]Department of Neurosurgery, Huashan Hospital Fudan University, Shanghai, China
[8]Department of Computer Science, University of Warwick, Coventry, UK
[9]These authors contributed equally
[10]Lead contact
*Correspondence: maoying@fudan.edu.cn (Y.M.), jianfeng64@gmail.com (J.-F.F.), wcheng@fudan.edu.cn (W.C.), jintai_yu@fudan.edu.cn (J.-T.Y.)
https://doi.org/10.1016/j.cell.2024.10.045

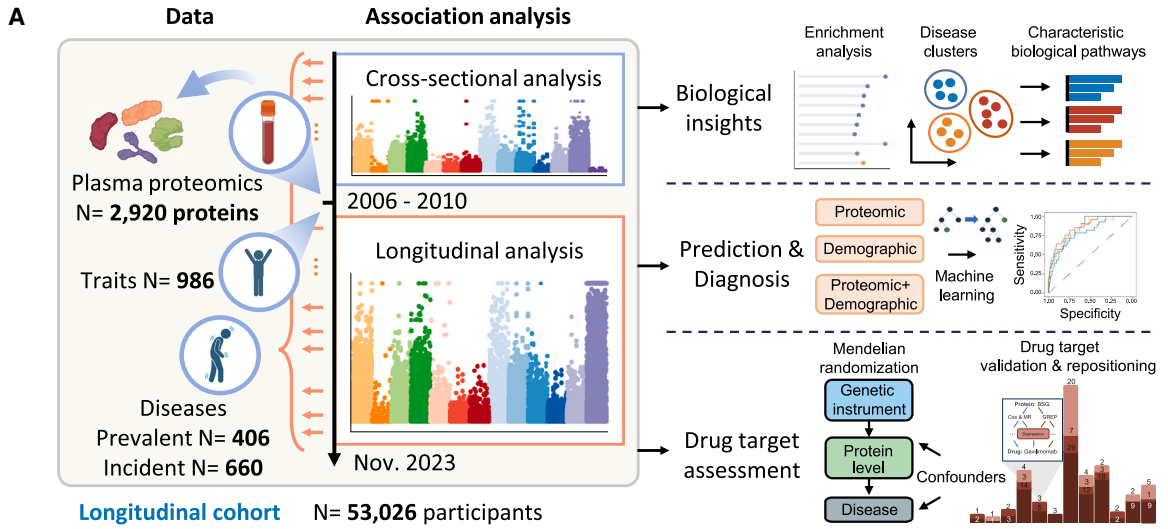
SUMMARY

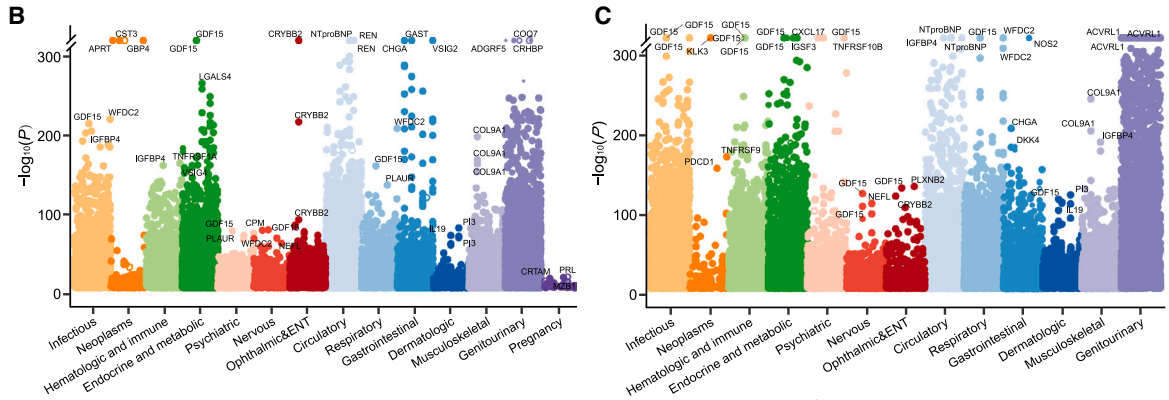
Large-scale proteomics studies can refine our understanding of health and disease and enable precision medicine. Here, we provide a detailed atlas of 2,920 plasma proteins linking to diseases (406 prevalent and 660 incident) and 986 health-related traits in 53,026 individuals (median follow-up: 14.8 years) from the UK Biobank, representing the most comprehensive proteome profiles to date. This atlas revealed 168,100 protein-disease associations and 554,488 protein-trait associations. Over 650 proteins were shared among at least 50 diseases, and over 1,000 showed sex and age heterogeneity. Furthermore, proteins demonstrated promising potential in disease discrimination (area under the curve [AUC] > 0.80 in 183 diseases). Finally, integrating protein quantitative trait locus data determined 474 causal proteins, providing 37 drug-repurposing opportunities and 26 promising targets with favorable safety profiles. These results provide an open-access comprehensive proteome-phenome resource (https://proteome-phenome-atlas.com/) to help elucidate the biological mechanisms of diseases and accelerate the development of disease biomarkers, prediction models, and therapeutic targets.

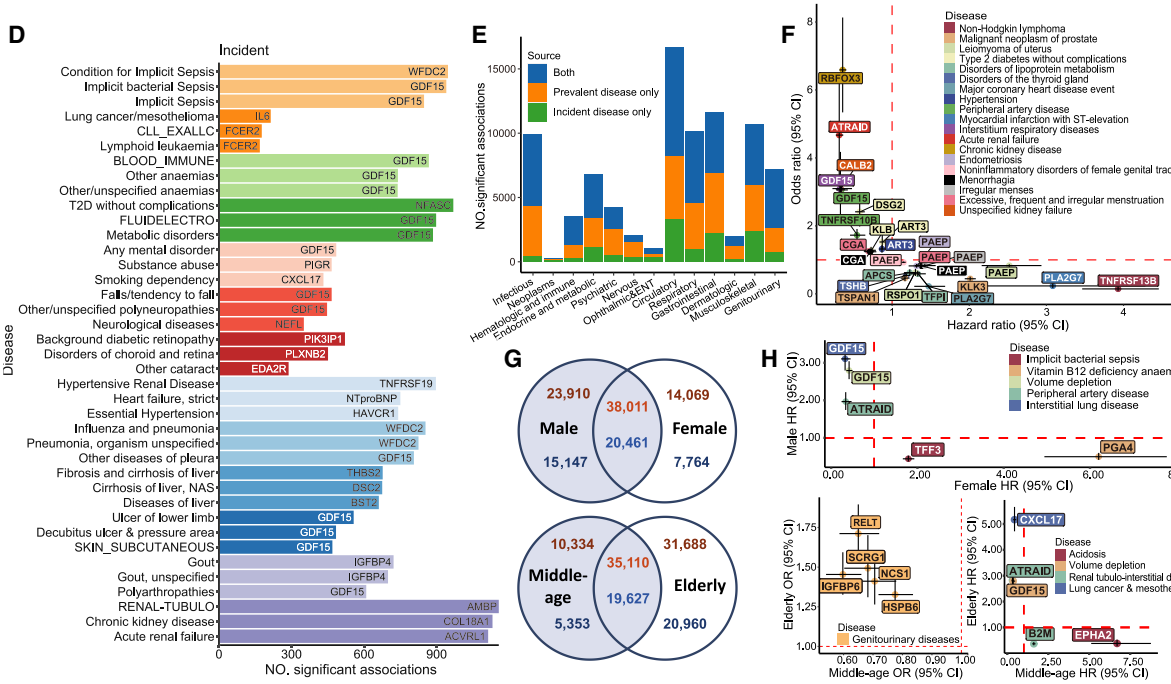
INTRODUCTION

With the population worldwide rapidly growing and aging, the demand to enhance health and alleviate disease burden is on the rise.[1] The challenges of disease prevention and treatment include the lack of reliable prediction models of individualized risk and variation in efficacy and adverse effects of existing treatments, which emphasizes the importance of precision medicine.[2,3] Currently, the implementation of precision medicine mainly focuses on identifying genomic underpinnings for human diseases and has shown initial effectiveness.[4–7] However, the complex and uncertain regulatory processes in the transcription and translation of genes obstructed the inference of causal genes, thus limiting mechanistic understanding and drug development based on genome-to-phenome associations.[8–10] Proteins, the ultimate biological effectors for genetic and environmental risk of diseases, directly reflect the biological processes and the pathophysiological changes in the human body. Elucidating protein-disease relationships holds the promise to characterize the biological signatures of different health states and disease conditions,[11,12] facilitating precision medicine with increased convenience and feasibility.

Technological advances in high-throughput proteomics have provided a remarkable opportunity to systematically interrogate the protein profiles of health states and diseases, facilitating mechanistic understanding,[13] biomarker identification,[14] risk prediction,[15] early detection of adverse drug reactions,[16] and aging.[17] Nevertheless, most current proteomics studies have only focused on limited disease outcomes.[18,19] While these studies revealed some disease-specific proteomic changes, the lack of a comprehensive human proteome-phenome atlas also raises many questions. For example, are the associated







(legend on next page)

proteins specific or shared among diseases and health-related traits, and can those proteomic profiles facilitate biological classification of human diseases? How do those plasma proteins contribute to minimally invasive assessment and tracking of hundreds of diseases? Are those associated proteins causally related to diseases, and do those causal proteins hold the potential to be therapeutic targets? Answering these questions is challenging, as the complexity of proteomics and phenomics has so far impeded a profound knowledge of human disease and health.

Here, we present a comprehensive atlas of proteome-phenome associations (https://proteome-phenome-atlas.com/) by systematically mapping 2,920 plasma proteins to the presence and onset of 720 diseases and 986 health-related traits in 53,026 individuals. This atlas provides insights into shared and characteristic biological mechanisms among diseases. The proteomic profiles, coupled with machine learning, identify useful biomarkers and prediction models for multiple health conditions simultaneously. Through integrating protein quantitative trait locus (pQTL) data, we illustrate the use of the atlas for causal protein discovery and further drug target prioritization. Our proteome-phenome atlas furnishes an extensive resource supporting future research in screening, diagnosis, and treatment of human diseases. The overall analytical workflow is presented in Figures 1A and S1.

RESULTS

Population characteristics and phenotypes

We included 53,026 participants with an average age of 56.8 years, comprising 53.9% women and 93.7% white individuals (Table S1). 2,920 proteins that satisfied quality control criteria were included in the subsequent investigation (Table S2). Two main categories of phenotypes comprising diseases and health-related traits were included. Prevalent disease endpoints were binary outcomes with more than 100 events before and at blood collection. Incident disease endpoints were organized as time-to-event data, with more than 100 events after blood collection during a median of 14.8 years of follow-up (Tables S3 and S4). Health-related traits contained continuous, binary, and ordered categorical variables processed by the PEACOK package[20] (Table S5).

In total, we incorporated 406 prevalent disease endpoints, 660 incident disease endpoints, and 986 health-related traits. Prevalent diseases were then categorized into 14 chapters, among which digestive diseases comprise the largest proportion (17.2%) (Figure S2A). The average number of prevalent disease cases per chapter varied between 286 and 865, with circulatory diseases ranking the highest (Figure S2B). Incident diseases were classified into 13 chapters (Figure S2A). The average number of incident disease cases per chapter varied between 483 and 1,508, of which circulatory diseases also ranked the highest (Figure S2B). Traits were classified into 11 chapters according to UK Biobank (UKB) paths, among which nuclear magnetic resonance (NMR) spectroscopy-derived metabolomic profiles accounted for the greatest proportion, amounting to 25.5% (Figure S2C). The average sample sizes for traits across chapters varied from 17,880 to 49,267 (Figure S2D).

Atlas of protein-disease associations

We first sought to understand the relationship between circulating levels of 2,920 proteins and 406 prevalent diseases and 660 incident diseases using logistic regression and Cox proportional hazards regression models, respectively (STAR Methods). We identified 60,942 protein-prevalent disease pairs that were significantly correlated at a stringent Bonferroni-corrected threshold of $p < 4.21 \times 10^{-8}$ ($p < 0.05/[2,920*406]$) (Figure 1B). Furthermore, 107,158 significant protein-incident disease associations were observed at a Bonferroni-corrected threshold of $p < 2.59 \times 10^{-8}$ ($p < 0.05/[2,920*660]$) (Figure 1C). As expected, well-established associations, including NTproBNP related to death due to cardiac causes[21] and GDF15 with diabetes,[22] were among the most significant protein-prevalent disease associations. WFDC2 was linked to the risk of incident respiratory diseases such as influenza and pneumonia, and GDF15 was linked to certain infectious and blood system diseases, including implicit sepsis and anemias, validating prior research[23–26] and confirming validity of our approach (Figure 1D). Notably, our

Figure 1. Summary of protein-disease association analysis results

(A) Schematic workflow of analyses. Data on plasma proteins and health-related traits were collected at baseline (2006–2010), while data on diseases were linked to UK electronic health records with detailed time of diagnosis spanning before and after baseline. Association analyses were performed based on cross-sectional data and time-to-event data, respectively, to uncover proteomic profiles of diverse phenotypes, followed by an in-depth exploration of biological insights, utility in prediction and diagnosis, and drug target assessment. Created in BioRender (BioRender.com/o97h873).

(B and C) Protein-disease associations revealed by (B) logistic regression and (C) Cox regression, colored by disease chapter. Only significant associations were plotted (B, $n = 60,942$; C, $n = 107,158$). Filled dots refer to positive associations (hazard ratios [HRs] > 1), while open dots indicate negative associations (HR < 1). ENT , ear, nose, and throat.

(D) Top three incident diseases with the largest number of significant associations in each disease chapter, colored by disease chapter. The proteins on the bars have the minimum p value for corresponding disease. CLL_EXALLC, chronic lymphocytic leukemia and small lymphocytic leukemia, excluding all cancers (controls excluding all cancers); BLOOD_IMMUNE, diseases of blood, blood-forming organs, and immune system; T2D, type 2 diabetes; FLUIDELECTRO, other disorders of fluid, electrolyte, and acid-base balance; NAS, unspecified; SKIN_SUBCUTANEOUS, diseases of the skin and subcutaneous tissue; RENAL-TUBULO, renal tubulo-interstitial diseases.

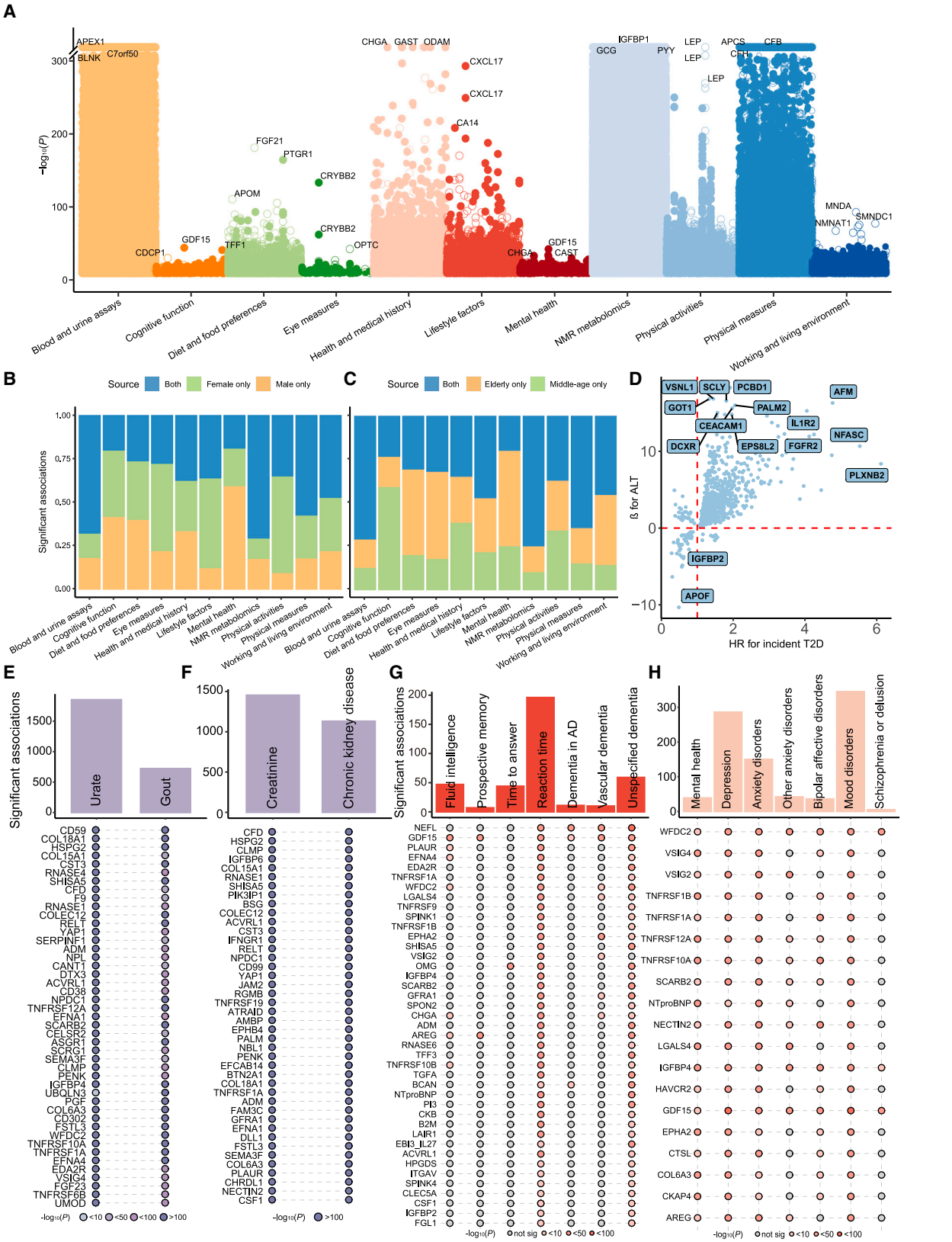
(E) Comparisons of number of significant associations in prevalent and incident diseases.

