Machine learning-guided deconvolution of plasma protein levels

Maik Pietzner[1,2,*], Carl Beuchel[1], Kamil Demircan[1,2], Julian Hoffmann 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)

Keywords Plasma proteomics/Biomarker/Drugs/Enrichment

ABSTRACT

Proteomic techniques now measure thousands of proteins circulating in blood 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 factors explaining variation in plasma levels of ~3,000 proteins among 43,240 participants of the UK Biobank that explain their links to ageing and diseases, and help guide protein biomarker and drug target discovery. We use machine learning to systematically identify a median of 20 factors (range: 1-37) out of >1,800 participant and sample charateristics that jointly explained an average of 19.4% (max. 100.0%) 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 (median: 10.0% vs 3.9%). We identify proteins for which the factors explaining varying levels in blood differed by sex (n=1374 proteins) or across ancestries (n=74 proteins). We establish a knowledge graph that integrates our findings with genetic studies and drug characteristics to guide identification of potential markers of drug 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 (https://github.com/comp-med/r-prodente).

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 proteins that are detectable but have no established role 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 based on comparatively cheap genotyping [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 systematically identify factors, including characterisitics of human health and disease but also technical measures, explaining variation in plasma protein levels building on our previous work [18]. We demonstrate that a relatively small number of factors (median: 20; range: 1-37) of all >1,800 tested explain a considerable part of protein variation (>25% for most protein targets) in plasma. Protein targets thereby segregated into distinct clusters explained by indicators of human health but also pre-analytical variation, such as accidental activation of platelets, and identify proteins that are best explained by different characteristics characteristics across the sexes and ancestral strata. We demonstrate how the collective knowledge generated here integrated with human genetic studies can guide 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

80 phenotype enrichment analyses in external studies (https://github.com/comp-med/r-
81 prodente).

82 **RESULTS**

83 We identified 411 diverse, modifiable and non-modifiable, participant and technical
84 characteristics (out of a total of 1879; **Fig. 1b** and **Tab. EV1**) to explain variation in plasma
85 levels of one or more of 2,853 protein targets (97.7%; **Tab. EV2**) among 30,268UK Biobank
86 participants (54.1% female; 5.2% Non-European) using regularised linear regression models
87 with stability selection [19]. A median of 20 characteristics (range: 1 - 37) was selected across
88 these protein targets and cumulatively explained on average 19.8% (range: 0.0005%-100.0%)
89 of variation in plasma protein levels in an independent validation set (n=12,972; **Fig. 1c and**
90 **Tab. EV3**). The highest amount of variance was explained for proteins actively secreted into
91 blood (median of 25.5%) and those demonstrated to be reliably detected using the assay
92 technology (median of 20.2% for 1,990 proteins with ≤5% of values below the limit of
93 detection). The ability to reliably detect mRNA levels of the protein coding gene in one or
94 more tissues was further associated with a higher amount of explained variance. (**Fig. EV1a-**
95 **h**).

96 We sought replication of our results in our previous work in a different cohort[18]. Despite
97 difference in the applied proteomic technology and available participant characteristics, the
98 total amount of explained variance was significantly correlated (r=0.28; p-value≤$6.7x10^{-37}$)
99 across 1,968 protein targets measured on both platforms (**Tab. EV4**). The correlation
100 coefficient improved to 0.46 (p-value≤$2.8x10^{-37}$) among 699 protein targets that have been
101 reported to correlate well across proteomic technologies[7], demonstrating good generasability
102 of our results. Improvements in explained variance by the present study were most strongly
103 associated with the inclusion of disease status (+14.4% per 1%-increase in explained variance
104 by disease status; p-value≤$4.8x10^{-17}$), genetically inferred ancestry (+8.5%; $1.3x10^{-31}$), or
105 considering parameters of general health (+18.4%; p-value≤$9.8x10^{-7}$). Results that showcased
106 the importance to expand the phenotypic space compared to our pioneering work.

107 ***A coordinate system for plasma protein variation***

108 Projecting the entire matrix of proteins times explaining factors into a lower-dimensional
109 space established a coordinate system along which proteins segregated into distinct clusters
110 according to shared major influences but also tissue and cell-type origin (**Fig. 1d**). The largest
111 cluster showed evidence for enrichment by multiple explanatory factors, whereas remaining
112 clusters distinguished by at most a few or even single characteristics explaining most of the
113 variance in plasma levels of proteins within the cluster (**Tab. EV5 and Fig. EV2**). For example,
114 one cluster contained 617 protein targets that were strongly enriched for participant and
115 sample characteristics indicating effects of technical variation and contamination by blood cell
116 activation. For example, recruitment centre (beta: 0.18; p-value<$4.9x10^{--08}$; mean explained
117 variance (MEV): 1.82%) or plateletcrit (PCT) (beta: 1.73; p-value<$2.8x10^{-264}$; MEV: 7.2%) were
118 most commonly selected for those proteins (**Fig. EV2k-l**). Our findings are in line with previous
119 studies [20–22] and are likely a result of platelet activation, and subsequent release of contained
120 proteins into plasma [21]. Collectively, these results provide evidence, that protein signatures in
121 blood can be partly deconvoluted into distinct origins. Modifiable participant characteristics
122 thereby outweighted non-modifiable ones, such as age, genetic sex, ancestry, or common
123 protein quantitative trait loci, on average (p<$7.5x10^{-47}$; two-sided Wilcoxon rank sum test; **Fig.**
124 **1d**).

125

126 ***Organ and cell-type contributions***

127 Proteins circulating in blood have diverse origins and we therefore integrated gene expression
128 data [3,23] to understand whether the selection of participant characteristics can be explained
129 by effects in certain tissues or cell-types. More than 80% (n=2,420) of all detected protein
130 targets showed evidence that the corresponding mRNA is preferentially expressed in one or
131 at most a few tissues or cell types enabling enrichment analysis in up to 13 tissues and 28 cell-
132 types (**Tab. EV2**). This identified 94 characteristics significantly more frequently selected for
133 proteins with tissue or cell-type specific mRNA expression **(Fig. 2; Fig. EV3 and Tab. EV6-7**).
134 Many of these enrichments were likely driven by tissue or cell damage. For example, proteins
135 with enhanced mRNA expression in the liver, specifically hepatocytes, were explained by
136 medications with adverse hepatic effects such as oral contraceptives [24] (e.g., conjugated
137 oestrogens; odds ratio: 5.62; p-value<$3.3x10^{-18}$) or carbamazepine (odds ratio: 8.98; p-

138 value<$1.3x10^{-13}$). Other medication associations more likely represented effects on target
139 organs/cells, such as proton pump inhibitors and stomach (e.g., esomeprazole; odds ratio:
140 237.10; p-value<$3.9x10^{-15}$) or immune suppressants and immune cells (e.g., azathioprine;
141 odds ratio: 13.01; p-value<$2.3x10^{-16}$).

142 On a cell population level, we observed yet unreported links between participant
143 characteristics and proteins with enhanced mRNA expression in fibroblasts that synthesize the
144 extra cellular matrix (ECM; **Fig. 2**). This included drug usage like beclomethasone dipropionate
145 (odds ratio: 14.24; p-value<$4.0x10^{-9}$), diseases like atrial fibrillation (odds ratio: 6.13; p-
146 value<$2.7x10^{-9}$), but also participant's age (odds ratio: 2.88; p-value<$6.7x10^{-10}$), body mass
147 index (BMI; odds ratio: 3.49; p-value<$7.7x10^{-13}$) and genetically inferred ancestry (odds ratio:
148 2.71; p-value<$7.2x10^{-9}$). These results provide evidence that changes in plasma protein levels
149 can to some extent be explained because of adverse, but also intentional, effects of drugs and
150 diseases on specific tissues or cell types..

151 ***Sex- and ancestral differential effects on the plasma proteome***

152 Genetically inferred ancestry (n=1,139) and participant's sex (n=1,199) were among the most
153 frequently selected non-modifiable participant characteristics. To better understand
154 potentially differential or specific effects across ancestries (41,369 White Europeans, 849
155 British Africans, and 823 British Central South Asians) and the sexes (23,601 females, 20,055
156 males), we repeated the feature selection procedure separately within each of the groups.

157 We identified significantly lower levels of explained variance in participants of British African
158 (median: -5.07%; IQR: -12.64% - -0.03%; p<$1.5x10^{-91}$) and to a much lesser extent British
159 Central South Asian ancestry (median: -0.03%; IQR: -4.36% - 3.90%; p<$1.3x10^{-2}$) compared to
160 White Europeans. Results, that were, at least among participants of British African ancestry,
161 not entirely explained by the lower sample size. We still observed a difference of ≥1.4% in
162 explained variance in plasma levels for more than half of the protein targets following
163 matching for the number of selected features across the White-European and British African
164 cohorts (IQR: -6.58% - 3.15%, p-value<$1.0x10^{-66}$; **Fig. 3a-b**) or when rerunning feature
165 selection in age- and sex-matched subsets of Europeans with the same sample size (median:
166 -2.62%; IQR: -9.34% - 0.00%; p-value<$6.1x10^{-29}$).

167 The single strongest contributor of differential estimates of explained variance in plasma
168 protein levels across ancestries was the contribution of cis- and trans-pQTL scores (**Tab. EV8**).
169 Possible explanations include that pQTLs are present at different minor allele frequencies
170 (MAF; **Fig. 3c-d**) across ancestries or may have differential effects on protein levels. We
171 identified 34 protein targets for which frequency-enriched ancestry-specific lead signals (non-
172 overlapping haplotype blocks; **Fig. 3d**) accounted for strong differences in explained variance.
173 For example, the missense variant rs2071421 (p.Asn352Ser) conferring arylsulfatase A pseudo
174 deficiency was most common in participants of British African ancestry (MAF_{AFR}=38.1%) but
175 less frequent in participants of other ancestries (MAF_{EUR}= 11.0%; MAF_{CSA}= 14.8%) and
176 accordingly explained considerably more variance (65.8% vs 5.9%) compared to the trans-
177 ancestral lead signal rs873697 (MAF_{AFR}=11.5%; MAF_{EUR}= 4.4%; MAF_{CSA}= 0.6%) in participants
178 of British African ancestry. However, most ancestral differential effects, including 135 out 137
179 cis-pQTLs, were due to the same pQTL having different effect sizes across ancestries with no
180 other nearby variant explaining those. The reason for different effects of the same genetic
181 variant across ancestries remains to be established.

182 We lastly identified few protein targets (n=22) that were much better explained (>4 s.d.) in
183 one but not the other genetically inferred sex (**Fig. EV4a-c**). We, however, noted that selected
184 participant characteristics differed for almost a third (n=895) of the protein targets (median
185 Jaccard index = 0.56; **Fig. EV4c**) indicating the need to consider sex-specific contributions to
186 plasma protein levels (**Tab. EV9**). Obvious examples included medications (n=1062 pairs, most
187 frequently anticontraceptives) and diseases (n=30 pairs) given/occurring only in one sex,
188 whereas abundant sex-differential effects were explained by associations with age,
189 biomarkers, or body mass index (**Fig. EV4d and Tab. EV9**). Notably, there were only few
190 examples of sex-differential genetic effects, e.g., the cis-pQTL score for plasma oxytocin (OXT;
191 female: 18.7%; male: 30.5%; $p\text{-value}_{inter}<2.8\times10^{-48}$; **Fig. EV4e**).

