Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Genetic regulation of human brain proteome reveals proteins implicated in psychiatric disorders

Jie Luo
Zhejiang Academy of Agricultural Sciences
Mingming Niu
St. Jude Children's Research Hospital
Ling Li
University of North Dakota https://orcid.org/0000-0001-6295-2128
Dehui Kong
University of North Dakota
Yi Jiang
Wenzhou Medical University
Annie Shieh
University of Chicago
Lijun Cheng
University of Chicago
Gina Giase
University of Chicago
Kay Grennan
University of Chicago
Kevin White
Tempus Labs, Inc. and Institute for Genomics & Systems Biology
Chao Chen
Central South University https://orcid.org/0000-0002-5114-2282
Sidney Wang
The University of Texas Health Science Center at Houston
Dalila Pinto
Icahn School of Medicine at Mount Sinai https://orcid.org/0000-0002-8769-0846
Yue Wang
Virginia Polytechnic Institute and State University https://orcid.org/0000-0002-1788-1102
Chunyu Liu
SUNY Upstate Medical University https://orcid.org/0000-0002-5986-4415
Junmin Peng
St Jude Children's Research Hospital https://orcid.org/0000-0003-0472-7648
Xusheng Wang (xusheng.wang@und.edu)

University of North Dakota https://orcid.org/0000-0002-1759-9588

Article

Keywords: Proteome, Quantitative Trait Loci, QTLs, Protein Expression, Transcriptome, Gene Expression, Genome-wide Association Study, GWAS, Mass Spectrometry, Proteomics, Gene Regulation, Colocalization, Mediation, Causality, Pleiotropy

Posted Date: May 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1633422/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Psychiatric disorders are highly heritable yet polygenetic, potentially involving hundreds of risk genes. Genome-wide association studies (GWAS) have identified hundreds of genomic susceptibility loci for psychiatric disorders, but how these loci contribute to the underlying psychopathology and etiology remains elusive. Here we generated a deep human brain proteome by quantifying 11,672 proteins across 288 subjects using 11-plex tandem mass tag (TMT) coupled with two-dimensional liquid chromatography-tandem mass spectrometry (LC/LC-MS/MS). We identified 788 *cis*-acting protein quantitative trait loci (*cis*-pQTLs) associated with 883 proteins at a genome-wide false discovery rate (FDR) < 5%. In contrast to expression at transcript level and complex diseases that are found to be mainly influenced by noncoding variants, we found protein expression level tends to be regulated by non-synonymous variants. We also provided evidence of 487 shared regulatory signals between gene expression (i.e., *e*QTL) and protein abundance (i.e., *p*QTLs). Mediation analysis revealed that for most (64%) of the colocalized genes, the expression level of their corresponding proteins are regulated by *cis*-pQTLs via gene transcription. Causality analysis by Mendelian Randomization (MR) revealed 4 *cis*-pQTLs and 19 *cis*-eQTLs causally controlling schizophrenia (SCZ) GWAS loci, respectively. We further integrated multiple omic data together with network analysis to prioritize candidate genes for SCZ GWAS loci. Collectively, our results underscore the potential of proteome-wide linkage analysis for mechanistic understanding of psychiatric disorders.

Introduction

Psychiatric disorders are polygenic and progressive diseases, characterized by brain dysfunction, that are influenced by both genetic and environmental factors [1,2]. Schizophrenia (SCZ) and bipolar disorder (BP) are the two most common psychiatric disorders, with 12-month prevalences of ~ 2.5%[3,4] and ~ 1.9% worldwise [5,6], respectively. Psychiatric disorders impose a considerable economic great burden on society due to the early age of onset, chronicity, and lack of efficient treatments or prevention strategies [7,8]. The disorders share a common genetic basis, with the heritability being estimated at 70 to 90% [9–12]. Current treatments, such as antidepressants, antipsychotics, and neurostimulation, are only partially effective [13], and the development of better treatments is hindered by limited understanding of the underlying molecular mechanisms of psychiatric disorders.

Over the past decade, genome-wide association studies (GWAS) have successfully identified hundreds of genomic loci associated with psychiatric disorders [14–17]. However, we have little understanding of molecular mechanisms affecting the disorders for most of these genomic loci. Gene expression quantitative trait locus (i.e., eQTLs) has been used to study the genetic regulation of molecular phenotypes to identify targets implicated in psychiatric disorders [18–20], and other endophenotypes (e.g., methylation and chromatin activity) are also used to understand the complex genetic basis of psychiatric disorders [20,21]. Recently, multi-omic [22] and cell-type-specific data [23] were employed to dissect the molecular mechanisms underlying the disorders. Proteins are the main regulators of other

endophenotypes and changes in mRNA and protein levels are often not correlated [24,25]. Protein expression is regulated at multiple levels, including elaborate transcriptional and post-transcriptional regulation that affects RNA stability, protein translation, and protein turnover and degradation. Each of these regulations can be influenced by genetic variation. However, the genetic landscape of proteome-wide regulation in psychiatric disorders remains largely unexplored.

Liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) technology has become a powerful platform for identifying and quantifying of proteins [26]. Several attempts have recently been made to proteome-wide define genomic loci associated with protein expression in human cell lines [27], plasma [28], and post-mortem brain tissues [29,30] from Alzheimer's disease (AD). The combination of proteomics and genetics studies has generated valuable insights into the complex processes influencing human diseases. Proteins are intermediate phenotypes that provide insight into how genetic variants are mechanistically linked to diseases [31]. However, little is known regarding the impact of genetic variants on psychiatric disorders by modulating protein expression in the human brain.

To gain a better understanding of how genetic variants influence protein expression in the brain and ultimately impact psychiatric disorders, we perform a deep proteome and transcriptome profiling of the post-mortem frontal cortex from a human cohort (Fig. 1a, b), followed by genetic analysis to identify genomic loci associated with gene expression (i.e., eQTL) and protein expression (i.e., pQTL) (Fig. 1c) and colocalization analysis of pQTLs and eQTLs (Fig. 1d). To understand how these pQTLs and eQTLs contribute to the pathogenesis of psychiatric disorders, we further integrate SCZ GWAS loci with pQTLs and eQTLs to identify risk genes that causally control SCZ (Fig. 1d). We finally integrate multi-omic bulk data and single-cell transcriptomic data to prioritize risk genes/proteins for SCZ GWAS loci (Fig. 1e).

Results

Profiling and analysis of human brain proteome and transcriptome

To identify genetic variants influencing protein expression and understand its link to psychiatric disorders, we generated deep proteomic data of the frontal cortex from a total of 288 human post-mortem brain samples (**Supplementary Table 1**), including 211 normal, 29 BP, and 48 SCZ samples. These samples were analyzed by extensive fractionation (two-dimensional LC) and high-resolution, accurate mass, tandem mass spectrometry (LC/LC-MS/MS) to generate a deep human brain proteome (**Fig. 2a**). We identified and quantified a total of 19,272 proteins (14,221 genes) in at least one sample at the protein FDR < 1% (**Fig. 2b**) using 29 batches of 11-plexes tandem mass tags (TMT) experiments. Of these, 11,672 proteins (8,321 genes) were detected across all 288 samples (**Fig. 2b; Supplementary Table 2**). The vast majority of proteins (78.07%; 15,045/19,272) were detected in more than 25 batches (**Fig. 2c**). To our best knowledge, this is the deepest human brain proteome data to date available for studying the genetic regulation of protein expression. After extensive quality controls (QC) [32], we focused on 268 samples of high quality for the subsequent proteome-wide genetic regulation analysis. In addition, we also performed RNA sequencing of the frontal cortex from 416 human brain samples (**Supplementary Table 1**), and high-

density genotyping using whole-genome sequencing for all samples. After stringent QC and sample identity verification [33] and normalization of gene expression quantifications and genotype imputation, we detected a total of 17,160 expressed genes (**Supplementary Table 3**) and ~8.1 million single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) > 5% as genotypes.

We next compared our proteomic to transcriptomic data genetated from matched 264 samples. The quantified proteins cover ~70% of the range of mRNA expression detected by RNA-seq, indicating a deep coverage of our proteomic data (**Fig. 2d**). Those undetected proteins whose corresponding RNAs also had low expression signals are likely due to either not being translated into proteins in the brain or the concentration under the detection limit by the mass spectrometer. A moderate positive correlation between expression levels of mRNAs and proteins (r = 0.44) was observed (**Fig. 2e**), which is consistent with previous findings [24,25,27]. A subset of proteins showed high variability in expression (**Fig. 2f**), which were found to be mainly enriched in functional terms related to extracellular matrix (ECM) organization, blood microparticle, plasma lipoprotein particle, and integrin binding (**Extended Data Fig. 1a**), whereas lowly variable proteins were enriched in terms associated with the housekeeping functions, such as proteasome complex, regulation of mRNA stability, and regulation of cell cycle (**Extended Data Fig. 1b**).

Human brain proteome and transcriptome reveal the genetic architecture of expression regulation

To characterize genetic variants influencing the expression level of genes and proteins, we performed both proteome-wide and transcriptome-wide association analyses. To increase statistical power and reduce false positives [34,35], we removed variously measured and unmeasured confounding factors, such as experimental and technical batch effects (**Extended Data Fig. 2a**). By using the probabilistic estimation of expression residuals (PEER) program [36], we captured 99% of the hidden variance in proteomic data with 13 controlling factors (**Extended Data Fig. 2b**). Further correlation analysis indicated that the effect of various covariates on protein expression variation has been well-controlled (**Extended Data Fig. 2c, d**).

We first performed a genome-wide linkage analysis (**Extended Data Fig. 3**) of the expression of 11,672 proteins using the QTLtools program, identifying 788 *cis*-acting (or local acting) genomic loci (i.e., *cis*-pQTLs; within ±1 Mb from the transcriptional start site for each tested gene) that modulate 883 protein expression (i.e., pGenes) at the genome-wide FDR < 5% using the Storey q value method (**Fig. 3a; Supplementary Table 4a**). Note that we will use the term FDR and "q value" interchangeably for the linkage analysis hereafter. Most of the significant *cis*-pQTLs (10.8%; n = 180) are associated with more than one protein as opposed to 3.1% of the *cis*-eQTLs, suggesting that *cis*-pQTLs tend to be more pleiotropic than *cis*-eQTLs. We observed that 316 and 567 *cis*-pQTLs involved positive and negative effects on protein expression (q value < 5%; see methods), respectively. Of these, 44 *cis*-pQTLs have relatively large effect size (β > 0.5), including 14 and 30 *cis*-pQTLs with positive and negative effect size, respectively (**Fig. 3b**). In addition, we found that a trend for alleles with low frequency has a stronger effect on *cis*-pQTLs (**Fig. 3c**). We also detected 256 *trans*-acting (or distal-acting) loci that regulate 511

proteins at the genome-wide FDR < 5% (**Supplementary Table 4b;** p value < 2.33×10^{-11}; Bonferroni Correction), of which 19 (3.9%) loci harbor both *cis*- and *trans*-pQTLs.

