Genetic control of the human brain proteome

Chloe Robins[1#], Aliza P. Wingo[2,3#], Wen Fan[1], Duc M. Duong[4], Jacob Meigs[1], Ekaterina S. Gerasimov[1], Eric B. Dammer[4], David J. Cutler[5], Philip L. De Jager[6,7], David A. Bennett[8], James J. Lah[1], Allan I. Levey[1,*], Nicholas T, Seyfried[2,*], Thomas S. Wingo[1,3,*]

[1] Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA

[2] Division of Mental Health, Atlanta VA Medical Center, Decatur, GA, USA

[3] Department of Psychiatry, Emory University School of Medicine, Atlanta, GA, USA

[4] Department of Biochemistry, Emory University School of Medicine, Atlanta, GA, USA

[5] Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA

[6] Cell Circuits Program, Broad Institute, Cambridge, MA, USA

[7] Center for Translational and Computational Neuroimmunology, Department of Neurology, Columbia University Medical Center, New York, NY, USA

[8] Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, Illinois, USA

[#] Contributed equally

[*] Corresponding author

Correspondence should be addressed to T.S.W. (thomas.wingo@emory.edu), N.T.S. (nseyfri@emory.edu), and A.I.L. (alevey@emory.edu)

Abstract

Alteration of protein confirmation and abundance is widely believed to be the hallmark of neurodegenerative diseases. Yet relatively little is known about the genetic variation that controls protein abundance in the healthy human brain. Genetic control of protein abundance is generally thought to parallel that of RNA expression, but heretofore there has been little direct evidence to support that belief. Here, we directly assessed single nucleotide variants (SNVs) that are associated with variation in brain protein abundance in healthy humans. We performed protein quantitative trait loci (pQTL) analyses using tandem mass spectrometry-based proteomic quantification of proteins from dorsolateral prefrontal cortex (dPFC) that identified 12,691 unique proteins (7,901 after quality control) and whole genome sequencing of 144 cognitively unimpaired older participants of the Religious Order Study (ROS) and Memory and Aging Project (MAP). Linear regression was used to test whether SNVs within a 100-kb window around the protein-coding sequence were associated with protein abundance. We identified 28,211 SNVs that were significantly associated with the abundance of 864 proteins (i.e. pQTLs), and the complete results are searchable at http://brainqtl.org. Brain pQTL sites were compared to expression quantitative trait loci (eQTL) analyses performed using RNA-sequencing from the dPFC of 169 cognitively unimpaired ROS/MAP participants, of which 81 were the exact same individuals whose protein abundance was measured. We found that strong pQTLs are generally only weak eQTLs, and that the majority of strong eQTLs are not detectable pQTLs. These results suggest that the genetic control of mRNA and protein abundance may be substantially distinct and provide additional evidence that inference concerning protein abundance made from mRNA studies should be treated with caution.

Keywords: brain pQTL, eQTL, proteome, transcriptome, quantitative trait locus, tandem mass spectrometry

Background

Proteins are known to play an important role in neurodegenerative disease. Alzheimer's disease (AD), for instance, is characterized by the abnormal accumulation of amyloid-beta and tau proteins in the brain. In addition to these hallmark proteins, the abundance of hundreds of other proteins have also been shown to correlate with neurodegenerative disease phenotypes such as rate of decline of cognition and diagnosis of AD [1, 2]. This suggests that the dysregulation of numerous proteins may contribute to disease.

While the genetic control of mRNA expression in the brain has been well studied [1, 3-7], little is known about how genetics influence protein abundance in the brain. Thousands of genetic variants have been reported to associate with variation in mRNA levels, known as expression quantitative trait loci (eQTLs). Oftentimes, these identified genetic effects on mRNA are used to help prioritize candidate causal genes identified by genome-wide association studies [8], and are assumed to translate to protein abundance. However, this relationship has yet to be tested with large-scale profiling of brain proteins. A better understanding of the genetic control of protein abundance will help us define the relationships between genetics, gene expression, and proteins, and lead to a better understanding of the molecular changes that underlie neurodegenerative and other brain diseases.

Here, we perform a large-scale investigation into the genetic control of the human brain proteome using high-throughput mass spectrometry-based protein quantification and whole genome sequencing from post-mortem brain samples of the dorsolateral prefrontal cortex of cognitively normal older adults. To understand the relationship between genetics, gene expression, and protein, we also investigate the genetic control of gene expression (i.e. mRNA) in the human brain from the same cohort and among the same individuals with brain proteomes. Finally, we compare the pQTLs identified with eQTLs from a recent meta-analysis of brain expression of over 1,433 brains [9]. The results of our analyses are available at http://brainqtl.org and serve as a resource for future investigations.

Methods

Study subjects

The subjects in this study are participants of the Religious Orders Study (ROS) and the Memory and Aging Project (MAP). ROS and MAP are longitudinal cohort studies of Alzheimer's disease and aging maintained by investigators at the Rush Alzheimer's Disease Center in Chicago, IL [10-12]. Both studies recruit participants without known dementia at baseline and follow them annually using detailed clinical evaluation. ROS recruits individuals from catholic religious orders from across the USA, while MAP recruits individuals from retirement communities as well as individual home visits in the Chicago, IL area. Participants in each study undergo annual medical, neurological, and neuropsychiatric assessments from enrollment to death, and neuropathologic evaluations at autopsy. Participants provided informed consent, signed an Anatomic Gift Act, and repository consent to allow their data and biospecimens to be repurposed. The studies were approved by an Institutional Review Board of Rush University Medical Center.

Clinical Diagnoses

For each ROS/MAP participant, a clinical diagnosis of dementia is rendered annually and at the time of death. The diagnosis rendered at death is based on all available clinical data and is given by a neurologist who is blinded to all postmortem data using the National Institute of Neurological and Communicative Disorders and Stroke and the AD and Related Disorders Association guidelines [13]. Case conferences including two neurologists and a neuropsychologist were used for consensus, as necessary, for select cases. Diagnoses of dementia status was coded as no cognitive impairment (NCI), mild cognitive impairment (MCI), or Alzheimer's dementia (AD). Here, we restricted the analyses to those with NCI at death to investigate the genetic control of the normal human brain proteome and transcriptome. ROS/MAP resources can be requested at www.radc.rush.edu.