192

193 ***Protein biomarker discovery and pruning for incident diseases***

194 We next systematically explored how our results can guide the identification of plasma protein
195 biomarkers in biobank scale studies. Among the 67,033 significant protein – disease
196 associations ($p<4.1\times10^{-8}$) observed here and reported in other studies [7,15], we observed a
197 more than 32-fold drop after regressing out characteristics explaining variation in plasma

198 protein levels for 424 incident diseases (**Fig. 4a** and **Tab. EV10**). Associations of more than
199 two-thirds (1,333 out of 2,080) of protein targets were almost completely attenuated. Even
200 among the 1,975 protein – disease associations with directionally concordant effects and
201 persisting significance, >80% (n=1,691) showed considerable attenuation of effect sizes in Cox-
202 models (≥20%). This suggests that preciser measurements of associated participant
203 characterisitcs are likely to lead to further statistical attenuation (**Fig. 4b**). Notably, a >5-fold
204 decrease in the number of significant protein – disease associations was achieved with as few
205 as five selected characteristics (67,033 to 13,307), demonstrating the importance of
206 understanding and considering protein determinants rather than solely statistical significance
207 for distinguishing biologically relevant from false positive findings and type 1 errors.
208 Robust protein – disease associations included established clinical screening markers such as
209 prostate-specific antigen (referred to by Olink as KLK3) for prostate cancer (HR 3.11; p-value
210 $5.3x10^{-254}$), and markers of early tissue damage, such as the lung-specific surfactant protein D
211 (SFTPD) for post-inflammatory pulmonary fibrosis [25], or the eye-specific protein crystallin beta
212 B2 (CRYBB2) for cataract (HR 1.41, p-value $3.8x10^{-97}$). Less-established links with strong effect
213 sizes even after accounting for all selected phenotypic characteristics included a 2.3-fold
214 increased risk (p-value $1.1x10^{-38}$) for abdominal aortic aneurysm (AAA) per 1 s.d. increase in
215 plasma levels of matrix metalloproteinase 12 (MMP12) (**Fig. 4b, Fig. EV5**). A putative role of
216 MMP12 in the progressive degradation of the ECM at the aortic wall, a hallmark of AAA, was
217 thereby supported by multiple lines of evidence. Firstly, we obtained evidence that the same
218 genetic variant (rs17368814) that increases plasma MMP12 levels by acting on its protein
219 coding gene also increased the risk for AAA (Posterior probability shared genetic signal: 97.5%;
220 Mendelian randomization estimate: hazard ratio=1.15; p-value=$3.3x10^{-16}$; **Fig. EV6**), providing
221 support that changes in plasma MMP12 precede AAA onset. Secondly, MMP12 was elevated
222 among patients diagnosed with AAA even after adjustment for all factors explaining MMP12
223 levels (beta=1.09 s.d. units, p-value<$2.1x10^{-21}$), potentially indicating ongoing degradation of
224 the ECM. Thirdly, our observational findings are in line with experimtental evidence linking
225 matrix metalloproteinases to the onset and progression of AAA[26], with MMP12 being highly
226 expressed at disease sites and suggested to contribute to the degradation of elastic fibres and
227 hence weakening and dilation of the aortic wall [27]. These findings seem to be context-specific,
228 since pharmacolocigal inhibition of MMP12 protected *Apoe*$^{-/-}$ deficient mice from AAA[28],

229 while, paradoxically, *Mmp12*$^{-/-}$/*Apoe*$^{-/-}$ mice were more susceptible to AAA and subsequent
230 rupture.
231 Identification of putative protein biomarkers has previously been proposed using genetic
232 imputation [14], suggesting that measuring common genomic variation can proxy measurement
233 of plasma protein levels and hence guide tailored biomarker assessement. However, even
234 scaled to the power of the entire UK Biobank cohort, we observed little overlap between
235 associations with genetically imputed and measured plasma protein levels, including those
236 that we robustly linked to disease outcomes (**Fig. 4c-d; Tab. EV10**). For example, the five most
237 significant genetically proxied protein – disease association were not among the twenty most
238 strongly associated measured protein – disease associations for two-thirds of all diseases
239 considered (268 out of 419). This included not only strongly differing effect estimates, and
240 hence power to identify people at high risk early, but also discordant results whether
241 considering cis- or trans-pQTLs for imputing plasma protein levels (**Fig. 4c-d**).
242
243 ***A protein foundation community resource***
244 Most plasma proteomic studies are done at small scale, with often incomplete metadata on
245 participants, adding to the complexity of the unknown sources of variation of most plasma
246 proteins. We therefore developed a phenotype enrichment framework that allows to test for
247 significant enrichment (correcting for multiple testing using the Bonferroni procedure) of
248 participant characteristics otherwise hidden in differentially expressed plasma proteomic
249 signatures (https://github.com/comp-med/r-prodente).
250 As a proof of principle, we observed proteins associated with the UKB characteristic 'fasting
251 time' to be more than 90-fold enriched (odds ratio: 91.0; p-value<1.6×10^{-5}; **Fig. 5a**) among
252 proteins significantly different following one day of complete caloric restriction in a well-
253 controlled intervention study[29]. However, a similar enrichment of a protein signature linked
254 to plasma bilirubin was not reported due to missing measurements but resembles the well-
255 known increased reuptake of bilirubin via the enterohepatic cycle due to lower gut motility
256 during prolonged fasting[30,31]. Another important application of phenotype enrichment is the
257 discovery of unknown confounders or imperfect matching in biomarker studies. For example,
258 we observed a more than 2-fold enrichment of proteins associated with smoking among those
259 differentially expressed between patients with ovarian carcinoma and selected controls[32] (**Fig.
260 5a**). Accordingly, smoking status was among the top three variables explaining plasma

261 variation in the most differentially expressed proteins in patients with cancer, Kunitz-type
262 protease inhibitor 1 (SPINT1) (3.0%; **Fig. 5b**). We observed similar residual phenotypic
263 enrichments, including smoking or socioeconomic factors, in a plasma proteomic model to
264 predict future coronary artery disease[33] (**Fig. 5b**), or among recently proposed proteomic
265 organ clocks[34]. Those proxying pancreatic or intestinal age being enriched for established
266 markers of muscle mass (odds ratio: 28.3; p-value:$3.6x10^{-10}$) or fasting time (odds ratio: 28.2;
267 p-value<$1.6x10^{-9}$), respectively (**Fig. 5c**).
268 Most notably, all studies showed strong depletion of proteins associated with technical
269 characteristics, such as study centre, suggesting that protein biomarkers emerging from well-
270 controlled settings are less likely to be driven by analytical artefacts. However, in study designs
271 with separate blood sampling of patients and controls, such as in the blood disease atlas from
272 the Human Protein Atlas project [3], we observed strong enrichment of such characteristics,
273 dominating biologically plausible findings (**Fig. EV7**).
274
275 ***A knowledge graph to triangulate evidence for clinical impact***

276 We finally integrated results from the multivariable variance estimation with drug target
277 annotations, pQTLs and potential effector genes, along with genetic disease associations
278 within a shallow knowledge graph to visually and dynamically illustrate examples of potential
279 clinical impact (**Fig. 6a; Tab. EV11-12**). This included 34 examples in which plasma protein
280 levels might act as readouts of successful drug target engagement (**Fig. 6b**). We identified
281 those as subnetworks that link high-confidence drug – protein associations from the present
282 study with genetic variants that mimick drug target modulation by changing the expression or
283 function of the drug target and are associated with the same protein(s) (**Fig. 6b**). For example,
284 plasma levels of desmoglein 2 were explained by pioglitazone intake (0.2%, beta=1.02 s.d.
285 units, p-value=$3.7x10^{-39}$) and the corresponding pQTL, rs1801282 (beta=0.06; p-value<$1.4x10^{-}$
286 ^{12}), maps to *PPARG* encoding the pioglitazone target peroxisome proliferator-activated
287 receptor gamma. Other such examples included plasma oxytocin as readout for oral
288 contraceptives or adiponectin as read out for atenolol treatment. The latter had previously
289 been suggested from small-scale trials in a specific target population [35], whereas a role of
290 desmoglein 2 in regulation of beta-cell activity has been proposed only recently [36].

291 In general, enrichment of proteins differentially expressed in disease states for associations
292 with germeline genetic variants, i.e., pQTLs, can point to a causal relationships. We observed
293 595 examples with significant support from enrichment analysis (**Tab. EV13-14**). For example,
294 cystatin E/M (CST6), endonuclease, poly(U) specific (ENDOU), and galanin (GAL) were
295 associated with acne in our study (e.g., galanin: beta=0.27 s.d. units, p-value<$1.2x10^{-13}$) and
296 six independent pQTLs, which themselves have been reported to increase the risk for acne
297 (**Fig. 6c and Tab. EV13**). The robust link between acne and the neuropeptide galanin aligns
298 with its expression in non-neural tissues such as skin [37] where it can modulate inflammation
299 in a context specific manner [38]. Although no link to acne has been made so far. Similar
300 triangulation with medication intake provided further strong support for the role of platelet
301 modulation on the plasma proteome with 41 genetic loci being significantly enriched for
302 proteins associated with the intake of the antiplatelet medication clopidogrel, almost all of
303 which have been previously linked to platelet characteristics (**Tab. EV14**). We observed similar
304 convergence of protein profiles for genetic predisposition to and effects of drugs prescribed
305 for osteoporosis (e.g., bisphosphonates like alendronic acid linked to 5 independent pQTLs) or
306 gastric ulcers (e.g., proton pump inhibitors like rabeprazole linked to 3 independent pQTLs)
307 (**Fig. 6c**), providing new avenues to understand drug effects. An interactive version of the
308 knowledge graph can be explored in a customized fashion to identify other such examples
309 (https://omicscience.org/apps/prot_foundation/).

310 **DISCUSSION**

311 Deep phenotyping of biobank-sized cohorts now provides unprecedented opportunities to
312 identify novel protein biomarkers for common and rare diseases [39] and to inform drug target
313 discovery [5]. However, emerging production-type protein biomarker association analyses are
314 already repeating failures of decades of biomarker research as they rarely translate into
315 actionable knowledge. We created a data-driven model of plasma protein variation that
316 revealed cell type-specific, but also systemic effects and provides a foundation to understand
317 how variation in blood protein levels is linked to human health. We demonstrated its
318 importance in several ways, including biomarker identification for the onset of severe
319 diseases, drug target engagement markers, and as a community resource to complement
320 pathway enrichment analysis providing otherwise unattainable insights. We created a
321 community resource to enable interactive exploration

322 (https://omicscience.org/apps/prot_foundation/) and incorporation in plasma proteomic
323 workflows (https://github.com/comp-med/r-prodente).

