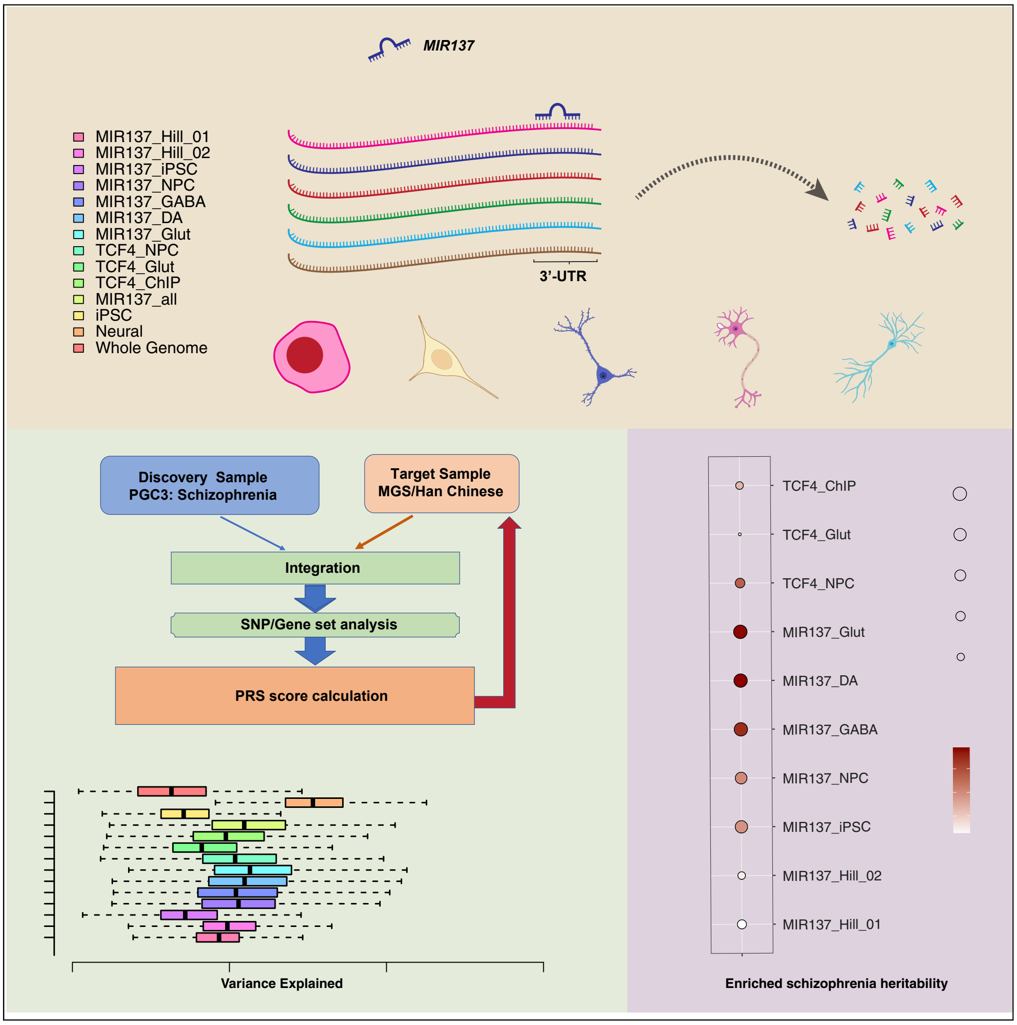
iScience

CellPress
OPEN ACCESS

Article

Cell type-specific and cross-population polygenic risk score analyses of *MIR137* gene pathway in schizophrenia



Yin Yao, Wei Guo, Siwei Zhang, ..., Alan R. Sanders, Weihua Yue, Jubao Duan

yin_yao@fudan.edu.cn (Y.Y.)
jduan@uchicago.edu (J.D.)

Highlights

PRS of neural *MIR137* target genes better explains schizophrenia (SZ) risk variance

SZ risk and SNP heritability explained by *MIR137* target genes is cell type-specific

MIR137 target genes explain a disproportionally larger SZ risk than genomic control

PRS of *MIR137* target genes better explains SZ risk in Europeans than in Han Chinese

Yao et al., iScience *24*, 102785
July 23, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.isci.2021.102785

iScience

CellPress
OPEN ACCESS

Article

Cell type-specific and cross-population polygenic risk score analyses of *MIR137* gene pathway in schizophrenia

Yin Yao,[1,*] Wei Guo,[2] Siwei Zhang,[3] Hao Yu,[4,5,6,7] Hao Yan,[4,5] Hanwen Zhang,[3] Alan R. Sanders,[3,8] Weihua Yue,[4,5,9] and Jubao Duan[3,8,10,*]

SUMMARY

Cell type-specific pathway-based polygenic risk scores (PRSs) may better inform disease biology and improve the precision of PRS-based clinical prediction. For *microRNA-137* (*MIR137*), a leading neuropsychiatric risk gene and a post-transcriptional master regulator, we conducted a cell type-specific gene set PRS analysis in both European and Han Chinese schizophrenia (SZ) samples. We found that the PRS of neuronal *MIR137*-target genes better explains SZ risk than PRS derived from *MIR137*-target genes in iPSC or from the reported gene sets showing *MIR137*-altered expression. Compared with the PRS derived from the whole genome or the target genes of *TCF4*, the PRS of neuronal *MIR137*-target genes explained a disproportionally larger (relative to SNP number) SZ risk in the European sample, but with a more modest advantage in the Han Chinese sample. Our study demonstrated a cell type-specific polygenic contribution of *MIR137*-target genes to SZ risk, highlighting the value of cell type-specific pathway-based PRS analysis for uncovering disease-relevant biological features.

INTRODUCTION

Genome-wide association studies (GWASs) of complex disorders are identifying an increasing number of common risk variants. The most recent schizophrenia (SZ) GWAS meta-analysis of Psychiatric Genomic Consortium (PGC3) with 69,369 patients with SZ and 236,642 controls identified 270 distinct risk loci (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020). Like other complex disorders, SZ is polygenic, i.e., there are possibly hundreds or even thousands of risk-associated genes, each contributing small effect. Beyond the polygenic model, an omnigenic model has also been hypothesized for complex disorders such as SZ, where most if not all genes, expressed in a relevant cell type may contribute to risk through interacting with a core set of genes (Boyle et al., 2017). The polygenic or omnigenic nature of SZ poses challenges to identifying causal variants/genes and better understanding pathogenesis. Polygenic risk scores (PRSs) derived from GWAS data of common variants in a large cohort have been commonly used to improve estimates of risk variance in a target sample and estimate individual risk. However, PRSs derived from genome-wide data can be noisy because they include many SNPs with no risk effect and many risk genes may not be expressed in a cell/tissue type relevant to the disorder. A more effective approach may be deriving PRS from more disorder relevant gene pathways and cell types that are more likely to harbor a higher proportion of causal SNPs (Baker et al., 2018; Zeng et al., 2017), an approach that has been understudied for SZ.

It has been increasingly recognized that many risk genes likely act as part of a gene network/pathway. It is thus imperative to identify core gene network(s) contributing to risk. As opposed to maximizing the predictive power of case-control status in PRS analyses by using genome-wide data, gene set PRS analyses can more specifically detect associations to some biologically informative features (Baker et al., 2018). For instance, a recent PRS study using SZ GWAS SNPs of gene sets highly expressed and dynamically modulated in placenta showed that considering environmental exposure (early-life complications) can increase the predicting accuracy of SZ risk (Ursini et al., 2018), although a later study did not replicate this finding (Vassos et al., 2021). Similarly, by incorporating cell type-specific or developmental stage-specific gene expression information, gene set PRS analysis may inform cell types or developmental stages that are more biologically relevant to a disorder. Furthermore, although genome-wide PRS of one specific ancestry

[1] Department of Computational Biology, Life Science Institutes and School of Life Science and Human Phenomics Institute, Fudan University, Shanghai 200438, China

[2] Genetic Epidemiology Research Branch, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA

[3] Center for Psychiatric Genetics, NorthShore University HealthSystem, Evanston, IL 60201, USA

[4] Peking University Sixth Hospital (Institute of Mental Health), Beijing 100191, China

[5] National Clinical Research Center for Mental Disorders & Key Laboratory of Mental Health, Ministry of Health (Peking University), Beijing 100191, China

[6] Department of Psychiatry, Jining Medical University, Jining, Shandong 272067, China

[7] Shandong Key Laboratory of Behavioral Medicine, Jining Medical University, Jining, Shandong 272067, China

[8] Department of Psychiatry and Behavioral Neurosciences, University of Chicago, Chicago, IL 60637, USA

[9] PKU-IDG/McGovern Institute for Brain Research, Peking University, Beijing 100191, China

[10] Lead contact

*Correspondence: yin_yao@fudan.edu.cn (Y.Y.), jduan@uchicago.edu (J.D.)

https://doi.org/10.1016/j.isci.2021.102785

has been commonly considered inappropriate for predicting risk variance of a different ancestry (Amariuta et al., 2020), a cell type-specific pathway-based PRS may be more portable between different populations. For example, PRS derived by prioritizing GWAS variants in the predicted cell type-specific regulatory elements has substantially increased trans-ancestry portability (Amariuta et al., 2020). Once a gene set is implicated by PRS analysis, the gene set risk scores for each individual can be used to stratify individuals for functional follow-up and to formulate a more tractable and testable hypothesis, e.g., by generating patient-specific cell models in those individuals with high PRS and perform pathway-specific genetic perturbation.

