Integrative analysis of proteome-wide association and transcriptome-wide association study identifies candidate brain proteins associated with insomnia

Peilin Meng
  Xi'an Jiaotong University   https://orcid.org/0000-0001-6297-0860

Bolun Cheng
  Xi'an Jiaotong University

Chun'e Li
  Xi'an Jiaotong University

Huijie Zhang
  Xi'an Jiaotong University

Shiqiang Cheng
  Xi'an Jiaotong University

Li Liu
  Xi'an Jiaotong University

Xuena Yang
  Xi'an Jiaotong University

Chujun Liang
  Xi'an Jiaotong University

Yao Yao
  Xi'an Jiaotong University

Jing Ye
  Xi'an Jiaotong University

Xiaomeng Chu
  Xi'an Jiaotong University

Chuyu Pan
  Xi'an Jiaotong University

Yujing Chen
  Xi'an Jiaotong University

Jingxi Zhang
  Xi'an Jiaotong University

Zhen Zhang
  Xi'an Jiaotong University

Yan Wen

Xi'an Jiaotong University

Yumeng Jia
Xi'an Jiaotong University

Xiong Guo
Xi'an Jiaotong University

Feng Zhang (✉ fzhxjtu@mail.xjtu.edu.cn)
Xi'an Jiaotong University

Research Article

Abstract

Great progress has been made in identifying risk loci for insomnia by genome-wide association studies (GWAS) analysis, but its association with human brain proteome is unclear. Two insomnia GWAS summary datasets were derived from deCODE (n = 113,006) and 23andMe (n = 1,331,010). Two human brain proteomic datasets were obtained from ROS/MAP and Banner, and two reference datasets were obtained from brain RNA-seq (CBR) and RNA-seq splicing (CBRS). Proteome-wide association study (PWAS) was first used to detect brain proteins associated with insomnia at the translation level. Transcriptome-wide association study (TWAS) was then used to verify the results at the DNA level. Finally, brain imaging GWAS was used to explore the brain functional areas related to the identified brain proteins and genes in insomnia. PWAS identified 4 and 1 common proteins shared by two human brain proteomic datasets in insomnia GWAS dataset 1 and 2, such as ME1 ($P_{Dataset1-Banner-full}$=1.08×10^{-2}, $P_{Dataset1-ROS/MAP-full}$=8.21×10^{-3}). Further TWAS identified 5 and 1 candidate genes shared by the two reference expression profiles in dataset 1 and 2, like ICA1L ($P_{Dataset2-CBR}$=3.01×10^{-2}, $P_{Dataset2-CBRS}$=3.24×10^{-2}). CAMLG was observed to associate with insomnia in both PWAS ($P_{Dataset2-ROS/MAP}$=2.94×10^{-2}) and TWAS ($P_{Dataset2-CBR}$=1.11×10^{-2}). Comparing the results of PWAS and TWAS, there are 9 common proteins and genes shared by both two datasets, such as INPP4A. Brain imaging analysis found that insomnia associated proteins and genes were functionally related to cortex. Our results may reinforce the understanding of the etiology and pathophysiology of insomnia and provide promising brain protein targets for further therapeutic and mechanistic studies.

Introduction

As the second-most prevalent mental disorder, insomnia is defined as difficulty initiating or maintaining sleep, accompanied by impaired daytime functioning, and occurs at least three times a week for at least three months (1). In the general population, the prevalence of insomnia is estimated at 8%-40%, and 4% of patients regularly using prescription sleep aids (2). Although insomnia often occurs as a chronic and primary disease, it can often be comorbid with other physical or mental disorder (3), such as depression (4) and hypertension (5). Insomnia can lead to frequent sick leave, incapacity and poor work performance (6), as well as a twofold increased risk of suicidal ideation and behavior (7). The serious long-term health consequences of insomnia place a huge burden on society and health care systems, directly or indirectly.

Family and twin studies showed a familial aggregation of insomnia, with an estimated heritability between 31% and 58% (8). While these findings suggest that insomnia is partly attributed to genetic factors, and a multiple genes are reported to involved in the pathogenesis of insomnia related to arousal regulation, sleep-wake processes, and brain functioning (9), these genotypes have rarely been replicated. Epigenetic mechanisms have been reported to be associated with the development and maintenance of insomnia (10), and exposure or re-exposure to some stressors may induce long-term changes in brain structures, such as hippocampus, which is a vulnerability factor in the development of insomnia (9). Brain proteome plays a critical role in sleep process and sleep problems. Current studies mainly

investigate and characterize the functional relevance between brain proteome and sleep deprivation (11). However, there is limited research in determining the association between insomnia and brain proteome.

Gene-determined protein sequence or abundance variability can provide a direct link between biochemical mechanisms and phenotypes, inferring that protein-level variation is more closely associated with phenotypic variation than with genome-level variation (12). Proteome Wide Association Study (PWAS) is a novel method to detect gene-phenotypic associations mediated by protein function alterations. PWAS applies machine learning and probabilistic models to aggregate the signal of all variants jointly that affect protein-coding genes and evaluate their overall impact on the protein function. These genes were then examined for functional variants related to individual target phenotypes (13). PWAS could capture complicated modes of heritability, showing outstanding capability in recovering causal protein-coding genes and seeking new correlations. PWAS has been widely used in the field of mental disorders. For example, Wingo et al. identified 19 and 11 genes that were consistent with the etiology of depression and Alzheimer's disease (AD) by performing PWAS (14). In addition, as an effective approach to identify significant expression-trait associations, transcriptome-wide association study (TWAS) could detect risk genes whose genetically regulated expression may be related to diseases by integrating GWAS associations with expression quantitative loci (eQTL) results (15).

In this study, PWAS were performed using the two large-scale GWAS datasets of insomnia and the human brain proteome aggregation dataset. PWAS results from two insomnia datasets were compared to identify insomnia associated genes in translation level. TWAS was then performed using the same two GWAS datasets for insomnia. The significant proteins and genes in PWAS and TWAS results were finally used for brain imaging analysis, respectively.

Results

PWAS results of Dataset 1 and 2

In Dataset 1, there are 16, 12, and 7 significant proteins associated with insomnia for full, female and male samples, respectively. For example, ME1 ($P_{Dataset1-Banner-full}$=1.08×10^{-2}, $P_{Dataset1-ROS/MAP-full}$=8.21×10^{-3}), and HINT1 ($P_{Dataset1-Banner-female}$=1.39×10^{-60}, $P_{Dataset1-ROS/MAP-female}$=9.72×10^{-3}). In Dataset 2, 18 candidate proteins were detected in Banner stage and ROS/MAP stage, such as PPP1R14C ($P_{Dataset2-Banner}$=3.30×10^{-3}, $P_{Dataset2-ROS/MAP}$=1.20×10^{-3}). Detailed information for all significant proteins identified in Dataset 1 and 2 are available in Table 1 and Figure 1.

