Machine learning-guided deconvolution of plasma protein levels

Maik Pietzner[1,2,*], Carl Beuchel[1], Kamil Demircan[1], Julian Hoffman Anton[2], Wenhuan Zeng[1], Werner Römisch-Margl[3], Summaira Yasmeen[1], Burulça Uluvar[1], Martijn Zoodsma[1], Mine Koprulu[2], Gabi Kastenmüller[3], Julia Carrasco-Zanini[2], Claudia Langenberg[1,2,*]

[1]Computational Medicine, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Berlin, Germany
[2]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK
[3]Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

*corresponding authors
Maik Pietzner (maik.pietzner[at]bih-charite.de)
Claudia Langenberg (claudia.langenberg[at]qmul.ac.uk)

ABSTRACT

Proteomic techniques now measure thousands of circulating proteins at population scale, driving a surge in biomarker studies and biological clocks. However, their potential impact, generalisability, and biological relevance is hard to assess without understanding the origins and role of the thousands of proteins implicated in these studies. Here, we provide a data-driven identification of the foundations of protein variation that underly their links to ageing and diseases, differ between sexes and ancestries, and help guide protein biomarker and drug target discovery. We use machine learning to systematically identify and quantify the foundations of plasma levels of ~3,000 protein targets among 43,240 participants of the UK Biobank. Out of >1,700 participant and sample characteristics, we identify a median of 19 factors (range: 1-36) that jointly explained an average of 23.7% (max. 79.9%) of the variance in plasma levels across protein targets. Proteins segregated into distinct clusters according to their explanatory factors, with modifiable characteristics explaining more variance compared to genetic variation (13.3% vs 9.8%). We identify proteins for which the factors underlying their variation differed by sex (n=1414 proteins) or across ancestries (n=86 proteins). We establish a knowledge graph that integrates our findings with genetic studies and drug characteristics to guide identification of potential markers of target engagement. We demonstrate the value of our resource 1) by identifying disease-specific biomarkers, like matrix metalloproteinase 12 for abdominal aortic aneurysm, and 2) by developing a framework for phenotype enrichment of protein signatures from independent studies to identify underlying sources of variation. All results are explorable via an interactive web portal (https://omicscience.org/apps/prot_foundation) and can be readily integrated into ongoing studies using an associated *R* package.

INTRODUCTION

High-throughput plasma proteomics is now fuelling a new wave of biomarker studies at unprecedented scale(*1*), but is about to repeat the failures of decades of previous studies with very few, if any, of thousands of tested candidates ever improving clinical practice (*2*). The ever-expanding content of proteomic assays, now surpassing half of the protein coding genome, further exceeds our ability to understand the origins and relevance of the many non-actively secreted proteins detectable in blood (**Fig. 1a**) (*3*, *4*).

Proteome-wide genome-wide association studies (*5*–*11*) provided evidence that changes in plasma protein abundance can reflect altered production in tissues and identified at least one protein quantitative trait locus (pQTL) for most protein targets. Some pQTLs explained jointly as much as 70% of the variance in plasma levels (*5*, *8*) and they have been widely advocated as instruments for causal inference or even to impute the plasma proteome (*12*–*14*). However, on average pQTLs explain relatively little of protein variation in plasma, partly because static germline genetic variation does not capture dynamic adaptions that indicate early disease states or worsening of conditions. Atlas-like efforts are now emerging that statistically associate changes in plasma protein levels with hundreds of diseases, ageing clocks, or health characteristics (*7*, *15*–*17*), but that do rarely translate into biological knowledge or understanding of the factors that underly disease associations or predictive models.

Here, we present a framework for the integration of multimodal data to derive foundations of variation in plasma protein levels building on our previous work (*18*). We demonstrate that a relatively small number of features (median: 19; range: 1-36) of all >1,700 tested explain a considerable part of protein variation (>25% for most protein targets) in plasma. Protein targets thereby segregated into distinct clusters according to major influences on human health and but also pre-analytical variation activating platelets. We provide evidence for differential contribution of participant characteristics across the sexes and ancestral strata and demonstrate how this knowledge guides the identification of disease-specific biomarkers and drug response markers from observational studies. We create a resource that is publicly available to the community to explore results (https://omicscience.org/apps/prot_foundation/) and a statistical framework to implement phenotype enrichment analyses in external studies (https://github.com/comp-med/r-prodente).

RESULTS

A coordinate system for plasma protein variation

We identified 384 diverse, modifiable and non-modifiable, participant and technical characteristics (out of a total of 1735; **Supplementary Tab. 1**) to explain variation in plasma levels of one or more of 2,853 protein targets (97.7%; **Supplementary Tab. 2**) among 43,240 UK Biobank participants (54.1% female; 5.2% Non-European) using a two-stage machine learning framework (*19*). A median of 19 characteristics (range: 1 - 36) was selected across these protein targets and explained on average 23.7% (range: 0.0004%-79.9%) of protein variation in plasma (**Fig. 1c and Supplementary Tab. 3**). The highest amount of variance was explained for proteins actively secreted into blood (32.1%) and those demonstrated to be reliably detected using the assay technology (27.8% for 1,990 proteins with ≤5% of values below the limit of detection), in addition to proteins for which the coding gene can reliably be detected through bulk RNA sequencing of tissue samples (**Supplementary Fig. 1a-h**).

Mapping the variance explained matrix into a lower-dimensional space established a coordinate system along which proteins segregated into distinct clusters according to shared major influences but also tissue and cell-type origin (**Fig. 1d**). Only two clusters showed contributions from diverse factors of human health, for example, a cluster enriched for proteins mainly produced by fibroblasts ('Fibroblasts-enhanced'; odds ratio: 3.79; p-value$<5.7\times10^{-11}$), whereas remaining clusters distinguished by at most a few or even single characteristics explaining most of the variance in plasma protein levels (**Supplementary Tab. 4 and Supplementary Fig. 2**). For example, one cluster contained 687 protein targets that were strongly enriched for participant and sample characteristics indicating effects of technical variation, such as recruitment/ study centre (odds ratio: 6.61; p-value$<1.2\times10^{-194}$; mean explained variance (MEV): 1.84%) and platelet count (odds ratio: 2.99; p-value$<2.5\times10^{-91}$; MEV: 9.40%) (**Supplementary Fig. 2k-l**). Our findings are in line with previous studies (*20*, *21*) and are likely a result of platelet activation, and subsequent release of contained proteins into plasma (*21*). Collectively, these results provide evidence, that protein signatures in blood can be partly deconvoluted into distinct origins with modifiable participant characteristics outweighing non-modifiable ones on average ($p<5.7\times10^{-61}$; two-sided Wilcoxon rank sum test; **Fig. 1d**).

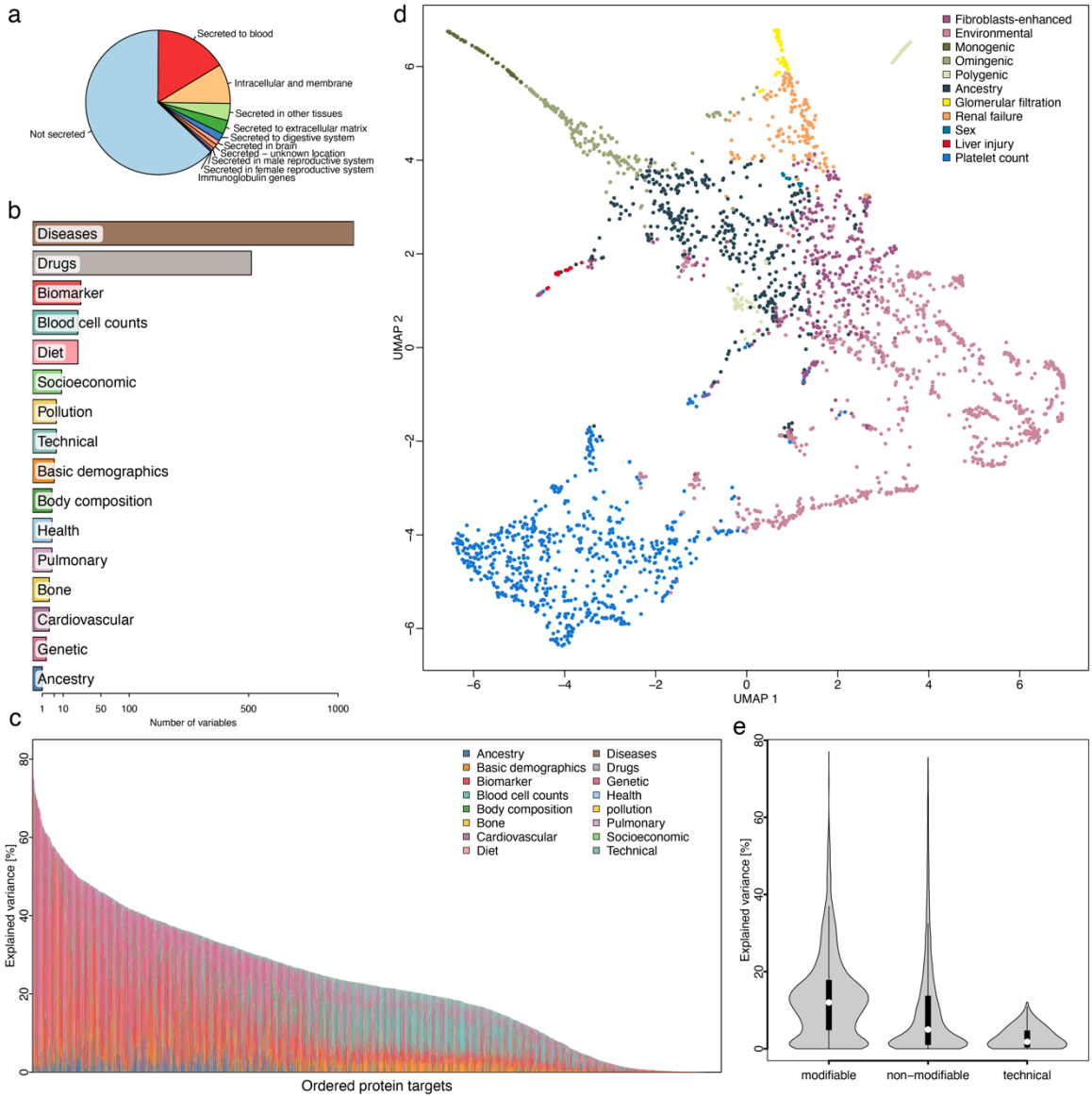

Figure 1 A coordinate system of plasma protein variation. a Secretome assignment[16]/12/2024 18:11:00 for 2919 protein targets analysed in the present study. **b** Number and type of UK Biobank participant characteristics considered in the study. **c** Stacked bar charts displaying the achieved variance explained for each protein target. Proteins are ordered by explained variance across all factors. Colours indicate domains of explaining factors. **d** Uniform Manifold Approximation and Projection (UMAP) mapping of the variance explained matrix across 2853 protein targets for which we identified at least one feature explaining the variance in plasma levels. Each protein has been assigned a cluster based on k-means clustering and is coloured accordingly. e Violin plots showing the explained variance across protein targets according to categories of selected characteristics. The centre of each violin plot is a boxplot giving median (white dot) and interquartile range.