MicroRNAs (miRNAs) and their targets are ideal candidates for gene set-based PRS analyses. This is because each miRNA often has many possible target genes, i.e., be a "master regulator." In addition, miRNAs are central to brain development and play an important role in the cause of neurodevelopmental disorders such as SZ (Guarnieri and DiLeone, 2008; Im and Kenny, 2012). Among GWAS-implicated SZ risk loci (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS)Consortium, 2011; Lee et al., 2012; Purcell et al., 2009; Shi et al., 2009; Stefansson et al., 2009; Ripke and Consortium, 2014), one of the most strongly associated ($p = 6.6 \times 10^{-16}$) (Schizophrenia Psychiatric Genome-Wide Association Study (GWAS)Consortium, 2011; Ripke and Consortium, 2014) spans *MIR137*. The SZ risk alleles of the *MIR137* locus also predict cognitive deficits (Green et al., 2012), and the deletion of the *MIR137* locus has been associated with intellectual disability (Willemsen et al., 2011). *MIR137* is highly expressed in brain and enriched at neuronal synapses (Willemsen et al., 2011). It regulates neuronal differentiation, migration, and dendritogenesis (Silber et al., 2008; Smrt et al., 2010; Sun et al., 2011; Szulwach et al., 2010; Volvert et al., 2012). Of interest, several SZ GWAS loci encompass *MIR137* targets; for instance, *MIR137* has been known to regulate other genes that harbor SZ-associated variants such as *ZNF804A*, *TCF4*, *CACNA1C*, *CSMD1*, and *C10orf26* (Collins et al., 2014; Kwon et al., 2013; Remmers et al., 2014; Wright et al., 2013), suggesting a central hub role for *MIR137* in an SZ gene network. A recent PRS analysis of genes whose expression was altered by artificial manipulation of *MIR137* expression in a neural progenitor cell (NPC) line uncovered an interesting association of the *MIR137* gene set PRS with cognitive function (Cosgrove et al., 2017). However, another study did not find a significant association between *MIR137* PRS and cortical thickness, surface area, or hippocampal volume measures linked to memory function (Cosgrove et al., 2018). Although interesting, these *MIR137* PRS analyses have several limitations: (1) the relative contribution of *MIR137* gene set PRS to the overall SZ risk variance has not been estimated; (2) because of the notorious weak expression correlation of *MIR137* and its targeting genes (Forrest et al., 2017; Topol et al., 2016), the gene set defined by artificially manipulating *MIR137* expression in a single cell line may not be representative of an actual specific *MIR137* gene network; and (3) NPCs used to define the gene sets for *MIR137* PRS analyses may not represent the most relevant cell type for SZ (Skene et al., 2018).

Here, by using PGC3 (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020) SZ GWAS summary statistics of European ancestry samples, we have compared the PRS performance of different gene sets of *MIR137* targets in explaining SZ risk variances in the Molecular Genetics of Schizophrenia (MGS) European ancestry cohort (Shi et al., 2009) and in a Han Chinese SZ case-control sample (Li et al., 2017; Yu et al., 2017) (Figure 1). We also compared with performances of PRSs derived from sets of target genes of *TCF4*, a transcriptional master regulator of SZ (Doostparast Torshizi et al., 2019). We discovered that the PRSs of neuronal *MIR137*-target gene sets explain disproportionally larger SZ risk variance, compared with the *TCF4*-target gene sets and genome-wide SNP sets. Furthermore, with human induced pluripotent stem cells (hiPSCs) as a model, we found that the PRS predictive efficacy of *MIR137*-target gene sets is neuron specific. Finally, we evaluated the PRS portability of *MIR137* gene sets from European Ancestry to Han Chinese samples. Our study underscores the value of cell type-specific gene set PRS analysis in understanding the biology of a polygenic or omnigenic disorder.

RESULTS

To evaluate the contribution of gene pathway(s) pertinent to *MIR137*, a leading noncoding SZ risk locus, to genetic risk of SZ, we addressed three specific questions in a set of PRS analyses: (1) Between genes perturbed by altering *MIR137* expression in a specific cell type (Cosgrove et al., 2017) and those predicted *MIR137* target genes specifically expressed in a disease-relevant cell type, which constitutes an optimal gene set of the *MIR137* pathway for PRS analysis? (2) What is the proportion of SZ risk variance that can be explained by disease risk variants in the *MIR137* gene pathway? (3) Is the *MIR137* gene pathway-based PRS portable across ancestry? To address these questions, we have compiled 10 gene sets (Tables 1 and S1), including the lists of genes perturbed by *MIR137* alteration and previously used for *MIR137* gene

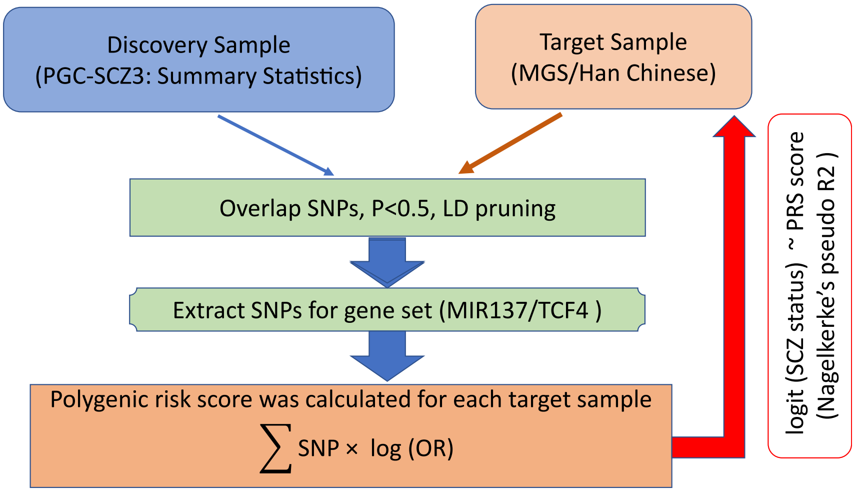

Discovery Sample
(PGC-SCZ3: Summary Statistics)
Target Sample
(MGS/Han Chinese)
Overlap SNPs, P<0.5, LD pruning
Extract SNPs for gene set (MIR137/TCF4)
Polygenic risk score was calculated for each target sample
∑ SNP × log (OR)
logit (SCZ status) ~ PRS score
(Nagelkerke's pseudo R2)

Figure 1. Workflow for PRS analysis

pathway PRS analysis by Hill et al. (Cosgrove et al., 2017), genes predicted to be *MIR137* targets and expressed in hiPSC or in various hiPSC-derived neuronal cell types (Forrest et al., 2017; Zhang et al., 2018, 2020). For a comparison, we also included sets of genes that were perturbed by knocking down a transcriptional master regulator, *TCF4* (also an SZ GWAS-implicated risk gene), in hiPSC-derived neuronal cells (Doostparast Torshizi et al., 2019), as well as empirical binding targets of *TCF4* in chromatin immunoprecipitation (ChIP-seq) (Forrest et al., 2017). For each gene set, the ''significant'' SNPs were selected based on seven predetermined significance thresholds ($P_d < 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5) in the discovery sample (the European ancestry PGC-SZ3 minus MGS) to derived PRS for each individual in two target samples, the European ancestry MGS (2,681 cases, 2,653 controls) (Shi et al., 2009) and a Han Chinese sample (4,288 cases, 5,512 controls) (Li et al., 2017; Yu et al., 2017). We estimated the proportion of variance explained by the aggregate risk score by each SNP/gene set in each sample (Figure 1).

Neuronal *MIR137* target genes better explain SZ risk variance than genes with expression altered *in vitro* by *MIR137*

Genes with expression altered by either over-expression or knocking-down of *MIR137* in NPCs have been recently used to correlate PRS to SZ-related cognitive phenotypes (Cosgrove et al., 2017, 2018). Given the known weak expression correlation of *MIR137* and its targeted genes (Forrest et al., 2017; Topol et al., 2016), we hypothesized that predicted *MIR137* gene targets specifically expressed in a disease-relevant cell type may constitute a better gene set for PRS analysis. As shown in Figure 2A (Table S3), although with similar number of genes (Table S2), PRS derived from the predicted *MIR137* target genes that are expressed in hiPSC-derived NPCs (N_{gene} = 925 in *MIR137*_NPC), GABA (N_{gene} = 986 in *MIR137*_GABA), DA neurons (N_{gene} = 1,034 of *MIR137*_DA), or Glut (N_{gene} = 991 of *MIR137*_Glut) neurons explains greater SZ risk variance in the MGS sample than PRS derived from genes expressed in hiPSCs (N_{gene} = 909 in MIR137_iPSC) or those with expression altered by *MIR137* in NPC (N_{gene} = 1,017 and 934 in *MIR137*_Hill_01 and *MIR137*_Hill_02, respectively). The PRS of gene set in Glut showed the best performance (Figure 2A and Table S3), consistent with the notion that it is among the most disease-relevant cell types for SZ (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020). We observed similar patterns across different GWAS *p*-value thresholds ($P_d < 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5) and at different gene intervals, i.e., 20 (Figure 2A) or 100 kb (Figure S1A) extended genic regions.

Although different gene sets in our above-described PRS analysis (Figure 2A) contain similar number of genes, the results may be cofounded by variable gene length (e.g., neuronal genes are often larger) (Zylka et al., 2015) and SNP numbers in different gene sets. To account for these confounding factors, we first estimated the relative PRS enrichment for each gene set in Figure 2A by correcting for the number of tested SNPs in each gene set (Figure S2). We found that *MIR137* target gene sets in iPSC-derived neuronal cells explained disproportionally more SZ risk variance than SNPs from *MIR137* target genes expressed in iPSC across all GWAS *p*-value thresholds (Figure S2). However, *MIR137* target gene sets in iPSC-derived

Table 1. Gene/SNP sets used in the PRS analysis

List no.	List name	Genes	Description	Reference
1	*MIR137*_Hill_01	1,017	Genes altered by over-expressing *MIR137* in an NPC cell line (previous study)	(Cosgrove et al., 2017; Hill et al., 2014)
2	*MIR137*_Hill_02	934	Genes altered by knocking down *MIR137* in an NPC cell line (previous study)	(Cosgrove et al., 2017; Hill et al., 2014)
3	*MIR137*_iPSC	909	TargetScan *MIR137* target genes expressed in hiPSC	(Zhang et al., 2020)
4	*MIR137*_NPC	925	TargetScan *MIR137* target genes expressed in hiPSC-derived NPCs	(Zhang et al., 2020)
5	*MIR137*_GABA	986	TargetScan *MIR137* target genes expressed in hiPSC-derived GABA neurons	(Zhang et al., 2018)
6	*MIR137*_DA	1,034	TargetScan *MIR137* target genes expressed in hiPSC-derived DA neurons	(Zhang et al., 2018)
7	*MIR137*_Glut	991	TargetScan *MIR137* target genes expressed in hiPSC-derived Glut neuron	(Forrest et al., 2017; Zhang et al., 2020)
8	*TCF4*_NPC	1,000	Top ranking genes altered by knocking down *TCF4* in hiPSC-derived NPCs	(Doostparast Torshizi et al., 2019)
9	*TCF4*_Glut	1,000	Top ranking genes altered by knocking down *TCF4* in hiPSC-derived Glut neurons	(Doostparast Torshizi et al., 2019)
10	*TCF4*_ChIP	1,000	Top ranking genes with promoter/enhancer bind by *TCF4* (previous study)	(Forrest et al., 2017)

The lists of *MIR137*_Hill_01 and *MIR137*_Hill_02 have been used by Hill et al. in previous PRS analysis (Cosgrove et al., 2017; Hill et al., 2014).

neuronal cells did not seem to outperform *MIR137*_Hill_01 and *MIR137*_Hill_02 in this analysis, especially at $P_d < 0.01$, 0.05, 0.1, 0.2, or 0.3.