Table1. Significant proteins identified by PWAS analysis in Insomnia dataset 1 and 2

	ID	Chromosome	*P-Value*
Dataset 1			
Full sample			
Banner dataset	ENSG00000112659.CUL9	6	3.57E-05
	ENSG00000198729.PPP1R14C	6	1.06E-04
	ENSG00000040933.INPP4A	2	1.12E-04
	ENSG00000115073.ACTR1B	2	8.66E-04
	ENSG00000091640.SPAG7	17	7.16E-03
	ENSG00000065833.ME1	6	1.08E-02
	ENSG00000080819.CPOX	3	1.34E-02
	ENSG00000157379.DHRS1	14	1.41E-02
ROS/MAP dataset	ENSG00000162104.ADCY9	16	3.43E-04
	ENSG00000116005.PCYOX1	2	5.89E-04
	ENSG00000102931.ARL2BP	16	4.71E-03
	ENSG00000130748.TMEM160	19	4.93E-03
	ENSG00000065833.ME1	6	8.21E-03
	ENSG00000169567.HINT1	5	1.07E-02
	ENSG00000091640.SPAG7	17	1.14E-02
	ENSG00000084073.ZMPSTE24	1	1.84E-02
Female sample			
Banner dataset	**ENSG00000169567.HINT1**	5	1.39E-60
	ENSG00000112659.CUL9	6	6.95E-03
	ENSG00000091640.SPAG7	17	3.11E-02
ROS/MAP dataset	ENSG00000072042.RDH11	14	5.78E-04
	ENSG00000102931.ARL2BP	16	2.24E-03
	ENSG00000162104.ADCY9	16	3.12E-03
	ENSG00000198931.APRT	16	4.76E-03
	ENSG00000169567.HINT1	5	9.72E-03
	ENSG00000091640.SPAG7	17	2.45E-02

	ENSG00000160803.UBQLN4	1	4.47E-02
	ENSG00000172375.C2CD2L	11	4.68E-02
	ENSG00000130748.TMEM160	19	4.70E-02
Male sample			
Banner dataset	ENSG00000104763.ASAH1	8	1.72E-04
	ENSG00000040933.INPP4A	2	2.43E-04
	ENSG00000198729.PPP1R14C	6	6.94E-04
	ENSG00000102781.KATNAL1	13	1.05E-03
	ENSG00000162390.ACOT11	1	1.44E-02
ROS/MAP dataset	ENSG00000179163.FUCA1	1	4.25E-08
	ENSG00000103061.SLC7A6OS	16	2.43E-02
Dataset 2			
Banner dataset	ENSG00000242028.HYPK	15	1.23E-11
	ENSG00000040933.INPP4A	2	1.50E-04
	ENSG00000080819.CPOX	3	2.40E-03
	ENSG00000198729.PPP1R14C	6	3.30E-03
	ENSG00000100580.TMED8	14	3.76E-03
	ENSG00000100445.SDR39U1	14	1.40E-02
	ENSG00000105255.FSD1	19	1.47E-02
ROS/MAP dataset	ENSG00000179295.PTPN11	12	8.10E-42
	ENSG00000103356.EARS2	16	5.40E-05
	ENSG00000198729.PPP1R14C	6	1.20E-03
	ENSG00000162104.ADCY9	16	1.82E-03
	ENSG00000198931.APRT	16	5.29E-03
	ENSG00000175183.CSRP2	12	5.32E-03
	ENSG00000116005.PCYOX1	2	9.12E-03
	ENSG00000164615.CAMLG*	5	2.94E-02
	ENSG00000130748.TMEM160	19	4.22E-02

ENSG00000137133.HINT2	9	4.48E-02
ENSG00000156052.GNAQ	9	5.04E-02

TWAS results for Dataset 1 and 2

In Dataset 1, 31, 23, and 45 candidate genes were identified to associate with insomnia in full, female and male samples, such as ZNF302 ($P_{\text{Dataset1-CBR-full}}$=1.33×10^{-3}, $P_{\text{Dataset1-CBRS-full}}$=5.06×10^{-3}), FRG1($P_{\text{Dataset1-CBR-female}}$=1.27×10^{-2}, $P_{\text{Dataset1-CBRS-female}}$=1.25×10^{-2}), and CCDC57($P_{\text{Dataset1-CBR-male}}$=3.23×10^{-3}, $P_{\text{Dataset1-CBRS-male}}$=5.57×10^{-3}). TWAS analysis detected 17 significant genes in the CBR reference dataset and 31 genes in the CBRS for dataset 2,such as ICA1L ($P_{\text{Dataset2-CBR}}$=3.01×10^{-2}, $P_{\text{Dataset2-CBRS}}$=3.24×10^{-2}). Interestingly, CAMLG was found to be significant in both PWAS-Banner ($P_{\text{Dataset2-ROS/MAP}}$=2.94×10^{-2}) and TWAS-CBR ($P_{\text{Dataset2-CBR}}$=1.11×10^{-2}) analysis. Significant genes detected by PWAS and TWAS analysis in Dataset 1 and 2 are shown in Table 2 and Figure 2.

Table2. Significant mRNAs identified by TWAS analysis in Insomnia dataset 1 and 2

	ID	Chromosome	*P-Value*
Dataset 1			
Full sample			
Brain RNA-seq (CBR) dataset	CEP250	20	9.54E-30
	RABGAP1L	1	1.62E-20
	KHDRBS3	8	1.50E-04
	SGCE	7	9.54E-04
	C8orf47	8	9.62E-04
	PRCC	1	1.00E-03
	ZNF302	19	1.33E-03
	HSD17B13	4	1.64E-03
	SSR1	6	2.05E-03
	MB21D1	6	2.70E-03
	MAP2K1	15	4.06E-03
	TP53INP1	8	6.25E-03
	ECI2	6	6.58E-03
	FRG1	4	1.17E-02
	PACRGL	4	1.73E-02
Brain RNA-seq splicing (CBRS) dataset	GAS5	1	8.29E-21
	CAMK2G	10	7.29E-06
	UNC50	2	2.49E-04
	ATP5H	17	6.19E-04
	MRPL3	3	1.16E-03
	DECR1	8	1.17E-03
	TTC28-AS1	22	1.48E-03
	GRM3	7	2.28E-03
	HSD17B11	4	2.53E-03
	ZNF559	19	3.34E-03
	ZNF302	19	5.06E-03
	GRIN2A	16	5.94E-03

