ARTICLE

Exome-wide screening identifies novel rare risk variants for major depression disorder

Shiqiang Cheng[1,2,3], Bolun Cheng[1,2,3], Li Liu[1,2,3], Xuena Yang[1,2,3], Peilin Meng[1,2,3], Yao Yao[1,2,3], Chuyu Pan[1,2,3], Jingxi Zhang[1,2,3], Chun'e Li[1,2,3], Huijie Zhang[1,2,3], Yujing Chen[1,2,3], Zhen Zhang[1,2,3], Yan Wen[1,2,3], Yumeng Jia[1,2,3] and Feng Zhang[1,2,3]✉

Despite thousands of common genetic loci of major depression disorders (MDD) have been identified by GWAS to date, a large proportion of genetic variation predisposing to MDD remains unaccounted for. By utilizing the newly released UK Biobank 200,643 exome dataset, we conducted an exome-wide association study to identify rare risk variants contributing to MDD. After quality control, 120,033 participants with MDD polygenic risk scores (PRS) values were included. The individuals with lower 30% quantile of the PRS value were filtered for case and control selecting. Then the cases were set as the individuals with upper 10% quantile of the PHQ depression score and lower 10% quantile were set as controls. Finally, 1612 cases and 1612 controls were included in this study. The variants were annotated by ANNOVRA software. After exclusions, 34,761 qualifying variants, including 148 frameshift variant, 335 non-frameshift variant, 33,758 nonsynonymous, 91 start-loss, 393 stop-gain, 36 stop-loss variants were imported into the SKAT R-package to perform single variants, gene-based burden and robust burden tests with minor allele frequency (MAF) < 0.01. Single variant association testing identified one variant, rs4057749 ($P = 5.39 \times 10^{-9}$), within *OR8B4* gene at an exome-wide significance level. The gene-based burden test of the exonic variants identified genome-wide significant associations in *OR8B4* ($P_{SKAT} = 6.23 \times 10^{-5}$, $P_{SKAT\ Robust} = 4.49 \times 10^{-5}$), *TRAPPC11* ($P_{SKAT} = 0.014$, $P_{SKAT\ Robust} = 0.015$), *SBK3* ($P_{SKAT} = 0.020$, $P_{SKAT\ Robust} = 0.025$) and *TNRC6B* ($P_{SKAT} = 0.026$, $P_{SKAT\ Robust} = 0.036$). We identified multiple novel rare risk variants contributing to MDD in the individuals with lower PRS of MDD. The findings can help to broaden the genetic insights of the MDD pathogenesis.

Molecular Psychiatry (2022) 27:3069–3074; https://doi.org/10.1038/s41380-022-01536-4

INTRODUCTION

The burden of major depressive disorder (MDD) continues to grow with significant impacts on health and major social, human rights and economic consequences in all countries of the world. According to the report of The Global Burden of Diseases (GBD), Injuries, and Risk Factors Study 2017, the estimated prevalence is 163 million for MDD [1]. Moreover, the psychiatric disorders have consistently formed more than 14% of age-standardized years lived with disability (YLDs) for nearly three decades, and have greater than 10% prevalence in all 21 GBD regions [1]. MDD often cause severe damage on patients' lives, such as low level of education, marital instability, occupational status, as well as high social costs. With the increasing disease burden of MDD, its prevention and treatment have become an urgent public health issue.

Previous studies have demonstrated that genetic factors play an indispensable role in the development and progression of MDD. Family and twin studies find that the heritability of major depression is likely to be in the range of 31–42%, and the level of heritability is likely to be substantially higher for clinical diagnosed MDD or for subtypes such as recurrent MDD [2]. MDD is a polygenic trait affected by many genetic variants, each with a small effect [3]. While recent genome-wide association studies (GWAS) have discovered multiple loci in the genome linked to depression [4, 5]. There are no conclusive effects of candidate gene on depression, either alone or in combination with life stress [6]. In addition, much of its genetic architecture is still unknown and few rare variants have been detected thus far.