Genetic data

Genotype data was generated from whole genome sequencing (WGS) of DNA that was extracted from cryopreserved peripheral blood mononuclear cells or frozen dorsolateral prefrontal cortex (dPFC) of ROSMAP subjects. WGS was performed as described in detail by De Jager et al. [14] and is available via Synapse (ID: syn10901595). Briefly, libraries were constructed using the KAPA Hyper Library Preparation Kit per the manufacturer's protocol and sequenced on an Illumina HiSeq X sequencer (v2.5 chemistry) using 150bp paired-end reads. Reads were aligned to the GRCh37 human reference genome using Burrows-Wheeler Aligner (BWA-MEM v0.7.8) [15] and processed using the GATK best-practices workflow, which includes marking duplicate reads by Picard tools v1.83, local realignment around indels, and base quality score recalibration by Genome Analysis Toolkit (GATK v3.4.0) [16, 17]. A multi-sample genomic variant call format (gVCF) was generated by merging results of HaplotypeCaller on each sample individually in gVCF mode (GATKv3.4.0) and batches of gVCF were merged into gVCFs processed by a joint genotyping step (GATK v3.2.2).

Annotation of the multi-sample VCF (n=1,196) was performed using Bystro [18] and supplemented by the Broad's ChromHMM annotation of dPFC tissue [19, 20]. A total of 1,133 samples passed all quality control measures and 63 samples were excluded for one or more of the following reasons. Samples with greater than five standard deviations for θ, silent:replacement sites, and transition:transversion ratio were excluded (n=7), and samples with greater than three standard deviations for genotype missingness, heterozygosity, or homozygosity were excluded (n=14). Samples that were discordant for sex based on heterozygosity of the X chromosome were also excluded (n=7). Cryptically related or duplicate samples were identified by identity-by-state sharing using PLINK [21] and removed (n=31). Unlinked ancestrally informative markers were used to infer eigenvectors for principal-component analysis using EIGENSTRAT [22] and over six standard deviation outliers (n=1) were removed. Before analysis, we also removed all non-SNVs (i.e. insertions and deletions), SNVs outside of Hady-Weinberg equilibrium, SNVs with missing data for over 10% of samples, and SNVs with a minor allele frequency less than 0.05.

Protein Abundance by Tandem MS-based Proteomics

Protein abundance from cortical microdissections of dPFC (Broadman area 9) of ROS/MAP subjects was generated using tandem mass tag (TMT) isobaric labeling mass spectrometry methods for protein identification and quantification. A brief description of these methods is provided below, and a detailed description is provided in the supplementary information (Section A).

Tissue homogenization was performed as described by [23], followed by protein digestion. For protein digestion, 100 µg of each sample was reduced with 1 mM dithiothreitol (DTT) at room temperature (RT) for 30 min, followed by 5 mM iodoacetamide (IAA) alkylation in the dark for another 30 min and overnight digestion with Lysyl endopeptidase (Wako) at 1:100 (w/w). Subsequently, samples were diluted 7-fold with 50 mM ammonium bicarbonate (AmBic) and digested with 1:50 (w/w) Trypsin (Promega) for another 16 h. The peptide solutions were acidified to a final concentration of 1% (vol/vol) formic acid (FA) and 0.1% (vol/vol) triflouroacetic acid (TFA), desalted with a 30 mg HLB column (Oasis). An equal amount of protein from each sample was aliquoted and digested in parallel to serve as the global pooled internal standard (GIS) in each TMT batch.

Prior to TMT labeling, all samples were randomized into 50 batches (8 samples per batch) based on age at death, sex, post-mortem interval, diagnosis, and measured neuropathologies. Peptides from each individual sample (*n*=400) and the GIS (*n*=100) were labeled using the TMT 10-plex kit (ThermoFisher). In each batch, TMT channels 126 and 131 were used to label GIS standards, while the 8 middle TMT channels were reserved for individual samples following randomization. The TMT labeling was performed as described by [23, 24], followed by high-pH fractionation performed as described by [25] with slight modification. Dried samples were re-suspended in high pH loading buffer (0.07% vol/vol NH_4OH; 0.045% vol/vol formic acid, 2% vol/vol acetonitrile) and loaded onto an Agilent ZORBAX 300Extend-C18 column (2.1mm x 150 mm with 3.5 µm beads). An Agilent 1100 HPLC system was used to carry out the fractionation. A total of 96 individual fractions were collected across the gradient and pooled into 24 fractions and dried.

All fractions were resuspended in equal volume of loading buffer (0.1% formic acid, 0.03% trifluoroacetic acid, 1% acetonitrile) and analyzed by liquid chromatography coupled to mass spectrometry as described by [26] with slight modifications. Peptide eluents were separated on a self-packed C18 (1.9 um Dr. Maisch, Germany) fused silica column (25 cm × 75 μM internal diameter (ID); New Objective, Woburn, MA) by a Dionex UltiMate 3000 RSLCnano liquid chromatography system (ThermoFisher Scientific) and monitored on an Orbitrap Fusion mass spectrometer (ThermoFisher Scientific). Sample elution was performed over a 180-min gradient with flow rate at 225 nL/min. The gradient goes from 3% to 7% buffer B in 5 mins, from 7% to 30% over 140 mins, from 30% to 60% in 5 mins, 60% to 99% in 2 mins, kept at 99% for 8 min and back to 1% for additional 20 min to equilibrate the column. The mass spectrometer was set to acquire in data dependent mode using the top speed workflow with a cycle time of three seconds. Each cycle consisted of one full scan followed by as many MS/MS (MS2) scans that could fit within the time window. The full scan (MS1) was performed with an m/z range of 350-1500 at 120,000 resolution (at 200 m/z) with AGC (automatic gain control) set at $4x10^5$ and maximum injection time of 50 msec. The most intense ions were selected for higher energy collision-induced dissociation (HCD) at 38% collision energy with an isolation of 0.7 m/z, a resolution of 30,000 and AGC setting of $5x10^4$ and a maximum injection time of 100 msec. Five of the 50 TMT batches were run on the Orbitrap Fusion mass spectrometer using the SPS-MS3 method as previously described by [23].

