Large Bi-Ethnic Study of Plasma Proteome Leads to Comprehensive Mapping of *cis*-pQTL and Models for Proteome-wide Association Studies

Authors: Jingning Zhang[1], Diptavo Dutta[1], Anna Köttgen[2,3], Adrienne Tin[2,4], Pascal Schlosser[2,3], Morgan E. Grams[2,5], Benjamin Harvey[1], CKDGen Consortium, Bing Yu[6], Eric Boerwinkle[6,7], Josef Coresh[1,2,5], Nilanjan Chatterjee[1,8*]

Affiliations:

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
2. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
3. Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany
4. MIND Center and Division of Nephrology, University of Mississippi Medical Center, Jackson, MS, USA
5. Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, US
6. Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX, USA
7. Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
8. Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[*]Corresponding author (nilanjan@jhu.edu)

Abstract

Improved understanding of genetic regulation of proteome can facilitate the identification of causal mechanisms for complex traits. We analyzed data on 4,657 plasma proteins from 7,213 European American (EA) and 1,871 African American individuals from the ARIC study, and further replicated findings on 467 AA individuals from the AASK study. We identified 2,004 plasma proteins in EA and 1,618 in AA, with majority overlapping, which showed significant genetic associations with common variants in *cis*-regions. Availability of AA sample led to smaller credible sets and identification of a significant number of population-specific *cis*-pQTLs. Estimates of *cis*-heritability for proteins were similar across EA and AA (median *cis*-h^2=0.09 for EA and 0.10 for AA) and tended to be lower than those of gene expressions. Elastic-net-based algorithms produced high accuracy for protein prediction in each population, but models developed in AA were more transportable to EA than conversely. An illustrative application of proteome-wide association studies (PWAS) to serum urate and gout, implicated several proteins, including *IL1RN,* revealing the promise of the drug anakinra to treat acute gout flares. Our study demonstrates the value of large and diverse ancestry study for understanding genetic mechanisms of molecular phenotypes and their relationship with complex traits.

Introduction

Genome-wide association studies (GWAS) to date have cumulatively mapped tens of thousands of loci containing common genetic variants associated with complex traits [1, 2]. As the majority of the variants are in non-coding regions [3, 4], researchers have focused on understanding the role of gene-expression regulation as a mechanism for complex trait genetic association [5-9]. There is known to be substantial overlap between genetic variations regulating gene expression and those influencing complex traits [7-9], but only a small fraction of GWAS heritability of complex traits can be explained by mediating effects through bulk gene-expression [10, 11]. While it is likely that future studies with more extensive cell-type specific gene expression measurements will lead to additional insights, comprehensive understanding of causal mechanisms for complex traits will ultimately require the integration of data from various types of genomic and molecular traits [11]. Proteins, the ultimate product of the transcripts, are subject to post-translational modifications and processing, and contain additional information that cannot be detected at the level of the transcriptome.

Recently, major opportunities have arisen to substantially increase our understanding of the causal role of proteins in complex traits due to availability of an accurate high throughput technology for measuring proteins in different types of samples [12, 13]. The plasma proteome has received particular attention as it can capture a wide variety of proteins that are active in different biological processes, including but not limited to circulation [14]. The proteome is often dysregulated by diseases, and it is highly amenable for drug targeting [15, 16]. A number of genetic studies have identified protein quantitative trait loci (pQTL), for plasma [15-20] as well as some other tissues [21-23], and noted that pQTLs are enriched for GWAS associations across an array of complex

70 traits [15-23]. Studies have used pQTLs as instruments in conducting Mendelian randomization (MR)
71 analysis to identify causative proteins, and hence potential therapeutic targets, across diverse
72 phenotypes [24-26], including COVID-19 related outcomes [19, 20].
73
74 In spite of substantial progress, understanding of the genetic architecture of the proteome and
75 its overlap with those of gene expressions and complex traits remains limited. While the sample
76 size for some studies [16-19] of the plasma proteome has involved thousands of individuals, it is
77 likely that identification of pQTLs remains incomplete, both due to inadequate sample size or/and
78 lack of comprehensive protein measurements. Further, existing proteomic studies have been
79 mostly restricted to samples of European ancestry, and thus cannot inform potential
80 heterogeneity by ancestry. Additionally, advanced tools for incorporating pQTL information for
81 exploring causal effects of proteins, such as those available for analysis of gene-expression [27, 28],
82 are lacking.
83
84 In this article, we report results from a comprehensive set of analyses of *cis*-genetic regulation of
85 the plasma proteome in the large European and African American cohorts of the Atherosclerosis
86 Risk in Communities (ARIC) study [29]. We focus on the identification of *cis*- associations, which
87 compared to *trans*-, have been shown to more replicable across different proteomic platforms [30]
88 and are less likely to be affected by horizontal pleiotropy that could pose additional challenge for
89 downstream Mendelian-randomization analyses [31]. We carry out a set of association and fine-
90 mapping analyses to identify common (minor allele frequency (MAF) > 1%) *cis*-pQTLs and
91 compare results across ethnic groups to explore shared and unique genetic architecture. For each
92 ethnic group, we characterize *cis*-heritability of the proteome due to common variants and build

93 models for genetically predicting levels of plasma proteins. Using these models, we then conduct
94 a proteome-wide association studies (PWAS) of serum urate [32], an important biomarker of purine
95 metabolism with high heritability and large available large GWAS summary statistics, and the
96 complex disease gout, which can result from high urate levels [32]. We create several data
97 resources for using our results to inform future studies (http://nilanjanchatterjeelab.org/pwas).
98
99 **Results**
100 *Identification of cis-pQTLs Across European and African American Populations*
101 We performed separate *cis*-pQTL analyses for the African American (AA) and European American
102 (EA) populations in the ARIC study, with total sample sizes of N=1,871 and N=7,213, respectively
103 (see **Methods**). We performed analyses based on plasma samples collected during the third visit
104 of the cohort [29] (see **Supplementary Table 1** for sample characteristics). Relative concentrations
105 of plasma proteins or protein complexes were measured by modified aptamers ('SOMAmer
106 reagents', hereafter referred to as SOMAmers) [12, 13].
107
108 We defined *cis*-regions to be +/- 500Kb of the transcription start site (TSS) in the *cis*-pQTL analysis.
109 After quality control (see **Methods**), we analyzed 4,657 SOMAmers, which tagged proteins or
110 protein complexes encoded by 4,435 genes, and 204 of them were tagged by more than one
111 SOMAmer. In the *cis*-regions, we analyzed 10,961,088 common (MAF>1%) single-nucleotide
112 polymorphisms (SNPs) for AA and 6,181,856 for EA with imputed or genotyped data after quality
113 filtering (see **Methods**). For identification of *cis*-pQTLs, we performed regression analyses of
114 protein levels after residualizing by a number of potential confounders, including sex, age, 10
115 genetic principal components (PCs) and the study sites at v3. In addition, similar to eQTL analyses

[8], we adjusted for Probabilistic Estimation of Expression Residuals (PEER) factors [33, 34] to account for hidden confounders that may influence clusters of proteins. We observed that the inclusion of PEER factors substantially improved power for *cis*-pQTL studies due to reduced residual variance (**Fig. 1a**, **Supplementary Table 2**). In all subsequent analyses, protein levels measured by SOMAmers were residualized with respect to these sets of PEER factors and then normalized by quantile-quantile transformation (see **Methods**).

In the ARIC study, we identified a total of 2,004 and 1,618 significant SOMAmers, i.e. SOMAmer with a significant (at false discovery rate (FDR)<5%)[9] *cis*-pQTLs near the putative protein's gene, in the EA and AA populations, respectively, with 1,447 of these overlapping across the populations (**Fig. 1b**, **Supplementary Tables 3.1 and 3.2).** Compared to plasma pQTL studies conducted in the past in European ancestry sample[16, 17], we almost tripled the number of significant SOMAmers with known *cis*-pQTLs [17, 18] (1,465 v.s. 508 using the same Bonferroni corrected genome-wide threshold for significance) (**Supplementary Table 3.1**) and we successfully replicated 99% (504/508) of previously identified *cis*-pQLTs (**Supplementary Table 4**).

We found 10% of the sentinel *cis*-pQTLs identified in EA were non-existent or rare, defined as two or less individuals carrying the variant, in the 1000Genome AA sample. In contrast, nearly one third of the variants identified in the AA population were non-existent or rare in the 1000Genome EA population, signifying the value of diverse ancestry data to identify population-specific *cis*-pQTLs (**Supplementary Tables 3.1 and 3.2**). *cis*-pQTLs which were identified through either of the two populations, but were common in (MAF>1%) in both, the effect-sizes showed

high degree of concordance across the populations with a correlation coefficient above 0.9 (**Supplementary Fig. 1**). We further carried out a replication study using data available on additional 467 individuals from the African American Study of Kidney Disease and Hypertension (AASK) [35] , which also ascertained proteins using the SOMAScan platform (see **Methods** and **Supplementary Table 1**). Among 1,398 sentinel *cis*-SNPs which were identified through the ARIC AA sample and which were genotyped or imputed in AASK, we found 93% showed effects in the same direction and 69% showed statistical significance at FDR<5% in the replication analysis (**Supplementary Tables 5.1 and 5.2)**.

Genotypic effect sizes for *cis*-pQTLs were inversely associated with minor allele frequencies even after accounting for bias due to power for detection [36](**Fig. 1c**). The *cis*-pQTLs appeared to be more concentrated near the TSS of corresponding pGenes in the AA than the EA population, and the genotypic effect sizes for *cis*-pQTLs decreased with distance from the TSS in both populations (**Fig. 1d**). Using stepwise regression [37, 38], we identified multiple conditional independent *cis*-SNPs for 1,398 (70%) and 1,021 (63%) of the significant SOMAmers in EA and AA populations, respectively (**Fig. 1e, Supplementary Tables 6.1 and 6.2**).

Protein altering variants (PAV) may result in apparent *cis*-pQTLs owing to altered epitope binding effects [16]. Following a procedure recommend earlier [16], we found that in the EA population while up to 65% (1,299 out of 2,004) of the sentinel pQTLs could be affected by LD with known PAVs, in the AA population the corresponding proportion drops to 47% (765 out of 1,618) (see **Supplementary Tables 3.1, 3.2 and 7**). However, large overlap observed between eQTL and pQTLs in colocalization analysis (see below) indicates they are driven by underlying causal

variants and reduces concerns for any large-scale effect of epitope artifacts in the detection of pQTLs.

Cis-eQTL Overlap and Functional Enrichment

To evaluate the extent to which the *cis*-pQTL variants were also involved in modulating transcriptional levels, we cross referenced the *cis*-pQTLs with significant *cis*-eQTLs (at FDR<5%) from the Genotype-Tissue Expression project (GTEx V8) [9] across 49 different tissues (See **Methods**). Since the GTEx cohort is primarily of European ancestry (85.3% EA), we restricted the analysis to the top *cis*-pQTLs identified in the EA cohort only. We found that, approximately 73.9% of the sentinel *cis*-pQTLs or variants in high LD ($r^2 > 0.8$) with them, were also significant *cis*-eQTLs for the same gene in at least one tissue (**Supplementary Fig. 2a**). Further, pairwise colocalization indicated that for 49.4% of the pGenes, *cis*-pQTLs colocalize with *cis*-eQTLs, in at least one of the GTEx tissues with high posterior probability (PP.H4 $\geqslant$ 80%) (**Supplementary Fig. 2b, Supplementary Tables 8.1 and 8.2**). Further, *cis*-pQTLs tended to be reported as significant *cis*-eQTLs across multiple tissues possibly because plasma protein level might contain signatures from multiple tissues (**Supplementary Fig. 3**).

