Schizophrenia Bulletin vol. 47 no. 5 pp. 1351–1363, 2021
doi:10.1093/schbul/sbaa167
Advance Access publication April 3, 2021

Spatial Expression Pattern of *ZNF391* Gene in the Brains of Patients With Schizophrenia, Bipolar Disorders or Major Depressive Disorder Identifies New Cross-Disorder Biotypes: A Trans-Diagnostic, Top-Down Approach

Hongyan Ren[1,2], Yajing Meng[1,2], Yamin Zhang[1,2], Qiang Wang[1,2], Wei Deng[1,2], Xiaohong Ma[1,2], Liansheng Zhao[1,2], Xiaojing Li[1,2], Yingcheng Wang[1,2], Pak Sham[3], and Tao Li*[,1,2,4,5,]

[1]Mental Health Center and Psychiatric Laboratory, the State Key Laboratory of Biotherapy, West China Hospital of Sichuan University, Chengdu, Sichuan, China; [2]West China Brain Research Center, West China Hospital of Sichuan University, Chengdu, Sichuan, China; [3]State Key Laboratory of Brain and Cognitive Sciences, Centre for Genomic Sciences, and Department of Psychiatry, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China; [4]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China; [5]Guangdong-Hong Kong-Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence, Guangzhou, China

*To whom correspondence should be addressed; West China Mental Health Centre, West China Hospital, Sichuan University, No. 28th Dianxin Nan Str., Chengdu, Sichuan 610041, China; tel: 86-28-85423561, fax: 86-28-85422632, e-mail: xuntao26@hotmail.com

The results generated from large psychiatric genomic consortia show us some new vantage points to understand the pathophysiology of psychiatric disorders. We explored the potential of integrating the transcription output of the core gene underlying the commonality of psychiatric disorders with a clustering algorithm to redefine psychiatric disorders. Our results showed that an extended MHC region was associated with the common factor of schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD) at the level of genomic significance, with rs7746199 (P = 4.905e-08), a cis-eQTL to the gene *ZNF391*, pinpointed as a potential causal variant driving the signals in the region. Gene expression pattern of *ZNF391* in the brain led to the emergence of 3 biotypes, independent of disorder. The 3 biotypes performed significantly differently in working memory and demonstrated different gray matter volumes in the right inferior frontal orbital gyrus (RIFOG), with a partial causal pathway arising from *ZNF391* to RIFOG to working memory. Our study illustrates the potential of a trans-diagnostic, top-down approach in understanding the commonality of psychiatric disorders.

Key words: cross-disorder/expression prediction/t-SNE/biotype/causality

Introduction

Thanks to the concerted efforts from multi-national consortia, remarkable progress has been made in the field of psychiatric genetics.[1–3] However, translating the findings from large-scale genome-wide association studies (GWAS) to clinical applications remains elusive. The 2 main reasons for such a dilemma are the phenotypic and genetic heterogeneity in the affected populations, due to the current lack of understanding of the pathogenesis of psychiatric disorders. Both linkage and genome-wide association/sequencing studies have observed a notable correlation/ co-segregation of multiple psychiatric disorders.[4,5] Such findings imply that a cross-disorder approach could increase statistical power to detect susceptibility genes and provide a fuller picture of the genetic relationships between psychiatric disorders.[6]

Moreover, a large proportion of the SNPs identified by GWAS so far are located at intergenic/non-coding regions, making direct interpretation of the association signals a difficult task.[7] However, with the establishment of gene expression data repositories, such as GTEx[8] and psychENCODE,[9] it has become possible to link many non-coding SNPs, potentially eQTLs (expression quantitative trait loci), to the expression levels of their target genes. By integrating the genotype information of eQTLs with tissue-specific expression data, methods have been developed to enable the prediction, in an independent sample, of gene expression level in a tissue of interest.[10]

Motivated by these new trends, this study seeks to (1) identify the core causal gene underlying the shared pathophysiology (common factor) of schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD), (2) evaluate the expression levels of the core gene in the brain regions, (3) define new biotypes by re-delineating patients based on the distinctive expression

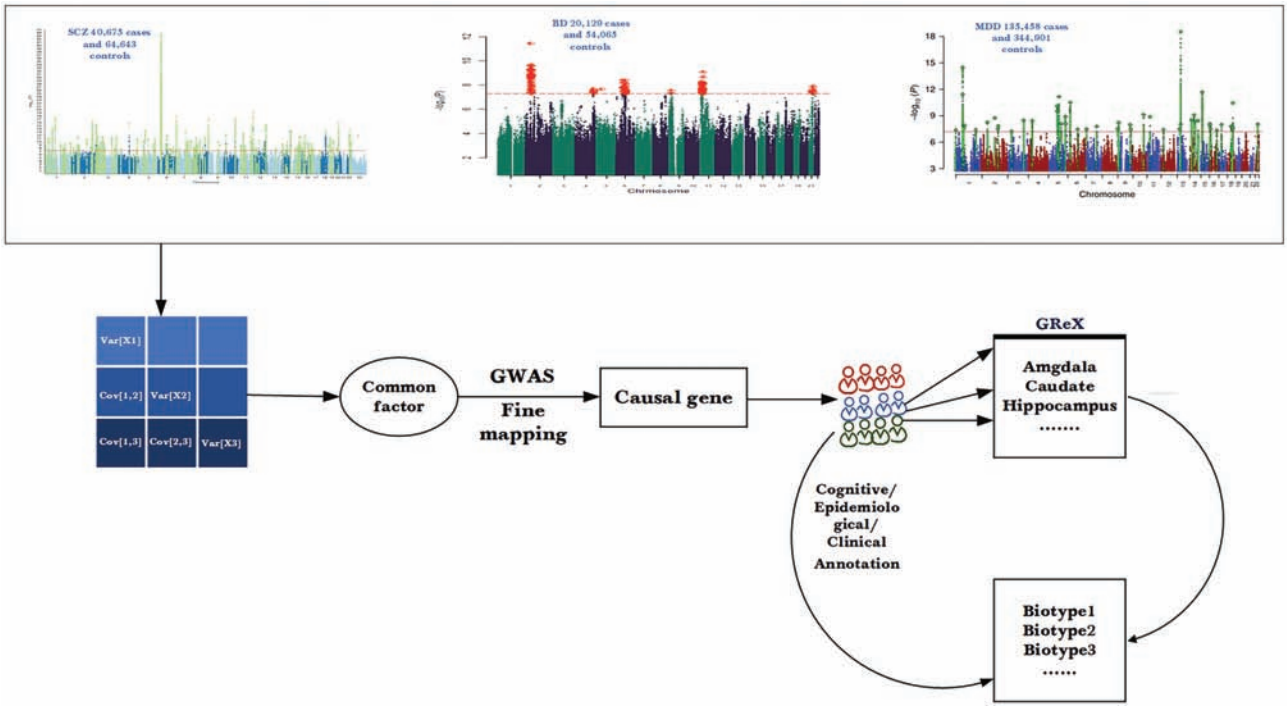

Fig. 1. A schematic for data analysis workflow of our study. We started by a genomic SEM modeling of genetic covariance matrix generated from the summary statistics of published genome-wide association studies (GWASs); then, we identified the causal gene of the common factor underlying 3 psychiatric disorders, followed by the imputation of genetically regulated expression (GReX) of the causal gene in the brain. Subsequently, we re-clustered the patient group based on their GReX profile and annotated the clusters in terms of epidemiological, cognitive, and neuroimaging indexes.

pattern of core gene in the brain regions, and (4) annotate the new biotypes in terms of scalable traits such as cognitive functions and gray matter volumes (GMV; the schematics of the study design is shown in figure 1).