Over the last decade, GWAS has been broadly applied in identifying genetic variants associated with complex traits and diseases, but those results accounted for a limited proportion of disease variability in the population [7, 8]. Furthermore, GWAS risk variants were mostly enriched in non-coding regulatory regions that affected gene expression [9]. The discovery of rare variants in genomic studies, especially the protein coding regions of the genome, is a more attractive and less expensive strategy for identifying rare variants of large effect on disease. For example, a whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants, which involved in immune response and transcriptional regulation [10]. Turcot et al. performed an exome-wide search for low frequency (MAF = 1–5%) and rare (MAF < 1%) single-nucleotide variants (SNVs) associated with BMI [11]. However, pinpointing the causal genes and rare variants of MDD remains a major challenge.

The genetic susceptibility of any disease may be caused by many common small effect genetic variants, which are captured

[1]Key Laboratory of Trace Elements and Endemic Diseases of National Health and Family Planning Commission, Xi'an Jiaotong University, Xi'an, China. [2]Key Laboratory of Environment and Genes Related to Diseases of Ministry of Education of China, Xi'an Jiaotong University, Xi'an, China. [3]Key Laboratory for Disease Prevention and Control and Health Promotion of Shaanxi Province, Xi'an Jiaotong University, Xi'an, China. ✉email: fzhxjtu@mail.xjtu.edu.cn

Received: 28 August 2021 Revised: 8 March 2022 Accepted: 16 March 2022
Published online: 1 April 2022

by polygenic risk scores (PRS), or rare variants of large effect, or a combination of the two [12]. It has been reported by a recent study that individuals with a disease but with low polygenic predisposition may be more likely to harbor a rare genetic variant of large effect [12]. Further, rare genetic variants are not generally captured in PRS and are generally not in linkage disequilibrium (LD) with common variants [13]. According to Zhou et al., GWAS-derived polygenic burden may be used for well-powered rare pathogenic variant discovery or as a sample prioritization tool for whole genome or exome sequencing [14]. In this study, by utilizing the large-scale exome cohort study of UK Biobank, we derived PRS of MDD, and identified pathogenic rare variant for MDD. We aimed to identify rare pathogenic variants in individuals with MDD but with low PRS score.

MATERIALS AND METHODS

UK Biobank cohort

We used the UK Biobank dataset, one of the largest health cohort study data, in this study. During 2006~2010, the UK Biobank recruited more than 500,000 participants aged between 40 and 69 years based on multiple assessment centers located in the UK, and collected a wide range of phenotypes and biological samples. UK Biobank had obtained ethics approval from the North West Multi-center Research Ethics Committee, which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. This research has been conducted using the UK Biobank Resource under Application Number 46478. The authors thank all UK Biobank participants and researchers who contributed or collected data.

UK Biobank genotyping

The UK Biobank conducted high-quality genome-wide genotyping and genotype imputation to the reference panel from the Haplotype Reference Consortium [15]. Briefly, DNA samples of all participants in the UK Biobank were genotyped using either the Affymetrix UK BiLEVE (807,411 markers) or Affymetrix UK Biobank Axiom (825,927 markers) array [15]. SNPs were imputed by IMPUTE2 against the reference panel of the Haplotype Reference Consortium, 1000 Genomes and UK10K projects. Full details regarding these data are available elsewhere [15].

UK Biobank exome-sequencing, genotype-calling and data processing

The UK Biobank exome dataset was downloaded for 200,643 subjects who had undergone exome-sequencing and genotype-calling by the UK Biobank Exome Sequencing Consortium using the GRCh38 assembly with coverage 20X at 95.6% of sites on average [16]. Briefly, exomes were captured using the IDT xGen Exome Research Panel v1.0 including supplemental probes. The OQFE pipeline was used to map raw reads (FASTQ) with BWA-MEM to the GRCh38 reference in a deterministic manner, retaining all supplementary alignments [16]. The OQFE CRAMs were then called for small variants with DeepVariant [17] to generate per-sample gVCFs. These gVCFs were aggregated and joint-genotyped with GLnexus to create a single multi-sample VCF (pVCF) for all UKB 200,643 samples. PLINK files were derived directly from this pVCF. We utilized the OQFE version of the UKB PLINK-formatted exome files (field 23155) in the subsequent analysis.