Results from the association analysis of molecular phenotypes, like plasma proteins, integrated with the functional and regulatory annotation of the genome offer a powerful way to understand the molecular mechanisms and consequences of genetic regulatory effects. The functional annotations were curated from several sources like variant effect predictor (VEP) [39], Loss-Of-Function Transcript Effect Estimator (LOFTEE) [40] and Ensembl Regulatory Build [41]. We found that *cis*-pQTLs were enriched for several protein altering functions which may be caused by epitope

binding effects noted earlier (**Supplementary Fig. 4a-b**). After adjusting for protein altering variants (PAVs), we found that independent top *cis*-pQTLs were enriched in a large spectrum of functional annotations including untranslated regions (5' and 3'), promoters and transcription factor binding sites, with a pattern that was consistent across the EA and AA populations (See **Methods**, **Supplementary Fig. 4c-d** and **Supplementary Table 9**).

Fine Mapping

To identify the causal variants underlying the significant *cis*-pQTLs for plasma proteins, we first conducted population-specific fine-mapping for the 1,447 significant SOMAmers that had at least one *cis*-pQTL both in EA and AA using SuSiE [42] (**Supplementary Tables 10.1 and 10.2**). Comparing the 95% credible sets, we found that the average number of variants in the credible sets were significantly smaller in AA compared to that in EA (21.29 in EA vs 12.11 in AA; p-value = 8.43×10^{-27}; **Fig. 2 a-b**). This is possibly driven in part by the lower average LD in AA, but also could be due to the lower sample sizes in AA compared to EA, resulting in lower statistical power. To demonstrate the added value of including two ethnic populations in identifying possibly shared causal variants, we further conducted a trans-ethnic meta-analysis using MANTRA [43], which accounts for effect heterogeneity among populations and constructed a 95% credible set of shared causal variants by ranking variants according to their Bayes factor (See **Methods**).

As an example of the fine mapping analysis, we illustrate the fine-mapped *cis*-region (+/-500Kb) for the *HBZ* pGene on chromosome 16p13.3 corresponding to the Hemoglobin subunit zeta protein (HBAZ; Uniprot ID: P02008), which is involved in oxygen transport and metal-binding mechanisms [44, 45] and has been associated with thalassemia [46]. Single variant pQTL association

results show that there are several significant *cis*-pQTL associations both in EA and AA populations (**Fig. 2c and 2e**). Fine-mapping within the EA individuals identifies a 95% credible set of seven variants (**Fig. 2d**) while that within the AA individuals identifies a smaller credible set of two variants only (**Fig. 2f**). Trans-ethnic meta-analysis using MANTRA further points to a single variant rs2541645 (16:161106 G>T) as the possible shared causal variant between EA and AA. This variant was in fact the most significantly associated *cis*-pQTL for *HBZ* pGene in AA but not in EA, and had some evidence of differences in minor allele frequency across the populations (MAF = 0.32 in EA vs 0.18 in AA). This SNP is a strong eQTL for *HBZ* expression in GTEx V8 whole blood (p-value = 6.7×10^{-80}), and associated with several erythrocyte related outcomes in the UK Biobank including mean corpuscular hemoglobin (p-value=1.1×10^{-14}) and reticulocyte fraction of red cells (p-value=3.2×10^{-9}) [47, 48]. Together, these findings suggest that rs2541645 might be a regulatory variant for *HBZ* protein levels and possibly warrant further study on downstream phenotypic consequences especially in the context of blood related mechanisms and thalassemia.

Analysis of Cis-Heritability of Proteins and Building Protein Imputation Model

We estimated *cis*-heritability (*cis*-h^2) of plasma proteins, i.e. the proportion of variance of protein levels that could be explained by all SNPs in *cis*-regions of their encoded genes, using the GCTA software [49]. We found 1,350 and 1,394 SOMAmers to have significant *cis*-h^2 (p-value < 0.01) for the EA and AA populations, respectively, and 1,109 of them overlapped (**Supplementary Table 11**). The majority of those significant *cis*-heritable SOMAmers also had *cis*-pQTLs identified in our study (96% for AA and 99% for EA, **Supplementary Table 12**). The *cis*-h^2 for significant SOMAmers (median *cis*-h^2 = 0.10 for AA, and 0.09 for EA) tended to be substantially smaller than those reported for gene-expression [50] in two related tissues [14] in liver and whole blood in GTEx V7 (**Fig.**

3a) and similar patterns were also observed when in GTEx V8 (**Supplementary Fig. 5**). The pattern is expected given the closer relationship of genetic variation to transcripts than to the encoded proteins, which are subject to additional processing including post-translational modifications.

Next, we built protein imputation models for *cis*-heritable SOMAmers using an elastic net machine learning method with 5-fold cross-validation as has been used for modeling gene-expression [27]. The median accuracy for the elastic-net models for protein predictions, evaluated as the prediction R^2 standardized by *cis*-heritability (R^2/*cis*-h^2), was 0.79 and 0.69 for the EA and AA populations, respectively. Compared with imputation models built only with the top *cis*-pQTL, the elastic net models gained, 36% and 40% of accuracy for the EA and AA populations, respectively (**Fig. 3b**, **Supplementary Table 13**). In cross-ethnic analysis, we found that models trained in the EA population performed worse in the AA population than the converse, in spite of a much smaller sample size in AA, again indicating the advantage of the latter population to identify causal pQTLs which are more likely to have robust effects across ethnic groups (**Fig. 3c**).

Cis-regulated Genetic Correlation between Plasma Proteome and Transcriptome across a variety of tissues

We then explored *cis*-regulated genetic correlation between plasma proteins and expression levels for the underlying genes across a variety of tissues. We used genotype data for Europeans from the Phase-3 1000 Genome Project (1000Genome) [51] to evaluate Pearson's correlation coefficients between genotypically-imputed protein levels, and genotypically-imputed expression levels, with the latter being computed based on models that have been previously built and published by Gusev *et al.* [28] based on data from the GTEx (V7) consortium

(**Supplementary Tables 13 and 14**). We also used models based on GTEx (V8) developed by the same group (available through personal communication), but because of their preliminarily nature we present the all analyses involving imputed gene-expression using the V7 models and present preliminary results from the V8 models as supplementary data. The analysis was restricted to the European population due to the lack of gene-expression imputation models for AA population. Overall, genetically imputed plasma proteins are only moderately correlated with those for gene expression levels (**Fig. 3d**). Consistent with previous studies [52], we find that plasma proteins show strongest genetic correlations with genes expression levels in the liver, the organ responsible for the synthesis of many highly abundant plasma proteins. The lowest genetic correlations were seen for brain-related tissues, which may be due to the blood-brain barrier. In GTEx (V8), that included a larger number of overlapping genes with our SOMAmers, we observed a similar pattern for high-/low-rank tissues for expression-protein genetic correlations although the absolute levels of these correlations were bit lower (**Supplementary Table 15.1**). The correlations between direct plasma protein measurements and imputed gene expression levels in ARIC showed similar trend but have generally much lower values as they account for additional variability of protein measurements due to non-genetic factors (**Supplementary Fig. 6**).

Proteome-wide Association Study (PWAS) of Complex Traits

We illustrate an application of the protein imputation model by conducting proteome-wide association studies for two related complex traits: (1) serum urate, a highly heritable biomarker of health representing the end product of purine metabolism in humans, and (2) gout, a complex disease caused by urate crystal deposition in the setting of elevated urate levels and the resulting inflammatory response. We obtained GWAS summary-statistics data for these traits generated

by the CKDGen Consortium [32] involving a total sample size of N=288,649 and N=754,056, respectively. As this GWAS was conducted primarily in EA population, we carried out the PWAS analysis using the model generated for the EA population.

We used a computational pipeline previously developed for conducting TWAS based on GWAS summary-statistics [28, 53] to carry out an analogous PWAS analysis. Simulation studies showed that type 1 error of PWAS analysis based on our protein imputation weights are well controlled (**Supplementary Fig. 7**). Among SOMEmers that showed significant *cis*-heritability, we identified 10 and 4 distinct loci containing genes for which the encoded proteins or protein complexes were found to be significantly (p-value < 3.7×10^{-5}) associated with serum urate and gout, respectively. We further examined whether the PWAS signals could be explained by *cis*-genetic regulation of the expression of nearby (1Mb region around) genes and vice versa by performing bivariate analysis conditioning on imputed expression values for nearby genes that are found to be significantly associated based on the TWAS analysis. Main results were based on GTEx V7 models (**Fig. 4**, **Table 1**, **Supplementary Fig. 8**), and further validated using GTEx V8 preliminary models (**Supplementary Table 16**). For the nearby TWAS analysis, we considered significance of genes based on two trait-relevant tissues available in GTEx V7, namely whole blood and liver, but also explored other tissues more broadly (see **Methods**).

The conditional PWAS analysis of serum urate revealed several interesting patterns (**Table 1a**). First, there were PWAS signals that could be largely explained by nearby TWAS signals for the corresponding transcript in relevant tissues (e.g., *INHBB* in liver, and *SNUPN* in whole blood). This may be indicative of genetic loci influencing serum urate through altered gene expression and

corresponding protein levels [54]. Second, there were also PWAS signals that could be largely explained by the TWAS signal of the corresponding transcript in other tissues (e.g. *B3GAT3* in brain), but not in whole blood or liver. Such examples support the notion that the evaluation of diverse potential tissues of action may be important to characterize these genetic loci. However, the effects of TWAS of *B3GAT3* in brain tissues are negative whereas the effect of its PWAS is positive. We found the opposite direction is consistent with their negative genetic correlation between plasma protein and gene-expression in those tissues. Future investigation for the complicated directions of effects is worthwhile. Third, for the locus around *INHBC*, the plasma PWAS signal for *INHBC* explains the most significant nearby TWAS signal *R3HDM2* in thyroid (conditional p-value of TWAS signal = 4.12×10^{-1}) but not *vice* versa (conditional p-value of PWAS signal = 6.84×10^{-34}). We also found the top TWAS gene-tissue signals detected using the V7 models show similar level of significance, direction and magnitude of association when the analyses were repeated using the V8 models and the corresponding conditional PWAS-TWAS conditional analysis show qualitatively similar results (**Supplementary Table 16**). For the significant PWAS signals, we further examined evidence of colocalization of with gene expressions across tissues (**Supplementary Tables 17.1 and 17.2**) and observed that whenever there was strong genetic correlation between plasma protein and gene expression there was also strong evidence of colocalization (e.g. *INHBB* in liver, and *B3GAT3* in brain), which gives the most confidence in those findings.

Finally, the PWAS of gout revealed a finding illustrating the potential to detect potential drug targets for treating gout based on the significant association with the Interleukin 1 Receptor Antagonist protein (IL1RN, p-value = 2.22×10^{-5}) (**Table 1b**). IL1RN binds to its target, the cell

surface interleukin-1 receptor (IL1R1), thereby inhibiting the pro-inflammatory effect of interleukin-1 signaling. Anakinra, an anti-inflammatory drug approved to treat rheumatoid arthritis, is a recombinant, slightly modified version of the IL1RN protein examined in our study that binds to IL1R1, blocking its actions (**Supplementary Fig. 9**). The observed association between higher levels of IL1RN protein and lower odds of gout are consistent with the beneficial effect of its synthetic analogue anakinra on other inflammatory diseases and suggest a repurposing opportunity for anakinra to treat acute gout flares. In fact, such evaluations are ongoing, with a recent randomized, double-blind, placebo-controlled trial of acute gout flares showing anakinra to be non-inferior to usual treatment [55]. While drug delivery to plasma proteins and their cell surface receptors is easier than to other molecules such as intra-nuclear proteins, druggabilty of any implicated protein in our study depends on various factors such as protein structure and biological functions, and needs to be evaluated on a case-by-case basis. A systematic connection of all *cis*-heritable proteins to active drug candidates is provided as an additional resource (**Supplementary Table 18**).