All raw files were analyzed using the Proteome Discoverer suite (version 2.3 ThermoFisher Scientific). MS2 spectra were searched against the canonical UniProtKB Human proteome database (Downloaded February 2019 with 20,338 total sequences). The Sequest HT search engine was used and parameters were specified as: fully tryptic specificity, maximum of two missed cleavages, minimum peptide length of six, fixed modifications for TMT tags on lysine residues and peptide N-termini (+229.162932 Da) and carbamidomethylation of cysteine residues (+57.02146 Da), variable modifications for oxidation of methionine residues (+15.99492 Da) and deamidation of asparagine and glutamine (+0.984 Da), precursor mass tolerance of 20 ppm, and a fragment mass tolerance of 0.05 Da for MS2 spectra collected in the Orbitrap (0.5 Da for the MS2 from the SPS-MS3 batches). Percolator was used to filter peptide spectral matches (PSM) and peptides to a false discovery rate (FDR) of less than 1%. Following spectral assignment, peptides were assembled into proteins and were further filtered based on the combined probabilities of their constituent peptides to a final FDR of 1%. In cases of redundancy, shared peptides were assigned to the protein sequence in adherence with the principles of parsimony. Reporter ions were quantified from MS2 or MS3 scans using an integration tolerance of 20 ppm with the most confident centroid setting.

The GIS channels 126 and 131 in each batch served as technical replicates. We compared the measured abundance of each protein from the two GIS, and found the measured proteome between the two GIS to be over 99% correlated for all batches (Figure S1). As a quality control measure, we removed protein abundance measurements with low correlations between the two GIS (outside the 95% confidence interval) in each batch before further analysis.

For this study, we included only cognitively normal subjects based on the clinical diagnosis of cognitive status rendered at death. To ensure the analysis of high-quality data we: 1) excluded proteins with missing data for over 50% of samples, 2) scaled each abundance value by a sample-specific total protein abundance measure to remove the effects of loading differences, and 3) transformed the data to the $\log_2$ scale. Outlier samples were then identified and removed through iterative principal component analysis. In each iteration we removed samples more than four standard deviations from the mean of the first or second principal component and then re-calculated all the principal components. Following outlier removal, we again removed proteins with missing data for over 50% of the samples. Finally, the abundance of each protein was residualized using a linear regression model to remove the effects of sex, age at death, post-mortem interval, study, batch, and MS2 versus MS3 reporter quantitation mode.

Gene Expression by RNA Sequencing

Gene expression was measured from the dPFC (Broadman area 46) by De Jager et al. [14]. Briefly, RNA was extracted from cortically dissected sections of dPFC grey matter and samples with

RNA integrity numbers (RIN) over 5 were used to prepare RNA-Seq libraries using strand-specific dUTP method with poly-A selection [27, 28] using the Illumina HiSeq with 101-bp paired-end reads to a target coverage of 50 million reads per library. Raw RNA-Seq reads were aligned to a GRCh38 reference genome and gene counts were computed using STAR [29] as described in reference [30]. We obtained RNA-Seq data from synapse (ID: syn17010685) and performed the following quality control measures in a subset of individuals with normal cognition defined by a clinical diagnosis of no cognitive impairment rendered at death. We excluded: 1) non-Caucasian samples, 2) outlier samples via the GTex expression outlier test (D-statistic below 0.9) [1, 31], and 3) genes with < 1 cpm in over 50% of samples. Subsequently, the filtered data was normalized using the varianceStabilizingTransformation function from the DESeq2 R package, which $\log_2$ transforms counts, normalizes for library size, and transforms counts to be approximately homoscedastic [32]. Lastly, the residual expression for each gene was estimated using linear regression to remove the effects of sex, sequencing batch, age at death, post-mortem interval, RIN, and study.

Estimation of confounders

To reduce confounding due to population structure, the first ten principal components derived from principal component analysis of the WGS data were added as model covariates in all relevant analyses. All ten of these principal components had significant Tracey-Widom statistics (p-value < 0.05).

Statistical Analyses

To identify genetic variants associated with protein abundance in the brain, we used linear regression to model protein abundance as a function of genotype. We reduced our computational and testing burden by investigating only the proximal genetic effects of common SNV variants by testing only SNVs within 100 Kb of each protein coding gene with a minor allele frequency (MAF) over 5%. The location of each protein coding gene was defined by the knownGene table (GRCh37/hg19 assembly) from the University of California, Santa Cruz (UCSC) table browser [33]. For each SNV-protein pair, we regressed genotype against protein abundance, assuming additive genetic effects and including the first ten genetic principal components as covariates. We also performed analyses that included cell type proportions and additional unmeasured confounders as covariates, but found their inclusion as covariates to increase λ, an estimate of test statistic inflation (see supplementary material section B and table S1). SNVs where genotype was significantly associated with protein abundance after False Discovery Rate (FDR) correction for multiple comparisons were declared protein quantitative trait loci (pQTLs; FDR < 0.05).

For each protein-coding gene, we also identified genetic variants associated with gene expression in the brain (i.e. mRNA expression). We used the same methods that were used to identify genetic variants associated with protein abundance. That is, we used linear regression to model mRNA abundance as a function of genotype for common SNVs (MAF > 0.05) within 100 kb of each protein-coding gene. To be able to compare SNVs associated with mRNA expression to those associated with protein abundance, we restricted our expression analyses to mRNA transcripts of genes that code for proteins present in our pQTL analyses. The location of each gene was defined by the Ensembl stable gene (EnsGene) table (GRCh37/hg19 assembly) from the University of California, Santa Cruz (UCSC) table browser. This table allowed us to match each EnsGene in our GRCh38-aligned RNAseq dataset with its GRCh37 location. For each SNV-mRNA pair, we regressed genotype against mRNA abundance, assuming additive genetic effects and including the first ten genetic principal components. We also performed analyses that included cell type proportions and additional unmeasured confounders as covariates, but found their inclusion as covariates to increase λ (see supplementary material section B and table S1). SNVs where genotype is significantly associated with gene expression after FDR correction for multiple comparisons were declared expression quantitative trait loci (eQTLs; FDR < 0.05).