(F) Protein-disease pairs with inconsistent directions in associations with prevalent and incident diseases, colored by disease. Error bar represents 95% confidence interval (CI).

(G) Comparisons of number of significant associations in sex (top) and age (bottom) subgroup analysis. The red font represents number of significant associations in longitudinal analysis, and the blue font represents that in cross-sectional analysis.

(H) Protein-disease pairs with inconsistent directions in sex (top) and age (bottom) subgroup analysis, colored by disease. Error bar represents 95% CI.

See also Figure S1.



(legend on next page)

results also unveiled protein-disease associations that have not been reported before. The top associations were mainly found in incident genitourinary diseases such as chronic kidney disease, containing both previously reported protein biomarkers[27] and unreported ones with high hazard ratios (HRs) including NBL1 (HR[95% confidence interval (CI)] = 17.055[15.566–18.686], $p < 1 \times 10^{-300}$), COLEC12 (HR [95% CI] = 16.320[14.985–17.774], $p < 1 \times 10^{-300}$), and others. Furthermore, we found 1,977 associations that were protective for both prevalent and incident diseases (Data S1). Among these proteins, EGFR exhibited the most extensive and significant protective effects, impacting 90 diseases. The largest protective effect was on hypertensive renal disease (odds ratio [OR][95% CI] = 0.020[0.009–0.044], $p = 5.97 \times 10^{-23}$; HR[95% CI] = 0.166[0.109–0.253], $p = 6.50 \times 10^{-17}$), supporting the pivotal role of EGFR signaling in kidney damage[28] (Data S1).

We then compared the protein rankings and direction of association between prevalent and incident diseases. Most protein-disease links were concurrently observed in both analyses (Figure 1E). We ranked the proteins based on their p values and calculated the number of diseases in which each protein had a first-place ranking. Among the top ten proteins with the highest number of first-place rankings, six were shared between prevalent and incident diseases (GDF15, WFDC2, NTproBNP, CHGA, COL9A1, and IGFBP4), indicating that the important proteins change both before and after disease onset (Data S1). Furthermore, the majority of protein-disease associations demonstrated consistent effects between prevalent and incident diseases, while 27 proteins exhibited different effects on prevalent diseases and incident diseases (Figure 1F). For instance, patients with prevalent type 2 diabetes (T2D) displayed higher DSG2, ART3, and KLB levels (OR[95% CI] = 2.415[1.982–2.943], 1.527[1.313–1.776], and 1.282[1.188–1.384], respectively), while those proteins were identified as protective factors for incident risk of T2D (HR[95% CI] = 0.586[0.527–0.652], 0.734 [0.676–0.796], and 0.879[0.844–0.915], respectively). As an example, DSG2, which is involved in cell adhesion and signaling, may initially protect islet cells and aid insulin secretion, but its elevated levels as T2D progresses might indicate a compensatory response to insulin resistance (Data S1). Thus, convergent protein-prevalent and protein-incident disease associations might highlight an important role across disease stages, whereas divergent associations provide additional insight into protein function in disease pathogenesis.

In sensitivity analysis, 80.5% of protein-prevalent disease and 74.9% of protein-incident disease associations remained significant (Bonferroni-corrected to 2,920 proteins at $p < 1.71 \times 10^{-5}$) when restricting the controls and further adjusting for comorbidity status for each prevalent and incident disease endpoint (STAR Methods). Meanwhile, additionally adjusting for age^2, age*sex, age^2*sex, and the first 10 genetic principal components (PCs) resulted in minimal changes in protein-disease associations, with 99.9% of protein-prevalent disease and 75.8% of protein-incident disease associations remaining significant. We performed subgroup analyses by sex and age (middle-aged: <60 and elderly: ≥60 years). More than half of the associations remained significant and exhibited consistent effect directions with the main analysis. Meanwhile, sex-specific associations were revealed, with 37,979 and 22,911 found in protein-incident and protein-prevalent disease associations, respectively (Figure 1G). The majority of associations maintained consistent directions across subgroups, and only 18 associations exhibited divergent effect directions in subgroups (Figure 1H).

Atlas of protein-trait associations

We next investigated the associations between proteins and 986 health-related traits. We identified 554,488 significant protein-trait associations involving 2,707 proteins and 782 traits after Bonferroni correction ($p < 1.71 \times 10^{-8}$) (Figure 2A). Findings in the protein-trait association analysis may further support the identified protein-disease associations. For instance, we found that GDF15 and CDCP1 exhibited the most significant associations with cognitive function traits. Notably, GDF15 (β[95% CI] = 14.464[12.423–16.506], $p = 9.32 \times 10^{-44}$) and CDCP1 (β [95% CI] = 7.997[6.248–9.745], $p = 3.29 \times 10^{-19}$) were both risk proteins for reaction time, providing additional evidence of their positive correlations with neuropsychiatric disorders observed in the protein-disease association analyses. In the sensitivity analysis adjusting for more covariates, 99.3% of the initially identified associations remained significant (STAR Methods).

In subgroup analysis, over 70% protein-trait associations remained significant. Protein associations with cognitive function and mental health-related traits presented higher subgroup specificity (Figures 2B and 2C). The directions of the shared associations across subgroups were largely consistent, except for 235 proteins that showed different effect directions by sex and 164 proteins by age (Data S1). For instance, OXT, a neuropeptide involved in social behavior and health,[29] showed a positive effect on testosterone levels in females (β[95% CI] = 0.136 [0.104–0.169], $p = 2.72 \times 10^{-16}$) while negative in males (β[95% CI] = −0.204[−0.240 to −0.167], $p = 8.06 \times 10^{-28}$).

Figure 2. Summary of protein-trait association analysis results and pleiotropy among diseases and traits

(A) Protein-trait associations, colored by trait category. Only significant associations are plotted (n = 554,488). Filled dots refer to positive associations (β > 0), while open dots indicate negative associations (β < 0).

(B and C) Proportion of shared and specific significant associations in (C) sex and (D) age subgroup analysis for traits.

(D) Shared proteins between ALT and incident T2D. Proteins with top ten p values or effect sizes in both ALT and T2D were labeled. ALT, alanine aminotransferase; T2D, type 2 diabetes.

(E and F) Shared proteins between (E) urate and gout and (F) creatinine and chronic kidney disease. The displayed proteins are those with top 30 p values and effect sizes in traits and diseases.

(G) Shared proteins between cognitive function traits and incident dementia subtypes. The displayed proteins are those significant with at least one disease and one trait.

(H) Shared proteins between mental health traits and incident psychiatric disorders. Only proteins associated with more than five diseases or traits were displayed. The shade of color represents the magnitude of p values.

(legend on next page)

The landscape of pleiotropy in protein-phenotype associations

Considering that the vast majority of proteins exhibited multi-phenotypic associations, we then focused on pleiotropic proteins due to their potential as prospective clinical targets.[30] There were 434 proteins (26.3% = 434/1,648) linking to over 50 prevalent diseases and 649 proteins (32.2% = 649/2,013) linking to over 50 incident diseases, which encompassed well-studied ones with versatile biological functions, such as GDF15, WFDC2, and tumor necrosis factor (TNF) family (Data S1). GDF15 was associated with the most diseases, containing 205 prevalent and 397 incident diseases, generally acting as a risk factor except for three incident diseases (respiratory diseases affecting the interstitium, disorders of magnesium metabolism, and peripheral artery disease). The TNF family, which mainly involves in inflammation and cellular differentiation,[31] had pronounced pleiotropy among various diseases. For instance, EDA2R, a type III transmembrane protein of this family, showed associations with incident circulatory system diseases ($n = 54$), followed by musculoskeletal ($n = 43$), digestive ($n = 35$), and endocrine ($n = 35$) system diseases. Additionally, this protein was closely linked to infectious diseases, with a p value as high as 6.92×10^{-266} for implicit sepsis.

In protein-trait associations, 365 proteins exhibited more than 300 significant associations. Notably, GDF15, similar to its high ranking with protein-disease pleiotropy, ranked second for protein-trait associations, boasting a substantial 428 associations. The majority of its most significant associations were with NMR metabolomics, especially lipid metabolites. Our observations were consistent with existing literature that has reported potential mechanisms by which GDF15 influences appetite, food intake, and gastric emptying,[22] and subsequently affecting lipid metabolism. These results suggested an extensive role of GDF15 involving in the pathogenesis of lipid-related outcomes encompassing circulatory, endocrine, and metabolic diseases.[32–34]

We then investigated whether specific proteins influenced certain diseases and disease-related traits simultaneously, focusing on three chapters including metabolic diseases and NMR metabolomics, dementia and cognitive function, and psychiatric diseases and mental health. We found specific proteins that exhibited a protective association with the disease as well as a favorable effect on the trait. For instance, IGFBP2 was correlated with lower levels of alanine aminotransferase (ALT) (β [95% CI] = −3.746[−3.909 to −3.583], $p < 1 \times 10^{-300}$), and lower risk of T2D (HR[95% CI] = 0.621[0.593–0.650], $p = 1.69 \times 10^{-93}$) (Figure 2D). IGFBP2 is a known biomarker of insulin sensitivity[35] and was confirmed as a protective protein for T2D in a longitudinal cohort.[36,37] Given that elevated levels of ALT have been genetically linked to an increased risk of T2D,[38,39] our observation revealing IGFBP2's favorable effect on ALT may further consolidate its protective role in T2D. Other metabolites and related diseases were also found to have overlap proteins, including urate and gout, and creatine and chronic kidney disease (Figures 2E and 2F). The fluid intelligence score and various types of dementia shared significant proteins such as NEFL and GDF15 (Figure 2G), further supporting the close relationships between these proteins and cognitive function. Mental disorders like anxiety disorders, depression, and mood disorders, along with mental health-related traits, also exhibited substantial protein similarities, including TNFRSF10A, GDF15, IGFBP4, WFDC2, and others (Figure 2H). Notably, IGFBP4 has been identified as a blood biomarker of mood disorders,[40] corroborating our findings.

Biological function of the disease-associated proteins

To refine our understanding of how the identified proteins participate in human diseases, we conducted a series of functional enrichment analyses of the identified proteins. Among the 660 incident diseases, 539 showed significant enrichment in at least one Reactome pathway, and the specific enriched pathways for each disease can be found on our website. Pathways related to the immune system were mostly enriched across human diseases, especially in the infectious and parasitic diseases and diseases of the blood and blood-forming organs, circulatory system, and respiratory system (Figure 3A). Specifically, the binding of TNFs to their physiological receptors was the most frequent pathways related to the immune system that participated in over half of diseases across various systems except for the nervous system. This is consistent with the broad pleiotropy of TNF family member proteins found in our protein-prevalent disease and protein-incident disease associations, emphasizing the important role of inflammation in human health. Pathways related to metabolism of proteins, including post-translational phosphorylation and regulation of insulin-like growth factor, were also enriched in a considerable proportion of diseases.

Comparison of biological pathways between different diseases refined our understanding of similarities and heterogeneity in disease pathophysiology. For example, we found that proteins related to Alzheimer's disease (AD) and vascular dementia (VaD) were enriched in shared pathways related to the nervous system, which included synapse maturation, neuron projection regeneration, intermediate filament organization, positive regulation of Schwan cell proliferation, and regulation of axon diameter (Figure 3B). Meanwhile, AD-specific pathways were mostly related to lipid metabolism, including regulation of lipid transport across the blood-brain barrier and intermediate-density lipoprotein particle clearance, while VaD-specific pathways were related to cardiac muscle, including the adenylate cyclase-activating adrenergic receptor signaling pathway and positive regulation of relaxation of cardiac muscle.

Disease clusters with characteristic biological features

We applied hierarchical clustering based on magnitudes of protein-disease associations (i.e., HRs) and grouped the 660 diseases into 40 clusters (Table S6). As expected, diseases with

Figure 3. Biological functions of the disease-associated proteins

(A) The results of Reactome pathway enrichment analysis classified by chapters of incident diseases, colored by the frequency of the item ranking top ten in each chapter.

(B) Comparison between the results of Gene Ontology (GO) biological process (BP) enrichment of proteins related to dementia in AD and VaD. All of these outcomes met the threshold of false discovery rate (FDR) < 0.05. The shared (top) and specific (bottom) pathways are displayed.

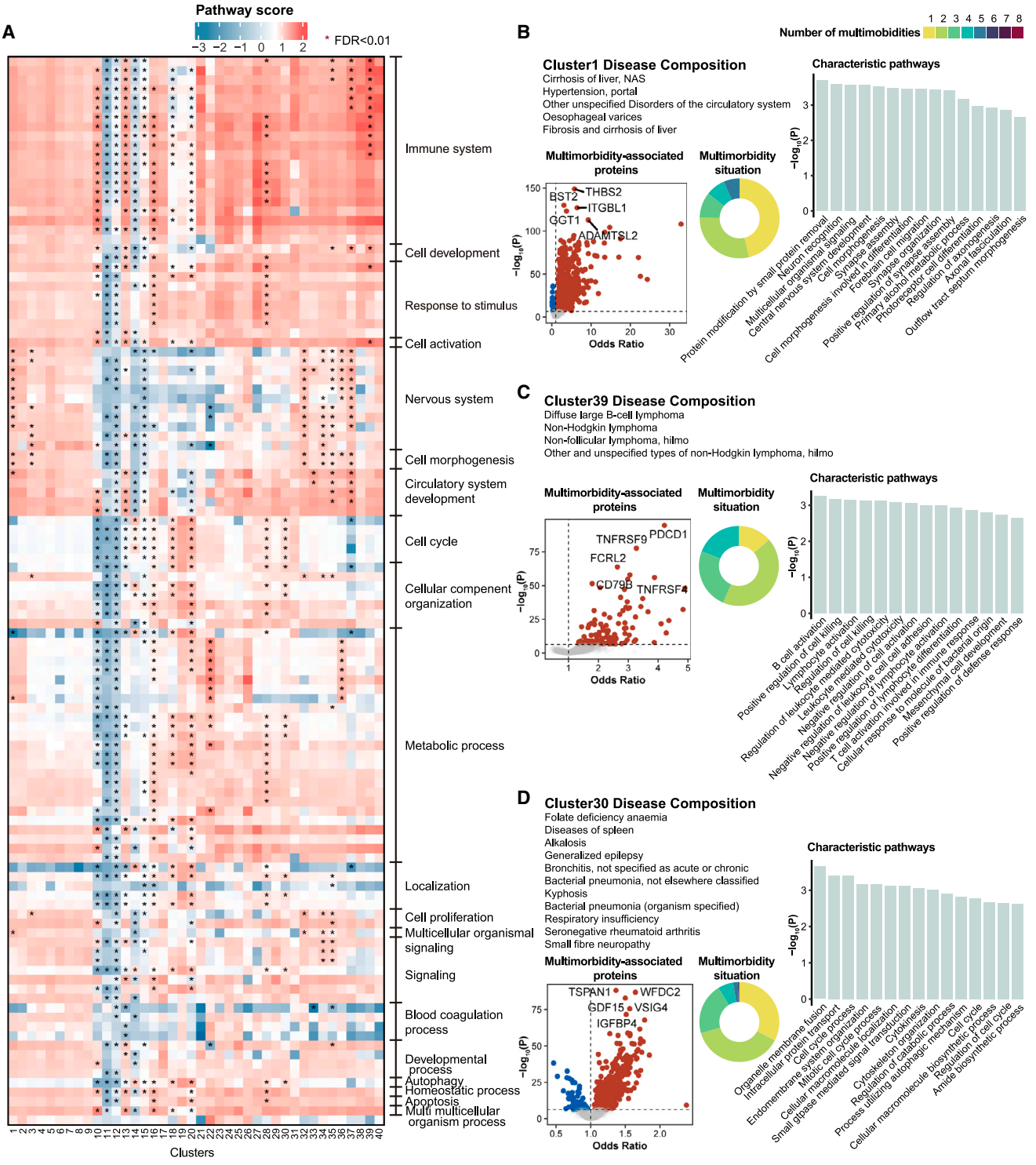


Figure 4. Characteristic biological features of 40 disease clusters and selected examples

(A) Proteomic classification of 660 diseases into 40 clusters. The heatmap shows the enrichment of GO BP gene sets in disease clusters detected by gene set enrichment analysis. Color of heatmap indicates the mean value of pathway scores for diseases in this cluster. Asterisk* depicts pathways differently regulated between one cluster and others (FDR < 0.01).