We next performed genome-wide linkage analysis for gene expression (i.e., eQTL) of 416 frontal cortex samples that include nearly all of the samples (264/268; **Supplementary Table 1**) used for proteome-wide analysis. Similarly, by removing hidden confounding factors for 17,160 expressed genes using the PEER program, we identified 9,333 significant *cis*-eQTLs that involve 9,960 genes (i.e., eGenes) at the genome-wide FDR < 5% (**Extended Data Fig. 5a; Supplementary Table 5a**), which include 569 *cis*-eQTLs with large effect size (β > 0.5; **Extended Data Fig. 5b**). We also found that expression level of a total of 271 genes is regulated by 259 *trans*-eQTLs (**Supplementary Table 5b**). Positional enrichment analysis showed that 56.28% (5,605/9,960) of a significant *cis*-eQTLs cluster within 10 kb of the transcription starting site (TSS) of its target genes (**Extended Data Fig. 5c**). In contrast, we found that 15.13% of *cis*-pQTLs are in exonic regions compared to 5.43% in *cis*-eQTLs (**Fig. 3d**). Interestingly, the vast majority (75.63%) of exonic *cis*-pQTLs are non-synonymous variants compared to 33.64% of non-synonymous exonic *cis*-eQTLs. The observation indicates that coding variants have a unique and great impact on protein expression, different from the finding that *cis*-eQTLs tend to be found in the TSS region observed in this study and previous GWAS and eQTL studies [37,38]. The observation that *cis*-pQTLs are enriched in coding variants is also consistent with recent published proteome-wide association studies in AD [39] and in lymphoblastoid cell lines (LCLs) [40].

One notable example of *cis*-pQTLs is C14orf159 (UniProt ID: Q7Z3D6) with a mapping nominal p value of 1.57×10^{-72} (q value = 5.20×10^{-43}; **Fig. 3e**), the second most significant *cis*-pQTL. Compared to the alternative allele (T/T), the reference homozygous allele (C/C) increases the protein expression by 4.59-fold (**Fig. 3e, inset**). C14orf159 is a mitochondrial matrix protein, involved in mitochondrial metabolism. Although C14orf159 has not been directly implicated in any psychiatric disorders, the protein-protein interaction (PPI) network of C14orf159 indicates that it could be associated with SCZ as it links to five proteins whose functions have been associated with SCZ, including PSMB4, GLS, GLS2, LRRN4CL, and C8orf82 (**Fig. 3f**). Among these five proteins, three of them were found to be a *cis*-eGene, including C15orf40 (p value = 1.51×10^{-11}, q value = 1.94×10^{-6}), GLS2 (p value = 3.96×10^{-6}, q value = 4.09×10^{-3}, and OPLAH (p value = 6.56×10^{-5}, q value = 3.57×10^{-2}). For example, GLS2 is an enzyme that catalyzes the hydrolysis of glutamine into glutamate and ammonia, which is highly expressed in the brain. GLS2 was also found to be hypermethylated in SCZ male samples [41] and its expression is altered in the pathology of SCZ [42].

Among 256 *trans*-pQTLs identified in this study, 11 were found to modulate the expression of more than five proteins (**Extended Data Fig. 4**). For example, a *trans*-QTL (*rs77546871*) in WW domain-containing oxidoreductase (WWOX) protein regulates the expression of 27 downstream proteins (**Extended Data Fig. 4**; blue lines). WWOX is also a significant *cis*-pGene (p value = 9.19×10^{-10}, q value = 1.57×10^{-72}). WWOX has been implicated in signaling pathways regulating the central nervous system (CNS) development and neural differentiation [43], and dysfunction of this gene has been found to result in reduced GABA-ergic

inhibitory interneuron numbers in mice [44]. GWA studies have also identified WWOX as a risk gene for common neurodegenerative conditions, such as SCZ [45], AD [46], and autism [47].

As expected, *cis*-eQTLs with larger effect sizes are enriched for risk genes [48]. We found that 11 out of 569 eGenes with large effect size ($\beta > 0.5$) were associated with SCZ (**Extended Data Fig. 5b**). For example, the lead *cis*-eQTL for *GSTM5* (Glutathione S-Transferase Mu 5) was strongly associated with both gene expression (p value = 2.78×10^{-74}, q value = 3.01×10^{-48}) and protein expression (p value = 8.63×10^{-28}, q value = 3.52×10^{-21}) (**Extended Data Fig. 5d, e**). *GSTM5* is a member of the antioxidant glutathione S-transferase family, playing an important role in protective mechanisms against oxidative stress in SCZ pathogenesis [49]. Genetic polymorphisms of several members of the glutathione S-transferase family, such as *GSTM1*, *GSTT1*, *GSTP1*, and *GSTA1*, have been associated with the pathophysiology of SCZ [50].

Extensive colocalization between *cis*-pQTLs and *cis*-eQTLs

To investigate whether the association signals regulating gene and protein expression levels are driven by the same genetic variant, we performed colocalization analysis for 883 *cis*-pQTLs and 9,960 *cis*-eQTLs using the coloc program [51]. The colocalization analysis estimates five posterior probabilities (PP_0, PP_1, PP_2, PP_3, and PP_4) (**see methods**). We identified 724 pairs of significant *cis*-pQTLs and *cis*-eQTLs within 1 Mb region (upstream or downstream), harboring 664 unique genomic loci. Among these, we found 465 pairs of colocalized QTLs (i.e., 423 genomic loci) with evidence of the same *cis*-pQTL and *cis*-eQTL signals ($PP_4 > 0.50$; **Fig. 4a; Supplementary Table 6a**). Over-representation analysis found a significant enrichment (Fisher exact test; p-value = 5.8×10^{-3}) of colocalized *cis*-QTL signals. These colocalized pGenes are enriched for pathways associated with psychiatric disorders, such as carbon and diverse amino acid metabolisms, and peroxisome (**Extended Data Fig. 6a**),

The vast majority (79.57%; 370/465) of co-localized signals have the same direction of effect, with a significant positive correlation of effect size ($r = 0.40$; p value = 2.2×10^{-16}; **Fig. 4b**). This observation largely supports the fact that an increase in effect size of *cis*-eQTLs is indicative of an increase that of *cis*-pQTLs. In contrast, only a moderate correlation of expression levels between colocalized proteins and genes ($r = 0.21$; p value = 7.1×10^{-4}) was observed (**Fig. 4c**). Furthermore, we found that the effect size of colocalized *cis*-pQTLs is smaller than that of its corresponding *cis*-eQTLs (**Extended Data Fig. 6b,c**), in agreement with the previous observation in AD [39]. In addition, we found 83 *cis*-pQTLs showing a different *cis*-eQTL signal within the *cis* window ($PP_3 > 0.5$; not colocalization), and 30 QTLs being either pQTL-specific or eQTL-specific ($PP_1 > 0.5$ or $PP_2 > 0.5$). As an example, we illustrate serine racemase (SRR) protein to show the colocalization of *cis*-eQTL and *cis*-pQTL signals, which had a large posterior probability of colocalization ($PP_4 = 0.99$). SRR was identified as a significant *cis*-pQTL (p value = 2.64×10^{-19}, q value = 3.71×10^{-13}), and a significant *cis*-eQTL (p value = 4.88×10^{-87}, q value = 5.85×10^{-75}) (**Fig. 4d-f**), supported by a significant positive correlation of expression at gene and protein levels (**Fig. 4g**). SRR is a highly expressed protein in brain acting as an endogenous ligand of N-methyl d-aspartate

(NMDA) receptors. Disruption of the SRR protein was shown to reduce the function of NMDA receptors and is associated with the susceptibility to SCZ [52].

Mediation analysis reveals mediation of protein regulation

To explore whether the protein expression is dependently regulated by its *cis*-pQTL via the corresponding mRNA transcription (**Fig. 5a**), we performed a conditional mapping for the protein expression using the corresponding gene expression as a co-variate (**Extended Data Fig. 7a**). For 411 colocalized pGenes, we found that expression level of more than half of the proteins (262/411; 63.75%) is regulated by transcription (**Fig. 5b; Supplementary Table 7**), suggesting that these protein regulation were largely regulated through transcriptional mechanisms. The remaining 149 transcription-independent pGenes might be mediated through other unknown genetic mechanisms, such as post-transcriptional processes, translation mechanisms, or protein degradation.

This transcription-dependent protein regulation was supported by a modest correlation (r = 0.34) between transcripts and proteins (**Fig. 5c**), which is significantly higher (p value = 2.2×10^{-16}) than those transcription-independent pGenes (r = 0.14). However, the effect size of transcription-mediated pGenes was significantly lower as compared to that of transcription-independent pGenes (**Fig. 5d**; 0.11 vs 0.18; p value = 6.5×10^{-6}), suggesting direct genetic effects on protein abundance tend to be stronger than the mediation effects. As expected, the vast majority of transcription-dependent pQTLs are found to be in the genomic regulatory regions, whereas transcription-independent pQTLs are enriched in exonic regions (**Extended Data Fig. 7b**).

To illustrate the transcription-dependent regulation, we highlighted an example of transient receptor potential cation channel subfamily V member 2 (TRPV2), an ion channel protein. TRPV2 showed a signficant *cis*-pQTL (p value = 2.93×10^{-45}, q value = 2.11×10^{-27}), but the signal was abolished after conditioning on gene expression as a co-variate. TRPV2 level at gene level is also regulated by a significant *cis*-eQTL (p value = 3.70×10^{-97}, q value = 2.69×10^{-58}) (**Fig. 5e**). The expression levels of transcript and protein are highly correlated (**Fig. 5e; inset**). For the transcription-independent regulation, a significant *cis*-pQTL was found to regulate DHTKD1 protein abundance without being regulated by its transcription (**Fig. 5f**). As expected, there is a lack of correlation between protein and transcript abundance (**Fig. 5f; inset**). Interestingly, DHTKD1 protein was found to be a transcription-mediated regulation in mouse liver [53].

Causal contribution of *cis*-pGenes and *cis*-eGenes to psychiatric disorders

We next sought to identify genomic loci that associate psychiatric disorders through genetic effects on gene and protein expression. We evaluated 313 SCZ genomic loci identified by a meta-analysis of data from a recent publication by the Psychiatric Genomics Consortium (PGC) [54]. We used the summary-based Mendelian randomization (SMR) analysis coupled with the heterogeneity independent instruments (HEIDI) test [55]. We identified 4 pGenes that passed both the HEIDI heterogeneity test ($P_{HEIDI} > 0.05$) and

the SMR significance threshold of $P_{SMR} < 6.3 \times 10^{-5}$ (0.05/790; p value = 0.05 corrected by the total number of pGenes) (**Fig. 6b; Supplementary Table 8a**). We also detected 19 eGenes that passed the HEIDI heterogeneity test ($P_{HEIDI} > 0.05$) and the significance threshold of $P_{SMR} < 5.3 \times 10^{-6}$ (0.05/9,495; corrected by the total number of eGenes) (**Fig. 6b; Supplementary Table 8b**). Among these pGenes and eGenes with significant *cis*-pQTL and *cis*-eQTL, 2 pGenes and 13 eGenes were also prioritized for the SCZ GWAS loci [54]. Note that SMR analysis cannot distinguish causality from pleiotropy. In addition, one protein (BTN2A1) and 21 genes showed $P_{HEIDI} < 0.05$ from the HEIDI test, suggesting that expression and GWAS are likely to be driven by different variants in the same linkage disequilibrium block.