We assessed whether the identified pQTLs and eQTLs are more likely to be a particular type of nucleotide substitution (i.e. synonymous or nonsynonymous) or in a particular genic location by testing the overlap between the sets of pQTL and eQTLs sites and the sets of sites annotated to each substitution type and genic location using Fisher's exact tests.

Results

Demographics

We analyzed genetic, proteomic, and transcriptomic data from 233 ROS/MAP participants, of which 144 have proteomic data, 169 have transcriptomic data, and 81 have both proteomic and transcriptomic data. Table S2 gives the demographic characteristics of these subjects. Participants had a high degree of education (median of 16 years), were all Caucasian, and were 63% women. The age at death ranged from 67 years to 102 years, with a median age at death of 86.5 years.

Brain QTL Mapping

To identify pQTLs, we analyzed proteomes from the dorsolateral prefrontal cortex and genotyping from WGS of 144 individuals. The genotypes of a total of 2,599,383 SNVs were tested against the abundance of 7,901 proteins. We found that 10.9% (864 of 7,901) of the proteins were associated with a total of 28,211 pQTLs (FDR < 0.05). Of the proteins with a pQTL, 78% (671 of 864 proteins) had multiple pQTLs, with an average of 33 pQTLs detected per protein.

To identify eQTLs and pQTLs from the same set of genes, we tested genotypes of 2,082,000 SNVs against the abundance of protein and mRNA from a restricted set of 5,739 genes found in both studies. We found that 10.7% (616 / 5,739) of genes had a pQTL and 14.7% (843 / 5,739) of genes had an eQTL. Fewer pQTLs were identified than eQTLs with a total of 21,034 and 35,054, respectively. Additionally, genes with a pQTL averaged 34 pQTLs per gene compared to 42 eQTLs per gene for genes with an eQTL. Only 199 genes had both a pQTL and an eQTL, which represents 32.3% (199 / 616) of genes with a pQTL and 23.6% (199 / 843) of genes with an eQTL. A total of 3,364 SNVs were identified as both a pQTL and an eQTL (i.e. eQTL/pQTLs) in 95 of the 199 genes with both a pQTL and an eQTL (see Table S3 for the top 10). Thus, only 16.0% (3,364 / 21,034) of all identified pQTLs are eQTLs and 9.6% (3,364 / 35,054) of all identified eQTLs are pQTLs (Figure 1A inset). These results were essentially unchanged when we: 1) limited the analysis to samples with complete proteomic and transcriptomic data (supplementary material section C); 2) used different R^2 thresholds for linkage disequilibrium (supplementary material section D); 3) varied the window of SNVs around the gene (50kb, 100kb, or 500kb; supplementary material section E); and 4) used the more stringent Bonferroni significance threshold to define pQTLs and eQTLs (supplementary material section F).

We tested the reproducibility of our analyses by comparing our eQTLs with those reported previously by [9] in a larger sample of 1,433 cognitively normal and cognitively impaired individuals, and saw a high replication rate of 93% (supplementary material section G). Furthermore, we assessed the relationship between the minor allele frequencies (MAF) of the tested variants and our declaration of pQTLs and eQTLs (supplementary material section H). We found that the percentage of the total pQTLs, eQTLs, and eQTL/pQTLs identified slightly increases with variant MAF. The smallest percentage of pQTLs, eQTLs and eQTL/pQTLs identified have a variant MAF between 0.05 and 0.1 (5%-8%), while the largest percentage of pQTLs, eQTLs, and eQTL/pQTLs have a variant MAF between 0.4 to 0.5 (26%-36%). This suggests a slight, but not substantial, dependence of pQTLs and eQTLs on MAF, likely because our power to detect QTLs is higher for variants with higher MAF. Finally, we compared our identified brain pQTLs to previously published pQTLs of human blood proteins [34-36], and found very few pQTLs from human brain to also be pQTLs of human blood proteins (supplementary material section I). This suggests that the genetic control of blood and brain proteins are largely distinct.

For each SNV that was identified as either an eQTL or a pQTL, we compared the effect of each variant on mRNA and protein abundance (Figure 1A). To facilitate the comparison, we defined strong QTLs as those with an effect estimate greater than two standard deviations from the mean effect estimate of eQTLs or pQTLs. Based on this definition, 7% (2,475/ 35,064) of the identified eQTLs are strong eQTLs, 11% (2,271 / 21,034) of the identified pQTLs are strong pQTLs., and 10% (324 / 3,364) of the sites that are both an eQTL and a pQTL (eQTL/pQTLs) are both a strong eQTL and a strong pQTL. We found that 64% (1,586 / 2,475) of sites with large effects on mRNA abundance (i.e. strong eQTLs) are not also strong pQTLs. Similarly, we found that 86% (1,947 / 2,271) of sites with large effects on protein abundance (i.e. strong pQTLs) are not also strong eQTLs. Finally, at sites that are both eQTL and pQTLs, only 2% (66 / 3,364) of the sites that have strong effects on mRNA have only weak effects on protein, and 13% (444 / 3,364) of the sites that have strong effects on protein have only week effects on mRNA. Furthermore, 94% (3,200 / 3,364) of these eQTL/pQTLs have effects on mRNA and protein abundance that are in the same direction (Figure 1A).

To help understand if these results are generalizable, we compared our pQTL results to eQTL results from a large meta-analysis using data from the dPFC of 1,433 samples from four cohorts [9]. For each SNV that we identified as a pQTL or was identified as an eQTL in the meta-analysis, we compared the effect of each variant on mRNA and protein abundance (Figure 1B). Even with the increase in sample size and power to detect eQTLs, we still see similar patterns between the effects of pQTLs and eQTLs. That is, the majority of strong eQTLs are not strong pQTLs, and vice versa. This suggests that most of the large genetic effects on protein abundance have only a small effect on mRNA levels, and that most of the large genetic effects on mRNA levels have only a small effect on protein abundance.