MDD cases definition

According to previous study, individuals with a disease but with low polygenic predisposition may be more likely to harbor a rare genetic variant of large effect [12]. The MDD cases in this study were defined as individuals with lower polygenic risk scores (PRS) of MDD but with a higher depression score. The individuals with lower polygenic risk scores (PRS) of MDD and lower depression score were set as controls. Briefly, depression score were defined according to patient health questionnaire (PHQ)−9 [18]. PHQ-9 is a total score (0–27) classification algorithm for screening and measuring depression severity, focusing on nine depressive symptoms and signs [18]. The depression score was adjusted by sex, age, and ten principal components of population structure. The MDD PRS of each individuals was calculated from the SNP genotype data for each UK Biobank subject according to the standard approach used by previous studies [19]. Briefly, a total of 21,510 SNPs were first obtained from the Polygenic Score (PGS) Catalog (PGS000145, https://www.pgscatalog.org/) [20], which were identified by a recent large GWAS of MDD [21]. Let PRSn denotes the PRS value of MDD for the nth subject, defined as $PRS_n = \sum_{i=1}^{I} \beta_i SNP_{in}$. β_i is the effect parameter of risk allele of the ith SNP associated with MDD, which was obtained from the published study [21]. I donates the total number of genetic markers. SNP_{in} is the dosage (0, 1, 2) of the risk allele of the ith SNP for the ith study subject. PLINK 2.0 software was used to perform the PRS calculation (http://www.cog-genomics.org/plink/2.0/) [22]. The participants who reported inconsistencies between self-reported gender and genetic gender, without ethic consents and imputation data and genetically related individuals were removed. After quality control, 120,033 participants with MDD PRS values were included. According to a recent study, the estimated power of sampling for apply collapsing tests or variance component score tests, such as the SKAT test was approximately 0.8 when $n = 3000$~4000 [23]. Therefore, the individuals with lower 30% quantile of the PRS value were filtered for case and control select. Then the cases were set as individuals with upper 10% quantile of the PHQ depression score and lower 10% quantile were set as controls. Finally, 1612 cases and 1612 controls were selected, respectively.

Variant filtering and annotation

The SNVs with MAF > 0.01, missing call rates <0.1 among all variants and samples were excluded from the subsequent analysis. All variants were annotated using the standard software packages ANNOVAR on human genome hg38 [24]. The most common method to group rare variants together in population-based genetic analyses is at the level of the gene, usually via a collapsing test [25]. For exome sequencing data, a natural unit for collapsing genetic variants is gene, and collapsing rare variants can increase the power of identifying disease-associated genes [26]. Burden tests assume all rare variants in the target gene have effects on the phenotype in the same direction and of similar magnitude [27]. If a large proportion of the rare variants in a region are truly causal and influence the phenotype in the same direction, then burden tests can have higher power [27]. All non-benign coding variants including frameshift variant, non-frameshift variant, nonsynonymous, start-loss, stop-gain and stop-loss were included in the gene-based burden test.

Single-variant and gene-based association analyses

The exome binary PLINK files were imported into the SKAT R-package to perform single variant test and gene-based burden test. For single rare variant association analysis, we used the SKATBinary_Single option, which computes p-values of single variant test using the firth and efficient resampling methods [28]. Gene-based tests examine the aggregate effect of variants within a region defined by gene annotations. The gene-based burden and robust test were conducted with the "SKATBinary.SSD.All" and "SKATBinary_Robust.SSD.All" option [29, 30]. We performed gene-based burden SKAT testing for genes with at least two qualifying variants contributing to the test. The minimum number of aggregated alleles (i.e., cumulative minor allele counts or cMAC) for a gene-based test was set at ten which has been adopted by previous study [10]. A conservative Bonferroni correction was set for the total number of tests done for the single variants and genes analyzed in this study. Only genes with at least two variants were retained and corrected for multiple testing by using Bonferroni correction. Sex, age, ever smoke, alcohol ever, TownSendIndex were used as covariates in the SKAT analysis. Bonferroni adjusted P-value was used to define statistical significance ($P_{Bonferroni\ adjust} < 0.05$).