Methods

Genomic SEM and in silico Fine-Mapping

We leveraged the summary statistics from large-scale GWAS of 3 disorders[11–13] to carry out a confirmatory factor analysis using genomic SEM.[14] Genomic SEM takes the genetic covariance matrix estimated from linkage disequilibrium (LD) score regression algorithm and fits a specified model to it; in our study, we specified a common factor (CF) which, to different extents, simultaneously explain the phenotypic variances and covariances of the 3 disorders. Following the emergence of the common factor (CF) underlying the disorders, we implemented a genome-wide scan to map loci conferring effect on CF. In brief, the same SEM model was extended to incorporate SNP effects on each disorder based on GWAS summary statistics, to estimate the effect of SNP on the CF, with corresponding standard error and *P*-value. The extent to which an SNP's effect on the 3 disorders is not mediated through CF is measured by a heterogeneity Q value, with a P-value to indicate the statistical significance of heterogeneity. The mapped loci were annotated using FUMA15; the estimation of the heritability explained by SNPs and the genetic correlation analysis between CF and other phenotypes were performed using the LD hub platform.[16]

The genome-wide scan from above identified an extended MHC region (CHR6: 262,663,11-293,566,87) to be associated with CF. As shown in figure 4 and aligned with the GRCh37 coordinate, the region locates many overlapped genes with multiple significantly associated SNPs. To further pinpoint the potential causal core SNP-gene pair, we carried out an *in silico* fine-mapping analysis in the associated region. We chose 2 algorithms ("FINEMAP"[17]and "iRIGS"[18]) with different theoretical bases to ensure robustness. "FINEMAP" is a stochastic search algorithm for identifying the casual configurations of SNPs in the genomic region of interest and generates, for each SNP, a substantial posterior probability (PP). "iRIGs" is a Bayesian framework which infers PP for the causal SNP-gene pair by integrating 2 layers of information: (1) evidence from multi-omics data, and (2) relationships among genes in biological networks. In our study, we define the SNP that reached a consensus PP of 0.9, obtained by averaging PPs from the 2 algorithms, as being causal. We then chose the gene meeting more than 2 of the following criteria to be causal: (1) the SNP of interest located in the intronic/exonic region thereof; (2) PP for

the casual gene arising from the iRIGs bigger than 0.9; and (3) expression of the gene regulated by the causal SNP as cis-eQTL in the brain according to at least one source of expression data repository.

Prediction of Genotype-Regulated Expression of ZNF391 in the Brain and Cluster Analysis

The fine-mapping above identified zinc finger protein 391 (*ZNF391*) as a potential causal core gene. To further explore the spatial expression pattern of this gene in the brain, we used PrediXcan[19] to predict its expression level in each individual of our in-house sample. In brief, the PrediXcan first used the gene expression data and the genotype data to train the elastic net models, with the expression level of the gene as the response variable and the genotypes of each SNP as the primary predictor variable. These models generated, for each SNP, a regularized weight of prediction. Subsequently, using genotype data from an independent sample, the predictive weights of SNPs within 1MB of the gene start or end (according to the GENCODE version 12) were aggregated using the following formula (1) to yield genetically regulated expression level (GReX) for the target gene in the target tissue:

$$\widehat{GReX} = \sum_k w_k x_k + \epsilon \quad (1)$$

In the equation above, x_k is the number of reference alleles for SNP k.

In our study, we used the pre-stored predictive weights for the expression of *ZNF391* in the 11 brain regions from GTEx (https://gtexportal.org) and the dorsal lateral prefrontal cortex (DLPFC) from the Common Mind Consortium (CMC),[20] to predict the expression of *ZNF391* for the participants in our in-house sample. To further guarantee the validity of genotypes we used for the GReX prediction of *ZNF391*, we randomly chose 20 individuals (5 for each group) for the genotyping of 50 SNPs included in the GReX prediction using a Sequenom MassARRAY platform according to manufacturer's protocol, and removed any SNPs the genotype of which indicative a inconsistency with genotyping and imputation. All participants signed the consent form, and the study was proved by the ethic committee of West China Hospital, Sichuan University.

Using the linear regression in R, we then identified the brain regions with differentiated expression of *ZNF391* between the cross-disorder patients and the controls (reference level), with the first 3 population PCs included as covariates.

Following the identification of regions with differential gene expression, we employed a method, t-distributed stochastic neighbor embedding (t-SNE),[21] for the clustering of patients according to their spatial pattern of *ZNF391* GReX (we denoted the clusters as biotypes hereafter). T-SNE uses random walks on neighborhood graphs to allow the implicit structure of all data to influence how a subset of the data is displayed. Compared to many other similar nonparametric techniques, t-SNE can better capture much of the local structure of the high-dimensional data while also revealing global structure such as the presence of clusters at several scales.[21] We iterated on different values of its super parameter, perplexity (range 50–150) to find the most optimal value which can, to the largest extent, distinguish between the clusters (biotypes).

Sensitivity Analysis

Following the identification of clusters in our sample, 2 layers of sensitivity analysis were conducted to ensure the robustness: firstly, we used an alternative method, K-means clustering, to validate our primary findings; secondly, we applied t-SNE clustering to an independent dataset including 84 patients with first-episode schizophrenia and who were genotyped using different genome-wide genotyping chip, aiming to replicate the biotype numbers identified in our primary analysis, with the sample summary, quality control and imputation of genomic raw data described in detail in the supplementary material.

Comparison of Cognitive Performance and Neuroimaging Measures Between Different ZNF391 Biotypes

Some post-hoc analyses were then implemented to further explain identified biotypes. Firstly, a Kruskal-Wallis

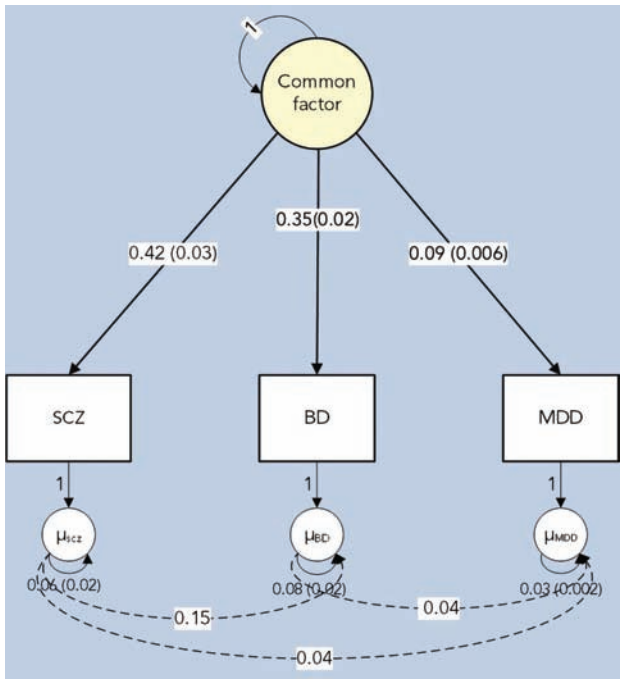

Fig. 2. Pathway diagram of the genomic SEM solution for a common factor (CF) underlying schizophrenia (SCZ), bipolar disorder (BD), and major depressive disorder (MDD).