Discussion

We present a comprehensive analysis of *cis*-genetic regulation of the plasma proteome based on a large discovery study that include both EA and AA individuals and an additional replication study based on AA individuals. Our study almost tripled the number of genes with identified *cis*-pQTL compared to previous reports [16, 17] and led to, for the first time, understanding of unique genetic architecture of plasma proteome in the AA population. We developed models for plasma protein imputation separately for EA and AA populations and make them publicly available to facilitate

346 future proteome wide association studies. Using large-scale GWAS summary-statistics from two
347 complex traits, we illustrate how PWAS can complement TWAS for the identification of causal
348 genes, protein products and inform potential drug targets. We have created a web resource for
349 downloading summary-statistics data and PWAS models with searchable options for
350 exploring/viewing various results from our analyses (http://nilanjanchatterjeelab.org/pwas).
351
352 Our analysis provides several important insights into the *cis*-genetic architecture of plasma
353 proteome. We observe that *cis*-heritability of protein levels tends to be smaller compared to
354 those of gene expression levels in related tissues (**Fig. 3a**), a pattern consistent with the central
355 dogma of DNA regulating the proteome through the transcriptome and the widespread presence
356 of post-translational modification. Further, while *cis*-heritability of plasma proteome is fairly
357 comparable across EA and AA populations, we observe important heterogeneity. We found
358 nearly 30% of the sentinel pQTLs detected in the AA population were non-existent or extremely
359 rare in the EA population, but the converse proportion was much more modest (~10%). We
360 further observe that the predictive performance of protein imputation model for the AA
361 population, in spite of its much smaller sample size, is comparable to that for the EA population
362 (**Fig. 3b**), and cross-population performance of such model is better from AA to EA population
363 than the converse (**Fig. 3c**). Further, fine-mapping analysis using SuSiE indicated that the size of
364 "credible set" for many genes is substantially smaller in the AA than the EA population. Taken all
365 together, our analysis demonstrates that similar to what has been reported earlier for more
366 complex traits [56], there are distinct advantages of including ethnically diverse samples in genetic
367 studies of molecular phenotypes.
368

While we increased the number of known *cis*-pQTLs by large margin, some of the patterns of associations we see have been noted earlier. For example, similar to ours, a prior study [16] has reported large overlap between eQTL and pQTLs. Further, the distributions of *cis*-pQTLs we observe in relationship to distance from gene transcription site and across various functional annotations have also been noted earlier. A study [25] has previously shown that pQTLs identified in the EA population largely replicates in non-EA Arabic and Asian population. Similarly, we found high degree of correlations in effect sizes for *cis*-pQTLs which are common across both EA and AA populations. However, we also showed that discovery analysis in the AA population itself leads to the identification of many unique *cis*-pQTLs and further fine-mapping analysis in this population leads to better resolution for the identification of causal variants.

We demonstrate applications of protein imputation models for conducting proteome-wide association studies (PWAS) for two related complex traits, resulting in the exemplary identification of the *IL1RN* protein which indicates potential promise for drug repurposing of anakinra to treat acute gout flares. Through multivariate analysis, we further explored relationship between plasma PWAS signals and those detected at the transcriptome level through complementary TWAS approach across various tissues. We found that while TWAS signals often exist in the same region, the underlying genes for which the strongest signals are seen can differ or/and the underlying tissue may not be closely related to plasma. As plasma proteins are easier target for drug delivery, we created an additional resource connecting all *cis*-heritable proteins to active drug candidates (**Supplementary Table 18**). In general, we believe the most promising target genes could be where there exists both PWAS and TWAS signals with underlying evidence of genetic correlation and colocalization.

Our study has several limitations. First, while the platform we used included SOMAmers for close to 5,000 proteins or protein complexes, it does not provide coverage for the entire plasma proteome. In the future, more comprehensive protein measurements across different tissues will be needed to further pinpoint target genes and tissues of actions. Second, the power of our PWAS analysis conditional on TWAS signals may be affected by small sample size of underlying eQTL datasets. Third, in this study, we have not carried out a joint analysis of the data across the two population and thus may have incurred some loss of power for the identification of shared pQTLs. Fourth, we have not explored effects of uncommon and rare variants, as well as complex trans-associations, all of which could have significant impact in explaining heritability, but substantial discovery is likely to need even larger sample size.

In conclusion, our study provides comprehensive and cross-population insight into *cis*-genetic architecture of plasma proteome. We generate several resources for utilizing our results for the mapping of causal protein-regulating variants, investigating the causal role of plasma proteins on complex traits and their drug repurposing potential. Future studies are merited to obtain more comprehensive coverage of proteome across different tissues and to comprehensively explore the role of rare variants and trans-effects on the variation of the proteome.

Author Contribution

J.Z, J.C. and N.C conceived the project. J.Z. and D.D. carried out all data analyses with supervision from N.C. B.H. developed online resources for data visualization and sharing, J.Z., D.D., A.K. and N.C. drafted the manuscript, and A.T., P.S., M.G. and B.Y. provided comments. All co-authors reviewed and approved the final version of the manuscript.

Competing interests

Proteomic assays in ARIC were conducted free of charge as part of a data exchange agreement with Soma Logic.

Acknowledgements

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, HHSN268201700004I). The authors thank the staff and participants of the ARIC study for their important contributions. SomaLogic Inc. conducted the SomaScan assays in exchange for use of ARIC data. The UK BioBank data was obtained under the UK BioBank resource application 17712. This work was supported in part by NIH/NHLBI grant R01 HL134320, NIH/NIDDK grant R01 DK124399, and NIH/NIDDK grant R01 DK108803. Research of J.Z., D.D. and N.C. was supported R01 grant from the National Human Genome Research Institute [1 R01 HG010480-01]. The work of A.K. was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 431984000 – SFB 1453. The work of P.S. was funded by the EQUIP Program for

433 Medical Scientists, Faculty of Medicine, University of Freiburg. We acknowledge Dr. Nicholas
434 Mancuso and Dr. Alexander Gusev for providing preliminary TWAS models built with GTEx V8
435 data.

436

437 **URLs**

438 Public data used in this study: 1000 Genomes Phase 3,
439 http://www.internationalgenome.org/category/phase-3/; UK Biobank,
440 https://www.ukbiobank.ac.uk/; GTEx: https://gtexportal.org/home/datasets; CKDGen,
441 http://ckdgen.imbi.uni-freiburg.de/; biomaRt,
442 https://www.bioconductor.org/packages/release/bioc/vignettes/biomaRt/inst/doc/biomaRt.ht
443 ml; GWAS Atlas: https://atlas.ctglab.nl/PheWAS.
444 Publicly available software used in this study: R statistical software, http://www.R-project.org/;
445 PLINK 2.0, https://www.cog-genomics.org/plink/2.0/; GCTA,
446 https://cnsgenomics.com/software/gcta/; QTLtools: https://qtltools.github.io/qtltools/; VEP,
447 https://useast.ensembl.org/info/docs/tools/vep/; LOFTEE, https://github.com/konradjk/loftee;
448 SuSIE, https://github.com/stephenslab/susieR; FUSION/TWAS,
449 http://gusevlab.org/projects/fusion/; TORUS: https://github.com/xqwen/torus/; coloc:
450 https://github.com/chr1swallace; MANTRA: Available on request from author Professor
451 Andrew P. Morris.

452

453 **Data availability**

Genome-wide summary statistics for all single-SNP *cis-pQTL analysis, irrespective of significance level,* are made available at http://nilanjanchatterjeelab.org/pwas. Additional data and codes required to perform PWAS are also made available from the website.

Plasma protein data availability: Pre-existing data access policies for each of the parent cohort studies (ARIC and AASK) specify that research data requests can be submitted to each steering committee; these will be promptly reviewed for confidentiality or intellectual property restrictions and will not unreasonably be refused. Please refer to the data sharing policies of these studies. Individual level patient or protein data may further be restricted by consent, confidentiality or privacy laws/considerations. These policies apply to both clinical and proteomic data.

Code availability

Example codes to perform PWAS are available at http://nilanjanchatterjeelab.org/pwas. All final codes in this study for data analysis of the protein data, including pQTL and PWAS analysis, will be posted through GitHub upon manuscript publication at https://github.com/nchatterjeelab/PlasmaProtein.

Methods

Study population. Our study was conducted using individual-level data from the Atherosclerosis Risk in Communities (ARIC) study [29]. The ARIC study is an ongoing community-based cohort study of individuals that initially enrolled 15,792 participants 1987 and 1989 from four communities across the US: Washington County, Maryland; suburbs of Minneapolis, Minnesota; Forsyth County, North Carolina; and Jackson, Mississippi. The third visit (v3) occurred in 1993-1995, when blood samples used for the measurement of the proteome were collected. A total of 9,084 participants with cleaned plasma protein data (1,871 African Americans (AA), 7,213 European Americans (EA)) after the exclusions of participants without genotype data (see below) were retained in the current study.

Plasma protein data and genetic data. The relative concentrations of plasma proteins or protein complexes from the blood samples were measured by SomaLogic Inc. (Boulder, Colorado, US) using an aptamer (SOMAmer)-based approach [12, 13]. Details for this approach and the SomaLogic normalization pipeline can be found in a technical white paper on the manufacturer's website, http://somalogic.com/wp-content/uploads/2017/06/SSM-002-Technical-White-Paper_010916_LSM1.pdf, and https://somalogic.com/wp-content/uploads/2017/06/SSM-071-Rev-0-Technical-Note-SOMAscan-Data-Standardization.pdf. Of the 4,877 SOMAmers measuring 4,697 unique proteins or protein complexes, we excluded 43 SOMAmers that mapped to multiple gene targets, 9 SOMAmers whose target proteins' encoding genes do not have position record in the biomaRt database [57], and 8 SOMAmers without any SNPs in *cis* region. By restricting analysis to plasma proteins or protein complexes encoded by autosomal genes, we further excluded 158

493 genes on the X chromosome, and 2 genes on the Y chromosome. In total, 4,657 SOMAmers
494 measuring 4,483 unique proteins or protein complexes encoded by 4,435 autosomal genes
495 passed quality control, and were retained in the current study.

496 Genotyping of ARIC samples was performed on the Affymetrix 6.0 DNA microarray and
497 imputed to the TOPMed reference panel (Freeze 5b) [58, 59]. The SNPs with imputation quality R^2 <
498 0.8, call rates <90%, Hardy-Weinberg equilibrium p-values < 10^{-6}, or minor allele frequencies <1%
499 were excluded. Genetic principal components show that the two self-reported ethnic subgroups,
500 European Americans (EA) and African Americans (AA) are well distinguished in terms of genetic
501 ancestry (**Supplementary Fig. 10**) [60].