	CCDC58	3	6.48E-03
	ALKBH8	11	8.59E-03
	FRG1	4	1.29E-02
	USP16	21	1.82E-02

Female sample

Brain RNA-seq (CBR) dataset	CEP250	20	7.10E-18
	COL11A2	6	4.00E-04
	SSR1	6	4.88E-04
	MFN1	3	1.29E-03
	ANKH	5	1.82E-03
	GLDC	9	2.47E-03
	SLC19A3	2	7.45E-03
	FRG1	4	1.27E-02
	ARHGAP1	11	1.31E-02
	ZNF502	3	2.13E-02
Brain RNA-seq splicing (CBRS) dataset	CCAR1	10	3.63E-04
	ATP13A4	3	6.73E-04
	NUP50	22	1.01E-03
	NUDT6	4	3.06E-03
	MTERFD1	8	3.27E-03
	ALKBH8	11	3.95E-03
	GRIN2A	16	4.32E-03
	URGCP	7	4.63E-03
	COG6	13	6.37E-03
	UGT8	4	8.45E-03
	FRG1	4	1.25E-02
	FAM63A	1	2.62E-02
	PTP4A2	1	3.16E-02

Male sample

Brain RNA-seq (CBR) dataset	SPRNP1	10	1.08E-07
	ZKSCAN4	6	2.56E-05
	VSIG10	12	1.12E-03
	PRCC	1	1.67E-03
	PMS2P4	7	2.40E-03
	MAP2K1	15	2.52E-03
	EPPK1	8	3.06E-03
	CCDC57	17	3.23E-03
	ZNF135	19	5.44E-03
	AACSP1	5	5.78E-03
	BCO2	11	9.28E-03
	PLA2G15	16	1.63E-02
Brain RNA-seq splicing (CBRS) dataset	SNTG1	8	7.52E-20
	CCNDBP1	15	1.84E-07
	CCNDBP1	15	8.82E-06
	IQCA1	2	2.15E-04
	ADCK3	1	3.71E-04
	GOLGA8B	15	4.53E-04
	GAS8	16	9.55E-04
	MRPL3	3	1.34E-03
	SDCCAG8	1	1.76E-03
	ME1	6	2.04E-03
	DCAF5	14	2.49E-03
	ZFAND1	8	2.78E-03
	MRPL3	3	2.99E-03
	ZFAND1	8	3.44E-03
	PPA2	4	4.65E-03
	UBE2Q2P1	15	5.23E-03
	DECR1	8	5.57E-03

	CCDC57	17	5.57E-03
	ATP5H	17	5.70E-03
	UBE2Q2P1	15	6.02E-03
	UBE2Q2P1	15	6.71E-03
	CCDC57	17	7.64E-03
	CCDC57	17	8.83E-03
	TTC28-AS1	22	9.81E-03
	RAB23	6	1.36E-02
	ZNF493	19	1.44E-02
	USP16	21	1.48E-02
	ZNF493	19	1.52E-02
	DNM1	9	1.57E-02
	FLOT1	6	3.09E-02
	FLOT1	6	3.22E-02
	FLOT1	6	3.26E-02
	FLOT1	6	3.30E-02

Dataset 2

Brain RNA-seq (CBR) dataset	ATG14	14	1.92E-18
	RILPL1	12	2.22E-08
	SCAP	3	4.81E-07
	SIL1	5	8.12E-07
	KLF12	13	3.15E-05
	HEXA	15	9.31E-05
	ECE2	3	5.06E-04
	DEF8	16	7.88E-04
	DMXL2	15	9.36E-04
	SLIT2	4	1.48E-03
	ADPRHL1	13	1.66E-03

	KIF27	9	1.70E-03
	SRPK1	6	3.58E-03
	TSPAN14	10	4.50E-03
	CGGBP1	3	9.61E-03
	CAMLG*	5	1.11E-02
	ICA1L	2	3.01E-02
Brain RNA-seq splicing (CBRS) dataset	GAS5	1	3.83E-18
	TCF12	15	2.53E-15
	ABI2	2	1.55E-13
	ZFP30	19	1.89E-08
	AKAP9	7	1.61E-05
	GON4L	1	3.30E-05
	LRP11	6	2.84E-04
	GUSBP11	22	5.00E-04
	PPA2	4	9.93E-04
	FCGR2A	1	1.31E-03
	GGA3	17	1.33E-03
	NUP50	22	1.84E-03
	C16orf13	16	1.94E-03
	GGA3	17	2.20E-03
	CSPP1	8	2.40E-03
	ST3GAL5	2	4.33E-03
	TMEM25	11	4.34E-03
	PPP6R2	22	4.52E-03
	NDUFV1	11	4.52E-03
	ALKBH8	11	7.16E-03
	GMCL1	2	9.02E-03
	DNAJC12	10	9.64E-03
	SPATS2L	2	9.83E-03

ZNRD1-AS1	6	1.40E-02
PPIP5K2	5	2.18E-02
PPIP5K2	5	2.26E-02
PPIP5K2	5	2.46E-02
ZNRD1-AS1	6	3.24E-02
ICA1L	2	3.24E-02
ICA1L	2	3.31E-02
ICA1L	2	3.63E-02

The common significant genes in Dataset 1 and 2

Comparing the PWAS results, there are 6 common significant proteins shared by the three samples of Dataset 1 and Dataset 2, INPP4A ($P_{\text{Dataset1-Banner-full}}=1.12\times10^{-4}$, $P_{\text{Dataset1-Banner-male}}=2.43\times10^{-4}$, $P_{\text{Dataset2-Banner}}=1.50\times10^{-4}$), CPOX ($P_{\text{Dataset1-Banner-full}}=1.34\times10^{-2}$, $P_{\text{Dataset2-Banner}}=2.40\times10^{-3}$), PPP1R14C ($P_{\text{Dataset 1-Banner-full}}=1.06\times10^{-4}$, $P_{\text{Dataset1-Banner-male}}=6.94\times10^{-4}$, $P_{\text{Dataset2-Banner}}=3.30\times10^{-3}$, $P_{\text{Dataset2-ROS/MAP}}=1.20\times10^{-3}$), ADCY9 ($P_{\text{Dataset1-ROS/MAP-full}}=3.43\times10^{-4}$, $P_{\text{Dataset 1-ROS/MAP-female}}=3.12\times10^{-3}$, $P_{\text{Dataset2-ROS/MAP}}=1.82\times10^{-3}$), PCYOX1 ($P_{\text{Dataset1-ROS/MAP-full}}=5.89\times10^{-4}$, $P_{\text{Dataset2-ROS/MAP}}=9.12\times10^{-3}$), TMEM160 ($P_{\text{Dataset1-ROS/MAP-full}}=4.93\times10^{-3}$, $P_{\text{Dataset1-ROS/MAP-female}}=4.70\times10^{-2}$, $P_{\text{Dataset2-ROS/MAP}}=4.22\times10^{-2}$),APRT ($P_{\text{Dataset1-ROS/MAP-female}}=4.76\times10^{-3}$, $P_{\text{Dataset2-ROS/MAP}}=5.29\times10^{-3}$).