324 Although not directly comparable, we prioritized ten-times less protein – phenotype
325 combinations (~50,000 vs >500,000) compared to similar efforts within UKB [7,15] emphasising
326 the need to consider and quantify the effect of different factors simultaneously rather than in
327 isolation to overcome confounding. Only such increase in specificity of phenotype – protein
328 associations enabled tangible phenotype enrichment analysis, and subsequently the
329 identification of disease-specific biomarkers that are not only statistically significant, but
330 relevant. The same applies to the considerable amount of sharedness of protein – disease
331 associations proposed previously [15], that we demonstrated to be largely driven by known
332 phenotypic characteristics. Notably, phenotype enrichment analysis of sparse predictive
333 biomarker signatures previously derived by our group[39] showed no evidence for widespread
334 enrichment of phenotypic characaterisitcs. It rather pointed to tissue damage markers or
335 refinement of imprecise phenotypic risk factors such as smoking or socioeconomic status (**Tab.**
336 **EV15**). At the same time, we provided evidence that purely genetically-anchored strategies [14]
337 are unlikely to discover powerful protein biomarkers as they lack the dynamic component.

338 Including >1,800 participant and technical characteristics explained on average about a
339 quarter of the variance in plasma protein levels across the population. Based on our findings,
340 we hypothesize that factors we had not considered or were not assessed in UK Biobank will
341 likely substantially increase the explained variance for at most few targets. Significant
342 improvements in the explained variance across all protein targets are more likely to be
343 achieved by improving assay performance and sample handling, but also by measuring
344 exposures more precisely. For example, quantifying the individual exposure to environmental
345 pollutants instead of regional measures or quantifying drug exposure using dosages instead
346 of self-report.

347 While we were underpowered to quantify ancestral-differential effects at scale, the proteins
348 we identified were either explained by varying effect allele frequencies or differential effect
349 sizes of the same genetic variant across ancestries. The former explains ancestral-enriched or
350 even -specific phenotypic consequences, such as for beta-thalassemia, whereas the latter
351 likely indicated the presence of yet to be identified gene – environment interactions. Larger

352 studies are needed to establish transferability of our results to different population subgroups
353 to ensure equitable biomarker research.

354 Pathway enrichment has been transformative in genome-wide differential gene expression
355 studies to enable interpretation of otherwise massive gene lists and has been almost
356 seamlessly adopted for proteomic studies. While such analyses can guide tissue or single cell
357 studies, proteins found in plasma are of diverse origin making the interpretation of such
358 findings difficult and possibly even misleading. We demonstrated across different applications,
359 how our results can guide the interpretation of plasma protein biomarker studies, including
360 novel insights but also pitfalls when differences in sample handling coincide with case –
361 control comparison. Future extensions may incorporate different proteomic technologies and
362 biofluids to enable widespread application.

363 The advances of our study have to be considered according to a number of limitations. Firstly,
364 our results remain restricted to the selection and precision of measurements of participant
365 and sample characteristics. Other data modalities, such as imaging of organs or exposure to
366 different environments or controlled challenge studies, will likely reveal additional
367 contributions to the plasma proteome. Secondly, we cannot entirely rule out that selected
368 participant characteristics may only approximate truly underlying reasons for variation in
369 plasma protein levels, and that observed associations, even if pertaining statistical significance
370 in multivariable linear regression models, do not necessarily represent causal associations.
371 Thirdly, we observed evidence that variation in plasma levels of proteins close to the limit of
372 detection were less well explained suggesting that more sensitive techniques for such protein
373 targets may improve estimations of explained variance. Lastly, while cis-pQTLs provide
374 reassurance in assay specificity, we cannot completely rule out that comparatively high
375 amounts of variance explained by such cis-pQTLs might be the result of measurement
376 artefacts. This may imply that variance estimates for such targets therefore relate to a specific
377 isoform of the protein target. However, our approach to mutually model genetic and non-
378 genetic effects delivered reliable estimates for the latter, even if the genetic part might have
379 been inflated.

380

381 **MATERIALS & METHODS**

382 ***Study participants***

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

392 ***Proteomic measurements***

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

409 ***Phenotyping***

410 We collated a large set of phenotypic and sample characteristics available for at least half of
411 UKB participants (**Tab. EV1**). Those included information on genetically inferred ancestry (see
412 below), basic demographics (e.g., age and sex, n=5), blood-based biomarkers (e.g., low-

density lipoprotein cholesterol, n=28), blood cell counts (n=29), body composition (n=20), indicators of bone health (n=7), cardiovascular risk factors (n=3), diet (n=23), pre-existing diseases (n=1,198), self-reported medication intake (n=704), generic indicators of health (n=4), indicators of pulmonary health (n=8), information on air pollution (n=10), socioeconomic indicators (n=9), and technical variables (n=5), following pruning as described below.

Measurement of blood-based biomarkers had been described in detail previously [42], and we further implemented standard quality control procedures to avoid strong influences of single values. We removed values more than 5 times the median absolute deviation away from the respective sample median (median of ≤0.05% values removed). Biomarkers with strong indications of skewed distributions were log-normalized (similar to protein values). We applied a similar quality control workflow to blood cell counts. We obtained information of when the blood sample was taken during the day, age of the sample when proteomics measurements were done, and duration since last meal ('fasting time') as technical variables. We compiled additional measures of body composition by augmenting Dual-X ray predictions of lean and fat mass using equations provided by Powell *et al.* [43] for each sex separately. In addition, we computed the waist-to-hip ratio as a common, readily available measure of body composition.

To identify participants with pre-existing diseases, we collated information from ICD-coded self-report, hospital episode statistics, cancer registry, and primary care (for 45% of the population). We parsed all records to exclude codes with a recorded date before or within the year of birth of the participant to minimize coding errors from electronic health records. We mapped different coding systems to a total of 1,198 'phecodes' that represent medical ontology terms intended to reduce redundancy among ICD-10 coding systems [44]. We recorded the earliest occurrence of each 'phecode', referred to hereafter as 'disease' from simplicity, and created binary variables indicating whether it had been recorded before the baseline examination of the volunteer.

We obtained mappings of self-reported brand names to ATC codes from previous work to generate corresponding medication variables [45]. We kept a total of 704 unique ATC codes mapping to medications taken by 10 or more participants.

If not otherwise indicated information was derived from basic questionnaires or the data portal (e.g., dietary habits and blood pressure readings) with minimal cleaning. We obtained

445 regional lead variants explaining variation in plasma protein levels from the trans-ancestral
446 meta-analysis performed by Sun *et al.* [5] to comput cis- and trans-pQTL scores. We used the
447 regression estimates provided to compute a weighted gentic risk score for each individual for
448 each protein with at least one reported pQTL.
449 After initial data collation, we filtered for variables with a correlation coefficient <0.85
450 retaining among highly correlated variables the one with the least missing values or manually
451 curated to maintain interpretability. This left us with a total of 1876 variables. Since feature
452 selection required a complete data matrix, we imputed missing values in phenotypic data
453 using Random-forest models that can cope with a mixture of different data types better than
454 usual linear regression techniques [46]. We used all selected phenotypic features for imputation
455 even if they had no missing values to ensure best predictive models for imputation. We further
456 imputed the entire data set before splitting into sets for feature selection and variance
457 estimation to ensure coherent estimates. We implemented these using the R package
458 miceRanger (v.1.5.0). To ensure reasonable computation times, we performed imputation
459 only once.
460 ***Genotyping and ancestral assignment***
461 The UK Biobank samples were genotyped using the Affymetrix UK BiLEVE or the Affymetrix
462 UK Biobank Axiom arrays. The following QC criteria was applied to the genotyping data (a)
463 routine quality checks carried out during the process of sample retrieval, DNA extraction, and
464 genotype calling; (b) checks and filters for genotype batch effects, plate effects, departures
465 from Hardy Weinberg equilibrium, sex effects, array effects, and discordance across control
466 replicates; and (c) individual and genetic variant call rate filters as previously described [47]. We
467 used the ancestry assignments as defined by the pan-UKB [48], and further assigned unclassified
468 individuals to their respective ancestries based on a k-nearest neighbour approach using
469 genetic principal components.
470 ***Statistical analysis***
471 To select approximately independent participant and sample characteristics that may
472 collectively best explain variation in plasma protein levels, we implemented stability selection
473 [19] and subsequent variance partition [49]. We first split UKB into a set for feature selection (70%)
474 and used the R package *stabs* (v.0.6.4) to implement feature selection using regularized linear
475 regression with the least absolute shrinkage and selection operator (LASSO). We controlled
476 the upper bound per-family error rate at 1, used a cut-off of 0.75, and a maximum of 500

477 iterations. We controlled the output from stability selection by artificial introducing 10
478 random variables, but observed consistent behaviour across the study (i.e., random variables
479 were not selected as important). Based on the features selected, we next performed variance
480 decomposition on 200 bootstrap samples of the remaining UKB set not used for feature
481 selection (30%) and computed the partial explained variance using the *etasq()* function of the
482 heplots R package (v.1.7.5) as recommended previously[16]. We finally used the median of
483 variance estimates across all 200 bootstrap samples as an estimate for the explained variance
484 in plasma protein levels and provide the 2.5th and 97.5th percentils as confidence intervals.
485 We repeated the same procedure using different strata of UKB (biological sexes and
486 genetically inferred ancestry). To establish whether differentially selected features indeed
487 corresponded to differential effects across strata, we additionally implemented interaction
488 testing by performing linear regression analysis with the respective protein target as outcome
489 and including the strata, the differentially selected feature, as well as an interaction effect
490 among both as explanatory variables. We subsequently report strata-differential effects if the
491 p-value for the interaction term passed multiple testing correction. This test was only done
492 for variables not specific to only one of the strata, e.g., no interaction analysis was done for
493 the selection of oral contraceptive across the genetically inferred sexes. We further created
494 four different subsamples of the White-European cohort to match sample sizes of British
495 African and British Central South Asian participants accounting for age and sex distributions.
496 We created four matched European subsamples for each ancestry and repeated the entire
497 feature selection. We took the median of the explained variance across all four subsamples as
498 summary measure.
499 To understand whether protein characteristics may have contributed to different levels of
500 explained variance, we collated information for all protein targets from UniProt via the R
501 package *queryup* (v.1.0.5) as well as the Human Protein Atlas (**Fig. EV1**). We binarized
502 information of protein characteristics from UniProt to facilitate numeric analysis. We further
503 computed summary measures of plasma protein level distributions based on untransformed
504 NPX values. We then used Boruta feature selection [50] to identify protein characteristics
505 significantly explaining variation in the levels of explained variance across protein targets
506 based on the entire UKB cohort. We used standard parameters apart from setting the
507 maximum number of iterations to 500 to ensure decisions on most characteristics.