Organ and cell-type contributions

Proteins circulating in blood have diverse origins and we therefore integrated gene expression data (*3*, *22*) to understand whether the selection of participant characteristics can be

122 explained by effects in certain tissues or cell-types. More than 80% (n=2,420) of all detected
123 protein targets showed evidence for preferential production in one or at most a few tissues or
124 cell types enabling enrichment analysis in up to 13 tissues and 28 cell-types (**Supplementary**
125 **Tab. 2**). This identified 93 characteristics significantly more frequently selected for proteins
126 with tissue or cell-type specific production **(Fig. 2; Supplementary Fig. 3 and Tab. 5-6**). Many
127 of these enrichments were likely driven by tissue or cell damage, for example, proteins with
128 enhanced production in the liver, specifically hepatocytes, were explained by medications
129 with adverse hepatic effects such as oral contraceptives (*23*) (e.g., conjugated oestrogens;
130 odds ratio: 5.62; p-value<1.6×10^{-16}) or tamoxifen (odds ratio: 7.55; p-value<2.9×10^{-11}). Similar
131 medication effects on target organs/cells were seen for proton pump inhibitors and stomach
132 (e.g., lansoprazole; odds ratio: 90.67; p-value<1.9×10^{-14}) or immune suppressants and
133 immune cells (e.g., azathioprine; odds ratio: 20.51; p-value<1.1×10^{-14}).

134 On a cell population level, we observed yet unreported links between participant
135 characteristics and proteins with enhanced expression in fibroblasts that synthesize the extra
136 cellular matrix (**Fig. 2**). This included drug usage like warfarin (odds ratio: 5.59; p-value<$5.4 \times 10^{-}$
137 ^{13}), diseases like atrial fibrillation (odds ratio: 7.07; p-value<2.1×10^{-11}), but also participant's
138 age (odds ratio: 3.00; p-value<1.8×10^{-10}), body mass index (BMI; odds ratio: 3.20; p-
139 value<1.6×10^{-11}) and genetically inferred ancestry (odds ratio: 2.98; p-value<2.6×10^{-10}). These
140 results provide evidence that changes in plasma protein levels can to some extend be
141 explained because of adverse effects of drugs and diseases on specific tissues rather than
142 necessarily the other way round.

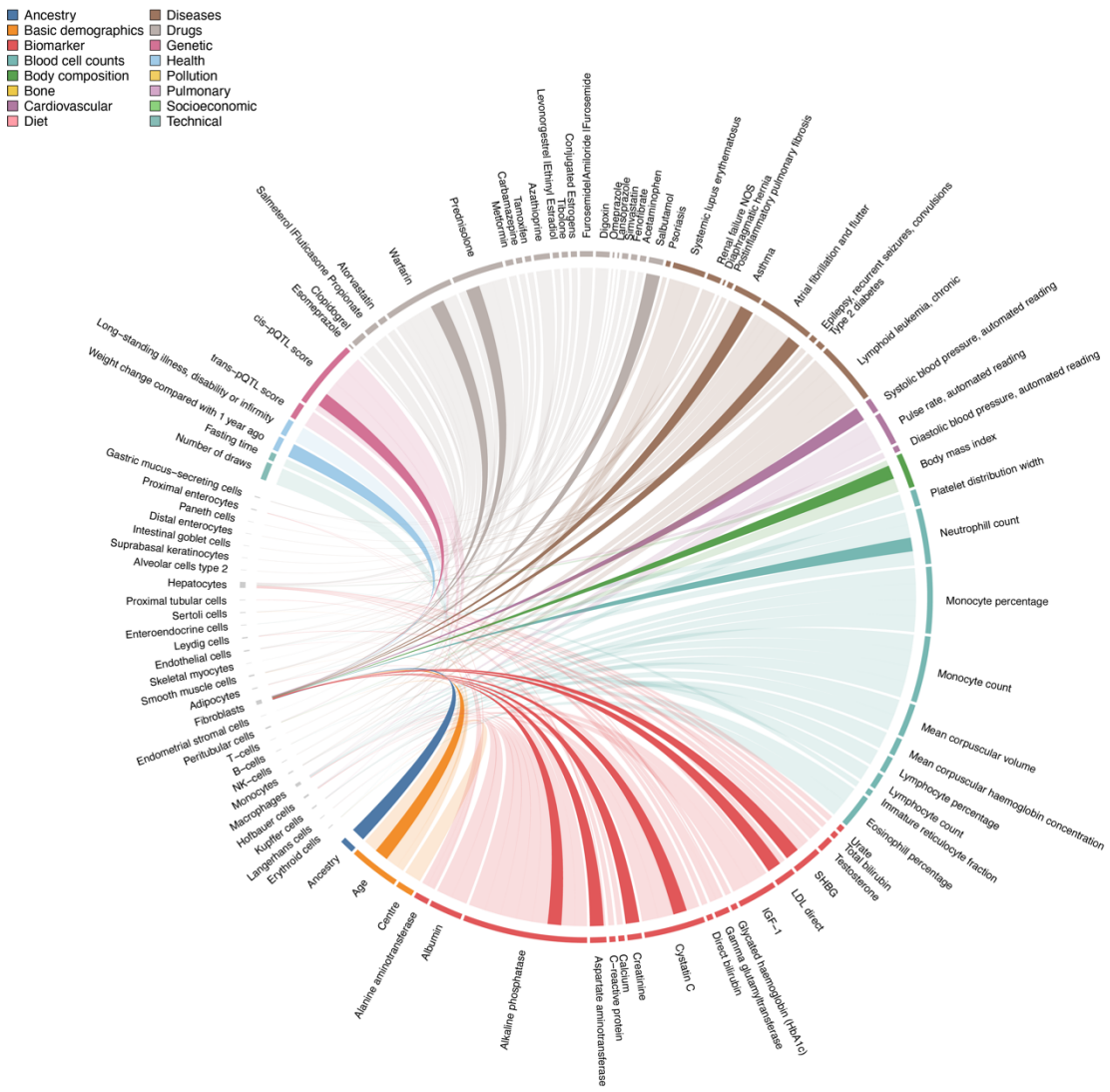

Figure 2 Plasma proteins link cell-types to indicators of health and disease. Chord diagram of phenotype associated protein signature enrichment among protein coding genes with enhanced cell type expression ('marker genes'). Each line represents a significant enrichment ($p<2.9x10^{-6}$) of proteins associated with a participant characteristic among protein coding genes with enhanced expression in a cell type. Enhanced expression estimates were derived from single-cell RNA sequencing data in the Human Protein atlas. Corresponding statistics can be found in Supplementary Table 6. Associations with marker genes of fibroblasts have been highlighted by darker colours.

Sex- and ancestral differential effects on the plasma proteome

Genetically inferred ancestry (n=1,157) and participant's sex (n=1,214) were among the most frequently selected non-modifiable participant characteristics. To better understand differential or -specific effects across ancestries (41,369 White Europeans, 849 British Africans, and 823 British Central South Asians) and the sexes (23,601 females, 20,055 males), we repeated the feature selection procedure separately within each of the groups.

We identified significantly lower levels of explained variance in participants of British African (median: -9.16%; IQR: -14.90% - -2.38%; $p<7.0x10^{-140}$) and to a lesser extent British Central South Asian ancestry (median: -2.73%; IQR: -8.04% - 0.00%; $p<2.3x10^{-29}$) compared to White Europeans. Results, that were, at least among participants of British African ancestry, not entirely explained by the lower sample size, as we still observed a difference of ≥4% in explained variance in plasma levels for more than half of the protein targets following matching of selected features (IQR: -8.05% - 0.00%, p-value$<1.5x10^{-113}$; **Fig. 3a-b**).

The single strongest contributor to ancestral variation was the contribution of cis- and trans-pQTLs (**Supplementary Tab. 7**), likely explained by a combination of different reasons. Firstly, genetic variants are present at different minor allele frequencies (MAF; **Fig. 3c-d**) across ancestries, and, secondly, may have differential effects on protein levels. We obtained evidence that differential effects of the same variant across ancestral groups most commonly accounted for differences and included 134 cis-pQTLs, almost all (n=130) of which could not be explained by different causal variants across ancestries. We also identified 33 protein targets for which ancestry-specific lead signals (non-overlapping haplotype blocks; **Fig. 3d**) accounted for strong differences in explained variance. For example, the missense variant rs2071421 (p.Asn352Ser) conferring arylsulfatase A pseudo deficiency was most common in participants of British African ancestry (MAF_{AFR}=38.1%) but less frequent in participants of other ancestries (MAF_{EUR}= 11.0%; MAF_{CSA}= 14.8%) and accordingly explained considerably more variance (5.9% vs 65.8%) compared to the trans-ancestral lead signal rs873697 (MAF_{AFR}=11.5%; MAF_{EUR}= 4.4%; MAF_{CSA}= 0.6%) in participants of British African ancestry.

We lastly identified very few protein targets (n=26) that were much better explained (>4 s.d.) in one but not the other sex (**Supplementary Fig. 4a**), but noted that selected participant characteristics differed for almost a third (n=880) of the protein targets (median Jaccard index = 0.56; **Supplementary Fig. 4b**) indicating the need to consider sex-specific contributions (**Supplementary Tab. 8**). Obvious examples included medications (n=550 pairs) and diseases (n=41 paris) given/occurring only in one sex, whereas abundant sex-differential effects were explained by associations with age, biomarkers, or body mass index (**Supplementary Fig. 4b and Supplementary Tab. 8**). Notably, there were only few examples of sex-differential genetic effects, e.g., the cis-pQTL score for plasma oxytocin (OXT; female: 16.1%; male: 31.1%; p-$value_{inter}<2.8x10^{-48}$; **Supplementary Fig. 4c**).