An example of the causality effect of pGene on SCZ is DARS2, a mitochondrial aspartyl-tRNA synthetase. The SMR analysis detected a significant association between DARS2 protein expression and SCZ ($P_{SMR} = 1.66 \times 10^{-5}$ and $P_{HEIDI} = 0.16$). DARS2 is a significant pGene (p value = 5.88×10^{-21}, q value = 4.93×10^{-15}) (**Fig. 6c**), which is a highly expressed in the brain and has been identified as the strongest causal gene of SCZ in an independent GWAS [17]. As an example of a significant causal association between eGenes and SCZ, CUL9 (cullin-9) exhibited a significant association between gene expression and SCZ, with $P_{SMR} = 1.78 \times 10^{-7}$ and $P_{HEIDI} = 0.11$ (**Fig. 6d**). CUL9 is a parkin-like ubiquitin ligase that has been prioritized as a candidate gene for an SCZ GWAS locus [54].

Integrative analysis prioritizes proteins for psychiatric disorders

Previous studies have shown that molecular endophenotypes with a significant QTL (e.g., eQTLs, methylation QTLs (mQTLs), and pQTLs) tend to influence complex diseases [38], and they can be harnessed to prioritize risk genes in GWAS [56]. Although it is currently difficult to pinpoint a causal gene for GWAS loci, the prioritized genes/proteins would be plausible candidates to unravel biological mechanisms underlying the GWAS associations. In this study, we attempt to establish a framework to systematically prioritize risk genes for 294 significant GWAS loci (p value < 5×10^{-8}) and 290 suggestive loci ($5 \times 10^{-8} < p$ value < 1×10^{-6}) with small effect size [53,55].

We sought to combine multiple data sets to prioritize genes/proteins for GWAS loci using order statistics (**Fig. 7a**). Five data sets were included for the prioritization, including pGenes ranked by *cis*-pQTL p values, eGenes ranked by *cis*-eQTL p values, co-localization between *cis*-pQTLs and *cis*-eQTLs ranked by PP_4 values, and disease relevance score with SCZ by the GeneCards database, and connectivity score ranked by the number of downstream SCZ risk genes in protein-protein interaction (PPI) network (**see Methods; Extended Data Fig. 8a**). To derive the PPI network connectivity score, we first extracted high-confidence PPI with a score ≥ 700 (mean score: 295, range: 150–999) and kept those nodes with *cis*-pGenes or *cis*-eGenes (**Extended Data Fig. 8b; Supplementary Table10**), yielding an SCZ network with 2,011 nodes and 3,118 protein-protein interactions (**Extended Data Fig. 8b**). We used order statistics to generate a final ranking score, followed by identifying candidate genes for GWAS loci. To further assess the cell-type-specific differential expression of these ranked proteins, we also downloaded single-cell transcriptomic data generated from 48 postmortem human prefrontal cortex samples, including 24

schizophrenia cases and 24 controls [57] and mapped differential expression events and expression abundance of 20 cell types to our ranked proteins (**Supplementary Table 10**). The final ranking result showed that the top-ranked 60 proteins include 8 candidate genes from the 313 significant PGC SCZ phase 3 GWAS loci (**Fig. 7b; Extended Data Fig. 9; Supplementary Table 9**) and 2 additional candidate genes for other SCZ GWAS loci (**Supplementary Table 10**). Single-cell transcriptomic data [57] support that 43 out of the top 60 proteins are schizophrenia differentially expressed genes (SZTR, defined as "schizophrenia transcriptional resilience").

A major challenge in GWAS is unable to detect loci with small effects due to low statistical power [58]. We next leveraged our prioritized proteins to identify candidate risk genes for 311 suggestive GWAS loci with a smaller effect size. We found 2 out of the top 60 ranked genes in suggestive GWAS loci (**Fig. 7b**), For example, PPP2R4 is prioritized as a candidate gene for a GWAS locus *rs6478858*. This is also supported by evidence that PPP2R4 is functionally associated with SCZ risk genes that include MAPT, PPP2R2A, FOXO3, AKT3, RERE, RSMO6, PSMA4, and MAD1L1 (**Fig. 7c**). PPP2R4 showed a colocalized significant *cis*-pQTL (p value = 2.70×10^{-15}, q value = 1.58×10^{-8}) and *cis*-eQTL (p value = 8.92×10^{-16}, q value = 2.31×10^{-12}) (**Fig. 7d**). The reference homozygous allele (G/G) increases protein and gene expression levels by 1.16-fold and 1.14-fold compared to the homozygous alternative allele (A/A), respectively (**Fig. 7e, f**). Single-cell transcriptomic data showed that PPP2R4 decreases the expression level in microglia and is differentially expressed in AD compared to control samples (**Extended Data Fig. 9**). These results suggest that our protein prioritization is a powerful strategy for identifying candidate genes in GWAS loci with small effects.

Discussion

In this study, we performed proteome-wide and transcriptome-wide association studies of post-mortem brain from a human cohort of controls and patients with psychiatric disorders. We characterized the genetic architecture of human gene and protein regulation by discovering 9,333 *cis*-eQTLs and 790 *cis*-pQTLs regulating gene and protein expression respectively. Our causality analysis highlighted eGenes and pGenes that are functionally implicated in psychiatric disorders. Prioritization analysis further revealed proteins as candidate risk genes for GWAS SCZ loci. Taken together, the findings of this study increases our understanding of the genetic regulation of gene and protein expression in the human brain, and shed light on the underlying molecular mechanisms involved in psychiatric disorders.

One of the strengths of this study was that we comprehensively defined the landscape of genetic regulation of protein expression by quantifying 11,672 unique proteins across all 288 human brain samples. We quantified a total of 19,160 unique proteins from at lease one batch of proteomic experiment (Fig. 2b). To the best of our knowledge, this is the deepest human brain proteome ever reported for such a large human cohort. Compared to a recent large-scale human brain proteomic study [30], our proteomic data detected ~ 43.08% (11,672 v.s. 8,356) more unique proteins. This deep proteomic data provides an opportunity to comprehensively evaluate genetic loci regulating protein expression even

with lowly expressed proteins, which otherwise remain undiscovered or poorly characterized with shallow proteomic data.

The availability of transcriptome and proteome of the matched tissue of the same human cohort in this study provided an excellent opportunity to investigate the regulations of gene and protein expression. We provided evidence that the vast majority (423/664) of colocalized cis-eQTLs and *cis*-pQTLs showed the same direction of the regulation (Fig. 4c), but the effect size of *cis*-pQTLs is generally smaller than *cis*-eQTLs (**Extended Data** Fig. 6b,c), which is in agreement with a previous report [40]. One of the explanations is that post-transcriptional buffering leads to less variation in protein expression levels [25]. We also identified three *cis*-QTLs showed an inconsistent direction of effect on eGenes and pGenes (i.e., TBC1D9B, NTPCR, and B3GAT3) (Fig. 5b; **Extended Data Fig. 10**). NTPCR had strong significant *cis*-eQTL (p value = 4.12×10^{-90}) and *cis*-pQTL (p value = 7.72×10^{-58}), but showed a negative correlation between gene expression and protein expression, suggesting a likelihood of the pleiotropic effect of the variant.

Another advantage of measuring gene and protein expression from the same tissue is that it allows us to investigate the mediation of protein regulation. Although protein expression often correlates poorly with transcript levels, we observed most of the pGenes (465/724) colocalized with eQTLs signals, and the expression level of most of these colocalized pGenes are modulated by gene transcription. This observation is consistent with a previous report that the vast majority of genetic variants controlling gene expression also influence protein abundance [53]. In addition, we also observed about some of the pGenes are regulated in the transcription-independent manner. For these pGenes, pos-transcriptional regulation often buffers differences in the genetic regulation of protein abundance from mRNA levels [25].

Only a limited number of proteome-wide association studies to date have been conducted in human brain tissue. For example, a proteome-wide association study (PWAS) identified 1,475 proteins associated with AD [38]. Although previous studies revealed shared genetic bases between neurodegenerative disorders (e.g., AD) and psychiatric disorders (e.g., SCZ and BP), an important question is the extent to which these diseases share the same gene and protein regulations. When comparing our 883 *cis*-pGenes (768 genes) with 1,469 proteins detected by PWAS and found that more than half (73.31%; 563/768) of the brain pGenes were detected in both data sets, indicating similar underlying mechanisms involved in the genetic regulation of protein expression in the brain despite the proteomic data being profiled from two different human cohorts.

While we measured the protein expression in the frontal cortex, a human brain region is still highly heterogeneous, containing different cell types [59]. Recent advances in single-cell transcriptomics have demonstrated the feasibility of identifying cell-type-specific eQTLs [60], which allows us to characterize the cellular specificity of genetic regulation of gene expression. In this study, we might be able to define cell-type-specific pQTLs by computationally deconvoluting sample-wise cell-type-specific expression from our bulk proteomic data. For example, CIBERSORTx [61] was developed for deriving a signature matrix and sample-wise deconvolution from the single-cell transcriptomic data. Although single-cell proteomics is

still in its infancy, several promising technologies are being explored, such as nanoTOPS and SCoPE-MS [62]. For example, nanoTOPS is capable of identifying ~ 2,000 proteins at 100-μm spatial resolution [63].

In summary, we provided a comprehensive resource of protein expression dynamics in the brain across a human cohort with psychiatric disorders. We defined the landscape of the genetic regulation of protein expression in brain, highlighting a large set of variants and targets involving in molecular mechanisms underlying psychiatric disorders. We developed a framework to investigate the causal link of eQTLs/pQTLs to genomic loci detected in the larger meta-GWAS study. We believe that integrating data of GWAS and large-scale omic data provide a new avenue to identify novel risk genes for GWAS loci, thereby providing important insights into the pathogenesis of psychiatric disorders.

Online Methods

Human postmortem brain tissue

For proteome profiling, a total of 288 well-characterized postmortem human brain samples (179 males, 109 females) from the Stanley Medical Research Institute (SMRI) and Banner Sun Health Research Institute (BSHRI) were used for this study. These samples were collected from 210 neurotypical controls, 49 individuals with schizophrenia (SCZ), and 29 individuals with bipolar disorder (BP). The samples include 282 Caucasians, 1 Hispanic, 1 African American, 3 Asian American, and 1 unknown. For transcriptome profiling, RNA-Seq data from 416 samples (262 males and 154 females). More detailed information about the specimens is provided in **Supplementary Table 1.**

Brain tissue lysis and protein quantification

Frozen tissues from the frontal cortex were obtained from controls and patients with SCZ and BP. The tissues were weighed and homogenized in lysis buffer (50 mM HEPES, pH 8.5, 8 M urea, and 0.5% sodium deoxycholate, 100 μl buffer per 10 mg tissue) with a 1×PhosSTOP phosphatase inhibitor cocktail (Sigma-Aldrich). The total protein concentration of each sample was measured by the BCA Protein Assay Kit (Thermo Fisher Scientific), and confirmed by Coomassie-stained short SDS gels.

Protein digestion and TMT labeling

We used our previously optimized protocol [64,65] for this analysis. In brief, quantified protein samples (~0.3 mg in the lysis buffer with 8 M urea) were proteolyzed with Lys-C (Wako, 1:100 w/w) at room temperature for 2 h, diluted 4-fold to reduce urea to 2 M, and digested by trypsin (Promega, 1:50 w/w) at room temperature overnight. The digestion was terminated by the addition of 1% trifluoroacetic acid, followed by centrifugation. The supernatant was desalted with Sep-Pak C18 cartridge (Waters), and then dried by speedvac. Each sample was resuspended in 50 mM HEPES, pH 8.5, labeled with 11-plex TMT reagents, mixed equally, and desalted again for subsequent fractionation. We used 0.1 mg protein per sample. A total of 29 batches of 11-plex TMT experiments were performed.