508 We used Uniform Manifold Approximation and Projection (UMAP) to visualize any potential
509 underlying structure in the plasma proteome according to factors contributing to the
510 explained variation in plasma protein abundances. UMAP is a dimensionality reduction
511 technique which enables better preservation of the data's global structure compared to
512 similar methods such as t-SNE. We applied UMAP on the matrix of explained variation by all
513 participant, environmental, and technical characteristics on all protein targets. We used
514 default values for most parameters used by the algorithm in the *umap* R package. Custom
515 configuration was done for the following parameters as follows: random_state (seed for
516 random number gnerator) = 10, metric = "pearson", n_epochs = 1000, init = "random".
517 To establish clusters of protein targets, we implemented a k-nearest neighbour clustering
518 based on the full variance explained matrix, but excluding features selected ≤10 times to
519 reduce dimensional burden. We selected the number of clusters by visual inspection of the
520 UMAP mapping and corroborating with marker analysis using hurdle models. The latter was
521 implemented using the *zeroinfl()* function of the *pscl* R package (v.1.5.9) and we modelled the
522 cluster assignement as exposure and the explained variance by phenotypic characterisitcs as
523 count data to account for thresholding of the explained variance between 0 and 100 and an
524 inflation of zeros (i.e., characteristics not being selected).
525 ***Ancestral differential variants***
526 For each cis-pQTL with significant evidence for ancestral-differential effects, we obtained
527 ancestral-specific GWAS results [5] and retained the strongest regional variant for further
528 testing. We declared and effect as ancestry-specific, if the regional lead variant within the
529 ancestry was not included among a list of highly correlated genetic variants with the other
530 ancestral regional lead variants (r^2>0.1), basically testing for overlapping haplotypes across
531 ancestries. We subsequently tested each ancestral regional lead variant for effects in other
532 ancestries and computed the explained variance. We further tested, whether the association
533 between the variant and the corresponding protein differed by ancestry (significant
534 interaction effects). The latter was important to establish whether different amounts of
535 explained variants across ancestries were due to different allele frequencies or a differential
536 effect sizes of similarly frequent alleles.
537 ***Generation of a knowledge graph***
538 We generated a shallow knowledge graph to facilitate triangulation of evidence for potential
539 causal relationships. The graph contained a total of 30,068 edges and 6,591 nodes. Nodes

540 included protein targets, pQTLs from Sun *et al.* [5], genes (either drug targets or effector genes
541 for pQTLs), drugs (based on ATC codes in the present study), and diseases. We linked proteins
542 to diseases (protein - disease) and drugs (protein - drug) based on evidence from the present
543 study. We included edges linking pQTLs to proteins, and further linked pQTLs to diseases based
544 on overlap with variants with genome-wide significance ($r^2 \geq 0.8$; $p < 5x10^{-8}$) reported in the
545 GWAS catalog (download: 22/05/2025). We retained only entries from the GWAS catalog that
546 we could match to diseases defined in UK Biobank by linking Experimental Factor Ontology
547 (EFO) terms to ICD-10 codes using the EMBL-EBI API. We obtained information from Open
548 Targets [51] to map drugs to targets (genes) and restricted drugs to the once available in UK
549 Biobank by mapping CHEMBL identifiers to ATC codes. We finally used effector gene
550 assignments from Sun *et al.* [5] to link pQTLs to effector genes. We implemented a user-friendly
551 interactive version of the knowledge using Python. More specifically, we implemented the
552 knowledge graph using the NetworkX Python package [52] (v.3.3), and we enabled the
553 knowledge graph to be interactively accessible using the Pyvis Python package [53] (v.0.3.1).
554 Notably, we customised the JavaScript generated by Pyvis to enhance the functionality of the
555 interactive knowledge graph, allowing the users to obtain information about the
556 neighborhood of the selected node and the shortest cycle path associated with it, computed
557 using the breadth-first search algorithm.

558 ***Protein – incident disease analyses***

559 We systematically investigated prospective associations of protein targets with 390 diseases
560 with more than 200 incident cases during a follow-up time of 10 years. For each protein –
561 disease pair, we employed separate Cox proportional hazards models to model the
562 associations between baseline plasma protein levels and the time-to-event, adjusting for age
563 and sex of the participant. In case of sex-specific diseases, we used only the relevant subset
564 of the population. We computed and compared Cox models with protein levels in two
565 different forms, i.e., crude protein levels and residuals of proteins after regressing out
566 identified explanatory factors. For each protein, we ran a multivariable linear model including
567 all selected characteristics as explanatory variables and subsequently used the residuals (the
568 fraction of variation in protein levels not explained) from this model as updated exposure in
569 Cox-models. We computed similar residual plasma protein levels accounting for the top 3, 5,
570 10, and 15 selected features.We entered the protein levels as continuous variables after
571 performing rank inverse normal transformation for each predictor to ensure comparability of

572 results. Hence, we report hazard ratios along with standard errors per increment in one
573 standard deviation for each protein. To account for potential reverse causality, we excluded
574 cases that occurred during the first 6 months. We considered associations significant when $p<$
575 $4.4 \times 10^{-8} = 0.05/(2919 \text{ proteins} \times 390 \text{ diseases})$. We computed time varying effects for each
576 protein – disease associations by computing Schoenfeld residuals, as well as by restricting
577 follow-up time to 2, 5, and 10 years. We used the R package *survival* (version 3.7.0) for these
578 analyses.

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

580 We programmatically downloaded tissue expression data from the Human Protein Atlas
581 (HPA) for all Olink proteins in JSON format (on 20.03.2024). We performed a two-sided
582 Fisher's exacttest to determine for any participant characteristic selected >5 times across all
583 proteins, whether there was an enrichment of associated protein targets for those showing
584 specific expression of corresponding mRNA levels in certain tissues or cell-types. We defined
585 tissue- or cell-type specific as 'enhanced', 'enriched', or 'group enriched' according to HPA
586 classification. We further performed two-sided Fisher's exact test to determine whether
587 proteins predicted to be secreted according to the HPA annotation were significantly enriched
588 among any of the clusters of proteins explained by the same biological influence.

589 ***A phenotype enrichment framework***

590 To enable phenotype enrichment, we first collated all results across all strata (i.e., all
591 participants, the sexes, and the three ancestries) tested into one results file (**Tab. EV3**). For
592 each stratum, we then implemented Fisher's exact test to test for an enrichment of a specified
593 list of protein targets among protein signatures associated with participant and sample
594 characteristics tested in our atlas work. In other words, for a given list of proteins, we ask the
595 question whether there are participant characteristics more frequently selected than by
596 chance to explain variation in their plasma levels. We restricted enrichment tests to
597 participant and sample characteristics with at least 5 associated, e.g., selected, protein
598 targets. We used dot plots to visualize results. All analyses have been implemented in the R
599 package *prodente* (https://github.com/comp-med/r-prodente) for dissemination. For plasma
600 proteomic studies not using the Olink Explore 3072 platform, we used a specific background
601 list of shared protein targets to minimize bias from differences in proteomic coverage. For the
602 purpose of enrichment tests, we obtained the lists of differentially expressed or selected
603 proteins as proposed by the authors of respective studies.

604

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), as well as by the DFG (German Research Foundation) to K.D. (Walter Benjamin Fellowship, Grant Number: 547107463). 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).

DISCLOSURE and COMPETING INTEREST STATEMENT

None of the authors declare a conflict of interest.

DATA AVAILABILITY

The datasets and computer code produced in this study are available in the following databases:

- UK Biobank data: All individual-level data is publicly available to bona fide researchers from the UK Biobank (https://www.ukbiobank.ac.uk/).
- GWAS statistics: GWAS catalog (https://www.ebi.ac.uk/gwas/)
- RNA expression: Human Protein Atlas (https://www.proteinatlas.org/)
- Drug characteristics: Open Targets (https://platform.opentargets.org/downloads)
- Modelling computer scripts: GitHub (https://github.com/comp-med/protein-foundations-ukb-olink-50k)

636 • Results can be obtained from https://omicscience.org/apps/prot_foundation/

637

REFERENCES

1. Topol, E. J. The revolution in high-throughput proteomics and AI. *Science* **385**, eads5749 (2024).
2. Ioannidis, J. P. A. & Bossuyt, P. M. M. Waste, Leaks, and Failures in the Biomarker Pipeline. *Clin. Chem.* **63**, 963–972 (2017).
3. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
4. Uhlén, M. *et al.* The human secretome. **0274**, 1–9 (2019).
5. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* (2023) doi:10.1038/s41586-023-06592-6.
6. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. *Nature* **622**, 339–347 (2023).
7. Sveinbjornsson, G., Magnusson, M. I., Helgason, A. & Oddsson, A. Large-scale plasma proteomics comparisons through genetics and disease associations. (2023) doi:10.1038/s41586-023-06563-x.
8. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
9. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
10. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* **22**, 19–37 (2021).
11. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

662 12. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the
663 plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).
664 13. Zuber, V. *et al.* Combining evidence from Mendelian randomization and colocalization:
665 Review and comparison of approaches. *Am. J. Hum. Genet.* **109**, 767–782 (2022).
666 14. Xu, Y. *et al.* An atlas of genetic scores to predict multi-omic traits. *Nature* **616**, 123–131
667 (2023).
668 15. Deng, Y.-T. *et al.* Atlas of the plasma proteome in health and disease in 53,026 adults.
669 *Cell* **0**, (2024).
670 16. Garcia, T. *et al.* Relationship of proteins and subclinical cardiovascular traits in the
671 population-based LIFE-Adult study. *Atherosclerosis* **398**, (2024).
672 17. Meyer, D. H. & Schumacher, B. Aging clocks based on accumulating stochastic variation.
673 *Nat. Aging* **4**, 871–885 (2024).
674 18. Carrasco-Zanini, J. *et al.* Mapping biological influences on the human plasma proteome
675 beyond the genome. *Nat. Metab.* 1–14 (2024) doi:10.1038/s42255-024-01133-5.
676 19. Meinshausen, N. & Bühlmann, P. Stability selection. *J R Stat. Soc B* **72**, 417–473 (2009).
677 20. Geyer, P. E. *et al.* Plasma Proteome Profiling to detect and avoid sample-related biases in
678 biomarker studies. *EMBO Mol. Med.* **11**, e10427 (2019).
679 21. Yunga, S. T. *et al.* Effects of ex vivo blood anticoagulation and preanalytical processing
680 time on the proteome content of platelets. *J. Thromb. Haemost.* **20**, 1437–1450 (2022).
681 22. Korff, K. *et al.* Pre-Analytical Drivers of Bias in Bead-Enriched Plasma Proteomics.
682 2025.05.07.652495 Preprint at https://doi.org/10.1101/2025.05.07.652495 (2025).
683 23. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci. Adv.* **7**,
684 (2021).