(B–D) Examples of clusters 1, 39, and 30 show the disease composition, multimorbidity situation, proteins associated with multimorbidity in this cluster, and a cluster's specific characteristic pathways in (A). The text in the upper left corner records the specific diseases included in this cluster. The volcano plot shows the multimorbidity-associated proteins, which were detected by ordinal regression with the number of incident diseases in this cluster as outcome.

(legend continued on next page)

similarities were grouped and exhibited characteristic biological features (Figure 4A). For example, cluster 1 included fibrosis and cirrhosis of liver and its complications, such as portal hypertension and esophageal varices (Figure 4B). These disorders were primarily enriched in protein modification by small protein removal, alcohol metabolism, and pathways involving the nervous system and cell morphogenesis, suggesting the potential mechanisms including deubiquitination,[41] alcohol consumption,[42] and epithelial-mesenchymal transition[43] and its impact on the nervous system.[44] Cluster 39, comprising of subtypes of non-Hodgkin lymphoma, was characterized by pathways including B cell activation (Figure 4C). Cluster 32, which mainly included neurological complications resulting from diabetes (Table S6), shows a substantial impact on pathways involving the nervous system (Data S1).

Clusters 10, 11, and 12 were distinct from others, as most pathways were downregulated, especially in cell cycle, cellular component organization, metabolic processes, localization, signaling, and autophagy. These pathways were associated with diseases in these clusters, like neurodegenerative diseases,[45–47] aortic aneurysm,[48] and obesity.[49,50] Intriguingly, cluster 11, including carcinoma *in situ* of the breast and other diseases common in women, showed downregulated pathways, which may indicate gene expression being reversed at different disease stages,[51] and further research is needed to elucidate the mechanisms.

Pathways involving immune system, cell development, response to stimulus, and cell activation exhibited consistent changes across most clusters (Figure 4A), which might be activated in most disorders.[52–60] Despite the consistent directional changes, different magnitudes of pathway scores among clusters imply cluster specificity. It is noteworthy that Figure 4A emphasized pathway significance through inter-cluster comparisons. Certain pathways, if consistent across most diseases, were not highlighted, even if they might be upregulated in specific diseases.

Intriguingly, 60% of clusters contained diseases from more than one disease category. Taking cluster 30 as an example (Figure 4D), it included diseases from the blood and blood-forming system, nervous system, respiratory system, musculoskeletal system, and connective tissue. Its characteristic pathways included protein transport, cell cycle processes, small GTP hydrolase (GTPase)-mediated signal transduction, catabolic processes, autophagy, and amide biosynthetic processes. This provides preliminary biological insights for re-understanding disease classification from biological profiles.

We calculated the multimorbidity status of each cluster, that was, the number of incident diseases in each cluster for each individual (STAR Methods). Then, ordinal regression models were applied to investigate the proteins associated with multimorbidity level (Table S7). For 36 clusters, more than half of proteins associated with multimorbidity level also significantly correlated with at least one disease within the cluster in the longitudinal analysis. This reflects the similar biological characteristics of diseases within the cluster from a population perspective and highlights the importance of the shared proteins.

Proteins contribute to disease diagnosis and prediction

By modeling the disease risk for each endpoint, we investigated the diagnostic and predictive value of proteins, demographics, and their integration (Tables S8 and S9). For disease prediction, the protein-based model demonstrated good areas under the curve (AUCs) exceeding 0.80 for 92 diseases (13.9% = 92/660) across 13 disease categories, with most found in endocrine and metabolic (n = 18 out of 42) and circulatory (n = 17 out of 65) diseases (Figure 5A). Of particular note, the protein-based model yielded excellent predictions (AUCs > 0.9) for 9 diseases, e.g., T2D with peripheral circulatory complications (AUC = 0.974 [0.963–0.982]), hypertensive renal disease (AUC = 0.951[0.934–0.967]), chronic nephritic syndrome (AUC = 0.925[0.899–0.946]), dialysis (AUC = 0.923[0.894–0.950]), and background diabetic retinopathy (AUC = 0.919[0.905–0.933]) (Table S8). The protein-based model showed significantly better accuracies than the demographic model in predicting 361 (54.7% = 361/660) specific diseases. Particularly in the prediction for diabetic nephropathy, coeliac disease, and hyperparathyroidism, the protein-based model (AUCs: 0.829∼0.885) demonstrated remarkable superiority over the demographic-based model (AUCs: 0.541∼0.616) but performed comparably to the model integrating both proteins and demographics (AUCs: 0.829∼0.887, $p_{\text{Delong test}} > 0.05$). When predicting diseases like constipation and diaphragmatic hernia, the protein-based model performed significantly better than the demographic model ($p_{\text{Delong test}} < 1 \times 10^{-4}$), although the AUC improvement ($\approx$0.02) was modest. Furthermore, adding plasma proteins to demographics substantially improved the predictive accuracies for 417 (63.2%) diseases with $p_{\text{Delong test}} < 0.05$ against demographic-based models (Table S8).

For disease diagnosis, the protein-based model showcased high AUCs surpassing 0.80 for 124 (30.5% = 124/406) diseases across 14 disease categories, with diseases of circulatory (n = 26 out of 37) and endocrine and metabolic (n = 12 out of 15) consistently showing good performance for disease diagnosis. In addition, the protein-based model achieved excellent AUCs beyond 0.9 for the diagnosis of 36 diseases, especially in type 1 diabetes (T1D), diabetic maculopathy, chronic kidney disease, T2D, hypertensive renal disease, myocardial infarction, and background diabetic retinopathy (AUCs: 0.952∼0.975) (Figure 5B). These results underscored the superior discriminative performance of the protein-based model when compared with the demographic-based model (AUCs in demographic-based models: 0.684∼0.840, $p_{\text{Delong test}}$=8.94 × 10^{-90}∼8.95 × 10^{-15}) (Table S9). Regarding the diagnosis of diseases like nerve, nerve root and plexus disorders, genitourinary diseases, and soft tissue disorders, the protein-based model performed significantly better than the demographic model ($p_{\text{Delong test}} < 1 \times 10^{-3}$),

Bonferroni-adjustment (0.05/[2,920 proteins*40 clusters]) was applied. The circular plot shows the proportion of individuals with different numbers of co-occurring diseases during follow-up, reflecting the multimorbidity situation of diseases in this cluster. The bar plot shows pathways that are significantly different between this cluster and others, which were adapted from (A).
See also Tables S6 and S7.

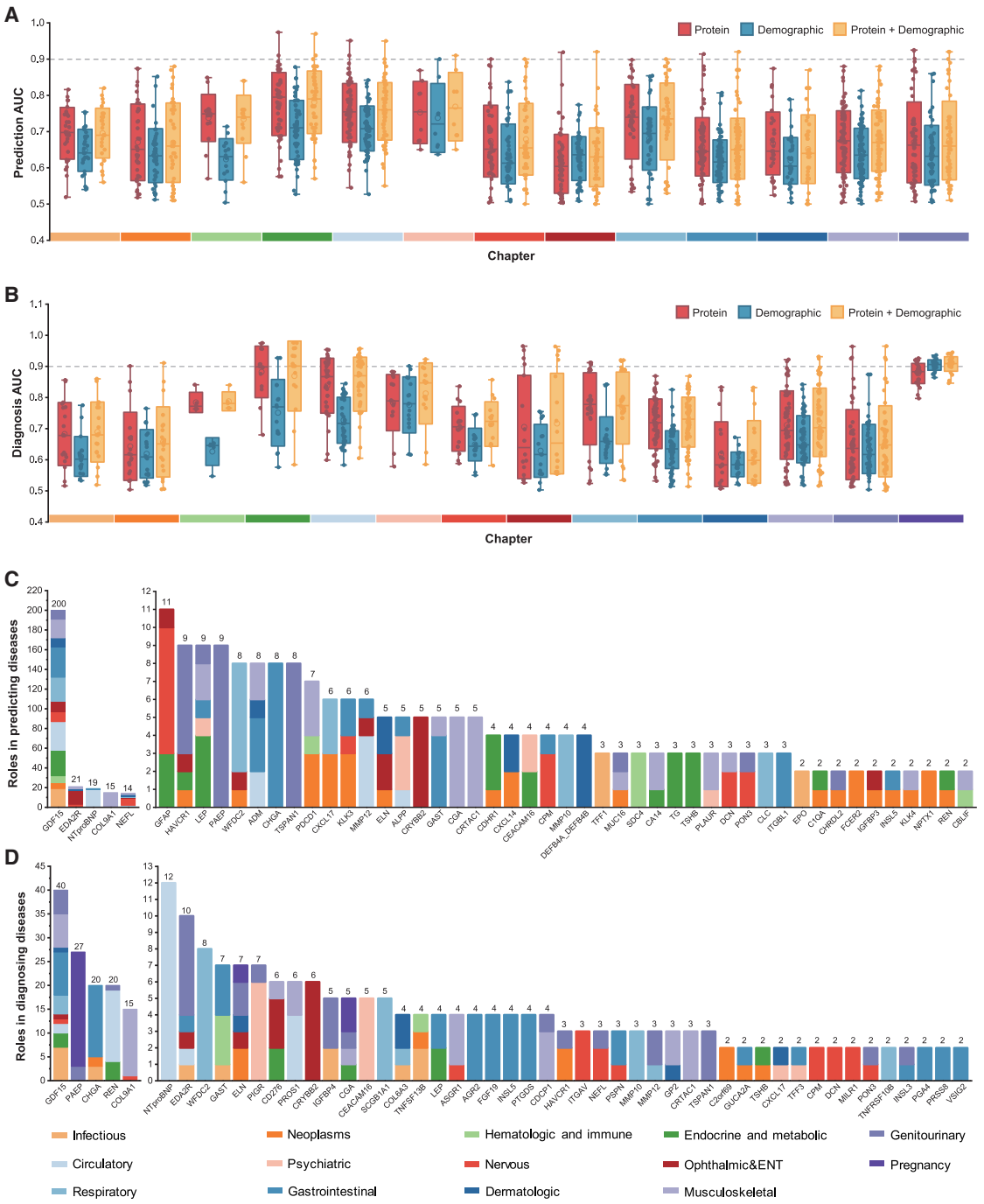
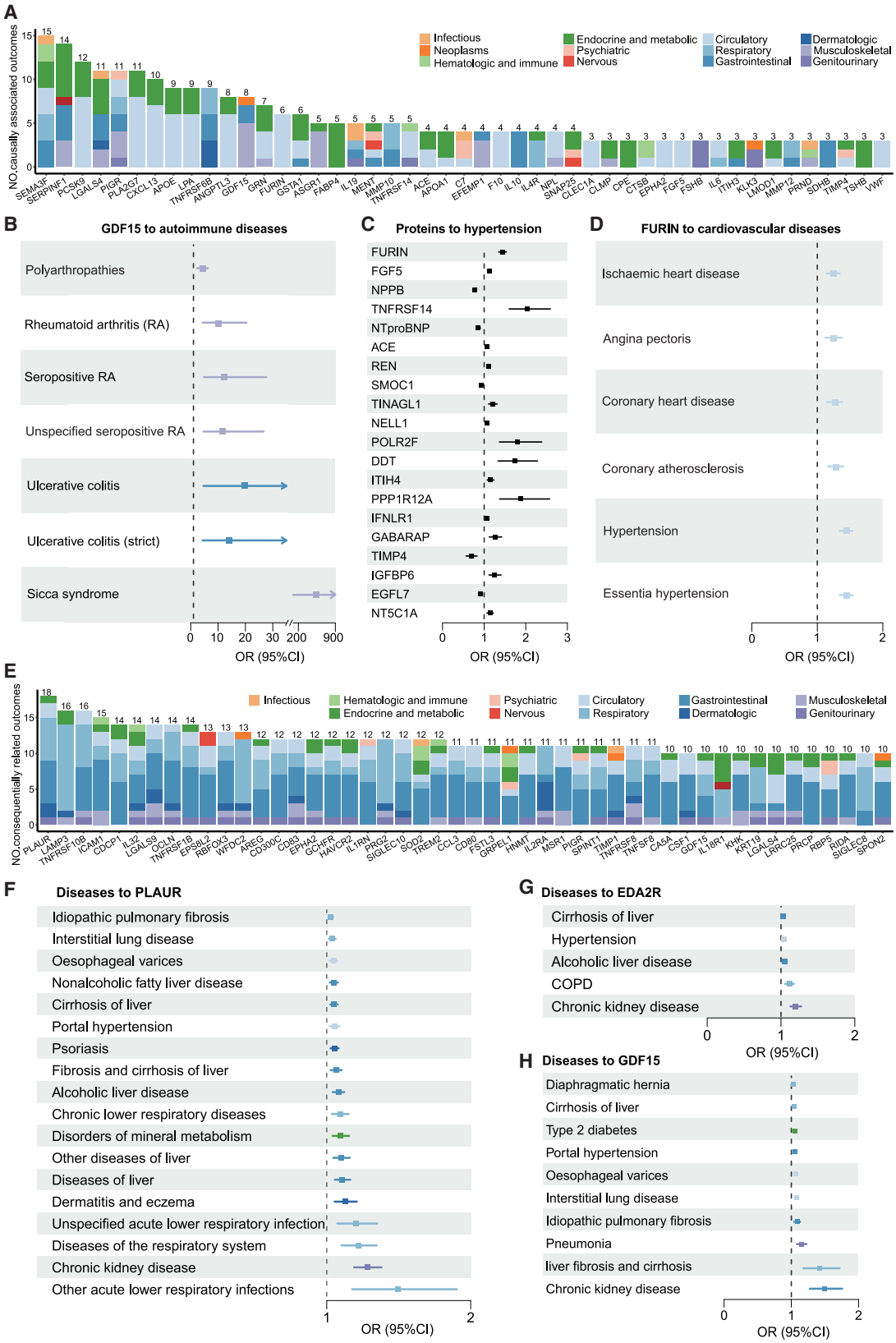


Figure 5. Contribution of proteins to disease prediction and diagnosis

(A and B) The discriminative performances of proteins in (A) prediction and (B) diagnosis (quantified by AUCs) based on three models: protein, demographic, and protein + demographic. Error bars represent the min-max, box ranges represent SD, and hollow circles represent mean values. Diseases with any AUCs < 0.5 derived from protein-only, demographic-only, or integrated models were excluded from the plot as they were considered uninformative. Specifically, 45 and 24 diseases were excluded for (A) (prediction) and (B) (diagnosis), respectively. AUC, area under the curve.

(C and D) Stacked bar chart of protein roles in (C) predicting and (D) diagnosing diseases, colored by disease chapter. Numbers above the bars represent the number of diseases in which the importance of this protein ranked first.

See also Tables S8 and S9.



(legend on next page)

although the AUC improvement (≈0.03) was modest. Furthermore, the protein-based model outperformed the demographic model in diagnosing 218 (53.7% = 218/406) specific diseases ($p_{Delong\ test} < 0.05$). Upon integrating plasma proteins into the demographic models, a notable enhancement in diagnostic accuracy was observed across 253 (62.3%) diseases with $p_{Delong\ test} < 0.05$ (Table S9).