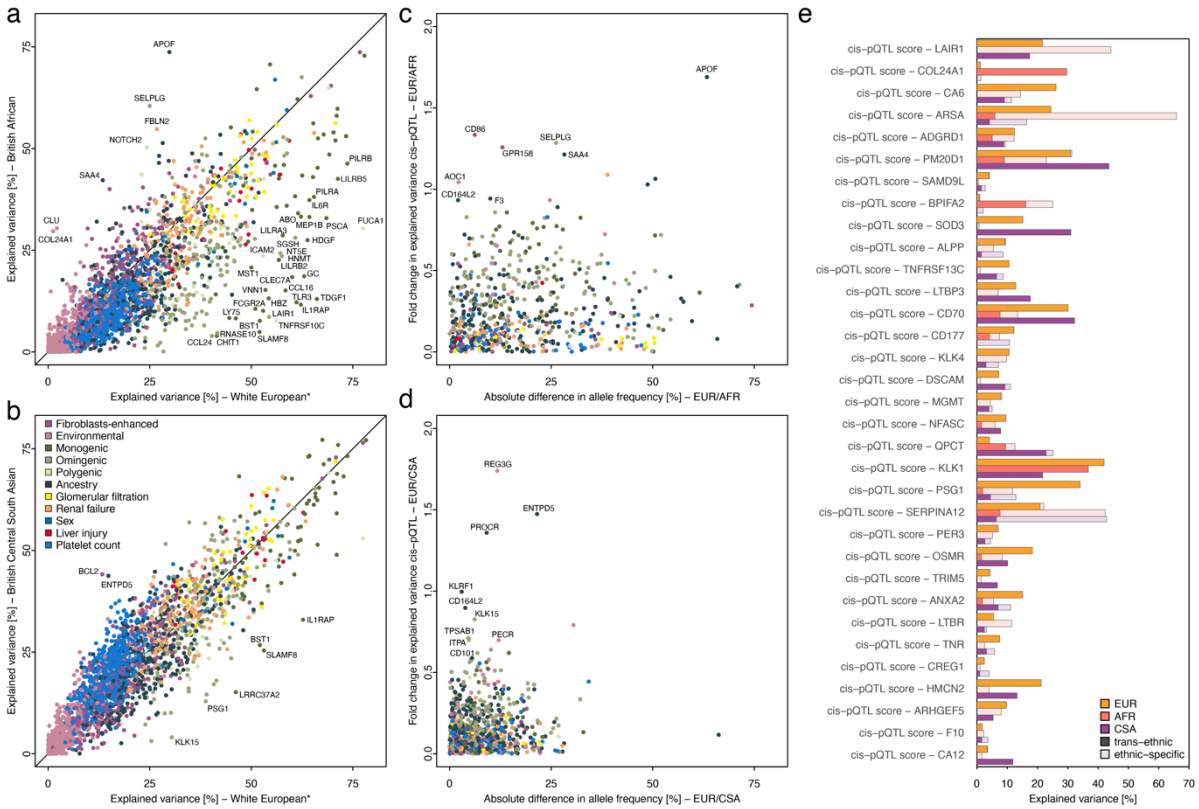

Figure 3 Summary of ancestral-stratified analysis. a Comparison of the achieved explained variance in plasma proteins levels within White Europeans compared to British Africans participants. Proteins are coloured based on cluster membership and those with extreme differences are annotated. *The number of selected features has been matched with the smaller ancestry to account for differences in the power of discovery. **b** Same as **a**, but now comparing to participants of British Central South Asian ancestry. **c-d** Contrasting differences in allele frequencies for cis-pQTL scores with changes in explained variance between White Europeans and British African (c, AFR) or British Central South Asian (d, CSA). **e** Variance explained estimates for 33 cis-pQTLs with evidence for distinct ancestral-lead signals. Darker colours indicate the effect of the trans-ancestral signal, whereas shades indicate the explained variance by the ancestral-specific lead signal. We note that the effect of ancestral-specific lead signals might be overestimated, since they have not been selected from an independent cohort.

Protein biomarker discovery and pruning for incident diseases

To transcend statistical mapping of protein – disease associations that are unlikely to inform prioritisation of biologically or clinically relevant insights, we next built a framework designed to provide the foundation for robust biomarker identification in biobank-scale studies. Among the 60,334 significant disease associations ($p<4.4x10^{-8}$) observed here and reported in other studies (*7*, *15*), we observed a more than 39-fold drop after regressing out foundations of plasma protein levels for 390 incident diseases (**Fig. 4a** and **Supplementary Tab. 9**). Associations of more than two-thirds (1,443 out of 2,081) of protein targets were almost completely attenuated. Even among the 1,560 protein – disease associations with directionally concordant effects and persisting significance, >80% (n=1,300) showed

210 considerable effect attenuation of ≥20%, suggesting residual confounding (**Fig. 4b**). A >5-fold
211 decrease in the number of significant protein – disease associations was achieved with as few
212 as five selected characteristics (60,334 to 11,185), demonstrating the importance of
213 understanding and considering protein determinants rather than solely statistical significance
214 for distinguishing biologically relevant from false positive findings and type 1 errors.
215 Robust protein – disease associations included established clinical screening markers such as
216 prostate-specific antigen (referred to by Olink as KLK3) for prostate cancer (HR 3.06; p-value
217 $3.5x10^{-223}$), and markers of early tissue damage, such as the lung-specific surfactant protein D
218 (SFTPD) for post-inflammatory pulmonary fibrosis (*24*), or the eye-specific protein crystallin
219 beta B2 (CRYBB2) for cataract (HR 1.47, p-value $5.6x10^{-103}$). Less-established links with
220 persistent considerable effect size included a 2.5-fold increased risk (p-value $1.5x10^{-36}$) for
221 abdominal aortic aneurysm per 1 s.d. increase in plasma levels of matrix metalloproteinase
222 12 (MMP12) (**Fig. 4b, Supplementary Fig. 5**). Matrix metalloproteinases have been implicated
223 in the progression of abdominal aortic aneurysm (*25*), with MMP12 being highly expressed at
224 disease sites and suggested to contribute to the degradation of elastic fibres and hence
225 weakening of the aortic wall (*26*). We further obtained evidence for a potentially causal effect
226 of high plasma levels of MMP12 on the risk of abdominal aortic aneurysm (hazard ratio=1.15;
227 p-value=$3.3x10^{-16}$) by establishing a shared genetic signal (Posterior probability:
228 rs17368814=97.5%, **Supplementary Fig. 6**).
229 Prioritisation of biomarker associations has previously been proposed using genetic
230 imputation (*14*), suggesting that measuring genomic variation can proxy measurement of
231 plasma protein levels. However, even scaled to the power of the entire UK Biobank cohort, we
232 observed little overlap between associations with genetically imputed and measured plasma
233 protein levels, including those that we robustly linked to disease outcomes (**Fig. 4c-d;**
234 **Supplementary Tab. 9**). This included not only strongly differing effect estimates, and hence
235 power to stratify people at risk, but also discordant results whether considering cis- or trans-
236 pQTLs for imputing plasma protein levels (**Fig. 4c-d**).

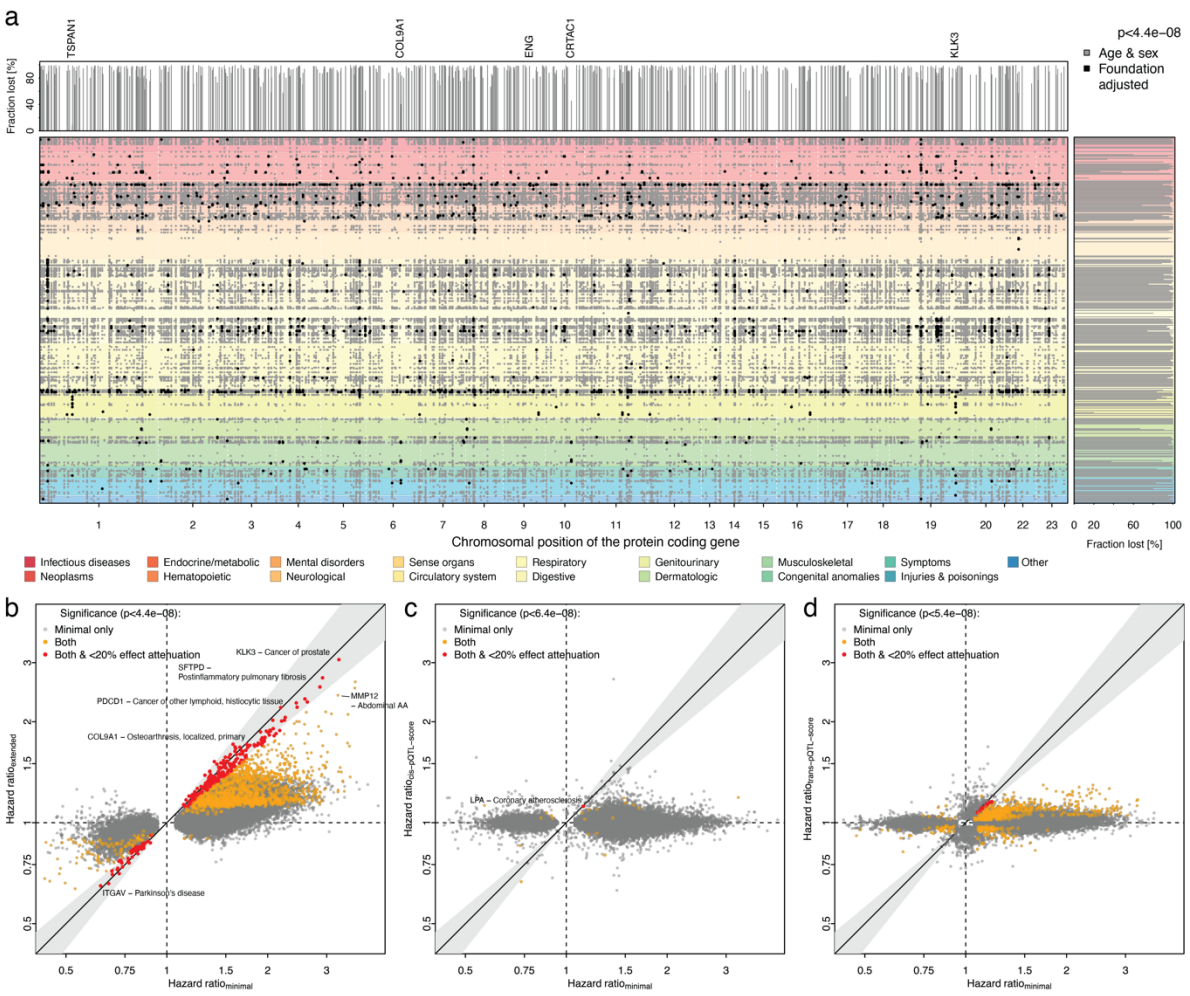

Figure 4. Protein – incident disease associations. a Summary of significant ($p<4.4x10^{-8}$) associations between plasma protein levels (x-axis indicated by the position of the protein-coding gene) and one or more of 390 diseases in UK Biobank using Cox-regression models. Each dot represents an association passing multiple testing, and black dots indicate those persisting after regressing out factors that explained plasma protein variation. The top panel illustrates the fraction of disease associations per protein that were no longer significant, similar to right panel for diseases. Proteins or diseases with minimal effect attenuation were annotated. **b** Scatterplot comparing hazard ratios per 1 s.d. increase in protein levels from Cox-regression models adjusting for age and sex or additionally accounting for factors explaining variation in plasma protein levels. Only protein – disease associations passing statistical significance in at least the minimal model are shown and coloured according to significance and effect attenuation in the extended model. **c-d** Scatter plots opposing effect estimates for significant (see legend) protein – disease associations comparing those using measured plasma protein levels (x-axis; n=43,647) to those based on genetically imputed plasma protein levels in the entire unrelated White-European UKB cohort (y-axis; n=342,240).