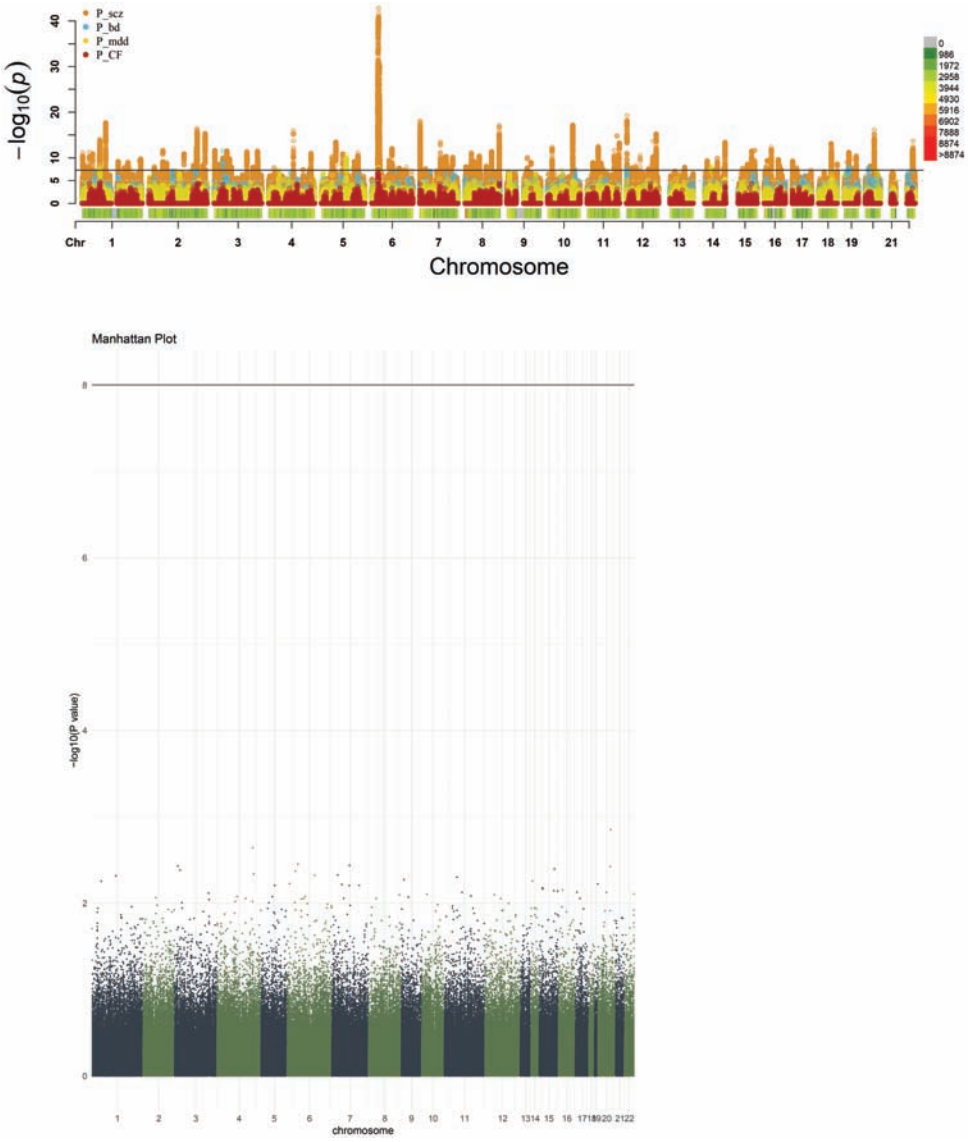

Fig. 3. (a) The Manhattan plot for the composite results from schizophrenia (SCZ), bipolar disorder (BD), major depressive disorder (MDD), and common factor genome-wide association study (CF GWAS); the P_scz: the P values of SCZ GWAS by Pardiñas, A.F. et al., P_bd: the *P* values of BD GWAS by Stahl, E. A. et al., P_mdd: the *P* values of MDD GWAS by Wray, N.R. et al., P_CF: the *P* values of CF GWAS (b) Manhattan plot of Q value for each SNP. In both plots, each point denotes the −log10 *P*-value for each included SNPs, the black line marks the threshold for genome-wide significance.

test was used to compare the mean of continuous demographic features, including age and years of education, between 3 biotypes; categorical features, including gender and disorder distribution, were compared using a Chi-square test. Univariate analysis of variance was conducted for the comparison of cognitive performance and GMV, with both cognitive tests and GMV being adjusted for age, gender, and years of educations before comparison. The details of cognitive performance evaluation and MRI scanning were described in the supplementary table S1. All of the statistical analysis was carried out in the R, version 3.5.3 (R foundation). Given the high correlation between measures of cognitive domains and GMV between brain areas, we chose a liberal threshold of $\alpha = 0.05$ as the significance level.

Exploratory Causal Analysis

The previous step uncovered significant differences between *ZNF391* biotypes in GMV of RIFOG, and in working memory measures (DMS_TC and DMS_TC_A). To further test the causal relationship from gray matter volume (GMV) in right inferior frontal orbital gyrus (RIFOG) to working memory, we carried out a MR analysis by setting the *ZNF391* biotype as the instrumental variable (IV) and using a 2-stage least square

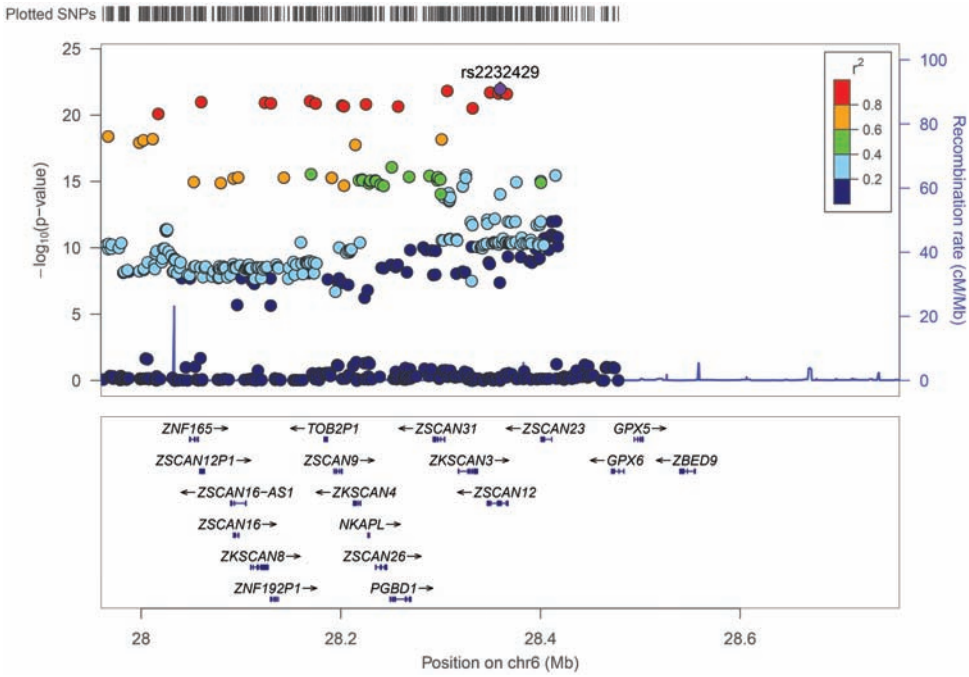

Fig. 4. A regional association plot of the top SNPs and their linkage disequilibrium, the diamond denotes rs2232429, which showed the strongest signal of association in CF GWAS.

method.[22] The analysis was carried out in the R (package "AER"). We set a nominal threshold of $\alpha = 0.05$ as the significance level.

Results

Identification of Common Factor (CF) and ZNF391 as its Causal Core Gene

As illustrated by the pathway diagram (figure 2), a confirmatory factor analysis of SCZ, BD and MDD using genomic SEM revealed a common factor (CF). The loadings of CF on the SCZ, BD, and MDD were 0.42, 0.35, and 0.09, respectively, with a comparative fit index (CFI) and a standardized root mean square residual (SRMR) indicative of a well-fitted model (CFI = 1, SRMR = 1.62×10^{-9}). Besides, consistent with previous studies,[23] the covariance between SCZ and BD is the largest among these 3 disorders (0.15).