Extensive two-dimensional LC/LC-MS/MS

The pooled TMT labeled samples were fractionated using offline basic pH reversed-phase chromatography (HPLC), and followed by acidic pH reverse phase LC-MS/MS analysis [66,67]. For the offline basic HPLC, we generated 40 concatenated fractions for each batch. We performed the offline LC run (~3 hr gradient) on an XBridge C18 column (3.5 μm particle size, 4.6 mm x 25 cm, Waters; buffer A: 10 mM ammonium formate, pH 8.0; buffer B: 95% acetonitrile, 10 mM ammonium formate, pH 8.0) [64]. For the acidic pH LC-MS/MS analysis, each fraction was run sequentially on a column (75 μm x 15-30 cm, 1.9 μm C18 resin from Dr. Maisch GmbH, 65º C to reduce backpressure) interfaced with a Orbitrap Fusion and Q Exactive HF MS (Thermo Fisher). Peptides were eluted by a 1.5-2 h gradient (buffer A: 0.2% formic acid, 5% DMSO; buffer B: buffer A plus 65% acetonitrile). MS settings included MS1 scans (60,000 resolution, 1 x 10^6 AGC and 100 ms maximal ion time) and 20 data-dependent MS2 scans (410-1600 *m/z*, 60,000 resolution, 1 x 10^5 AGC, ~105 ms maximal ion time, HCD, 38% normalized collision energy, 1.0 *m/z* isolation window with 0.2 *m/z* offset, and ~15 s dynamic exclusion).

Identification of proteins by database search with JUMP software

We performed peptide identification with the JUMP search engine to improve the sensitivity and specificity [68]. JUMP searched MS/MS raw data against a composite target/decoy database [69] to evaluate FDR. The target human protein sequences (83,955 entries) were downloaded from the UniProt database. The decoy database was generated by reversing to generate a decoy database that was concatenated to the target database. FDR was estimated by the ratio of the number of decoy matches and the number of target matches. Major parameters included precursor and product ion mass tolerance (± 15 ppm), full trypticity, static mass shift for the TMT tags (+229.16293) and carbamidomethyl modification of 57.02146 on cysteine, dynamic mass shift for Met oxidation (+15.99491) and Ser/Thr/Tyr phosphorylation (+79.96633), maximal missed cleavage (*n* = 2), and maximal modification sites (*n* = 3). Putative PSMs were filtered by mass accuracy and then grouped by precursor ion charge state and filtered by JUMP-based matching scores (Jscore and ΔJn) to reduce FDR below 1% for proteins during the whole proteome analysis. If one peptide could be generated from multiple homologous proteins, based on the rule of parsimony, the peptide was assigned to the canonical protein form in the manually curated Swiss-Prot database.

Protein quantification by JUMP software suite

Protein quantification was carried out using the following steps [70]. We first extracted TMT reporter ion intensities of each PSM and corrected the raw intensities based on isotopic distribution of each labeling reagent. We discarded PSMs with low intensities (i.e., the minimum intensity of 1,000 and median intensity of 5,000). After normalizing abundance with the trimmed median intensity of all PSMs, we calculated the mean-centered intensities across samples (e.g., relative intensities between each sample and the mean) and summarized protein relative intensities by averaging related PSMs. Finally, we derived

protein absolute intensities by multiplying the relative intensities by the grand mean of the three most highly abundant PSMs. Log_2-transformed data were used for the subsequent PEER factor analysis.

Genotype data

The raw genotype data were called and merged from three different platforms: whole-genome sequencing, microarray, and RNA-seq. The quality-controlled genotype data were then prepared by prephasing using Eagle2 [71]. The prephased data were imputed using Minimac3 [72] and the Haplotype Reference Consortium (HRC) [73]. After imputation, we filtered genotypes with MAF > 5%. Finally, a total of 8,101,465 SNPs were used for subsequent QTL mapping.

Transcriptome profiling by RNA-seq

We used different RNA preparation techniques for human brain samples from the SMRI and the BSHRI collections. For SMRI brain samples, total RNA was isolated for SMRI samples through organic extraction. Briefly, approximately 50-60 mg of brain tissue was homogenized by polytron probe in Trizol. Total RNA was precipitated with isopropanol at room temperature, pelleted, washed with 75% ethanol, and resuspended in DEPC treated water. Quantification was performed by obtaining OD at A260, and quality assayed by agarose gel electrophoresis. For BSHRI brain samples, total RNA was mixed with ethanol and applied to a miRNeasy mini-column. Columns were treated with the RNase-free DNase digestion set (Qiagen), then washed with the appropriate miRNeasy mini kit buffers. Total RNA was eluted with RNase-free water. All total RNA samples that passed QC for library generation had a concentration of ≥ 100 ng/uL, assayed by the Qubit 2.0 RNA BR Assay or Xpose, and a RIN score ≥ 5.5, assayed by the Bioanalyzer RNA 6000 Nano assay kit.

RNA-seq data analysis

All FASTQ files were trimmed for adapter sequence and low base call quality (Phred score < 30 at ends) using cutadapt (v1.12) and then aligned to the GRCH37 (i.e., hg19) reference genome with STAR (2.4.2a) [74] using GENCODE gene annotations. BAM files were sorted using samtools (v1.3) [75]. Gene expression levels were quantified using RSEM (v1.2.29) [76].

PEER factor analysis

We employed the Probabilistic Estimation of Expression Residuals (PEER) method [36] to remove hidden batch effects and other confounding effects for both transcriptomic and proteomic data. This PEER method was designed to maximize the number of *cis*-eQTLs. A total of 30 and 13 covariate factors were identified in transcriptomic and proteomic data, respectively. These covariant factors captured ~99% of the total variance in both transcriptomic and proteomic data. We used the inverse normal-transformed PEER-processed residuals for downstream QTL mapping.

Association analysis

We performed eQTL/pQTL mapping for both gene and protein expression using the QTLtools program (Version 1.2) [77] with the permutation number of 10,000. The top variant was selected as the QTL for the protein/gene. eQTLs/pQTLs were defined as *cis* (local) if the QTL was within 1 Mb on either side of the TSS, whereas eQTLs/pQTLs were defined as *trans* (distal) if the peak association was at least 5 Mb outside of the exon boundaries. We used beta distribution-adjusted empirical *p* values to estimate the *q* value by using the qvalue R package. Significant *cis*-eQTLs and *cis*-pQTLs were controlled by the *q* value < 5%. Due to the large number of analyses for calculating *trans*-eQTL and *trans*-pQTL, we used the conservative Bonferroni-corrected *p* value of 0.05. Positive effect means the increases of expression level in the presence of the reference allele, whereas the negative effect indicates the decrease of the expression level in the presence of the alternative allele.

Functional annotation of QTLs

ANNOVAR [78] was used to functionally annotate the leading SNP of a QTL. RefSeq (hg19_refGene.txt) was used to annotate SNPs. The functional consequence (synonymous, missense) of coding SNP was also determined. We summed stop-gain, stop-loss, and start-lost as missense mutation.

Co-localization analysis

We used the coloc R package [51] to analyze the colocalization between *cis*-eQTls and *cis*-pQTLs. A window size of 500 kb on either side of the pQTL was used. The coloc program uses a Bayesian model to determine posterior probabilities for five mutually exclusive hypotheses: no association of any variant in the region with either *cis*-pQTL and *cis*-eQTL (H_0); association with *cis*-pQTL but not *cis*-eQTL (H_1), association with *cis*-eQTL but not *cis*-pQTL (H_2), two different QTLs (H_3); and a shared QTL for both gene and protein expression (H_4). These hypotheses were tested to produce the posterior probabilities, PP_i ($i \in [0,4]$). We consider $PP_4 > 0.5$ to be significant evidence of colocalization.

Causal/pleiotropic analysis of the effect of pGenes/eGenes on SCZ GWAS

We applied SMR [55] to test the causal/pleoitropic effect between genes/proteins and diseases using summary statistics from GWAS. In this study, SMR used SCZ GWAS loci as instrument variables and gene/protein expression levels as exposure to test whether the causal effect of a specific variant on the SCZ GWAS signal acts via a specific gene/protein. The SCZ GWAS loci were downloaded from the recently published Psychiatric Genomics Consortium (PGC) [54]. SMR first performs the HEIDI (heterogeneity in dependent instruments) test to exclude the GWAS and pQTLs/eQTLs caused by genetic linkage. The HEIDI threshold ($P_{HEIDI} > 0.05$) and the SMR FDR-corrected threshold (adjusted $P_{SMR} < 0.05$) were used.

Mediation analysis

The mediation analysis was performed to identify eGenes that are likely to be a causal mediator between the pQTL and protein expression it regulates. We implemented this analysis with Perl language based on

the conditional mapping function provided in the QTLtools. If the expression level of a protein is regulated by its *cis*-pQTL via the corresponding eGene as a mediator, the *p* value in the conditional pQTL mapping model (transcript as a covariate) should significantly decrease or abolish the pQTL effect. To assess whether the *p*-value significantly drops for a given mediator on a *cis*-pQTL, a null distribution of *p* values was estimated by randomly permuting sample labeling of the eGene. For each protein, this analysis produces 1 *cis*- nominal *p*-value and 1,000 permutated nominal *p*-values. The combined *p*-values are then converted into z-scores. We consider a potential causal mediator with a z-score ≤ -4.26 (*p* value = 1 x 10^{-5} significance level; 0.01 / 1,000 multiple correction tests).

PPI network of SCZ risk genes

GWASs have identified hundreds of GWAS loci associated with SCZ and BP. To manually curate a catalog of SCZ and BP risk genes, we extracted a total of 1,194 risk genes in GWAS loci reported by a list of 9 papers (**Supplementary Table 10**). Detailed information about the studied subjects, diagnosis, genotyping, quality control, and statistical analyses is provided in the original papers. To create a PPI network of SCZ risk genes, we downloaded the STRING database [55] and extracted physically binding protein-protein interactions with a score ≥ 700, yielding a network of SCZ risk genes

Prioritization analysis

We employed order statistics to integrate multiple datasets for prioritizing genes/proteins in GWAS loci [79,80]. A total of 5 individual data sets with were used for this analysis: (1) pQTL data, ranked by the nominal *p* value; (2) pQTL data, ranked by the nominal *p* value; (3) colocalization between pQTLs and eQTLs; ranked by the PP_4 values generated by the coloc program; (4) GeneCards disease-relevant score, ranked by the scores provided by GeneCards [81]; (5) Interaction with known SCZ-GWAS genes: ranked by the number of SCZ proteins and/or genes were connected to it. The final integrative protein ranking was generated by the order statistics.

Declarations

Acknowledgments

This study was supported by NIH R01MH110920 grant (C.L.). This work was also partially UND Vice President for Research & Economic Development (VPRED) seed program (X.W.). J.L. was supported by funding from Natural Science Foundation of Zhejiang Province (LY20C150004) and Key Research and Development Program of Zhejiang (2021C02052). C.C. was supported by National Natural Science Foundation of China (82022024 and 31970572).

Author Contributions

X.W., conceived and led the project with C.L., and J.P.. J.L., L.L., and D.K. developed and implemented all scripts for data analyses. M.N. and J.P. performed the mass spectrometry-based proteomic experiment.

