Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders

Chengran Yang1,2,3, Fabiana H. G. Farias1,2,3, Laura Ibanez1,2,3, Adam Suhy1,2,3, Brooke Sadler4, Maria Victoria Fernandez1,2,3, Fengxian Wang1,2,3, Joseph L. Bradley1,2,3, Brett Eiffert1,2,3, Jorge A. Bahena1,2,3, John P. Budde1,2,3, Zeran Li1,2,3, Umber Dube1,2,3, Yun Ju Sung1,2,3, Kathie A. Mihindukulasuriya1,2,3, John C. Morris3,5,6, Anne M. Fagan3,5,6, Richard J. Perrin3,5,6,7, Bruno A. Benitez1,2,3, Herve Rhinn8, Oscar Harari1,2,3,6,9 and Carlos Cruchaga1,2,3,6,9

Understanding the tissue-specific genetic controls of protein levels is essential to uncover mechanisms of post-transcriptional gene regulation. In this study, we generated a genomic atlas of protein levels in three tissues relevant to neurological disorders (brain, cerebrospinal fluid and plasma) by profiling thousands of proteins from participants with and without Alzheimer’s disease. We identified 274, 127 and 32 protein quantitative trait loci (pQTLs) for cerebrospinal fluid, plasma and brain, respectively. cis-pQTLs were more likely to be tissue shared, but trans-pQTLs tended to be tissue specific. Between 48.0% and 76.6% of pQTLs did not co-localize with expression, splicing, DNA methylation or histone acetylation QTLs. Using Mendelian randomization, we nominated proteins implicated in neurological diseases, including Alzheimer’s disease, Parkinson’s disease and stroke. This first multi-tissue study will be instrumental to map signals from genome-wide association studies onto functional genes, to discover pathways and to identify drug targets for neurological diseases.

Genetic studies have been successful in identifying genetic regions associated with complex traits, including diabetes, cardiovascular disease and neurodegenerative diseases, among others. However, the studies have fallen short in promoting understanding of the biological mechanisms underlying those traits. Most genome-wide association studies (GWASs) identify multiple disease loci rather than functional variants or genes, which makes it difficult to biologically interpret association results and to identify novel biomarkers and drug targets. By leveraging gene expression and genetic data generated by multiple studies, including the Genotype-Tissue Expression (GTEx) project, it has been possible to identify the functional variants or genes driving some GWAS signals. GTEx and others have shown that there are more tissue-specific expression QTLs (eQTLs) in trans than in cis, and that, to identify disease-relevant functional genes, it is important to interrogate the tissue of interest for the specific trait in question.

However, eQTL mapping has not been able to fully identify the functional variants and genes driving GWAS signals. One explanation is that many genetic variants alter protein levels without affecting transcript levels. Several published studies analyzed the genetic architecture of protein levels, but most focused on a single tissue, mainly plasma. A few studies with smaller sample sizes investigated cerebrospinal fluid (CSF) or brain tissue. These studies suggested that a sizeable proportion of pQTLs are not eQTLs and that additional GWAS signals can be mapped to protein levels. Integration of pQTLs with Mendelian randomization (MR) has identified pathways and biomarkers for complex traits as well as potential therapeutics that could be repurposed.

In this study, we combined high-throughput proteomics from multiple tissues with genetic data to determine the genetic architecture of protein levels in neurologically relevant tissues (brain, CSF and plasma). This integration led to the identification of tissue-shared and tissue-specific pQTLs that are critical for the understanding of the biology of complex traits, particularly neurological diseases.

Results

Discovery of multi-tissue pQTLs. We measured the abundance of 1,305 proteins using an aptamer-based platform in CSF (n = 971), plasma (n = 636) and brain (n = 458) samples (Extended Data Fig. 1 and Supplementary Table 1). We included multiple technical and biological replicates to confirm the replicability and reproducibility of our proteomic measurements (Extended Data Fig. 2). We performed stringent quality control (QC) steps for the proteomic data (Methods). After QC, 835 CSF samples and 713 proteins, 529 plasma samples and 931 proteins and 380 brain samples and 1,079 proteins were included in the analyses (Supplementary Tables 1 and 2). The cohort included individuals with Alzheimer’s disease (AD)
and cognitively normal individuals of European ancestry (Table 1). To identify pQTLs within each tissue (Fig. 1a), we performed genome-wide association analyses of 14.06 million imputed autosomal common variants (minor allele frequency (MAF) ≥ 0.02) against protein levels in each tissue. We defined cis signals as those where the single-nucleotide polymorphism (SNP) fell within 1 Mb upstream or downstream of the gene start site, which might not include all enhancers for the corresponding gene. Trans signals were defined as those where the SNP fell outside of the 2-Mb window. To correct for multiple tests, we used a stringent genome-wide significance threshold.

We found 96.8% and 96.9% of plasma and brain pQTLs that were associated with gene expression (GTEx; multi-tissue P = 0.05) in the same direction. We identified two additional genome-wide and nominal associations. We were unable to test for replication of five (1.8%) pQTLs, as protein data were not available. Replication and meta-analyses using independent datasets. To replicate our pQTL findings, we analyzed several publicly available datasets (Fig. 3a–c and Extended Data Fig. 3a,b). We also performed meta-analyses and cross-tissue replication. For CSF (Fig. 3a), we found 274 independent pQTLs, of which 223 (81.3%) were cis associations, and 6% were trans associations. We were able to replicate more than 90.1% of the CSF pQTLs, which is higher than in previous studies (56.3%) pQTLs that showed at least a nominal association in brain (P < 0.05) in the same direction. We identified an additional 16 (5.8%) pQTLs that were reported as pQTLs in other tissues (plasma studies from AddNeuroMed20, INTERVAL10, KORA11, SCALLOP12, the PhenoScanner database13, a brain mass spectrometry-based pQTL study14 and in our plasma or brain pQTL data). We were unable to test for replication of five (1.8%) pQTLs, as protein levels were not available in these other studies. In summary, we were able to replicate more than 90.1% of the CSF pQTLs, which is higher than in previous studies. Twenty-two (8.1%) pQTLs are still pending replication, as current CSF studies with smaller sample sizes do not provide enough statistical power. However, based on our validation with plasma and brain pQTLs, we estimate that more than 90% of those pQTLs are real. This is supported by the fact that we have been able to replicate 96.8% and 96.9% of plasma and brain pQTLs, respectively (see below).

For plasma (Fig. 3b), we found 127 independent pQTLs, of which 17 were novel. For replication of plasma pQTLs, we used the five studies mentioned above. We were able to replicate 96.8% of our 127 pQTLs. We were unable to test two (1.6%), as they were not measured in those studies. For brain (Fig. 3c), we found 32 independent pQTLs, of which 27 were novel. As there were no published studies using the same aptamer-based proteomic method, we matched our proteins with a mass spectrometry-based pQTL dataset15. We were able to replicate five (15.6%) signals at genome-wide significance, eight (25%) signals at a nominal (P < 0.05) association and 18 (56.3%) pQTLs that showed at least a nominal association in brain or CSF. Only one (3.1%) pQTL was not replicated.

To increase statistical power and to identify additional genomewide significant pQTLs, we performed meta-analyses that included all CSF cohorts as well as multi-tissue analyses. We first performed

### Table 1 | Demographics of the baseline cohort

|          | CSF   | Plasma | Brain |
|----------|-------|--------|-------|
| n        | 835   | 529    | 380   |
| Age (mean ± s.d.) | 69.4 ± 9.3 | 69.8 ± 9.4 | 83.3 ± 10 |
| Female (%) | 53%   | 54%    | 57%   |
| % CDR=0 | 47.3% | 68.24% | 11.57% |
| APOE e4 (%) | 38%   | 41%    | 48%   |

Characteristics of the baseline cohort after QC, including age, gender, AD status (as clinical dementia rating (CDR)) and APOE e4 allele percentage. For CSF, age denotes age at lumbar puncture; for plasma, age denotes age at plasma draw; for brain, age denotes age at death. Values are reported in years (mean ± s.d.). For basic demographics of the entire cohort before QC, see Supplementary Table 1.
a CSF meta-analysis including the 596 common proteins shared among our study, PPM19 and the Sasayama et al. study (Extended Data Fig. 3c,d). Owing to the increased sample size, we identified 425 pQTLs for 310 proteins, compared to 250 pQTLs for 185 proteins identified in our CSF cohort alone. This represents a nearly two-fold increase in pQTL signals by increasing the sample size by 25%, suggesting that more pQTLs will be identified with larger sample sizes. We observed a similar increase in the number of pQTLs when performing a multi-tissue meta-analysis. For these analyses, we included 342 proteins that passed QC in all three tissue types as well as the PPM19 and Sasayama et al. study. We found 253 pQTLs compared to the 139 that were found in our CSF cohort alone (Extended Data Fig. 3c,d).

Because our study includes cognitively normal older individuals and participants with AD, we performed additional analyses to determine if any of the pQTLs are disease or age specific. To investigate a disease-specific effect, we first included disease status as a covariate. Next, we performed case-only and control-only analyses. Finally, we compared the effect sizes of the genome-wide significant pQTLs from the initial analyses with the beta of these analyses. We found an extremely high correlation (Pearson’s r = 0.98; Extended Data Fig. 4 and Supplementary Tables 8–10), indicating that the associations of the genetic variants with protein levels are not disease specific. To investigate an age-specific effect, we performed separate analyses in participants younger than or older than the average age of our cohort and compared the effect sizes of the genome-wide significant pQTLs from the initial analyses with the beta of these analyses. We found an extremely high correlation (Pearson’s r = 0.98; Supplementary Table 11), indicating that few pQTLs are age specific.