Functional analysis of pQTLs and eQTLs

The pQTLs identified were more likely to be located in a genic region (5' UTR OR: 1.97, FDR adjusted p-value: $1.3\text{x}10^{-10}$, exons OR: 1.56, FDR adjusted p-value: $2.3\text{x}10^{-18}$, introns OR: 1.22, FDR adjusted p-value: $3.0\text{x}10^{-44}$, Figure 1C) and less likely to be located in an intergenic region by Fisher's exact test (OR: 0.77, FDR adjusted p-value: $8.3\text{x}10^{-73}$, Figure 1C). Furthermore, the identified pQTLs are more likely to be non-synonymous nucleotide substitutions (OR: 1.53, FDR adjusted p-value: $6.8\text{x}10^{-6}$), and less likely to be synonymous nucleotide substitutions (OR: 0.65, FDR adjusted p-value: $6.8\text{x}10^{-6}$). The identified eQTLs were more likely to be located in exons (OR: 1.48, FDR adjusted p-value: $2.9\text{x}10^{-22}$) , 5' UTRs (OR: 1.86 FDR adjusted p-value: $1.6\text{x}10^{-13}$), 3' UTRs (OR: 1.44, FDR adjusted p-value: $3.3\text{x}10^{-20}$), and enhancers (OR: 1.74, FDR adjusted p-value: $7.5 \text{ x } 10^{-5}$), and less likely to be located in introns (OR: 0.98, FDR adjusted p-value: 0.04) and in an intergenic region (OR: 0.96, FDR adjusted p-value: 0.00017) (Figure 1D). Additionally, identified eQTLs are not significantly more likely to be either non-synonymous (OR: 1.04, adjusted p-value: 0.60) or synonymous nucleotide substitutions (OR: 0.96, adjusted p-value: 0.60). Furthermore, we found pQTLs to be significantly enriched in GWAS results (supplementary information section J), suggesting that pQTLs may play a role in disease susceptibility.

Correlation between mRNA and protein abundance

To understand the relationship between mRNA and protein abundance, we examined the correlation between mRNA and protein levels for five sets of genes: 1) genes with sites that are both an eQTL and a pQTL (i.e. eQTL/pQTLs) (n=95); 2) genes with both pQTLs and eQTLs but no eQTL/pQTLs (n=104); 3) genes with pQTLs but no eQTLs (n = 417); 4) genes with eQTLs but no pQTLs (n = 642); and, 5) all tested genes (n=5,739). Remarkably, genes with sites that are both eQTLs and pQTLs (eQTL/pQTLs) have the highest average correlation between mRNA and protein level (0.21), while all other genes have much lower average correlations (0.04 to 0.07) (Figure 1E). Additionally, we found the distributions of correlations between mRNA and protein level to be significantly different between the genes with eQTL/pQTLs, genes with both pQTLs and eQTLs but no eQTL/pQTLs, genes with pQTLs but no eQTLs, and genes with eQTLs but no pQTLs (Kruskal-Wallis test, chi-square =74.6, p = 4.5x10^{-16}) (Figure 1F). Specifically, the median correlation between mRNA and protein levels was found to be significantly higher for genes with eQTL/pQTLs than genes with both pQTLs and eQTLs but not eQTL/pQTLs (0.21 vs. 0.04, Wilcoxon test, p-value:1.5x10^{-12}), genes with only pQTLs (0.21 vs. 0.07, Wilcoxon test, p-value: 1.9x10^{-14}), and genes with only eQTLs (0.21 vs. 0.06, Wilcoxon test, p-value: 4.1x10^{-16}) (Figure 1F).

Brain QTL Resource

Our pQTL and eQTL results are available online at http://brainqtl.org . This website visually displays the results of our pQTL and eQTL analyses and provides summary statistics on individual variants tested.

Discussion

We performed the first unbiased large-scale investigation into the genetic control of the human brain proteome and presented evidence for thousands of pQTLs influencing abundance of hundreds of brain proteins. Our most striking result was that the majority of strong pQTLs are weak eQTLs, and vice versa. Additionally, only a small minority of protein-coding genes have a SNV with an appreciable association with both protein and mRNA abundance that can be defined as both a pQTL and an eQTL. This result is consistent with work in other tissues that also found few sites that are both a pQTL and an eQTL [34, 37, 38].

For the minority of genes that have a SNVs that is both a pQTL and an eQTL, protein abundance may be limited by transcript availability. In contrast, genes that have pQTLs but no eQTLs or only weak eQTLs, protein abundance may be more likely to be regulated by post-transcriptional regulation (e.g., miRNA), localization (e.g., membrane, intra- or extracellular), translational regulation, or post-translational modifications. This is supported by our observation of higher correlations between protein and mRNA abundance for genes with SNVs that are both a pQTL and an eQTL than for genes with either a pQTL or an eQTL alone. Additionally, the depletion and enrichment of pQTLs for non-synonymous and synonymous sites, respectively, also suggests translational regulation. Together these results suggest that caution is needed when inferring the effect of an eQTL on protein abundance in the human brain.

We identified 1.7 times more eQTLs than pQTLs. This is consistent with work in other tissues from both humans and mice that also found fewer pQTLs than eQTLs [34, 35, 37-39]. The difference in the number of pQTLs and eQTLs may be due to greater tolerance for differences in mRNA abundance [40]. Furthermore, the protein lifecycle has considerable variability, and includes numerous contributing processes beyond transcription. In many cases, the effect of genetic variants on protein abundance may be lessened by post-transcriptional mechanisms [37]. This is compatible with both our observation of more genetic variants associated with mRNA abundance than protein abundance and our observation of differing effect sizes for the genetic control of protein and mRNA abundance.