24. Kalman, S. M. Effects of Oral Contraceptives. *Annu. Rev. Pharmacol. Toxicol.* **9**, 363–378 (1969).
25. Ikeda, K. *et al.* Serum surfactant protein D predicts the outcome of patients with idiopathic pulmonary fibrosis treated with pirfenidone. *Respir. Med.* **131**, 184–191 (2017).
26. Hellenthal, F. A. M. V. I., Buurman, W. A., Wodzig, W. K. W. H. & Schurink, G. W. H. Biomarkers of AAA progression. Part 1: extracellular matrix degeneration. *Nat. Rev. Cardiol.* **6**, 464–474 (2009).
27. Curci, J. A., Liao, S., Huffman, M. D., Shapiro, S. D. & Thompson, R. W. Expression and localization of macrophage elastase (matrix metalloproteinase-12) in abdominal aortic aneurysms. *J. Clin. Invest.* **102**, 1900–1910 (1998).
28. Di Gregoli, K., Atkinson, G., Williams, H., George, S. J. & Johnson, J. L. Pharmacological Inhibition of MMP-12 Exerts Protective Effects on Angiotensin II-Induced Abdominal Aortic Aneurysms in Apolipoprotein E-Deficient Mice. *Int. J. Mol. Sci.* **25**, 5809 (2024).
29. Pietzner, M. *et al.* Systemic proteome adaptions to 7-day complete caloric restriction in humans. *Nat. Metab.* 1–14 (2024) doi:10.1038/s42255-024-01008-9.
30. Gambino, S. R. SERUM-BILIRUBIN AND FASTING. *The Lancet* **300**, 1369 (1972).
31. Barrett, P. V. D. Hyperbilirubinemia of Fasting. *JAMA* **217**, 1349–1353 (1971).
32. Qian, L. *et al.* Proteomic landscape of epithelial ovarian cancer. *Nat. Commun.* **15**, 6462 (2024).
33. Helgason, H. *et al.* Evaluation of Large-Scale Proteomics for Prediction of Cardiovascular Events. *JAMA* **330**, 725–735 (2023).
34. Oh, H. S.-H. *et al.* Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).

35. Pöyhönen-Alho, M. K. *et al.* Central sympatholytic therapy has anti-inflammatory properties in hypertensive postmenopausal women. *J. Hypertens.* **26**, 2445 (2008).
36. Myo Min, K. K. *et al.* Desmoglein-2 is important for islet function and β-cell survival. *Cell Death Dis.* **13**, 1–15 (2022).
37. Kofler, B. *et al.* Expression of Neuropeptide Galanin and Galanin Receptors in Human Skin. *J. Invest. Dermatol.* **122**, 1050–1053 (2004).
38. Lang, R. *et al.* Physiology, Signaling, and Pharmacology of Galanin Peptides and Receptors: Three Decades of Emerging Diversity. *Pharmacol. Rev.* **67**, 118–175 (2015).
39. Carrasco-Zanini, J. *et al.* Proteomic signatures improve risk prediction for common and rare diseases. *Nat. Med.* **30**, 2489–2498 (2024).
40. Sudlow, C. *et al.* 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. Assarsson, E. *et al.* Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PLoS ONE* **9**, (2014).
42. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* (2021) doi:10.1038/s41588-020-00757-z.
43. Powell, R. *et al.* 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. Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu. Rev. Biomed. Data Sci.* **4**, 1–19 (2021).
45. Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, (2019).

46. Stekhoven, D. J. & Bühlmann, P. MissForest - nonparametric missing value imputation for mixed-type data. Preprint at https://doi.org/10.48550/arXiv.1105.0828 (2011).
47. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
48. Karczewski, K. J. *et al.* Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. 2024.03.13.24303864 Preprint at https://doi.org/10.1101/2024.03.13.24303864 (2024).
49. Friendly, M. HE Plots for Multivariate Linear Models. *J. Comput. Graph. Stat.* **16**, 421–444 (2007).
50. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
51. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Res.* **51**, D1353–D1359 (2023).
52. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in 11–15 (Pasadena, California, 2008). doi:10.25080/TCWV9851.
53. Perrone, G., Unpingco, J. & Lu, H. Network visualizations with Pyvis and VisJS. Preprint at https://doi.org/10.48550/arXiv.2006.04951 (2020).
54. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

755 **FIGURE LEGENDS**
756
757 **Figure 1 A coordinate system of plasma protein variation.**
758 **a** Secretome assignment[4] for 2919 protein targets analysed in the present study.
759 **b** Number and type of UK Biobank participant characteristics considered in the study.
760 **c** Stacked bar charts displaying the achieved variance explained for each protein target.
761 Proteins are ordered by explained variance across all factors. Colours indicate domains of
762 explaining factors.
763 **d** Uniform Manifold Approximation and Projection (UMAP) mapping of the variance explained
764 matrix across 2853 protein targets for which we identified at least one feature explaining the
765 variance in plasma levels. Each protein has been assigned a cluster based on k-means
766 clustering and is coloured accordingly.
767 **e** Violin plots showing the explained variance across protein targets according to categories of
768 selected characteristics. The centre of each violin plot is a boxplot giving median (white dot)
769 and interquartile range.
770 Data information: In E, data are presented as boxplots, indicating median, interquartile range
771 (IQR), and whiskers for >1.5 times IQR
772
773 **Figure 2 Plasma proteins link cell-types to indicators of health and disease.**
774 Chord diagram of phenotype associated protein signature enrichment among protein coding
775 genes with enhanced cell type expression ('marker genes'). Each line represents a significant
776 enrichment (Fisher's test; $p<2.9x10^{-6}$) of proteins associated with a participant characteristic
777 among protein coding genes with enhanced expression in a cell type. Enhanced expression
778 estimates were derived from single-cell RNA sequencing data in the Human Protein atlas.
779 Corresponding statistics can be found in Tab. EV7. Associations with marker genes of
780 fibroblasts have been highlighted by darker colours.
781
782 **Figure 3 Summary of ancestral-stratified analysis.**
783 **a** Comparison of the achieved explained variance in plasma proteins levels within White
784 Europeans compared to British Africans participants. Proteins are coloured based on cluster
785 membership and those with extreme differences are annotated. *The number of selected
786 features has been matched with the smaller ancestry to account for differences in the power
787 of discovery.
788 **b** Same as **a**, but now comparing to participants of British Central South Asian ancestry.
789 **c-d** Contrasting differences in allele frequencies for cis-pQTL scores with changes in explained
790 variance between White Europeans and British African (c, AFR) or British Central South Asian
791 (d, CSA).
792 **e** Variance explained estimates for 34 cis-pQTLs with evidence for distinct ancestral-lead
793 signals. Darker colours indicate the effect of the trans-ancestral signal, whereas shades
794 indicate the explained variance by the ancestry-specific lead signal. We note that the effect of

ancestry-specific lead signals might be overestimated, since they have not been selected from an independent cohort.

Data information: In (A-B), point estimates were derived as variance explained (r^2) from a multivariable linear regression model; In E, partial variance explained was computed for each genetic variant based on a multivariable linear regression model

Figure 4. Protein – incident disease associations.

a Summary of significant (Cox-proportional hazard model; $p<4.1x10^{-8}$) associations between plasma protein levels (x-axis indicated by the position of the protein-coding gene) and one or more of 424 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 still significant after accounting for associated protein characteristics, 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).

Data information: In (A-D), each dot reflects a significant association between plasma protein levels and disease onset based on Cox-proportional hazard models

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 [32], and 3) a proteome score to predict onset of coronary artery disease CAD [33]. Enrichment was computed using Fisher's exact test and only factors passing corrected statistical significance are shown (Bonferroni-correction).

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 [34].

Data information: In (A and B), enrichment statistics were derived based on Fisher's exact test; In B; partial variance estimates were derived from a multivariable linear regression model

836 **Figure 6 Integrated knowledge graph of gene, protein, drug, and disease relationships.**
837 **a** Full knowledge graph connecting single nucleotide polymorphisms (SNPs – pQTLs) to
838 proteins and diseases from previous publications[5] and the GWAS catalog[54]. Each SNP
839 association represents a genome-wide significant finding. Drugs were mapped to target genes
840 based on Open Targets[51]. Protein – disease and protein – drug associations were derived from
841 the present study. The graph was visualised using cytoscape v.3.10.3.
842 **b-c** Specific subnetworks derived from the knowledge graph illustrating evidence for drug
843 target engagement markers (**b**) by linking pQTLs to drug target genes, and non-random pQTL
844 – drug protein profiles (**c**). An interactive version of (**a**) can be found at
845 https://omicscience.org/apps/prot_foundation/.
846
847

848 **EXPANDED VIEW FIGURE LEGENDS**
849
850 **Figure EV1 Summary of protein and assay characteristics associated with the variance**
851 **explained achieved in plasma protein levels.**
852 **a** Variable importance of protein and assays characteristics based on Boruta feature selection
853 predicting the variance explained achieved for each protein target. Boxplots indicate the
854 distribution of the variable importance across 500 iterations. Darker colours indicate features
855 passing corrected statistical significance ($p<0.01$).
856 **b** Scatterplot opposing the skewness of individual plasma protein distributions with the
857 variance explained.
858 **c – h** Variance explained according to different criteria deemed important by the Boruta
859 feature selection.
860
861 **Figure EV2 Major foundations of plasma protein variation.**
862 **a** Uniform Manifold Approximation and Projection (UMAP) mapping of the variance explained
863 matrix across 2853 protein targets for which we identified at least one feature explaining the
864 variance in plasma levels. Each protein has been assigned a cluster based on k-means
865 clustering and is coloured accordingly.
866 **b** Number of protein targets included in each cluster.
867 **c – l** Same UMAP plot but coloured according to the variance explained by the factor given on
868 top of each plot. Proteins with strong contributions (>1%) are highlighted. pQTL = protein
869 quantitative trait loci.
870
871 **Figure EV3 Chord diagram of phenotype associated protein signature enrichment among**
872 **protein coding genes with enhanced tissue expression.** Each line represents a significant
873 enrichment ($p<6.9x10^{-6}$) of proteins associated with a given participant characteristic among
874 protein coding genes with enhanced expression in a given tissue. Enhanced expression
875 estimates were derived from the Human Protein atlas. Corresponding statistics can be found
876 in Table EV7. Colouring was done according to phenotype categories as introduced in Figure
877 1 in the main text.
878
879 **Figure EV4 Summary of sex-differential feature selection and contribution.**
880 **a-b** Scatterplots comparing the variance explained achieved for plasma protein levels among
881 the entire UK Biobank population (x-axis) compared to what was achieved in females (left) and
882 males (right) alone. Proteins that deviated the most (>4 s.d.) were annotated.
883 **c** Comparison of cumulative variance explained when stratifying the UK Biobank population
884 by sex. Proteins are coloured by cluster assignments as introduced in Fig. 1. Proteins with
885 strong differences are annotated with gene names. The inlet depicts the distribution of the
886 Jaccard index of overlapping features for the same protein across the sexes.
887 **d** Protein – Feature combinations with significant evidence for sex-differential effects
888 ($p<9.0x10^{-7}$). Only combinations with a difference of more than 10% are shown.