We then performed permutation (N = 1,000) to randomly select the same number of SNPs in each gene set for calculating PRS (Table S3). As shown in Figures 2B and S1B, we found that *MIR137* target gene sets expressed in hiPSC-derived NPCs or neurons explain more SZ risk variance in the MGS sample than the genes expressed in non-neuronal hiPSCs. Overall, these neural *MIR137* target gene sets also explained disproportionally more SZ risk variance than SNPs of the two gene sets with expression altered by artificially changing *MIR137* expression in NPCs (*MIR137*_Hill_01 and *MIR137*_Hill_02), especially with SZ GWAS *p*-value cut-offs of <0.3, <0.4, or <0.5 (Figure 2B). Although this assessment may still be confounded by the number of SNPs in each permutation at some SZ GWAS p value cut-offs, we noted that the SNP numbers in the permutation tests were overall similar between gene sets, especially with exactly the same number of SNPs (n = 1,500) at SZ GWAS *p*-value cut-off <0.5 (Figure S3), suggesting a greater relative enrichment of variance explained by *MIR137* target gene sets from hiPSC-derived neural cells (versus those from iPSC, *MIR137*_Hill_01 or *MIR137*_Hill_02). We also compared the SZ PRS of these SNP sets with an equal number of SNPs from all the predicted *MIR137* target genes independent of whether they were expressed in any particular cell type (*MIR137*_all) and found that SNPs from *MIR137*_all explained similar SZ risk variance as SNPs from neuronal gene sets, but significantly more variance than genes expressed in iPSC (Figure 2B and Table S4), suggesting a major contribution of neuronal *MIR137* target genes to SZ risk (also see below). It is noteworthy that *MIR137*_Hill_01 and *MIR137*_Hill_02 gene sets, although explaining proportionally less variance than the neuronal *MIR137* target gene sets, still explain proportionately more variance than the control genomic SNP set (Figure 2B and Table S4). The statistical significance of a Student's *t*-tests was shown in Figure 3 and Table S4 on 1,000 permuted Nagelkerke's pseudo R^2 values between any two gene sets. Altogether, these results suggest the importance of PRS analysis using cell type-specific gene sets.

To independently validate the results from gene set-specific PRS analysis, we further conducted a partitioned SZ heritability analysis by a stratified linkage disequilibrium score regression (sLDSC) (Bulik-Sullivan et al., 2015; Finucane et al., 2015) for different gene sets. With PGC SZ wave_3 GWAS summary statistics, we estimated the proportion of explained SZ SNP heritability (h^2) and the proportion of SNPs in each gene set (Figure 2C and

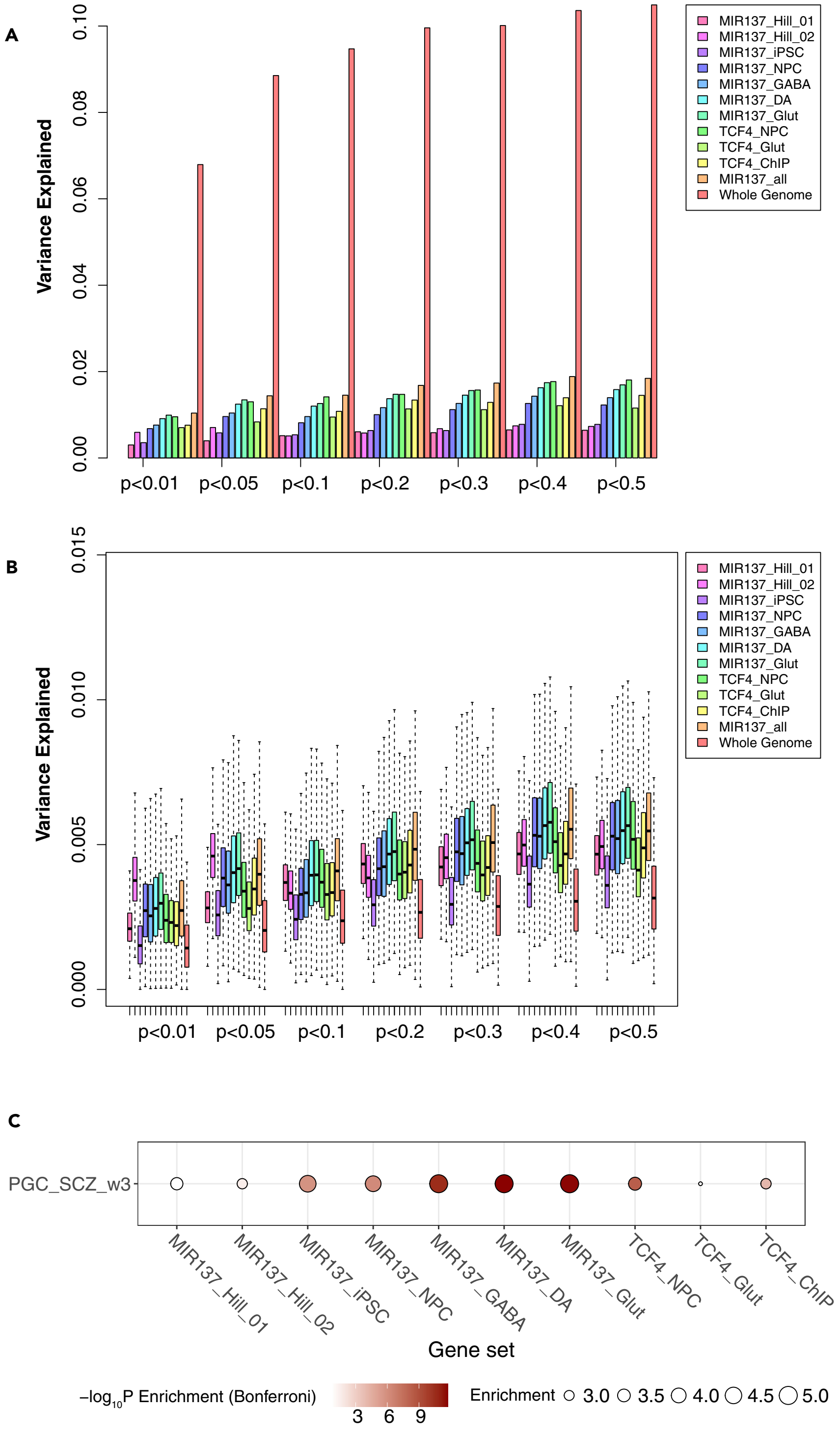


Figure 2. PRS analysis of different lists of target genes of *MIR137* and *TCF4* annotated in a 20-kb window for the MGS sample

Gene sets in (A) and (B) were described in Table 1.

(A) PRS result using LD-pruned SNPs ($r^2 < 0.2$) for each gene list. The variance explained in the target sample is based on risk scores derived from an aggregated sum of weighted SNP risk allele effect sizes estimated from the discovery samples

Figure 2. *Continued*

at seven significance thresholds ($p < 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5). The y axis indicates the percentage of phenotypic variance explained by the PRS (Nagelkerke's pseudo R^2).

(B) The permutation (N = 1,000) PRS results based on randomly selected 1,500 SNPs for each gene set. Data are presented as mean ± SEM.

(C) Stratified LDSC analysis showing stronger enrichment of SZ GWAS SNP heritability in neural *MIR137*-target gene sets. Shown are the folds of enrichment of SNP heritability (proportion of explained heritability h^2 normalized by the proportion of SNPs in each gene set) and the statistical significance (Bonferroni-corrected *p* values; in $-\log_{10}$ scale).

Table S5). We found that all the tested *MIR137* target gene sets are enriched (3- to 5-fold) for SZ SNP heritability, with neuronal gene sets, especially those expressed in Glut neurons, showing strongest enrichment, whereas *MIR137*_Hill_01 and *MIR137*_Hill_02 gene sets showed the least enrichment (Figure 2C and Table S5). On comparison with the enrichment score of each gene set, the coefficient from the sLDSC analysis gave even stronger support for proportionally more SZ SNP heritability explained by neuronal *MIR137* target gene sets than those expressed in iPSC or *MIR137*_Hill_01 and *MIR137*_Hill_02 gene sets (Table S5). The result from the sLDSC analysis is thus largely consistent with our gene set-specific SZ PRS analysis using randomly selected SNPs from each gene set, supporting that neuronal *MIR137* target genes better explain SZ risk variance/heritability than non-neuronal genes or those with expression altered *in vitro* by *MIR137*.

SZ PRS explained by *MIR137* target genes is mainly attributed to a small number of neuron-specific genes

Given that the *MIR137* target gene sets expressed in hiPSC-derived neurons explain 2- to 3-fold more SZ variance than those in hiPSCs (Figures 2A and 2B), we speculated that those *MIR137* target genes specifically expressed in neurons contributed to the difference. As shown in Figure 4A, most genes (N = 784; Table S1) were expressed in both hiPSC and hiPSC-derived NPCs, DA, GABA, and Glut neurons. Therefore, we hypothesize that the PRS of a very small subset of *MIR137* target genes expressed only in NPCs and/or neurons explained a larger proportion of the SZ risk variance. These genes include those at genome-wide significant SZ risk loci, such as *CACNA1C* uniquely expressed in Glut/DA/GABA neurons, *CSMD1* and *DPYD* expressed only in Glut neurons, and *GRIA1* expressed in all neuronal cells. It is noteworthy that the dysregulated synaptic expression of *GRIA1* in dendritic protrusions of Glut neurons was found associated with the SZ risk allele of *MIR137* (Forrest et al., 2017). Some other contributing neuronal genes are not among those top-ranking SZ GWAS risk genes, e.g., *PDE10A* (Table S1), a gene that was recently shown to be a key target that mediates the *MIR137* function in heterozygous *MIR137* knockout mice (Cheng et al., 2018).