For the TWAS results of the three samples of Dataset 1 and Dataset 2, we detected 4 candidate common genes, including GAS5($P_{\text{Dataset1-CBRS-full}}=8.29\times10^{-21}$, $P_{\text{Dataset2-CBRS}}=3.83\times10^{-18}$), ALKBH8($P_{\text{Dataset1-CBRS-full}}=8.59\times10^{-3}$, $P_{\text{Dataset1-CBRS-female}}=3.95\times10^{-3}$, $P_{\text{Dataset2-CBRS}}=7.16\times10^{-3}$), PPA2($P_{\text{Dataset1-CBRS-male}}=4.65\times10^{-3}$, $P_{\text{Dataset2-CBRS}}=9.93\times10^{-4}$), and NUP50($P_{\text{Dataset1-CBRS-female}}=1.01\times10^{-3}$, $P_{\text{Dataset2-CBRS}}=1.84\times10^{-3}$).

The results of brain imaging analysis of significant genes in PWAS and TWAS

Brain imaging analysis found that insomnia associated proteins and genes were functionally related to cortex. For example, CUL9 was found to be belonged to the categories of cortical thickness and cortical area, with the phenotype of aparc-a2009s lh thickness S-precentral-inf-part and aparc-Desikan rh area paracentral; ACTR1B also belongs to the categories of cortical thickness with the phenotype of aparc-a2009s lh thickness S-precentral-sup-part. The detailed information about brain imaging analysis of significant proteins and genes in PWAS and TWAS has been listed in Table 3.

Table 3. The results of brain imaging analysis of significant proteins and genes in PWAS and TWAS analysis

Proteins and Genes	Category	Phenotype	Loci < 10^{-6}	Top variant	Top p-value
CUL9	cortical thickness	aparc-a2009s lh thickness S-precentral-inf-part	12	6:43,156,160 T / A (rs10498754)	3.4e-10
	cortical area	aparc-Desikan rh area paracentral	31	6:43,156,160 T / A (rs10498754)	9.5e-10
PPP1R14C	rfMRI connectivity	rfMRI connectivity ICA100 edge 312	1	6:150,526,573 C / A (rs55749434)	3.8e-7
ACTR1B	cortical thickness	aparc-a2009s lh thickness S-precentral-sup-part	19	2:98,275,354 G / A (rs11692435)	1.9e-10
KATNAL1	rfMRI connectivity	rfMRI connectivity ICA100 edge 1089	4	13:30,818,071 T / C (rs9578160)	1.9e-7
FSD1	rfMRI connectivity	rfMRI connectivity ICA100 edge 287	2	9:97,171,109 G / T (rs2993957)	1.1e-7
KHDRBS3	rfMRI connectivity	rfMRI connectivity ICA100 edge 1285	4	8:136,314,418 C / CTTTTTTTTTTTTTTTT (rs550710091)	1.9e-7
MRPL3	rfMRI connectivity	rfMRI connectivity ICA25 edge 16	2	17:73,901,143 G / A (rs55805916)	2.4e-8
	tfMRI activation	IDP tfMRI 90th-percentile zstat shapes	2	21:26,811,769 C / T (rs145024673)	1.5e-7
	tfMRI activation	IDP tfMRI median zstat shapes	3	21:26,811,769 C / T (rs145024673)	2.1e-7
GRM3	rfMRI connectivity	rfMRI connectivity ICA100 edge 1435	3	7:86,394,223 T / C (rs17160953)	2.2e-9
	rfMRI connectivity	rfMRI connectivity ICA100 edge 848	2	7:86,291,446 T / A (rs802421)	3.8e-7
	rfMRI connectivity	rfMRI connectivity ICA25 edge 172	4	7:85,650,184 G / A (rs35134543)	4.9e-7
GRIN2A	rfMRI connectivity	rfMRI connectivity ICA100 edge 1308	4	16:10,024,957 C / A (rs35056140)	2.5e-7
ZKSCAN4	regional and tissue volume	ThalamNuclei lh volume AV	27	6:28,221,264 G / A (rs1679732)	2.5e-13
MRPL3	rfMRI connectivity	rfMRI connectivity ICA25 edge 16	2	17:73,901,143 G / A (rs55805916)	2.4e-8
	tfMRI activation	IDP tfMRI 90th-percentile zstat shapes	2	21:26,811,769 C / T (rs145024673)	1.5e-7

Proteins and Genes	Category	Phenotype	Loci < 10^{-6}	Top variant	Top p-value
	tfMRI activation	IDP tfMRI median zstat shapes	3	21:26,811,769 C / T (rs145024673)	2.1e-7
FLOT1	cortical thickness	aparc-DKTatlas lh thickness insula	15	6:30,699,022 G / A (rs3095327)	6.4e-13
	WM tract diffusivity	IDP dMRI TBSS L1 Fornix cres Stria terminalis R	20	6:30,704,982 C / CTGTA (rs746908817)	7.2e-11
	cortical thickness	aparc-Desikan lh thickness insula	14	6:30,699,022 G / A (rs3095327)	5.1e-10
	rfMRI connectivity	rfMRI connectivity ICA100 edge 28	2	5:14,910,099 A / AAG (rs76909499, rs397757597)	8.5e-7
UGT8	rfMRI connectivity	rfMRI connectivity ICA100 edge 56	2	4:115,352,162 T / A (rs11733369)	8.1e-7
SCAP	rfMRI connectivity	ANKH	2	15:76,871,034 T / C (rs72738853)	5.4e-7
SLIT2	cortical area	aparc-DKTatlas lh area parahippocampal	16	4:20,121,121 G / A (rs13150272)	1.2e-12
	cortical area	aparc-Desikan lh area parahippocampal	19	4:20,121,121 G / A (rs13150272)	1.4e-11
	cortical area	aparc-pial lh area parahippocampal	21	4:20,121,121 G / A (rs13150272)	1.0e-9
	cortical area	aparc-pial lh area isthmuscingulate	17	4:20,547,578 G / C (rs12501870)	1.0e-8

Discussion

In recent years, great progress has been made in the study of the etiology and pathophysiology of insomnia, and several candidate genetic variants of insomnia have been identified by GWAS analysis. In this study, we performed PWAS analysis on two GWAS datasets of insomnia and human brain proteome respectively, and found the characteristics of insomnia at the translation level. TWAS was then used to verify the proteomic results of PWAS at the transcriptional level. Our results provide important information for pathogenic and therapic studies of insomnia.