502

503 **Plasma protein data processing.** Additional variation in high-throughput gene expression data
504 which is not due to genetic variants has been found to impact the power of eQTL discoveries [8, 9].
505 The fluctuations of internal environment, experimental deviations, and batch effects can all have
506 large influence on high throughput measurements [33]. To study whether this type of variance
507 exists in our high-throughput plasma protein data measured by the SOMAmers, we performed
508 analysis of variance (ANOVA) test for non-genetic factors to the first 10 principal components
509 (PCs) of log-transformed relative abundance of SOMAmers. Non-genetic factors include common
510 covariates (age, sex, and study sites at v3), as well as batch effects (plate run date, scanner ID,
511 plate position, and subarray). (**Supplementary Table 19**).

512 To account for those non-genetic variances, which may obscure genetic association signals, we
513 used the Probabilistic Estimation of Expression Residuals (PEER) method to estimate a set of
514 latent covariates, and put them linearly in the model [34]. The number of PEER factors for each

ancestry was selected to maximize the number of significant SOMAmers, i.e. SOMAmers with a significant *cis*-pQTL near the putative protein's gene.

The log-transformed relative abundance of SOMAmers were adjusted in a linear regression model including PEER factors and the covariates sex, age, study site, and 10 genetic principal components (PCs). The residuals from this linear regression were then rank-inverse normalized to avoid the influence of extreme values, and were used as the corrected-protein quantification in the analysis. By analyzing up to 200 PEER factors in increments of 10, the maximum of number of significant SOMAmers were achieved at 90 and 80 PEER factors for EA and AA respectively (**Fig. 1a**). Thus, the corrected-protein quantifications adjusted for 90 and 80 PEER factors were used as phenotypes in the analysis of the EA and AA populations, respectively.

Significant SOMAmers discovery. Significant SOMAmer is defined as SOMAmer with a significant *cis*-pQTL near the putative protein's gene. For all primary analyses, we defined the mapping window as 500-kb upstream and downstream of the target protein-coding genes' transcription start site (TSS). In a secondary analysis, we found that *cis*-heritability of SNPs within +/- 500Kb and +/- 1Mb of the TSS to be quite similar, indicating that vast majority of *cis*-pQTLs for the larger region to be concentrated within +/- 500Kb window (**Supplementary Table 20**). Gene position of GRCh38 reference genome was obtained from Ensembl BioMart database [57]. Common linear regression procedures for association tests using the Bonferroni correction to p-values usually proves to be overly stringent and results in many false negatives [38]. To overcome this issue, adaptive permutation approach implemented in QTLtools were applied [37]. We used one hundred permutations to empirically characterize the null distribution of the strongest signal which is fitted by a Beta distribution. The p-values of association adjusted for the number of variants

538 tested in *cis* given by the fitted beta distribution were used to calculate gene-level q-values. By
539 controlling the false discovery rate (FDR) threshold < 5%, significant SOMAmers were identified.
540
541 **Comparison with previous identified *cis*-pQTL.** A list of existing pQTL studies were summarized
542 by Karsten Suhre (http://www.metabolomix.com/a-table-of-all-published-gwas-with-
543 proteomics/) [25]. We focus on two recent European-ancestry pQTL studies with large sample size
544 and proteins assayed by SOMAscan. The first was performed in the INTERVAL study with UK blood
545 donors [16]. The other was performed in the AGES-RS cohort [17]. To make fair comparison, we
546 compared identified *cis*-pQTLs across the two analyses using the same standard -- sentinel *cis*-
547 associations (+/-500Kb) for common SNPs (MAF>0.01) and Bonferroni corrected genome-wide
548 threshold for significance. Using these criteria, the two previous studies identified a total of 508
549 unique significant SOMAmers (304 and 422 respectively) and we identified 1,465 significant
550 SOMAmers. We then tested replication of their sentinel SNPs in our ARIC EA sample (Bonferroni
551 corrected p-value < $0.05/726 = 6.89 \times 10^{-5}$, where 726 = 218x2 + 204 + 86. There were 218
552 SOMAmers discovered in both studies, 204 discovered only in AGES-RS and 86 discovered only in
553 INTERVAL). If a significant SOMAmer's sentinel SNPs was not available in ARIC, we used their LD
554 proxies and the r^2 was calculated from the 1000Genome European individuals.
555
556 **Replication of *cis*-pQTL identified in AA.** We replicated *cis*-pQTLs discovered in the ARIC AA in
557 the African American Study of Kidney Disease and Hypertension (AASK), a clinical trial of alternate
558 blood pressure lowering regimen and goals [35]. Enrollment occurred from 1995 to 1998, with the
559 original trial population consisting of 1094 African American participants with chronic kidney
560 disease. Blood samples used for the measurement of the proteome were collected at baseline.

A total of 467 participants with serum protein data and genotype data were retained in the current study. Proteomic profiling was performed using the SomaScan technology using the V4.1 platform. Genotyping was conducted using the Infinium Muti-Ethnic Global BeadChip array (Illumina, GenomeStudio) and imputed to the TOPMed reference panel (Freeze 5 on GRCh38).

Independent *cis*-pQTL mapping. It is likely that the significant SOMAmers have multiple proximal *cis*-SNPs which have independent effects. To identify independent signals for them, we performed independent *cis*-pQTL mapping using the conditional pass implemented in QTLtools [37].The algorithm first uses permutations to derive a nominal p-value per SOMAmer, then it uses a forward-backward stepwise regression to select the conditional independent signals. In this process, it automatically learns the number of independent signals per SOMAmer using forward selection, and then determines the best candidate SNP per signal using backward selection controlling for the remaining signals. If no SNP is significant at the previous nominal p-value threshold, the candidate signal will be dropped; otherwise, the SNP with smallest backward-p-value will be chosen as the lead SNP for this candidate signal. In some cases, the same SNP during the backward selection can explain multiple independent signals that were detected during the forward selection. In the reporting our results (**Supplementary Table 6.1 and 6.2**), we show the rank of all the SNPs selected by the forward selection step that is explained by a given lead SNP selected during the final backward selection step.

To account for power for detection in **Fig. 1c**, we adjusted the SNP effect sizes by assigning a weight of the inverse of statistical power. The statistical power can be derived as following.

The SNP effect is chi-square distributed with one degree of freedom (df). It is a central chi-square distribution under the null, and a non-central chi-square distribution under the alternative

hypothesis. The non-centrality parameter (NCP), λ, is $\frac{N(1-2f(1-f)\beta^2)}{2f(1-f)\beta^2}$, where N is the number of samples in study, f is the MAF of the SNP, and β is the SNP effect [61, 62]. The significance threshold for the test statistic under the central chi-square distribution of df 1 and the SOMAmer's nominal p-value cut-off, p_0, is $t_0 = F^{-1}(1-p_0, 1)$, where $F(\cdot, 1)$ is the cumulative distribution function (CDF) of a central chi-square distribution of df 1. The statistical power can be computed by $Pr(T > t_0 \mid H_a) = 1 - G(t_0, \lambda, 1)$, where T is the test statistics and $G(\cdot, \lambda, 1)$ is the CDF of the non-central chi-square distribution with NCP of λ and df 1. The weight assigned to SNP effect is $\left(1 - G(t_0, \lambda, 1)\right)^{-1}$.

Investigation of epitope-binding effects. SOMAscan assay relies on aptamer binding which may be influenced by the change of protein structure. Protein altering variants (PAV) may result in *cis*-pQTLs by altering binding affinity, instead of protein abundance. Following a procedure recommend earlier [16], we cataloged all *cis*-pQTLs that were not in LD (r^2<0.1) with any PAV in the *cis* region or those in LD (0.1$\leqslant r^2 \leqslant$0.9) but remain significant in a conditional analysis after adjusting for PAVs. We annotated variants with variant effect predictor (VEP) [39], Loss-Of-Function Transcript Effect Estimator (LOFTEE) [40] and Ensembl Regulatory Build [41]. Variants were considered to be PAV if annotated as coding sequence, frameshift, in-frame deletion, in-frame insertion, missense, splice acceptor, splice donor, splice region, start lost, stop gained, or stop lost variants. LD-pruned (r^2>0.9) PAVs were included as covariates for association testing.

***Cis*-eQTL overlap.** We cross referenced the identified *cis*-pQTLs against *cis*-eQTLs identified in the overall analysis of GTEx (V8) data across different tissues. For each SOMAmer, we first extracted the sentinel *cis*-pQTLs, meaning the variants having most significant association for a pGene along

with all the variants in high LD (r^2 > 0.8). Using this list of variants across 2,004 SOMAmers which had at least one *cis*-pQTL in EA, we calculated the percentage overlap with the set of significant *cis*-eQTLs (at FDR<5%, as defined by GTEx consortium) for the same gene identified in each tissue of GTEx V8 [9]. Since the GTEx cohort is primarily of European ancestry, we restricted this analysis to EA only.

Colocalization. Colocalization analysis was performed to investigate whether the same variants were likely to be causal for variation in protein levels and gene expression levels. We used publicly available overall *cis*-eQTL summary statistics from GTEx consortium (V8). For testing whether *cis*-eQTL and *cis*-pQTL associations for the same gene colocalize, we used coloc package in R with the default setting [63]. Evidence for colocalization was assessed using the posterior probability (PP) for the hypothesis that there is an association for both protein levels and gene expression levels, and they are driven by the same causal variant (PP.H4). Since we tested across a large number of tissues, we chose a stringent cut-off of 0.8 and pGenes with PP.H4 > 0.8 were identified as likely to have a shared causal variant for the *cis*-eQTL and *cis*-pQTL associations. As before, we restricted our analysis to the 2,004 pGenes identified in EA.

Function annotations enrichment. We performed an enrichment analysis of the *cis*-pQTLs for known regulatory elements in the genome to identify the broad functions of the *cis*-pQTLs. The functional annotations were curated from variant effect predictor (VEP) [39], Loss-Of-Function Transcript Effect Estimator (LOFTEE) [40] and Ensembl Regulatory Build as was reported in the recent GTEx analysis. For each SOMAmer, we used sentinel *cis*-pQTLs, meaning the variants having the most significant association and variants in high LD (r^2 > 0.8) for evaluating functional

enrichment. With these annotations, we used torus [64] to perform functional enrichment for each functional category. To remove effect of potential epitope binding effects associated with the PAVs, we also investigated functional enrichment among sentinel *cis*-pQTLs (and variants in high LD) that showed significant effects independent of the PAVs (See previous section for details).

Fine-mapping analysis. To identify the set of possibly causal variants regulating plasma protein levels we performed fine-mapping [65] using the *cis*-variants for each of the 1,447 SOMAmers that had at least one *cis*-pQTL in both EA and AA using SuSiE [42]. For a given SOMAmer and corresponding variants in the *cis*-regulatory region, SuSiE outputs a number of single effect components or credible sets that have 95% probability to contain a variant with non-zero causal effect. We set the maximum number of such singlet effect components to be 10, meaning broadly we allow for the possibility that a SOMAmer can be regulated by 10 causal variants at best. Further, SuSiE also outputs the posterior inclusion probability for each variant. This corresponds to the probability of the variant to be included in one of the credible sets.

To perform trans-ethnic meta-analysis, we used MANTRA [43] which is based on a computationally intensive Bayesian partition accounting for the shared similarity in closely related populations assuming the same underlying allelic effect. It models the effect heterogeneity among distant populations by clustering according to the shared ancestry and allelic effects. MANTRA outputs the Bayes factor for association of a variant across ancestries. Using this, we constructed the posterior probability [66] of the kth variant (π_k) as:

$$\pi_k = \frac{\delta_k}{\sum \delta_k}$$

where δ_k is the Bayes factor for association of the kth variant obtained using trans-ethnic meta-analysis in MANTRA and the sum in the denominator is across all the variants in the *cis*-region.