The genome-wide scanning mapped one genomic region in chromosome 6 (262,663,11-293,566,87) containing 39 SNPs to the common factor (CF) at a genome-wide significant level (5×10^{-8}, the Manhattan plot displayed in figure 3a. The top SNP, rs2232429, is located at an intronic region of *ZSCAN12* (P =2.06×10^{-8} , figure 4). Besides, no SNP was found to have significant heterogeneity, with the Manhattan plot of Q values for each SNP displayed in figure 3b. The tissue enrichment analysis using MAGMA in FUMA exhibited a remarkable enrichment of the associated genes in the brain regions (supplementary figure S1). In addition, the SNP heritability of CF was estimated to be 0.2071 (0.0144) and the subsequent partitioned heritability analysis detected the Bonferroni-adjusted ($0.05/75 = 6.66 \times 10^{-4}$) significant enrichment of heritability in 4 cell types (figure 5): Nucleotide_Diversity_10kbL2_0 (P_enrichment = 3.39×10^{-10}), GERP.NSL2_0 (P_enrichment = 9.51×10^{-7}), MAF_Adj_LLD_AFRL2_0 (P_enrichment = 1.73×10^{-5}) and Backgrd_Selection_StatL2_0 (P_enrichment = 5.79×10^{-5}) . The genetic correlation analysis using LDSC indicated that the CF was mainly correlated with psychiatric, cognitive and behavior traits (supplementary figure S2).

The results arising from *in silico* fine-mapping are summarized in table 1. The rs7746199 ($\beta = 0.081$, $P = 4.905 \times 10^{-8}$) was prioritized with the highest consensus PP (0.96); in particular, iRIGS showed that *ZNF391,* 95 kb upstream from which rs7746199 was located according to, as the potential core gene. Based on the calculation and visualization from a centralized platform using S-PrediXcan (https://phenviz.navigome.com/gene_phenotypes/ENSG00000124613.html),[24] the predicted expression of *ZNF391* decreases by at least 2 standard deviations in the mental and behavioral traits (supplementary figures S3 and S4).

Brain Regions Indicative of a Differentiated ZNF391 Expression and Novel Biotypes Arising From Differentiated Expression Pattern

After the sample quality control, population structure analysis, and imputation, we included 501 patients (118 SCZ, 224 BD, 159 MDD) and 250 healthy controls,

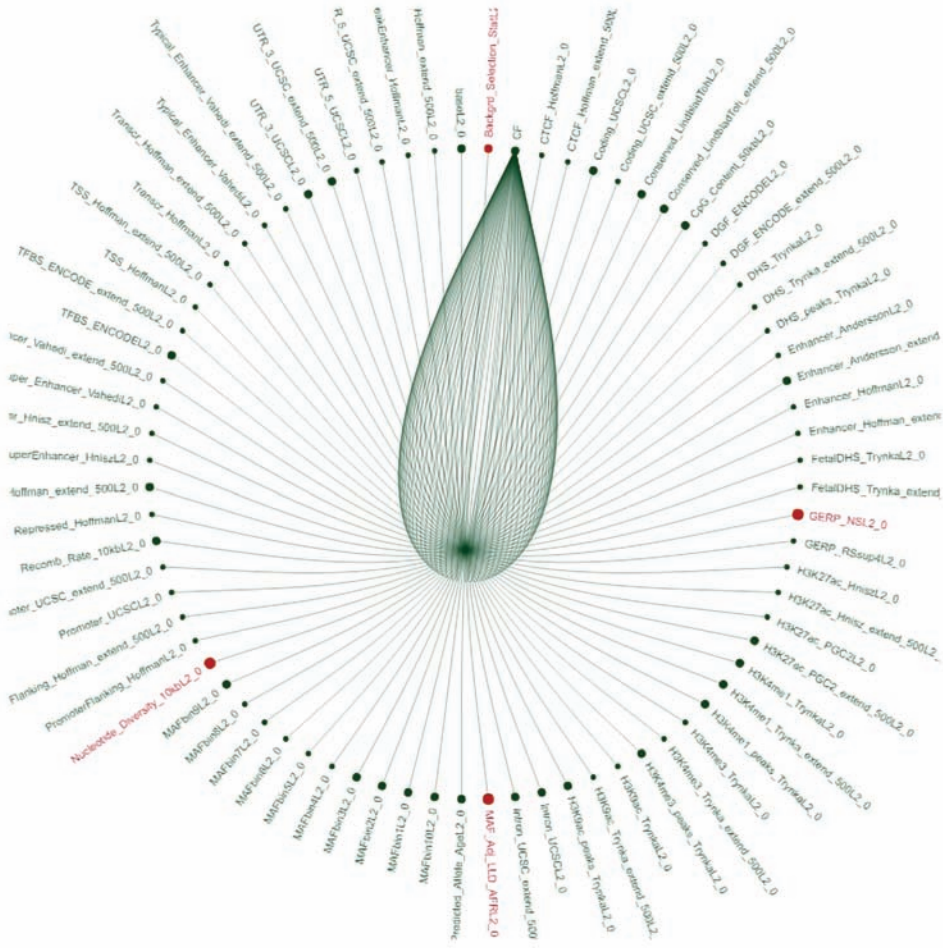

Fig. 5. The analysis result of heritability enrichment in different cell types; the four largest dots (Nucleotide_Diversity_10kbL2, MAF_Adj_LLD_AFRL2_0, GERP.NSL2_0, and Backgrd_Selection_StatL2_0) represent the cell type contributing significantly to the heritability of CF, with the significance surviving a Bonferroni correction for multiple comparisons.

each with 6 230 803 high-quality SNPs, in our subsequent expression prediction of *ZNF391* GReX (consort diagram for the quality control of genotypes shown in supplementary figure S5; demographic characteristics of the included sample in table 2). In total, 50 SNPs were included for the prediction of *ZNF391* GReX, with the imputation details listed in supplementary table S2. According the MAF information from 1000 Genomes, the MAF of these 50 SNPs is significantly correlated between European and East Asian populations ($\rho = 0.95$, $P < .001$), validating the use of expression weights derived from brain sample of European ancestry in our in-house data. A further genotyping of 50 SNPs identified 7 SNPs to fail the genotype consistency between imputation and genotyping, we therefore removed these 7 SNPs from the further analysis, leaving 43 SNPs for the prediction expression of *ZNF391*.

As demonstrated in table 2, significant difference existed among different diagnostic groups in age and years of education. Further population structure analysis indicated, displayed in supplementary figure S6, that no population stratification was detected in our sample, all included individuals clustering in the East Asian population.

The result of the comparison of *ZNF391* GReX was illustrated in figure 6 and the supplementary table S3; 8 brain regions showed significantly differentiate GReX ($P < 0.05$) between cross-disorder patient group and controls, but not between groups of SCZ, BD and MDD, indicative a shared pathology underpinned by these brain regions.

Three clusters (denoted as biotypes henceforth) arose from the clustering analysis of the spatial profile of *ZNF391* GReX in the 8 brain regions; as indicated in figure 7, the most optimal inter-cluster distance could be achieved by setting the hyper-parameter, perplexity, as 96 (supplementary figure S7). For these 3 newly identified biotypes, a χ^2 test did not find significant difference in the distribution of disorders between them(figure 7b).