RESULTS

General population characteristics

After quality control, a total of 1612 cases (60% female; mean age, 55.01 years) and 1612 controls (76% female; mean age, 48.41 years) were included in this study, respectively. The detailed characteristics of those individuals included in the current study are presented in Table 1. The basic characteristics of the individuals included for PRS analysis were summarized in Supplementary Table 1.

Distribution of identified variants

After exclusions, 64,901 variants with MAF < 0.01 were annotated, including 148 frameshift variant, 335 non-frameshift variant,

33,758 nonsynonymous, 91 start-loss, 393 stop-gain, 36 stop-loss, 29,772 synonymous, and 368 unknown variants. The benign coding variants including 29,772 synonymous and 368 unknown variants were excluded thus leaving 34,761 qualifying variants in the subsequent single rare variants and gene-based burden test. The distribution of the included variants was presented in Supplementary Fig. 1.

Single-variants test result

We performed single variant analyses for the 34,761 variants with MAF ≤ 0.01 and a combined minor allele count of at least ten copies across all participants. Single variant association testing identified one variant, rs4057749 ($P = 5.39 \times 10^{-9}$), within *OR8B4* gene at an exome-wide significance level ($P < 1.44 \times 10^{-6}$) (Fig. 1).

Gene-based burden test result

Gene-based analysis of 34,761 qualifying non-begin variants with MAF ≤ 0.01 map to 2698 genes with at least two variants and cMAC > 10. The gene-based burden and robust test of the exonic variants identified genome-wide significant associations in *OR8B4* ($P_{SKAT\ Bonferroni\ adjust} = 6.23 \times 10^{-5}$, $P_{SKAT\ Robust\ Bonferroni\ adjust} = 4.49 \times 10^{-5}$), *TRAPPC11* ($P_{SKAT\ Bonferroni\ adjust} = 0.014$, $P_{SKAT\ Robust\ Bonferroni\ adjust} = 0.015$), *SBK3* ($P_{SKAT\ Bonferroni\ adjust} = 0.020$, $P_{SKAT\ Robust\ Bonferroni\ adjust} = 0.025$), and *TNRC6B* ($P_{SKAT\ Bonferroni\ adjust} = 0.026$, $P_{SKAT\ Robust\ Bonferroni\ adjust} = 0.036$). The detailed description of the significant genes was presented in Table 2 and Fig. 2. In addition, we have searched those genes for evidence involving brain development, neurological, psychiatric, cognitive, and behavioral traits from the published literature and added the relevant references in Table 2.

DISCUSSION

In this study, in order to identify rare pathogenic variants in individuals with MDD but with lower PRS score and increase our knowledge and understanding of the genetics of MDD, we conducted single variants and gene-based exome-wide association study by utilizing the large-scale exome cohort study of UK Biobank. We identified one single variant and four new candidate genes for MDD in the individuals with low MDD PRS. Trinucleotide Repeat Containing 6B (*TNRC6B*) is a protein coding gene that belongs to the GW182 family of proteins, which are necessary for micro-RNA gene silencing in animal cells [31]. Previous study have demonstrated that these proteins could control circadian behavior in *Drosophila* [32] and bound to known circadian transcription factors in mouse liver [33]. *TNRC6B* heterozygous truncating variants have been reported in three patients from large cohorts of subjects with autism [34, 35]. In a clinical and molecular characterization study performed on 17 patients with *TNRC6B* variants, Granadillo et al. reported pathogenic variants in *TNRC6B* could cause a genetic disorder which characterized by developmental delay/intellectual disability and a spectrum of neurobehavioral phenotypes including autism and attention deficit and hyperactivity disorder [36]. Seven of the patients also have other behavioral abnormalities, such as anxiety, depression, aggressiveness and impulsivity [36]. Ackerman et al. have found that two individuals, one exposed to antidepressants in utero and one not, had a *TNRC6B* likely gene-disrupting mutation [37]. T*NRC6B* was found to be a candidate gene identified by this study which have been reported to be associated neurobehavioral phenotypes by previous study. Further mechanism-based studies are warranted to explore its underlying role in MDD.