Several candidate proteins identified by PWAS analysis are strongly associated with brain development and mental disorder, such as ME1, HINT1 and EARS2. ME1, also known as Tcf12 or HEB, encodes the bHLH E-protein ME1 in the complex orchestration of neural tube formation and brain development (16). It is expressed in the cerebellar primordium, but not sustained in aged cerebellum, and maintains a pattern

of decreased expression in the aging hippocampus (17). Histidine triad nucleotide-binding protein 1 (HINT1), a member of the histidine triad (HIT) protein nucleotide transferase and hydrolase superfamily, is widely expressed in several tissues including the central nervous system, and is involved in the transcriptional regulation of multiple proteins. Studies have demonstrated that HINT1 is associated with many neuropsychiatric diseases, such as schizophrenia, anxiety, drug addiction, and pain (18). EARS2 is important to protein translation in mitochondria, which encodes mitochondrial glutamyl t-RNA synthetase and attach glutamate to its homologous mitochondrial t-RNA. Mutations in the EARS2 lead to a neurological disorder defined as leukoencephalopathy with thalamus and brainstem involvement and high lactate(19). These studies suggest that ME1, HINT1, and EARS2 may play an important role in the occurrence and development of the population.

Some significant genes related to neurodegeneration and neurodevelopmental disorder were detected by TWAS analysis. For example, RABGAP1L, a GTPase activated protein (GAP) related to regulation of intracellular membrane trafficking, plays a critical role in is important in brain development(20), and involved in the development of Huntington's disease which is characterized by selective neurodegeneration of the striatum(21). Hexosaminidase-A (HEXA) could produce the HexA enzyme, loss or inactivation in this enzyme are necessary for massive accumulation of GM2 ganglioside in neurons which is the underlying cause of neurodegenerative disorders including Tay−Sachs disease(22). Notably, calcium modulating cyclophilin ligand (CAMLG) has been identified to be involved in regulating locomotor activity and wakefulness in mice through interaction with Tmub1, and this modulation may be an interesting avenue for sleep research (23). The impact of CAMLG shRNA is to target the GABAA receptors, which has been widely accepted in sleep physiology.

Interestingly, there were several common proteins and genes involved in brain development and function in the two GWAS datasets, such as INPP4A, ADCY9 and GAS5. Inositol polyphosphate phosphatase 4A (INPP4A) is a suppressor of glutamate excite-toxicity in the central nervous system and the first signaling protein that suppress excite-toxic cell death in neurons. It is implicated in motor and cognitive behaviors and neuronal subset survival in mice (24). Adenylate cyclase (AC) type 9 gene (ADCY9) is a widely distributed adenylyl cyclase that plays a critical role in neuronal signaling, and has been reported to be involved in mood disorder (25). Additionally, ACs catalyze the conversion of intracellular ATP to cyclic AMP (cAMP), and changes in cAMP signaling have been associated with major depression, bipolar disorder, and suicide. The fundamental roles and observed dysregulation of long non-coding RNAs (lncRNAs) in the brain development, functional diversification, and neurodegenerative diseases has been recognized by independent studies (26). Growth arrest-specific5 (GAS5) is a lncRNA that hosts a number of small nucleolar RNAs, which can promote the recovery of learning and memory in rats with cholinergic nerve injury (27). It has been reported that GAS5 silencing can protect hippocampal neurons from hypoxic/ischemic-induced brain injury (28). Our results suggest that these common genes may affect brain function and cause insomnia through modifications of AC and lncRNA.

Much progress has been made in recent years to determine the roles of brain proteins in brain and associated diseases, such as autism spectrum disorder and schizophrenia (29). Several significant genes

in this study have been found to be associated with brain-related disorders. For instances, there is a close relationship between ALKBH8 and neurodevelopmental disorder including autosomal recessive non-syndromic intellectual disability, and non-syndromic x-linked intellectual disability (30). Besides, INPP4A has a strong expression in the central nervous system and plays a vital role in the pathogenesis of intellectual disability, mental retardation syndrome and mental disorder (31).

State-of-the-art brain imaging techniques have accelerated progress in the study and understanding of insomnia in humans(32). In our brain imaging analysis, an interesting observation was that some significant common proteins and genes identified in PWAS and TWAS were associated with cortical function. In exploring the neurobiological underpinnings for sleep, increasing researchers are paying more attention to the role of cortex. Evidence from Fernandez-Mendoza and colleagues supports the idea that functional cortical alteration may manifest in insomnia symptoms in adolescents (33). Disrupted cortical networks has been proved to be related to high sleep reactivity, which is often manifested as a severe insomnia phenotypes (34). Neuroanatomical alterations in primary insomnia mostly involve the hippocampus, anterior cingulate cortex and orbitofrontal cortex (35). More research is needed to confirm the link between insomnia and cortical function.

However, the limitations of this study should not be ignored. First, since the two GWAS data applied in this study were driven from previous studies, the accuracy of our results may be affected by previous studies. Second, all subjects recruited in our study were European ancestry, and caution should be exercised when applying our results to other ethnic group.

In conclusion, we systematically explored candidate proteins and genes that may be involved in the development of insomnia by integrating GWAS data and human brain proteome and conducting PWAS and TWAS analysis. We then compared the results and observed several common protein and genes shared in difference datasets. Our results may provide a novel insight for pathology of insomnia and reveal the roles of brain proteins in development of insomnia.

Methods

The GWAS dataset of insomnia in deCODE (Dataset 1)

A recent large-scale summary of GWAS data was recruited here, including 12,444,916 SNPs in the total population, 12,428,592 SNPs in males and 12,432,937 SNPs in females (36). Briefly, this dataset of insomnia was obtained from 113,006 individuals in UK Biobank and 7,565 participants in deCODE. For UK Biobank sample, genotyping was conducted on the UK BiLEVE custom array and the UK Biobank Axiom array. Imputation was performed with the UK10K haplotype panel and the 1000 Genomes Project Phase 3 reference panel. Association tests were performed in SNPTEST using logistic regression model. SNPs with imputation quality of total sample < 0.8 and minor allele frequency (MAF) < 0.001 were excluded. For deCODE sample, the Icelandic GWAS dataset was generated from whole genome sequencing, chip genotyping and long range phasing of Icelandic population samples. Logistic

regression was applied to association tests using software developed by deCODE genetics. The detailed information of GWAS, meta-analysis, imputation, standard quality-control procedures, and analytic procedure were described in the previous study (36).