To explore the hypothesis that a small subset of neuronal *MIR137* target genes show a higher enrichment of SZ risk variance, we restricted the PRS analysis to the neural *MIR137* target genes (N = 182; denoted as "Neural") or those genes also expressed in iPSC (N = 784; denoted as "iPSC") (Figure 4B). Similar to the SNP permutation test in Figure 2B, we compared the SZ variance explained by each subset of genes by using the same number of 1,500 SNPs in each permutation for each gene set at GWAS *p*-value threshold of <0.5 (Figure 4B). We found that the subset of neural *MIR137* target genes (N = 182) disproportionally explained much greater SZ risk variance than those also expressed in iPSC (N = 784) or any other gene set tested in Figure 2B (Figures 3 and 4B).

We further conducted Hi-C-coupled MAGMA (H-MAGMA) analysis (de Leeuw et al., 2015; Sey et al., 2020) to orthogonally validate the enrichment of SZ GWAS risk in the small subset of neuronal *MIR137* target genes (N = 75 in NPC to N = 161 in Glut) as opposed to genes expressed in both iPSC and neural cells (N = 784; Figure 4A). H-MAGMA improved the prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles (Sey et al., 2020), specifically the Hi-C data from iPSC-derived neurons (Rajarajan et al., 2018) for the current study (Figure 4C). Consistent with the result from our PRS analysis (Figure 4B), we found that *MIR137* target genes specifically expressed in each subtype of neural cells, but not those also expressed in iPSC, showed robust enrichment of SZ GWAS risk (enrichment *p*-values ranging from 3.2×10^{-4} in GABA to 9.6×10^{-7} in Glut) (Figure 4C). Taken together, these analyses suggest that SZ PRS explained by *MIR137* target genes is mainly attributed to a small number of neuron-specific genes.

SZ PRS of *MIR137* target genes accounts for a disproportionally larger SZ risk variance of MGS sample

The maximum SZ risk variance explained by PRS of neuronal *MIR137* target gene sets in the MGS sample was ~2%, whereas the whole genome SNP set explains up to ~10% SZ risk variance (Figure 2A). Considering

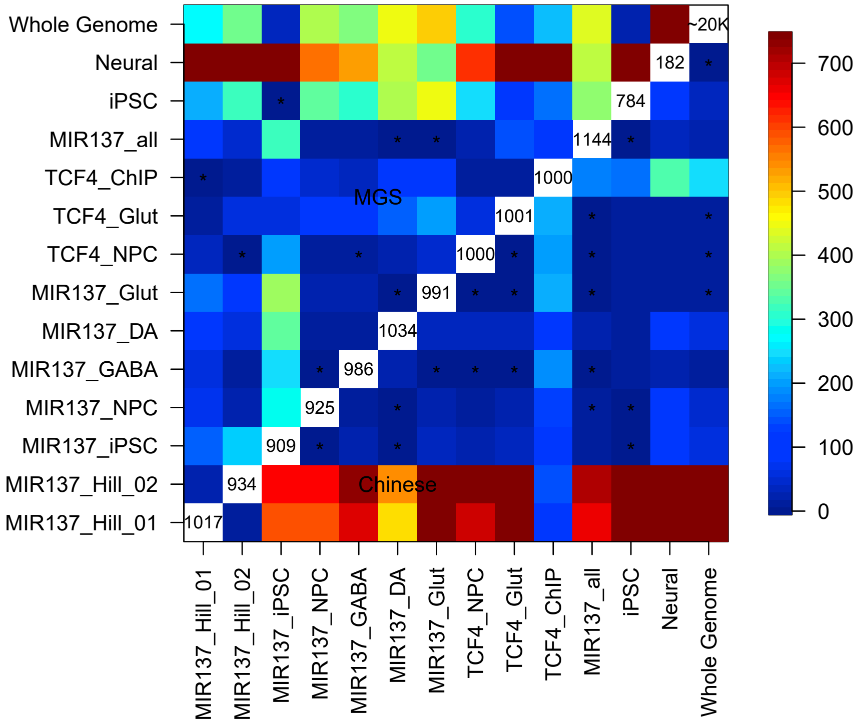


Figure 3. -log(*p*-values) of *t*-tests for comparing the variance explained by the equal number of randomly selected SNPs (1,000 permutations) of two gene sets

The upper triangle indicates the results in the MGS sample, and the lower triangle indicates the results in the Han Chinese sample, both with SNPs within 20-kb flanking regions of a gene. The number of genes of each gene set is shown in the diagonal grid. The cells with * indicate non-significant *p*-value > 0.0009 (from two-sample Student's *t*-test; Bonferroni corrected).

the small number of *MIR137* target genes (N = ~1,000) in each gene set in comparison with the number of all the RefSeq genes (N = ~20,000), the variance explained by neuronal *MIR137* target gene sets suggests a disproportionally larger (4-fold) contribution of neuronal *MIR137* target gene sets to genetic etiology of SZ. This observation is consistent with the ''master regulator'' role of *MIR137* in post-transcriptional regulation. The larger proportion of SZ risk variance explained by neuronal *MIR137* target gene sets was further supported by a comparative PRS analysis of gene targets of *TCF4*, a transcriptional ''master regulator'' (and also a prominent SZ GWAS risk locus): PRS of *MIR137* target gene sets explained higher SZ risk variance in the MGS sample (Figures 2A and S1, and Table S3).

Permutation tests of gene set-specific PRS analysis accounting for SNP number in each gene set also demonstrated an overall better performance of the PRS of neuronal *MIR137* target genes in explaining SZ risk variance (Table S4), compared with the PRS of *TCF4* target gene sets or the genome-wide SNP set (Figures 2B and S1B). Our independent partitioned heritability analysis further showed neuronal *MIR137* target gene sets explained disproportionally more SZ heritability (with PGC3 SZ GWAS summary statistics) than TCF4 target gene sets (Figure 2C and Table S5).

The European ancestry-based PRS of *MIR137* target gene sets explains smaller SZ risk variance of Han Chinese in cell type-specific manner

Although genome-wide PRS is usually considered to not be portable between samples of different ancestry, PRS of a disease-relevant gene pathway/set may perform better in explaining disease risk variance across populations (Baker et al., 2018). Given the strong genetic association of the *MIR137* locus with SZ in both European and Han Chinese samples, as well as our observed large contribution of neuronal *MIR137* target genes to SZ risk variance in the MGS sample (Figure 2), we conducted a PRS analysis in a Han Chinese SZ sample for the same gene lists and by using the same European ancestry-based PGC-SZ3-minus-MGS summary GWAS statistics. As expected, we observed that the PGC-SZ3 genome-wide PRS

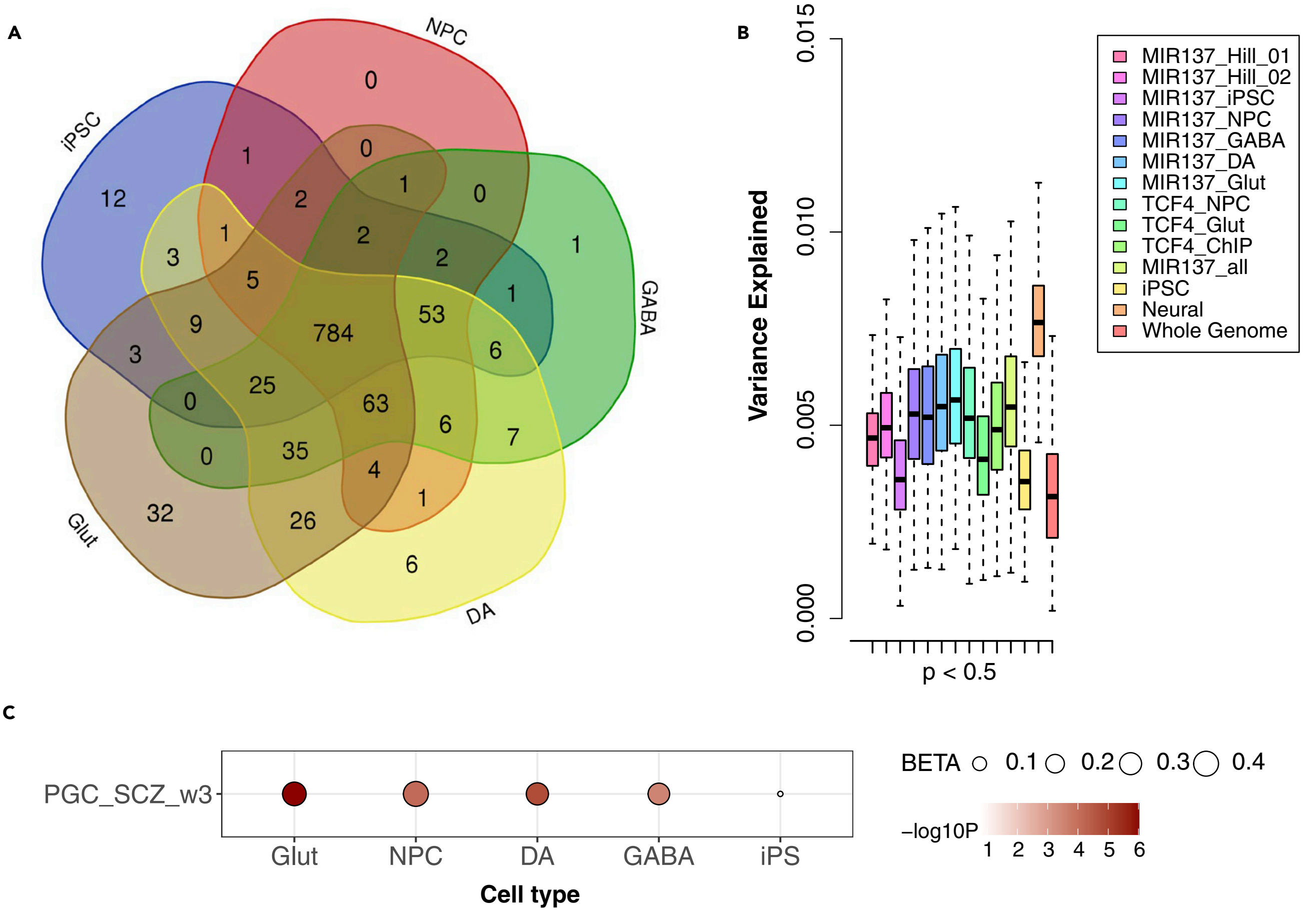


Figure 4. SZ PRS explained by *MIR137* target genes (in Figure 2) is mainly attributed to a small number of neuron-specific genes

(A) Venn diagram of the five lists of *MIR137* target genes expressed in different cell types derived from hiPSC. NPC, neuron progenitor cells; GABA, GABAergic neurons; DA, dopaminergic neurons; Glut, glutamatergic neurons.