For prediction and diagnosis, the added value of demographics on top of proteomics was not significant for 305 (46.2% = 305/660) and 185 (45.6% = 185/406) diseases respectively, such as interstitial lung disease, coeliac disease, and idiopathic pulmonary fibrosis. This suggests that substantial parts of the discriminatory information on demographics might be shared with proteomic data.[61] In addition, the demographic- and protein-based models exhibited poor performance (AUC < 0.5) in a small number of diseases (<5%) like benign neoplasms and diseases of sense organs, skin, and subcutaneous tissues (Data S1).

We calculated the importance of the plasma proteins in predicting and diagnosing diseases (Tables S8 and S9). This facilitated the identification of key discriminators (top 30) associated with each condition. The protein GDF15 attracted our attention due to its pivotal role in both predicting and diagnosing multiple diseases. Specifically, GDF15 stood out as the predictive protein with the highest ranking (top 1) among the largest number of diseases (n = 200). Following GDF15 were EDA2R, NTproBNP, COL9A1, and NEFL, each claiming the top position in 21, 19, 15, and 14 diseases, respectively (Figure 5C). As for disease diagnosis, GDF15 also emerged as the top-ranked protein across the largest number of diseases (n = 40), followed by PAEP, CHGA, REN, and COL9A1, which secured the top position among 27, 20, 20, and 15 diseases, respectively (Figure 5D). In addition, among the top ten proteins with the highest number of first-place rankings, five (i.e., GDF15, WFDC2, NTproBNP, EDA2R, and PAEP) overlapped between diagnosis and prediction models, emphasizing their favorable discriminative performance in both diagnosing and predicting diseases. Interestingly, six proteins (i.e., GDF15, WFDC2, NTproBNP, NEFL, COL9A1, and GFAP) among the top ten proteins in prediction models were also ranked at the top ten in protein-incident disease analyses, indicating consistency in the identification of disease biomarkers of Cox models and machine learning approaches.

Potential causal proteins of diseases

Taking advantage of extensive genetic information in UKB, we investigated whether the proteins associated with prevalent and incident diseases played causal roles in disease processes or were a consequence of a disease, clarifying which can contribute to the understanding of disease pathogenesis and the establishment of potential drug targets. Mendelian randomization (MR) analysis was conducted for the significant protein-disease associations using pQTL data and genome-wide association study (GWAS) summary data of diseases. In *cis*-MR analysis using *cis* pQTL as "exposure" and disease GWAS as "outcome," 178 protein-prevalent disease and 185 protein-incident disease pairs demonstrated potential causal relationships that attained an FDR-corrected p value less than 0.05 (equal to $p < 1.63 \times 10^{-4}$ and $p < 9.55 \times 10^{-5}$, respectively). The *trans*-MR analysis also identified 198 and 199 potential causal pairs from protein-prevalent disease and protein-incident disease associations, respectively. After excluding redundant pairs and pairs that were also significant in reverse MR (STAR Methods), we determined 474 unique potential causal protein-disease pairs, among which seven proteins showed ten or more potential causal pairs, including SEMA3F (n = 15), SERPINF1 (n = 14), and PCSK9 (n = 12) (Figure 6A).

These results provided causal evidence for the established protein-disease associations and found relevant genetic variants. For instance, GDF15, the protein with pleiotropic effects, was causally associated with several autoimmune diseases, including ulcerative colitis and rheumatoid arthritis (Figure 6B). An autoimmune pleiotropic SNP, rs4728142,[62] was associated in *trans* with higher plasma levels of GDF15, supporting the hypothesis that GDF15 may contribute to the pathogenesis of autoimmune diseases, extending our prior epidemiological evidence.[63] In addition, the majority (52.7% = 250/474) of causal proteins were identified in diseases of the circulatory system and endocrine and metabolic diseases, with hypertension ranking at the top (n = 20) (Figure 6C). Protein FURIN showed the most significant association with hypertension (OR = 1.438, 95% CI = [1.347–1.536], $p = 1.57 \times 10^{-27}$), followed by angina pectoris, coronary heart diseases, and ischemic heart diseases (Figure 6D), which was in line with recent findings on the role of FURIN in cardiovascular diseases.[64] Full results of MR analyses are available on our website.

In addition to offering clues for pathogenesis of diseases through investigation of potentially causal associations, we identified 4,014 disease-protein pairs where protein changes were possibly a consequence of certain diseases (Figure 6E). The higher plasma level of PLAUR was found to be associated with 18 diseases across six systems (Figure 6F), among which seven were liver diseases such as fibrosis and cirrhosis of the liver. During the progression of liver fibrosis, PLAUR, the urokinase plasminogen activator surface receptor, engages in the inflammatory response, vascular homeostasis, and immune regulation,[65] which is also reflected by the significant associations between plasma PLAUR levels and leukocyte count ($\beta = 0.103$, $p < 1 \times 10^{-300}$) and CRP ($\beta = 0.073$, $p < 1 \times 10^{-300}$) in our study. Interestingly, we found that EDA2R and GDF15 might also be a consequence of diseases including liver cirrhosis, chronic

Figure 6. Summary of potential causal and consequential proteins

(A) Stacked bar chart of potential causal proteins, colored by disease chapter. Numbers above the bars represent the number of causally associated diseases.

(B–D) Significant results of the MR analysis for (B) GDF15 and autoimmune diseases, (C) proteins and hypertension, and (D) FURIN and cardiovascular diseases, colored by disease chapter. Data are all represented as OR ± 95% CI.

(E) Stacked bar chart of proteins that can be a consequence of certain diseases, colored by disease chapter.

(F–H) Significant results of the MR analysis for diseases and (F) PLAUR, (G) EDA2R, and (H) GDF15. Data are all represented as OR ± 95% CI.

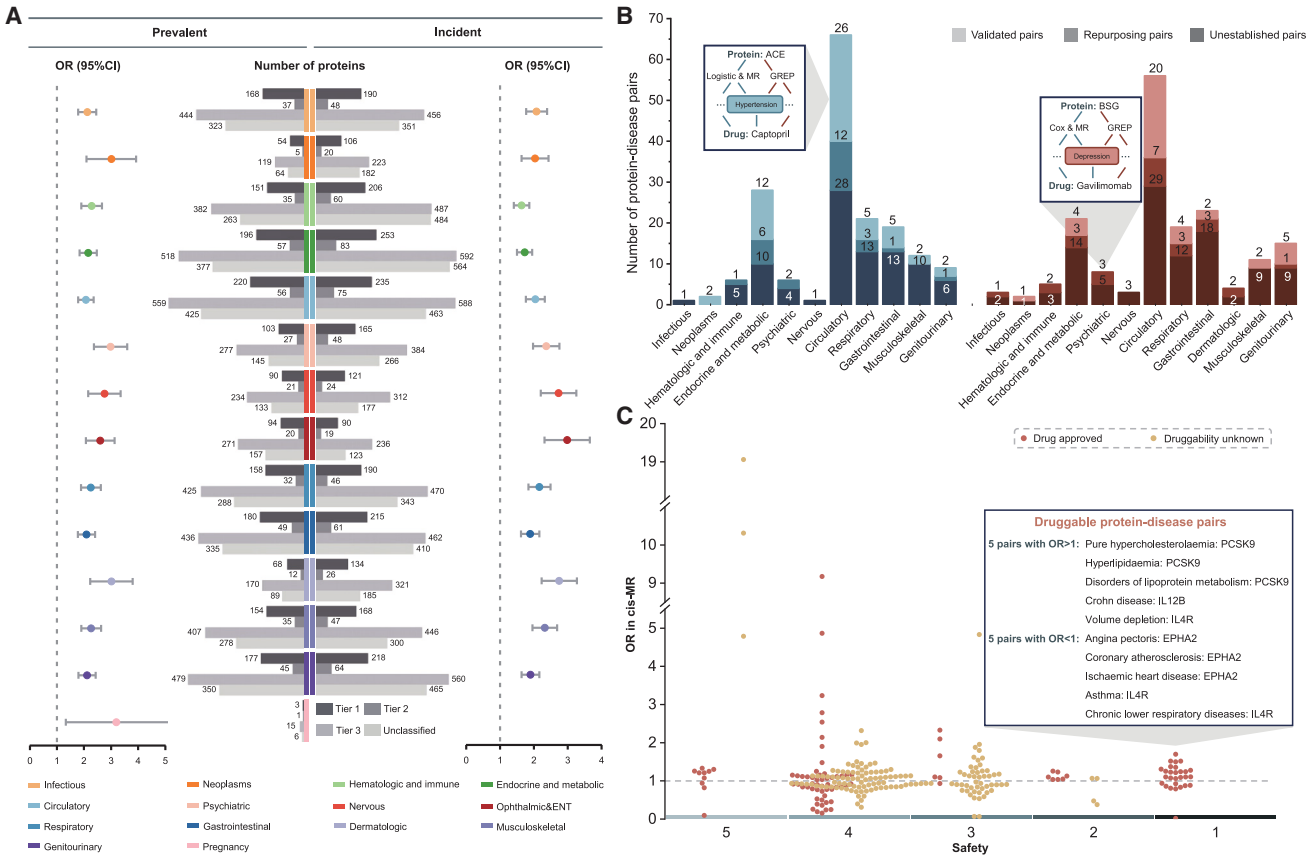


Figure 7. Drug target validation, repositioning, and identification

(A) Enrichment of proteins with the druggable genome. Data are all represented as OR ± 95% CI in the Fisher's exact test.

(B) Comparison of the *cis*-MR findings with information on target-indication pairs in DrugBank and Therapeutic Target Database. The left panel is the pairs identified from the *cis*-MR for the significant protein-prevalent disease associations, while the right panel is that for protein-incident disease associations. The numbers above the bars represent the number of validated, repurposing, and unestablished pairs that belonged to each category. The unestablished pairs are pairs in which proteins do not have established drugs. The content in the left box is an example of a validated protein-disease pair and the drug that targets this protein, while the content in the right box is an example of a repurposing pair.

(C) The safety, causality, and druggability of potential targets. The box lists examples of druggable protein-disease pairs with safety = 1, showing five examples with the strongest positive and negative causality as indicated by odds ratios.

See also Table S10.

obstructive pulmonary disease (COPD), and chronic kidney disease (Figures 6G and 6H), reflecting the pathology in certain organs.[66]

Drug target validation and repositioning

Given that plasma proteins are the major source of drug targets, we sought to mine our identified disease-associated proteins for promising targets in future drug development. Among the 1,648 prevalent disease-associated and 2,013 incident disease-associated proteins, 1,029 (62.4%) and 1,124 (55.8%) overlapped with the druggable genome reported by Finan and colleagues,[67] respectively. Moreover, the set of genes encoding these disease-associated proteins was enriched for druggable genes (ORs of 1.74 for prevalent diseases and 1.32 for incident diseases). In Figure 7A, we showed the enrichment results of disease-associated proteins classified according to disease category. The considerable overlap between these disease-associated proteins and druggable genes (especially for tier 1, which included efficacy targets of approved small molecules and biotherapeutic drugs as well as clinical-phase drug candidates) suggested the substantial potential for translating our findings into therapeutic opportunities of various diseases.[67]

Previous studies have highlighted the opportunities of using genetics for drug development and repositioning.[68,69] Therefore, we further compared those with causal evidence from *cis*-MR analysis to target-indication pairs in two drug databases, namely DrugBank[70] and Therapeutic Target Database.[71] Of 171 protein-prevalent disease pairs and 170 protein-incident disease pairs with causal evidence, 32% (54 pairs) and 22.4% (38 pairs) had approved or clinical trial drugs, respectively (Figure 7B; Table S10). For example, dozens of drugs (e.g., captopril, enalaprilat, and fosinopril) targeting ACE have been approved to

treat hypertension. In addition, we discovered 37 repurposing opportunities for 25 established drug targets, such as BSG for depression.

Safety is also a key aspect of target assessment and drug development.[72] Using an open-access database (AD Knowledge Portal[73,74]), we evaluated the safety of 123 potential targets prioritized by *cis*-MR analysis. Ten targets (e.g., EPHA2) have the lowest risk (with drug in phase IV trial; safety scale of 1). Six targets (e.g., MMP12) have lower risk (no major issues found from gene expression or genetic or pharmacological profiling, but they have not been extensively tested in humans; safety scale of 2). Twenty-six targets (e.g., SEMA3F) have potential risks (with two or fewer of high off-target gene expression, cancer driver, essential gene, associated deleterious genetic disorder, human phenotype ontology [HPO] phenotype associated gene, or black box warning on clinically used drug; safety scale of 3). Seventy-six targets (e.g., BSG) have probable risks (with more than two of high off-target gene expression, cancer driver, essential gene, associated deleterious genetic disorder, HPO phenotype associated gene, or black box warning on clinically used drug; safety scale of 4). Five targets (e.g., F10) are potentially unsafe in humans (with on-target adverse drug reactions and withdrawn drug; safety scale of 5) (Figure 7C; Table S10). Notably, our *cis*-MR findings prioritized 26 unestablished potential therapeutic targets with favorable safety profiles (safety scale $\leq$ 3) for future drug development.

Interactive webtool enables in-depth exploration of proteome-phenome atlas

To facilitate in-depth exploration of detailed results in this study, we developed an interactive webtool that provides effortless access to all summary statistics. Structured into four distinct sections, the webtool covers epidemiological associations (association analysis between proteins with diseases and traits), biomedical insights (enrichment analysis), diagnosis and prediction (discriminative analysis), and genomic associations (MR analysis). The web tool offers a comprehensive resource for future research on the role of proteins in the pathogenesis, screening, diagnosis, and treatment of human diseases, which can be accessed at https://proteome-phenome-atlas.com/ under the CCBY-NC-ND 4.0 license (for non-commercial use only). We provide selected examples to further highlight the scientific opportunities arising from this resource in Data S1. In particular, we identified that (1) BSG is likely involved in the upstream mechanisms leading to depression supported by causal evidence and the existing anti-BSG antibody, named meplazumab, can facilitate the timely execution of foundational experiments to validate this finding; and (2) the importance of proteins including GDF15 and EDA2R in diagnosing and predicting comprehensive human diseases.

DISCUSSION

The role of proteomics in categorizing and predicting health and disease represents an effective and rich biologically relevant resource to implement precision medicine. Existing studies have previously been limited to individual diseases. Here, we have performed the largest plasma proteomics investigation on a comprehensive collection of health and disease phenotypes in 53,026 individuals, revealing a total of 168,100 protein-disease associations and 554,488 protein-trait associations. Critically, the capacity for disease diagnosis and prediction models based on plasma proteins showcased markedly superior or comparable performance compared with established demographic variables across about 70% of disease endpoints. In addition, by combining pQTL data, we determined 474 potential causal proteins that overlapped with the druggable genome, thus providing promising therapeutic targets and extended applications for existing drugs. Our findings are publicly available, which we expect will help guide the development of future clinical diagnosis, prediction, and intervention strategies.

While previous studies have explored the phenotypic consequences of proteins, they have mainly been limited in scale, focusing on individual diseases and genetic associations.[12,75,76] This study is distinguished by a systematic integration of the comprehensive health and disease profiles and an in-depth investigation on disease discrimination performance capacity, causality, and the potential for therapeutic application of the identified proteins. Furthermore, the traditional disease classification strategy usually depends on similar clinical symptoms and phenotypic traits, overlooking the shared molecular etiology.[77] Through clustering diseases based on their proteomic profiles, we contribute to a reexamination of the boundaries and subtypes of disorders by anchoring the convergence of diseases in their shared biological properties. Connecting biologically related diseases could help explain why seemingly unrelated symptoms occur simultaneously in patients and further aid in mechanistic understanding and effective therapeutics.[76]

By implementing an analytical approach that incorporates comprehensive health-related phenotypes and assessing their associations with plasma protein levels uniformly, proteins exhibiting multiple significant associations were identified. Key proteins identified include those with broad hazardous effects like GDF15 and with protective effects such as EGFR. These findings shed mechanistic light by anchoring on potential shared biological pathways explaining common comorbid presentations and overlaps between disease phenotypes. More importantly, these proteins hold the opportunity to become ideal markers of systemic health status and common therapeutic targets for multi-disease. Moreover, we demonstrated that the shared associations were minimally confounded by comorbidity status, which remained largely significant in the sensitive analysis systematically correcting comorbidity status and restricting participants with comorbidities. Furthermore, the heterogeneity of protein effects merits additional focus. For instance, we show that protein DSG2 exhibited opposite directions of impact on prevalent and incident T2D. Very little is known about the precise biological mechanisms of these proteins, while our findings suggest potentially divergent roles during disease pathogenesis.