The GWAS dataset of insomnia from 23andMe (Dataset 2)

Independent significant SNPs in Dataset 2 were derived from UKB and 23andMe datasets, including a large genetic association samples of 1,331,010 participants (37). Genotype was conducted in Illumina HumanHap550+ BeadChip platform. HapMap CEU subset was applied to perform imputation. SNP associations were tested with linear or logistic regression models based on the sleep phenotype. Detailed description of meta-analysis, genotype, imputation, and quality control are available in the published study (37).

Human brain proteomes of ROS/MAP and Banner datasets

The human brain proteomes of ROS/MAP and Banner datasets were derived from a published study (14). 376 donators of European ancestry from the ROS/MAP and 152 participants of European ancestry recruited by the Banner Sun Health Research Institute were included in human brain proteomes studies (38). Isobaric tandem mass tag (TMT) peptide labeling and liquid chromatography coupled to mass spectrometry (MS) were applied to perform proteomic sequencing and peptides analysis. Proteome Discoverer suite and tandem mass spectrometry spectra were then used to match the peptide profiles, so that the combination probability FDR of the constituent peptides was less than 1%. Proteins without reliable measurement were removed from further analysis. Genotyping was obtained by using Illumina OmniQuad Express or Affymetrix GeneChip 6.0 platform for whole genome sequencing or complete genotyping. Only the sites included in the link disequilibrium reference panel were used for confirmation of PWAS.

Proteome-wide association study of Insomnia

FUSION was applied to estimate protein weights in the discovery and confirmation dataset. Following the standard approach from previous study (14), we applied FUSION here for combine the genetic effect of the two datasets of insomnia (insomnia GWAS z-score) with the protein weights by calculating the linear sum of z-score × weight for the independent SNPs at the locus to perform the PWAS of the two datasets of insomnia. 5,000 permutations were performed to obtain the empirical distributions of PWAS statistics of each gene, from which the P value was calculated. The permutated empirical P value < 0.05 was considered as statistically significant in PWAS analysis.

Transcriptome-wide association study of Insomnia

FUSION software was applied for TWAS analysis to validate the proteomic findings for both the two datasets of insomnia (15). Brain RNA-seq (CBR) reference and brain RNA-seq splicing (CBRS) reference are two expression reference datasets used for TWAS. In brief, all five models of FUSION were used to estimate RNA weights and to select the most predictive model. Next, TWAS was performed for the two

datasets of insomnia by combining the genetic effect (insomnia GWAS z-score) and the mRNA expression weights. Genes with P-value<0.05 was regarded as statistically significant.

Brain imaging analysis of significant genes for PWAS and TWAS analysis

PheWeb is an easy-to-use open-source web-based tool for visualization, navigation, and sharing GWAS and PheWAS results (https://github.com/statgen/pheweb.). PheWeb has been used to explore associations for large datasets, such as the UK Biobank, the Oxford Brain Imaging Genetics Project, and the Michigan Genomics Initiative (39). In this study, we applied PheWeb to obtain brain imaging-derived phenotypes of genes for PWAS and TWAS analysis based on expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank(40)

Declarations

References

1. Battle DE. Diagnostic and Statistical Manual of Mental Disorders (DSM). CoDAS. 2013;25(2):191–2.

2. Chong Y, Fryer CD, Gu Q. Prescription sleep aid use among adults: United States, 2005-2010. NCHS data brief. 2013(127):1–8.

3. Fernandez-Mendoza J, Vgontzas AN. Insomnia and its impact on physical and mental health. Current psychiatry reports. 2013;15(12):418.

4. Baglioni C, Battagliese G, Feige B, Spiegelhalder K, Nissen C, Voderholzer U, et al. Insomnia as a predictor of depression: a meta-analytic evaluation of longitudinal epidemiological studies. Journal of affective disorders. 2011;135(1-3):10–9.

5. Jarrin DC, Alvaro PK, Bouchard MA, Jarrin SD, Drake CL, Morin CM. Insomnia and hypertension: A systematic review. Sleep medicine reviews. 2018;41:3–38.

6. Kucharczyk ER, Morgan K, Hall AP. The occupational impact of sleep quality and insomnia symptoms. Sleep medicine reviews. 2012;16(6):547–59.

7. Bjørngaard JH, Bjerkeset O, Romundstad P, Gunnell D. Sleeping problems and suicide in 75,000 Norwegian adults: a 20 year follow-up of the HUNT I study. Sleep. 2011;34(9):1155–9.

8. Gehrman PR, Meltzer LJ, Moore M, Pack AI, Perlis ML, Eaves LJ, et al. Heritability of insomnia symptoms in youth and their relationship to depression and anxiety. Sleep. 2011;34(12):1641–6.

9. Riemann D, Nissen C, Palagini L, Otte A, Perlis ML, Spiegelhalder K. The neurobiology, investigation, and treatment of chronic insomnia. The Lancet Neurology. 2015;14(5):547–58.

10. Palagini L, Biber K, Riemann D. The genetics of insomnia–evidence for epigenetic mechanisms? Sleep medicine reviews. 2014;18(3):225–35.

11. Noya SB, Colameo D. The forebrain synaptic transcriptome is organized by clocks but its proteome is driven by sleep. 2019;366(6462).

12. Okada H, Ebhardt HA, Vonesch SC, Aebersold R, Hafen E. Proteome-wide association studies identify biochemical modules associated with a wing-size phenotype in Drosophila melanogaster. Nature communications. 2016;7:12649.

13. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study-linking genes and phenotypes by functional variation in proteins. Genome biology. 2020;21(1):173.

14. Wingo AP, Liu Y, Gerasimov ES, Gockley J, Logsdon BA. Integrating human brain proteomes with genome-wide association data implicates new proteins in Alzheimer's disease pathogenesis. 2021;53(2):143–6.

15. Gusev A, Ko A. Integrative approaches for large-scale transcriptome-wide association studies. 2016;48(3):245–52.

16. Uittenbogaard M, Chiaramello A. Expression of the bHLH transcription factor Tcf12 (ME1) gene is linked to the expansion of precursor cell populations during neurogenesis. Brain research Gene expression patterns. 2002;1(2):115–21.

17. Uittenbogaard M, Chiaramello A. Expression of the basic Helix-Loop-Helix ME1 E-protein during development and aging of the murine cerebellum. Neuroscience letters. 1999;274(3):191–4.

18. Liu P, Liu Z, Wang J, Ma X, Dang Y. HINT1 in Neuropsychiatric Diseases: A Potential Neuroplastic Mediator. 2017;2017:5181925.