Using genotyping, 6 out of 50 SNPs sensitivity analyses using K-means clustering and the analyses using an independent sample confirmed the primary

Table 1. Genome-Wide Scanning and Fine-Mapping Results for the Top Loci ($P < 5 \times 10^{-8}$) Conferring Risk to the Common Factor

									FINEMAP		iRIGS		
rsid	CHR	Position	Allele1	Allele2	MAF	β	SE	*P*	PP	log10bf	Gene	PP	Consensus PP
rs7746199	6	27261324	T	C	0.17	.08	0.01	4.90E-08	1	7.93	*ZNF391*	0.91	0.96
rs6904596	6	27491299	A	G	0.11	.11	0.01	3.32E-08	0.53	1.44	*ZNF184*	1.00	0.77
rs17693963	6	27710165	A	C	0.10	.10	0.01	4.22E-08	0.58	1.52	*HIST1H2BN*	0.46	0.52
rs13195636	6	27509493	A	C	0.10	.11	0.01	4.01E-08	0.01	−0.46	*ZNF184*	1.00	0.51
rs13209332	6	27520752	A	T	0.10	.11	0.01	3.78E-08	0.01	−0.46	*ZNF184*	1.00	0.51
rs13191227	6	27390115	C	G	0.09	.11	0.01	3.83E-08	0.01	−0.52	*ZNF184*	1.00	0.51
rs10484399	6	27534528	A	G	0.10	.11	0.01	3.97E-08	0.01	−0.55	*ZNF184*	1.00	0.51
rs13207689	6	27369704	C	G	0.10	.11	0.01	3.90E-08	0.01	−0.53	*ZNF184*	0.97	0.49
rs13207082	6	27251379	A	G	0.09	.11	0.01	4.01E-08	0.01	−0.48	*ZNF391*	0.92	0.47
rs13195040	6	27413924	A	G	0.10	.11	0.01	3.97E-08	0.01	−0.55	*ZNF184*	0.90	0.46
rs13196692	6	27379119	T	C	0.09	.11	0.01	4.22E-08	0.01	−0.55	*ZNF184*	0.89	0.45
rs13204012	6	28201531	A	G	0.10	.11	0.01	3.20E-08	0.02	−0.29	*PGBD1*	0.76	0.39
rs13205211	6	28203056	T	G	0.10	.11	0.01	3.27E-08	0.02	−0.29	*PGBD1*	0.74	0.38
rs13197574	6	28060239	T	C	0.10	.11	0.01	2.93E-08	0.01	−0.45	*ZSCAN16*	0.72	0.37
rs13200214	6	28017250	T	C	0.10	.11	0.01	4.05E-08	0.01	−0.57	*ZSCAN16*	0.71	0.36
rs13201308	6	28130089	T	C	0.10	.11	0.01	3.02E-08	0.01	−0.49	*ZSCAN16*	0.70	0.36
rs13205911	6	28124114	T	C	0.10	.11	0.01	2.96E-08	0.01	−0.45	*ZSCAN16*	0.65	0.33
rs13208096	6	28225311	C	G	0.10	.11	0.01	3.10E-08	0.02	−0.29	*PGBD1*	0.64	0.33
rs2232429	6	28359632	T	G	0.10	.12	0.01	2.06E-08	0.19	0.74	*PGBD1*	0.47	0.33
rs13195291	6	28169241	A	G	0.10	.11	0.01	2.84E-08	0.02	−0.44	*ZSCAN16*	0.64	0.33
rs17750424	6	27701122	T	C	0.10	.11	0.01	3.83E-08	0.02	−0.36	*HIST1H2BL*	0.62	0.32
rs13193542	6	27702425	T	G	0.10	.11	0.01	3.37E-08	0.02	−0.44	*HIST1H2BL*	0.62	0.32
rs13193480	6	27702561	A	G	0.10	.11	0.01	3.74E-08	0.02	−0.36	*HIST1H2BL*	0.60	0.31
rs13197633	6	28174757	A	G	0.10	.11	0.01	3.03E-08	0.01	−0.44	*ZSCAN16*	0.60	0.31
rs2232423	6	28366151	A	G	0.10	.12	0.01	2.34E-08	0.13	0.53	*PGBD1*	0.47	0.30
rs2232426	6	28360659	C	G	0.10	.12	0.01	2.22E-08	0.13	0.53	*PGBD1*	0.47	0.30
rs13211507	6	28257377	T	C	0.10	.11	0.01	3.29E-08	0.02	−0.41	*PGBD1*	0.56	0.29
rs13213152	6	28349698	A	G	0.10	.11	0.01	2.25E-08	0.09	0.35	*PGBD1*	0.46	0.27
rs13213986	6	28358009	A	T	0.10	.12	0.01	2.33E-08	0.13	0.53	*PGBD1*	0.41	0.27
rs13214023	6	28332141	A	G	0.10	.11	0.01	3.47E-08	0.03	−0.13	*PGBD1*	0.50	0.27
rs13217619	6	28306671	T	C	0.10	.11	0.01	2.17E-08	0.04	0.01	*PGBD1*	0.46	0.25
rs17750747	6	27730334	T	C	0.10	.11	0.01	3.83E-08	0.02	−0.36	*HIST1H2BN*	0.44	0.23
rs13194781	6	27815639	A	G	0.10	.11	0.01	4.35E-08	0.02	−0.43	*HIST1H2BN*	0.34	0.18
rs13199906	6	27834139	C	G	0.10	.11	0.01	4.14E-08	0.02	−0.43	*HIST1H2BN*	0.32	0.17
rs17751184	6	27775028	T	C	0.10	.11	0.01	3.83E-08	0.02	−0.36	*HIST1H2BN*	0.31	0.16
rs17763089	6	27835218	A	G	0.10	.11	0.01	4.28E-08	0.02	−0.43	*HIST1H2BN*	0.30	0.16
rs17695758	6	27837183	T	C	0.10	.11	0.01	4.19E-08	0.02	−0.43	*HIST1H2BN*	0.30	0.16
rs13199772	6	27834085	A	G	0.10	.11	0.01	4.09E-08	0.01	−0.59	*HIST1H2BO*	0.29	0.15

Table 2. Demographic Characteristics of the Participants Involved in the Inference of *ZNF391* GReX

	SCZ	BD	MDD	HC	Statistics*	*P**
N	118	224	159	292	NA	NA
Age	22.32 ± 7.26	25.43 ± 8.48	29.43 ± 10.17	26.82 ± 8.27	52.744	<0.001
Gender (%male)	49.15%	43.30%	36.48%	40.40%	4.88	0.18
Years of education	11.81 ± 2.82	13.24 ± 3.09	13.39 ± 3.37	14.97 ± 2.82	91.28	<0.001

Note: SCZ, schizophrenia; BD, bipolar disorder; and MDD, major depressive disorder.
*Denotes the statistics and *P*-values derived from the comparison between the patients from 3 disorder groups (SCZ + BD + MDD) and controls.

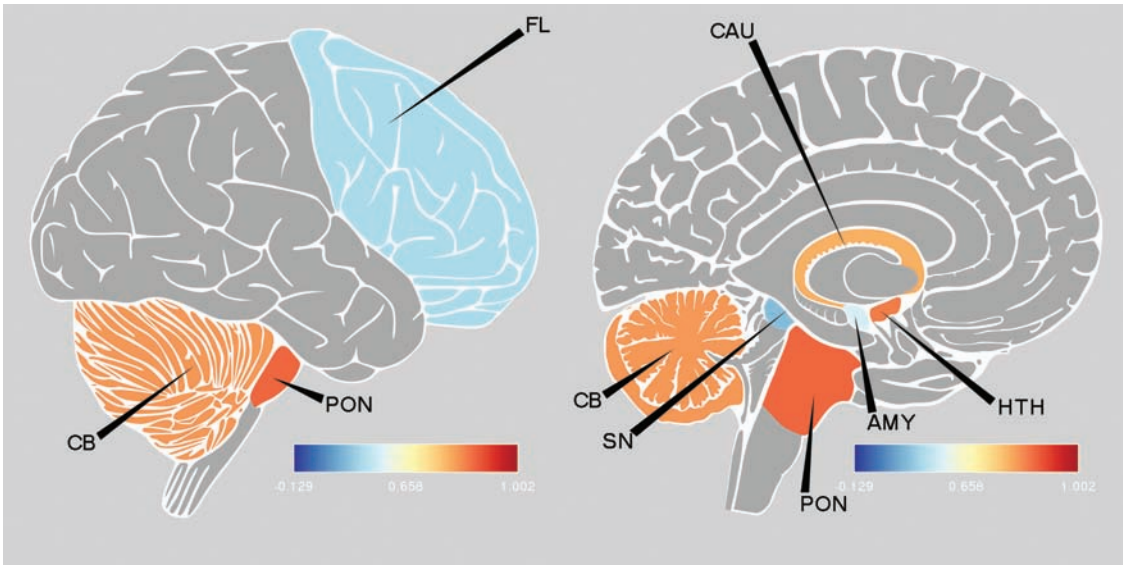

Fig. 6. The brain regions indicating the difference in *ZNF391* GReX between patients and controls; The brain regions correspond to the ones from the human brain transcriptome (HBT) database; FL, frontal lobe; CB, cerebellum; CAU, caudate; AMY, amygdala; HTH; hypothalamus; SN, Substantia nigra.

clustering findings (perplexity = 9, supplementary figure S8).