A protein foundation community resource

Most plasma proteomic studies are done at small scale, with often incomplete metadata on participants, adding to the complexity of the unknown sources of variation of most plasma proteins. We therefore developed a phenotype enrichment framework that allows to test for

256 relevant participant characteristics otherwise hidden in differentially expressed plasma
257 proteomic signatures (https://github.com/comp-med/r-prodente).
258 As a proof of principle, we observed proteins associated with the UKB characteristic 'fasting
259 time' to be more than 80-fold enriched (odds ratio: 88.7; p-value<1.7×10^{-5}; **Fig. 5a**) among
260 proteins significantly different following one day of complete caloric restriction in a well-
261 controlled intervention study (*27*). However, a similar enrichment of a protein signature
262 linked to plasma bilirubin was not reported due to missing measurements but resembles the
263 well-known increased reuptake of bilirubin via the enterohepatic cycle due to lower gut
264 motility during prolonged fasting (*28*, *29*). Another important application of phenotype
265 enrichment is the discovery of unknown confounders or imperfect matching in biomarker
266 studies. For example, we observed a more than 2-fold enrichment of proteins associated with
267 smoking among those differentially expressed between patients with ovarian carcinoma and
268 selected controls (*30*) (**Fig. 5a**). Plasma levels of one of the most differentially expressed
269 proteins in patients with cancer, Kunitz-type protease inhibitor 1 (SPINT1), was accordingly
270 best explained by smoking in UKB (6.9%; **Fig. 5b**). We observed similar residual phenotypic
271 enrichments, including smoking or socioeconomic factors, in a plasma proteomic model to
272 predict future coronary artery disease (*31*) (**Fig. 5b**), or among recently proposed proteomic
273 organ clocks (*32*) with those proxying pancreatic or brain age being enriched for established
274 markers of muscle mass (odds ratio: 27.2; p-value:6.3×10^{-10}) or plasma glucose (odds ratio:
275 4.4; p-value<7.8×10^{-6}), respectively (**Fig. 5c**).
276 Most notably, all studies showed strong depletion of proteins associated with technical
277 characteristics, such as study centre, suggesting that significant findings in well-controlled
278 settings are less likely to be driven by analytical artefacts. However, in study designs with
279 separate blood sampling of patients and controls, such as in the blood disease atlas from the
280 Human Protein Atlas project (*3*), we observed strong enrichment of such characteristics,
281 dominating biologically plausible findings (**Supplementary Fig. 7**).
282

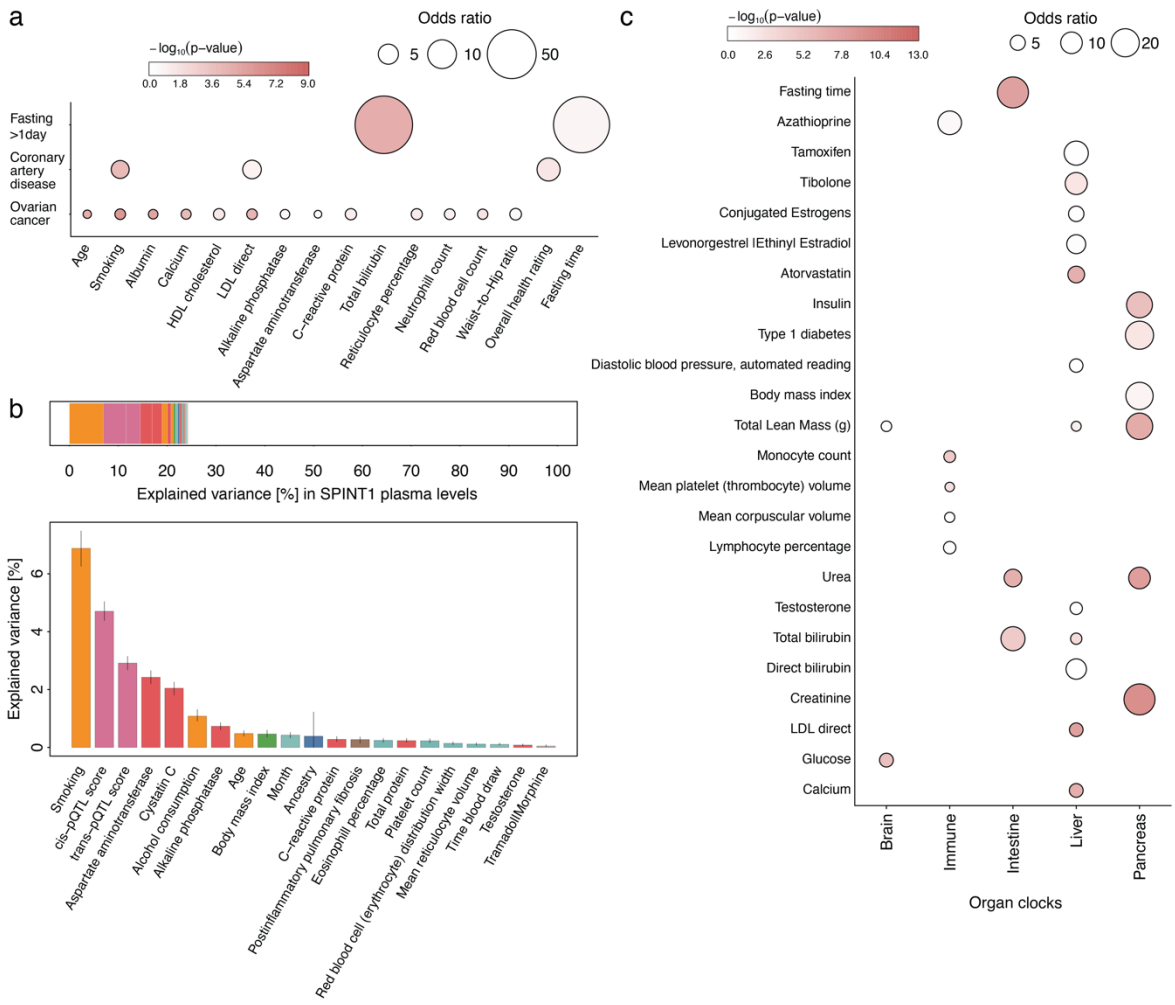

Figure 5 Phenotype enrichment of plasma proteomic signatures. a Phenotype enrichment for plasma proteomic signatures that 1) differed after one day of complete caloric restriction, 2) differentiated healthy women from women with ovarian cancer (*30*), and 3) a proteome score to predict onset of coronary artery disease CAD (*31*). Enrichment was computed using Fisher's exact test and only factors passing corrected statistical significance are shown. **b** Factors explaining variance in plasma levels of serine peptidase inhibitor, Kunitz type 1 (SPINT1), one of the most differential plasma proteins described for ovarian cancer. **c** Phenotype enrichment among plasma proteomic signatures proposed to track organ age (*32*).

A knowledge graph to triangulate evidence for clinical impact

We finally triangulated evidence from our current study with drug target annotations, pQTLs and potential effector genes, along with genetic disease associations within a shallow knowledge graph to visually and dynamically illustrate examples of potential clinical impact (**Fig. 6a; Supplementary Tab. 10-11**). This included 29 examples in which plasma protein levels might act as readouts of successful drug target engagement (**Fig. 6b**). We identified those as subnetworks that link drug – protein associations from the present study with genetic variants that change expression or function of the drug target and are associated with the same

299 protein(s), hence mimicking target modulation (**Fig. 6b**). For example, plasma levels of
300 desmoglein 2 were explained by pioglitazone intake and the corresponding pQTL, rs1801282,
301 maps to *PPARG* encoding the pioglitazone target peroxisome proliferator-activated receptor
302 gamma. Other such examples included plasma oxytocin as readout for oral contraceptives or
303 adiponectin as read out for atenolol treatment. The latter had previously been suggested from
304 small-scale trials in a specific target population (*33*), whereas a role of desmoglein 2 in
305 regulation of beta-cell activity has been proposed only recently (*34*).

306 In general, instances in which multiple genetic variants show similar associated protein
307 signatures with a disease, or drug may point to a causal relationship of the latter on protein
308 profiles. We observed 487 examples with support from enrichment analysis (**Supplementary**
309 **Tab. 12-13**). For example, cystatin E/M and galanin were associated with acne and three
310 independent pQTLs, which themselves have been reported to increase the risk for acne (**Fig.**
311 **6c and Supplementary Tab. 12**). The robust link between acne and the neuropeptide galanin
312 aligns with its expression in non-neural tissues such as skin (*35*) where it can modulate
313 inflammation in a context specific manner (*36*), although no link to acne has been made so
314 far. Similar triangulation with medication intake provided further strong support for the role
315 of platelet modulation on the plasma proteome with 41 genetic loci being significantly
316 enriched for proteins associated with the intake of the antiplatelet medication clopidogrel,
317 almost all of which have been previously linked to platelet characteristics (**Supplementary**
318 **Tab. 13**). We observed similar convergence of protein profiles for genetic predisposition to
319 and effects of drugs prescribed for osteoporosis (e.g., bisphosphonates like alendronic acid
320 linked to 7 independent pQTLs) or gastric ulcers (e.g., proton pump inhibitors like rabeprazole
321 linked to 3 independent pQTLs) (**Fig. 6c**), providing new avenues to understand drug effects.
322 An interactive version of the knowledge graph can be explored in a customized fashion to
323 identify other such examples (https://omicscience.org/apps/prot_foundation/).