We performed MANTRA using the variants common to EA and AA and subsequently calculated the posterior probabilities.

***Cis*-SNP heritability estimation.** *Cis*-SNP heritability (*cis*-h^2) of SOMAmers were estimated using the REML algorithm implemented in GCTA [49]. Genotypes of SNPs in a *cis*-window around the encoding gene of the corresponding target protein of a SOMAmer were used to estimate genetic relatedness matrix (GRM). Corrected-protein quantifications and the estimated GRM were input to the GCTA to estimate *cis*-h^2 using the REML algorithm (option --reml --reml-no-constrain). A maximum number of 100 iterations was set to determine the convergence of the estimation algorithm. The nonzero *cis*-heritability was tested using a likelihood-ratio test for the first genetic variance component (option --reml-lrt 1) with significance level of 0.01. Plasma protein SOMAmers with negative estimate *cis*-h^2 estimates were excluded. *Cis* window size of +/- 500Kb and 1Mb were examined, and there were no significant differences between the heritability estimations (**Supplementary Table 20**). Therefore, throughout the paper, we defined +/- 500Kb window size which is same as those used for TWAS models we used.

Imputation models trained jointly with *cis*-SNPs. Using the TWAS / FUSION software [28], we built imputation models for 1,394 (AA) and 1,350 (EA) SOMAmers with significant non-zero *cis*-h^2. Imputation model for a SOMAmer was trained jointly by elastic net using *cis*-SNPs in +/-500Kb around the TSS of the encoding gene of the target protein. The performance of models was evaluated by adjusted prediction accuracy which was defined as the 5-fold cross-validated R^2 between predicted and true values standardized by *cis*-h^2. The imputation models built only with the top *cis*-pQTL was used as a baseline comparison.

Trans-ethnic prediction capacity. To study the trans-ethnic prediction performance, we applied the genetic imputation models to the genotypes of individuals from their opposite races in ARIC. The cross-ethnic prediction performance is evaluated by the R^2 between predicted and true values standardized by *cis*-h^2.

***Cis*-regulated genetic correlation between plasma proteome and transcriptome across a variety of tissues.** To study the *cis*-regulated genetic correlation between plasma protein and expression levels for underlying genes across a variety of tissues, we computed the Pearson's correlation coefficients between genotypically-imputed plasma proteins and genotypically-imputed gene expressions for the same gene for individuals from Phase-3 1000 Genome Project (1000Genome) [51] by applying weights of their imputation models to the genotype data. For primary analyses, we used established gene expression imputation models available based on GTex V7 dataset across different tissues (http://gusevlab.org/projects/fusion/#reference-functional-data (see **Supplementary Table 13** for the full list, **Supplementary Table 14** for their prediction accuracies). Here we only studied for genes significant *cis*-heritable (p-value of *cis*-h^2 from GCTA < 0.01) for both gene expression levels and plasma protein levels (**Supplementary Tables 15.1 and 15.2**). Since the gene expression imputation models were derived using participants predominantly from European ancestry from GTEx V7, the plasma protein imputation models here were restricted to EA-derived only. If multiple transcripts or SOMAmers were measured for the same gene, the sum of their imputed levels was used to represent "the total level of the gene" in terms of gene expression or plasma protein level. We also obtained preliminary gene-expression imputation models trained based on GTEx V8 dataset (obtained based personal communication

699 with Gusev lab) and used them to conduct several secondary/validation analyses for comparison
700 of results with V7.
701
702 **Proteome-wide association studies (PWAS).** As an analog of TWAS, weights in the imputation
703 models of SOMAmers can be applied to summary level data using the test statistics derived in
704 TWAS / FUSION. The mathematical derivation can be found in the original paper [28]. The type 1
705 error of PWAS is well-controlled in simulation using null phenotypes simulated from UK Biobank
706 using 337,484 unrelated European ancestry individuals [67]. Note that the enet model coefficients
707 for 9 proteins in AA and 2 proteins in EA were all zero. These proteins were excluded in PWAS
708 analysis, and therefore, 1,385 (AA) and 1,348 (EA) imputation models were available in PWAS.
709 The significance level for PWAS loci identification is adjusted by of the total number of imputation
710 models for significant *cis*-heritable plasma proteins or protein complexes (p-value <
711 $0.05/1{,}348=3.7\text{x}10^{-5}$ in EA which was used in our PWAS of serum urate and gout). As discussed in
712 a recent TWAS paper [50], multiple SOMAmers, whose encoding genes of their target proteins or
713 protein complexes locate closely in a locus, were sometimes identified at the same time. To
714 identify distinct loci, a 1Mb region (+/- 500Kb of TSS) was defined around each encoding gene of
715 the target protein of significant SOMAmers, and overlapping regions were merged. The sentinel
716 association in each locus was selected to be the top PWAS candidate hit for this region
717 (**Supplementary Tables 21.1 and 21.2**).
718 We obtained standardized estimate for the causal effect ($\hat{\gamma}_P$) and standard error ($se(\hat{\gamma}_P)$), and
719 thereby confidence intervals, of the underlying proteins on the complex traits (Y) by slightly
720 extending S-PrediXcan [68]. We derived these as

721 $$\hat{\gamma}_P = \frac{Cov(\hat{P}, Y)}{Var(\hat{P})} = \frac{Cov(\sum_{l=1}^{M} w_{Pl} X_l, Y)}{\hat{\sigma}_P^2} = \frac{\sum_{l=1}^{M} w_{Pl} Cov(X_l, Y)}{Var(\sum_l w_{Pl} X_l)} = \frac{\sum_{l=1}^{M} w_{Pl} \hat{\beta}_l \sigma_l^2}{\boldsymbol{W_P^T \Gamma W_P}}$$

$$se(\hat{\gamma}_P)^2 = \frac{\hat{\sigma}_Y^2}{N}\frac{1-R_P^2}{\hat{\sigma}_P^2} \approx \frac{1}{M}\sum_{l=1}^{M}\left(\frac{se\left(\widehat{\beta}_l\right)^2\sigma_l^2}{1-R_l^2}\right)\frac{1-R_P^2}{\hat{\sigma}_P^2} \approx \frac{\sum_{l=1}^{M} se\left(\widehat{\beta}_l\right)^2\sigma_l^2}{M}\frac{1}{\boldsymbol{W_P^T \Gamma W_P}}$$

where $\widehat{\beta}_l$ is SNP l's summary statistics for the complex trait, w_{Pl} is SNP l's weight in the imputation model for protein P, σ_l^2 is the variance of SNP l which can be computed from allele frequency, and $\boldsymbol{\Gamma}$ is the LD (correlation) matrix for all M SNPs in the imputation model. We used the same formulae to derive corresponding causal effects, standard errors and confidence intervals for results from TWAS analyses.

Druggability of PWAS genes. PWAS genes were annotated based on the therapeutic target database [69]. Only drugs that were actively pursued were retained in the database and discontinued, terminated or withdrawn drugs were excluded. Additionally, druggability tiers from Finan et al. [70] were mapped via gene symbols (**Supplementary Table 18**).

Bivariate conditional analysis for PWAS and TWAS. For each significant PWAS loci, we searched all TWAS genes nearby (+/-500Kb around) whose TSS locate within 500Kb of the TSS of its sentinel PWAS gene, and selected the one with the smallest TWAS p-value. The position of genes in TWAS (based on GTEx V7 based on genome build GRCh37) and PWAS (based on genome build GRCh38) were matched using the UCSC genome browser webtool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) [71].

We first performed the nearby TWAS in two trait-relevant tissues, whole blood and liver, for serum urate and gout. Note that kidney is also a trait-relevant tissue, but there is no imputation model trained with GTEx V7 data available on TWAS / FUSION for kidney. The significance of the

743 nearby TWAS hit was determined by significance level after Bonferroni Correction ($0.05/$
744 $\sum_{relevant\ tissues}$ #transcripts with imputation models).

745 Using z-scores (z_P for PWAS gene and z_T for TWAS gene) and the *cis*-regulated genetic
746 correlation (ρ) of each PWAS gene and the most significant TWAS gene nearby, we performed
747 conditional analysis [72] to study the potential underlying mechanism of gene expressions in tissue
748 or proteins in plasma. The *cis*-regulated genetic correlation was computed from the Pearson's
749 correlation coefficients between genotypically-imputed plasma proteins and genotypically-
750 imputed gene expressions for individuals from 1000Genome by applying weights of their
751 imputation models to the genotype data. The least-squares estimate of the PWAS z-score
752 conditional on TWAS z-score is

754 $$z_P|z_T = z_P - \rho z_T$$

753 and its variance is

755 $$\mathrm{var}(z_P|z_T) = \mathrm{var}(z_P) - \mathrm{var}(\rho z_T) = 1 - \rho^2$$

756 So the conditional z-score of the PWAS gene is

758 $$z_{P|T} = \frac{z_P - \rho z_T}{\sqrt{1-\rho^2}}$$

757 Similarly, the conditional z-score of the nearby TWAS gene is

759 $$z_{T|P} = \frac{z_T - \rho z_P}{\sqrt{1-\rho^2}}$$

760 We then performed the same procedure for all nearby TWAS genes in *all* GTEx V7 tissues. Using
761 Bonferroni Correction for the total number of transcripts with imputation models ($0.05/$
762 $\sum_{all\ GTex\ tissues}$ #transcripts with imputation models), we identified the tissues which have at
763 least one significant TWAS gene in the PWAS significant loci. The most significant TWAS gene in
764 this region and its corresponding tissue were recorded, and then used to perform conditional

765 analysis (**Supplementary Tables 22.1 and 22.2**). We further validated the top gene-tissue
766 combination identified through TWAS models in V7 using preliminary models that were available
767 to us based on V8.

768 References

769 1. Buniello, A. *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association
770 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005-D1012
771 (2019).

772 2. Visscher, P. M. *et al*. 10 years of GWAS discovery: biology, function, and translation. *The*
773 *American Journal of Human Genetics* **101**, 5-22 (2017).

774 3. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**,
775 R102-R110 (2015).

776 4. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome
777 engineering to understand the functional relevance of SNPs in non-coding regions of the human
778 genome. *Epigenetics & chromatin* **8**, 1-18 (2015).

779 5. Musunuru, K. *et al*. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol
780 locus. *Nature* **466**, 714-719 (2010).

781 6. Kumar, V. *et al*. Human disease-associated genetic variation impacts large intergenic non-
782 coding RNA expression. *PLoS Genet* **9**, e1003201 (2013).

783 7. Lonsdale, J. *et al*. The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45**, 580-585
784 (2013).

785 8. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204-
786 213 (2017).

787 9. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
788 tissues. *Science* **369**, 1318-1330 (2020).

789 10. Torres, J. M. *et al*. Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a
790 complex trait. *The American Journal of Human Genetics* **95**, 521-534 (2014).

791 11. Hormozdiari, F. *et al*. Leveraging molecular quantitative trait loci to understand the genetic
792 architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041-1047 (2018).

793 12. Gold, L. *et al*. Aptamer-based multiplexed proteomic technology for biomarker discovery.
794 *Nature Precedings*, 1 (2010).

795 13. Williams, S. A. *et al*. Plasma protein patterns as comprehensive indicators of health. *Nat.*
796 *Med.* **25**, 1851-1857 (2019).

797 14. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and
798 diagnostic prospects. *Molecular & cellular proteomics* **1**, 845-867 (2002).