**Pleiotropic loci.** We found that some loci were associated with the levels of more than one protein and up to 13 different proteins in the case of genetic variants in the APOE region. Previous findings reported that the APOE locus is a pleiotropic region using plasma, and we found this in CSF as well. Genetic variants in the APOE gene region were associated with 13 different CSF proteins, including apo E2 and 14-3-3 (Fig. 3d,e, Extended Data Fig. 5 and Supplementary Table 12). The genes encoding these proteins were located on different chromosomes, indicating that this is not just cis regulation. Variants in the APOE locus are the strongest genetic risk factors for AD. Several studies found that the 14-3-3 protein is a marker of non-specific neuronal death22–24, and our results indicated that 14-3-3 protein was also regulated by the APOE locus. For CSF, we found 59 pleiotropic regions where a single locus was associated with two or more proteins. In plasma, we replicated the known pleiotropy of the ABO locus for seven different proteins, including E-selectin (Fig. 3f, Extended Data Fig. 6 and Supplementary Table 13), which was implicated in stroke risk by a recent study25. Additional studies are needed to establish how these genetic variants are associated with multiple proteins. For example, genetic variants in the SPCS3-VEGFC region regulated brain levels of five different proteins, including angiopoietin-1 and growth hormone receptor (Fig. 3g, Extended Data Fig. 7 and Supplementary Table 14). We found an additional 32 pleiotropic regions in plasma and nine in brain.

Several published studies on eQTLs and pQTLs did not identify pleiotropic loci, as they analyzed only cis associations16,26–27. Our results indicate that one protein is regulated in coordination with other independent proteins, which are likely part of the same signaling pathway. To understand the biological processes of health and disease, it is important to identify which proteins are regulated by the same genetic factors. This study identifies tissue-specific pleiotropic effects, highlighting the complex mechanisms that regulate protein levels. Identification of additional tissue-specific cis, trans and pleiotropic regions will lead to discovery of novel pathways relevant to pathogenesis.

**Tissue specificity investigation.** Our unique study design, which included protein measurements in multiple tissues linked to genetic data, enables us to investigate the overlap of the genetic architecture of protein levels across tissues. To identify tissue specificity, we performed mashr analyses on our multi-tissue pQTL results using a P-value threshold of < 0.05 with proteins (n = 411) that passed QC in all three tissues (Fig. 4a). Given a local false sign rate (LFSR) <0.05 and a z-score of at least a two-fold difference, we found that 15–26% of cis-pQTLs were tissue specific (Fig. 4b), whereas 77–91% of trans-pQTLs were tissue specific (Fig. 4c). This analysis indicated that cis-pQTLs were more likely to be shared across tissues than trans-pQTLs. For example, SIG14 only had a cis-pQTL shared across all three tissues (Fig. 4d). We performed additional analyses by comparing the proportion of pQTLs shared across tissues with different P-value thresholds, reaching the same conclusion (Supplementary Materials, Extended Data Figs. 8 and 9 and Supplementary Tables 15–17). CSF and plasma each shared more than 70% of brain pQTLs, suggesting that CSF and plasma were informative tissues for studying brain-related disorders such as AD.

**Functional annotation and biological mechanisms of pQTLs.** Previous studies discovered that most eQTLs are non-coding variants, leading to the hypothesis that most eQTLs modulate transcription factor binding or chromatin structure. However, it remains elusive if this is the case for pQTLs. For this reason, we performed bioinformatic functional annotation and statistical analyses to determine if pQTLs are enriched in specific regions, such as untranslated regions and downstream or upstream of genes, introns, exons, splice sites, non-coding RNA (ncRNA) splice sites, ncRNA_introns, ncRNA_exons or intergenic regions. We found that the strength of the association (−log10 (P value)) for cis signals decreased with distance from the transcription start site (Extended Data Fig. 10a), similar to what has been previously reported for cis-eQTLs 4. This effect was found in all three tissues, suggesting that this is a common biological event. There was an inverse relationship between absolute value of the effect size (beta) and MAF (Extended Data Fig. 10b), which is consistent with previous protein-level GWASs26–29. However, both cis- and trans-pQTLs were strongly enriched for synonymous and non-synonymous exonic variants (odds ratio (OR) = 3.71, 5.25 and 4.19 for CSF, plasma and brain, respectively; Extended Data Fig. 10c and Supplementary Fig. 1). For 42–53% of cis-pQTLs (95 of 226 in CSF, 44 of 97 in plasma and 16 of 30 in brain), the association can be explained by a coding variant, whereas, for cis-eQTLs only 2–5% of signals are located in coding regions. This indicates that pQTLs are significantly enriched.
for coding variants. These results suggest a role for additional regulatory mechanisms (including post-transcriptional changes), as protein levels might not correlate with mRNA levels.

The enrichment of coding variants for pQTLs in cis and trans (Extended Data Fig. 10c and Supplementary Tables 3–5) suggests that protein levels are likely to be regulated post-transcriptionally.
rather than by regulating mRNA level\(^{16}\). cis-pQTLs could lead to changes in protein levels by affecting the signal or the cleavage peptide. For trans-pQTLs, these variants could alter function of a receptor of the protein, affect the machinery that cleaves proteins from the membrane or regulate the function of a gene encoding a transcription factor. This coding variant enrichment observation might be confounded by aptamer-binding effects inherent to the aptamer-based platform. In multiple cases, the most significant signal was a coding variant in a gene that affects protein cleavage or secretion (cis signal, as in the case of IL6R or YKL-40; Extended Data Fig. 10c and Supplementary Table 3), or a coding variant in the receptor of the protein that is likely to modify protein receptor binding (trans signal, as in the case of variants in the APOC4 gene region associated with the BAFF receptor; Extended Data Fig. 10c and Supplementary Table 3). In line with the hypothesis that coding variants have a greater effect size and that pQTLs are enriched for coding variants, we found that pQTLs explained a large proportion of the variation in protein levels. The median variation in protein levels explained by pQTLs (adjusted \(R^2\) calculated using the top variant as the only predictor and again using a nested model accounting for the covariates) was 9–14.9% (interquartile range (IQR): 13.2–15%; Extended Data Fig. 10d). However, there were some extreme cases in which the top pQTL explained more than half of the variability in protein levels. For example, rs2075803 (p.K100E) explained 81% of CSF Siglec-9; rs5498 (p.K469E) explained 74.4% of plasma sICAM-1; and rs5498 (p.K469E) explained 67.4% of brain PPAC (Extended Data Fig. 10d and Supplementary Tables 3–5). CSF Siglec-9, plasma sICAM-1 and brain PPAC were replicated in other studies\(^{16,29,31,32}\) using a different proteomic approach, which indicates that these findings might not be driven by platform.

**Co-localization of pQTLs with other molecular QTLs.** eQTL mapping and co-localization analyses have been instrumental in identifying the functional genes in genetic studies of complex traits\(^{5,33}\). However, it is known that changes in transcript level do not necessarily translate to changes in protein level. To identify
the most likely gene underlying the GWAS signals, it is vital to go beyond eQTL and use other types of molecular QTLs. To determine if pQTLs provide additional information to that of eQTLs, we performed co-localization analyses on pQTLs with eQTLs.

Co-localization analyses indicated that only 13.3–33.3% of pQTLs had a posterior probability of hypothesis-4 (PPH4) > 0.8 for the gTeX cortex eQTLs or 14.2–17.5% for gTeX whole-blood eQTLs (Supplementary Tables 18 and 19). These results were further replicated on the much larger study of blood eQTLs from the eQTlGen consortium (n = 31,684 samples; Supplementary Table 20). In brain, 28% of our pQTLs were co-located in gTEks, which is within a similar scale to the Robins et al. study in which the authors used mass spectrometry-based proteomics to measure 7,901 proteins in 144 samples. This might also be explained by the fact that there are only 323 brain samples in gTEks compared to 380 in this study, which represents the largest brain pQTL analysis performed so far. In plasma, 16% of our pQTLs co-localized with eQTLs, which is similar to previous studies. Previous studies have analyzed the co-localization between pQTLs and eQTLs in CSF.

Although the mapping has been the most common approach to identify the genes driving the GWAS signals, there are other types of molecular QTLs (splicing, DNA methylation or histone acetylation) that could be useful. We also wanted to determine the overlap of pQTLs found in this study with splicing, DNA methylation and histone acetylation QTLs (Fig. 4e). Our co-localization analyses indicated that just 3.1–16.7% of pQTLs were splicing QTLs (Supplementary Tables 21 and 22). Between 1.03% and 10% of pQTLs co-located with DNA methylation QTLs, and fewer than 2% co-located with histone acetylation QTLs (Supplementary Tables 23 and 24). Overall, between 48% (brain) and 76.6% (CSF) of pQTLs identified in this study did not co-localize with QTLs for expression, splicing, DNA methylation or histone acetylation, suggesting that protein level might help explain some missing heritability of disease when integrated with other molecular traits.

MR and drug repositioning. As a large proportion of pQTLs are not eQTLs or other types of QTLs, it is possible that the pQTLs identified in this study can help to identify the gene explaining GWAS signals. Using the MR framework, it is possible to identify functional genes and to prioritize proteins involved with complex traits for further analyses. To identify proteins implicated in AD and other neurodegenerative traits, we performed MR analysis by using pQTLs as instrumental variables and AD risk, progression and onset; Parkinson’s disease (PD); frontotemporal dementia (FTD); PD risk. This locus contains more than one independent signal and likely affects microglia activity and innate immune response. Our analyses also indicated that CD33 is implicated in AD (Supplementary Table 29). CD33 is a microglia-specific gene and likely affects microglia activity and innate immune response. An ongoing clinical trial has been using antibodies targeting CD33 as a potential therapy for AD. In summary, this analysis validated CD33 as a key protein to AD; it is also part of the TREM2-MS4A4A pathway.