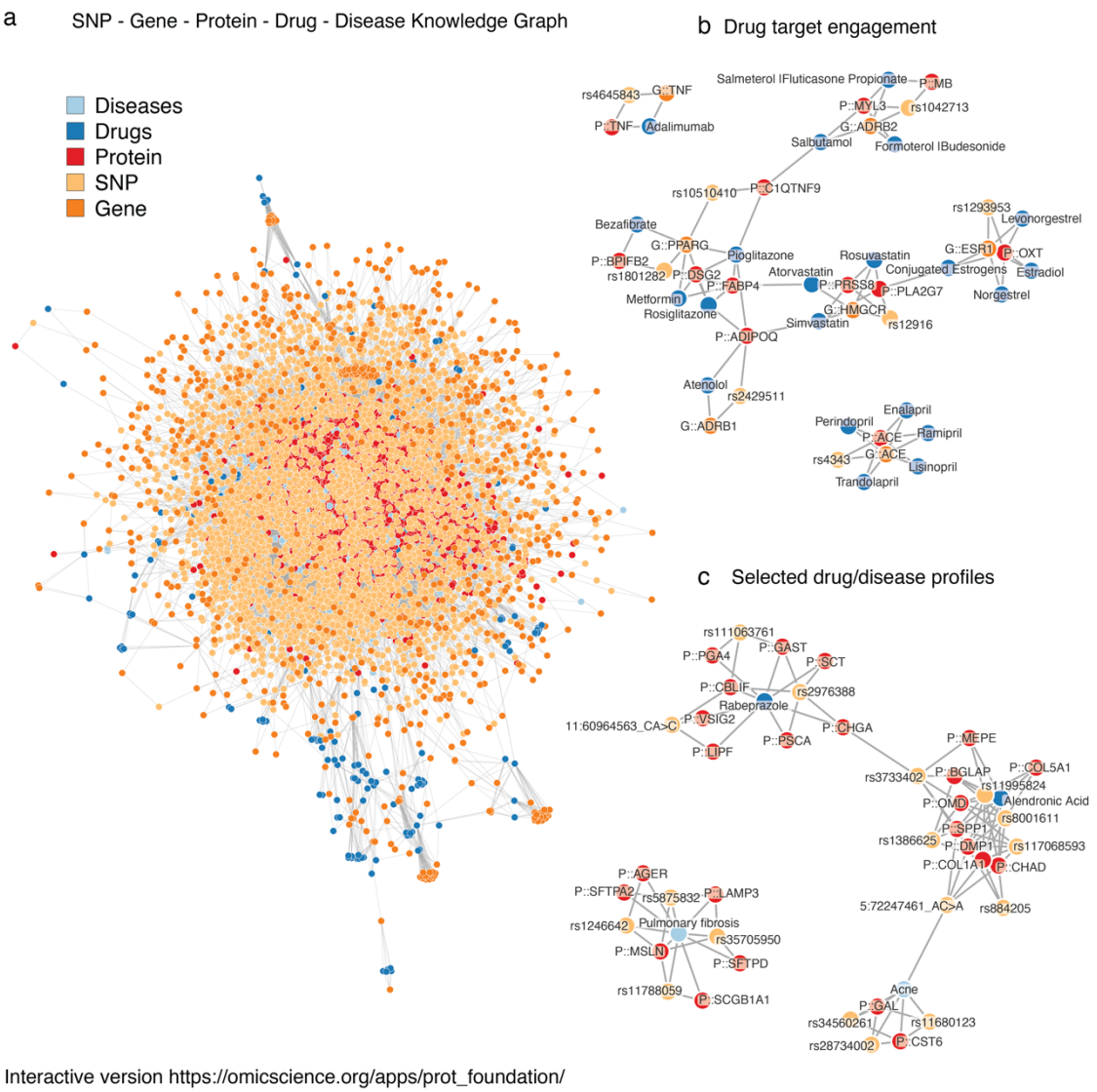

Figure 6 Integrated knowledge graph of gene, protein, drug, and disease relationships. **a** Full knowledge graph connecting single nucleotide polymorphisms (SNPs - pQTLs) to proteins and diseases from previous publications(*5*) and the GWAS catalog(*37*). Each SNP association represents a genome-wide significant finding. Drugs were mapped to target genes based on Open Targets(*38*). Protein – disease and protein - drug associations were derived from the present study. **b-c** Specific subnetworks derived from the knowledge graph illustrating evidence for drug target engagement markers (**b**) by linking pQTLs to drug target genes, and non-random pQTL – drug protein profiles (**c**). An interactive version of (**a**) can be found at https://omicscience.org/apps/prot_foundation/.

DISCUSSION

Deep phenotyping of biobank-sized cohorts now provides unprecedented opportunities to identify novel protein biomarkers for common and are diseases (*39*) and to inform drug target discovery (*5*). However, emerging production-type protein biomarker association analyses are already repeating failures of decades of biomarker research as they rarely translate into actionable knowledge. We created a data-driven model of plasma protein variation that

340 revealed cell type-specific, but also systemic effects and provides a foundation to understand
341 how variation in blood protein levels is linked to human health. We demonstrated its
342 importance in several ways, including biomarker identification for severe disease onset, drug
343 target engagement markers, and as a community resource to complement pathway
344 enrichment analysis providing otherwise unattainable insights. We created a community
345 resource to enable interactive exploration
346 (https://omicscience.org/apps/prot_foundation/) and incorporation in plasma proteomic
347 workflows (https://github.com/comp-med/r-prodente).

348 Although not directly comparable, we prioritized ten-times less protein – phenotype
349 combinations (~50,000 vs >500,000) compared to similar efforts within UKB (*7*, *15*)
350 emphasising the need to consider and quantify the effect of different factors simultaneously
351 rather than in isolation to overcome confounding. Only such increase in specificity of
352 phenotype – protein associations enabled tangible phenotype enrichment analysis, and
353 subsequently the identification of disease-specific biomarkers that are not only statistically
354 significant, but relevant. The same applies to the considerable amount of sharedness of
355 protein - disease associations proposed previously (*15*), that we demonstrated to be largely
356 driven by known phenotypic characteristics. At the same time, we provided evidence that
357 purely genetically-anchored strategies (*14*) are unlikely to discover powerful protein
358 biomarkers as they lack the dynamic component.

359 While we were underpowered to quantify ancestral-differential effects at scale, the proteins
360 we identified were either explained by varying effect allele frequencies or differential effect
361 sizes of the same genetic variant across ancestries. The former explains ancestral-enriched or
362 even -specific phenotypic consequences, such as for beta-thalassemia, whereas the latter
363 likely indicated the presence of yet to be identified gene - environment interactions. Larger
364 studies are needed to establish transferability of our results to different population subgroups
365 to ensure equitable biomarker research.

366 Pathway enrichment has been transformative in genome-wide differential gene expression
367 studies to enable interpretation of otherwise massive gene lists and has been almost
368 seamlessly adopted for proteomic studies. While such analyses can guide tissue or single cell
369 studies, proteins found in plasma are of diverse origin making the interpretation of such

370 findings difficult and possibly even misleading. We demonstrated across different applications,
371 how our results can guide the interpretation of plasma protein biomarker studies, including
372 novel insights but also pitfalls when differences in sample handling coincide with case –
373 control comparison. Future extensions may incorporate different proteomic technologies and
374 biofluids to enable widespread application.

375 The advances of our study have to be considered according to a number of limitations. Firstly,
376 our results remain restricted to the selection and precision of measurements of participant
377 and sample characteristics. Other data modalities, such as imaging of organs or exposure to
378 different environments or controlled challenge studies, will likely reveal additional
379 contributions to the plasma proteome. Secondly, we cannot entirely rule out that selected
380 participant characteristics may still approximate truly underlying reasons for variation in
381 plasma protein levels, and the causality of most of the observed associations remains to be
382 assessed experimentally. Thirdly, we observed evidence that variation in plasma levels of
383 proteins close to the limit of detection were less well explained suggesting that more sensitive
384 techniques for such protein targets may improve estimations of explained variance. Lastly,
385 while cis-pQTLs provide reassurance in assay specificity, we cannot completely rule out that
386 comparatively high amounts of variance explained by such cis-pQTLs might be the result of
387 measurement artefacts. This may imply that variance estimates for such targets therefore
388 relate to a specific isoform of the protein target. However, our approach to mutually model
389 genetic and non-genetic effects delivered reliable estimates for the latter, even if the genetic
390 part might have been inflated.

391

392 **METHODS**

393 ***Study participants***

394 UKB is a large-scale, population-based cohort with deep genetic and phenotypic data with the
395 full cohort consisting of approximately 500,000 participants (*40*). The participants were
396 recruited across centres in United Kingdom and were aged 40 to 69 years at the time of
397 recruitment (*40*). Ethics approval for the UKB study was obtained from the North West Centre
398 for Research Ethics Committee as a Research tissue biobank (REC reference 11/NW/0382)(*40*)
399 and all participants provided informed consent. We used the subset of individuals from UKB
400 where both genotype and proteomic measurements were available after excluding ancestry
401 outliers or samples which have failed genomic or proteomics quality control (n=43,240). This
402 research has been conducted using the UK Biobank Resource under Application Number
403 44448.

404 ***Proteomic measurements***

405 The UKB proteomic measurements were conducted for plasma samples of ~54,000
406 participants using the antibody-based Olink technology, Explore 3072 platform which uses the
407 Proximity Extension Assay technology (*41*). In summary, each protein is targeted by two
408 unique antibodies with unique complimentary oligonucleotides, which only hybridize when
409 they come into close proximity. This is subsequently quantified by next-generation
410 sequencing. Normalized protein expression (NPX) units, which are reported on a log2 scale,
411 were generated by normalization of the extension control and further normalization of the
412 plate control. Further details about antibody-based proteomic measurements and QC have
413 been described elsewhere, including the exclusion of samples due to poor quality and
414 selective measurements with assay warnings (*5*). After initial quality control checks (e.g.,
415 principal component analysis), we only retained individuals with at least 50% valid protein
416 measurements (n=43,240). Notably, UKB provided proteomic data with values flagged as
417 assay warnings by Olink as 'NA', which meant that each protein measurement had differing
418 numbers of valid values and we decided not to impute values for the purpose of this study to
419 minimize skewed estimates of features. We included a total of 2919 unique protein targets in
420 our analysis.

421 ***Phenotyping***

422 We collated a large set of phenotypic and sample characteristics available for at least half of
423 UKB participants (**Supplementary Tab. 1**). Those included information on genetically inferred

424 ancestry (see below), basic demographics (e.g., age and sex, n=5), blood-based biomarkers
425 (e.g., low-density lipoprotein cholesterol, n=25), blood cell counts (n=22), body composition
426 (n=4), indicators of bone health (n=3), cardiovascular risk factors (n=3), diet (n=22), pre-
427 existing diseases (n=1,106), self-reported medication intake (n=513), generic indicators of
428 health (n=4), indicators of pulmonary health (n=4), information on air pollution (n=6),
429 socioeconomic indicators (n=9), and technical variables (n=6), following pruning as described
430 below.
431 Measurement of blood-based biomarkers had been described in detail previously (*42*), and
432 we further implemented standard quality control procedures to avoid strong influences of
433 single values. We removed values more than 5 times the median absolute deviation away from
434 the respective sample median (median of ≤0.05% values removed). Biomarkers with strong
435 indications of skewed distributions were log-normalized (similar to protein values). We
436 applied a similar quality control workflow to blood cell counts. We obtained information of
437 when the blood sample was taken during the day, age of the sample when proteomics
438 measurements were done, and duration since last meal ('fasting time') as technical variables.
439 We compiled additional measures of body composition by augmenting Dual-X ray predictions
440 of lean and fat mass using equations provided by Powell *et al.* (*43*) for each sex separately. In
441 addition, we computed the waist-to-hip ratio as a common, readily available measure of body
442 composition.
443 To identify participants with pre-existing diseases, we collated information from ICD-coded
444 self-report, hospital episode statistics, cancer registry, and primary care (for 45% of the
445 population). We parsed all records to exclude codes with a recorded date before or within the
446 year of birth of the participant to minimize coding errors from electronic health records. We
447 mapped different coding systems to a total of 1,106 'phecodes' that represent medical
448 ontology terms intended to reduce redundancy among ICD-10 coding systems (*44*). We
449 recorded the earliest occurrence of each 'phecode', referred to hereafter as 'disease' from
450 simplicity, and created binary variables indicating whether it had been recorded before the
451 baseline examination of the volunteer.
452 We obtained mappings of self-reported brand names to ATC codes from previous work to
453 generate corresponding medication variables (*45*). We kept a total of 629 unique ATC codes
454 mapping to medications taken by 10 or more participants.