Following the identification of new clustering patterns (biotypes) existing in the patient group, we compare the biotypes in terms of demographic characteristics, cognitive performance and GMV. While no significant difference was found in age and gender between 3 biotypes, significant difference in the working memory capabilities, measured by using Delay-matching-to-sample (DMS), was detected between biotypes. As demonstrated in figure 8a and 8b, biotype 3 performed significantly better than other 2 biotypes (biotype 3 > biotype 2 > biotype 1) in the total correct rate (DMS_TC, $F = 4.691$, $P = 0.00969$), and the total correct rate-all delays (DMS_TC_A, F = 4.659, $P = 0.01$). Further, the comparison of GMV identified the GMV in RIFOG to be significantly different between 3 biotypes, in the same order as that of DMS comparisons ($F = 9.697$, $P_{anova} = .002$, figure 8c).

A Partial Causal Relationship Existing Between GMV of RIFOG and Working Memory

GReX is mainly derived from genotype information, which makes GReX a potential instrumental variable (IV) in exploring the causal relationship between phenotypes. Indeed, with *ZNF391* biotypes as an IV, a causal path, albeit a partial one, could be drawn from RIFOG GMV to working memory (causal $\boldsymbol{\beta}_{DMS\_TC} = 4.12$, $P = 0.044$; causal $\boldsymbol{\beta}_{DMS\_TC\_A} = 4.51$, $P = 0.036$). As demonstrated in figure 9, intriguingly, the signal of association with *ZNF391* GReX decreased gradually from gray matter volume to working memory; while the hypothetical total effect of pathway (*ZNF391* GReX ↬ RIFOG ↬ working memory) was estimated to be 2.24% on average, the empirical estimation based on a linear model implied an effect around 8.3%, emphasizing the complexity of the biological mechanism linking gene to the complex trait (figure 9).

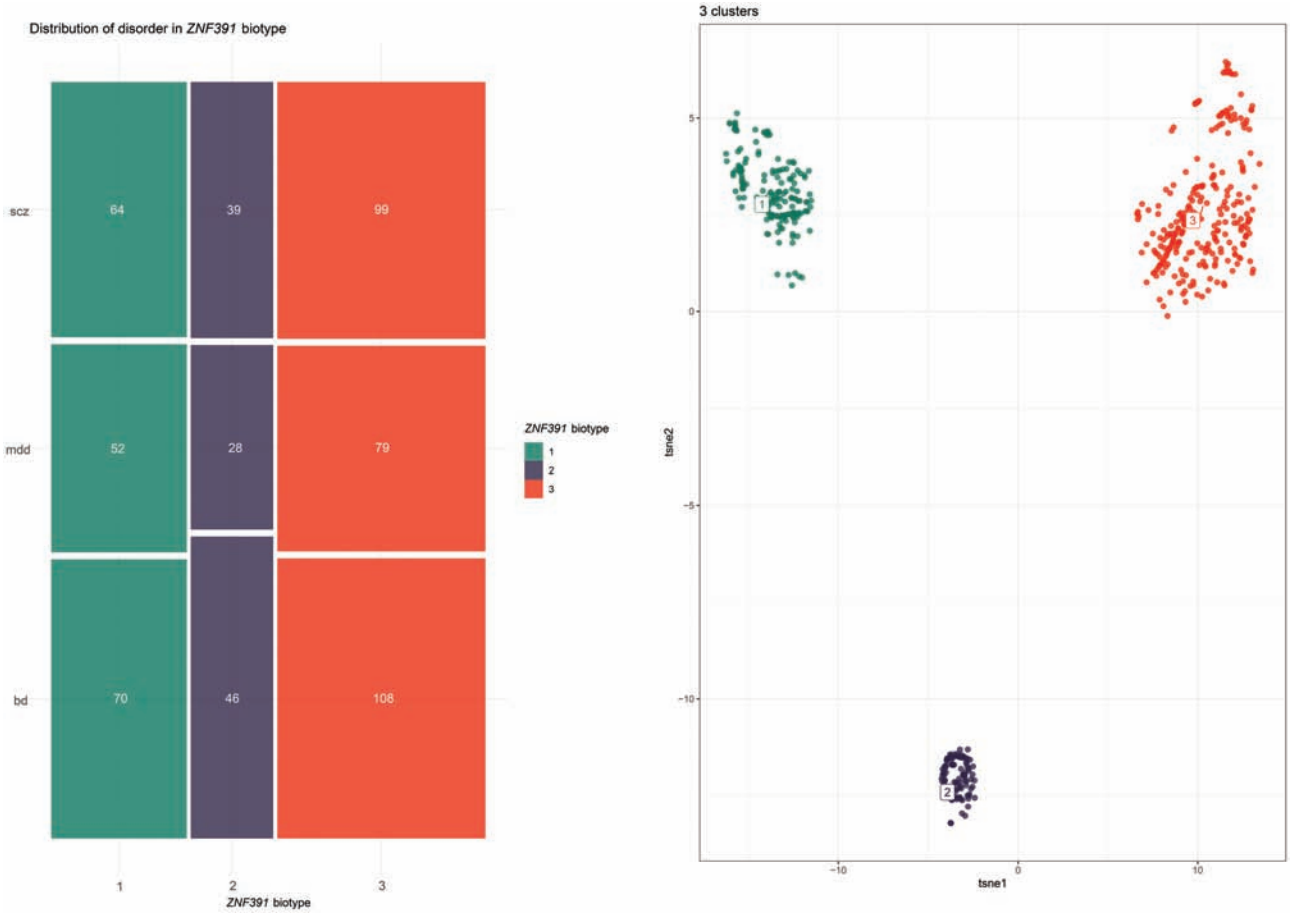

Fig. 7. (a) t-SNE results on the transformed 2-dimensional dataset with a perplexity parameter of 120; (b) the distribution of patients with each disorder in each *ZNF391* biotype.

Discussion

Consistent with previous studies,[5] our study detected a common factor (CF) underlying 3 psychiatric disorders (SCZ, BD, and MDD), and the CF was associated with a segment at the extended major histocompatibility complex (MHC) region. Multiple large-population GWASs of major psychiatric disorders, to a different extent, all pinpointed to the implication of the major histocompatibility complex (MHC) region in the pathogenesis of disorder.[13,25] MHC plays an active role in immune function, especially the adaptive one. Studies using various methods such as cytokines and gene expression of postmortem brains also indicated a critical role that immune components play in psychiatric disorders.[26,27] One seminal study by Sekar et al identified the variations of C4 in the MHC region as the driving signal of association with schizophrenia; when looking beyond the MHC region, Sekar et al detected a pattern of bimodal associations in the sense that 2 peaks of association existed in this region, one was C4, another was the extended region harboring the signal of interest in our study, *ZNF391*.[28] However, due to its complex variation pattern and LD structure, it still remains a challenge to investigate, at a large population scale, the association between the genetic variants in the region hereof and complex diseases. In contrast, the methods, such as Predixcan and FUSION, predict the gene expression level by integrating the in-house genotype information and external tissue expression data, circumventing the pitfalls facing the genetic association study in the highly polymorphous region.