Our results should be interpreted with respect to the strengths and limitations of this study. To our knowledge, the brain proteomic sequencing performed here are among the deepest thus far profiled with a total of 12,691 unique proteins (12,415 detected genes), which reflects about 79% of all expressed transcripts in the human brain. To achieve throughput and proteomic depth of sequencing, we used TMT isobaric labeling coupled with high-pH offline fractionation following well-established protocols [25]. A limitation of the proteomic data was the use of MS2 acquisition, which can suffer from the presence of co-isolated and co-fragmented interfering ions that can obscure quantification [41]; however, high-pH offline fractionation largely mitigates this issue [25]. Another potential limitation is that the observed differences in the number of pQTLs and eQTLs may be in part due to technical differences in the proteomic and transcriptomic profiling methods. However, we note that plasma-based pQTL analysis made a similar observation in blood and used the SOMAscan technology, which differs from the mass spectrometry-based proteomics used here [34]. Another potential limitation is that the number of pQTLs and eQTLs may be limited by the power to detect them, as suggested by our sensitivity analysis (supplementary information section C). However, our study was still sufficiently powered to detect thousands of pQTLs and eQTLs and these potential power issues do not appear to influence our main conclusion regarding the control of brain protein and RNA expression (supplementary information section C). Another potential limitation of this study is that both the gene expression and protein data were generated from bulk tissue that is composed of a mixture of cell types. We estimated cell type proportions for both data types, but found the inclusion of estimated cell type proportions in our analyses to increase test-statistic inflation (supplementary information section B). This suggests that available methods for cell type deconvolution are not sufficiently able to remove confounding by cell type for human brain proteomic data. Despite these limitations, our study is the first to examine genetic control of human brain proteins and reveals key differences in pQTLs versus eQTLs in the human brain and provides a web-based resource to enable researchers to explore the genetic control of the human brain proteome.

FUNDING

Support for this research was provided by the following grants from the US National Institutes of Health: T32 NS007480-18 (C.R.), R01 AG056533 (A.P.W. and T.S.W.), R01AG036042 (D.A.B and P.L.D), R01AG017917 (D.A.B.), R01AG015819 (D.A.B.), P30AG10161 (D.A.B.), U01 AG46152 and U01AG061356 (P.L.D. & D.A.B.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGEMENTS

We thank the participants of the ROS, MAP studies. The results published here are in whole or in part based on data obtained from the AMP-AD Knowledge Portal (doi:10.7303/syn2580853). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, U01AG61356 , the Illinois Department of Public Health, and the Translational Genomics Research Institute. A.P.W. was also supported by a grant from the US Veterans Administration (BX003853) and US National Institute of Health U01 MH115484. T.S.W. was also supported by US National Institute of Aging (P50 AG025688, R56 AG062256, R56 AG060757, U01 AG061357, U01 AG057195, and RF1 AG051633). We thank Thanneer Perumal, PhD for creating the estimated cell type proportions from RNA-sequencing data.

REFERENCES

1. Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, et al. A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease. Cell Systems. 2017;4(1):60-72.e4. doi: 10.1016/j.cels.2016.11.006.

2. Wingo AP, Dammer EB, Breen MS, Logsdon BA, Duong DM, Troncosco JC, et al. Large-scale proteomic analysis of human brain identifies proteins associated with cognitive trajectory in advanced age. Nat Commun. 2019;10(1):1619. Epub 2019/04/10. doi: 10.1038/s41467-019-09613-z. PubMed PMID: 30962425; PubMed Central PMCID: PMC6453881.

3. Ramasamy A, Trabzuni D, Guelfi S, Varghese V, Smith C, Walker R, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014;17(10):1418-28. Epub 2014/09/01. doi: 10.1038/nn.3801. PubMed PMID: 25174004; PubMed Central PMCID: PMC4208299.

4. O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, et al. Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. Genome Biol. 2018;19(1):194. Epub 2018/11/14. doi: 10.1186/s13059-018-1567-1. PubMed PMID: 30419947; PubMed Central PMCID: PMC6231252.

5. Kim S, Cho H, Lee D, Webster MJ. Association between SNPs and gene expression in multiple regions of the human brain. Transl Psychiatry. 2012;2:e113. Epub 2012/07/27. doi: 10.1038/tp.2012.42. PubMed PMID: 22832957; PubMed Central PMCID: PMC3365261.

6. Kim Y, Xia K, Tao R, Giusti-Rodriguez P, Vladimirov V, van den Oord E, et al. A meta-analysis of gene expression quantitative trait loci in brain. Transl Psychiatry. 2014;4:e459. Epub 2014/10/08. doi: 10.1038/tp.2014.96. PubMed PMID: 25290266; PubMed Central PMCID: PMC4350525.

7. Gamazon ER, Badner JA, Cheng L, Zhang C, Zhang D, Cox NJ, et al. Enrichment of cis-regulatory gene expression SNPs and methylation quantitative trait loci among bipolar disorder

susceptibility variants. Mol Psychiatry. 2013;18(3):340-6. Epub 2012/01/04. doi: 10.1038/mp.2011.174. PubMed PMID: 22212596; PubMed Central PMCID: PMC3601550.

8. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592-9. Epub 2019/03/31. doi: 10.1038/s41588-019-0385-z. PubMed PMID: 30926968.

9. Sieberts SK, Perumal T, Carrasquillo MM, Allen M, Reddy JS, Hoffman GE, et al. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. 2019. doi: 10.1101/638544.

10. Bennett DA, Schneider JA, Buchman AS, Barnes LL, Boyle PA, Wilson RS. Overview and findings from the rush Memory and Aging Project. Curr Alzheimer Res. 2012;9(6):646-63. Epub 2012/04/05. PubMed PMID: 22471867; PubMed Central PMCID: PMC3439198.

11. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. J Alzheimers Dis. 2018;64(s1):S161-S89. Epub 2018/06/06. doi: 10.3233/JAD-179939. PubMed PMID: 29865057; PubMed Central PMCID: PMC6380522.

12. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012;9(6):628-45. Epub 2012/04/05. PubMed PMID: 22471860; PubMed Central PMCID: PMC3409291.

13. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology. 1984;34(7):939-44. PubMed PMID: 6610841.

14. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142. Epub 2018/08/08. doi: 10.1038/sdata.2018.142. PubMed PMID: 30084846; PubMed Central PMCID: PMC6080491.