455 If not otherwise indicated information was derived from basic questionnaires or the data
456 portal (e.g., dietary habits and blood pressure readings) with minimal cleaning. We obtained
457 regional lead variants explaining variation in plasma protein levels from the trans-ancestral
458 meta-analysis performed by Sun *et al.* (*5*).
459 After initial data collation, we filtered for variables with a correlation coefficient <0.85
460 retaining among highly correlated variables the one with the least missing values or manually
461 curated to maintain interpretability. Since feature selection required a complete data matrix,
462 we imputed missing values in phenotypic data using Random-forest models that can cope with
463 a mixture of different data types better than usual linear regression techniques (*46*). We
464 implemented these using the R package miceRanger (v.1.5.0). To ensure reasonable
465 computation times, we performed imputation only once.
466 ***Genotyping and ancestral assignment***
467 The UK Biobank samples were genotyped using the Affymetrix UK BiLEVE or the Affymetrix
468 UK Biobank Axiom arrays. The following QC criteria was applied to the genotyping data (a)
469 routine quality checks carried out during the process of sample retrieval, DNA extraction, and
470 genotype calling; (b) checks and filters for genotype batch effects, plate effects, departures
471 from Hardy Weinberg equilibrium, sex effects, array effects, and discordance across control
472 replicates; and (c) individual and genetic variant call rate filters as previously described (*47*).
473 We used the ancestry assignments as defined by the pan-UKB (*48*), and further assigned
474 unclassified individuals to their respective ancestries based on a k-nearest neighbour
475 approach using genetic principal components.
476 ***Statistical analysis***
477 To select approximately independent participant and sample characteristics that may
478 collectively best explain variation in plasma protein levels, we implemented stability selection
479 (*19*) and subsequent variance partition (*49*). We first split UKB into a set for feature selection
480 (70%) and used the R package *stabs* (v.0.6.4) to implement feature selection using regularized
481 linear regression with the least absolute shrinkage and selection operator (LASSO). We
482 controlled the upper bound per-family error rate at 1, used a cut-off of 0.75, and a maximum
483 of 500 iterations. We controlled the output from stability selection by artificial introducing 10
484 random variables, but observed consistent behaviour across the study (i.e., random variables
485 were not selected as important). Based on the features selected, we next performed variance
486 decomposition on 200 bootstrap samples of the remaining UKB set (30%) implemented in a

487 mixed effect regression framework using the R package *variancePartition* (v.1.32.5)(*49*). We
488 modelled categorical variables as random effects and continuous variables as fixed effects. We
489 finally used the median of variance estimates across all 200 bootstrap samples as an estimate
490 for the explained variance in plasma protein levels.
491 We repeated the same procedure using different strata of UKB (biological sexes and
492 genetically inferred ancestry) as well as omitting blood-based biomarkers as one of the most
493 dominant explanatory factors. To establish whether differentially selected features indeed
494 corresponded to differential effects across strata, we additionally implemented interaction
495 testing by performing linear regression analysis with the respective protein target as outcome
496 and including the strata, the differentially selected feature, as well as an interaction effect
497 among both as explanatory variables. We subsequently report strata-differential effects if the
498 p-value for the interaction term passed multiple testing correction. This test was only done
499 for variables not specific to only of the strata, e.g., no interaction analysis was done for the
500 selection of oral contraceptive across the biological sexes.
501 To understand whether protein characteristics may have contributed to different levels of
502 explained variance, we collated information for all protein targets from UniProt via the R
503 package *queryup* (v.1.0.5) as well as the Human Protein Atlas (**Supplementary Fig. 1**). We
504 binarized information of protein characteristics from UniProt to facilitate numeric analysis. We
505 further computed summary measures of plasma protein level distributions based on
506 untransformed NPX values. We then used Boruta feature selection (*50*) to identify protein
507 characteristics significantly explaining variation in the levels of explained variance across
508 protein targets based on the entire UKB cohort. We used standard parameters apart from
509 setting the maximum number of iterations to 500 to ensure decisions on most characteristics.
510 We used Uniform Manifold Approximation and Projection (UMAP) to visualize any potential
511 underlying structure in the plasma proteome according to factors contributing to the
512 explained variation in plasma protein abundances. UMAP is a dimensionality reduction
513 technique which enables better preservation of the data's global structure compared to
514 similar methods such as t-SNE. We applied UMAP on the matrix of explained variation by all
515 participant, environmental, and technical characteristics on all protein targets. We used
516 default values for most parameters used by the algorithm in the *umap* R package. Custom
517 configuration was done for the following parameters as follows: random_state (seed for
518 random number gnerator) = 10, metric = "pearson", n_epochs = 1000, init = "random".

519 To establish clusters of protein targets, we implemented a k-nearest neighbour clustering
520 based on the full variance explained matrix. We systematically tested for silhouette
521 coefficients by varying k from 2 to 25 and selected k=11 as it minimized the within sum of
522 squares. We performed Fisher's exact test to test whether clusters were enriched for protein
523 targets explained by certain participant and sample characteristics.

524 ***Ancestral differential variants***

525 For each cis-pQTL with significant evidence for ancestral-differential effects, we obtained
526 ancestral-specific GWAS results (*5*) and retained the strongest regional variant for further
527 testing. We declared ancestral-specific effects, if the regional lead variant within the ancestry
528 was not included among a list of highly correlated genetic variants with the other ancestral
529 regional lead variants (r^2>0.1), basically testing for overlapping haplotypes across ancestries.
530 We subsequently tested each ancestral regional lead variant for effects in other ancestries and
531 computed the explained variance as well as tested for a significant interaction effects. The
532 latter was important to establish whether ancestral-differential explained variance was due to
533 different allele frequency or a differential effect size of similarly frequent alleles.

534 ***Generation of a knowledge graph***

535 We generated a shallow knowledge graph to facilitate triangulation of evidence for potential
536 causal relationships. The graph contained a total of 27,224 edges and 6,267 nodes. Nodes
537 included protein targets, pQTLs from Sun *et al.* (*5*), genes (either drug targets or effector genes
538 for pQTLs), drugs (based on ATC codes in the present study), and diseases. We linked proteins
539 to diseases (protein - disease) and drugs (protein - drug) based on evidence from the present
540 study. We included edges linking pQTLs to proteins, and further linked pQTLs to diseases based
541 on overlap with variants with genome-wide significance ($r^2 \geq 0.8$; $p<5x10^{-8}$) reported in the
542 GWAS catalog (download: 20/05/2024). We retained only entries from the GWAS catalog that
543 we could match to diseases defined in UK Biobank by linking Experimental Factor Ontology
544 (EFO) terms to ICD-10 codes using the EMBL-EBI API. We obtained information from Open
545 Targets (*38*) to map drugs to targets (genes) and restricted drugs to the once available in UK
546 Biobank by mapping CHEMBL identifiers to ATC codes. We finally used effector gene
547 assignments from Sun *et al.* (*5*) to link pQTLs to effector genes. To maximise the advantages
548 of the knowledge graph, we implemented a user-friendly interactive version using Python.
549 More specifically, we implemented the knowledge graph using the NetworkX Python package
550 (*51*) (v.3.3), and we enabled the knowledge graph to be interactively accessible using the Pyvis

551 Python package (*52*) (v.0.3.1). Notably, we customised the JavaScript generated by Pyvis to
552 enhance the functionality of the interactive knowledge graph, allowing the users to obtain
553 information about the neighborhood of the selected node and the shortest cycle path
554 associated with it, computed using the breadth-first search algorithm.

555 ***Protein – incident disease analyses***

556 We systematically investigated prospective associations of protein targets with 390 diseases
557 with more than 200 incident cases during a follow-up time of 10 years. For each protein –
558 disease pair, we employed separate Cox proportional hazards models that were adjusted for
559 age and sex. In case of sex-specific diseases, we used only the relevant subset of the
560 population. We computed and compared Cox models with protein levels in four different
561 forms, i.e., crude protein levels, residuals of proteins after regressing out all non-genetic
562 variance explaining factors, cis-pQTL scores, and trans-pQTL scores. We computed the protein
563 residuals from linear regression models containing variance explaining features applied to
564 each protein target. We entered the protein levels as continuous variables after performing
565 rank inverse normal transformation for each predictor to ensure comparability of results.
566 Hence, we report hazard ratios along with standard errors per increment in one standard
567 deviation for each protein. To account for potential reverse causality, we excluded cases that
568 occurred during the first 6 months. We considered associations significant when $p < 4.4 \times 10^{-8}$
569 $=0.05/(2919 \text{ proteins} \times 390 \text{ diseases})$. We computed time varying effects for each protein –
570 disease associations by computing Schoenfeld residuals, as well as by restricting follow-up
571 time to 2, 5, and 10 years. We used the R package *survival* (version 3.7.0) for these analyses.

572 ***Mapping to tissues and cell-types***

573 We programmatically downloaded tissue expression data from the Human Protein Atlas
574 (HPA) for all Olink proteins in JSON format (on 20.03.2024). We performed a two-sided
575 Fisher's exact test to determine whether tissue-specifically expressed proteins were enriched
576 among those best explained by the same major biological factors. We defined tissue- or cell-
577 type specific as 'enhanced', 'enriched', or 'group enriched' according to HPA classification. We
578 further performed two-sided Fisher's exact test to determine whether proteins predicted to
579 be secreted according to the HPA annotation were significantly enriched among any of
580 clusters of proteins best explained by the same biological influence.