Another example is plasma IDUA for PD risk. The IDUA locus is the third most significant locus in the largest GWAS studying PD risk. This locus contains more than one independent signal associated with PD, and it was unclear which gene was functional. We found 42.5% of protein–disease associations co-localized with GWAS loci, including plasma Siglec-3 (also known as CD33) and AD risk (Supplementary Tables 29 and 30).

We found that variants in the CD33 locus, associated with plasma CD33 protein levels, are also known to be associated with AD risk. MR analyses indicated that CD33 is implicated in AD (Supplementary Table 29). CD33 is a microglia-specific gene and likely affects microglia activity and innate immune response. Our analyses also indicated that CD33 is a target of AVE9633, an anti-CD33 antibody used to treat acute myeloid leukemia, and, therefore, this antibody could be used in the context of AD.

The third example is CSF carbonic anhydrase IV for ALS risk, which is supported by MR analysis (FDR = 7.61 × 10⁻⁴; Supplementary Table 27). We further inferred acetazolamide as a potential drug to treat ALS risk. Acetazolamide, used to treat glaucoma, epilepsy and altitude sickness, is known to inhibit carbonic anhydrase.
Moreover, a recent in vitro study supported this drug as a potential treatment for ALS. The fourth example is CSF or plasma E-selectin protein. E-selectin is a known stroke biomarker, and our MR plus co-localization analyses indicated that this protein is genetically associated with the disease (Supplementary Tables 27 and 30). Carvedilol was identified from our drug repositioning analysis and has been reported to provide neuroprotection in animal models.

In summary, our study is useful for mapping additional GWAS loci and identifying drug targets. We integrated our MR results with drug databases to identify potential drugs that could be repurposed for neurological diseases. It is known that compounds targeting a

b

CSF

Plasma

Brain

d

| Tissue | Number of pleiotropic regions (loci with ≥ 2 protein pairs) | Top1 locus ranked by number of unique proteins |
|--------|----------------------------------------------------------|-------------------------------------------|
| CSF    | 59                                                      | APOE-TOMM40 (13 proteins)                |
| Plasma | 34                                                      | ABO and HRG (7 proteins)                 |
| Brain  | 10                                                      | SPCS3-VEGFC (5 proteins)                 |
Fig. 4 | Summary of the tissue-specificity analyses and co-localization of pQTLs with other molecular QTLs. a, Venn diagrams of proteins passing QC across all three tissues. b, Bar plot of tissue specificity percentage inferred from mashr on all cis-pQTLs across all three tissues given a P < 0.05 threshold. c, Bar plot of tissue specificity percentage inferred from mashr on all trans-pQTLs across all three tissues given a P < 0.05 threshold. d, Manhattan plots of the SIG14-chr19:52158316 within each tissue as an example of three-tissue-shared cis-pQTL. The dark red line represents P = 5 × 10⁻⁸. Actual P values (two sided) without multiple comparison adjustments for each variant–protein pair were estimated using an additive linear regression model. e, Upset plots for co-localization investigation on pQTLs versus expression QTLs versus splicing QTLs versus DNA-methylation QTLs versus histone acetylation QTLs for each tissue in cis and (bottom) with the percentage of remaining pQTLs not co-localized.
proteins that are supported by genetic data are more likely to work than those without such support\textsuperscript{16}. This study goes beyond AD or PD, and the data generated here can be applied to other traits.

**Limitations.** Our study has several limitations. An important note is that we used an aptamer-binding platform to profile proteomics, which might be confounded by recognition difference instead of real protein abundance change, as already addressed by Sun et al\textsuperscript{10}. Mass spectrometry can be a complementary platform\textsuperscript{30} that will not be affected by this problem. Our observation of an enrichment of coding variants in pQTLs could be largely biased by the platform. However, we detected 83–90\% of our non-coding pQTL and 10–17\% coding pQTL, respectively, depending on tissues, in a mass spectrometry-based pQTL study\textsuperscript{14} (Supplementary Table 32). The percentages for coding variants were similar (proportional test: \( P > 0.05 \) in all three tissues), indicating that this enrichment was not exclusively due to the bias of a single technology.

Our data suggest that pQTLs are located in coding regions more often than are eQTLs, and our results could be biased, because coding pQTLs have stronger effects than non-coding pQTLs. Therefore, non-coding pQTLs might remain undetected due to a lack of statistical power. However, the sample size of our study is similar to or larger than GTEx for brain-relevant tissues, and eQTL studies from GTEx do not see the same amount of coding variant enrichment\textsuperscript{5}. Our data suggest that around half of pQTLs do not co-localize with other molecular QTLs, but these data cannot be extrapolated to all coding genes. GTEx and eQTLGen include eQTLs for more

---

**Table 2 | Number of significant protein–trait associations from MR analyses**

| Outcome                  | CSF | Plasma | Brain |
|--------------------------|-----|--------|-------|
| AD risk                  | 3   | 13     | 7     |
| AD progression           | 18  | 25     | 6     |
| AD age at onset          | 8   | 2      | 20    |
| PD risk                  | 13  | 15     | 35    |
| ALS risk                 | 1   | 4      | 3     |
| FTD risk                 | 1   | 8      | 5     |
| Stroke risk              | 7   | 8      | 24    |
| Asthma risk (non-neuro)  | 14  | 4      | 2     |

Within each tissue, the table contains the number of significant proteins (false discovery rate (FDR) \( < 0.05/24 \)) and associations with the seven neurological traits: (1) AD risk; (2) AD progression; (3) AD age at onset; (4) PD risk; (5) ALS risk; (6) FTD risk; and (7) stroke risk and a non-neurological trait: asthma risk.
than 34,000 RNA molecules, but not all of those RNA molecules code for proteins or are expressed in all tissues. It is clear that we cannot anticipate an eQTL/pQTL co-localization for a large proportion of coding genes. Another reason for a potential lack of eQTL/pQTL co-localization in brain is because our brain pQTLs are derived from parietal lobe cortex, and the brain eQTLs from GTEx are from whole cortex. The gene regulation might be region specific and, thus, complicates the overlapping between two QTL types. The small proportion of overlap between eQTL and pQTL data can also be caused by measurements of mRNA expression and protein level in different cohorts, sample sizes or other technical details.

A standard limitation of MR analyses is the inability to fully account for horizontal pleiotropy, although confidence in our results might be bolstered by the fact that we performed MR after removing all pleotropic regions reported in this study within each tissue. Another limitation of MR interpretation is that our observation on associations between proteins and neurological diseases might not be disease specific, as we observed several associations (Table 2 and Supplementary Table 33) between proteins and asthma risk, an example of non-neurological diseases.

Additional studies performed in more tissues and using approaches that include an even larger number of proteins are necessary to better understand the genetic architecture of protein levels and to replicate these findings. To our knowledge, this is the first study to address the single-tissue bias when comparing pQTLs to results of multi-tissue eQTLs.

Discussion

In this study, we generated a detailed map of multi-tissue pQTLs that will be crucial for understanding the tissue-specific genetic architecture of protein levels. By leveraging these data, we have been able to map some additional GWAS loci, identify additional causal proteins implicated in disease pathogenesis, identify novel biomarkers and reposition drugs. To fully understand the genetic architecture of complex traits, it is necessary to understand the genetic architecture of protein levels. Until now, multi-tissue pQTL mapping has been constrained by the limited availability of large-scale proteomic analyses in multiple tissues. We present the largest brain and CSF pQTL analyses to date, the first neurologically relevant multi-tissue pQTL study and a unique resource for leveraging multi-tissue pQTL to understand neurological traits. These data can be further used to perform MR analysis on other complex traits with available GWASs. All results can be downloaded from https://www.niagads.org/datasets/ng00102 and interactively accessed through a PhWeb43-based website, the Online Neurodegenerative Trait Integrative Multi-Omics Explorer (ONTIME; https://ontime.wustl.edu/).

To achieve personalized medicine, these results highlight the need to implement additional functional genomic approaches beyond gene expression toward understanding the biology of complex traits and to identify potential drug targets for those traits. We predict that this and other studies including additional omic layers (for example, epigenomics and metabolomics) will be instrumental in advancing the field.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00886-6.