799 15. Folkersen, L. *et al*. Genomic and drug target evaluation of 90 cardiovascular proteins in
800 30,931 individuals. *Nature metabolism* **2**, 1135-1148 (2020).

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

802 17. Emilsson, V. *et al*. Human serum proteome profoundly overlaps with genetic signatures of
803 disease. *BioRxiv* (2020).

804 18. Yao, C. *et al*. Genome-wide mapping of plasma protein QTLs identifies putatively causal
805 genes and pathways for cardiovascular disease. *Nature communications* **9**, 1-11 (2018).

806 19. Pietzner, M. *et al*. Genetic architecture of host proteins involved in SARS-CoV-2 infection.
807 *Nature communications* **11**, 1-14 (2020).

808 20. Zhou, S. *et al*. A Neanderthal OAS1 isoform protects individuals of European ancestry
809 against COVID-19 susceptibility and severity. *Nat. Med.*, 1-9 (2021).

810 21. Yang, C. *et al*. Genomic and multi-tissue proteomic integration for understanding the
811 biology of disease and other complex traits. *medRxiv* (2020).

812 22. He, B., Shi, J., Wang, X., Jiang, H. & Zhu, H. Genome-wide pQTL analysis of protein
813 expression regulatory networks in the human liver. *BMC biology* **18**, 1-16 (2020).

814 23. Wingo, A. P. *et al*. Integrating human brain proteomes with genome-wide association data
815 implicates new proteins in Alzheimer's disease pathogenesis. *Nat. Genet.*, 1-4 (2021).

816 24. Bretherick, A. D. *et al*. Linking protein to phenotype with Mendelian Randomization detects
817 38 proteins with causal roles in human diseases and traits. *PLoS genetics* **16**, e1008785 (2020).

818 25. Suhre, K. *et al*. Connecting genetic risk to disease end points through the human blood
819 plasma proteome. *Nature communications* **8**, 1-14 (2017).

820 26. Zheng, J. *et al*. Phenome-wide Mendelian randomization mapping the influence of the
821 plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122-1131 (2020).

822 27. Gamazon, E. R. *et al*. A gene-based association method for mapping traits using reference
823 transcriptome data. *Nat. Genet.* **47**, 1091 (2015).

824 28. Gusev, A. *et al*. Integrative approaches for large-scale transcriptome-wide association
825 studies. *Nat. Genet.* **48**, 245-252 (2016).

826 29. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC
827 investigators. *Am. J. Epidemiol.* **129**, 687-702 (1989).

828 30. Pietzner, M. *et al*. Cross-platform proteomics to advance genetic prioritisation strategies.
829 *bioRxiv* (2021).

830 31. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in
831 Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195-R208 (2018).

832 32. Tin, A. *et al*. Target genes, variants, tissues and transcriptional pathways influencing human
833 serum urate levels. *Nat. Genet.*, 1-16 (2019).

834 33. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-
835 genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput*
836 *Biol* **6**, e1000770 (2010).

837 34. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of
838 expression residuals (PEER) to obtain increased power and interpretability of gene expression
839 analyses. *Nature protocols* **7**, 500 (2012).

840 35. Gassman, J. J. *et al*. Design and statistical aspects of the African American Study of Kidney
841 Disease and Hypertension (AASK). *Journal of the American Society of Nephrology* **14**, S154-S165
842 (2003).

843 36. Park, J. *et al*. Distribution of allele frequencies and effect sizes and their interrelationships
844 for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*
845 **108**, 18026-18031 (2011).

846 37. Delaneau, O. *et al*. A complete tool set for molecular QTL discovery and analysis. *Nature*
847 *communications* **8**, 1-7 (2017).

848 38. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL
849 mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485 (2016).

850 39. McLaren, W. *et al*. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

851 40. Karczewski, K. J. *et al*. The mutational constraint spectrum quantified from variation in
852 141,456 humans. *Nature* **581**, 434-443 (2020).

853 41. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl
854 regulatory build. *Genome Biol.* **16**, 56 (2015).

855 42. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable
856 selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical*
857 *Society: Series B (Statistical Methodology)* **82**, 1273-1300 (2020).

858 43. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet.*
859 *Epidemiol.* **35**, 809-822 (2011).

860 44. He, Z., Song, D., van Zalen, S. & Russell, J. E. Structural determinants of human ζ-globin
861 mRNA stability. *Journal of hematology & oncology* **7**, 35 (2014).

862 45. He, Z. & Russell, J. E. Effect of ζ-globin substitution on the O2-transport properties of Hb S in
863 vitro and in vivo. *Biochem. Biophys. Res. Commun.* **325**, 1376-1382 (2004).

864 46. Lafferty, J. D. *et al*. A Multicenter Trial of the Effectiveness of ζ-Globin Enzyme-Linked
865 Immunosorbent Assay and Hemoglobin H Inclusion Body Screening for the Detection of α0-
866 Thalassemia Trait. *Am. J. Clin. Pathol.* **129**, 309-315 (2008).

867 47. Watanabe, K., Stringer, S., Polderman, T. & Posthuma, D. *A global view of the genetic*
868 *architecture in human complex traits* (HUMAN GENOMICS Ser. 12, BIOMED CENTRAL LTD 236
869 GRAYS INN RD, FLOOR 6, LONDON WC1X 8HL, ENGLAND, 2018).

870 48. Astle, W. J. *et al*. The allelic landscape of human blood cell trait variation and links to
871 common complex disease. *Cell* **167**, 1415-1429. e19 (2016).

872 49. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex
873 trait analysis. *The American Journal of Human Genetics* **88**, 76-82 (2011).

874 50. Wainberg, M. *et al*. Opportunities and challenges for transcriptome-wide association
875 studies. *Nat. Genet.* **51**, 592-599 (2019).

876 51. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*
877 **526**, 68-74 (2015).

878 52. Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in
879 human liver. *Electrophoresis* **18**, 533-537 (1997).

880 53. Mancuso, N. *et al*. Integrating gene expression with summary association statistics to
881 identify genes associated with 30 complex traits. *The American Journal of Human Genetics* **100**,
882 473-487 (2017).

883 54. Köttgen, A. *et al*. Genome-wide association analyses identify 18 new loci associated with
884 serum urate concentrations. *Nat. Genet.* **45**, 145-154 (2013).

885 55. Janssen, C. A. *et al*. Anakinra for the treatment of acute gout flares: a randomized, double-
886 blind, placebo-controlled, active-comparator, non-inferiority trial. *Rheumatology* **58**, 1344-1352
887 (2019).

888 56. Wojcik, G. L. *et al*. Genetic analyses of diverse populations improves discovery for complex
889 traits. *Nature* **570**, 514-518 (2019).

890 57. Durinck, S. *et al*. BioMart and Bioconductor: a powerful link between biological databases
891 and microarray data analysis. *Bioinformatics* **21**, 3439-3440 (2005).

892 58. Kowalski, M. H. *et al*. Use of> 100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed)
893 Consortium whole genome sequences improves imputation quality and detection of rare
894 variant associations in admixed African and Hispanic/Latino populations. *PLoS genetics* **15**,
895 e1008500 (2019).

896 59. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation
897 method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529
898 (2009).

899 60. Price, A. L. *et al*. Principal components analysis corrects for stratification in genome-wide
900 association studies. *Nat. Genet.* **38**, 904-909 (2006).

901 61. Wang, M. & Xu, S. Statistical power in genome-wide association studies and quantitative
902 trait locus mapping. *Heredity* **123**, 287-306 (2019).

903 62. Schmid, A. B. *et al*. Genetic components of human pain sensitivity: a protocol for a genome-
904 wide association study of experimental pain in healthy volunteers. *BMJ open* **9**, e025530 (2019).

905 63. Giambartolomei, C. *et al*. Bayesian test for colocalisation between pairs of genetic
906 association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).

907 64. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false
908 discovery rate control. *Annals of Applied Statistics* **10**, 1619-1638 (2016).

909 65. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal
910 variants by statistical fine-mapping. *Nature Reviews Genetics* **19**, 491-504 (2018).

911 66. Mahajan, A. *et al*. Trans-ethnic fine mapping highlights kidney-function genes linked to salt
912 sensitivity. *The American Journal of Human Genetics* **99**, 636-646 (2016).

913 67. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature*
914 **562**, 203-209 (2018).

915 68. Barbeira, A. N. *et al*. Exploring the phenotypic consequences of tissue specific gene
916 expression variation inferred from GWAS summary statistics. *Nature communications* **9**, 1-20
917 (2018).

918 69. Wang, Y. *et al*. Therapeutic target database 2020: enriched resource for facilitating research
919 and early development of targeted therapeutics. *Nucleic Acids Res.* **48**, D1031-D1041 (2020).

920 70. Finan, C. *et al*. The druggable genome and support for target identification and validation in
921 drug development. *Science translational medicine* **9** (2017).

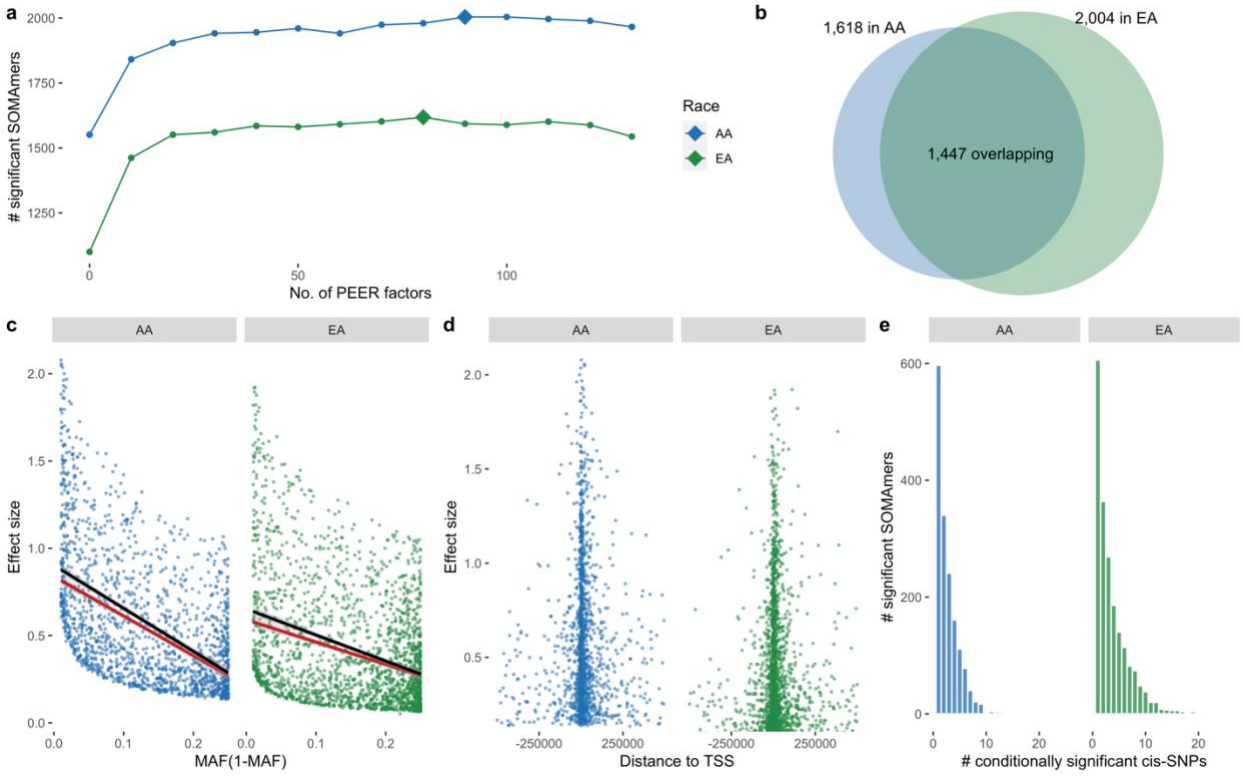
922 71. Navarro Gonzalez, J. *et al*. The UCSC Genome Browser database: 2021 update. *Nucleic Acids*
923 *Res.* **49**, D1046-D1057 (2021).