Our deep "omics" profiling of multi-disease outcomes suggests the promising diagnostic and predictive capabilities of plasma proteomics across a spectrum of diseases. Herein we highlight the potential for a real-world application of proteomics. Blood proteomics, which can be obtained through a single blood sample, shows potential both as a diagnostic supportive tool and in refining multi-disease risk estimation. Enhanced prediction

and diagnosis, even sometimes by a modest improvement, can lead to earlier disease detection, better patient stratification, and more effective personalized treatment plans, thus contributing to better health outcomes and more efficient healthcare delivery. Leveraging proteomic profiling as a single-domain and readily accessible assay is not only clinically relevant but further allows a more holistic mechanistic understanding of human disease susceptibility.[61] This not only streamlines diagnostic procedures but also opens promising avenues for proactive disease prevention and personalized interventions.

A major challenge of observational studies investigating disease-associated proteins is identifying the causal proteins that can motivate therapeutic target discovery. To achieve this, we integrated pQTL and disease GWAS data to perform proteome-by-phenome MR, which offers a data-driven approach to drug discovery using population-level data.[78,79] Using pQTLs as genetic instruments for thousands of proteins, we evaluated the potential effects of modifying protein levels on hundreds of disease phenotypes and quantified the strength of evidence for causation. Our *cis*-MR findings showed a partial consistency with established target-indication pairs in two drug databases, such as ACE for hypertension, confirming the idea that genetically supported targets are more likely to be successful in drug development.[69] Moreover, the discovery of unestablished target-disease associations suggests potential drug repurposing opportunities. For example, we observed an association between depression and BSG, a protein that has been a clinical trial target for the treatment of liver cancer and graft-versus-host disease. Furthermore, in addition to validating the potential for repurposing known targets, our findings also offered insights into identifying promising and safe therapeutic targets (e.g., FCRLB, IFNLR1, and SEPTIN8), which may provide directive significance for future drug development.

Our proteomics atlas holds several future directions. While we have unveiled the detailed protein-disease and protein-trait associations, initiating basic experiments in animals or humans is warranted to explore the specific mechanisms by which proteomics and diseases are linked. This research represents one of the largest proteome-phenome studies to date. As the population biobanks containing comprehensive spectrum of health data keep mounting, we aspire to validate our findings such as the pleiotropy of EDA2R from large prospective external cohorts, especially more diverse biobanks. Meanwhile, it is anticipated to obtain far more biological information by expanding proteomic coverage to encompass cell and organ-specific splice isoforms and posttranslational modifications. Moreover, the drug targets supported by genetic evidence, such as BSG for depression, carry a superior likelihood of success,[80] which potentially constitutes an attractive source for drug discovery programs. Surveillance of patients for markers in the future may help to monitor the efficacy of drug interventions and guide individualized treatments. Finally, reforming the taxonomy of human diseases based on biological molecular signatures is paramount for patient enrollment in future clinical trials and for the implementation of precision medicine.[77]

In summary, our study symbolizes major strides toward achieving a comprehensive understanding of the plasma proteomic atlas for human health and disease, with clinically actionable insights to integrate the advantages of the proteome across disease diagnosis, prediction, and treatment. Moving forward, the research community will benefit from this open-access proteomics atlas to allow a deeper understanding of disease pathogenesis and promote the effective development of biomarkers, predictive models, and therapeutic targets.

Limitations of the study

Several limitations of this study should be acknowledged. First, our current findings rely on proteomic data from plasma samples. Although more than 2,000 proteins enter the bloodstream through secretion, cell damage, or cell death, which could inform organ aging status[9] and health and disease status of different organs,[17,75] assessing the role of protein levels from diseased tissue may provide more insight into disease pathogenesis. Further combination with large-scale proteomic data from other tissues will be able to reveal the effects of tissue-enriched or tissue-specific proteins on relevant diseases and traits. Second, the comorbidities might confound the results of protein-disease associations.[81] However, such comorbidities are common and often difficult to eliminate in population-based studies, particularly in multi-disease analyses.[20,61,80,82,83] Sensitivity analysis was performed on each disease with rigorous quality controls, which suggested that most associations remained significant, demonstrating the robustness of the current analytical approach. Third, we performed GWAS of 75 diseases (for which no summary statistics were provided by the FinnGen study) using UKB subjects without proteomic data. These MR findings need to be validated by using disease GWAS from other sources in future research. Furthermore, since prevalent disease information may be collected before baseline protein data, causal effects from the two-sample MR analysis should be interpreted cautiously. Finally, the individuals included in this study were predominantly white Europeans. The insufficient sample size of other ancestries in the UKB limits the power of extending the current discoveries to the whole populations, emphasizing the necessity of further proteomic studies in large-scale non-European ancestral cohorts.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jian-Feng Feng (jianfeng64@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All results of protein-disease associations, protein-trait associations, enriched biological pathways, prediction and diagnosis performance, and genetic associations have been deposited through an interactive portal (https://proteome-phenome-atlas.com/) and are publicly available as the date of publication. UKB data are publicly available to bona fide researchers upon application at http://www.ukbiobank.ac.uk/using-the-resource/. The main data used in this study were accessed from the UKB (https://biobank.ndph.ox.ac.uk/) under application numbers

202239 and 19542. This study also utilized other public resources, and the DOIs are listed in the key resources table.

- All software and analytical methods used in this study are publicly available, as listed in the key resources table. The analysis code of this study has been uploaded to GitHub as listed in the key resources table.

ACKNOWLEDGMENTS

We thank Barbara J. Sahakian and Valerie Voon for critical reading and helpful discussions of the manuscript. This study used the UK Biobank Resource under application numbers 202239 and 19542. We want to thank all the participants and researchers from the UK Biobank. We want to acknowledge the participants and investigators of the FinnGen study.[84] The computations in this research were performed using the Computing for the Future at Fudan (CFFF) platform of Fudan University. Meanwhile, the results published here are in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org, Data Version: syn13363442, v11), a platform initially developed by the NIA-funded AMP-AD consortium.[73,74]

This study was supported by grants from the STI2030-Major Projects (2022ZD0211600 to J.-T.Y.), National Natural Science Foundation of China (82071201, 82271471, and 92249305 to J.-T.Y.; 82071997 and 82472055 to W.C.; 82402381 and 82471940 to J.Y.), Shanghai Municipal Science and Technology Major Project (2023SHZDZX02 to J.-T.Y. and 2018SHZDZX01 to J.-F.F.), Research Start-up Fund of Huashan Hospital (2022QD002 to J.-T.Y.), Excellence 2025 Talent Cultivation Program at Fudan University (3030277001 to J.-T.Y.), Program of Shanghai Academic Research Leader (23XD1420400 to J.-T.Y.), 111 Project (B18015 to J.-F.F.), Humboldt Research Award (to J.-F.F.), National Key Research and Development Program of China (2023YFC3605400 to W.C.), National Postdoctoral Program for Innovative Talents (BX20230087 to S.-D.C., BX20230089 to Y.-R.Z., and BX20240073 to Y.G.), Shanghai Pujiang Talent Program (23PJD006 to J.Y.) and ZHANGJIANG LAB, Tianqiao and Chrissy Chen Institute, the State Key Laboratory of Neurobiology and Frontiers Center for Brain Science of Ministry of Education, and Shanghai Center for Brain Science and Brain-Inspired Technology, Fudan University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

All authors had full access to the data in the study and accepted responsibility to submit it for publication. Conceptualization, J.-T.Y., W.C., Y.M., and J.-F.F.; methodology, Y.-T.D. and J.Y.; formal analysis, Y.-T.D., J.Y., Y.H., Y.Z., H.-Y.L., Z.-W.L., and Y.-L.C.; data curation, J.Y., Z.-Y.L., and L.Y.; writing – original draft, Y.-T.D., Y.Z., and X.-R.W.; writing – review & editing, Y.-T.D., Y.G., Y.-R.Z., S.-D.C., Y.-J.G., Y.-Y.H., L.-M.S., and Y.M.; visualization, Y.-T.D., Y.Z., and J.-Y.C.; supervision, J.-T.Y., W.C., and J.-F.F.; project administration, W.C. and J.-F.F.; funding acquisition, Y.G., Y.-R.Z., S.-D.C., J.-T.Y., J.Y., W.C., and J.-F.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Disease definition
  - Health-related trait
  - Proteomics
  - Covariates
  - Statistical analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Associations of proteins with diseases
  - Associations of proteins with traits
  - Sensitivity analysis
  - Subgroup analysis
  - Pathway enrichment analysis
  - Clustering analysis
  - Gene set enrichment analysis (GSEA)
  - Ordinal regression models
  - Prediction and diagnostic modelling
  - Mendelian randomization (MR)
  - Enrichment of proteins with druggable genome
  - Drug target validation and repositioning
  - Safety assessment
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.10.045.

Received: March 24, 2024
Revised: July 17, 2024
Accepted: October 24, 2024
Published: November 22, 2024

REFERENCES

1. GBD 2019 Diseases and Injuries Collaborators (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet *396*, 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9.
2. Howes, O.D., Thase, M.E., and Pillinger, T. (2022). Treatment resistance in psychiatry: state of the art and new directions. Mol. Psychiatry *27*, 58–72. https://doi.org/10.1038/s41380-021-01200-3.
3. National Research Council Committee on A Framework for Developing a New Taxonomy of Disease (2011). The National Academies Collection: Reports funded by National Institutes of Health. In Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease (National Academies Press) https://doi.org/10.17226/13284.
4. Denny, J.C., and Collins, F.S. (2021). Precision medicine in 2030-seven ways to transform healthcare. Cell *184*, 1415–1419. https://doi.org/10.1016/j.cell.2021.01.015.
5. Cohen, J.C., Boerwinkle, E., Mosley, T.H., Jr., and Hobbs, H.H. (2006). Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N. Engl. J. Med. *354*, 1264–1272. https://doi.org/10.1056/NEJMoa054013.
6. Lambert, G., Sjouke, B., Choque, B., Kastelein, J.J.P., and Hovingh, G.K. (2012). The PCSK9 decade. J. Lipid Res. *53*, 2515–2524. https://doi.org/10.1194/jlr.R026658.
7. Joy, T.R. (2012). Novel therapeutic agents for lowering low density lipoprotein cholesterol. Pharmacol. Ther. *135*, 31–43. https://doi.org/10.1016/j.pharmthera.2012.03.005.
8. Barbeira, A.N., Bonazzola, R., Gamazon, E.R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2021). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. Genome Biol. *22*, 49. https://doi.org/10.1186/s13059-020-02252-4.
9. Abell, N.S., DeGorter, M.K., Gloudemans, M.J., Greenwald, E., Smith, K.S., He, Z., and Montgomery, S.B. (2022). Multiple causal variants underlie genetic associations in humans. Science *375*, 1247–1254. https://doi.org/10.1126/science.abj5117.
10. Koprulu, M., Carrasco-Zanini, J., Wheeler, E., Lockhart, S., Kerrison, N.D., Wareham, N.J., Pietzner, M., and Langenberg, C. (2023). Proteogenomic

links to human metabolic diseases. Nat. Metab. *5*, 516–528. https://doi.org/10.1038/s42255-023-00753-7.

11. Emilsson, V., Ilkov, M., Lamb, J.R., Finkel, N., Gudmundsson, E.F., Pitts, R., Hoover, H., Gudmundsdottir, V., Horman, S.R., Aspelund, T., et al. (2018). Co-regulatory networks of human serum proteins link genetics to disease. Science *361*, 769–773. https://doi.org/10.1126/science.aaq1327.
12. Walker, K.A., Chen, J., Zhang, J., Fornage, M., Yang, Y., Zhou, L., Grams, M.E., Tin, A., Daya, N., Hoogeveen, R.C., et al. (2021). Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. Nat Aging *1*, 473–489. https://doi.org/10.1038/s43587-021-00064-0.
13. Johnson, E.C.B., Dammer, E.B., Duong, D.M., Ping, L., Zhou, M., Yin, L., Higginbotham, L.A., Guajardo, A., White, B., Troncoso, J.C., et al. (2020). Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat. Med. *26*, 769–780. https://doi.org/10.1038/s41591-020-0815-6.
14. Lam, M.P.Y., Ping, P., and Murphy, E. (2016). Proteomics Research in Cardiovascular Medicine and Biomarker Discovery. J. Am. Coll. Cardiol. *68*, 2819–2830. https://doi.org/10.1016/j.jacc.2016.10.031.
15. Guo, Y., You, J., Zhang, Y., Liu, W.S., Huang, Y.Y., Zhang, Y.R., Zhang, W., Dong, Q., Feng, J.F., Cheng, W., et al. (2024). Plasma proteomic profiles predict future dementia in healthy adults. Nat Aging *4*, 247–260. https://doi.org/10.1038/s43587-023-00565-0.
16. Williams, S.A., Murthy, A.C., DeLisle, R.K., Hyde, C., Malarstig, A., Ostroff, R., Weiss, S.J., Segal, M.R., and Ganz, P. (2018). Improving Assessment of Drug Safety Through Proteomics: Early Detection and Mechanistic Characterization of the Unforeseen Harmful Effects of Torcetrapib. Circulation *137*, 999–1010. https://doi.org/10.1161/CIRCULATIONAHA.117.028213.
17. Oh, H.S.H., Rutledge, J., Nachun, D., Pálovics, R., Abiose, O., Moran-Losada, P., Channappa, D., Urey, D.Y., Kim, K., Sung, Y.J., et al. (2023). Organ aging signatures in the plasma proteome track health and disease. Nature *624*, 164–172. https://doi.org/10.1038/s41586-023-06802-1.
18. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M., and Smoller, J.W. (2013). Pleiotropy in complex traits: challenges and strategies. Nat. Rev. Genet. *14*, 483–495. https://doi.org/10.1038/nrg3461.
19. Udler, M.S., Kim, J., von Grotthuss, M., Bonàs-Guarch, S., Cole, J.B., Chiou, J., Christopher D. Anderson on behalf of METASTROKE and the ISGC, Boehnke, M., Laakso, M., Atzmon, G., et al. (2018). Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. PLoS Med. *15*, e1002654. https://doi.org/10.1371/journal.pmed.1002654.
20. Wang, Q., Dhindsa, R.S., Carss, K., Harper, A.R., Nag, A., Tachmazidou, I., Vitsios, D., Deevi, S.V.V., Mackay, A., Muthas, D., et al. (2021). Rare variant contribution to human disease in 281,104 UK Biobank exomes. Nature *597*, 527–532. https://doi.org/10.1038/s41586-021-03855-y.
21. Troughton, R., Michael Felker, G., and Januzzi, J.L., Jr. (2014). Natriuretic peptide-guided heart failure management. Eur. Heart J. *35*, 16–24. https://doi.org/10.1093/eurheartj/eht463.
22. Wang, D., Day, E.A., Townsend, L.K., Djordjevic, D., Jørgensen, S.B., and Steinberg, G.R. (2021). GDF15: emerging biology and therapeutic applications for obesity and cardiometabolic disease. Nat. Rev. Endocrinol. *17*, 592–607. https://doi.org/10.1038/s41574-021-00529-7.
23. Wei, X., Su, J., Yang, K., Wei, J., Wan, H., Cao, X., Tan, W., and Wang, H. (2020). Elevations of serum cancer biomarkers correlate with severity of COVID-19. J. Med. Virol. *92*, 2036–2041. https://doi.org/10.1002/jmv.25957.
24. Tanno, T., Bhanu, N.V., Oneal, P.A., Goh, S.H., Staker, P., Lee, Y.T., Moroney, J.W., Reed, C.H., Luban, N.L.C., Wang, R.H., et al. (2007). High levels of GDF15 in thalassemia suppress expression of the iron regulatory protein hepcidin. Nat. Med. *13*, 1096–1101. https://doi.org/10.1038/nm1629.
25. Tamary, H., Shalev, H., Perez-Avraham, G., Zoldan, M., Levi, I., Swinkels, D.W., Tanno, T., and Miller, J.L. (2008). Elevated growth differentiation factor 15 expression in patients with congenital dyserythropoietic anemia type I. Blood *112*, 5241–5244. https://doi.org/10.1182/blood-2008-06-165738.
26. Desmedt, S., Desmedt, V., De Vos, L., Delanghe, J.R., Speeckaert, R., and Speeckaert, M.M. (2019). Growth differentiation factor 15: A novel biomarker with high clinical potential. Crit. Rev. Clin. Lab. Sci. *56*, 333–350. https://doi.org/10.1080/10408363.2019.1615034.
27. Lanktree, M.B., Perrot, N., Smyth, A., Chong, M., Narula, S., Shanmuganathan, M., Kroezen, Z., Britz-Mckibbin, P., Berger, M., Krepinsky, J.C., et al. (2023). A novel multi-ancestry proteome-wide Mendelian randomization study implicates extracellular proteins, tubular cells, and fibroblasts in estimated glomerular filtration rate regulation. Kidney Int. *104*, 1170–1184. https://doi.org/10.1016/j.kint.2023.08.025.
28. Wozniak, J., Floege, J., Ostendorf, T., and Ludwig, A. (2021). Key metalloproteinase-mediated pathways in the kidney. Nat. Rev. Nephrol. *17*, 513–527. https://doi.org/10.1038/s41581-021-00415-5.
29. Carter, C.S., Kenkel, W.M., MacLean, E.L., Wilson, S.R., Perkeybile, A.M., Yee, J.R., Ferris, C.F., Nazarloo, H.P., Porges, S.W., Davis, J.M., et al. (2020). Is Oxytocin "Nature's Medicine"? Pharmacol. Rev. *72*, 829–861. https://doi.org/10.1124/pr.120.019398.
30. Cortes, A., Albers, P.K., Dendrou, C.A., Fugger, L., and McVean, G. (2020). Identifying cross-disease components of genetic risk across hospital data in the UK Biobank. Nat. Genet. *52*, 126–134. https://doi.org/10.1038/s41588-019-0550-4.
31. Dostert, C., Grusdat, M., Letellier, E., and Brenner, D. (2019). The TNF Family of Ligands and Receptors: Communication Modules in the Immune System and Beyond. Physiol. Rev. *99*, 115–160. https://doi.org/10.1152/physrev.00045.2017.
32. Breit, S.N., Brown, D.A., and Tsai, V.W.W. (2021). The GDF15-GFRAL Pathway in Health and Metabolic Disease: Friend or Foe? Annu. Rev. Physiol. *83*, 127–151. https://doi.org/10.1146/annurev-physiol-022020-045449.
33. Zhang, Y., Zhao, X., Dong, X., Zhang, Y., Zou, H., Jin, Y., Guo, W., Zhai, P., Chen, X., and Kharitonenkov, A. (2023). Activity-balanced GLP-1/GDF15 dual agonist reduces body weight and metabolic disorder in mice and non-human primates. Cell Metab. *35*, 287–298.e4. https://doi.org/10.1016/j.cmet.2023.01.001.
34. Yang, L., Chang, C.C., Sun, Z., Madsen, D., Zhu, H., Padkjær, S.B., Wu, X., Huang, T., Hultman, K., Paulsen, S.J., et al. (2017). GFRAL is the receptor for GDF15 and is required for the anti-obesity effects of the ligand. Nat. Med. *23*, 1158–1166. https://doi.org/10.1038/nm.4394.
35. Hedbacker, K., Birsoy, K., Wysocki, R.W., Asilmaz, E., Ahima, R.S., Farooqi, I.S., and Friedman, J.M. (2010). Antidiabetic effects of IGFBP2, a leptin-regulated gene. Cell Metab. *11*, 11–22. https://doi.org/10.1016/j.cmet.2009.11.007.
36. Luo, H., Bauer, A., Nano, J., Petrera, A., Rathmann, W., Herder, C., Hauck, S.M., Sun, B.B., Hoyer, A., Peters, A., et al. (2023). Associations of plasma proteomics with type 2 diabetes and related traits: results from the longitudinal KORA S4/F4/FF4 Study. Diabetologia *66*, 1655–1668. https://doi.org/10.1007/s00125-023-05943-2.
37. Thorand, B., Zierer, A., Büyüközkan, M., Krumsiek, J., Bauer, A., Scheder-ecker, F., Sudduth-Klinger, J., Meisinger, C., Grallert, H., Rathmann, W., et al. (2021). A Panel of 6 Biomarkers Significantly Improves the Prediction of Type 2 Diabetes in the MONICA/KORA Study Population. J. Clin. Endocrinol. Metab. *106*, e1647–e1659. https://doi.org/10.1210/clinem/dgaa953.
38. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK Biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.