Received: 1 September 2020; Accepted: 3 June 2021; Published online: 8 July 2021

References

1. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. Science 322, 881–888 (2008).
2. Morris, A. P. et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat. Genet. 44, 981–990 (2012).
3. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. 51, 414 (2019).
4. Clausenbütter, M. et al. A brief history of human disease genetics. Nature 577, 179–189 (2020).
5. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330 (2020).
6. van der Wijst, M. G. P. et al. The single-cell eQTLGen consortium. eLife 9, e52155 (2020).
7. Aguet, F. et al. Genetic effects on gene expression across human tissues. Nature 550, 204–213 (2017).
8. Gamazon, E. R., Zwierd Mister, A. H., Cox, N. J., Denys, D. & Derks, E. M. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. Nat. Genet. 51, 933–940 (2019).
9. Yósa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. Preprint at bioRxiv https://doi.org/10.1101/447367 (2018).
10. Sun, B. B. et al. Genomic atlas of the human plasma proteome. Nature 558, 73–79 (2018).
11. Sasaki, K. et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun. 8, 14357 (2017).
12. Folktersen, L. et al. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. 13, e1006706 (2017).
13. Deming, Y. et al. Genetic studies of plasma analytes identify novel potential biomarkers for several complex traits. Sci. Rep. 6, 18092 (2016).
14. Sasayama, D. et al. Genome-wide quantitative trait loci mapping of the human cerebrospinal fluid proteome. Hum. Mol. Genet. 26, 44–51 (2017).
15. Kauwe, J. S. K. et al. Genome-wide association study of CSF levels of 59 Alzheimer’s disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation. PLoS Genet. 10, e1004758 (2014).
16. Robins, C. et al. Genetic control of the human brain proteome. Preprint at bioRxiv https://doi.org/10.1101/816652 (2019).
17. Gold, L. et al. Apteram-based multiplexed proteomic technology for biomarker discovery. PLoS ONE 5, e15004 (2010).
18. Haddick, P. C. G. et al. A common variant of IL-6R is associated with elevated IL-6 pathway activity in Alzheimer's disease brains. J. Alzheimer's Dis. 56, 1037–1054 (2017).
19. Marek, K. et al. The Parkinson Progression Marker Initiative (PPMI). Prog. Neurobiol. 95, 629–635 (2011).
20. Lovestone, S. et al. AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Ann. N. Y. Acad. Sci. 1180, 36–46 (2009).
21. Kamat, M. A. et al. PhenoScanner V2: an expanded tool for searching human genotype–phenotype associations. Bioinformatics 35, 4851–4853 (2019).
22. Jayaratnam, S., Khoo, A. K. L. & Basir, D. Rapidly progressive Alzheimer's disease and elevated 14-3-3 proteins in cerebrospinal fluid. Age Ageing 37, 467–469 (2008).
23. Foote, M. & Zhou, Y. 14-3-3 proteins in neurological disorders. Int. J. Biochem. Mol. Biol. 3, 152–164 (2012).
24. Ibanez, L. et al. Overlay in the genetic architecture of stroke risk, early neurological changes, and cardiovascular risk factors. Stroke 50, 1339–1345 (2019).
25. Lourdesamy, A. et al. Identification of cis-regulatory variation influencing protein abundance levels in human plasma. Hum. Mol. Genet. 21, 3719–3726 (2012).
26. Walker, R. L. et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. Cell 179, 750–771 (2019).
27. Orozco, L. D. et al. Integration of eQTL and a single-cell atlas in the human eye identifies causal genes for age-related macular degeneration. Cell Rep. 30, 1246–1259 (2020).
28. Urbot, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet. 51, 187–195 (2019).
29. Hillary, R. F. et al. Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. Nat. Commun. 10, 3160–3160 (2019).
30. Shuere, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. Nat. Rev. Genet. 22, 19–37 (2020).
31. Yao, C. et al. Genome-wide mapping of plasma protein QTLs identifies putative causal genes and pathways for cardiovascular disease. Nat. Commun. 9, 3268 (2018).
32. Paré, G. et al. Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet.* **4**, e1000118 (2008).

33. Ndungu, A., Payne, A., Torres, J. M., van de Bunt, M. & McCarthy, M. I. A multi-tissue transcriptome analysis of human metabolites guides interpretability of associations based on multi-SNP models for gene expression. *Am. J. Hum. Genet.* **106**, 188–201 (2020).

34. Cruchaga, C. et al. Cerebrospinal fluid APOE levels: an endophenotype for genetic studies for Alzheimer's disease. *Hum. Mol. Genet.* **21**, 4558–4571 (2012).

35. Kibinge, N. K., Relton, C. L., Gaunt, T. R. & Richardson, T. G. Characterizing the causal pathway for genetic variants associated with neurological phenotypes using human brain-derived proteome data. *Am. J. Hum. Genet.* **106**, 885–892 (2020).

36. Del-Aguila, J. L. et al. A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. *Alzheimer's Res. Ther.* **11**, 71 (2019).

37. Alector Inc. First in human study for safety and tolerability of AL003. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT03822208 (2019).

38. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

39. Bethea, J. W. *Clinical Anesthesia*, 6th Edition. *Anesthesiology* **112**, 767–768 (2010).

40. Camerino, G. M. et al. Elucidating the contribution of skeletal muscle ion channels to amyotrophic lateral sclerosis in search of new therapeutic options. *Sci. Rep.* **9**, 3185 (2019).

41. Savitz, S. I. et al. The novel beta-blocker, carvedilol, provides neuroprotection in transient focal stroke. *J. Cereb. Blood Flow. Metab.* **20**, 1197–1204 (2000).

42. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

43. Gagliano Taliun, S. A. et al. Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat. Genet.* **52**, 550–552 (2020).

44. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

45. Del-Aguila, J. L. et al. Assessment of the genetic architecture of Alzheimer's disease risk in rate of memory decline. *J. Alzheimers Dis.* **62**, 745–756 (2018).

46. Huang, K. et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nat. Neurosci.* **20**, 1052–1061 (2017).

47. van Rheenen, W. et al. Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).

48. Ferrari, R. et al. Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* **13**, 686–699 (2014).

49. Malik, R. et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524 (2018).

50. Demenais, F. et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50**, 42–53 (2018).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021
Apptmer-based proteinomics data sample collection. This study included 1,537 participants from Washington University School of Medicine in St. Louis. All participants provided informed consent to allow their data and biospecimens to be included. The study was approved by an institutional review board at Washington University School of Medicine in St. Louis. Samples from participants include three tissue types: CSF; plasma and brain (parietal lobe cortex). CSF samples were collected the morning after an overnight fast, processed and stored at −80 °C. Plasma samples were collected the morning after an overnight fast, immediately centrifuged and stored at −80 °C. Brain tissues (−500 mg) were collected from fresh frozen human parietal lobes.

For CSF tissue, there were 971 unique participants (including 249 patients with AD, 127 cognitively normal controls and 134 with unknown or other status (for example, VTD or other neurological diseases)) and 459 samples in total. Age is distributed with a mean of 69.9 years and an s.d. of 9.3 years, and 53% of the samples are from women.

For plasma tissue, there were 636 unique participants (including 230 patients with AD, 401 cognitively normal controls and five with unknown status) and 648 samples in total. Age is distributed with a mean of 70.4 years and an s.d. of 9.8 years, and 56% of the samples are from women.

For brain tissue, there were 458 unique participants (including 297 patients with AD, 27 cognitively normal controls and 134 with unknown or other status (for example, VTD or other neurological diseases)) and 459 samples in total. Age is distributed with a mean of 82.2 years and an s.d. of 12 years, and 60% of the samples are from women.

The donor overlap across three tissues are shown as a Venn diagram in Extended Data Fig. 2b: nine donors were shared across all three tissues; 481 donors were shared by both CSF and plasma; 29 donors were shared by plasma and brain; and 481, 117 and 420 were exclusively for CSF, plasma and brain, respectively. The Venn diagram was drawn using the VennDiagramR package (v1.6.20).

These recruited participants were evaluated by clinical personnel from Washington University. AD severity was determined by the Clinical Dementia Rating scale at the time of lumbar puncture (LP) (LP for CSF samples) or blood previously described by Gold et al.17. In brief, modified single-stranded DNA aptamer-based technology was highly reproducible.

After QC, we kept 1,079 proteins and 435 samples. Of these 435 samples, only one replicate of the participant outliers remained. Thus, we compared two technical replicates using the values before QC across all 1,305 proteins. The overall Pearson correlation coefficient between fasted and non-fasted samples is 0.907, with a 95% confidence interval from 0.904 to 0.911 (Extended Data Fig. 2a,d).

For plasma samples, we included one participant with two biological replicates: one collected in 1997, the other in 2007. Both samples passed QC. The overall Pearson correlation coefficient between these two biological replicates is 0.938, with a 95% confidence interval from 0.929 to 0.945 (Extended Data Fig. 2a,e).

For brain samples, we included one participant with two technical replicates. After QC, we kept 1,079 proteins and 435 samples. Of these 435 samples, only one replicate of the participant outliers remained. Thus, we compared two technical replicates using the values before QC across all 1,305 proteins. The overall Pearson correlation coefficient between these two replicates is 0.976, with a 95% confidence interval from 0.976 to 0.9812 (Extended Data Fig. 2a,f).

We processed the proteinomics data using SonomaDataIO (v1.8.0) and Biobase (v2.42.0). Proteins were mapped to UniProt® identifiers and Entrez Gene symbols. Mapping to Ensemble gene IDs and genomic positions (start and end coordinates) was performed using gencode version 30 liftover to hg19/GRCh37.

Reproducibility investigation via comparisons between biological or technical replicates. To measure the reproducibility of the aptamer-based platform, we compared the replicates for the same participant given each tissue.

For plasma samples, we included 11 participants with two measures, one as fasted and the other as non-fasted. After QC, we kept 931 proteins in 633 samples. The overall Pearson correlation coefficient between fasted and non-fasted samples is 0.907, with a 95% confidence interval from 0.904 to 0.911 (Extended Data Fig. 2a,d).

For CSF samples, we designed 329 participants with two measures, one as baseline (LP date 1) and the other as longitudinal (LP date 2). After QC, we kept 713 proteins and 1,270 samples. Of these 1,270 samples, 321 participants with two measures remained in the analysis (Extended Data Fig. 2a,c). The average time difference between the two LP dates was 6.14 years, and an s.d. was 2.98 years. The overall Pearson correlation coefficient between two LP dates was 0.995, with a 95% confidence interval from 0.99518 to 0.99526 (Extended Data Fig. 2c).

The overall high correlations within each tissue indicated that the aptamer-based technology was highly reproducible.

Genomic data QC process. Most of the samples with proteinomic profiling were collected with genotyping data (Extended Data Fig. 1d). For CSF, 965 of 971 unique participants had both genotype and proteomic data. For plasma, 633 of 636 unique participants had both genotype and proteomic data. For brain tissue, 450 of 458 unique participants had both genotype and proteomic data.