Y.J. and S. W. performed quality control on genotypic data. A.S, L.C., D.F., M.B., G.G., K.G., K.W. carried out transcriptomic experiments. X.W., C.L., C.C., S.W., D.P., and Y.W. contributed to the general scientific discussion and interpretation of the results. C.L. and K.W. provided human specimens and genotypic data. X.W., L.L., and J.L. wrote the manuscript.

Declaration of Interests

The authors declare no competing interests.

Data availability

The raw mass spectrometry data and RNA-seq used in this study are available in the Synapse database under accession code syn26231732.

Code availability

Data analyses were performed in LINUX shell, Perl (v5.18.4), and R (v4.0.4; https://www.r-project.org/). RNA-seq data were mapped to the human reference genome (GRCH37) using the following software tools: cutadapt (v1.12) (https://github.com/marcelm/cutadapt),

STAR (2.4.2a) (https://github.com/alexdobin/STAR/releases), samtools (v1.3) (https://sourceforge.net/projects/samtools/files/samtools/1.3/), RSEM (v1.2.29) (https://github.com/deweylab/RSEM/releases/tag/v1.2.29). Genotype were processed by Eagle2 (https://alkesgroup.broadinstitute.org/Eagle/), Minimac3, (https://genome.sph.umich.edu/wiki/Minimac3). Annotation using ANNOVAR (https://annovar.openbioinformatics.org/en/latest/user-guide/startup/), Proteomic data were processed with JUMP software (https://github.com/JUMPSuite/JUMP). QTL mapping were performed using PEER (https://github.com/PMBio/peer), QTLtools (https://qtltools.github.io/qtltools/), coloc (https://github.com/cran/coloc), SMR (https://yanglab.westlake.edu.cn/software/smr/), and qvalue (https://github.com/StoreyLab/qvalue). STRING database for the network analysis (https://string-db.org/).

References

1. Sullivan, P. F. *et al.* Psychiatric genomics: An update and an Agenda. American Journal of Psychiatry **175**, 15–27 (2018).
2. Lee, P. H. *et al.* Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. Cell **179**, 1469–1482.e11 (2019).
3. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. PLoS Medicine vol. 2 0413–0433 (2005).
4. Pilon, D. *et al.* Prevalence, incidence and economic burden of schizophrenia among Medicaid beneficiaries. Current Medical Research and Opinion **37**, 1811–1819 (2021).

5. Merikangas, K. R. *et al.* Prevalence and correlates of bipolar spectrum disorder in the World Mental Health Survey Initiative. Archives of General Psychiatry **68**, 241–251 (2011).
6. Carvalho, A. F., Firth, J. & Vieta, E. Bipolar Disorder. New England Journal of Medicine **383**, (2020).
7. Cloutier, M. *et al.* The Economic Burden of Schizophrenia in the United States in 2013. J Clin Psychiatry **77**, (2016).
8. Cloutier, M., Greene, M., Guerin, A., Touya, M. & Wu, E. The economic burden of bipolar I disorder in the United States in 2015. J Affect Disord **226**, (2018).
9. Gordovez, F. J. A. & McMahon, F. J. The genetics of bipolar disorder. Molecular Psychiatry vol. 25 544–559 (2020).
10. Guan, J., Cai, J. J., Ji, G. & Sham, P. C. Commonality in dysregulated expression of gene sets in cortical brains of individuals with autism, schizophrenia, and bipolar disorder. Translational Psychiatry **9**, (2019).
11. Lichtenstein, P. *et al.* Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet **373**, 234–239 (2009).
12. Escudero, I. & Johnstone, M. Genetics of Schizophrenia. Current Psychiatry Reports vol. 16 (2014).
13. Krystal, J. H. & State, M. W. Psychiatric disorders: Diagnosis to therapy. Cell vol. 157 201–214 (2014).
14. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. Nature **511**, 421–427 (2014).
15. Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics **50**, 381–389 (2018).
16. Geschwind, D. H. & Flint, J. Genetics and genomics of psychiatric disease. Science vol. 349 1489–1494 (2015).
17. Ruderfer, D. M. *et al.* Genomic dissection of bipolar disorder and Schizophrenia, including 28 subphenotypes. Cell **173**, 1705–1715.e16 (2018).
18. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nature Neuroscience vol. 19 1442–1453 (2016).
19. Huckins, L. M. *et al.* Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. Nature Genetics **51**, 659–674 (2019).
20. Gusev, A. *et al.* Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nature Genetics **50**, 538–548 (2018).
21. Perzel Mandell, K. A. *et al.* Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. Nature Communications **12**, (2021).
22. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. Science (1979) **362**, (2018).
23. Bryois, J. *et al.* Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders Introduction. doi:10.1101/2021.10.09.21264604.

24. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. Molecular and Cellular Biology **19**, 1720–1730 (1999).
25. Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. Cell **165**, 535–550 (2016).
26. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. Nature **422**, 198–207 (2003).
27. Wu, L. *et al.* Variation and genetic control of protein abundance in humans. Nature **499**, 79–82 (2013).
28. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. Nature **558**, 73–79 (2018).
29. Wingo, T. S. *et al.* Brain proteome-wide association study implicates novel proteins in depression pathogenesis. Nature Neuroscience **24**, 810–817 (2021).
30. Wingo, A. P. *et al.* Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. Nature Genetics **53**, 143–146 (2021).
31. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. Nature Reviews Genetics **22**, 19–37 (2021).
32. Li, L. *et al.* SMAP is a pipeline for sample matching in proteogenomics. Nature Communications **13**, 744 (2022).
33. Jiang, Y. *et al.* DRAMS: A tool to detect and re-align mixed-up samples for integrative studies of multi-omics data. PLoS Computational Biology **16**, (2020).
34. Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis HHS Public Access. Nat Neurosci **21**, 1117–1125 (2018).
35. Lipska, B. K. *et al.* Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia. Biol Psychiatry **60**, (2006).
36. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nature Protocols **7**, 500–507 (2012).
37. Elkon, R. & Agami, R. Characterization of noncoding regulatory DNA in the human genome. Nature Biotechnology **35**, 732–746 (2017).
38. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (1979) **348**, 648–660 (2015).
39. Robins, C. *et al.* Genetic control of the human brain proteome. The American Journal of Human Genetics **108**, 400–410 (2021).
40. Battle, A. *et al.* Impact of regulatory variation from RNA to protein. Science (1979) **347**, 664–667 (2015).
41. Mill, J. *et al.* Epigenomic Profiling Reveals DNA-Methylation Changes Associated with Major Psychosis. The American Journal of Human Genetics **82**, 696–711 (2008).
42. Bruneau, E. G., McCullumsmith, R. E., Haroutunian, V., Davis, K. L. & Meador-Woodruff, J. H. Increased expression of glutaminase and glutamine synthetase mRNA in the thalamus in schizophrenia.

Schizophrenia Research **75**, 27–34 (2005).

43. Serin, H. M., Simsek, E., Isik, E. & Gokben, S. WWOX-associated encephalopathies: identification of the phenotypic spectrum and the resulting genotype-phenotype correlation. Neurological Sciences **39**, 1977–1980 (2018).
44. Hussain, T. *et al.* Wwox deletion leads to reduced GABA-ergic inhibitory interneuron numbers and activation of microglia and astrocytes in mouse hippocampus. Neurobiology of Disease **121**, 163–176 (2019).
45. Greenwood, T. A. *et al.* Genome-Wide Linkage Analyses of 12 Endophenotypes for Schizophrenia From the Consortium on the Genetics of Schizophrenia. American Journal of Psychiatry **170**, 521–532 (2013).
46. Kunkle, B. W. *et al.* Novel Alzheimer disease risk loci and pathways in African American individuals using the African genome resources panel. JAMA Neurology **78**, 102 (2021).
47. Bacchelli, E. *et al.* An integrated analysis of rare CNV and exome variation in Autism Spectrum Disorder using the Infinium PsychArray. Scientific Reports **10**, 3198 (2020).
48. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. PLoS Genetics **6**, e1000888 (2010).
49. Mohanty, G., Jena, S. R., Nayak, J., Kar, S. & Samanta, L. Proteomic Signatures in Spermatozoa Reveal the Role of Paternal Factors in Recurrent Pregnancy Loss. The World Journal of Men's Health **38**, 103 (2020).
50. Gravina, P. *et al.* Genetic polymorphisms of glutathione S-transferases GSTM1, GSTT1, GSTP1 and GSTA1 as risk factors for schizophrenia. Psychiatry Research **187**, 454–456 (2011).
51. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genetics **10**, e1004383 (2014).
52. Morita, Y. *et al.* A Genetic Variant of the Serine Racemase Gene Is Associated with Schizophrenia. Biological Psychiatry **61**, 1200–1203 (2007).
53. Chick, J. M. *et al.* Defining the consequences of genetic variation on a proteome-wide scale. Nature **534**, 500–505 (2016).
54. Consortium, T. S. W. G. of the P. G., Ripke, S., Walters, J. T. R. & O'Donovan, M. C. Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *medRxiv* 2020.09.12.20192922 (2020) doi:10.1101/2020.09.12.20192922.
55. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics **48**, 481–487 (2016).
56. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics **48**, 214–220 (2016).
57. Ruzicka, W. B. *et al.* Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. *medRxiv* 2020.11.06.20225342 (2020) doi:10.1101/2020.11.06.20225342.

58. Ehrenreich, I. M. *et al.* Dissection of genetically complex traits with extremely large pools of yeast segregants. Nature **464**, 1039–1042 (2010).
59. Melé, M. *et al.* The human transcriptome across tissues and individuals. Science (1979) **348**, 660–665 (2015).
60. Kim-Hellmuth, S. *et al.* Cell type specific genetic regulation of gene expression across human tissues. Matthew Stephens **11**, 13.
61. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology **37**, 773–782 (2019).
62. Perkel, J. M. Single-cell proteomics takes centre stage. Nature **597**, 580–582 (2021).
63. Piehowski, P. D. *et al.* Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-µm spatial resolution. Nature Communications **11**, (2020).
64. Bai, B. *et al.* Deep profiling of proteome and phosphoproteome by isobaric labeling, extensive liquid chromatography, and mass spectrometry. in 377–395 (2017). doi:10.1016/bs.mie.2016.10.007.
65. Pagala, V. R. *et al.* Quantitative protein analysis by mass spectrometry. in 281–305 (2015). doi:10.1007/978-1-4939-2425-7_17.
66. Wang, H. *et al.* Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. Journal of Proteome Research **14**, 829–838 (2015).
67. Xu, P., Duong, D. M. & Peng, J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. Journal of Proteome Research **8**, 3944–3950 (2009).
68. Wang, X. *et al.* JUMP: A tag-based database search tool for peptide identification with high sensitivity and accuracy. Molecular and Cellular Proteomics **13**, 3663–3673 (2014).
69. Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. Journal of Proteome Research **2**, 43–50 (2003).
70. Niu, M. *et al.* Extensive Peptide Fractionation and y1 Ion-Based Interference Detection Method for Enabling Accurate Quantification by Isobaric Labeling and Mass Spectrometry. Analytical Chemistry **89**, 2956–2963 (2017).
71. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics **48**, 1443–1448 (2016).
72. Das, S. *et al.* Next-generation genotype imputation service and methods. Nature Genetics **48**, 1284–1287 (2016).
73. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics **48**, 1279–1283 (2016).
74. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics **29**, 15–21 (2013).
75. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078–2079 (2009).

76. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **12**, 323 (2011).
77. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. Nature Communications **8**, (2017).
78. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research **38**, e164–e164 (2010).
79. Aerts, S. *et al.* Gene prioritization through genomic data fusion. Nature Biotechnology **24**, 537–544 (2006).
80. Bai, B. *et al.* Deep Multilayer Brain Proteomics Identifies Molecular Networks in Alzheimer's Disease Progression. Neuron **105**, 975–991.e7 (2020).
81. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database* 2010, baq020–baq020 (2010).

Figures

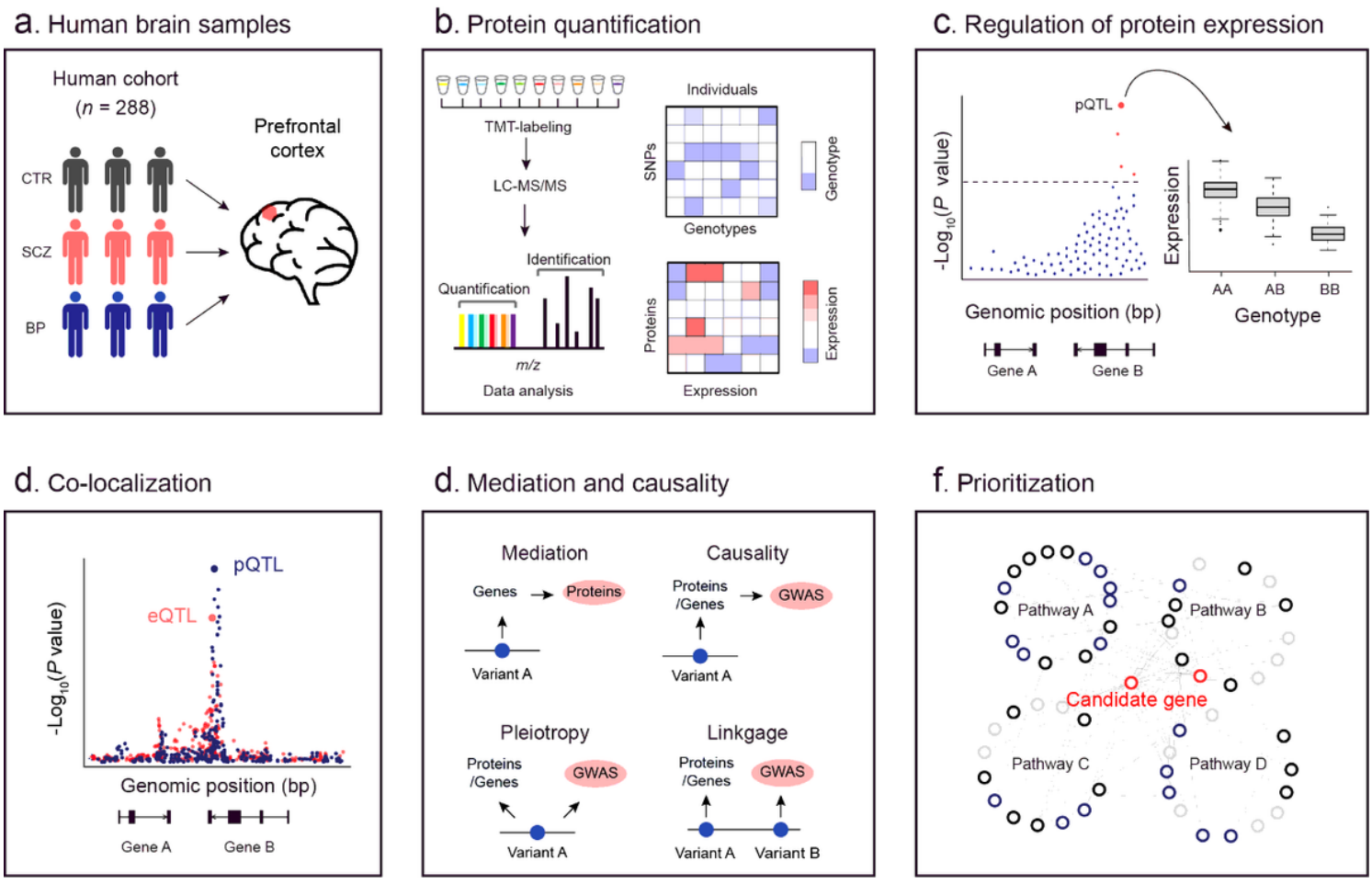

Figure 1

Schematic diagram showing the experimental design and analysis pipeline used in this study. **a**, Postmortem brain samples from a human cohort with 288 participants were used, including 210 normal individuals (CTR), 49 patients with schizophrenia (SCZ), and 29 patients with bipolar (BP). **b**, Deep brain proteome was profiled by 11-plex TMT-based proteomics, followed by extensive quality control and data analysis. Brain proteomic data and comparable genotype data were prepared for subsequent linkage analysis. **c**, Genome-wide association analysis to identify genetic regulations of protein expression and gene expression. **d**, Co-localization analysis to investigate the same variant underlying *cis*-eQTLs and *cis*-pQTLs. **e**, Mediation analysis to identify transcript-dependent and -independent regulations and causality analysis to link eGenes and pGenes to SCZ GWAS loci. **f**, Prioritization of proteins for SCZ GWAS loci.

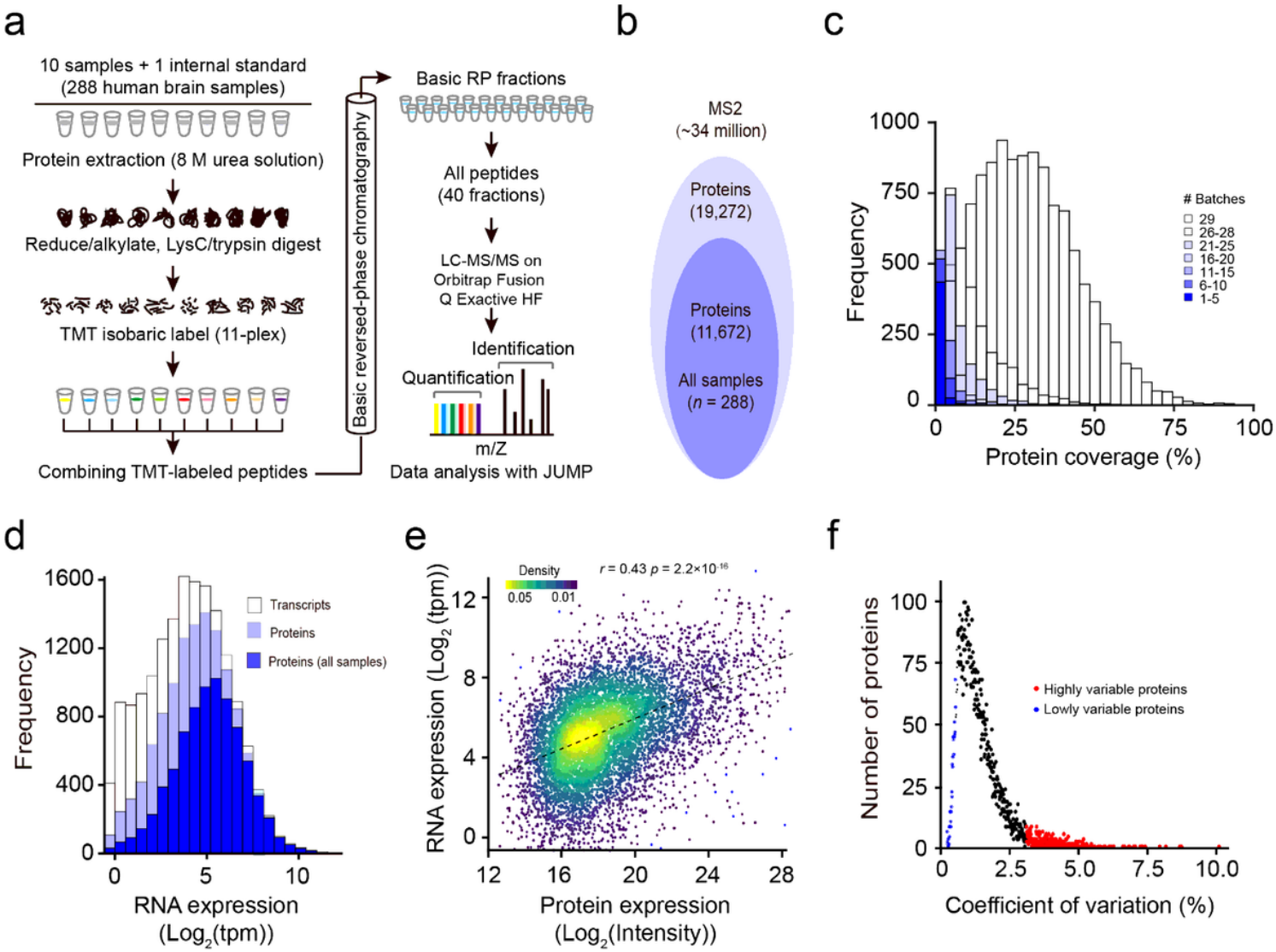

Figure 2

Deep profiling of human brain proteome. a, Workflow of 11-plex TMT-based proteome analysis. A total of 10 samples and 1 internal standard (i.e., 10 pooled samples) were analyzed by LC/LC-MS/MS. MS raw data were analyzed using JUMP software. **b**, Stacked Venn diagram showing the numbers of proteins identified in all 288 samples. **c**, Histogram showing the coverage of quantified proteins across 29 batches

of TMT experiments. **d**, Histogram showing the coverage of proteomic data compared to RNA-seq data. The open bar represents the distribution of protein-coding genes detected by RNA-seq, the light blue bar indicates the distribution of protein-coding genes from proteomic data, and the navy bar indicates the distribution of protein-coding genes from no missing value proteomic data. **e**, Scatter plot showing a comparison of gene expression levels and protein abudance. The expression is defined as the average expression across all samples. **f**, Distribution of coefficient of variation (CV) for all proteins across all samples.