OR8B4 belongs to olfactory receptor gene family which interact with odorant molecules in the nose to initiate a neuronal response that triggers the perception of a smell [38]. *TRAPPC11* variants have been associated with movement disorder and neurological abnormalities including cerebral atrophy, ataxia and intellectual disability and several cases were reported with seizures [39–41]. A genome-wide association study showed that *SBK3* is associated with other psychiatric, cognitive and behavioral traits after genome-wide correction [42]. According to those studies, the

Table 1. Basic characteristics of individuals included in this study.

	N	Sex (female)	Age ± SD	Depression score ± SD
Total	3224	2,186 (67.80%)	51.71 ± 7.41	2.26 ± 6.23
Case	1612	966 (59.93%)	55.01 ± 8.03	7.84 ± 3.89
Control	1612	1220 (75.68%)	48.41 ± 4.85	−3.33 ± 0.30

[a]standard deviation SD.

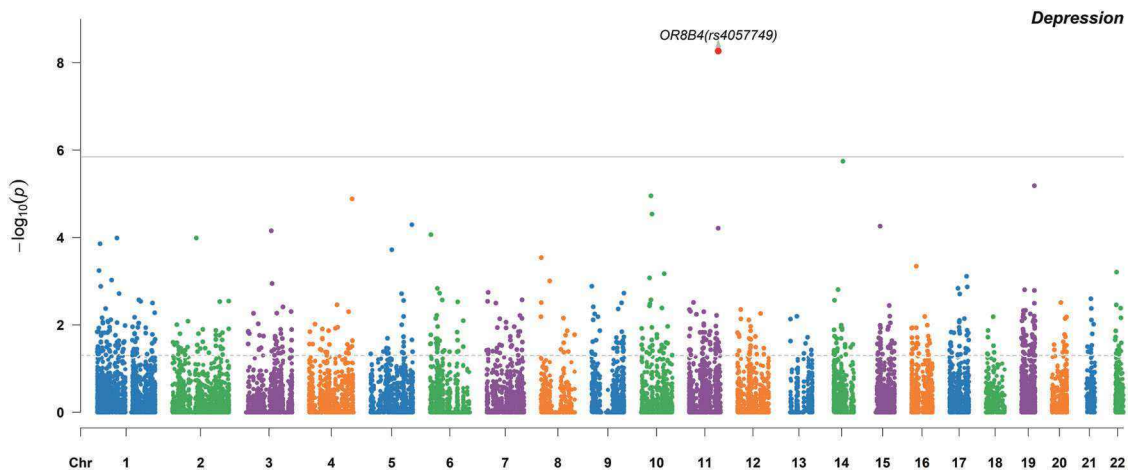

Fig. 1 Manhattan plots of the exome-wide association results for single rare variants test. *Manhattan plot showing the exome-wide association results for single variants. The plot shows the *p*-values against their genomic position for association with depression. Each point represents a *p*-value from SKAT single rare variants (MAF < 0.01). Only variants with a combined minor allele count of >10 were included.

Table 2. Gene-based association analysis results.

Gene	Number of Marker Test	Minor allele count	$P_{SKAT\ Bonferroni\ adjust}$	$P_{SKAT\ Robust\ Bonferroni\ adjust}$	Evidence of function
OR8B4	2	74	6.23×10^{-5}	4.49×10^{-5}	Neuronal response to smell (Malnic et al. [38])
TRAPPC11	3	95	0.014	0.015	Movement disorder and neurological abnormalities including cerebral atrophy, ataxia, intellectual disability and seizures (Bögershausen et al., Koehler et al., Matalonga et al. [39–41])
SBK3	3	85	0.020	0.025	Psychiatric, cognitive and behavioral traits (Wetherill et al. [42])
TNRC6B	2	40	0.026	0.036	Circadian behavior (Zhang et al., Koike et al. (25–26)); autism, developmental delay, intellectual disability and attention deficit and hyperactivity disorder (Babbs et al., Iossifov et al., Granadillo et al. (27–29)); anxiety, depression, aggressiveness and impulsivity (Granadillo et al. [36]).

genes identified in this study were functionally related to neurological, psychiatric, cognitive and behavioral traits. Further studies are warranted to explore their potential roles in the development of MDD.