Samples were genotyped on multiple genotyping platforms from Illumina. Stringent quality thresholds were applied to the genotype data for each platform separately. SNPs were kept if they passed all of the following criteria: (1) genotyping successful rate ≥ 98% per SNP or per individual; (2) MAF ≥ 0.01; and (3) Hardy-Weinberg equilibrium (HWE) (P ≥ 1 × 10−6).

After removing low-quality SNPs and samples, genotype imputation was performed using the Impute2 program18 with haplotypes derived from the 1000 Genomes Project (released June 2012). Genotype imputation was performed separately based on the genotype platform used. SNPs with an info-score quality of less than 0.3 reported by Impute2, with a MAF < 0.02 or out of HWE were removed. After imputation and QC, the different imputed PLINK files were merged. A total of 1,405,245 imputed and directly genotyped SNPs and 1,530 individuals were used for final analyses. To adjust for population substructure (Extended Data Fig. 1d), principal component analysis (PCA) was conducted using the PLINK1.9 (v1.90b6) subcommand pca. HapMap samples (CEU; Caucasian Europeans from Utah; JPT; Japanese in Tokyo; YRI; Yoruba in Ibadan, Nigeria) were included in the analyses to remove outliers and confirm self-reported ethnicity. Samples within the CEU cluster were kept. To identify unanticipated duplicates and cryptic relatedness using pairwise genome-wide estimates of proportion identity by descent (IBD) (Extended Data Fig. 1d), we used the
subcommand IBD from PLINK1.9 (v1.0b6.4)10. When duplicate samples or a pair of samples with cryptic relatedness (PL_HAT ≥ 0.5) were identified, we removed one sample from the pair. A total of 835 CSF, 529 plasma and 380 brain samples passed filters on both genomic and proteomic QC.

**pQTL identification.** To test for the association between genetic variants and protein levels, we performed a linear regression (additive model), including age, sex, principal component (PC) factors from population stratification and genotype platform as covariates:

\[
\log_{10}(\text{Protein Level}) \sim \beta_0 + \beta_1 \times \text{SNP dosage} + \\
\beta_2 \times \text{age} + \beta_3 \times \text{gender} + \beta_4 \times \text{PC1} + \beta_5 \times \text{PC2} + \\
\sum \beta_j \times \text{genotype Platform} + \epsilon
\]

**cis-pQTL mapping.** We conducted cis-pQTL mapping within each of the three tissues. Only proteins passing QC were included in the analyses. Protein level was log10 transformed to approximate the normal distribution. For these tests, data distribution was assumed to be normal, but this was not formally tested. Within each tissue, cis-pQTLs were identified by linear regression, as implemented in PLINK2 (v2.00a2LM)60, adjusting for sex, age, the first two genotype-based PCs (genetically very homogeneous) and genotyping platforms (for example, Omni1, Omni2.5 and NeuroX). We restricted our search to variants within 1 Mb upstream and downstream of the gene start site by which each protein is coded. Actual P values for each variant–protein pair were estimated using an additive linear regression model. To identify the list of all significant variant–protein pairs, variants with an actual P value below the genome-wide significance (5 × 10−8) level were considered significant and included in the final list.

**trans-pQTL mapping.** PLINK2 (v2.00a2LM)60 was used to test all autosomal variants (MAF ≥ 0.02) using the same QC pipelines as cis-pQTL mapping but was restricted to variants and proteins encoded by the genes locating outside the 2-Mb window in each tissue independently using an additive linear model. For trans-pQTL mapping, we tested variants or association with the same protein list as for cis-pQTL mapping. We included the covariates of the first two genotype PCs (genetically very homogeneous), age, sex and genotyping platforms when performing association tests. The correlation between variant and protein levels was evaluated using an additive linear regression model. For these tests, data distribution was assumed to be normal, but this was not formally tested. To identify the list of all significant variant–protein pairs, variants with an actual P value below the genome-wide significance (5 × 10−8) level were considered significant and included in the final list.

**Disease-specific analyses.** To investigate a disease-specific effect on pQTLs, we performed linear regression on the same protein–loci pairs (before conditioning on top variants) identified from the above default model using three additional models: (1) joint analysis including disease status as another covariate (CO versus non-CO); (2) AD case (CA) only using the same covariates as the default model; and (3) cognitive unimpaired (CO) only using the same covariates as the default model. Using scatter plots, we visualized the correlation between each of the additional models and our default model. Using model 1 for comparison, we observed a Pearson correlation coefficient of 0.999, 0.999 and 0.999 for CSF, plasma and brain, respectively. Using model 2 for comparison, we observed a Pearson correlation coefficient of 0.991, 0.989 and 0.998 for CSF, plasma and brain, respectively. Using model 3 for comparison, we observed a Pearson correlation coefficient of 0.999, 0.998 and 0.602 (P = 0.002) for CSF, plasma and brain, respectively. The relatively low correlation seen when using model 3 for comparison with controls only in brain samples was due to a much smaller sample size.

**Age-specific analyses.** To confirm that none of the findings was age specific, we performed separate analyses in participants younger and older than the average age of our cohort and compared the β1 of SNP dosage for all significant pQTLs to identify any age-specific effect.

**Meta-analyses on pQTLs.** Meta-analysis was also performed on all significant pQTLs from round_0 using PLINK2 (v2.00a2LM)60 with the --condition or --condition-list option. We used the same significance threshold of P = 5 × 10−8 used for the univariable analysis on identifying independent local pQTLs within a window size of 2 Mb.

Conditional analyses were performed as follows. Before conditioning (row-1), no SNPs were used as a covariate given one region. For round_1 (row-2) conditioning, the top SNP from the before-conditioning stage and the top SNP from round_1 stage were used as additional covariates in the default model. Both SNPs were within the same region. This iteration was continued for each round by adding one more top SNP from the prior round until no variants passed the genome-wide significance threshold given the same region. For CSF and plasma samples, in total, four rounds of conditional analyses were performed. For brain samples, three rounds of conditional analyses were performed. The results were visualized using LocusZoom version 1.3 (ref. 33).

**Replication strategy for CSF pQTL.** To identify all previously reported pQTLs from large-scale protein-level GWASs, we performed CSF replication analyses. We first searched the recent pQTL work by Sasayama et al. (SOMAscan-based, CSF); PPMI19 (SOMAscan-based, CSF); and a meta-analysis of these two previous studies. Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pQTL studies.

**Reprocessing pQTL using Sasayama et al. CSF SOMAscan individual data.** To replicate CSF pQTLs, we performed linear regression on all proteins using the individual-level proteomic and genotype data from Sasayama et al.16. We decided to reprocess the pQTL analyses because the original studies used unimputed genotype data. We performed imputation in the Sasayama et al. study using the same SOMAscan platform. We used a similar number of SNPs across studies. For proteomics QC, only the IQR strategy was used, as neither calibrator nor positive/negative control values were provided. QC Positive is a technical sample provided with the SOMAscan platform for use as a positive control. Similarly, QC Negative is a technical sample used as a negative control. We performed replication analyses with the additional covariates of age and the IQR strategy, including participants older and younger than 75 years.

**Reprocessing pQTL using PPMI19 CSF SOMAscan data.** To replicate CSF pQTLs, we performed linear regression on all 709 shared proteins using the proteomic and genotype data from the PPMI cohort19. We performed protein QC, genotype imputation and QC and analyses using the same protocols as those used for the data generated in this study and described above. A total of over 7 million (5,187,563) imputed and directly genotyped SNPs and 154 individuals were used for final analyses. Population substructure analyses were performed as described above, except samples kept were within the IPR cluster. A total of 132 CSF samples from the study by Sasayama et al. passed filters on both genomics and proteomics QC. We performed linear regression (additive model), including sex and first two PC factors from population stratification as covariates.
pGWAS findings from EBI-NHGRI using PhenoScannerV2 with proxy SNPs ($r^2 > 0.5$).

Reprocessing pQTLs using AddNeuroMed plasma SOMAscan data. To replicate plasma pQTLs, we performed linear regression on all proteins using the proteomic and genotype data from the AddNeuroMed consortium. A total of over 7 million (7,313,640) imputed and directly genotyped SNPs and 343 plasma samples from the study by AddNeuroMed passed filters on both genomics and proteomics QC. We performed a linear regression (additive model), including age, gender, first two PC factors from population stratification, centers, status, visit cohorts and batch effects as covariates.

Replication strategy for brain pQTLs. To identify all previously reported pQTLs from large-scale protein-level GWASs, we performed the brain replication analyses using the following strategies. We first searched the processed pQTL results using results from the published brain findings8. Next, we queried all plasma pGWAS findings from EBI-NHGRI using PhenoScanner V2 and from our CSS findings.

For each locus, we investigated whether the sentinel SNP or a proxy ($r^2 > 0.5$) was associated with the same target protein (or aptamer) in our study at different defined significance thresholds. For the known category in our primary assessment, we used a $P$-value threshold of $5 \times 10^{-8}$. For the replicated category in our primary assessment, we used a $P$-value threshold of $5 \times 10^{-7}$.

Identification of tissue-specific/shared pQTLs. We first performed mash statistical analysis of our multi-tissue pQTL results given 411 proteins shared in gene-based annotation mode. Genomic features and variants affecting the nearest genes were used for downstream analyses. The bar plots were drawn using the ggplot2 (ref. 5) R package.

Testing for genomic feature enrichment. We used Fisher's exact test (two sided) for testing the enrichment. The null set was set using $P > 5 \times 10^{-08}$ and permuted with the same amount of variants as the positive set. The various groupings were determined by ANNOVAR version 2018-04-16. We also tested whether our pQTLs were enriched for functional and regulatory characteristics using GARFIELD v2 (ref. 6). GARFIELD is short for GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction, and it is a method to test feature enrichment by integrating GWAS findings and a large set of regulatory or functional annotations (1,005 features in total), mainly from ENCODE and Roadmap epigenomics consortia. It takes into account the LD while annotating and calculating enrichments. The enrichment is quantified as OR.