15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. Epub 2009/05/20. doi: 10.1093/bioinformatics/btp324. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234.

16. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-8. Epub 2011/04/12. doi: 10.1038/ng.806. PubMed PMID: 21478889; PubMed Central PMCID: PMC3083463.

17. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297-303. Epub 2010/07/21. doi: 10.1101/gr.107524.110. PubMed PMID: 20644199; PubMed Central PMCID: PMC2928508.

18. Kotlar AV, Trevino CE, Zwick ME, Cutler DJ, Wingo TS. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. Genome Biol. 2018;19(1):14. Epub 2018/02/08. doi: 10.1186/s13059-018-1387-3. PubMed PMID: 29409527; PubMed Central PMCID: PMC5801807.

19. De Jager PL, Srivastava G, Lunnon K, Burgess J, Schalkwyk LC, Yu L, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17(9):1156-63. Epub 2014/08/19. doi: 10.1038/nn.3786. PubMed PMID: 25129075; PubMed Central PMCID: PMC4292795.

20. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215-6. Epub 2012/03/01. doi: 10.1038/nmeth.1906. PubMed PMID: 22373907; PubMed Central PMCID: PMC3577932.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75. doi: 10.1086/519795. PubMed PMID: 17701901; PubMed Central PMCID: PMC1950838.

22. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12):e190. doi: 10.1371/journal.pgen.0020190.

23. Ping L, Duong D, Yin L, Gearing M, Lah JJ, Levey AI, et al. Global Quantitative Analysis of the Human Brain Proteome in Alzheimer's and Parkinson's Disease. Scientific Data. 2018;in press. Epub in press.

24. Johnson ECB, Dammer EB, Duong DM, Yin L, Thambisetty M, Troncoso JC, et al. Deep proteomic network analysis of Alzheimer's disease brain reveals alterations in RNA binding proteins and RNA splicing associated with disease. Mol Neurodegener. 2018;13(1):52. Epub 2018/10/06. doi: 10.1186/s13024-018-0282-4. PubMed PMID: 30286791; PubMed Central PMCID: PMC6172707.

25. Mertins P, Tang LC, Krug K, Clark DJ, Gritsenko MA, Chen L, et al. Reproducible workflow for multiplexed deep-scale proteome and phosphoproteome analysis of tumor tissues by liquid chromatography-mass spectrometry. Nat Protoc. 2018;13(7):1632-61. Epub 2018/07/11. doi: 10.1038/s41596-018-0006-9. PubMed PMID: 29988108; PubMed Central PMCID: PMC6211289.

26. Wingo TS, Duong DM, Zhou M, Dammer EB, Wu H, Cutler DJ, et al. Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. J Proteome Res. 2017;16(9):3336-47. Epub 2017/07/12. doi: 10.1021/acs.jproteome.7b00324. PubMed PMID: 28691493; PubMed Central PMCID: PMC5698003.

27. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010;7(9):709-15. Epub 2010/08/17. doi: 10.1038/nmeth.1491. PubMed PMID: 20711195; PubMed Central PMCID: PMC3005310.

28. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods. 2013;10(7):623-9. Epub 2013/05/21. doi: 10.1038/nmeth.2483. PubMed PMID: 23685885; PubMed Central PMCID: PMC3821180.

29. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. Epub 2012/10/30. doi: 10.1093/bioinformatics/bts635. PubMed PMID: 23104886; PubMed Central PMCID: PMC3530905.

30. Logsdon B, Perumal TM, Swarup V, Wang M, Funk C, Gaiteri C, et al. Meta-analysis of the human brain transcriptome identifies heterogeneity across human AD coexpression modules robust to sample collection and methodological approach. bioRxiv. 2019. doi: 10.1101/510420.

31. Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60. Epub 2015/05/09. doi: 10.1126/science.1262110. PubMed PMID: 25954001; PubMed Central PMCID: PMC4547484.

32. Proitsi P, Lupton MK, Velayudhan L, Newhouse S, Fogh I, Tsolaki M, et al. Genetic predisposition to increased blood cholesterol and triglyceride lipid levels and risk of Alzheimer disease: a Mendelian randomization analysis. PLoS Med. 2014;11(9):e1001713. Epub 2014/09/17. doi: 10.1371/journal.pmed.1001713. PubMed PMID: 25226301; PubMed Central PMCID: PMC4165594.

33. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(Database issue):D493-6. Epub 2003/12/19. doi: 10.1093/nar/gkh103. PubMed PMID: 14681465; PubMed Central PMCID: PMC308837.

34. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science. 2018;361(6404):769-73. Epub 2018/08/04. doi: 10.1126/science.aaq1327. PubMed PMID: 30072576; PubMed Central PMCID: PMC6190714.

35. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73-9. Epub 2018/06/08. doi: 10.1038/s41586-018-0175-2. PubMed PMID: 29875488.

36. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. Connecting genetic risk to disease end points through the human blood plasma proteome. Nat Commun. 2017;8:14357. Epub 2017/02/28. doi: 10.1038/ncomms14357. PubMed PMID: 28240269; PubMed Central PMCID: PMC5333359.

37. Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. Variation and genetic control of protein abundance in humans. Nature. 2013;499(7456):79-82. Epub 2013/05/17. doi: 10.1038/nature12223. PubMed PMID: 23676674; PubMed Central PMCID: PMC3789121.

38. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Genomic variation. Impact of regulatory variation from RNA to protein. Science. 2015;347(6222):664-7. Epub 2015/02/07. doi: 10.1126/science.1260793. PubMed PMID: 25657249; PubMed Central PMCID: PMC4507520.

39. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016;534(7608):500-5. Epub 2016/06/17. doi: 10.1038/nature18270. PubMed PMID: 27309819; PubMed Central PMCID: PMC5292866.

40. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. Primate transcript and protein expression levels evolve under compensatory selection pressures. Science. 2013;342(6162):1100-4. Epub 2013/10/19. doi: 10.1126/science.1242379. PubMed PMID: 24136357; PubMed Central PMCID: PMC3994702.