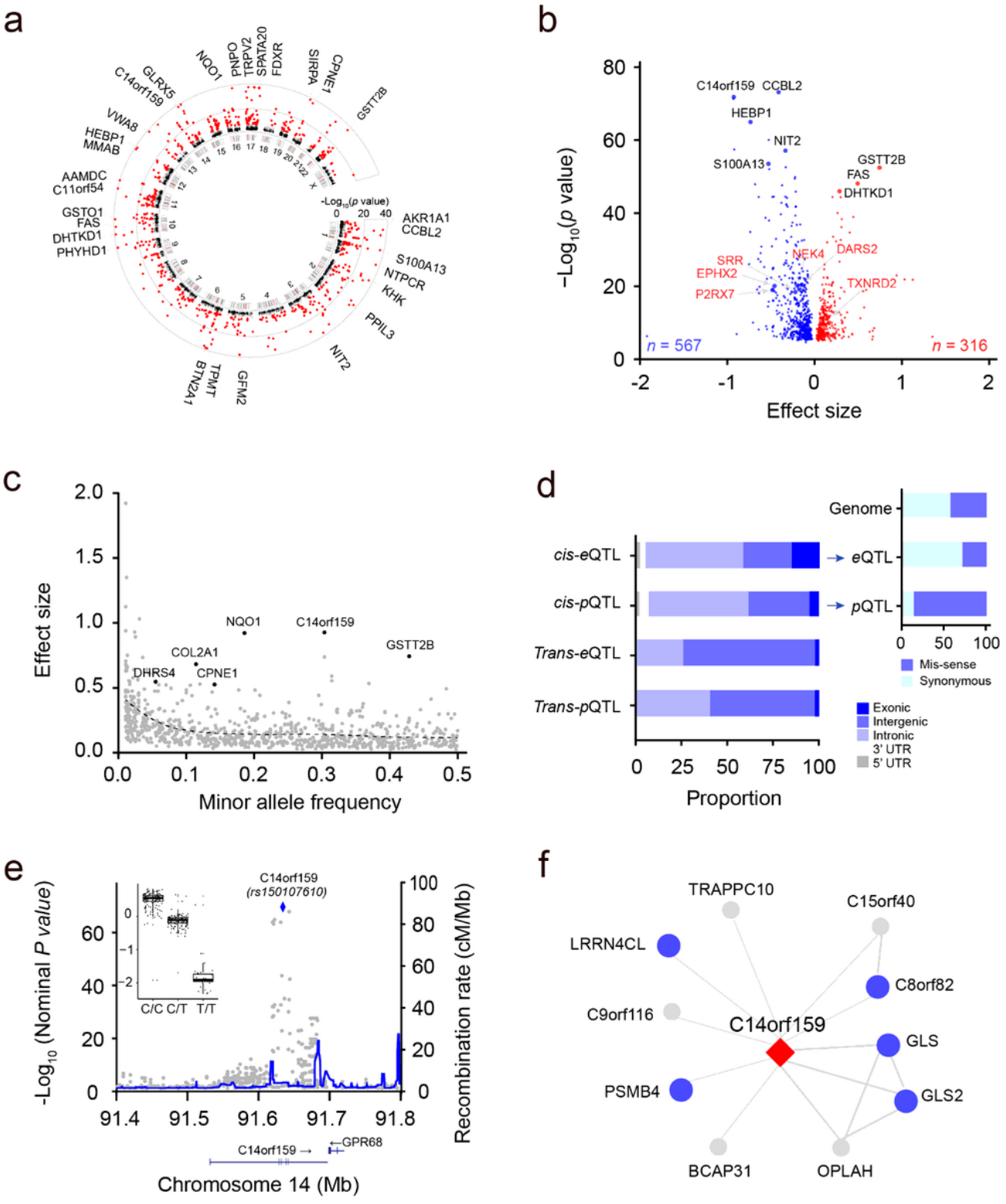

Figure 3

Genetic regulation of the brain proteome. a, Circos plot showing genome-wide *cis*-pQTLs. Significant cis-pQTLs (*q* value < 0.05) are highlighted in red color. **b**, Volcano plot showing effect size and $-\log_{10}$ nominal *p* values in *cis*-pQTLs. SCZ risk genes with a *cis*-pQTL that has large effect size ($\beta > 0.5$) and norminal *p* value < 10^{-15} are labeled in the plot. **c**, Scatter plot showing the relationship between minor allele frequency (MAF) and effect size of significant *cis*-pQTLs. SCZ risk genes with large effect size ($\beta > 0.5$) are also labeled in the plot. **d**, Stacked bar chart illustrating the proportions of each class of QTLs found in different genomic regions. **e**, LocusZoom plot showing that a *cis*-pQTL (i.e., *rs150107610*) is associated with C14orf159 protein expression. The inset box plot shows normalized C14orf159 protein expression and *rs150107610* allele dosage. **f**, Protein-protein interaction network highlights C14orf159 protein as a candidate SCZ risk gene.

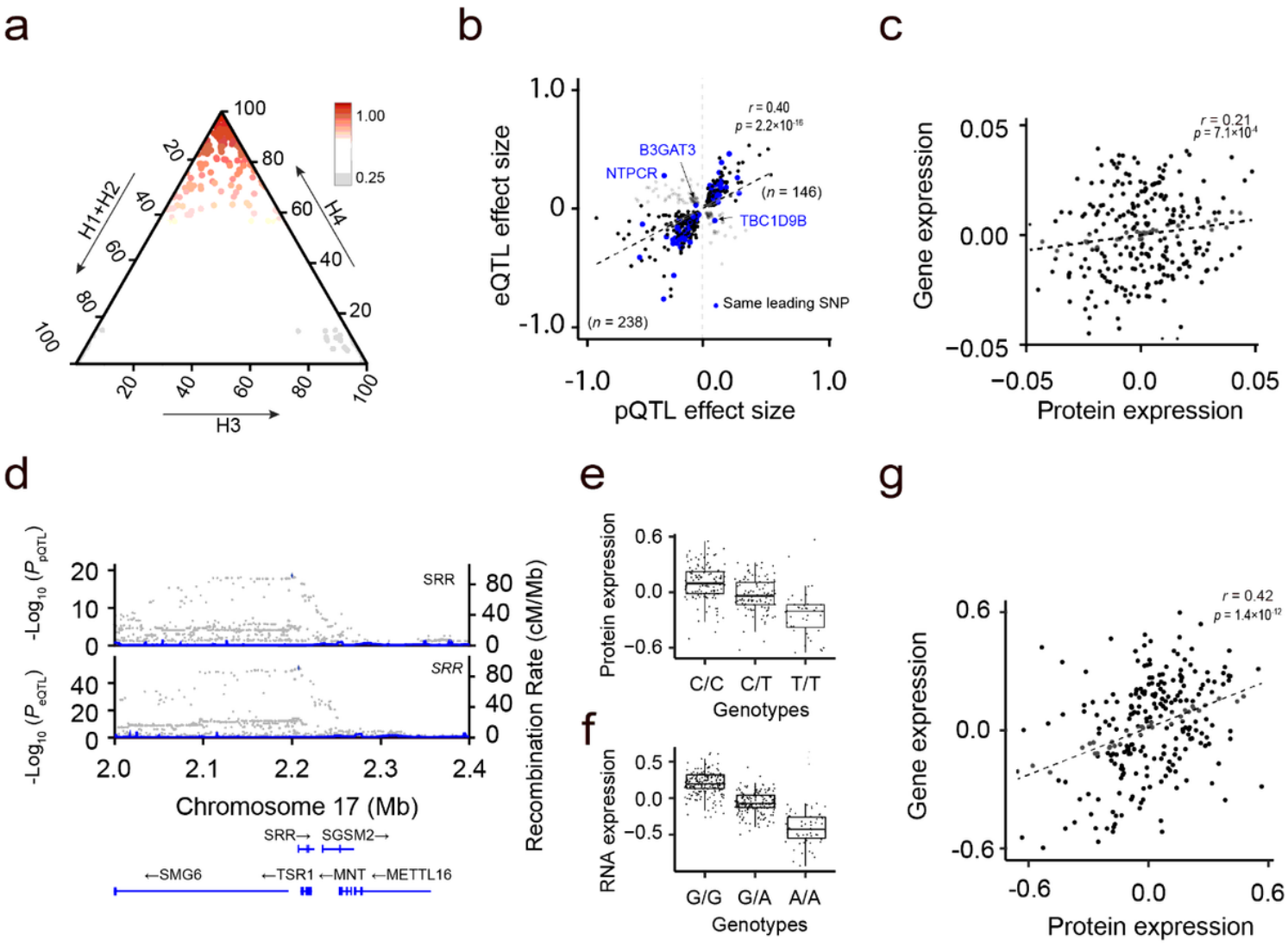

Figure 4

Co-localization of QTLs regulating the expression level of genes and proteins. a, Ternary plot showing colocalization posterior probabilities of QTLs of RNA and protein expression. We considered $H_0 + H_1 + H_2$

as evidence for the lack of test power. H_0: no causal variant, H_1: causal variant for PD GWAS only, H_2: causal variant for QTL only, H_3: two distinct causal variants, H_4: one common causal variant. **b**, Scatter plot showing the distribution of effect sizes of colocalized *cis*-eQTLs and *cis*-pQTLs. **c**, Correlation of expression levels of colocalized *cis*-eGenes and *cis*-pGenes. **d**, LocusZoom plot showing a colocalized QTL regulating SRR gene and protein expression. **e**, Box plot showing normalized SRR protein expression and its *cis*-pQTL allele dosage. **f**, Box plot showing normalized SRR protein expression and its *cis*-eQTL allele dosage. **g**, Box plot showing normalized SRR gene and protein expression levels and *rs150107610* allele dosage.