924 72. Yang, J. *et al*. Conditional and joint multiple-SNP analysis of GWAS summary statistics
925 identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369-375 (2012).

926

927

928 **Fig. 1: *Cis*-pQTL analysis**

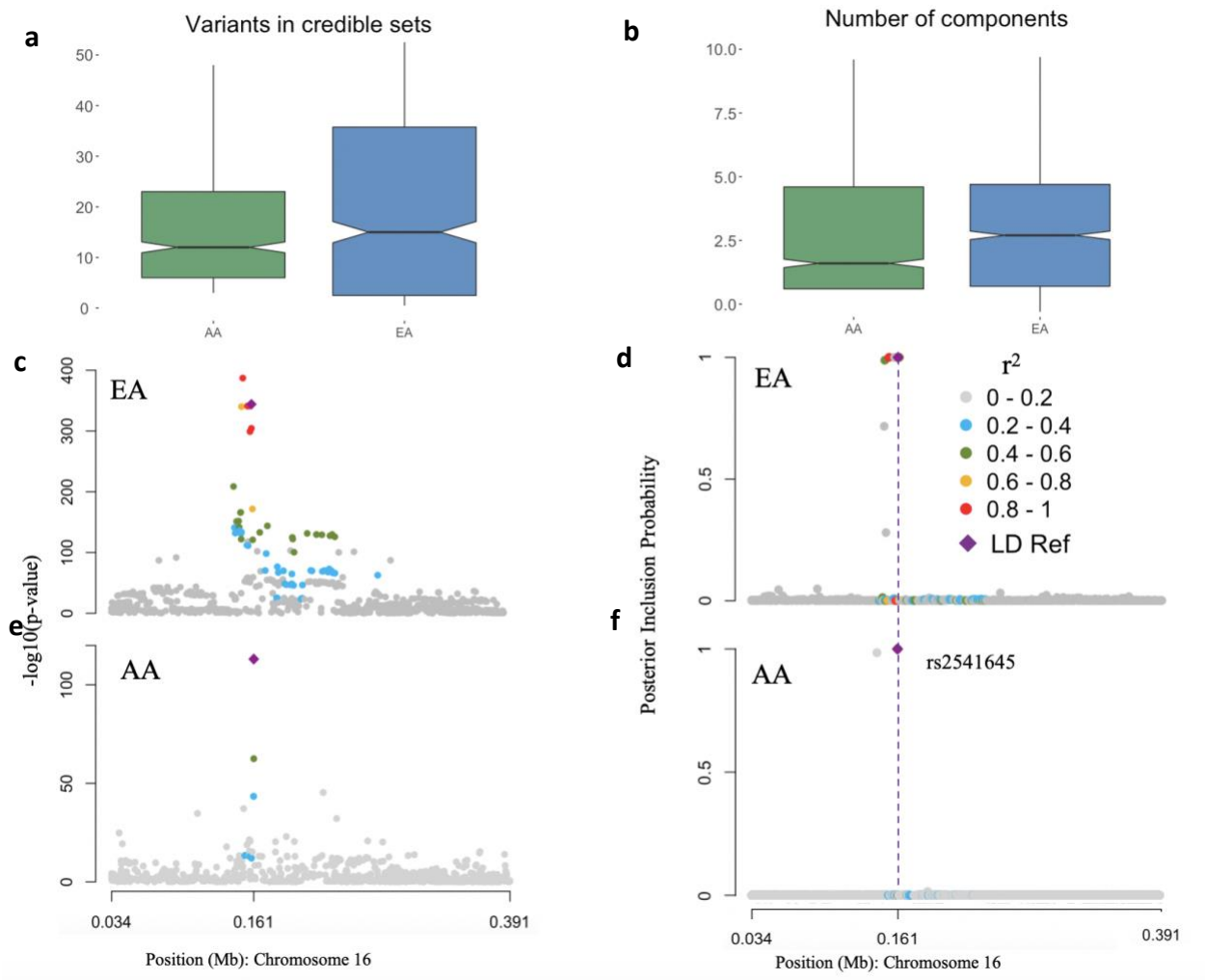


929

930 (**a**) Number of SOMAmers detected to have significant *cis*-pQTLs versus number of PEER factors
931 used in models. Diamonds mark the numbers of PEER factors used in the following analysis which
932 identify maximal number of significant SOMAmers. (**b**) Venn diagram of significant SOMAmers in
933 EA and AA. (**c**) Effect sizes of sentinel *cis*-SNPs of pQTLs vs. minor allele frequencies (MAF(1-
934 MAF)). Lines are fitted with (red) and without inverse-power weighting (black). (**d**) Effect sizes of
935 sentinel *cis*- SNPs of pQTLs v.s. distance to TSS. (**e**) Number of conditional independent *cis*-pQTLs
936 per significant SOMAmer.

937 **Fig. 2: Fine-mapping analysis**

938



939

940 (**a**) Distribution of size of credible sets and (**b**) that of number of independent SuSIE clusters

941 across 1,447 proteins that had at least one significant *cis*-pQTL both in European American and

942 African American populations. The power of fine-mapping using data from two populations is

943 further illustrated using the example of *HBZ*. Regional Manhattan plots based on single SNP p-

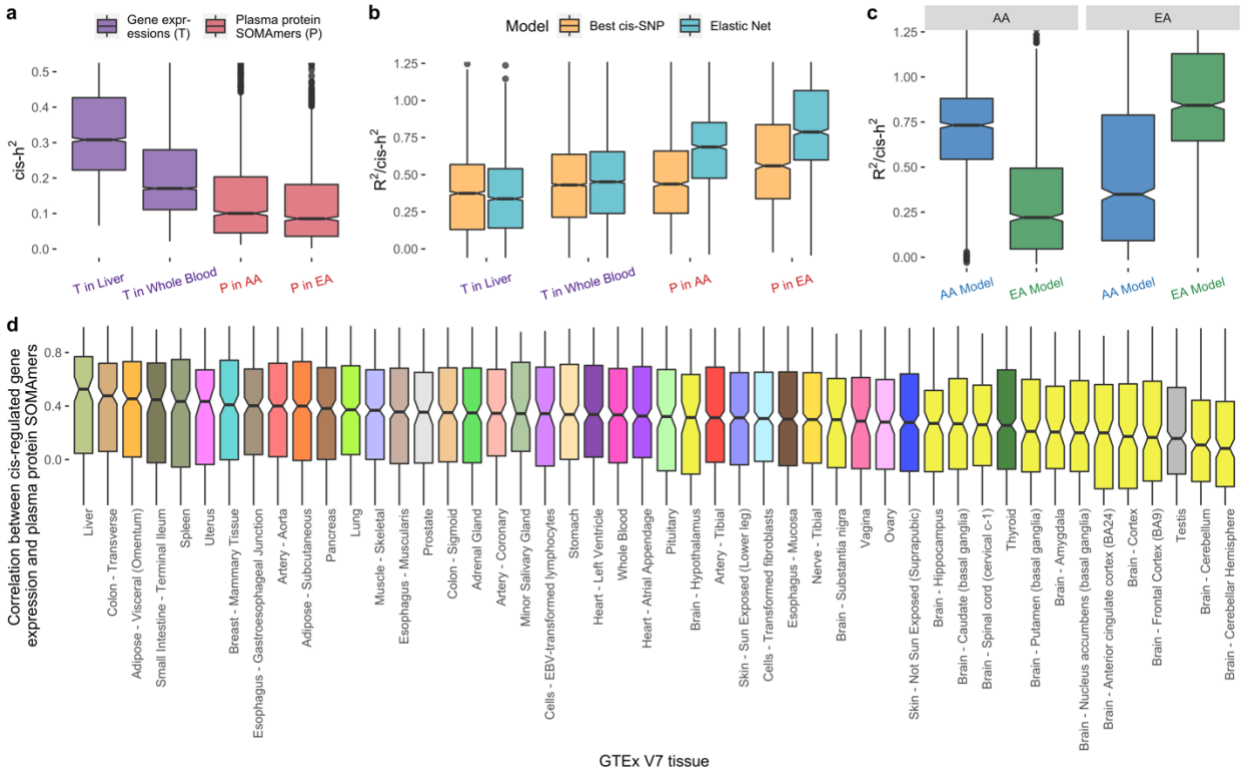
944 value and SuSIE posterior probabilities are shown for EA (Panel **c** and **d**) and AA (Panel **e** and **f**).

945 The SNP rs2541645 (chr16: 161106; marked in diamond shape throughout) is detected as the

946 shared causal *cis*-pQTL across the two ancestries by MANTRA. This variant has been used as the

947 LD reference variant throughout.

948 **Fig. 3: *Cis*-heritability and evaluation of models for genetic prediction of proteins**



949

950 (**a**) Estimated *cis*-h^2 for gene expression levels and plasma protein levels with significant

951 heritability (p-value < 0.01). (**b**) Prediction R^2 standardized by estimated *cis*-h^2 (R^2/*cis*-h^2) using

952 prediction models trained by: the most significant *cis*-SNP; and Elastic Net using all *cis*-SNPs. (**c**)