39. De Silva, N.M.G., Borges, M.C., Hingorani, A.D., Engmann, J., Shah, T., Zhang, X., Luan, J., Langenberg, C., Wong, A., Kuh, D., et al. (2019). Liver Function and Risk of Type 2 Diabetes: Bidirectional Mendelian Randomization Study. Diabetes *68*, 1681–1691. https://doi.org/10.2337/db18-1048.

40. Le-Niculescu, H., Kurian, S.M., Yehyawi, N., Dike, C., Patel, S.D., Edenberg, H.J., Tsuang, M.T., Salomon, D.R., Nurnberger, J.I., Jr., and Niculescu, A.B. (2009). Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol. Psychiatry *14*, 156–174. https://doi.org/10.1038/mp.2008.11.

41. Baek, J.H., Kim, M.S., Jung, H.R., Hwang, M.S., Lee, C.H., Han, D.H., Lee, Y.H., Yi, E.C., Im, S.S., Hwang, I., et al. (2023). Ablation of the deubiquitinase USP15 ameliorates nonalcoholic fatty liver disease and nonalcoholic steatohepatitis. Exp. Mol. Med. *55*, 1520–1530. https://doi.org/10.1038/s12276-023-01036-7.

42. Roerecke, M., Vafaei, A., Hasan, O.S.M., Chrystoja, B.R., Cruz, M., Lee, R., Neuman, M.G., and Rehm, J. (2019). Alcohol Consumption and Risk of Liver Cirrhosis: A Systematic Review and Meta-Analysis. Am. J. Gastroenterol. *114*, 1574–1586. https://doi.org/10.14309/ajg.0000000000000340.

43. Lee, S.J., Kim, K.H., and Park, K.K. (2014). Mechanisms of fibrogenesis in liver cirrhosis: the molecular aspects of epithelial-mesenchymal transition. World J. Hepatol. *6*, 207–216. https://doi.org/10.4254/wjh.v6.i4.207.

44. Ardizzone, G., Arrigo, A., Schellino, M.M., Stratta, C., Valzan, S., Skurzak, S., Andruetto, P., Panio, A., Ballaris, M.A., Lavezzo, B., et al. (2006). Neurological complications of liver cirrhosis and orthotopic liver transplant. Transplant. Proc. *38*, 789–792. https://doi.org/10.1016/j.transproceed.2006.01.039.

45. Muddapu, V.R., Dharshini, S.A.P., Chakravarthy, V.S., and Gromiha, M.M. (2020). Neurodegenerative Diseases - Is Metabolic Deficiency the Root Cause? Front. Neurosci. *14*, 213. https://doi.org/10.3389/fnins.2020.00213.

46. Joseph, C., Mangani, A.S., Gupta, V., Chitranshi, N., Shen, T., Dheer, Y., Kb, D., Mirzaei, M., You, Y., Graham, S.L., et al. (2020). Cell Cycle Deficits in Neurodegenerative Disorders: Uncovering Molecular Mechanisms to Drive Innovative Therapeutic Development. Aging Dis. *11*, 946–966. https://doi.org/10.14336/AD.2019.0923.

47. Guo, F., Liu, X., Cai, H., and Le, W. (2018). Autophagy in neurodegenerative diseases: pathogenesis and therapy. Brain Pathol. *28*, 3–13. https://doi.org/10.1111/bpa.12545.

48. Wang, L., Liu, S., Pan, B., Cai, H., Zhou, H., Yang, P., and Wang, W. (2020). The role of autophagy in abdominal aortic aneurysm: protective but dysfunctional. Cell Cycle *19*, 2749–2759. https://doi.org/10.1080/15384101.2020.1823731.

49. Lluch, A., Latorre, J., Serena-Maione, A., Espadas, I., Caballano-Infantes, E., Moreno-Navarrete, J.M., Oliveras-Cañellas, N., Ricart, W., Malagón, M.M., Martin-Montalvo, A., et al. (2023). Impaired Plakophilin-2 in obesity breaks cell cycle dynamics to breed adipocyte senescence. Nat. Commun. *14*, 5106. https://doi.org/10.1038/s41467-023-40596-0.

50. Zhou, Y., Qiu, L., Xiao, Q., Wang, Y., Meng, X., Xu, R., Wang, S., and Na, R. (2013). Obesity and diabetes related plasma amino acid alterations. Clin. Biochem. *46*, 1447–1452. https://doi.org/10.1016/j.clinbiochem.2013.05.045.

51. Wang, Y., Liang, F., Zhou, Y., Qiu, J., Lv, Q., and Du, Z. (2021). Sharp Downregulation of Hub Genes Associated With the Pathogenesis of Breast Cancer From Ductal Carcinoma In Situ to Invasive Ductal Carcinoma. Front. Oncol. *11*, 634569. https://doi.org/10.3389/fonc.2021.634569.

52. Fernández-Ruiz, I. (2016). Immune system and cardiovascular disease. Nat. Rev. Cardiol. *13*, 503. https://doi.org/10.1038/nrcardio.2016.127.

53. DeMaio, A., Mehrotra, S., Sambamurti, K., and Husain, S. (2022). The role of the adaptive immune system and T cell dysfunction in neurodegenerative diseases. J. Neuroinflammation *19*, 251. https://doi.org/10.1186/s12974-022-02605-9.

54. Gutierrez-Arcelus, M., Rich, S.S., and Raychaudhuri, S. (2016). Autoimmune diseases — connecting risk alleles with molecular traits of the immune system. Nat. Rev. Genet. *17*, 160–174. https://doi.org/10.1038/nrg.2015.33.

55. Eizirik, D.L., Szymczak, F., and Mallone, R. (2023). Why does the immune system destroy pancreatic β-cells but not α-cells in type 1 diabetes? Nat. Rev. Endocrinol. *19*, 425–434. https://doi.org/10.1038/s41574-023-00826-3.

56. Dunn, G.P., Koebel, C.M., and Schreiber, R.D. (2006). Interferons, immunity and cancer immunoediting. Nat. Rev. Immunol. *6*, 836–848. https://doi.org/10.1038/nri1961.

57. Sosa, V., Moliné, T., Somoza, R., Paciucci, R., Kondoh, H., and Lleonart, M.E. (2013). Oxidative stress and cancer: an overview. Ageing Res. Rev. *12*, 376–390. https://doi.org/10.1016/j.arr.2012.10.004.

58. Seen, S. (2021). Chronic liver disease and oxidative stress - a narrative review. Expert Rev. Gastroenterol. Hepatol. *15*, 1021–1035. https://doi.org/10.1080/17474124.2021.1949289.

59. Münzel, T., Gori, T., Bruno, R.M., and Taddei, S. (2010). Is oxidative stress a therapeutic target in cardiovascular disease? Eur. Heart J. *31*, 2741–2748. https://doi.org/10.1093/eurheartj/ehq396.

60. Forbes, J.M., Coughlan, M.T., and Cooper, M.E. (2008). Oxidative stress as a major culprit in kidney disease in diabetes. Diabetes *57*, 1446–1454. https://doi.org/10.2337/db08-0057.

61. Buergel, T., Steinfeldt, J., Ruyoga, G., Pietzner, M., Bizzarri, D., Vojinovic, D., Upmeier Zu Belzen, J., Loock, L., Kittner, P., Christmann, L., et al. (2022). Metabolomic profiles predict individual multidisease outcomes. Nat. Med. *28*, 2309–2320. https://doi.org/10.1038/s41591-022-01980-3.

62. Wang, Z., Liang, Q., Qian, X., Hu, B., Zheng, Z., Wang, J., Hu, Y., Bao, Z., Zhao, K., Zhou, Y., et al. (2023). An autoimmune pleiotropic SNP modulates IRF5 alternative promoter usage through ZBTB3-mediated chromatin looping. Nat. Commun. *14*, 1208. https://doi.org/10.1038/s41467-023-36897-z.

63. Tonkic, A., Kumric, M., Akrapovic Olic, I., Rusic, D., Zivkovic, P.M., Supe Domic, D., Sundov, Z., Males, I., and Bozic, J. (2024). Growth differentiation factor-15 serum concentrations reflect disease severity and anemia in patients with inflammatory bowel disease. World J. Gastroenterol. *30*, 1899–1910. https://doi.org/10.3748/wjg.v30.i13.1899.

64. Wichaiyo, S., Koonyosying, P., and Morales, N.P. (2024). Functional Roles of Furin in Cardio-Cerebrovascular Diseases. ACS Pharmacol. Transl. Sci. *7*, 570–585. https://doi.org/10.1021/acsptsci.3c00325.

65. Kanno, Y. (2023). The uPA/uPAR System Orchestrates the Inflammatory Response, Vascular Homeostasis, and Immune System in Fibrosis Progression. Int. J. Mol. Sci. *24*, 1796. https://doi.org/10.3390/ijms24021796.

66. Cai, Z., Deng, X., Jia, J., Wang, D., and Yuan, G. (2021). Ectodysplasin A/Ectodysplasin A Receptor System and Their Roles in Multiple Diseases. Front. Physiol. *12*, 788411. https://doi.org/10.3389/fphys.2021.788411.

67. Finan, C., Gaulton, A., Kruger, F.A., Lumbers, R.T., Shah, T., Engmann, J., Galver, L., Kelley, R., Karlsson, A., Santos, R., et al. (2017). The druggable genome and support for target identification and validation in drug development. Sci. Transl. Med. *9*, eaag1166. https://doi.org/10.1126/scitranslmed.aag1166.

68. Sanseau, P., Agarwal, P., Barnes, M.R., Pastinen, T., Richards, J.B., Cardon, L.R., and Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. Nat. Biotechnol. *30*, 317–320. https://doi.org/10.1038/nbt.2151.

69. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860. https://doi.org/10.1038/ng.3314.

70. Knox, C., Wilson, M., Klinger, C.M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E.L., Strawbridge, S.A., et al. (2024). DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Res. *52*, D1265–D1275. https://doi.org/10.1093/nar/gkad976.

71. Zhou, Y., Zhang, Y., Zhao, D., Yu, X., Shen, X., Zhou, Y., Wang, S., Qiu, Y., Chen, Y., and Zhu, F. (2024). TTD: Therapeutic Target Database describing target druggability information. Nucleic Acids Res. *52*, D1465–D1477. https://doi.org/10.1093/nar/gkad751.

72. Emmerich, C.H., Gamboa, L.M., Hofmann, M.C.J., Bonin-Andresen, M., Arbach, O., Schendel, P., Gerlach, B., Hempel, K., Bespalov, A., Dirnagl, U., et al. (2021). Improving target assessment in biomedical research: the GOT-IT recommendations. Nat. Rev. Drug Discov. *20*, 64–81. https://doi.org/10.1038/s41573-020-0087-3.

73. Greenwood, A.K., Montgomery, K.S., Kauer, N., Woo, K.H., Leanza, Z.J., Poehlman, W.L., Gockley, J., Sieberts, S.K., Bradic, L., Logsdon, B.A., et al. (2020). The AD Knowledge Portal: A Repository for Multi-Omic Data on Alzheimer's Disease and Aging. Curr. Protoc. Hum. Genet. *108*, e105. https://doi.org/10.1002/cphg.105.

74. Britton, J.S., Wiley, J.C., Beck, J., Yi, L., Bradic, L., Do, K., Grosenbacher, N., Simon, S., Hodgson, J., and Greenwood, A.K. (2023). Agora: An open-access platform for the exploration of nascent targets for Alzheimer's disease therapeutics. Alzheimers Dement. *19*, e079328. https://doi.org/10.1002/alz.079328.

75. Williams, S.A., Kivimaki, M., Langenberg, C., Hingorani, A.D., Casas, J.P., Bouchard, C., Jonasson, C., Sarzynski, M.A., Shipley, M.J., Alexander, L., et al. (2019). Plasma protein patterns as comprehensive indicators of health. Nat. Med. *25*, 1851–1857. https://doi.org/10.1038/s41591-019-0665-2.

76. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D., et al. (2021). Mapping the proteo-genomic convergence of human diseases. Science *374*, eabj1541. https://doi.org/10.1126/science.abj1541.