In most cases, the rare causal variants do not explain the association signals that were previously identified by GWAS with common and predominantly non-functional variants [13]. There are some strategies to further study the genes identified in current study. Firstly, study designs exploiting unique characteristics of different populations by other deep sequencing approaches (e.g., whole genome, target gene resequencing) would boost the power of association studies of rare variants identified in current study. Secondly, our findings provided novel clues into disease mechanisms and targets of biological experiments to gain understanding about the role of the identified genes in MDD pathogenesis. Further cellular and animal experimental studies (e.g., knocking out) will be needed to identify causal variations which account for contribution of structural variants (e.g., larger insertions and deletions, copy number variants, etc.) to MDD risk.

There are some advantages of the current study. Firstly, based on the "common disease-common variant" hypothesis, thousands of genetic loci of MDD have been identified by GWAS to date. However, the common variants identified by GWASs is limited for its effect in explaining the corresponding disease heritability [43]. Whole genome exome sequencing technology enable focused explorations on the contribution of low frequency and rare variants to human traits. According to a recent study, individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening [12]. In this study, the individuals with lower polygenic risk scores (PRS) of MDD, which was computed from the common variants identified by GWAS, but higher depression score were set as cases to identify novel rare genetic basis of MDD. The selected cases are more likely to carry pathogenic rare variants for MDD, which may help to improve clinical care by tailoring diagnostic and/or treatment strategies. Secondly, according to previous studies, the strategy to increase efficiency in GWAS or exome-wide association study is to sequence individuals who are at both ends of a phenotype distribution (those with extreme phenotypes) [44, 45]. This study was conducted using an extreme phenotype design in which exome-wide association study was carried out on patients with upper and lower 10% MDD scores to further enrich pathogenic rare variants, which may increase the statistic efficiency. The findings can help to broaden the genetic insights of the MDD pathogenesis. Two limitations of this study should be noted. Firstly, all subjects included in this study are from European ancestry. Therefore, it should be careful to apply our study results to other ethnic groups. Secondly, the GWAS used for computing PRS in our study defined MDD by using different tools or methods, including 'minimal' and 'strictly' definition. We used depression score which were derived according to (PHQ)−9 from self-reported results to define MDD cases in the subsequent exome-wide association study as phenotype, which may increase the possibility of measurement error and recall bias.

In summary, by utilizing the large-scale exome cohort study of UK Biobank, we derived PRS of MDD, and identified pathogenic rare variant for MDD. The single variant and genes identified in this study were related to neurological, psychiatric, cognitive and behavioral traits. We hope that our findings will provide novel insights into the future etiology study of MDD and serve as a fundamental resource for understanding the role of rare variants on the development of MDD. With the continuous decrease of sequencing cost and a growing effort to build large biobanks and cohorts, rare variant association analysis will be increasingly applied to complex traits and diseases. By exploring rare pathogenic variants in the individuals with depression but with low polygenic risk scores (PRS) by using the large cohorts of UK Biobank with sufficient sample size, we attempt to identify novel