581 ***A phenotype enrichment framework***

To enable phenotype enrichment, we first collated all results across all strata tested into one results file (**Supplementary Tab. 3**). For each stratum, we then implemented Fisher's exact test to test for an enrichment of a specified list of protein targets among protein signatures associated with participant and sample characteristics tested in our atlas work. We restricted enrichment tests to participant and sample characteristics with at least 5 associated, e.g., selected, protein targets. We used dot plots to visualize results. All analyses have been implemented in the R package *prodente* (https://github.com/comp-med/r-prodente) for dissemination. For plasma proteomic studies not using the Olink Explore 3072 platform, we used a specific background list of shared protein targets to minimize bias from differences in proteomic coverage. For the purpose of enrichment tests, we obtained the lists of differentially expressed or selected proteins as proposed by the authors of respective studies.

ACKNOWLEDGEMENT

The authors acknowledge the Scientific Computing of the IT Division at the Charité - Universitätsmedizin Berlin for providing computational resources that have contributed to the research results reported in this paper (https://www.charite.de/en/research/research_support_services/research_infrastructure/science_it/#c30646061). This work was supported by the DZHK (German Centre for Cardiovascular Research) and the BMBF (German Ministry of Education and Research) to C.L. (Grant Number: 81X2100281). Co-funded by the European Union (ERC, GenDrug, 101116072) to M.P.. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This work was supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) funded by the German Federal Ministry of Education and Research (BMBF) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

AUTHOR CONTRIBUTION

Conceptualization: MP, JCZ, CL

Data curation/Software: MP, CB, KD, JHA, WZ, WR, SY, BU, MZ

Formal Analysis: MP, CB, KD

Methodology: MP, KD, MK, JCZ, CL

Visualization: MP, KD, WZ, CL

Funding acquisition: MP, CL

Project administration: GK, CL

Supervision: MP, CL

Writing – original draft: MP, KD, CL

Writing – review & editing: MP, CB, KD, WZ, WR, MZ, MK, GK, JCZ, CL

CONFLICT OF INTEREST

None of the authors declare a conflict of interest.

DATA and CODE AVAILABILITY

All individual-level data is publicly available to bona fide researchers from the UK Biobank (https://www.ukbiobank.ac.uk/). Results can be obtained from https://omicscience.org/apps/prot_foundation/ and analysis code will be made available upon publication.

REFERENCES

1. E. J. Topol, The revolution in high-throughput proteomics and AI. *Science* **385**, eads5749 (2024).

2. J. P. A. Ioannidis, P. M. M. Bossuyt, Waste, Leaks, and Failures in the Biomarker Pipeline. *Clin. Chem.* **63**, 963–972 (2017).

3. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

4. M. Uhlén, M. J. Karlsson, A. Hober, A. Svensson, J. Scheffel, D. Kotol, W. Zhong, A. Tebani, L. Strandberg, F. Edfors, E. Sjöstedt, J. Mulder, A. Mardinoglu, A. Berling, S. Ekblad, M. Dannemeyer, S. Kanje, J. Rockberg, M. Lundqvist, M. Malm, A. Volk, P. Nilsson, A. Månberg, T. Dodig-crnkovic, E. Pin, M. Zwahlen, P. Oksvold, K. V. Feilitzen, R. S. Häussler, M. Hong, C. Lindskog, F. Ponten, B. Katona, J. Vuu, E. Lindström, J. Nielsen, J. Robinson, B. Ayoglu, D. Mahdessian, D. Sullivan, P. Thul, F. Danielsson, C. Stadler, E. Lundberg, G. Bergström, A. Gummesson, B. G. Voldborg, H. Tegel, S. Hober, B. Forsström, J. M. Schwenk, L. Fagerberg, Å. Sivertsson, The human secretome. **0274**, 1–9 (2019).

5. B. B. Sun, J. Chiou, M. Traylor, C. Benner, Y.-H. Hsu, T. G. Richardson, P. Surendran, A. Mahajan, C. Robins, S. G. Vasquez-Grinnell, L. Hou, E. M. Kvikstad, O. S. Burren, J. Davitte, K. L. Ferber, C. E. Gillies, Å. K. Hedman, S. Hu, T. Lin, R. Mikkilineni, R. K. Pendergrass, C. Pickering, B. Prins, D. Baird, C.-Y. Chen, L. D. Ward, A. M. Deaton, S. Welsh, C. M. Willis, N. Lehner, M. Arnold, M. A. Wörheide, K. Suhre, G. Kastenmüller, A. Sethi, M. Cule, A. Raj, Alnylam Human Genetics, AstraZeneca Genomics Initiative, Biogen Biobank Team, Bristol Myers Squibb, Genentech Human Genetics, GlaxoSmithKline Genomic Sciences, Pfizer Integrative Biology, Population Analytics of Janssen Data Sciences, Regeneron Genetics Center, L. Burkitt-Gray, E. Melamud, M. H. Black, E. B. Fauman, J. M. M. Howson, H. M. Kang, M. I. McCarthy, P. Nioi, S. Petrovski, R. A. Scott, E. N. Smith, S. Szalma, D. M. Waterworth, L. J. Mitnaul, J. D. Szustakowski, B. W. Gibson, M. R. Miller, C. D. Whelan, Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*, doi: 10.1038/s41586-023-06592-6 (2023).

6. R. S. Dhindsa, O. S. Burren, B. B. Sun, B. P. Prins, D. Matelska, E. Wheeler, J. Mitchell, E. Oerton, V. A. Hristova, K. R. Smith, K. Carss, S. Wasilewski, A. R. Harper, D. S. Paul, M. A. Fabre, H. Runz, C. Viollet, B. Challis, A. Platt, D. Vitsios, E. A. Ashley, C. D. Whelan, M. N. Pangalos, Q. Wang, S. Petrovski, Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).

7. G. Sveinbjornsson, M. I. Magnusson, A. Helgason, A. Oddsson, Large-scale plasma proteomics comparisons through genetics and disease associations. doi: 10.1038/s41586-023-06563-x (2023).

8. M. Pietzner, E. Wheeler, J. Carrasco-Zanini, A. Cortes, M. Koprulu, M. A. Wörheide, E. Oerton, J. Cook, I. D. Stewart, N. D. Kerrison, J. Luan, J. Raffler, M. Arnold, W. Arlt, S. O'Rahilly, G. Kastenmüller, E. R. Gamazon, A. D. Hingorani, R. A. Scott, N. J. Wareham, C. Langenberg, Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

9. B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, P. Surendran, C. Oliver-Williams, M. A. Kamat, B. P. Prins, S. K. Wilcox, E. S. Zimmerman, A. Chi, N. Bansal, S. L. Spain, A. M. Wood, N. W. Morrell, J. R. Bradley, N. Janjic, D. J. Roberts, W. H. Ouwehand, J. A. Todd, N. Soranzo, K. Suhre, D. S. Paul, C. S. Fox, R. M. Plenge, J. Danesh, H. Runz, A. S. Butterworth, Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

10. K. Suhre, M. I. McCarthy, J. M. Schwenk, Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).

11. V. Emilsson, M. Ilkov, J. R. Lamb, N. Finkel, E. F. Gudmundsson, R. Pitts, H. Hoover, V. Gudmundsdottir, S. R. Horman, T. Aspelund, L. Shu, V. Trifonov, S. Sigurdsson, A. Manolescu, J. Zhu, Ö. Olafsson, J. Jakobsdottir, S. A. Lesley, J. To, J. Zhang, T. B. Harris, L. J. Launer, B. Zhang, G. Eiriksdottir, X. Yang, A. P. Orth, L. L. Jennings, V. Gudnason, Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

12. J. Zheng, V. Haberland, D. Baird, V. Walker, P. C. Haycock, M. R. Hurle, A. Gutteridge, P. Erola, Y. Liu, S. Luo, J. Robinson, T. G. Richardson, J. R. Staley, B. Elsworth, S. Burgess, B. B. Sun, J. Danesh, H. Runz, J. C. Maranville, H. M. Martin, J. Yarmolinsky, C. Laurin, M. V. Holmes, J. Z. Liu, K. Estrada, R. Santos, L. McCarthy, D. Waterworth, M. R. Nelson, G. D. Smith, A. S. Butterworth, G. Hemani, R. A. Scott, T. R. Gaunt, Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

13. V. Zuber, N. F. Grinberg, D. Gill, I. Manipur, E. A. W. Slob, A. Patel, C. Wallace, S. Burgess, Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am. J. Hum. Genet.* **109**, 767–782 (2022).

14. Y. Xu, S. C. Ritchie, Y. Liang, P. R. H. J. Timmers, M. Pietzner, L. Lannelongue, S. A. Lambert, U. A. Tahir, S. May-Wilson, C. Foguet, Å. Johansson, P. Surendran, A. P. Nath, E. Persyn, J. E. Peters, C. Oliver-Williams, S. Deng, B. Prins, J. Luan, L. Bomba, N. Soranzo, E. Di Angelantonio, N. Pirastu, E. S. Tai, R. M. van Dam, H. Parkinson, E. E. Davenport, D. S. Paul, C. Yau, R. E. Gerszten, A. Mälarstig, J. Danesh, X. Sim, C. Langenberg, J. F. Wilson, A. S. Butterworth, M. Inouye, An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123–131 (2023).

15. Y.-T. Deng, J. You, Y. He, Y. Zhang, H.-Y. Li, X.-R. Wu, J.-Y. Cheng, Y. Guo, Z.-W. Long, Y.-L. Chen, Z.-Y. Li, L. Yang, Y.-R. Zhang, S.-D. Chen, Y.-J. Ge, Y.-Y. Huang, L.-M. Shi, Q. Dong, Y. Mao, J.-F. Feng, W. Cheng, J.-T. Yu, Atlas of the plasma proteome in health and disease in 53,026 adults. *Cell* **0** (2024).

16. T. Garcia, A. Petrera, S. M. Hauck, R. Baber, K. Wirkner, H. Kirsten, J. Pott, A. Tönjes, S. Henger, M. Loeffler, A. Peters, M. Scholz, Relationship of proteins and subclinical cardiovascular traits in the population-based LIFE-Adult study. *Atherosclerosis* **398** (2024).

17. D. H. Meyer, B. Schumacher, Aging clocks based on accumulating stochastic variation. *Nat. Aging* **4**, 871–885 (2024).

18. J. Carrasco-Zanini, E. Wheeler, B. Uluvar, N. Kerrison, M. Koprulu, N. J. Wareham, M. Pietzner, C. Langenberg, Mapping biological influences on the human plasma proteome beyond the genome. *Nat. Metab.*, 1–14 (2024).

19. N. Meinshausen, P. Bühlmann, Stability selection. *J R Stat. Soc B* **72**, 417–473 (2009).

20. P. E. Geyer, E. Voytik, P. V. Treit, S. Doll, A. Kleinhempel, L. Niu, J. B. Müller, M. Buchholtz, J. M. Bader, D. Teupser, L. M. Holdt, M. Mann, Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies. *EMBO Mol. Med.* **11**, e10427 (2019).