The tissue expression enrichment analysis using MAGMA in FUMA showed a notable enrichment of associated genes in the brain areas, especially the ones essential to the cognitive capacity, such as frontal cortex BA9, cortex and cerebellar hemisphere. It has been widely recognized that cognitive impairment is the core feature of schizophrenia, BD and depression.[29,30] The project, like Research Domain Criteria (RDoC), aimed to redefine psychiatric nosology to better reflect the cognitive commons and differences of mental disorders. Our study utilized both results from consortia-based studies and in-house data to link the *ZNF391* to working memory capacity as a potential standard pathophysiological change underlying the 3 disorders herein. Although future studies are still required to validate our current findings, our study provides another vantage point to the molecular mechanism driving the cognitive deficits in major psychiatric disorders.

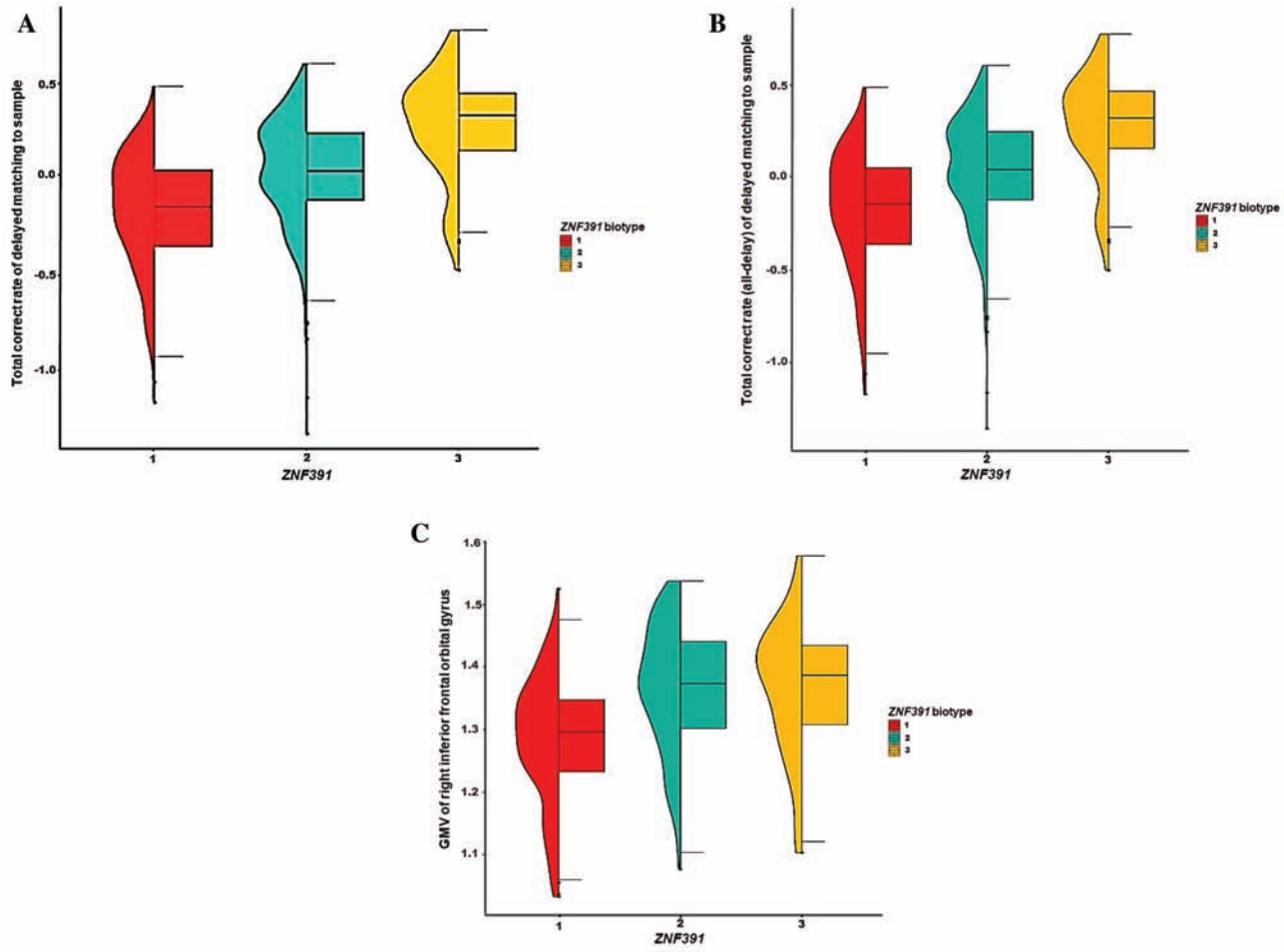

Fig. 8. The comparisons of cognitive measures in delay-matching-to-sample (a,b) and gray matter volume (GMV) of right inferior frontal orbital gyrus (RIFOG, c) between 3 *ZNF391* biotypes.

ZNF391 is highly expressed in the brain, was implicated in depressive phenotypes,[25] schizophrenia,[31] neuroticism.[32] Of 8 brain regions showing difference between *ZNF391 biotypes*, 6 are subcortical regions, 2 are cortical regions, which corroborates the previous findings that psychiatric disorders implicated many brain regions.[33] It is worth noting that the *ZNF391* GReX did not differ significantly between many diagnostic groups, indicative a potential role *ZNF391* might play in the shared pathophysiology of 3 disorders. In the light of current findings, our study took the results to the next level by defining new biotypes in the patient group based on the spatial profile of *ZNF391* GReX in these 8 brain regions. To date, we believe our study is the first one to explore the role of *ZNF391* biotypes in refined phenotypic levels with results showing a significant difference between the biotypes in working memory measures and GMV of RIFOG. Although future replication using an independent sample is needed, our study provides another piece of evidence for the mechanism by which *ZNF391*, even MHC confers risk for major psychiatric disorders.

In the terms of study design, we argue that our top-down approach could both reduce the risk of false-positive discovery existing in the multiple comparisons ("Winner's curse") and provide the robustness for subsequent investigation of the biological mechanism at a higher level of granularity. PredixCan chose a sparse algorithm, elastic net, which could improve predictive efficacy. Further, Mogil et al.[34] showed that best population-specific predicted gene has a highly correlated performance across populations, echoing the findings from another study that 83% of genes differentially expressed among individuals, and 17% differentially expressed among populations, with the most variation coming from within-population.[35] All these studies justify the application of expression-level data in the across-population studies.

Limitations

The current study should be interpreted in light of the following limitations: (1) the present study used the largest available gene expression project, GTEx, to infer GReX, and we are fully aware that most of the donors in GTEx are of European ancestry, which could potentially bias the prediction accuracy in the trans-ethnic samples. Although

Fig. 9. Application of Mendelian randomization for identifying the causal relationship between gray matter volume (GMV) of RIFOG and working memory.

studies of population-specific genetic expression hinted at a minor genetic and expression differentiation of *ZNF391* in studied populations,[36,37] the limited conclusion emphasizes the need to include more ancestrally diverse individuals in publicly accessible -omic database to enhance the interpretation accuracy of precision medicine. (2) In the current study, we defined one gene, the expression of which is regulated by the casual SNP as the core gene. When Prichard *et al.* first proposed the term "omnigenic model",[38] Prichard et al refers to the gene which tends to have biologically interpretable roles in disease and the damage of which by loss of function can have the strongest effect on the disease risk as a "core gene." Unlike some other complex diseases, such as hypertension, cancer and diabetes, it is far more exclusive for psychiatric disorders, due to the lack of biomarker in their diagnosis,

to demarcate between core genes and peripheral genes. Arbitrary as our definition may sound, we argue that it is probably the most scientific way to pinpoint a gene for the further exploration of its role in the pathogenesis of psychiatric disorders and a future study in a larger population with more enrich phenotype profiles is warranted. (3) We chose a lenient threshold of nominal significance to report our findings. While we argue it is a more appropriate approach when phenotypes are highly correlated with each other, we fully recognize the risk of false-positive findings. Future repliactions in a larger sample with more refined information are required. (4) Meanwhile, the weights we used to predict the gene expression of *ZNF391* are derived from the postmortem samples with multiple underlying conditions, including psychiatric ones (CMC); although the original paper developing the prediction model has shown reliable prediction accuracy both across the ethnic groups and across clinical diagnoses, we fully acknowledge the possible selection bias existing in analysis and future replication by using weights generated from large-scale, highly homogenous tissue samples is required.