Identification of pleiotropic regions. To identify unique non-overlapping regions associated with a given aptamer, we first defined a 2-Mb region 1 Mb upstream and 1 Mb downstream of each significant variant for that aptamer. Within the 2-Mb region containing the variant with the smallest $P$ value, any overlapping regions were then grouped into the same locus. Owing to the complexity of the major histocompatibility region, we assigned a genetic region spanning from 25.5 Mb to 34.0 Mb on chromosome 6 as one region. To identify whether a region was associated with multiple aptamers, we next used an LD-based clumping approach (LD block from the 1000 Genomes Project implemented into the RHOGED R package (v0.1), as we also used the 1000 Genomes Project as our imputation reference panel). Variants were collapsed into a single region per LD (EUR)-defined loci. Any loci associated with more than one protein were identified as pleiotropic regions. The cytoband ID was also annotated. Circos plots were generated using functions from the R package circlize7. To measure variance explained by each pleiotropic region, we also calculated the adjusted $R^2$ value (Supplementary Figs. 5–8).

Performing MR using TwoSampleMR R package. MR is a method of using measured variation in genes of known function to examine the causal effect of a modifiable exposure on disease. This method obtains unbiased estimates of the effects of a putative causal variable without conducting a traditional randomized trial. We used the R package TwoSampleMR version 0.4.22 (ref. 8). For each single SNP that remained after clumping, the most basic method, the Wald ratio, was used. This package also implements the harmonization steps before performing MR, and these steps are: (1) correcting the wrong effect/non-effect alleles; (2) correcting the strand issues; (3) fixing the palindromic SNPs; and (4) removing the SNPs with incompatible alleles. The SNPs selected for the analysis were based on a suggestive threshold of $1 \times 10^{-5}$. The beta coefficients and standard errors for the selected variants (pQTLs) from this study were used as input of instrumental variables. These instrumental variables were also extracted from the summary statistics from the latest GWAS for the outcome on neurological disease-related traits. (For details, see Supplementary Table 26. Briefly, the AD-risk GWAS was published in 2019 (ref. 9); the AD-progression GWAS was published in 2018 (ref. 10); the AD-age at onset GWAS was published in 2017 (ref. 11); the PD-risk GWAS was published in 2019 (ref. 12); the ALS-risk GWAS was published in 2016 (ref. 13); the T2D-risk GWAS was published in 2014 (ref. 14); and the stroke-risk GWAS was published in 2018 (ref. 15).)

To check the specificity of protein-neurological disease associations, we also chose the asthma-risk GWAS as an outcome of non-neurological disease. To test that the directionality of exposure causing outcome is valid, we used the directionality test function from the same R package. The method confirms whether the exposure (protein) and outcome (trait) directions are correct.

Co-localization analyses. We performed Bayesian co-localization analysis using the coloc.abf function from the coloc R package (ref. 16) version 3.1. We used the default priors with $P_1 = 1 \times 10^{-4}$ and $P_2 = 1 \times 10^{-6}$. Evidence for co-localization was assessed using the PHH4 (indicating that there is an association for both protein and disease and that they are driven by the same causal variant(s)). We used PHH4 $> 0.8$ as a threshold to suggest that associations were highly likely to co-localize.

For co-localization of pQTLs with disease status. We downloaded and used the full GWAS summary statistics for each disease/trait from their original publications as the same for MR analysis.

For co-localization of cis-pQTLs with cis-eQTLs, cis-sQTLs from GTEx v8 release. We downloaded and used the significant cis-eQTLs and cis-sQTLs summary statistics for two single tissues—cortex and whole blood—from GTEx (https://gtexportal.org/home/datasets). For cis-sQTLs, we used gene-level sQTL results rather than exon-level sQTLs.

For co-localization analysis of plasma pQTLs with eQTLs from qTLGen. We downloaded and used the significant cis- and trans-eQTL summary statistics for blood from eQTLGen (https://www.eqtlgen.org/index.html). In both cases, we analyzed cis- and trans-QTLs.

For co-localization of cis-pQTLs with cis DNA methylation qTLs and cis histone acetylation qTLs from ROSMAP. We downloaded and used the significant DNA methylation qTL summary statistics for brain tissue from ROSMAP (http://mostafavilah.stat.ubc.ca/qTLServer/). We downloaded the significant histone acetylation qTL summary statistics (assigning to up to 10 Mb upstream of the transcription start site given the same gene) for brain tissue from ROSMAP as well. To ensure that DNA methylation qTLs affecting pQTLs are mediated by eQTLs, we further subset the DNA methylation qTLs/pQTLs co-localization result with the eQTLs/pQTL co-localization result.

For co-localization of cis-pQTLs with cell-type-specific cis-eQTLs from ROSMAP. We identified the neuron, oligodendrocyte, microglia and astrocyte eQTL data using a pseudo-bulk strategy on single-nucleus RNA sequencing (n = 48) from ROSMAP data. In total, we recreated the expression matrices on five cell types (microglia, excitatory neurons, inhibitory neurons, oligodendrocytes and astrocytes). We then identified cis-eQTLs for each cell type using FastQTLv2.0 (ref. 17) after integrating with the whole-genome sequencing data from ROSMAP (n = 39). Using both the neuron and microglia and to the common environmental and genotypic effects, we followed the significant Gene calling approach from the GTEx pipeline. We used different priors because the pseudo-bulk-derived cell-type-specific eQTLs were underpowered compared to bulk-level pQTLs with $P_1 = 1 \times 10^{-4}$, $P_2 = 1 \times 10^{-2}$ and $P_3 = 1 \times 10^{-0.2}$.

Results can be found in Supplementary Fig. 9 and Supplementary Table 35.

Overlapping of proteins with pQTLs and drug targets. To obtain information on drugs that target proteins with pQTLs from this study, we used the DrugBank database (as of 3 January 2020). This is a manually curated database that maintains profiles for more than 15,000 drugs (including Food and Drug Administration-approved and experimental drugs). For our analysis, we focused on the protein target for each drug. For each protein assayed, we identified all drugs in the DrugBank with a matching protein target based on UniProt ID, annotated via https://www.uniprot.org/database/DB-0019. We further integrated the MR results on proteins as drug targets.

Randomization. Data collection for the proteomics was randomized. All samples were randomly assigned to a profiling pool before proteomics measurement, considering age, sex and disease status. Data collection for the genotypic was not randomized or blocked.

Sample sizes. No statistical methods were used to predetermine sample sizes within each tissue type, but our sample sizes are much larger than or similar to those in previous publications (supplementary references).

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

51. Chen, H. VennDiagram: generate high-resolution Venn and Euler plots. https://rdrr.io/cran/VennDiagram/ (2018).
52. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. Neurology 43, 2412–2414 (1993).
53. Mirra, S. S. et al. The Consortium to Establish a Registry for Alzheimer’s Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer’s disease. Neurology 41, 479–486 (1991).
54. Khachaturian, Z. S. Diagnosis of Alzheimer’s disease. Arch. Neurol. 42, 1097–1105 (1985).
55. Sattlecker, M. et al. Alzheimer’s disease biomarker discovery using SOMAscan multiplexed protein technology. Alzheimers Dement. 10, 724–734 (2014).
56. Williams, S. A. et al. Plasma protein patterns as comprehensive indicators of health. Nat. Med. 25, 1851–1857 (2019).
57. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12, 115–121 (2015).
58. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 47, D506–D515 (2019).
59. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
60. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4, 7 (2015).
61. Pruim, R. J. et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26, 2336–2337 (2010).
62. Miller, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
63. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38, e164 (2010).
64. Wickham, H. ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag, 2009).
65. Iotchkova, V. et al. GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. Nat. Genet. 51, 343 (2019).
66. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. Am. J. Hum. Genet. 100, 473–487 (2017).

Acknowledgements

We thank all the participants and their families as well as the many involved institutions and their staff. Funding: This work was supported by grants from the National Institutes of Health (NIH) (R01AG044546 (C.C.), P01AG03991 (C.C. and J.C.M.), R01AG053303 (C.C.), R01AG058591 (C.C.), U01AG058922 (C.C.), R01NS118146 (B.A.B.) and R01AG057777 (O.H.)) and the Alzheimer Association (NIBR-11-20110 (C.C.), BAND-14-338165 (C.C.), AARG-16-441560 (C.C.) and BFG-15-362540 (C.C.)). This work was supported by access to equipment made possible by the Hope Center for Neurological Disorders and the Departments of Neurology and Psychiatry at Washington University School of Medicine. The recruitment and clinical characterization of research participants at Washington University were supported by NIH P50AG05681 (J.C.M.), P01AG03991 (J.C.M.) and P01AG026276 (J.C.M.).

Author contributions

C.Y. performed the analyses, interpreted the results and wrote the manuscript. E.H.G.F., L.I., M.V.F., E.W., J.L.R., Z.L., U.D., Y.S., K.M. and J.P.B. contributed to data collection, data processing, quality control and cleaning. J.C.M., A.M.F. and R.J.P. contributed samples and/or data. B.S. wrote the manuscript. I.A.R., B.E. and O.H. developed the PheWeb browser. B.A.B. interpreted the results. H.R., O.H. and C.C. designed the study, collected the data, supervised the analyses, interpreted the results and wrote the manuscript. C.Y., A.S. and C.C. addressed the comments from peer review and updated the manuscript. All authors read and contributed to the final manuscript.