19. Steenweg ME, Ghezzi D, Haack T, Abbink TE, Martinelli D, van Berkel CG, et al. Leukoencephalopathy with thalamus and brainstem involvement and high lactate 'LTBL' caused by EARS2 mutations. Brain: a journal of neurology. 2012;135(Pt 5):1387–94.

20. Qu F, Lorenzo DN, King SJ, Brooks R, Bear JE, Bennett V. Ankyrin-B is a PI3P effector that promotes polarized α5β1-integrin recycling via recruiting RabGAP1L to early endosomes. 2016;5.

21. Yao Y, Cui X, Al-Ramahi I, Sun X, Li B, Hou J, et al. A striatal-enriched intronic GPCR modulates huntingtin levels and toxicity. eLife. 2015;4.

22. Dersh D, Iwamoto Y, Argon Y. Tay-Sachs disease mutations in HEXA target the α chain of hexosaminidase A to endoplasmic reticulum-associated degradation. Molecular biology of the cell. 2016;27(24):3813–27.

23. Zhang W, Savelieva KV, Suwanichkul A, Small DL, Kirkpatrick LL, Xu N, et al. Transmembrane and ubiquitin-like domain containing 1 (Tmub1) regulates locomotor activity and wakefulness in mice and interacts with CAMLG. PLoS One. 2010;5(6):e11261.

24. Nystuen A, Legare ME, Shultz LD, Frankel WN. A null mutation in inositol polyphosphate 4-phosphatase type I causes selective neuronal loss in weeble mutant mice. Neuron. 2001;32(2):203–12.

25. Toyota T, Hattori E, Meerabux J, Yamada K, Saito K, Shibuya H, et al. Molecular analysis, mutation screening, and association study of adenylate cyclase type 9 gene (ADCY9) in mood disorders. American journal of medical genetics. 2002;114(1):84–92.

26. Andersen RE, Lim DA. Forging our understanding of lncRNAs in the brain. 2018;371(1):55–71.

27. Zhao H, Jin T, Cheng X, Qin J, Zhang L, He H, et al. GAS5 which is regulated by Lhx8 promotes the recovery of learning and memory in rats with cholinergic nerve injury. Life sciences. 2020;260:118388.

28. Zhao RB, Zhu LH, Shu JP, Qiao LX, Xia ZK. GAS5 silencing protects against hypoxia/ischemia-induced neonatal brain injury. Biochemical and biophysical research communications. 2018;497(1):285–91.

29. Bousman CA, Luza S, Mancuso SG, Kang D, Opazo CM, Mostaid MS. Elevated ubiquitinated proteins in brain and blood of individuals with schizophrenia. 2019;9(1):2307.

30. Saad AK, Marafi D. Neurodevelopmental disorder in an Egyptian family with a biallelic ALKBH8 variant. 2021;185(4):1288–93.

31. Banihashemi S, Tahmasebi-Birgani M, Mohammadiasl J, Hajjari M. Whole exome sequencing identified a novel nonsense INPP4A mutation in a family with intellectual disability. European journal of medical genetics. 2020;63(4):103846.

32. Viallon M, Cuvinciuc V, Delattre B, Merlini L, Barnaure-Nachbar I, Toso-Patel S, et al. State-of-the-art MRI techniques in neuroradiology: principles, pitfalls, and clinical applications. Neuroradiology. 2015;57(5):441–67.

33. Fernandez-Mendoza J, Li Y, Vgontzas AN, Fang J, Gaines J, Calhoun SL, et al. Insomnia is Associated with Cortical Hyperarousal as Early as Adolescence. Sleep. 2016;39(5):1029–36.

34. Kalmbach DA, Anderson JR. The impact of stress on sleep: Pathogenic sleep reactivity as a vulnerability to insomnia and circadian disorders. 2018;27(6):e12710.

35. O'Byrne JN, Berman Rosa M, Gouin JP, Dang-Vu TT. Neuroimaging findings in primary insomnia. Pathologie-biologie. 2014;62(5):262–9.

36. Hammerschlag AR, Stringer S. Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. 2017;49(11):1584–92.

37. Jansen PR, Watanabe K. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. 2019;51(3):394–403.

38. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. Journal of Alzheimer's disease: JAD. 2018;64(s1):S161-s89.

39. Gagliano Taliun SA, VandeHaar P. Exploring and visualizing large-scale genetic associations by using PheWeb. 2020;52(6):550–2.