Conclusion

Taken together, in our current study, we tried to define new cross-disorder biotypes by using a cross-disorder top-down approach. The findings of this study suggest a common pathological mechanism may underlie the 3 disorders. Our results showed that a general liability underlies SCZ, BD and MDD, with *ZNF391* being a potential causal core gene conferring risk of such a general liability. Moreover, we redefined patient group into 3 biotypes based on their expression profiles of *ZNF391* in the brain. The subsequent analysis led to the linkage between *ZNF391*, working memory and gray matter volume (GMV) of RIFOG. Although future studies are required to delve deep into the biological mechanism linking them together, our study provides an example vis-à-vis how to increase the granularity of genetic study to further our understanding of etiology of psychiatric disorders by incorporating the knowledge from the big data and prediction algorithm into the real-world data.

Supplementary Material

Supplementary material is available at *Schizophrenia Bulletin* online.

Figure S1. The results of tissue expression enrichment using MAGMA in FUMA; the red bars are the tissues significantly enriched by the genes associated CF and the black dashed line indicates the Bonferroni-corrected significant threshold

Figure S2. The genetic correlation between common factor and the relevant traits in LD hub database

Figure S3. Multi-tissue plot for effect of rs7746199 on ZNF391 and corresponding posterior possibility, downloaded from https://gtexportal.org/home/snp/rs7746199

Figure S4. Inferred *ZNF391* GReX in the brain regions using the summary statistics of traits of different categories, downloaded from https://phenviz.navigome.com/gene_phenotypes/ENSG00000124613.html

Figure S5. Quality control steps to filter genotyped individuals

Figure S6. Population scatter plot of PC1 and PC2, stratified by disease status (a), and in the context of populations in the 1000 Genomes (b)

Figure S7. Clustering pattern using different perplexity value in the t-SNE analysis

Table S2. Summary of 50 SNPs included in the expression of *ZNF391* GreX

Funding

This work was partly supported by National Nature Science Foundation of China Key Project (81630030 and 81920108018 to T.L.); Special Foundation for Brain Research from Science and Technology Program of Guangdong (2018B030334001); National Key R & D Program, Ministry of Science and Technology, China 2016YFC0904300; 2021 Project for Hangzhou Medical Disciplines of Excellence & Key Project for Hangzhou Medical Disciplines; 1.3.5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZY2016203, ZY2016103, ZYGD200004).

Acknowledgment

The authors have declared that there are no conflicts of interest in relation to the subject of this study.

References

1. Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014; 511(7510): 421.
2. Howard DM, Adams MJ, Clarke TK, et al.; 23andMe Research Team; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci*. 2019;22(3):343–352.
3. Okbay A, Baselmans BM, De Neve JE, et al.; LifeLines Cohort Study. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet*. 2016;48(6):624–633.
4. St Clair D, Blackwood D, Muir W, et al. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet*. 1990;336(8706):13–16.
5. Consortium C-DGotPG. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *The Lancet* 2013;381(9875):1371–1379.

6. Lee PH, Anttila V, Won H, et al. Genome wide meta-analysis identifies genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *bioRxiv* 2019:528117. doi: 10.1101/528117
7. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–753.
8. Consortium G. Genetic effects on gene expression across human tissues. *Nature* 2017;550(7675):204.
9. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science*. 2018;362(6420):eaat8464.
10. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245–252.
11. Pardiñas AF, Holmans P, Pocklington AJ, et al.; GERAD1 Consortium; CRESTAR Consortium. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet*. 2018;50(3):381–389.
12. Wray NR, Ripke S, Mattheisen M, et al.; eQTLGen; 23andMe; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet*. 2018;50(5):668–681.
13. Stahl EA, Breen G, Forstner AJ, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *BioRxiv* 2018:173062. doi: 10.1101/173062
14. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic SEM provides insights into the multivariate genetic architecture of complex traits. *BioRxiv* 2018: 305029. doi:10.1101/305029
15. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8(1):1826.
16. Bulik-Sullivan BK, Loh PR, Finucane HK, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291–295.
17. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016;32(10):1493–1501.
18. Wang Q, Chen R, Cheng F, et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat Neurosci*. 2019;22(5):691–699.
19. Gamazon ER, Wheeler HE, Shah KP, et al.; GTEx Consortium. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091–1098.
20. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci*. 2016;19(11):1442–1453.
21. Maaten Lvd, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(Nov):2579–2605.
22. Angrist JD, Pischke J-S. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press; 2008.
23. Lee PH, Anttila V, Won H, et al. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell*. 2019;179(7):1469–1482.
24. Barbeira AN, Dickinson SP, Bonazzola R, et al.; GTEx Consortium. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825.
25. Howard DM, Adams MJ, Shirali M, et al. Genome-wide association study of depression phenotypes in UK Biobank (n = 322,580) identifies the enrichment of variants in excitatory synaptic pathways. *BioRxiv* 2017:168732. doi: 10.1101/168732
26. Barnes J, Mondelli V, Pariante CM. Genetic contributions of inflammation to depression. *Neuropsychopharmacology*. 2017;42(1):81–98.
27. Gandal MJ, Haney JR, Parikshak NN, et al.; CommonMind Consortium, PsychENCODE Consortium, iPSYCH-BROAD Working Group, Steve Horvath. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Focus (Am Psychiatr Publ)*. 2019;17(1):66–72.
28. Sekar A, Bialas AR, de Rivera H, et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium. Schizophrenia risk from complex variation of complement component 4. *Nature*. 2016;530(7589):177–183.
29. Guo JY, Ragland JD, Carter CS. Memory and cognition in schizophrenia. *Mol Psychiatry*. 2019;24(5):633–642.
30. Rock PL, Roiser JP, Riedel WJ, Blackwell AD. Cognitive impairment in depression: a systematic review and meta-analysis. *Psychol Med*. 2014;44(10):2029–2040.
31. Li H, Chang H, Song X, et al. Integrative analyses of major histocompatibility complex loci in the genome-wide association studies of major depressive disorder. *Neuropsychopharmacology*. 2019;44(9):1552–1561.
32. Nagel M, Jansen PR, Stringer S, et al.; 23andMe Research Team. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet*. 2018;50(7):920–927.
33. Drysdale AT, Grosenick L, Downar J, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat Med*. 2017;23(1):28–38.
34. Mogil LS, Andaleon A, Badalamenti A, et al. Genetic architecture of gene expression traits across diverse populations. *PLoS Genet*. 2018;14(8):e1007586.
35. Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM. Gene-expression variation within and among human populations. *Am J Hum Genet*. 2007;80(3):502–509.
36. Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;369(6509):1318–1330.
37. Gay NR, Gloudemans M, Antonio ML, et al.; GTEx Consortium. Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol*. 2020;21(1):233.
38. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. *Cell*. 2017;169(7):1177–1186.