Competing interests

C.C. receives research support from Biogen, Eisai, Alzceutics and Parabon. C.C. is a member of the advisory board of Vivid Genomics, Halia Therapeutics and ADx Healthcare. The remaining authors declare no competing financial interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00886-6.
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00886-6.
Correspondence and requests for materials should be addressed to C.C.
Peer review information Nature Neuroscience thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.
Extended Data Fig. 1 | QC pipeline. QC on both proteins (a to c) and samples (d) were described as follows: a, Flowchart of CSF protein level QC, starting from 1305; after step-1, Limit Of Detection VS 2-StDeviation, 807 proteins were kept with a pass-rate \( > 85\% \); after step-2, given Max Difference of Scale Factor \( < 0.5 \), 749 proteins were kept; after step-3, given Coefficient of Variation (of calibrator) \( < 0.15 \) & step-4, given IQR, sum(outliers) \( < 15\% \), 746 proteins were kept. After step-5, 713 proteins that shared by \( < 30 \) samples (shared by \( \sim 80\% \) of the subject outliers) were kept. b, Flowchart of plasma protein level QC, starting from 1305; after step-1, 1301 proteins were kept with a pass-rate \( > 85\% \); after step-2, 956 proteins were kept; after step-3 & step-4, 955 proteins were kept. After step-5, 931 proteins that shared by \( < 10 \) samples were kept. c, Flowchart of brain protein level QC, starting from 1305; after step-1, 1109 proteins were kept with a pass-rate \( > 85\% \); after step-2, 1107 proteins were kept; after step-3 & step-4, given IQR, sum(outliers) \( < 15\% \), 1106 proteins were kept. After step-5, 1079 proteins that shared by \( < 21 \) samples were kept. d, Table of sample size after each step of QC in genotype and proteomics. Within each tissue (1st column), we profiled proteomics from 1300 CSF, 648 plasma and 459 samples (2nd column). From unique donors in proteomics data (3rd column), we first kept donors with genotyping array data (4th column). We next kept only the donors with a European ancestry after checking principal components (5th column). Moreover, we kept donors that were not close with each other (PI_HAT \( < 0.05 \)) after checking identity by descent (6th column). Finally, the samples remained only passing both the genotype and protein data QC (7th column).
Extended Data Fig. 2 | Reproducibility of proteomic data. a, Table of total sample size for each tissue before and after QC, including the biological and technical replicates. b, Venn diagram on the designed donor overlap across tissues. c, Scatterplot of 321 subjects with both longitudinal and baseline samples from CSF indicates a Pearson correlation coefficient of 0.995 (95% confidence interval from 0.995 to 0.995). d, Scatterplot of 11 subjects with both fasted and nonfasted samples from plasma indicates a Pearson correlation coefficient of 0.907 (95% confidence interval from 0.904 to 0.911). e, Scatterplot of one subject with both longitudinal and baseline samples from plasma indicates a Pearson correlation coefficient of 0.938 (95% confidence interval from 0.930 to 0.945). f, Scatterplot of one subject with two technical replicates from brain indicates a Pearson correlation coefficient of 0.976 (95% confidence interval from 0.976 to 0.981). All statistical tests used were two-sided from (c) to (f).
Extended Data Fig. 3 | Overview of the sample size and number of pQTLs from pQTL studies mentioned in this paper and the summary statistics from the meta-analyses. a, Scatter plot of sample size (log10-scaled) and number of total pQTLs after clumping or unique proteins when no clumping was performed (log10-scaled). Dot color represents the tissue type; dot size represents total number of proteins profiled. b, Table of these nine datasets listed the exact numbers for drawing the scatter plot. c, Table of three different combinations of meta-analyses: 2) meta2_WUcsf_PPMI19_JP17: meta-analysis on all three CSF studies by Sasayama and colleagues published in 2017, by PPMI released in 2019, and by Washington University cohort (this study); 3) meta3_WUplasma_WUbrain: meta-analysis on all three-tissue findings from CSF, plasma and brain respectively by Washington University cohort (this study); 4) meta4_WUplasma_WUbrain_PPMI19_JP17: meta-analysis on both the CSF studies by Sasayama and colleagues published in 2017 and by PPMI released in 2019 plus all three-tissue findings from CSF, plasma and brain respectively by Washington University cohort (this study). The columns include number of proteins in common, number of protein-level GWAS hits after meta-analysis, number of protein-level GWAS hits before meta-analysis using only the common proteins within each tissue for each combination. d, Stacked Manhattan plots for all three different combinations of meta-analyses. The darkred line represents $P = 5 \times 10^{-8}$. 

| meta_ID | Study included | Tissue type | Total sample size | Proteins in common | Number of pGWAS hits after meta-analysis | GWAS hits in CSF | GWAS hits in WU plasma | GWAS hits in WU brain |
|---------|----------------|-------------|-------------------|-------------------|-----------------------------------------|-----------------|-----------------------|----------------------|
| 2       | CSF, PPMI19 + CSF, JP17 + CSF_WU | CSF         | 1030              | 596               | 310 proteins with -225 loci             | 165 proteins with 200 loci | 90 proteins with 107 loci | 36 proteins with 64 loci |
| 3       | CSF, WU plasma + brain_WU, Multi-tissue | 1744         | 411               | 150 proteins with -200 loci             | 120 proteins with 157 loci | 92 proteins with 109 loci | 36 proteins with 42 loci |
| 4       | CSF, PPMI19 + CSF, JP17 + CSF, WU plasma + brain_WU, Multi-tissue | 2007         | 342               | 180 proteins with -253 loci             | 104 proteins with 130 loci | 77 proteins with 92 loci | 32 proteins with 37 loci |
Extended Data Fig. 4 | Disease stratified analysis on comparing pQTLs effect size. To investigate of disease status effect on pQTLs, we performed linear regression on the same protein-loci pairs (before conditioning on top variants) identified from above default model using three additional models: a, joint analysis but with disease status as another covariate (CO vs non-CO). Pearson correlation coefficient was 0.999 (p-value < 2.2 × 10^-16, 95%CI = 0.999 to 0.999), 0.999 (p-value = 4.3 × 10^-202, 95%CI = 0.999 to 0.999), 0.999 (p-value = 9.5 × 10^-52, 95%CI = 0.999 to 0.999) for CSF, plasma, and brain respectively. Sample size for this joint analysis was 835, 529, and 380 for CSF, plasma, and brain respectively. b, AD case (CA) only using the same covariates as default model. Pearson correlation coefficient of 0.991 (p-value = 3.9 × 10^-160, 95%CI = 0.988 to 0.993), 0.989 (p-value = 1.8 × 10^-83, 95%CI = 0.983 to 0.992), 0.998 (p-value = 2.4 × 10^-29, 95%CI = 0.995 to 0.999) for CSF, plasma, and brain respectively. Sample size for this AD case (CA) only analysis was 217, 168, and 248 for CSF, plasma, and brain respectively. c, Cognitive unimpaired (CO) only using the same covariates as default model. Pearson correlation coefficient of 0.999 (p-value = 5.2 × 10^-234, 95%CI = 0.998 to 0.999), 0.998 (p-value = 1.17 × 10^-202, 95%CI = 0.997 to 0.999), 0.602 (p-value = 0.002, 95%CI = 0.262 to 0.809) for CSF, plasma, and brain respectively. Sample size for this cognitive unimpaired (CO) only analysis was 614, 357, and 24 for CSF, plasma, and brain respectively. The relatively low correlation in default model comparison with control only in brain samples was due to much smaller sample size as a control for brain samples. All statistical tests used were two-sided from (a) to (c).
Extended Data Fig. 5 | Global view of pleiotropic regions in CSF. In total, 59 Pleiotropic regions passing genome-wide significance threshold ($5 \times 10^{-8}$) in CSF (sample size = 835). Unique non-overlapping regions associated with a given SOMAmer were first defined as 1-Mb region upstream and downstream of each significant variant for that SOMAmer. Within the region (2 Mb) containing the variant with the smallest $P$ value, any overlapping regions were then merged into the same locus. Next, an LD-based clumping approach was adapted to identify whether a region was associated with multiple SOMAmers. Variants were combined into a single region per LD (EUR) defined loci. Any loci associated with more than one protein were identified as pleiotropic regions. Genomic locations of pQTLs were visualized by a squared-Manhattan plot. Dark-green represents cis-pQTLs; gold represents trans-pQTLs. X-axis indicates the positions of the top variant; and Y-axes indicates the gene encoding the protein. All pleiotropic genomic regions are annotated at the top of each plot along the X-axis.
Extended Data Fig. 6 | Global view of pleiotropic regions in plasma. In total, 34 pleiotropic regions passing genome-wide significance threshold ($5 \times 10^{-8}$) in plasma (sample size = 529). Genomic locations of pQTLs were visualized by a squared-Manhattan plot, same as Extended Data Fig. 5.
Extended Data Fig. 7 | Global view of pleiotropic regions in brain. In total, 10 pleiotropic regions passing genome-wide significance threshold (5 × 10^{-8}) in brain (sample size = 380). Genomic locations of pQTLs were visualized by a squared-Manhattan plot, same as Extended Data Fig. 5.
Extended Data Fig. 8 | Tissue specificity exploration with permissive thresholds. To determine whether our tissue-specificity results were biased by statistical power, we performed similar analyses with two more permissive p-values on the 411 proteins. **a**, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold ($5 \times 10^{-8}$) for all three tissues. **b**, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold for one tissue and 0.001 for the other two tissues. For example, when checking CSF pQTLs shared in plasma or brain, we chose $5 \times 10^{-8}$ as threshold for CSF and 0.001 for plasma or brain. **c**, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold for one tissue and 0.05 for the other two tissues. For example, when checking CSF pQTLs shared in plasma or brain, we chose $5 \times 10^{-8}$ as threshold for CSF and 0.05 for plasma or brain.
Extended Data Fig. 9 | Tissue specificity exploration with plasma result from INTERVAL study. To further demonstrate that tissue-specificity findings are not a product of different sample size, we performed similar comparisons by analyzing the plasma pQTLs from the INTERVAL study on 616 proteins that passed QC in our CSF, brain and plasma INTERVAL. a, Venn diagrams of proteins passing QC across all three tissues: CSF and brain results are from WashU cohort, plasma result is from INTERVAL study. b, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold (5 × 10⁻⁸) for all three tissues. c, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold for one tissue and 0.001 for the other two tissues. For example, when checking CSF pQTLs shared in plasma or brain, we chose 5 × 10⁻⁸ as threshold for CSF and 0.001 for plasma or brain. d, Venn diagrams of all pQTLs across all three tissues by fixing genome-wide significance threshold for one tissue and 0.05 for the other two tissues. For example, when checking CSF pQTLs shared in plasma or brain, we chose 5 × 10⁻⁸ as threshold for CSF and 0.05 for plasma or brain.
Extended Data Fig. 10 | See next page for caption.
Extended Data Fig. 10 | Properties of pQTLs. a, Dot plots of -log10(P) from all significant associations (via linear regression) against the distance of sentinel SNPs from TSS within each tissue. b, Dot plots of absolute effect size associated with MAF within each tissue. c, Forest plot of enrichment on the predicted functional annotation classes of pQTLs versus null sets of variants from permutation within each tissue (Data are presented as mean values of Odds Ratio +/- 95% confidence interval from Fisher’s Exact Test) and Bar plots of the proportion of variants annotate in each class. (Note: Features on exonic_splicing/ncRNA_splicing/splicing/UTR5_UTR3 are not shown due to not all tissues have these features). d, Histograms of variance explained by conditionally independent variants within each tissue. For CSF, the mean = 0.141, standard deviation = 0.144, mode = 0.061; For plasma, the mean = 0.157, standard deviation = 0.125, mode = 0.188; For brain, the mean = 0.208, standard deviation = 0.151, mode = 0.092.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
- Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen’s d, Pearson’s r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used for data collection.