(B) The PRS results from permutation (N = 1,000) for the small subset of *MIR137* target genes expressed only in neuronal cell type (N = 182, denoted as "Neural") or shared with hiPSC (N = 784, denoted as "iPSC"), using randomly selected 1,500 SNPs for each gene set at GWAS *p*-value threshold of <0.05. Data are represented as mean ± SEM.

(C) H-MAGMA analysis showing enrichment of SZ risk (PGC_wave3) in *MIR137* target genes specific to each neuronal cell type but not in genes shared with iPSC (n = 784 from A). Shown in bubble plot are corresponding effect size (BETA) and enriched *p*-value (Bonferroni corrected; in -log10 scale) of each gene set.

explained much smaller (~2%) SZ risk variance in the tested Han Chinese sample (Figure 5A, Figure S4A, and Table S3) (versus ~10% in MGS). Similarly, PRS of *MIR137* target genes also explained much less variance of the Han Chinese sample (up to ~0.3% with extended 20 kb; Figure 5A) than of the MGS sample (Figure 2A).

We further examined whether the PRS of neuronal *MIR137* target genes also explain larger SZ risk variance of Han Chinese sample than that of genes with expression altered *in vitro* by *MIR137* or the *TCF4* target gene sets as we observed for MGS sample. For gene sets with the 20-kb extended genic region, although we observed a better predictive performance of PRS of the entire SNP sets of neuronal *MIR137* genes at $p < 0.01$ (Figures 5A and 3), permutation testing accounting for SNP number in each gene list did not provide support (Figures 5B and 3). However, for gene sets with a 100-kb extended genic region at some SZ GWAS *p*-value cut-offs ($p < 0.01$, <0.05, and <0.5) both the PRS analysis of the entire SNP sets and the permutation test supported that the PRS of neuronal *MIR137* target genes overall explain larger SZ risk variance of the Han Chinese sample than that of non-neuronal *MIR137* target genes or genes simply altered by changing *MIR137* expression (Figures S4 and 3). As observed for MGS, the PRS of neuronal *MIR137* target genes with the extended 100-kb genic region overall explained larger SZ risk variance of the Han Chinese sample compared with the PRS of *TCF4* target genes or over genome-wide random

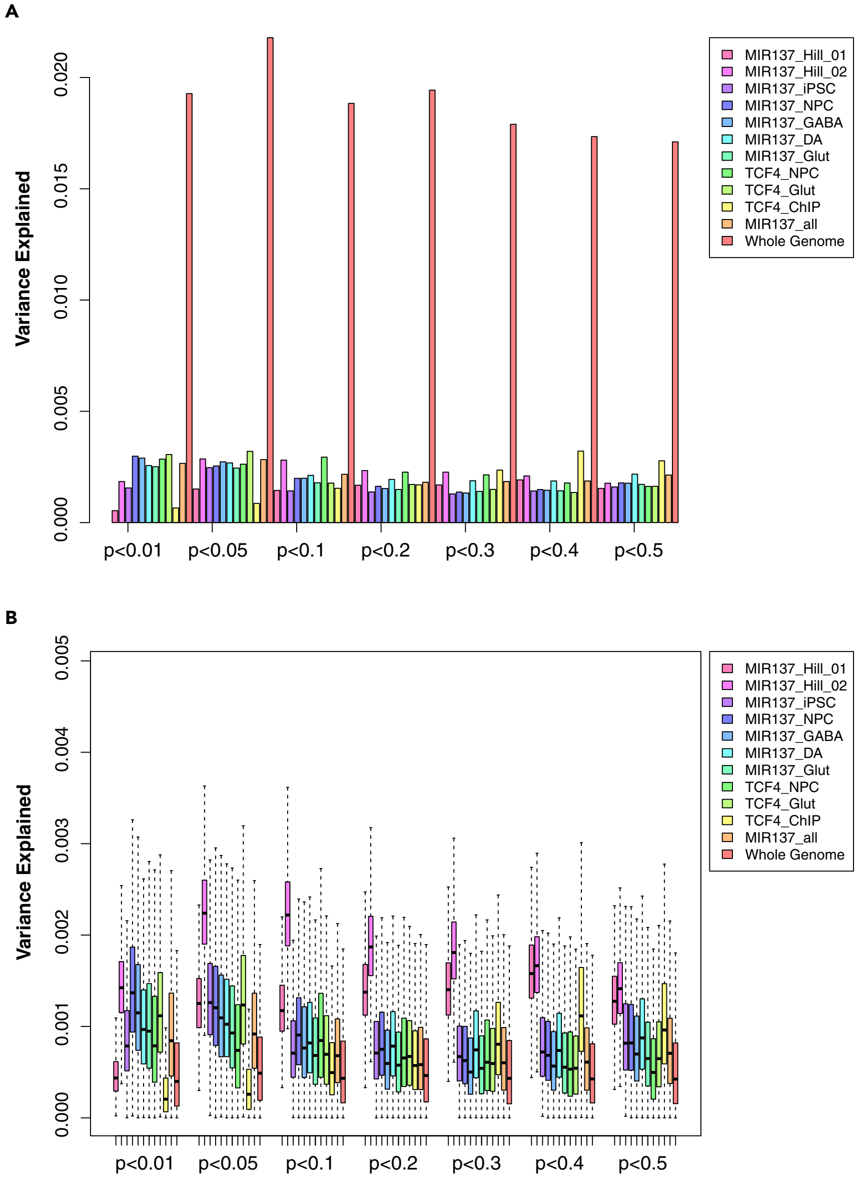


Figure 5. PRS analysis of different SNP lists of *MIR137* and *TCF4* target genes annotated in a 20-kb window for the Han Chinese sample

(A) PRS result using LD-pruned SNPs ($r^2 < 0.2$) for each gene list. The variance explained in the target sample is based on risk scores derived from an aggregated sum of weighted SNP risk allele effect sizes estimated from the discovery samples at seven significance thresholds ($p < 0.01$, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5). The y axis indicates the percentage of phenotypic variance explained by the PRS (Nagelkerke's pseudo R^2).

(B) The permutation (N = 1,000) PRS results based on randomly selected 1,500 SNPs for each gene set. Data are presented as mean ± SEM. Gene sets in (A) and (B) were described in Table 1.

SNP sets (Figures S4B and 3). However, the same pattern was not observed for the gene sets with the 20-kb extended genic region, even with the same number of randomly selected SNPs in each gene set (Figure 5B), which may be due to the different LD structure in the Han Chinese sample that makes the PRS analysis very sensitive to the SNP numbers in each gene set. Altogether, these results suggest that, compared with the MGS sample of European ancestry, there was only a more modest proportional advantage of neuronal *MIR137* gene sets over non-neuronal gene sets or genomic controls in explaining SZ risk variance in Han Chinese.

DISCUSSION

Gene pathway- or set-based PRS analysis in a large cohort can help inform disease biology and identify core gene networks that contribute most risk to a polygenic or omnigenic disorder (Baker et al., 2018). Furthermore, because a disorder-relevant gene pathway may contain SNPs that show more convergent cross-ancestry genetic associations, the pathway-based PRS in a sample of one ancestry may be somewhat more portable for predicting individual risks in a population of another ancestry. Here, for *MIR137*, a leading SZ GWAS risk gene and a post-transcriptional master regulator, our gene pathway-based PRS analysis suggested a cell type-specific effect of PRS in explaining SZ risk variance and a disproportionally large contribution of *MIR137* target gene set to SZ risk in a European ancestry sample and to a lesser degree in a Han Chinese sample.

MIR137 is a leading SZ risk gene that is predicted by TargetScan to post-transcriptionally regulate more than 1,000 genes. However, the relative genetic contribution of the *MIR137* gene pathway to SZ etiology has not been examined. Through cell type-specific PRS analyses, we have found that neuronal *MIR137* target genes can explain ~2% of SZ risk variance in the MGS European ancestry sample (Figure 2), which accounts for one-fifth of the total risk variance explained by the genome-wide SNP set. Despite an overall much smaller proportion (as expected) of SZ risk variance of Han Chinese explained by the European ancestry-based SZ PRS, a comparable proportion of PRS contribution from neuronal *MIR137* target gene pathway was identified for Han Chinese. The cross-population convergent PRS contribution of the neuronal *MIR137* target gene sets is not unexpected, because *MIR137* is also a leading SZ GWAS locus in the Han Chinese sample (Li et al., 2017). The same GWAS SNP at the *MIR137* locus has also been associated with brain imaging phenotypes in the Han Chinese population (Liu et al., 2014). The relatively large contribution of the *MIR137* gene pathway to SZ risk in both populations (Figures 2B, 2C, 5B, and S4B) presents a clear contrast to the target gene sets of *TCF4*, another leading GWAS risk gene and a transcriptional master regulator (Doostparast Torshizi et al., 2019). These results suggest a possibly shared disease pathway involving *MIR137* and its target genes across populations.

Our cell type-specific *MIR137* target gene set PRS contribution to SZ risk also implies the importance of considering the most disease-relevant cell types in PRS analysis. Gene sets with expression altered by either over-expressing or knocking down *MIR137* expression have been used in previous PRS analysis of SZ-relevant cognitive performance, but with mixed outcomes (Cosgrove et al., 2017, 2018). By comparing with the exact set of *MIR137* gene sets used by these studies (Cosgrove et al., 2017, 2018), we have shown here that PRSs of neuronal *MIR137* target gene sets rather than those genes simply altered by manipulating *MIR137* expression better explain SZ risk variance in a target sample (Figure 2). The improved performance of neuronal *MIR137* target gene sets was not unexpected, because many genes showing gene expression changes upon artificially changing *MIR137* expression may just reflect the noise of transcriptomic data, i.e., not specific to *MIR137* function. The importance of cell type specificity in PRS analysis is echoed by a recent study that shows PRSs derived from cell type-specific regulatory elements can substantially increase its trans-ancestry portability (Amariuta et al., 2020). Beyond disorder-relevant cell types, other considerations that reflect the convergent disease biology, such as developmental stage or variant functionality, may further improve the applicability of PRS.