41. McAlister GC, Nusinow DP, Jedrychowski MP, Wuhr M, Huttlin EL, Erickson BK, et al. MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. Anal Chem. 2014;86(14):7150-8. Epub 2014/06/14. doi: 10.1021/ac502040v. PubMed PMID: 24927332; PubMed Central PMCID: PMC4215866.

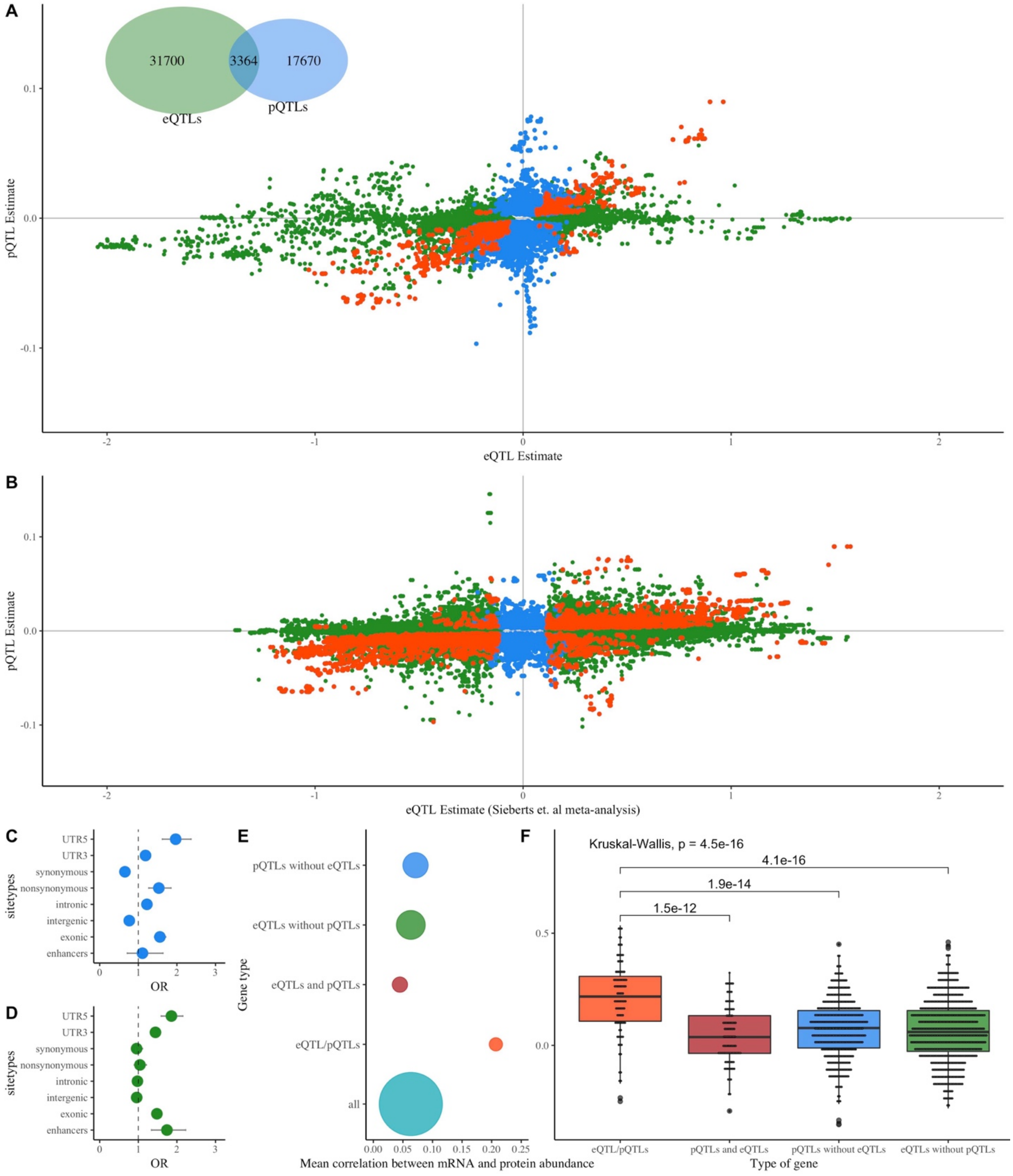

Figure 1. Protein and RNA Quantitative Locus Results. This figure summarizes the direction of effect and genomic annotation for pQTL and eQTL sites. (A) Comparison of eQTL and pQTL estimates. Each point represents one SNV tested against the abundance of the mRNA and protein of a single gene. eQTLs

(defined based on False Discovery Rate (FDR) < 0.05) are shown in green, pQTLs (defined based on FDR < 0.05) are shown in blue, and sites that are both an eQTL and a pQTL (i.e. eQTL/pQTLs) are shown in orange. (B) Comparison of Sieberts *et al.* meta-analysis eQTL estimates (N=1433) and our pQTL estimates. Each point represents one SNV tested against the abundance of the mRNA and protein of a single gene. eQTLs (defined based on FDR < 0.05) are shown in green, pQTLs (defined based on Bonferroni correction < 0.05) are shown in blue, and sites that are both an eQTL and a pQTL (i.e. eQTL/pQTLs) are shown in orange. (C) Results of Fischer's exact tests assessing the overlap of pQTLs and genic locations. Odds ratio (OR) estimates are shown with 95% confidence intervals. (D) Results of Fischer's exact tests assessing the overlap of eQTLs and genic locations. OR estimates are shown with 95% confidence intervals. (E) Mean correlation between mRNA and protein abundance for all genes, genes with eQTL/pQTLs (i.e. sites that are both an eQTL and a pQTL), genes with pQTLs and eQTLs but no eQTL/pQTLs, genes with pQTLs and no eQTLs, and genes with eQTLs and no pQTLs. The size of the point reflects the relative number of proteins within each gene type. (F) Comparison of genes with eQTL/pQTLs, genes with both eQTLs and pQTLs but no eQTL/pQTLs, genes with pQTLs and no eQTLs, and genes with eQTLs and no pQTLs. P-values from the significant pairwise comparisons are shown.