21. S. T. Yunga, A. J. Gower, A. R. Melrose, M. K. Fitzgerald, A. Rajendran, T. A. Lusardi, R. J. Armstrong, J. Minnier, K. R. Jordan, O. J. T. McCarty, L. L. David, P. A. Wilmarth, A. P. Reddy, J. E. Aslan, Effects of ex vivo blood anticoagulation and preanalytical processing time on the proteome content of platelets. *J. Thromb. Haemost.* **20**, 1437–1450 (2022).

22. M. Karlsson, C. Zhang, L. Méar, W. Zhong, A. Digre, B. Katona, E. Sjöstedt, L. Butler, J. Odeberg, P. Dusart, F. Edfors, P. Oksvold, K. von Feilitzen, M. Zwahlen, M. Arif, O. Altay, X. Li, M. Ozcan, A. Mardinoglu, L. Fagerberg, J. Mulder, Y. Luo, F. Ponten, M. Uhlén, C. Lindskog, A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7** (2021).

23. S. M. Kalman, Effects of Oral Contraceptives. *Annu. Rev. Pharmacol. Toxicol.* **9**, 363–378 (1969).

24. K. Ikeda, M. Shiratori, H. Chiba, H. Nishikiori, K. Yokoo, A. Saito, Y. Hasegawa, K. Kuronuma, M. Otsuka, G. Yamada, H. Takahashi, Serum surfactant protein D predicts the outcome of patients with idiopathic pulmonary fibrosis treated with pirfenidone. *Respir. Med.* **131**, 184–191 (2017).

25. F. A. M. V. I. Hellenthal, W. A. Buurman, W. K. W. H. Wodzig, G. W. H. Schurink, Biomarkers of AAA progression. Part 1: extracellular matrix degeneration. *Nat. Rev. Cardiol.* **6**, 464–474 (2009).

26. J. A. Curci, S. Liao, M. D. Huffman, S. D. Shapiro, R. W. Thompson, Expression and localization of macrophage elastase (matrix metalloproteinase-12) in abdominal aortic aneurysms. *J. Clin. Invest.* **102**, 1900–1910 (1998).

27. M. Pietzner, B. Uluvar, K. J. Kolnes, P. B. Jeppesen, S. V. Frivold, Ø. Skattebo, E. I. Johansen, B. S. Skålhegg, J. F. P. Wojtaszewski, A. J. Kolnes, G. S. H. Yeo, S. O'Rahilly, J. Jensen, C. Langenberg, Systemic proteome adaptions to 7-day complete caloric restriction in humans. *Nat. Metab.*, 1–14 (2024).

28. S. R. Gambino, SERUM-BILIRUBIN AND FASTING. *The Lancet* **300**, 1369 (1972).

29. P. V. D. Barrett, Hyperbilirubinemia of Fasting. *JAMA* **217**, 1349–1353 (1971).

30. L. Qian, J. Zhu, Z. Xue, Y. Zhou, N. Xiang, H. Xu, R. Sun, W. Gong, X. Cai, L. Sun, W. Ge, Y. Liu, Y. Su, W. Lin, Y. Zhan, J. Wang, S. Song, X. Yi, M. Ni, Y. Zhu, Y. Hua, Z. Zheng, T. Guo, Proteomic landscape of epithelial ovarian cancer. *Nat. Commun.* **15**, 6462 (2024).

31. H. Helgason, T. Eiriksdottir, M. O. Ulfarsson, A. Choudhary, S. H. Lund, E. V. Ivarsdottir, G. Hjorleifsson Eldjarn, G. Einarsson, E. Ferkingstad, K. H. S. Moore, N. Honarpour, T. Liu, H. Wang, T. Hucko, M. S. Sabatine, D. A. Morrow, R. P. Giugliano, S. R. Ostrowski, O. B. Pedersen, H. Bundgaard, C. Erikstrup, D. O. Arnar, G. Thorgeirsson, G. Masson, O. Th. Magnusson, J. Saemundsdottir, S. Gretarsdottir, V. Steinthorsdottir, G. Thorleifsson, A. Helgadottir, P. Sulem, U. Thorsteinsdottir, H. Holm, D. Gudbjartsson, K. Stefansson, Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–735 (2023).

32. H. S.-H. Oh, J. Rutledge, D. Nachun, R. Pálovics, O. Abiose, P. Moran-Losada, D. Channappa, D. Y. Urey, K. Kim, Y. J. Sung, L. Wang, J. Timsina, D. Western, M. Liu, P. Kohlfeld, J. Budde, E. N. Wilson, Y. Guen, T. M. Maurer, M. Haney, A. C. Yang, Z. He, M. D. Greicius, K. I. Andreasson, S. Sathyan, E. F. Weiss, S. Milman, N. Barzilai, C. Cruchaga, A. D. Wagner, E. Mormino, B. Lehallier, V. W. Henderson, F. M. Longo, S. B. Montgomery, T. Wyss-Coray, Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).

33. M. K. Pöyhönen-Alho, K. Manhem, P. Katzman, A. Kibarskis, R. L. Antikainen, R. U. Erkkola, J. O. Tuomilehto, P. E. Ebeling, R. J. Kaaja, Central sympatholytic therapy has anti-inflammatory properties in hypertensive postmenopausal women. *J. Hypertens.* **26**, 2445 (2008).

34. K. K. Myo Min, D. Rojas-Canales, D. Penko, M. DeNichilo, M. P. Cockshell, C. B. Ffrench, E. J. Thompson, O. Asplund, C. J. Drogemuller, R. B. Prasad, L. Groop, S. T. Grey, H. E. Thomas, T. Loudovaris, T. W. Kay, M. G. Mahoney, C. F. Jessup, P. T. Coates, C. S. Bonder, Desmoglein-2 is important for islet function and β-cell survival. *Cell Death Dis.* **13**, 1–15 (2022).

35. B. Kofler, A. Berger, R. Santic, K. Moritz, D. Almer, C. Tuechler, R. Lang, M. Emberger, A. Klausegger, W. Sperl, J. W. Bauer, Expression of Neuropeptide Galanin and Galanin Receptors in Human Skin. *J. Invest. Dermatol.* **122**, 1050–1053 (2004).

36. R. Lang, A. L. Gundlach, F. E. Holmes, S. A. Hobson, D. Wynick, T. Hökfelt, B. Kofler, Physiology, Signaling, and Pharmacology of Galanin Peptides and Receptors: Three Decades of Emerging Diversity. *Pharmacol. Rev.* **67**, 118–175 (2015).

37. A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, H. Parkinson, The NHGRI-EBI GWAS Catalog of published

genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

38. D. Ochoa, A. Hercules, M. Carmona, D. Suveges, J. Baker, C. Malangone, I. Lopez, A. Miranda, C. Cruz-Castillo, L. Fumis, M. Bernal-Llinares, K. Tsukanov, H. Cornu, K. Tsirigos, O. Razuvayevskaya, A. Buniello, J. Schwartzentruber, M. Karim, B. Ariano, R. E. Martinez Osorio, J. Ferrer, X. Ge, S. Machlitt-Northen, A. Gonzalez-Uriarte, S. Saha, S. Tirunagari, C. Mehta, J. M. Roldán-Romero, S. Horswell, S. Young, M. Ghoussaini, D. G. Hulcoop, I. Dunham, E. M. McDonagh, The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).

39. J. Carrasco-Zanini, M. Pietzner, J. Davitte, P. Surendran, D. C. Croteau-Chonka, C. Robins, A. Torralbo, C. Tomlinson, F. Grünschläger, N. Fitzpatrick, C. Ytsma, T. Kanno, S. Gade, D. Freitag, F. Ziebell, S. Haas, S. Denaxas, J. C. Betts, N. J. Wareham, H. Hemingway, R. A. Scott, C. Langenberg, Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**, 2489–2498 (2024).

40. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).

41. E. Assarsson, M. Lundberg, G. Holmquist, J. Björkesten, S. B. Thorsen, D. Ekman, A. Eriksson, E. R. Dickens, S. Ohlsson, G. Edfeldt, A. C. Andersson, P. Lindstedt, J. Stenvang, M. Gullberg, S. Fredriksson, Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9** (2014).

42. N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. R. Venkataraman, M. Wainberg, H. M. Ollila, T. Kiiskinen, A. S. Havulinna, J. P. Pirruccello, J. Qian, A. Shcherbina, F. Rodriguez, T. L. Assimes, V. Agarwala, R. Tibshirani, T. Hastie, S. Ripatti, J. K. Pritchard, M. J. Daly, M. A. Rivas, Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.*, doi: 10.1038/s41588-020-00757-z (2021).

43. R. Powell, E. de Lucia Rolfe, F. R. Day, J. R. B. Perry, S. J. Griffin, N. G. Forouhi, S. Brage, N. J. Wareham, C. Langenberg, K. K. Ong, Development and validation of total and regional body composition prediction equations from anthropometry and single frequency segmental bioelectrical impedance with DEXA. *medRxiv* **44**, 1–15 (2020).

44. L. Bastarache, Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).

45. Y. Wu, E. M. Byrne, Z. Zheng, K. E. Kemper, L. Yengo, A. J. Mallett, J. Yang, P. M. Visscher, N. R. Wray, Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10** (2019).

46. D. J. Stekhoven, P. Bühlmann, MissForest - nonparametric missing value imputation for mixed-type data. arXiv arXiv:1105.0828 [Preprint] (2011). https://doi.org/10.48550/arXiv.1105.0828.

47. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D. Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G. McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

48. K. J. Karczewski, R. Gupta, M. Kanai, W. Lu, K. Tsuo, Y. Wang, R. K. Walters, P. Turley, S. Callier, N. N. Shah, N. Baya, D. S. Palmer, J. I. Goldstein, G. Sarma, M. Solomonson, N. Cheng, S. Bryant, C. Churchhouse, C. M. Cusick, T. Poterba, J. Compitello, D. King, W. Zhou, C. Seed, H. K. Finucane, M. J. Daly, B. M. Neale, E. G. Atkinson, A. R. Martin, Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. medRxiv [Preprint] (2024). https://doi.org/10.1101/2024.03.13.24303864.

49. G. E. Hoffman, E. E. Schadt, variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).

50. M. B. Kursa, W. R. Rudnicki, Feature selection with the boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).

51. A. A. Hagberg, D. A. Schult, P. J. Swart, "Exploring Network Structure, Dynamics, and Function using NetworkX" (Pasadena, California, 2008; https://doi.curvenote.com/10.25080/TCWV9851), pp. 11–15.

52. G. Perrone, J. Unpingco, H. Lu, Network visualizations with Pyvis and VisJS. arXiv arXiv:2006.04951 [Preprint] (2020). https://doi.org/10.48550/arXiv.2006.04951.