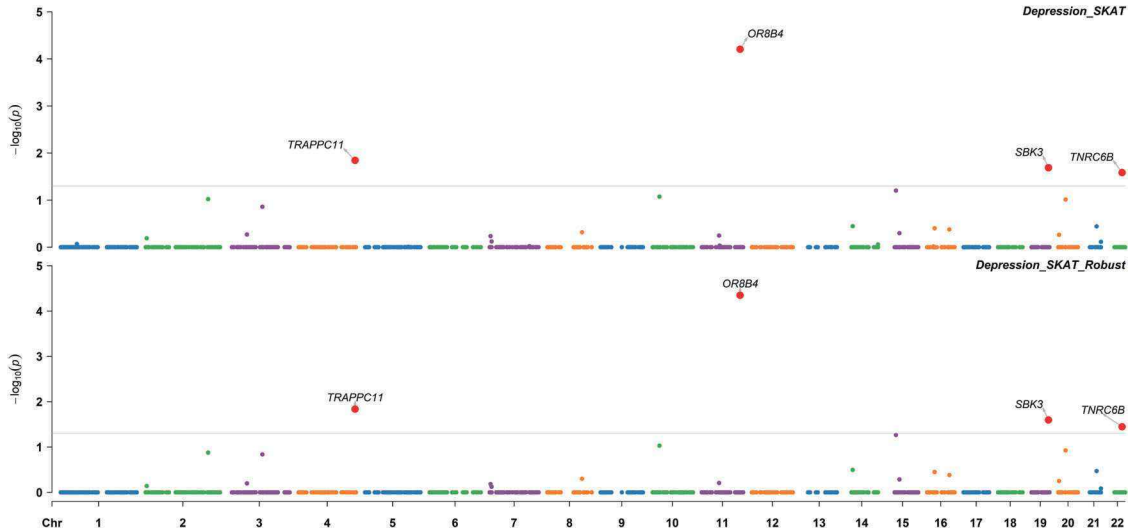

Fig. 2 Manhattan plots of the exome-wide association results for gene-based tests of rare variants. *The plots show the gene-based Bonferroni adjusted p-values against their genomic position for association with depression. Each point represents a *p*-value from SKAT burden test or SKAT robust burden test aggregating rare variants (MAF < 0.01) by gene. Only genes with a cumulative minor allele count of >10 were included.

rare risk variants contributing to MDD, and hopefully providing novel clues for broadening the genetic structure of MDD that is difficult to be captured by common variants identified by previous GWAS.

DATA AVAILABILITY

The UK Biobank data are available through the UK Biobank Access Management System https://www.ukbiobank.ac.uk/. We will return the derived data fields following UK Biobank policy; in due course, they will be available through the UK Biobank Access Management System.

CODE AVAILABILITY

All scripts used to generate the SKAT analyses are available from the authors upon request.

REFERENCES

1. James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2018;392:1789–858.
2. Sullivan PF, Neale MC, Kendler KS. Genetic epidemiology of major depression: review and meta-analysis. Am J Psychiatry. 2000;157:1552–62.
3. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, et al. A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry. 2013;18:497–511.
4. Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nat Genet. 2018;50:668–81.
5. Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shirali M, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat Neurosci. 2019;22:343–52.
6. Culverhouse RC, Saccone NL, Horton AC, Ma Y, Anstey KJ, Banaschewski T, et al. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. Mol Psychiatry. 2018;23:133–42.
7. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
8. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. Am J Hum Genet. 2017;101:5–22.
9. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.
10. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2020;25:1859–75.
11. Turcot V, Lu Y, Highland HM, Schurmann C, Justice AE, Fine RS, et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat Genet. 2018;50:26–41.
12. Lu T, Zhou S, Wu H, Forgetta V, Greenwood CMT, Richards JB. Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. Genet Med: Off J Am Coll Med Genet. 2021;23:508–15.
13. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 2017;18:77.
14. Zhou D, Yu D, Scharf JM, Mathews CA, McGrath L, Cook E, et al. Contextualizing genetic risk score for disease screening and rare variant discovery. Nat Commun. 2021;12:4418.
15. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562:203–9.
16. Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet. 2021;53:942–8.
17. Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, et al. GLnexus: joint variant calling for large cohort sequencing. bioRxiv 2018:343970.
18. Kroenke K, Spitzer RL, Williams JBW, Löwe B. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. Gen Hosp Psychiatry. 2010;32:345–59.
19. Euesden J, Lewis CM, Oreilly PF. PRSice: polygenic risk score software. Bioinformatics. 2015;31:1466–68.
20. Lambert SA, Gil L, Jupp S, Ritchie SC, Xu Y, Buniello A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53:420–5.
21. Cai N, Revez JA, Adams MJ, Andlauer TFM, Breen G, Byrne EM, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet. 2020;52:437–47.
22. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
23. Bjørnland T, Bye A, Ryeng E, Wisløff U, Langaas M. Powerful extreme phenotype sampling designs and score tests for genetic association studies. Stat Med. 2018;37:4234–51.
24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164–e64.
25. Cirulli ET, White S, Read RW, Elhanan G, Metcalf WJ, Tanudjaja F, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat Commun. 2020;11:542.