Data analysis

- We processed the proteomics data using SomaDataIO (v1.8.0) and Biobase (v2.42.0).
- We processed the genotype data using PLINK1.9 (v1.90b6.4).
- We performed linear regression for identifying pQTL using PLINK2 (v2.00a2LM). The stacked manhattan plot was drawn using ggplot2 (v3.2.1). The interactive manhattan plots and QQ-plots can be visualized using pheweb (v1.1.14).
- We performed conditional analyses using PLINK2 (v2.00a2LM). The regional plot was visualized using locusZoom (v1.3).
- We performed meta-analyses using METAL (version released on 2011-03-25)
- We queried EBI-NHGRI GWAS catalog using phenoscanerV2.
- We identified tissue-specific/shared pQTLs using mashr (v0.2.21).
- We visualized the tissue-specific and tissue-shared pQTLs using VennDiagram R package (v1.6.20).
- We annotated significant pQTLs using gene-based annotation from ANNOVAR (version 2018-04-16).
- We tested whether significant pQTLs enriched for functional and regulatory characteristics using GARFIELD v2.
We identified pleiotropic regions using the LD block information from 1000 Genome Project implemented into the RHOGE R package (v0.1). Circos plots were generated using functions from R package circlize (v0.4.7).

We performed Bayesian colocalization analysis using coloc R package (v3.1).

We performed Mendelian Randomization using TwoSampleMR package (v0.4.22).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. Github). See the Nature Research guidelines for submitting code & software for further information.

**Data**

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Three from Knight ADRC dataset for discovery: Both summary statistics and individual-level data have been uploaded to NIAGADS (The National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site) repository at https://www.niagads.org/datasets/ng00102. Summary statistics (pQTL) data is freely available, as the data exceeds 500Gb, so please email niagads@pennmedicine.upenn.edu to set up an FTP transfer of the data. Summary association results can also be explored through Online Neurodegenerative Trait Integrative Multi-Omics Explorer, ONTIME (https://ontime.wustl.edu/), a PheWeb-based browser.

CSF-Sasayama2017 dataset for replication: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83711
Plasma-AddNeuroMed dataset for replication: https://www.synapse.org/#!Synapse:syn4988768

Drug targets were queried using DrugBank database collected via UniProtKB (as of 1/3/2020) at https://www.uniprot.org/database/DB-0019.

Circos plots were generated using functions from R package circlize (v0.4.7).

We identified pleiotropic regions using the LD block information from 1000 Genome Project implemented into the RHOGE R package (v0.1). Circos plots were generated using functions from R package circlize (v0.4.7).

We performed Bayesian colocalization analysis using coloc R package (v3.1).

We performed Mendelian Randomization using TwoSampleMR package (v0.4.22).

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

**Field-specific reporting**

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- **Life sciences**
- **Behavioural & social sciences**
- **Ecological, evolutionary & environmental sciences**

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

**Life sciences study design**

All studies must disclose on these points even when the disclosure is negative.

**Sample size**

Samples from participants include three tissue types: CSF, plasma, and brain (parietal lobe cortex). This is a discovery study, therefore we did not perform power calculation to detect a given effect size, but to our knowledge, we are well-powered as our study is much larger than previous studies on CSF (n=835), compared to Sasayama et al., and brain (n=380), compared to Robins et al., for discovery analysis, and similar scale on plasma, compared to Deming et al.

**Data exclusions**

Samples, proteins and genotypes which failed QC were excluded as detailed in Methods section. Briefly, QC was performed at the sample and protein levels using control aptamers (positive and negative controls) and calibrator samples. To QC the proteomics datasets, the protein outliers were first removed by applying four criteria: 1) Minimum detection filtering. 2) Flagging analytes based on the scale factor difference. 3) CV of calibrators lower than 0.15. 4) Interquartile range (IQR) strategy. 5) An orthogonal approach was used to call subject outliers based on IQR. Genotype QC was performed based on low quality SNPs and individuals. Individuals within the CEU cluster and without unanticipated duplicates and cryptic relatedness were kept. All exclusion criteria were pre-established.

**Replication**

For CSF, 90.1% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using Sasayama et al., 2017 (SOMAscan-based, CSF); PPMI19, unpublished (SOMAscan-based, CSF); and meta-analysis of these two prior studies. Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI. For plasma, 96.8% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using AddNeuroMed (SOMAscan-based, plasma). Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI. For brain, 96.9% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using AddNeuroMed (SOMAscan-based, plasma). Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI. For plasma, 96.9% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using AddNeuroMed (SOMAscan-based, plasma). Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI. For brain, 96.9% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using AddNeuroMed (SOMAscan-based, plasma). Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI. For brain, 96.8% were replicated using other published pQTL results: We first searched the reprocessed pQTL results using AddNeuroMed (SOMAscan-based, plasma). Next, we checked summary statistics from INTERVAL (SOMAscan-based, plasma). Finally, we queried other plasma pGWAS findings from EBI-NHGRI.

**Randomization**

Samples were classified into experimental groups on the basis of neuropathological analysis/diagnosis and all samples were randomly assigned to a profiling pool prior to proteomics measurement. In brief, analyses were controlled for individual-level covariates including: age (details in Table-1), sex, first two principal components for genetic ancestry, genotyping array types.

**Blinding**

Laboratory staff were blinded to sample status during protein measurement. The investigators were not blinded to group allocation during data collection and/or analysis. We found pQTLs were not disease-specific, thus blinding was not relevant to these experiments.
Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

| Involved in the study | n/a |
|-----------------------|-----|
| Antibodies            | ☑   |
| Eukaryotic cell lines | ☑   |
| Palaeontology and archaeology | ☑   |
| Animals and other organisms | ☑   |
| Human research participants | ☑   |
| Clinical data        | ☑   |
| Dual use research of concern | ☑   |

### Methods

| Involved in the study |
|-----------------------|
| ☑ ChiP-seq            |
| ☑ Flow cytometry      |
| ☑ MRI-based neuroimaging |

### Human research participants

Policy information about [studies involving human research participants](#).

**Population characteristics**

Covariate-relevant population characteristics of the human donors of three tissues analyzed in this study are presented in Table 1 and S1. Descriptions are provided in the text and below.

- **Age**: For CSF, age denotes age at lumbar puncture; For plasma, age denotes age at plasma draw date; For brain, age denotes age at death.
- **Sex**: sex of the individual.
- **Clinical Dementia Rating**: the individual’s clinical dementia severity (please see: [https://knightadrc.wustl.edu/cdr/cdr.htm](https://knightadrc.wustl.edu/cdr/cdr.htm) for details).
- **APOE4**: the number of APOE4 alleles the individual possessed.

**Recruitment**

1) Participant recruitment information for the Knight-ADRC - [https://knightadrc.wustl.edu/default.htm](https://knightadrc.wustl.edu/default.htm). Very briefly: Subjects are recruited if they over the age of 60yrs. They can be demented or nondemented. Women and minorities are included in this research.

2) Participant recruitment information for the DIAN - [https://dian.wustl.edu/](https://dian.wustl.edu/). Very briefly: Families with early-onset AD (onset <65) and that carry a known pathogenic mutation in APP, PSEN1, or PSNE2 are recruited for this study. Any family member independently of the mutation or clinical status is recruited if older than 18.

Because our study includes cognitively normal older individuals and AD cases, there is a self-selection bias of disease and age effect. However, we found our pQTLs were not disease- or age-specific.

**Ethics oversight**

Washington University in St. Louis

Note that full information on the approval of the study protocol must also be provided in the manuscript.