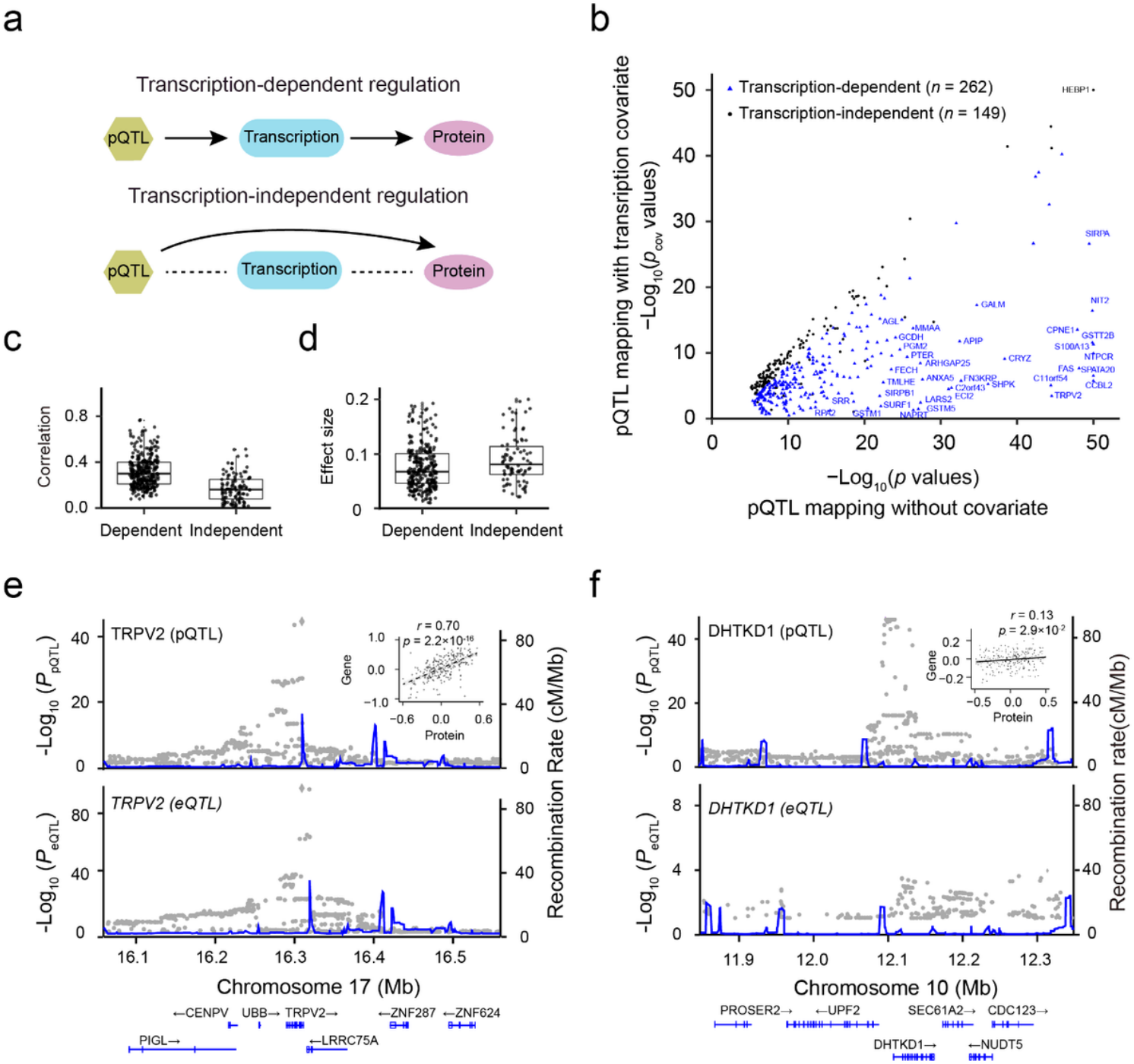

Figure 5

Genetic regulation of protein expression mediated by mRNA. a, Two mediation models of protein expression: transcription-dependent protein regulation and transcription-independent protein regulation. **b**, Scatter plot showing negative log-transformed *p* values of *cis*-pQTL before and after conditioning on mRNA. **c**, Box and whisker plot showing Pearson correlation coefficient between expression levels of proteins and transcripts in both transcript-mediated and transcript-independent groups. The plot shows the mean (horizontal lines), 5^{th}–95^{th} percentile values (boxes), and SEM (whiskers). **d**, Box and whisker plot showing effect sizes of transcription-dependent and transcription-independent regulations. **e**, An example of transcription-dependent regulation is exemplified by TRPV2. LocusZoom plots show a significant localization of *cis*-pQTL (top) and *cis*-eQTL (bottom). The inset shows the scatter plot of a high correlation (r = 0.70) between the expression of gene and protein. **f**, Transcription-independent regulation is exemplified by DHTKD1. LocusZoom plots show a significant *cis*-pQTL but not a *cis*-eQTL. The inset shows the scatter plot of a low correlation (r = 0.13) of DHTKD1 expression levels between gene and protein.

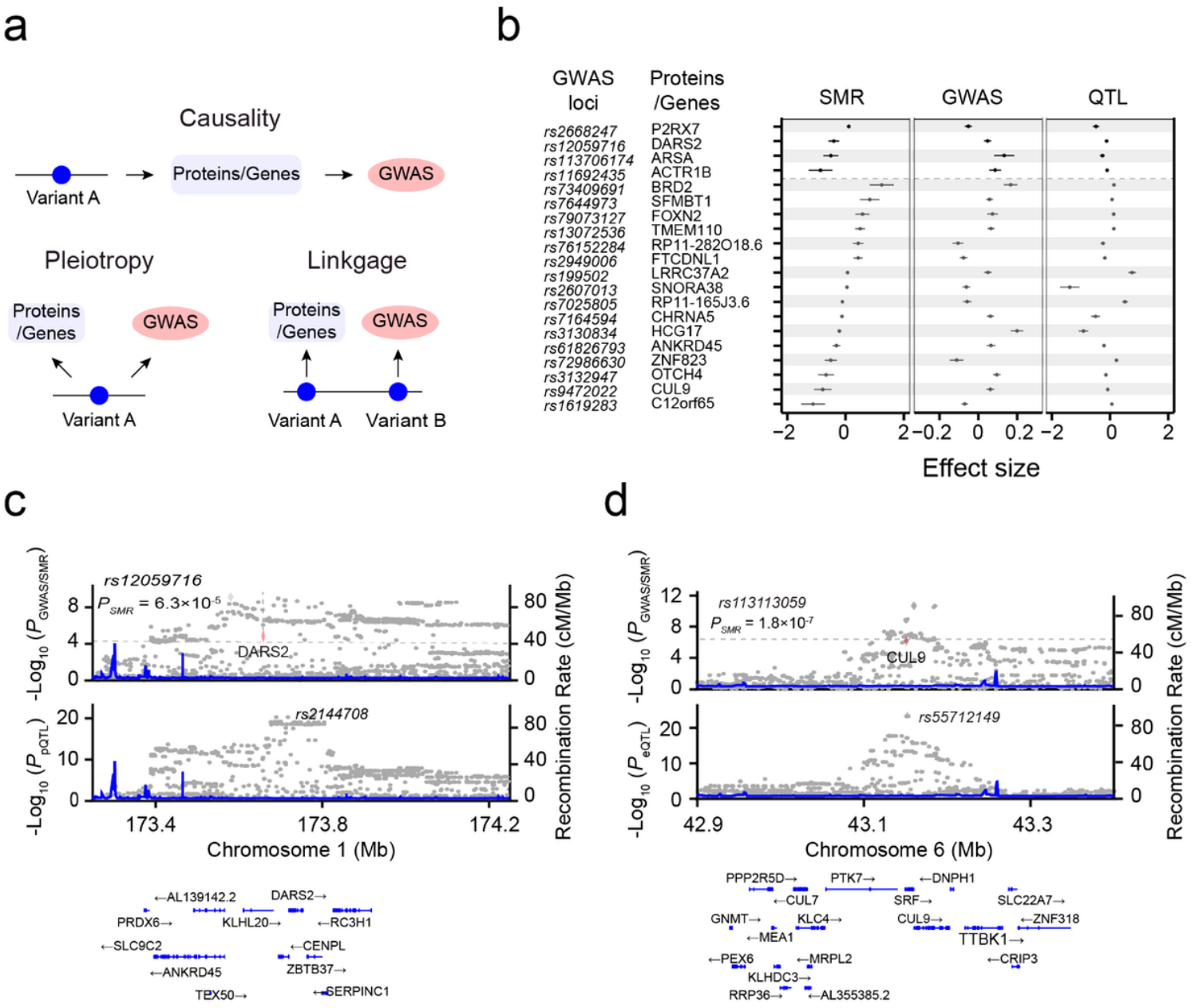

Figure 6

Causal relationship between pGenes and SCZ. a, Schematic diagram showing three putative mechanistic controls of a QTL: causality, pleiotropy, and genetic linkage. **b**, Forest plots showing effect sizes of 4 and 19 SCZ GWAS loci causally controlled by pGenes and eGenes, respectively. The causality relationship was estimated by the SMR/HEIDI method. Center values mark effect size point estimates, error bars the 95% confidence intervals. **c**, LocusZoom plot showing an example of an SCZ GWAS is controlled by a *cis*-pQTL. **d**, LocusZoom plot showing an example of an SCZ GWAS is controlled by a *cis*-eQTL.

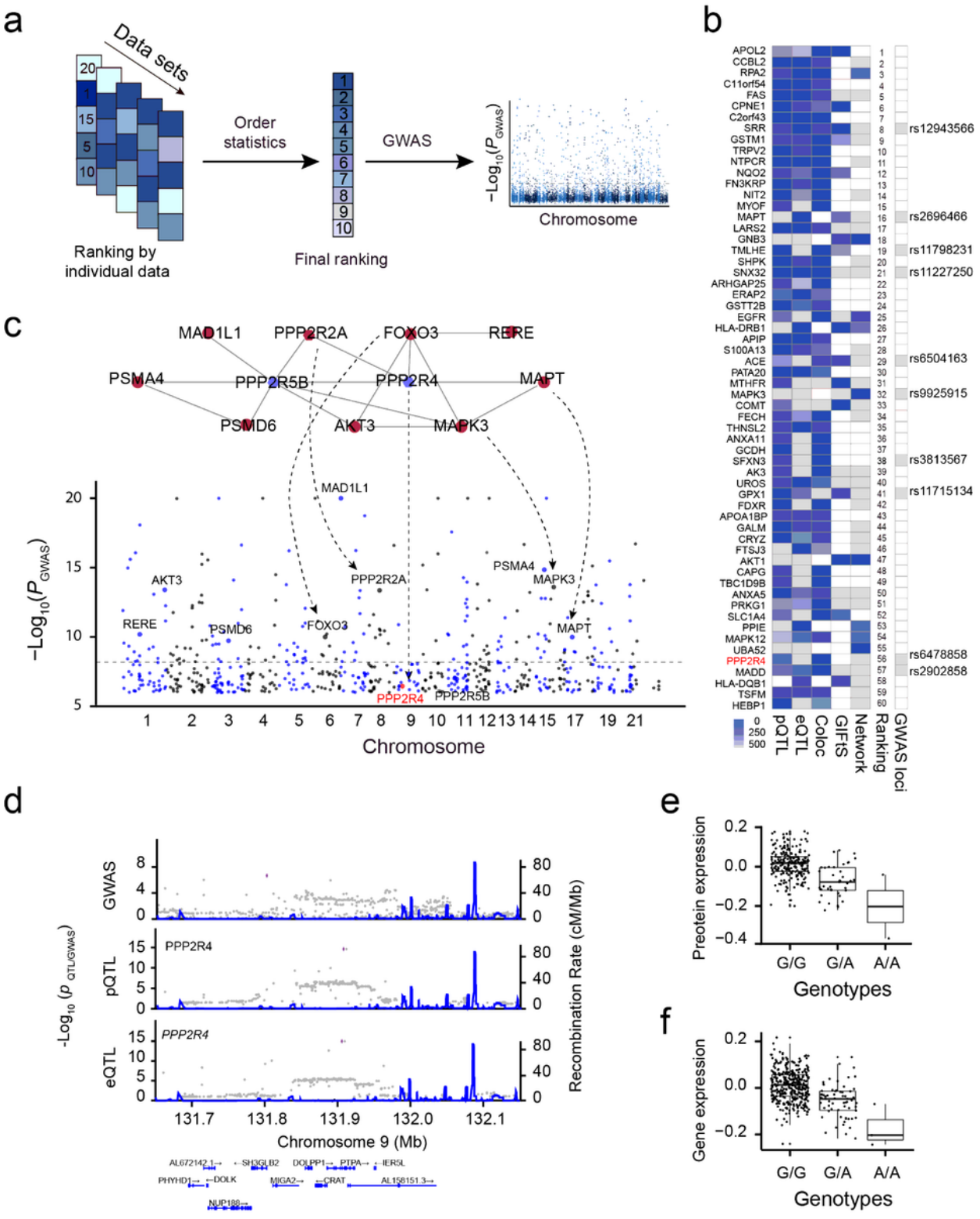

Figure 7

Prioritization of candidate genes for SCZ GWAS loci by integrating multiple data sets. a, Schematic diagram of candidate gene prioritization using order statistics. **b**, Heatmap showing the top 60 proteins ranked by combining five data sets. The missing values are indicated by white boxes. **c**, Network-based reprioritizing candidate genes for SCZ GWAS associations with small effect. Sub-network (top) were derived from the STRING PPI network. Significant GWAS risk genes are indicated by red nodes, whereas

candidate genes (PPP2R4 and PPP2R5B) for suggestive GWAS loci are indicated by blue nodes. **d**, LocusZoom plot showing a SCZ GWAS locus (*rs6478858*), a colocalized QTL regulating PPP2R4 gene and protein expression. **e**, Box plot showing normalized SRR protein expression and its *cis*-pQTL allele dosage. **f**, Box plot showing normalized SRR gene expression and its *cis*-eQTL allele dosage.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.docx
- SupplementaryTablev3.6.xlsx