40. Smith SM, Douaud G. An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. 2021;24(5):737–45.
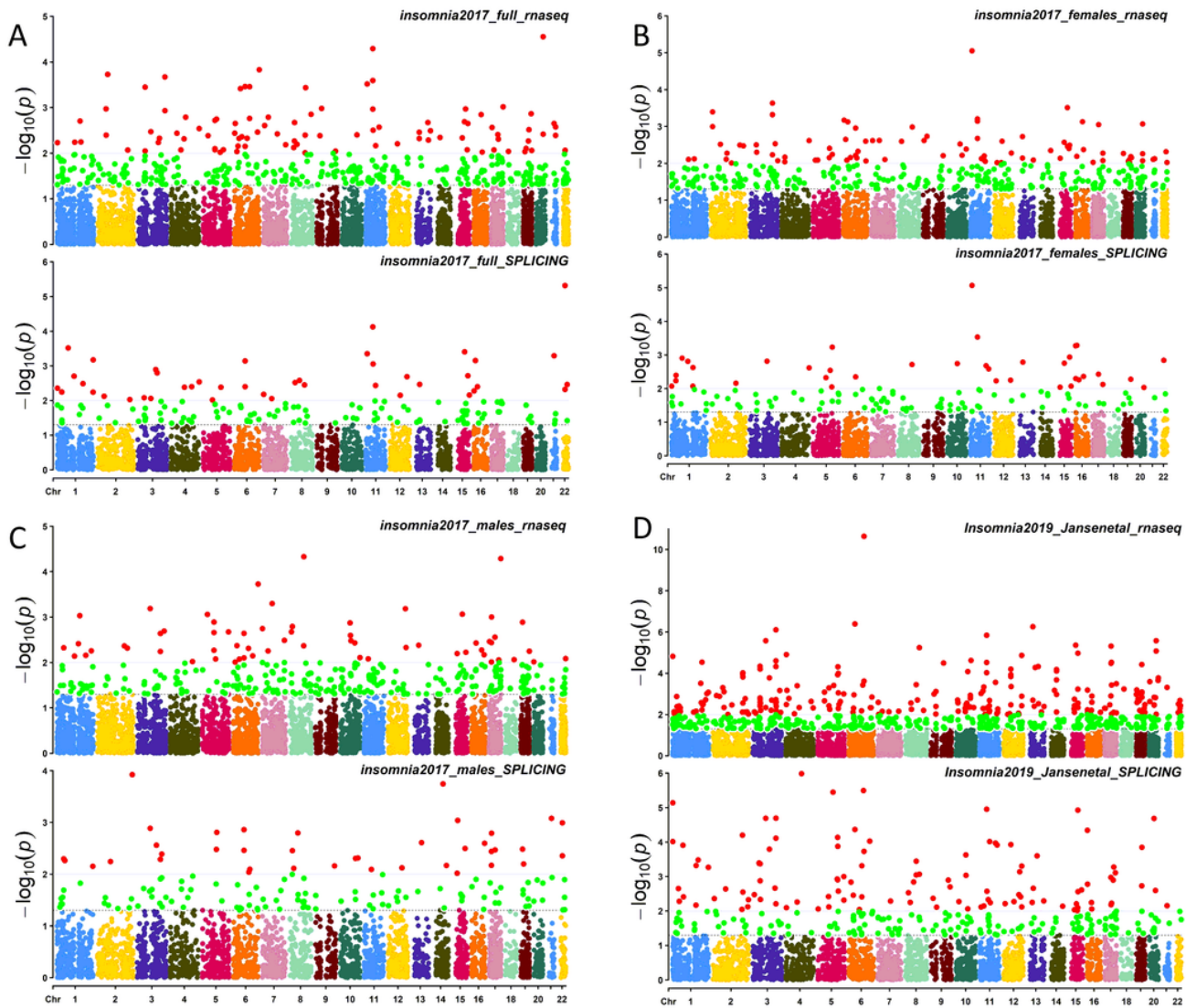
Figures
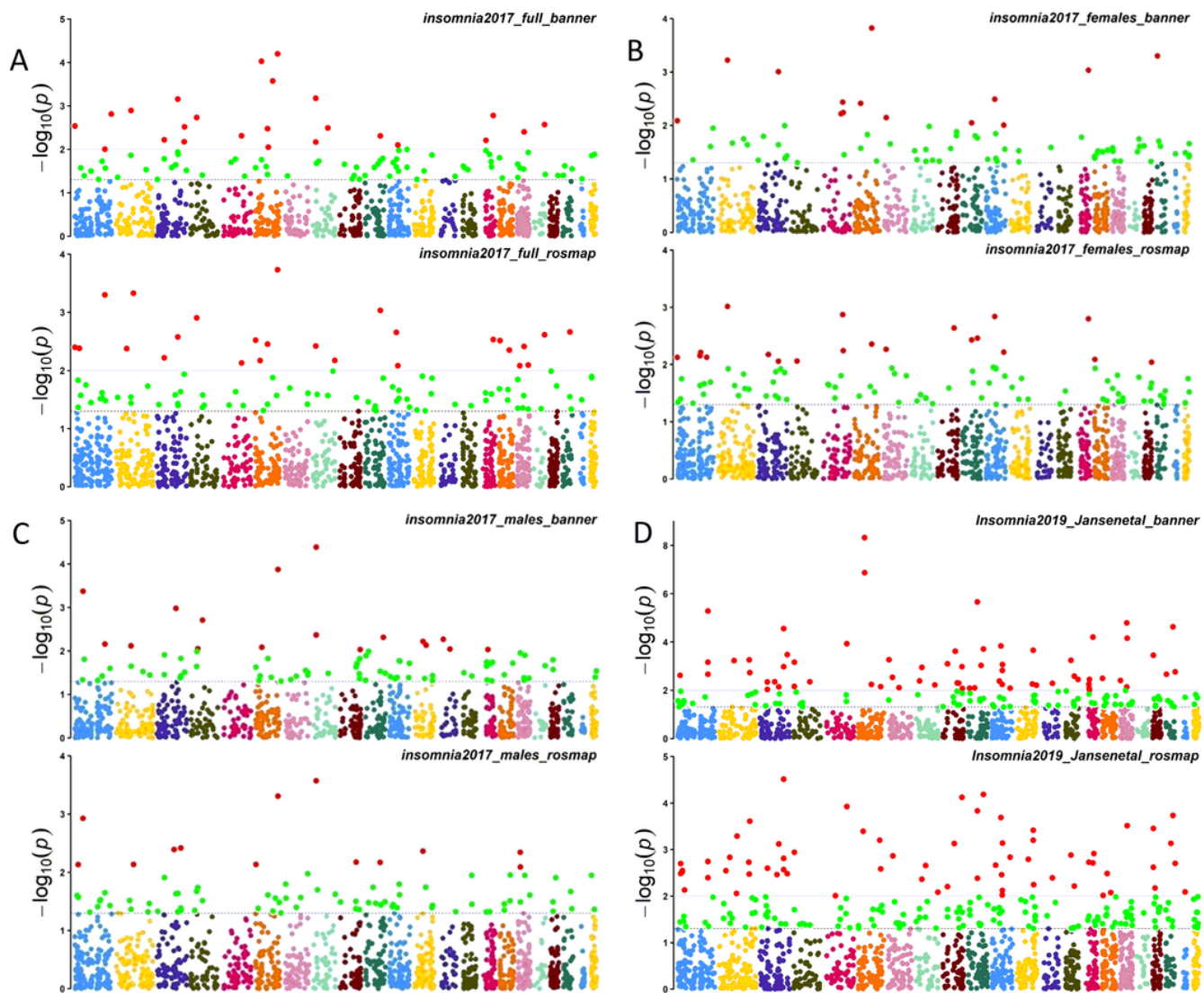
Figure 1

Manhattan Plots of significant human brain proteins identified in PWAS analysis for Insomnia dataset 1 and 2 Note: A refers to results for the full sample in Dataset 1, B refers to results for the female sample in Dataset 1, C refers to results for the male sample in Dataset 1, and D refers to results for Dataset 2. ROS/MAP and Banner means human brain proteomes for PWAS analysis. Each point represents a single association test between a gene and phenotype, in order of genomic position on the x-axis and strength of association −log10(P) on the y-axis. The horizontal line indicates significance (0.05).

Figure 2

Manhattan Plots of significant mRNAs identified in TWAS analysis for Insomnia dataset 1 and 2 Note: A refers to results for the full sample in Dataset 1, B refers to results for the female sample in Dataset 1, C refers to results for the male sample in Dataset 1, and D refers to results for Dataset 2. CBR and CBRS means two datasets of human brain gene expressions for TWAS analysis. Each point represents a single association test between a gene and phenotype, in order of genomic position on the x-axis and strength of association −log10(P) on the y-axis. The horizontal line indicates significance (0.05).