Identifying the core gene networks and the functionally relevant cell types for a polygenic disorder has been a challenge. Leveraging the transcriptomic profiles of hiPSC-derived relatively pure (>80%) neuronal cultures (Forrest et al., 2017; Zhang et al., 2018, 2020), both our cell type-specific PRS analyses of *MIR137* target gene sets with the same number of randomly selected SNPs (Figure 2B) and our independent partitioned heritability analysis for each gene set (Figure 2C) indicated that Glut, DA, and GABA neurons are all relevant cell types for *MIR137* gene pathway to function and contribute SZ risk, which has not been previously demonstrated. Indeed, the implicated specific cell types where the PRS of *MIR137* target gene sets better explain SZ risk variance are consistent with our previous studies that showed high-level expression of *MIR137* in DA and Glut neurons, but not as much in NPCs or iPSCs (Forrest et al., 2017; Shi et al., 2014). Moreover, our observation that the PRS of the Glut neuronal *MIR137* target gene set explained the highest SZ risk variance in MGS (Figure 2) is consistent with our recent report of a negative correlation of *MIR137* expression and its target gene *GRIA1* in dendritic protrusions (or spines) of hiPSC-derived Glut neurons (Forrest et al., 2017). Identifying disease relevant cell type(s) is important for understanding the biology of SZ (Skene et al., 2018). The significance of our current analysis is the identification of SZ-relevant cell types for a specific gene pathway through cell type-specific and cross-population PRS analyses (Figures 2 and 3), providing a more tractable mechanistic insight on understanding the biology of SZ.

Limitations of the study

We acknowledge the limitations in our analyses. First, although the number of genes in each list are comparable (N~1,000), the number of SNPs (and gene lengths) in each cell type or population may vary, which may confound the SZ risk variance explained by a specific gene/SNP set. Nevertheless, our permutation tests using the same number SNPs randomly selected from each gene list and our independent partitioned heritability analysis (Figures 2B, 2C, S1B, and 3B), which was expected to account for the varying SNP number and gene length in each gene list, revealed consistent results with the observation directly from all SNPs in each gene set. Second, some other cell types such as astrocytes and microglia might also be relevant to *MIR137* function, which was not investigated in our study; however, compared with glutamatergic and GABAergic neurons, these cell types are less genetically relevant to SZ (Skene et al., 2018). Third, iPSC-derived neurons, although relatively pure, are commonly known to represent immature neurons (during the prenatal stages), and our observations may thus not reflect the disorder biology at the adult stage of SZ. Nevertheless, SZ is known to have a strong neurodevelopmental pathogenic aspect and hiPSC models have been effectively used for understanding SZ-relevant functional genomics and disease biology (Amiri et al., 2018; Forrest et al., 2017; Rajarajan et al., 2018). Finally, although *MIR137* post-transcriptionally functions through targeting 3′-UTRs of its target genes, our gene set PRS analysis was conducted with genetic risk variants not only at 3′-UTRs of these *MIR137* target genes but also in other parts of the genes such as enhancers. More comprehensive PRS analysis or partitioned heritability analysis of the *MIR137* gene pathway in different cell types at various developmental stages, and by variant location in different genomic/epigenomic features, may shed further mechanistic insight onto how *MIR137* confers individual susceptibility to SZ.

In summary, by combining pathway-based PRS analysis with cell type-specific transcriptomic profiles in hiPSC-derived neurons, we discovered that *MIR137* target genes expressed in three major SZ-relevant neuronal cell types explain a larger proportion of SZ risk variance in European and, to a lesser degree, Han Chinese samples. Despite the limitations, our analysis informed the SZ-relevant cell types for *MIR137* to function, highlighting the value of pathway-based PRS analysis in uncovering disorder-relevant biological features. Knowledge on cross-population high-risk gene pathways may not only help improve the specificity of PRS-based individual risk prediction but also have the potential to facilitate the translation of genetic findings into pathway-based therapeutic interventions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Genetic data sets
  - PRS analysis
  - Gene-sets
  - Permutation to account for variable SNP number
  - Partitioned heritability analysis by stratified LDSC
  - Gene-set GWAS risk enrichment analysis by H-MAGMA
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102785.

ACKNOWLEDGMENTS

MGS includes P. V. Gejman, A. R. Sanders, J. Duan (NorthShore University HealthSystem and University of Chicago, IL, USA); D. F. Levinson (Stanford University, CA, USA); J. Shi (National Cancer Institute, MD, USA); N. G. Buccola (Louisiana State University Health Sciences Center, LA, USA); B. J. Mowry (Queensland Center for Mental Health Research, Brisbane and Queensland Brain Institute, The University of Queensland, Australia); R. Freedman, A. Olincy (University of Colorado Denver, CO, USA); F. Amin (Atlanta Veterans

Affairs Medical Center and Emory University, GA, USA); D. W. Black (University of Iowa Carver College of Medicine, IA, USA); J. M. Silverman (Mount Sinai School of Medicine, NY, USA); W. F. Byerley (University of California at San Francisco, CA, USA); and C. R. Cloninger, D. M. Svrakic (Washington University, MO, USA). We thank the study participants of MGS. MGS was mainly supported by R01MH059571, R01MH081800, and U01MH079469 (to P.V. Gejman.); and other NIH grants for other MGS sites (R01MH067257 to N. G. Buccola, R01MH059588 to B. J. Mowry, R01MH059565 to R. Freedman, R01MH059587 to F. Amin, R01MH060870 to W. F. Byerley, R01MH059566 to D. W. Black, R01MH059586 to J. M. Silverman., R01MH061675 to D. F. Levinson, R01MH060879 to C. R. Cloninger, U01MH046276 to C. R. Cloninger, and U01MH079470 to D. F. Levinson). We thank Dr. P. V. Gejman for helping with the interpretation of the MGS dataset. We also thank the PGC-SZ group for providing early access of wave-3 (PGC-SZ3) summary statistics. This study was supported by Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01) and by Department of Computational Biology and Institute of Life Sciences and Human Phenomics Institute (to Y.Y.), and NIH grants R21MH102685, R01MH106575, R01MH116281, and R01AG063175 (to J.D.).

AUTHOR CONTRIBUTIONS

W.G. performed the PRS analysis and wrote the manuscript; S.Z. prepared most genomics datasets and performed partitioned heritability and gene set enrichment analyses; H. Yu and H. Yan analyzed the Han Chinese GWAS dataset; H.Z. produced the iPSC-derived neuronal expression data; A.R.S. helped with the interpretation of MGS dataset and wrote the manuscript; W.Y. supervised the analysis of the Han Chinese GWAS and wrote the manuscript; Y.Y. supervised the overall PRS analyses and wrote the manuscript; J.D. conceived the study, coordinated and supervised the analyses, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

Received: January 20, 2021
Revised: May 18, 2021
Accepted: June 23, 2021
Published: July 23, 2021

REFERENCES

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K.K., Matsuda, K., Murakami, Y., Price, A.L., Kawakami, E., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat. Genet. *52*, 1346–1354.

Amiri, A., Coppola, G., Scuderi, S., Wu, F., Roychowdhury, T., Liu, F., Pochareddy, S., Shin, Y., Safi, A., Song, L., et al. (2018). Transcriptome and epigenome landscape of human cortical development modeled in organoids. Science *362*, eaat6720.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Baker, E., Schmidt, K.M., Sims, R., O'Donovan, M.C., Williams, J., Holmans, P., Escott-Price, V., and Consortium, W.T.G. (2018). POLARIS: polygenic LD-adjusted risk score approach for set-based analysis of GWAS data. Genet. Epidemiol. *42*, 366–377.

Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of complex traits: from polygenic to omnigenic. Cell *169*, 1177–1186.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., and Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Cheng, Y., Wang, Z.-M., Tan, W., Wang, X., Li, Y., Bai, B., Li, Y., Zhang, S.-F., Yan, H.-L., Chen, Z.-L., et al. (2018). Partial loss of psychiatric risk gene Mir137 in mice causes repetitive behavior and impairs sociability and learning via increased Pde10a. Nat. Neurosci. *21*, 1689–1703.

Collins, A.L., Kim, Y., Bloom, R.J., Kelada, S.N., Sethupathy, P., and Sullivan, P.F. (2014). Transcriptional targets of the schizophrenia risk gene MIR137. Transl. Psychiatry *4*, e404.

Schizophrenia Psychiatric Genome-Wide Association Study (GWAS)Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

The Schizophrenia Working Group of the Psychiatric Genomics Consortium, Ripke, S., Walters, J.T.R., and O'Donovan, M.C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. MedRxiv. https://doi.org/10.1101/2020.09.12.20192922.

Cosgrove, D., Harold, D., Mothersill, O., Anney, R., Hill, M.J., Bray, N.J., Blokland, G., Petryshen, T., Richards, A., Mantripragada, K., et al. (2017). MiR-137-derived polygenic risk: effects on cognitive performance in patients with schizophrenia and controls. Transl. Psychiatry *7*, e1012. https://doi.org/10.1038/tp.2016.286.

Cosgrove, D., Mothersill, D.O., Whitton, L., Harold, D., Kelly, S., Holleran, L., Holland, J., Anney, R., Richards, A., Mantripragada, K., et al. (2018). Effects of MiR-137 genetic risk score on brain volume and cortical measures in patients with schizophrenia and controls. Am. J. Med. Genet. B Neuropsychiatr. Genet. *177*, 369–376.

Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat. Methods *9*, 179–181.

Doostparast Torshizi, A., Armoskus, C., Zhang, H., Forrest, M.P., Zhang, S., Souaiaia, T., Evgrafov, O.V., Knowles, J.A., Duan, J., and Wang, K. (2019). Deconvolution of transcriptional networks identifies TCF4 as a master regulator in schizophrenia. Sci. Adv. *5*, eaau4139.

Euesden, J., Lewis, C.M., and O'Reilly, P.F. (2015). PRSice: polygenic risk score software. Bioinformatics *31*, 1466–1468.

Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.

Forrest, M.P., Zhang, H., Moy, W., McGowan, H., Leites, C., Dionisio, L.E., Xu, Z., Shi, J., Sanders, A.R., Greenleaf, W.J., et al. (2017). Open chromatin profiling in hiPSC-derived neurons prioritizes functional noncoding psychiatric risk variants and highlights neurodevelopmental loci. Cell Stem Cell *21*, 305–318.e8.

Green, M.J., Cairns, M.J., Wu, J., Dragovic, M., Jablensky, A., Tooney, P.A., Scott, R.J., and Carr, V.J. (2012). Genome-wide supported variant MIR137 and severe negative symptoms predict membership of an impaired cognitive subtype of schizophrenia. Mol. Psychiatry *18*, 774–780.

Guarnieri, D.J., and DiLeone, R.J. (2008). MicroRNAs: a new class of gene regulators. Ann. Med. *40*, 197–208.