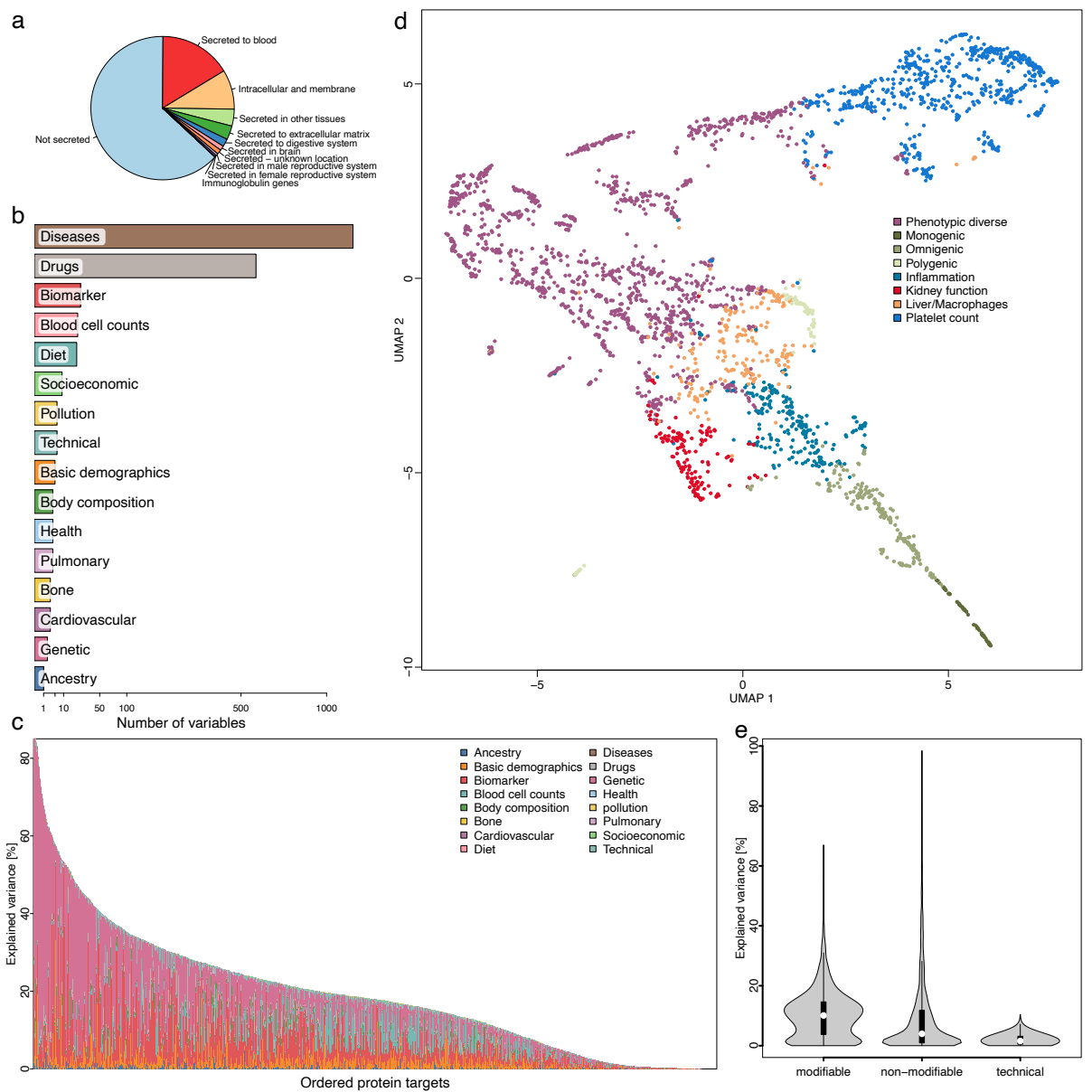
e Individual variance explained estimates for plasma oxytocin levels by sex, ordered by the variance explained among females. Proteins in b and c are coloured according to feature categories as introduced in Fig. 1.

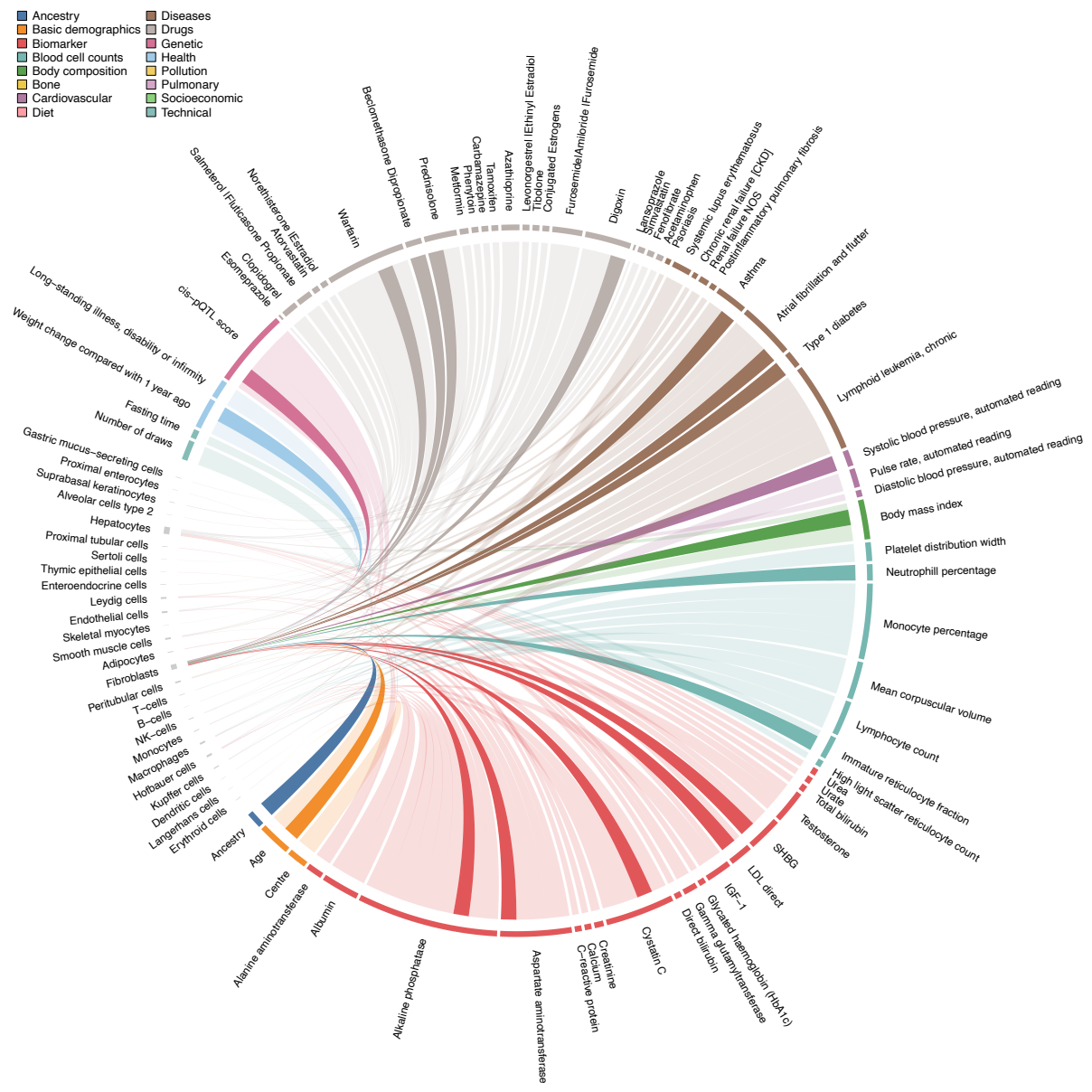
Figure EV5 Volcano plots for protein - abdominal aortic aneurysm associations, with adjustment for age and sex (left panel), or comprehensive adjustment based on results from the feature selection algorithm for each protein target (right panel).

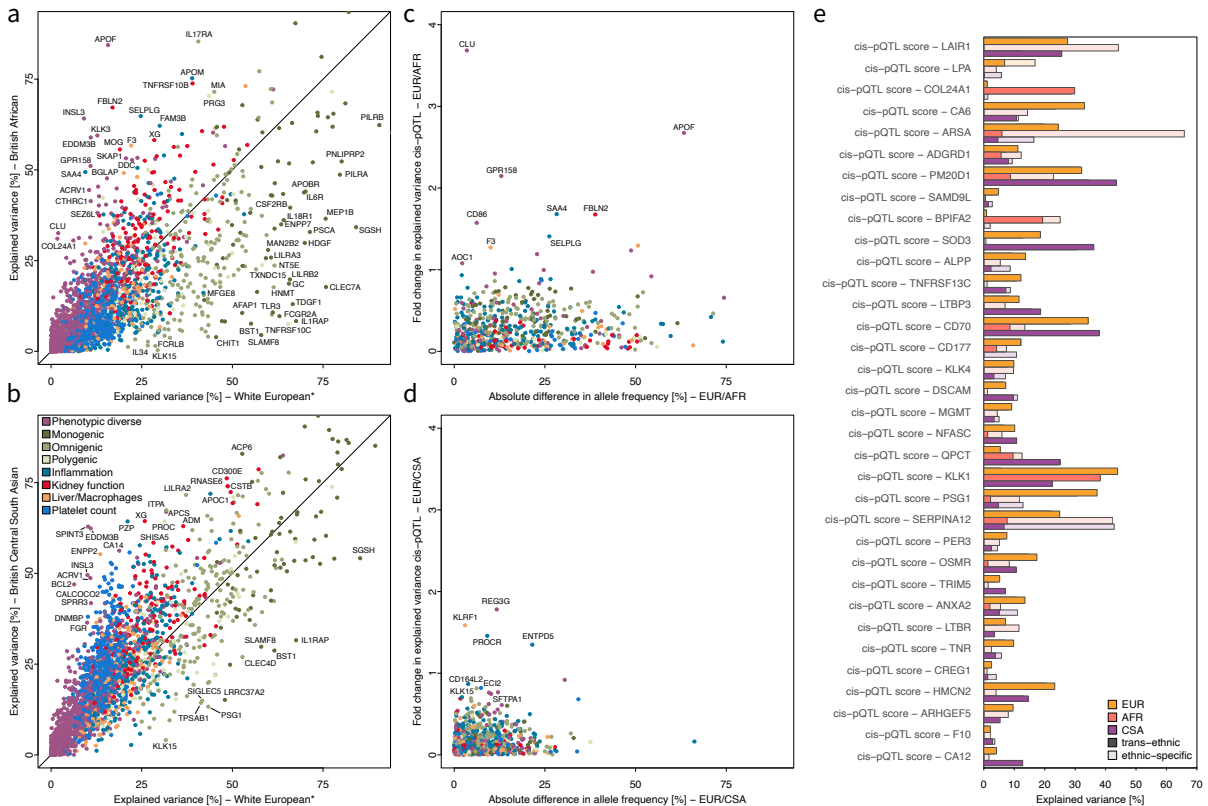
Figure EV6 Regional association plot of the *MMP12* locus for MMP12 protein plasma levels (top) and abdominal aortic aneurysm (AAA, bottom). One SNP (rs17368814) have been prioritized as shared genetic signal. Summary statistics from logistic regression models for AAA are publicly available from the AAAgen consortium and summary statistics from linear regression models for MMP12 plasma levels are based on UK Biobank. Coloring indicates linkage disequilibrium to the lead variant of the same color code. Probability for a shared genetic signal (PP-H_4) is given as $PP_{rs17368814}$=97.5%.