953 Cross-ethnic prediction accuracy by applying prediction models to individuals from their opposite

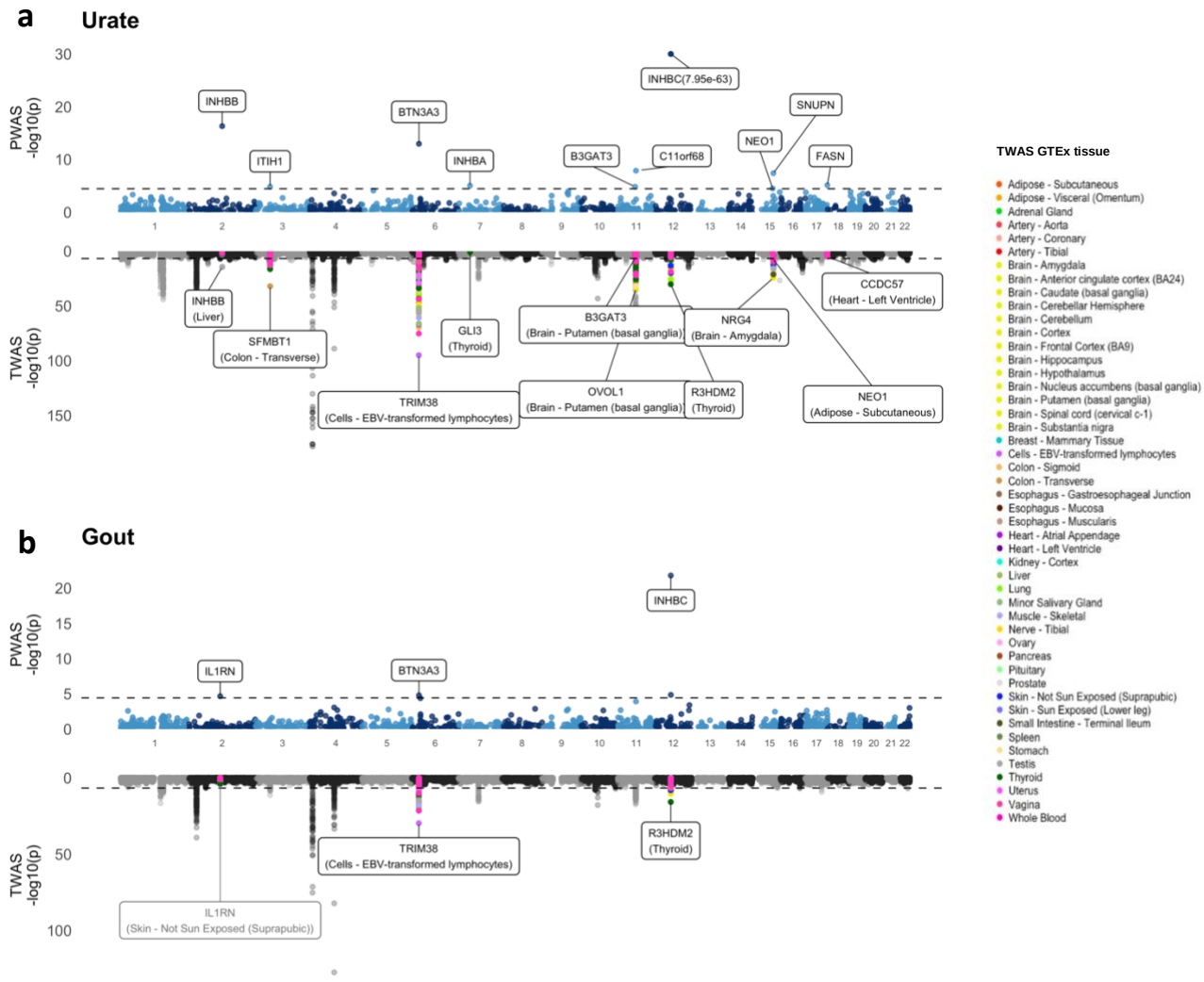
954 races (**d**) *Cis*-regulated genetic correlation between plasma proteins and expression levels for

955 underlying genes across all GTEx (V7) tissues. All results involving gene-expression levels are

956 reported based on established models developed using data from GTEx V7, and additional results

957 using preliminary models available from GTEx V8 can be found in **Supplementary Table 16**.

Fig. 4: Miami plot for PWAS and TWAS analyses for serum urate level and gout



Miami plot for PWAS (upper) and TWAS (lower) of (**a**) urate and (**b**) gout. Each point represents a test of association between the phenotypes and the *cis*-genetic regulated plasma protein or expression level of a gene, ordered by genomic position on the x axis and the -log10(p-value) for the association strength on the y axis. The black horizontal dash lines are the significance threshold after Bonferroni correction for the total number of imputation models (p-value = 3.7×10^{-5} for PWAS and 1.3×10^{-6} for TWAS). Figure is truncated in the y-axis at $-\log_{10}$(p-value) = 30 for PWAS and $-\log_{10}$(p-value) = 150 for TWAS. Nearby TWAS genes (+/- 500Kb) for significant PWAS hits are colored by GTEx tissues. The most significant nearby-TWAS hit is labelled with its

969 gene name and corresponding tissue. The TWAS of *IL1RN* does not reach TWAS significance
970 threshold and thereby was labeled with grey. All primary TWAS analyses were conducted
971 based on established models developed using data from GTEx V7, and results for the identified
972 top genes/tissue combinations were further validated using preliminary models available from
973 GTEx V8 (**Supplementary Table 16**).

Table 1. Proteome-wide association analysis of Serum Urate Level and and Gout. Analysis was done based on external summary statistics data from GWAS of serum urate level (N=288,649) and gout (N=754,056) and the imputation models for plasma proteome built from the ARIC study for a total of 1,348 plasma proteins with significant *cis*-heritability. Results are also shown for most significant genes from Transcriptome Wide Association Studies around +/- 500kb region of the TSS of sentinel protein for two specific trait-relevant tissues (whole blood and liver) and across all tissues. Further result from bivariate analysis of genetically imputed level of the plasma protein and that of the expression for most significant gene from the TWAS analysis are reported in terms of conditional p-values. All TWAS analyses were performed based on models available from the GTEX V7 datasets. Results for identified top genes/tissue combinations were further validated using preliminary models available from GTEx V8 (**Supplementary Table 16**).

A: Urate

Plasma PWAS		Relevant-tissue TWAS***						All-tissue TWAS						
Gene	**P-value**	**Relevant tissue**	**Gene**	**P-value**	**cor (P, T)**	**pval(P|T)**	**pval(T|P)**	**Most significant tissue**	**Gene**	**P-value**	**cor (P, T)**	**pval(P|T)**	**pval(T|P)**	**# of signifi-cant tissues****
INHBB (2q14.2)	5.01×10^{-17}	Blood	*RALB*	1.32×10^{-2}	0.04	1.05×10^{-16}	3.07×10^{-2}	Liver*	*INHBB**	3.92×10^{-15}	0.97	1.68×10^{-3}	2.58×10^{-1}	2
		Liver*	*INHBB**	3.92×10^{-15}	0.97	1.68×10^{-3}	2.58×10^{-1}							
ITIH1 (3q21.1)	1.53×10^{-5}	Blood*	*MUSTN1**	2.47×10^{-12}	-0.67	6.06×10^{-1}	3.11×10^{-8}	Colon - Transverse*	*SFMBT1**	1.08×10^{-32}	-0.48	1.18×10^{-1}	3.86×10^{-29}	48
		Liver*	*SERBP1P3**	3.21×10^{-9}	-0.12	3.86×10^{-7}	8.62×10^{-11}							
BTN3A3 (6p22.2)	1.13×10^{-13}	Blood*	*TRIM38**	5.83×10^{-76}	0.41	8.50×10^{-1}	5.85×10^{-64}	Cells - EBV-transformed lymphocytes*	*TRIM38**	1.19×10^{-95}	0.09	2.61×10^{-8}	2.16×10^{-90}	48
		Liver*	*BTN3A2**	2.74×10^{-14}	0.74	8.40×10^{-3}	1.81×10^{-3}							
INHBA (7p14.1)	9.93×10^{-6}	Blood	*NA*	*NA*	*NA*	*NA*	*NA*	Thyroid	*GLI3*	1.40×10^{-1}	-0.08	1.57×10^{-5}	2.53×10^{-1}	0
		Liver	*NA*	*NA*	*NA*	*NA*	*NA*							
C11orf68 (11q13.1)	1.40×10^{-8}	Blood*	*MAP3K11**	1.05×10^{-22}	0.20	1.68×10^{-4}	9.53×10^{-19}	Brain - Putamen (basal ganglia)*	*OVOL1**	5.55×10^{-35}	0.27	1.54×10^{-2}	3.15×10^{-29}	45
		Liver*	*EFEMP2**	2.99×10^{-7}	-0.12	3.21×10^{-7}	7.00×10^{-6}							
B3GAT3 (11q12.3)	1.56×10^{-5}	Blood*	*INTS5**	4.02×10^{-5}	-0.94	1.76×10^{-1}	8.78×10^{-1}	Brain - Putamen (basal ganglia)*	*B3GAT3**	3.32×10^{-7}	-0.82	7.93×10^{-1}	6.32×10^{-3}	1
		Liver	*BSCL2*	7.79×10^{-1}	0.27	1.06×10^{-5}	3.70×10^{-1}							
INHBC (12q13.3)	7.64×10^{-63}	Blood*	*MARS**	1.4×10^{-19}	0.52	5.18×10^{-45}	6.82×10^{-1}	Thyroid*	*R3HDM2**	7.52×10^{-31}	-0.72	6.84×10^{-34}	4.12×10^{-1}	28
		Liver	*METTL21B*	4×10^{-5}	-0.04	1.09×10^{-61}	6.72×10^{-4}							
SNUPN (15q24.2)	4.25×10^{-8}	Blood*	*SNUPN**	2.67×10^{-10}	0.74	2.26×10^{-1}	7.58×10^{-4}	Brain - Amygdala*	*NRG4**	4.63×10^{-25}	0.21	6.08×10^{-4}	4.73×10^{-21}	42
		Liver*	*UBE2Q2**	5.5×10^{-12}	-0.19	1.92×10^{-5}	2.26×10^{-9}							
NEO1 (15q24.1)	3.29×10^{-5}	Blood	*NEO1*	5.62×10^{-4}	-0.02	4.33×10^{-5}	7.47×10^{-4}	Adipose-Subcutaneous*	*NEO1**	1.71×10^{-7}	0.49	6.92×10^{-2}	2.53×10^{-4}	4
		Liver	*NA*	*NA*	*NA*	*NA*	*NA*							
FASN (17q25.3)	7.73×10^{-6}	Blood*	*CCDC57**	2.71×10^{-5}	-0.91	1.15×10^{-1}	7.59×10^{-1}	Heart - Left Ventricle	*CCDC57*	1.27×10^{-5}	-0.92	2.36×10^{-1}	4.98×10^{-1}	0
		Liver	*ARL16*	5.68×10^{-3}	-0.05	1.43×10^{-5}	1.09×10^{-2}							

B: Gout

Plasma PWAS		Relevant-tissue TWAS						All-tissue TWAS						
Gene	**P-value**	**Relevant tissue**	**Gene**	**P-value**	**cor (P, T)**	**pval(P|T)**	**pval(T|P)**	**Most significant tissue**	**Gene**	**P-value**	**cor (P, T)**	**pval(P|T)**	**pval(T|P)**	**# of signif-cant tissues****
IL1RN (2q14.1)	2.22×10^{-5}	Blood	*DDX11L2*	2.09×10^{-1}	-0.03	1.90×10^{-5}	1.71×10^{-1}	Skin - Not Sun Exposed (Suprapubic)	*IL1RN*	9.30×10^{-5}	-0.46	5.75×10^{-3}	2.67×10^{-2}	0
		Liver	*PAX8*	5.91×10^{-2}	0.03	2.86×10^{-5}	7.91×10^{-2}							
BTN3A3 (6p22.2)	1.66×10^{-5}	Blood*	*TRIM38**	2.98×10^{-22}	0.41	7.34×10^{-1}	3.33×10^{-18}	Cells - EBV-transformed lymphocytes*	*TRIM38**	2.13×10^{-30}	0.09	1.04×10^{-3}	1.07×10^{-28}	35
		Liver*	*BTN3A2**	6.77×10^{-6}	0.74	1.52×10^{-1}	5.24×10^{-2}							
INHBC (12q13.3)	1.63×10^{-22}	Blood*	*MARS**	1.84×10^{-5}	0.52	1.13×10^{-18}	3.52×10^{-1}	Thyroid*	*R3HDM2**	1.54×10^{-16}	-0.72	3.95×10^{-8}	8.76×10^{-2}	16
		Liver	*STAC3*	8.69×10^{-2}	-0.12	6.05×10^{-22}	5.60×10^{-1}							

*Genes and tissues that are significant in TWAS after Bonferroni correction of all GTEx V7 transcripts across all tissues (p-value < 0.05 / 37,366 = 1.34×10^{-6}).

**A tissue is defined significant if there is at least one transcript near the PWAS signal for which the TWAS is significant at p-value < 1.34×10^{-6}.

****NA*s in the table means no applicable TWAS model for transcripts in the defined region