Hill, M.J., Donocik, J.G., Nuamah, R.A., Mein, C.A., Sainz-Fuertes, R., and Bray, N.J. (2014). Transcriptional consequences of schizophrenia candidate miR-137 manipulation in human neural progenitor cells. Schizophr. Res. *153*, 225–230.

Howie, B.N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. *5*, e1000529.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. G3 (Bethesda) *1*, 457–470.

Im, H.I., and Kenny, P.J. (2012). MicroRNAs in neuronal function and dysfunction. Trends Neurosci. *35*, 325–334.

Kwon, E., Wang, W., and Tsai, L.H. (2013). Validation of schizophrenia-associated genes CSMD1, C10orf26, CACNA1C and TCF4 as miR-137 targets. Mol. Psychiatry *18*, 11–12.

Lee, S.H., DeCandia, T.R., Ripke, S., Yang, J., Sullivan, P.F., Goddard, M.E., Keller, M.C., Visscher, P.M., and Wray, N.R. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat. Genet. *44*, 247–250.

de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219.

Li, Z., Chen, J., Yu, H., He, L., Xu, Y., Zhang, D., Yi, Q., Li, C., Li, X., Shen, J., et al. (2017). Genome-wide association analysis identifies 30 new susceptibility loci for schizophrenia. Nat. Genet. *49*, 1576–1583.

Liu, B., Zhang, X., Hou, B., Li, J., Qiu, C., Qin, W., Yu, C., and Jiang, T. (2014). The impact of MIR137 on dorsolateral prefrontal-hippocampal functional connectivity in healthy subjects. Neuropsychopharmacology *39*, 2153–2160.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E.A., Kassim, B., et al. (2018). Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science *362*, eaat4311.

Remmers, C., Sweet, R.A., and Penzes, P. (2014). Abnormal kalirin signaling in neuropsychiatric disorders. Brain Res. Bull. *103C*, 29–38.

Ripke and Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

Sey, N.Y.A., Hu, B., Mah, W., Fauni, H., McAfee, J.C., Rajarajan, P., Brennand, K.J., Akbarian, S., and Won, H. (2020). A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. Nat. Neurosci. *23*, 583–593.

Shi, J., Levinson, D.F., Duan, J., Sanders, A.R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P.A., Whittemore, A.S., Mowry, B.J., et al. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature *460*, 753–757.

Shi, S., Leites, C., He, D., Schwartz, D., Moy, W., Shi, J., and Duan, J. (2014). MicroRNA-9 and microRNA-326 regulate human dopamine D2 receptor expression and the microRNA-mediated expression regulation is altered by a genetic variant. J. Biol. Chem. *289*, 13434–13444.

Silber, J., Lim, D.A., Petritsch, C., Persson, A.I., Maunakea, A.K., Yu, M., Vandenberg, S.R., Ginzinger, D.G., James, C.D., Costello, J.F., et al. (2008). miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. BMC Med. *6*, 14.

Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D., Miller, J.A., Munoz-Manchado, A.B., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nat. Genet. *50*, 825–833.

Smrt, R.D., Szulwach, K.E., Pfeiffer, R.L., Li, X., Guo, W., Pathania, M., Teng, Z.Q., Luo, Y., Peng, J., Bordey, A., et al. (2010). MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase mind bomb-1. Stem Cells *28*, 1060–1070.

Stefansson, H., Ophoff, R.A., Steinberg, S., Andreassen, O.A., Cichon, S., Rujescu, D., Werge, T., Pietilainen, O.P., Mors, O., Mortensen, P.B., et al. (2009). Common variants conferring risk of schizophrenia. Nature *460*, 744–747.

Sun, G., Ye, P., Murai, K., Lang, M.F., Li, S., Zhang, H., Li, W., Fu, C., Yin, J., Wang, A., et al. (2011). miR-137 forms a regulatory loop with nuclear receptor TLX and LSD1 in neural stem cells. Nat. Commun. *2*, 529.

Szulwach, K.E., Li, X., Smrt, R.D., Li, Y., Luo, Y., Lin, L., Santistevan, N.J., Li, W., Zhao, X., and Jin, P. (2010). Cross talk between microRNA and epigenetic regulation in adult neurogenesis. J. Cell Biol. *189*, 127–141.

Topol, A., Zhu, S., Hartley, B.J., English, J., Hauberg, M.E., Tran, N., Rittenhouse, C.A., Simone, A., Ruderfer, D.M., Johnson, J., et al. (2016). Dysregulation of miRNA-9 in a subset of schizophrenia patient-derived neural progenitor cells. Cell Rep. *15*, 1024–1036.

Ursini, G., Punzi, G., Chen, Q., Marenco, S., Robinson, J.F., Porcelli, A., Hamilton, E.G., Mitjans, M., Maddalena, G., Begemann, M., et al. (2018). Convergence of placenta biology and genetic risk for schizophrenia. Nat. Med. *24*, 792–801.

Vassos, E., Kou, J., Tosato, S., Maxwell, J., Dennison, C.A., Legge, S.E., Walters, J.T.R., Owen, M.J., O'Donovan, M.C., Breen, G., et al. (2021). Lack of support for the genes by early environment interaction hypothesis in the pathogenesis of schizophrenia. Schizophr. Bull. sbab052. https://doi.org/10.1093/schbul/sbab052.

Volvert, M.L., Rogister, F., Moonen, G., Malgrange, B., and Nguyen, L. (2012). MicroRNAs tune cerebral cortical neurogenesis. Cell Death Differ. *19*, 1573–1581.

Willemsen, M.H., Valles, A., Kirkels, L.A., Mastebroek, M., Olde Loohuis, N., Kos, A., Wissink-Lindhout, W.M., de Brouwer, A.P., Nillesen, W.M., Pfundt, R., et al. (2011). Chromosome 1p21.3 microdeletions comprising DPYD and MIR137 are associated with intellectual disability. J. Med. Genet. *48*, 810–818.

Wright, C., Turner, J.A., Calhoun, V.D., and Perrone-Bizzozero, N. (2013). Potential impact of miR-137 and its targets in schizophrenia. Front. Genet. *4*, 58.

Yu, H., Yan, H., Li, J., Li, Z., Zhang, X., Ma, Y., Mei, L., Liu, C., Cai, L., Wang, Q., et al. (2017). Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. Mol. Psychiatry *22*, 954–960.

Zeng, Y., Navarro, P., Fernandez-Pujals, A.M., Hall, L.S., Clarke, T.K., Thomson, P.A., Smith, B.H., Hocking, L.J., Padmanabhan, S., Hayward, C., et al. (2017). A combined pathway and regional heritability analysis indicates NETRIN1 pathway is associated with major depressive disorder. Biol. Psychiatry *81*, 336–346.

Zhang, S., Moy, W., Zhang, H., Leites, C., McGowan, H., Shi, J., Sanders, A.R., Pang, Z.P., Gejman, P.V., and Duan, J. (2018). Open chromatin dynamics reveals stage-specific transcriptional networks in hiPSC-based neurodevelopmental model. Stem Cell Res. *29*, 88–98.

Zhang, S., Zhang, H., Zhou, Y., Qiao, M., Zhao, S., Kozlova, A., Shi, J., Sanders, A.R., Wang, G., Luo, K., et al. (2020). Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants. Science *369*, 561–565.

Zylka, M.J., Simon, J.M., and Philpot, B.D. (2015). Gene length matters in neurons. Neuron *86*, 353–355.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
PGC3 SZ GWAS summary statistics	The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020	www.med.unc.edu/pgc/pgc-workgroups/schizophrenia
Chinese Han SZ GWAS summary statistics	Li et al., 2017; Yu et al., 2017	Accession number 10638 https://doi.org/10.18170/DVN/IHROUE
RNA-seq data from cell lines of 8 donors per cell type (iPSC, NPC, DA, and GABA)	Zhang et al., 2020; Forrest et al., 2017; Zhang et al., 2018	GSE129017 and GSE132757
Software and algorithms		
IMPUTE2 software program (version 2.1.2)	Howie et al., 2011; Howie et al., 2009	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
SHAPEIT2 (version v2.r644)	Delaneau et al., 2011	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html
LDSC	Bulik-Sullivan et al., 2015; Finucane et al., 2015	https://github.com/bulik/ldsc
H-MAGMA	Sey et al., 2020	https://github.com/thewonlab/H-MAGMA
Other		
1,000 Genomes Phase 3 panel (October 2014 Data Release)	Auton et al., 2015	https://www.internationalgenome.org/category/phase-3/
1000G_EUR_Phase3_baseline v2.2	Bulik-Sullivan et al., 2015; Finucane et al., 2015	https://github.com/bulik/ldsc/issues/96

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jubao Duan (jduan@uchicago.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The PGC3 SZ GWAS summary statistics can be accessed through ''schizophrenia data access portal'' (https://www.med.unc.edu/pgc/pgc-workgroups/schizophrenia/). The Chinese Han SZ GWAS summary statistics have been deposited in the Peking University Open Research Data repository with accession number 10638 (https://doi.org/10.18170/DVN/IHROUE). The gene expression data used to generate the list of *MIR137*-target genes expressed in different neural cell types can be accessed through GEO: GSE129017 and GSE132757. This paper does not report original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Genetic data sets

PGC-SZ3. The PGC-SZ3 (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al., 2020) dataset was downloaded through approved access from PGC SZ workgroup (www.med.unc.edu/pgc/pgc-workgroups/schizophrenia). This specific dataset is the summary statistics of PGC-SZ samples of European ancestry leaving out MGS. This PGC-SZ3 dataset contains the odds ratios (ORs) and p values that were calculated on 7,651,682 SNPs in the meta-analysis of all 75 European cohorts (except the MGS data), and was used as a discovery sample to identify SZ risk variants.

MGS. MGS was used as a target sample of European ancestry. This MGS dataset contains GWAS data for 2,681 SZ cases and 2,653 controls (3,134 males, 2,200 females; mean age = 46.8 with 14.7 SD) (Shi et al., 2009). Quality control (QC) performed on the genotype data was previously reported (Shi et al., 2009). Genotype imputation was conducted using the IMPUTE2 software program (version 2.1.2) (Howie et al., 2009, 2011), using haplotypes from all 2,504 individuals in the 1,000 Genomes Phase 3 panel (October 2014 Data Release) as a reference panel (Auton et al., 2015). The haplotypes were phased using SHAPEIT2 (version v2.r644) (Delaneau et al., 2011), which can perform the pre-phasing step for the study genotypes to produce 'best-guess' haplotypes. The imputed SNPs with low quality were excluded (IMPUTE2 info <0.6, IMPUTE2 certainty <0.8, or minor allele frequency (MAF) < 0.05). A total of 5,132,475 imputed SNPs was retained in the follow-up PRS analyses. The NorthShore University HealthSystem IRB approved the study.