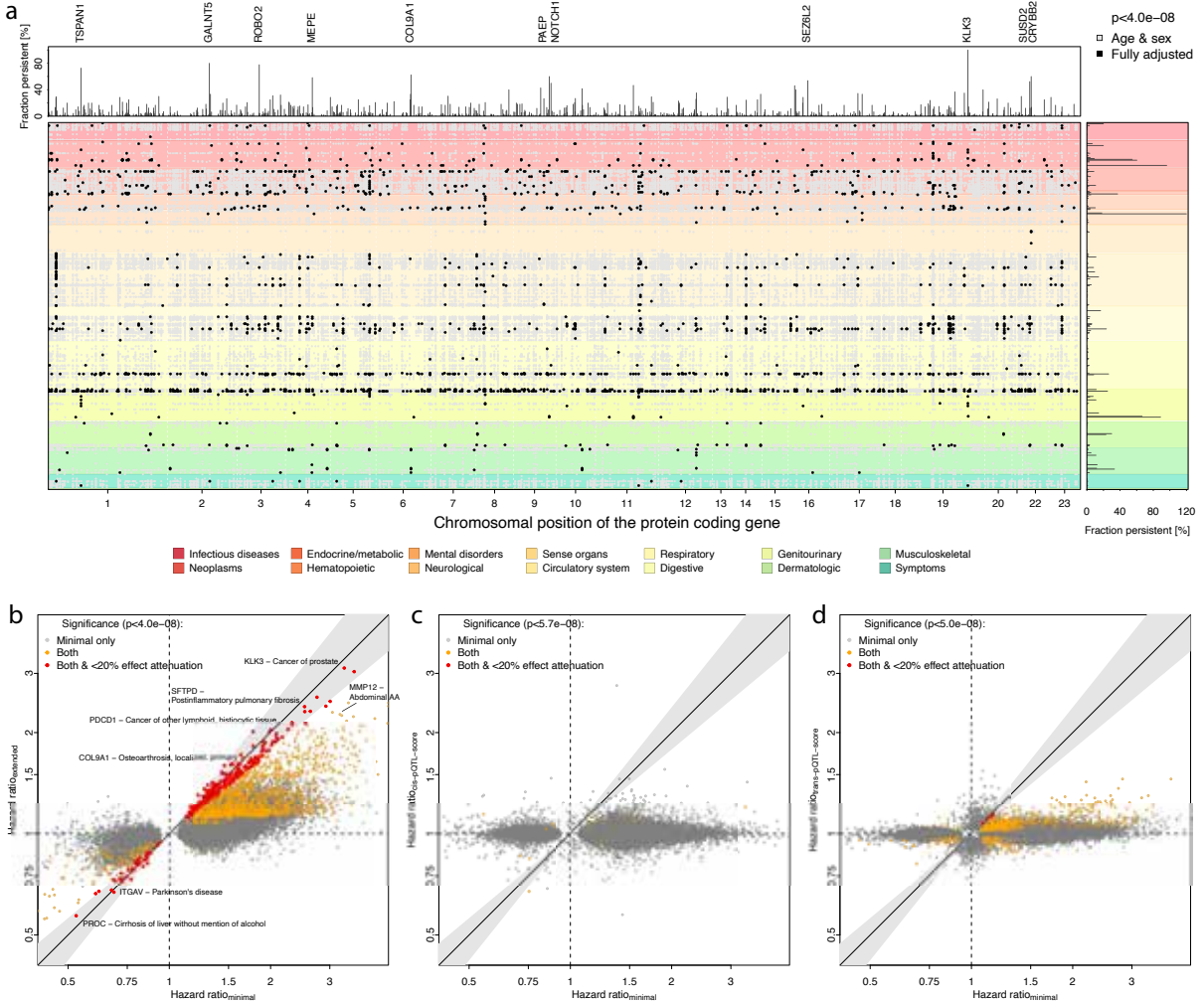
Figure EV7 Phenotype enrichment among differentially expressed plasma proteins between a random set of controls and different patient groups from the blood protein atlas section of the Human Protein Atlas.

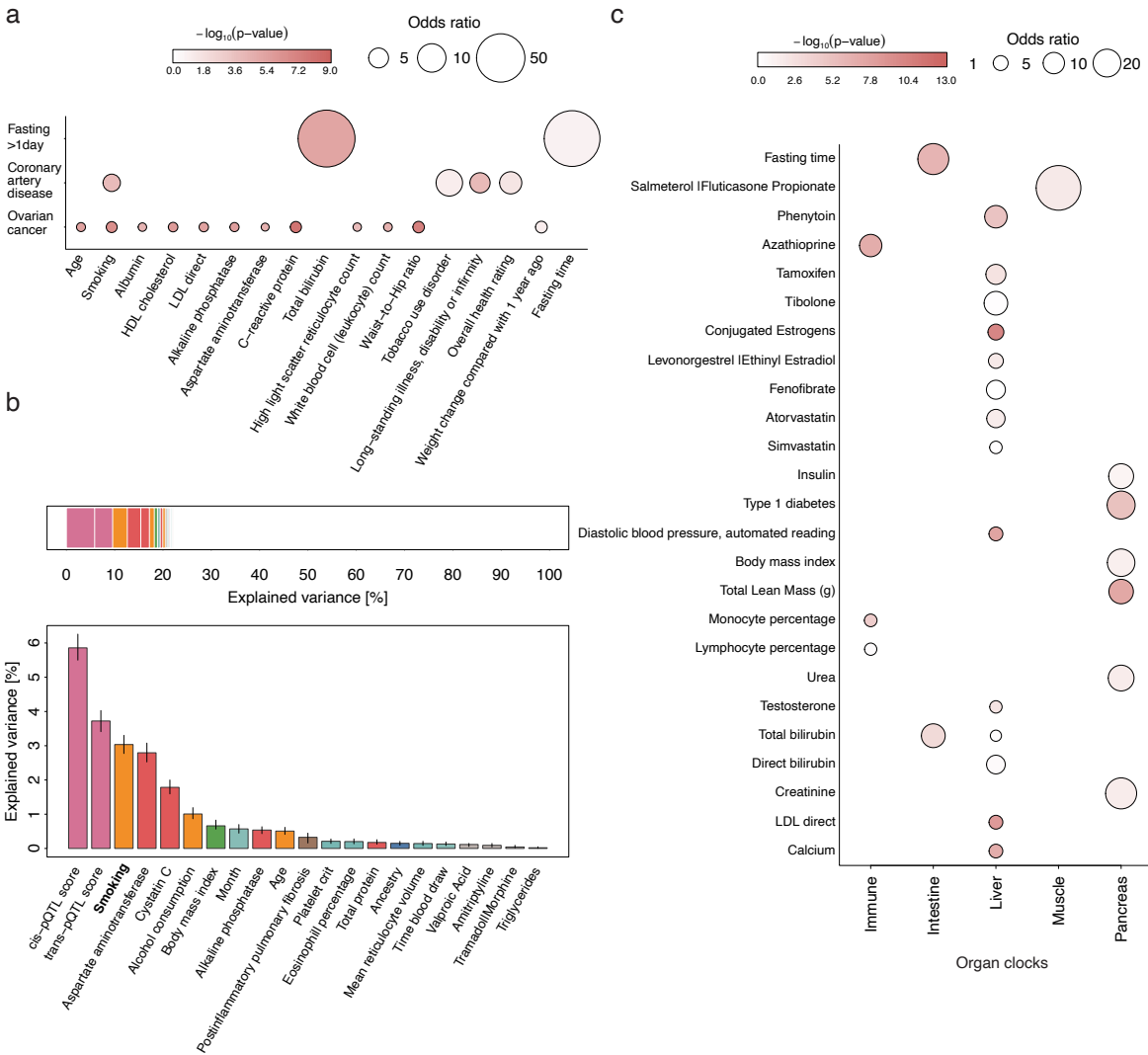
Each row refers to a protein signature significantly differential (corrected p-value < 0.05 and |log-fold change| > 0.5) in plasma of diseased patients. Each column refers to phenotype associated protein signature from the protein atlas that was significantly ($p<4.7x10^{-6}$) enriched among the respective disease – protein signatures. Significant findings are highlighted by dots, with colour representing transformed p-values and size reflecting odds ratios.


a
Secreted to blood
Intracellular and membrane
Secreted in other tissues
Secreted to extracellular matrix
Secreted to digestive system
Secreted in brain
Secreted – unknown location
Secreted in male reproductive system
Secreted in female reproductive system
Immunoglobulin genes
Not secreted
b
Diseases
Drugs
Biomarker
Blood cell counts
Diet
Socioeconomic
Pollution
Technical
Basic demographics
Body composition
Health
Pulmonary
Bone
Cardiovascular
Genetic
Ancestry
1
10
50
100
500
1000
Number of variables
d
Phenotypic diverse
Monogenic
Omnigenic
Polygenic
Inflammation
Kidney function
Liver/Macrophages
Platelet count
UMAP 2
UMAP 1
c
Ancestry
Basic demographics
Biomarker
Blood cell counts
Body composition
Bone
Cardiovascular
Diet
Diseases
Drugs
Genetic
Health
pollution
Pulmonary
Socioeconomic
Technical
Explained variance [%]
Ordered protein targets
e
Explained variance [%]
modifiable
non−modifiable
technical