77. Kola, I., and Bell, J. (2011). A call to reform the taxonomy of human disease. Nat. Rev. Drug Discov. *10*, 641–642. https://doi.org/10.1038/nrd3534.

78. Bretherick, A.D., Canela-Xandri, O., Joshi, P.K., Clark, D.W., Rawlik, K., Boutin, T.S., Zeng, Y., Amador, C., Navarro, P., Rudan, I., et al. (2020). Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. PLoS Genet. *16*, e1008785. https://doi.org/10.1371/journal.pgen.1008785.

79. Zheng, J., Haberland, V., Baird, D., Walker, V., Haycock, P.C., Hurle, M.R., Gutteridge, A., Erola, P., Liu, Y., Luo, S., et al. (2020). Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nat. Genet. *52*, 1122–1131. https://doi.org/10.1038/s41588-020-0682-6.

80. Sun, B.B., Kurki, M.I., Foley, C.N., Mechakra, A., Chen, C.Y., Marshall, E., Wilk, J.B., Biogen Biobank Team, Chahine, M., Chevalier, P., et al. (2022). Genetic associations of protein-coding variants in human disease. Nature *603*, 95–102. https://doi.org/10.1038/s41586-022-04394-w.

81. Mielke, M.M., Dage, J.L., Frank, R.D., Algeciras-Schimnich, A., Knopman, D.S., Lowe, V.J., Bu, G., Vemuri, P., Graff-Radford, J., Jack, C.R., et al. (2022). Performance of plasma phosphorylated tau 181 and 217 in the community. Nat. Med. *28*, 1398–1405. https://doi.org/10.1038/s41591-022-01822-2.

82. Karczewski, K.J., Solomonson, M., Chao, K.R., Goodrich, J.K., Tiao, G., Lu, W., Riley-Gillis, B.M., Tsai, E.A., Kim, H.I., Zheng, X., et al. (2022). Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. Cell Genom. *2*, 100168. https://doi.org/10.1016/j.xgen.2022.100168.

83. Julkunen, H., Cichońska, A., Tiainen, M., Koskela, H., Nybo, K., Mäkelä, V., Nokso-Koivisto, J., Kristiansson, K., Perola, M., Salomaa, V., et al. (2023). Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nat. Commun. *14*, 604. https://doi.org/10.1038/s41467-023-36231-7.

84. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518. https://doi.org/10.1038/s41586-022-05473-8.

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

86. Millard, L.A.C., Davies, N.M., Gaunt, T.R., Davey Smith, G., and Tilling, K. (2018). Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. Int. J. Epidemiol. *47*, 29–35. https://doi.org/10.1093/ije/dyx204.

87. Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., et al. (2021). Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. Mol. Cell. Proteomics *20*, 100168. https://doi.org/10.1016/j.mcpro.2021.100168.

88. Eldjarn, G.H., Ferkingstad, E., Lund, S.H., Helgason, H., Magnusson, O.T., Gunnarsdottir, K., Olafsdottir, T.A., Halldorsson, B.V., Olason, P.I., Zink, F., et al. (2023). Large-scale plasma proteomics comparisons through genetics and disease associations. Nature *622*, 348–358. https://doi.org/10.1038/s41586-023-06563-x.

89. Dhindsa, R.S., Burren, O.S., Sun, B.B., Prins, B.P., Matelska, D., Wheeler, E., Mitchell, J., Oerton, E., Hristova, V.A., Smith, K.R., et al. (2023). Rare variant associations with plasma protein levels in the UK Biobank. Nature *622*, 339–347. https://doi.org/10.1038/s41586-023-06547-x.

90. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature *622*, 329–338. https://doi.org/10.1038/s41586-023-06592-6.

91. Elliott, P., and Peakman, T.C.; UK Biobank (2008). The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int. J. Epidemiol. *37*, 234–244. https://doi.org/10.1093/ije/dym276.

92. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. Omics *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

93. Yu, G., and He, Q.Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol. Biosyst. *12*, 477–479. https://doi.org/10.1039/c5mb00663e.

94. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

95. Pietzner, M., Stewart, I.D., Raffler, J., Khaw, K.T., Michelotti, G.A., Kastenmüller, G., Wareham, N.J., and Langenberg, C. (2021). Plasma metabolites to profile pathways in noncommunicable disease multimorbidity. Nat. Med. *27*, 471–479. https://doi.org/10.1038/s41591-021-01266-0.

96. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat. Genet. *53*, 1712–1721. https://doi.org/10.1038/s41588-021-00978-w.

97. Sakaue, S., and Okada, Y. (2019). GREP: genome for REPositioning drugs. Bioinformatics *35*, 3821–3823. https://doi.org/10.1093/bioinformatics/btz166.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Association results	This paper	https://proteome-phenome-atlas.com/
UK Biobank data	UK Biobank	https://www.ukbiobank.ac.uk/
GWAS summary statistics	FinnGen	https://www.finngen.fi/en/access_results
DrugBank	Knox et al.[70]	https://www.drugbank.ca/
Therapeutic Target Database	Zhou et al.[71]	https://db.idrblab.net/ttd/
Agora	the NIA-funded AMP-AD consortium[73,74]	https://www.synapse.org/Synapse:syn13363443
Software and algorithms		
Python	Python Software Foundation	https://www.python.org/
R	The R Foundation	https://www.r-project.org
GCTA	Yang et al.[85]	https://yanglab.westlake.edu.cn/software/gcta/#fastGWA-GLMM
Other		
code	This paper	https://github.com/jasonHKU0907/proteome-phenome-atlas

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Data assessed in this study were extracted from the UKB, a population-based cohort of around 500,000 participants aged 40–69 years at baseline. The participants were enrolled from 2006 to 2010 in 22 recruitment centers across the UK. The UK Biobank Pharma Proteomics Project (UKB-PPP) conducted blood-based proteomic profiling in a randomized subset of UKB participants. In this study, individuals missing over 20% of proteomics data were excluded, resulting in a final sample of 53,026 participants who had a median follow-up of 14.8 years until November 2023. The study was conducted following the Declaration of Helsinki, and all study participants provided informed consent. Demographic characteristics are shown in detail in Table S1.

METHOD DETAILS

Disease definition

The diagnosis data in the UKB were linked to UK electronic health records, and disease were ascertained and classified according to the International Classification of Diseases (ICD)-10 codes (Tables S3 and S4). These codes were extracted from hospital inpatient records (category 2000, fields 41270 and 41280). Prevalent and incident diseases were respectively defined as events that occurred before and after the date of participants' baseline visits, the same time that blood samples and other clinical information were collected. The incident disease outcomes are processed as time-to-event data. Participants with a prevalent event of a disease were excluded from the analysis for that incident disease. Participants' follow-up period commenced on the date of their initial visit and concluded upon the earliest occurrence of disease diagnosis, death, or the final available date from hospital inpatient records (November 2023), whichever occurred first.

We utilized the FinnGen disease endpoints code (https://www.finngen.fi/) and adhered strictly to FinnGen's guidelines for quality control (QC). This included pre-defined conditions based on sex or age, as well as control exclusions for particular diseases. Detailed QC criteria can be found in Tables S3 and S4. Notably, several diseases originally coded with two decimal places were rounded to one decimal due to the availability of only one decimal in the ICD-10 codes provided by UKB. In our analysis of incident events, individuals diagnosed with the disease prior to the baseline were excluded, while for prevalent events, those diagnosed during follow-up were similarly excluded. For each disease, controls were defined for the rest of participants without that disease. Furthermore, we excluded endpoints with fewer than 100 cases, resulting in a total of 660 incident and 406 prevalent disease endpoints for the study population with proteomic data available.

Health-related trait

The UKB resource includes physical measures such as eye and body composition measures, questionnaire data about diet, physical activities, lifestyle, living environment and health status, as well as blood and urine assays. Although these are not disease endpoints

based on hospital inpatient records, they are closely related to human health status by reflecting disease incidence and progression, being risk factors for diseases, and others. To process these trait phenotypes, we parsed the UKB data of traits collected at baseline with codes basically from PEACOK (https://github.com/astrazeneca-cgr-publications/PEACOK) R package,[20] a updated version of the PHESANT package.[86] Variables algorithmically defined by UKB were ultimately categorized into one of the three data types: continuous, categorical, and binary. The utilized parameters for each type of variable are provided in the Data S2.

After merging with participants with protein data, variables with sample sizes less than 10,000 were excluded. A minimum of 50 cases or controls were required for a binary trait to be studied. In total, we studied 453 continuous traits, 331 ordered categorical traits, and 202 binary traits. To allow for more delineated chapter-based analyses, each trait has been assigned to its corresponding secondary or tertiary UKB path manually. Chapter mappings for each trait are shown in Table S5.

Proteomics

The UKB-PPP consortium conducted the generation of blood-based proteomic data. Blood samples were collected in EDTA vacutainers, and promptly centrifuged at 2500g for 10 minutes at 4°C to separate the plasma. The plasma samples were preserved at -80°C before being transported on dry ice to Olink Analytical Services in Sweden. Subsequently, samples underwent quantification of 2,923 unique proteins, utilizing the Olink Explore™ Proximity Extension Assay and next-generation sequencing.[87] Stringent quality control (see details in biobank.ndph.ox.ac.uk/ukb/ukb/docs/PPP_Phase_1_QC_dataset_companion_doc.pdf) was implemented to ensure the measurement of proteins across four panels encompassing cardiometabolic, inflammation, neurology, and oncology proteins. To control for batch effects and technical variabilities and to enhance the accuracy of measuring low-abundance proteins, Normalized Protein eXpression (NPX) values transforming from raw protein values were utilized, following previous practice[88–90] and the suggestions of Olink manufacturer.[87] Further details regarding sample selection, processing, and quality control are provided in previous publications.[90,91] In our study, we included a collective of 2,920 proteins after excluding those with missing data of over 50%. A list of all proteins included in this study can be found in Table S2.

Covariates

Within this study, covariates were selected on the basis of (1) demographic variables, including age, sex, ethnicity (classified as White, Asian, Black, and mixed and others), and Townsend deprivation index (TDI); and (2) potential factors associated with the measurement of plasma proteins, including time fasted at blood collection, season of sample collection (classified as summer/autumn and winter/spring), sample age (time from sampling to protein measurement), body mass index (BMI), and smoking status (classified as never, former, and current). The median values were employed to impute covariates with missing values in smoking status, ethnicity, TDI, time fasted at blood collection, and BMI.

Statistical analysis

The analytical methods used in this paper are elaborated in detail in the next section. Corresponding code can be found in the key resources table.

QUANTIFICATION AND STATISTICAL ANALYSIS

Associations of proteins with diseases

Cross-sectional analysis was performed for prevalent diseases using logistic regression models, adjusting for the covariates listed above. The logistic regression models were performed utilizing the statsmodels (v0.13.1) in a Python environment (v3.9.16). The significant associations were defined by the ones that passed a stringent Bonferroni correction ($p < 0.05/[2,920*406]$). For incident diseases, we applied a longitudinal study design, using Cox proportional hazards models with the same covariate adjustments to investigate the relationship between baseline plasma protein levels and risks of incident diseases. The longitudinal data for incident diseases refer to disease diagnosis data that are collected from baseline to the end of follow-up, which is processed as time-to-event data as described in 'Disease definition'. We then applied a stringent Bonferroni correction ($p < 0.05/[2,920*660]$) to evaluate significant associations for each outcome. The p values and directions of odds ratios (ORs) and hazard ratios (HRs) for the significant associations are displayed in Figures 1B and 1C. The specific p values, ORs, HRs, confidence intervals (CIs), and sample sizes can be found in our website. The Cox proportional hazards models were executed using the "CoxFitter" function provided by the lifelines (v0.27.4) and were implemented within a Python environment (v3.9.16).

Associations of proteins with traits

Diverse regression models were utilized according to the variable type of health-related traits. A linear regression model was leveraged for the analysis of continuous traits. A logistic regression model was used for binary traits. The proportional odds logistic regression model was employed for ordered categorical traits. Importantly, all regression models were adjusted with the covariates mentioned above. We used multiple testing corrected threshold of $p < 1.71 \times 10^{-8}$ ($P<0.05/[2,920*$approximately 1,000 traits]) to define significant associations. The significant associations and corresponding p values and direction of β are displayed in Figure 2A. The specific p values, β, standard errors, and sample sizes can be found in our website. These regression models were conducted by the 'lm', 'glm' and the 'polr' function from the R package 'MASS' (v4.2.0).

Sensitivity analysis

To further investigate the influence of comorbidity, we conducted a sensitivity analysis with rigorous quality controls as follows. In cross-sectional analysis, for each prevalent disease, individuals diagnosed with any diseases from the same disease category were removed from the control group and the model was additionally adjusted with the baseline multimorbidity of the other disease categories. In longitudinal analysis, for each incident disease, anyone diagnosed with diseases from the same disease category before baseline was removed and we further excluded, from the control group, participants who had incident diseases from the same disease category; in addition, baseline multimorbidity of the other disease categories were adjusted in Cox models. The adjusted multimorbidity, defined as 14 binary variables based on 14 disease categories (Tables S3 and S4), each represented whether an individual diagnosed with any disease belonged to that specific category before baseline. Moreover, to enhance robustness and reliability, we conducted a sensitivity analysis that additionally adjusted for age^2 age*sex, age^2*sex, and the first 10 genetic principal components (PCs). Bonferroni correction ($p < 0.05/2{,}920$) was applied to define significant association. Full results can be accessed on our website.

Subgroup analysis

Furthermore, to ascertain whether the associations between proteins and health phenotypes exhibited variations across different individual characteristics, we executed subgroup analyses stratified by sex and age (middle-aged: <60 and elderly: $\geq$60 years). The same set of regression models were employed in subgroup analyses. Sex-stratified subgroup analyses were conducted by accommodating all the aforementioned covariates except for sex, while age-stratified subgroup analyses were accomplished by controlling for all the previously mentioned covariates. We applied a stringent Bonferroni correction (for prevalent diseases, $p < 0.05/[2{,}900*406*2]$; for incident diseases, $p < 0.05/[2{,}900*660*2]$; for traits, $p < 0.05/[2{,}900*1000*2]$) to evaluate significant associations for each phenotype. The ORs, HRs and CIs of several significant associations are displayed in Figure 1H, while full results including sample sizes can be found in our website.

Pathway enrichment analysis

To explore the biological insights, pathway and enrichment analyses were first performed using disease-associated proteins for prevalent diseases and incident diseases. For each disease, enrichment analyses on Gene Ontology (GO) biological process (BP) terms and Reactome pathways were conducted. Notably, for diseases without significant proteins, we utilized the top 30 proteins with the smallest *P* values for the pathway enrichment analysis. The ClusterProfiler R package (v4.10.0)[92] was employed to uncover over-represented biological processes based on the GO database. Then, Reactome pathway enrichment analysis was performed using the ReactomePA R package (v1.47.0).[93] The heatmap in Figure 3A contains the frequency of the Reactome enrichment item ranking top ten by FDR in each disease chapter. Specific pathways and corresponding *P* values and Fold enrichment can be found in our website. Since multiple pathways were examined concurrently in a single disease, the Benjamini-Hochberg method was implemented to account for multiple testing. False discovery rate (FDR) < 0.05 was established as the statistical significance threshold.

Clustering analysis

Hierarchical clustering was employed to group incident diseases (n = 660) based on the magnitude of associations (hazard ratios generated from Cox proportional hazards models) of all analyzed plasma proteins. Each incident disease was represented by the hazard ratios and a condensed distance matrix (660 × 2,920) was then formed using such disease-protein associations. The hazard ratios were pre-normalized before the analysis. The Ward's linkage was employed during clustering. The dendrogram and heatmap were computed, allowing diseases to cluster on the foundation of plasma protein association profiles. The cluster of each disease can be found in Table S6. Clustering analysis was implemented through "hierarchy" function in Scipy (v1.9.0) in Python.

Gene set enrichment analysis (GSEA)

To further uncover the biological features of disease clusters, we conducted GSEA[94] to find pathway changes among the 40 disease clusters. To determine the normalized pathway enrichment score for each individual disease, we initially ranked proteins based on the Z-values generated in Cox models. Then, the ranked list was submitted to GSEA using R package clusterProfiler (v4.10.0)[92] and GO BP gene sets (subcategory in C5, MSigDB database v2023.2.Hs) with at least 10 overlapping genes. The top 10 pathways that were significant in any disease after FDR adjustment were kept. To further elucidate the biological features of disease clusters, we performed Wilcoxon ranked-sum tests to identify pathways differently regulated between one cluster and the others. Specifically, we compared the pathway enrichment scores of diseases in one cluster against the scores of all other clusters. The top five pathways for each cluster are shown in Figure 4A and the heatmap contains the average of normalized pathway enrichment score in each cluster. *P* values were adjusted using the Benjamini-Hochberg method.