26. Sun YV, Sung YJ, Tintle N, Ziegler A. Identification of genetic association of multiple rare variants using collapsing methods. Genet Epidemiol. 2011;35: S101–6.
27. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13:762–75.
28. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. Biostatistics. 2016;17:1–15.
29. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91:224–37.
30. Zhao Z, Bi W, Zhou W, VandeHaar P, Fritsche LG, Lee S. UK biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. Am J Hum Genet. 2020;106:3–12.
31. Kalmbach DA, Schneider LD, Cheung J, Bertrand SJ, Kariharan T, Pack AI, et al. Genetic basis of chronotype in humans: insights from three landmark GWAS. Sleep. 2016;40.
32. Zhang Y, Emery P. GW182 controls Drosophila circadian behavior and PDF-receptor signaling. Neuron. 2013;78:152–65.
33. Koike N, Yoo SH, Huang HC, Kumar V, Lee C, Kim TK, et al. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. Science. 2012;338:349–54.
34. Babbs C, Lloyd D, Pagnamenta AT, Twigg SR, Green J, McGowan SJ, et al. De novo and rare inherited mutations implicate the transcriptional coregulator TCF20/SPBP in autism spectrum disorder. J Med Genet. 2014;51:737–47.
35. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515:216–21.
36. Granadillo JL, A PAS, Guo H, Xia K, Angle B, Bontempo K, et al. Pathogenic variants in TNRC6B cause a genetic disorder characterised by developmental delay/intellectual disability and a spectrum of neurobehavioural phenotypes including autism and ADHD. J Med Genet. 2020;57:717–24.
37. Ackerman S, Schoenbrun S, Hudac C, Bernier R. Interactive effects of prenatal antidepressant exposure and likely gene disrupting mutations on the severity of autism spectrum disorder. J Autism Developmental Disord. 2017;47:3489–96.
38. Malnic B, Godfrey PA, Buck LB. The human olfactory receptor gene family. Proc Natl Acad Sci. 2004;101:2584–9.
39. Bögershausen N, Shahrzad N, Chong JX, von Kleist-Retzow JC, Stanga D, Li Y, et al. Recessive TRAPPC11 mutations cause a disease spectrum of limb girdle muscular dystrophy and myopathy with movement disorder and intellectual disability. Am J Hum Genet. 2013;93:181–90.
40. Koehler K, Milev MP, Prematilake K, Reschke F, Kutzner S, Jühlen R, et al. A novel TRAPPC11 mutation in two Turkish families associated with cerebral atrophy, global retardation, scoliosis, achalasia and alacrima. J Med Genet. 2017;54:176–85.
41. Matalonga L, Bravo M, Serra-Peinado C, García-Pelegrí E, Ugarteburu O, Vidal S, et al. Mutations in TRAPPC11 are associated with a congenital disorder of glycosylation. Hum Mutat. 2017;38:148–51.
42. Wetherill L, Lai D, Johnson EC, Anokhin A, Bauer L, Bucholz KK, et al. Genome-wide association study identifies loci associated with liability to alcohol and drug dependence that is associated with variability in reward-related ventral striatum activity in African- and European-Americans. Genes, Brain Behav. 2019;18:e12580.
43. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45:D896–d901.
44. Emond MJ, Louie T, Emerson J, Zhao W, Mathias RA, Knowles MR, et al. Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis. Nat Genet. 2012;44:886–9.
45. Lanktree MB, Hegele RA, Schork NJ, Spence JD. Extremes of unexplained variation as a phenotype. circulation: cardiovascular. Genetics. 2010;3:215–21.

ACKNOWLEDGEMENTS

This study was conducted using the UK Biobank Resource (Application 46478).

AUTHOR CONTRIBUTIONS

SC and FZ conceived and designed the study, and wrote the manuscript; SC and FZ collected the data and carried out the statistical analyses; BC, LL, XL, PM, YY, CP, JZ, CL, HZ, YC, ZZ, YW, and YJ made preparations for the manuscript at first.

FUNDING

This study was supported by the National Natural Scientific Foundation of China (81922059).

COMPETING INTERESTS

The authors declare no competing interests.

ETHICS APPROVAL

Ethical approval of UK Biobank study was granted by the National Health Service National Research Ethics Service (reference 11/NW/0382).

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41380-022-01536-4.

Correspondence and requests for materials should be addressed to Feng Zhang.

Reprints and permission information is available at http://www.nature.com/reprints

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