Han Chinese. The Han Chinese dataset contains 4,288 SZ cases and 5,512 controls (4,395 males, 5,405 females; mean age = 31 with 10.03 SD) (Li et al., 2017; Yu et al., 2017). Genotyping of these samples was conducted using different types of Illumina Genome-Wide Arrays: GWAS1 dataset using Human-Hap610-Quad BeadChips, GWAS2 dataset using Human660W-Quad BeadChip, and GWAS3 and GWAS4 datasets using Illumina Human OmniZhongHua BeadChips specifically designed for the Chinese population. The genotyping QC filters included: genotype call rate >98%, MAF >0.01, difference in SNP missingness between cases and controls <0.02, Hardy-Weinberg disequilibrium $p > 1 \times 10^{-6}$. Samples were excluded for gender discordance and genetic ancestry outliers. Genotypes were first phased using SHAPEIT (v2.r727) (Delaneau et al., 2011), and imputation was then performed over each 3 Mb-interval centered on all index SNPs using IMPUTE (v2.3.0) software (Howie et al., 2009, 2011). Haplotypes derived from Phase 1 of the 1000 Genomes Project (release v3) were used as reference data. After imputation, we excluded SNPs with an imputation quality score below a set threshold (info<0.6), call rate <0.98 in either cases or controls, MAF<0.01, Hardy-Weinberg disequilibrium $p < 1 \times 10^{-6}$.

PRS analysis

To explore the genetic relationship between the discovery sample (PGC-SZ3 without MGS) and target samples (MGS, Han Chinese) datasets, we used PRSice to conduct a standard PRS analysis (Euesden et al., 2015) as in the PRS workflow (Figure 1). Let P_d be the p values of the GWAS study in the discovery sample. Specifically, 'significant' SNPs with P_d<0.5 were pruned based on linkage disequilibrium (LD) ($r^2 < 0.2$) in MGS data, and then selected based on seven predetermined significance thresholds (P_d<0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5) in PGC-SCZ3. Within each pruned SNP set under each significance threshold, a quantitative aggregate risk score was calculated for each individual in the target samples, defined as a sum across SNPs of the number of reference alleles (0, 1, or 2) at that SNP multiplied by the effect size measures (log of OR) for that SNP estimated from the discovery sample. Association of aggregate risk score and actual SZ in the target samples were performed with logistic regression adjusted for gender, age, and 10 principal components to control for population stratification. We then calculated the percentage of phenotypic variance explained by the aggregate risk score (Nagelkerke's pseudo R^2) of different SNP/gene-sets (see below).

Gene-sets

We have used the same lists of 900–1,000 genes previously used for *MIR137* PRS analysis (Cosgrove et al., 2017; Hill et al., 2014) for comparison to our own gene sets. These lists of genes are those with expression altered by overexpressing *MIR137* or knocking down *MIR137* in a single NPC cell line, selected by the fold-change of gene expression and their statistical significance (Cosgrove et al., 2017; Hill et al., 2014).

For testing lists of cell type-specific *MIR137* target genes (900–1,000 genes per list), we assembled them by cross-referencing the predicted *MIR137* target genes (target prediction score >0.1 by TargetScan 6.2) with their expression value FPKM (fragments per kilobase per million reads) in hiPSC and hiPSC-derived NPCs, dopaminergic neurons (DA), GABAergic neurons (GABA), and glutamatergic neurons (Glut) (Forrest et al., 2017; Zhang et al., 2018, 2020). Briefly, RNA-seq data (GENCODE v28) from cell lines of 8 donors (genetic backgrounds) per cell type (iPSC, NPC, DA, and GABA) (Zhang et al., 2020) were used to estimate the average gene expression level. For Glut neurons, we used NGN2-induced excitatory neurons at day-27 post neuronal induction, and expression values were calculated as the average of all available samples (Forrest et al., 2017; Zhang et al., 2018, 2020). An expressed gene in each cell type was defined using a cut-off value of log2 FPKM >0.5. Subsequently, the lists of expressed genes were cross-referenced with predicted target gene list of *MIR137* in each cell type.

For targets of *TCF4*, we have compiled 3 lists of ~1,000 genes each. These include genes with expression either up- or down-regulated by knockdown *TCF4* expression in NPCs and in early-stage Glut neurons (Doostparast Torshizi et al., 2019). For genes in both cell types, we only selected the top 1,000 genes showing differential expression upon *TCF4* knockdown. The direct binding targets of *TCF4* were extracted from an empirical chromatin immune-precipitation (ChIP) study, and the top ranking 1,000 gene were selected according to their promoter and intergenic ChIP peak significance (Forrest et al., 2017).

Permutation to account for variable SNP number

The number of genes varies from 909 to 1,017 among 10 gene sets from both *MIR137* and *TCF4* (Table 1 and Table S1). Furthermore, the number of SNPs for each gene set varies in both MGS data and Chinese data before/after LD pruning (Table S2). Therefore, to adjust for different SNP number in each gene set, PRS were evaluated based on the same number of SNPs that were randomly selected within each gene set in 1,000 permutations. A total of 3,000 and 1,500 SNPs (P_d<0.5) was randomly selected in the permutation for both the MGS and the Han Chinese datasets annotated within 20kb and 100kb windows (i.e., extended genic regions), respectively. The detailed procedures were:

1) Obtain GWAS summary statistics (P_d values and ORs) in the discovery sample (PGC-SZ3 without MGS).
2) Cross-reference with target sample data with both genetic and phenotype data and identify SNPs in common between the discovery and target samples.
3) Account for association redundancy due to LD by SNP pruning (LD-r^2 = 0.2 as a cut-off) among all overlapped SNPs with P_d<0.5.
4) Extract SNPs located within a specific gene-set, and randomly select N_{random} SNPs, where N_{random} was selected based on the number of SNPs after LD pruning (Table S2), and N_{random} = 1500/3000 for data within 20kb and 100kb window, respectively.
5) Restrict to SNPs with various thresholds (P_d<0.01, 0.05, 0.1, 0.2, 0.3, 0.4, and 0.5).
6) Construct PRS = sum of risk alleles weighted by log (OR) from regression.
7) Regress trait in target sample onto PRS. Evaluate strength of this association adjusted by age, sex, and 10 principal components of ancestry (Nagelkerke's pseudo R^2).
8) Repeat steps 2-8 for 1,000 permutations.

These procedures were repeated for all the gene-sets including the whole-genome SNP set for both MGS and Han Chinese samples as target samples. Two-sample Student's *t*-tests were further performed on 1,000 permutated Nagelkerke's pseudo R^2 values to test for the mean difference between any two gene-sets.

Partitioned heritability analysis by stratified LDSC

To independently validate the gene-set specific PRS analysis and account for potential confounding factors (e.g., numbers of SNPs) in PRS analysis, we performed a sLDSC analysis (Bulik-Sullivan et al., 2015; Finucane et al., 2015) for different gene sets using PGC SCZ wave_3 GWAS summary statistics. For calculating partitioned LD scores, a 20 kb window size (similar to what was used in Figures 2A and 2B) was applied to all genes within the ten gene sets. For partitioned heritability (h^2), we used baseline model LD scores from the 1,000 Genomes project European cohort (1000G_EUR_Phase3_baseline v2.2). Finally, we used the cell-type group analysis algorithm as described in the LDSC documentation (Bulik-Sullivan et al., 2015; Finucane et al., 2015) to estimate the proportion of heritability and identify the enrichment of h^2 in each gene set.

Gene-set GWAS risk enrichment analysis by H-MAGMA

To account for gene sizes and overlapping genetic variants in different gene sets as well as other potential cofounding factors, we performed H-MAGMA analysis using MAGMA version 1.08 (Sey et al., 2020) to evaluate the enrichment for SZ GWAS risk. Specifically, we first started by generating the MAGMA-required gene analysis data files using pre-compiled gene-SNP annotation files. Compared to the original MAGMA analysis, H-MAGMA improved the prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles (de Leeuw et al., 2015; Sey et al., 2020). Since we defined neural gene sets using iPSC-derived neurons, we used the gene-SNP annotation file pre-compiled with Hi-C data from iPSC-derived neurons (Rajarajan et al., 2018). With the gene-SNP annotation file, we then performed gene-level analysis on SNP *p*-values using the reference SNP data of 1,000 Genomes European panel (g1000_eur, –bfile) and the pre-computed SNP *p*-values from the PGC SZ GWAS wave 3 dataset (daner_PGC_SCZ_w3, –pval). The sample size (ncol =) was directly taken from the column of sample sizes per SNP column of the PGC3 SZ dataset. Subsequently, the result files (–gene-results) from the gene-level analysis were read in for competitive gene-set analysis (–set-annot), where we used default setting ('correct = all') to control for gene sizes in number of SNPs and the gene density (a measure of within-gene LD). The gene-set analysis produced output file (gsa.out) with competitive gene-set analysis results that contained the effect size (BETA) and the statistical significance of the enrichment of each gene set for SZ GWAS risk.

QUANTIFICATION AND STATISTICAL ANALYSIS

To quantify the SZ risk variance explained, we used PRSice to conduct a standard PRS analysis (Euesden et al., 2015). A quantitative aggregate risk score was calculated for each individual in the target samples. The percentage of phenotypic variance explained by the PRS (Nagelkerke's pseudo R^2) was presented in Figures 2, 3, 4, and 5. To compare the SZ risk variance explained by different gene sets, we accounted for SNP number used in each gene set by permutation test (N = 1,000) using the randomly selected 1,500 SNPs for each gene-set. Data are presented as mean ± SEM in Figures 2 and 5. To further determine whether the difference of the SZ risk variance explained by the equal number of randomly selected SNPs (1,000 permutations) of two gene-sets are statistically significant, we used two-sample Student's *t*-test and corrected *p*-values for multiple testing by Bonferroni method (Figure 3). To quantify the enrichment of SZ GWAS SNP heritability (h^2) for each gene set, we used sLDSC analysis (Bulik-Sullivan et al., 2015; Finucane et al., 2015) to determine the folds of enrichment of SNP heritability (proportion of explained heritability h^2 normalized by the proportion of SNPs in each gene set) and derive the statistical significance (Bonferroni-corrected *p*-values) in Figure 2.