Ordinal regression models

To determine the multimorbidity level of each disease cluster in the population, we summed the number of incident diseases for each person into ordinal variables.[95] Figures 4B–4D provide the multimorbidity level for three example disease clusters. Ordinal regression models were then employed to investigate proteins associated with multimorbidity in each cluster. The models were adjusted for age,

sex, ethnicity, TDI, smoking status, BMI, fasting time, season, and sample age. Participants with any specific disease in the cluster at baseline and missing values for covariates were excluded from the analyses for this cluster. Bonferroni-adjustment (0.05/[2,920 proteins*40 clusters]) was applied. The results for ordinal regression are recorded in Table S7.

Prediction and diagnostic modelling

For each disease, two models, a prediction model and a diagnostic model, were developed using a machine learning algorithm, named light gradient boosting machine (LightGBM). Specifically, the prediction model (number of diseases=660) aimed to determine whether a baseline healthy participant would develop a certain disease (predicted as class 1) or stay healthy (predicted as class 0), and any individuals who had prevalent diagnosis of that disease were excluded. As for diagnostic model (number of diseases = 406), it aimed to discriminate whether a participant had experienced or is currently experiencing a certain disease (predicted as class 1) versus those who have not (predicted as class 0). The models were established based on the top-30 important proteins ($\approx$ 1% of all 2,920 proteins), which was determinate based on information gain, an inherent algorithm within the LightGBM, that measured the extent to which a specific protein can impact the model. We calculated the protein importance using information gain, an inherent algorithm with LightGBM, that measured the extent to which a specific protein can impact the model.

For comparison purposes, we established models by utilizing participants' clinical-demographic information, namely, age, sex, ethnicity, TDI, BMI, systolic blood pressure, and status of smoking and alcohol consumption. Furthermore, an integrated model was established by combining proteins and clinical-demographic data. We then compared the discriminative performance between protein-based model versus demographic-based model. In addition, we also explored the additive values of proteins by comparing the demographic-based model versus integrated model. We examined the significance through DeLong test.

Models were trained and optimized through a nested leave-one-region-out cross-validation strategy. Specifically, the data was partitioned into 10 folds based on the geographical locations of participants' recruitment centers. The geographical locations included East Midlands, London, North-East, North-West, Scotland, South-East, South-West, Wales, West Midlands, and Yorkshire and Humber. Each time, nine folds of data (training set) were used to develop models, including proteins selection, hyperparameters tuning and model training, and the rest one was used as a testing set. After 10 iterations, all folds of data were used as testing sets, and they were aggregated for evaluation. A bootstrap sampling method with replacement, iterated 1,000 times, was applied to report the median and 95% confidence intervals of the areas under the curve (AUCs), accuracy, sensitivity, specificity, precision, Youden index, and F1 score. The optimal hyperparameter was tuned within the training set, and it was performed using a random partitioned fivefold cross-validation through grid search within a hyperparameter space of 100 candidate combinations. Notably, the testing test was kept untouched and was merely used for evaluations. The AUCs, top 30 important proteins, and *P* values of DeLong test for each disease are reported in Figure 5 and Tables S8 and S9. Model development and evaluations were implemented through lightgbm (v3.3.2) and scikit-learn (v1.0.2) under Python (v3.9.16).

Mendelian randomization (MR)

Instrumental selection

Instruments to proxy for changed protein abundance were variants associated in cis (within 1Mb of the transcription start sites) and in trans, separately, at genome-wide significance ($p < 5 \times 10^{-8}$) extracted from the protein genome-wide association study (GWAS) summary statistics.[90] Instruments to proxy for disease incidence were variants at $p < 5 \times 10^{-8}$ extracted from the disease GWAS summary statistics. Most of the disease GWAS were from the FinnGen study release DF9 (https://www.finngen.fi/en/access_results)[84] and the remaining disease GWAS (for which no summary statistics were provided by the FinnGen study) were calculated using generalized linear mixed models (GLMM) with Genome-wide Complex Trait Analysis (GCTA,[85] v1.94.0) (https://yanglab.westlake.edu.cn/software/gcta/#fastGWA-GLMM) in the subset of British White participants without protein data from the UKB. Linkage disequilibrium clumping ($r^2 < 0.01$) was conducted using the European 1000 Genomes Project phase 3 as the reference panel. We removed instruments associated with more than five proteins to minimize pleiotropic effects and instruments with an F-statistic of less than 10 to reduce weak instrument bias.

Mendelian randomization

For significant protein-disease associations identified in the epidemiological analysis, the genomic associations were further explored using bidirectional two-sample MR analysis with instruments of proteins and diseases. The Wald ratio was used to estimate MR effects if only a single instrument was available and the inverse variance weighted (IVW) method was used if two or more instruments were available. We defined two kinds of relationship between proteins and diseases, as detailed in results.[96] First, a protein was causally associated with a disease, which is evidenced by a significant association (defined by FDR < 0.05) identified in the protein-to-disease direction as well as an insignificant association revealed in the disease-to-protein direction. Second, the altered plasma level of a protein was a consequence of a disease, as evidenced by a significant association in the disease-to-protein direction while insignificant in the protein-to-disease direction. In Figures 6B–6D, we reported examples of potentially causal associations and their ORs and 95% CIs. In Figures 6F–6H, we showed examples of associations where the change of protein levels might result from the incidence of diseases. Results for all examined associations are reported in our website. The "TwoSampleMR" R package was used to conduct MR analysis (R v4.2.0).

Enrichment of proteins with druggable genome

Previously, Finan et al. constructed an updated compendium of druggable genomes to validate drug targets and accurately match disease indications, which incorporated 4,479 genes.[67] The research categorizes druggable gene sets into three tiers based on their drug development pipeline. The 1,427 genes corresponding to efficacy targets of clinical-phase drug candidates, approved biotherapeutics, and small-molecule drugs are in tier 1. The 682 genes corresponding to targets of drug-like compounds and those closely related to known drug targets are in tier 2. The 2,370 genes corresponding to druggable genes not incorporated in the aforementioned two tiers, as well as extracellular or secreted proteins less similar to approved drug targets, are in tier 3[67] We separately evaluated whether proteins screened by Cox regression models and logistic regression models overlapped with druggable genome genes (4,479 genes). Fisher's exact test was employed for the enrichment analysis and the related calculations were performed in R v.4.3.1. The numbers of disease-associated proteins overlapped with druggable genome and ORs and 95% CIs of enrichment analysis are shown in Figure 7A.

Drug target validation and repositioning

Utilizing the GREP (Genome for REPositioning drugs) software,[97] we performed an enrichment analysis of the proteins prioritized by cis-MR in the drug targets of clinical indications categorized by disease chapter and captured potentially repositionable drugs. Information on approved drug targets or clinical trial targets was collected from two publicly accessible and regularly updated databases, DrugBank[70]: https://www.drugbank.ca/ and Therapeutic Target Database[71]: https://db.idrblab.net/ttd/. The validated, repurposing and unestablished protein-disease pairs are listed in Table S10 and the counts are shown in Figure 7B.

Safety assessment

Agora, an AD Knowledge Portal results browser (https://agora.adknowledgeportal.org/),[73,74] was employed to perform safety assessment of protein targets. Targets were placed into buckets ranked according to therapeutic antibody feasibility, safety and small molecule drug development preferences, with the safety of targets being categorized into six levels (1: lowest risk, 2: lower risk, 3: potential risks, 4: probable risks, 5: potentially unsafe in humans, and 6: unknown). Smaller bucket numbers were generally considered to be more beneficial for successful drug development. The safety scale of 123 potential targets identified by cis-MR analysis can be found in Table S10.

ADDITIONAL RESOURCES

We developed an interactive webtool to provide effective access to the results (https://proteome-phenome-atlas.com/), which is described in results section "interactive webtool enables in-depth exploration of proteome-phenome atlas".

CellPress
OPEN ACCESS

Supplemental figures

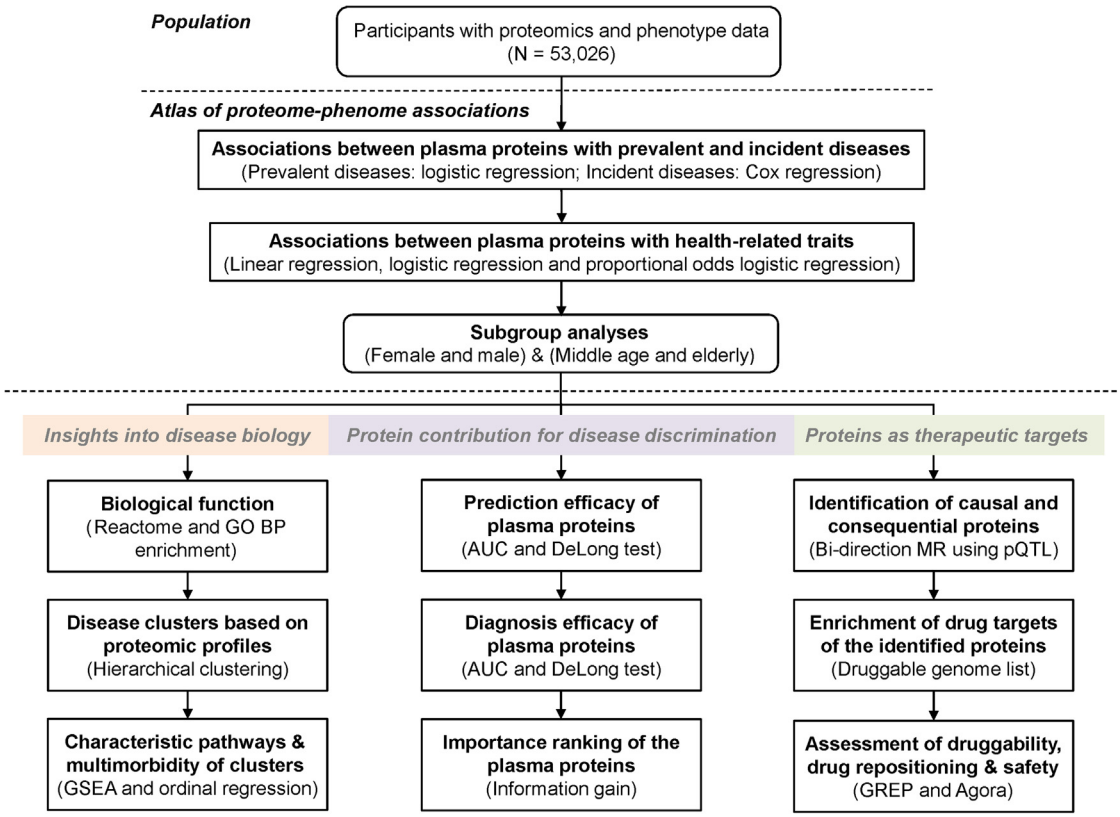


Figure S1. Analytical workflow, related to Figure 1

First, we included 53,026 participants with available proteomics and phenotype data. Specifically, we extracted data of 2,920 proteins, 406 prevalent disease endpoints, 660 incident disease endpoints, and 986 traits for the main analysis. The associations between plasma proteins and diseases were investigated with logistic regression models (prevalent diseases) and Cox models (incident diseases), separately. The associations between proteins and traits were investigated with linear regression models (continuous traits), logistic regression models (binary traits), and proportional odds logistic regression models (ordered categorical traits), separately. Subgroup analyses for different sex and age (middle age: 39–59 years; the elderly, $\geq$60 years) were conducted. Second, we investigated insights into disease biology provided by the protein-disease associations. Third, we investigated the protein contribution for disease discrimination. The prediction and diagnosis efficacy of identified proteins were estimated based on three models (i.e., the protein-only model, the demographics-only model, and the integrated model), and the DeLong test was used to compare the performance of the three models. Finally, we investigated the potential causal relationship between proteins and diseases. For the causal proteins, we estimated their potential value as therapeutic targets.

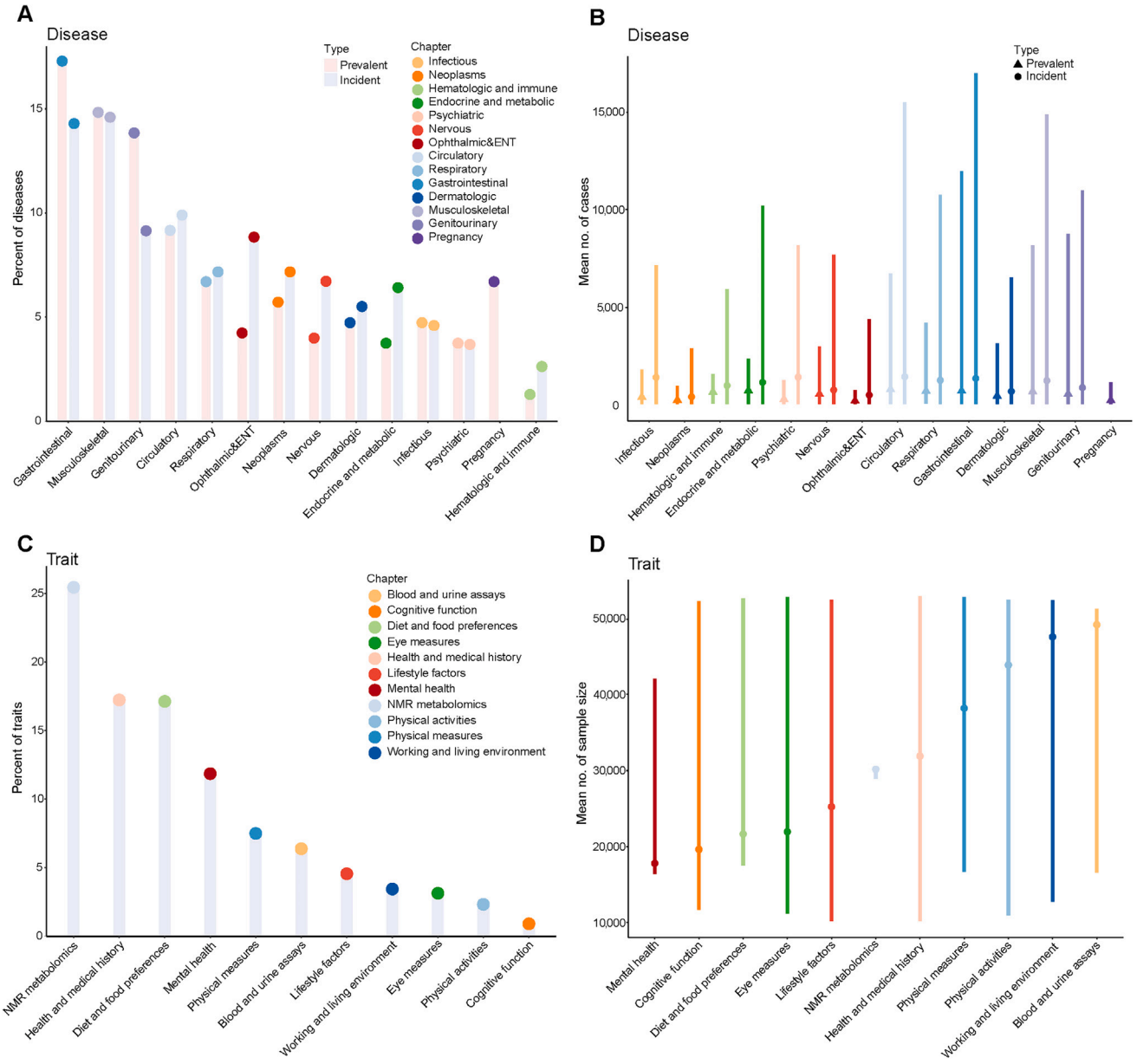


Figure S2. Distribution and sample sizes of the included phenotypes, related to Figure 1

(A) The percentage of prevalent and incident diseases assessed in the study classified by ICD-10-based chapter.

(B) The mean number of cases for each prevalent and incident disease stratified by ICD-10-based chapter. Bars represent the min-max range of the case numbers.

(C) The percentage of health-related traits assessed in the study classified by UKB-based chapter.

(D) The mean number of cases per trait stratified by chapter. Bars represent the min-max range of the case numbers.