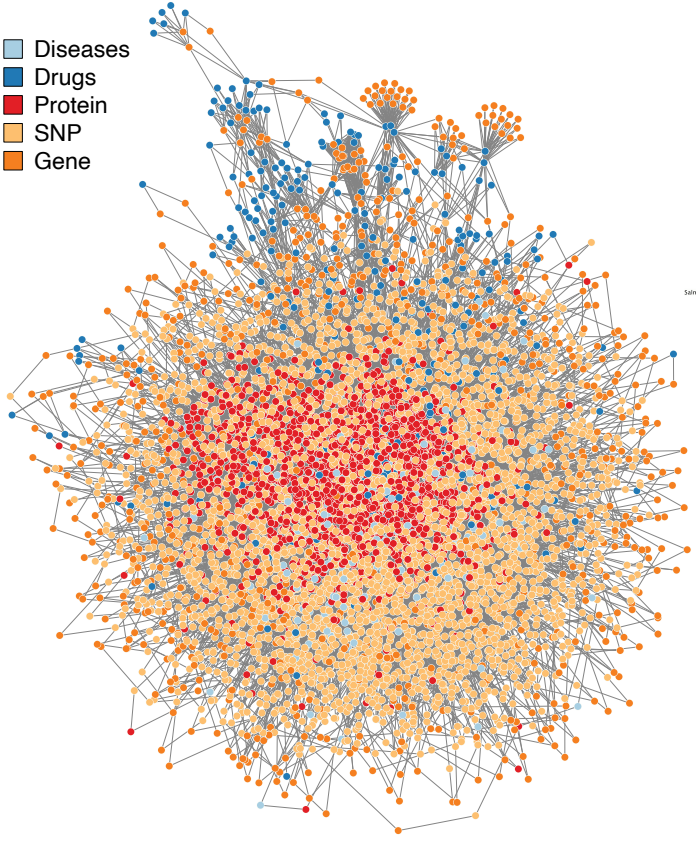

Ancestry
Basic demographics
Biomarker
Blood cell counts
Body composition
Bone
Cardiovascular
Diet
Diseases
Drugs
Genetic
Health
Pollution
Pulmonary
Socioeconomic
Technical
Salmeterol /Fluticasone Propionate
Norethisterone /Estradiol
Atorvastatin
Clopidogrel
Esomeprazole
Warfarin
Beclomethasone Dipropionate
Prednisolone
Metformin
Phenytoin
Carbamazepine
Tamoxifen
Azathioprine
Levonorgestrel /Ethinyl Estradiol
Tibolone
Conjugated Estrogens
Furosemide/Amiloride /Furosemide
Digoxin
Lansoprazole
Simvastatin
Fenofibrate
Acetaminophen
Psoriasis
Systemic lupus erythematosus
Chronic renal failure [CKD]
Renal failure NOS
Postinflammatory pulmonary fibrosis
Asthma
Atrial fibrillation and flutter
Type 1 diabetes
Lymphoid leukemia, chronic
Systolic blood pressure, automated reading
Pulse rate, automated reading
Diastolic blood pressure, automated reading
Body mass index
Platelet distribution width
Neutrophill percentage
Monocyte percentage
Mean corpuscular volume
Lymphocyte count
Immature reticulocyte fraction
High light scatter reticulocyte count
Urea
Urate
Total bilirubin
Testosterone
SHBG
LDL direct
IGF−1
Glycated haemoglobin (HbA1c)
Gamma glutamyltransferase
Direct bilirubin
Cystatin C
Creatinine
Calcium
C−reactive protein
Aspartate aminotransferase
Alkaline phosphatase
Albumin
Alanine aminotransferase
Centre
Age
Ancestry
Erythroid cells
Langerhans cells
Dendritic cells
Kupffer cells
Hofbauer cells
Macrophages
Monocytes
NK−cells
B−cells
T−cells
Peritubular cells
Fibroblasts
Adipocytes
Smooth muscle cells
Skeletal myocytes
Endothelial cells
Leydig cells
Enteroendocrine cells
Thymic epithelial cells
Sertoli cells
Proximal tubular cells
Hepatocytes
Alveolar cells type 2
Suprabasal keratinocytes
Proximal enterocytes
Gastric mucus−secreting cells
Number of draws
Fasting time
Weight change compared with 1 year ago
Long−standing illness, disability or infirmity
cis−pQTL score


a
Explained variance [%] – British African
Explained variance [%] – White European*
APOF
IL17RA
APOM
TNFRSF10B
MIA
PRG3
FBLN2
SELPLG
INSL3
FAM3B
PILRB
KLK3
XG
EDDM3B
MOG
F3
PNLIPRP2
GPR158
SKAP1
DDC
SAA4
BGLAP
PILRA
ACRV1
APOBR
CTHRC1
IL6R
CSF2RB
SEZ6L
IL18R1
MEP1B
ENPP7
CLU
PSCA
SGSH
COL24A1
MAN2B2
HDGF
LILRA3
NTSE
TXNDC15
LILRB2
GC
CLEC7A
MFGE8
HNMT
AFAP1
TLR3
TDGF1
FCGR2A
IL1RAP
BST1
TNFRSF10C
FCRLB
CHIT1
SLAMF8
IL34
KLK15
b
Phenotypic diverse
Monogenic
Omnigenic
Polygenic
Inflammation
Kidney function
Liver/Macrophages
Platelet count
Explained variance [%] – British Central South Asian
Explained variance [%] – White European*
ACP6
CD300E
RNASE6
CSTB
LILRA2
APOC1
TPA
XG
APCS
ADM
PZP
PROC
SPINT3
EDDM3B
SHISA5
CA14
ENPP2
SGSH
INSL3
ACRV1
BCL2
CALCOCO2
SPRR3
DNMBP
FGR
SLAMF8
IL1RAP
BST1
CLEC4D
SIGLEC5
LRRC37A2
TPSAB1
PSG1
KLK15
c
Fold change in explained variance cis–pQTL – EUR/AFR
Absolute difference in allele frequency [%] – EUR/AFR
CLU
APOF
GPR158
CD86
SAA4
FBLN2
F3
SELPLG
AOC1
d
Fold change in explained variance cis–pQTL – EUR/CSA
Absolute difference in allele frequency [%] – EUR/CSA
REG3G
KLRF1
PROCR
ENTPD5
CD164L2
ECI2
KLK15
SFTPA1
e
cis–pQTL score – LAIR1
cis–pQTL score – LPA
cis–pQTL score – COL24A1
cis–pQTL score – CA6
cis–pQTL score – ARSA
cis–pQTL score – ADGRD1
cis–pQTL score – PM20D1
cis–pQTL score – SAMD9L
cis–pQTL score – BPIFA2
cis–pQTL score – SOD3
cis–pQTL score – ALPP
cis–pQTL score – TNFRSF13C
cis–pQTL score – LTBP3
cis–pQTL score – CD70
cis–pQTL score – CD177
cis–pQTL score – KLK4
cis–pQTL score – DSCAM
cis–pQTL score – MGMT
cis–pQTL score – NFASC
cis–pQTL score – QPCT
cis–pQTL score – KLK1
cis–pQTL score – PSG1
cis–pQTL score – SERPINA12
cis–pQTL score – PER3
cis–pQTL score – OSMR
cis–pQTL score – TRIM5
cis–pQTL score – ANXA2
cis–pQTL score – LTBR
cis–pQTL score – TNR
cis–pQTL score – CREG1
cis–pQTL score – HMCN2
cis–pQTL score – ARHGEF5
cis–pQTL score – F10
cis–pQTL score – CA12
EUR
AFR
CSA
trans–ethnic
ethnic–specific
Explained variance [%]

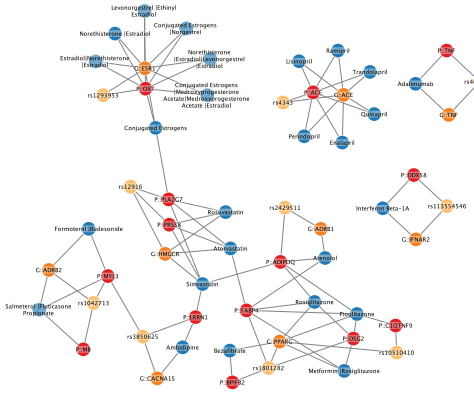

a
TSPAN1
GALNT5
ROBO2
MEPE
COL9A1
PAEP
NOTCH1
SEZ6L2
KLK3
SUSD2
CRYBB2
p<4.0e−08
Age & sex
Fully adjusted
Fraction persistent [%]
Chromosomal position of the protein coding gene
Infectious diseases
Neoplasms
Endocrine/metabolic
Hematopoietic
Mental disorders
Neurological
Sense organs
Circulatory system
Respiratory
Digestive
Genitourinary
Dermatologic
Musculoskeletal
Symptoms
b
Significance (p<4.0e−08):
Minimal only
Both
Both & <20% effect attenuation
KLK3 – Cancer of prostate
SFTPD – Postinflammatory pulmonary fibrosis
MMP12 – Abdominal AA
PDCD1 – Cancer of other lymphoid, histiocytic tissue
COL9A1 – Osteoarthrosis, localized
ITGAV – Parkinson's disease
PROC – Cirhosis of liver without mention of alcohol
Hazard ratio minimal
c
Significance (p<5.7e−08):
Minimal only
Both
Both & <20% effect attenuation
Hazard ratio minimal
d
Significance (p<5.0e−08):
Minimal only
Both
Both & <20% effect attenuation
Hazard ratio minimal


a
−log10(p−value)
0.0 1.8 3.6 5.4 7.2 9.0
Odds ratio
5 10 50
Fasting >1day
Coronary artery disease
Ovarian cancer
Age
Smoking
Albumin
HDL cholesterol
LDL direct
Alkaline phosphatase
Aspartate aminotransferase
C−reactive protein
Total bilirubin
High light scatter reticulocyte count
White blood cell (leukocyte) count
Waist−to−Hip ratio
Tobacco use disorder
Long−standing illness, disability or infirmity
Overall health rating
Weight change compared with 1 year ago
Fasting time
b
0 10 20 30 40 50 60 70 80 90 100
Explained variance [%]
Explained variance [%]
0 1 2 3 4 5 6
cis−pQTL score
trans−pQTL score
Smoking
Aspartate aminotransferase
Cystatin C
Alcohol consumption
Body mass index
Month
Alkaline phosphatase
Age
Postinflammatory pulmonary fibrosis
Platelet crit
Eosinophil percentage
Total protein
Ancestry
Mean reticulocyte volume
Time blood draw
Valproic Acid
Amitriptyline
Tramadol|Morphine
Triglycerides
c
−log10(p−value)
0.0 2.6 5.2 7.8 10.4 13.0
Odds ratio
1 5 10 20
Fasting time
Salmeterol |Fluticasone Propionate
Phenytoin
Azathioprine
Tamoxifen
Tibolone
Conjugated Estrogens
Levonorgestrel |Ethinyl Estradiol
Fenofibrate
Atorvastatin
Simvastatin
Insulin
Type 1 diabetes
Diastolic blood pressure, automated reading
Body mass index
Total Lean Mass (g)
Monocyte percentage
Lymphocyte percentage
Urea
Testosterone
Total bilirubin
Direct bilirubin
Creatinine
LDL direct
Calcium
Immune
Intestine
Liver
Muscle
Pancreas
Organ clocks

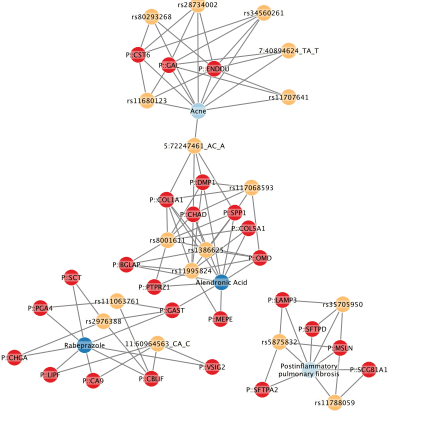
a SNP - Gene - Protein - Drug - Disease Knowledge Graph



b Drug target engagement



c Selected drug/disease profiles



Interactive version https://omicscience.org/apps/prot_foundation/